An Ensemble Voting Approach for Dropout Student Classification Using Decision Tree C4.5, K-Nearest Neighbor and Backpropagation

[1]Daffa Nur Cholis, [2]Nurissaidah Ulinnuha
[1,2]Departement of Mathematics, Faculty of Science and Technology,
UIN Sunan Ampel Surabaya, Indonesia
Email: [1]daffanurcholis282@gmail.com, [2]nuris.ulinnuha@uinsby.ac.id

Article Info

Article history:
Received Mar 9th, 2023
Revised Apr 20th, 2023
Accepted May 26th, 2023

Keyword:
Classification
Drop Out
Ensemble Learning
Ensemble Voting
Student

ABSTRACT (10 PT)

Many factors cause drop out in students. This study classified active students and drop out students using 1092 student data consisting of 557 active student data and 535 drop out student data. The independent variables used are Semester, Semester Credit Units (SKS), Semester Grade Point Average (IPS), Grade Point Average (IPK), admission pathways and Single Tuition Fee (UKT). Classification is carried out using the Ensemble Voting method where the method will combine the Decision Tree C4.5, KNN and Backpropagation methods as a single method. In addition to knowing the classification of active students and drop out students, this study aims to prove whether the Ensemble Voting method is able to get better results than the single method. This classification using a comparison of training and testing data of 90:10 to build model. Classification results from a single method will be included in the Ensemble Voting method. The Decision Tree C4.5 method gets 95.45% accuracy, 98.03% precision and 92.59% recall. KNN gets 96.36% accuracy, 100% precision and 92.59% recall. Backpropagation gets 90.90% accuracy, 95.83% precision and 95.18% recall. Meanwhile, the Ensemble Voting rule used is Ensemble Soft Voting with a weight of (2,1,1). Ensemble Voting with Ensemble Soft Voting rules is able to improve the accuracy, precision and recall values with 98.18% accuracy, 100% precision and 96.29% recall.

Corresponding Author:
Daffa Nur Cholis,
Departement of Mathematics, Faculty of Science and Technology, UIN Sunan Ampel Surabaya
Ahmad Yani 117 Surabaya 60237, Indonesia
Email: daffanurcholis282@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v6i1.23412

1. INTRODUCTION

Drop out student data is one of indicators that lowes the higher education accreditation [1]–[5]. Students who experience dropping out will cause losses both for themselves and for the college. In 2019, students dropped out in Indonesia by 7% where there were 602,208 students in Indonesia who dropped out. This percentage is smaller than the percentage in 2018 with 8%. Many factors influence drop out students so that it attracts many researchers to carry out factor analysis and classify drop out students using various methods. One of them was carried out at Pembangunan Panca Budi University with an accuracy value of 59.58%. Utilizing the multilayer perceptron and radial basis function, Alban and Mauricio attained accuracy values of 96.3% and 96.8% respectively [6] .

Many factors cause student drop out and each university has different regulations regarding the provisions for student drop out [7]–[11]. Of the many studies on the classification of drop out students, there are several studies using the Ensemble Learning method. Ensemble Learning collects classification results from several models that will be put together in one place in the hope of getting better results. Ensemble learning of multiple machine learning is expected to have better general performance than a single machine learning, especially under varying conditions or over a long period of time. There are many methods used in Ensemble

Learning including Bagging, Voting, Adaboost and Stacking. By using Ensemble Learning, it is hoped that the shortcomings of a single machine learning method can be overcome. The following is a description of the compensation for the bias (error) and variance of a single method during data training and testing.

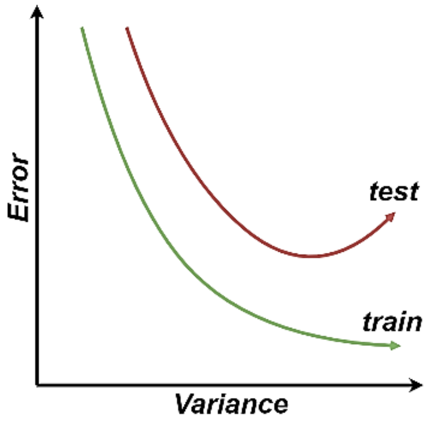

Figure 1. The relationship between model complexity and error values [12]

The figure above shows that during the training process the complexity value of a single model will be higher but the prediction error value will be smaller. During the testing phase, it was expected that the error value would decrease. However, contrary to expectations, the error value actually increased. [12]. This is the reason why various classification methods yield different outcomes, even when applied to the same dataset. These methods employ different learning techniques and possess their own strengths and weaknesses.

Ensemble Learning on the Decision Tree and Forest Tree methods has been used to predict dropout students. The results show that Ensemble Learning has a marginal increase in classification performance with the best accuracy value being 69.22% with a passing precision of 56.8%, a dropping out precision of 77.44%, a passing recall of 62.49% and a dropping out recall of 73.04% [13]. In addition, Ensemble Learning on the Naive Bayes method, SVM and Logistic Regression has also been used for grouping dropout students showing that Ensemble can improve predictive performance significantly with 94.24% accuracy, 95.62% precision, 93.12% recall, 94.1% f-measure and 0.91% AUC [14]. Individual models may have high bias (underfitting) or high variance (overfitting), but by aggregating their predictions, ensemble methods can achieve a better balance between the two. By averaging or voting over multiple models, ensemble methods can smooth out individual model's errors and improve generalization to unseen data.

The Decision Tree C4.5 algorithm demonstrates its capability to handle missing data, process both discrete and numeric data types, and generate easily interpretable rules. Meanwhile, K-nearest neighbor (KNN) excels in generalization, even with relatively small training datasets, due to its ease of implementation and capability to handle diverse types of data. [15]. On the other hand, backpropagation is quite reliable in solving problems as it train neural networks by iteratively adjusting the weights based on error gradients, enabling effective learning of complex patterns in the data [16], [17].

Taking into account the advantages of Decision Tree C4.5, KNN, and Backpropagation, this study attempts to combine them using Ensemble Voting. Setti and team implemented Ensemble Hard Voting on Multiclassifier Ensemble Learning with KNN, Naive Bayes, and Random Forest methods and got an accuracy value of 99.68%. Kajornrit and team apply Ensemble Voting on the Linear Regression, Backpropagation, SVM and KNN and get the results that Ensemble Voting can increase the accuracy value with an average MAE of 0.115% and RMSE of 0.175%. This research aims to demonstrate whether the Ensemble Voting method can improve the accuracy level of Decision Tree C4.5, KNN, and Backpropagation. We will compare the classification results of the Decision Tree C4.5, KNN, Backpropagation, and Ensemble Voting methods in classifying students who are at risk of drop out.

2. RESEARCH METHOD

2.1 Literature Review

2.1.1. Normalization

This study uses several variables including semester, the number of credits, IPK, IPS, student admission pathways and nominal of UKT. These data have different ranges between variables [18]. Researchers use the min-max normalization technique to change the data range. Data with different ranges will be converted into the range 0 to 1 using the min-max normalization formula as follows:

$$\boldsymbol{Norm(x)} = \frac{\boldsymbol{x - min}}{\boldsymbol{max - min}} \tag{1}$$

Which:

x : data to be normalized
min : minimum value of each attribute
max : maximum value of each attribute

2.1.2. Decision Tree C4.5

Decision Tree C4.5 is a classification model that uses a tree structure [19]. The output of this model is a decision and leaf nodes that contains rules. The decision tree construction process start with the selection of the root attribute. To determine the optimal attribute, we calculate the Gain Ratio measure, which takes into account both the information gain and the split information. The information gain quantifies the reduction in uncertainty provided by an attribute, while the split information measures the amount of information required to encode the split resulting from that attribute.

$$Entropy(S) = -\sum_{i=1}^{n} p_i * log_2\, p_i \tag{2}$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \left|\frac{S_i}{S}\right| * Entropy(S_i) \tag{3}$$

$$\text{Split Info}(S, A) = -\sum_{i=1}^{n} \left|\frac{S_i}{S}\right| * log_2\left(\left|\frac{S_i}{S}\right|\right) \tag{4}$$

$$\text{Gain Ratio}(S, A) = \frac{Gain(S,A)}{Split\ Info(S,A)} \tag{5}$$

Which:

S : set of training data
A : attribute training data
i : index on S (i = 1,2,3,...,n)
S_i : denotes the ith subset in sample S.
n : the number of S partitions
p_i : Proportion of S_i to S

Having identified the root attribute, we proceeded to split the dataset based on its unique values, generating child nodes corresponding to each branch. This splitting process continued recursively for each child node until terminating conditions were met. These conditions include the homogeneity of labels within a subset or the absence of remaining attributes to further partition the data.

Throughout the decision tree construction, various heuristics and pruning techniques were employed to enhance the generalization capabilities of the model. These measures aimed to prevent overfitting and improve the tree's ability to accurately classify unseen instances.

2.1.3. K-Nearest Neighbor (KNN)

KNN is a simple classification method using the basic principle of calculating the shortest distance [20]. KNN gives equal weight to each attribute. In KNN, we will find the distance between the two nearest k neighbors using the Euclidean formula. The accuracy of the KNN method is greatly influenced by the selection of the number of K. The formula for calculating the Euclidean distance is as follows [21].

$$Dist(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6}$$

Which:

Dist(x_i, y_i) : distance between objects (Euclidean Distancing) i^{-th}
xi : data sample
yi : test data
i : index on data (i = 1, 2, 3, ..., n)
n : amount of data

2.1.4. Backpropagation

In the first phase of Backpropagation, the network's weights and biases are randomly initialized, serving as initial values for the learning process. The forward pass is then executed, where input samples are fed into the network, and the activations of neurons in each layer are computed using the sigmoid activation function as equation 7.

$$f(x) = \frac{1}{1+e^{-x}} \tag{7}$$

Subsequently, the loss is calculated by comparing the predicted output with the actual output. Following the forward pass, error gradients are calculated for each weight and bias in the network, signifying their contribution to the overall error. The weights and biases of the network are updated to minimize the error by multiplying the gradients with a learning rate and subtracting the result from the current values.

This iterative process of forward pass, loss calculation, backpropagation, and weight update is repeated for each input sample in the training dataset. Multiple epochs are typically performed to refine the network's weights and enhance its performance on the training data. The convergence of training relies on stopping criteria, such as reaching a maximum number of epochs or achieving a desired performance level. This process showcases the iterative and incremental nature of learning in these sophisticated models.

2.1.5. Ensemble Voting

Ensemble Voting is one type of Ensemble Learning, a method that works by combining several machine learnings in the hope of getting more accurate results. Ensemble Voting that used in this research is included in the model mixing ensemble where this method will combine and train several models with different hyperparameter settings. Figure 2 shows a general description of how Ensemble Voting works.

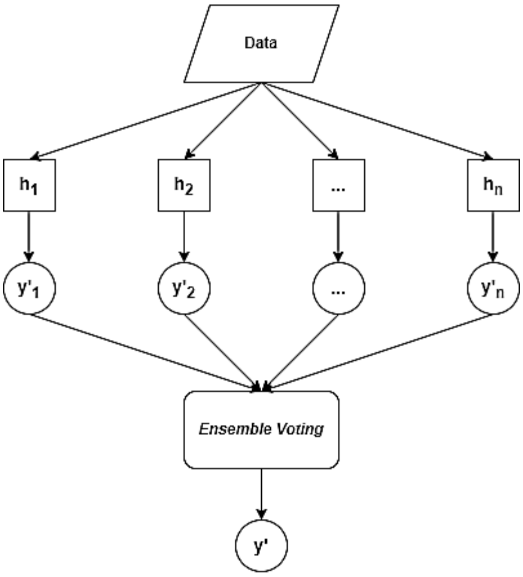

Figure 2. Ensemble Voting [22]

The image above shows the stages of the calculation process of the Ensemble Voting method. The data ready for calculation will be inputted into the model n times. In each of those models, it will produce y'_n. The results from those models will be inputted into the Ensemble Voting model. Optimization is performed on Ensemble Voting to obtain the result y'. In this study, we tried to use Ensemble Hard Voting and Ensemble Soft Voting [23].

a. Ensemble Hard Voting

The label on the dependent variable is obtained from the mode value in each classifier h_j.

$$\hat{y} = \text{mode}\{h_1(x), h_h(x), \dots, h_n(x)\} \tag{8}$$

b. Ensemble Soft Voting

The label on the dependent variable is obtained based on the predicted probability p for the classifier using the following equation 9.

$$\hat{y} = arg\ max\ i \sum_{j=1}^{m} w_j p_{ij} \tag{9}$$

where wj is the weight that can be assigned to the j-th classifier

2.1.6 Confusion Matrix

A confusion matrix is a table that is used to evaluate the performance of a classification model. It provides a summary of the predictions made by the model against the actual ground truth values. There are 4 sections in the confusion matrix table which contain true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values [24]. The formula for model evaluation that taken from confusion matrix is explained as follows [25].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \tag{10}$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \tag{11}$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \tag{12}$$

Accuracy measures the overall correctness of the model's predictions. Precision, on the other hand, evaluates the model's ability to avoid false positive errors. Recall, also known as sensitivity or true positive rate, assesses the model's ability to detect positive instances correctly.

2.2 Methodology

2.2.1. Stages of Research Methods

We used data on student status as the dependent variable (Y) and semester, number of credits, IPK, IPS, student admission pathways and nominal of UKT as independent variables (X). Figure 3 shows the stages of the research method.

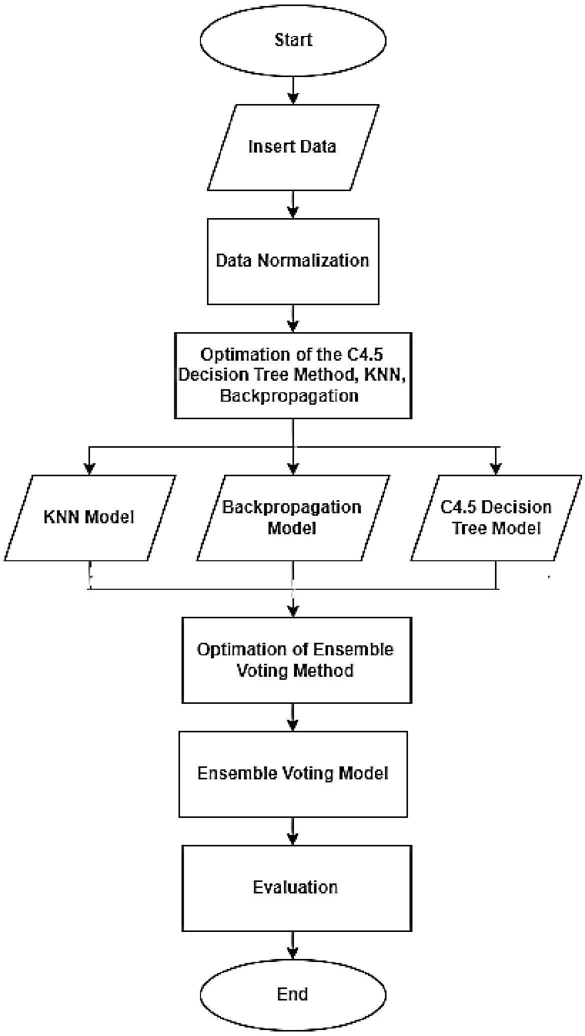

Figure 3. Stages of Research Methods

The first stage in this research is to input student data from the student database. The data will be selected, and complete and suitable data will be extracted for use in the research. A total of 1092 student data were obtained from several study programs in each faculty. Data normalization is performed because there are variables that have a wide range. Normalization is carried out with the expectation of obtaining good results. The data that has undergone pre-processing will be inputted into each individual method to obtain an optimal model. The results from each individual method that have been optimized will be combined into one method called Ensemble Voting. In Ensemble Voting, optimization will be conducted using rule and weight experimentation. When the best rule and weight have been found, classification will be performed using Ensemble Voting. The classification results from the Ensemble Voting method will be evaluated using accuracy, precision, and recall values. After obtaining accuracy, precision, and recall values, it will be possible to determine whether the results of the model are good or not.

3. RESULTS AND ANALYSIS

3.1. Research data

Classification was carried out using 1092 student data from all faculties at UIN Sunan Ampel Surabaya [26]. Researchers took student data from several study programs in the faculty. Table 1 shows the research data used.

Table 1. Research Data

No	Status	Semester	SKS	IPK	IPS	UKT	Admission Pathways
1	Active	3	40	3.62	0.00	2405000	Report Card
2	Active	3	47	3.77	3.77	2405000	Test
3	Active	3	40	3.35	3.35	3605000	Report Card
4	Active	3	40	3.57	0.00	3605000	Report Card
5	Active	4	91	3.56	3.63	3605000	Report Card
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1088	Drop Out	4	21	1.65	0.00	1325000	Test
1089	Drop Out	4	22	1.65	0.00	2935000	Test
1090	Drop Out	2	0	0.15	0.00	2215000	Report Card
1091	Drop Out	2	0	0.00	0.00	2215000	Report Card
1092	Drop Out	2	0	0.00	0.00	4030000	Report Card

The table above shows the data of 1092 student data with 557 active student data and 535 drop out student data. Table 2 shows the research data distribution.

Table 2. Distribution of research data

Data	Status	Semester	SKS	IPK	IPS	UKT	Admission Pathways
Minimum	0	2	0	0	0	150000	0
Mean	0.489	3.753	52.141	2.762	1.713	3004805	2.362
Maximum	1	4	96	3.9	3.9	9260000	4

The highest value and lowest value of each of these attributes will be used as a reference in normalizing data in the form 0 to 1. By using the formula in Equation (1), normalized data results displayed in the Table 3. The normalized dataset will be divided into training and testing data.

Table 3. Normalization of research data

No	Status	Semester	SKS	IPK	IPS	UKT	Admission Pathways
1	0.0	0.5	0.416667	0.928205	0.000000	0.247530	0.75
2	0.0	0.5	0.489583	0.966667	0.966667	0.247530	1.00
3	0.0	0.5	0.416667	0.858974	0.858974	0.379254	0.75
4	0.0	0.5	0.416667	0.915385	0.000000	0.379254	0.75
5	0.0	1.0	0.947917	0.912821	0.930769	0.379254	0.75
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1088	1.0	1.0	0.218750	0.423077	0.000000	0.128979	1.00
1089	1.0	1.0	0.229167	0.423077	0.000000	0.305708	1.00
1090	1.0	0.0	0.000000	0.038462	0.000000	0.226674	0.75
1091	1.0	0.0	0.000000	0.000000	0.000000	0.226674	0.75
1092	1.0	0.0	0.000000	0.000000	0.000000	0.425906	0.75

3.2. Results of the Single Methods Optimization

Both the three single methods and the ensemble method used a 90:10 data ratio with the aim of building an optimal model. Before entering the Ensemble Learning method, optimization is carried out for each single method. In Decision Tree C4.5, parameter testing was not carried out. In KNN and Backpropagation, tests were carried out on the value of k and the number of hidden nodes of two layers. Table 4 shows the optimization result on a single method:

Table 4. Single method optimization

Method		Accuracy	Precision	Recall
Decision Tree C4.5		95.45%	98.03%	92.59%
KNN	k=3	95.45%	98.03%	92.59%
	K=5	95.45%	96.22%	94.44%
	K=7	96.36%	100%	92.59%
	K=9	96.36%	100%	92.59%
Backpropagation	Hidden Nodes (10,10)	50.90%	0%	0%
	⋮	⋮	⋮	⋮
	Hidden Nodes (200,10)	90.90%	95.83%	95.18%
	⋮	⋮	⋮	⋮
	Hidden Nodes (500,500)	50.90%	0%	0%

The table above shows the accuracy, precision and recall values of each single method. In Decision Tree C4.5 we did not conduct parameter tests and obtained an accuracy value of 95.45%. In the KNN method, researchers tested the K value and obtained the best accuracy value of 96.36% using a K of 7. In the Backpropagation method, the best Hidden Nodes parameter is obtained with a value of (200.10) with an accuracy value of 90.90%. Of the 3 single methods above, the best accuracy value is obtained from the KNN method using the parameter k = 7 with the resulting accuracy value of 96.36%. Each model from a single method will be combined into the Ensemble Voting method in the hope of getting a better accuracy value.

3.3. Results of the Ensemble Voting method

In this ensemble voting method, we tested the types of ensemble voting rules, which consisted of Ensemble Hard Voting and Ensemble Soft Voting. We also conducted trials on the weight values used. The weight values to be tested are (1,1,1), (2,1,1), (1,2,1) and (1,1,2). The following is a table of test results for rules and weights for ensemble voting:

Tabel 5. Trial of Ensemble Voting Rules and Weight Parameters

Rule	Weight	Accuracy	Precision	Recall
Ensemble Hard Voting	-	96.36%	100%	92.59%
Ensemble Soft Voting	(1,1,1)	97.27%	98.11%	96.29%
	(2,1,1)	98.18%	100%	96.29%
	(1,2,1)	97.27%	98.11%	96.29%
	(1,1,2)	95.45%	98.03%	92.59%

From the test table for the rules and weight parameters above, the best accuracy value is the Ensemble Soft Voting rule using the weight parameters (2,1,1) with accuracy value of 98.18%. After getting the best weight rules and parameters, the accuracy values obtained will be compared with the single methods that have been optimized previously.

3.4. Evaluation of the Single Methods and the Ensemble Voting

Table 6 shows the accuracy, precision and recall results of the three single methods and the Ensemble method.

Tabel 6. Evaluation of Decision Tree C4.5, KNN, Backpropagation and Ensemble Voting

Metode	Accuracy	Precision	Recall
Decision Tree C4.5	95.45%	98.03%	92.59%
KNN	96.36%	100%	92.59%
Backpropagation	90.90%	95.83%	85.18%
Ensemble Soft Voting	98.18%	100%	96.29%

From the table above, it can be proven that ensemble voting is able to improve not only the accuracy value of each single method but also improve the precision value and recall value. Before classifying the 1092 data, we conducted a comparative trial of training and testing data. We conducted a data comparison trial with trials of 50:50, 60:40, 70:30, 80:20 and 90:10:

Tabel 7. Accuracy Value in Data Comparison Trial

Method	Comparison of Training Data: Test Data				
	50:50	60:40	70:30	80:20	90:10
Decision Tree C4.5	95.42%	96.56%	96.95%	97.26%	95.45%
KNN	95.05%	94.96%	94.81%	94.52%	96.36%
Backpropagation	86.81%	89.70%	86.89%	88.12%	90.90%
Ensemble Soft Voting	96.15%	96.79%	97.25%	97.26%	98.18%

From the above Table 7, the researchers used a 90:10 data ratio comparison because it had the highest accuracy of 98.18%. With these results, it is evident that the Ensemble Voting method can improve the accuracy level compared to the Decision Tree C4.5, KNN, and Backpropagation methods. As done by Assiri in 2020, a classification of Breast Tumor Using an Ensemble Machine Learning Method was performed by combining Logistic Regression, SVM with Backpropagation using Ensemble Hard Voting, showing an accuracy of 99.42%, precision of 99.40%, and recall of 99.40% [22].

4. CONCLUSION

Classification was carried out using 1092 student data with 557 active student data and 535 dropped out student data. By using 6 independent variables (X), namely student semester data, number of credits taken, IPK, IPS, student adminission pathways and UKT nominal, we combines 3 machine learning algorithms in one method, namely Ensemble Voting. We conducted several trials including a comparison trial of training and testing data, a trial of K values on the KNN method, a trial of hidden nodes values on the Backpropagation method and a trial of weight values on the Ensemble Voting method. After optimizing each single method, we combined the 3 methods into an Ensemble Voting which was divided into 2 rules, namely Ensemble Hard Voting and Ensemble Soft Voting. Single method Decision Tree C4.5 gets 95.45% accuracy, 98.03% precision and 92.59% recall. With a K parameter of 7 the KNN method gets 96.36% accuracy, 100% precision and 92.59% recall. By using the hidden layer (200.10) the Backpropagation method gets 90.90% accuracy, 95.83% precision and 95.18% recall. Trials were also carried out on comparisons of training and testing data, we used a comparison of training and testing data of 90:10 because of its highest accuracy. In the Ensemble Voting, rule that has the best accuracy results is Ensemble Soft Voting with a weight (2,1,1). Ensemble Voting is able to improve the accuracy, precision and recall of each single method. Ensemble Soft Voting obtains 98.18% accuracy, 100% precision and 96.29% recall. Suggestions for further research include conducting experimentation with a wider range of weight values in Ensemble Soft Voting, such as (0.1, 0.2, 0.8), (0.7, 0.3, 0.4), (0.9, 0.7, 0.5), and many others. In the next research, additional data can be included as independent variables, such as exploring other factors like parental occupation, parental income, the status of whether the student receives a scholarship or not, and other factors that may contribute to student dropouts.

REFERENCES

[1] A. Armansyah, “Prototipe Jaringan Syaraf Tiruan Multilayer Perceptron Untuk Prediksi Mahasiswa Dropout,” *J. Nas. Komputasi dan Teknol. Inf.*, vol. 4, no. 4, pp. 265–271, 2021, doi: 10.32672/jnkti.v4i4.3171.

[2] R. Manrique, B. P. Nunes, O. Marino, M. A. Casanova, and T. Nurmikko-Fuller, “An Analysis Of Student Representation, Representative Features And Classification Algorithms To Predict Degree Dropout,” *ACM Int. Conf. Proceeding Ser.*, pp. 401–410, 2019, doi: 10.1145/3303772.3303800.

[3] U. S. Aesyi, A. R. Lahitani, T. W. Diwangkara, and R. T. Kurniawan, “Deteksi Dini Mahasiswa Drop Out Menggunakan C5.0,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 6, no. 2, pp. 113–119, 2021, doi: 10.14421/jiska.2021.6.2.113-119.

[4] H. Aldowah, H. Al-Samarraie, A. I. Alzahrani, and N. Alalwan, “Factors Affecting Student Dropout In MOOCs: A Cause And Effect Decision-Making Model,” *J. Comput. High. Educ.*, vol. 32, no. 2, pp. 429–454, 2020, doi: 10.1007/s12528-019-09241-y.

[5] K. Coussement, M. Phan, A. De Caigny, D. F. Benoit, and A. Raes, “Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model,” *Decis. Support Syst.*, vol. 135, p. 113325, 2020, doi: 10.1016/j.dss.2020.113325.

[6] M. Alban and D. Mauricio, “Neural Networks To Predict Dropout At The Universities,” *Int. J. Mach. Learn. Comput.*, vol. 9, no. 2, pp. 149–153, 2019, doi: 10.18178/ijmlc.2019.9.2.779.

[7] M. Laufer and M. Gorup, “The Invisible Others: Stories Of International Doctoral Student Dropout,” *High. Educ.*, vol. 78, no. 1, pp. 165–181, 2019, doi: 10.1007/s10734-018-0337-z.

[8] G. A. S. Santos, K. T. Belloze, L. Tarrataca, D. B. Haddad, A. L. Bordignon, and D. N. Brandao, “Evolved Tree: Analyzing Student Dropout In Universities,” *Int. Conf. Syst. Signals, Image Process.*, vol. 2020-July, pp. 173–178, 2020, doi: 10.1109/IWSSIP48289.2020.9145203.

[9] D. Olaya, J. Vásquez, S. Maldonado, J. Miranda, and W. Verbeke, “Uplift Modeling for preventing

student dropout in higher education," *Decis. Support Syst.*, vol. 134, p. 113320, 2020, doi: 10.1016/j.dss.2020.113320.

[10] L. Bäulke, C. Grunschel, and M. Dresel, "Student dropout at university: a phase-orientated view on quitting studies and changing majors," *Eur. J. Psychol. Educ.*, vol. 37, no. 1, pp. 853–876, 2021, doi: 10.1007/s10212-021-00557-x.

[11] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: A machine learning approach," *Eur. J. High. Educ.*, vol. 10, no. 1, pp. 28–47, 2020, doi: 10.1080/21568235.2020.1718520.

[12] X. Dong and Z. Yu, "A Survey On Ensemble Learning," *Front. Comput. Sci.*, vol. 10, no. 1, pp. 1–18, 2019, doi: 10.1007/s11704-019-8208-z A.

[13] F. F. Patacsil, "Survival Analysis Approach For Early Prediction Of Student Dropout Using Enrollment Student Data And Ensemble Models," *Univers. J. Educ. Res.*, vol. 8, no. 9, pp. 4036–4047, 2020, doi: 10.13189/ujer.2020.080929.

[14] V. Senthil Kumaran and B. Malar, "Distributed Ensemble Based Iterative Classification For Churn Analysis And Prediction Of Dropout Ratio In E-Learning," *Interact. Learn. Environ.*, vol. 0, no. 0, pp. 1–16, 2021, doi: 10.1080/10494820.2021.1956547.

[15] R. Agrawal, *Smart Intelligent Computing And Applications*, vol. 104. Springer Singapore, 2019. doi: 10.1007/978-981-13-1921-1.

[16] I. S. Purba *et al.*, "Accuracy Level Of Backpropagation Algorithm To Predict Livestock Population Of Simalungun Regency In Indonesia," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012014.

[17] S. Setti, A. Wanto, M. Syafiq, A. Andriano, and B. K. Sihotang, "Analysis Of Backpropagation Algorithms In Predicting World Internet Users," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012018.

[18] M. Raharjo, M. Napiah, J. L. Putra, and M. Mustofa, "Prediksi Pengaruh Matakuliah Terhadap Peminatan Outline Tugas Akhir Mahasiswa Dengan Jaringan Syaraf Tiruan," *J. Infortech*, vol. 2, no. 1, pp. 78–83, 2020, doi: 10.31294/infortech.v2i1.7965.

[19] H. Sulistiani and A. A. Aldino, "Decision Tree C4.5 Algorithm For Tuition Aid Grant Program Classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia)," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, pp. 40–50, 2020, doi: 10.21107/edutic.v7i1.8849.

[20] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, "Transforming Big Data Into Smart Data: An Insight On The Use Of The K-Nearest Neighbors Algorithm To Obtain Quality Data," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 2, pp. 1–24, 2019, doi: 10.1002/widm.1289.

[21] O. Nurdiawan, D. A. Kurnia, D. Solihudin, T. Hartati, and T. Suprapti, "Comparison of the K-Nearest Neighbor algorithm and the decision tree on moisture classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012031, 2021, doi: 10.1088/1757-899x/1088/1/012031.

[22] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast Tumor Classification Using An Ensemble Machine Learning Method," *J. Imaging*, vol. 6, no. 6, 2020, doi: 10.3390/JIMAGING6060039.

[23] GitHub, "EnsembleVoteClassifier: A majority voting classifier," 2022. https://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier/ (accessed Jan. 09, 2023).

[24] D. Irmayanti, Y. Muhyidin, and D. A. Nurjaman, "Prediksi Mahasiswa Berpotensi Drop Out Dengan Metode Iteratif Dichotomiser 3 (ID3)," *J. Teknol. Inf.*, vol. 5, no. 2, pp. 103–113, 2021, doi: 10.36294/jurti.v5i2.2054.

[25] R. Sudiyarno, A. Setyanto, and E. T. Luthfi, "Peningkatan Performa Pendeteksian Anomali Menggunakan Ensemble Learning Dan Feature Selection," *Creat. Inf. Technol. J.*, vol. 7, no. 1, p. 1, 2021, doi: 10.24076/citec.2020v7i1.238.

[26] UINSA, "SIM Akademik Universitas Islam Negeri Sunan Ampel Surabaya," *2023*, 2023. https://sinau.uinsby.ac.id/siakad/data_mahasiswa/detail (accessed Jan. 09, 2023).

BIBLIOGRAPHY OF AUTHORS

Daffa Nur Cholis was born in Surabaya and is currently a graduate of Mathematics from UIN Sunan Ampel Surabaya. He has interest and experience in the field of Data Mining. His research focus is Machine Learning and Data Mining.

Her research interests revolve around the application of machine learning techniques in various domains, including computer vision and natural language processing.