Available Online : https://proceeding.researchsynergypress.com/index.php/cset/index
RSF Conference Series: Engineering and Technology
ISSN 2809-6843 (Online) | 2809-6878 (Print)
Volume 1 Number 1 (2021): 407-413

The Comparison of Tree-Based Ensemble Machine Learning for Classifying Public Datasets

Nur Heri Cahyana[1], Yuli Fauziah[2], Agus Sasmito Aribowo[3]

[1, 2, 3]Informatics Department, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Indonesia

Abstract

This study aims to determine the best methods of tree-based ensemble machine learning to classify the datasets used, a total of 34 datasets. This study also wants to know the relationship between the number of records and columns of the test dataset with the number of estimators (trees) for each ensemble model, namely Random Forest, Extra Tree Classifier, AdaBoost, and Gradient Bosting. The four methods will be compared to the maximum accuracy and the number of estimators when tested to classify the test dataset. Based on the results of the experiments above, tree-based ensemble machine learning methods have been obtained and the best number of estimators for the classification of each dataset used in the study. The Extra Tree method is the best classifier method for binary-class and multi-class. Random Forest is good for multi-classes, and AdaBoost is a pretty good method for binary-classes. The number of rows, columns and data classes is positively correlated with the number of estimators. This means that to process a dataset with a large row, column or class size requires more estimators than processing a dataset with a small row, column or class size. However, the relationship between the number of classes and accuracy is negatively correlated, meaning that the accuracy will decrease if there are more classes for classification.

Keywords: Tree-Based Ensemble Machine Learning, Public Datasets, Estimators, Accuracy

INTRODUCTION

Classification is an important part of data mining to solve business and technical problems. Classification using machine learning is supervised. This algorithm requires knowledgeable data to build a model. One measure of the success of the classification model is the accuracy of the model. The accuracy of the model depends on several things, ranging from the condition of the dataset, pre-processing techniques, feature extraction methods, and machine learning algorithms. The characteristic of supervised learning is a learning method that reuses data and outputs that have been entered by users or done by the system in the past. Some examples of basic classification algorithms that apply supervised learning methods are the Naïve Bayes algorithm, Decision Tree, Support Vector Machine, and K-Nearest Neighbour.

Many researches on machine learning algorithms for public dataset classification have been carried out, including the iris dataset (Vatshayan, 2019). These basic algorithms have their characteristic limitations. Especially for tree-based methods, its limitations are sensitive to outliers, unstable and allow overfitting to occur. The development of a tree-based algorithm is an ensemble tree. Tree based ensemble machine learning are machine learning methods that apply bagging and boosting techniques using tree-based algorithms. Bagging is a method that can improve the results of machine learning classification algorithms by combining prediction classifications from several models. It is used to overcome instability in complex models with relatively small data sets. Bagging is one of the earliest and simplest and most effective ensemble- based algorithms. Bagging is best suited for problems with relatively small training datasets. Bagging adopts bootstrap distribution in order to generate different base learners, to obtain subset data. bagging also adopts a base learner output aggregation strategy, namely the voting method for classification cases and averaging for regression cases. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement (Sewell, 2009).

Corresponding author:
sasmito.skom@upnyk.ac.id
DOI: 10.31098/cset.v1i1.412

Boosting is a way to generate multiple models or classifiers for prediction or classification, and also combine predictions from multiple models into a single prediction. Boosting is an iterative approach to generate strong classifiers and minimal training errors from a group of weak classifiers. Boosting is generally designed for binary class problems.

Research related to the use of ensemble tree-based algorithms is (Raghavendra & Santosh Kumar, 2020) about the performance of random forest in prediction of diabetes on Pima Indian dataset. The experiment was conducted using R studio platform and achieved classification accuracy of 84.1%. Another research is (Rajendar et al., 2020) on classifier models for early prediction of diabetes. In this research, random forest gives accuracy 78.6%. Research from (Yu et al., 2020) used five datasets including Blood, Haberman, Iris, Seeds and Wine from the UCI dataset and used modified random forest for machine learning. The result of this research is the classification accuracy of each dataset, namely Blood (75.94%), Haberman (72.37%), Iris (73.90%), Seeds (71.15%), and Wine (75.64%). Research from (Sharma et al., 2020) used the iris dataset to be classified by Random Forest with very satisfactory results. Research from (Cahyana et al., 2019) uses seven datasets to be classified using Gradient Boosting. The results of the classification accuracy of each dataset are for Mammography (0.78), Liver Disorders (0.77), Pima Indian (0.79), Indian Liver (0.71), Haberman (0.62), and Immunotherapy (0, 83). This study succeeded in increasing the accuracy of around 2-11% after oversampling. Research (Prasetiyowati et al., 2020) describes the performance of the Random Forest method on high-dimensional datasets such as the Parkinson, CNAE-9, and Urban Land Cover datasets. The average accuracy using K-fold Cross Validation for Parkinson (86.66%), CNAE-9 (93.72%), and Urban Land Cover (85.08%). The study (Ahmad et al., 2021) used HbA1c and FPG labelled datasets and the accuracy for HbA1c (81.48%) and FPG (88.27%). The studies above do not explain how many best estimators (trees) are used to produce high accuracy. This knowledge is needed to determine the complexity of the machine learning process which can ultimately be useful for considering the tree-based ensemble machine learning method to be chosen.

This study aims to determine the best methods of tree-based ensemble machine learning to classify the datasets used, a total of 34 datasets. This study also wants to know the relationship between the number of records and columns of the test dataset with the number of estimators (trees) for each ensemble model. There are 4 tree-based ensemble machine learning methods to be compared in this study, namely Random Forest, Extra Tree Classifier, AdaBoost, and Gradient Bosting. The four methods will be compared to the maximum accuracy and the number of estimators when tested to classify the test dataset. Specifically, the Random Forest and Extra Tree Classifier methods have been studied before, for example for sentiment analysis (A S Aribowo et al., 2020) and cross domain sentiment analysis (Agus Sasmito Aribowo et al., 2021). To measure maximum accuracy, this study will try out estimator values from the lowest to the highest on the four tree-based ensemble machine learning methods.

RESEARCH METHOD

This research is an experimental study to determine the best tree-based ensemble machine learning method for several public datasets. The research starts from getting the dataset, cleaning the dataset, eliminating the damaged data, and imputing the missing data.

The datasets used in this study are public datasets available in several repositories such as the UCI Machine Learning Repository (https://archive.ics.uci.edu/), datahub.io, and several other sites that allow the dataset to be freely downloaded. . We conducted trials on 34 types of datasets that varied in the number of records, the number of columns, and classes. Our datasets are divided into two groups, namely binary-class and multi-class datasets as shown in Table 1.

Table 1. Datasets for Experiment

Binary-Class					Multi-Class				
No.	Dataset	Record Number	Column Number	Class Number	No.	Dataset	Record Number	Column Number	Class Number
1	Sonar	208	61	2	1	Seeds	210	8	3
2	WDBC	360	31	2	2	Wine	178	14	3
3	Iono	351	35	2	3	Iris	150	6	3
4	WPBC	194	34	2	4	Glass	214	11	6
5	Haber	306	4	2	5	Libras	360	91	15
6	Musk1	476	167	2	6	CNAE	1080	857	9
7	Musk2	6598	167	2	7	Vowel	990	14	10
8	PIMA	768	9	2	8	DUser	258	6	4
9	Park	195	23	2	9	Ecoli	307	9	4
10	Climate	540	21	2	10	Letter	20000	17	26
11	Trans	748	5	2	11	Robot	5456	5	4
12	plrx	182	13	2	12	Cardio	2126	23	10
13	QSAR	1055	42	2	13	Opt	5620	65	10
14	Bank	1372	5	2	14	Landsat	6435	37	6
15	Hill	1212	101	2	15	Page	5473	10	5
16	Thyroid	2000	25	2	16	Pen	10992	17	10
					17	CNAE	1080	857	9
					18	WRed	1599	12	6

The research will try out four tree-based ensemble machine learning on all datasets in this study. Each model will also be tested to find out how many bootstraps (trees) are optimally providing the highest accuracy. So we created a number of models for each bootstrap test. The number of bootstrap is also known as the number of estimators. The number of estimators will be tried starting from 10 estimators, 20 estimators, and so on up to 200 estimators. More details can be found in the research steps.

Research steps

The research steps are the sequences of the research process to answer the research questions. The question is what is the best tree-based ensemble machine learning method for each dataset and how many estimators to achieve it. The unit of measurement is classification accuracy. All datasets are used in research and get trials from all tree-based ensemble machine learning with the number of each estimator varying between 10 estimators to 200 estimators.
The research steps were :

1. Normalize data in every dataset. Handling missing values using imputation methods.
2. READ dataset X
    a. Split dataset into data training and data testing with ratio 8:2
    b. Set estimator, N = 10..200, STEP 10 :
        i. Creating 4 models using ensemble machine learning and data training
            1. Model 1: Using Random Forest(estimator = N)
            2. Model 2: Using AdaBoost(estimator = N)
            3. Model 3: Using Gradient Boosting(estimator = N)
            4. Model 4: Using Extra Tree Classifier(estimator = N)
        ii. Validate data testing using Model 1, Model2, Model 3 and Model4.
    c. Get Max(Accuracy) and N from Validation Model1, Model2, Model3, and Model4.

The results of the research are the best accuracy figures from each tree-based ensemble machine learning in each dataset. Another research result is the number of estimators to get the highest accuracy.

RESULTS

The result of the research is the maximum accuracy rate obtained for each type of tree-based ensemble machine learning ensemble per dataset after trying estimators ranging from 10, 20, to 200. The output of each experiment is the maximum accuracy and the estimator for each type of tree-based ensemble machine learning per dataset. Table 2 is a summary of the experimental results for all binary-class datasets. Table 3 is a summary of the experimental results for all multi-class datasets.

Table 2. Summary of The Experimental Results for All Binary-Class Datasets

No	Dataset	Record number	Column Number	RF		ADA		GRA		EXT	
				Estim	Acc	Estim	Acc	Estim	Acc	Estim	Acc
1	Sonar	208	61	130	0,889	**20**	**0,905**	140	0,889	70	0,905
2	WDBC	360	31	50	0,982	10	0,972	110	0,963	**30**	**0,991**
3	Iono	351	35	20	0,943	**20**	**0,962**	20	0,953	50	0,962
4	WPBC	194	34	50	0,797	**10**	**0,831**	200	0,797	190	0,814
5	Haber	306	4	30	0,674	10	0,652	20	0,663	**80**	**0,685**
6	Musk1	476	167	200	0,888	**130**	**0,902**	190	0,888	160	0,909
7	Musk2	6598	167	130	0,976	**190**	**0,981**	190	0,971	80	0,976
8	PIMA	768	9	170	0,784	30	0,753	40	0,784	**160**	**0,801**
9	Park	195	23	30	0,966	190	0,983	40	0,966	**50**	**0,983**
10	Climate	540	21	**10**	**0,944**	10	0,926	50	0,938	30	0,926
11	Trans	748	5	100	0,729	**120**	**0,769**	140	0,764	40	0,729
12	plrx	182	13	20	0,709	30	0,727	130	0,691	**20**	**0,727**
13	QSAR	1055	42	**80**	**0,893**	60	0,868	140	0,886	70	0,890
14	Bank	1372	5	10	0,993	50	1,000	110	0,995	**20**	**1,000**
15	Hill	1212	101	190	0,544	10	0,536	90	0,508	**190**	**0,552**
16	Thyroid	2000	25	20	0,990	30	0,988	**10**	**0,990**	130	0,982
Information : Estim = estimators, Acc = Accuracy											
RF=Random Forest, EXT = Extra Tree, ADA=AdaBoost, GRA=Gradient Boosting											

Table 2 shows that the best accuracy for the classification of binary-class datasets is the extra tree classifier, followed by AdaBoost. Extra Tree classifier is very good at classifying Bank, WDBC and Park datasets. While AdaBoost is better at classifying Sonar, Iono, WPBC, and Musk1 datasets. This assessment is carried out by observing the accuracy value and the number of estimators. The fewer estimators, the algorithm works more efficiently and processing time will certainly be faster. Table 2 also shows that there is a relationship between the number of dataset records and the number of estimators.

Table 3. Summary of the Experimental Results for All Multi-Class Datasets

No	Dataset	Rec Num	Column Number	Class Num	RF		ADA		GRA		EXT	
					Estim	Acc	Estim	Acc	Estim	Acc	Estim	Acc
1	Seeds	210	8	3	**10**	**0.952**	30	0.937	110	0.952	**10**	**0.952**
2	Wine	178	14	3	40	1.000	10	0.926	20	0.963	**30**	**1.000**
3	Iris	150	6	3	**10**	**1.000**	10	0.978	10	0.978	**10**	**1.000**
4	Glass	214	11	6	20	0.985	10	0.831	10	0.985	**30**	**1.000**

5	Libras	360	91	15	190	0.872	10	0.101	30	0.752	**60**	**0.899**
6	CNAE	1080	857	9	60	0.914	10	0.531	70	0.907	**60**	**0.926**
7	Vowel	990	14	10	140	0.973	70	0.182	190	0.906	**40**	**0.990**
8	DUser	258	6	4	40	0.962	10	0.808	180	0.949	**70**	**0.974**
9	Ecoli	307	9	4	30	0.978	10	0.839	**20**	**0.989**	110	0.978
10	Letter	20000	17	26	150	0.962	20	0.329	200	0.946	180	0.971
11	Robot	5456	5	4	**10**	**1.000**	10	0.850	180	0.999	80	0.995
12	Cardio	2126	23	10	**30**	**0.912**	10	0.484	130	0.906	80	0.908
13	Opt	5620	65	10	130	0.988	20	0.675	180	0.979	**140**	**0.990**
14	Landsat	6435	37	6	130	0.919	30	0.775	200	0.909	**110**	**0.919**
15	Page	5473	10	5	**70**	**0.971**	10	0.920	190	0.968	30	0.970
16	Pen	10992	17	10	180	0.993	40	0.281	160	0.991	**40**	**0.995**
17	CNAE	1080	857	2	90	0.914	10	0.531	50	0.907	**140**	**0.932**
18	WRed	1599	12	9	190	0.688	10	0.592	120	0.650	**120**	**0.696**
Information : Estim = estimators, Acc = Accuracy												
RF=Random Forest, EXT = Extra Tree, ADA=AdaBoost, GRA=Gradient Boosting												

Table 3 shows that the best accuracy for the classification of multiclass datasets is extra tree classifier, followed by Random Forest. Extra Tree classifier is very good at classifying almost all datasets. Meanwhile, Random Forest is very good for classifying Robot and Cardio datasets. This assessment is carried out by observing the accuracy value and the number of estimators. Likewise in experiments on binary-class datasets, the fewer the number of estimators, the more efficient the algorithm works and the faster processing time. Table 3 also shows that there is a relationship between the number of dataset records and the number of estimators.

Then it will be calculated how much is the relationship between the number of records and the number of the best method estimators for binary-class and multi class. Table 4 shows the relationship between the number of dataset records and the number of estimators in the binary-class datasets.
Table 5. Relationship between the number of records, columns and the estimator count in multi-class datasets.

Table 4. Relationship between the Number of Dataset Records, Column and Number of Estimators in the Binary-Class Datasets

No	**Dataset**	**Record number**	**Column Number**	**Estimator Number**	**Best Accuracy**	**Best Method**
1	Sonar	208	61	20	0,905	ADA
2	WDBC	360	31	30	0,991	EXT
3	Iono	351	35	20	0,962	ADA

4	WPBC	194	34	10	0,831	ADA
5	Haber	306	4	80	0,685	EXT
6	Musk1	476	167	130	0,902	ADA
7	Musk2	6598	167	190	0,981	ADA
8	PIMA	768	9	160	0,801	EXT
9	Park	195	23	50	0,983	EXT
10	Climate	540	21	10	0,944	RF
11	Trans	748	5	120	0,769	ADA
12	plrx	182	13	20	0,727	EXT
13	QSAR	1055	42	80	0,893	RF
14	Bank	1372	5	20	1,0	EXT
15	Hill	1212	101	190	0,552	EXT
16	Thyroid	2000	25	10	0,990	GRA
Method Information :						
RF=Random Forest, EXT = Extra Tree, ADA=AdaBoost, GRA=Gradient Boosting						

If it is calculated using Pearson correlation, the variable number of rows and the estimator is correlated with a score of 0.497, where the Pearson correlation coefficient (r) is a value ranging from -1 to 1 to indicate the strength of the association. This value indicates that the two variables are positively connected. Similarly, the number of columns and the number of estimators are also connected with the Pearson correlation score of 0.579.

Then it will be calculated how much the relationship between the number of records and the number of the best method estimators for multi-class will be calculated. Table 5 shows the relationship between the number of dataset records and the number of classes in multi-class datasets, with the number of estimators and accuracy values.

Table 5. Relationship between the Number of Dataset Records, Column, Class and Number of Estimators in the Multi-Class Datasets

No	Dataset	Record Number	Column Number	Class Number	Estimator Number	Best Accuracy	Best Method
1	Seeds	210	8	3	10	0,952	EXT
2	Wine	178	14	3	30	1,000	EXT
3	Iris	150	6	3	10	1,000	EXT
4	Glass	214	11	6	30	1,000	EXT
5	Libras	360	91	15	60	0,899	EXT
6	CNAE	1080	857	9	60	0,926	EXT
7	Vowel	990	14	10	40	0,990	EXT
8	DUser	258	6	4	70	0,974	EXT
9	Ecoli	307	9	4	20	0,989	GRA
10	Letter	20000	17	26	180	0,971	EXT
11	Robot	5456	5	4	10	1,000	RF
12	Cardio	2126	23	10	30	0,912	RF
13	Opt	5620	65	10	140	0,990	EXT
14	Landsat	6435	37	6	110	0,919	EXT
15	Page	5473	10	5	70	0,971	RF

16	Pen	10992	17	10	40	0,995	EXT
17	CNAE	1080	857	2	140	0,932	EXT
18	WRed	1599	12	9	120	0,696	EXT
Method Information :							
RF=Random Forest, EXT = Extra Tree, ADA=AdaBoost, GRA=Gradient Boosting							

In Table 5, if calculated using Pearson correlation, the number of rows and estimators is connected to the Pearson correlation score of 0.549, and the relationship between the number of columns and the estimator variable is connected to the Pearson correlation score of 0.528. This value indicates that the number of rows and the number of columns is positively related to the number of estimators. The relationship between the number of classes and the number of estimators is also positively correlated with a Pearson correlation score of 0.528, and the relationship between the number of classes and the negative Pearson correlated accuracy is -0.354. This means that the number of classes and accuracy will be inversely proportional.

Conclusion

Based on the results of the experiments above, tree-based ensemble machine learning methods have been obtained and the best number of estimators for the classification of each dataset used in the study. The Extra Tree method is the best classifier method for binary-class and multi-class. Random Forest is good for multi-classes, and AdaBoost is a pretty good method for binary-classes. Then the knowledge is obtained that the number of rows, columns and data classes is positively correlated with the number of estimators. This means that to process a dataset with a large row, column or class size requires more estimators than processing a dataset with a small row, column or class size. However, the relationship between the number of classes and accuracy is negatively correlated, meaning that the accuracy will decrease if there are more classes for classification.

REFERENCES

Ahmad, H. F., Mukhtar, H., Alaqail, H., Seliaman, M., & Alhumam, A. (2021). Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. *Applied Sciences (Switzerland)*, *11*(3), 1–18. https://doi.org/10.3390/app11031173

Aribowo, A S, Basiron, H., Herman, N. S., & Khomsah, S. (2020). An evaluation of preprocessing steps and tree-based ensemble machine learning for analysing sentiment on Indonesian youtube comments. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(5), 7078–7086. https://doi.org/10.30534/ijatcse/2020/29952020

Aribowo, Agus Sasmito, Basiron, H., Yusof, N. F. A., & Khomsah, S. (2021). Cross-Domain Sentiment Analysis Model On Indonesian Youtube Comment. *International Journal of Advances in Intelligent Informatics*, *7*(1), 12–25. https://doi.org/10.26555/ijain.v7i1.554

Cahyana, N., Khomsah, S., & Aribowo, A. S. (2019). Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting. *Proceeding - 2019 5th International Conference on Science in Information Technology: Embracing Industry 4.0: Towards Innovation in Cyber Physical System, ICSITech 2019*, 217–222. https://doi.org/10.1109/ICSITech46713.2019.8987499

Prasetiyowati, M. I., Maulidevi, N. U., & Surendro, K. (2020). Feature selection to increase the random forest method performance on high dimensional data. *International Journal of Advances in Intelligent Informatics*, *6*(3), 303–312. https://doi.org/10.26555/ijain.v6i3.471

Raghavendra, S., & Santosh Kumar, J. (2020). Performance evaluation of random forest with feature selection methods in prediction of diabetes. *International Journal of Electrical and Computer Engineering*, *10*(1), 353–359. https://doi.org/10.11591/ijece.v10i1.pp353-359

Rajendar, S., Thangaraj, R., Palanisamy, J., & Kaliappan, V. K. (2020). Comparative Analysis of Classifier

Models for the Early Prediction of Type 2 Diabetes. *International Journal of Advanced Science and Technolo*, *29*(7), 2184–2194.

Sewell, M. (2009). Ensemble learning. *Academia-UCL Department of Computer Science*, *4*(1), 2776. https://doi.org/10.4249/scholarpedia.2776

Sharma, K., College, S., & Rajhasthan. (2020). Classification of IRIS Dataset using Weka. *International Journal of Computer Applications & Information Technology*, *12*(1), 287–291.

Vatshayan, S. (2019). Performance Evaluation of Supervised Learning for Iris Flower Species. *INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY*, *April*.

Yu, Y., Wang, L., Huang, H., & Yang, W. (2020). An Improved Random Forest Algorithm. *Journal of Physics: Conference Series*, *1646*(1). https://doi.org/10.1088/1742-6596/1646/1/012070

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*, 308–324. https://doi.org/10.1016/j.trc.2015.02.019