International Journal of Engineering and Computer Science Applications (IJECSA)
Vol. 1, No. 1, Maret 2022, pp.35~41
DOI: 10.30812/IJECSA.v1i1.1805

Ensemble Implementation for Predicting Student Graduation with Classification Algorithm

Ria Rismayati[1,*], Ismarmiaty[2], Syahroni Hidayat[3]

[1] *Department Of Computer Science, Faculty of Engineering, Universitas Bumigora,Mataram, Indonesia*

[2] *Department Of Information System, Faculty of Engineering, Universitas Bumigora,Mataram, Indonesia*

[3]*Lembaga Pengembangan Ilmu Pengetahuan dan Teknologi, Sekawan Insitute*

[*]*Corresponding author: riris@universitasbumigora.ac.id*

(Received Feb 25 2022 ; Revised Feb 28, 2022 ; Accepted Mar 01, 2022)

***Abstract*─ Graduating on time at the higher education level is one of the main targets of every student and university institution. Many factors can affect a student's length of study, the different character of each student is also an internal factor that affects their study period. These characters are used in this study to classify data groups of students who graduated on time or not. Classification was chosen because it is able to find a model or pattern that can describe and distinguish classes in a dataset. This research method uses the esemble learning method which aims to see student graduation predictions using a dataset from Kaggle, the data used is a IPK dataset collected from a university in Indonesia which consists of 1687 records and 5 attributes where this dataset is not balanced. The intended target is whether the student is predicted to graduate on time or not. The method proposed in this study is Ensemble Learning Different Contribution Sampling (DCS) and the algorithms used include Logistic Regression, Decision Tree Classifier, Gaussian, Random Forest Classifier, Ada Bost Classifier, Support Vector Coefficient, KNeighbors Classifier and MLP Classifier. From each classification algorithm used, the test value and accuracy are calculated which are then compared between the algorithms. Based on the results of research that has been carried out, it is concluded that the best accuracy results are owned by the MLPClassifier algorithm with the ability to predict student graduation on time of 91.87%. The classification model provided by the DCS-LCA used does not give better results than the basic classifier of its constituent, namely the MLPClassifier algorithm of 91.87%, SVC of 91.64%, Logistic Regression of 91.46%, AdaBost Classifier of 90.87%, Random Forest Classifier of 90.45% , and KNN of 89.80%.**

***Keywords* : Esemble Learning, Ensemble Learning Different Contribution Sampling, Classification, Graduation.**

I. INTRODUCTION

Graduating on time at the higher education level is one of the main targets of every student and university institution. Universities are expected to be able to produce people who have morals like people who are educated and competent in their fields. To achieve this goal, the quality of higher education is influenced by the accreditation assessment based on the rules issued by the National Accreditation Board for Higher Education (BAN – PT). One aspect of the assessment of the many benchmarks by BAN - PT is the graduation of students on time. Student graduation on time benefits both parties, where students who graduate on time will help assess the accreditation of a college as well as the student will not pay tuition fees again in the semester concerned and can quickly work after graduating from college [1].

The large number of students at a university that has several study programs makes it difficult for early detection to see the potential for student graduation. Early detection aims to be able to study student patterns and behaviors so that they can minimize tardiness and increase student graduation. The punctuality of graduation for each level of students has different criteria in each program, where students from the D3 (Diploma) program are said to have graduated on time if they can complete less than or equal to 3 years of study. S1 (Bachelor) program students are said to graduate on time if they can complete less than or equal to 4 years of study. Likewise, S2 (Master) program students are said to graduate on time if they can complete less or the same two years of study and S3 (Doctoral) if they can complete less or the same three years of study [2]

Many factors can affect a student's length of study, the different character of each student is also an internal factor that affects their study period. From the character that is an internal factor in each student which is implemented in the lecture bench, it produces a value in each semester and is referred to as the Indeks Prestasi Kumulatif (IPK). Apart from the IPK, this research also utilizes the database owned by each university to store data in the form of academic data and student biodata, so that hidden information can be known by processing student data [3]. Based on these two characters, this research was developed, the IPK values used were in the

first, second, third and fourth semesters, and the second character used was academic recap regarding students who graduated on time and did not.

The characters used in this study will then be used to classify groups of students who graduate on time or not. Classification was chosen because it is able to find a model or pattern that can describe and distinguish classes in a dataset, aiming that the resulting model can be used in predicting objects with unknown class labels and the model is based on training data analysis which then the results can be used to predict future data trends [4] .

There are previous studies related to graduation predictions using the Classifier method, as in the Prediction of On-Time Student Graduation with the C4.5 Algorithm with Particle Swarm Optimization at XYZ University, obtained a graduation prediction accuracy of 84.72% with graduation predictions and 195 true graduation data. [5], while in the research on Prediction of Student Graduation On Time Using the Decision Tree and Artificial Neural Network Methods, an accuracy of 74.51% was obtained for the decision tree and 79.74% for the artificial neural network. [1].

There are many variants of the classification algorithm, but it is difficult to determine the appropriate algorithm for a particular classification task, because each algorithm has advantages and disadvantages [6]. According to (Raschka, 2015) in research [6] to combine several classification algorithms, there are certain rules in such a way that the strengths of each single classifier algorithm can be combined and have a better generalization performance than a single classification using the Ensemble machine learning method. Some classification algorithms implemented in machine learning, have weaknesses in dealing with class imbalances [7]. Acording to research [8] in research [4] that class imbalance is a situation where there is a significant difference between the number of minority class instances and the number of majority class instances, which causes real world domain problems (real word problems) to often appear in the field of data mining. The existence of an unbalanced class distribution can affect the performance of a classification algorithm, where the classification algorithm works by assuming the class distribution in the dataset is relatively balanced and the cost of misclassification is the same according to research [9] in research [4]. This is according to research [10] in research [4] which may pose a risk of misclassification of the dataset, which results in the performance of a classification algorithm being not optimal. Therefore, it is necessary to apply the ensemble learning method to the combination of classification algorithms in the following research..

In this research, several classification algorithms are used for graduation predictions such as Logistic Regression, Decision Tree Classifier, Gaussian, Random Forest Classifier, Ada Bost Classifier, Support Vector Coefficient, KNeighbors Classifier and MLP Classifier. Cross-validation method is used for testing and confusion matrix. to calculate the accuracy of each of these algorithms.

II. MATERIALS AND METHODS

This research method uses the esemble learning method which aims to see predictions of student graduation using a dataset from Kaggle. The data used is the GPA dataset collected from a university in Indonesia which consists of 1687 records and 5 attributes where this dataset is not balanced. The intended target is whether the student is predicted to graduate on time or not. The method proposed in this research is Ensemble Learning Different Contribution Sampling (DCS) and the algorithm that we will use is Logistic Regression, Decision Tree Classifier, Gaussian, Random Forest Classifier, Ada Bost Classifier, Support Vector Coefficient, KNeighbors Classifier and MLP Classifier. From each classification algorithm used, the test value and accuracy are calculated which are then compared between the algorithms.

The data collection carried out in this study has gone through a cleaning process including filling in empty data, eliminating data duplication, checking data inconsistencies and correcting errors in the data, and the dataset used does not contain missing values or empty data so that it can immediately proceed to the next stage. The attributes used include:IP1 (1^{st} semester Achievement Index) IP2 (2^{nd} semester Achievement Index), IP3 (3^{rd} semester Achievement Index), IP4 (4^{th} semester Achievement Index), and 0 target (did not pass on time), 1 (pass on time). The student graduation dataset has 5 attributes of numeric type, which consist of 4 attribute classes and 1 class that is a label. The snippet of the dataset can be seen in Table 1.

Table 1
Dataset Snippet

No	IP1	IP2	IP3	IP4	Target
1	2.3	1.97	1.8	1.56	0
2	1.81	1.68	1.57	1.86	0
3	3.07	3	2.75	3.21	0
4	2.71	2.33	2.61	1.98	0
5	3.17	3.02	3.28	2.96	0
6	3.16	3.45	3.02	3.06	0
7	2.72	2.5	2.92	3	0
8	2.97	3.27	2.9	2.83	0
9	2.72	2.61	2.64	2.46	0
10	2.78	2.85	3.08	3.35	0
11	2.5	1.56	3.17	3.33	0
12	2.19	2.92	2.54	2.67	0
…	…	…	…	…	…
1685	3.31	3.25	3.44	3.52	1
1686	3.44	3.35	3.5	3.5	1
1687	3.18	3.05	3.05	3.27	1

2. 1 Research Stages

In carrying out this research, the stages in Figure 1. are carried out.

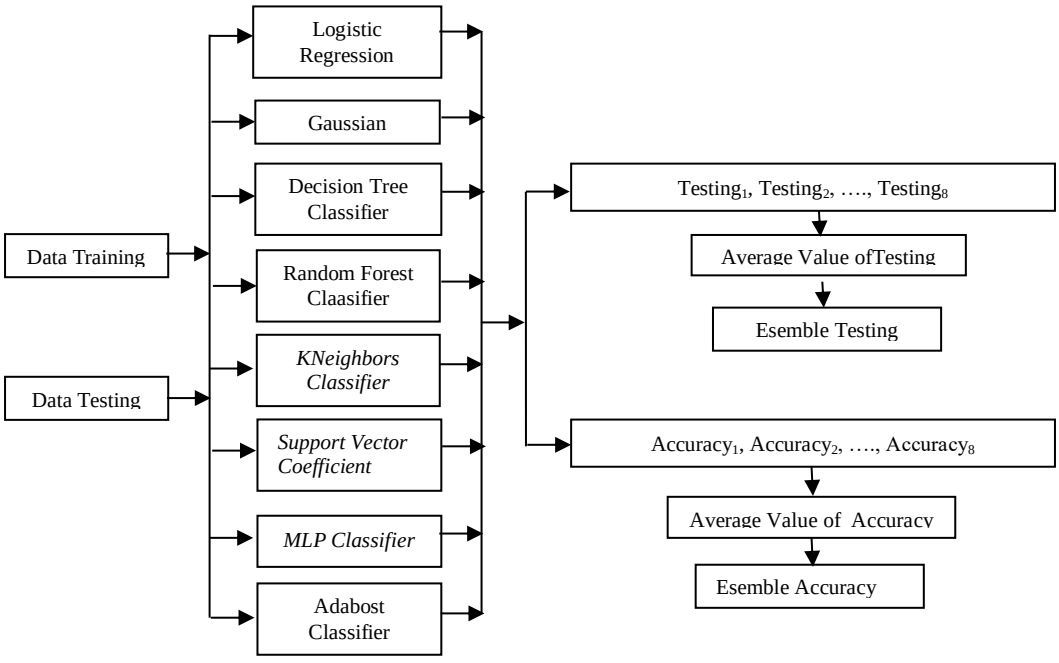

Figure 1. Research Stages

2.2 Classification and Models

At this stage the algorithms that will be used are Logistic Regression, Gaussian, Decision Tree (DT), Random Forest, KNN, Support Vector Machine (SVM), Multi Layer Perceptron (MLP) and Adaptive Boosting (Adaboost)

a) Logistic Regression
Is a non-linear regression, used to explain the relationship between X and Y which is not linear, Y distribution is not normal. The diversity of responses is not constant which cannot be explained by the ordinary linear regression model. [11],
[12] Logistic Function = $Logistic\ function = \frac{1}{1+e^{-x}}$

b) Gaussian
Gaussian-Naïve Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of strong (nave) independence between features [13]. The NGB algorithm is very simple and easy to implement and does not require too much training data [14]

c) Decision Tree
A decision tree is a method that converts very large facts into a decision tree that represents the rules. Useful for exploring data, finding hidden relationships between a number of potential input variables and a target variable [15]

d) Random Forest
The random forest algorithm is divided into two parts, namely the part for making "n" trees (trees) to form a random forest and the part for making predictions from Random Forests. [16]. A random forest contains a collection of classification trees, for example {h(x,Θk), k = 1, …} where {Θk) is a vector that is independently identically distributed and each tree chooses the most class of data (majority vote). The expected result is a collection of classification trees that have a small correlation between trees, so it will reduce the results of random forest prediction errors [13]

e) K-Nearest Neighbor (k-NN)
The k-NN algorithm is a classification algorithm that works based on the calculation of the distance or similarity between data [17]

$$y' = \underset{v}{argmax} \sum_{i=1}^{n} x_i y_i \in D_z I(v = y_i)$$

A rare calculation commonly used is the Euclidean method [18], where the result is stored in the value D, and choose the value $D_z \in D$ which is the k nearest neighbor of z. while v is the number of data entered in class y_i.

$$d(X_{ij}) = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2}$$

f) Support Vector Machine (SVM)
The concept of SVM is to find the best hyperlane that has a function as a separator from two data classes, positive opinion (+1) and negative opinion (-1), such as the following equation [19] :

$$w.x + b = 0$$

w is a vector weight that is w = {w_1,w_2, …, w_n}; n is the number of attributes and b is a scalar called bias.

g) Adaptive Boosting (Adaboost)
AdaBoost Algorithm from Freund and Schapire (1995) in research [4] explained that the first practical amplifier algorithm is and remains one of the most widely used and studied, with applications in various fields. Boosting can be combined with other algorithmic classifiers to improve intuitive classification performance and combine different models with each other. Adaboost and its variants have been successfully applied because of a strong theoretical basis, accurate predictions and simplicity. AdaBosst stages in research [20] include::

i. Input: A research sample set with the label{(x_i,y_i), …, (x_N,Y_N)}, a component learn algorithm, the number of cycles T.
ii. Initialize : Weight of a training sample w1i = 1/N, for all i = 1, …, N
iii. Do for t = 1, …, T
iv. Use the component learn algorithm to train a classification component, ht, on the training weight sample
v. Calculate training error on h_t : $\varepsilon_t = \Sigma^N_i = 1w^t_i$, $y_i \neq h_t(x_i)$
vi. Set weight for *component classifier* $h_t = \alpha_t = ½\text{In}\ (^{1-\varepsilon t}/\varepsilon t)$
vii. Update training sample weight $w^{t+1}_i = w^t_i \frac{\exp\{-\alpha_t y_i h_t(x_i)\}}{C_t}$, $i = 1, \ldots, N$ C_t is a normalization constant
viii. *Output* : f(x) =*sign* ($\Sigma^T_t = {}_1\alpha_t\, h_t(x)$).

h) Algoritma MLP Classifier
Is MultiLayer Perceptron (MLP), is an algorithm that adopts the workings of neural networks in living things that are reliable because the learning process can be carried out in a directed manner. Learning this algorithm is done by updating the back weight (backpropagation) and determining the optimal weight will lead to the right classification results. [21]

2.3 Model Evaluation

i. Cross Validation
Is a statistical method for evaluating and comparing learning algorithms by dividing the data into two segments, one segment is used for learning or training data and the other for validating the model, according to (Refaelzadeh, Tang & Liu, 2009) in[22].

ii. Confusion Matriks
In this study, the Confusion matrix was used to measure Accuracy, which according to research [23] that the Confusion matrix is a predictor of classification problems. The number of correct and incorrect predictions is summarized by calculating the value and achieved for each class, thus providing complete information which includes not only the errors made by the classifier but more importantly the types of errors made.

Table 2. Confusion Matrix

	Class 1 : Positive	Class 2 : Negative
Class 1 : Positive	TP	FN
Class 2 : Negative	FP	TN

Information :
Class 1 : Positive
Class 2 : Negative
True Positive (TP), False Negatave (FN), True Negative (TN), False Positive (FP).
Formula untuk menghitung accuracy sebagai berikut :

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

III. RESULTS AND DISCUSSION

The existence of an ensemble learning algorithm helps to dynamically combine basic machine learning models (Multiple Classifier System) to solve classification problems according to (Garcia, et all) in research [24] and DCS applies a technique where only the basic classification algorithm that obtains the highest level of competence will be selected for the classification learning process. From the results of the classification of each classification algorithm up to the number of folds to 4, it is shown in table 2.

Table 2
Test Results Up to the 4th Fold

Model		0	1	2	3	Average
Logistic Regression:	**0**	0.902844	0.914692	0.919431	0.921615	0.914646
Decision Tree Classifier:	**1**	0.755924	0.85545	0.876777	0.855107	0.835815
Gaussian NB:	**2**	0.604265	0.867299	0.917062	0.931116	0.829935
Random Forest Classifier:	**3**	0.881517	0.905213	0.909953	0.921615	0.904574
AdaBoost Classifier:	**4**	0.883886	0.914692	0.919431	0.916865	0.908719
SVC:	**5**	0.907583	0.917062	0.919431	0.921615	0.916423
KNeighbors Classifier:	**6**	0.872038	0.890995	0.909953	0.91924	0.898056
MLP Classifier	**7**	0.914692	0.919431	0.919431	0.921615	0.918792

The calculation of the 4th fold number for each classification is averaged to obtain the best value, and up to the 4th fold test the best classification algorithm is in the MPL Classifier algorithm with an average value of 0.921615. From the prediction results of student graduation in the form of a confusion matrix using the DCS-LA model by using a combination of several classification algorithms which are then calculated the values of Accuracy (Acc), Precision (Pr), Sensitivity (Se), and Specificity (Sp) up to the 4th fold, in table 3.

Table 3
Confusion Matrix DCS-LA

Cross validation	Accuracy	Precission	Recal	F-Score
Fold 4	0.858	0.06	0.05	0.05

From the results of calculations carried out, combining several classification algorithms using DCS-LA from the 4th fold, the classification accuracy value is 0.858, and from these results, after being compared with the Logistic Regression algorithm, the accuracy is 91.4%, Decision Tree 83.5%, Gaussian 82.9%. , Random Forest 90.4%, AdaBosst 90.8%, SVC 91.6 %, KNN 89.8%, MPL 91.8 %, the final result does not exceed other classification algorithms.

IV. CONCLUSION

Based on the results of research, the following conclusions are obtained. Research has been carried out on student graduation predictions at one of the universities in Indonesia using the Logistic Regression classification algorithm, Decision Tree Classifier, Gaussian, Random Forest Classifier, Ada Bost Classifier, Support Vector Coefficient, KNeighbors Classifier and MLP Classifier and Ensemble Learning DCS-LCA. The best accuracy results are owned by the MLPClassifier algorithm with the ability to predict student graduation on time of 91.87% .

The classification model provided by the DCS-LCA used did not give better results than the basic classifiers that formed it, namely the MLPClassifier algorithm of 91.87%, SVC of 91.64%, Logistic Regression of 91.46%, AdaBost Classifier of 90.87%, Random Forest Classifier of 90.45 %, and KNN of 89.80%.

This research can still be developed in the future to be able to predict the length of the study period, or predict graduation scores and perform hyperparameter tuning for the ensemble learning model that will be used.

REFERENCES

[1] E. P. Rohmawan, "PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU MENGGUNAKAN METODE DECISION TREE DAN ARTIFICIAL NEURAL NETWORK," 2007.
[2] M. B. Cart, J. Matematika, U. Bengkulu, and J. W. R. Supratman, "Analisis ketepatan waktu lulus mahasiswa dengan menggunakan bagging cart," pp. 155–166, 2019.
[3] L. Setiyani, "ANALISIS PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU MENGGUNAKAN METODE DATA MINING NAÏVE BAYES : SYSTEMATIC REVIEW," vol. 13, no. 1, pp. 35–43, 2020.
[4] Y. Pristianto, "PENERAPAN METODE ENSEMBLE UNTUK MENINGKATKAN KINERJA ALGORITME KLASIFIKASI PADA IMBALANCED DATASET," vol. 13, no. 1, pp. 11–16, 2019.
[5] R. Maulida, "Prediksi Kelulusan Mahasiswa Tepat Waktu dengan Algoritma C4 . 5 dengan Particle Swarm Optimization pada Univeristas XYZ," vol. 1, no. 3, pp. 138–144, 2020.
[6] A. Ikhlas and D. Y. Prasetyo, "MESIN PEMBELAJARAN ENSEMBLE UNTUK IDENTIFIKASI VARIETAS PADI Ensemble Machine Learning for Rice Varieties Identification," 2020.
[7] R. S. Wahono, U. Dian, N. Semarang, N. Suryana, and S. Ahmad, "Neural Network Parameter Optimization Based

on Genetic Algorithm for Software Defect Prediction," no. October, 2014.
[8] A. Saifudin, U. Pamulang, R. S. Wahono, U. Dian, and N. Semarang, "Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan Kelas pada Prediksi Cacat Software," no. May 2015, 2019.
[9] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," vol. 40, pp. 3358–3378, 2007.
[10] L. M. Rushi Longadge, Snhlata S.Dongre, "Class Imbalance Problem in Data Mining : Review," vol. 2, no. 1, 2013.
[11] R. Hendayana, "ADOPSI TEKNOLOGI PERTANIAN Application Method of Logistic Regression Analyze the Agricultural Technology Adoption," no. 2, pp. 1–9, 2012.
[12] O. D. Thomas W, Manz, *Research Methods for Cyber Security*. 2017.
[13] L. Breiman, *RANDOM FOREST*. 2001.
[14] E. K. Ampomah, G. Nyame, Z. Qin, P. C. Addo, E. O. Gyamfi, and M. Gyan, "Stock market prediction with gaussian naïve bayes machine learning algorithm," *Inform.*, vol. 45, no. 2, pp. 243–256, 2021.
[15] F. Dwi Meliani Achmad, Budanis, Slamat, "Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree," *J. IPTEK*, vol. 16, no. 1, pp. 18–23, 2012.
[16] J. Han, *Data Mining Concepts and Techniques 3rd Edition - 2012.pdf*. 2012.
[17] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *AICCSA 08 - 6th IEEE/ACS Int. Conf. Comput. Syst. Appl.*, no. December, pp. 108–115, 2008.
[18] N. C. S. N. I. M Anbarasi, E Anupriya, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5370–5376, 2010.
[19] Y. Astuti, U. A. Yogyakarta, and L. D. Farida, "ALGORITMA SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI SIKAP POLITIK," no. August, 2019.
[20] A. M. Listiana, Eka; Muslim, "Penerapan AdaBosst untuk Klasifikasi Support Vector Machine Guna Mengingkatkan Akurasi pada Diagnosa Chronic Kidney Disease," *SNATIF*, vol. 4, no. ISBN 978-602-1180-50-1, 2017.
[21] A. Muliantara and I. M. Widiartha, "PENERAPAN MULTI LAYER PERCEPTRON," pp. 9–15.
[22] I. Mahendro, "PENERAPAN DATA MINING UNTUK PREDIKSI KELULUSAN UKP," pp. 155–159.
[23] A. M. Siregar, S. Faisal, and A. Fauzi, "Klasifikasi untuk Prediksi Cuaca Menggunakan Esemble Learning," vol. 13, no. 2, pp. 138–147, 2020.
[24] D. C. S. L. C. A. Method, "Sistem Pendeteksi Kerusakan Buah Mangga Menggunakan Sensor Gas Dengan Metode DCS - LCA (Mango Damage Detection System Using Gas Sensor With," vol. 3, no. 4, pp. 186–194, 2022.