International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 3
ISSN: 2088-5334

An Improved Accuracy of Multiclass Random Forest Classifier with Continuous Attribute Transformation Using Random Percentile Generation

Ronny Susetyoko [a,*], Elly Purwantini [b], Budi Nur Iman [b], Edi Satriyanto [a]

[a] *Department of Informatics and Computer Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, 60111 Indonesia*
[b] *Department of Electrical Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, 60111 Indonesia*
*Corresponding author: *rony@pens.ac.id*

Abstract—**This study aims to improve classification accuracy by transforming continuous attributes into categories by randomly generating percentile values as categorization limits. Four algorithms were compared for the generation of percentile values and selected based on the small variability of the percentile values and the distribution of the highest revenue expectations. The distribution of testing and training data classification accuracy becomes the second consideration. Random forest (RF) classification is modeled from selected percentiles with three transformation variations. The results of the ANOVA test, the algorithm with three variations of the transformation, has a mean that is not significantly different from the best model and the original dataset model. However, in some variations of training data, RF classification with continuous attribute transformation was superior to the original dataset model. The effectiveness of this continuous attribute transformation algorithm was very well applied to the LR, MLP, and NB methods. In the tuition fee dataset, the application of the algorithm for the three methods each had an accuracy of 0.178, 0.204, and 0.318. The results of the attribute transformation give a significant increase in accuracy to 0.967, 0.949, and 0.594 for each method, respectively. In the date fruits dataset, the attribute transformation was effective in the MLP method with an accuracy of 0.193 (original attribute) to 0.690 (continuous attribute transformation). The transformation results are effectively applied to the LR, MPL, and NB methods for datasets with continuous and categorical mixed attributes.**

Keywords— **Random forest; continuous attribute transform; random percentile generation; accuracy; revenue expectation.**

Manuscript received 27 Oct. 2022; revised 24 Dec. 2022; accepted 24 Jan. 2023. Date of publication 30 Jun. 2023.

I. Introduction

RF is an ensemble of decision trees based on bagging and random subspace concepts. The strength of unstable learners and the diversity among them are the ensemble models' core strengths [1]. The algorithm can reduce the impact of outliers and the possibility of overfitting with high accuracy [2] and demonstrate competitive accuracy across a wide range of tasks [3]. Decision trees as a classifier have many advantages, such as handling categorical data and dealing with outliers and noisy or missing data [4]. The RF algorithm has significant advantages: high generalizability and community noise, overfitting prevention, and fast processing of large-scale data [5].

RF was used to classify heart rate [6]. This method has also been used to classify single tuition fees. The accuracy of the method was compared with logistic regression (LR), Naïve Bayes (NB), and Multilayer Perceptron (MLP). The results showed that the RF method had the highest average accuracy (97.9%) [7]. This technique has proven very powerful, and many related algorithms have appeared over the years [8].

RF was also used to compare the multiclass classification of three crops: rice, sugar cane, and peanuts. The Improved Mahalanobis Taguchi System (IMTS) was the multiclass model based on normal observations and Mahalanobis distance for agricultural development. IMTS accuracy is 100%, and RF is 93.3%. Other methods, namely NB, J48, PART, AdaBoost, and Decision table, each have an accuracy of 80%, 73.3%, 73.3%, 73.3%, and 66.6% [9].

RF was used for the online classification of soil types based on waveforms. An RF algorithm has been implemented to clarify footprints into five types: flat, building, terrace, forest, and mountain. RF performance compared to linear support vector machine (linear-SVM), radial base function SVM (RBF-SVM), LR, K-nearest neighbor (KNN), and NB. As a result, RF had the best performance of the other four types of land. LR performs better than RBF-SVM, linear-SVM, and KNN. While NB has the worst performance. Classification accuracy for all methods was worse when the waveform is more complex [10].

Hybrid iris images transformed by RF algorithm on iris biometric identification have shown superior iris activation detection with 99.95% accuracy. The proposed transformation hybridization for feature extraction has demonstrated the ability to identify all nine types of iris spoofing attacks and proved robust [11].

The best, average, and worst accuracy RF model, called Best Average Worst (BAW), is used to determine the characteristics of the best, average, and worst accuracy functions in multiclass classification. The results of the ANOVA test showed no difference in the mean classification accuracy in the variation of the percentage level of the training data. However, there are significant differences in the mean accuracy for several variations in the number of attributes. With the polynomial regression model approach, there is a linear effect of the percentage of training data on the mean accuracy. And there was a linear or quadratic effect of the number of attributes on the mean accuracy. The results of this research, using 7 attributes, resulted in a classification accuracy of 96.9 - 97.6% for the tuition dataset and a classification accuracy ranging from 74.0% - 75.0% for the date fruits dataset [12]. BAW was used as a comparison in this study.

Besides RF, LR was also widely applied for classification. The LR classifier has the lowest error rate among the three related methods (k-nearest neighbor classifier, linear discriminant analysis classifier, and RF classifier) [13]. LR was also used for disease prediction based on clinical tumor stage data and total expression of the constructed Long noncoding RNA (LncRNA) transcript [14]. This method was also implemented for feature selection in a predictive model of recovery from hemorrhagic shock (HS) with resuscitation using blood in multiple rat animal protocols with better accuracy (84%) than the baseline classifier using only measured the heart rate (HR) and the mean arterial pressure (MAP) [15]. The LR classification model was used to select new scholarship recipients in the Indonesia Smart College Card or known as Kartu Indonesia Pintar Kuliah [16]. LR Classifier (LRC) neural network-based Convolution (CNN) is also used to predict obstetric ultrasound output with increased maternal and perinatal mobility [17].

Another application of the LR method was used to investigate the decision-making characteristics of drivers in overtaking on the highway. This regression model approach shows driving behavior with accurate estimates without the need for prior knowledge and contributes to various driving actions in dynamic environments [18]. The method was also used as a base model on the weakly supervised object localization problem using weighted regions due to its good performance in multi-instance settings [19].

An additive semi-supervised logistic regression model was also implemented to detect corporate credit anomalies based on the proportion of unlabeled sample information covering financial and non-financial variables. The results reveal the main financial variables that are correlated with the detection of firm credit anomaly and verify that the non-financial variables significantly improve the firm's credit anomaly model prediction accuracy [20].

Several machine learning (ML) Support Vector Machine (SVM), LR, RF, Extreme Gradient Enhancement (XGBoost), Decision Tree (DT), and Extra Tree (ET) compared their performance in detecting credit card fraud. This ML algorithm is combined with Adaptive Boosting Technique (AdaBoost) to improve the classification quality. The model was evaluated using accuracy, recall, precision, Matthews Correlation Coefficient (MCC), and Area Under the Curve (AUC). The experimental results show that using AdaBoost has a positive impact and is superior to other methods [21].

The same research was also carried out using ensemble classifier neural networks and hybrid data re-sampling methods. The ensemble classifier was obtained by using the Long Short-Term Memory (LSTM) neural network as a basic lesson in the adaptive boosting technique (AdaBoost). Meanwhile, hybrid re-sampling was achieved using the synthetic minority oversampling technique and the closest neighbor editing method (SMOTE-ENN). The experimental results show that the classifier performs better when trained with re-sampled data, and the LSTM ensemble outperforms other algorithms by obtaining a sensitivity and specificity of 0.996 and 0.998, respectively [22].

MaLCaDD (Machine Learning-based Cardiovascular Disease Diagnosis Framework) is proposed to predict cardiovascular disease with high precision. Feature Importance technique is used for feature selection. The LR Classifier Ensemble Model and K-Nearest Neighbor (KNN) are proposed for prediction with higher accuracy. Framework validation was carried out through three benchmark data sets (i.e.. Framingham, Heart Disease, and Cleveland), and the accuracy was achieved at 99.1%, 98.0%, and 95.5%, respectively. Comparative analysis proves that MaLCaDD prediction is more accurate (with less feature set) [23].

The Bayesian model was used as a recommendation system as well as a prediction system. This model is a user-based and item-based collaborative filtering approach, which recommends items using user information and similar items, respectively. Experiments conducted using four data sets gave good results compared to some advanced baselines, achieved the best performance using the Normalized Discounted Cumulative Gain (nDCG) measure of quality, and improved prediction accuracy across multiple datasets [24].

Hardware Naive Bayes classifier (NBC) real-time implemented in the field programmable gate array (FPGA). There are multiple processing element arrays (PEs) in the accelerator where each PE in the array runs in parallel, which speeds up the classification process. Experiments prove that the proposed accelerator has much better real-time efficiency than common processors [25].

Several classification techniques are also used for automated fall detection machines with approaches based on wearable sensors, ambient devices, and computer vision. Machine learning classifier methods such as LR, K-Nearest Neighbor, Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), RF, and MLP, show that the proposed approach is very effective. Classification accuracy and F1 scores can reach as high as 99% and 96%, respectively [26]. The NB under-sampling method approach is also used to improve classification performance for unbalanced datasets [27].

Several transformation techniques were used to increase the accuracy of the classification model, such as prototype selection, normalization, and feature mapping aimed at reducing the complexity and increasing the accuracy of the classification model [28]. The available datasets are generally not always normally distributed (Gaussian), and the distribution of variables tends to be skewed. Data normalization or transformation plays an important role in machine learning-based intrusion detection systems to achieve a high detection rate. Several methods are used to normalize data attributes before training model classification [29].

Invariant-scaling-based weight normalization can also speed up convergence and lower computations. Experiments show that the method can consistently improve the performance of various network architectures in large-scale datasets [30]. In other studies, the Deep Learning (DL) method is becoming more popular because of its outstanding performance in the field of disease detection. However, the performance of the DL method is affected by limitations such as dimensions, sparsity, and feature dominance. The normalization method was used to overcome the limitations. This technique was combined with ranking transformation and feature selection to further improve model performance [31]. Often there is a significant difference between the minimum and maximum values in different features, so Min-Max normalization was used to scale features within a range. This technique was applied to predict credit default [32].

Min-max normalization in video content was used in the associative knowledge graph. This step is done by considering the length of the different video content. As a result, the performance of the resulting association knowledge graph is better than the conventional association rule method [33]. Attribute transformations are also used in mobile applications to reduce memory and computational use [34].

This study aims to improve the performance of multiclass classification accuracy in RF classification by transforming continuous attributes into categories (ordinal). In this study, an ANOVA test is conducted to determine whether there was a significant effect of continuous attribute transformation results on the accuracy of the classification model. The two datasets used are the tuition fee and date fruits datasets [12]. Continuous type attributes are transformed into categorical attributes using the random generation of percentile values. There are 4 algorithms used for percentile generation.

The evaluation of the algorithm for the tuition fee dataset is based on the metrics of the distribution of revenue expectations, the distribution of testing data classification accuracy, and the distribution of training data accuracy. From the four algorithms, one algorithm is chosen as the best in one set of percentiles with the highest revenue expectation. Furthermore, the best algorithm compared with the original data classification model, the results of the best RF classification at [12], and several classification models with several variations of transformations on attributes that are not of continuous type are used for modeling. As a comparison, this algorithm is also applied to the date fruits dataset. The effectiveness of this algorithm was also tested on several other methods, namely LR, AdaBoost, Naïve Bayes, and MLP, using the same two datasets.

II. Materials and Method

A. Dataset

The dataset used in this study is the same as that carried out in [12], namely the tuition fee dataset (Uang Kuliah Tunggal) and the date fruits dataset. The first dataset is 873 raws, namely applicants who was accepted on Seleksi Bersama Masuk Politeknik Negeri (SBMPN) at Politeknik Elektronika Negeri Surabaya (PENS) in 2019 - 2021. The attributes used in the RF classification model are single tuition (as a label), income, homeownership, number of motorcycles, cars, electric capacity, other assets, and children. The second dataset of 897 rows, namely date fruits data, was taken from Kaggle. The attributes used are date type (as a label), equivalent area diameter, perimeter solidity, main axis convex area, minor axis area, eccentricity aspect ratio, equivalent diameter, and solidity.

B. Methodology

To complete this study using several stages, as seen in Fig. 1. After data collection and preprocessing, attributes of continuous type are transformed into categorical data using 4 algorithms. Algorithm-1 is performed by determining (k – 1) percentile values with equal distances. In algorithm-2 and Algorithm-4, the 2nd and 6th percentile values are determined, and other percentile values are determined randomly. While Algorithm-3, (k – 1) percentile values are determined randomly.

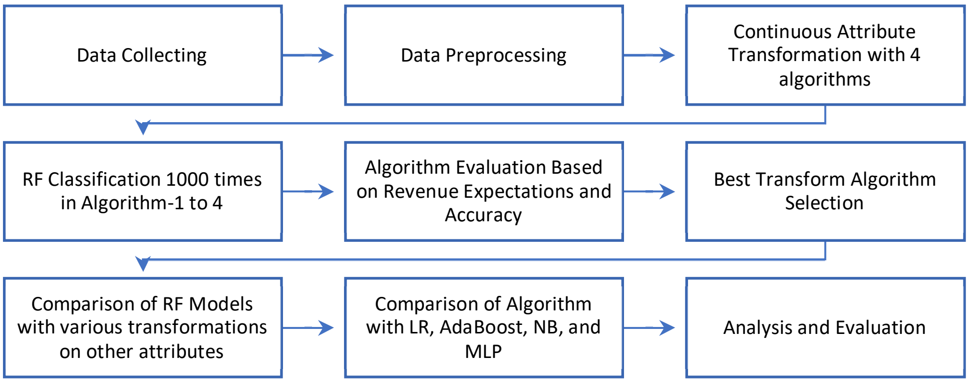
Data Collecting
Data Preprocessing
Continuous Attribute Transformation with 4 algorithms
RF Classification 1000 times in Algorithm-1 to 4
Algorithm Evaluation Based on Revenue Expectations and Accuracy
Best Transform Algorithm Selection
Comparison of RF Models with various transformations on other attributes
Comparison of Algorithm with LR, AdaBoost, NB, and MLP
Analysis and Evaluation

Fig. 1 Methodology

At the next stage, RF classifier modeling is carried out on Algorithm-1 (only once). While Algorithm-2, Algorithm-3, and Algorithm-4 are 1000 times each. The evaluation of the transformation method is based on the training data classification's accuracy, the testing data classification, and the value of the revenue expectations. From the best algorithms selected compared to the best models on [12]. Then the model is also compared with several other models with some transformation variations for attributes of continuous type. To determine the effectiveness of the algorithm, at the last stage, a comparison of RF with several other methods is also carried out, namely LR, AdaBoost, NB, and MLP. Furthermore, analysis and evaluation of the model obtained is carried out.

C. Random Percentile Generation Algorithm

The determination of (k – 1) percentile values as the limitation of the transformation of continuous attributes to categorical attributes is illustrated in Fig. 2.

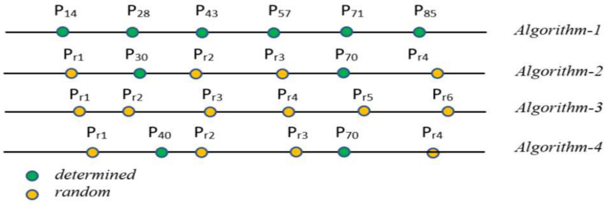

Fig. 2 Illustration of Random Generation of Percentile Values

While the 4 algorithms are as follows.

1) Algorithm-1: The steps of this algorithm are as follows:

- Determine (k – 1) percentiles with equal distances.
- Transformation of the selected continuous attribute into a category attribute (k category) based on the specified threshold of the percentile value.
- Model a new dataset using the RF Classification.
- Determine the value of model accuracy and distribution of revenue from classification results.
- Save the result.

2) Algorithm-2: The steps of this algorithm are as follows:

- For i = 1, 2, ..., 1000
  a. Determine (k – 1) a random number.
  b. Sort those random numbers from the smallest value to the largest value. Then each is divided by the total of the random numbers as (k – 1) percentile values.
  c. Replace the 2nd percentile with P_{30} and the 6th percentile with P_{70}.
  d. The selected continuous attribute is transformed into a categorical attribute (k category) based on the specified threshold of the percentile value.
  e. Model a new dataset using The RF Classification.
  f. Determine the value of model accuracy and distribution of revenue from classification results.
- Save the result.

3) Algoritma-3: The steps of this algorithm are as follows:

- For i = 1, 2, ..., 1000
  a. Determine (k – 1) a random number.
  b. Sort those random numbers from the smallest value to the largest value. Then each is divided by the total of the random numbers as (k – 1) percentile values.
  c. Transformation of the selected continuous attribute into a category attribute (k category) based on the specified threshold of the percentile value.
  d. Model a new dataset using the RF Classification.
  e. Determine the value of model accuracy and distribution of revenue from classification results.
- Save the result.

4) Algoritma-4: The steps of this algorithm are as follows:

- For i = 1, 2, ..., 1000
  a. Determine (k – 1) a random number.
  b. Sort those random numbers from the smallest value to the largest value. Then each is divided by the total of the random numbers as (k – 1) percentile values.
  c. Replace the 2nd percentile equals P40 and the 6th percentile equals P70.
  d. The selected continuous attribute is transformed into a categorical attribute (k category) based on the specified threshold of the percentile value.
  e. Model a new dataset using The RF Classification.
  f. Determine the value of model accuracy and distribution of revenue from classification results.
- Save the result.

D. Random Forest (RF)

RF is an ensemble method involving the construction of multiple CART via bootstrap sampling. The growth of a single CART in the RF is as follows:

1) The training set of each tree is generated by bootstrap sampling: n is the number of original training samples set, and randomly selected samples from the original training are determined using the bootstrap sampling method.

2) The internal nodes of each tree are selected from a randomly selected subset of candidate features: let the number of features in the original dataset be M and be positive the integer Mtry M is predefined; on each internal node, M try randomly selected feature of all M features as feature candidate, from this Mtry feature, we choose the best feature that can separate data sets.

3) Each tree is allowed to grow without being pruned.

The steps of a random forest are shown as follows.

- For i = 1: nTree
  a) Using the bootstrap method, each tree is given a training set with the size of n.
  b) Randomly select Mtry features at nodes, compare, and select the best features.
  c) Recursively generate each decision tree without pruning.

The classification is determined by majority voting [35]. The RF algorithm is shown in Fig. 3.

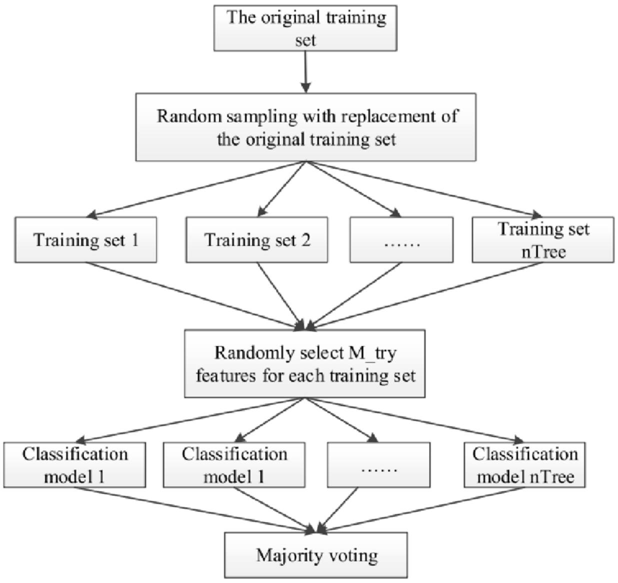
The original training set
Random sampling with replacement of the original training set
Training set 1
Training set 2
……
Training set nTree
Randomly select M_try features for each training set
Classification model 1
Classification model 1
……
Classification model nTree
Majority voting

Fig. 3 The Random Forest [35]

III. Result and Discussion

A. Distribution of Income

Of the seven attributes used for the RF classification, only the income attribute is of the continuous type. The distribution of income attributes is indicated as a histogram in Fig. 4. From the figure. The income data has a positive slope and a relatively high density of opportunities in income below Rp. 10 million.

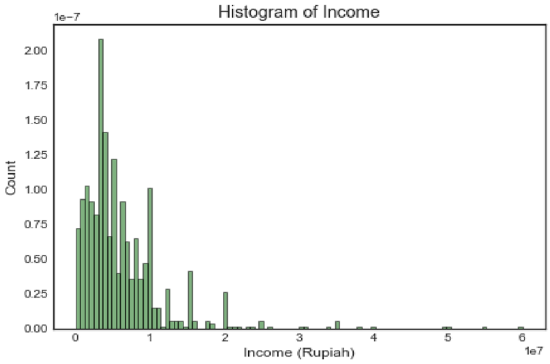

Fig. 4 Histogram of Income

Table I is a descriptive statistic of income attributes. The average income is Rp. 6,614,197 with a very large standard deviation of Rp. 6,183,685. The minimum income is 0 rupiah, and the maximum income is 60 million. The quartile-1 or Percentile-25 (P_{25}) value is Rp. 2,500,000, the Median (P_{50}) is Rp. 4,500,000, and quartile-3 (P_{75}) is Rp. 8,000,000. The income distribution has a positive slope of 3.53 with a tapered peak (leptokurtic) with a kurtosis of 19.78.

TABLE I
Descriptive Statistic of Income

Mean	6,164,197	Q1	2,500,000
Std Dev	6,183,685	Median	4,500,000
Minimum	0	Q3	8,000,000
Maximum	60,000,000	Skewness	3.53
Coef Variation	100	Kurtosis	19.78

QQ plot income attribute is shown in Fig. 5. Income is not distributed normally because the data distribution does not follow the red line as a reference. This is also proven in the Kolmogorov Smirnov Test (KS) results, a p-value of 0.00 with a value of KS = 0.973.

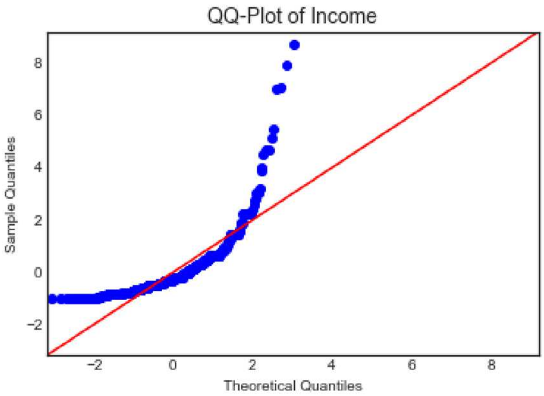

Fig 5 QQ-Plot of Income

B. Percentile Generation Results

Statistics of 1000 percentile generation results in Algorithm-2, Algorithm-3, and Algorithm-4 can be seen in Table II. In Algorithm-2, the values of P-1 (P_{30}) and P-4 (P_{70}) remain, amounting to IDR. 3 million and IDR. 7 million, respectively.

TABLE II
Statistics of Percentile Generation Results

Percentile	Statistics	Algorithm		
		Algo-2	Algo-3	Algo-4
P-1	Mean	2209347	392344	2761685
	Std Dev	407502	415086	436540
P-2	Mean	3000000	1232638	3788620
	Std Dev	0	499901	0
P-3	Mean	3959367	2197566	4894974
	Std Dev	647585	598479	792296
P-4	Mean	5282686	3304741	6708400
	Std Dev	892089	597865	1136677
P-5	Mean	7000000	4802547	7000000
	Std Dev	0	690959	0
P-6	Mean	8983966	7950671	14396005
	Std Dev	906739	908124	7333676

In Algorithm-4, the values of P-2 (P_{40}) and P-6 (P_{70}) are also fixed, amounting to IDR. 3.78862 million and IDR. 60 million, respectively. The variability of each percentile of random generation results in Algorithm-2 is between IDR. 407,502 – IDR. 906,739. In Algorithm-3, percentile variability ranges from IDR. 415,086 – IDR. 908,124. While in Algorithm-4, percentile variability is between IDR. 448,227 – IDR. 1,148,735. Fig. 6 indicates the percentile distribution of generation results in Algorithm-2. P-2 (P30) and P-5 (P70) have the highest density because the value is fixed. Percentile variability in P-1 is relatively smaller compared to variability in other percentiles.

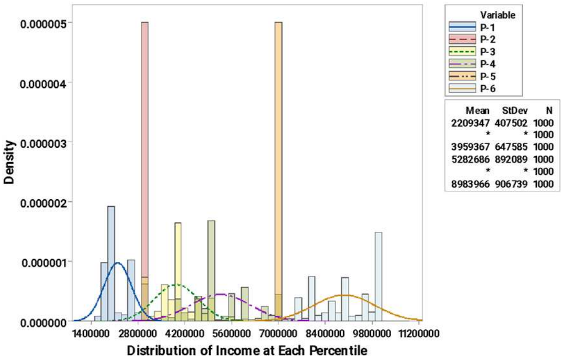

Fig. 6 Histogram of Percentile Distribution in Algorithm-2

C. *Algorithm Evaluation*

Based on the 4 percentile determination algorithms used, an algorithm evaluation was carried out based on the characteristics of the accuracy of the training data classification, the accuracy of the testing data classification, and the revenue expectation obtained. Fig. 7 shows the distribution of training data accuracy from Algorithm-2, Algorithm-3, and Algorithm-4. Of the 1000 modeling times, Algoritma-4 has an average accuracy of 0.994 with a standard deviation of 0.001. Algorithm-2 has an average accuracy of 0.991 with a standard deviation of 0.002. While Algorithm-3 has an average accuracy of 0.897 with a standard deviation of 0.002. Of the two statistics, Algorithm-4 is better than other algorithms. But it is also worth looking at some of the performances of the others.

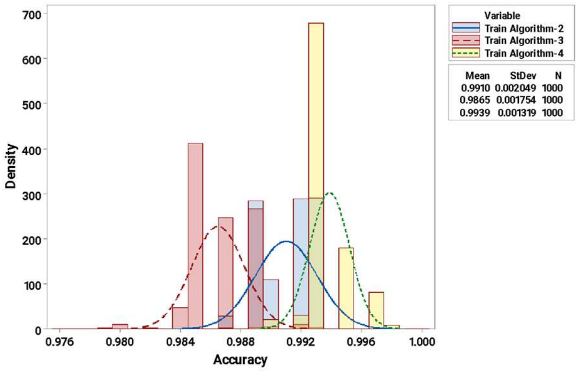

Fig. 7 Histogram of Training Accuracy

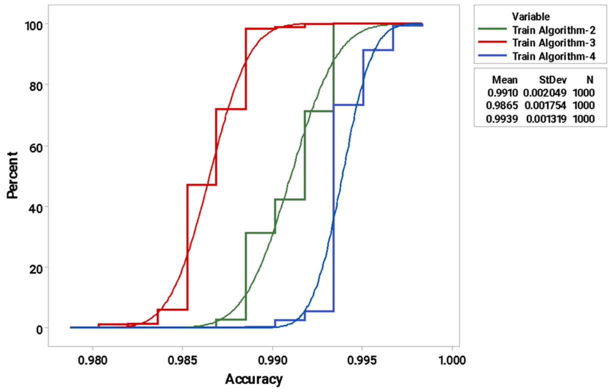

Fig. 8 Empirical CDF of Training Accuracy

The cumulative distribution function (CDF) of the training data classification accuracy of the 3 algorithms compared to the normal distribution CDF is shown in Fig. 8. Generally, the accuracy values are clustered and distributed at certain points only. This is also clearly visible on the Fig histogram. 6. When each of these CDFs is compared to a normal CDF curve has a considerable deviation. This indicates that the accuracy of the training data classification for the three algorithms is not normally distributed.

The results of normality testing using Kolmogorov-Smirnov (KS test) also concluded that the accuracy of the classification of training data from the three algorithms was not normally distributed. A summary of the statistics and results of the KS test are shown in Table III.

TABLE III
STATISTICS AND KS TEST OF TRAINING ACCURACY

Statistics	Method B	Method C	Method D
Mean	0.991	0.987	0.994
Mode	0.993	0.985	0.993
Median	0.992	0.987	0.993
KS test	0.226	0.232	0.369
P-value	<0.01	<0.00	<0.01

Fig. 9 shows the distribution accuracy of the classification of testing data from Algorithm-2, Algorithm-3, and Algorithm-4. With the same number of looping (1000 times modeling), Algoritma-4 has an average accuracy of 0.963 with a standard deviation of 0.007. Algorithm-2 has an average accuracy of 0.953 with a standard deviation of 0.005. While Algorithm-3 has an average accuracy of 0.949 with a standard deviation of 0.005. Of the two statistics, Algorithm-2 has an accuracy below Algorithm-4 but has a smaller variability.

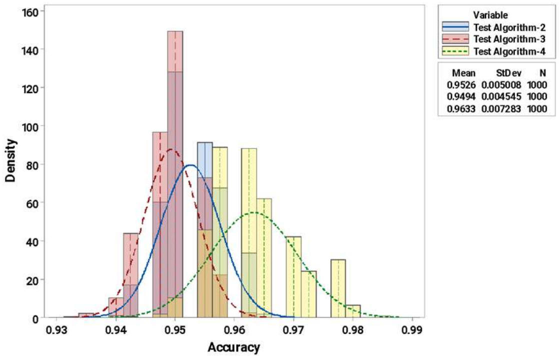

Fig. 9 Histogram of Testing Accuracy

Likewise, on Fig. 10 each CDF of testing data accuracy is also clustered and distributed at certain points only. The results of the KS test, the three distributions of testing data accuracy are not normally distributed with KS and p-values respectively being 0.189 (<0.01), 0.201 (<0.00), and 0.168 (<0.01). Fig. 11 indicates the distribution of the revenue expectations. Algorithm-2 and Algorithm-3 have the same average revenue expectation as IDR. 3,954 billion (per 1000 students).

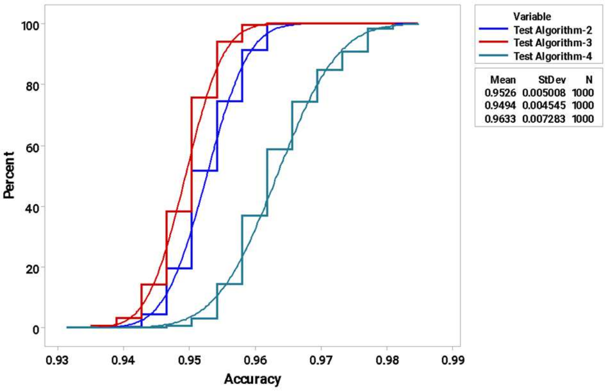

Fig. 10 Empirical CDF of Testing Accuracy

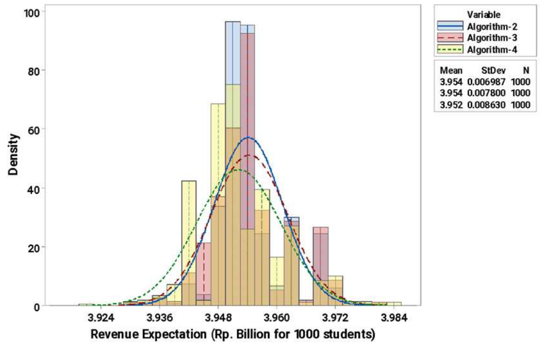

Fig. 11 Histogram of The Revenue Expectation

Meanwhile, the standard deviation of Algorithm-2 is IDR. 0.007 billion (per 1000 students) and Algorithm-3 is IDR. 0.008 billion (per 1000 students). So, Algorithm-2 has a smaller variability of the revenue expectations. Meanwhile, Algorithm-4 has a lower average revenue expectation, which is IDR. 3,952 billion (per 1000 students) with a standard deviation of IDR. 0,009 billion. On the Fig.12, the revenue expectations' histogram has somewhat different characteristics, where the distribution is more evenly distributed with a peak point in the middle. From Fig. 11 The CDFs of each earnings expectation are close to the normal CDF.

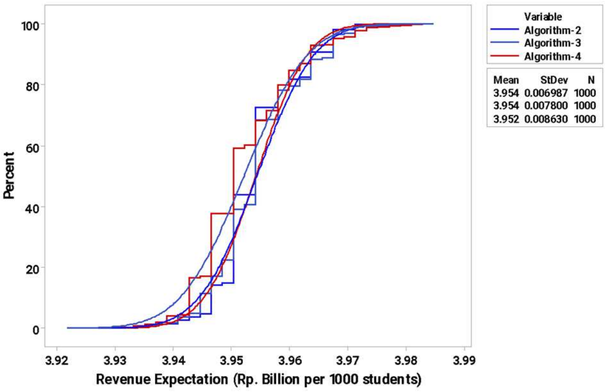

Fig. 12 Empirical CDF of The Revenue Expectation

However, from the results of the KS test, the revenue expectations of the three algorithms are also not normally distributed with KS and p-values respectively being 0.227 (<0.01), 0.955 (<0.01), and 0.189 (<0.01). The result of looping 1000 times when sorted from the smallest to largest revenue expectations is shown in Fig. 13. The highest frequency of the three algorithms is estimated to be between IDR. 3.95 million – IDR. 3.96 million.

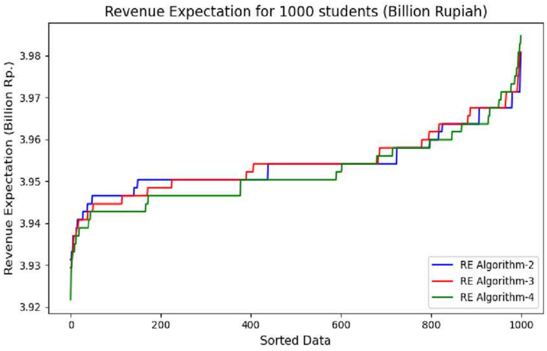

Fig. 13 The Revenue Expectation for 1000 Students (Billion Rupiah)

The distribution of tuition classes is shown on Fig. 14. From the 4 algorithms tried, class I to class IV obtained the same percentage. The difference lies in class V to class VII. Algorithm-2 can be chosen as the best model because the percentage of tuition in class VI is higher than algorithm-3. While in class VII it has the same percentage.

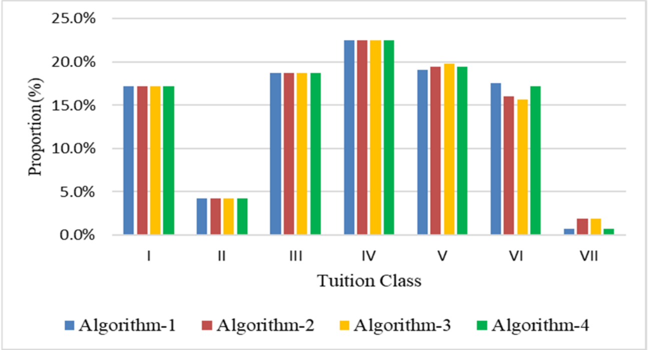

Fig. 14 Distribution of Single Tuition Class (%)

From the results of the evaluation of the distribution of training data accuracy, testing data accuracy, and revenue expectations, Algorithm-2 was chosen as the best model because the variability in each percentile tends to be smaller. Another consideration is that the 6th percentile value (P-6) is between P-6 in Algorithm-3 and Algorithm-4.

D. Comparison of Models

In the next stage of the 1000 iterations algorithm 2, a set of ordered percentiles is selected based on the highest testing data classification accuracy, the highest revenue expectations, and the highest training data classification accuracy. The RF model formed from the selected percentile is compared to the original model (without attribute transformation) and the best model on [12]. Algorithm-2 Model also has 3 variations tried, namely with a continuous attribute transformation (Model 1A), an Algorithm-2 model with a continuous attribute transformation and all other attributes with a MinMaxScaler() transformation called Model 1B, and Algorithm-4 with a continuous attribute transformation and all other attributes with a StandardScaler() transformation called Model 1C.

Classification accuracy with training data size of 50% to 90% with an increase of 5% is shown in Fig. 15.

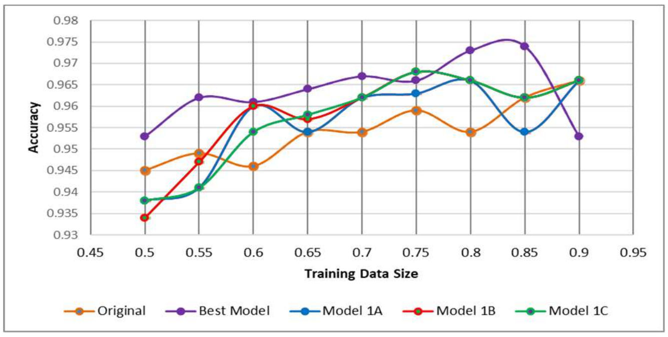

Fig. 15 Testing Model Accuracy (based on Training Data Size)

From Fig. 15 can be seen that in the original data model (preprocessing results) the accuracy of the monotony increases but the size of the training data is 60% - 80% lower than other RF models. Best model [12] tends to have the highest accuracy on the training data size as much as 50% - 70%, 80%, and 85%. However, the training data size as much as 90% is the lowest compared to other models. Approaches with polynomial equations from Model 1A, Model 1B, and Model 1C are shown in Table IV.

TABLE IV
PARAMETER TEST AND TOTAL VARIATIONS OF POLYNOMIAL REGRESSION

Parameter	Model 1A	Model 1B	Model 1C
β_o	0.014	0.001	0.000
β_1	0.051	0.004	0.002
β_2	0.077	0.006	0.003
Total variation (%)			
R^2	75.78	91.16	94.77
R^2-Ajd	67.71	88.20	93.03

With a 95% confidence level, the β_1 and β_2 parameters on the Model 1A are insignificant in the model. Of the three models, the total variation on the Model 1C was the highest with R^2 at 94.77% and R^2-Ajd at 93.03%. Model 1B also has a high total variation, namely R^2 at 91.16% and R^2-Ajd at 88.20%.

But on Fig. 16, boxplots of the 5 models tend to have the same mean. In the one-way ANOVA Test the value of F-value = 1.24 and p-value of 0.309 so it was concluded that there was no mean difference between the original model, the best model [12], and the C model, both C-1, C-2 and C-3.

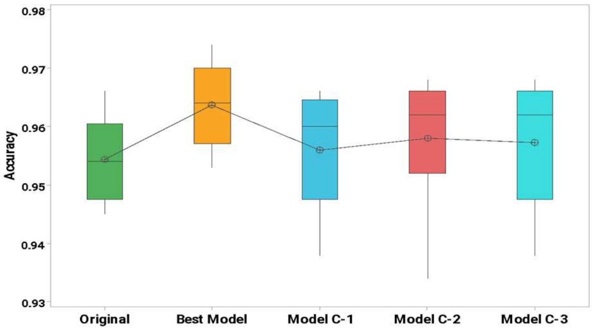

Fig. 16 Boxplot of Accuracy (Testing Data)

However, if you must choose from the 3 new models, the Model C-2 can be selected to apply. So of the three models developed, the Model C-2 was chosen as the best model in the case of the ukt dataset classification. In the next step, correlation testing is carried out, to see whether there is a significant relationship between the number of categories to the accuracy value. The scatter diagram between the number of categories with the accuracy of the Model C-1 and the Model C-2 is shown on Fig. 17. The correlation value between the number of categories and the classification accuracy of model C-1 is -0.223 with a p-value of 0.719. While the correlation value between the number of categories and the classification accuracy of model C-2 is -0.184 with a p-value of 0.767. This means that there is no significant correlation between the number of categories and the classification accuracy of both Model C-1 and Model C-2.

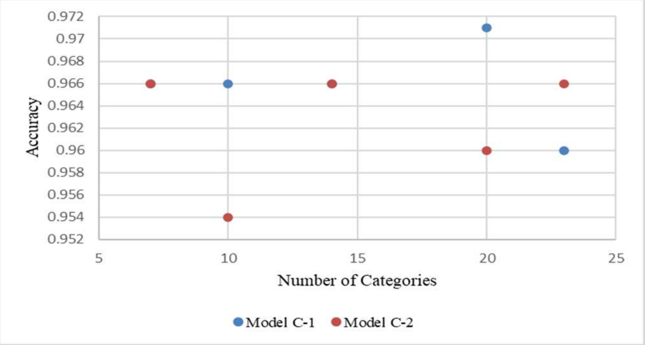

Fig. 17 Scatter Plot Number of Categories vs Accuracy

To find out the effectiveness of transforming continuous attributes into categorical categories of this algorithm, Fig. 18 shows the accuracy of the test data in the date fruits dataset with variations in the percentage of training data from 0.5 to 0.9. Models developed using Algorithm-2 (Model 1F) have relatively lower accuracy than the original and best models [12]. However, the model is still better when compared to the discretization model of each continuous attribute by rounding the multiplication/division results of the original attribute (Model 2F).

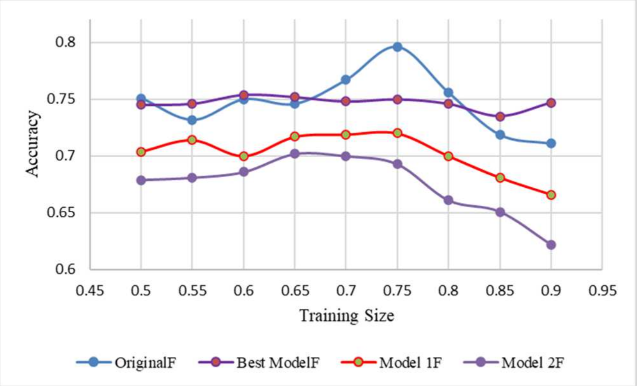

Fig. 18 Testing Model Accuracy (based on Training Data Size) for Date Fruits Dataset

E. Comparison of Methods

In the last stage, RF performance is compared to several other methods, namely LR, AdaBoost, NB, and MLP, both for the tuition fee and date fruits datasets. The results of modeling using some of these methods for the tuition fee dataset are shown in Table V.

TABLE V
ACCURACY FOR SEVERAL METHODS (ORIGINAL VS TRANSFORM)
DATASET: TUITION FEE

Method	**Test Size**				
	0.1	**0.2**	**0.3**	**0.4**	**0.5**
RF Ori	1	0.989	0.985	0.986	0.949
RF Trans	0.966	0.96	0.985	0.957	0.934
RL Ori	0.148	0.16	0.209	0.183	0.188
RL Trans	0.966	0.977	0.966	0.966	0.961
Ada Ori	0.295	0.246	0.259	0.277	0.281
Ada Trans	0.432	0.577	0.218	0.189	0.199
NB Ori	0.318	0.274	0.351	0.377	0.271
NB Trans	0.625	0.611	0.569	0.591	0.572
MLP Ori	0.079	0.251	0.244	0.257	0.189
MLP Trans	0.966	0.96	0.927	0.957	0.936

One-way ANOVA is used to test the difference in the mean of classification accuracy in the testing data. ANOVA test results on the tuition fee dataset, F-value = 154.18, p-value = 0.000 and R^2 = 97.20%. That is, there is at least one significantly different means of accuracy.

Table VI compares the mean with the Tukey method for all methods with a 95% confidence level. The table shows that the RF classification by the original attribute and the transformation result attribute (specific to the continuous attribute) does not differ significantly. For LR and MLP methods, there is a very significant difference/increase in the accuracy value of the original attribute classification with the transformed result attribute. The original accuracy on LR and MLP was 0.178 and 0.204, respectively. However, after the transformation of continuous attributes, they became 0.967 and 0.949, respectively. The AdaBoost method has no difference before and after it is transformed. While using the Naïve Bayes method, there is a significant increase in its accuracy value, which is from 0.318 to 0.594.

TABLE VI
TUKEY PAIRWISE COMPARISONS (DATASET: TUITION FEE)

Factor	**Mean**	**Grouping**
RF Ori	0.982	A
RL Trans	0.967	A
RF Trans	0.960	A
MLP Trans	0.949	A
NB Trans	0.594	B
Ada Trans	0.323	C
NB Ori	0.318	C
Ada Ori	0.272	C D
MLP Ori	0.204	C D
RL Ori	0.178	D

The results of modeling using some of these methods for the date fruits dataset are shown in Table VII.

TABLE VII
ACCURACY FOR SEVERAL METHODS (ORIGINAL VS TRANSFORM)
DATASET: DATE FRUITS

Method	Test Size				
	0.1	0.2	0.3	0.4	0.5
RF Ori	0.678	0.767	0.752	0.733	0.755
RF Trans	0.622	0.661	0.704	0.683	0.677
RL Ori	0.633	0.656	0.696	0.653	0.648
RL Trans	0.667	0.706	0.726	0.728	0.715
Ada Ori	0.422	0.639	0.637	0.342	0.664
Ada Trans	0.611	0.25	0.474	0.303	0.448
NB Ori	0.622	0.6	0.648	0.644	0.646
NB Trans	0.667	0.689	0.741	0.694	0.704
MLP Ori	0.367	0.089	0.196	0.125	0.187
MLP Trans	0.611	0.711	0.726	0.694	0.708

One-way ANOVA test results on the date fruits dataset, F-value = 32.82, p-value = 0.000 and R^2 = 84.27%. That is, there is at least one significantly different means of accuracy. The comparison test of the mean with the Tukey method for all methods with a confidence level of 95% is shown in Table VIII. From the table it can be explained that the classification of RF with the original attribute and the attribute of the result of the transformation differs significantly. The result of the transformation decreases the accuracy of the classification. For the LR method, there is no significant difference/increase in the accuracy value of the original attribute classification with the transform result attribute, although the accuracy value is different. The MLP method in the original dataset has an accuracy of 0.193 after being transformed to 0.690 (significantly different). In the AdaBoost and NB methods, there was no significant difference in the accuracy mean value of the original dataset with the transformed dataset.

TABLE VIII
TUKEY PAIRWISE COMPARISONS (DATASET: DATE FRUITS)

Factor	**Mean**	**Grouping**
RF Ori	0.737	A
RL Trans	0.708	A
NB Trans	0.699	A B
MLP Trans	0.690	A B
RF Trans	0.669	A B
RL Ori	0.657	A B
NB Ori	0.632	A B
Ada Ori	0.541	B C
Ada Trans	0.517	C
MLP Ori	0.193	D

IV. CONCLUSION

The percentile value generation algorithm is used to determine the limits of categorization of continuous attributes that are not normally distributed or distribution-free. At the algorithm evaluation stage, Algorithm-2 was chosen as the best algorithm on the grounds that the variability of each percentile is relatively smaller than that of the other algorithms. Another main consideration is that Algorithm-2 has the revenue expectation distribution with the highest mean value of IDR. 3.94 billion (per 1000 students), and the smallest standard deviation of IDR. 0.007 billion (per 1000 students). Meanwhile, the next considerations are the distribution of the accuracy of the classification of testing data and the accuracy of training data successively. In Algorithm-2, the mean and standard deviation of the testing data's accuracy and training data's accuracy are 0.953 (0.005) and 0.991 (0.002), respectively.

With the one-way ANOVA test, some models built from Algorithm-2 with this continuous attribute transformation have a mean that does not differ significantly from the best model at [12] and models with the original dataset attribute. However, when observed at some variations in the training data level, RF classification with continuous attribute transformations is superior to models with native attribute datasets.

At the final stage, a comparison of the RF method with several other methods is carried out to determine the effectiveness of the continuous attribute transformation algorithm. The results are quite encouraging. The transformation of continuous attributes into category types (ordinals) is very effective when some attributes have ordinal/category types, as in the case of tuition fee datasets. Some classification methods have a very significant change in accuracy, where LR rose from 0.178 to 0.967 (an increase of 443.3%), MLP rose from 0.204 to 0.949 (an increase of 365.2%) and NB rose from 0.318 to 0.594 (an increase of 86.8%). In the case of the date fruits dataset with all attributes of continuous type, the continuous attribute transformation algorithm is effective only on the MLP method, which is from 0.193 to 0.690 (an increase of 257.5%). For the rest of the methods there are no significant differences.

The technique of transforming continuous attributes into categories is less effective to apply to RF classification. However, the technique is effectively applied to the classification of LR, MLP, and NB, especially if the types of attributes used in the classification model are diverse. In the application of other datasets, the selection of percentile sets can be based on the accuracy of the classification of training data and testing data, also considering other metrics according to the problems encountered.

Acknowledgment

We thank the Center for Research and Community Service of Politeknik Elektronika Negeri Surabaya which has provided funding and facilities for this research.

References

[1] M. A. Ganaie, M. Tanveer, P. N. Suganthan, and V. Snasel, "Oblique and rotation double random forest," *Neural Networks*, vol. 153, pp. 496–517, 2022, doi: 10.1016/j.neunet.2022.06.012.

[2] L. Linhui, J. Weipeng, and W. Huihui, "Extracting the Forest Type from Remote Sensing Images by Random Forest," *IEEE Sens. J.*, vol. 21, no. 16, pp. 17447–17454, 2021, doi: 10.1109/JSEN.2020.3045501.

[3] A. Dmitry Devyatkin and G. Oleg Grigoriev, "Random Kernel Forests," *IEEE Access*, vol. 10, no. July, pp. 77962–77979, 2022, doi: 10.1109/ACCESS.2022.3193385.

[4] M. Gencturk, A. Anil Sinaci, and N. K. Cicekli, "BOFRF: A Novel Boosting-based Federated Random Forest Algorithm on Horizontally Partitioned Data," *IEEE Access*, vol. 10, no. August, pp. 89835–89851, 2022, doi: 10.1109/ACCESS.2022.3202008.

[5] Y. Zhu and H. Peng, "Multiple Random Forests Based Intelligent Location of Single-phase Grounding Fault in Power Lines of DFIG-based Wind Farm," *J. Mod. Power Syst. Clean Energy*, vol. 10, no. 5, pp. 1152–1163, 2022, doi: 10.35833/mpce.2021.000590.

[6] C. Zou *et al.*, "Heartbeat Classification by Random Forest With a Novel Context Feature: A Segment Label," *IEEE J. Transl. Eng. Heal. Med.*, vol. 10, no. August 2022, doi: 10.1109/JTEHM.2022.3202749.

[7] R. Susetyoko, W. Yuwono, E. Purwantini, and N. Ramadijanti, "Perbandingan Metode Random Forest , Regresi Logistik , Naïve Bayes , dan Multilayer Perceptron Pada Klasifikasi Uang Kuliah Tunggal (UKT)," vol. 7, no. 1, 2022.

[8] J. Biedrzycki and R. Burduk, "Weighted scoring in geometric space for decision tree ensemble," *IEEE Access*, vol. 8, no. 3, pp. 82100–82107, 2020, doi: 10.1109/ACCESS.2020.2990721.

[9] N. Deepa, M. Z. Khan, B. Prabadevi, D. R. P. M. Vincent, P. K. R. Maddikunta, and T. R. Gadekallu, "Multiclass model for agriculture development using multivariate statistical method," *IEEE Access*, vol. 8, pp. 183749–183758, 2020, doi: 10.1109/ACCESS.2020.3028595.

[10] X. Liu, X. Liu, Z. Wang, G. Huang, and R. Shu, "Classification of Laser Footprint Based on Random Forest in Mountainous Area Using GLAS Full-Waveform Features," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 2284–2297, 2022, doi: 10.1109/JSTARS.2022.3151332.

[11] S. Khade, S. Gite, S. D. Thepade, B. Pradhan, and A. Alamri, "Detection of Iris Presentation Attacks Using Hybridization of Discrete Cosine Transform and Haar Transform with Machine Learning Classifiers and Ensembles," *IEEE Access*, vol. 9, pp. 169231–169249, 2021, doi: 10.1109/ACCESS.2021.3138455.

[12] R. Susetyoko, W. Yuwono, E. Purwantini, and B. N. Iman, "Characteristics of Accuracy Function on Multiclass Classification Based on Best, Average, and Worst (BAW) Subset of Random Forest Model," pp. 410–417, 2022, doi: 10.1109/ies55876.2022.9888374.

[13] Q. Lei, H. Zhang, H. Sun, and L. Tang, "Fingerprint-Based Device-Free Localization in Changing Environments Using Enhanced Channel Selection and Logistic Regression," *IEEE Access*, vol. 6, pp. 2569–2577, 2017, doi: 10.1109/ACCESS.2017.2784387.

[14] B. Wang and J. Zhang, "Logistic Regression Analysis for LncRNA-Disease Association Prediction Based on Random Forest and Clinical Stage Data," *IEEE Access*, vol. 8, pp. 35004–35017, 2020, doi: 10.1109/ACCESS.2020.2974624.

[15] A. Lucas, A. T. Williams, and P. Cabrales, "Prediction of Recovery from Severe Hemorrhagic Shock Using Logistic Regression," *IEEE J. Transl. Eng. Heal. Med.*, vol. 7, no. June, pp. 1–9, 2019, doi: 10.1109/JTEHM.2019.2924011.

[16] R. Susetyoko, Wiratmoko Yuwono, and Elly Purwantini, "Model Klasifikasi Pada Seleksi Mahasiswa Baru Penerima KIP Kuliah Menggunakan Regresi Logistik Biner," *J. Inform. Polinema*, vol. 8, no. 4, pp. 31–40, 2022, doi: 10.33795/jip.v8i4.914.

[17] Z. Zhang and Y. Han, "Detection of Ovarian Tumors in Obstetric Ultrasound Imaging Using Logistic Regression Classifier with an Advanced Machine Learning Approach," *IEEE Access*, vol. 8, pp. 44999–45008, 2020, doi: 10.1109/ACCESS.2020.2977962.

[18] J. C. Nwadiuto, S. Yoshino, H. Okuda, and T. Suzuki, "Variable Selection and Modeling of Drivers' Decision in Overtaking Behavior Based on Logistic Regression Model with Gazing Information," *IEEE Access*, vol. 9, pp. 127672–127684, 2021, doi: 10.1109/ACCESS.2021.3111753.

[19] L. Wang, T. Wang, and X. Hu, "Logistic regression region weighting for weakly supervised object localization," *IEEE Access*, vol. 7, pp. 118411–118421, 2019, doi: 10.1109/ACCESS.2019.2935011.

[20] S. Han, "Semi-supervised learning classification based on generalized additive logistic regression for corporate credit anomaly detection," *IEEE Access*, vol. 8, pp. 199060–199069, 2020, doi: 10.1109/ACCESS.2020.3035128.

[21] E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.

[22] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A Neural Network Ensemble with Feature Engineering for Improved Credit Card Fraud Detection," *IEEE Access*, vol. 10, pp. 16400–16407, 2022, doi: 10.1109/ACCESS.2022.3148298.

[23] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," *IEEE Access*, vol. 9, pp. 106575–106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

[24] P. Valdiviezo-Diaz, F. Ortega, E. Cobos, and R. Lara-Cabrera, "A Collaborative Filtering Approach Based on Naïve Bayes Classifier," *IEEE Access*, vol. 7, pp. 108581–108592, 2019, doi: 10.1109/ACCESS.2019.2933048.

[25] Z. Xue, J. Wei, and W. Guo, "A Real-Time Naive Bayes Classifier Accelerator on FPGA," *IEEE Access*, vol. 8, pp. 40755–40766, 2020, doi: 10.1109/ACCESS.2020.2976879.

[26] T. Le Minh, L. Van Tran, and S. V. T. Dao, "A Feature Selection Approach for Fall Detection Using Various Machine Learning Classifiers," *IEEE Access*, vol. 9, pp. 115895–115908, 2021, doi: 10.1109/ACCESS.2021.3105581.

[27] C. K. Aridas, S. Karlos, V. G. Kanas, N. Fazakis, and S. B. Kotsiantis, "Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers under Imbalanced Data Sets," *IEEE Access*, vol. 8, pp. 2122–2133, 2020, doi: 10.1109/ACCESS.2019.2961784.

[28] J. Ortiz-Bejar, E. S. Tellez, M. Graff, D. Moctezuma, and S. Miranda-Jimenez, "Improving k Nearest Neighbors and Naïve Bayes Classifiers

through Space Transformations and Model Selection," *IEEE Access*, vol. 8, pp. 221669–221688, 2020, doi: 10.1109/ACCESS.2020.3042453.

[29] M. A. Siddiqi and W. Pak, "An Agile Approach to Identify Single and Hybrid Normalization for Enhancing Machine Learning-Based Network Intrusion Detection," *IEEE Access*, vol. 9, pp. 137494–137513, 2021, doi: 10.1109/ACCESS.2021.3118361.

[30] Q. Yuan and N. Xiao, "Scaling-Based Weight Normalization for Deep Neural Networks," *IEEE Access*, vol. 7, pp. 7286–7295, 2019, doi: 10.1109/ACCESS.2018.2890373.

[31] M. Mulenga, S. A. Kareem, A. Q. M. Sabri, and M. Seera, "Stacking and Chaining of Normalization Methods in Deep Learning-Based Classification of Colorectal Cancer Using Gut Microbiome Data," *IEEE Access*, vol. 9, pp. 97296–97319, 2021, doi: 10.1109/ACCESS.2021.3094529.

[32] T. M. Alam *et al.*, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201173–201198, 2020, doi: 10.1109/ACCESS.2020.3033784.

[33] H. J. Kim, J. W. Baek, and K. Chung, "Associative Knowledge Graph Using Fuzzy Clustering and Min-Max Normalization in Video Contents," *IEEE Access*, vol. 9, pp. 74802–74816, 2021, doi: 10.1109/ACCESS.2021.3080180.

[34] J. H. Soeseno, D. S. Tan, W. Y. Chen, and K. L. Hua, "Faster, Smaller, and Simpler Model for Multiple Facial Attributes Transformation," *IEEE Access*, vol. 7, pp. 36400–36412, 2019, doi: 10.1109/ACCESS.2019.2905147.

[35] C. Zhang, X. Wang, S. Chen, H. Li, X. Wu, and X. Zhang, "A modified random forest based on Kappa measure and binary artificial bee colony algorithm," *IEEE Access*, vol. 9, pp. 117679–117690, 2021, doi: 10.1109/ACCESS.2021.3105796.