P-ISSN: 2715-2448 | E-ISSN: 2715-7199
Vol.5 No.1 Januari 2024
Buana Information Technology and Computer Sciences (BIT and CS)

Implementation of Orange Data Mining to Predict Student Graduation on Time at Pringsewu Muhammadiyah University

Roby Novianto[1*], Bambang Triraharjo[2], Baskoro[3]
[1,2,3] Sistem dan Teknologi Informasi, Universitas Muhammadiyah Pringsewu
[1]robynovianto@umpri.ac.id, [2]bambangtriraharjo@umpri.ac.id, [3]baskoro@umpri.ac.id

Abstract
The process of monitoring and evaluating the graduation of Muhammadiyah Pringsewu University (UMPRI) students really needs to be done because the student graduation rate is an element of accreditation assessment that is very important for each Study Program. Data Mining can be used to classify student graduation accuracy. This study aims to apply the orange data mining application using the K-Nearest Neighbor (K-NN), Decision Tree and Naive Bayes models and will then evaluate the accuracy of each of these models. This research was conducted at Pringsewu Muhammadiyah University in several batches, then student data will be analyzed using the orange data mining application using the K-NN, Decision Tree and Naive Bayes models. The data testing process applies K-Fold Cross Validation (K=5), while the evaluation model used is the Confusion Matrix and ROC. The results of the comparison of the three models are as follows, K-NN has an accuracy rate of 75.7%, Decision Tree has an accuracy rate of 78.1%, and Naive Bayes has an accuracy rate of 77.8%. Therefore, for classifying the graduation rate of Muhammadiyah University students, Pringsewu recommends the Decision Tree model because it has a better level of accuracy than K-NN and Naive Bayes.

Keywords: *Graduation, Prediction, Data Mining, C4.5, Naïve Bayes*

I. Introduction
The development of the world of education in Indonesia has had the impact of very tight competition. This was triggered by the increasingly advanced education in universities. One of the impacts of competition is producing quality graduates. The criteria for quality graduates include being able to complete mass learning on time. The timely learning period greatly influences the quality of higher education [16]. The ability of universities to produce graduates who are able to solve learning problems on time is a factor that influences higher education accreditation.

This is in accordance with the National Accreditation Board for Higher Education regulations Number 3 of 2019 concerning Higher Education Accreditation Instruments which states that one of the indicators for accreditation assessment is the percentage of graduates on time for each program from a tertiary institution. Therefore, it is necessary to monitor the student's study period. The average study period for students studying at Pringsewu Muhammadiyah University is still over 4 years so it is necessary to try an evaluation using the student classification method using the orange application with three models, namely K-Nearest Neighbor (KNN), Decision Tree and Naive Bayes. Length of study is the period of time required for students to complete their education. The duration of student study has been regulated in the Ministry of Education and Culture's decree regarding the undergraduate program (S1) education system which has a semester credit load that must be taken between 144 and 160 credits with a length of study on campus of between 8 and 10 semesters or the equivalent of between 4 and 5 years.

Information on students' grades for each semester and student graduation information can be processed to create data that is useful for analyzing the accuracy of students' study progress [1]. Based on information obtained from Muhammadiyah Pringsewa University, in the 2015 to 2020 class year with an average number of graduates of 63%, data obtained that the average student study period was

still over 4 years [6]. Therefore, it is necessary to try to analyze the factors that support the punctuality and delays in the student's study period. Previous research related to predicting student graduation and classification mostly used K-NN, SVM, Neural Network and Naive Bayes modes. [2],[14]. On the other hand, many previous studies have reviewed the results of comparative analyzes of several data mining models used to classify poor people [1], UMKM income classification [3], nutritional classification [8], as well as classifications for the agricultural industry and public health [12].

This research aims to classify the timeliness of the study period of Muhammadiyah Pringsewu University students by applying three methods, namely KNN, Naive Bayes and Decision Tree. Next, a comparative analysis of the three models will be carried out by applying Confusion Matrix and ROC analysis to ensure the level of accuracy of the three methods. This research contribution really helps the management of Muhammadiyah University of Pringsewu to develop strategies to minimize students who are not on time in completing their studies and contributes to determining the accuracy performance of several data mining methods, including KNN, Naive Bayes and Decision Tree.

II. Methods

1. Research Flow

This research aims to carry out a comparative analysis of the KNN, Naive Bayes and Decision Tree methods used to classify graduating students at Muhammadiyah University of Pringsewu. The application used for the simulation is Orange Data Mining, an open source data mining application that has been proven to be able to help researchers analyze their data. The process stages in this research can be seen in Figure 1.

Figure 1 Research flow

According to Figure 1, the first step is problem identification, formulation and literature review. This is done first to develop research objectives and research contributions [10]. Second is the process of collecting data, namely compiling training data and test data as a source of data classification. Third is the process of designing the orange data mining widget for the student graduation classification process and method comparison. Fourth is the process of classifying graduating students from Muhammadiyah University of Pringsewu using the KNN, Decision Tree and Naive Bayes models. Fifth is the process of evaluating the performance of classification methods and analyzing the comparison results of these methods.

a. K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) is a supervised method which means it requires training information to classify objects that are very close. The working principle of K-NN is to find the shortest distance between the information to be evaluated and k neighbors in the training data [11]. The dataset is grouped manually according to the type of student data Pringsewu Muhammadiyah University. The dataset used as a reference is 35 student data for which the classification model will be tested. Next, the formula calculates the similarity of the dataset vector to each training dataset that has been classified. The K-NN theorem for calculating distance universally is as follows:

$$di = \sqrt{\sum_{i=1}^{n} (x_{1j} - p_j)} \quad ^2$$

Information:
d_i = Sample distance
x_{ij} = Knowledge sample data
p_j = Data input var ke-j
n = Number of samples
The stages of the process of implementing the K-NN method are as follows:

1) Determine the parameter k (number of closest neighbors).
2) Calculates the square of the object's Euclidean distance to the given training data.
3) Sort result number 2 in ascending order (in order from high to low)
4) Collecting Y categories (nearest neighbor classification based on k value)
5) By using the nearest neighbor category with the majority, the object category can be predicted.

b. Decision tree (C4.5)

Decision Tree based on C4.5 algorithm is a commonly used classification technique to extract relevant relationships in data. The C4.5 algorithm is a program that creates a decision tree based on a labeled input data set. The advantage is that the model can be easily interpreted and implemented with both continuous and discrete values. The C4.5 algorithm divides training data with the help of information acquisition [17]. Attributes that have high frequencies are considered to separate data based on the information available in the dataset. When calculating the gain value, you need to know the entropy value, namely using the following formula:

$$Entropy\ (i) = \sum_{j=1}^{m} \quad f(i,j).\ 2f\ [(i,j)]$$

1) Gain value using the formula:

$$gain = -\sum_{i=1}^{p} \quad .\ IE\ (i)$$

2) To calculate the gain ratio, you need to know a new term called Split Information with the formula:

$$SplitInformation = -\sum_{t=1}^{c} \quad \frac{S1}{S} log2 \frac{S1}{S}$$

3) Next, calculate the gain ratio

$$Gainratio\ (S, A) = \frac{Gain(S,A)}{SplitInformation\ (S,A)}$$

4) Repeat step 2 until all records have been split. The decision tree splitting process ends when:
   1) All tuples in node record m are of the same class.
   2) The attributes in the dataset are not further divided.
   3) An empty branch has no records

c. Naïve Bayes

Bayesian classification is a statistical classification that can be used to predict the probability of membership of a class discovered by the British scientist Thomas Bayes [4]. Naive Bayes is a classification algorithm that is quite simple and easy to implement so this algorithm is very effective when tested with the correct data set, especially if Naive Bayes is combined with function selection, so Naive Bayes can reduce redundancies in the data, besides that Naive Bayes shows good results when combined with clustering methods. Naive Bayes is proven to have high accuracy compared to support vector machines.

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

Then X is evidence, H is hypothesis, P(H|X) is probability that hypothesis H is true, evidence of true or hypothesis H or Posterior

$$P(C|F1 \ldots . Fn) = \frac{P(C)P(F1.Fn|C)}{P(F1 \ldots Fn)}$$

So variable C explains the class, while variables F1...Fn explain the character of the instructions in carrying out the classification. Where this formula explains the probability that the sample enters a special character in class C (Posterior), namely the probability that it comes out of class C (before entering the sample, many priors are made), multiplied by the probability of the sample character appearing in class (also called likelihood), divided by the probability of the character appearing global examples (also called evidence). The formula above can be made simply as follows

$$Posterior = \frac{Prior\ x\ likelihod}{evidence}$$

For continuous data classification, the Gaussian Density formula is used:

$$P(Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu i)^2}{2\sigma ij}}$$

Where: P: Opportunity
Xi: Atributke i
xi: The value of the i attribute
Y: Class sought
yi: Subclass Y is sought
μ: mean, explains the average of all attributes
σ: Standard deviation, explained variance across attributes.

2. Evaluasi Kinerja

a. Confusion matrix

This method only uses a matrix table as in Table 1, if the dataset only consists of two classes, one class is considered positive and the other negative [5]. Evaluation with the confusion matrix produces accuracy, precision and recall values.

Table 1 Confusion Matrix

Correct Classification	*Classified as*	
	+	-
+	*True positives*	*False negatives*
-	*False positives*	*True negatives*

True Positive is the number of positive records that are classified as positive, false positive is the number of negative records that are classified as positive, false negative is the number of positive records that are classified as negative, true negative is the number of negative records that are classified as negative,

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

$$P = \frac{TP}{TP+FP}$$

$$Sn = \frac{TP}{TP+FN}$$

$$Sp = \frac{TN}{TN+FP}$$

b. Kurva ROC

The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold varies [2]. This method was originally developed for military radar receiver operators starting in 1941, giving rise to its name. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. True positive rate is also known as sensitivity, recall, or probability of detection. The false positive rate is also known as the false alarm probability and can be calculated as (1 - specificity). The ROC can also be thought of as a plot of the power as a function of the Type I Error of the decision rule (when performance is calculated only from a sample of the population, it can be thought of as an estimator of this quantity). AUC accuracy performance can be classified into several groups, namely [7]:

1. 0.90 – 1.00 = *Excellent classification*
2. 0.80 – 0.90 = *Good Classification*
3. 0.70 – 0.80 = *Fair Classification*
4. 0.60 – 0.70 = *Poor Classification*
5. 0.50 – 0.60 = *Failure Classification*

III. Results and Discussions

1. Data Mining Process

In analyzing the performance of several classification models in the Orange tool, a comparison of several data mining methods was carried out to select the best method with high accuracy, in classifying the Pringsewu Muhammadiyah University graduation status dataset as shown in Figure 3.

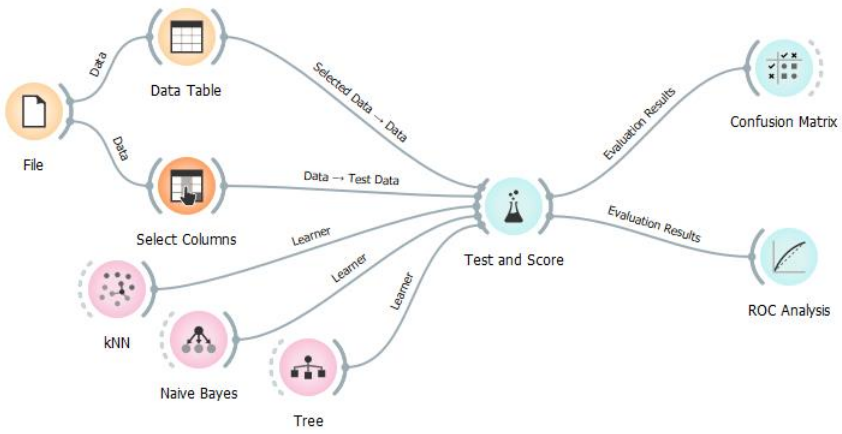

Figure 3. Widget design for student graduation status classification model

In Figure 3, a widget is designed using a classification model in Orange data mining software in the form of K-NN, Decision Tree and Naive Bayes which is input by a dataset that has been previously processed. Then the dataset is processed into classification mode.

2. Classification Model Testing Process

In the process of testing the classification model that has been created previously, a collection of test data is needed to determine the classification results as shown in Figure 4.

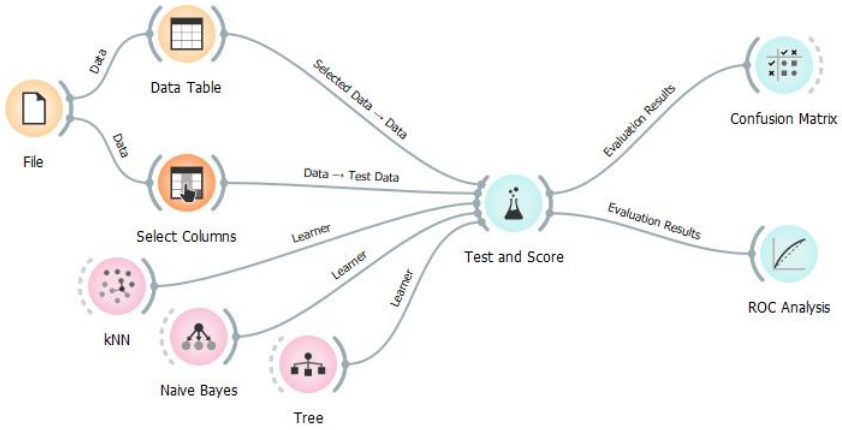

Figure 4. Widget design for student graduation status dataset classification model

In Figure 4 is a widget design that has been added to the classification testing process for the classification model. In the red box image is a set of trial data that is entered into the classification process to find out the results of the graduation classification of Muhammadiyah University Pringsewu students.

3. Evaluation Process of Classification Model Comparison Results

The next process is to carry out a classification model comparison process using Test and Score which is needed to calculate the success rate between each classification model in Orange data mining as shown in Figure 5.

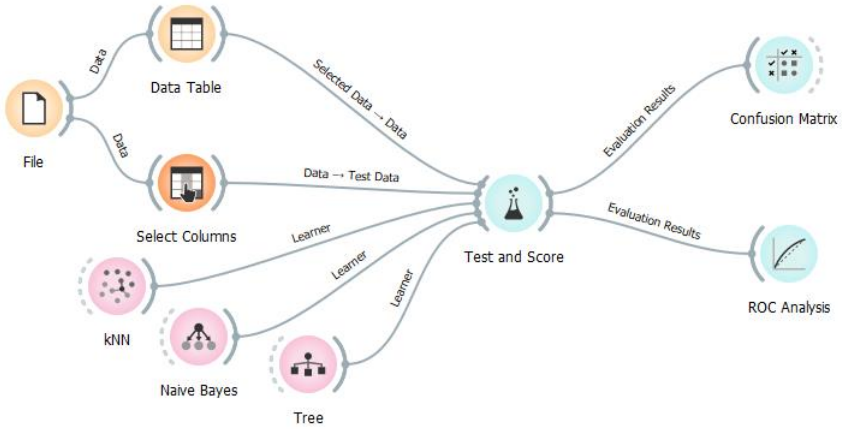

Figure 5. Widget design for calculating the success of the classification model

In Figure 5 is a widget design that has been added to the process of calculating the success rate of the classification model using the Test and Score widget, which will then be evaluated for accuracy using Confusion Matrix and ROC Analysis.

4. Simulation results of 3 classification models

The simulation results of the classification model were carried out using a test data set with 1 attribute as the target, 10 numeric attributes, namely NIM, Study Program, Gender, Marital Status, Employment Status, 3rd Sem GPA, 4th Sem GPA, 5th Sem GPA, Total SKS, Origin Students, Information, so that test score results are obtained as shown in Figure 6.

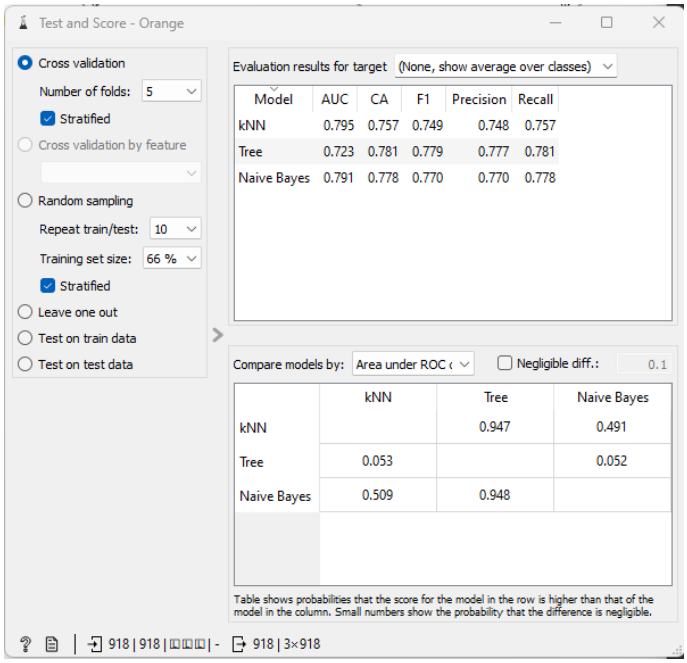

Figure 6. Test and score widget results

Based on student data that has been tested, the calculation results of precision, recall and accuracy for each model are obtained as shown in Figure 6. Model classification resultsK-NN, Decision Tree as well asNaive Bayes shows that the accuracy valueDecision Tree the highest is 78%. Based on Figure 6 which also shows a comparison of 3 AUC models, it is known that the highest AUC value is the methodK-NN namely 0.795. AUC is used to measure discriminatory performance by estimating the probability of output from randomly selected examples from a positive or negative population. The greater the AUC, the better the classification results used.

5. Evaluation Results with Confusion Matrix

Confusion Matrix is a performance measurement for machine learning classification problems where the output can be in the form of 2 or more classes. Confusion Matrix is a table with 4 different mixtures of predicted values and actual values. The evaluation results for each classification model can be seen in Figure 7 for the K-NN model, while the Confusion Matrix results for the Decision Tree model can be seen in Figure 8 and the Confusion Matrix values for the Naive Bayes model can be seen in Figure 9.

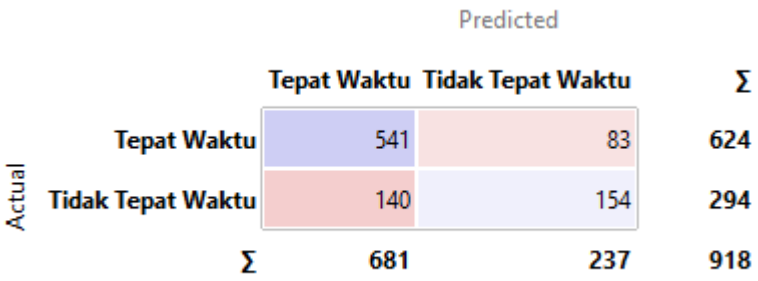

Figure 7. Confusion Matrix value of the K-NN method

Figure 7 shows that the value of True Positive (TP) is 541, True Negative (TN) is 154, False Positive (FP) is 83, and False Negative (FN) is 140. So the Accuracy, Precision and Recall values of the K method -NN is as follows:

$$Accuracy = \frac{541 + 154}{541 + 154 + 83 + 140} x100\% \qquad Maka\ nilai\ accuracy = 75\%$$

$$Precision = \frac{541}{541 + 140} x100\% \qquad Maka\ nilai\ Precision = 74\%$$

$$Recall = \frac{541}{541 + 83} x100\% \qquad Maka\ nilai\ Recall = 75\%$$

		Predicted		
		Tepat Waktu	**Tidak Tepat Waktu**	**Σ**
Actual	**Tepat Waktu**	535	89	**624**
	Tidak Tepat Waktu	112	182	**294**
	Σ	**647**	**271**	**918**

Figure 8. Confusion Matrix value for the Decision Tree method

Figure 8 shows that the value of True Positive (TP) is 535, True Negative (TN) is 182, False Positive (FP) is 89, and False Negative (FN) is 112. So the Accuracy, Precision and Recall values of the Decision method Tree is as follows:

$$Accuracy = \frac{535 + 182}{541 + 182 + 89 + 112} x100\% \qquad Maka\ nilai\ accuracy = 78\%$$

$$Precision = \frac{535}{535 + 112} x100\% \qquad Maka\ nilai\ Precision = 77\%$$

$$Recall = \frac{535}{535 + 89} x100\% \qquad Maka\ nilai\ Recall = 78\%$$

		Predicted		
		Tepat Waktu	**Tidak Tepat Waktu**	**Σ**
Actual	**Tepat Waktu**	552	72	**624**
	Tidak Tepat Waktu	132	162	**294**
	Σ	**684**	**234**	**918**

Figure 9. Confusion Matrix value of the Naive Bayes method

Figure 9 shows that the value of*True Positif (TP)* is 552, *True Negative (TN)* is 162*False Positive (FP)* is 72, and*False Negative (FN)* is 132. Then the value*Accuracy, Precision dan Recall* from the method*Naive Bayes* are as follows:

$$Accuracy = \frac{552 + 162}{552 + 162 + 72 + 132} x100\% \qquad Maka\ nilai\ accuracy = 77\%$$

$$Precision = \frac{552}{552 + 162} x100\% \qquad Maka\ nilai\ Precision = 77\%$$

$$Recall = \frac{552}{552 + 72} x100\% \qquad Maka\ nilai\ Recall = 77\%$$

Based on the results of evaluation and validation using*Confusion Matrix* comparative values are obtained*Accuracy, Precision dan Recall* of 3 methods*K-NN, Naive Bayes*, and *Decision Tree* as seen in Table 2.

Table 2. Performance Comparison

Metode	Accuracy	Precision	Recall
K-NN	75,7%	74,8%	75,7%
Naive Bayes	77,8%	77%	77,8%
Decision Tree	78,1%	77,7%	78,1%

Based on Table 2, it can be seen that the performance of the model*Naive Bayes* better than the model*K-NN* and*Decision Tree*. Classification accuracy cannot achieve perfect results because there must be error values. This is influenced by the amount of test data and training data used in the simulation process carried out.

6. Evaluation Results with ROC Curve

Manual accuracy values can be done by looking at the ROC curve comparison visualized from the Confusion Matrix. Model viewing ROC curves are the most easily visible way to graphically compare the accuracy values of each classification model. The graphical results of the ROC can be seen in Figures 10 and 11. Figure 10 shows that the ROC analysis results for student graduation at Muhammadiyah University of Pringsewu are CORRECT in each model as follows: (1)*K-NN* is 0.500, (2) *Naive Bayes* is 0.500, and (3) Decision Tree is 0.600. Therefore, for this case study, the model that has the best accuracy value is*Naive Bayes* and*K-NN* because the curve approaches the point 0.1.

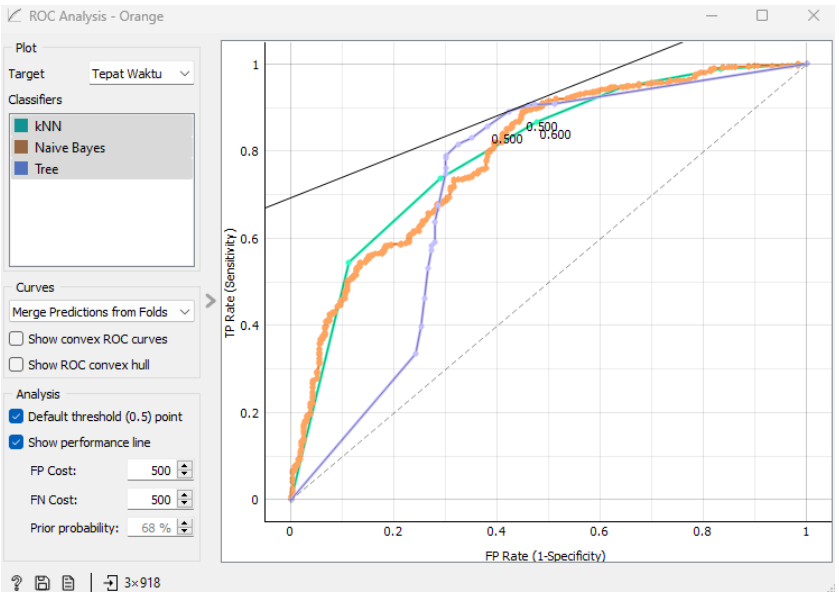

Figure 10. ROC analysis with the TRUE student graduation target

Figure 11 shows that the results of the ROC analysis of LATE graduation of Pringsewu Muhammadiyah University students in each classification model are as follows: (1)*K-NN* is 0.500, (2) *Naive Bayes* is 0.500, and (3) *Decision Tree* is 0.600. Therefore, classification research using 3 models with a study from Muhammadiyah University of Pringsewu is highly recommended using models*Naive Bayes* and*K-NN* because the curve approaches the point 0.1.

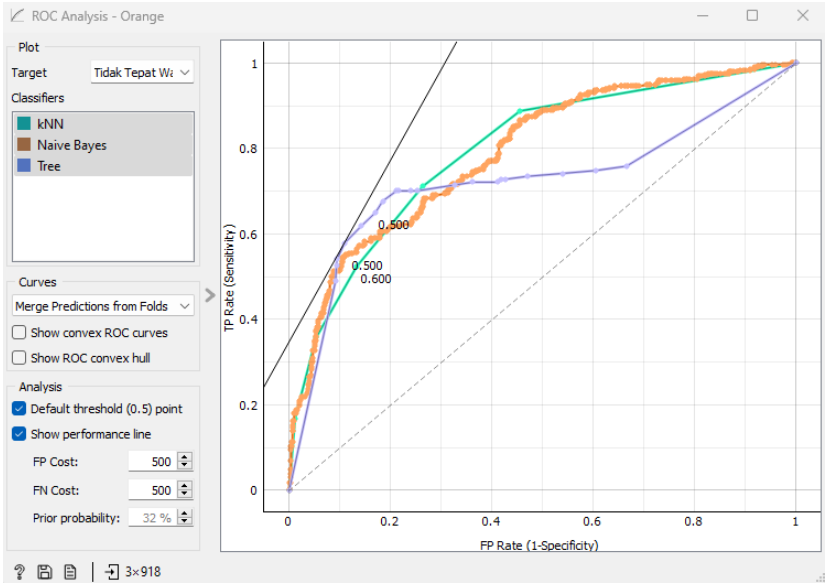

Figure 11. ROC analysis with the target of LATE student graduation

Based on the test results above, in this research the decision tree method has a slightly higher level of accuracy compared to the naive Bayes method. There are several analyzes that cause the decision tree method to have higher accuracy, including the following:

a. Decision Trees can work well on fairly large datasets without requiring excessive computing time. If the dataset is large enough, Decision Tree may be more efficient then Decision Tree provides an easy to interpret decision tree structure, which can assist in understanding the factors that most influence student graduation.
b. Naive Bayes assumes feature independence, and normality of distribution. If these assumptions are not fully met in the dataset, the performance of Naive Bayes can be affected. Decision Trees, in some cases, are more resistant to this assumption.

Future research in the context of predicting student graduation on time using data mining methods can explore a number of aspects to improve the accuracy and sustainability of the model, including exploring the use of deep learning methods such as neural networks to understand more complex and non-linear patterns in the data and then considering factors time in analysis, such as changes in student behavior over time, curriculum changes, or changing campus policies and can develop algorithms that can provide better explanations for model decisions, especially in the context of academic decisions that can have major implications.

IV. Conclusions

The resuts of this research show that after using the K-Nearest Neighbor, Decision Tree and Naive Bayes models to classify the graduation status of students at Muhammadiyah University of pringsewu, the results obtained were that Decision Tree's performance was superior to K-Nearest Neighbor and Naive Bayes. It is proven that the data used by Naive Bayes has an accuracy value of 77.8%, a precision of 77%, while K-Nearest Neighbor has an accuracy value of 75.7%, a precision of 74% and the Decision Tree has an accuracy value of 77.9% and a precision of 78%. The contribution of this research can be used by the management of Muhammadiyah University of pringsewu to detect early the condition of students so that their graduation is not too late and affect the accreditation score of Muhammadiyah University of Pringsewu.

References

[1] Alim, S. (2021a). Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes Orange Data Mining Implementation For Student Graduation Classification Using K-Nearest Neighbor, Decision Tree And Naive Bayes Models. In Jurnal Ilmiah Nero (Vol. 6, Issue 2).

[2] Amra, I. A. A., & Maghari, A. Y. A. (2017). Students performance prediction using KNN and Naïve Bayesian. *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, 909–913. doi: 10.1109/ICITECH.2017.8079967

[3] Annur, H. (2018). Klasifikasi Masyarakat Miskin Menggunakan Metode Naïve Bayes. In Agustus (Vol. 10, Issue 2).

[4] Berrar, D. (2019). Bayes' Theorem and Naive Bayes Classifier. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), Encyclopedia of Bioinformatics and Computational Biology (pp. 403–412). Oxford: Academic Press. doi: https://doi.org/10.1016/B978-0-12-809633-8.20473-1

[5] Caelen, O. (2017). *A Bayesian Interpretation of the Confusion Matrix*.

[6] Eko Prasetiyo Rohmawan. (2018). *Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Desicion Tree Dan Artificial Neural Networ*.

[7] Forsyth, D. (2018). *Probability and Statistics for Computer Science*.

[8] Hafizan, H., & Putri, A. N. (2020). *Penerapan Metode Klasifikasi Decision Tree Pada Status Gizi Balita Di Kabupaten Simalungun* (Vol. 1, Issue 2).

[9] Kartini, D., Nugroho, R. A., & Faisal, M. R. (2017). Klasifikasi Kelulusan Mahasiswa Menggunakan Algoritma Learning Vector Quantization. In Jurnal Positif (Vol. 3, Issue 2).

[10] Mikut, R., & Reischl, M. (2011). Data mining tools. *WIREs Data Mining and Knowledge Discovery*, *1*(5), 431–443. doi: https://doi.org/10.1002/widm.24

[11] Parteek Bhatia. (2019). *Data Mining and Data Warehousing*.

[12] Sari Dewi. (2016). Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan.

[13] Seref, B., & Bostanci, E. (2018). Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework. *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–7. doi: 10.1109/ISMSIT.2018.8567243

[14] Wati, E. F., & Rudianto, B. (2022). Universitas Bina Sarana Informatika 1 Teknik Informatika, Universitas Nusa Mandiri ,2 Jl. Kramat Raya No.98, Senen, Jakarta Pusat 10450 1 Jl. In Raya Jatiwaringin (Vol. 11, Issue 2). Teknik Dan Informatika.

[15] Wojtek J. Krzanowski, D. J. H. (2009). *ROC Curves for Continuous Data*.

[16] Kartini, K., Sujanto, B., & Mukhtar, M. (2017). The influence of organizational climate, transformational leadership, and work motivation on teacher job performance. IJHCM (International Journal of Human Capital Management), 1(01), 192-205.

[17] Galit Shmueli, P. C. B. I. Y. N. R. P. K. C. L. Jr. (2018). *Data Mining For Business Analytics*.