Comparative Study of KNN, SVM and Decision Tree Algorithm for Student's Performance Prediction

Slamet Wiyono, Dega Surono Wibowo, M. Fikri Hidayatullah and Dairoh

***Abstract*—Students who are not-active will affect the number of students who graduate on time. Prevention of not-active students can be done by predicting student performance. The study was conducted by comparing the KNN, SVM, and Decision Tree algorithms to obtain the best predictive model. The model making process was carried out by the following steps: data collecting, pre-processing, model building, comparison of models, and evaluation. The results show that the SVM algorithm has the best accuracy in predicting with a precision value of 95%. The Decision Tree algorithm has a prediction accuracy of 93% and the KNN algorithm has a prediction accuracy value of 92%.**

***Index Terms*—KNN, SVM, decision tree.**

I. Introduction

IMPROVING the quality of education and accreditation of departments is always endeavored by every college department. Timeliness of graduating students is one of the elements for accreditation assessment [1]. The accreditation will be better if more students graduate on time. Students who are not-active will affect the number of students who graduate on time. Thus, the more students who graduate not on time will the lower the department's accreditation.

Prevention of not-active students can be done by predicting student performance. Several studies on student performance had been conducted. Some studies use Data Mining algorithm. Data Mining algorithm was used to perform student performance analysis system (SPAS) [2], to analyze student performance using clustering techniques [3], and to predict student performance (poor, average, good, and excellent) using educational data [4]. Other research by applying Decision Tree algorithms such as: predictions of drop-out students from college based on GPA [5], analysis to predict the accuracy of 4-year studies of student [6]. Other research to predict student performance at the beginning of joining a course program [7], predicting the student performance in distance higher education using active learning [8], predictions of student performance correlated with course activities [9], and predicting student performance using advanced learning analytics to compare features [10]. In addition to the Data Mining algorithm, using the Fuzzy method is also done to predict student performance. Fuzzy Support System method was used for evaluation of student performance in laboratory [11], and an application of fuzzy logic for evaluation of student academic performance [12].

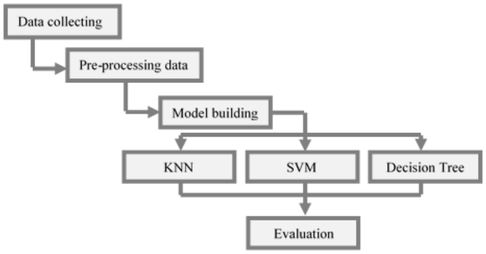

Fig. 1: Step of the research process.

Research by comparing several algorithms to obtain the best predictions has been done. Among had been done is; comparing Simple Logistic Classifier and SVM algorithms to predict athlete's win [13] at, comparative analysis between SVM and KNN classifier for EMG signal classification [14], compare KNN, SVM, and Random Forest algorithms for facial expression classification [15]. Comparative algorithm research for predicting student performance had also been carried out. Among them have been done are; look for classification algorithm that can be used to predict student performance [16], comparing Bayesian algorithm and Decision Tree [17], compare Apriori and K-Means algorithms [18], and compare Neural Network, SVM, and Decision Tree algorithms [19]. Comparison of KNN, SVM, and Decision Tree algorithms [20]. Recent research comparing the KNN, SVM, and Decision Tree algorithms concludes that the SVM algorithm has the best accuracy. The study used $K = 5$ on the KNN algorithm. Further research is deemed necessary to proceed with trying out different K. This paper is a continuation of previous research, which compares the accuracy of the KNN, SVM, and Decision Tree algorithms by changing the K value in the KNN algorithm.

II. Methods

This research had been done using several Machine Learning algorithms, namely KNN, SVM, and Decision Tree. The tools used are R Studio. The library used in the R Studio is the Caret package. Machine Learning processing through several processes: data collecting, pre-processing, model building, comparison of models, and evaluation [21]. The research process is shown in Figure 1.

Manuscript received October 8, 2018; accepted August 5, 2020.

The authors are with the Politeknik Harapan Bersama, Jl. Mataram No. 9, Pesurungan Lor, Kota Tegal, Indonesia. E-mail: slamet2wiyono@gmail.com

This research was supported by Ministry of Research, Technology and Higher Education (RISTEKDIKTI) Indonesia.

TABLE I: Detail of Dataset

No	Feature Title	Variable Data Type	Feature Categorization
1	GP	Continuous	0 - 4
2	GPA	Continuous	0 - 4
3	hometown	Categorical	1: city close from campus 0: city near from campus
4	type of school	Categorical	1: public school 0: private school
5	major	Categorical	1: computer/informatics 2: science major 3: others
6	parents job	Categorical	1: civil servant 2: employee 3: entrepreneur 4: farmer / fisherman 5: others
7	aktif	Categorical	1: active 0: non-active

Data collection is conducted by combining all data into one with the same attributes. The data used are: GP (grade point), GPA (grade point average), hometown, type of school, majors, parent's work, and student performance (active/non-active). Pre-processing is used to improve the data before building a Machine Learning model. Problems in data are usually like different attributes, missing values, etc. Pre-processing is also done by splitting the data into training and testing. Training data is used to build models. The model that has been built is then tested using data testing to determine the accuracy of the prediction. The next step is to compare several models that have been built, namely the model of the KNN , SVM, and Decision Tree algorithm. The final step is to evaluate to determine the best algorithm for predicting student's performance based on the model obtained.

III. Results

Student academic data of Informatics Engineering Department Politeknik Harapan Bersama are used in this paper. The dataset consists of 1530 rows and 7 attributes data. First 6 variables had been used for predicting the 7th variable. Table I shows all the details of data.

GP (Grade Points) is the average score of learning outcomes in every semester, 0 means the lowest score and 4 means the highest score. GPA (Grade Points Average) is the cumulative average point value of all semesters that have been passed, 0 means the lowest score and 4 means the highest score. Hometown is the hometown of students, 0 means student coming from a city that is near from campus and 1 means student coming from a city far away from campus. Type of school is a type of high school, 0 means students come from private schools and 1 means students come from public schools. Major is majors when high school, 1 means students come from the computer/informatics department, 2 means students come from natural science majors, and 3 mean students come from other than both. Parents jobs are jobs from student parents, 1 means parents work as civil servants, 2 means as private employees, 3 mean as entrepreneurs, 4 means as farmers/fishermen, and 5 mean other than that. Active is student performance, 0 means students are not active and 1 means students are active.

TABLE II: Model Result

Algorithm	Result	Accuracy
KNN	$k = 3$	94.50%
SVM	value $C = 1$	95.09%
Decision Tree	cp = 0.6689113	95.65%

TABLE III: Confusion Matrix for KNN

Prediction	Reference	
	active	non-active
active	309	14
non-active	7	52

TABLE IV: Confusion Matrix for SVM

Prediction	Reference	
	active	non-active
active	311	13
non-active	5	53

TABLE V: Confusion Matrix for Decision Tree

Prediction	Reference	
	active	non-active
active	308	18
non-active	4	48

A. Model Result

Before the data is processed, the data set is split into two parts by a ratio of 75:25, which 75% to training and 25% to testing. Training data used to construct the model. Training data used were 1148 samples, 6 predictor, and 2 classes, with cross-validation 10 fold and repeated 3 times. Output of training data is a model used for classification. The model that had been built is shown in Table II.

The model was then tested used testing data to know how accurate that model. Table III shows a matrix of the testing result for KNN algorithm, Table IV is testing result for SVM algorithm, and Table V is testing result for Decision Tree algorithm.

B. Classification Results

Classification result is obtained from the model that has been tested. Table VI shows the comparison of the testing result between KNN, SVM, and Decision Tree algorithm on the confusion matrix. Figure 2 shows the comparison accuracy between algorithm based on classes.

The final result is a comparison of model classification to see which algorithm has the best accuracy. Table VII shows the comparison of the classification model obtained.

IV. Discussions

The best model for KNN algorithm to predict student performance is $k = 3$ (kernel) with accuracy 94.5%, value

TABLE VI: Comparison of Confusion Matrices

Prediction		KNN	SVM	Decision Tree
Active	TRUE	96%	96%	94%
	FALSE	4%	4%	6%
Non-Active	TRUE	88%	91%	92%
	FALSE	12%	9%	8%

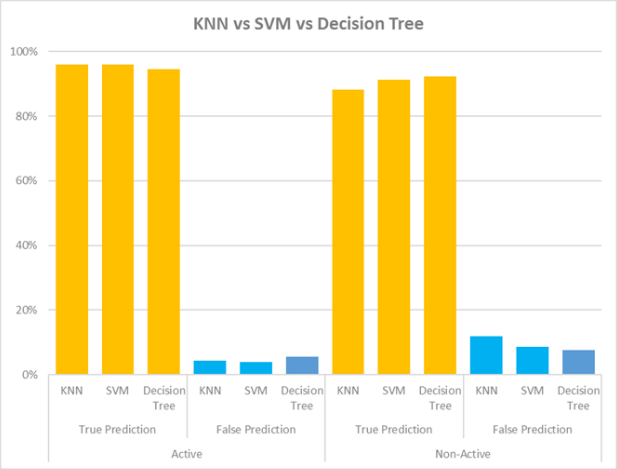

Fig. 2: Comparison of testing accuracy.

TABLE VII: Classification Accuracy Comparison

Accuracy		
KNN	SVM	Decision Tree
94.5%	95%	93%

$C = 1$ for SVM algorithm with accuracy 95.09%, and cp = 0.6689113 for Decision Tree algorithm with 95.65% accuracy. The comparison of the three algorithms shows that the best accuracy is the Decision Tree algorithm. This model has not been tested yet. After testing, it turns out that the SVM model can predict better than the KNN algorithm and Decision Tree. It can be seen that the SVM algorithm can predict exactly 311 active students and 53 non-active students, while the KNN algorithm only predicts exactly 309 active students and 52 non-active students, and the Decision Tree algorithm can only predict exactly 308 active students and 48 non-active students. If not testing the model, the Decision Tree is the best predictive accuracy model compared to SVM and KNN. Whereas if the model testing is done, SVM algorithm is the best accuracy model compared to KNN and Decision Tree.

Comparison with matrix confusion shows different things from the results of previous comparisons. SVM algorithm has the best accuracy to predict active students (96%) compared to KNN (96%) and Decision Tree (92%). However, the Decision Tree algorithm has the best accuracy for predict non-active students (92%) compared to SVM (91%) and KNN (88%). Although Decision Tree algorithm has the best accuracy in predicting non-active students, but only 1% difference from SVM algorithm. While for predicting the accuracy of active students, SVM has a 4% difference from Decision Tree and KNN. It could be said that the SVM algorithm still occupies the best position compared to KNN and Decision Tree. This is corroborated after the overall accuracy calculation is performed, it is found that SVM has the best classification accuracy of 95% while KNN has 94.5% accuracy and Decision Tree has 93% accuracy. Thus, the best algorithm for predicting student performance is by using the SVM algorithm.

V. Conclusions

KNN algorithm can predict student performance well with $k = 3$. The best model of SVM algorithm to predict student's performance is by using the value of $C = 1$. Whereas if using the Decision Tree algorithm, the best predictions if using the model cp = 0.6689113. Comparison of three algorithm machine learning (KNN, SVM, and Decision Tree) shows that SVM has the best accuracy (95%) compared to KNN (94.5%) and Decision Tree (93%) in predicting student's performance.

References

[1] B. A. N. P. Tinggi, "Buku i naskah akademik. akreditasi institusi perguruan tinggi. jakarta," 2019.

[2] C. Sa, D. Ibrahim, E. Hossain, and M. bin Hossin, "Student performance analysis system (spas)," in *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, 2014, pp. 1–6.

[3] I. Singh, A. Sabitha, and A. Bansal, "Student performance analysis using clustering algorithm," in *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, 2016, pp. 294–299.

[4] T. Devasia, T. Vinushree, and V. Hegde, "Prediction of students performance using educational data mining," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 2016, pp. 91–95.

[5] M. Quadri and N. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global Journal of Computer Science and Technology*, 2010.

[6] R. Asif, A. Merceron, S. Ali, and N. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, pp. 177–194, 2017.

[7] A. Vihavainen, "Predicting students' performance in an introductory programming course using data from students' own programming process," in *2013 IEEE 13th International Conference on Advanced Learning Technologies*, 2013, pp. 498–499.

[8] G. Kostopoulos, A.-D. Lipitakis, S. Kotsiantis, and G. Gravvanis, "Predicting student performance in distance higher education using active learning," in *International Conference on Engineering Applications of Neural Networks*, 2017, pp. 75–86.

[9] R. Conijn, A. Van den Beemt, and P. Cuijpers, "Predicting student performance in a blended mooc," *Journal of Computer Assisted Learning*, vol. 34, no. 5, pp. 615–628, 2018.

[10] A. Daud, N. Aljohani, R. Abbasi, M. Lytras, F. Abbas, and J. Alowibdi, "Predicting student performance using advanced learning analytics," in *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 415–421.

[11] Z. Yıldız and A. Baba, "Evaluation of student performance in laboratory applications using fuzzy decision support system model," in *2014 IEEE Global Engineering Education Conference (EDUCON)*, 2014, pp. 1023–1027.

[12] N. Meenakshi and N. Pankaj, "Application of fuzzy logic for evaluation of academic performance of students of computer application course," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 3, no. 10, pp. 260–7, 2015.

[13] E. Rainarli and A. Romadhan, "Perbandingan simple logistic classifier dengan support vector machine dalam memprediksi kemenangan atlet," *Journal of Information Systems Engineering and Business Intelligence*, vol. 3, no. 2, pp. 87–91, 2017.

[14] Y. Paul, V. Goyal, and R. Jaswal, "Comparative analysis between svm & knn classifier for emg signal classification on elementary time domain features," in *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*, 2017, pp. 169–175.

[15] R. Nugrahaeni and K. Mutijarsa, "Comparative analysis of machine learning knn, svm, and random forests algorithm for facial expression classification," in *2016 International Seminar on Application for Technology of Information and Communication (ISemantic)*, 2016, pp. 163–168.

[16] V. Patil, S. Suryawanshi, M. Saner, V. Patil, and B. Sarode, "Student performance prediction using classification data mining techniques," *International Journal of Scientific Development and Research*, vol. 2, no. 6, pp. 163–167, 2017.

[17] A. Khasanah and A. Harwati, "A comparative study to predict students performance using educational data mining techniques," in *IOP Conference Series: Materials Science and Engineering*, vol. 215, no. 2, 2017, pp. 1–7.

[18] G. Gowri, R. Thulasiram, and M. Baburao, "Educational data mining application for estimating students performance in weka environment," in *IOP Conference Series: Materials Science and Engineering*, vol. 263, 2017, p. 032002.

[19] M. Ciolacu, A. Tehrani, R. Beer, and H. Popp, "Education 4.0fostering student's performance with machine learning methods," in *2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME)*, 2017, pp. 438–443.

[20] S. Wiyono and T. Abidin, "Comparative study of machine learning knn, svm, and decision tree algorithm to predict students performance," *International Journal of Research - Granthaalayah*, vol. 7, no. 1, pp. 190–196, 2019.

[21] B. Lantz, *Machine learning with R*. Packt publishing ltd, 2013.