Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 4, No. 2, September 2021, pp. 62 – 71
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

C4.5, K-Nearest Neighbor, Naïve Bayes and Random Forest Algorithms Comparison to Predict Students' On Time Graduation

[1]Gunawan, [2]Hanes, [3]Catherine
[1] Department of Informatics Engineering, STMIK Mikroskil
[2,3] Department of Information Systems, STMIK Mikroskil
Email: [1]gunawan@mikroskil.ac.id, [2]hanes@mikroskil.ac.id, [3]catherine.yang@mikroskil.ac.id

Article Info

Article history:
Received Sept 25th, 2020
Revised March 4th, 2021
Accepted May 28th, 2021

Keyword:
C4.5
Data Mining Classification
K-Nearest Neighbor
Naïve Bayes
Random Forest
Timely Graduation

ABSTRACT

Study program performance can be seen from the achievement of accreditation status, where one of the assessment instruments related to the graduate profile is the study length. Graduation on time is one indicator of student's success in obtaining a bachelor's degree and is an important attribute, because by being able to predict the period of study, universities can minimize student graduation failures by making more intensive planning, study escort, and guidance. Data mining classification techniques can be used to predict students graduation on time. Many data mining classification algorithms can be used, so it is necessary to make comparisons to determine the level of accuracy of each algorithm. The algorithms that will be compared in this study are C4.5, K-Nearest Neighbor, Naive Bayes, and Random Forest. The data used were 2,022 graduates from Informatics Engineering and Information Systems Undergraduate Study Programs of STMIK Mikroskil Medan from 2011 to 2014, in which the attributes used include gender, region of origin, study time, grade of entrance examination, and Grade Point Average (GPA). The results of the classification process are evaluated using cross validation and confusion matrix to determine the most accurate data mining classification algorithm for predicting students graduation on time, where the K-Nearest Neighbor and Random Forest algorithms have the highest accuracy of 72,651%, followed by the C4.5 algorithm of 72,453%, and the Naïve Bayes algorithm of 71,860%.

Corresponding Author:
Gunawan
Department of Informatics Engineering,
STMIK Mikroskil,
Jl. Thamrin No. 112 Medan, Sumatera Utara, Indonesia.
Email: gunawan@mikroskil.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v4i2.10833

1. INTRODUCTION

The process of digitizing the flow of information that flows quickly is tailored to the needs of today's users. Continuity of business processes in companies and higher education requires the use of a lot of data. In the field of education, many users use the internet to get good quality information. Students get a digital learning process that is getting faster through the support of internet technology so they can attend lectures well. In addition, students can monitor and evaluate their learning methods so that the level of knowledge absorption is higher. In fact, parents are able to keep up with the development of student knowledge and learning experiences that are better and more actual. With the development of information technology, there are many forms of education that encourage students to be better prepared for the increasingly advanced industrial revolution. From the use of e-learning media, academic portals, e-mail, to the application of higher education which is facilitated by the internet, it is able to develop the potential of students which will later affect the community.

The higher the implementation of information technology followed by the student learning process, the higher the level of educational data storage that is accessed every time. Every educational institution in every country has different curriculum rules and must be followed by each student. Curriculum items followed by students are stored and managed in an academic database. The academic database has the distribution or distribution of information for each attribute that is related to the student entity. Academic databases not only store Grade Point Average (GPA) scores per semester and Grade Point Average (GPA) scores, but also store data to support other student performance and consider the attributes of student entities, such as gender, age, interest in learning, and so on. Seeing the importance of student attributes in each semester for the timeliness of completing the study period, the concept of data mining is needed to process the data.

The use of EDM is very important to deal with current education problems. Students have difficulty learning something new through curriculum changes. Academic procedures that must be adjusted to the achievement of student goals can be one of the obstacles to education. The distribution of knowledge and experience of each lecturer varies which can cause students to be unable to respond as a whole. Differences in student potential and competence that are not explored effectively and the less controlled control of the learning system can have a negative impact on student characteristics. As a result, the GPA scores obtained by students per semester have decreased significantly. Changes in lecture rules experienced by students will later affect the graduation rate on time. This will have a negative effect on student performance. Therefore, it is necessary to implement EDM in academic databases to determine the best estimate of the pattern of students passing on time. With the EDM to predict the timely graduation rate of students who are collected, detected, and measured at a certain time, so that it will approach the actual value of the accumulated value of student performance. The form of transformation of student academic data can produce educational patterns for the use of case simulations as a standard or better curriculum rule. It is very important to do EDM processing to predict the passing rate on time to determine how high the student's performance is.

Prediction techniques for student graduation on time that consider student entities in each semester can use the classification method. The results of calculating the predictions of graduation on time can provide new opportunities for educational institutions to minimize student laziness in completing their studies. The classification method is a predictive technique that utilizes machine learning to retrieve sensitive data that is directly related to the entity [1][2]. Classification can be used for the filtering process of large, unprocessed data sets resulting in an estimated value as an intelligible representation form accurately [3][4][5]. Classification technique consists of a training part which involves a training set containing all the attributes and classes as well as a testing part that defines a new class record for an unknown attribute pattern [6][7]. There are several algorithms in the implementation of classification techniques, such as Decision Tree, Naïve Bayes, Rule-Based, Neural Network, Support Vector Machine, K-Nearest Neighbor, Random Forest, Random Tree, and so on [6][8][9][7][4][10]. Classification can also be used to predict failure rates that may occur during system implementation [11].

Based on the description above, there are many data mining classification algorithms that can be used, but not all of these algorithms have good performance, so it is necessary to make comparisons to determine the level of accuracy of each algorithm. The results of the predictions made by each algorithm differ depending on the determination of the multi-attribute class. This study focuses on comparing the performance of the C4.5, K-Nearest Neighbor, Naïve Bayes, and Random Forest classification algorithms to predict students' timely graduation using case studies of academic data from graduates of the STMIK Mikroskil Informatics Engineering and Information Systems Undergraduate Study Programs.

2. METHODOLOGY

Research methodology is a method used to carry out the steps to be carried out in research and writing, because it reflects the interrelation of the steps so that writing will be easier, more directed, and systematic. The stages carried out in this study can be seen in the following figure.

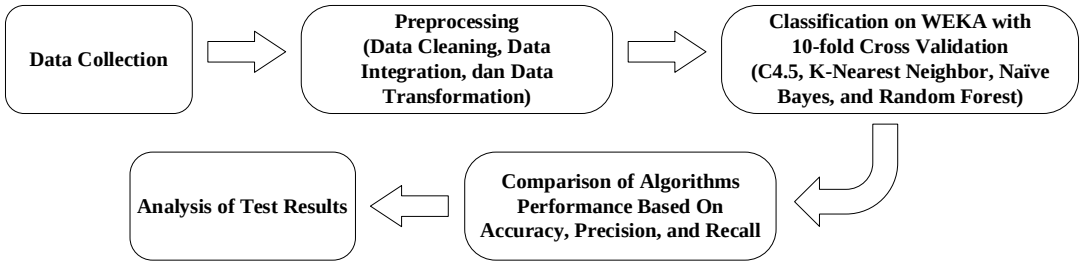

Figure 1. Research Design

2.1. Literature Review

2.1.1. Data Mining and EDM

Data mining is part of the Knowledge Discovery in Database (KDD) which has a process of exploitation and extraction of large data sets that are formulated into a format so that they can be processed in more detail [12][8][13]. Data mining has a supervised nature that is in line with the model and data type for analysis needs and unsupervised properties that form a pattern according to the type of data used [12]. In general, data mining has two categories of tasks, namely descriptive and predictive tasks, where predictive modeling is more suitable for determining target variables from a set of unknown data [8][14][15]. Data mining is more effective at finding patterns and capturing consumption trends to support organizational decision-making processes [16][17]. There are data mining techniques that can be used to form information building, such as decision trees, neural networks, clustering, sequence mining, regression, relationship mining, descriptive statistics, visualization, and so on [12][15][18].

In the education world, performance, knowledge, or predicted values have an important role as one of the attributes of the student entity through the management of Educational Data Mining (EDM) [18]. EDM is taken from academic databases which can be processed further to predict the timely graduation of each student. The digital learning process and the provision of online student assignment instructions allow for optimal EDM in data collection. The curriculum improvement per semester and each grade given by the lecturer is an important factor in calculating the prediction of graduation on time. In addition, the involvement of student, lecturer, and administrative support entities has different EDM perspectives which are further analyzed to produce knowledge of the prediction of passing scores close to the actual value [9][10]. Just like data mining in general, EDM consists of preprocessing, data mining, and post-processing processes with the scope of improving student and domain models, learning software skills, and scientific research that focuses on the learning process [15][18]. This facilitates the prediction technique of student graduation on time which refers to an increase in student academic performance [21].

2.1.2. C4.5 Algorithm

The C4.5 algorithm is a continuation of the Iterative Dichotomiser (ID3) and the application of a decision tree with a heuristic technique approach in the form of quantitative and categorical data [12][8][21][22]. This algorithm makes use of the J48 rule that determines the related attributes in dealing with problems and setting pruning parameters on and off [12][23][24]. The C4.5 algorithm utilizes the divide and conquer method for decision tree fit construction and applies the binarization process to the data partition so that there is an increase in the estimated error and speed [19][10][25]. As part of a supervised algorithm, C4.5 considers a subset of training, input, and output expectations, as well as optimizing the resulting solutions [19][3][22]. For the calculation of the C4.5 algorithm, it is determined from the gain ratio value which is concerned with the position of the node [4][25][26][27][28][29]. In the calculation formulation, the gain information also considers the missing attribute value and performs splitting of several attributes, weighting the characteristics of user interest, and other similarities of attributes [3][22][23].

2.1.3. K-Nearest Neighbor Algorithm

K-Nearest Neighbor (KNN) classification is the simplest and easiest learner algorithm for managing datasets [26][30][31][32]. KNN can produce more detailed estimation pattern data and matters relating to tertiary education [33]. When executing the algorithm, there is a memory-based technique that determines the weight of the training point as the top priority and emphasizes the distance traveled between the data points being examined [3][26][34][35]. The pseudo code to process the KNN of training data [3][32][34]:

1. Identify the value of K in a class
   K → number of nearest neighbor
2. Calculate the Euclidean distance of each class
   For Each Object Z from Test Set Do
   Calculate distance Y(i,j) between i and every object j in training set neighborhood, then K neighbor closest to i
   i.Class → SelectClass (Neighborhood)
   For the calculation of Euclidean distance using the following formula.

$$dist(y_i - y_j) = \sqrt{\sum_{z=1}^{n}(y_{iz}, y_{jz})^2} \quad (1)$$

3. Determine the K Neighbor attribute with the lowest Euclidean distance value

The processes in the KNN algorithm can compress the Nearest Feature Line (NFL), K-d tree, Principal Axis Search Tree, Ball Tree, and Orthogonal Search Tree to produce the closest k points in the formation of new observations [3][26].

2.1.4. Naïve Bayes Algorithm

Naïve Bayes (NB) uses Bayes probability theory and likelihood estimation where the parameter value has an independent character but is dependent on other parameters to form a new next probability [19][6][7][29][30][36]. Generally, the level of efficiency and ease of use of Naïve Bayes can be done to handle real-world conditions, such as detecting spam, determining the best recommendation system, classifying data, and so on [13][1][26][36]. Similar to the Nearest Neighor algorithm, Naïve Bayes can be evaluated through the OneR attribute which is a strong and simple classifier [28][30]. The higher the level of accuracy obtained if the database size is larger for the Naïve Bayes calculation which affects the interpretation of the output [13][3][6]. Naïve Bayes classification calculation formula by considering the probability of each class [3][7][26][29].

$$P(x_i|y) = \frac{p(y|x_i)\, p(x_i)}{p(y)} \tag{2}$$

When executing descriptive and predictive training data, it is possible to add new datasets [1][3].

2.1.5. Random Forest Algorithm

Random Forest (RF) is an ensemble method in terms of learning that is used for classification, regression, and other tasks. RF was first proposed by Tin Kam Ho and further developed by Leo Breiman and Adele Cutler. RF is a method that can improve accuracy results, because generating child nodes for each node is done randomly [37]. The advantages of RF include being able to produce lower errors, provide good classification results, be able to efficiently handle very large amounts of training data, and an effective method for estimating missing data [38]. This method is used to build a decision tree consisting of root nodes, internal nodes, and leaf nodes by taking attributes and data randomly according to the applicable provisions. Root node is the node that is located at the top, or commonly referred to as the root of the decision tree. Internal nodes are branch nodes, where this node has at least two outputs and only one input. Meanwhile, a leaf node or terminal node is the last node that has only one input and no output. The decision tree begins by calculating the entropy value as a determinant of the level of impurity attributes and the value of information gain. To calculate entropy, a formula is used as in the following equation.

$$Entropy\ (Y) = -\sum_i p(c|Y)\ \log_2 p(c|Y) \tag{3}$$

where Y is the set of cases and p (c | Y) is the proportion of Y values to class c. Meanwhile, to calculate the value of information gain the following equation is used.

$$Information\ Gain\ (Y, a) = Entropy\ (Y) - \sum_{v \in Values(a)} \frac{|Y_v|}{|Y_a|}\ Entropy\ (Y_v) \tag{4}$$

where Values (a) are all possible values in the set of cases a. Yv is a subclass of Y where class v corresponds to class a. Yes are all values corresponding to a [39].

2.2. Data Collection

Data obtained from secondary data in the form of Higher Education Information System database which is managed by the STMIK Mikroskil Information Systems Center. In a previous study [40], researchers have examined several variables that can be used to predict students graduation on time, namely gender, region of origin, time of study, grade of entrance examination, and Grade Point Average (GPA). In this study, the authors compared again the same variables used in the previous study.

The main data source used in this study is the graduates data from Informatics Engineering and Information Systems Undergraduate Study Programs from 2011 to 2014. In this database, there are a number of tables, but not all of these tables were included in this study. The tables that are involved depend on the data to be used in this research, namely Student Identification Number, student name, gender, region of origin, grade of entrance examination, study program, study time, Grade Point Average (GPA), and study length. Gender, region of origin, study time, grade of entrance examination, and GPA will be used as attributes, while the study length will be used as a class.

2.3. Preprocessing

At this stage, data preparation is carried out by preprocessing, so that the data is ready to be processed, namely for training and testing.

2.3.1. Data Cleaning

Data cleaning is carried out to clean noise in the data, in the form of missing values, data inconsistencies, and data redundancy. All attributes specified in the previous section will be selected to obtain attributes that contain relevant values, are not missing value, and are not redundant, where these three conditions are the initial requirements that must be done in data mining so that a clean dataset will be obtained to be used in data mining stage. In this study, the data cleaning process has been arranged in such a way that any data about students who have graduated, including Student Identification Number, student name, gender, region of origin, grade of entrance examination, study program, study time, GPA, and study length are not empty and consistent.

2.3.2. Data Integration

Data integration is carried out with the aim of moving all cleaned data into one file so that it will facilitate the data mining process. At this stage, all the tables involved will be integrated to form a single file as a dataset that will be used for the data mining analysis process, namely student graduate data. In this study, the integration of data from a number of tables involved was carried out using the Structured Query Language (SQL) and the results were stored in a single file in Microsoft Excel format.

2.3.3. Data Transformation

Data transformation is performed to convert data into values with a certain format. In addition, there are data whose coverage is too broad so that it needs to be grouped into several small groups. Data on the origin of the region and the city of origin of the school is data that has a broad coverage. The region of origin will be divided into 2 (two) groups, namely "Medan" and "Outside Medan". The GPA will be divided into 3 (three) groups, namely "Cum Laude", "Very Good", and "Good". The scores per group are divided based on the STMIK Mikroskil Academic Regulations. For students with a GPA of more than 3.50 will be classified as "Cum Laude", students with GPA between 3.01 to 3.50 will be classified as "Very Good", and students with a GPA of less than 3.01 will be classified as "Good". In this study, a student's graduation will be divided into 2 (two) groups, namely "On Time" and "Not On Time".

2.4. Classification

Classification is implemented using the Waikato Environment for Knowledge Analysis (Weka) data mining tool [41] with the 10-fold cross validation test method (the number of standard folds found in WEKA) using the C4.5, K-Nearest Neighbor, Naïve Bayes, and Random Forest classification algorithms. The 10-fold cross validation test method will divide the training data and test data as much as 10 (ten) parts of the data. The data used were randomly divided into 10 (ten) subsets of the same size. The training and testing process will be repeated 10 (ten) times.

The dataset used for both the training and testing process was 2,022 records, which came from the data from the undergraduate students in Informatics Engineering and Information Systems Study Programs. The data that comes from the transformed data file saved with the Comma Separated Values (CSV) extension is first converted to the .arff format known by the Weka tool.

2.5. Comparison of Algorithms Performance and Analysis of Test Results

Classification performance measurement is done by evaluating the test results using a confusion matrix. Confusion matrix is a method for evaluating classification models to estimate whether objects are true or false. The following table provides a two-class confusion matrix.

Table 1. Two-Class Confusion Matrix

Classification	*Predicted Class*	
	Class = on-time	*Class = not-on-time*
Class = on-time	a (*True Positive*)	b (*False Negative*)
Class = not-on-time	c (*False Positive*)	d (*True Negative*)

In the above table, true positive (TP) is the number of positive records classified as positive, false positive (FP) is the number of negative records classified as positive, false negatives (FN) is the number of positive records classified as negative, while true negatives (TN) is the number of negative records that are classified as negative. The higher the TP and TN values, the better the classification level for accuracy,

precision, and recall. Accuracy is the level of closeness between the predicted value and the true value. Precision shows the level of accuracy or precision in classification. Meanwhile, recall serves to measure the proportion of actual positives that are correctly identified.

3. RESULTS AND ANALYSIS

The dataset to be used in the training and testing process in this study has passed the preprocessing stage so that the data is ready to be processed. In this study, the classification process used the Weka v3.8.3 tool. Training data and testing data use data from undergraduate students of Informatics Engineering and Information Systems Study Programs from 2011 to 2014.

For performance estimation based on the training model that has been formed for the calculation of the dataset, it is k-fold cross validation with a fold value of 10. Classification by processing the C4.5 algorithm uses the J48 method. Classification by processing the K-Nearest Neighbor algorithm uses the IBk method. Classification by processing the Naïve Bayes algorithm uses the Naïve Bayes method. Classification by processing the Random Forest algorithm uses the RandomTree method.

Testing is done using confusion matrix, which is based on accuracy, precision, and recall value parameters. The test method used in the Weka tool is 10-fold cross validation. The dataset used for the training and testing process was 2,022 records, which came from the data of undergraduate students of STMIK Mikroskil Informatics Engineering and Information Systems Study Programs from 2011 to 2014. In addition, training and testing were also carried out on the dataset of each study program, where the Informatics Engineering Undergraduate Study Program consisted of 896 records, while the Information Systems Undergraduate Study Program consisted of 1,126 records. After implementing the four classification algorithms, namely C4.5, K-Nearest Neighbor, Naive Bayes, and Random Forest on the Weka tool using the 10-fold cross validation method, the results of the performance measurement of each classifier are obtained as shown in the following table.

Table 2. Comparison Results of WEKA Classification Performance (Informatics Engineering)

Algoritma (Pengklasifikasi)	*Accuracy* (%)	*Precision*	*Recall*
C4.5	66.4063	0.662	0.664
K-*Nearest Neighbor*	64.5089	0.645	0.645
Naïve Bayes	63.0580	0.631	0.631
Random Forest	65.4018	0.652	0.654

Based on the results of the performance comparison for the Informatics Engineering Undergraduate Study Program dataset, the C4.5 algorithm has the best performance or is in the top position indicated by the highest accuracy value, namely 66.406%, followed by the Random Forest algorithm of 65.402%, the K-Nearest Neighbor algorithm of 64.509%, and the Naïve Bayes algorithm of 63.058%.

Table 3. Comparison Results of WEKA Classification Performance (Information Systems)

Algoritma (Pengklasifikasi)	*Accuracy* (%)	*Precision*	*Recall*
C4.5	78.1528	0.745	0.782
K-*Nearest Neighbor*	79.3073	0.761	0.793
Naïve Bayes	78.8632	0.762	0.789
Random Forest	79.2185	0.761	0.792

Based on the results of the performance comparison for the Information Systems Undergraduate Study Program dataset, the K-Nearest Neighbor algorithm has the best performance or is in the top position as indicated by the highest accuracy value, namely 79.307%, followed by the Random Forest algorithm of 79.219%, the Naïve Bayes algorithm of 78,863 %, and the C4.5 algorithm of 78.153%.

Table 4. Comparison Results of WEKA Classification Performance (Informatics Engineering and Information Systems)

Algoritma (Pengklasifikasi)	*Accuracy* (%)	*Precision*	*Recall*
C4.5	72.4530	0.707	0.725
K-*Nearest Neighbor*	72.6508	0.711	0.727
Naïve Bayes	71.8595	0.701	0.719
Random Forest	72.6508	0.711	0.727

Based on the performance comparison results for the dataset of the Informatics Engineering and Information Systems Undergraduate Study Programs, the K-Nearest Neighbor and Random Forest algorithms

have the best performance or are in the top position indicated by the highest accuracy value, namely 72.651%, followed by the C4.5 algorithm of 72.453% and the Naïve Bayes algorithm of 71.860%.

From the results of the data classification, it was found that students with morning study time are not able to graduate on time when compared to those who study in the evening. This is suspected to be because students who study in the evening that are already working have a better understanding of the course material presented, because it is related to their daily work. However, the on-time graduation of students who study in the evening is also related to other attributes, namely grade of entrance examination, gender, and region of origin. In fact, a high grade of entrance examination does not guarantee that students can pass on time. This is suspected because it is possible for prospective students with a grade of entrance examination of A entered the university through special route that is not seen from academic factors. Likewise, what happened to gender, where male students were more able to graduate on time than female students. This is suspected because in the case study used in this research (Informatics Engineering and Information Systems Undergraduate Study Program) is an engineering study program, not a social study program, so it is more dominated by male gender whose interests are more in technical matters. While region of origin and GPA, namely students from within the city (Medan) can graduate on time than students from outside the city (outside Medan), and high GPA is more able to pass on time. A high GPA is also correlated with a high GPA per semester of students, so it indirectly affects the GPA score.

4. CONCLUSION

In this study, a model was made using a classification algorithm, namely C4.5, K-Nearest Neighbor, Naïve Bayes, and Random Forest using a dataset of graduates of the STMIK Mikroskil Informatics Engineering and Information Systems Undergraduate Study Programs from 2011 to 2014. For the dataset of graduates of the Informatics Engineering Undergraduate Study Program from 2011 to 2014, the C4.5 algorithm has higher accuracy than the other three algorithms. For the dataset of graduates of the Information Systems Undergraduate Study Program from 2011 to 2014, the K-Nearest Neighbor algorithm has higher accuracy than the other three algorithms. As for the dataset for graduates of the Informatics Engineering and Information Systems Undergraduate Study Programs from 2011 to 2014, the K-Nearest Neighbor and Random Forest algorithms have higher accuracy than the other two algorithms. Each algorithm has been able to predict the on-time graduation of STMIK Mikroskil students into two classes, namely on time and not on time. This research is limited to statistical analysis or based on existing data in the field, so that external factors are needed which can be used as additional attributes. In addition, it can be re-identified other attributes from the academic database that are likely to affect the predictions of student graduation on time so that better accuracy can be obtained, such as to use the grades of each courses of the students' transcript as features and some questionnaire related to their difficulties when studying in their fields.

REFERENCES

[1] Bhardwaj, B., & Pal, S. (2011, April). Data Mining: A Prediction for Performance Improvement Using Classification. International Journal of Computer Science and Information Security (IJCSIS), 9(4), 1-5.

[2] Manek, A. S., Shenoy, P., Mmohan, M., & R, V. K. (2017). Aspect Term Extraction for Sentiment Analysis in Large Movie Review Using Gini Index Feature Selection Method and SVM Classifier. World Wide Web, 20(2), 135-154. doi:10.1007/s11280-015-0381-x

[3] Nikam, S. (2015). A Comparative Study of Classification Techniques in Data Mining Algorithms. Oriental Journal of Computer Science & Technology, 8(1), 13-19.

[4] Qabajeh, I., Thabtah, F., & Chiclana, F. (2015). A Dynamic Rule-Induction Method for Classification in Data Mining. Journal of Management Analytics, 2(3), 233-253. doi:10.1080/23270012.2015.1090889

[5] Seidlova, R., Pozivil, J., & Seidl, J. (2019). Marketing and Business Iintelligence with Help of Ant Colony Algorithm. Journal of Strategic Marketing, 27(5), 451-463. doi:10.1080/0965254X.2018.1430058

[6] Ahmad, F., Ismail, N., & Aziz, A. (2015). The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. Apllied Mathematical Sciences, 9(129), 6415-6426. Retrieved from 10.12988/ams.2015.53289

[7] Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and Predicting Student's Academic Performance Using Data Mining Techniques. International Journal of Modern Education and Computer Science, 8(11), 36-42. doi:10.5815/ijmecs.2016.11.05

[8] Cheewaprakobkit, P. (2013). Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. Proceedings of the International MultiConference of Engineers and Computer Scientests (IMECS), (pp. 13-15). Hongkong.

[9] Saa, A. A. (2016). Educational Data Mining & Students' Performance Prediction. International Journal of Advanced Computer Science and Applications, 7(5), 212-220.

[10] Polaka, I., & Borisov, A. (2010). Clustering-Based Decision Tree Classifier Construction. Technological and Economic Development of Economy, 16(4), 765-781. doi:10.3846/tede.2010.47

[11] Dindarloo, S. R., & Siami-Irdemoosa, E. (2016). Data Mining in Mining Engineering: Results of Classification and Clustering of Shovels Failures Data. Journal of Mining, Reclamation and Environment, 31(2), 105-118. doi:10.1080/17480930.2015.1123599

[12] Abad, F. M., & Lopez, A. A. (2016). Data-Mining Techniques in Detecting Factors Linked to Academic Achievement. School Effectiveness and School Improvement, 28(1), 39-55. doi:10.1080/09243453.2016.1235591

[13] Hussain, S., Dahan, N., Ba-Alwib, F., & 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. doi:10.11591/ijeecs.v9.i2

[14] Mishra, T., Kumar, D., & Gupta, S. (2014). Mining Students' Data for Performance Prediction. 2014 Fourth International Conference on Advanced Computing & Communication Technologies, (pp. 255-263). doi:10.1109/ACCT.2014.105

[15] Hashemi, F. S., Ismail, M. R., Yusop, M. R., Hashemi, M. S., Shahraki, M. H., Rastegari, H., . . . Aslani, F. (2017). Intelligent Mining of Large-Scale Bio-Data: Bioinformatics Applications. Bioinformatics Applications, Biotechnology & Biotechnological Equipment, 32(1), 10-29. doi:10.1080/13102818.2017.1364977

[16] Gallagher, C., Bruton, K., & O'Sullivan, D. T. (2016). Utilising the Cross Industry Standard Process for Data Mining to Reduce Uncertainty in the Measurement and Verification of Energy Savings. Springer (pp. 48-58). Cham: Springer International Publishing Switzerland. doi:10.1007/978-3-319-40973-3_5

[17] Pivk, A., Vasilecas, O., Kalibatiene, D., & Rupnik, R. (2013). On Approach for the Implementation of Data Mining to Business Process Optimisation in Commercial Companies. Technological and Economic Development of Economy, 19(2), 237-256. doi:10.3846/20294913.2013.796501

[18] Cho, M.-H., & Yoo, J. (2017). Exploring Online Students' Self-regulated Learning With Self-reported Surveys and Log Files: A Data Mining Approach. Interactive Learning Environments, 25(8), 970-982. doi:10.1080/10494820.2016.1232278

[19] Osmanbegovic, E., & Suljic, M. (2012). Data Mining Approach for Predicting Student Performance. Economic Review: Journal of Economics and Business, 10(1), 3-12. Retrieved from http://hdl.handle.net/10419/193806

[20] Rao, K., Swapna, N., & Kumar, P. (2018). Educational Data Mining for Student Placement Prediction Using Machine Learning Algorithm. International Journal of Engineering & Technology, 7(1.2), 43-46.

[21] Shingari, I., Kumar, D., & Khetan, M. (2017). A Review of Application of Data Mining Ttechniques for Prediction of Students' Performance in Higher Education. Journal of Statistics and Management Systems, 20(4), 713-722. doi:10.1080/09720510.2017.1395191

[22] Sathyadevan, S., & Nair, R. R. (2015). Comparative Analysis of Decision Tree Algorithm: ID3, C4.5 and Random Forest. Computational Intelligence in Data Mining, 1, 549-562. doi:10.1007/978-81-322-2205-7_51

[23] Li, Y., Jiang, Z. L., Yao, L., Wang, X., Yiu, S., & Huang, Z. (2017). Outsourced Privacy-Preserving C4.5 Decision Tree Algorithm Over Horizontally and Vertically Partitioned Dataset Among Multiple Parties. Cluster Computing, 1-13. doi:10.1007/s10586-017-1019-9

[24] Nagra, A., Han, F., Ling , Q., Abubaker, M., Ahmad, F., Mehta, S., & Apasiba, A. (2019). Hybrid Self-Inertia Weight Adaptive Particle Swarm Optimisation with Local Search Using C4.5 Decision Tree Classifier for Feature Selection Problems. Connection Science, 1-21. doi:10.1080/09540091.2019.1609419

[25] Cherfi, A., Nouira, K., & Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm. Applied Artificial Intelligence, 32(2), 119-137. doi:10.1080/08839514.2018.1447479

[26] Anuradha, C., & Velmurugan, T. (2015, July). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. Indian Journal of Science and Technology, 8(15), 1-12. doi:10.17485/ijst/2015/v8i15/74555

[27] Tahi, M., Miloudi, A., Dron, J., & Bouzouane, B. (2018). Decision Tree and Feature Selection by Using Genetic Wrapper for Fault Diagnosis of rotating machinery. Australian Journal of Mechanical Engineering, 1-9. doi:10.1080/14484846.2018.1552355

[28] Koutina, M., & Kermanidis, K. (2011). Predicting Postgraduate Students' Performance Using Machine Learning Techniques. IFIP International Federation for Information Processing, 159-168.

[29] Ratnaningsih, D. J., & Sitanggang, I. S. (2015). Comparative Analysis of Classification Methods in Determining Non-Active Student Characteristics in Indonesia Open University. Journal of Applied Statistics, 43(1), 87-97. doi:10.1080/02664763.2015.1077940

[30] Alam, F., Mehmood, R., Katib, I., & Albeshri, A. (2016). Analysis of Eight Data Mining Algorithms for Smarter Internet of Things (IoT). Procedia Computer Science, 98, 437-442. doi:10.1016/j.procs.2016.09.068

[31] Adeniyi, D., Wei, Z., & Yongquan, Y. (2016). Automated Web Usage Data Mining and Recommendation System Using K-Nearest Neighbor (KNN) Classification Method. Applied Computing and Informatics, 12(1), 90-108. doi:10.1016/j.aci.2014.10.001

[32] Kanj, S., Abdallah, F., Denoeux, T., & Tout, K. (2016). Editing Training Data for Multi-label Classification with the K-nearest Neighbor Rule. Pattern Analysis and Applications, 19(1), 145-161. doi:10.1007/s10044-015-0452-8

[33] Shahiri, A., Husain, W., & Rashid, N. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414-422. doi:10.1016/j.procs.2015.12.157

[34] Tarapitakwong, J., Chartrungruang, B., Tantranont, N., & Smhom, S. (2017). A Classification Model for Predicting Standard Levels of OTOP's Wood Handicraft Products by Using the K-Nearest Neighbor. Tthe International Journal of the Computer, the Internet and Management, 25(2), 135-141.

[35] Erkayaoglu, M., & Dessureault, S. (2018). Improving Mine-to-Mill by Data Warehousing and Data Mining. International Journal of Mining, Reclamation and Environment, 1-16. doi:10.1080/17480930.2018.1496885

[36] Ashraf, N., Ahmad, W., & Ashraf, R. (2018). A Comparative Study of Data Mining Algorithms for High Detection Rate in Intrusion Detection System. Annals of Emerging Technologies in Computing (AETiC), 2(1), 49-57.

[37] Oktanisa, I., & Supianto, A. (2018). Perbandingan Teknik Klasifikasi Dalam Data Mining untuk Bank Direct Marketing. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK), 5(5), 567-576.

[38] Nidhomuddin, & Otok, B. (2015). Random Forest dan Multivariate Adaptive Regression Spline (MARS) Binary Response untuk Klasifikasi Penderiat HIV/ AIDS. Statistika, 1(3), 567-576.

[39] Nugroho, Y., & Emiliyawati, N. (2017). Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen terhadap Mobil Menggunakan Metode Random Forest. Jurnal Teknik Elektro, 9(1), 24-29.

[40] Gunawan, Hanes, & Catherine. (2020). Information Systems Students' Study Performance Prediction Using Data Mining Approach. IEEE Xplore. doi: 10.1109/ICIC47613.2019.8985718

[41] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18.

BIBLIOGRAPHY OF AUTHORS

Gunawan, is currently as a lecturer from Informatics Engineering Undergraduate Study Program at STMIK Mikroskil.

Hanes, is currently as a lecturer from Information Systems Undergraduate Study Program at STMIK Mikroskil.

Catherine, is currently as a lecturer from Information Systems Undergraduate Study Program at STMIK Mikroskil.