Comparison of Decision Tree Classification Methods and Gradient Boosted Trees

Arif Rinaldi Dikananda [1], Sri Jumini [2], Nafan Tarihoran [3], Santy Christinawati [4], Wahyu Trimastuti [4], Robbi Rahim [5]

[1]*STMIK IKMI Cirebon, Indonesia*
[2]*Universitas Sains Al-Qur'an Indonesia, Wonosobo, Indonesia*
[3]*Universitas Islam Negeri Sutan Maulana Hasanuddin Banten, Indonesia*
[4]*Politeknik Piksi Ganesha Bandung, Bandung, Indonesia*
[5]*Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia*

Abstract **– The purpose of this research is to analyze the C4.5 and Random Forest algorithms for classification. The two methods were compared to see which one in the classification process was more accurate. The case is the success of university students at one of the private universities. Data is obtained from the https://osf.io/jk2ac data set. The attributes used were gender, student, average evaluation (NEM), reading session, school origin, and presence as input and success as a result (label). The process of analysis uses Rapid Miner software with the same test parameters (k-folds = 2, 3, 4, 5) with the same type of sample (stratified sample, linear sample, shuffled sampling). The first result shows that the sample type test k-fold (stratified sampling) achieved an average accuracy of 55.76 percent (C4,5) and 5618 percent (Random Forest). The second result showed that the k-fold (linear sampling) sample test achieved an average precision of 58.06 percent (C4.5) and 6506 percent. (Random Forest). The third result shows that the k-fold test with the sampling type has averaged 58.68 per cent (C4,5) and 60,76 per cent (shuffled sampling) precision (Random Forest). From the three test results, in the case of student success at a private university, the Random Forest method is better than C4.5.**

Keywords **– Comparison, Data mining, Classification, C4.5, Random Forest, Accuracy.**

DOI: 10.18421/TEM111-39
https://doi.org/10.18421/TEM111-39

Corresponding author: Robbi Rahim,
Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia.
Email: usurobbi85@zoho.com

Received: 04 November 2021.
Revised: 03 February 2022.
Accepted: 08 February 2022.
Published: 28 February 2022.

1. Introduction

Data mining is one of the methods used for extracting knowledge or finding patterns from large data. Data mining is the process of extracting important information from data implicit and previously unknown [1], [2], [3], [4], [5]. Some of the data extraction roles can be played by estimating, predicting, classifying, clustering, and assembly [6, 7], [8], [9], [10]. There are several well-known data mining algorithms, including C4.5, Random Forest and others [11],[12]. The choice to use the C4.5 and Random Forest algorithms is based on several reasons, which can both be easily implemented and which in the case of classification both produce good results [13]. Several previous studies have analyzed these two algorithms, for example [14] on the performance comparison of decision tree algorithms and random forest: application on health expenditure songul for OECD countries. Two comparisons of the C4.5 and Random Forest methods are presented in this paper using 50 trees. The results showed that the Random Forest exceeded C4.5, namely AUC = 0.98 and AUC = 0.90, with classification accuracy. In addition, [15] research on the comparison between the Decision Tree and Random Forest in Type 2 diabetes cases. This paper compares the accuracy, sensitivity, speciality and area under the ROC curve between these two models. The results have shown that in terms of accuracy, sensitivity, specificity and area within the ROC curve, the Random Forest method outperforms C4.5. Further research was conducted [16] on rainfall prediction weather data analysis using the Rattle-R GUI tool. This paper

proposes two algorithms for classification, C4.5 and Random Forest. The results show that a smaller redistribution error rate than C4.5 is shown in the Random Forest algorithm. Lan, 2020 [17] conducted a study of the comparison of the decision tree method with the Random Forest method to identify the F-spread. The paper proposes three automatic identification processes for F-dispersion by Decision Tree, Random Forest, and CNN. The results show that the Random Forest method results in the exact identification of monograms with the CNN F distribution better than that of the Decision Tree method.

The C4.5 and Random Forests algorithms were used mostly in previous studies. However, each study cannot determine which model in classification is more accurate and faster. This is because previous researchers' data sets are not the same, because each data treatment is also different. The more complex the data, the data noise and the inconsistent data naturally affect the performance of the classification algorithm [13], [18], [19], [20]. This study aims to answer the question, which model is more exact in classifying student success in tertiary institutions, based on identifying these problems.

2. Methodology

2.1. Dataset Input

In a comparative study of decision tree and random forest classification methods, we used the AMIK Tunas Bangsa Student dataset of the study (Luvia et al., 2017). There are five attributes used as input and one attribute used as output. The following attribute data is shown in the following table:

Table 1. Predicate of Graduation

Field Name	Data Class Type	Data Class Used
Predicate of Success	Nominal	Cum laude, Very Good, Good, Enough, Less
Gender	Nominal	Man / Woman
Student Evaluation Average Score (NEM)	Nominal	NEM ≤ 20, NEM> 20
Lecture Session	Nominal	Morning afternoon Evening
School Origin	Nominal	Pematangsiantar, Outside the Region
Presence	Nominal	Attendance <50, Attendance> 50

The dataset used to calculate the algorithm when comparing the Decision Tree Classification Methods and the Random Forest in the case of student success stories can be accessed via the https://osf.io/jk2ac.

2.2. Import Dataset

The first thing to do is to import the dataset into the RapidMiner software, which is to provide the dataset that has been saved in the.xls format. Then import it to the Read Excel tool using the Operator menu. One can use the view command to display datasets that have been imported into RapidMiner.

2.3. K-Fold Cross Validation

K-Fold Cross Validation is a method for assessing the performance of an algorithm [21]. K is to fold the data as much as K and iterate as K so that the algorithm has a data accuracy value. The decision tree algorithm being evaluated at this point is the C4.5 algorithm [7] and the Random Forest [22]. What is being done in each model is to make a fold and use the best number of folds to assess the validity using 5-fold cross-validation in the model [23, 24]. The C4.5 algorithm and the Random Forest recursively visit each decision node, selecting the optimal branch until no more branches are generated.

2.4. The Random Forest

The random forest algorithm estimates the error rate more accurately in relation to decision trees [25, 26]. More specifically, the error rate has always converged with the increase in the number of trees [22]. The steps in the random forest classification are as follows [25]:

a) A set of decision trees has been created from the training set. In the present work, 100 trees have been grown.
b) Each tree in the dataset has been grown by randomly selecting attributes.
c) The "m" features are randomly selected from the "M" features in the dataset, $m = \sqrt{M}$ is this work, where M is the total number of features in the dataset.
d) Attribute selection was done using Gini index score between 0 and 1, where 0 indicates the most interesting information and 1 indicates the least interesting information.
e) Trees shall be grown to the maximum depth (all selected attributes).
f) When a test instance (obtained from a 10-fold cross-validation) was given to a constructed random forest, all the trees in the forest will have their resultant class. The final class is decided on the basis of the majority vote.
g) The accuracy of the classifier is calculated where the accuracy is the percentage of the test set correctly classified by the classifier.

2.5. C4.5 Decision Tree Algorithm

C4.5 Decision tree-based upon ID3 Decision Tree is expanded, and does not directly use information gain, but adds gain as one of the best partition

standard attributes is selected [27]. In this case, C4.5 uses the expected class value to create a higher tree decision node [1].

3. Results and Discussion

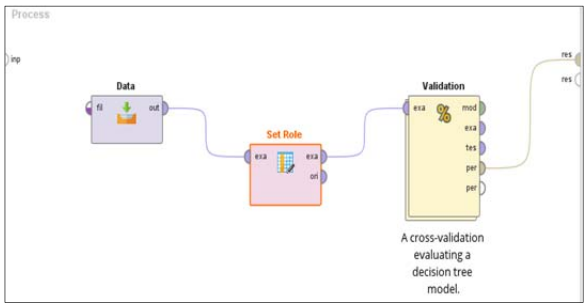
The comparative analysis design process of the two classification methods (C4.5 and Random Forest) uses the help of the Rapid Miner cross validation software (k-fold = 2, 3, 4, 5) where the results of the two methods are compared to the accuracy values of the classification results. The following is a model created with the Rapid Miner software.

3.1. Creating a C4.5 Model Tree with Cross Validation

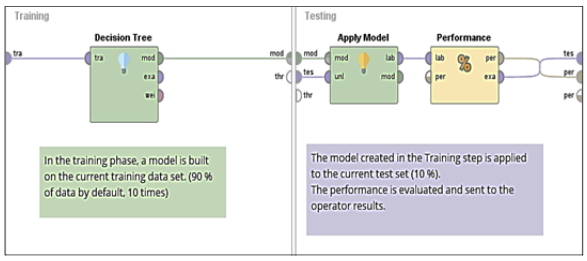
In making the C4.5 model using the RapidMiner software, the data entry process uses an import dataset (.xls) where after importing the dataset; the role set process is carried out to determine the output of the classification model created. In this case the output attribute is Predicate of Success (Table 1). The testing process uses cross validation with k-fold = 2, 3, 4, 5. For training and testing the same parameters were used, namely:

a) maximal dept : 10
b) apply pruning : confidence= 0.1
c) minimal gain : 0.01
d) minimal leaf size : 1
e) minimal size for split : 1
f) number of prepuning : 1

The following is the design of the C4.5 model using cross validation as shown in Figure 1.



(a)



(b)

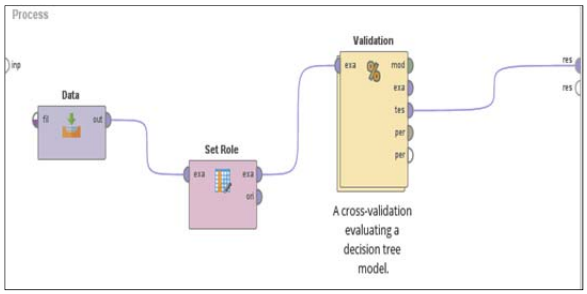
Figure 1. The C4.5 Model Tree with Cross Validation (a)(b)

3.2. Creating a Random Forest Model with Cross Validation

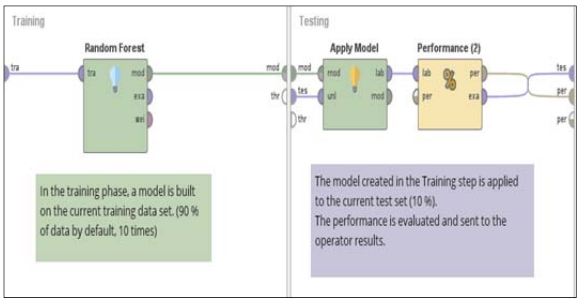
The data entry process uses an import dataset (.xls) to create the Random Forest model using the RapidMiner Software, where the role set process is performed after importing the dataset to determine the output of the classification model. In this case the output attribute is a success indicator (Table 1). The test process uses k-fold = 2, 3, 4, 5 for cross-validation. For training and testing with the same parameters:

a) maximal dept : 10
b) apply pruning : confidence= 0.1
c) minimal gain : 0.01
d) minimal leaf size : 1
e) minimal size for split : 1
f) number of prepuning : 1

The following is the design of the Random Forest model using cross validation as shown in Figure 2.



(a)



(b)

Figure 2. The Random Forest Model Tree with Cross Validation (a)(b)

3.3. Results of Model Tree C4.5 with Cross Validation

Following are the results of the analysis of the C4.5 method based on different sampling types with predetermined cross validation as shown in the following table:

Table 2. sampling type use stratified sampling

Method	Cross Validation	accuracy
Decision Tree	K-Folds= 2	26.67 percent
	K-Folds= 3	72.22 percent
	K-Folds= 4	70.83 percent
	K-Folds= 5	53.33 percent
Average		55.76 percent

In Table 2 the highest accuracy value is in cross validation with (k-folds = 3) which is 72.22 percent. The mean score for accuracy was 55.76 percent.

Table 3. sampling type use linear sampling

Method	Cross Validation	accuracy
Decision Tree	K-Folds= 2	36.67 percent
	K-Folds= 3	72.22 percent
	K-Folds= 4	50.00 percent
	K-Folds= 5	73.33 percent
Average		58.06 percent

In table 3 the highest accuracy value is in cross validation with (k-folds = 5) which is 73.33 percent. The mean score for accuracy was 58.06 percent.

Table 4. sampling type use shuffled sampling

Method	Cross Validation	accuracy
Decision Tree	K-Folds= 2	36.67 percent
	K-Folds= 3	72.22 percent
	K-Folds= 4	62.50 percent
	K-Folds= 5	63.33 percent
Average		58.68 percent

In Table 4 the highest accuracy value is in cross validation with (k-folds = 3) which is 72.22 percent. The mean score for accuracy was 58.68 percent.

3.4. Results of the Random Forest Model with Cross Validation

The results of the Random Forest method analysis based on different sample types with pre-determined cross-validation as shown in the table below are as follows:

Table 5. sampling type use stratified sampling

Method	Cross Validation	accuracy
Random Forest	K-Folds= 2	36.67 percent
	K-Folds= 3	72.22 percent
	K-Folds= 4	62.50 percent
	K-Folds= 5	53.33 percent
Average		56.18 percent

The highest accuracy in the Random Forest model test in Table 5 is in cross-validation with (k-folds = 3) 72.22 percent. The mean accuracy score was 56.68 percent.

Table 6. sampling type use linear sampling

Method	Cross Validation	accuracy
Random Forest	K-Folds= 2	57.67 percent
	K-Folds= 3	63.89 percent
	K-Folds= 4	75.33 percent
	K-Folds= 5	63.33 percent
Average		65.06 percent

The highest accuracy in the Random Forest model test in Table 6 is in cross-validation with (k-folds = 4) 75.33 percent. The mean accuracy score was 65.06 percent.

Table 7. sampling type use shuffled sampling

Method	Cross Validation	accuracy
Random Forest	K-Folds= 2	53.33 percent
	K-Folds= 3	72.22 percent
	K-Folds= 4	54.17 percent
	K-Folds= 5	63.33 percent
Average		60.76 percent

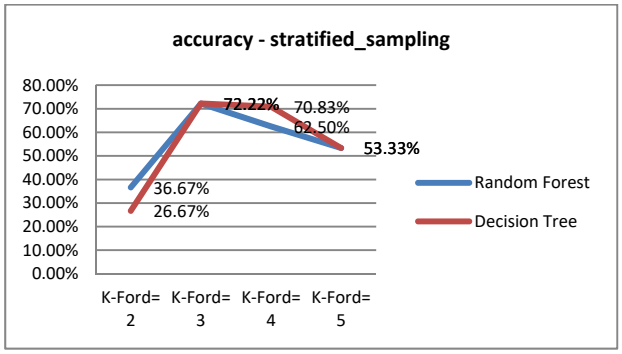
The highest accuracy in the Random Forest model test in Table 7 is in cross-validation with (k-folds = 3) 72.22 percent. The mean accuracy score was 60.76 percent.

3.5. Discussion

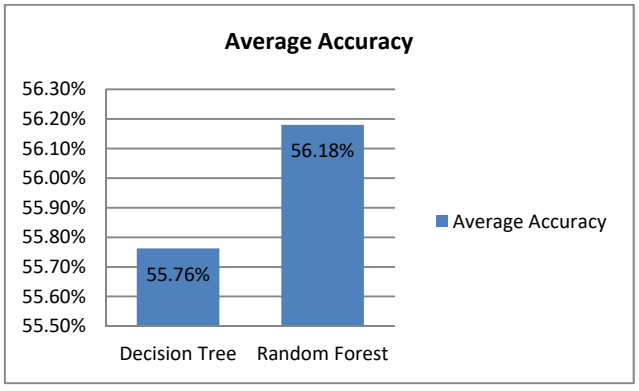
A summary and graph of the accuracy values for the two methods at different sampling types can be seen in the following table and figure:

Table 8. Accuracy values for both methods on the sampling type use stratified sampling

Method	K-Folds= 2	K-Folds= 3	K-Folds= 4	K-Folds= 5
Random Forest	36.67 percent	72.22 percent	62.50 percent	53.33 percent
Decision Tree	26.67 percent	72.22 percent	70.83 percent	53.33 percent



(a)



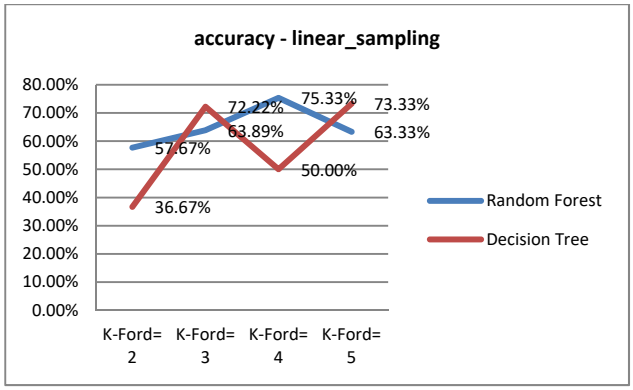
(b)

Figure 3. Graph The accuracy value of both methods on the sampling type use stratified sampling (a)(b)

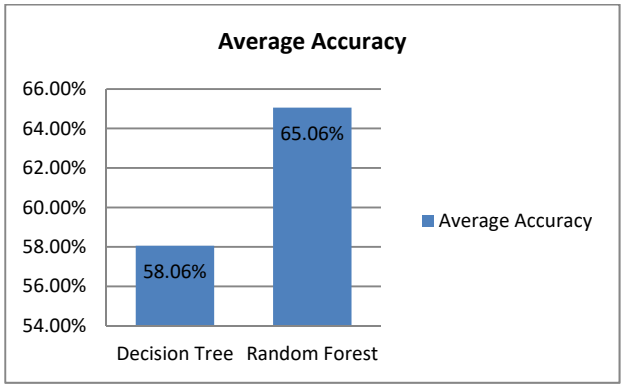
It can be explained in Table 8 and Figure 3 that the Random Forest method is better than the C4.5 method when using a stratified_sampling type sample with 56.18 percent exactness.

Table 9. Accuracy values for both methods on the sampling type use linear sampling

Method	K-Folds= 2	K-Folds= 3	K-Folds= 4	K-Folds= 5
Random Forest	57.67 percent	63.89 percent	75.33 percent	63.33 percent
Decision Tree	36.67 percent	72.22 percent	50.00 percent	73.33 percent



(a)



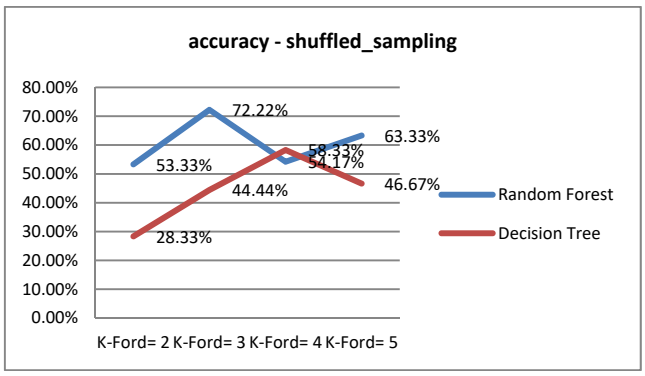
(b)

Figure 4. Graph The accuracy value of both methods on the sampling type use linear sampling (a)(b)

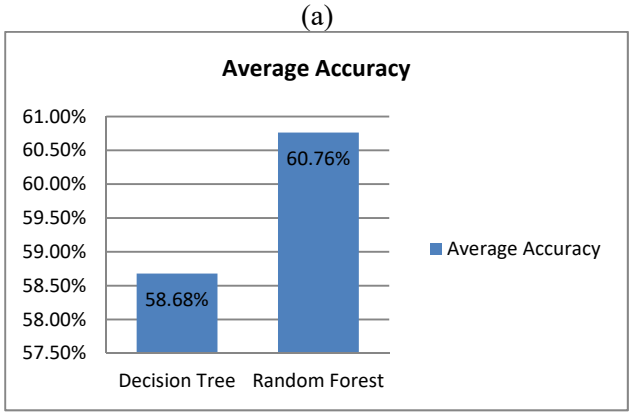
It can be explained in Table 9 and Figure 4 that the Random Forest method is better than the C4.5 method when using a linear sampling type sample with 65.06 percent exactness.

Table 10. Accuracy values for both methods on the sampling type use shuffled sampling

Method	K-Folds= 2	K-Folds= 3	K-Folds= 4	K-Folds= 5
Random Forest	53.33 percent	72.22 percent	54.17 percent	63.33 percent
Decision Tree	28.33 percent	44.44 percent	58.33 percent	46.67 percent



(a)



(b)

Figure 5. Graph The accuracy value of both methods on the sampling type use shuffled sampling (a)(b)

It can be explained in Table 10 and Figure 5 that the Random Forest method is better than the C4.5 method when using a linear sampling type sample with 60.76 percent exactness.

4. Conclusion

On the basis of the results of the research, it is explained that the Random Forest method in terms of classification accuracy using cross validation is better than the C4.5. The first result shows that the test type k-fold (stratification) of the sample achieved a mean precision of 55.76% (C4.5) and 5618% (Random Forest). The second result demonstrated that the k-fold (linear sample) test achieved an average accuracy of 58.06% (C4.5) and 6506%. (Forest Random). The third finding shows that a k-fold test has an average of 58.68% (C4.5) and 60.76% (shuffled) precision (Random Forest). Of the three test results, the Random Forest method is better than C4.5 in the case of student success at a private university.

References

[1]. Elacio, A. A., Lacatan, L. L., Vinluan, A. A., & Balazon, F. G. (2020). Machine Learning Integration of Herzberg's Theory using C4. 5 Algorithm. *Int. J. Adv. Trends Comput. Sci. Eng*, *9*(1.1), 57-63. doi:10.30534/ijatcse/2020/1191.12020

[2]. Agarwal, S. (2013, December). *Data mining: Data mining concepts and techniques*. In 2013 international conference on machine intelligence and research advancement (pp. 203-207). IEEE.

[3]. Muhajir, M., & Efanna, B. R. (2015). Association Rule Algorithm Sequential Pattern Discovery using Equivalent Classes (SPADE) to Analyze the Genesis Pattern of Landslides in Indonesia. *International Journal of Advances in Intelligent Informatics*, *1*(3), 158-164. doi:10.26555/ijain.v1i3.50

[4]. Triyanto, Y., Sepriani, Y., Mustamu, N. E., Siregar, R. A., & Rambe, B. H. (2021, June). Implementation of PROMETHEE Method for Potential Suitability of Land Oil Palm Plant. In *Journal of Physics: Conference Series* (Vol. 1933, No. 1, p. 012060). IOP Publishing. doi:10.1088/1742-6596/1933/1/012060

[5]. Setiawan, M. I., Hasyim, C., Kurniasih, N., Abdullah, D., Napitupulu, D., Rahim, R., ... & Wajdi, M. B. N. (2018, April). E-Business, the impact of regional growth on the improvement of Information and Communication Development. In *Journal of Physics: Conference Series* (Vol. 1007, No. 1, p. 012044). IOP Publishing. doi:10.1088/1742-6596/1007/1/012044

[6]. Proud, R., Mangeni-Sande, R., Kayanda, R. J., Cox, M. J., Nyamweya, C., Ongore, C., ... & Brierley, A. S. (2020). Automated classification of schools of the silver cyprinid Rastrineobola argentea in Lake Victoria acoustic survey data using random forests. *ICES Journal of Marine Science*, *77*(4), 1379-1390. doi:10.1093/icesjms/fsaa052.

[7]. Praveena, M.; Bhavana, N. (2019). Prediction of chronic kidney disease using C4.5 algorithm. *International Journal of Recent Technology and Engineering*, *7*(6), 721-723.

[8]. Susanto, R., & Rachmadtullah, R. (2019). Model of pedagogic competence development: Emotional intelligence and instructional communication patterns. *International Journal of Scientific and Technology Research*, *8*(10), 2358-2361.

[9]. Raharja, N. M., Prasojo, I., & Tanane, O. (2021). Empowerment of msmes during the covid-19 pandemic with information technology. *Jurnal Pengabdian dan Pemberdayaan Masyarakat Indonesia*, *1*(1), 1-8.

[10]. Mujanah, S., Ardiana, I. D. K. R., Nugroho, R., Candraningrat, C., Fianto, A., & Arif, D. (2022). Critical thinking and creativity of MSMEs in improving business performance during the covid-19 pandemic. *Uncertain Supply Chain Management*, *10*(1), 19-28. doi:10.5267/J.USCM.2021.10.014

[11]. Hermanto, H., Kuryanti, S. J., & Khasanah, S. N. (2019). Comparison of Naïve Bayes Algorithm, C4. 5 and Random Forest for Classification in Determining Sentiment for Ojek Online Service. *Sinkron: jurnal dan penelitian teknik informatika*, *3*(2), 266-274.

[12]. Nugraheni, I. A. (2021). Implementation of environmental care character for elementary school students through verticultural culture techniques. *Jurnal Pengabdian dan Pemberdayaan Masyarakat Indonesia*, *1*(2), 59-66.

[13]. Suryanto, A.; Alfarobi, I.; Tutupoly, T. A. (2018). Komparasi Algoritma C4.5, Naive Bayes Dan Random Forest Untuk Klasifikasi Data Kelulusan Mahasiswa Jakarta, Mitra Dan Teknologi Pendidikan, Vol. iv nomor 1, 2–14

[14]. Çinaroğlu, S. (2016). Comparison of Performance of Decision Tree Algorithms and Random Forest An Application on OECD Countries Health Expenditures. *International Journal of computer applications*, *138*(1). doi:10.5120/ijca2016908704.

[15]. Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., & Amirabadizadeh, A. (2018). A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *Journal of research in health sciences*, *18*(2), 412. doi:10.34172/jrhs183777

[16]. Prajwala, T. R. (2015). A comparative study on decision tree and random forest using R tool. *International journal of advanced research in computer and communication engineering*, *4*(1), 196-199. doi:10.17148/ijarcce.2015.4142.

[17]. Lan, T., Hu, H., Jiang, C., Yang, G., & Zhao, Z. (2020). A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification. *Advances in Space Research*, *65*(8), 2052-2061. doi:10.1016/j.asr.2020.01.036

[18]. Tang, J., Alelyani, S., & Liu, H. (2014). Data classification: algorithms and applications. *Data Mining and Knowledge Discovery Series*, 37-64. doi:10.1201/b17320

[19]. Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792. doi:10.1890/07-0539.1

[20]. Ali, S., & Smith, K. A. (2006). On learning algorithm selection for classification. *Applied Soft Computing*, *6*(2), 119-138. doi:10.1016/j.asoc.2004.12.002

[21]. Rumahorbo, A. C., & Sekarwati, K. A. (2020). Penerapan Data Mining Dengan Menggunakan Algoritma C4. 5 Pada Klasifikasi Fasilitas Kesehatan Provinsi Di Indonesia. *Jurnal Ilmiah Komputasi*, *19*(1), 27-38.

[22]. Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3-29. doi:10.1177/1536867X20909688

[23]. Syukmana, F., Wahyudi, E., Gata, W., Wahono, H., Febianto, N. I., Kuntoro, A. Y., ... & Sulaeman, O. R. (2020, February). Predicting Relegation Clubs in Italian Serie A with Method based C4. 5 Decision Tree Algorithm. In *Journal of Physics: Conference Series* (Vol. 1471, No. 1, p. 012016). IOP Publishing. doi:10.1088/1742-6596/1471/1/012016.

[24]. Agustina, N., & Rozali, Y. A. (2020). Analysis of the application of the pedagogical competency model case study of public and private primary schools in West Jakarta Municipality, Dki Jakarta Province. *Ilkogretim Online*, *19*(3), 167-182. doi:10.17051/ILKONLINE.2020.03.114

[25]. Christo, V. E., Nehemiah, H. K., Brighty, J., & Kannan, A. (2020). Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest. *IETE Journal of Research*, 1-14. doi:10.1080/03772063.2020.1713917/

[26]. Bhosle, N., & Kokare, M. (2020). Random forest-based active learning for content-based image retrieval. *International Journal of Intelligent Information and Database Systems*, *13*(1), 72-88. doi:10.1504/ijiids.2020.10030218

[27]. Cuan, Y., Wang, Z., & Han, J. (2020). Research on TV imaging casing damage detection and classification method based on C4. 5 decision tree. In *Journal of Physics: Conference Series* (Vol. 1437, No. 1, p. 012132). IOP Publishing. doi:10.1088/1742-6596/1437/1/012132