ISSN Online 3025-1605
Journal of Statistical Methods and Data Science
Volume 01 Number 2 December 2023
https://jurnalmipa.unri.ac.id/jsmds

IMPLEMENTATION OF THE RANDOM FOREST METHOD FOR PREDICTING STUDENTS' LENGTH OF STUDY

Ali Akbar[1], Zul Indra[2], Yanti Andriyani[3], Tisha Melia[4*]
[1,2,3,4]Computer Science Department, Universitas Riau, Riau, 28293, Indonesia

***e-mail**: tisha.melia@lecturer.unri.ac.id

Article Info:
Received: 20-03-2024
Accepted: 23-03-2024
Available Online: 05-04-2024

Keywords:
cross validation
duration of study
feature importance
random forest

Abstract: *Predicting a student's duration of study is essential for universities to ensure students complete their studies on time. This research aims to develop an effective prediction model for determining the length of study based on related factors. To overcome the complexity and diversity of student data, the Random Forest method was chosen. The results indicate that the Random Forest method is an effective tool for predicting the duration of study for university students. A study was conducted on 1,535 graduates from the five departments at the Faculty of Mathematics and Natural Sciences, Riau University. The study employed cross-validation techniques to measure model performance. The model's accuracy was evaluated using a confusion matrix, which revealed that the Random Forest model had an average accuracy of 95.12%. Additionally, feature importance analyses identified grade point average in the eighth semester as a major contributor to the prediction outcome.*

1. INTRODUCTION

The duration required for a student to complete their educational program is referred to as the student's study period. The Ministry of Education and Culture has regulated the study period through the Directorate General of Higher Education, which stipulates that undergraduate students (S1) must complete a mandatory load of 144-160 semester credit units (SKS) within 8-10 semesters or approximately 4-5 years. Considering the mandatory workload that must be completed within a certain timeframe, an approach is needed that can provide accurate predictive results to assist educational institutions in monitoring and improving student study efficiency.

2. LITERATURE REVIEW

2.1. Study duration

Study duration refers to the time required for an individual to complete a specific program or level of education. It is one of the factors evaluated in higher education accreditation. The duration of undergraduate studies is generally four years, and students are expected to complete their education within this time frame to meet the standards of graduate competency (Hasan et al., 2022). The Ministry of Education and Culture has regulated the study duration. To meet the graduation competency standards, students must complete a minimum of 144-160 credit points over 8-10 semesters or 4-5 years. Therefore, the duration of a student's study is an important consideration for the university's study program (Endang Etriyanti, 2021).

2.2. Random Forest

Random forest is a combination of the Bagging and Random Subspaces methods. In recent years, this method has proven successful in solving regression and classification problems, and has become one of the best machine learning algorithms used in various fields (Apriliah et al., 2021). According to Adrian, Random Forest is a combination of various decision tree techniques that are merged into a single model. The decision trees that have been formed will make a decision, and the final decision will be determined based on the results of the majority voting (Adrian et al., 2021).

2.3. Data Mining

Data mining is the process of collecting, extracting, and analysing data to discover patterns and relationships that are hidden within it. Data mining is the process of collecting, extracting, and analysing data to discover patterns and relationships that are hidden within it. It can be used for various purposes. According to Asana et al. (2022), data mining is the process of identifying and extracting useful information and knowledge from large datasets using statistical, mathematical, artificial intelligence and machine learning techniques.

2.4. Confusion Matrix

The confusion matrix, also known as the error matrix, is a method used to calculate accuracy in the process of classification or supervised learning (Purbolaksono et al., 2021). It is commonly used to evaluate the performance of a classification model. According to Marutho (2019), the confusion matrix consists of four combinations of predicted and actual values, as depicted in Table 1.

Table 1. Confusion Matrix

Predict	**Actual**	
	Positive	**Negative**
Positive	TP	FP
Negative	FN	TN

true positive (TP) refers to data that is predicted to be positive and is actually positive, true negative (TN) refers to data that is predicted to be negative and is actually negative, false positive (FP) refers to data that is predicted to be positive but is actually negative and false negative (FN) refers to data that is predicted to be negative but is actually positive.

Below is the formula for calculating the accuracy, specificity, and sensitivity of a classification model using a confusion matrix:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$Specificity = (TN)/(TN + FP) \quad (2)$$

$$Sensitivity = (TP)/(TP + FN) \quad (3)$$

2.5. Model Evaluation

Model evaluation is a critical step in developing prediction or classification models. It involves assessing the extent to which the built model is capable of providing reliable and accurate results when applied to new or unseen data. One commonly used evaluation method is K-fold cross-validation.

K-Fold Cross Validation is a step to validate the accuracy of predictions. K-Fold Cross Validation is a step to validate the accuracy of predictions. It is a model validation method used

to evaluate the accuracy of analysis results. In this method, preprocessed data is divided into training and testing subsets used for the classification process (Windarto et al., 2021). The main concept of K-Fold Cross-Validation involves dividing the dataset into 'K' subsets (usually 'K' is an integer, such as 5 or 10), where one subset is used as the test data and the rest as the training data. This is done in cycles, where each subset takes turns being the test data while the others are the training data. Based on the 10-fold cross-validation, it can be concluded that the dataset was divided into 10 folds, each containing 1 part for testing and 9 parts for training. The total number of tests was 10, and the results can be seen in Figure 1.

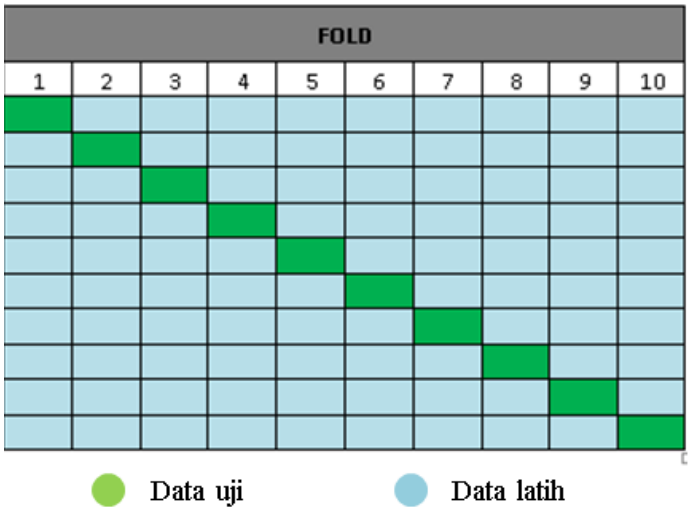

Figure 1. Ilustration of 10-Fold Cross Validation

3. METHODOLOGY

3.1. Data and Sources

The data used for this study consists of information on students who have graduated from the Computer Science, Mathematics, Chemistry, Physics, and Biology undergraduate programmes at the Faculty of Mathematics and Natural Sciences (FMIPA), University of Riau. The data set contains 14 variables that are used as attributes for the analysis. The data set includes a total of 1,535 students who graduated between 2014 and 2016. The data set contains 14 variables that are used as attributes for the analysis.

3.2. Data Analysis Method

This process is divided into two stages: data training and data testing. The method used to create the Random Forest model involved K-fold cross validation. The model was created using K-fold cross validation twice. The first model used 10-fold cross validation, while the second model used cross validation based on the division of departments in FMIPA.

The evaluation of the Random Forest model involves the use of the confusion matrix method. This step evaluates the model using a confusion matrix table, which provides accuracy, specificity, and sensitivity rates for the 10 tested sections calculated based on equations 1, 2, and 3.

3.3. Analysis Stage

i. Problem Identification Technique
The initial process of recognizing a problem is problem identification. In this study, the problem is the need for predicting the duration of a student's study to help improve educational institution efficiency.

ii. Data Collection
This study requires data that includes information on the length of study, attributes, and factors that affect the length of study taken from the relevant Faculty and Study Program at the University of Riau.

iii. Prepocessing Data
The obtained data underwent a cleaning process to remove missing values, outliers, or input errors. Additionally, data transformation was performed if necessary

iv. Random Forest Model Creation.
The next step is to implement the Random Forest algorithm by building decision trees using a subset of the training data.

v. Prediction Using Random Forest.
Then, use the trained Random Forest model to predict the duration of study for students in the testing data.

vi. Evaluate The Model.
The performance of the Random Forest model was evaluated using a Confusion Matrix. The predicted results were compared with the actual values of the length of study of students in the testing dataset. The variables that have the most significant impact on the length of study of students in FMIPA, Universitas Riau were then determined. In addition, researchers also sought to determine the accuracy, specificity, and sensitivity values of the resulting metrics.

vii. Visualization of The Result
The predicted results are visualised by displaying data in a plot based on the variables used, using the Shiny package in RStudio.

4. RESULTS AND DISCUSSION

4.1 Preprocessing Data

Data cleaning was performed manually by removing defective data or data without attributes. Meanwhile, filtering of unused data variables was done using a code program. For example, in this study, the variable NIM was not used, so the code program was 'data$NIM <- NULL'. The necessary variables will be extracted and generalized if necessary, while the unnecessary variables will be deleted. This process resulted in test-ready data with a total of 1,333. For numerical data such as the number of credits taken (SKS), IP, and cumulative IP (IPK), researchers seek the maximum, minimum, and mean values as shown in Table 2.

Table 2. Variable Data

Variable	Types	Min	Max	Average	Values
Economic Status	Categorical	-	-	-	High and Low
Scholarship	Categorical	-	-	-	Yes, No
Grade Point Average (GPA)	Numerical	0	3.93	2.79	-
Credits Completed	Numerical	0	159	135.10	-

Variable	Types	Min	Max	Average	Values
High School Concentration	Categorical	-	-	-	High School – Natural Science, High School – Social Science, Vocational High School, and others
Semester 1 GPA (GPA1)	Numerical	0	4	2.98	-
Semester 2 GPA (GPA2)	Numerical	0	3.97	2.89	-
Semester 3 GPA (GPA3)	Numerical	0	4	2.75	-
Semester 4 GPA (GPA4)	Numerical	0	4	2.71	-
Semester 5 GPA (GPA5)	Numerical	0	4	2.74	-
Semester 6 GPA (GPA6)	Numerical	0	4	2.88	-
Semester 7 GPA (GPA7)	Numerical	0	4	2.81	-
Semester 8 GPA (GPA8)	Numerical	0	4	1.97	-
Category of High School Concentration	Categorical	-	-	-	Science, Not Science

4.2 Random Forest Model Creation

The model was created using the K-fold Cross Validation method to minimize overfitting. The K-fold cross-validation method was used twice to create the model. The process of creating a Random Forest model is divided into two stages: data training and data testing. The first model used 10-fold cross-validation, while the second model used cross-validation based on the division of departments in FMIPA.

i. 10-Fold Cross Validation

10-fold cross-validation is a cross-validation method that divides data into 10 equal parts. The choice of 10 in data division is considered to provide a more accurate and efficient estimate of accuracy than other data division methods (Hestie et al, 2001). Each part is then used as test data to test the model trained on the other 9 parts. If section 1 is chosen as the test data, sections 2 to 10 will be used as training data. The same applies if section 2 is chosen as the test data and so on.

ii. Cross validation Based on Majors

Similar to the first method, data is divided into several parts. In this case, the data is divided into 5 parts, each corresponding to the available departments. Each major's data will be the test data while the other parts will be the training data.

Each experiment used either 50 or 100 decision trees. The number of decision trees was chosen based on considerations of efficiency, effectiveness, and accuracy. The optimal number of decision trees was determined through experimentation, as shown in Figure 2.

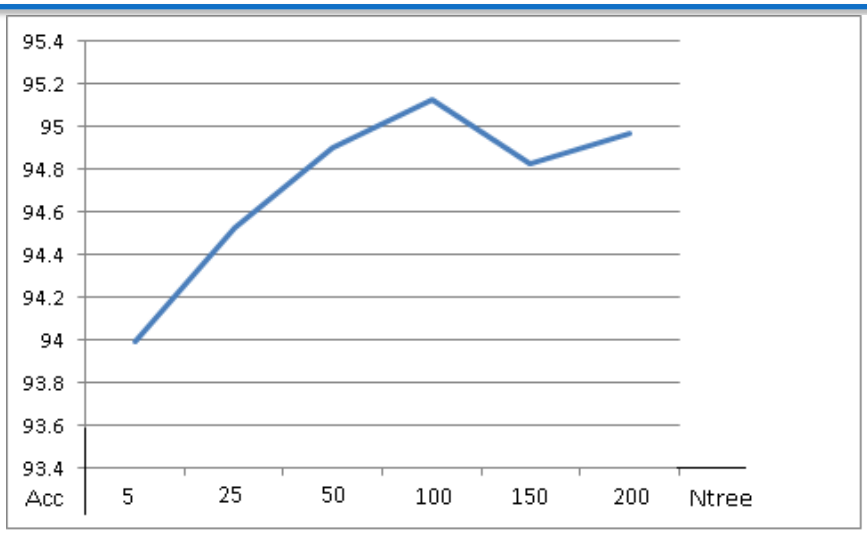

Figure 2. Accuracy level of the number of *ntree*

From Figure 2, it can be seen that the highest level of accuracy was achieved using 100 decision trees with an average accuracy of 95.12%.

4.3 Prediction Using Random Forest

After creating the random forest model, it was tested using the prepared data. Predictions were made using the two methods previously described. The prediction results are from the test data of the 10-fold cross-validation experiment with 50 ntree, as shown in Table 3 (not all columns and rows are shown).

Table 3. Predicted Data Results

No	Economic Status	Scholarship	GPA	High School Concentration	GPA 1	…	GPA 8	Category of High School Concentration	Class Label	Prediction
1	High	No	3.29	High School Science	3.29	…	4.00	Science	Untimely	Timely
2	High	No	3.59	High School Science	3.65	…	0	Science	Untimely	Untimely
3	Low	No	2.57	Vocational High School	2.38	…	3.45	Science	Untimely	Untimely
4	High	No	2.70	High School Science	2.53	…	2.75	Science	Untimely	Untimely
5	High	No	3.23	High School Science	3.1	…	0	Science	Untimely	Untimely
6	High	No	3.63	High School Science	3.72	…	4.00	Science	Timely	Timely
7	High	No	2.22	Other	2.29	…	2.36	Not Science	Untimely	Untimely
…	…	…	…	…	…	…	…	…	…	…
137	High	No	2.81	High School Science	2.41	…	3.19	Science	Untimely	Untimely

The complete data can be viewed on the website https://my.unri.ac.id/vPGx6F.

4.4 Variables Contributing to Prediction

Each variable in the dataset has a contribution or influence on the prediction outcome. The level of contribution of each variable can be seen through the 'importance' feature. The 'importance' feature in the randomForest library in R is used to measure how important each feature is in the decision tree classification model. This feature can be used to understand the

model and identify the most influential features on the classification results. In this study, the contribution level of each variable to the prediction results can be seen in Table 4.

Table 4. Contribution Levels of Each Variable

Variable	Contribution Level
GPA8	0.383876312
GPA	0.088975591
GPA4	0.062535408
GPA5	0.039795558
GPA3	0.035884341
GPA2	0.033053905
GPA6	0.022142244
GPA7	0.021496254
GPA1	0.017362073
High School Concentration	0.010758068
Credits Completed	0.005653797
Economic Status	0.000567515
Scholarship	0.000447151
Category of High School Concentration	0.000274495

Table 4 presents the feature importance data of the random forest model used to predict the duration of a student's study. According to the table, IPS8 (semester 8 grade point average) is the most important feature in the classification model, with a MeanDecreaseGini value of 0.3838763118. This indicates that IPS8 plays a crucial role in influencing the prediction results.

4.5 Model Evaluation

At this stage, the model will be evaluated using a confusion matrix. The results will include accuracy, specificity, and sensitivity of the 10 tested parts, calculated based on equations 1, 2, and 3. The evaluation results using 10-fold cross-validation with ntree values of 50 and 100 can be seen in Tables 5 and 6.

Table 5. 10-fold cross validation with ntree value of 50

Ntree 50	Accuracy	Specificity	Sensitivity
Fold-1	94.73	94.38	95.45
Fold-2	96.24	97.00	93.93
Fold-3	94.73	96.55	91.30
Fold-4	97.74	98.96	94.44
Fold-5	92.48	96.15	87.27
Fold-6	95.48	94.73	97.36
Fold-7	95.48	95.04	96.87
Fold-8	94.73	94.5	95.23
Fold-9	94.73	94.73	94.73
Fold-10	92.64	94.25	89.79
Average	94.89	95.62	93.63
SD	1.54	1.52	3.21

Table 6. 10-fold cross validation with ntree value 100

Ntree 100	Accuracy	Specificity	Sensitivity
Fold-1	94.73	94.38	95.45
Fold-2	96.99	97.00	96.87
Fold-3	94.73	96.55	91.30
Fold-4	98.49	100.00	94.59
Fold-5	92.48	96.15	87.27
Fold-6	96.24	96.73	95.12
Fold-7	95.48	95.04	96.87
Fold-8	94.73	94.5	95.23
Fold-9	94.73	94.73	94.73
Fold-10	92.64	94.25	89.79
Average	95.12	95.93	93.72
SD	1.82	1.77	3.19

From both tables, it can be concluded that the accuracy, specificity, and sensitivity of the random forest model increase with an increase in the number of decision trees, although the increase is limited. This may be due to the model being more complex and better able to distinguish data.

The evaluation results of the model, including accuracy, specificity, and sensitivity using cross-validation for each department with the same ntree value, are presented in Tables 7 and 8.

Table 7. Cross validation based on majors with ntree value of 50

Ntree 50	Accuracy	Specificity	Sensitivity
D.Test: Mathematics D.Train: Chemistry, Physics, Biology, Computer Science	95.27	95.87	93.82
D.Test: Chemistry D.Train: Physics, Biology, Computer Science, Mathematics	93.07	98.02	81.39
D.Test: Biology D.train: Computer Science, Mathematics, Physics, Chemistry	93.79	95.12	91.81
D.Test: Physics D.train: Biology, Computer Science, Mathematics, Chemistry	90.38	87.79	95.95
D.Test: Computer Science D.Train: Mathematics, Chemistry, Physics, Biology	98.36	99.31	94.73
Average	94.17	95.22	91.54
Standard Deviation	2.93	4.47	5.87

Based on Tables 7 and 8, the decision tree classification model with a sample size of 50 resulted in an average accuracy of 94.17%, average sensitivity of 91.54%, and average specificity of 95.22%. Similarly, the decision tree classification model with a sample size of

100 resulted in an average accuracy of 94.23%, average sensitivity of 91.88%, and average specificity of 95.09%. In general, decision tree classification models with sample sizes of 50 and 100 demonstrate good performance. However, models with a sample size of 100 yield slightly higher accuracy compared to those with a sample size of 50.

Table 8. Cross validation based on majors with ntree value of 100

Ntree 50	Accuracy	Specificity	Sensitivity
D.Test: Mathematics D.Train: Chemistry, Physics, Biology, Computer Science	95.27	95.87	93.82
D.Test: Chemistry D.Train: Physics, Biology, Computer Science, Mathematics	93.07	98.02	81.39
D.Test: Biology D.train: Computer Science, Mathematics, Physics, Chemistry	93.79	95.12	91.81
D.Test: Physics D.train: Biology, Computer Science, Mathematics, Chemistry	90.38	87.79	95.95
D.Test: Computer Science D.Train: Mathematics, Chemistry, Physics, Biology	98.36	99.31	94.73
Average	94.17	95.22	91.54
Standard Deviation	2.93	4.47	5.87

When the specificity value is higher than the sensitivity value, it suggests that identifying students who will not graduate on time (true negatives) is easier than predicting students who will graduate on time (true positives). This could be attributed to the existence of other features that are associated with the classification. Nevertheless, our current sensitivity scores are satisfactory (low 90s). Another observation that the researcher can draw from Tables 7 and 8 is the small difference in accuracy between departments. This indicates that the factors influencing the length of study for students in different departments of FMIPA are the same. The accuracy of the Random Forest model does not decrease when predicting the length of study for a particular department using a model trained with data from students outside that department.

4.6 Visualization of the Results

Visualization was performed by creating a web-based dashboard generated using the RShiny framework. The purpose of this application is to display the data resulting from the Random Forest model prediction, which has been trained and presented in Table 3. The data is presented using plots with inputs consisting of student attributes used in the study, allowing for the visualization of the predicted data distribution using the selected attributes.

4.6.1 Representation of Visualization

Each predicted data has its own colour. The colours in the plot have different meanings according to each combination in the confusion matrix, namely:

i. Green, True Positive (TP): The prediction is that graduation will occur on time, according to the facts,
ii. Blue, True Negative (TN): The prediction is that graduation will not be on time, according to the facts,

iii. Orange, False Positive (FP): The prediction is that graduation will occur on time, but in reality, they do not,
iv. Red, False Negative (FN): The prediction is that graduation will not be on time, but it is actually on time.

For instance, to visualize the prediction results of numerical and categorical data, refer to Figures 3 and 4.

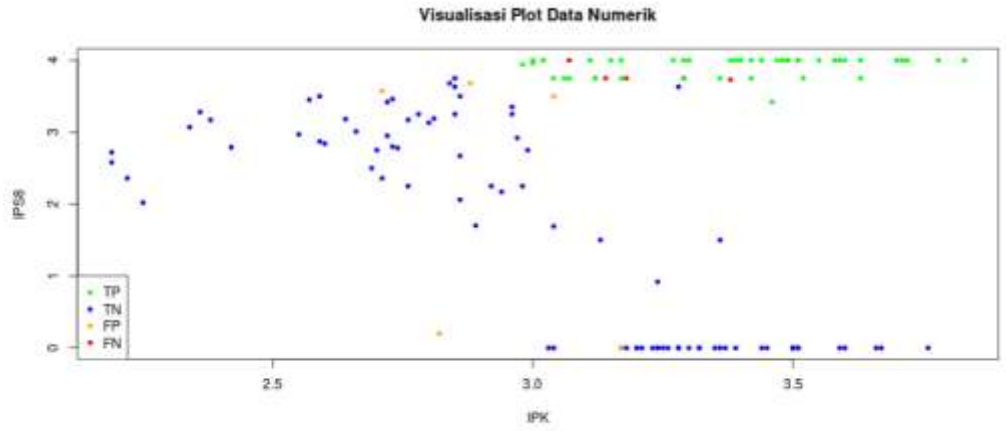

Figure 3. The visualization of numerical data plots

The green-coloured data points indicate true positives (TP) and are scattered only among the data points that have an 8th-semester IP above 3.0 and a GPA above 3.0. In Figure 3, the plot visualises numerical data using the variables of 8th-semester IP and IPK. Therefore, it can be concluded that, on average, students who achieve these IP and IPK scores graduate on time, as predicted by the model and actual data.

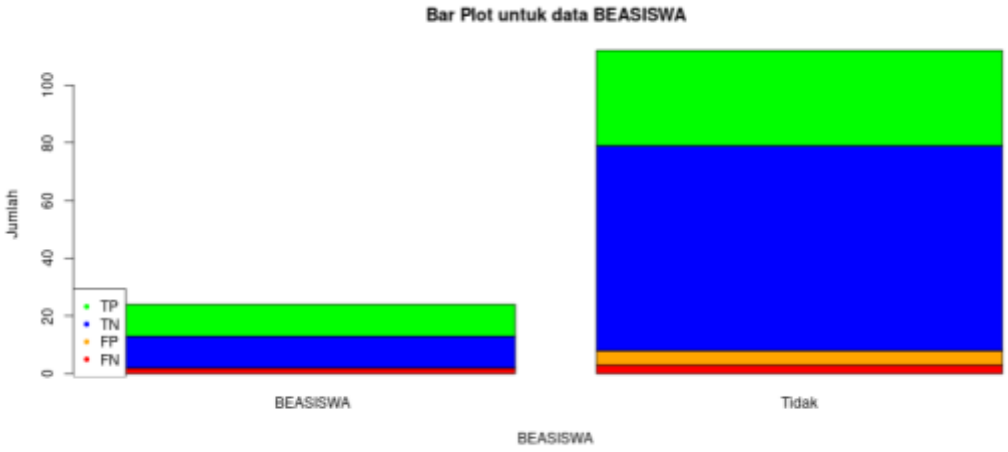

Figure 4. The visualization of categorical data plots

The visualization of categorical data plots, such as the example in Figure 4 using the variable Scholarship (BEASISWA), shows that there are more students without scholarships than those with scholarships. Additionally, the blue colour or TN (true negative) on the bar of

students with scholarships is more dominant than other colours. This means that most students without scholarships are predicted to graduate late, which is consistent with the actual data.

5. CONCLUSION

Based on the conducted research, the average accuracy, specificity, and sensitivity of predicting the length of study for students in the Faculty of Mathematics and Natural Sciences at the University of Riau using 10-fold cross-validation and 100 decision trees can be considered high, namely 95.12, 95.93, and 93.72. There are slight differences in the prediction results, one of which is that the group of students with physics major has a lower accuracy value. This could be due to differences in the level of contribution of variables to certain majors. The accuracy of the Random Forest model does not decrease significantly when predicting the study period of students in a specific major using a model trained with data from students other than that major.

REFERENCES

Adhani, M. H. R., & Iswari, L. (2022). Pengembangan Aplikasi Berbasis Web dengan R Shiny untuk Analisis Data Menggunakan Algoritma PCA. *Journal.Uii.Ac.Id*, *3*(1), 1–18.

Adrian, M. R., Putra, M. P., Rafialdy, M. H., & Rakhmawati, N. A. (2021). Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB. *Jurnal Informatika Upgris*, *7*(1), 36–40. https://doi.org/10.26877/jiu.v7i1.7099

Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Sistemasi*, *10*(1), 163–171. https://doi.org/10.32520/stmsi.v10i1.1129

Asana, I. M. D. P., Sudipa, I. G. I., Mayun, A. A. T. W., Meinarni, N. P. S., & Waas, D. V. (2022). Aplikasi Data Mining Asosiasi Barang Menggunakan Algoritma Apriori-TID. *INFORMAL: Informatics Journal*, *7*(1), 38. https://doi.org/10.19184/isj.v7i1.30901

Endang Etriyanti. (2021). Perbandingan Tingkat Akurasi Metode Knn Dan Decision Tree Dalam Memprediksi Lama Studi Mahasiswa. *Jurnal Ilmiah Binary STMIK Bina Nusantara Jaya Lubuklinggau*, *3*(1), 6–14. https://doi.org/10.52303/jb.v3i1.40

Hasan, I. K., Resmawan, R., & Ibrahim, J. (2022). Perbandingan K-Nearest Neighbor dan Random Forest dengan Seleksi Fitur Information Gain untuk Klasifikasi Lama Studi Mahasiswa. *Indonesian Journal of Applied Statistics*, *5*(1), 58. https://doi.org/10.13057/ijas.v5i1.58056

Hastie, T., Friedman, J., Tibshirani, R. (2001). Model Assessment and Selection. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-21606-5_7

Mariko, S. (2019). Aplikasi website berbasis HTML dan JavaScript untuk menyelesaikan fungsi integral pada mata kuliah kalkulus. *Jurnal Inovasi Teknologi Pendidikan*, *6*(1), 80–91. https://doi.org/10.21831/jitp.v6i1.22280

Marutho, Dhendra. 2019. "Perbandingan Metode Naïve Bayes, KNN, Decision Tree Pada Laporan Water Level Jakarta." *Jurnal Ilmiah Infokam* 15 (2): 90–97.

Orpa, E. P. K., Ripanti, E. F., & Tursina. (2019). *Model Prediksi Awal Masa Studi Mahasiswa Menggunakan Algoritma Decision tree c4.5*. JUSTIN (Jurnal Sistem dan Teknologi Informasi), *7*(4), 272–278.

Purbolaksono, M. D., Tantowi, M. I., Hidayat, A. I., & Adiwijaya. (2021). Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes. *Jurnal.Iaii.or.Id*, *1*(10), 393–399.

Qadrini L, Sepperwali A, & Aina A. (2021). Decision Tree Dan Adaboost Pada Klasifikasi Penerima Program Bantuan Sosial. *Jurnal Inovasi Penelitian*, *2*(7), 1959–1966.

Saadah, S., & Salsabila, H. (2021). Prediksi Harga Bitcoin Menggunakan Metode Random Forest. *Jurnal Politeknik Caltex Riau*, *7*(1), 24–32.

Sari, I. P., Syahputra, A., Zaky, N., Sibuea, R. U., & Zakhir, Z. (2022). Perancangan Sistem Aplikasi Penjualan dan Layanan Jasa Laundry Sepatu Berbasis Website. *Jurnal.Ilmubersama.Com*, *1*(1), 32–37.

Windarto, A. P., Defit, S., & Wanto, A. (2021). Optimalisasi Parameter dengan Cross Validation dan Neural Back-propagation Pada Model Prediksi Pertumbuhan Industri Mikro dan Kecil. *Jurnal Sistem Informasi Bisnis*, *11*(1), 34–42. https://doi.org/10.21456/vol11iss1pp34-42