Predicting the Academic Performance of the Engineering Students Using Decision Trees

Editha Rivera Jorda

Abstract— **Data mining is an integral part of knowledge discovery in database and a process that converts raw data into useful information. Once applied in education, it is called Educational Data Mining (EDM). EDM is a field of scientific inquiry for the developments of method to discover unique kind of data in educational settings, and using this method to understand better the students and their learning environment. One of the current popular methods in EDM is prediction. Prediction is used to detect student's behavior, and predicting or understanding student outcome.**

Based on the data gathered, many engineering students in Technological University of the Philippines were either dropouts or dismissed from the engineering program they enrolled in. The dismissal or dropping out of students resulted to wastage of the scarce resources of the government and deprived the opportunity of the other students.

As such, the paper aimed to develop and validate a predictive model to serve as a framework in predicting the academic performance of the engineering students in Technological University of the Philippines Manila (TUPM) based on Mathematics and Physics courses towards retention policy and identify academically at-risk engineering students for early intervention.

The research design of the paper is descriptive-quantitative. The data of the engineering students from school year 2008 - 2015 were gathered from the Electronics Registration System of TUPM. It contained the students' final grades in College Algebra, Plane and Spherical Trigonometry, Solid Mensuration, Advance Algebra, Analytic Geometry, Differential and Integral Calculus, and Physics 1 and 2. C5.0 and Chi-squared Automatic Interaction Detection (CHAID), two of the decision tree algorithms provided by IBM SPSS Modeler were used to develop and validate the model and t-test was used if the two models were significantly different. C5.0 suited best for the model based on accuracy and 10-fold cross validation for identifying students who were likely to be retained in the program and those who were academically at-risk. Lastly, the two models were significantly different based on the level of accuracy of prediction about the academic performance of the engineering students in TUPM.

Keywords— **Data mining, Educational Data Mining, Decision Tree, C5.0.**

I. Introduction

Among the many sectors of our society, the education sector receives the biggest allocation from the government to alleviate poverty. Unfortunately, many students dropped out or were dismissed from the program just after a few semesters. At Technological University of the Philippines-Manila (TUPM), the same thing happens especially in the engineering program. Hence, [32] argued that SUCs must use efficiently their resources to achieve their intended purpose. One possibility is for TUPM to predict with high accuracy the right students in the engineering program.

One strategy is the use of Mathematical Modeling in predicting the academic performance of engineering students. It forms an abstract model of mathematical language to describe a complex behavior expressed in differential equation and partial differential equations or sometimes it can be conveniently expressed as a set of rules [34]. Its capability is enhanced more using a software tool such as an IBM SPSS Modeler that can store large volume of data and extract intelligent information about the performance of the students to support future decision making.

The aforementioned process is called Data Mining that determines valid, useful and understandable patterns in data on the academic performance of the students by applying pattern recognition (PR) and machine learning principles in different data sets. In education, the technique is called Educational Data Mining (EDM) wherein the data from education is explored [45]. The data may come from traditional face-to-face classroom environments, educational software, online courseware or summative/high stakes tests. One of the popular methods in EDM is Prediction. The Prediction Model attempts to determine what the output value would be in context where it is not desirable to directly obtain a label for that construct [28]) Its accuracy is affected by at least two factors: Selection of predictors and the mathematical techniques in developing the predictive model. The accuracy of a predictive model also changes with different predictors.

In the study, the Prediction Model was used because the study aimed to develop a model that can determine a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). It determined the patterns in student retention at TUPM. The Predictor Variables were the final grades in mathematics and physics of the engineering course to evaluate the engineering students' academic performance. The final grades were based on course structure, assessment mark, final exam score and also extracurricular activities.

Based on TUPM Student Manual, a student is on probation once he has acquired two failing grades and he is dismissed once he has acquired three failing grades in any subject. The student often acquired these deficiencies in mathematics and Physics

Editha Rivera Jorda, Centro Escolar University, Technological University of the Philippines, Manila, Philippines

subjects. Consequently, an engineering student is required to have a strong mathematical knowledge to keep him motivated to progress in the engineering program [18]. Without it, the engineering student may eventually drop out or be dismissed from the Program.

As such, the purpose of the study is to develop and validate a Mathematical Model to serve as a framework in predicting the academic performance of the engineering students toward an improved retention policy in TUPM. Specifically, it described how the predictive model was developed using the four degree programs offered in TUPM namely: Civil, Electrical, Electronics, and Mechanical; the nine subjects of the Program such as: College Algebra, Trigonometry, Advanced Algebra, Analytic Geometry, Solid Mensuration, Differential and Integral Calculus, Physics 1 and 2 as predictors; and the Decision Tree as its Data Mining Technique and what predicting model was utilized.

The predicted models were based on the quantitative data of students' academic performance from school year 2008 – 2015. The criteria used to evaluate and compare the models were also defined. It is hoped that the findings of the study could reduce the big number of students who dropped out, on probation or dismissed from the College of Engineering (COE) at TUPM.

In the study on students' failure in their courses, students who have a good understanding of the content being taught are more motivated and have a positive attitude, so they have a greater chance of doing well in their schoolwork [36] Furthermore, students knew that they need support from their college and instructors to keep them on track. This means that there is a need for a university to develop a comprehensive strategy to determine the academic readiness of the engineering students. Once a university has identified it, there is a chance that it can prepare a remedial plan for engineering students who are at risk and bring them back to the mainstream program. However, considering the huge volume of data about the students, traditional methods of prediction are not enough. They should be enhanced with other techniques such as the use of a Mathematical Model.

A Mathematical Model is a quantitative model that uses a mathematical language. One of which is the Knowledge Discovery in Database (KDD) that converts big volume of data to simplify and extract relevant information that can guide the decision-making process of school administrators [22]; [24]; and [23].

Recently, according to [21], KDD is a process of iterative sequence with the following steps: 1) data cleaning, 2) Data integration, 3) Data Selection, 4) Data transformation 5) Data mining, 6) Pattern evaluation, and 7) Knowledge presentation. Based on [21], [5] used it to generate licensure examination performance models using PART and JRips classifiers of WEKA. Likewise, [10] adapted the steps of [21] for extracting knowledge from data to describe students' performance in end semester examination. Hence, KDD process applies to many issues related to the students with a high level of accuracy.

One integral part of KDD is Data Mining which is a process that converts raw data into useful information [23]. In education, it is called Educational Data Mining (EDM) which is a scientific inquiry for the development of methods to discover unique kinds of data in educational settings, and using these methods to understand better the students and their learning environment [28]. It includes traditional face-to-face classroom environments, educational software, online courseware or summative/high stakes tests [45].

One popular methods of EDM is Prediction. It aims to develop a model to infer a single aspect of a data or predicted variable from some combinations of other aspects of the data. It is used to model continuous–valued functions, i.e., predict unknown or missing values. It is also used to detect students' behavior, predicting or understanding students' educational outcome [11]; [43] and [17]

One of the three types of prediction is classification that predicts variable in binary or nominal categories. Some of the classification methods include Decision Tree, Regression, Neural Networks, Support Vector Machine and Bayesian network. A classification model based on the technique of decision tree was applied by [1]. This technique provided a guideline that help students and school management to choose the right track of study for a student. On the other hand, [15] compared the Bayesian network classifiers to predict the student's academic performance to help in identifying the drop outs and students who need special attention and allow the teacher to provide appropriate counseling/ advising. Likewise, [16] investigated the application of Bayes Network to predict causal relationship in a dataset that captures several demographic and academic features of a group of students from a four-year university.

Each technique employs a learning algorithm to identify the model that best fits the relationship between the attribute set and class label of the input data. Thus, a key objective of the learning algorithm is to build models that accurately predict the class labels of previously unknown records, that is, models with good generalization capability.

[42] proposed a framework to predict the students' academic performance using the Decision tree, Naïve Bayes, and Rule Based classification techniques. The experiment revealed that the Rule Based technique is the best model with a high accuracy value of 71.3%.

Another paper by [39] tried to find out if there were patterns in the available data that could be useful to predict the students' performance using decision tree (C4.5, J48), Bayesian Classifiers (Naïve Bayes and Bayes Net), A Nearest Neighbour algorithm and Two Rule Learners (OneR and JRip). The results revealed that decision tree classifier (J48) performs best with a high accuracy, followed by the rule learner (JRip). However, all tested classifiers had an overall accuracy below 70% which means that the error rate was high and the predictions were not reliable.

A. Decision Tree

A decision tree is a flowchart tree structure wherein each

internal (non-leaf node) denotes a test on an attribute. Each branch represents an outcome of the test, and each leaf node (or terminal node) holds a label. The top node in a tree is the root node [21] For the decision tree used for classification, a given tuple, X, for which the associated class label is unknown , the attributes values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Thus, decision tree can easily be converted to classification rules. Some of the decision classifiers are ID3 (J48), C4.5, C5.0, Classification and Regression Tree (C&RT), Chi-Squared Automatic Interaction Detection (CHAID) and Quick, Unbiased, Efficient Statistical tree (QUEST), Random Forest.

Based on the definition of [31], a decision tree model allows developing classification systems that predicts or classify future observations based on the set of decision rules. This approach is also known as rule induction. It has several advantages such as: the reasoning behind the model is clearly evident when browsing the tree and the process automatically includes in its rule, the attributes that are really important in making a decision. Attributes that do not contribute to decision making is ignored.

According to [23] the Decision trees classifiers are popular because the construction of decision trees classifiers do not require any domain knowledge or parameter setting, thus it is appropriate for exploratory knowledge. The Decision tree also handles multidimensional data. Its representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The study of [43] indicated that the results of decision tree and rule induction are important because the classification model given by these two methods is user friendly as it represents rules which are easily interpretable by humans and useful in making policies.

[19] concluded in their experiment that simple classifiers such as decision trees (CART and J48) give a useful result with accuracies between 75% and 80% that is hard to beat with other sophisticated models. The study of [39] revealed that the decision tree classifier (J48) performed best with the highest overall accuracy for predicting student performance.

B. Students' Academic Performance

The Student Academic Performance (SAP) helps Higher Education Institutions (HEIs) to study what attributes are important for prediction, as well as extract the hidden information in students' data [28]. What EDM does is to predict or describe the significant patterns of the many data about the academic performance of the engineering students. In the Predictive Task, it determines the particular value of a particular attribute. The attribute to be predicted is called the target or dependent variable, while the attribute used for making the prediction is known as the explanatory or independent variable [23]. For the Predictive Task, the EDM technique used is often the Classification Technique, because it finds a model (or function) that describes and distinguishes data classes or concepts about the academic performance of the engineering students unlike the other techniques. Under the Classification Technique, there are various models which are as follows: 1) Ruled-based classifier (IF-THEN), 2) Decision Tree, 3) Bayes Classification (Naïve Bayesian), 4) Neural network, 5) K-Nearest Neighbor, and 6) Support Vector Machine.

Various models were generated to predict the performance of the engineering students based on the studies of the following authors: [3]; [7]; [38]; [41]; [35]; and [37]. They differ in specific attributes in predicting performance of engineering students

However, one its the key features is that the process can be repeated, managed and measured to increase the level of accuracy of predicting the academic performance of the engineering students, and at the same time false data mining results are checked and validated.

The sequence seems impossible to do using the traditional method. However, EDM can do the sequence, manage the data, measure the results, and repeat the sequence over and over again, because it is technology driven that combines the traditional data analysis methods with sophisticated algorithms to process large volume of data. It has already been applied in many big businesses and has produced many positive results. According to [23], Data Mining is proficient in the business industry because it is built upon methodology and algorithms. Many studies have already applied it in education and it produced similar positive results.

Most studies used the Classification Technique but differ on its algorithm and software that ranges from ID3 and J48, Simple CART/and software WEKA [3]; C4.5 and ID3, and software WEKA [7] ;k-NN, IBk, decision trees, naïve Bayes and Rapidminer software, [38]; C4.5, Naïve Bayes, K-NN, Support vector machine, neural network and Rapid miner version 6.1 [37]

They also differ in predicted model ranging from J48, ID3 and C4.5, Naïve Bayes, Radial Basis Function (RBF) network, and Support Vector Machine (SVM), and Neural Network.

Despite the specific differences in the studies, all the studies concluded that they were able to achieve the specific goals of their study namely: Predicting the students' performance using the decision tree algorithm applied on engineering students' past performance to generate the model [3]; Assisting the low academic achievers in engineering [7]; Obtaining a model to predict new students' academic performance taking into account socio-demographic and academic variables [38]; Developing a validated set of mathematical models to predict student academic performance in engineering dynamics [41]; Predicting students' grade in three major courses [35] and Predicting the performance of the engineering students in the core engineering courses[37].

As such, EDM has all the potentials to predict the academic performance of the engineering students that can work well with the traditional method, because it has a wide range of applications of the real world problem in education.

C. Retention Policy

Based on the aforementioned discussion, EDM is an important tool to a University, since it has to achieve its vision,

mission, goals and objectives, and sustain its quality education. And it could not do it without a clear retention policy wherein every student is provided with a learning environment that would give all types of students an equal opportunity to develop their full potential and guide them to the right path of their career.

A retention policy is a measure of the quality of a University's overall product, retention and graduation rates. Many retention experts claimed that a University's ability to demonstrate a student success and its ability to attract and recruit new students are intertwined [13] and [33].

In any form of learning process, other students will naturally excel while others would lag behind, so the university needs to have a good retention policy that is student-centered. Students who might be at risk are properly assisted and given a chance to cope with their academic requirements.

A good retention practice should be based on intrusive and intentional interventions that are focused on student engagement and intellectual involvement; and it should emphasize general quality enhancements of educational programs and services. A good retention rate is essentially the bi-product of improved quality of student life and learning on college campuses [9]. Many researches confirmed that Universities with higher retention outcomes conduct sound educational practices [13].

One good retention practice is for the students to know their chances of finishing their respective academic program and the areas they need to improve before they enroll in their respective academic program. A student is more likely to persist and graduate in settings that provide frequent and early feedback about his possible performance. The use of early warning systems by a University created an impact in providing a student the much needed information about his performance, so he can adjust his performance in order to persist and finish his program.

According to Tinto (2000), a student who learns is the student who stays. A student who is actively involved in learning, that is, who spends more time on task with others is more likely to learn, and in turn more likely to stay (Tinto, 1997).

Henceforth, a predictive model is a valuable tool for a University, since from the data gathered from interesting patterns, it can design and develop management and classroom practices that will help the University and students persist and finish their respective academic program.

II. Methodology

A. 2.0 Data Collection

The subject of the study is composed of engineering students who are officially enrolled in Mechanical, Civil, Electrical, Electronics and Communication Engineering at TUPM who were not dismissed, dropped out, or on probation before their 3rd year status in the program. The data of the engineering students from school year 2008 - 2016 were collected from the ERS of TUPM that contained their final grades in College Algebra, Plane and Spherical Trigonometry, Solid Mensuration, Analytic Geometry, Advance Algebra, Differential and Integral Calculus, Physics I and 2. A total of 3 765 students qualified in the criteria, broken down as follows:

Table I: Respondents' Profile per Course

Course	Number of Students before their 3rd year
CE	1042
ECE	1144
EE	725
ME	854
Total	**3765**

B. Predictive Model Development

The development of the predictive model was adapted from Han et. al (2011) and Ahmed et. al (2015). The stages involved in developing a predictive model were as follows: 1) Data Collection, 2) Data Transformation, and 3) Pattern Extraction. Figure 1 illustrates the Input – Process - Output (IPO) in developing the predictive model.

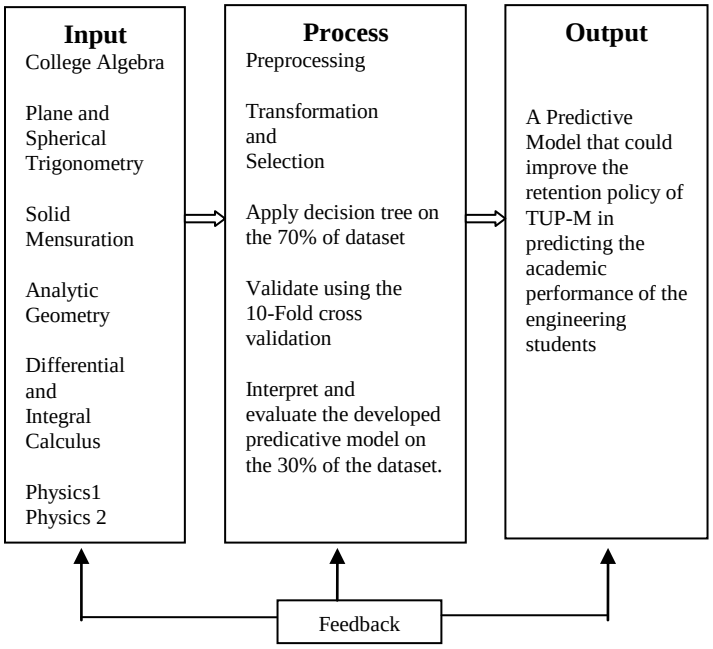


Fig.1 Development of a Predictive Model

Based on figure 1, the application of Data Mining in education is an iterative cycle of hypothesis formation, testing, and refinement that consists of several steps until the proper model with a high level of accuracy of prediction is developed.

First, in the Input Stage, the academic performance of the engineering students on the following subjects were gathered namely: Algebra, Plane and Spherical Trigonometry, Advanced Algebra, Analytic Geometry, Solid Mensuration, Differential Calculus, Integral Calculus, and Physics 1 & 2.

Second, in the Process Stage, the Predictive task was performed wherein the gathered data were transformed and the interesting patters were extracted (Han, et. al., 2011 and Ahmad, et. al., 2015). Data transformation covering the final grades of the engineering students in Mathematics and Physics were selected. Data was cleaned by removing engineering students who dropped out, on probation, or dismissed before their 3rd year status in the program. The cleaned dataset were

encoded and stored in Microsoft excel. While in Pattern extraction, using the commercial software tool IBM SPSS Modeler Version 18.0, the data from external source such as Microsoft was extracted and read; Fill out the table that specified field properties such as measurement level (the type of data that the field contains), a category that indicated the type of data in the field such as nominal, ordinal, and continuous and the role of each field as a target or input in modeling. Table II shows for each field, the type node which specifies a role to indicate the part that each field plays in modeling.

Based on table II, the columns were divided as follows: Field of course code, description of each course, measurement level such as continuous and nominal, value for each field, and its role is set to input or target. The input fields are also known as predictors or whose values were used by the modeling algorithm to predict the value of the target field while target indicates whether or not the engineering students were retained or not in the degree program.

The dataset was divided into training set and test set. Two-thirds of the data set belonged to the training set and used to build the model while one-third of the dataset belonged to the test set to evaluate the model.

C. Training Set

Two-third of the dataset was used as training set. The training set was mined using the decision tree models namely: C5.0, Chi-squared Automatic Interaction Detection (CHAID). The top two in decision tree models were used based on the auto classifier, a built-in classifier in the software that rank the models based on their overall accuracy. Each model indicated its prediction importance, validation, decision tree, and the mined pattern namely: coincidence matrix, data, analysis and graph. The Decision Tree Models implemented in IBM SPSS consists of definition, requirements, strengths and the methods used for splitting. Table III shows the two decision tree models.

TABLE II. SETTING THE TARGET AND INPUT FIELDS WITH THE TYPE NODE

Field	Description	Measurement	Value	Role
Math 1	College Algebra	continuous	[1.00 – 5.00]	Input
Math 2	Plane and Spherical Trigonometry	continuous	[1.00 – 5.00]	Input
Math 3	Solid Mensuration	continuous	[1.00 – 5.00]	Input
Math 4	Analytic Geometry	continuous	[1.00 – 5.00]	Input
Math 5	Differential Calculus	continuous	[1.00 – 5.00]	Input
Math 6	Integral Calculus	continuous	[1.00 – 5.00]	Input
Math 10	Advance Algebra	continuous	[1.00 – 5.00]	Input
Physics 1	General Physics	continuous	[1.00 – 5.00]	Input
Physics 2	Fluids, Thermodynamics and Electromagnetism	continuous	[1.00 – 5.00]	Input
Course	CE, ECE, EE, and ME	Nominal		Input
Retain		Nominal		Target

TABLE III. TWO DECISION TREES IMPLEMENTED IN [31]

Model	Definition	Requirements	Strengths	Method used for splitting
C 5.0	The node builds either decision tree or a rule set. The model works by splitting the sample based on the field that provides the maximum information gain at each level.	To train a model, there must be one categorical (nominal or ordinal) target field, and one or more input field/s of any type.	C5.0 model are quite robust in the presence of the problems such as missing data and large number of inputs. It is does not require long time to estimate and tends to be easier to understand than some other type since the rules derived from the model have a very straight forward interpretation	C5.0 used an information theory, the information gains ratio.
Chi-squared Automatic Detection (CHAID)	It is classification method building decision tree by using chi-squared statistics to identify optimal splits. It can generate non-binary tree where some splits have more than two branches	Target and Input fields can be continuous or categorical nodes that can be split into two or more subgroups at each level.	CHAID can generate non-binary. Therefore it tends to create a wider tree than the binary growing methods. It works for all types of inputs	Chi-squared uses a chi-squared test. To calculate chi-squared statistics for categorical target, two methods are used: Pearson, where it provides faster calculation but should be used with caution on small samples. Likehood, more robust than Pearson, but takes longer to calculate.

Predictor importance was used to fine tune the model. Predictor importance chart indicated the significance of each predictor (attributes) in estimating the model and consider ignoring those that matter least.

To validate the model, a 10-fold cross validation was used. Data from school year 2008 to 2013 were partitioned into 10 subsets for 10 - fold cross validation. The initial data were randomly partitioned into ten mutually exclusive subsets or "folds, $D_1, D_2, \ldots, D_{10}$, each of approximately equal size. The data in the Training Set were partitioned again into training set and testing set where cross - validation was performed ten times. In iteration i, partition D_i was reserved as the test set, and the remaining partitions were collectively used to train the model. Thus, in cross validation, each sample was used the same number of times for training and once for testing. For classification, the accuracy estimate was the overall number of correct classifications from the 10 iterations, divided by the total number of tuples in the initial data.

The mined pattern of the generated model included classification (coincidence) matrix for categorical (nominal) targets that showed the pattern of matches between generated (predicted) field and its target field for categorical targets. A table was displayed with rows defined by the actual values and columns defined by the predicted values. Each cell in the table contained the number of true positive that were labeled correctly by the classifier; the number of true negatives, negative tuples that were correctly labeled by the classifier; number of false positive, negative tuples that were incorrectly labeled as positive; the number of false negatives, positive tuples that are mislabeled as negative. A data wherein a list of students who was likely to be retained or not in the predicted data of the built model have the corresponding actual data to match it. To find exactly how many predictions were correct, there was matching of the data of students retained or not retained in the predicted data and the students retained or not retained in the actual data. That is, the analysis allowed to test the model against data for which the actual data was already known. The graphical representation of the classification result was interpreted through a Receiver Operator Characteristic (ROC) chart. ROC curves generally have the shape of cumulative gains chart (it always starts at 0% and end at 100% as the line go from left to right.). If the graph that rises steeply towards the (0, 1) coordinate and levels off then it indicated a good classifier. The classifier with the optimum threshold of classification was located closest to the (0, 1) coordinate, or upper left corner, of the chart. This location represented a high number of instances that were correctly classified as yes (retained), and a low number of instances that were incorrectly classified as no (not retained). Points above the diagonal represented good classification results. Points below the diagonal line represented poor classification results even worse, if the instances were classified at random. ROC chart with points above the diagonal indicated that it has a good classification results.

D. Testing Set

Two-third of the data set were used in the study as training set, while the remaining one-thirds were used as its test set. The test set contained data of students enrolled during school year 2014 – 2015 to estimate the model's accuracy. The goal of modeling with the target field (retained or not retained) was to study the data to which the outcome was known and dentify the patterns of the outcomes that were not known. Evaluation of accuracy was done by comparing the data predicted, whether the student will be retained or not in the degree program with the created model to the actual result.

. Finally, in the output, the developed Predictive Model was the predictor of the academic performance of the engineering students and at the same time an instructor to identify the academically-at-risk engineering students.

Since the EDM is an iterative cycle of hypothesis formation, testing, and refinement, the feedback mechanism provided input to the level of accuracy of the Predictive Model. It determined the desired level of accuracy of the Predictive Model in predicting the academic performance of the engineering students in TUPM.

III. Results and Discussion

A. Building and Validation of Models

Data of students who entered the university from school years 2008 – 2013were entered as training data because they have the actual data whether they were retained or not retained in the degree program. As for the objective of building a model to predict academic performance of the engineering students based on the following:

- Final grades in Math 1, Math 2, Math 3, Math 4, Math 5, Math 6, Math 10, Physics 1 and Physics 2
- Courses (CE, ECE,EE,ME)

The table IV listed the two decision tree(predictive) models according to auto classifier of the IBM SPSS Modeler based on their build time, overall accuracy, number of fields used and area under curve.

Table IV. The Two Predictive Models

Model	Build Time (min)	Overall Accuracy (%)	Number Field Used	Area Under Curve
C 5.0	< 1	86.93	10	0.78
CHAID	< 1	83.68	9	0.81

Based on table IV, both predictive models have less than one minute to build the models. The overall accuracy indicated the percentage of records that is correctly predicted by the model relative to the total number of records. Obviously, C5.0 is slightly higher in percentage 86.93% compared to CHAID with 83.68%. C5.0 ranked model by using 10 input fields in contrast with CHAID. However, CHAID 's area under the curve slightly higher than C 5.0 which indicates the curve lies further above the reference line (IBMSPSS Modeler Version 18).

Table V gave the comparison of the engineering students who were retained and not retained in the engineering program based on their overall accuracy.

Table V. Overall Accuracy of the Two Predictive Models in Training Set

	N	Total (%)
Retain	2408	86.93
Not Retain	362	13.07
Total	2770	100.00

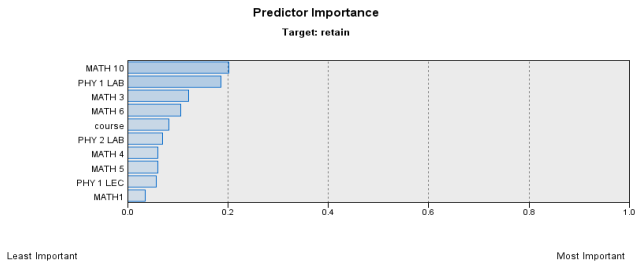
(a) C5.0

	N	Total (%)
Retain	2318	83.68
Not Retain	452	16.32
Total	2770	100.00

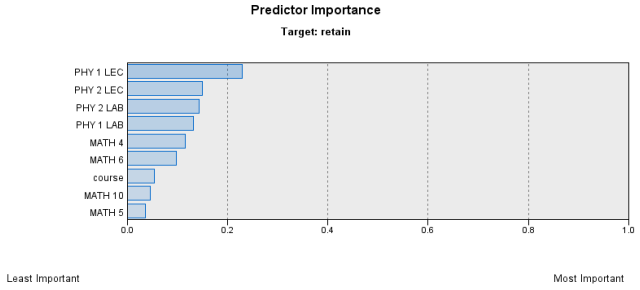
(b) CHAID

Based on table V, there were 2408 students out 2770 or 86.93% who are retained in the engineering program based on C 5.0. On the other hand, 2318 out of 2770 or 83.68% students who are retained in the engineering program based on CHAID.

Figure 2 shown the predictor importance chart which indicates the significant of each predictor in estimating the model.



(a) C5.0



(b) CHAID

Fig. 2(a) and 2(b). The Predictor Importance Chart

Based on the figures 2(a) and 2(b), the predictors of C5.0 and CHAID list down the predictors according to the most important to the least important. In C5.0, the most important is Math 10 while the least important is Math 1 in contrast with the most important, Physics 1Lec and least important, in CHAID. However, predictor importance does not relate to model accuracy. It indicates the importance of each predictor in making a prediction, however, it does not matter whether or not the prediction is correct [31].

To validate the two models, the10-fold cross validation was used. 10-cross validation is used when the training set data were randomly partitioned into 10 mutually exclusive subsets or folds. Table VI showed the accurate and error estimate. The accurate and error estimate is the overall number of correct classifications from the 10-iterations, divided by the total number of tuples in the training data.

TABLE VI. 10-CROSS FOLD VALIDATION

	Model	Mean	N	Standard Deviation
Pair 1	C 5.0 error	13.0670	10	3.83468
	CHAID error	16.3070	10	3.08778
Pair 2	C 5.0 accuracy	86.9330	10	3.83468
	CHAID accuracy	83.6930	10	3.08997

Based on table VI, C5.0 showed the highest (lowest) accuracy (error) in 10-fold cross validation compared to CHAID. Also, C5.0 standard deviation is higher than CHAID which means the accuracy of each fold is nearer to the mean of C5.0.

The model selection is selecting one model over another. Table VII shows the tests of statistical significance whether the difference in accuracy (error) between models is due to chance.

TABLE VII. MODEL SELECTION USING T-TEST

Model	Mean	Standard Deviation	t	p - value	Sig
C 5.0 error CHAID error	3.24000	1. 45383	−7.047	$p = 0.000 < 0.05$	S
C 5.0 accuracy CHAID accuracy	3.25000	1.47541	6.966	$p = 0.000 < 0.05$	S

Based on the table VII, the two models has p-value that is less than 0.05 which means that the mean of C5.0 and CHAID are the same (3.2500); therefore, there is statistically significant difference between the models. Based on overall accuracy (error) accuracy and 10-fold validation, the best predictive model between the two models is C5.0.

B. Evaluation of the Predictive Model

To evaluate the Predictive model, one-third of the data is allocated to the testing set. The evaluation of the performance of the predictive model is based on the account of test records correctly and incorrectly predicted by the model. The table shown the overall accuracy of the predictive model, C 5.0.

TABLE VIII. OVERALL ACCURACY OF C5.0 IN TESTING SET

	N	Total (%)
Retain	843	84.72
Not Retain	152	15.28
Total	995	100.00

Based on table VIII, the overall accuracy of C 5.0 is 84.72% which means that 843 out 995 students were retained and 152 out 995 are not retained in engineering programs. The accuracy of the C 5.0 which is above 80% indicates a good model.

The coincidence matrix analyzes how well the model can recognized tuples of different classes. True positive and true negative indicates that the model is getting things right. On the other hand, False positive and false negative indicates that the model is misleading. The accuracy of the model on a given set is the percentage of test tuples that are correctly classified by the model. The table IX showed the coincidence matrix of the C 5.0 model which explain the difference between the actual value and the predicted value.

TABLE IX. COINCIDENCE MATRIX OF C 5.0

		Predicted Value		
		Retain	Not Retain	Total
Actual Value	Retain	828	53	879
	Not Retain	99	15	114
	Total	927	68	995

Chi-squared = 8.086, df =1, probability = 0.004

Based on table IX, the actual value of the engineering students who retained in the degree program of is 879 while the predicted value is 927. While the actual value of engineering students who are not retained is 114 while the predicted value is 68. Since the p - value is less than 0.05, it indicated the actual value and predicted value is significantly different

The result of the coincidence matrix is validated using Error Rate (ERR) and Accuracy Rate (ACC) as shown in the foregoing tables. ERR is equal to the number of incorrect prediction divided by the total number of dataset. The best and worst accuracy respectively is 0.0 and 1.0. ACC is the equal to the number of correct prediction divided by the total number of prediction. The best and worst error rate respectively is 1.0 and 0.0.

TABLE X. THE ERROR AND ACCURACY RATE OF C 5.0

Rate	Computed Value	Accepted Value
Accuracy	0.8472	$0 < ACC \leq 1.0$
Error	0.1528	$-1.0 < ERR \leq 0.0$

Based on the table X, the computed value for ACC (ERR) is within the acceptable value. Hence, C 5.0 is suited for the predicting the academic performance of the engineering students.

The graphical representation of the C 5.0 can be interpreted through a Receiver Operator Characteristics (ROC) chart. The figure shown is ROC chart with the curve starts at (0, 0) coordinate and ends at the (1, 1).

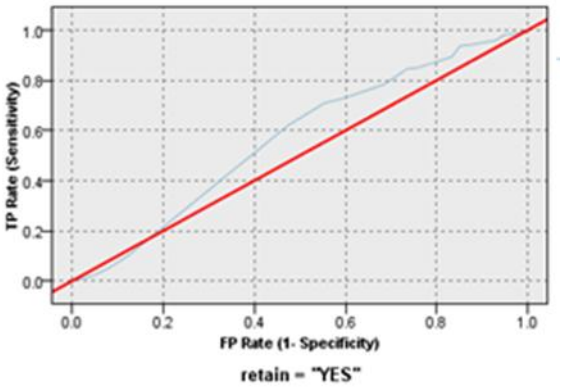


Fig. 3 ROC Chart

Based on fig.3, the vertical axis represents the True Positive (TP) and the horizontal the False Positive (FP). The points above the diagonal line represent the good classification result. Thus, ROC chart of the C5.0 indicates that it has a good classification result.

C. Developed (Predictive) Model

Since C 5.0 has the best accuracy. The predictive model of the decision tree is C 5.0. The decision tree model classifies records and predicts an outcome using a series of decision rules. The decision tree nodes in IBM SPSS Modeler provide access to the tree building algorithm. The algorithm constructed a decision rule by recursively splitting the data into smaller and smaller subgroups. The figure below represents the predictive model in the form of decision tree.

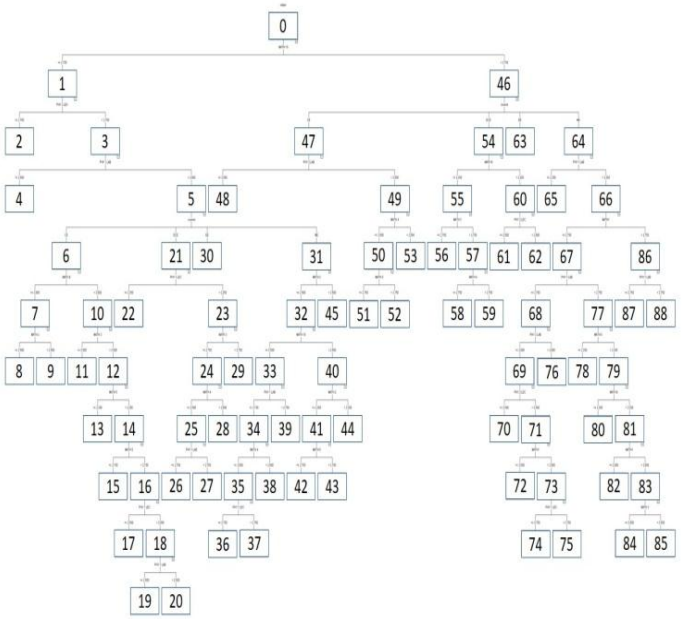
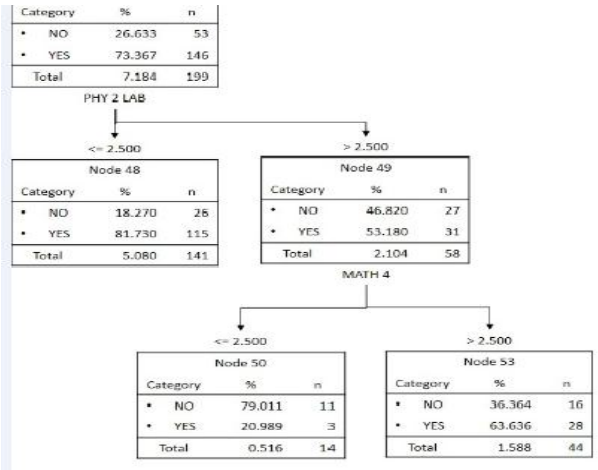


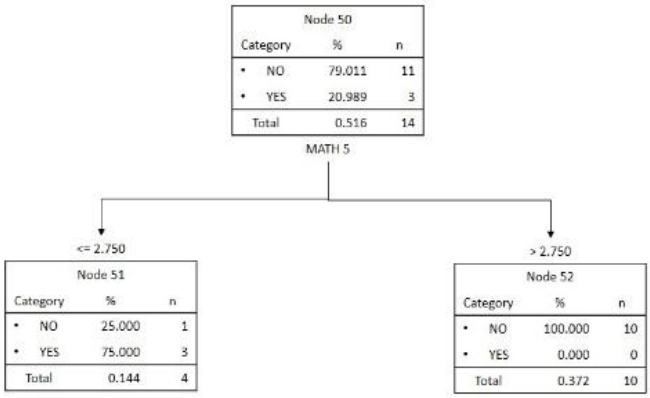
Fig.4 Decision Tree Mapping

Based on fig.4, the decision tree has 86 nodes (trees). The first node, (node 0) represents a summary for all the records in the dataset. The first split, node 1 and node 46 are called child nodes (tree) which indicate recursively splitting data into smaller subgroup. On the other hand, node 2 and node 63 are called terminal nodes which indicate no more splitting occur.

The tree created by the model is large but important rules can be noted as in the following figures:



(a)



(b)

Fig. 5(a) and 5(b) are the extracted tree of the CE course using C 5.0 model

Based on gig. 5(a), in node 47, if the grade of CE students in Math 10 ≥2.750, 73.36% will be retained and 26.633 will not be retained in the program. Hi is a high proportion, there must be some factors responsible for that proportion. So, the course is recursively split into smaller subgroups. The first split is by Phy2 Lab, if the Phy2 Lab ≤2.500, there are increase in engineering students who will be retained in the program from 73. 367% to 81.730. Thus it means students whose grades are greater than or equal to 2.500 will be retained in the program. However, if Phy2 Lab >2.500, the engineering students who will be retained in the program decreased to 53.180%, and an increase of 46.820% for engineering students who will not retain in the program. Therefore, there is a high proportion for academically at-risk students. Because of that high risk factor, the second split will be Math 4. If the grade in Math 4≤2.500, the students who will be retained decreased to 20.909 and if Math 4>2.500, there is an increased from 53.180 to 63.636. Since there is a very high.

IV. Conclusion

Based on the result of the previous section, C5.0 is suited best as the predicted model based on the training set with an overall accuracy of 86.93% compared to CHAID with value of 83.68%. In terms of 10-fold validation, C5.0 has a mean value of 86.9330 and standard deviation of 3.83468 which is higher than CHAID. The two models are significantly different, hence the difference of the models accuracy is not due to chance. The developed model is evaluated by using the one-thirds of the dataset. The evaluation is based on overall accuracy, coincidence matrix and accuracy (error) rate. The overall accuracy is 84.72% which means that 843 out of 995 students were retained in the engineering program. The coincidence matrix indicated the actual value and predicted value which is significantly different since the p-value (0.004) is less than 0.05. The actual value of the engineering students who are retained in the program is 879 which is less than the predicted value, 927. Meanwhile, the actual value of engineering students who are not retained is equal to 114 which are higher to the predicted value of 68. The result of the coincidence matrix is validated using the ACC (ERR) rate. The ACC (ERR) rate is equal to 0.8472(0.1528) which is within the accepted value.

The developed predictive model, C5.0 generates the algorithm wherein it constructed a decision rule by recursively splitting the data into smaller subgroups. The decision tree of C 5.0 will be the framework in identifying those students who would likely be academically at-risk and early detection can help the students for early intervention. A student is more likely to persist and graduate in settings that provide frequent and early feedback about his possible performance. The use of early warning systems by a University created an impact in providing a student the much needed information about his performance, so he can adjust his performance in order to persist and finish his program.

References

[1] Q.A . Al-Radaideh, E.M, Al-Shawakfa, A. A. Ananbeh, "A classification model for predicting the suitable study track for school students", *InternationaI Journal of Research and Review in Applied Science*, vol.8(2), pp. 247-252., 2011.

[2] A. Astin, "Student involment : a development theory for higher education," *Journal of College Student Personnel*, vol. 25, pp. 297 – 308, 1984.

[3] R. S. Bichkar, R. R. Kabra," Performance prediction of engineering students using decision tree," *InternationalJournal of Computer Applications*, vol. 36(11), pp. 8 – 11, 2011.

[4] A.C. Lagman, "Predictive decision support system using logistic regression and decision tree model for student graduation success determination," *An International Referred Journal of Applied Technology in Education*. vol.1(1), 2015.

[5] B.Gerardo, B.T. Tanguilig III, I.Tarun, "Generating licensure examination performance models using Part and JRip classifiers: a data mining application in education," *International Journal of Computer and Communication Engineering*, vol.3(3),2014.

[6] E.Osmanbegovic, M.Suljic, "Data mining approach for predicting student performance," *Journal of Economics and Business*, vol.10(1), pp. 3 – 12, 2012

[7] S. Pal, S.K. Yadav, "Data mining: a prediction for performance improvement of engineering students using classification," *World ofComputer Science and Information Technology Journal*, vol. 2(2), pp.51 – 56 ,2012.

[8] N.Menzies, A. Nandeshwar, A. Nelson, "Learning pattern of university student retention," *Expert Systems with Application*, vol 38(12), pp 14890 – 14896, 2011.

[9] L.Noel, R Levitz, D. Saluri," *Increasing Student Retention*," San Francisco, CA: Jossey-Bass, 1985.

[10] B.K Baradwaj, S.Pal, "Mining educational data to analyze students' performance," *International Journal of Advanced Computer Science and Applications*, vol. 2(6), pp. 63 – 68, (2011).

[11] D. l. la Red Martinez, C. E. Podesta Gomez, "Contributions from data mining to study academic performance of students of a tertiaryinstitute," *American Journal of Educational Research*, vol.2(9), pp.713-726, 2014. https://doi.org/10.12691/education-2-9-3

[12] L.Noel, R.Levitz, "Student success,retention, and graduation: Definitions, theories,practices,patterns,and trends," *Noel- Levitz Retention Codifications*, pp. 1 – 22, 2008.

[13] E.T. Pascarella, P.T. Terenzini, "*How College Affects Students: AThird Decade of Research, "*San Francisco, CA: Jossey – Bass ,2005.

[14] A.A.Sayficar, S. Talebi, "Using educational data mining (edm) to predict and classify students" *International Journal of Engineeringand Computer Science*, vol.3(12), pp.7242– 9395, 2014.

[15] P.V.Praveen Sundar, "A comparative study for predicting student's academic performance using bayesian network classifiers" *JournalOrganization of Science Research Journal of Engineering*, vol.3(2),pp.37 – 42, 2013.

[16] K.Chandra, N. Misiunas, A. Oztekin, M.Raspopovic, "Sensitivity of predictors in education data : a bayesian network," *Proceedings of the 2015 INFORMS Workshop on Data Mining and Analytics*, pp.1 – 6, 2015.

[17] S.A. Aher, L.M.R. J. Lobo, "Data mining in educational system using WEKA," *IJCA Proceeding on International Conference on Emerging Technology Trends*, vol.3, pp.20 -25, 2011.

[18] I .Asshaari, N.A Ismail, Z. M Nopiah, Othman, H., N.M.Tawil, A. Zaharim, "Mathematical performance of engineering students in Universiti Kebangsaan Malaysia (UKM)," *Procedia-Social and Behavioral Sciences* vol.60 pp. 206-2012, 2012 https://doi.org/10.1016/j.sbspro.2012.09.369.

[19] G.Dekker, M. Pechenizkiy, J. Vleeshouwers, "Predicting students drop out; a case study" *2nd International Educational Data Mining Conference*, pp. 41 – 50,2009.

[20] V. Tinto, "Classrooms as communities: Exploring the educationalcharacter of student persistance," *Journal of Higher Education, vol.* 68(6), pp.599 – 623, 1997. https://doi.org/10.1016/j.sbspro.2012.09.369

[21] J.Han, M.Kamber, J. Pei,. *Data Mining Concepts and Techniques* 3rded. Morgan Kaufmann Publishers, 225 Wyman Strreet Waltham, MA 02451, USA, 2011.

[22] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, P. From data mining to knowledge discovery: an overview. *Advances in Knowledge Discovery and Data Mining*, pp. 1 – 34. AAAI Press, 1996.

[23] V. Kumar, M. Steinbach, P. Tan, P. *Introduction to data mining* 1sted.Pearson Education, Inc., 2006.

[24] R. Pressman, *Software Engineering: A Practioner's Approach.,* McGraw-Hill, New York, 2005.

[25] V.Tinto, *Leaving college: Rethinking the causes and cures of student attrition. 2nd. ed*. Chicago: The University of Chicago Press, 1993.

[26] J. Fleming, *Blacks in college.* San Francisco: Jossey-Bass Inc. J.,1984.

[27] S. Hurtado, D.F. Carter, Latino students' sense of belonging in thecollege community: Rethinking the concept of integration on campus. *In College Students: The Evolving Nature of Research.* Needham Heights, MA: Simon&Schuster Publishing, 1996.

[28] R. Baker, M.Pechenizkiy, C. Romero, S. Ventura, (Eds). *Handbook ofeducational data mining.* Boca Raton, Florida: Taylor and Francis Group, LLC, 2011.

[29] W. Hämäläinen, M. Vinni, M., "Classifiers for educational data mining," In Baker, R., Pechenizky, M., Romero, C., & Ventura (Eds), *Handbook of Educational Data Mining.* Boca Raton, Florida: Taylor and Francis Group, LLC 2011, pp. 57 – 74.

[30] C. Romero, S. Ventura, S. ,A. Zafra, "Multi-Instance learnining versus single-instance learning for predicting the student'sperformance," In Baker, R., Pechenizkiy, M., Romero, C., &Ventura, S. (Eds), *Handbook of Educational Data Mining*Raton,Florida: Taylor Boca and Francis Group, LLC, 2011, pp. 187 – 200.

[31] IBM SPSS Modeler Version 18 Modeling Nodes, IBM Corporation, 2016.

[32] J. S. Cuenca, J. S. ,"Efficiency of state universities and colleges in the phillipines: a date envelopment analysis," In R.G. Manasan (Ed), *Analysis of the President's Budget for 2012: Financing of State Universities and Colleges* Philippine Institute for Development Studies, Makati City, Philippines. 2013, pp. 126 – 146..

[33] V. Tinto, "Linking learning and leaving: Exploring the role of the college classroom in student departure," In J. Braxton(ed.) *Reworking the student departure puzzle.* Nashville:vanderbilt University Press, 2000.

[34] X. Yang, (ed). Mathematical modeling. *Mathematical Modeling with Multidisciplinary Applications.* United States Of America: A John Wiley & Sons, Inc., 2013.

[35] M. Atanasov, H. Darabi, F.Karim, A. Sharabiani, A. Sharabiani, (2014). An Enhancedbayesian network model for prediction of students' academic performance in engineering program. *Institute of Electrical and Electronics Engineers Global Engineering Conference*, pp. 832 – 837. doi: 978-1-4799-3190-3/14.

[36] G. E. Adams, A.A. Cherif, A.A, F. Movahedzadeh (2013). Why dostudents fail? student's perspective.http://www.researchgate.net/publication/256319939

[37] D. Jaithavil, M. Pracha, W.,Punlumjeak,N.S. Rugtanom. (2015). A prediction of engineering students performance from core engineering course using classification. *Lecture Notes in Electrical Engineering* 339, 649-656. doi:10.1007/978-3- 662-46578-3_7

[38] E.P.I Garcia, P.M. Mora. (2011). Model prediction of academic performance for first year students. , *Institute of Electrical and Electronics Engineers.* doi: 10.11109/MICA.2011.28.

[39] D. Kabakchieva. (2013). Predicting student performance by using data mining methods for classification. *CYBERNETICS AND INFORMATION TECHNOLOGIES, 13(1),* 61 – 72. doi:10.2478/cait-2013-0006 https://doi.org/10.2478/cait-2013-0006

[40] L.A. Kurgan, P. Musilex (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review, 21(1), 1 – 24.* doi: 10. 1017/S0269888906000737 https://doi.org/10.1017/S0269888906000737

[41] N. Fang, S. Huang. (2013). Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical model. Computer &Education 61, 133 – 145.doi.org/10.1016/comedu.2012.08.015

[42] Ahmad, F., Aziz, A.A., Ismail, N.H. (2015). The prediction ofstudents' academic performance using classification data mining techniques. *Applied Mathematical Sciences, 9(129),64*15– 6426. doi:10.12988/ams.2015.53289. https://doi.org/10.12988/ams.2015.53289

[43] R. Asif, A.Merceron, M.K. Pathan. (2014). Predicting studentacademic performance at degree level: a case study. *I.J. Intelligent Systems andApplications, 01,* 49-61. doi.10.5815/ijisa.2015.01.05. https://doi.org/10.5815/ijisa.2015.01.05

[44] P. Ranada. (2017). Will it help Duerte fulfill his promise? *http:/www. rappler.com/authorprofile/ pia-ranada.*

[45] C. Romero, S. Ventura. (2005). Educational data mining: a survey from 1995 to 2005. *Expert System with Application* 33(1),135-146. doi: 10.1016/j.eswa. 2006.04.005.

[46] Technet, Microsoft. (2017). Testing and validation(data mining). docs.microsoft.com.