

## DESIGN OF STUDENT GRADUATION PREDICTION SYSTEM USING NAIVE BAYES AND WEBSITE-BASED DECISION TREE

Muhammad Isnaini Syaifudin<sup>1</sup>, Bagus Mulyawan<sup>2</sup>, Novario Jaya Perdana<sup>3</sup>

<sup>1</sup> Department of Computer Science, , Tarumanagara University, Jakarta, Indonesia  
Email: muhammad.535200070@stu.untar.ac.id

<sup>2</sup> Department of Computer Science, , Tarumanagara University, Jakarta, Indonesia  
Email: bagus@fti.untar.ac.id

<sup>3</sup> Department of Computer Science, , Tarumanagara University, Jakarta, Indonesia  
Email: novariojp@fti.untar.ac.id

Submitted: 27-09-2023, Revised: 27-10-2023, Accepted: 11-12-2023

---

### ABSTRACT

*In college, the aspect that affects quality is the length of time students study. Therefore, a system is needed to predict student graduation whether they graduate at the specified time or late. This research aims to design and implement a website-based Student Graduation Prediction System Using Naive Bayes and Decision Tree Methods. The data used to predict graduation in the form of assessments obtained in the form of Assignment Scores, Midterm Exam Scores, Final Exam Scores, Total Scores, and Gender. In this research, the Naive Bayes method is used to calculate the probability of student graduation based on the attributes in the dataset. In addition, the Decision Tree algorithm, specifically the C4.5 algorithm, is applied to build a decision tree model that can efficiently predict student graduation. This system is web-based using PHP programming language, the database used is MySQL. System evaluation is done through accuracy, precision, and recall measurements, which provide an overview of the extent to which the system can predict graduation correctly. In addition, the prediction results of student graduation consist of two, namely On Time and Late. The results of this study indicate that the Student Graduation Prediction System using data on Gender, Assignment Scores, Midterm Exam Scores, Final Exam Scores, and Total Scores is able to provide a fairly accurate prediction of student graduation.*

**Keywords:** Prediction, Naïve Bayes, C4.5, Web-Based System, PHP

### 1. INTRODUCTION

Student graduation is a key aspect that all higher education institutions must pay attention to because graduation represents the outcome of the educational process. In the context of education, the meaning of graduation is highly relevant to the understanding of a student's success in navigating the educational process. This success involves passing through various educational procedures, including evaluations or exams for each course, including the thesis course, which is a mandatory course for almost all undergraduate programs (Bachelor's degree) [1]. Technology can be employed to predict various cases, utilizing different calculation methods. The Naive Bayes algorithm is a classification method based on probability and statistics. It predicts a set of possibilities in the future by summing up variables, frequencies, and combinations of the dataset used [2]. On the other hand, the Decision Tree method is a technique for finding a set of patterns or functions that describe and separate data classes from one another, indicating which category an object belongs to based on the behavior and attributes of the defined groups. It offers flexibility, enhancing decision quality, and simplifying complex situations [3].

## 1.1 Problem Formulation

The website design aims to create a system capable of predicting student graduation using data obtained from assessments during the study period of Computer Science students at Tarumanagara University, class of 2018-2019. This prediction is made using the Naive Bayes and Decision Tree methods with the C4.5 Algorithm. The system employs attributes such as Gender, Assignment Scores, Midterm Exam Scores, Final Exam Scores, and Total Scores to determine whether a student will graduate on time or be delayed. The system will provide a prediction result indicating whether the graduation will be on time or delayed.

## 2. RESEARCH METHOD

In this research, two classification methods were used, namely Naive Bayes and Decision Tree using the C4.5 algorithm. Then an experiment was carried out comparing the two methods to predict student graduation.

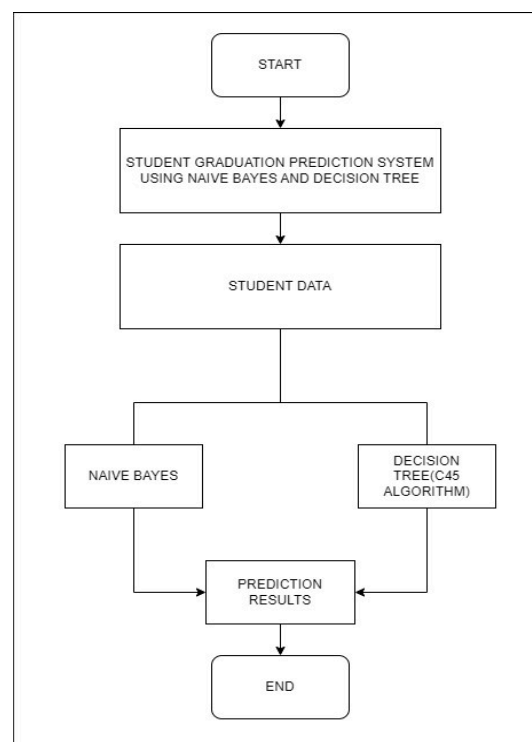


Figure 1 Research Design

Based on the research design in Figure 1, prediction of student graduation carried out by collecting student data, Then the data is processed separately using the Naive Bayes method and the decision tree method using the C4.5 algorithm, then the results between the two methods will be compared.

## 2.1 Method of collecting data

Data collection methods are used to obtain information about what needs to be done during system development. At this data collection stage, several steps are carried out, including:

### a. Literature study

Literature study is a method of collecting data by searching for information in books, magazines and other works to serve as a theoretical basis. In this research, a search was carried out for similar systems that already existed before and were used as a reference in developing the system to be created.

### b. Observation

Observation is an activity carried out by observing the object being researched. This observation is carried out by observing data on Gender, Assignment Grades, UTS Grades, UAS Grades, and Total Grades on data from the 2018-2019 Informatics Engineering students at Tarumanagara University.

### c. Interview

Interviews are activities carried out to search for and collect information, data needed to build a system, activities carried out with the assistance of one of the Tarumanagara University Faculty of Information Technology study programs.

## 2.2 Systems Analysis

System design is useful in identifying problems and analyzing existing problems, then designing a student graduation prediction system. System design starts from illustrating the system with Context Diagrams, DFD, ERD, and Relationships Between Tables.

### a. Context Diagram

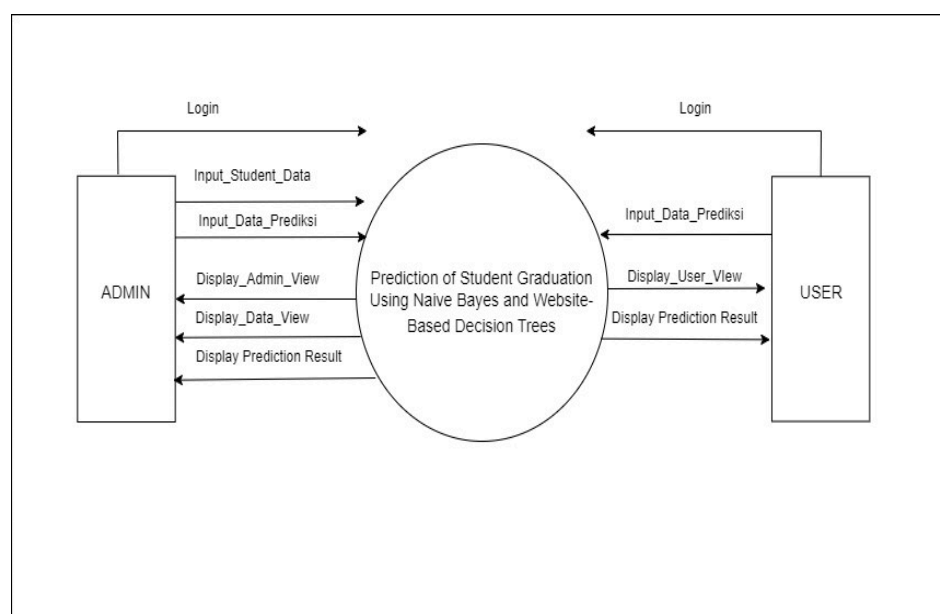


Figure 2 Context Diagram

b. DFD (Data Flow Diagram)

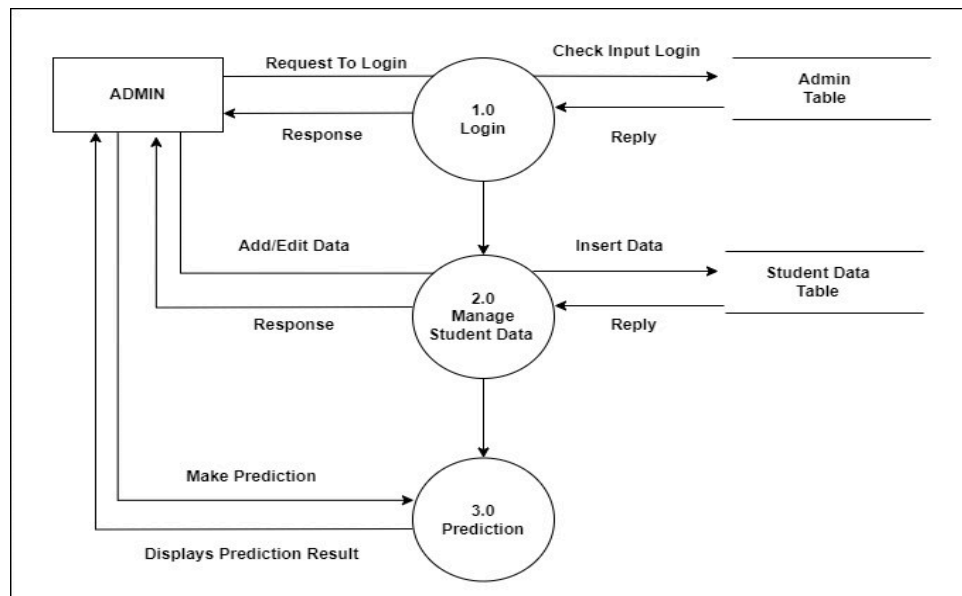


Figure 3 Data Flow Diagram

c. ERD (Entity Relationship Diagram)

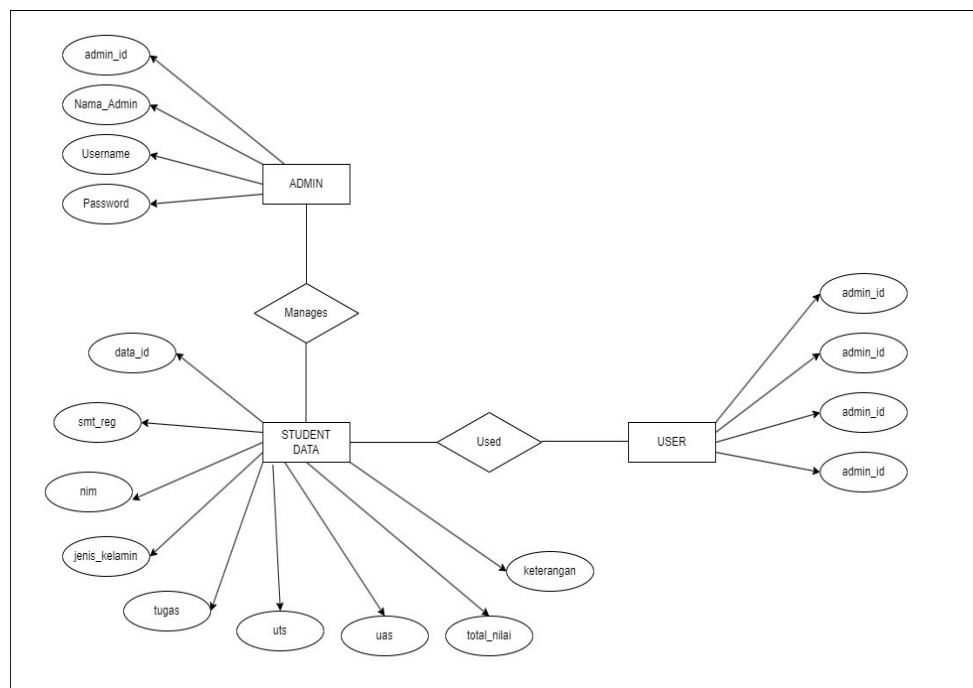


Figure 4 ER Diagram

#### d. Relationships Between Tables

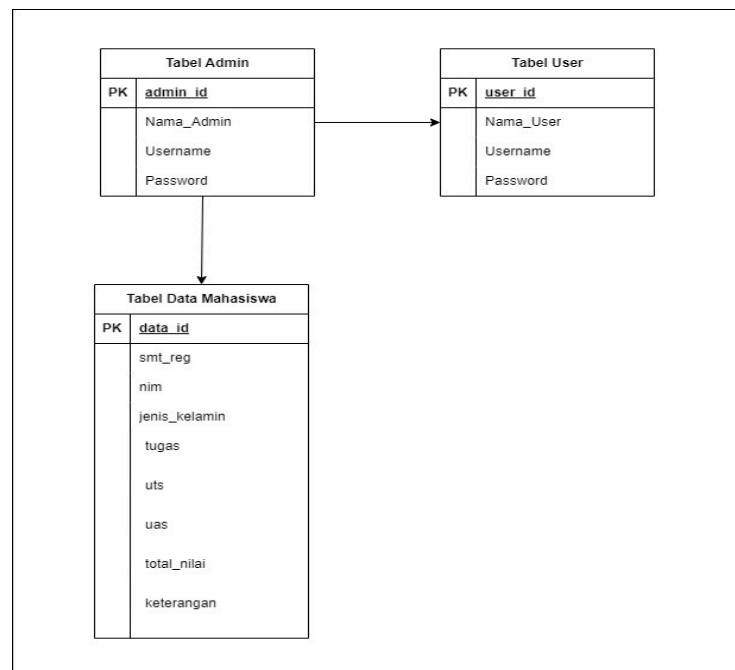


Figure 5 Relationships Between Tables

### 2.3 Naive Bayes

Naive Bayes is a probabilistic classification method. This method calculates a set of probabilities by summing the frequency and combination of values from the given dataset. The Naive Bayes method assumes that all attributes in each category are independent of each other, meaning they do not depend on one another [4].

Here is the basis of the theorem used [5]:

$$(H|X) = \frac{P(X|H) P(H)}{P(X)} \quad (2.1)$$

Explanation:

X: Data with an unknown class.

H: Hypothesis that data X belongs to a specific class.

P(H|X): Probability of hypothesis H given condition X (Probability that hypothesis H is true given that X is true).

P(H): Probability of hypothesis H (Probability that H is true regardless of X).  
P(X|H): Probability of X given condition H (Probability that X is true if H is true).  
P(X): Probability of X (Probability that X occurs regardless of the hypothesis).

## 2.4 C4.5 Algorithm

The C4.5 algorithm is used to construct a decision tree. A decision tree is valuable for exploring data and uncovering hidden relationships among several candidate input variables and a target variable.

In general, the steps for building a decision tree using the C4.5 algorithm are as follows: [6]

1. Select the attribute to be used as the root: Choose an attribute that best splits the data into distinct classes.
2. Create branches for each value: For the chosen attribute, create branches for each unique value of that attribute.
3. Partition the cases within branches: Divide the cases in each branch based on the values of the selected attribute.
4. Repeat the process: Recur this process for each branch until all cases in the branch have the same class.

To choose an attribute as the root, it is based on the highest gain value among all available attributes. The formula used to calculate gain is as follows:

$$Gain(s, a) = Entropy(s) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \quad (2.2)$$

Explanation:

S: Set of cases (data)

A: Attribute

n: Number of partitions of attribute A

|S<sub>i</sub>|: Number of cases in partition i

|S|: Total number of cases in set S

Before obtaining the Gain value, it is necessary to calculate the Entropy. Entropy is used to determine how informative an attribute input is in generating an attribute. The basic formula for Entropy is as follows:

$$Entropy(S) = - \sum_{i=1}^n p_i * \log_2 p_i \quad (2.3)$$

Explanation:

S: Set of cases (data)

n: Number of partitions in S

p<sub>i</sub>: Proportion of S<sub>i</sub> to S

The C4.5 algorithm is an improvement over ID3 that uses Gain Ratio to refine information gain. The formula for updating information gain using Gain Ratio is as follows: [6]

$$\text{Gain ratio} = \frac{\text{Information gain (S.A)}}{\text{split info(S.A)}} \quad (2.4)$$

Explanation:

S: Space/Data Sample used for training data.

A: Attribute.

Gain (S.A): Information Gain in attribute A

Split Info (S.A): Split Information in attribute A

The attribute with the highest Gain Ratio value is then selected as the test attribute for the node. This approach normalizes the Information Gain by using what is called Split Information. The formula for this is:

$$\text{SplitInfo (S.A)} = - \sum_{i=1}^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (2.5)$$

Explanation:

S: Sample data used for training.

A: Attribute.

Si: Number of samples for attribute i

### 3. RESULT AND DISCUSSION

#### 3.1 System Implementation

##### a. Module Login

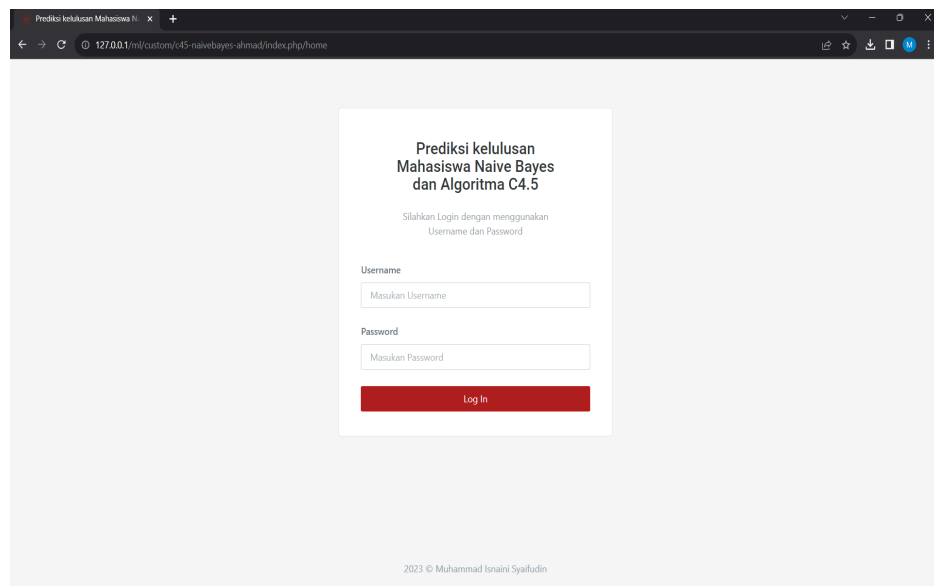


Figure 7 Module Login

## b. Module Dashboard

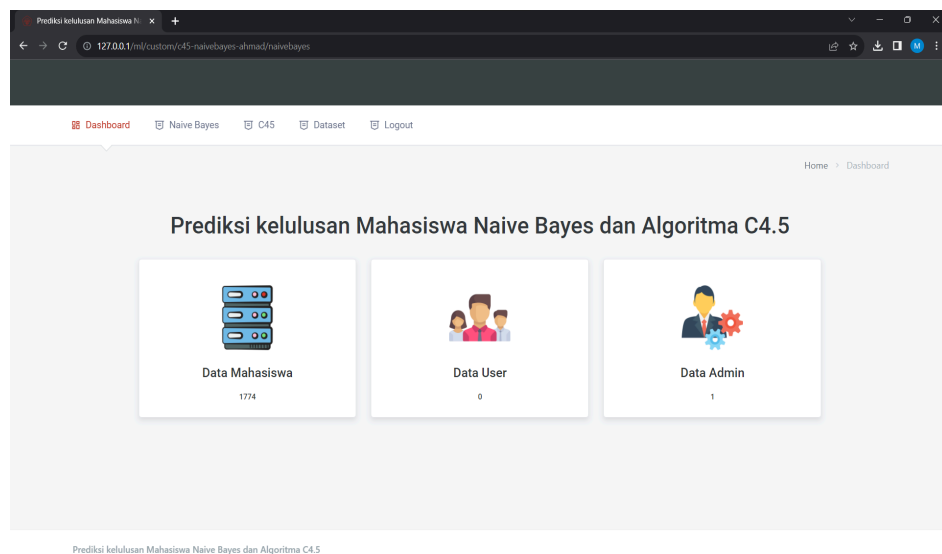
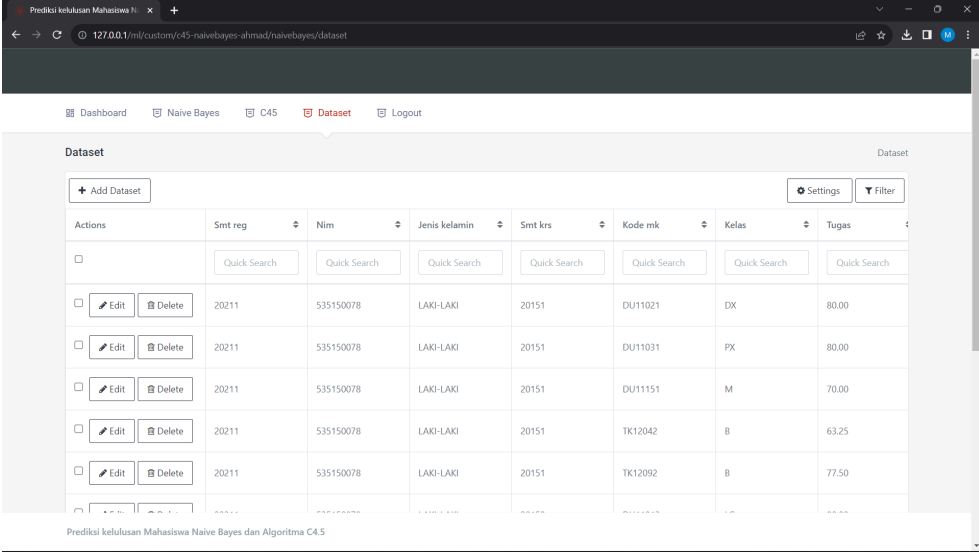


Figure 8 Module Dashboard

## c. Module Dataset



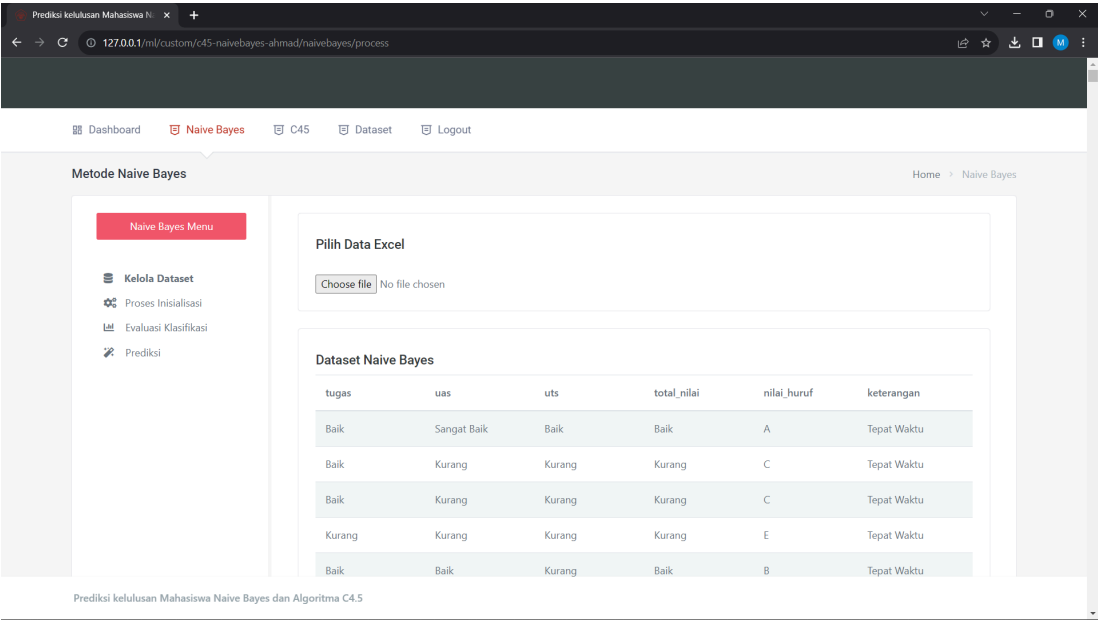


Prediksi kelulusan Mahasiswa Naive Bayes dan Algoritma C4.5

Actions	Smt reg	Nim	Jenis kelamin	Smt krs	Kode mk	Kelas	Tugas
<input type="checkbox"/>	Quick Search	Quick Search	Quick Search	Quick Search	Quick Search	Quick Search	Quick Search
<input type="checkbox"/> Edit Delete	20211	535150078	LAKI-LAKI	20151	DU11021	DX	80.00
<input type="checkbox"/> Edit Delete	20211	535150078	LAKI-LAKI	20151	DU11031	PX	80.00
<input type="checkbox"/> Edit Delete	20211	535150078	LAKI-LAKI	20151	DU11151	M	70.00
<input type="checkbox"/> Edit Delete	20211	535150078	LAKI-LAKI	20151	TK12042	B	63.25
<input type="checkbox"/> Edit Delete	20211	535150078	LAKI-LAKI	20151	TK12092	B	77.50

Figure 9 Module Dataset

#### d. Module Naive Bayes



Prediksi kelulusan Mahasiswa Naive Bayes dan Algoritma C4.5

Naive Bayes Menu

- Kelola Dataset
- Proses Inisialisasi
- Evaluasi Klasifikasi
- Prediksi

Pilih Data Excel

Choose file No file chosen

Dataset Naive Bayes

tugas	uas	uts	total_nilai	nilai_huruf	keterangan
Baik	Sangat Baik	Baik	Baik	A	Tepat Waktu
Baik	Kurang	Kurang	Kurang	C	Tepat Waktu
Baik	Kurang	Kurang	Kurang	C	Tepat Waktu
Kurang	Kurang	Kurang	Kurang	E	Tepat Waktu
Baik	Baik	Kurang	Baik	B	Tepat Waktu

Figure 10 Module Naive Bayes

#### e. Module C4.5

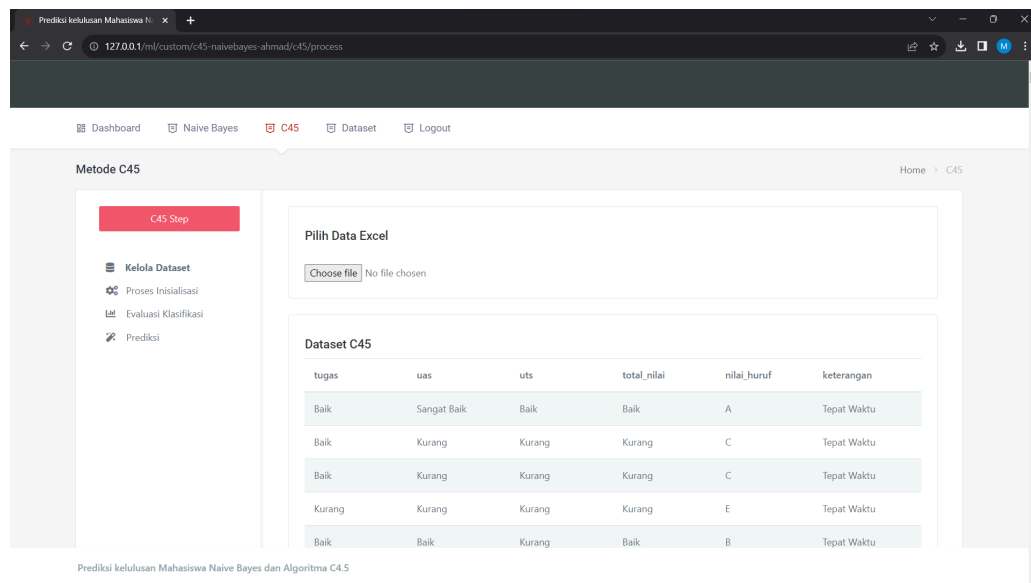


Figure 11 Module C4.5

### 3.2 Test Result

This test was carried out to determine the accuracy results of the graduation prediction system which was built to analyze student graduation predictions on time or late. The data used is data from Tarumanagara University students class 2018-2019 semesters one to four. The data used is assignment scores, Midterm Exam Scores, Final Exam Scores, Total Scores and gender. The testing system is obtained based on the overall existing data by dividing the data into three data divisions, 50% training data: 50% test data, 70% training data: 30% test data, 90% training data: 10% test data.

Table 1. Accuracy Testing

Proportion	Akurasi Naive Bayes	Akurasi C4.5
50:50	84%	62%
70:30	82,3%	88,3%
90:10	79,1%	96,3%

## 4. CONCLUSIONS AND RECOMMENDATIONS

### 4.1 CONCLUSIONS

Based on research that has been carried out, changes in accuracy values depend on the data used. The uneven distribution of data is the most influential thing in predicting student admissions. Variables that have evenly distributed data tend to make a better contribution to prediction

accuracy, especially in the Naive Bayes method. Then, in Naive Bayes, the more testing data/tests, the higher the level of accuracy, whereas by using the C4.5 algorithm, the less test data, the higher the accuracy. Even and accurate data collection is essential to improve prediction accuracy in both methods.

## 4.2 RECOMMENDATIONS

The suggestions addressed to readers for developing this system include:

1. Collecting good data can influence research results so that data predictions become more accurate.
2. Carry out development or research using other classification methods.

## 4.3 Acknowledgement

Praise and gratitude to Allah SWT submitted by me for the blessings and bounty such as the ideas, great health and good people around me so that I finally finished my graduating paper. Without it, I would not be here to write an acknowledgment of my paper. I would like to Thank you to my Prophet Muhammad SAW, who gave us the light in the dark age. The writer is also expressing her extremely grateful to the following people:

- Bagus Mulyawan S.Kom., M.M., As a Computer Science Lecturer.
- Novario Jaya Perdana S.Kom., M.T. As a Computer Science Lecturer.
- Fathin Dzakiyah. As a student at UIN Walisongo and someone who helped by providing encouragement in the process of creating this paper.

## REFERENCE

- [1] Ghuftron Afif Pratama. (2021). "Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Teknik Universitas Islam Riau Menggunakan Algoritma Naive Bayes". Pekanbaru: Universitas Islam Riau.
- [2] Bustami. (2013). "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi". TECHSI: Jurnal Penelitian Teknik Informatika. 3(2), 127-146.
- [3] Syamsu, S., Muhajirin, M., & Wijaya, N. S. (2019). "Rules Generation Untuk Klasifikasi Data Bakat dan Minat Berdasarkan Rumpun Ilmu Dengan Decision Tree". Inspiration: Jurnal Teknologi Informasi Dan Komunikasi. 9(1), 40.
- [4] Nafalski, A., Dan Wibawa, A. P. (2016). "Machine Translation with Javanese Speech Levels' Classification". Informatics, Control, Measurement in Economy and Environment Protection. 6(1), 21–25.
- [5] Suriana, Bintari, E.D., Kurniawan, D. (2017). "Desain Aplikasi Klasifikasi Kelulusan Mahasiswa Menggunakan Metode Naive Bayes Dan Algoritma C4.5". Journal Of Big Data Analytic and Artificial Intelligence. 3(1), 31-41.
- [6] Kamagi, D.H., Hansun, S. (2014). "Implementasi Data Mining Dengan Algoritma C4.5 Untuk Memprediksi Tingkat Kelulusan Mahasiswa". ULTIMATICS., 6(1), 15-20.
- [7] Arie Suharyadi. (2022). "Analisa Sistem Prediksi Penyakit Stroke Menggunakan Decision Tree C4.5". Jakarta: universitas Trisakti.