Universal Journal of Educational Research 8(9): 4036-4047, 2020 http://www.hrpub.org
DOI: 10.13189/ujer.2020.080929

Survival Analysis Approach for Early Prediction of Student Dropout Using Enrollment Student Data and Ensemble Models

Frederick F. Patacsil

College of Computing, Pangasinan State University-Urdaneta City Campus, Philippines

Received May 20, 2020; Revised July 2, 2020; Accepted July 20, 2020

Cite This Paper in the following Citation Styles

(a): *[1] Frederick F. Patacsil , "Survival Analysis Approach for Early Prediction of Student Dropout Using Enrollment Student Data and Ensemble Models," Universal Journal of Educational Research, Vol. 8, No. 9, pp. 4036 - 4047, 2020. DOI: 10.13189/ujer.2020.080929.*

(b): Frederick F. Patacsil (2020). Survival Analysis Approach for Early Prediction of Student Dropout Using Enrollment Student Data and Ensemble Models. Universal Journal of Educational Research, 8(9), 4036 - 4047. DOI: 10.13189/ujer.2020.080929.

Abstract The Universal Access to Quality Tertiary Education Act is a law in the Philippines that provides college students with free tuition and other fees in Philippine state universities and local universities. People's tax is used to finance this law and the government should ensure that student retention or persistence is attained throughout the duration of their stay. To effectively decrease student dropout, it is necessary to understand which students are at risk of dropping out. In addition, this study proposed a model that detects and predicts student success in tertiary education through the right selection of the suitable program utilizing the enrollment data that may have a significant on the study outcome of the students. This study experimented single classifier and added ensemble approach classifiers to propose a predictive model to detect early dropout of first-year college students. The study utilized tree algorithms and then applied the ensemble algorithm to identify student attributes that distinguish potential dropouts from college. The result reveals a very interesting prediction that if their average grade is less than 85, there is a high tendency of dropping any program they enrolled in. Evaluation results in the final stage of the model construction process reveal that applying bagging ensemble into j-48 tree attained the highest accuracy as matched with other tree algorithms, however, forest tree algorithm achieved the highest value in terms of dropout precision and graduated recall. The result also shows that applying ensemble approaches have a marginal increase in classification performance.

Keywords Dropout, Survival Analysis, Ensemble Model, Prediction

1. Introduction

The Universal Access to Quality Tertiary Education Act, which both Philippine Senate and Congress ratified, forwarded to the Office of the Philippine President was signed by President Rodrigo Roa Duterte covers a total of 190 colleges and universities which comprise 112 state universities and colleges (SUCs), and 78 local universities and colleges (LUCs) nationwide gives full free tuition and other miscellaneous fees to all enrolled students. People tax is used to finance this law and the Philippine government should ensure that student retention or persistence is attained throughout the duration of their stay. This will assure that the money invested will not be wasted.

With this scenario, the Philippine higher education generation for the next years will reverse the current situation from 80 percent of college students enrolled in private schools and 20 percent in state universities and

colleges (SUCs) to 20 percent, private colleges, and 80 percent SUCs (Macha, Mackie, & Magaziner, 2018). Free tuition also translates the increasing enrollment rates among students in the SUC and the drop backdrop is the government is spending large amounts of money per student enrolled. However, the drop-out rates revealed an alarming 83.7 percent, meaning the country produces 2.13 million college drop-out annually. Therefore, college student dropout is a major concern in the Philippine education system and first-year student dropout is of particular importance, as the Philippine state set aside more than 16 billion for higher education for full-time first-year students seeking baccalaureate degrees who do not return for a second year. Dropping out at State University and Colleges (SUCs) is a serious problem that may result in kick out or force students to leave from the university thus, it may deny the individual fundamental human right of students to their education and the financial waste particularly of public funds is higher for those students who eventually drop out without completing their degree.

With this current state, student retention and graduation have become very significant than ever to SUCs in terms of accountability and recent national initiatives have focused higher education feel increasingly forced to outline and implement interventions/strategies to increase student retention and student success. Whatever the reasons behind the drop out phenomenon at SUCs, one thing is certain. If this problem is not addressed, it will continue to cost the students, parents, and the larger public, especially our taxpayers, and to the Philippines government, it will mean a waste of scarce resources with very little to benefit to show. Increasing the student graduation rate and decreasing the drop out rates is a long term goal of state universities and colleges and the Commission on Higher Education (CHED). From the point of view of the parents and students, timely and successful graduation is vital as these two factors would strongly affect their employability rate and their intention to help their family financially.

Many of the students studying at the state university face several difficulties during the first year and thus the performance of the first year has been recognized as a significant predictor of timely graduation rate. In terms of keeping the students in the university, educators and researchers extensively studied the factors of retention rate and suggest that an early identification of the students at high risk of failing will enable a timely intervention with the necessary measures/intervention by the educators would increase the graduation rate. According to Mallincrodt and Sedlacek (1987) in terms of keeping the students in the university, “the retention rate is a factor that should have been studied extensively”.

One reason for high drop out rates as reported in some researches were poor career choices and lack of personal interest (Kotsiantis & Pintelas, 2005). Yet, few researchers have conducted to investigate and analyze the success of career paths used by the students in the Philippines, especially the factors that affect the career or program choice of Filipino students. With this scenario, parents and students have limited information on how to help them identify their proper career options and course choice they have to pursue in the future.

Pangasinan State University, first-year students have significantly increased during recent years, thousands of students are admitted to study at universities every year, but after the first year of their study, some of them decided not to continue to go in the school. Thus, conducting studies on the monitoring and supporting the student at risk is a topic that needs attention at many educational institutions in the Philippines. CHED and SUCs would like to have 100% students' success to finish their studies, however, it is hard to recognize these students in their early stages. It is important to explore effective approaches for predicting a student drop out as well as identifying the factors affecting it with sufficiently high accuracy. As an example of the massive dropout of first-year students can be seen in Urdaneta City Campus of Pangasinan State University, for the academic year 2012 – 2016, the dropout and falling rate of freshmen is more than 50% in the first year and it increases to more or less 10% after passing two years of study which shows the importance of modeling student drop out early during their study.

Several researches indicate that one of the important factors of student dropout rate is the initial program/course studied at university as well as the secondary school academic records. Indeed, the dropout rate is higher among students in engineering programs, and among students with relatively low performing levels during their high school (Di Pietro & Cutillo, 2008; Tumapon, 2017; Min et al, 2011).

The study will propose a student dropout model using data gathered from the PSU-Urdaneta registrar office. This study has explored and to understand the key determinants of drop out, to accurately identify students likely to give up their schooling and to recommend policy interventions to eliminate or lessen student attrition.

This study has experimented and analyzed the enrollment data that may have an impact on the study outcome of the students and explore the performance accuracy and efficiency of a single classifier and try to improve classification performance by applying the ensemble approach to propose a predictive model to detect early dropout of first-year college students. A set of rules was obtained from the experimented predictive model to predict/identify a suitable program for a certain student based on their enrollment data set. Furthermore, this study was proposed to experiment, investigate, and compare the use of different tree classifiers and combining ensemble

model techniques to improve the results of different tree classifiers. The purpose of this paper is to analyze the causes of the first-year students' dropout rates in the institutions using the real data of Pangasinan State University Urdaneta Campus.

2. Related Literature

2.1. Student High School Profiles

The determinants of the success of college students and their academic performance were analyzed taking into account the student's pre-collegiate endowment of knowledge and various factors associated with the high school profile.

2.2. Predicting Factors

High school General Point Average (GPA), admissions test scores, gender, and race/ethnicity are the factors in which the researcher has consistently found to be significant predictors of retention and/or dropout.2.3 High School Grade Point Average.

The variable GPA used in this study was the weighted grade-point average of all the subjects of a student. Many of the previous studies have demonstrated that GPA is a reliably better predictor of college performance than other variables (Schmitt, Keeney, Oswald, Pleskac, Billington, Sinha & Zorzie, 2009; Klopfenstein & Thomas, 2009.; Estrera *et al.*; Shovon, Islam & Haque, 2012.; Al-Barrak & Al-Razgan, 2016; Angeline; Quadri & Kalyankar, 2010; Osmanbegovic & M. Sulji, 2012; Ham et al, 2006; Tair & El-Halees; 2012; Huang & Fang, 2013; Mayilvaganan & Kalpanadevi, 2014; Al-Twijri & Noaman, 2015; Christian & Ayub, 2014; Asif, Merceron, & Pathan, 2014 ; Jishan et al, 2015). The main idea of previous researchers in using GPA as predicting attributes is the tangible value for future educational and student career mobility. It can also be considered as an indication of realizing academic potential (Mat et. al., 2015). Further, Grade Point Average is an indicator used to evaluate the performance and success of the students during their high school years.

2.3. Grade in Mathematics, Science, English and TLE

Some of the predictor attributes that were considered in this study were high school Grade Point Average (GPA) and average grades in Science, Mathematics, English and Technology and Livelihood Education (TLE). Other researchers utilized the high school final grades in English, Mathematics, and major subjects as predictor attributes Hayward et. al. (2014). Beaulac & Rosenthal (2018) observed that grades in Mathematics, Biology, and Chemistry are consistently among the most important variables for predicting if a student will complete their program. Studies varied in identifying factors that affect student retention the most of their freshman year. Zhang (2004), Veenstra (2008) found out that students GPA in high school and grades in Mathematics, science subjects like Chemistry, and Physics, were all strong predictors for Engineering student retention (Cengiz & Uka, 2014; Al-Radaideh *et al*., 2011). The study of Aulck, Velagapudi, Blumenstock, & West (2010), Mesarić & Šebalj (2016) revealed that GPAs in Chemistry, Mathematics, and English were among the most influential individual predictors of student retention.

2.4. Gender

Kovacic, 2010; Aulck et al (2016) conducted a research utilizing social and demographic variables that may influence determination or drop-out pre-identifying successful and unsuccessful students. Cengiz & Uka (2014) found out that the students' high school GPA and gender were the most important/influential predicting attributes based on their research utilizing data mining techniques.

2.5. Machine Learning Building Predictive Model

There is a substantial number of classification algorithm/ learning machine available applying different approaches to determine student performance which was reported by many researchers. One of the most popular algorithms is decision tree which was experimented to generate a predictive model that will predict the performance of students. Several researchers (Kabra & Bichkar, 2011; Kovacic, 2010; Al-Barrak & Al-Razgan, 2016; Kabakchieva, 2012) utilized decision trees in educational data mining to predict the performance of students using their past performance data. The model predicts to identify the students who are likely to fail/drop-out/unsuccessful and allow the teacher to provide appropriate interventions.

Kabakchieva (2012) also applied different decision tree algorithms for predicting the academic success of students and found that the algorithm obtained a high accuracy prediction rate.

However, some studies utilized other different classification algorithms/ learning machine to create their predictive model.

Smooth Support Vector Machine (SSVM) classification algorithm and kernel k-means clustering was utilized by Sembiring, Zarlis, Hartama, Ramliana, & Wani (2011), Shovon & Haque (2012), Strecht, Cruz, Soares & Mendes-Moreira (2015) to determine/predict the success of students and produce a model to determine the predictors of the student performance.

Yadav and Pal (2012) applied and experimented decision tree algorithms to classify engineering students to

predict their performance in the final exam. The dataset comprised 90 student records with 16 attributes. Among the decision tree algorithms, C4.5 produced the best accuracy standing at 67.78%. (Kovačić, 2012; Simeunović & Preradović, 2014; Cheewaprakobkit, 2015; Shah, 2012; El Zeweidy, Osman & Elhennawy, 2013) conducted a study in predicting student success using the different classifying learning machine. Their study found out that decision tree obtained the highest accuracy in terms of correct classifications among growing method of classifying learning machines. Decision tree was utilized in this study because it is so simple to understand the results and easy to make good interpretations. In addition, it is easier to be understood by a reader of this study. The majority of the previous studies have used this algorithm because of its simplicity and comprehensibility to uncover small or large data structure and predict the value (Zhang *et al*., 2004; Cengiz & Uka, 2014; Shmueli & Koppius, 2011). Further, based on the previous studies, the decision tree is one of the most commonly used techniques in predicting student's performance (Pandey & Sharma, 2013). However, the main disadvantage of decision trees is that they tend to overfit, but there are ensemble methods to counteract this weakness.

3. Materials and Methods

The educational data mining (EDM) is defined as: "Educational Data Mining is a new field in research that concerned with developing methods for exploring the unique and increasingly large-scale data derived from educational settings, and utilizing those methods to better understand students, and the settings which they learn" (International Educational Data Mining Society, 2019).

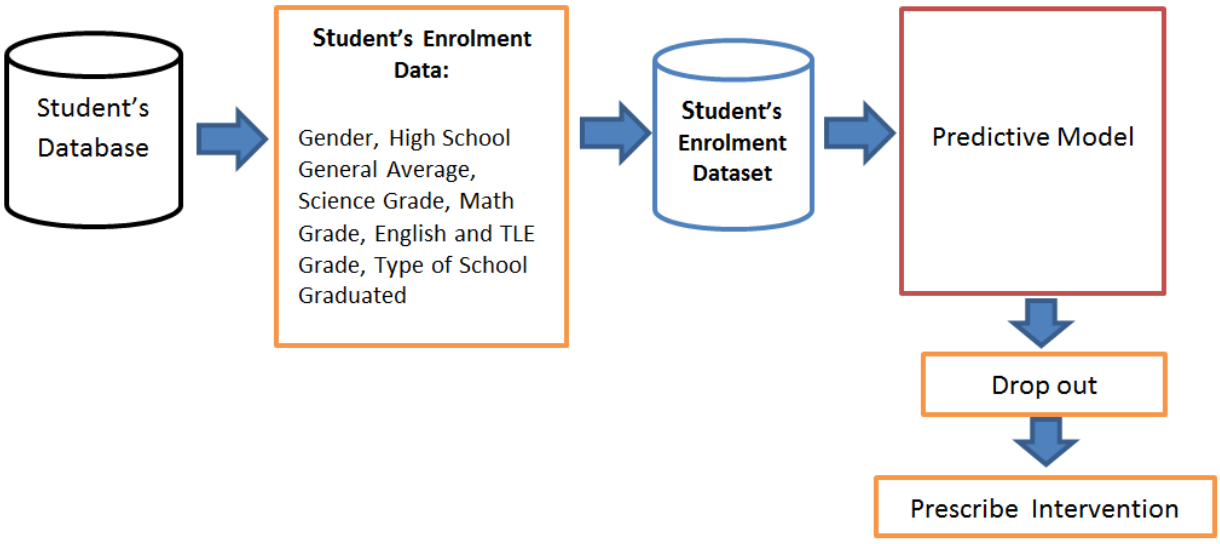

Figure 1. Conceptual Model

3.1. Predictive Analytics

Predictive analytics aimed at predicting future events/outcomes and behaviors present in previously unknown data, using a model built from historical data and analytic techniques (Nyce & Cpcu, 2007; Shmueli & Koppius, 2011). The method used previously collected data, a machine learning algorithm finds the relations between different properties of the data. The result is a predictive model that will be able to predict future outcomes based on the properties of the collected data. The data were collected through the enrollment form filled by the student at the time of admission. The student fills-up attributes in the enrollment form such as their demographic data such as the gender and the type of school graduated (public and private), past academic performance such as their grade in Mathematics, English, Science, TLE, and General Point Average (GPA). From this information, the attributes that possibly influence their results are selected as shown in Table 1.

Most of the attributes reveal the historical performance of the students. The reasons behind concentrating on the past performance data are 1. Data is easily available in the Registrars Office of the campus. 2. If a student has performed well in high school, it is most likely that he will perform better during their college years as well or the other way around.

3.2. Data Selection and Preprocessing Data

The data were composed of 2401 admission forms/records of 956 graduated and 1445 dropout students of Pangasinan State University - Urdaneta City Campus was collected who finish their schooling from the year 2012-16 as shown in table 1.

Table 1. List predicting variables

Variable	Description	Type	Values
Course	Program the students enrolled	Character	ABEL - AB English Archi -BS Architecture BEED - BS Elementary Education CE - BS Civil Engineering COE - BS Computer Engineering EE - BS Electrical Engineering IT - BS Information Technology Mathematics - BS Mathematics ME - BS Mechanical Engineering BSED - BS Secondary Education
Gender	Students Gender (Male or / Female)	Nominal	1 - Male 2 - Female
Grade in Science	Average Grade in Science	Numeric	0 - 100
Grade in Mathematics	Average Grade in Mathematics	Numeric	0 - 100
Grade in English	Average Grade in English	Numeric	0 - 100
Grade in TLE	Average Grade in TLE	Numeric	0 - 100
High School Grade Point Average	high-school grade point average (HSGPA)	Numeric	0 -100
Type School Graduated	Type of School Graduated	Character	1 - Public 2 - Private

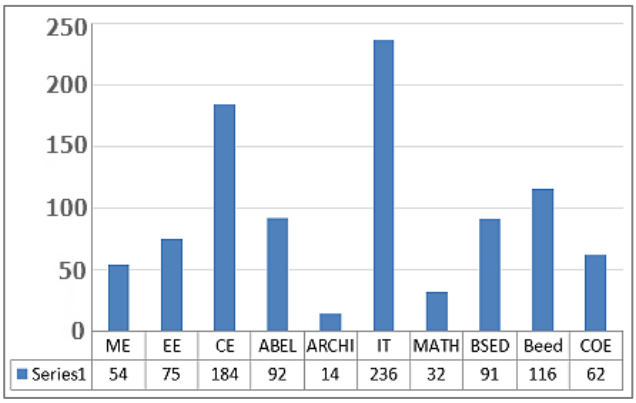

Figure 2. Total Number of dropouts per program

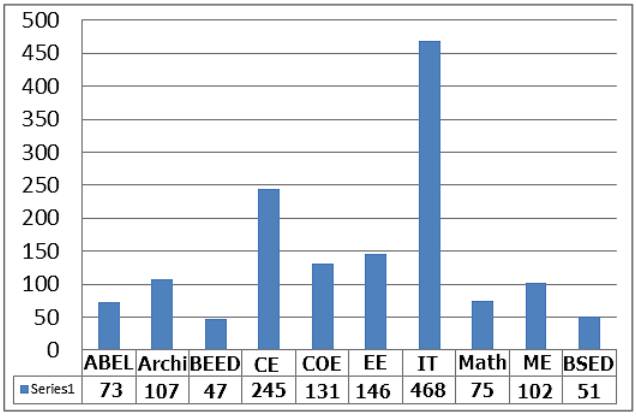

Figure 3. Total Number of Graduated per program

3.3. Predictor Variables

The main predictor variables considered in the study were high-school grade-point average (HSGPA) and grade average in the subjects, English, Mathematics, Science and TLE. The HSGPA used in this analysis was weighted grade-point average, that is, a HSGPA at 85. This is the minimum GPA needed to be admitted in the majority of programs offered at the campus. Grades in Science, Mathematics and TLE were also considered in the analysis because these are considered to be prerequisites in the subjects in college. These subjects are considered preparatory subjects where students have to take general courses in Mathematics, and English communication. In addition, this study considered the school where students graduated and their gender.

3.4. Model the Predictive Building

3.4.1. Classifying and Predictive Tool

Previous research reveals that the most popular learning machine used and supervised classification technique was the decision tree. It involves simple steps, very fast, and it is easy to apply therefore, very intuitive and easy to discuss and explain. Decision Tree can be applied to any domain (Lakshmi *et al.*, 2013). The main goal and the objective of the decision tree is to produce a tree model that will predict the value of a target variable by applying several input attributes. In addition, the output of the decision tree can be utilized as a decision support tool that utilized a tree-like graph with several predicted possible outcomes. A decision tree is a classifier in the form of a tree structure where each node is either: Leaf node- Leaf node is an indicator of the value of the target attribute (class) of examples, or a decision node- A decision node specifies all possible tests on a single attribute-value, with one branch and sub-tree for each possible outcome of the test (Chahal, 2013). Sample result is shown in figure 4.

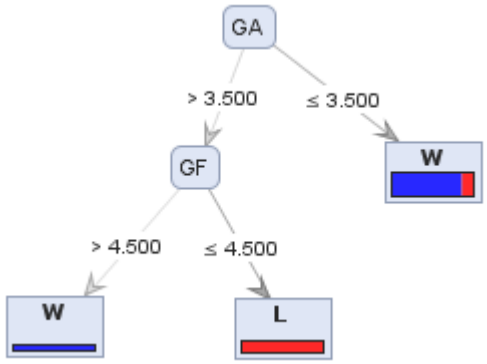

Figure 4. Sample tree structure result

In this study, a node in the tree is a predicting attribute, and its branches are drawn on the basis of predicting suitable program for each student. Every node provides a decision, predicting the success of students. However, to strengthen the predicting performance of the proposed tree model, ensemble methods were applied through combining several decision tree classifiers, boosting decision tree classifiers and bagging decision tree classifiers.

3.5.1. Building Model Process

Building the predictive/classification model is the next step. In this particular process, the selection of appropriate decision tree algorithms and ensemble models were finalized as classifiers under the cross-validation method. The proposed classification models used 8 input variables, as shown in Table 1. Data was collected from enrollment information of all the students who graduated from the school years 2013 - 2017. The attribute having maximum gain ratio value is the basis of splitting the nodes and the process will continue until it produces the complete decision tree. RapidMiner was utilized as a tool kit for experimentation and construction of the decision trees. In terms of the software tool kit, Moghimipour & Ebrahimpour (2014) study reveals Rapidminer obtained the highest accuracy as compared to three Data Mining Software. Figure 5 shows the decision tree construction. The decision tree result has a leaf node that is represented by the rectangle and oval representing the root node/splitting node.

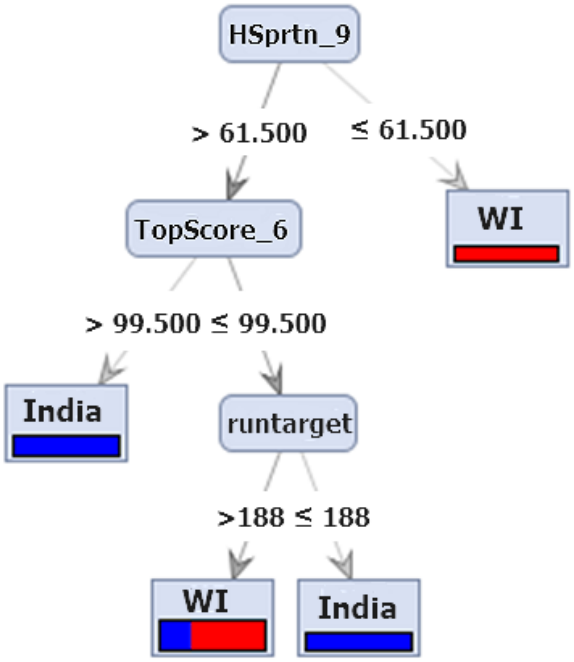

Figure 5. Decision Tree construction

This study used decision tree algorithms to generate predictive models to suggest programs based on the high school academic records of the students. According to Rokach & Maimon (2014) decision tree provides many advantages and some of which are the following:

- it is simple and can be clearly understood by the reader, end-user and analyst.
- it can accommodate different kinds of input data such as textual, nominal, and numeric.
- it can continue to process data that are erroneous or missing or uncompleted values.
- with a minimal amount of effort and time, it produces a high level of performance.
- it can work in data mining applications over a multi-variety of platforms.

However, there are drawbacks of decision tree classifiers such as the following:

- There is a high probability of overfitting in decision trees.
- Generally, it produces low prediction accuracy for a dataset as compared to other machine learning algorithms.
- Information gain in a decision tree with categorical variables gives a biased response for attributes with greater no. of categories.
- Calculations can become complex when there are many class labels (Gupta et al, 2017).

However, the research experimented and applied ensemble models to reverse the main weaknesses of decision tree classifiers and to form a strong learner machine, thus, increasing the accuracy of the model.

3.5.2. Ensemble Model

The ensemble model is composed of meta-algorithms that creates *n* learners from one algorithm sequentially or combine several machine learning techniques into one predictive model in order to decrease variance, bias and improve predictions (Hamilton, 2009). Furthermore, ensemble methods combine several decision tree classifiers to produce better predictive performance than a single decision tree classifier. The main principle behind the applying ensemble model is to group of weak learner algorithms combines to form a strong learning algorithm, thus, increasing the accuracy of the form model (Garg, 2018). Previous research results clearly show that the applying ensemble model obtained better accuracies and reliable rules as compared with other classifying component models, other conventional forecasting tools, and other combination schemes (Garg, 2018; Satyanarayana & Nuckowski, 2016). In addition, the study of Adejo (2018) reveals that heterogeneous ensemble techniques are more efficient and very accurate in the prediction of student performance and very useful in the proper identification of students at risk of attrition.

3.5.3. Vote (Stacking)

The Vote (stacking) ensemble operator is a nested ensemble operator that has a subprocess operator with at least two or more classification algorithms, called base learners. Furthermore, all the operators in the subprocess operator of the vote accept the given data set and generate a combined single classification model.

This study explored and experimented ensemble (vote) to handle imbalance data (graduate / not graduated data) and to improve the performance to suggest and identify suitable programs for a certain student. This study experimented with three tree classification algorithms and then combining the results into a single score in order to improve the accuracy of predictive analytics and data mining applications.

3.5.4. Bagging

Bagging is an ensemble model that creates *n* learners from one algorithm sequentially to decrease bias and variance results. The dataset randomly sampled with replacement and created *n* dataset in a given ratio. There can be data points that are misclassified by a given learner (Sewwandi, 2018). The error of the previous classifier is considered and a new weight is given to the misclassified data element letting that data element appear in new datasets more often. Therefore, bagging is used to help reduce the bias because it reduces result variance and overfitting. According to Breiman (1996), for unstable learning algorithms and unbalance data, bagging is an effective ensemble technique where there is a big change in prediction results with small changes in the training data set. In addition, Hasan *et al*., (2015) research reveals that

applying bagging ensemble model into decision tree classifiers improves the predicting accuracy performance of their study.

E.4 AdaBoost

AdaBoost a part of a family of ensemble algorithms that are able to convert and enhance weak machine learners to become strong machine learners. The main principle of adaboosting is to fit a sequence of weak machine learner models that are only slightly better than random guessing, such as small decision trees− to weighted versions of the data (Patel, 2017).

According to Patel (2017), Adaboost is found to be the best ensemble model for predicting the student's result based on the academic marks obtained in the current semester. Furthermore, Smolyakov (2017) found that applying adaboost improves the performances of tree algorithms.

3.5.5. Model Evaluation and Interpretation

To evaluate the predictive models 10 fold cross validation and percentage, split methods have been applied. 10 Fold Cross validation was utilized because the data is small and it is not feasible to split into two subsets. Thus, in order to minimize the bias, it makes full use of the dataset for training and for testing. Results were presented using confusion matrix that contains information about the actual and predicted classification done by the predictive models (Hamilton, 2009). In terms of comparing their performance, the confusion matrix that contains the precision, recall, and accuracy (Smolyakov, 2017) was employed.

Table 2. Confusion matrix for two-class classifier

Predicted Class	Graduated	Drop out	Class Precision
Graduated	tn	fp	Not Graduated
Not Graduated	fn	tp	Graduated
Class Recall	Drop out	Graduated	

Accuracy:

- The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using equation (1) :

$$AC = \frac{tn + tp}{tn + fp + fn + tp} \quad (1)$$

- The recall (in the case of positive cases) is the proportion of positive cases that were correctly identified, as calculated using equation (2):

$$Recall = \frac{tp}{tp + fp} \quad (2)$$

- The precision (in the case of positive cases) is defined as the proportion of negative cases that were classified correctly, as calculated using equation (3):

$$Precision = \frac{tp}{tp + fn} \quad (3)$$

In addition, ensemble models were applied in the tree algorithms and compared the result of other models without applying ensembles. This procedure provides a mechanism to gradually modify, correct noisy, and overfitting data may lead to a better classification.

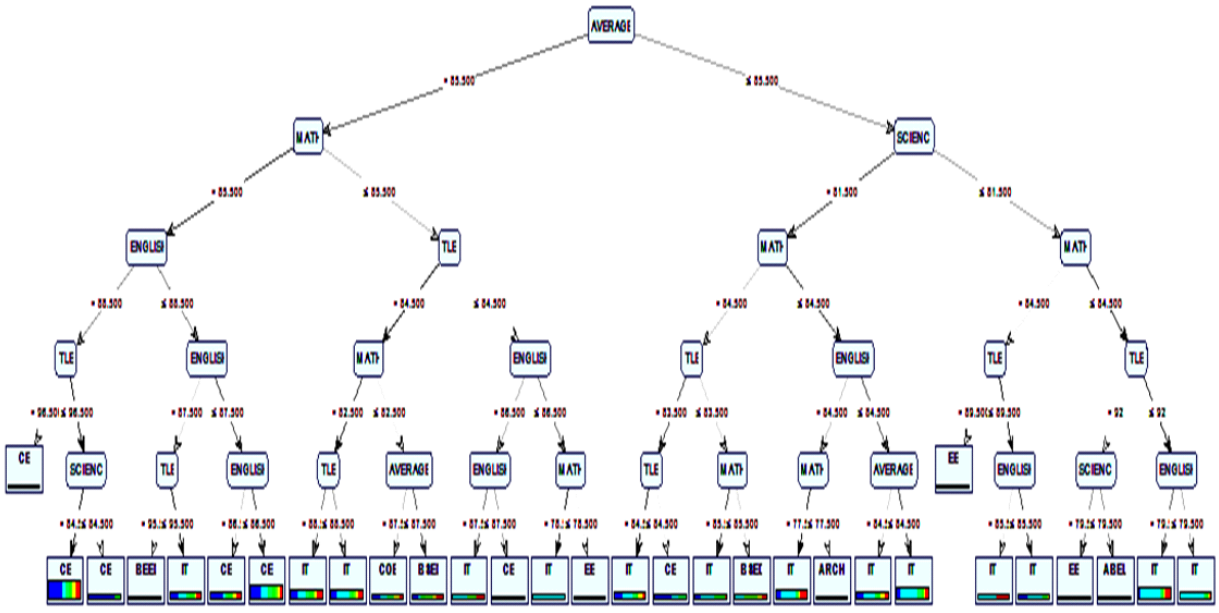

Figure 6. Decision tree produces by J-48 Algorithm

4. Results and Discussion

The model in figure 6 gives interesting information about students and provides guidance to students not to choose a track that is not suitable for them. The result indicates that the average grade is the best predictor of whether a student is going to graduate successfully or my dropped in their college year. In addition, grades in Mathematics, Science, and English become the next best predictor as shown in the figure above.

- For the Mathematics program. Students with an average grade of less than 85.5 and a grade of less than 84.5 in Mathematics have the tendency to drop or cannot finish their enrolled program. The results indicate students should have an average grade higher than 85.5 and Mathematics grade higher than 85 to be able to finish the Mathematics program.
- For the Information Technology program. The average grade is the main predictor of IT students and their average grade is less than 85.5 and TLE is another predicting variable. IT students should have a higher general average grade (GPA) of 85.5 in every subject and TLE.
- For the ABEL and Education programs. High School General Average Grade Point and English is the main predictor for ABEL and BS Education courses. Students should have a High School General Average Grade Point of 91 and an average grade of 87 in English. The result obviously reveals that English majors and future teachers should have a high grade in English.
- For the Engineering and Architecture programs. Students with a general average grade of less than 85 should not enroll in an Engineering program. The result also reveals that Science is also an important predictor for engineering courses. This result confirmed the study of Hayward et. al. (2014), Beaulac & Rosenthal (2018) that the confidence in quantitative skills were significant predictors for engineering students' success.

Rules Extraction A set of rules can be extracted from the decision tree. These rules are used to predict and classify the suitable program for each student. The class label in this tree acts as the suitable classified program after the end of a high school. The set of extracting rules from the decision tree is shown in Table 3.

Table 3. Rules extracted from the decision models

Rules Extracted	Programs
AVERAGE ≤ 85.500 MATHEMATICS < 84.500 : ME	Mechanical Engineering
AVERAGE ≤ 85.500 TLE < 84.500: IT	Information Technology
AVERAGE ≤ 85.500 SCIENCE > 81.500 MATHEMATICS > 84.500: CE	Civil Engineering
AVERAGE ≤ 85.500 SCIENCE > 81.500 MATHEMATICS ≤ 82.500 : COE	Computer Engineering
AVERAGE ≤ 85.500 MATHEMATICS ≤ 77.500: ARCHI	Architecture
AVERAGE ≤ 85.500 SCIENCE > 79.500 MATHEMATICS ≤ 78.500: EE	Electrical Engineering
AVERAGE ≤ 85.500 MATHEMATICS ≤ 85.500: BSED	BS Secondary Education
AVERAGE > 85.500 ENGLISH ≤ 88.500: BEED	BS Elementary Education
AVERAGE ≤ 85.500 ENGLISH ≤ 84.500:ABEL	AB English Language

The result reveals a very interesting prediction that the average grade in their high school is a very important predictor for all programs. In addition, Engineering course average grades in Mathematics and science are also a dominant predicting variable. In the case of IT, the average grade in TLE is also an important predicting variable. Variables gender and type of school were graduated are not influential variables for predicting student success as shown in the result.

Accuracy Results of Different Tree Classifiers and Applying Ensemble Approaches

Experimentation and evaluation results in the final stage in the model construction are shown in Table 3. The table reveals that applying the bagging ensemble to j-48 tree classifier obtained the highest accuracy as compared with other tree classifiers. Furthermore, a significant increase was seen in dropout recall and graduated precision measures, however, in terms of precision (dropout) and recall (graduated) applying bagging to J-48 did not show any increase in its performance as compared with single j-48 classifier.

Table 4. Comparison Accuracy Results of Tree Classifier and Applying Ensemble Approach

Classifier	Precision		Recall		Accuracy
	Graduated	Drop out	Graduated	Drop out	
Decision Tree	60.98	74.23	61.05	74.2	68.97
Forest Tree	56.8	77.44	62.49	73.04	69.22
J-48	58.16	76.82	62.4	73.51	69.39
Ensemble Model (Vote)	62.32	73.41	57.95	76.82	69.31
Ensemble Model (Bagging + Decision Tree)	57.19	71.20	55.75	72.39	65.76
Ensemble Model (Bagging + Forest Tree)	61.29	72.64	55.65	76.75	68.35
Ensemble Model (Bagging+j-48)	64.31	73.93	58.05	78.69	70.47
AdaBoost+Decision Tree	53.65	72.06	62.24	64.43	63.56
AdaBoost+Forest Tree	60.54	71.62	54.08	76.68	67.68
AdaBoost+j-48	61.57	73.00	57.32	76.33	68.73

The forest tree algorithm produces an impressive result in terms of classifying performance of 76.62 which is the highest value in the case dropout precision and obtained the highest predicting results in the case of a graduated recall. The table also reveals that the ensemble model, did not perform well in the classification task that it obtained a minimal increase in their classification performances in all measured areas. This result indicates that combining the ensemble model to these machine learning algorithms has a minimal increase in classification performance, which can be attributed to the average in performances of true/strong models and bad/weak classifiers. According to Smolyakov (2017) that building ensemble models with the most accurate models may not result in better ensemble models. Furthermore, experiments by Baba, Makhtar, Fadzli, & Awang (2015) show that in the case of substantial classification noise, bagging is much better than boosting and sometimes better than randomization.

5. Conclusions

This study has explored and analyzed the enrollment data that may have an impact on the study outcome of the students and proposes a tree classification algorithm to help them to choose a suitable program when they enter PSU-Urdaneta City Campus. The experimentation of tree classification algorithms and applying the ensemble approach shows that applying bagging into j-48 tree classifier attained the highest accuracy as compared with other models. In addition, the forest tree classifier achieves the highest value in the case of graduate recall and dropout precision. The experimental result revealed a minimal increase in their performance in all measure areas when the ensemble approach was combined/applied to other classification algorithms.

The experimental result shows a very interesting prediction that the general average grade (GPA) is a significant predictor for all programs. In addition, variable gender and type of school were not also significant predictors. The study shows the potential of data mining in higher education, especially when used to improve students' performance and detect early predictor of their success.

REFERENCES

[1] Macha, W., Mackie, C., and Magaziner, J., Education in the Philippines, https://wenr.wes.org/2018/03/education-in-the-Philippines.

[2] Kotsiantis, S., & Pintelas, P. (2005). Local voting of weak classifiers. International Journal of Knowledge-based and Intelligent Engineering Systems, 9(3), 239-248.

[3] Min, Y., Zhang, G., Long, R. A., Anderson, T. J., & Ohland, M. W. (2011). Nonparametric Survival Analysis of the Loss. Journal of Engineering Education, 100(2), 349–373. Retrieved from http://www.jee.org

[4] Tumapon, T., Creating a culture of student engagement, http://www.manilatimes.net/creating-culture-student-engagement/347968/, Sept. 2017

[5] Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. Journal of Applied Psychology, 94(6), 1479.

[6] Klopfenstein, K., & Thomas, M. K. (2009). The link between advanced placement experience and early college success. Southern Economic Journal, 873-891.

[7] Estrera, P. J. M., Natan, P. E., Rivera, B. G. T., & Colarte, F. B. Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School.

[8] M. N. Quadril and N. V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance
Using Decision Tree Techniques," Global Journal of Computer Science and Technology, vol. 10, no. 2, pp. 2 - 5, April 2010.

[9] Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: A case study. International Journal of Information and Education Technology, 6(7), 528.

[10] Angeline, D. M. D. (2013). Association rule generation for student performance analysis using apriori algorithm. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), 1(1), 12-16.

[11] Shovon, M., Islam, H., & Haque, M. (2012). An Approach of Improving Students Academic Performance by using k means clustering algorithm and Decision tree. arXiv preprint arXiv:1211.6340.

[12] Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business, 10(1), 3-12.

[13] W. Ham ¨al ¨ainen, ¨M. Vinni, Comparison of machine learning methods for intelligent tutoring systems, in: Intelligent Tutoring Systems, Springer, 2006, pp. 525–534.

[14] Abu Tair, M. M., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. Mining educational data to improve students' performance: a case study, 2(2).

[15] Taruna, S., & Pandey, M. (2014, February). An empirical analysis of classification techniques for predicting academic performance. In 2014 IEEE International Advance Computing Conference (IACC) (pp. 523-528). IEEE.

[16] Al-Twijri, M. I., & Noaman, A. Y. (2015). A new data mining model adopted for higher institutions. Procedia Computer Science, 65, 836-844.

[17] T. M. Christian, M. Ayub, Exploration of classification using nbtree for predicting students' performance, in: Data and Software Engineering (ICODSE), 2014 International Conference on, IEEE, 2014, pp. 1–6.

[18] Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting student academic performance at degree level: a case study. International Journal of Intelligent Systems and Applications, 7(1), 49.

[19] S. T. Jishan, R. I. Rashu, N. Haque, R. M. Rahman, Improving accuracy of students final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique, Decision Analytics 2 (1) (2015) 1–25.

[20] Strecht, P., Cruz, L., Soares, C., & Mendes-Moreira, J. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. International Educational Data Mining Society.

[21] Using Decision Trees to Predict Student Placement and Course Success CAIR Conference November 21, 2014.

[22] Beaulac, C., & Rosenthal, J. S. (2018). Predicting University Students' Academic Success and Choice of Major using Random Forests. arXiv preprint arXiv:1802.03418

[23] Zhang, G., Anderson, J., Ohland, M., Carter, R., & Thorndyke, B. (2004). Identifying Factors Influencing Engineering Student Retention: A Longitudinal and Cross-Institutional Study using Multiple Logistic Regression Models. Retrieved in, 18(1).

[24] Veenstra, C. P. (2008). Modeling Freshman Engineering Success.

[25] Cengiz, N., & Uka, A. (2014). Prediction of Student Success Using Enrolment Data. KOS, 14(17), 45-2.

[26] Al-Radaideh, Q. A., Al Ananbeh, A., & Al-Shawakfa, E. (2011). A classification model for predicting the suitable study track for school students. Int. J. Res. Rev. Appl. Sci, 8(2), 247-252.

[27] Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

[28] Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.

[29] Kabra, R. R., & Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. International Journal of Computer Applications, 36(11), 8-12.

[30] Kabakchieva, D. (2012). Student performance prediction by using data mining classification algorithms. International Journal of Computer Science and Management Research, 1(4), 686-690.

[31] Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011, April). Prediction of student academic performance by an application of data mining techniques. In International Conference on Management and Artificial Intelligence IPEDR (Vol. 6, No. 1, pp. 110-114).

[32] Shovon, M. H. I., & Haque, M. (2012). Prediction of student academic performance by an application of k-means clustering algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, 2(7).

[33] Yadav, S. K. and Pal, S. (2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. World of Computer Science and Information Technology Journal (WCSIT), 2(2), 51-56.

[34] Kovačić, Z. (2012). Predicting student success by mining enrolment data. Research in Higher Education Journal, 15(1).

[35] Simeunović, V. and Preradović, Lj. (2014). Using data mining to predict success in studying. Croatian Journal of Education, 16(2), 491-523.

[36] Cheewaprakobkit, P. (2015). Predicting student academic achievement by using the decision tree and neural network techniques. Catalyst, 12(2), 34-43

[37] Shah, N. S. (2012). Predicting Factors That Affect Students' Academic Performance by Using Data Mining Techniques. Pakistan Business Review, 13(4), 631-638.

[38] El Zeweidy, M., Osman, E., & Elhennawy, M. E. (2013). A comparative analysis of Techniques for Predicting Academic Performance. Journal of ACS: Advances in Computer Science, 294(1872), 1-42.

[39] Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. Mis Quarterly, 553-572.

[40] Mayilvaganan, M., & Kalpanadevi, D. (2014, December). Comparison of classification techniques for predicting the

performance of students academic environment. In 2014 International Conference on Communication and Network Technologies (pp. 113-118). IEEE.

[41] International Educational Data Mining, 2019, Society http://educationaldatamining.org/

[42] Nyce, C., & Cpcu, A. (2007). Predictive analytics white paper. American Institute for CPCU. Insurance Institute of America, 9-10.

[43] Lakshmi, T. M., Martin, A., Begum, R. M., & Venkatesan, V. P. (2013). An analysis on performance of decision tree algorithms using student's qualitative data. International Journal of Modern Education and Computer Science, 5(5), 18.

[44] Hemlata Chahal, "ID3 Modification and Implementation in Data Mining", International Journal of Computer Applications (0975-8887), Volume 80-No7, October 2013.

[45] Moghimipour, I., & Ebrahimpour, M. (2014). Comparing decision tree method over three data mining software. International Journal of Statistics and Probability, 3(3), 147.

[46] L. Rokach and O. Maimon. "Data mining with decision trees: theory and applications." World scientific, 2014.

[47] Gupta, Bhumika, Aditya Rawat, Akshay Jain, Arpit Arora, and Naresh Dhami. "Analysis of various decision tree algorithms for classification in data mining." Int J Comput Appl 8 (2017): 15-9.

[48] Hamilton, H. (2009). Computer Science 831: Knowledge discovery in databases, Retrieved from http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html

[49] Garg, R. (2018), A Primer to Ensemble Learning – Bagging and Boosting, https://www.analyticsindiamag.com/primer-ensemble-learning- bagging-boosting/ \

[50] Adhikari, R., Verma, G., & Khandelwal, I. (2015). A model ranking based selective ensemble approach for time series forecasting. Procedia Computer Science, 48, 14-21.

[51] Satyanarayana, A., & Nuckowski, M. (2016). Data mining using ensemble classifiers for improved prediction of student academic performance.

[52] Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61-75.

[53] Sewwandi, U., (2018) How to create ensemble models using rapid miner, https://towardsdatascience.com/how-to-create-ensemble-models-using-rapid-miner-72a12160fa51

[54] Breiman L.1996. Bagging predictors, Machine Learning, 24(2):123-140.

[55] Hasan, M. R., Siraj, F., & Sainin, M. S. (2015, December). Improving ensemble decision tree performance using Adaboost and Bagging. In AIP Conference Proceedings (Vol. 1691, No. 1, p. 030008). AIP Publishing.

[56] Patel, S., (2017). Machine Learning 101, https://medium.com/machine-learning-101/https-medium-com-savanpatel-chapter-6-adaboost-classifier-b945f330af06

[57] Shanthini A., G. Vinodhini and R.M. Chandrasekaran (2018), Predicting Students' Academic Performance in the University Using Meta Decision Tree Classifiers, Journal of Computer Science

[58] Smolyakov, V., (2017) "Ensemble Learning to Improve Machine Learning Results How ensemble methods work: bagging, boosting and stacking", https://blog.statsbot.co/ensemble-learning-d1dcd548e936.

[59] Baba, N. M., Makhtar, M., Fadzli, S. A., & Awang, M. K. (2015). Current issues in ensemble methods and its applications. Journal of Theoretical and Applied Information Technology, 81(2), 266.

[60] Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40

[61] Mallinckrodt, B., & Sedlacek, W. E. (1987). Student retention and the use of campus facilities by race. NASPA Journal, 24, 28-32.

[62] Di Pietro, G., & Cutillo, A. (2008). Degree flexibility and university drop-out: The Italian experience. Economics of Education Review, 27(5), 546-555.

[63] Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive Mathematicsematical models. Computers & Education, 61, 133-145.

[64] Hayward, C., Hetts, J. Sorey, K., Willett, T. (2014) Using Decision Trees to Predict Student Placement and Course Success CAIR Conference.