INTERNATIONAL JOURNAL OF SOCIAL SCIENCE HUMANITY & MANAGEMENT RESEARCH

ISSN (print) 2833-2172, ISSN (online) 2833-2180

Volume 03 Issue 06 June 2024

DOI: 10.58806/ijsshmr.2024.v3i6n13 ,Impact Factor: 5.342

Page No. 671-678

Development of a Model for Predicting Students' Achievement

Lukman Nadjamuddin[1], Arwansyah Arwansyah[2], Sukmawati Sukmawati[3], Windayanti[4]

[1, 4]Department of History Education, Faculty of Teacher Training and Education, Tadulako University, Palu, Indonesia.

[2]Department of Chemistry Education, Faculty of Teacher Training and Education, Tadulako University, Palu, Indonesia.

[3]Department of Civic Education, Faculty of Teacher Training and Education, Tadulako University, Palu, Indonesia.

ABSTRACT: In this study, data mining was implemented to find the common variables such as "gender", "age", "GPA first semester", "GPA second semester", "organization activity", "part time job", "living place", "family income", "father education", "mother education" influencing the grade point of average (GPA) score of the 3nd-semester student at the Department of History Education. Three methods, including logistic regression (LR), decision tree (DT), and support vector machine (SVM) were employed. According to the validation results, the best algorithm method is found in the model developed by decision tree with the Accuracy 0.96 and all models provide sufficient data since the AUC value for all classes is greater than 0.5. This finding proved that above variables are linked to student achievement. As a result, concern to those aspects is critical for improving academic performance.

INTRODUCTION

Education is one of the main factors for the progress and independence of the nation. According to Brighouse (2006), Pérez-Ordás et al. (2021), Sindre et al. (2018), education is the learning of a group of people's information, skills, and habits that are passed down from generation to generation through teaching, training, and even research activities. One of the key elements in the development and independence of a country is education. The improvement of education is correlated with a country's quality and independence. The goal of Indonesia's national education system is to help students realize their potential as people who believe in and fear God Almighty, have good character, knowledgeable, capable, creative, self-sufficient, democratic, and responsible citizens. This is done by helping students develop the skills, character, and civilization of a dignified nation.

The government and the private sector collaborate to achieve the goal of education by pursuing a variety of initiatives to improve education quality, such as developing and improving curriculum and evaluation systems, improving educational facilities, developing and acquiring teaching materials and training. Given the significance of learning quality in the context of educating the nation's life. Academic accomplishment is one sign of the standard and management of educational institutions (Nasim et al., 2020; Sciarelli et al., 2020; Straus, 1996). As a result, the government has made efforts to raise student achievement a top priority in order to fulfill the goals of Indonesian national education.

Numerous research teams have looked into a number of aspects that can affect a student's academic success, such as family income, the learning environment at school, motivation, student behaviour, social media, past successes, how much time is spent studying and playing each day, etc. The authors of the paper by Christenson and colleagues (Christenson et al., 1992) suggested that there is a convergence of familial factors that are significant to students' performance; as a result, recommendations are made for school psychologists on how to enhance student learning opportunities by offering direct and indirect services to families.

In a different article, Morrissey and colleagues suggested that low family wealth was related to children's poor academic performance because of the relationship with the student facilities (Morrissey et al., 2014). Oguguo group has reported the impact of social media (Oguguo et al., 2020). The findings showed that social media is frequently used by students to make new friends, obtain resources for their assignments, and do research. Relying on these statements, the current study was undertaken to create a model that can show how some of the aforementioned variables affect academic achievement, allowing one to predict it and look for an appropriate method to boost student performance.

Data mining was used in the current inquiry to assess various elements influencing the academic performance of the second-semester student at the Department of Historical Education. Data mining is the process of sifting through or gathering important information from massive data sets (Borkar et al., 2019; Gordan et al., 2019; Mughal, 2018). Artificial intelligence technology, statistical analysis, and mathematical techniques are frequently used in the process.

Development of a Model for Predicting Students' Achievement

Due to the availability of vast amounts of data that can be transformed into useful information based on the patterns of the examined data, data mining has recently caught the interest of scholars, practitioners, and even the industry (Dehkordi & Sajedi, 2019; Huber et al., 2019; Uska et al., 2020). In this study, data mining is utilized to forecast data patterns based on a number of factors, such as family income, gender, past accomplishments (senior high scholar report), the amount of time spent studying and playing each day, and the student's most recent grade point average (GPA).

Materials and Methods Data Preparation and Collection

This study used quantitative approach to analyze the collected datasets. The place of the research is at the at Department of Historical Education, Tadulako University, Palu, Central Sulawesi, Indonesia. The duration of research and data collection were conducted from December 2022 to March 2023. The subject of the research is the 3nd-semester students of Department of Historical Education in the academic year 2022/2023 with a total of 106 students. The Google form application was used to collect the data such as gender, age, GPA first semester, GPA second semester, organization activity, part time job, living place, family income, father education, mother education and the latest GPA (target).

Machine Learning Algorithms

The student accomplishment of the Historical Education Department's third-semester students was predicted using three algorithms: logistic regression, decision trees, and support vector machines. A statistical technique called logistic regression (LR) is frequently used to examine data that includes both the response variable and one or more predictor variables (Abonazel & Ibrahim, 2018; Denœux,
2019; Shah et al., 2020). The response variable from logistic regression is a binary variable with only 1 (yes) and 0 (no) values, so the resulting response variable will follow the Bernoulli distribution using the probability function below:

$f(y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}$ (1) with $y_i = 0,1$

Based on equation (1), the model for LR can be written according to the following equation:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1+\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)} \quad (2)$$

From equation (2), the $g(x)$ function used to predict the regression parameter can be estimated according to the equation:

$$g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (3)$$

One of the most common categorization techniques for a predictive model employing a tree structure or hierarchical structure is the decision tree (DT) (Figure 1). Using this technique, a tree made up of decision nodes connected by branches from the root node to the leaf node is constructed (end). At the decision node, the qualities will be assessed, and each result will produce a branch. Each branch is directed to a different node or the end node to reach a decision (Charbuty & Abdulazeez, 2021; Hasan et al., 2018; Patel & Prajapati, 2018).

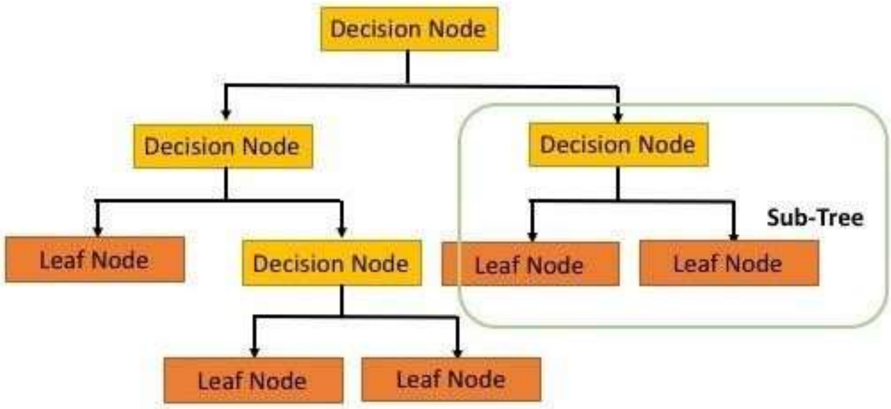


Figure 1. The representation of decision workflow

The flow of the analysis process in the decision tree is worked by changing the shape of the data (table) into a tree model, generating the rules, and simplifying the created rules (pruning). In this study, the data retrieved from a population of 106 students will be used to make a prediction model. The prediction accuracy was estimated on the created model.

The Support Vector Machine (SVM) is a supervised learning method that is commonly applied for classification (for example, Support Vector Classification) and regression (Support Vector Regression) (Cervantes et al., 2020; Cortes & Vapnik, 1995; Pisner & Schnyer, 2020). SVM has a more sophisticated and explicit mathematical notion than other classification algorithms in the classification model. SVM can also cause regression issues with linear and non-linear models. By maximizing the distance between classes, SVM is used to find the optimum hyperplane. A hyperplane is a function that can be used to divide classes. The function for classification between classes in two dimensions (2-D) is called "a line whereas," while the function for classification between classes in three dimensions (3-D) is called "plane similarly." Hyperplane is the classification function used in a higher-dimensional class space.

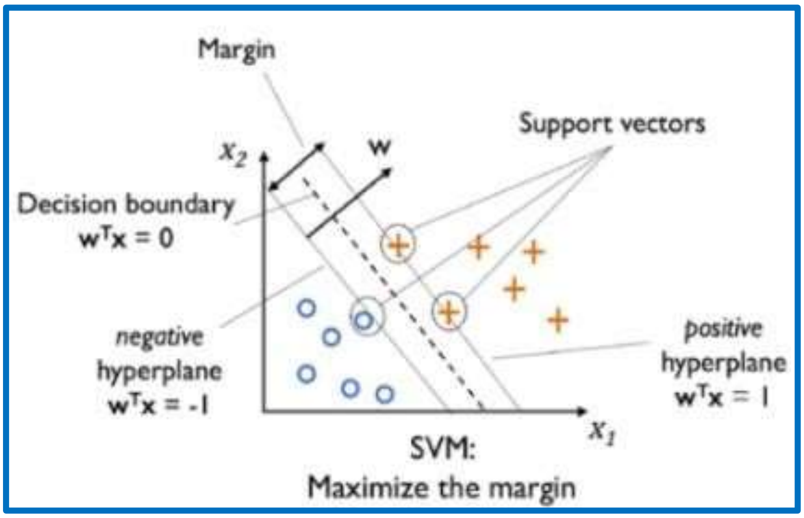


Figure 2. The representation of hyperplane and margin in SVM. Hyperplane was separated by two classes, i.e., positive (+1) and negative (-1) classes.

Figure 2 shows the hyperplane discovered via SVM. The middle position indicates that the distance between the hyperplane and data items is different from the adjacent (outermost) class, which is marked with an empty and positive round. A support vector is the outermost data object closest to the hyperplane in SVM. Because of their nearly overlapping placements with other classes, support vectors are the most difficult to classify. Only this support vector is evaluated by the SVM method to obtain the most optimal hyperplane due to its crucial character. In SVM, a classifier from available samples is constructed by avoiding misclassification in future predictions. In the classifier, the separating hyperplane is written as $\vec{W} \cdot x + b = 0$, which denotes to the equation of $y_i(\vec{W} \cdot \vec{x} + b) \geq 1, i = 1, \ldots, N$. Along the training process, SVM will search for the best separating hyperplane by minimizing $\left(\frac{1}{2}\right) \| \vec{W} \|^2$ in response to the constraint. $\| \vec{W} \|^2$ refers to the Euclidean norm of $\vec{w}$ that optimizes the distance between the hyperplane and support vectors. Using Lagrange multipliers, the SVM training technique is transformed into a convex Quantum Programming (QP) problem. The QP problem's solution will be implemented as a global optimal, which is written as:

$$\vec{W} = \sum_{i=1}^{N} y_i \alpha_i \cdot x_i \quad (4)$$

where: x_i denotes to the support vector when $\alpha_i > 0$. After the training process, the function of the decision applied in forecast is formulated by:

$f(x) = sgn\left(\sum_{i=1}^{N} y_i \alpha_i \cdot x \cdot x_i + b\right)$ (5) where, $sgn()$ denotes to the given sign function.

Furthermore, Cortes and Vapnik represent the slack variable (ζ) through $\zeta_i > 0 = 0, i = 1, \ldots, N$ to consent errors along the training (Cortes & Vapnik, 1995). This method is recognized as a soft margin, which is effective in avoiding overfitting. Focus on the slack variable, the relaxed separation constraint is expressed as:

$$y_i(\vec{W} \cdot \vec{x}_i + b) \geq 1 - \zeta_i = 1, \ldots, N \quad (6)$$

and the ideal hyperplane is gained by minimizing

$$\frac{1}{2} \| \vec{W} \|^2 + C \sum_{N_i} \zeta_i \quad (7)$$

where: C denotes a regularization parameter which controls a trade-off between the optimal margin and training error. Also, to select the best hyperplane, the input vector was projected into a higherdimensional Hilbert space, with the kernel function controlling the operation. RBF, linear, and polynomial kernel functions are the most frequent kernel functions utilized in the SVM model. The polynomial kernel function is written as follows:

$$K(x, y) = (\langle x, y \rangle + 1)^E \quad (8)$$

where: The exponent value is represented by E. The value of the exponent value for the linear kernel is to be 1, while the RBF kernel function can be written as follows:

$$K(x, y) = e^{-(y \cdot \langle x - y, x - y \rangle^2)} \quad (9)$$

Development of a Model for Predicting Students' Achievement

Hyperperparameter Tuning

To improve the development of model, hyperparameter tuning was carried out (Kurniawan et al., 2020; Schratz et al., 2019). Grid search cross-validation was applied to scan parameters during the tuning phase (grid search CV). Table 1 shows a set of parameters and the option value for the LR, DT, and SVM algorithms. In LR, parameter types such as penalty and regularization parameter (C) were used. For DT, the optimum parameters consisted of criterion, samples leaf, and min samples split, while SVM used regularization parameter (C) and kernel coefficient (gamma).

Table 1. The parameters and the option value for the LR, DT, and SVM algorithms.

Method	Parameters	Value options
LR	penalty	['11', '12']
	C	(0.1, 0.3, 0.5, 0.7, 1)
DT	criterion	['gini, 'entropy']
	min_sample_leaf	[1, 2, 3, 4, 5]
	min_sample_split	[2, 4, 8, 10]
SVM	C	[0.1, 1, 10, 100, 1000]
	gamma	[1, 0.1, 0.01, 0.001, 0.0001]

Model validation

Multiple metrics produced from a confusion matrix were utilized to validate the created models. In this letter, true positive (TP) refers to the number of given variables affecting student achievement, whereas true negative (TN) refers to the number of given variables which not affect the student achievement. Several validation measures, including precision (PR), accuracy (Q), specificity (SP), F-score, sensitivity (SE), Area Under Curve (AUC) and Receiver Operating Characteristic (ROC), were estimated to be (Ho et al., 2019; Kurniawan et al., 2020; Luque et al., 2019; Muschelli, 2020). The ROC curve is generated by perusing the TPR and FPR values. Afterward, the area under the ROC curve reflects the AUC value. The equations of those validation metrics are provided in the equation below:

$$Q = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

$$PR = \frac{TP}{TP + FP} \quad (11)$$

$$SE = \frac{TP}{TP + FN} \quad (12)$$

$$SP = \frac{TN}{TN + FP} \quad (13)$$

$$F-score = 2 \times \frac{PR \times SE}{PR + SE} \quad (14)$$

RESULTS DATA ANALYSIS AND SELECTION

The approach of *knowledge data discovery* (KDA) (Cios et al., 1998; Fayyad et al., 1996; Palacios et al., 2021) was applied to analyze the collected data by following several steps such as (1) *Data Selection*, in this step, the data was selected based on the suitable variables; (2) *Preprocessing/Cleaning*, before the machine learning process can be applied, a cleaning process on the data is required for removing the redundant data, finding the inconsistent data, and correcting errors in data, such as typographical errors; (3) *Transformation* is the process of transforming the data that has been chosen so that the data is sufficient for the machine learning process; (4) *Data analysis* was used to estimate the statistical quantity from the data selection; (5) *Pattern Evaluation* is an evaluation stage to identify the data patterns that represent knowledge based on existing data sources. In this present study, several variables, including "gender", "age", "GPA first semester", "GPA second semester", "organization activity", "part time job", "living place", "family income", "father education", "mother education" and "the latest GPA (target)" are considered to find the main factor affecting to the GPA of the students (target). Those variables are common factors that correspond to the academic achievement. From the collected variable, we analyze the correlation between those variables and the target presented in Figure 3. From this figure, all variables contributed to the GPA with different values. However, it is not easy to determine which main factor influences the student performance. Thus, several data mining algorithms such as logistic regression (LR), decision tree (DT), and support vector machine (SVM) are utilized to create the correlation model between the chosen variables and target.

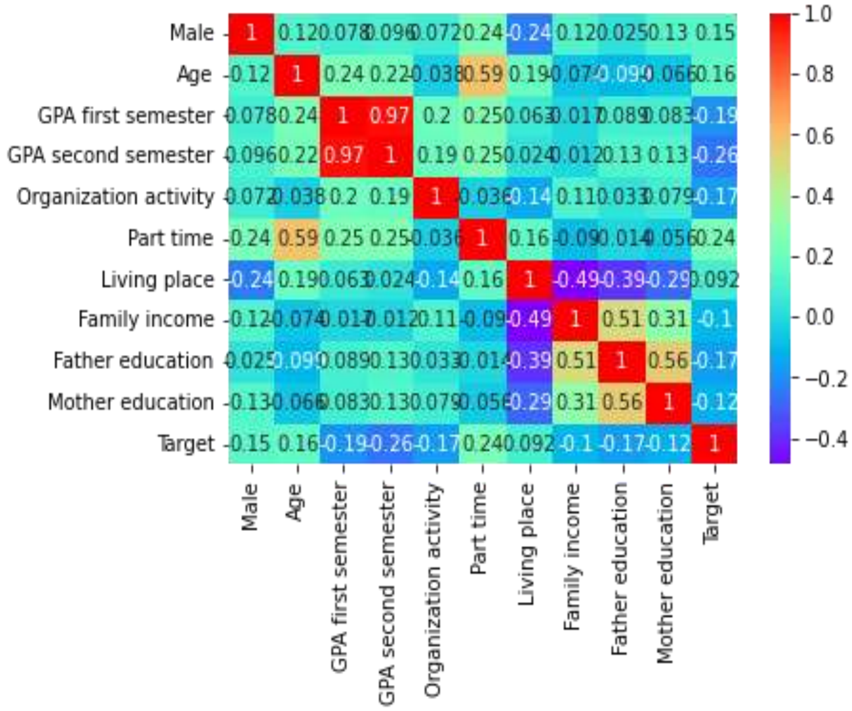


Figure 3. Heatmap of the correlation matrix of targets and the variables

Model Development

Hyperparameter tuning is applied to attain the best parameters for each constructed models. In the LR models, we found that the penalty and C were ['11'] and (0.1), respectively. For DT, the parameters of criterion, min_sample_leaf, and min_sample_split were found as ['entropy'], [2], and [4], respectively. Meanwhile, C and gamma of SVM were to be [0.1] and [1], respectively. The summary of hyperparameter tuning is listed in Table 2.

Figure 4 shows the comparison of accuracy (Q) between non- tuned and tuned models. We observed that only LR has different score between the tuned and non-tuned models. For other models DT and SVM, no difference was found, implying the non-tuned model in the default parameter for those both models are enough to find the best value of each algorithm for the prediction of student achievement.

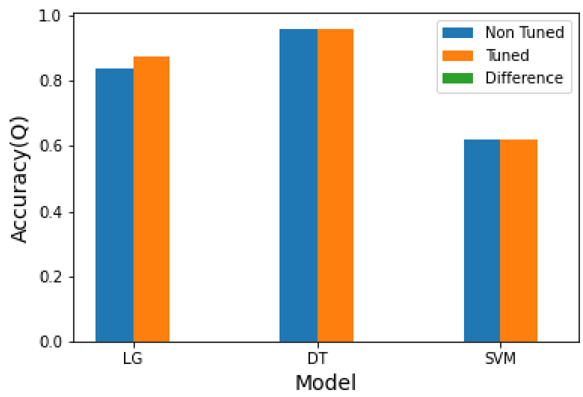


Figure 4. Comparison of accuracy (Q) of Tuned and Non- Tuned Model

Table 2. Parameters and the option values utilized in hyperparameter tuning.

Method	Parameters	Best Value
LR	penalty	['11']
	C	(0.1)
DT	criterion	['entropy']
	min_sample_leaf	[2]
	min_sample_split	[4]
SVM	C	[0.1]
	gamma	[1]

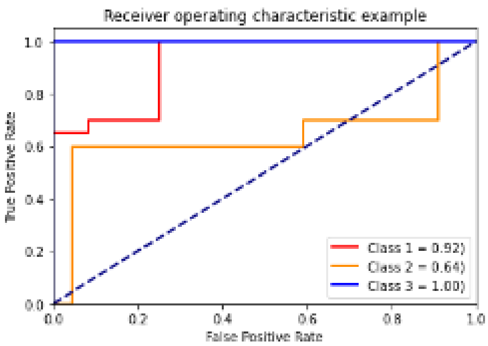
Model Validation

The summary of the model validation is listed in Table 3. The best model was determined based on several validation metrics, i.e., accuracy (Q), precision (PR), sensitivity (SE), and F-score. In the training set, we found that the best metrics are from models developed by decision tree (DT) and support vector machine (SVM), it's because these models have a high value of the metrics compared to results obtained from logistic regression (LR). In the test set, the Q and SE of all algorithms are slight difference values, indicating all models may be acceptable for the model development. The higher scores of accuracies were found for DT model (0.96) suggesting the DT algorithm could be the best model based on those validation metrics.

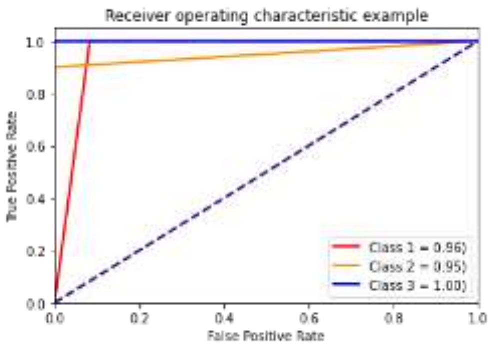
To validate the developed models estimated by the machine learning algorithm, we also consider the AUC of classes 1, 2, and 3. The AUC was estimated by utilizing the ROC presented in Figure 5. The ROC score of all classes for LR and SVM is better than DT. The difference in ROC values of LR and SVM can be observed in classes 2 and 3, respectively. Overall, all models are sufficient data since the AUC value is higher than 0.5 for all classes.

Table 3. Summary of Validation Results obtained from LR, DT and SVM estimations

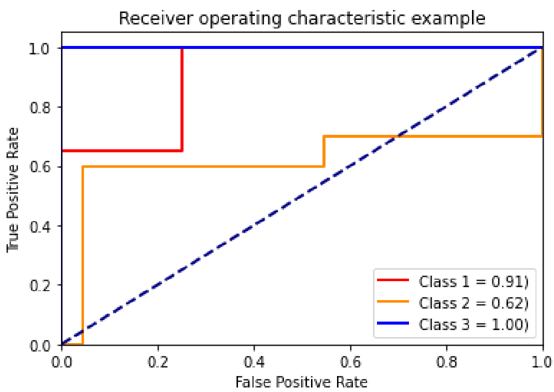
Model	Q	PR	SE	F-score
		Training set		
LR	0.82	0.82	0.82	0.81
DT	1	1	1	1
SVM	1	1	1	1
		Test set		
LR	0.84	0.88	0.84	0.83
DT	0.96	0.97	0.97	0.97
SVM	0.62	0.39	0.62	0.48



(a)



(b)



(c)

Figure 5. ROC Curves for all classes. (a) logistic regression (LR), (b) decision tree (DT), and (c) support vector machine (SVM)

DISCUSSION

One of the goals of higher education is to produce graduates who have superior human resource capabilities. The Grade Point Average (GPA) is a measure of a student's academic ability or achievement in a certain period which is calculated based on the credits that have been taken (Jamelske, 2009). Information regarding predictions of student academic achievement can give an idea of whether the student will succeed in obtaining the expected academic achievement or not. If a student's academic achievement can be known beforehand, even during the selection process, of course it can help universities make decisions (Browning & Rigolon, 2019; Tus, 2020).

Development of a Model for Predicting Students' Achievement

At the time of selection of prospective students, the information can be taken into consideration in deciding whether to be accepted or not. Information and knowledge to support the business processes of a university is increasingly needed. Coupled with the data that is already very large, of course it can motivate to turn this data into useful information and knowledge by extracting or mining knowledge from the data. This information and knowledge can be extracted by utilizing information technology, especially applying machine learning (ML) analysis. This analysis is a technique for making inferences about data with a mathematical approach (Rastrollo-Guerrero et al., 2020; Sekeroglu et al., 2019; Zeineddine et al., 2021). In essence, machine learning is a technique used to create (mathematical) models that describe patterns in data. The inference referred to in machine learning focuses more on the relationships between attributes. In addition, machine learning is a form of depicting data/science/knowledge in a mathematical formulation model. It is called a mathematical model because machine learning science is derived from mathematical and statistical formulas. Machine learning is like a "tool", which is synonymous with a mathematical formula. The way to use it depends on the area of the problem at hand. The use of machine learning has at least 2 objectives: predicting the future (unobserved event): and/or acquiring knowledge (knowledge discovery/discovering unknown structure) (Ray, 2019; Yang & Shami, 2020). Hence, in relation in relation to the student academic performance, machine learning can also be employed to predict the main variables connecting to the student achievement.

The student achievement in academics is very crucial point in education institution since it is a benchmark for the success of a university (Kim et al., 2019; Wu & Shen, 2022). Therefore, investigation of student academic performance is major required by universities to take initial action to improve student performance. Many factors are associated with the academic student that can be investigated. In this research, we considered several variables, i.e., gender, age, GPA first semester, GPA second semester, organization activity, part time job, living place, family income, father education, mother education and the latest GPA (target) connecting to the student academic performance using several algorithms including logistic regression (LR), decision tree (DT), and support vector machine (SVM). From our calculation, the models developed by decision tree (DT) and support vector machine (SVM) produce the finest metrics in the training set due to the fact that these models produce metrics with a higher value than those obtained from logistic regression (LR). In the test set, the Q and SE values for all algorithms are comparable, indicating that all models may be suitable for model development. The DT model had the highest accuracy scores (0.96), suggesting that the DT algorithm may be the best model based on these validation metrics.

We also examine the AUC of classes 1, 2, and 3 in order to validate the developed models estimated by the machine learning algorithm. The AUC was calculated using the ROC displayed in Figure 5. All LR and SVM classes have a higher ROC value than DT. In classes 2 and 3, respectively, the disparity in ROC values between LR and SVM can be observed. Overall, all models provide considerable data, as the AUC value for all classes is greater than 0.5. This finding demonstrated that the developed model from the above variables is associated with student achievement. Therefore, attention to these factors is essential for increasing the student academic performance.

CONCLUSIONS

Three methods, i.e., logistic regression (LR), decision tree (DT), and support vector machine (SVM) were employed to investigate the several factors affecting the student's achievement. The number of variables was determined and documented at the 3nd-semester student at the history education. We found that the performance improvement of the models between the non-tuned and tuned models is similar value. This finding implied that the default setting of the parameter is enough to obtain the best value of each algorithm for the prediction of student achievement. Moreover, according to the validation results, the best algorithm method is found in the model developed by the decision tree with accuracy 0.96. This score is the highest scores compared with other values obtained by LR (0.84) and SVM (0.62), respectively. This outcome revealed that those variables are related to student achievement. As a consequence, focusing on such elements is important for enhancing the academic achievement.

ACKNOWLEDGMENTS

The authors express thanks to Tadulako University for financial support to work on this research.

CONFLICTS OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1) Abonazel, M. R., & Ibrahim, M. G. (2018). On estimation methods for binary logistic regression model with missing values. *International Journal of Mathematics and Computational Science*, *4*(3), 79–85.
2) Browning, M. H. E. M., & Rigolon, A. (2019). School green space and its impact on academic performance: A systematic literature review. *International Journal of Environmental Research and Public Health*, *16*(3), 429.
3) Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189–215.

4) Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28.
5) Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (1998). Data mining and knowledge discovery. In *Data mining methods for knowledge discovery* (pp. 1–26). Springer.
6) Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018
7) Denœux, T. (2019). Logistic regression, neural networks and Dempster–Shafer theory: A new perspective. *Knowledge-Based Systems*, *176*, 54–67.
8) Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37.
9) Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S., & Sarker, K. U. (2018). Student academic performance prediction by using decision tree algorithm. *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, 1–5.
10) Ho, D. S. W., Schierding, W., Wake, M., Saffery, R., & O'Sullivan, J. (2019). Machine learning SNP based prediction for precision medicine. *Frontiers in Genetics*, *10*, 267.
11) Jamelske, E. (2009). Measuring the impact of a university first-year experience program on student GPA and retention. *Higher Education*, *57*, 373–391.
12) Kim, H. J., Hong, A. J., & Song, H.-D. (2019). The roles of academic engagement and digital readiness in students' achievements in university e-learning environments. *International Journal of Educational Technology in Higher Education*, *16*(1), 1–18.
13) Kurniawan, I., Rosalinda, M., & Ikhsan, N. (2020). Implementation of ensemble methods on QSAR Study of NS3 inhibitor activity as anti-dengue agent. *SAR and QSAR in Environmental Research*, *31*(6), 477–492.
14) Luque, A., Carrasco, A., Martín, A., & de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231.
15) Muschelli, J. (2020). ROC and AUC with a binary predictor: a potentially misleading metric. *Journal of Classification*, *37*(3), 696–708.
16) Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy*, *23*(4), 485.
17) Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. *International Journal of Computer Sciences and Engineering*, *6*(10), 74–78.
18) Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101– 121). Elsevier.
19) Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, *10*(3), 1042.
20) Ray, S. (2019). A quick review of machine learning algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39.
21) Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120.
22) Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student performance prediction and classification using machine learning algorithms. *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, 7–11.
23) Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research*, *5*(1), 1–16.
24) Tus, J. (2020). Academic stress, academic motivation, and its relationship on the academic performance of the senior high school students. *Asian Journal of Multidisciplinary Studies*, *8*(11), 29–37.
25) Wu, H., & Shen, J. (2022). The association between principal leadership and student achievement: A multivariate meta-meta-analysis. *Educational Research Review*, *35*, 100423.
26) Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316.
27) Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, *89*, 106903.