Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 4 No. 2 2024, 65 - 75
DOI: *https://doi.org/10.22487/27765660.2024.v4.i2.17158*
eISSN: 2776-5660

EXPLORATION OF STUDENTS INTERESTS IN MBKM AT RIAU UNIVERSITY USING A MACHINE LEARNING APPROACH

Nuraini Safitri[1*], Lathifah Zahra[2], Melanie Maria Lafina[3], Gustriza Erda[4], Anne Mudya Yolanda[5]

[1, 2, 3, 4, 5]Statistics Study Program, Riau University

[*]***e-mail:*** *nuraini.safitri2206@student.unri.ac.id*

ABSTRACT

This study aims to analyze the factors that have a significant influence on the interest of Riau University students in the Merdeka Belajar Kampus Merdeka (MBKM) program using a machine learning approach. MBKM is an innovation initiated by the Ministry of Education and Culture with the aim of improving student competence through its various programs. The Riau University as one of the universities supports this program by providing opportunities for its students to participate in various activities provided in the MBKM program. This study will specifically use a machine learning approach by utilizing several methods to analyze significant factors that have not been analyzed in depth by previous studies. The methods used in this analysis are logistic regression, decision trees, random forests, and naive bayes by utilizing secondary data on the level of interest of Riau University students to participate in the MBKM program in 2023. The variables used in this study include gender, generation, faculty, knowledge, self-confidence, feeling benefits, family support, friend support, lecturer support, self-ability, and facilities as independent variables and MBKM interest as a dependent variable. The results of the analysis of several methods show that the logistic regression method provides the best performance in modeling with an accuracy level of 95%. Variables that have a significant influence on students' interest in the MBKM program have also been successfully identified. The variables that have a significant effect are self-ability and family support. The development strategy of MBKM at the University of Riau can be optimized by paying attention to and focusing on these variables. The optimization of this strategy aims to make the implementation of the program more effective and efficient. Supportive policies such as workshops for the development of students' soft skills can be one of the strategic steps to improve students' abilities to the maximum.

Keywords: *Decision Tree, MBKM, Naive Bayes, Random Forest, Logistic Regression*

Cite: Safitri, N., Zahra, L., Lafina, M. M., Erda, G., & Yolanda, A. M. (2024). *Exploration of Students' Interests in MBKM at Riau University Using a Machine Learning Approach. Parameter: Journal of Statistics, 4(2), 65-75,* https://doi.org/10.22487/27765660.2024.v4.i2.17158.

[*]Received: June 2024; Reviewed: June 2024; Published: Dec 2024

INTRODUCTION

The Merdeka Belajar Kampus Merdeka (MBKM) is an innovation program initiated by the Ministry of Education and Culture and regulated in the Regulation of the Minister of Education and Culture Number 3 of 2020 concerning National Standards and Higher Education. This program is carried out with the concept of independence and freedom for students and universities in determining the best learning method to be used (Afida, 2021). According to Nadiem Makarim, MBKM will be an opportunity for students to improve their abilities through various activities that suit their interests. Students will also gain experience to be able to jump directly into the world of work so that students will be better prepared to face professional challenges. In addition, the development of the character of Pancasila students is one of the aspects focused on in this independent learning program (Kurniati, 2022).

The MBKM program aims to improve students' skills, knowledge, and experience by providing freedom and opportunities to study and carry out off-campus activities, such as internships. MBKM programs are designed to help students improve their self-competence, leadership skills, entrepreneurship, and international experience to increase student competitiveness (Nafisah, et al. 2023). As one of the State Universities, the University of Riau also participates in the implementation of this MBKM program. In line with the enthusiasm of the University's participation, the interest of University of Riau students in MBKM continues to increase from time to time. However, in understanding the factors that affect this interest, it is still necessary to understand and study more about the factors that affect this interest, in order to optimize the MBKM program in the future.

Previous research conducted by students of the University of Muhammadiyah Jakarta showed that the MBKM program still faces several obstacles such as lack of socialization and understanding of the MBKM program offered (Maulana, et al., 2022). Therefore, improvements and evaluations must still be made regarding the MBKM program policy itself. In addition, the dissemination of information about MBKM is generally obtained through social media in line with the progress of technological and information developments (Laga, et al., 2022). The University of Flores also conducted research on student interest in the MBKM program. The results of this study show that University of Flores students have a high level of interest in the MBKM program because it has a significant influence on improving self-competence.

The machine learning approach can be used to explore and analyze the interest of University of Riau students to participate in the MBKM program, especially by utilizing internal factors such as self-ability, self-confidence, and family support. Some of the methods that will be used in this approach are: logistic regression, decision tree, random forest, and naïve bayes. These methods will be able to identify hidden patterns and trends through traditional analysis so that their use will produce more comprehensive and accurate information about what variables are the main factors that students are interested in participating in the MBKM program. By identifying the main factors of student interest, the University of Riau will be able to optimize the strategy and increase the effectiveness of the MBKM program at the University of Riau.

The use *of decision trees* in this study is due to its interpretation that is easy to understand, and can be used for tabular data. (Albreiki et al., 2021) researched the decision tree method implemented to predict factors that affect student performance. The results of this study show that the accuracy of the decision tree results in a fairly high score in predicting student dropout based on several variables that are factors in student performance. Another study was also conducted by (Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, Mishra R, Pillai S, 2020) Iwendi, which discussed the prediction of the severity of COVID-19 patients. In this study, Iwendi used the random forest method which is believed to provide accurate prediction accuracy on an unbalanced dataset. According to Breiman (2001), Random Forest is a method that tests the best randomly selected subset of possible tests that can improve the accuracy of predicting the success of a study.

(Nurhidayati, M., & Hendayanti, N, P, 2020) have used logistic regression to research on "Binary Logistics Regression in Determining the Accuracy of Classification of Poverty Depth Levels in Provinces in Indonesia". The regression model and the resulting values will then be used to measure the proportion of variants contained in the independent variables and their bound variables. Binary logistic regression is used to model the relationship between the binary response variable (Y) (0.1) and the free variable (X) (Hosmer & Lemeshow, 2000). (Umar & M. Adnan Nur, 2022) have also conducted research to determine the performance of these approach methods, including the naïve Bayes method. In his research, the naïve bayes method was able to provide high accuracy results when implemented to analyze data (Arif, 2013). In the naïve bayes method, one of the methods is multinomial naïve bayes which is often used to predict the probability of the frequency of an event (Yuyun, et al., 2021).

The purpose of this study is to evaluate the effectiveness of the methods that have been used in subsequent studies in predicting the interest of University of Riau students to participate in the Merdeka Belajar Kampus Merdeka (MBKM) program. The machine learning approach will be used to identify factors that have a significant influence on the level of interest (Kamalia & Andriansyah, 2021). The results of this study are expected to provide insight for the University of Riau in evaluating factors that do not support the increase in student interest in participating in the MBKM program. In addition, the results of this study can also be used as a reference in formulating regulations or policies related to requirements or other matters related to the MBKM program at the University of Riau. Optimizing the MBKM program at the University of Riau can be done by considering factors that have a significant effect on increasing student interest. That way, the results of this research will be able to contribute to optimizing the implementation of the MBKM program at the University of Riau more effectively and relevantly.

MATERIAL AND METHOD

This research focuses on analyzing the interest of University of Riau (UNRI) students in the Merdeka Belajar Kampus Merdeka (MBKM) program using UNRI MBKM data in 2023. In this data there are 396 rows and 12 columns. The research variables analyzed included gender, generation, faculty, knowledge, self-confidence, feeling benefits, family support, friend support, lecturer support, self-ability, facilities, and MBKM interests. The target variable is the interest of MBKM which presents whether a student is interested or not interested in participating in the MBKM program. Furthermore, data analysis was carried out using a supervised learning approach, including logistic regression, decision trees, random forests, and naïve bayes. This approach allows the identification of the main factors that influence students' interest in the MBKM program with a high degree of accuracy and provides in-depth insight into the relationship between variables. The following are the stages of the data processing method using *Python*:

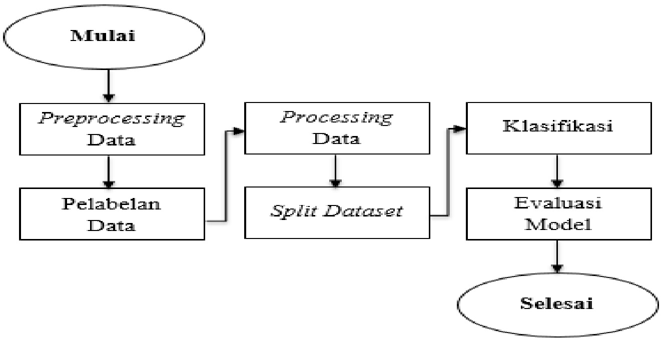

Figure 1. Data Processing Procedure

1. Classification

(a) Logistic Regression (RL)

According to Hosmer, et al. (2013), classification by logistic regression is carried out through a mathematical model that connects response variables with independent variables. Logistic regression is an analysis that uses quantitative independent variables to predict the probability of the occurrence of binary dependent variables (Dowdy, et al., 2004). The target variable in this method can use categorical or numeric variables. In binary logistic regression, the probability that an event will have two possibilities, namely 0 and 1. The purpose of the logistic regression model is to predict the probability of events of a target variable based on the value of the feature variable and to find out the relationship between the feature variable and the target variable (Larasati, 2003). Modeling data with logistic regression in python can use the syntax *library 'from sklearn.linear_model import LogisticRegression'* then *'sm_logit_train = sm. Logit'* for classification tasks. Logistic regression does not require the assumption of normality of the target variables. When the target variable is binary categorical, then the distribution is binomial.

The model of logistic regression is as follows:

$$P(Y = 1|X) = \text{or } P(Y=1|X)\frac{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)} = \frac{\exp(g(x))}{1+\exp\,(g(x))} \quad (1)$$

Information:

P(Y = 1|X) = Chance of an event
Exp = 2.71828183
g(x) = Logit function of the logistic regression model
Where $g(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

Logit-Regression *Test*

According to Hosmer and Lemeshow (2000), testing of model parameters can be carried out simultaneously and partially. Simultaneous testing of model parameters using a likelihood ratio test, with the following hypotheses:
H0: All regression coefficients are equal to zero ($\beta_1=\beta_2=...=\beta_k=0$).
H1: There is at least one non-zero regression coefficient ($\beta_i \neq 0$ for at least one i).
Test statistics: $G = -2\left[\frac{L_0}{L_p}\right]$ (2)
Information:
L0 is *the likelihood* of a model with only *intercepts* (null models).
L1 is the *likelihood* of a model with all predictors (full model).

The assumption H0 is correct, if $p - value > alpha$ (0.05) or $G^2 < x^2_{p(\alpha)}$will follow the chi squared distribution with the p free degree. If the value $p - value < alpha$ (0.05) or $G^2 > x^2_{p(\alpha)}$then the result is minus H0.

Wald Test

Testing of partial model parameters using the wald test, with the hypothesis:
H0: βi = 0 (There was no significant influence between variable X and variable Y (interest of MBKM UNRI students)).
H1: βi ≠0 (There is a significant influence between variable X and variable Y (interest of MBKM UNRI students))
Test statistics: $W_i = \frac{\widehat{\beta_i}}{\hat{S}E(\beta_i)}$ (3)

The H0 assumption is correct, if the Wald test statistics will follow the standard normal spread. If Wi > $alpha$ (0.05) or $p - value < \ alpha$ (0.05) , then the result is rejected H0

Odds Ratio *(OR)*

Next, calculate the odds ratio which is formulated by:
$\psi = e^{\beta_1}$ (4)

This aims to find out the feature variables that are significant to the target variables. If the value $\psi = 1$, then the variable is not significant to the target variable. If the value $\psi < 1$, then between the two variables there is a negative relationship. If the value $\psi > 1$, then the two variables have a positive relationship.

(b) Decision Tree (DT)

According to Rokach and Maimon (2008), a decision tree is a predictive model that can be used to represent classification and regression models in operations for strategic decision-making that are most likely to achieve goals. Essentially, a decision tree is a hierarchy of if/else questions that lead to a decision. The purpose of this method is to create a model that can be used so that it can predict the target value of the data that has not been seen with the decision rules obtained from the training data. The structure of the decision tree consists of 3 parts, namely the root node (the top node of the tree or a representative of all datasets), *the internal node* (the *node* that represents an attribute in the dataset), and the leaf node (the node that will provide the result of the target value prediction). To do data modeling with this method in python can use the syntax *library 'from sklearn.tree import DecisionTreeClassifier'* then *'clf = DecisionTreeClassifier'* for the classification task. The procedure for the *decision tree* method is as follows:

(1) Restrict or split data. It is used to decide how to divide the data on each internal *node* of the tree in order to find the best separation that minimizes the variance of the target variable within each node. The commonly used separation criteria for decision trees (classification) are '*gini'* and *'entropy'*.

(2) Calculate the *gain* value based on the '*gini'* and *'entropy'* values. Then, choose the best gain value (has the highest *gain* value) which will be the first root of the tree. The roots of the tree are taken from the selected attributes.

(3) Repeat steps 1 and 2 until certain criteria until the decision tree creation process cannot continue. There are 3 reasons why the decision tree cannot be continued, including:
- When a node contains only one class.
- The number of observations or attributes before data separation can no longer be defined.
- The depth of the tree has reached its maximum.

(c) Random Forest (RF)

The random forest method is an algorithm that is often used to classify large amounts of data because of its high level of accuracy and prediction, and based on the number of trees (Polamuri, 2017). According to Farnaazz and Jabbar (2016), the random forest method has a low error rate compared to other classification algorithms. This method aims to improve the accuracy of predictions and reduce overfitting that may occur in the decision tree. The random forest method will work by building multiple decision trees that work independently and then combining the prediction results from each tree to make a final prediction. To model data with this method in python can use the syntax *library 'from sklearn.ensemble import RandomForestClassifier'* then *'rf=RandomForestClassifier'* for classification tasks. In the random forest method, bootstrapping techniques are used to reduce *overfitting.* According to Breiman (2001) and Breiman & Culter (2003), the steps to carry out the random forest method include:

(1) Perform random sampling of n sizes with returns. This stage is the *bootstrap stage*.

(2) Using the *bootstrap sample*, the tree is built until it reaches its maximum size (no pruning). The construction of the tree is carried out by applying the random feature selection method in each selection process, namely k explanatory variables are randomly selected.

(3) Repeat steps 1 and 2, so that a forest consisting of several trees is formed.

(d) Multinomial Naive Bayes (MNB)

The multinomial naïve Bayes method is used in the classification of documents and texts developed from Bayesian algorithms. Multinomials take into account the number of times each word appears in a given document. According to Kurniawan et al. (2017), this method is a branch of the naïve bayes algorithm that takes into account the number of occurrences of words in documents. The purpose of this method is to predict the class of a data instance based on the probability distribution of the given features. Multinomial naïve bayes can be formulated with (Rahman and Doewes, 2017):

$$P(c) = \frac{N_c}{N} \tag{5}$$

Information:
P(c) = Initial chance of class c
Nc = Total class c on all documents
N = Total of all documents

Furthermore, calculating the probability that the nth word belongs to a certain class can be formulated by:

$$P(tn|c) = \frac{W_{ct}+1}{(\Sigma W'€VW'_{ct}+B'} \tag{6}$$

Information:
W_{ct} = Value *TF-IDF* in category C
$\Sigma W'€VW'_{ct}$ = Total *W* in category c
B' = Value *Idf* that is not multiplied *tf*

After that, calculate the probability of a document:

$$P(c|d) = P(c) \times P(t_1|c) \times P(t_2|c) \times \ldots \times P(t_n|c) \tag{7}$$

Information:
$P(c|d)$ = Probability of a class c document
$P(c)$ = Prior probability in class c

$P(t_n|c)$ = Probability of the nth word class c
t_n = Nth word in the document

2. Model Evaluation

The last stage is model evaluation which involves calculating accuracy by comparing data before optimization and after optimization to produce the best model performance. By using a confusion matrix which has 4 components, including true positive (the prediction model is indeed positive), true negative (the prediction model is indeed negative), false positive (the prediction model is positive but should be negative), and false negative (the prediction model is negative but should be positive). According to Makhmud (2019), the accuracy value is calculated by the formula:

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{8}$$

Information:
TP = True positive
TN = True negative
FP = False positive
FN = False negative

At this stage, variables that affect the target variables are also checked. One way to check variables affects *feature importance* (for random forest models). In the decision tree model, determining the influencing variables can be seen in the decision tree that has been formed. For the logistic regression model, the determination of the influencing variable can be obtained by performing a logit-regression test, wald test, and odds ratio test.

RESULTS AND DISCUSSION

1. Preprocessing and Processing Data

The first step before processing data is to import the data into the application that will be used in processing the data. This research uses *python* as an application. The dataset input is named 'df'. Then, the next stage is data cleaning by checking for missing values or values that are likely to be missing in the dataset. This aims to be a step in the analysis carried out so that the data is more accurate and ready to be used for the next step. From the output cleaning data, it is stated that there is no missing value in the dataset.

The second stage is labeling or categorizing variables that are not numerical. One of the variables that is categorized is the MBKM knowledge variable for students who have knowledge of the MBKM program are labeled 1, while for those who do not have knowledge of the MBKM program are labeled 0. The third stage is data *processing*. For this stage, the target variable and the feature variable will be defined. The target variable in this study is MBKM Interest while the rest are feature variables.

The next stage is the separation of data into 2 parts, namely data *training* and data *test*. Before splitting the dataset, make sure that he *'from sklearn.model_selection import train_test_split' library* is used. Test size is the amount of *test* data that will be used in processing data. In this study, the number of test data was 0.2 or 20%, while the rest (0.8 or 80%) was training data. After that, perform the data separation process so that the model's performance can be measured objectively.

2. Classification

Table 1. Report on Logistics Regression Model

Featured	*Coefficient (P-value)*	*Featured*	*Coefficient (P-value)*
Gender	-0.0465 (0.949)	Family Support	1.4325 (0.036)
Force	0.1880 (0.439)	Friend Support	1.3536 (0.031)
Faculty	0.2072 (0.092)	Lecturer Support	0.3427 (0.606)
MBKM Knowledge	-0.8002 (0.557)	Self-Abilities	2.0847 (0.006)
Self-Confidence	2.6811 (0.004)	Facilities	-0.1665 (0.899)
Experiencing the Benefits of the Program	-1.7149 (0.305)		

Based on the stages of applying machine learning for the analysis of MBKM interest of University of Riau students using Logistic Regression analysis, Decision Tree, Random Forest, and Multinomial Naive Bayes, the accuracy percentage of each model is obtained, including:

a. Logistic Regression (RL)

By using *python, a* logistic regression model is obtained, as follows:

P(Y=y) = $\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_{11}x_{11})}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots+\beta_{11}x_{11})}$

P(Y=1)= $\frac{exp(-2.3898-0.0465\beta_1+0.1880\beta_2+0.2072\beta_3-\ldots-0.1665\beta_{11}}{1+exp(-2.3898-0.0465\beta_1+0.1880\beta_2+0.2072\beta_3-\ldots-0.1665\beta_{11}}$

P(Y=0) = 1- $\frac{exp(-2.3898-0.0465\beta_1+0.1880\beta_2+0.2072\beta_3-\ldots-0.1665\beta_{11}}{1+exp(-2.3898-0.0465\beta_1+0.1880\beta_2+0.2072\beta_3-\ldots-0.1665\beta_{11}}$

Based on the analysis of the logistics regression algorithm, the value G^2 is 3.354^{-13}. From the test statistics, due to the $G^2 < alpha$ (0.05), the decision was to reject H_0 (accept H_1). That is, at a significant level of 5%, sufficient evidence is obtained to state that there is at least one $\beta i \neq 0$ with i = gender, generation, faculty, knowledge, self-confidence, feeling benefits, family support, friend support, lecturer support, self-ability, and facilities.

If the value $G^2 > alpha$ (0.05), then the decision is not to reject H0 (accept H0) or to obtain sufficient evidence to state that there is a significant difference between the feature variable and the target variable.

Based on the wald test, the tests of the variables of self-confidence, family support, friend support, and self-ability resulted in a *p-value* of < *alpha* (0.05). The result obtained was to reject H0 (accept H1). That is, at the significance level of 5%, sufficient evidence is obtained to state that there is a significant influence between feature variables on target variables. Meanwhile, in the test of the remaining variables, the *p-value* > *alpha* (0.05) was produced. This means that at the significant level of 5%, there is not enough evidence to state that there is a significant influence between the variables tested.

Table 2. Odds Ratio Logistic Regression Model

Featured	***Odds Ratio***	***Featured***	***Odds Ratio***
Gender	1.0392	Family Support	4.7799
Force	0.9956	Friend Support	2.5856
Faculty	1.1547	Lecturer Support	1.8243
MBKM Knowledge	2.2240	Self-Abilities	4.2416
Self-Confidence	3.8707	Facilities	0.4927
Experiencing the Benefits of the Program	0.2573		

In the OR test table above, it was obtained that the odds ratio value of the variables gender, faculty, MBKM knowledge, self-confidence, family support, friend support, lecturer support, and self-ability was more than 1. This means that these variables tend to have a greater influence on the target variable. The highest observation value was found in the family support variable of 4.78, while the lowest observation value was owned by the variable of feeling the benefits of the program at 0.26.

Table 3. Report on Logistics Regression Model

Accuracy	*Precision*	*Recall*
0.95	0.95	1.00

From the output above, the classification accuracy value is obtained by 95% and the classification error in this model is 5%. This means that the Logistic Regression model is categorized as a good classification because it has a very high accuracy value. This model has excellent performance in classifying data.

b. *Decision Tree* (*DT*)

Table 4. Report Model Decision Tree

Report	**Before Optimization**	**After Optimization**
Accuracy	0.95	0.94
Precision	0.95	0.96
Recall	1.00	0.97

Based on the above accuracy output, the initial accuracy shows a very high accuracy level of 95%. After optimization, the accuracy percentage becomes 94%. Despite a 1% drop from the initial accuracy percentage, the model still belongs to the category of very high accuracy levels. There are several factors that cause a decrease in accuracy after optimization, including due to *overfitting* in the model before optimization or suboptimal parameter selection (Patlisan, P., & Rusdah, 2023). If the model is too suitable for *training data*, it can cause poor model performance in the test data or also known as *overfitting*.

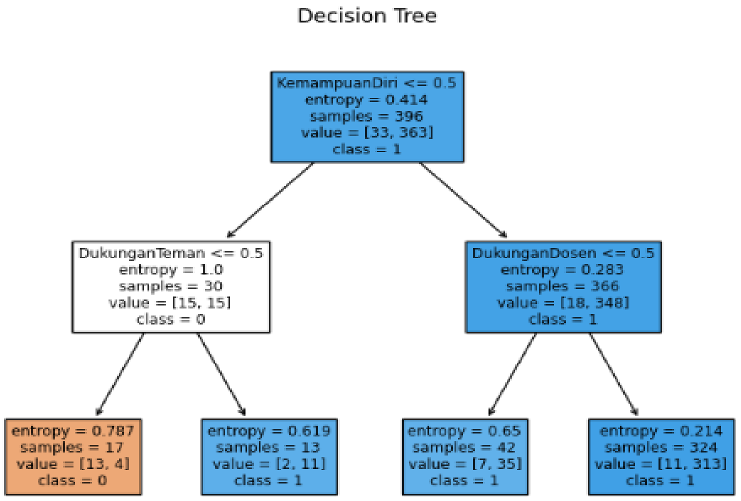

Figure 2. Decision Tree Structure

Based on the decision tree structure above, it was obtained that the variables that affected the target variables consisted of variables of self-ability, peer support, and lecturer support. It can be concluded that, from all the variables included in the modeling, it is obtained that the factor that most affects students' interest in participating in MBKM is the self-ability possessed by the students themselves.

c. Random Forest (RF)

Table 4. Report Model Random Forest

Report	**Before Optimization**	**After Optimization**
Accuracy	0.90	0.89
Precision	0.92	0.92
Recall	0.97	0.96

From the accuracy table before and after optimization on the *RL* model, it is obtained that the initial accuracy is 90%. This means that the level of accuracy in this model is categorized as very high so that the model has good performance in classifying existing data. However, after optimization on *the RL* model, there was a decrease of 1%. Despite the decline, the accuracy level of the model is still relatively high.

Table 5. *Feature Importance* Model *Random Forest*

Featured	***Importance***	***Featured***	***Importance***
Faculty	0.2069	Lecturer Support	0.0489
Force	0.1771	MBKM Knowledge	0.0304
Self-Abilities	0.1407	Facilities	0.0158
Family Support	0.1154	Experiencing the Benefits of the Program	0.0137
Self-Confidence	0.1228		
Friend Support	0.0889		
Gender	0.0495		

To determine the variables that affect prediction or accuracy, it can be seen from the importance value of each variable. In the *RF* model, if the *importance* value of a variable is x > *alpha* (0.05), then the variable has an effect in processing the MBKM Interest data. In the *output* above, the highest importance value is owned by the faculty variable, which is 0.21. This means that these variables are very influential in processing MBKM Interest data. The next highest importance value was followed by the variables of generation (0.18), self-ability (0.14), family support (0.12), self-confidence (0.11), and friend support (0.08). While the rest consisting of 5 variables (gender, lecturer support, MBKM knowledge, facilities, and experiencing the benefits of the program) had no effect on the MBKM Interest variable. The variable that has no effect on the target variable is the variable that feels the benefits of the program with *an importance* value of 0.01.

d. ***Multinomial Naive Bayes* (*MNB*)**

Table 6. Report Multinomial Model Naive Bayes

Report	**Before Optimization**	**After Optimization**
Accuracy	0.91	0.91
Precision	0.83	0.83
Recall	0.91	0.91

From the *accuracy output* of the *MNB model*, the accuracy before optimization was obtained of 91%. The accuracy is categorized as a very high level of accuracy. After optimization is carried out on the model, the accuracy remains the same value as the initial accuracy. If the accuracy level of a model is high enough and has the same value, then the model's performance is good in processing the tested data and the model before optimization is quite optimal (Wilindia, A. S., Dasuki, M., & Fitriyah, 2021). However, this does not rule out the possibility of *underfitting* or the data used is not complex.

Table 7. Feature Importance Model Multinomial Naive Bayes

Featured	***Importance***	***Featured***	***Importance***
Family Support	0.7384	Force	-02080
Friend Support	0.7344	Facilities	-0.2147
Self-Abilities	0.3633	Experiencing the Benefits of the Program	-0.2215
Self-Confidence	0.2460	Gender	-0.2888
Lecturer Support	0.1959	MBKM Knowledge	-0.2902
Faculty	0.0039		

To determine the variables that affect the target variables, *the MNB* model uses '*SelectBest*' in order to be able to select the best variable based on *chi-squared*. In this model, calculating the *importance* value aims to be the difference in probability logs between classes. The probability log value in *the MNB* model describes how much the variable has an effect. So, it has nothing to do with *alpha values*. From the *output* above, it is known that the most influential variable is the family support variable of 0.74. Meanwhile, the variable that has the least influence is the MBKM knowledge variable with *an importance value* of -0.29.

3. Model Evaluation Comparison

Table 8. Report Every Model

Type	**Accuracy**	***Precision***	***Recall***
Logistic Regression	95%	95%	100%
Decision Tree	94%	96%	97%
Random Forest	89%	92%	96%
Multinomial Naive Bayes	91%	83%	91%

In this analysis, four models were used, namely logistic regression, decision tree, random forest, and multinomial naïve bayes. Of the four models produced, the logistic regression model is the best model with very high accuracy, precision, and *recall* values (Sembiring, 2023). High accuracy, precision, and *recall* values ensure that no data is missed in the analysis. Then in second place is the decision tree model which has accuracy, precision, and recall values that tend to be high and stable even though they are not as high as logistic regression (Tanujaya, L. B. C., Susanto, B., & Saragih, 2020). The random forest and multinomial naïve bayes models are models that have lower performance when compared to logistical regression models and decision trees. The use of different models in the analysis will produce different results as well as in determining the variables that affect determining the interest of UNRI students in participating in the MBKM program.

- In logistic regression, the variables that affect are: self-confidence, family support, friend support, and self-ability.
- In the decision tree, the variables that affect are: self-ability, peer support, and lecturer support.
- In random forests, the influencing variables were: faculty, generation, self-ability, family support, self-confidence, and friend support.
- In *the* multinomial naïve bayes, the most influential variable is family support. Meanwhile, the variable whose influence is small or insignificant is the MBKM knowledge variable.

From each model used, different results will appear. But there are variables that always appear in every model used, these variables are the variables of self-ability and peer support. This shows that these two variables are the main factors for UNRI students to be interested in participating in MBKM activities.

CONCLUSION

In research conducted using logistic regression, decision tree, random forest, and multinomial naïve bayes, it was found that University of Riau students have an interest in participating in the MBKM program. Using the 95% accuracy level produced by the logistics regression model, it can be concluded that almost all students of the University of Riau have an interest and interest in participating in this program. In addition, from this analysis, it was obtained that there are two variables that greatly influence the determination of students' interest in participating in the MBKM program, namely the variables of self-ability and peer support. The logistic regression model is the model that has the best performance compared to other models used, so in this study, the logistic regression model is appropriate and accurate to predict students' interest in participating in the MBKM program.

REFERENCES

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student’s Performance Prediction Using Machine Learning Techniques. *Education Sciences*, *11*(9).

Arif, N. (2013). Comparison of Logistic Regression Model and Radial Model Basis Function Neural Network for Classification of Binary Response Variables. *Doctoral Dissertation, Universitas Brawijaya*.

Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, Mishra R, Pillai S, J. O. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Front Public Health*, *8*.

Kamalia, P. U., & Andriansyah, E. H. (2021). Independent Learning-Independent Campus (MBKM) in Students’ Perception. *Jurnal Kependidikan: Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan, Pengajaran Dan Pembelajaran,* *7*(4), 857. https://doi.org/10.33394/jk.v7i4.4031

Larasati, N. (2003). Comparison of Logistic Regression and Random Forest in the Weather Classification of the Central Java Region. *AXIOMS: Journal of Mathematics and Mathematics Education*, *14*(2), 172–181.

Nurhidayati, M., & Hendayanti, N, P, N. (2020). Binary Logistics Regression in Determining Accuracy in Classification of Poverty Depth Levels in Provinces in Indonesia. *AMSET IAIN Batusangkar and IAIN Batusangkar Press*, *12*(2), 63–70.

Patlisan, P., & Rusdah, R. (2023). Decision Tree Model Accuracy Optimization Using Random Forest Regression to Predict the Purchase Quantity of Goods in Manufacturing Companies. *Symmetrical: Journal of Mechanical Engineering, Electrical and Computer Science*, *14*(2), 217–228.

Sembiring, P. (2023). Logistic regression analysis to determine the factors that affect the welfare of the regency/city community on Nias Island. *FARABI: Journal of Mathematics and Mathematics Education*, *6*(1), 25–31.

Tanujaya, L. B. C., Susanto, B., & Saragih, A. (2020). The comparison of logistic regression methods and random forest for spotify audio mode featurre classification. *Indonesian Journal of Data and Science*, *1*(3), 63–78.

Umar, N., & M. Adnan Nur. (2022). Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(4), 585–590. https://doi.org/10.29207/resti.v6i4.4179

Wilindia, A. S., Dasuki, M., & Fitriyah, N. Q. (2021). Implementation of Naïve Bayes’ Multinomial Algorithm for Twitter Sentiment Analysis on the Policy of Freedom of Learning. *Journal of Smart Technology*, *1*(1), 100–102.