Ingénierie des Systèmes d'Information
Vol. 27, No. 3, June, 2022, pp. 495-501
Journal homepage: http://iieta.org/journals/isi

Student Performance Prediction in Sebelas Maret University Based on the Random Forest Algorithm

Maulida Gusnina, Wiharto*, Umi Salamah

Department of Informatics, Universitas Sebelas Maret, Surakarta 57126, Indonesia

Corresponding Author Email: wiharto@staff.uns.ac.id

https://doi.org/10.18280/isi.270317

Received: 19 May 2022
Accepted: 6 June 2022

Keywords:
random forest, classification, information gain, student academic performance, support vector machine

ABSTRACT

Students who have low levels of academic performance may result in such students having drop out. Various factors influence the level of academic performance of such students. Preventive action would be better to cope with the drop out. This study aims to conduct prediction of students' academic performance at Sebelas Maret University based on three categories of factors namely social, economic, and academic factors. Methods used include, data acquisition stages, data preprocessing, feature selection, classification, and analysis of results. Feature selection uses information gain (IG) and Random Forest (RF) classification algorithms, with 10-fold cross validation. The test results showed an accuracy of 90.7%, such performance outperforms support vector machine (SVM) and Decision Tree (DT) algorithms.

1. INTRODUCTION

1.1 Research background

Based on Ristekdikti's [1] 2018 statistics, 5% of all students in Indonesia experienced drop out caused by resignation, dropping out of college, and being expelled. Various studies were conducted to find out what factors could affect students' academic performance. According to Bhardwaj and Pal [2], the academic performance of students depends on personal, social, economic and environmental factors. Feng et al. [3] conducted a study on the effect of social media use on academic performance on 100 students at a University in Hong Kong, the result showing a positive relationship between the high social media use on the low academic performance rate of students. Research conducted by Syed and Raza Naqvi [4] on 300 students of Punjab University of Pakistan, drew the conclusion that the economic factors of parents of students as well as the education level of parents greatly affected students' performance in institutions. A similar study was also conducted by Jayaprakash et al. [5], academic performance is greatly influenced by economic and personal factors.

Referring from conducted studies related to academic performance prediction, a number of influential factors can be shown, namely personal, social media, parents' economy, the environment, and parents' education levels. Unfortunately a number of studies have been conducted using only a few factors [2-5], so it cannot be shown which factors are actually the most dominant of all the factors that have been used in a number of such studies.

The quick development of machine learning also brought about an impact in the development of academic performance prediction models. Deepika on her research predicted students' academic performance, obtaining the result that the influential variable is important i.e., distance and time traveled to school, by the proposed machine learning method RFBT-RF (Relief-F and Based Tree Random Forest). It was obtained that the submitted algorithm was superior to SVM, kNN, naïve Bayesian classifier (NBC), and DT [6]. Hussain on his research regarding student performance prediction, with 300 data and 24 variables, his research results show Random Forest methods have a higher accuracy rate of 99% compared to DT which is only 74% [7]. On another study by Zhou regarding train cross accident prediction, it states that Random Forest can improve the accuracy of prediction especially reducing false alarms. Random Forest could process unbalanced data and handle over fitting more effectively [8].

Another study was conducted by Karabuluta et al. [9] on the performance comparison of Information Gain, Gain Ratio, Relief-F, and Chi-square feature selection methods used in conjunction with Decision Tree classification methods. Research results show selection of Information Gain features improves Decision Tree accuracy by 1.93% [9]. On a similar study conducted by Sharma and Dey [10] perform also a comparison of a number of feature selection methods. Comparison results show that the Information Gain feature selection method is more stable as well as its performance is not affected by the amount of data used.

Referring from a number of studies that have been conducted to show that academic performance is greatly influenced by several factors, including social, economic and academic factors on the development of machine learning-based predictive models, these factors need to be re-analyzed for their contribution to predicting academic performance. Thus, in this study proposed an academic performance prediction model by using a combination of information gain to analyze its contribution in the model. The conclusion of the prediction is done using a Random Forest (RF) algorithm. The proposed model is validated by using k-folds cross-validation, with performance parameters of accuracy, precision, F1-Score and Recall.

This paper will be presented in several sections, part I which is the introduction and background of the study, part II which is the method of the study conducted, part III which is the

result of the study along with its discussion, and the latter part IV contains the conclusions drawn from the results of this study.

1.2 Literature review

Machine learning algorithms are widely used to resolve problems in various fields, including academic performance prediction. Hussain conducted research on student academic performance classification, comparing a number of machine learning algorithms, namely J48, PART, Random Forest, and Bayes Network Classifiers [7]. Hussain's research in predicting using socioeconomic, personal, and academic historical data with. In other Hussain studies for the improvement of students' academic performance, with the application of Deep Learning Hussain underlined the variables of the semester exam as an indicator of whether to perform further actions on those students [11].

Another study conducted by Feng using the ANOVA method emphasized the variables of social media use as the most influential variable on the low academic performance of college students [3]. Deepika on her research uses several machine learning algorithms, i.e., by comparing SVM, KNN, NBC, DT, and RFBT-RF. The study emphasized the distance and time variables of students to university as the most significant factor [6]. Subsequent research was conducted by Yang, on his research on student performance classification using Back-Propagation Neural Network (BP-NN). In the study, it proved that students with similar variable scores have a tendency to get equivalent results of academic performance [12].

2. MATERIAL AND METHOD

The study was conducted with several stages as Figure 1 shows. Figure 1 shows research divided into several stages i.e., data collection, preprocessing, feature selection, classification, and analysis of system performance.

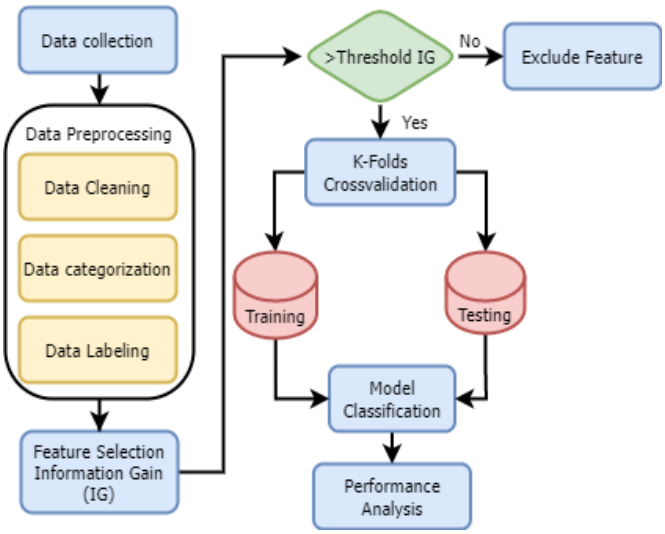

Figure 1. Proposed model

2.1 Data collection

Data were collected through online forms filled in by active students of March Eleven University. The data is divided into 12 Faculties and evenly distributed for each Faculty. The data collected were a total of 223 university students. The attributes used a number of 21, following attributes in full are shown in Table 1.

Table 1. Attributes

Code	Attribute
jk	Gender
tt	Residence type
jt	Distance of residence to college
status	Marriage status
pda	Paternal education
pdi	Maternal education
org	Organization
jb	Study hours
hp	Smartphone usage
ukt	Tuition fee
kja	Father's employment
kji	Mother's employment
pdp	Family income
tgg	Family dependents
smp	Junior high school final exam score
sma	Senior high school final exam score
ip	1st semester GPA
ipk	GPA
jm	College entrance
ab	Class attendance
pre	Academic achievement

2.2 Data preprocessing

Data preprocessing represents a process of data extraction that includes the elimination of duplicate data, checking for inconsistent data, and correcting errors or deficiencies in the data. On this process also an enrichment or process enriches the data with other relevant information [13].

2.3 Feature selection

The feature selection stage is to perform feature selection or attributes that affect student academic performance prediction. The feature selection method used is information gain. Information Gain is one method of feature selection to select the best feature. This method performs an outreach of the feature reduction results. An Information Gain value obtained from the entropy value before separation is subtracted by the entropy value after separation. entropy describes the amount of information needed to encode a class. This value measurement is used only as an initial stage for determination of attributes that will later be used or discarded. Attributes that meet the defection criteria that will later be used in the classification process of an algorithm [13]. The entropy attribute and gain calculation are as shown in Eqns. (1) and (2).

$$\text{Entropy } (Y)=\sum p(c|Y)\log_2 p(c|Y) \quad (1)$$

$$\text{Gain } (Y,a)= \text{Entropy } (Y)-\sum v \in \text{Values}(a) \frac{|Yv|}{|Ya|} \text{Entropy}(Yv) \quad (2)$$

where, Y is the sum of the entire feature, a is the category, Yv is the number of samples for the value v, v is the possible value for the category a, $p(c|Y)$ is the proportion of the value Y against c.

2.4 Classification

Classification is a technique of grouping data based on a particular behavior of the attributes of a predefined group. The results of this technique constitute a grouping rule of the classified data [13]. Some classification algorithms are Decision Tree, K-Nearest Neighbors, Naïve Bayes and one of the embedded algorithms is Random Forest.

Random Forest represents the development classification method of Decision Tree. On the process its classification is based on the most votes from the returned decision tree. Random Forest is constructed by using bagging with random attribute selection.

The Classification and Regression Tree (CART) method is used to grow the decision tree. Such decision trees grow to the maximum number and produce a collection of trees that are then called forests (forest) [14].

The Random Forest algorithm process can be detailed as follows: suppose that the training data cluster we have is n in size and consists of d explanatory modifiers (predictors), the stages of preparation and assessment using Random Forest as follows:

(1) Create bootstrap data, draw random instances with n-sized formulations of the training data cluster.
(2) Perform random sub-setting, draw a decision tree based on that data, on every separation process randomly select the number of predictor variables (m)<d explanatory modifications, and perform the best separation.
(3) Repeat the first and second steps k times so that k number decision trees is obtained.
(4) Perform a joint presumption based on the k number of the trees (e.g. using a majority vote untuk classification cases or mean for regression cases).

2.5 System performance analysis

At the classification algorithm, the resulting output of such a training process is named a model. On this model an evaluation is needed to determine whether or not such a built-in machine is good enough [15].

Train-Test Split

In this method datasets are divided into two types of data i.e., training data and test data. The percentage share of this data set is determined by the researcher, suppose for training data it is determined as much as 80% and 20% of the data will be used as test data.

k-folds cross validation

This method divides and group the data set as many k values as specified. The evaluation was performed a number of k with the test data used i.e. one of the data groups that had been divided earlier. As many as k iterations, data used as test data continuously being tested to get the results that represent the entire data set.

Confusion matrix

Confusion matrix gives an assessment of the performance of the classification model based on the number of correctly and incorrectly predicted objects [16]. The confusion matrix illustrates a comparison between prediction results with reality. True Positive (TP) is that the prediction and the actual value are both correct. True Negative (TN) is the prediction value and the actual value are both incorrect. False Positive (FP), the prediction value is correct, and the actual value is incorrect. False Negative (FN) prediction value is incorrect, and the actual value is correct. The calculation of accuracy, precision, sensitivity and f1-score can be shown in Eqns. (3)-(6).

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}\times 100\% \quad (3)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (5)$$

$$\text{F-1 score} = 2\times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (6)$$

3. RESULTS AND DISCUSSION

The research was conducted using a python programming language library with jupyter notebook editor. The hardware specifications used are laptops with AMD Ryzen 3 2200U CPU @ 2.50GHz, 8GB RAM and AMD Radeon™ Vega 3 Graphics GPUs.

3.1 Data preprocessing

The steps of data preprocessing are data clearing, data labeling, data categorization, and attribute selection. In the data clearing step, the marriage status variable is eliminated because all values are the same i.e., "unmarried". In addition to the incomplete elimination of the data, a number of 3 data have null variable values so the 3 data are eliminated. The number of data after the elimination of 1 variable and 3 data are 220 data with 19 variables as attributes and 1 variable as classes.

Table 2. Gain value

Attribute	Gain
Ip	0.1399
Kja	0.0640
Pda	0.0527
Pdi	0.0510
ab	0.0478
tgg	0.0466
ukt	0.0449
sma	0.0369
kji	0.0359
jm	0.0336
pre	0.0303
jb	0.0293
hp	0.0206
jk	0.0200
pdp	0.0157
tt	0.0099
jt	0.0085
org	0.0045
smp	0.0022

Further, the data conversion from numeric to categorical is performed. Some data are of numerical value such as achievement index, UN value and smartphone usage. At the achievement index and UN grade the grade categories are based on the curriculum. On smartphone usage variable, it is categorized lightly when the value is <5 hours and categorized heavily when the value is > 5 hours [17]. Then labeling the classes on the data, the data will be grouped into three classes

i.e., excellent, good, and poor that are based on the GPA variable.

In the next step i.e., feature selection using Information Gain, the result of ranking the attribute gain value is obtained as in Table 2. To calculate the entropy attribute and gain calculation, Eqns. (1) and (2) are used.

To sort out attributes deemed less relevant an attribute elimination shall be performed by means of limiting the attribute gain values by various threshold values as in Table 3.

Table 3. Applied threshold

Threshold	Selected Attributes
0.004	ip, kja, pda, pdi, ab, tgg, ukt, sma, kji, jm, pre, jb, hp, jk, pdp, tt, jt, org
0.005	ip, kja, pda, pdi, ab, tgg, ukt, sma, kji, jm, pre, jb, hp, jk, pdp, tt, jt
0.01	ip, kja, pda, pdi, ab, tgg, ukt, sma, kji, jm, pre, jb, hp, jk, pdp
0.02	ip, kja, pda, pdi, ab, tgg, ukt, sma, kji, jm, pre, jb, hp, jk
0.03	ip, kja, pda, pdi, ab, tgg, ukt, sma, kji, jm, pre
0.04	ip, kja, pda, pdi, ab, tgg, ukt
0.05	ip, kja, pda, pdi

3.2 Random Forest implementation

On this experiment a comparison of Random Forest (RF) classification accuracy rates without Information Gain (IG) feature selection and with various threshold value limits of IG feature selection as listed in Table 3. The results of such accuracy value comparison experiments are as shown in Table 4.

Table 4. Comparison of RF accuracy against various IG threshold

Threshold IG	Accuracy		
	train-test	5-folds cv	10-folds cv
Without IG	0.886	0.904	0.890
0.05	0.795	0.772	0.772
0.04	0.909	0.881	0.872
0.03	0.841	0.904	0.904
0.02	0.795	0.881	0.877
0.01	0.864	0.904	0.896
0.005	0.932	0.908	0.907
0.004	0.886	0.904	0.896

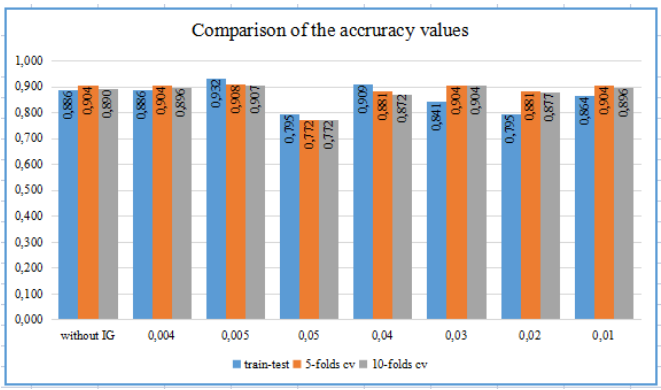

Figure 2. Comparison of RF accuracy against IG threshold

Based on the comparison experiment above, the result is that the limitation threshold value of 0.005 has a higher accuracy value compared with the accuracy value at the others limitation threshold value. At threshold 0.005 the attributes used as determinants are ip, kja, pda, pdi, ab, tgg, ukt, sma, kji, jm, pre, jb, hp, jk, pdp, tt, jt, while org and smp attributes have been eliminated from the data set.

Table 5. The odds of attributes grouped within a class

Attributes	Value	Classes odds		
		Good	Excellent	Poor
jk	F	0.60	**0.66**	-
	M	**0.39**	0.33	1
tt	Dorm	**0.42**	0.36	-
	House	0.57	**0.60**	1
pda	Highschool	**0.51**	0.38	0.25
	Degree	0.46	**0.53**	0.75
pdi	Highschool	**0.43**	0.390	0.5
	Degree	0.390	**0.398**	0.5
kja	No	0.05	**0.10**	-
	Retirement	-	**0.07**	-
	Laborer	0.04	**0.09**	-
	Self-employed	**0.53**	0.36	0.25
	Employee	0.13	**0.20**	0.25
	Civil employee	**0.23**	0.14	0.5
kji	No	0.298	**0.39**	0.5
	Laborer	-	**0.031**	-
	Self-employed	**0.287**	0.257	-
	Employee	**0.13**	0.070	-
	Civil employee	**0.287**	0.250	0.5
pdp	Low	0.46	**0.562**	0.25
	Moderate	**0.36**	0.343	0.5
	High	**0.19**	0.093	0.25
tgg	Big	**0.26**	0.062	-
	Moderate	0.54	**0.601**	0.5
	Small	0.19	**0.335**	0.5
ukt	BM	0.15	**0.242**	-
	I	**0.06**	0.046	-
	II	**0.08**	0.039	-
	III	**0.16**	0.156	-
	IV	**0.22**	0.179	0.75
	V	0.07	**0.132**	-
	VI	0.09	**0.117**	0.25
hp	VII	**0.10**	0.046	-
	VIII	**0.07**	0.039	-
	Light	**0.78**	0.593	1
	Heavy	0.22	**0.406**	-
sma	A	**0.57**	0.546	0.5
	B	0.40	**0.429**	0.5
	C	0.020	**0.023**	-
ip	Excellent	0.30	**0.734**	-
	Good	**0.70**	0.265	0.75
	Poor	-	-	0.25
jm	snmptn	0.26	**0.429**	-
	sbmptn	**0.40**	0.351	1
	mandiri	**0.33**	0.218	-
ab	Poor	-	**0.015**	-
	Acceptable	**0.56**	0.351	-
	Good	0.43	**0.687**	1
pre	Yes	0.18	**0.414**	-
	No	**0.81**	0.585	1
jb	Low	**0.04**	0.007	-
	Moderate	**0.45**	0.289	0.75
	High	0.49	**0.703**	0.5

In Figure 2 it is seen that at a threshold limitation of 0.005 both using train-test and 5-fold cross validation methods and 10-fold cross validation have higher accuracy rates compared to those in other limits.

The trends of each attribute that goes into a class are as presented in Table 5. The comparison is calculated based on the percentage of the number of values an attribute appears on

a class category against the overall data on that class.

Based on Table 5 the trends of each attribute that can be grouped into "Excellent" classes are as follows:

(1) Gender

Female students tend to get better academic performance results than male students. According to Montolio, female students tend to get higher results when not working under pressure than male students, such as in a freer university academic environment without schedule pressure like in middle school [18].

(2) Residence

Students who reside in homes go into a condition of Excellent academic performance than students who reside in dormitory. According to Noh, the role of parents in accompanying the student's learning process can increase the student's confidence and focus [19].

(3) Parent's education and employment

Students whose parents one has a bachelor's degree/diploma and one of them working as an employee have a tendency to fall into the performance category of the academy Very Good. Parental assistance was influential in college students understanding his studies, with parents having higher levels of education rated more stable in providing assistance [20].

(4) Family income

The conditions of families with moderate incomes and the number of dependents of families that do not have much tend to have Excellent academic performance. It is said that students' families are in a capable and simple condition according to BPS. The modest family is thought to be able to provide more time in student learning assistance [21].

(5) Family dependents

Students with families with small to moderate family dependents tend to have better academic performance. According to Shi, students who have siblings have a better growth in attitude and discipline and thus have an effect on academic performance [22].

(6) Tuition fee

Students who have excellent performance is in the middle class of tuition fee, namely III to VI and BM students. The amount of tuition fee in this class is obtained by students with economically capable and simple conditions.

(7) Smartphone usage for entertainment purpose

Students with Excellent academic performance tends to be light in smartphone use. The heavy use of smartphones for entertainment purposes is in a straight line with the low academic performance of students, this is because the concentration of students will be disrupted by this activity [3].

(8) Senior high school final exam score

In the results of grouping students' academic performance classes in a straight line with high school exam scores, with Category A exam results the most in the Very Good class.

(9) 1st semester GPA

The results of grouping students' academic performance classes are in a straight line with the GPA in the first semester of the student course.

(10) College entrance

Students grouped in Excellent class tend to enter college through SNMPTN path or selection using school grades, this is in accordance with the results of a study showing that high school exam scores are in a straight line with student performance.

(11) Class attendance

Students with good attendance percentage or above 90% tend to fall into the Excellent student performance category.

(12) Academic achievement

Students who have academic achievement tend to get an Excellent academic performance.

(13) Study hours

Students who take extra study time for more than 2 hours outside the course hours have a greater chance of falling into the category of Excellent academic performance. Learning time between 2 to 3 hours and interspersed with other activities is considered to have a better effect on student performance because it is not very tiring [23].

3.3 Comparison with Decision Tree and SVM methods

In comparison of the performance of Random Forest classification methods, experiments were conducted on existing datasets using Decision Tree and SVM classification methods. On the Decision Tree method, the number of nodes in use is 100. In the Decision Tree preliminarily determined root nodes, branch nodes, and leaves using gain values as well as entropy obtained based on calculations with formulas such as those in Eq. (1) and Eq. (2). On the SVM method, the kernel used on this experiment, which is the RBF kernel, the RBF kernel has two parameters namely gamma and C. Used GridSearchSV to obtain the best gamma and C parameter values on the model used. The gamma values that are chosen are 1, 0.1, 0.01, and 0.001 while the choices for C values are 0.1, 1, 10, and 100.

The results are as examined in Table 6, on testing using both split train-folds and k-folds cross validation methods (k=5 and k=10), the accuracy of Random Forest is higher compared to the accuracy of Decision Tree and SVM.

Table 6. The comparison of RF, DT, and SVM accuracy

Classification method	Train-Test	5-Folds CV	10-Folds CV
RF	**0.932**	**0.908**	**0.907**
DT	0.886	0.854	0.867
SVM	0.728	0.699	0.721

A comparison graph of the accuracy of the three methods can be seen in Figure 3.

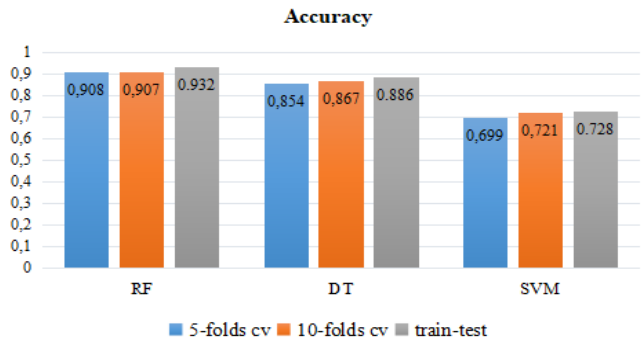

Figure 3. RF, DT, and SVM accuracy comparison

Based on these results, the factors most influencing students' academic performance included the achievement index, father's job, father's education level, mother's education level, absent class, family dependency, UKT, high school exam scores, mother's job, average number of hours of study, academic achievement, college entrance, average number of hours of smartphone use for entertainment purposes, gender,

family income, residence type, and residence distance to college.

Meanwhile, the factors irrelevant to the academic performance of students according to this study are organizational factors and junior high school final exam score factors. According to Rangkuti, students' activities in organizing have a positive but insignificant influence on the student's academic achievement [24]. Meanwhile, the difference between junior high school and high school final exam scores can be attributed to the length of time between junior high school and college levels, as well as differences in the degree of complexity of the subject matter test made the factor of the junior high school exam not very significant. The length and sleep patterns that students at age 15 have with students entering age 17 are also different, such differences could affect the activity and physical condition of students [25, 26].

3.4 Evaluation

An evaluation is performed using the confusion matrix. Obtained the confusion matrix from the experiment conducted as in Table 7.

Table 7. Confusion matrix

	Poor	Good	Excellent
Poor	12	0	1
Good	0	1	0
Excellent	2	0	28

Then a precision, recall, as well as f-1 score calculation was performed using Eqns. (3)-(6), the results as examined in Table 8.

Table 8. Precision, recall, F-1 score

Class	Precision	Recall	F-1 score
Poor	0.857	0.923	0.888
Good	1	1	1
Excellent	0.966	0.933	0.949

4. CONCLUSIONS

Based on research on the classification of student academic performance at Sebelas Maret Universities using the Information Gain feature selection method and Random Forest classification, it can be concluded that the factors that most influence student academic performance sequentially include the achievement index, father's job, father's education level, tingkat maternal education, class attendance, number of family dependents, tuition fee, high school exam scores, maternal employment, average number of hours of study, academic achievement, college entrance, average number of hours on smartphone for entertainment purposes, gender, family income, residence, and residence distance to college.

The use of the Information Gain feature selection method improved the accuracy rate of the Random Forest classification method by 4.6% at train-test evaluation and 0.4% as well as 1.8% at k-fold cross validation evaluation. The Accuracy rate of Random Forest method is 93.2%, higher compared to other classification methods such as Decision Tree which has 88.6% accuracy and SVM which has 72.8% accuracy. On testing using the k-folds cross validation method the accuracy of Random Forest was 90.8% and 90.7% higher compared to Decision Tree methods with 85.4% and 86.7% accuracy and SVM methods with 69.9% and 72.1% accuracy for k=5 and k=10 values respectively.

On this study for relationships between variables has not been discussed in depth, suggestions for subsequent studies that can be considered namely adding other methods such as Association Rules to examine the correlation of each variable to the final result obtained. Adds other variables that may affect the academic performance level of students.

REFERENCES

[1] Ristekdikti, Statistik Pendidikan Tinggi Indonesia. (2018). Pangkalan Data Pendidikan Tinggi. available: https://pddikti.kemdikbud.go.id/asset/data/publikasi/Statistik%20Pendidikan%20Tinggi%20Indonesia%202018.pdf.

[2] Bhardwaj, B.K., Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. Int. J. Comput. Sci. Eng., 9(3): 5. https://doi.org/10.48550/arXiv.1201.3418

[3] Feng, S., Wong, Y.K., Wong, L.Y., Hossain, L. (2019). The internet and Facebook usage on academic distraction of college students. Computers & Education, 134: 41-49. https://doi.org/10.1016/j.compedu.2019.02.005

[4] Syed, T.H., Raza Naqvi, S.M.M. (2006). Factors affecting students' performance: A case of private colleges. Bangladesh e-Journal of Sociology, 3(1): 1-10.

[5] Jayaprakash, S., Krishnan, S., Jaiganesh, V. (2020). Predicting students academic performance using an improved random forest classifier. In 2020 international conference on emerging smart computing and informatics (ESCI), 238-243. https://doi.org/10.1109/ESCI48226.2020.9167547

[6] Deepika, K., Sathyanarayana, N., Sathyanarayana, N. (2019). Relief-F and budget tree random forest based feature selection for student academic performance prediction. International Journal of Intelligent Engineering and Systems, 12(1): 30-39. https://doi.org/10.22266/ijies2019.0228.04

[7] Hussain, S., Dahan, N.A., Ba-Alwib, F.M., Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2): 447-459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

[8] Zhou, X., Lu, P., Zheng, Z., Tolliver, D., Keramati, A. (2020). Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. Reliability Engineering & System Safety, 200: 106931. https://doi.org/10.1016/j.ress.2020.106931

[9] Karabulut, E.M., Özel, S.A., Ibrikci, T. (2012). A comparative study on the effect of feature selection on classification accuracy. Procedia Technology, 1: 323-327. https://doi.org/10.1016/j.protcy.2012.02.068

[10] Sharma, A., Dey, S. (2012). A comparative study of feature selection and machine learning techniques for sentiment analysis. In Proceedings of the 2012 ACM Research in Applied Computation Symposium, pp. 1-7. https://doi.org/10.1145/2401603.2401605

[11] Hussain, S., Muhsin, Z., Salal, Y., Theodorou, P.,

Kurtoğlu, F., Hazarika, G. (2019). Prediction model on student performance based on internal assessment using deep learning. International Journal of Emerging Technologies in Learning, 14(8): 4-22. https://doi.org/10.3991/ijet.v14i08.10001
[12] Yang, F., Li, F.W. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. Computers & Education, 123: 97-108. https://doi.org/10.1016/j.compedu.2018.04.006
[13] Larose, D.T., Larose, C.D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining, 4. John Wiley & Sons. Inc. 2005
[14] Dicky, N., Nurcahyo, G.W. (2015). Algoritma Data Mining Dan Pengujian. Yogyakarta, Indonesia: Deepublish.
[15] Brownlee, J. Master Machine Learning Algorithms: Discover How They Work and Implement Them from Scratch. Jason Brownlee, 2016.
[16] Han, J.W., Pei, J. (2012). Data Mining Concepts and Techniques. Morgan Kaufman Publisher.
[17] Gorunescu, F. (2011). Data Mining: Concepts, models and techniques, 12. Springer Science & Business Media.
[18] Alwagait, E., Shahzad, B., Alim, S. (2015). Impact of social media usage on students academic performance in Saudi Arabia. Computers in Human Behavior, 51: 1092-1097. https://doi.org/10.1016/j.chb.2014.09.028
[19] Montolio, D., Taberner, P.A. (2021). Gender differences under test pressure and their impact on academic performance: a quasi-experimental design. Journal of Economic Behavior & Organization, 191: 1065-1090. https://doi.org/10.1016/j.jebo.2021.09.021
[20] Noh, M. (2022). Effect of parental financial teaching on college students' financial attitude and behavior: The mediating role of self-esteem. Journal of Business Research, 143: 298-304. https://doi.org/10.1016/j.jbusres.2022.01.054
[21] Zhang, X., Hu, B. Y., Ren, L., Zhang, L. (2019). Family socioeconomic status and Chinese children's early academic development: Examining child-level mechanisms. Contemporary Educational Psychology, 59: 101792. https://doi.org/10.1016/j.cedpsych.2019.101792
[22] Shi, J., Li, L., Wu, D., Li, H. (2021). Are only children always better? Testing the sibling effects on academic performance in rural Chinese adolescents. Children and Youth Services Review, 131: 106291. https://doi.org/10.1016/j.childyouth.2021.106291
[23] Frederick, G.M., O'Connor, P.J., Schmidt, M.D., Evans, E.M. (2021). Relationships between components of the 24-hour activity cycle and feelings of energy and fatigue in college students: A systematic review. Mental Health and Physical Activity, 21: 100409. https://doi.org/10.1016/j.mhpa.2021.100409
[24] Rangkuti, S.D.S. (2020). Pengaruh keaktifan mahasiswa dalam organisasi dan motivasi belajar terhadap prestasi akademik mahasiswa akuntansi universitas pembangunan panca budi medan. Kumpulan Karya Ilmiah Mahasiswa Fakultas Sosial Sains, 2(02).
[25] Stefansdottir, R., Rognvaldsdottir, V., Gestsdottir, S., Gudmundsdottir, S.L., Chen, K.Y., Brychta, R.J., Johannsson, E. (2020). Changes in sleep and activity from age 15 to 17 in students with traditional and college-style school schedules. Sleep Health, 6(6): 749-757. https://doi.org/10.1016/j.sleh.2020.04.009
[26] Sheetal, A.P., Ravindranath, K. (2021). High efficient virtual machine migration using glow worm swarm optimization method for cloud computing. Ingénierie des Systèmes d'Information, 26(6): 591-597. https://doi.org/10.18280/isi.260610