PHILIPPINE ENGINEERING JOURNAL
PEJ 2021; Vol. 42, No. 2: 75-98

Received: 03 August 2021
Accepted: 04 December 2021

Modelling Student Dropout using AdaBoost and Survival Analysis

Mikayla Alexis D. Sagun*, Patricia Dolores M. Soriano,
Jhoanna Rhodette I. Pedrasa, and Darvy P. Ong
Electronics and Electrical Engineering Institute, College of Engineering,
University of the Philippines, Diliman, Quezon City, 1101, Philippines
**Corresponding Author: mikayla.sagun@eee.upd.edu.ph*

Abstract— *The average graduation rate of UPD COE freshmen admitted between 2009 and 2013 is 66.89%. The UPD COE graduation rate is quite low compared to other schools, indicating that it is important to investigate the dropout rates of students as well. Existing studies made use of several different models in order to predict student dropout. These studies made use of both pre-enrollment data and data on student performance per semester. Out of the different models used, the AdaBoost model and the Cox models consistently performed well. For this study, the AdaBoost model and time-varying Cox model were used to predict whether a student drops out, predict when a student will dropout, and analyze the features that lead to student dropout. Hazard ratios from the Cox model allow us to know whether the features increase or decrease risk of dropout. Pre-enrollment data and post-enrollment data was used to analyze student dropout. Higher number of semesters of absence without leave increase the risk while high school GWA and getting accepted in the student's first or second choice degree program decrease the risk of dropout. These features were found to be significant factors that affect dropout risk for both 4-Year and 5-Year programs. Of the two models, the AdaBoost model performed better at predicting student dropout and drop time. The results of the models can be used to help identify at-risk students as early as possible and guide them with regards to their specific needs..*

Keywords— adaboost, machine learning, student dropout, student retention, survival analysis

I. INTRODUCTION

1.1 Background of the Study

The University of the Philippines Diliman College of Engineering (UPD COE) is the largest college on campus. Every year, thousands of new students enter the college with the hopes of completing an undergraduate degree in engineering. Despite this, the average graduation rate of UPD COE freshmen admitted between 2009 and 2013 is only 66.89%. Comparing UPD COE's 66.89% 5-year graduation rate to the Massachusetts Institute of Technology's 4-year graduation rate of 85.9%, the difference is quite hard to ignore [1]. Graduation data also shows that 39.01% of the 4796 freshmen admitted between 2009 and 2013 dropped out from their original degree programs before they graduated. In this study, dropout is defined as a student who did not finish the initial degree that they enrolled in under the UPD COE.

The total population of the UPD COE studied in this project is 4796 students comprised of 4176 students from the 5-Year Programs and 620 students from the 4-Year Program. Table 1 presents the breakdown of dropouts in the UPD COE for the freshmen of A.Y. 2009-2013. For both the 4 and 5-Year programs, the most common reason for a student to be considered as a dropout is shifting out of the college.

Table 1. Breakdown of Dropout Students in the College of Engineering

Program	Shift Within the College	Shift Out of the College	Honorable Dismissal	Absence Without Leave	Total
4-Year	41 (6.61%)	164 (26.45%)	22 (3.55%)	65 (10.48%)	292 (47.10%)
5-Year	417 (9.99%)	621 (14.71%)	132 (3.16 %)	409 (9.79%)	1579 (37.81%)

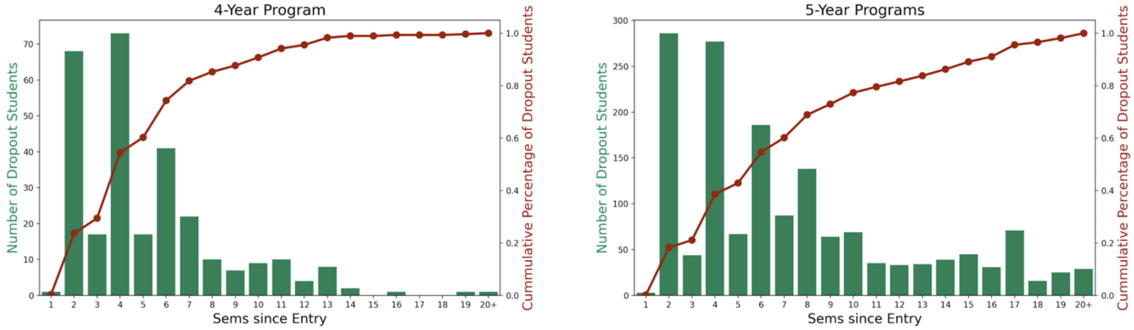

Figure 1. Dropout Histogram for the 4 and 5-Year programs

Figure 1 presents the histogram of dropout for students over time in the 4-Year program and the 5-Year programs. The percentage of students who drop out greatly increases at the second semester of each school year compared to the first semester. This is as expected since most colleges within the university accept shiftees for the first semester. By the on-time graduation semester, 77.33% of all dropouts have left the college for the 5-Year programs and 85.27% for the 4-Year program. Additionally, students mostly drop out within the first half of the on-time completion period, with 54.45% of all dropout students leaving before the 5th semester in the 4-Year program and 54.65% leaving before the 7th semester in the 5-Year programs.

In this study, machine learning and statistical analysis were used to model student dropout. The results of the modelling can be used to address concerns students may have regarding retention. The objectives of the project are as follows:

- Identify and analyze the factors that affect a student's risk of dropping out
- Develop prediction models to identify students who are at risk of dropping out of their course and when they are likely to drop out

1.2 Literature Review

Several studies have used machine learning techniques and Survival Analysis to study and model student retention and attrition. Berens et. al. first developed models to predict student dropout in a state university (SU) and a private university (PU) using Logistic Regression (logit), Neural Networks, and Decision Trees, specifically Bagging Random Forest(BRF) [2]. After combining the predictive power of the three models using the AdaBoost algorithm, the AdaBoost accuracy and recall rates for the SU are 75% and 87%, respectively, in the first semester and in the fourth semester, 82% and 93%, respectively. This is an improvement from the BRF's accuracy of 75% in the first semester and 81% in the fourth semester while recall is 86% in the first semester and 91% in the fourth semester.

Ameri et. al. used the Cox Proportional-Hazards (Cox) model and a time-varying Cox model to predict student dropout [3]. They compared the performance of these models to logistic regression, AdaBoost, and decision tree. From the two experiments they conducted, Cox based models had higher accuracy with an approximate of 9% increase in accuracy when predicting student dropout. In addition to student dropout prediction, they also estimated semester of dropout of students. For this experiment, the performances of Cox, Support Vector Regression (SVR), and linear regression were compared. Out of the three models, the Cox model performed the best.

Similarly, Chen, Johri, and Rangwala compared the performance of their Survival Analysis approach to the performance of other machine learning algorithms like Logistic Regression, Decision tree, Random Forest, Naive Bayes, and AdaBoost [4]. They found that Survival Analysis did well when there was less than two semesters' worth of data. Logistic Regression and AdaBoost use features with high predictive power to get more accurate predictions.

II. METHODOLOGY

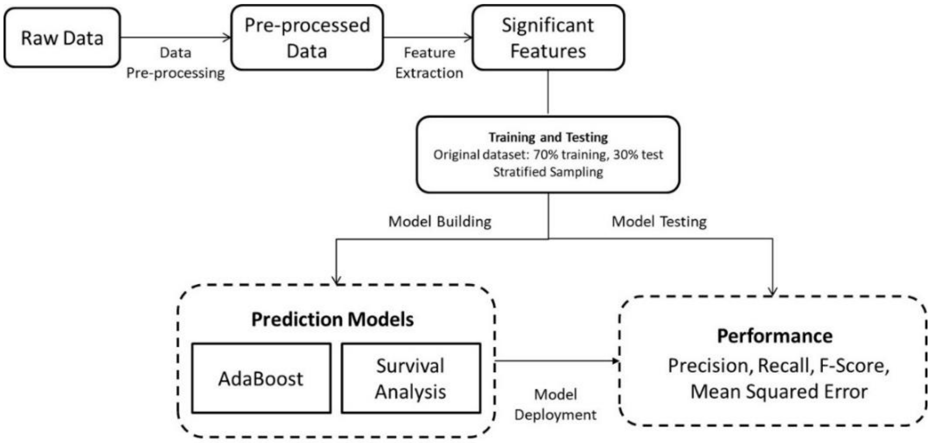

Figure 2. Design Overview

Figure 2 shows the overview of the project methodology. The proposed design for the project is split into multiple phases. First is data pre-processing, where the collected data is analyzed and explored. The next phase is identifying the important factors by computing for the correlations of the factors and creating the dataset to be used by the models by splitting the data. The final phases deal with creating and refining the models for predicting student dropout. All stages make use of Python as its main programming language while both Python and R programming language were used in creating and training the prediction models.

2.1 Data Pre-Processing

The dataset used in this study contains 4796 entries of student data for admitted freshmen from AY 2009-2013. The dataset was separated into two subsets, one for the 4-Year program and one for the 5-Year programs. Due to the differences in program duration and dropout characteristics between the two, it was decided that they will be analyzed and modelled separately. Dropout was determined based on whether the student shifted out of their initial program, honorable dismissal, and graduation status.

The Socialized Tuition and Financial Assistance Program (STFAP) and Socialized Tuition System (STS) provides additional subsidy through tuition discount and other forms of financial assistance. The amount of subsidy a student receives is based on the family income and the socioeconomic characteristics of the household of the student [6]. For the years 2009-2010, the default bracket of the STFAP system was "Bracket B," while for 2011 onwards was "Bracket A." In 2013, the system changed from STFAP to STS. To unify the system used in the dataset, the STS brackets were converted to the STFAP brackets based on Table 1. Missing values for the financial bracket were replaced with the default values of the STFAP/STS System.

Table 2. STFAP and ST System Comparison of Income Cutoffs [6]

STFAP Bracket		ST System Partial Discount	
A	Over P1 Million	ND	Over P1.3 Million
B	P250,001 to P500,000	33%	P650,001 to P1 Million
C	P135,001 to P250,000	60%	P325,001 to P650,000
D	P80,001 to P135,000	80%	P135,001 to P325,000
E1	P80,000 and below	100%	P80,001 to P135,000
E2	P250,001 to P500,000	100%	P80,000 and below

For this study, the data was analyzed and modeled per semester with dropout being the event of interest. Features that can change every semester were also split semester-wise to accommodate the time-varying features Scholarship Status, STFAP/STS Bracket, number of AWOL semesters, and number of LOA semesters.

The dataset contained multiple entries that had incomplete or missing data and was preprocessed as follows:

- Entries that had incomplete data for almost all features were removed.
- Data entries that did not exit the college, through graduation or dropout, were assumed to be absent without leave (AWOL) after their last enrolled semester.

- Maximum allowable semesters of AWOL or leave of absence (LOA) is set to 4 semesters for 4-Year programs and 5 semesters for 5-Year programs for this study based on the maximum residence rule (MRR) in UP Diliman [5].

To have all students on the same timeline, the measurement of time is the number of semesters since entering the university. The following adjustments were made:

- The first year, first semester of every student is 1, no matter what year they entered.
- Semester of exit from the student's initial program refers to the student's last semester in the program before exit.

Additionally, categorical features need to be encoded into numerical data for data analysis and prediction modelling. The following encoding processes were done:

- Features with only two categories were converted 1 or 0.
- Features with location, such as Home Province, were considered as within Metro Manila or not and converted to 1 or 0.
- One-hot encoding was used on the other categorical features that were non-numerical

Table 3 shows the features that remained after preprocessing the data and were used in the next phase of the study.

Table 3. Continuous and Categorical Factors

Continuous Factors			
Code	Description	Code	Description
AGE	Age Upon Entry	*HS-GWA*	High School GWA
UPCAT-AVE	UPCAT Score Average T-Score	*UPG*	UP Predicted Grade – Overall
UP-MPG	UP Math Predicted Grade	*UPCAT-MATH*	UPCAT Math T-Score
UPCAT-LP	UPCAT Language Proficiency T-Score	*UPCAT-RC*	UPCAT Reading Comprehension T-Score
UPCAT-SCI	UPCAT Science T-Score		
Categorical Factors			
Code	Description	Code	Description
SEX	Sex	*REL*	Religion
RH	Lived in a Residence Hall	*HS-Type*	High School Type
HP	Home Province	*HS-P*	High School Province
DEG	Degree Program Upon Entry	*CH-DEG*	Accepted in 1st/2nd Choice Degree Program
Categorical Factors: Semestral			
Code	Description	Code	Description
STFAP/STS	STFAP/STS Bracket	*SCHOLAR*	Scholarship Status
AWOL-Sems	No. of Semesters AWOL	*LOA-Sems*	No. of Semesters LOA

2.2 Feature Extraction

In order to identify which features can be removed, correlation tests were performed to determine if the features are independent from each other. For continuous features, Pearson's coefficient was used. The Pearson's coefficient is a value between -1 and 1 that describes the linear relationship between two variables. If the correlation is -1, the two variables are negatively linear related. On the other hand, a correlation of 1 indicates that the variables are positively linear related. A correlation value of 0 indicates that the variables are independent of each other [7]. The correlation coefficient between continuous factors x and y is computed using Eq. 1 where n is the sample size.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2] - [n\sum y^2 - (\sum y)^2]}} \tag{1}$$

As for categorical features, the Cramer's V Coefficient was used to determine the strength of the association between two categorical variables. The Cramer's V coefficient is a value between 0 and 1 wherein a coefficient of 0 indicates minimal association and a coefficient of 1 means a high association between the two categorical variables being tested [8]. The equation for the Cramer's V coefficient can be seen in Eq. 2, where χ^2 refers to the Chi-Square Test Statistic, *L* refers to the smaller number between the categories of variable A and the categories of variable B, and *n* is the total sample size. To get the Chi-Square Test Statistic, Eq. 3 is used where n_{rc} is the number of observations at the intersection of category *r* from variable A and category *c* from variable B [9].

$$V = \sqrt{\frac{X^2}{n(L-1)}} \tag{2}$$

$$X^2 = \sum \frac{\left(n_{rc} - \frac{n_r * n_c}{n}\right)^2}{\frac{n_r * n_c}{n}} \tag{3}$$

For this study, pairs with a correlation coefficient of 0.5 or higher in magnitude were considered as features with a strong relationship. Features that were removed were chosen to lessen association between all features and features that had more categories than other features they were highly associated with were also removed.

2.3 Dataset Preparation

After the final list of features were selected using the correlation tests, stratified sampling was used to split the data into the respective training sets and testing sets for the AdaBoost and Cox models. The training datasets consisted of 70% of the pre-processed dataset and the remaining 30% for the testing dataset. Since not all semesters were under study, all feature data was cut-off at Semester 10 for the 4-Year program data while the 5-Year program data was cut-off at Semester 12. In order to accommodate predicting both the dropout status and semester of dropout, the AdaBoost model has 2 variations that need datasets for each goal. On the other hand, the time-varying Cox (TV-Cox) model only requires the changes for the time-varying features and time of change.

The distribution of datapoints for the overall, training, and testing sets for the 4-Year and 5-Year programs are found in Table 4, where strata used in splitting the data is Dropout Status.

Table 4. Distribution of Dropout Status Outputs Per Dataset

	4-Year Program			5-Year Program		
	Overall	Training	Testing	Overall	Training	Testing
Dropout	265 (42.74%)	185 (42.63%)	80 (43.01%)	1289 (30.87%)	900 (30.81%)	386 (30.83%)
Not Dropout	355 (57.26%)	249 (57.37%)	106 (56.99%)	2887 (69.13%)	2021 (69.19%)	866 (69.17%)

Table 5. Distribution of Dropout Time Outputs Per Set

Semester of Dropout	5-Year Program		4-Year Program	
	Training	Testing	Training	Testing
2	200 (22.22%)	86 (22.28%)	47 (25.54%)	21 (26.25%)
3	31 (3.44%)	13 (3.37%)	12 (6.52%)	5 (6.25%)
4	194 (21.56%)	83 (21.5%)	51 (27.72%)	22 (27.5%)
5	47 (5.22%)	20 (5.18%)	12 (6.52%)	5 (6.25%)
6	130 (14.44%)	56 (14.51%)	29 (15.76%)	12 (15%)
7	61 (6.78%)	26 (6.74%)	15 (8.15%)	7 (8.75%)
8	97(10.78%)	41 (10.62%)	7 (3.80%)	3 (3.75%)
9	45 (5%)	19 (4.92%)	5 (2.72%)	2 (2.5%)
10	48 (5.33%)	21 (5.44%)	6 (3.26%)	3 (3.75%)
11	24 (2.67%)	11 (2.85%)		
12	23 (2.56%)	10 (2.59%)		

The AdaBoost Dropout Status model and TV-Cox model training and testing datasets consist of all dropout and not dropout entries. The training and testing sets for the AdaBoost Dropout Time model, however, consist of only the entries that were labelled as dropouts after data pre-processing and cutting off semesters not under study. Because dropout counts in the 1st semester in both the 4-Year and 5-Year Programs are very low, entries with dropout time 1 were excluded from training and testing for the Dropout Time model. The training and testing dataset distributions for the Dropout Time model can be seen in Table 5.

2.3.2 Time-Varying Dataset for Cox Model

The counting process data layout is used to format the dataset for the TV-Cox model. Two columns are allotted for time for each student to indicate the start and stop times. The start and stop times indicate when these changes were recorded. In addition, it is possible for a student to have multiple rows. Every change in the time-varying feature will generate an additional row for the student. Table 6 presents an example of the counting process format [10] where the time-varying features STS/STFAP, SCHOLAR, AWOL-Sems, LOA-Sems, as well as the start and stop times are presented.

Table 6. Example of Counting Process Format [10]

ID	STS/STFAP	SCHOLAR	AWOL-Sems	LOA-Sems	Start	Stop	Dropout Status
1	B	1	0	0	0	4	FALSE
1	A	1	0	0	4	5	FALSE
1	A	0	0	0	5	7	TRUE
2	B	0	0	0	0	8	FALSE

In order to properly assess the model for each semester, a different testing dataset is generated for each semester. The data is cut off at each semester in order to include time-varying data until the specified semester only. For example, when training the model for semester 4, only the first row for ID 1 is included and stop time for student 2 is 4.

2.4 Prediction Models

AdaBoost and Survival Analysis are the prediction models to be used for this study. AdaBoost is chosen for its accuracy and transparency in classification, while Survival Analysis is chosen for its ability to show how specific factors can influence student dropout and answer the question of when a student will dropout.

2.4.1 AdaBoost Model

The AdaBoost model has two implementations each for 4-Year Program dataset and 5-Year Program dataset:

- Dropout Status - outputs a 'Yes'(1) or a 'No'(0) when predicting if a student is likely to dropout from their initial degree program or not.
- Dropout Time - predicts the semester when a student is likely to dropout from their initial degree program. The model produces a number between 2 to 12 for the 5-Year Program and a number between 2 to 10 for the 4-Year Program to indicate the possible semester of dropout.

The main idea behind the AdaBoost algorithm is that it combines weak or base learners to make a strong classifier by learning from mistakes made in previous rounds during training. According to Schapire [11], for each training iteration a weak learning algorithm is fitted on the training data to find a weak classifier h_t with the lowest weighted error ϵ_t. Since the goal is to minimize classification error, the classifier's weight α_t, which determines its significance in the final prediction, is determined such that $\alpha_t h_t$ minimizes the exponential loss. The more accurate a classifier is the more say it has on the final output. Before the next iteration, the sample weights are updated and normalized such that wrongly classified samples will be given more weight in the next round to find the new h_t. The final output or prediction is made by getting the sign of the sum of the weighted prediction of each classifier as seen in Eq. 4.

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{4}$$

Different kinds of learning algorithms can be taken as the base learning algorithm, such as decision trees, neural networks, etc. For the AdaBoost implementation in this study, Decision Trees were used as the weak learner because aside from the fact that decision trees are easier to understand and interpret, the default base learner for AdaBoost in Python is a Decision Tree Classifier. Other parameters involved in the development of the model are the depth of the Decision Trees, the number of base estimators, and the learning rate [12].

To find the optimal set of hyperparameters to improve the performance of the AdaBoost models, Grid Search was used in hyperparameter tuning. Grid Search trains and validates multiple models with every possible combination of the specified hyperparameters and finds the best combination by k-fold cross validation. The model that has the best cross validation score would be used to fit on the data. For this study, $k = 5$ was used for k-fold cross validation in the Grid Search implementation. The hyperparameters used in tuning the models and their values are listed in Table 7.

Table 7. GridSearch Hyperparameters

	4-Year Program		5-Year Program	
	Dropout Status	Dropout Time	Dropout Status	Dropout Time
max_depth	1, 2	1, 3, 5	1, 2, 3	1, 3, 5
n_estimators	25, 50, 75			
learning_rate	0.1, 1		0.01, 0.1, 1	

The implementation of the AdaBoost models also has the ability to rank features based on their Gini importance, or the total decrease in node impurity weighted by the number of samples it splits and then averaged over all trees of the ensemble. The higher the Gini importance, the more important the feature [12].

2.4.2 Cox Proportional-Hazards Model

One of the models in survival analysis is the Cox model. The Cox model is a semi-parametric model that can process both categorical and continuous features. The baseline hazard represents the hazard when the predictor features are set to 0. It is considered semi-parametric because it does not specify the form of its baseline hazard function, unlike parametric models where they are fully specified [10]. However, one assumption for the Cox model is that the predictor values do not change over time or the change is very minimal. In the case that there are time-varying predictors needed for the model, such as this study, the Cox model can be extended to accommodate these predictors [10]. Eq. 5 is the extended formula for the hazard function where X_i are the time independent predictors and $X_j(t)$ are the time dependent predictors. The hazard function provides the instantaneous rate that an event, in this case dropout, will occur at a specified time t [10].

$$h(t, X(t)) = h_0(t)exp\left[\sum_{i=0}^{p1} \beta_i X_i + \sum_{j=0}^{p2} \delta_j X_j(t)\right] \tag{5}$$

One important aspect to look at in the hazard function are the hazard ratios. These show the relationship of the feature to the event. The hazard ratio for the TV-Cox model can be computed using the Eq. 6 where X^*_i is the placebo group and X_i is the treatment group.

$$HR = exp\left[\sum_{i=0}^{p1} \beta_i (X_i^* - X_i) + \sum_{j=0}^{p2} \delta_j X_j^*(t) - X_j(t)\right] \tag{6}$$

A hazard ratio above 1 indicates an increase of risk, while a value below 1 indicates a decreased risk [13]. The hazard ratios were used to identify the features that affect dropout and how they affect dropout.

For this study, the output of the survival function is a probability value ranging from 0 to 1 at specified semesters. A cutoff value determines whether it will be classified as 0 or 1, similar to logistic regression. The Youden index was maximized in order to find the optimal cutoff point. The Youden index is a performance metric that takes into account the sensitivity and specificity of a binary classification model [14]. Eq. 7 presents the formula to compute Youden Index. Predicted probabilities are generated for each row in the dataset of the specified semester. In addition, predicted probabilities of each semester were generated separately. For example, predictions for semester 2 do not affect predictions for semester 3.

$$YJS = \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1 \tag{7}$$

2.4.3 Performance Metrics

Precision, recall, f-measure, and mean squared error were used to measure the performance of the models. Recall measures how much of the positive cases were predicted correctly. Eq. 8 presents the equation for recall.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

Precision measures how much of the predicted cases are correct. Eq. 9 presents how precision is computed.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Precision and recall give focus to the positive class, which works well when working with imbalanced data. However, an increase in precision may mean a decrease in recall and vice versa. To provide a way for both precision and recall being handled equally, F-score is used [15]. F-score is the harmonic mean of precision and recall. Eq. 10 presents how the F-score is computed.

$$F - Score = \frac{2 * recall * precision}{recall + precision} \tag{10}$$

Mean Squared Error (MSE) measures how close the predictions are to the actual values [16]. This was used to compare the performance of the two models on predicting the semester of dropout and actual semester of dropout. Eq. 11 presents the formula for computing the MSE.

$$MSE = \frac{\sum_{i=1}^{n}(Actual_i - Predicted_i)^2}{n} \tag{11}$$

The performance measures for the models will mainly be the precision, recall, F-score, and MSE. A Precision-Recall (PR) curve is a plot of recall versus precision that is used to show the tradeoff between precision and recall for different probability thresholds. PR curves are used to visually compare precision-recall behavior of the models developed in this study.

III. RESULTS AND DISCUSSION

This chapter provides the results of the study. The first section presents the resulting factors to be used for modelling. The second section provides descriptive analysis of the dataset after pre-processing of data and factors for modelling are chosen. The third section presents the results of the models used. The fourth section presents the comparison of the two models.

3.1 Feature Selection

For both the 4-Year and 5-Year programs, the features UP-MPG, UPG, and the UPCAT-AVE all had coefficients of greater than 0.5 or less than -0.5 with other continuous features. These relatively high correlations are understandable since these three features are calculated using the UPCAT subject-related test scores.

Table 8. Remaining Features after Feature Selection

Continuous Factors	Categorical Features	Categorical Features: Semestral
HS-GWA	SEX	STFAP/STS
AGE	REL	LOA-Sems
UPCAT-RC	CH-DEG	AWOL-Sems
UPCAT-MATH	HP	SCHOLAR
UPCAT-SCI	HS-Type	
UPCAT-LP	RH	
	DEG (for 5-Year Programs)	

HP and HS-P have a strong association with a V value greater than 0.5, which is understandable since going to high schools that are within the areas of their home is common for most students. There were also associations within the semestral features such as

SCHOLAR and STFAP/STS. This is expected since time-varying features basically represent the same feature for each semester. However, because there is no association strong enough with other categorical features that are non-time-varying, all semestral features were kept. The final list of categorical and continuous features used in the prediction models are listed in Table 8.

3.2 Prediction Models

3.2.1 AdaBoost

Table 9 presents the test performance metrics results for the 4-Year and 5-Year Dropout Status models. The confusion matrices for both models can be seen in Table A.1a and Table A.1b. For the 4-Year program, the best hyperparameter values were *max_depth* is 1, *n_estimators* is 50, and *learning_rate* is 1. The Dropout Status model for the 4-Year program has a precision score of 97.4%, a recall score of 92.5%, a F-Score of 94.9%, and a MSE of 0.043.The best hyperparameter values for the 5-Year program Dropout Status model were *max_depth* is 1, *n_estimators* is 75, and *learning_rate* is 1. The 5-Year program Dropout Status model has a precision rate of 96.1%, a recall rate of 95.9%, a F-score of 96%, and a MSE of 0.025.

Table 9. Performance of AdaBoost Dropout Status Model for 4-Year and 5-Year Programs

Program	Precision	Recall	F-Score	MSE
4-Year	0.974	0.925	0.949	0.043
5-Year	0.961	0.959	0.960	0.025

Significant features that are common and rank high for both programs are the features *AWOL-Sems*, *HS-GWA*, *AGE* and *UPCAT scores*, specifically Science and Reading Comprehension T-scores.

The best hyperparameter values found after performing Grid Search for the 4-Year program Dropout Time model were *max_depth* is 5, *n_estimators* is 75, and *learning_rate* is 1. For the 5-Year program Dropout Time model, the optimal hyperparameters were found to be *max_depth* is 5, *n_estimators* is 50, and *learning_rate* is 0.01. Test performance metric results are discussed in more detail in Section 3.4.

Significant features that are common to both programs are the features *AWOL-Sems*, *UPCAT scores*, *HS-GWA*, and semestral features that pertain to *STFAP/STS Brackets*. The Gini importance values for the top 10 ranked features for the models can be seen in Appendix B.

3.2.2 Time-varying Cox Model

Table 10 presents the reference levels used for the multi-level categorical features. The reference level serves as the baseline for computation of the hazard ratios. For DEG and HS-Type, the reference level was based on the lowest percentage of dropout. On the other hand, bracket A was chosen as the reference for STFAP/STS since it is the highest bracket and indicates no need for financial assistance. Roman Catholic was chosen as the reference level for religion as a majority of the students belonged to this group. For binary categorical features and continuous features, the reference level is the lowest value or 0.

Table 10. Reference Levels for Multi-Level Categorical Features

Feature	5-Year Programs	4-Year Program
HS-Type	UP	State
REL	Roman Catholic	Roman Catholic
STFAP/STS	Bracket A	Bracket A
DEG	Program E	

Table 11 presents the significant features of the model and the corresponding hazard ratios. For both the 4 and 5-Year programs, features that decrease risk of dropout is being in the student's 1st or 2nd choice degree program and having a scholarship. On the other hand, a higher number of semesters spent AWOL or LOA increases the risk of dropout. Additionally, for the 4-Year program, older age of entry, female students, and students in brackets D, E1, and E2 have higher risks of dropout. For 5-Year programs, students in programs A, C, F, G, H, I, and J, compared to students in program E, and Agnostic students have an increased risk of dropout. On the other hand, higher HS-GWA and UPCAT-MATH and students residing in Metro Manila, where the university is located, have lower risks of dropout.

Table 11. List of Significant Features and Hazard Ratios

4-Year Program					
Feature	HR	Effect	Feature	HR	Effect
AGE	3.864	Increase	STFAP/STS - Bracket D	2.010	Increase
SEX- F	1.791		STFAP/STS - Bracket E1	4.526	
AWOL-Sems	2.195		STFAP/STS - Bracket E2	34.046	
LOA-Sems	3.280		Scholar	0.507	Decrease
			CH-DEG	0.482	
5-Year Programs					
Feature	HR	Effect	Feature	HR	Effect
DEG - Prog A	4.368	Increase	DEG - Prog I	2.323	Increase
DEG - Prog C	2.275		DEG - Prog J	1.967	
DEG - Prog F	2.229		AWOL-Sems	2.353	
DEG - Prog G	2.300		LOA-Sems	1.915	
DEG - Prog H	1.593		REL – AGNOSTIC	1.597	
HP	0.846	Decrease	HS-GWA	0.435	Decrease
Scholar	0.690		CH-DEG	0.590	
UPCAT-Math	0.450				

3.3 Model Comparison

Table 12. List of Common Significant Features for AdaBoost and Cox models

Features for the 4-Year Program		
AWOL-Sems	AGE	HS-GWA
CH-DEG	SEX- F	
Features for the 5-Year Programs		
AWOL-Sems	LOA-Sems	CH-DEG
HS GWA	HP	UPCAT-Math

Table 12 presents the list of features that were found to be significant in both the AdaBoost and TV-Cox models. Number of semesters AWOL, getting into the 1st or 2nd choice program, and high school GWA are the features that are significant to dropout for both 4-Year and 5-Year programs.

Table 13. Comparison of Model Performance for the 4-Year Program

Sem	Precision		Recall		F-Score		MSE	
	AB	CM	AB	CM	AB	CM	AB	CM
2	0.955	0.213	1	0.684	0.977	0.325	0.013	0.290
3	1	0.197	0.600	0.542	0.750	0.289	0.025	0.291
4	0.750	0.329	0.818	0.511	0.783	0.400	0.125	0.293
5	0.500	0.321	0.400	0.490	0.444	0.388	0.063	0.296
6	0.733	0.324	0.917	0.407	0.815	0.361	0.063	0.290
7	0.667	0.313	0.571	0.299	0.615	0.305	0.063	0.300
8	0.500	0.328	0.333	0.310	0.400	0.319	0.038	0.299
9	0.667	0.328	1	0.253	0.800	0.286	0.013	0.295
10	1	0.387	0.333	0.300	0.500	0.338	0.025	0.297

Table 13 presents the side-by-side comparison of the test results for the dropout time models for the 4-Year program. The confusion matrix for the AdaBoost Dropout Time model is presented in Table A.2a. The precision and MSE of the TV-Cox model generally increase while recall decreases as the number of semesters increases. One factor that may affect this is the initial size of the test dataset, which had 186 entries at semester 2 and 318 entries at semester 10, and the format of the dataset to accommodate the time-varying features. As for the AdaBoost model, the model performs the best at semester 2 for all performance metrics but there is no consistent pattern for the rest of the semesters except that the model had a harder time classifying dropout semester since the number of dropouts per semester were few. AdaBoost had a much higher precision and lower MSE compared to the TV-Cox model. For recall and F-Score, AdaBoost also had a higher recall and F-Score in general. Overall, the AdaBoost model performed better compared to the TV-Cox model.

Table 14 presents the model performance of the AdaBoost and TV-Cox models for the 5-Year programs. The confusion matrix for the AdaBoost Dropout Time model is presented in

Table A.2b. Compared to the 4-Year program TV-Cox model, the precision of the TV-Cox model for the 5-Year programs is much lower. However, the recall values are much higher in the 5-Year programs. The test results for the AdaBoost model show that even though all semesters have a low MSE, the model has a better score for the other performance metrics for semesters with relatively higher number of dropouts compared to semesters that have lower dropout counts. Similar to the 4-Year program models, AdaBoost had better performance compared to the TV-Cox model. However, there were some semesters where the TV-Cox model had a higher recall value.

Table 14. Comparison of Model Performance for the 5-Year Programs

Sem	Precision		Recall		F-Score		MSE	
	AB	CM	AB	CM	AB	CM	AB	CM
2	0.977	0.088	0.977	0.403	0.977	0.144	0.010	0.293
3	1	0.084	0.769	0.427	0.870	0.140	0.008	0.296
4	0.906	0.180	0.928	0.530	0.917	0.269	0.036	0.296
5	0.647	0.177	0.550	0.532	0.595	0.266	0.039	0.298
6	0.787	0.214	0.857	0.554	0.821	0.309	0.054	0.292
7	0.615	0.228	0.308	0.563	0.410	0.325	0.060	0.294
8	0.609	0.241	0.683	0.565	0.644	0.338	0.080	0.294
9	0.722	0.246	0.684	0.561	0.703	0.342	0.029	0.295
10	0.556	0.245	0.714	0.613	0.625	0.350	0.047	0.296
11	0.333	0.248	0.364	0.615	0.348	0.353	0.039	0.297
12	0.455	0.259	0.500	0.620	0.476	0.365	0.029	0.296

Figures 3 and 4 present the PR curves for the AdaBoost and TV-Cox models. Comparing the PR curves, the performance of the AdaBoost model is much higher compared to the TV-Cox model in classifying drop time.

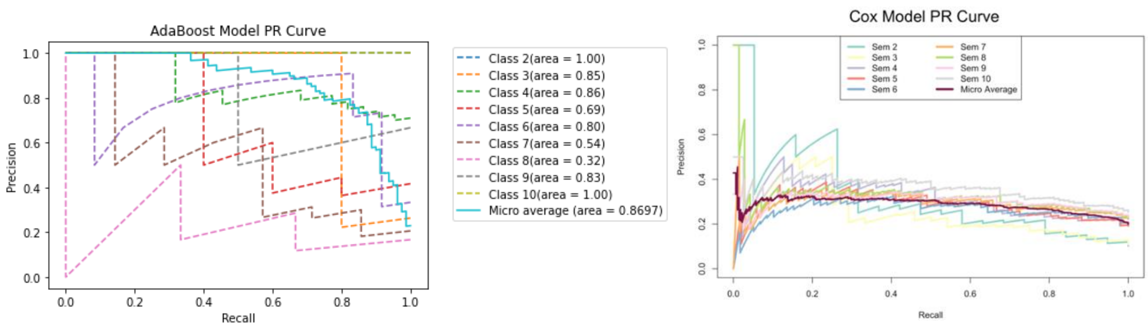

Figure 3. PR Curves for the 4-Year Program

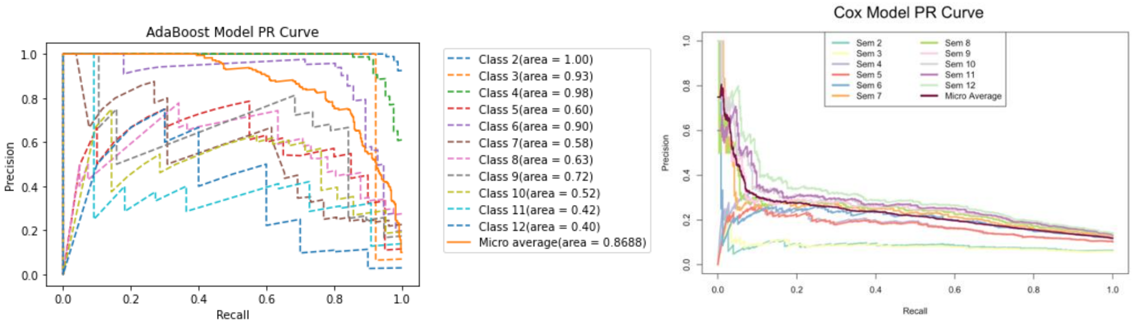

Figure 4. PR Curves for the 5-Year Program

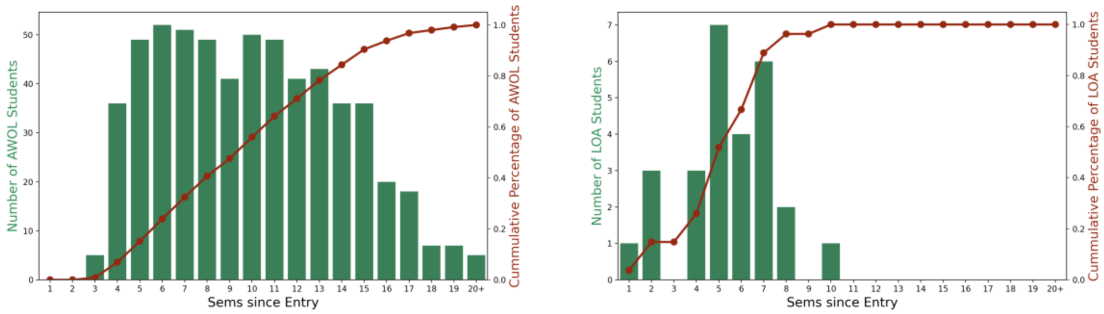

Figure 5. AWOL and LOA Data for the 4-Year Program

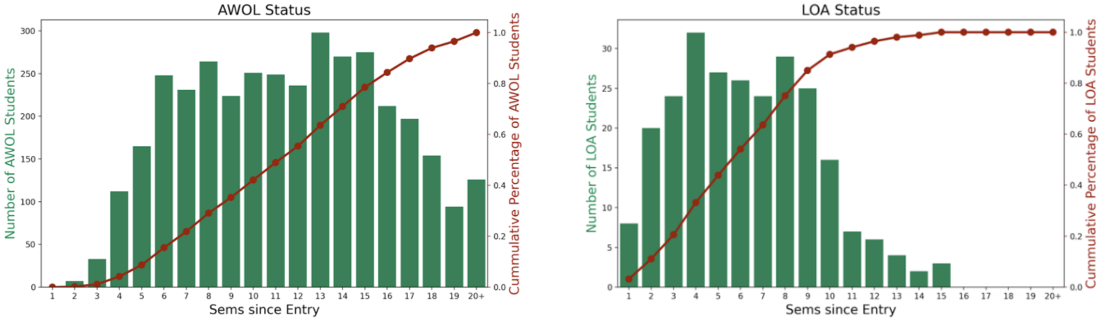

Figure 6. AWOL and LOA Data for the 5-Year Programs

3.4 Descriptive Analysis

Absence Without Leave and Leave of Absence. Figures 5 and 6 present the LOA and AWOL data per semester of the 4-Year and 5-Year programs. For the 4-Year program, 59.33% of students who AWOL do so after the on-time graduation at semester 8. On the other hand, for students in the 4-Year program 96.3% of students who LOA do so by the semester of on-time graduation. Similarly, for the 5-Year programs, 57.9% of students who AWOL do so after semester 10 and 91.3% of students who LOA do so before semester 10.

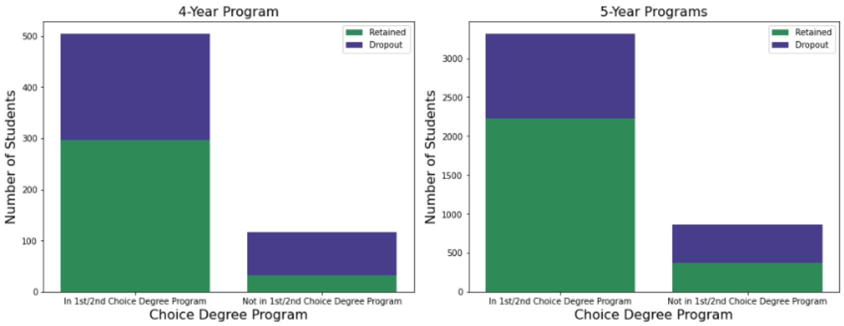

Figure 7. Admission into 1st/2nd Choice Degree Program

Accepted In 1st/2nd Degree Choice Program. Figure 7 shows that for both the 4-Year and 5-Year programs, students who were admitted into their 1st or 2nd choice degree program were more likely to retain and finish their initial degree programs. Students who did not get into their 1st choice nor their 2nd choice degree program had a higher tendency to drop out.

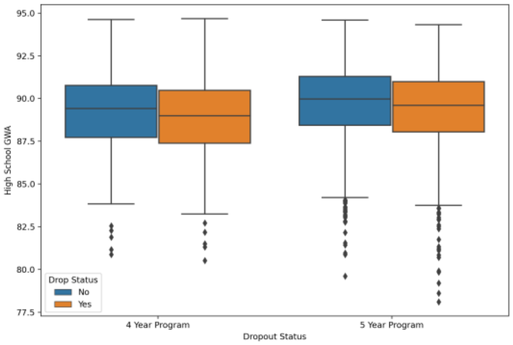

Figure 8. High School GWA

High School GWA. Figure 8 presents the HS-GWA of all freshmen students in the college for academic years 2009-2013. Higher high school GWA slightly reduces the chance of dropout.

Age of Entry. Figure 9 presents the histogram on the age of entry into the university for the 4 and 5-Year programs. For both programs, most students enter the college at age 17. Dropout percentage is similar for all age groups except for age 20. In the 5-Year programs, there were only 2 students who entered at 20 years old and both were considered dropout.

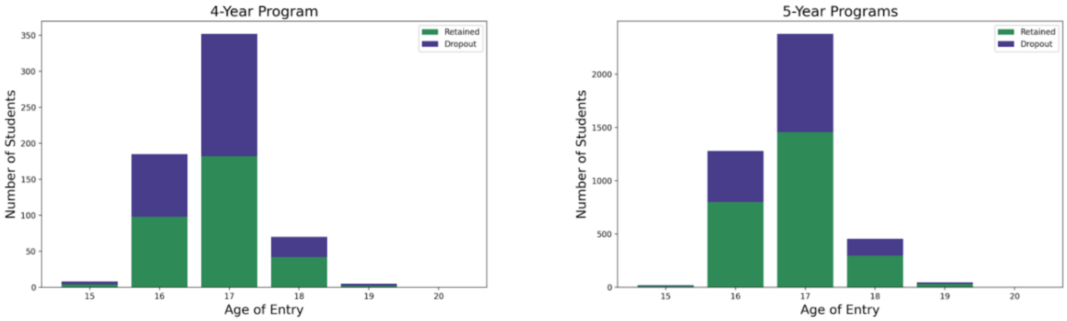

Figure 9. Age of Entry

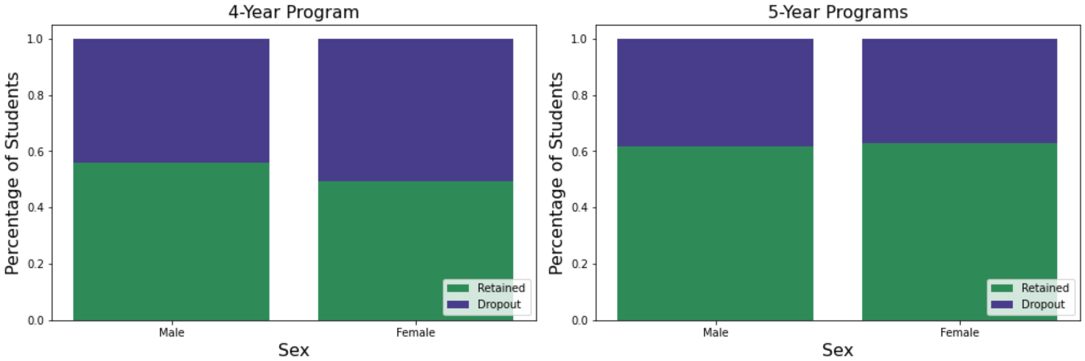

Figure 10. Sex

Sex. Over half of the students of the college in both the 4-Year and 5-Year programs are male. It can be seen in Figure 10 that even with the significant difference in the frequency of male and females in the 4-Year and 5-Year programs, females still have a higher record of dropping out than males in the 4-Year program and the margin between dropouts between male and females in the 5-Year program is small even if there were significantly more males than females.

Home Province. Students who had permanent addresses within Metro Manila were more likely to retain their course since they're closer to the college and students from provinces had more likelihood to drop out compared to students that had residences within Metro Manila.

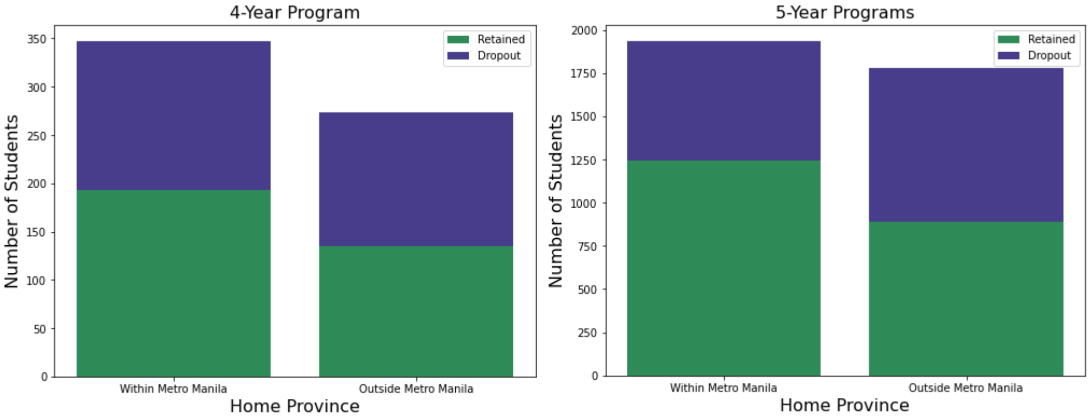

Figure 11. Home Province

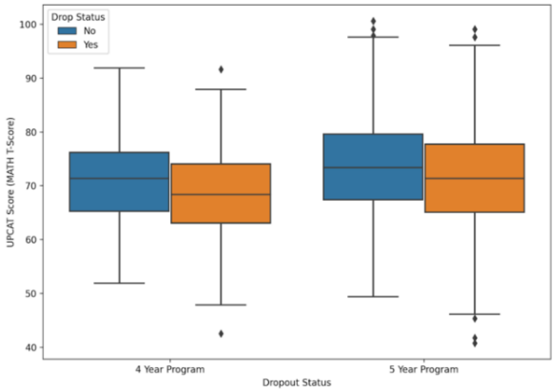

Figure 12. UPCAT Math T-Score

UPCAT Math T-Score. Having a higher math score decreases the chance of dropout for both 4-Year and 5-Year programs.

IV. CONCLUSION AND RECOMMENDATION

4.1 Conclusion

The graduation rates in the UPD COE are quite low compared to other schools, indicating that it is important to investigate the dropout rates of students as well. In order to address this problem, machine learning was used to model student dropout and identify factors that affect a student's risk of dropping out.

The input features of the prediction models for the 4-Year and 5-Year programs were selected using the Pearson's Correlation and Cramer's V tests for continuous and categorical features, respectively. All continuous factors come from pre-enrollment data. After performing the Cramer's V test, the features left contained pre-enrollment, demographic, and semestral data.

High school GWA, number of semesters AWOL, and being in the student's first or second choice degree program were seen to influence dropout for both the 4-Year and 5-Year program of both models. For the 4-Year program, AGE and being a female student were shown to be significant to student dropout risk. In addition, home province, number of semesters LOA, and UPCAT Math T Score were also significant to dropout in the 5-Year programs.

Two models were used for this study, the AdaBoost model and the TV-Cox model. The AdaBoost model had two variations to predict student dropout and dropout time separately, with both models using Decision Trees as base learners. For the TV-Cox model, the optimal cutoff point that maximizes the Youden index was used to classify whether the predicted survival rate is dropout or retained. This cutoff point was adjusted so that the TV-Cox model would have an accuracy of at least 70%. Between the two models, AdaBoost performed better in predicting student dropout and drop time. However, for some semesters where the number of student dropouts are low, the TV-Cox model and the AdaBoost model had similar recall values.

4.2 Recommendation

The performance of the models, especially for the TV-Cox model, can be improved by using a larger and more balanced dataset, especially for the 4-Year program. In addition, more post-enrollment data can help improve accuracy. Features such as semestral GWA, number of units passed and taken per semester, and other academic-related data might give a clearer idea on student performance per semester.

For the analysis of the dataset, additional tests or methods can be used to have a clearer insight on how these features affect dropout. Additionally, it may be good to use the effects of these features from the dataset with respect to dropout and compare it to the hazard ratios from the TV-Cox model.

V. ACKNOWLEDGEMENT

We would like to express our sincerest gratitude to our advisers, Dr. Jhoanna Rhodette Pedrasa and Sir Darvy Ong, for guiding us throughout this study. We wouldn’t be able to accomplish this project without their help and insight.

REFERENCES

[1] Massachusetts Institute of Technology Graduation & Retention. Retrieved from https://www.collegefactual.com/colleges/massachusetts-institute-of-technology/academic-life/graduation-and-retention/ on 24 Nov 2020.

[2] Bernes J, Schneider K, Gortz S, Oster S, Burghoff J. Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. Journal of Educational Data Mining. 11(3):1-41.

[3] Ameri S, Fard M, Chinnam R, Reddy C. 2016. Survival Analysis Based Framework for Early Prediction of Student Dropouts. 25th ACM International on Conference on Information and Knowledge Management; Indianapolis, IN, USA. ACM. P. 903-912. doi: https://doi.org/10.1145/2983323.2983351

[4] Chen Y, Johri A, Rangwala H. 2018. Running out of STEM: A Comparative Study across STEM Majors of College Students at-Risk of Dropping out Early . 8th International Conference on Learning Analytics and Knowledge; Sydney, Australia. ACM. doi: https://doi.org/10.1145/3170358.3170410

[5] Maximum Residence. Retrieved from https://our.upd.edu.ph/files/acadinfo/MAXIMUM/RESIDENCE/_Undergraduate.pdf on 31 Mar 2021.

[6] A Primer on the UP Socialized Tuition System. Retrieved from https://ovpaa.up.edu.ph/wp-content/uploads/2017/03/A-Primer-on-UPs-Socialized-Tuition-System-27Feb17-2.pdf on 22 May 2021.

[7] Mukaka MM. September 2012. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. Malawi Medical Journal. 24(3): 69-71.

[8] Prematunga RK. August 2012. Correlational Analysis. Australian Critical Care. 25(3):195-199.

[9] Iversen GR, Gergen M. 1997. Statistics: The Conceptual Approach. New York (NY): Springer New York. p. 345-360

[10] Kleinbaum DG, Klein M. 2012. Survival Analysis: A Self-learning Text. New York, NY: Springer New York.

[11] Schapire R. 2013. Explaining AdaBoost. In: Schölkopf B., Luo Z., Vovk V. (eds) Empirical Inference. Berlin: Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_5

[12] Pedregosa F, et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 12(85):2825-2830.

[13] Etikan I, Babatope O. Survival Analysis: A Major Decision Technique in Healthcare Practices. IJRSM Human Journals. 8(4).

[14] Faraggi D, Reiser B. Estimation of the Youden Index and its Associated Cutoff Point. Biometric Journal 47 (4): p. 458-472.

[15] Tharwat A. Classification assessment methods. Applied Computing and Informatics. 17 (1).

[16] Wackerly D, Mendenhall W III, Schaeffer R. 2014. Mathematical Statistics with Applications. 7th ed. Belmont (CA): Cengage Learning. p. 392-393

APPENDIX A

AdaBoost Confusion Matrices
A.1 Dropout Status Model

Table A.1: Dropout Status Model Confusion Matrices for 4-Year and 5-Year Programs

(a) 4-Year Program

	Predicted	
True	0	1
0	104	2
1	6	74

(b) 5-Year Program

	Predicted	
True	0	1
0	851	15
1	16	371

Table A.2: Dropout Time Model Confusion Matrices for 4-Year and 5-Year Programs

(a) 4-Year Program

	4-Year Program								
	Predicted								
True	2	3	4	5	6	7	8	9	10
2	**21**	0	0	0	0	0	0	0	0
3	1	**3**	1	0	0	0	0	0	0
4	0	0	**18**	2	2	0	0	0	0
5	0	0	2	**2**	1	0	0	0	0
6	0	0	1	0	**11**	0	0	0	0
7	0	0	2	0	1	**4**	0	0	0
8	0	0	0	0	0	2	**1**	0	0
9	0	0	0	0	0	0	0	**2**	0
10	0	0	0	0	0	0	1	1	**1**

(b) 5-Year Program

	Predicted										
True	2	3	4	5	6	7	8	9	10	11	12
2	**84**	0	0	2	0	0	0	0	0	0	0
3	0	**10**	3	0	0	0	0	0	0	0	0
4	1	0	**77**	1	2	0	1	1	0	0	0
5	0	0	2	**11**	5	0	0	2	0	0	0
6	0	0	0	1	**48**	3	2	1	0	1	0
7	1	0	2	2	1	**8**	10	0	2	0	0
8	0	0	1	0	3	0	**28**	0	5	3	1
9	0	0	0	0	1	0	1	**13**	2	1	1
10	0	0	0	0	0	2	2	1	**15**	0	1
11	0	0	0	0	0	0	2	0	2	**4**	3
12	0	0	0	0	1	0	0	0	1	3	**5**

APPENDIX B

AdaBoost Significant Features
B.1 Dropout Status Model

Table B.1: Dropout Status Model Feature Importances for 4-Year and 5-Year Programs

4-Year Program			5-Year Program		
Rank	Feature	Gini Importance	Rank	Feature	Gini Importance
1	No. of Sems AWOL	0.320000	1	No. of Sems AWOL	0.146667
2	STFAP/STS Bracket 9th Sem	0.180000	2	STFAP/STS Bracket 12th Sem	0.106667
3	STFAP/STS Bracket 10th Sem	0.160000	3	UPCAT Score (Reading Comprehension T-Score)	0.080000
4	UPCAT Score (SCIENCE T-Score)	0.100000	4	UPCAT Score (Language Profiency T-Score)	0.080000
5	STFAP/STS Bracket 6th Sem	0.080000	5	STFAP/STS Bracket 9th Sem	0.053333
6	High School GWA	0.040000	6	STFAP/STS Bracket 4th Sem	0.053333
7	STFAP/STS Bracket 1st Sem	0.040000	7	STFAP/STS Bracket 6th Sem	0.040000
8	Age of Entry to UP	0.020000	8	STFAP/STS Bracket 10th Sem	0.040000
9	Dorm Status	0.020000	9	STFAP/STS Bracket 8th Sem	0.040000
10	UPCAT Score (Reading Comprehension T-Score)	0.020000	10	UPCAT Score (Math T-Score)	0.040000

B.2 Dropout Time Model

Table B.2: Dropout Time Model Feature Importances for 4-Year and 5-Year Program

4-Year Program			5-Year Program		
Rank	Feature	Gini Importance	Rank	Feature	Gini Importance
1	STFAP/STS Bracket 3rd Sem	0.104093	1	No. of Sems AWOL	0.251581
2	UPCAT Score (Reading Comprehension T-Score)	0.093773	2	STFAP/STS Bracket 5th Sem	0.103189
3	UPCAT Score (MATH T-Score)	0.072280	3	STFAP/STS Bracket 3rd Sem	0.091376
4	No. of Sems AWOL	0.070920	4	UPCAT Score (Language Profiency T-Score)	0.083494
5	UPCAT Score (SCIENCE T-Score)	0.068648	5	STFAP/STS Bracket 6th Sem	0.082991
6	High School GWA	0.062698	6	STFAP/STS Bracket 4th Sem	0.074161
7	STFAP/STS Bracket 7th Sem	0.057039	7	STFAP/STS Bracket 7th Sem	0.053323
8	UPCAT Score (Language Proficiency T-Score)	0.056246	8	STFAP/STS Bracket 9th Sem	0.034862
9	STFAP/STS Bracket 6th Sem	0.044995	9	High School GWA	0.031990
10	STFAP/STS Bracket 4th Sem	0.040277	10	UPCAT Score (Math T-Score)	0.030009