Classification Algorithm Accuracy Improvement for Student Graduation Prediction Using Ensemble Model

Ace C. Lagman, Lourwel P. Alfonso, Marie Luvett I. Goh, Jay-ar P. Lalata, Juan Paulo H. Magcuyao, and Heintjie N. Vicente

***Abstract*—According to National Center for Education Statistics, almost half of the first-time freshmen full time students who began seeking a bachelor's degree do not graduate. The imbalance between the student enrolment and student graduation can be solved by early predicting and identifying students who are prone of not having graduation on time, so proper remediation and retention policies can be formulated and implemented by institutions. The study focused on the application of the ensemble models in predicting student graduation. Ensemble modeling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications. The study recorded an increase of classification accuracy in predicting student graduation using ensemble models and combining multiple algorithms.**

***Index Terms*—Machine learning, ensemble model, student graduation, predictive analytics.**

I. Introduction

One main research focus of educational data mining is student graduation [1]. The student graduation rate is the percentage of a school's first-time, first-year undergraduate students who complete their program successfully. Most students' first year freshmen enrolled in tertiary level failed to graduate. According to National Center for Education Statistics, almost half of the first-time freshmen full time students who began seeking a bachelor's degree do not graduate. Addressing this problem is crucial as colleges and universities consisting of high leaver rates go through loss of fees and potential alumni contributors [2]. Most researchers already developed multiple decision-based models for modeling drop outs and retentions of students however only few considered the power of ensemble models in prediction. The study aimed to determine the accuracy of ensemble models and algorithm combination in student graduation prediction.

II. Literature Review

Data Mining is application of a specific algorithm in order to extract patterns from data and transform the information into a comprehensible structure for further use [3]. KDD has become a very important process to convert this large wealth of data in to business intelligence, as manual extraction of patterns has become seemingly impossible in the past few decades [4].

Data Mining is a step inside the KDD process, which deals with identifying patterns in data in a large dataset [5]. It is only the application of a specific algorithm based on the overall goal of the KDD process, which is to extract hidden patterns or develop predictive models using machine-learning techniques [6].

Educational data mining is one of the main applications of machine learning where it analyzes students' behaviors, and performance so proper interventions can be provided [7]. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data [8].

The Ensemble classification is based on the philosophy that a group of experts gives more accurate decisions as compared to a single expert. Literature reveals that prediction from composite tests give more better result to a single prediction. This section describes the ensemble techniques used in this paper [9].

Boosting boosts the performance of the weak classifier to a strong level. It generates sequential learning classifiers using resampling (reweighting) the data instances. Initially equal uniform weights are assigned to all the instances. During each learning phase a new hypothesis is learned, and the instances are reweighted such that correctly classified instance having lower weight and system can concentrates on instances that have not been correctly classified during this phase having higher weights. It selects the wrongly classified instance, so that they can be classified correctly during the next learning step. This process continuous tills the last classifier construction. Finally, the results of all the classifiers are combined using majority voting to find the final prediction. AdaBoost is a more general version of the Boosting algorithm [10].

III. Methodology

In order to solve research objectives, the researcher used Knowledge Discovery in Databases to extract hidden patterns form the data.

A. *Knowledge Discovery in Databases*

The researchers used the modified steps of Knowledge Discovery in Databases indicated in the Fig. 1 below.

Manuscript received March 5, 2020; revised June 1, 2020.

Ace C. Lagman, Lourwel P. Alfonso, Marie Luvett I. Goh, Jay-ar P. Lalata, Juan Paulo H. Magcuyao, and Heintjie N. Vicente are with the FEU Institute of Technology, P. Paredes St. Sampaloc, Manila, Philippines (e-mail: lpalfonso@feutech.edu.ph, migoh@feutech.edu.ph, aclagman@feutech.edu.ph, jplalata@feutech.edu.ph, jhmagcuyao@feutech.edu.ph, hnvicente@feutech.edu.ph).

doi: 10.18178/ijiet.2020.10.10.1449

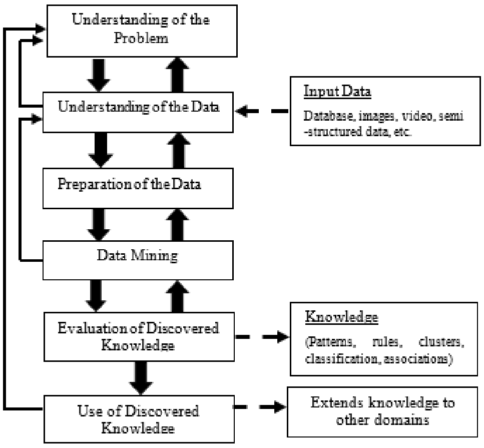

Understanding of the Problem
Understanding of the Data
Input Data
Database, images, video, semi -structured data, etc.
Preparation of the Data
Data Mining
Evaluation of Discovered Knowledge
Knowledge
(Patterns, rules, clusters, classification, associations)
Use of Discovered Knowledge
Extends knowledge to other domains

Fig. 1. The six-step KDD model.

The modified version of the KDD consists of six steps. The six phases include understanding the problem and data, data preparation, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge.

B. Problem and Data Understanding

This section entails the researcher to understand the problem and what possible solutions that can be proposed. This section determines the rational of the research and potential of the data to achieve researchers' goals.

C. Data Preparation

This section provides after examining the data, what necessary data preprocessing techniques that are necessary to improve the accuracy of the algorithm. The researcher used discretization and imputation techniques to normalize the values that is easier to extract patterns from the students' data

D. Bootstrap Algorithm

There is a method to increase the accuracy of k learned models; this method is called ensemble methods or methods that use a combination of models. Bagging or Bootstrap and stacking are the most used ensemble methods developed to increase the accuracy of the learned model [6].

Bootstrap is a method of increasing accuracy; the new test sets of data was evaluated by the learning scheme of the logistic regression. The bootstrap algorithm created an ensemble of models for a learning scheme where each model gives an equally-weighted prediction.

Algorithm

```
 Input:
$D$, a set of $d$ training tuples;
$k$, the number of models in the ensemble;
a learning scheme

 Output: A composite model, M_.
 Method:
(1) for $i$ = 1 to $k$ do // create $k$ models:
(2) create bootstrap sample, $Di$, by sampling $D$ with replacement;
(3) use $Di$ to derive a model, $Mi$;
(4) endfor
To use the composite model on a tuple, $\boldsymbol{X}$:
(1) if prediction then
(2) let each of the $k$ models classify $\boldsymbol{X}$ and return equally weighted prediction
```

1) Learned model prediction combination

There is a method to increase the accuracy of k learned models; this method is called ensemble methods or methods that use a combination of models. Bagging or Bootstrap and stacking are the most used ensemble methods developed to increase the accuracy of the learned model [6].

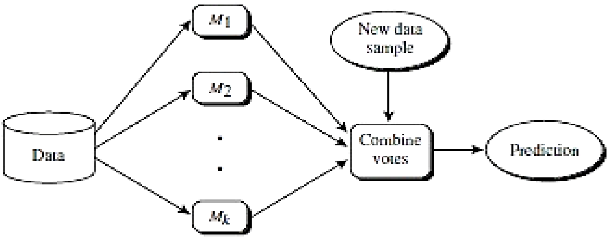


Fig. 2. Learned model prediction combination.

Fig. 2 shows to increase model accuracy: gagging and boosting were used. Each generate a set of classification or prediction models, M1, M2, : : : , Mk which refers to sets of classifiers. Voting strategies are used to combine the predictions for a given unknown tuple.

They are examples of ensemble methods, or methods that use a combination of models. Each combines a series of k learned models (classifiers or predictors), M1, M2, : : : , Mk, with the aim of creating an improved composite model, M_. Both bagging and boosting can be used for classification as well as prediction

2) Performance measure

TABLE I: CONFUSION MATRIX TABLE

Predicted	**Actual Graduation**		
		Yes	**No**
	Yes	True Positive	False Positive
	No	False Negative	True Negative

The confusion matrix is a useful tool for analyzing how well your classifier can recognize tuples of different classes [7]. A confusion matrix for two classes given m classes, a confusion matrix is a table of at least size m by m. An entry, CMi, j in the first m rows and m columns indicates the number of tuples of class i that were labelled by the classifier as class j as seen in Table I.

The confusion matrix table illustrates a tabular display that evaluates the forecasting precision of a predictive model.

The main objective of a predictive model is to maximize the correctly classified instances. For binary classification scenarios, the misclassification rate gives the overall model performance with respect to the exact number of categorizations in the training data.

To determine the accuracy level of the classification table of the algorithms the formula was used

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

where true positive (TP) refers to as number of actual outcomes of graduation yes accurately classified as predicted graduation yes and true negative (TN) refers to as number of actual outcomes of graduation 'no' accurately classified as predicted graduation 'no'.

IV. Results and Discussion

A. Accuracy of the Algorithm

The summary of accuracy rate from the different algorithm as reveals in Table I reveals that logistic regression algorithm has the best accuracy rate in predicting student graduation with 87.4 accuracy rate. Thereof, the values in the coefficients table were used to derive data models or equations in predicting student graduation in new test sets of data.

To cross validate the results, knowledge flow of Weka was used to determine the best algorithm that can predict the accuracy of student graduation.

TABLE II: Logistic Regression Values in the Equation

Logistic Regression		
Observed Value		Percentage Corrected
Graduate	Not Graduated	95.85
	Graduated	49.74
Average Percentage		87.4
Neural Network		
Observed Value		Percentage Corrected
Graduate	Not Graduated	99.89
	Graduated	5.66
Average Percentage		84.27
Decision Tree		
Observed Value		Percentage Corrected
Graduate	Not Graduated	97.73
	Graduated	31.61
Average Percentage		86.77
Naive Bayes		
Observed Value		Percentage Corrected
Graduate	Not Graduated	90.21
	Graduated	39.86
Average Percentage		85.23

The data sets were tested simultaneously with the different algorithms which include decision tree, Naïve Bayes, Logistic Regression, Multilayer Perceptron and Neural Network. Cross validation technique was used. The cross-validation technique divides the data set into ten equal parts where each part can be calculated by total number of instance over the number of fold validations which resulted to 116.4 data instances. The 9 out of 10 sets will be used as the training set – this set will be used to train the classifier and the remaining set will be used to estimate the error rate of the trained classifier. The text viewer generates the prediction accuracy results from different algorithms.

The researchers used the Knowledge Flow "data-flow" inspired interface of Weka. At present, all of Weka's classifiers and filters are available in the Knowledge Flow along with some extra tools. The flow presents results of the multiple algorithms which include Naïve Bayes, Neural Network, Decision Tree, J48 and Logistic Regressions in one output using classifier performance evaluation. This function evaluates the performance of incrementally trained classifiers. The table below indicates the result or the performance of the algorithm.

TABLE III: Summary of Results in Knowledge Flow

Algorithm	Accuracy	Error	Precision	Recall
Naive Bayes	85.30	14.69	0.85	0.85
Decision Tree	83.44	16.56	0.82	0.83
Neural Network	83.33	15.67	0.82	0.83
Logistic	87.45	12.55	0.86	0.87

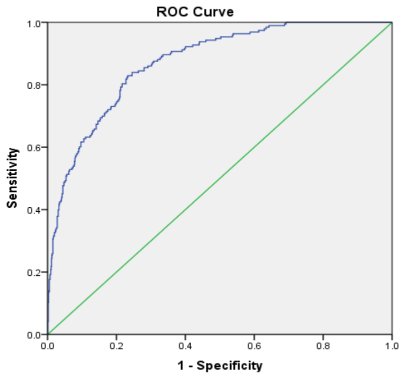
Table III reveals that Logistic Regression has predicted more in student graduation compared with other algorithms using Knowledge Flow in Weka.

B. Data Model of Logistic Regression in Predicting Test Sets

The highest accuracy in the lists of data models in the logistic regression was used its predictive accuracy in the training sets of data. The derived equation was shown below.

$$prob(graduated) = \frac{1}{1+e^{-(-5.716C+.888g*X_1-.991*X_2+.307Ve*X_3+.250Ab*X_4+.289A*X_5+.430F*X_6+.567*X_7+.423W*X_8)}} \qquad (2)$$

To determine and evaluate the goodness-of-fit of a logistic regression model it will be tested based on the simultaneous measure of sensitivity (True positive) and specificity (True negative) to possible cut of points through receiver operating characteristic curve.



Diagonal segments are produced by ties.

Fig. 3. ROC Curve of logistic regression model.

Results in the Table IV reveals that output which shows the ROC curve. The area under the curve is .872 with 95% confidence interval (.846, .897). Also, the area under the curve is significantly different from 0.5 since p-value is .000 meaning that the logistic regression classifies the group significantly better than by chance.

Since the model classifies group significantly better by chance, the generated data model of the logistic regression (Fig. 3) was then tested to new testing sets of data. The table below illustrates the prediction of the model in the test sets.

TABLE IV: Test Results Area under the Curve

Area	Std. Error	Asymptotic Sig.	Asymptotoic 95% Confidence Level	
			Lower Bound	Upper Bound
0.872	0.13	0	0.848	0.897

Table V reveals that the data model of logistic regression recorded an accuracy rate of 86.04 in the new test sets of data. The datasets was tested to multiple processes to increase the accuracy result of the algorithm which include inclusion of ensemble model. To improve the accuracy rate of the correctly classified graduated status, bootstrap technique was used to improve the learned model of the logistic regression.

TABLE V: CLASSIFICATION TABLE OF LOGISTIC REGRESSION IN TESTING DATA (N=129)

Observed Value		Predicted Not Graduated	 Graduated	Percentage Corrected
Graduate	Not Graduated	98	2	97.00
	Graduated	16	13	44.82
Average Percentage				**86.04**

C. Improving Data Model of Logistic Regression by Applying

1) Bootstrap algorithm

Bootstrap aggregating, often abbreviated as bagging, boosting and stacking using majority of votes popularly known as ensemble model were used to increase the accuracy of logistic regression in the test sets. The logistic model equation derived from the training set was tested in the new sets of data.

TABLE VI: BOOTSTRAP IN THE EQUATION

Attributes	B	Bootstrap^{a} Bias	 Std. Error	 Sig. (2-tailed)
Gender	0.888	0.002	0.031	0.002
Scholarship	-0.991	-0.001	0.013	0.002
Verbal_quivalent	0.307	-0.001	0.011	0.002
Abstract_quivannt	0.25	0	0.009	0.002
Algebra	0.289	0	0.011	0.002
IT_funda	0.43	0	0.019	0.002
Programming	0.567	0	0.007	0.002
Vedu	0.423	-0.001	0.024	0.002
Constant	-5.716	0.002	0.108	0.002

Bootstrap technique was used by the researcher to increase the accuracy rate of the logistic regression in the test sets. After running a bootstrap in the logistic regression, the lower value coefficients of the predictors were used to derive an equation in evaluating the testing sets of data. The coefficients were used to derive a new learned model.

2) Boosted logistic regression

$$prob(graduated) = \frac{1}{1+e^{-(-5.934+.824*X_1-1.012*X_2+.288*X_3+.234*X_4+.267*X_5+.397*X_6+.553*X_7+.381*X_8)}} \quad (3)$$

To determine the accuracy of the new derived learned model generated by logistic regression, the model was tested using new testing sets of data. The classification table results were shown as follows.

TABLE VII: BOOSTED LOGISTIC MODEL RESULT

Profile		Predicted Graduate	 Not Graduated	Total
Graduation Status	Graduate	16	13	55.17%
	Non-Graduate	4	96	93.20%
Total				86.82171

The boosted logistic model has classified 3 out of 16 misclassified instances from the initial logisitc regression data model. From 44. 82 accuracy rate of the graduated status it becomes 55.17. Table VI reveals that the after using the boostraping technique under logistic regresion model the accuracy rate of testing sets has increased to 86.82 %. The boosted logistic model has also a parallel testing using weka that accumulated also a performance of 86.82 accuracy rate as seen in Table VII.

D. Improving Accuracy by Combining Data Model Predictions

To improve accuracy rate of the test sets, prediction of data models of Naïve Bayes, Logistic Regression, Decision Tree and Neural Network were combined in predicting student graduation using majority of votes.

To improve accuracy rate of test sets instances of logistic regression, combinations of predictions of set of classifiers were tested. The experiments were done using WEKA using majority of votes. The results of the experiment were shown below.

Combination of predictions of data models reveals that all combinations have increased the accuracy rate of logistic regression from 86.04 to 87.6, 86.82 and 86.62 respectively. Noticeably, the most cited improvement was recorded to the combinations of logistic regression and naïve bayes with 87.6 accuracy rate. The data models combination has classified 17 out of 29 instances of student who graduated. The accuracy rate of the student graduation using models combination boosted to 87.60%

E. Improving Data Model of Logistic Regression by Combining Rule Sets

To improve accuracy rate of the correctly classified of the graduated status the 16 instances (58.62) underwent to three rules sets generated by the decision tree algorithm.

The derive rule sets from the decision tree algorithm in predicting student graduation were shown as follows.

TABLE VIII: RULE SET OF DECISION TREE FOR GRADUATES

Rule	IT Fundamentals	Scholarship	Gender
1	>2.50 and <=3	1	
2	>3	1	
3	>3	2	2

TABLE IX: RULE SET OF DECISION TREE FOR GRADUATES RESULTS

No.	Rule1	Rule2	Rule3
1	False	False	False
2	False	True	False
3	False	False	False
4	False	False	False
5	False	False	False
6	False	False	False
7	False	False	False
8	False	False	False
9	True	False	False
10	False	False	False
11	False	False	False
12	False	False	True
13	False	False	False
14	False	False	False
15	False	False	False
16	False	False	False

Table IX presents that there were three instances were correctly classified by the rules sets generated by the decision tree model, hence it contributes in the increase of the logistic regression.

The rule sets generated from the decision tree algorithm has classified 3 out of 16 misclassified instances from the logisitc regression data model. From 44. 82 accuracy rate of the graduated status it becomes 55.17 after combining the prediciton of the decision tree rule sets. Table X reveals that the after combining the prediction of data model of logistic regresion and rule set of decision tree, the accuracy rate of testing sets has increased to 88.3

TABLE X: LOGISTIC REGRESSION (EQUATION) + DECISION TREE (RULE SET) ACCURACY RATE

Observed Value		**Predicted**		**Percentage Corrected**
		Not Graduated	Graduated	
Graduate	Not Graduated	98	2	97.00
	Graduated	13	16	55.17
Average Percentage				**88.3**

V. CONCLUSION

Adding Boosting technique in logistic regression made a significant increase of accuracy in predicting student graduation. Result reveals that after using the boostraping technique under logistic regression model, the accuracy rate of testing sets has increased to 86.82 %. The boosted logistic model has also a parallel testing using weka that accumulated also a performance of 86.82 accuracy rate. Model combination also is very efficient in increasing the accuracy of the classifier. Multiple test experiment should be considered by the researcher to determine the right combination test of an algorithm to give a more accurate prediction.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this work.

REFERENCES

[1] S. Wanjau and G. Muketha, *Improving Student Enrollment Prediction Using Ensemble Classifiers*, 2018.

[2] W. A. Olugbenga and C. Thomas, *Predicting Student Academic Performance Using Multi-model Heterogeneous Ensemble Approach*, 2018.

[3] I. Natthakan and B. Tossapon, *Improved Student Dropout Prediction in Thai University Using Ensemble of Mixed-Type Data Clusterings*, 2015.

[4] A. Ahmed, "Data mining: A prediction for student's performance using classification method," *World Journal of Computer Application and Technology*, vol. 2, no. 2, pp. 43-47, 2014.

[5] C. Clifton. Encyclopædia britannica: Definition of data mining. [Online]. Available: http://www.britannica.com/EBchecked/topic/1056150/data-mining

[6] J. Han, M. Kamber, and J. Pie, *Data Mining Concepts and Techniques*, 2006.

[7] A. Seidman, *College Student Retention: Formula for Student Success*, Westport, CT., 2005.

[8] U. Fayyad. G. Piatetsky-Shapiro, and P. Smyth. (1996). From data mining to knowledge discovery in databases. [Online]. Available: https://en.wikipedia.org/wiki/Gregory_Piatetsky-Shapiro

[9] M. Pandey and S. Taruna, "Comparative study of ensemble methods for students' performance modeling," *International Journal of Computer Applications*, vol. 103, no. 8, 2014.

[10] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proc. ICML'96*, 1996, pp. 148–156.

Ace C. Lagman finished his doctorate degree in AMA University as Summa Cum Laude. He already finished his post-doctoral degree in information technology in Royal Institute in Singapore as fellow in Royal Institute in Information Technology (FRIIT). Currently he is taking up another doctorate degree phd in computer science at Technological Institute of the Philippines. Dr. Lagman was awarded as Outstanding Alumnus of Baliuag University in Science and Technology and he was also recognized as this year's one the Dangal ng Baliuag awardees. Both awards mentioned are the highest awards given by Baliuag University and Municipality of Baliuag. He also bagged multiple international achievements in terms of research and international certifications. He already published numerous research papers in machine learning algorithms and other computing research areas. He consistently serves as a technical committee member to various local and international research conferences.

Marie Luvett I. Goh is a full-time faculty member of Information Technology Department at FEU Institute of Technology-Manila. She earned her doctor of technology degree at Technological University of the Philippines. She took her degrees in master in information technology and bachelor of science in computer engineering at Adamson University. Her research interests are embedded systems, sensors technology and iot applications.

Jay-ar P. Lalata is a full-time faculty member of Information Technology Department at FEU Institute of Technology. He is a DOST-scholar and a cum-laude graduate of B.S. computer engineering from Adamson University where he also took his master in information technology. He is a candidate for the degree, doctor in information technology at Technological Institute of the Philippines - Quezon City. His areas of interest include data mining, natural language processing and embedded systems.