ISSN 1412-2715

Prediction model of teacher candidate student graduation status: Decision Tree C4.5, Naive Bayes, and k-NN

Kartianom [a,1,*], Arpandi [a,2,*], Gulzhaina K. Kassymova [b,2], Oscar Ndayizeye [c,3]

[a] Institut Agama Islam Negeri Bone, Jl. HOS Cokroaminoto, Bone, Indonesia
[b] Abai Kazakh National Pedagogical University, Kazakhstan
[c] Hebei Foreign Studies University, China
[1] kartianom@gmail.com*; [2] arpandi@gmail.com; [3] gulzhainak@gmail.com; [4] oscarndayiz@gmail.com
* Korespondensi Penulis

ARTICLE INFO

Article history
Received December 12, 2022
Revised December 14, 2022
Accepted December 15, 2022
Available Online December 16, 2022

Keywords
Data Mining
Decision Tree C4.5
Naïve Bayes
k-NN
Student Graduation Status

ABSTRACT

This study aims to determine the prediction model of the graduation status of prospective teacher students at IAIN Bone in terms of attributes, accuracy levels, and differences in the level of accuracy produced in the attributes of decision tree C4.5, Naïve Bayes, and k-NN data mining algorithms. This research uses a quantitative approach by adopting the Data Mining method. This research was conducted at IAIN Bone. The data collection process in this study used documentation techniques in the form of data on alumni of the Tarbiyah Faculty of IAIN Bone. The data analysis used was a descriptive analysis using decision tree C4.5, Naive Bayes, and k-NN data mining algorithms assisted by the RapidMiner application. The results of this study show that (1) model prediction of the graduation status of prospective teacher students in IAIN Bone in terms of attributes generated in the Decision Tree C4.5 and Naïve Bayes data mining algorithms consist of gender, age, Semester 1 IP, Semester 2 IP, Semester 3 IP, Semester 4 IP, and GPA, while the attributes produced in k-NN data mining algorithm consists of gender, regional origin, number of siblings, age, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, and GPA; (2) model prediction of graduation status of iain bone teacher candidate students in terms of the accuracy rate generated in the Decision Tree C4.5 data mining algorithm of 93.90%, Naïve Bayes by 90.24%, and k-NN of 92.07%; and (3) there was no significant difference between the accuracy rate produced by decision tree's data mining algorithm. C4.5 and Naïve Bayes (p-value = 1.00); Decision Tree C4.5 and k-NN (p-value = 1.00); as well as Naïve Bayes and k-NN (p-value = 1.00) in predicting the graduation status of iain bone teacher candidate students.

1. Introduction

Over the past 5 years there has been an imbalance between the number of incoming students and the number of students graduating on time in most Indonesian universities. At UIN Syarif Hidayatullah Jakarta, starting from 2014-2020, the number of graduates on time is less than 50% (Wirawan, 2020). At STMIK AMIKOM Surakarta, there was a decrease in the percentage of on-time student graduation which was previously in 2017 by 90% to 70% in 2018. These two conditions are also experienced by IAIN Bone, especially in the (Mulyadi & Sugiarto, 2021)Madrasah Ibtidaiyah Teacher Education Study Program (PGMI Study Program). Based on the results of the Self-Evaluation Report (LED) of the PGMI Study Program in 2020, information was obtained that the percentage of the number of students who graduated on time in 2018 was 50%, in 2019 it was 52%, and in 2020 it was 28%. This indicates a problem with the strategies used in increasing the percentage of students graduating on time.

Previous research on predictive models of student graduation timeliness tended to analyze three things. First, the accuracy of the predictive model of the timeliness of graduating students. This kind of research manages to provide information about the percentage of accuracy of some prediction models (Rohmawan, 2018; Zainuddin, 2019). Second, the development of prediction models. This kind of research focuses on uncovering the design of a model for predicting the graduation status of students of the faculty of informatics engineering using certain widely used algorithms, such as Naive Bayes(Fadrial, 2021; Haryanto & Saputra, 2018; Qisthiano et al., 2021; Testiana, 2018) . Third, factors affecting the timely graduation of students. This kind of research focuses on finding those variables that have an influence (Nastiti et al., 2019; Nursyafti & Purwanto, 2021). Of the three previous research trends, no one has discussed the model of predicting the graduation status of prospective teacher students using the Decision Tree C4.5 algorithm, Naive Bayes, and k-NN assisted by the RapidMiner Application.

This study is based on an argument that the imbalance between the number of incoming students and the number of students who graduate on time is influenced by the strategies used. The wrong strategy if it continues to be allowed, it will have an impact on the management of study programs that are ineffective due to Over-Capacity (Mulyadi & Sugiarto, 2021). The use machine learning and data mining algorithms is one of the strategies that can be used to find patterns or measurable prediction models from a set of big data (Bulut & Yavuz, 2019a). The accuracy of the pattern or prediction model is closely related to the accuracy of the algorithm used. Instructions on the use of measurable (accurate) strategies in increasing the percentage of student graduation on time (≥ 50%) have also been explained in the Regulation of the National Accreditation Board for Higher Education Number 5 of 2019 concerning Study Program Accreditation Instruments.

This study seeks to fill the gaps from previous studies by providing information related to student graduation status prediction models by utilizing student graduate data for prospective teachers, sophisticated machines, and accurate algorithms. It is hoped that from the results of this study, a website-based prediction model can be developed that can later be accessed by students every semester to find out their graduation status in the future based on their current condition. In addition, the results of this study can also be used by study program managers as a strategy in increasing the percentage of student graduation on time which is integrated with student data. Therefore, this research becomes very important to be carried out considering that the indicators of student graduation status on time have a very large influence in the accreditation needs of study programs criterion 9.

2. Method

2.1. Types of Research

This research was approached quantitatively by adopting the Data Mining method. Current data mining methods can be categorized into two main types according to the availability of target variables (dependent variables) in the data, namely supervised (supervised, commonly also known as predictive) and unsupervised (unsupervised, commonly also known as descriptive). This research uses method data mining which is supervised because it wants to predict certain target variables that are already available in the data (Aldowah et al., 2019). Research with supervised data mining methods often provides higher prediction accuracy than traditional methods, such as linear regression and multiple logistics (Koon & Petscher, 2015; Spikol et al., 2018). This research focuses on two data mining algorithms that can be used for classification and regression problems, namely Decision Tree,

Naive Bayes, and K-NN. The selection of these two data mining algorithms is based on the results of previous studies that stated that the Decision Tree, Naive Bayes, and k-NN algorithms are the best algorithms to answer most real-world classification problems (Fernández-Delgado et al., 2014).

2.2. Time and Place of Research

This research will be carried out for 3 months, starting from March to May 2022. This research was carried out in Kabupaten Bone. Data collection was carried out at UTIPD IAIN Bone.

2.3. Data Sources and Research Objects

The data of this study are alumni of the Tarbiyah Faculty of IAIN Bone who graduated in 2018, 2019, 2020, and 2021, namely 1,795 alumni. The object of this study is data on 1,795 alumni of the Tarbiyah Iain Bone Faculty students related to gender, hometown, school test scores, parental investigation, parents' occupations, study program names, semester achievement index (IPS), cumulative achievement index (GPA), and study period (graduation status).

2.4. Research Variables

There are variables in this study, namely target variables or result variables and predictor variables. The target variable of this study is the graduation status of students (on time or not on time) obtained from data on the length of the alumni / graduate study period. The length of this period of study is obtained from the results of the confinement of the year of graduation and the year of entry. If the study period of alumni / graduates when completing S1 studies ≤ 4 years, then the graduation status is categorized on time. While alumni /graduates when completing their S1 studies > 4 years, their graduation status is categorized as late.

For this type of research with a quantitative approach, a description of the validity and reliability of the instrument is very important, since it describes the quality of the information produced. Likewise with a qualitative approach, data triangulation information needs to be described. The predictor variables in this study are variables related to gender, regional origin, age, semester 1 achievement index (IPS 1), semester 2 achievement index (IPS 2), semester 3 achievement index (IPS 3), semester 4 achievement index (IPS 4), semester 5 achievement index (IPS 5), semester 6 achievement index (IPS 6), cumulative achievement index (IPK), and study period (graduation status). For more details the predictor variables are presented in Table 1.

Table 1. List of Predictor Variables of Graduation Status of IAIN Bone Teacher Prospective Students

Variable	Data Type	Description
Gender	Category	Men; Woman
Regional Origin	Category	Bone District; Outside Bone District
Age	Category	0-24 Years; More than 24 Years
IPS 1	Category	1st semester achievement index score
IPS 2	Category	2nd semester achievement index score
IPS 3	Category	3rd semester achievement index score
IPS 4	Category	4th semester achievement index score
IPK	Category	Cumulative achievement index value
Graduation Status	Category	On Time and Late

2.5. Data Collection Techniques

The data of this study is in the form of ex-post facto data obtained through documentation techniques by submitting requests for the data needed in this study to the Head of UTIPD IAIN Bone. What will be documented is in the form of data on the response of alumni /graduates of the Tarbiyah Iain Bone Faculty students who graduated in 2018, 2019, and 2020. The selection of alumni/graduate data in 2018, 2019, and 2020 as data for this study is because there were no graduates in 2016 and 2017. The data collected are in the form of gender data, region of origin, school test scores, parents' work, parental education, study program names, semester 1 achievement index (IPS 1), semester 2 achievement index (IPS 2), semester 3 achievement index (IPS 3), semester 4 achievement index (IPS 4), cumulative achievement index (GPA), and study period (graduation status).

2.6. Data Analysis Techniques

The data analysis used was a descriptive analysis using decision tree C4.5, Naive Bayes, and k-NN data mining algorithms assisted by the RapidMiner application. In the implementation of the analysis using a decision tree C4.5 data mining algorithm, Naive Bayes, and k-NN assisted by RapidMiner, there are several stages that need to be done, namely datasets, data mining methods, compilers, and evaluation. At the dataset stage, data cleaning, data integration, data reduction, and data transformation processes will be carried out. At the data mining method stage, what method will be chosen that matches the available data (in this study using the classification data mining method). At the knowledge stage, the data will be studied using algorithms from classification data mining methods (Decision Tree C4.5, Naive Bayes, and k-NN) which will then produce a prediction model of student graduation status. At the evaluation stage, the student graduation status prediction model is seen at the level of accuracy of the prediction. Unlike traditional statistical methods, data mining methods require researchers to build multiple models, evaluate the results of each model, adjust the model, and refine the model until an acceptable level of accuracy is achieved.

3. Results and Discussion

3.1. Result

3.1.1 Data Preparation

This study analyzes secondary data (ex-post facto) alumni/graduates of Tarbiyah IAIN Bone Faculty students who graduated in 2018, 2019, 2020, and 2021. The selection of alumni/graduate data in 2018, 2019, 2020, and 2021 because the data in 2016 and 2017 had no graduates. Alumni/graduate data in this study were obtained from the IAIN Bone database. There are three types of data obtained, namely data related to Students, Alumni, and Study Result Cards (KHS). The three data are then combined using a unique code "Student ID". The data collected was 1,643 student data. The recapitulation of student data based on the timeliness of graduation, year of graduation, and gender is presented in Table 2. To ensure data quality, it is necessary to carry out Data Preparation activities. There are three Data Preparation activities, namely Data Cleaning, Data Reduction, and Data Transformation.

Table 2. Recapitulation of Student Graduation Time for Prospective Teachers by Class and Gender

Graduate Years	Gender				Sum
	Man		Woman		
	On time	Late	On time	Late	
2018	53	19	265	40	377
2019	78	27	280	19	404
2020	46	3	221	18	288
2021	84	16	452	22	574

Data cleaning aims to identify, fill in or delete incomplete data (missing value), typo or wrong input (noisy), and have an extreme value (outlier). Of the 1,643-student data and 14 attributes used in this study, there were 582 incomplete data on IPS 1 attributes, 670 incomplete data on IPS 2 attributes, 679 incomplete data on IPS 3 attributes, 638 incomplete data on IPS 4 attributes, and 1 incomplete data on gender attributes. To overcome the problem of missing value data, replace the missing value with the average method with the help of the RapidMiner application.

Data reduction aims to obtain a dataset with a smaller number of attributes and records but informative. There are two ways that can be done for data reduction, namely dimensionality reduction and numerosity reduction. Dimensionality reduction itself is twofold, namely feature extraction (principal component analysis, independent component analysis, and others) and feature selection (filter approach, wrapper approach, and embedded aproach). In this study, data reduction was carried out by means of feature selection, especially the filter approach. The filter approach activity utilizes the Rapidminer application with the select attribute operator. The result of this data reduction activity is in the form of accuracy information in decision making.

Data transformation is used to improve the accuracy and efficiency of algorithms. The attributes transformed in this study are related to the study period. The type of data from the study period

attribute is integer data that is transformed into categorical (nominal) data. The results of the transformation of the attribute data of this study period are named the attributes of student graduation status with two categories, namely on time and late. The data transformation process itself is carried out through the help of the RapidMiner application using the generate attribute operator.

3.1.2 IAIN Bone Teacher Candidate Student Graduation Status Prediction Model

3.1.2.1Decision Tree C4.5

Visualization of the prediction model resulting from the Decision Tree C4.5 method assisted by the Rapidminer application in the form of a decision tree. In addition to the decision tree, the information generated is in the form of a description or decision-making rule (rule). Based on Figure 6, information was obtained that age is the strongest attribute compared to the other 6 attributes in explaining the prediction model of student graduation status resulting from the Decision Tree C4.5 method. There are 12 decisions resulting from the Decision Tree C4.5 method. However, of the 12 decisions, it can be made simpler so that it can be more easily interpreted. To make it easier to interpret the decision tree from Figure 1, you can use the description information as shown in Figure 2.

Based on Figure 2, information was obtained that the decision in prioritizing the graduation status of students was based on the age category of students. For students with an age of more than 24 years, the decision making is seen from the IP score of Semester 1, while for students with an age of less than or equal to 24 years, the decision making is seen from the student's GPA score in semester 4. The GPA value itself has four branches in its decision-making, namely a GPA of less than 2.75; GPA between 2.75 and 2.99; GPA between 3.00 and 3.49; and a GPA of more than 3.49. If the student's GPA score is less than 2.75, it is necessary to look at the student's Semester 1 IP score (Semester 1 IP is less than 2.75 is predicted to graduate late, while Semeter 1 IP is between 3.00 to 3.49 is predicted to graduate on time). If the student's GPA score is between 2.75 and 2.99, the decision-making pays attention to the gender of the student concerned (Men are predicted to graduate on time, while women are predicted to graduate late). If the student's GPA score is between 3.00 to 3.49, it is predicted to graduate on time, while students with a GPA of more than 3.49, the decision-making needs to be seen from the grades of Semester 1 IP, Semester 2 IP, Semester 3 IP, and Semester 4 IP from the student concerned. Students with a GPA of more than 3.49 accompanied by IP semester 1, IP Semester 2, IP Semester 3, and IP Semester 4 which are also more than 3.49 are predicted to graduate late. On the other hand, students with a GPA of more than 3.49 accompanied by IP Semester 1, IP Semester 2, and IP Semester 3 which are also more than 3.49, but the 4th semester IP score of the student concerned has decreased or is in the range of 3.00-3.49, so it is predicted to graduate late.

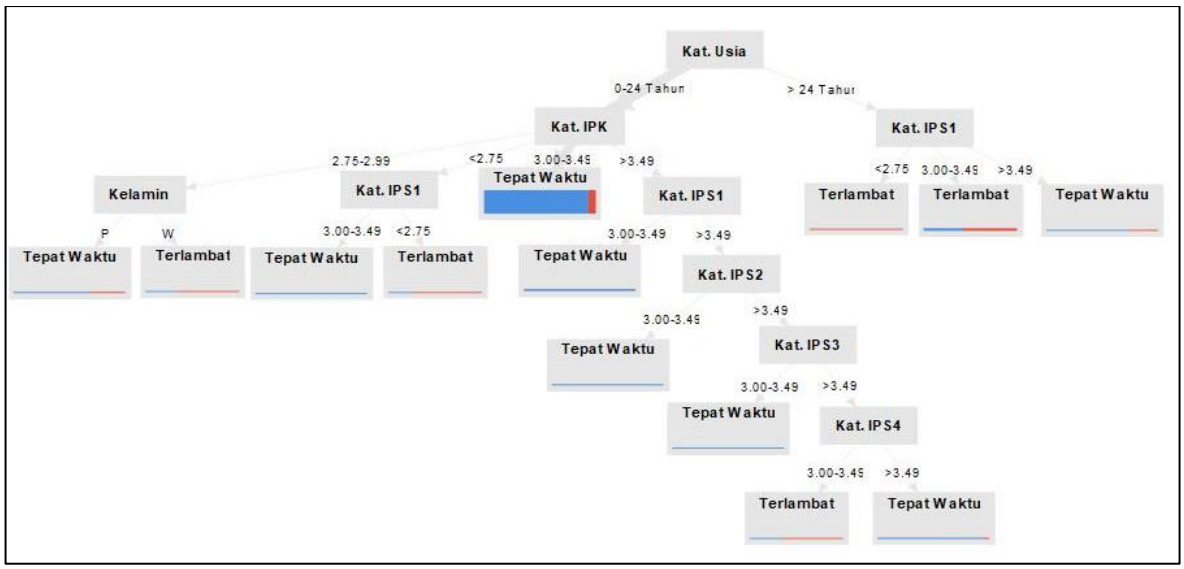

Fig. 1. Visualization of Decision Tree C4.5 Prediction Model

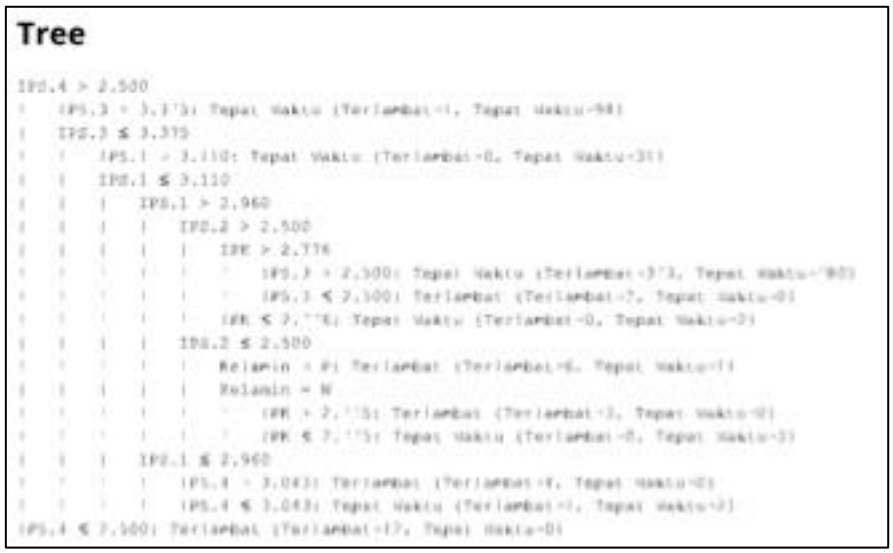

Fig. 2. Description *Decision Tree* C4.5 Prediction Model

3.1.2.2Naïve Bayes

Visualization of the prediction model resulting from the Naïve Bayes method assisted by the Rapidminer application in the form of a distribution table. There are four variables presented in the Naïve Bayes distribution table, namely attributes, parameters, timely probability, and late probability. The visualization of the Naïve Bayes distribution table is presented in Figure 3.

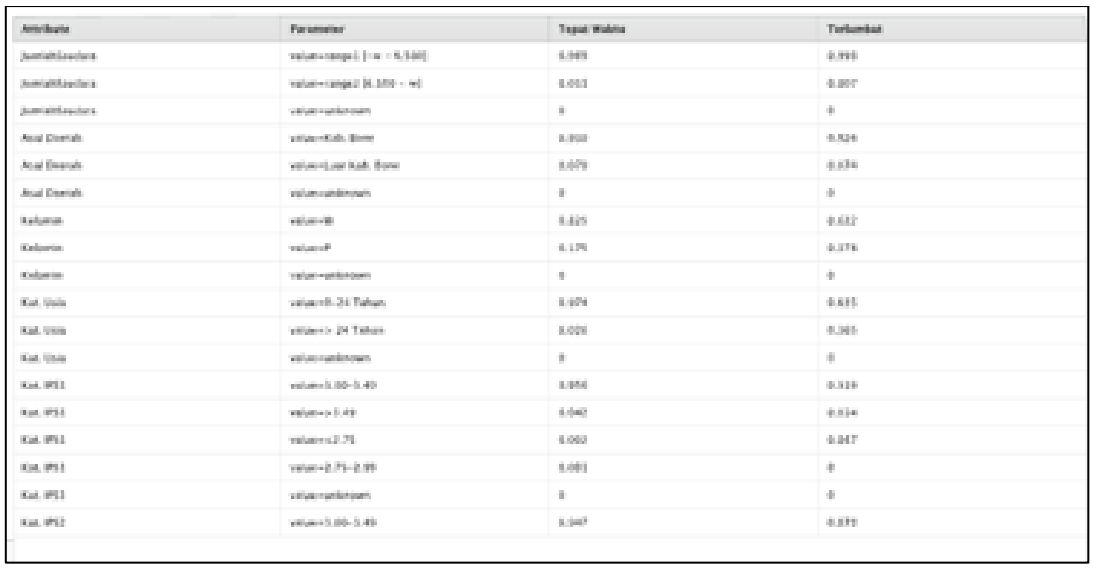

Fig. 3. Naïve Bayes Prediction Model Visualization

Based on Figure 3, information was obtained that the decision in predicting student graduation status was based on nine attributes used in this study, namely gender, regional origin, number of siblings, age, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, and GPA. However, when viewed from the probability value, the attributes of the number of relatives and regional origins are not able to provide precise prediction results because they are unable to accurately distinguish students who graduate on time and late. This is shown from the probability value of college students graduating on time and graduating late on the attributes of the number of brothers who are less than 6.5 are equally more than 99%. The same information was also obtained from the attributes of regional origin, students from Bone Regency and outside Bone Regency both had the probability value of graduating on time and the longest as big as 93% and 7%. On the other hand, attributes such as gender, age, Semester 1 IP, Semester 2 IP, Semester 3 IP, Semester 4 IP, and GPA can accurately distinguish students who graduate in time and late.

3.1.2.3k-NN

Visualization of prediction models resulting from the k-NN method assisted by the Rapidminer application in the form of prediction tables from everyone. In the k-NN prediction table as one needs to note is the confidence (On Time) and confidence (Late) columns. If the value in the confidence column (on time) is greater than the confidence value (late) then the resulting decision in the prediction column (Status) is "On Time", and vice versa. The data presented from the results of the analysis of the k-NN method are testing data that are predicted based on the results of analysis from training data (training data to make prediction models) which amount to 164 student data (10% of 1,643 student data). The k-NN prediction table is presented in Figure 4.

Based on Figure 4, information was obtained that the decision in predicting student graduation status was based on nine attributes used in this study, namely gender, regional origin, number of siblings, age, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, and GPA. All these attributes were able to accurately predict the graduation status of students by 92.07% (151 out of 164). This can be seen from the Status and prediction (Status) columns which have the same result.

Fig. 4. Visualization of k-NN Prediction Model

3.1.3 Evaluation of the Model of Predicting the Graduation Status of Iain Bone Prospective Teacher Students

The prediction model for the graduation status of prospective teacher students was obtained from the results of the analysis using the Decision Tree C4.5 and Random Forest methods. The Random Forest method has better accuracy than the Decision Tree C4.5 method (70.12% > 69.82%). The accuracy level of both methods is presented in Figure 5.

accuracy: 93.90%

	true Tepat Waktu	true Terlambat	class precision
pred. Tepat Waktu	147	9	94.23%
pred. Terlambat	1	7	87.50%
class recall	99.32%	43.75%	

Decision Tree C4.5

accuracy: 90.24%

	true Tepat Waktu	true Terlambat	class precision
pred. Tepat Waktu	144	12	92.31%
pred. Terlambat	4	4	50.00%
class recall	97.30%	25.00%	

Naive Bayes

accuracy: 92.07%

	true Tepat Waktu	true Terlambat	class precision
pred. Tepat Waktu	148	13	91.93%
pred. Terlambat	0	3	100.00%
class recall	100.00%	18.75%	

k-NN

Fig. 5. Comparison of Prediction Model Accuracy

Figure 5 shows that the Decision Tree C4.5 method can correctly investigate 7 students graduating late and 147 students graduating on time, the rest there was a prediction error. For the Naive Bayes method, it was able to properly educate 4 students who graduated late, and 144 students graduated on time, the rest there was a prediction error. While the K-NN method was able to correctly educate 5 students graduating late and 146 students graduating on time, the rest there was a prediction error. If you look further, the three methods do not show a significant difference in the degree of accuracy. This is seen to be reinforced by the results of the t-test of the comparison of the accuracy levels of the two methods presented in Figure 6.

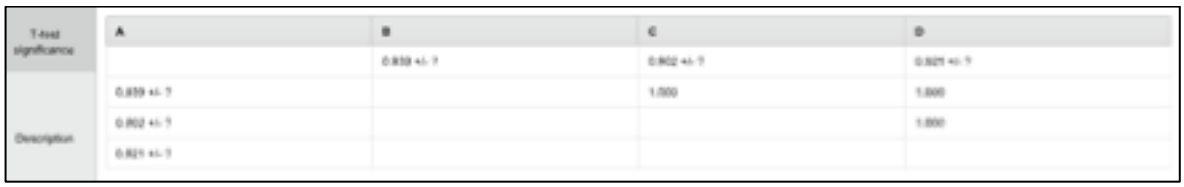

Fig. 6. T-Test Results Comparison accuracy of Prediction Models

3.2. Discussion

This study aims tofind out the prediction model, the level of accuracy of the prediction model, and the form of implementation of the prediction model of the timely graduation status of iain bone teacher candidate students in terms of the decision tree C4.5, Naive Bayes, and k-NN data mining algorithms. Based on the results of the analysis of research data, information was obtained that the Decision Tree C4.5 data mining algorithm had the highest percentage of accuracy (93.90%) followed by k-NN (92.07%) then Naive Bayes (90.24%). The results of this research are in line with the research of Effendi & Setiawan (2021) who said that the Decision Tree C4.5 method has a higher percentage of accuracy (98.99%) than the Naive Bayes method (97.42%). The results of the same study were also shown by Putri & Sulistyowati (2020) that the Decision Tree C4.5 method has a higher percentage of accuracy (80.90%) than the Naive Bayes method (77.90%). On the other hand, the results of this study have differences with the results of research from Pattiasina & Rosiyadi (2020) where the Decision Tree C4.5 method still has a higher percentage of accuracy (99.60%) but the accuracy of the k-NN method (83.00%) is not higher than the Naive Bayes method (94.07%).

The Decision Tree C4.5 method produces information that the graduation status of prospective teachers (on time and late) is based on the attributes of age, GPA, IPS 1, IPS 2, IPS 3, IPS 4, and gender. This information is in line with the results of previous studies that stated that age attributes (Arifin & Hadiana, 2019); GPA and gender jelnis (Gunawan et al., 2019; D. Y. Putri et al., 2018); and the student Semester Achievement Index (Fitrani, 2019) are attributes that affect the classification of a student's graduation status. The average level of accuracy resulting from these seven attributes in classifying a student's graduation status is more than 75%. A prediction model with a high degree of accuracy (more than 75%) will produce high information with a very small error rate. This is in line with what Bucos et al said. (2018) that prediction models with an accuracy rate of more than 75% can transform traditional data into data that can be used in education policy decision making.

The Naïve Bayes method generates information that the graduation status of prospective teacher students (on time and late) is based on the attributes of gender, age, Semester 1 IP, Semester 2 IP, Semester 3 IP, Semester 4 IP, and GPA. This information has similarities with the information generated in decision tree method C4.5. Although in terms of the level of accuracy of the prediction model, the two have differences, the differences are not significant. This is in line with the results of the research produced by Hardiani (2021) that the accuracy rate of the prediction model produced from the Naïve Bayes and Decision Tree C4.5 methods did not differ significantly (82.62% and 88.82%). With mikian we can assume that the attributes of gender, age, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, and GPA are attributes that are very relevant and have high information in predicting or classifying the graduation status of prospective teacher students with the same level of accuracy.

The k-NN method produces information that the graduation status of prospective teacher students (on time and late) is based on all attributes used in this study, namely gender, regional origin, number of siblings, age, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, and GPA. In carrying out the classification of data, the k-NN method is strongly influenced by the value of k (Altayef et al., 2022) . The value of k itself is the value of the nearest neighbor which later based on that value will

result in a decision of the predicted result (on time or late) based on the status that appears the most (on time or late) from the result of the dispute of the three nearest or smallest values. In this study, the indigo k used was 3 with a resulting accuracy rate of 92.07%, while the k value of less than 3 and more than 3 accuracy levels produced was less than 92.07%. The results of this study are different from the results of research from Hidayati & Hermawan (2021) which showed that the maximum k value with the highest accuracy was 7 with an accuracy rate of 85.25%. However, the difference in results is more due to the use of different measure types and mixed measures. In this study, measure types "mixed measures" and mixed measures "mixed euclidean distance" were used, while the research conducted by Hidayati & Hermawan (2021) used measure types "numerical measures" and numerical measures "euclidean distance".

In addition to the accuracy rate, the Decision Tree C4.5, Naive Bayes, and k-NN methods also provide information on the percentage of accuracy and divinity in predicting. The information is presented in the form of a confusion matrix. In the Decision Tree C4.5 method, the resulting model successfully predicted 147 students who graduated on time appropriately, 7 students who graduated late appropriately, 9 students who graduated on time incorrectly, and 1 student who graduated late incorrectly. In the Naive Bayes method, the resulting model successfully predicted 144 students who graduated on time appropriately, 4 students who graduated late appropriately, 12 students who graduated on time incorrectly, and 4 students who graduated late incorrectly. In the k-NN method, the resulting model successfully predicted 146 students who graduated on time appropriately, 5 students who graduated late appropriately, 11 students who graduated on time incorrectly, and 2 students who graduated late incorrectly. Overall, there was no significant difference between the Decision Tree C4.5, Naive Bayes, and k-NN methods in terms of confusion matrix accuracy. This is in line with the results of research from Altayef et al. (2022) which says that there is no significant difference in terms of the accuracy level of the Decision Tree C4.5, Naive Bayes, and k-NN methods.

Overall, the results of this study show that the Decision Tree C4.5, Naive Bayes, and k-NN methods have a very high level of accuracy in predicting the graduation status of prospective teacher students. The Decision Tree C4.5 method has a high level of accuracy with a prediction model that is simpler and easier to implement in predicting student graduation status than the Naive Bayes and k-NN methods. A web-based student graduation status prediction model using Decision Tree C4.5, Naïve Bayes, or k-NN needs to be developed given that Drop-Out (DO) issues have a negative impact in various aspects. As Rechkoski et al said. (2018) said that college students who are in a Drop-Out (DO) without receiving that degree can have very high human and monetary costs throughout life. On the other hand, the flexibilityin the overall study process and having relevant and timely student guidance in the process of selecting the study program, can have a great impact on the satisfaction of the study, the level of motivation and ultimately even reduce the dropout rate (Roy & Garg, 2017).

4. Conclusion

Based on the results of the analysis of research data on the prediction model of graduation status of prospective teacher students at IAIN Bone that has been carried out, it can be concluded that (1) model prediction of timely graduation status of iain bone prospective teacher students in terms of attributes generated in the decision tree C4.5 and Naïve Bayes data mining algorithms consists of gender, age, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, and GPA; while the attributes generated in the k-NN data mining algorithm consist of gender, regional origin, number of siblings, age, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, and GPA; (2) Model prediction of the status of timely graduation of iain bone teacher candidate students in terms of the accuracy rate generated in the Decision Tree C4.5 data mining algorithm of 93.90%, Naïve Bayes by 90.24%, and k-NN of 92.07%; and (3) There was no significant difference between the accuracy rate produced by decision tree data mining algorithm C4.5 and Naïve Bayes (p-value = 1.00); Decision Tree C4.5 and k-NN (p-value = 1.00); as well as Naïve Bayes and k-NN (p-value = 1.00) in predicting the graduation status of IAIN bone teacher candidate students.

This study aims to find a model for predicting the timely graduation status of prospective teacher students at IAIN Bone. This study has discussed information about attributes and accuracy levels generated from decision tree C4.5, Naïve Bayes, and k-NN data mining algorithms in predicting the timely graduation status of prospective teacher students at IAIN Bone. However, this study still has limitations, especially in terms of the use of data mining algorithms which are limited to only three

algorithms and the number of test attributes is still too small. For further research, use other data mining algorithms such as Random Forest in predicting student graduation status. In addition, develop one android-based prediction model (it can be with decision tree data mining algorithms C4.5, Naïve Bayes, or k-NN).

Reference

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics and Informatics*, *37*, 13–49.

Altayef, E., Anayi, F., & Packianather, M. (2022). A new enhancement of the k-NN algorithm by Using an optimization technique. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 24–31. https://doi.org/10.1109/ICACITE53722.2022.9823537

Arifin, D., & Hadiana, A. (2019). Computer-based Techniques for Predicting the Failure of Student Studies Using the Decision Tree method. *IOP Conference Series: Materials Science and Engineering*, *662*(2), 022112. https://doi.org/10.1088/1757-899X/662/2/022112

Bucos, M., Journal, B. D.-T., & 2018, undefined. (2018). Predicting student success using data generated in traditional educational environments. *Ceeol.Com*, *7*(3), 617. https://doi.org/10.18421/TEM73-19

Bulut, O., & Yavuz, H. C. (2019). Educational data mining: A tutorial for the rattle package in R. *International Journal of Assessment Tools in Education*, *6*(5), 20–36. https://doi.org/10.21449/ijate.627361

Effendi, M. M., & Setiawan, A. (2021). Menentukan prediksi kelulusan siswa dengan membandingkan algoritma C4. 5 dan naive bayes studi kasus SMKN. 1 Cikarang Selatan. *Jurnal SIGMA*, *10*(3), 183–190.

Fadrial, Y. E. (2021). Algoritma naive bayes untuk mencari perkiraan waktu studi mahasiswa. *Journal of Information Technology and Computer Science (INTECOMS)*, *4*(1).

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Fitrani, A. S. (2019). Prediction of Study Period Students (Bachelor Degree) Muhammadiyah University of Sidoarjo Based on Decision Tree Method using C4.5 Algorithm. *Journal of Physics: Conference Series*, *1179*(1), 012033. https://doi.org/10.1088/1742-6596/1179/1/012033

Gunawan, Hanes, & Catherine. (2019). Information Systems Students' Study Performance Prediction Using Data Mining Approach. *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 1–8. https://doi.org/10.1109/ICIC47613.2019.8985718

Hardiani, T. (2021). Comparison of Naive Bayes Method, K-NN (K-Nearest Neighbor) and Decision Tree for Predicting the Graduation of 'Aisyiyah University Students of Yogyakarta. *International Journal of Health Science and Technology*, *2*(1), 75–85. https://doi.org/10.31101/ijhst.v2i1.1829

Haryanto, K. W., & Saputra, R. A. (2018). Aplikasi prediksi masa studi mahasiswa menggunakan algoritma naïve bayes classifier (NBC) (Studi kasus: di STMIK Yadika Bangil). *Jurnal SPIRIT*, *10*(1), 5–12.

Hidayati, N., & Hermawan, A. (2021). K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation. *Journal of Engineering and Applied Technology*, *2*(2). https://doi.org/10.21831/jeatech.v2i2.42777

Koon, S., & Petscher, Y. (2015). Comparing Methodologies for Developing an Early Warning System: Classification and Regression Tree Model versus Logistic Regression. REL 2015-077. *Regional Educational Laboratory Southeast*.

Mulyadi, C., & Sugiarto, L. (2021). Penggunaan algoritma naïve bayes untuk prediksi ketepatan waktu lulus mahasiswa diploma 3 STMIK Cipta Darma Surakarta. *TEKNOMATIKA*, *11*(01), 21–30.

Nastiti, V. R. S., Azhar, Y., & Pramudita, A. E. (2019). Penerapan algoritma C5. 0 pada analisis faktor-faktor pengaruh kelulusan tepat waktu mahasiswa Teknik Informatika UMM. *Jurnal Repositor*, *1*(2), 131–140.

Nursyafti, Y., & Purwanto, W. (2021). Faktor-faktor penghambat kelulusan tepat waktu mahasiswa D3 jurusan Teknik Otomotif Fakultas Teknik Universitas Negeri Padang tahun masuk 2016 dan 2017. *MSI Transaction on Education*, *2*(3), 2021.

Pattiasina, T., & Rosiyadi, D. (2020). Comparison of data mining classification algorithm for predicting the performance of high school students. *Jurnal Techno Nusa Mandiri*, *17*(1), 22–30. https://doi.org/10.33480/techno.v17i1.1226

Putri, D. I., & Sulistyowati, N. (2020). Comparison of student graduation classification analysis based on study length using naïve-bayes and C4. 5 algorithms. *International Research Journal of Advanced Engineering and Science*, *5*(1), 233–238.

Putri, D. Y., Andreswari, R., & Hasibuan, M. A. (2018). Analysis of Students Graduation Target Based on Academic Data Record Using C4.5 Algorithm Case Study: Information Systems Students of Telkom University. *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 1–6. https://doi.org/10.1109/CITSM.2018.8674366

Qisthiano, M. R., Kurniawan, T. B., Negara, E. S., & Akbar, M. (2021). Pengembangan model untuk prediksi tingkat kelulusan mahasiswa tepat waktu dengan metode naïve bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(3), 987–994. https://doi.org/10.30865/mib.v5i3.3030

Rechkoski, L., Ajanovski, V. v., & Mihova, M. (2018). Evaluation of grade prediction using model-based collaborative filtering methods. *2018 IEEE Global Engineering Education Conference (EDUCON)*, 1096–1103. https://doi.org/10.1109/EDUCON.2018.8363352

Rohmawan, E. P. (2018). Prediksi kelulusan mahasiswa tepat waktu menggunakan metode desicion tree dan artificial neural network. *Jurnal Ilmiah MATRIK*, *20*(1).

Roy, S., & Garg, A. (2017). Predicting academic performance of student using classification techniques. *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, UPCON 2017*, *2018-Janua*, 568–572. https://doi.org/10.1109/UPCON.2017.8251112

Spikol, D., Ruffaldi, E., Dabisias, G., & Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *Journal of Computer Assisted Learning*, *34*(4), 366–377.

Testiana, G. (2018). Perancangan model prediksi kelulusan mahasiswa tepat waktu pada UIN Raden Fatah. *JUSIFO (Jurnal Sistem Informasi)*, *4*(1), 49–62.

Wirawan, C. (2020). Teknik data mining menggunakan algoritma decision tree C4. 5 untuk memprediksi tingkat kelulusan tepat waktu. *Applied Information Systems and Management*, *3*(1), 47–52.

Zainuddin, Moh. (2019). Perbandingan 4 algoritma berbasis particle swarm optimization (PSO) untuk prediksi kelulusan tepat waktu mahasiswa. *Jurnal Ilmiah Teknologi Informasi Asia*, *13*(1).