International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 6 (2019), pp. 917-928
© International Research Publication House. http://www.irphouse.com

Predictive Model for the Academic Performance of the Engineering Students Using CHAID and C 5.0 Algorithm

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

Avelina R. Raqueno
Centro Escolar University

Abstract

Many engineering students in Technological University of the Philippines Manila (TUPM) 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. TUPM needs to increase the retention rate to lessen the number of students who will drop out, on probation or be dismissed from College of Engineering (COE). Predictive modeling could be one of them. It is used to detect student behavior, predicting or understanding student educational outcome. It is one of the current popular method in Educational Data Mining (EDM). EDM is a field of scientific inquiry for the development of method to discover unique kind of data in educational settings, and using this method to understand better the students and their learning environment. As such, the study aimed to develop and validate a predictive model that will serve as a framework in predicting the academic performance of the engineering students towards an improved retention rate at TUPM. The research design of the paper was descriptive-quantitative. The data of the engineering students' final grades from school year 2008 - 2015 were gathered from the Electronics Registration System of TUPM. The dataset was divided into two sets: training and testing set. The training set was used to build and validate two decision tree algorithms namely, C5.0 and Chi-squared Automatic Interaction Detection (CHAID), using IBM SPSS Modeler Version 18.0 based on their overall accuracy and ten-fold cross validation. To determine their significant difference, t-test was used. Furthermore, the testing set was used to evaluate C5.0 and CHAID on their overall accuracy, sensitivity and specificity. Based on the result of the overall accuracy it was found out that C 5.0 was slightly higher than CHAID and both were valid. However, the predicted model is CHAID based on the evaluation of the two algorithms. Hence, CHAID was the best early warning system for TUPM to detect the students who are academically at risk. As such, the study concludes that CHAID modeling algorithm suited best as the predictive model for identifying students who were likely to be retained in the COE program and those who were academically at-risk. .

Keywords: C5.0, CHAID, Decision Tree, Educational Data Mining, Prediction Model

I. INTRODUCTION

Many students dropped out or were dismissed from the engineering program at Technological University of the Philippines-Manila (TUPM); hence, [9] it is urged to use efficiently its resources to achieve their intended purpose. One possibility is the use of Data Mining that determines valid, useful and understandable patterns on the data on the academic performance of the students by applying pattern recognition (PR) and machine learning principles in different data sets called Educational Data Mining (EDM). One popular method of EDM is Prediction. The Prediction Model determines the output value in context where it is not desirable to directly obtain a label for that construct [6].

One of the three types of Prediction is Classification. It 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 [5]. 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, [18] 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 counselling / advising. Likewise, [8] 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. [3] 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 [14] 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 which means that the error rate was high and the predictions were not reliable.

The Prediction Model was used in the study, because it aimed to develop and validate a predictive model that will serve as a framework in predicting the academic performance of the engineering students towards an improved retention rate 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.

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 [2]. 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.

The paper is organized as follows: Section II review of related works of authors in building, validating and evaluating different algorithms as framework in predictive models; Section III discussed the methodology in developing, validating and evaluating the modelling algorithms; The recommended predictive models, decision tree mapping and the IF-THEN generated rules are presented in Section IV; lastly, Section V concludes based on the results, significant of the study and recommended study for future works.

II. REVIEW OF RELATED LITERATURE AND STUDIES

This section presents various related literature and studies which are guidelines to the framework of predictive models.

Classification Techniques

Classification is defined as a data mining task that maps data into predefined groups and classes,[11]. It has a two –steps. First, a model is built by analyzing the data tuples from the training data which have a set of attributes. And classification techniques(algorithms) applied in training set to create a model. Secondly, test data is used to check the accuracy of the model [11]. Some of the classification techniques(algorithms) are Decision Tree, Regression, Neural Networks, Support Vector Machine and Bayesian network.

Decision Tree

A decision tree develops classification systems that predict or classify future observations based on a set of decision rules Based on [13] some of the decision trees algorithms are: The Classification and Regression Tree (C&R) Tree, Chi-squared Automatic Interaction Detection (CHAID), C 5.0 and, Quick, Unbiased, Efficient, Statistical Tree (QUEST)

The study of [16] aimed to develop a model using the decision algorithms, C 5.0 and CHAID to estimate the financial failure and/or success of the manufacturing company. The result indicated that CHAID algorithm rate of accuracy and its sensitivity (rate for successful companies) are higher than the rates obtained from C 5.0. However, CHAID algorithm gave lower results than the C 5.0 in predicting its specificity (rate for unsuccessful companies). Furthermore, the model formed applied to whole data set by C 5.0 and CHAID had an overall accuracy of 85.15 percent and 87.37 percent respectively. Hence, study concluded that developed models based on C5.0 and CHAID algorithms can be used to classify both successful and unsuccessful firms based at acceptable level. And both models classify firms based on fundamental ratios related to leverage, liquidity, profitability, and cash flows.

[10] objective is to classify a tumor in breast whether it is benign or malignant based on cell descriptions compound by the microscopic examination using decision tree. The five models namely: C&RT, Quest, C5.0, CHAID, and SVM were measured in terms of classification accuracy, sensitivity, and specificity. The dataset were partitioned into training and testing set by the ratio 70:30 percent respectively. The findings indicated that SVM with 99.976 percent is the best in accuracy for training but the CHAID with 99.074 percent is the best for sensitivity of testing dataset followed by C5.0 and SVM with 98.198 percent. Moreover, SVM and the Decision tree models (C&RT, Quest, C5.0 and CHAID) can be effectively used for breast cancer diagnosis to help physicians and oncologists.

Educational Data Mining (EDM)

Educational Data Mining (EDM) which is a field of 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 [6]. One of the current popular methods of EDM is prediction. It is also used to detect student behavior, predicting or understanding student educational outcome.

The focus of the study of [1] was to identify the optimal decision tree algorithm for predicting students' performance in a computer programming course based on their Mathematics and Physics courses as their attributes. The study used C4.5, CART, and Best- Free Tree as the decision tree algorithms. The 10 - fold validation was used to compare the results of the three algorithms. The obtained result showed that C 4.5 had the highest prediction accuracy of 70.37 percent and that the essential attributes of students' performance in a computer programming were previous knowledge in Mathematics and Physics courses. Likewise, the study of [4] showed that prior knowledge in Mathematics and Physics courses were vital for students' proficiency in computer programming.

[7] aimed to analyse three separate predictors; demographics, study habits, and technology familiarity to identify university students' characteristics and the relationship between each of the predictors with student achievement. The data gathered were analyzed using the CHAID algorithm. The study revealed that relationships involving university students' demographics, study habits, and familiarity with technology were correlated with their self-reported GPAs. Hence, it implied that gender,

study habits, and familiarity with technology were important factors that may affect university students' achievement.

The purpose of [15] paper was to understand the external factors that will contribute to the student loyalty and predicting the pattern of loyal (success) students. The dataset were based on the following external factors (predictor variables): personal information of students, student academic status, types of their previous university, finances, and occupational and educational status of the parents. The algorithms used were CART, C5.0 and CHAID in predicting students' loyalty. The results revealed that data mining techniques can predict loyal (success) students wherein CART is the best model with 91.42 percent accuracy, followed by C5.0 with 88.57 percent. However, CHAID produced the lowest prediction accuracy of 80.95 percent. Furthermore, to estimate the prediction accuracy for each model, sensitivity analysis were performed to identify the relative importance of the predictor variables. The results implied that the most important predictor variables were educational background (previous university, number of terms) and parent's educations.

[17] developed a model to achieve a measurable student progress monitoring process that will provide quick results. It focused on performance monitoring of students' continuous assessment through tests and examination scores in order to predict the students' final status upon graduation. Several data mining techniques such as: ANN, C&RT C5.0 and CHAID were utilized. C 5.0 algorithm was the best representative since it determined which of the various attributes represents best the division of the training sets. The classification techniques of ANN, C&RT, C5.0, and CHAID were compared in terms of training, testing, and validation datasets for model performance The results indicated that C5.0 had the highest average of 97.30 percent compared to other four while CHAID had the lowest average of 57.77 percent. The paper concluded that data mining techniques provided effective monitoring tool for student academic performance and fine tuning derived variables improves rules quality producing improved performance.

III. METHODOLOGY

The research design of the paper was descriptive-quantitative. The subject of the study was composed of engineering students who were officially enrolled in, Civil, Electrical, Electronics and Communication Engineering, and Mechanical 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 - 2015 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 1and (Lec & Lab). [1] and [4] used also the Mathematics and Physics courses as their predictor variables to showed that prior knowledge in Mathematics and Physics courses were vital for students' proficiency in computer programming.

A total of 3 765 students qualified in the criteria, broken down as follows:

Table 1: RESPONDENT'S PROFILE PER COURSE

Degree Program (Course)	Number of Students Before their 3^{rd} year
CE	1042
ECE	1144
EE	725
ME	854
Total	**3765**

Predictive Model Development

The development of the predictive model was adapted from [11] and [3]. 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 three main stages of the framework of a predictive model by [3].

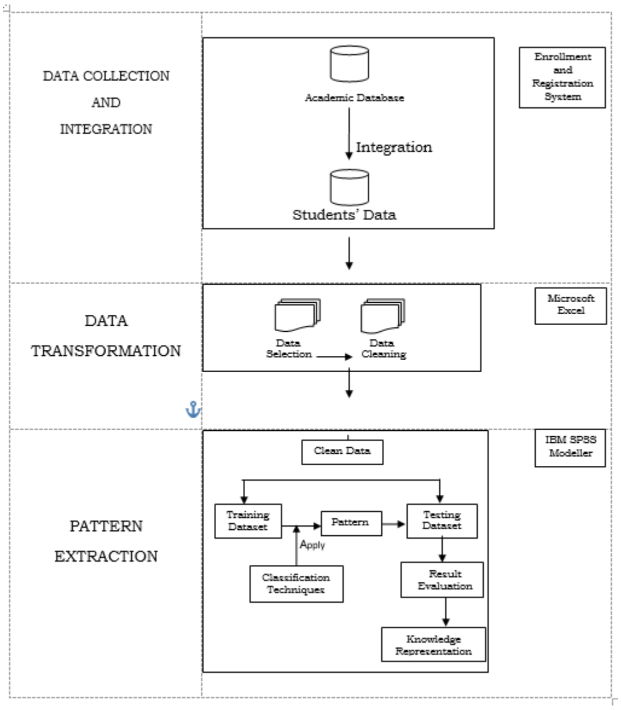


Figure 1. The Framework of the Predictive Model

Based on Figure 1, in the Data Collection and Integration Stage, data are gathered from ERS Database of the Registrar. Meanwhile, in the Data Transformation Stage, where the quality of the input data is improved in order to produce quality result. Only 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. In the Pattern Extraction Stage, a data mining tool is used to conduct the process of extracting the pattern among the input using the IBM SPSS Modeller Version 18. It consists of five steps namely: 1) Training Set, 2) Pattern, 3) Testing, 4) Result Evaluation, and 5) Knowledge Representation.

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. The cleaned data is divided into two datasets: 1) Training set and 2) Testing set. The training dataset is used to build the model and mined by data mining techniques to extract patterns, while the testing dataset is used to evaluate the model. In order to determine the interesting patterns from the datasets, the result evaluation is conducted. Finally, the mined knowledge is presented to the users through visualization and knowledge techniques in the Knowledge Representation [11].

Training Set

Two-third of the dataset was used as training set. Figure 2 showed the process of pattern extraction in building and validating the model using the software IBM SPSS Modeler version 18. The software builds stream to create a model. The stream needs three elements namely, a source node, type node, a modelling node.

The training set was mined using the decision tree algorithms namely: C5.0, Chi-squared Automatic Interaction Detection (CHAID). The top two in decision tree algorithms were used based on the auto classifier, a built-in classifier in the software that rank the models based on their overall accuracy and number of fields. Each algorithm indicated its overall accuracy, prediction importance chart and validation.

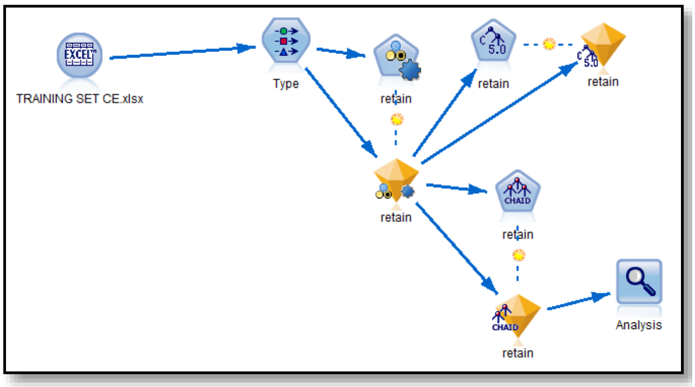


Fig 2. Screenshot of how to build and validate the modelling algorithms in Training Set

As exhibit in figure 2, the source node reads the training dataset from the external node, Microsoft Excel. A connector links the source node to the Type Node that specified field properties. Then it connected to the modelling node that generated the model nugget when the stream is run. The generated model nugget will be based on the build in auto classifier of the IBM SPSS Modeler Version 18.0. The auto classifier listed down the algorithms based on their overall accuracy and number of fields [13]. Each algorithm (orange colour, called jewel) was links to Analysis for overall accuracy, predictor importance chart and cross-fold validation. Table 2, shown the type node that specified field properties.

Table 2. THE TARGET AND PREDICTORS WITH THE TYPE NODE

Field (Predictor Variables)	Description	Measurement	Value	Role
Math 1	College Algebra	Continuous	[1.00 – 300]	Input
Math 2	Plane and Spherical Trigonometry	Continuous	[1.00 – 3.00]	Input
Math 3	Solid Mensuration	Continuous	[1.00 – 3.00]	Input
Math 4	Analytic Geometry	Continuous	[1.00 – 3.00]	Input
Math 5	Differential Calculus	Continuous	[1.00 – 3.00]	Input
Math 6	Integral Calculus	Continuous	[1.00 – 3.00]	Input
Math 10	Advance Algebra	Continuous	[1.00 – 3.00]	Input
Physics 1	General Physics (Lec)	Continuous	[1.00 – 3.00]	Input
Physics 1	General Physics (Lab)	Continuous	[1.00 – 3.00]	Input
Physics 2	Fluids, Thermodynamics and Electromagnetism (Lec)	Continuous	[1.00 – 3.00]	Input
Physics 2	Fluids, Thermodynamics and Electromagnetism (Lab)	Continuous	[1.00 – 3.00]	Input
Degree Program (Course)	CE, ECE, EE, and ME	Nominal	None	Input
Retain		Nominal	None	Target

Based on table 2, the columns were divided as follows: Field of course codes, 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 role of each predictor variables will be the Input field whose values were used by the modelling algorithm to predict the value of the target field while Retain role will be the Target field. It indicates whether or not the engineering students were retained or not retained in the degree programs of COE.

Table 3 showed the auto classifier which estimates and compares algorithms for either nominal sets or binary targets.

In ranking algorithm for a nominal target it is restricted to overall accuracy and number of fields. Note that, the area under the curve provides only an index for the performance of the model [13].

Table 3. AUTOCLASSIFIER OF THE ALGORITHMS

Algorithm	Build Time (mins)	Overall Accuracy (%)	No. Fields Used	Area Under Curve
C5.0	< 1	85.931	12	0.779
CHAID	< 1	83.682	9	0.814
Logistic regression	< 1	83.538	12	0.783
C&R Tree	< 1	83.032	12	0.500
Neural Network	< 1	82.924	12	0.794
Bayesian Network	< 1	82.058	12	0.798
Discriminant	< 1	63.032	11	0.735
Decision List	< 1	43.574	7	0.648

It was observed in Table 3 that the top two algorithms are C5.0 and CHAID with an overall accuracy of 85.931 percent and 83.682 percent respectively. Although, the area of the curve of CHAID is slightly higher than C5.0. It just an indication that the area under the ROC curve of CHAID is a little further than the reference line. It was also observed that number of fields of C5.0 is greater than CHAID. The number of fields represent the predictor importance. CHAID applied calculation stopping rule and probability values were used for the computation of the predictor importance (IBM SPSS Modeler, 2016).

Classification Techniques

The two classification techniques under Decision Tree algorithms identified by the auto classifiers based on their overall accuracy were C.50 and CHAID. [13] defines C 5.0 as the node that builds either in decision tree or a rule set. The modeling algorithm used information theory for splitting. It splits the dataset based on the field that provides the maximum information gain at each category. On the other hand, CHAID used chi-squared statistics to identify optimal splits. It can generate non-binary tree where other splits have more than two branches. It involves three steps that is done iteratively such as: merging, splitting, and stopping.

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 modelling with the target field (retained or not retained) was to study the data to which the outcome was known and identify the patterns of the outcomes that were not known.

Evaluation of two modelling algorithms was done base on overall accuracy, sensitivity and specificity. Figure 2 showed the process evaluating the modelling algorithms using the software IBM SPSS Modeler Version 18. Likewise, the stream for testing set also needs three elements namely, a source node, type node, a modelling node. The figure 3 illustrates the process of evaluating the modelling algorithms.

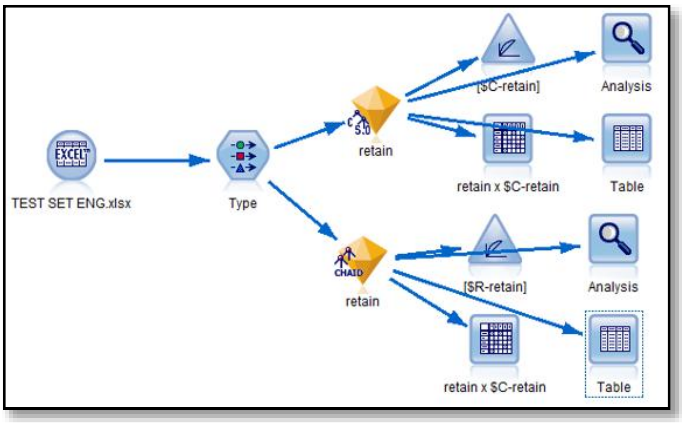


Fig 3. Screenshot of how to evaluate the modelling algorithms in Testing Set

As observed in Figure 3, the source node reads the testing dataset from the external node, Microsoft Excel. A connector links the source node to the Type Node that specified field properties. Then it connected to the modelling algorithm (model nugget) nodes. The model nugget based on the build in auto classifier of [13] were C5.0 and CHAID. Each modelling algorithm was links to table, coincidence matrix, Analysis and ROC for the evaluation. The Table listed down the comparison between the students who were predicted to be retained in the degree programs which was created by the modelling algorithms and those students who were actually retained in the degree programs. In the Analysis node, it compared the number of correctly predicted, that is, students who were retained and not retained in the degree programs with incorrectly predicted, that is, students who were misclassify as retained but not retained in actual result or students who were not retained but were retained in actual result. On the other hand, the Coincidence (confusion) matrix analyses how will the modelling algorithms can recognized tuples of different classes such as: true positive, true negative, false positive , and false negative. True positive and true negative indicated that the modelling algorithm correctly labelled the students who were retained and who were not retained respectively in the degree program. One the other hand, false positive indicated that the algorithms incorrectly labelled the students as retained but in actual value, not retained and false negative the student labelled as not retained but in actual value, is retained. The graphical representation for comparing the two modelling algorithms is through Receiver Operating Characteristics (ROC).

Result Evaluation

The predicative model for the academic performance of the engineering students will depended on the evaluation of the two modelling algorithms. The evaluation will be based on overall accuracy, sensitivity and specificity.

Knowledge Presentation

The recommended predictive model will be presented in the two forms: decision tree mapping and IF-THEN rule. A decision tree is a flowchart-like tree structure, where each parent denotes a test on an attribute, each child node represents an outcome of the test, and leaf node (terminal node) holds a class label [11]. The topmost node in a tree is the root node which summarizes who are retained and not retained in the degree program of COE. On the other hand, in the IF-THE rules, the IF part of a rule is called as precondition, and the THEN part is the conclusion. In the precondition, the condition consists of one or more predictor variables that uses the connective AND. The rule's conclusion contains the predictive variable (target), that is, whether the student will be retained or not retained in the degree programs of COE.

IV. RESULTS AND DISCUSSION

The result and discussion of the study that aimed to develop and validate a predictive model that will serve as a framework in predicting the academic performance of the engineering students towards an improved retention rate at TUPM were as follows:

A. Building and Validation of Modeling Algorithms

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 (Lec and Lab)
- Degree Programs namely: CE, ECE, EE, ME

Table 4 listed the two decision trees 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 4. THE TWO DECISION TREES

Algorithm	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 4, both algorithms have less than one minute to build the models. The overall accuracy indicated the percentage of records that is correctly predicted, that is, students who were retained and not retained in the degree programs of the College of Engineering by the algorithms 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 area under the curve slightly higher than C 5.0 which indicates the curve lies further above the reference line [13].

Overall Accuracy in Training Set

Table 5 gave the comparison of the algorithms in terms of correctly predicted, students who were retained and not retained in the engineering program and incorrectly predicted that is, students who were misclassify as retained but in actual the students were not retained. Likewise, students who were misclassify as not retain but they were retained in the program.

Table 5. OVERALL ACCURACY OF THE TWO ALGORITHMS IN TRAINING SET

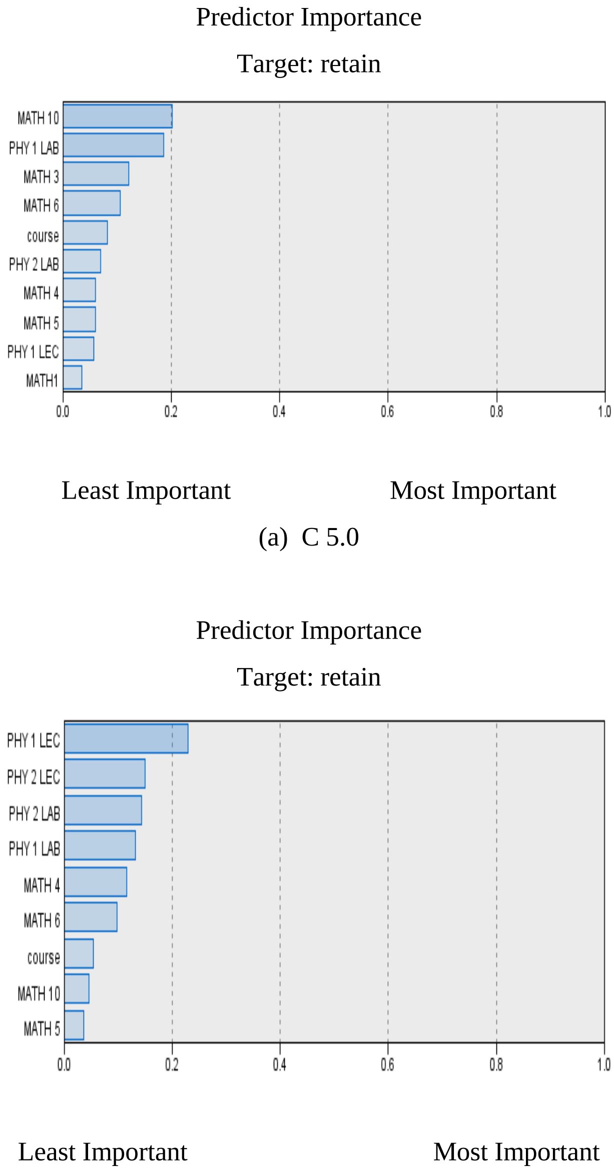
Algorithm	C 5.0		CHAID	
	N	Percentage (%)	N	Percentage (%)
Correctly Predicted	2408	86.93	2318	83.68
Incorrectly predicted	362	13.07	452	16.32
Total	2770	100.00	2770	100.00

It was observed in table 5 that C5.0 was slightly higher than CHAID in correctly predicting students who were retained and not retained in the degree programs. Meanwhile, CHAID percentage in misclassifying students was higher than C5.0. The findings of the study is the same with that of [19] that compared the performance of C 5.0 and CHAID. In their paper, it was found that C5.0 had a better performance in terms of over accuracy than CHAID. Similarly, the paper of [15] whose aim was to predict loyal students, that is, students who have decided to continue studying in their respective programs using the three models: CART, C5.0, and CHAID.

The result showed that CHAID had the lowest prediction accuracy with 80.95 percent compared to CART with 91.42 percent and C5.0 with 88.57% in predicting loyal student.

Predictor Importance

Figure 3 shown the predictor importance chart which indicates the significant of each predictor in the algorithms



(a) C 5.0

(b) CHAID

Fig. 3(a) and 3(b). The Predictor Importance Chart

Based on the figures 3(a) and 3(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. On the other hand, CHAID listed Physics 1Lec as the most important and least important, is Math 5. However, predictor importance does not relate to model accuracy. It indicates the importance of each predictor in making a prediction, but, it does not matter whether or not the prediction is correct [13].

Ten-Fold Cross Validation

To validate the two models, the ten-fold cross validation was used. 10-fold cross validation is used when the training dataset were randomly partitioned into 10 mutually exclusive subsets or folds. Table 6 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 6. TEN-CROSS FOLD VALIDATION

	Algorithm	Mean	Standard Deviation
Pair 1	C 5.0 error	13.0670	3.83468
	CHAID error	16.3070	3.08778
Pair 2	C5.0 accuracy	86.9330	3.83468
	accuracy	83.6930	3.08778

Based on table 6, C 5.0 showed the highest (lowest) accuracy (error) in ten-cross fold evaluation 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. It also indicated that there were homogeneity in the mean error (accuracy) of both pairs.

t-test

The t-test was utilized in the model selection to show that the difference between the two algorithms in accuracy (error) is not due to chance [11]. Table 7 shows the test of significant of the two algorithms.

Table 7. MODEL SELECTION USING t-test

Algorithm		Mean	Standard Deviation	t	p-value	Significant
C5.0	error	3.240	1.454	-7.047	p=0.000< 0.01	Very Significant
CHAID	error					
C5.0	accuracy	3.250	1.475	6.966	p = 0.000 < 0.01	Very Significant
CHAID	accuracy					

As noted on table 7, the two models has p-value that is less than 0.01 which means that C5.0 and CHAID has very significant difference in their error (accuracy) estimate in the prediction of the students who will not retained and not retained in the engineering programs of TUP. It indicates the result of their difference is not due to chance [11].

B. Evaluation of the C5.0 and CHAID

To evaluate C5.0 and CHAID, one-third of the data is allocated to the testing data. The evaluation of the performance of the two algorithms were based on overall accuracy, sensitivity, and specificity.

Overall accuracy of the Testing Set

Table 8 illustrated the overall accuracy of the algorithms on a testing set . It is the percentage of the students that were correctly predicted and incorrectly predicted by the two algorithms

It was observed that CHAID is slightly higher this time than C 5.0 in overall accuracy. It means that CHAID had higher percentage of correctly predicted students who were retained and not retained in the program. However, it is in contrast with the finding of [17], wherein the overall accuracy of the C5.0 in testing data is greater than that of CHAID in evaluation of the performance of the two algorithms.

The coincidence matrix analyzes how well the model can recognized tuples of different classes. True positive (TP) and true negative (TN) indicates that the model is getting things right. On the other hand, false positive (FP) and false (FN) 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. Table 9 showed the coincidence matrix of the C 5.0 and CHAID which explain the difference between the actual value and the predicted value

Table 8. OVERALL ACCURACY OF C5.0 AND CHAID IN TESTING SET

Algorithm	C 5.0		CHAID	
	N	Percentage (%)	N	Percentage (%)
Correctly Predicted	843	84.72	870	87.44
Incorrectly Predicted	152	15.28	125	12.56
Total	995	100.00	995	100.00

Table 9. COINCIDENCE MATRIX OF C 5.0 AND CHAID

Algorithm			Predicted Value					
			Retain		Not Retain		Total	
C 5.0	Actual Value	Retain	828	TP	53	FN	881	
		Not Retain	99	FP	15	TN	114	
		Total	927		68		995	
			Predicted Value					
			Retain		Not Retain		Total	
CHAID	Actual Value	Retain	868	TP	13	FN	881	P
		Not Retain	112	FP	2	TN	114	N
		Total	980		15		995	

C 5.0 model has predicted correctly that 828 (TP) out of 881 Engineering students were retained and 15(TN) out of 114 who were not retained in the degree program of COE; however, it had incorrectly predicted that 53 (FN) students were not retained in the program, but they were actually retained in the program. Furthermore, C 5.0 had predicted incorrectly that 99(FP) students were retained in the program, but were actually not retained in the program. On the other hand, CHAID had predicted correctly that 868 (TP) out of 881 students were retained in the program and 2 (TN) out 114 were not retained in the degree program of COE; however, it had predicted incorrectly that 13(FN) students were not retained in the program, but they were actually retained in the program. Furthermore, CHAID had predicted incorrectly that 112(FP) students were retained in the were actually not retained in the program.

Sensitivity and Specificity

The sensitivity also known as true positive rate is the proportion of students who are retained that are correctly identified in actual value and specificity is the true negative rate is the proportion of students who are not retained in the actual value that are correctly identified. They are defined in [11] as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{P}} \quad \text{and Specificity} = \frac{\text{FP}}{\text{N}} \qquad (1)$$

Table 10 illustrated the sensitivity and specificity of the algorithms.

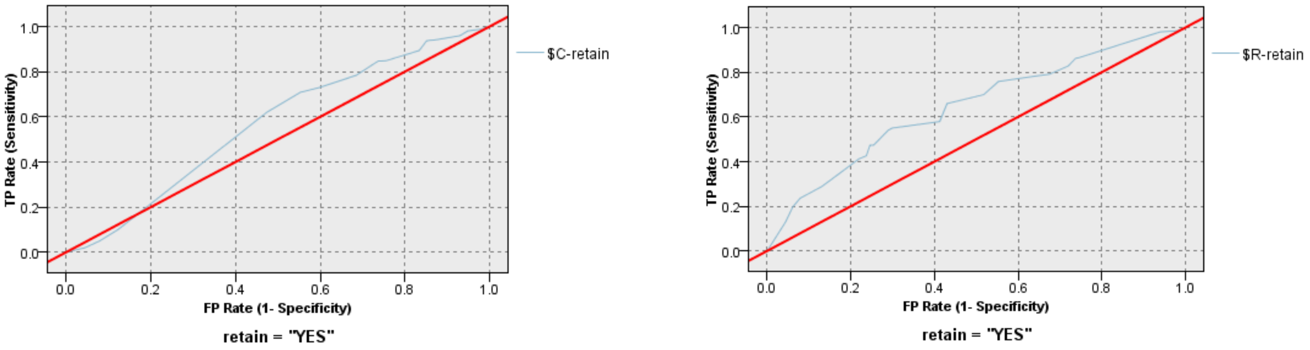
Table 10. Evaluation of Algorithms Based on Sensitivity and Specificity

Algorithm	Sensitivity (%)	Specificity (%)
C 5.0	93.98	13.16
CHAID	98.52	1.75

Based on the results, CHAID had a higher sensitivity than C5.0 but low specificity compared to C 5.0 in the estimate of students who were retained and not retained in the Engineering program of TUPM. Hence, CHAID had had a higher proportion of correctly identified students who were retained in the program (Sensitivity), However, CHAID had low proportion of correctly identified students who were not retained in the program (Specificity) compared to C 5.0 model.

Receiver Operating Characteristics (ROC)

The graphical representation of the C 5.0 can be interpreted through a Receiver Operating Characteristics (ROC) chart. Figure 3 shown is ROC chart with the curve starts at (0,0) coordinate and ends at the (1, 1). The vertical axis and horizontal axes represent the True Positive (TP) and the False Positive (FP) respectively.



(a) C5.0

(b) CHAID

Fig.3 ROC Chart

As exhibited in Figure 3, C 5.0 had some points below and above the line in reference line. This implies that false positive were more significant and the curve indicated less accurate prediction since there were less area under the curve (0.568). On the other hand, CHAID had all points above the reference line. However, the line tails off to the right earlier and the area under the curve (0.643) is a little less, it implied that the risk of false positive increased. Between the two algorithms, CHAID had a better classification result because CHAID area under the curve was higher than C 5.0. It implied that CHAID can correctly predicted students who were retained and not retained in the degree program of COE. Moreover, identify the students who were academically at-risk.

C. Developed (Predictive) Model

Based on the findings of the research, CHAID is recommended in the study to predict who will be retained and not retained in the engineering program of TUPM, because CHAID has a high accuracy and sensitivity compared to C 5.0.

CHAID evaluates all of the values of a potential predictor fields (variables) using the significance of a statistical test as criterion, [12]. CHAID algorithm was developed and introduced by Kass in 1980. CHAID algorithm moves from root node towards the bottom of tree called the terminal node. At each steps, CHAID examines the crosstabulations between each of the predictor fields also (called an input fields) and the target(retain) using a chi-sqaured independence test. It merges values that are statistically similar with respect to the target. If more than one of the values is dissimilar, CHAID will select predictor field with smallest p-value. The predictor field with the smallest p-value will be the first split. Each final category of a predictor field will represent child node if predictor field is used to split the node. The category –merging process stops when all categories differ at the specified testing level. Figure 4 shown the step in creating the decision tree of CHAID.

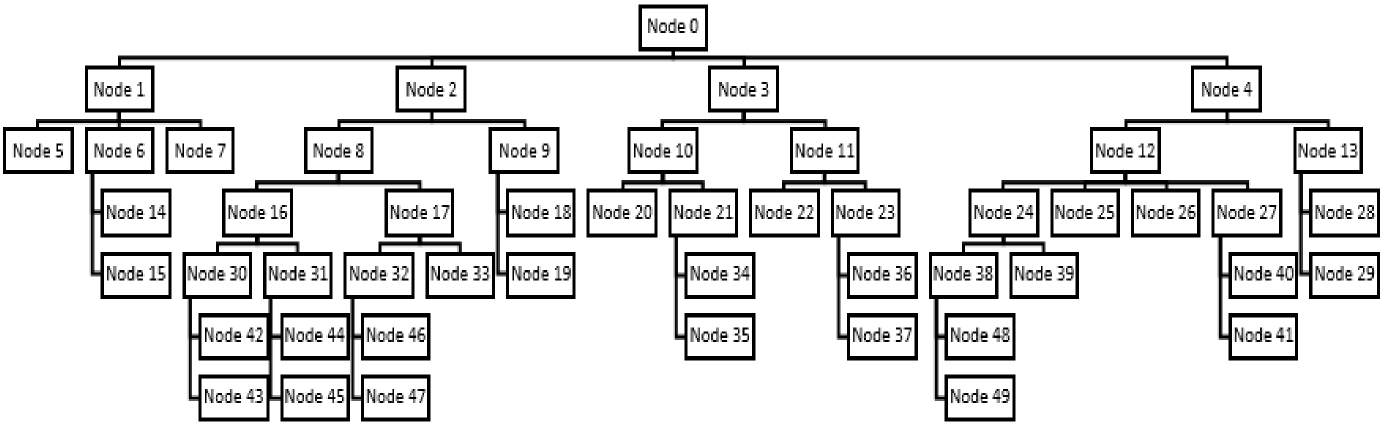
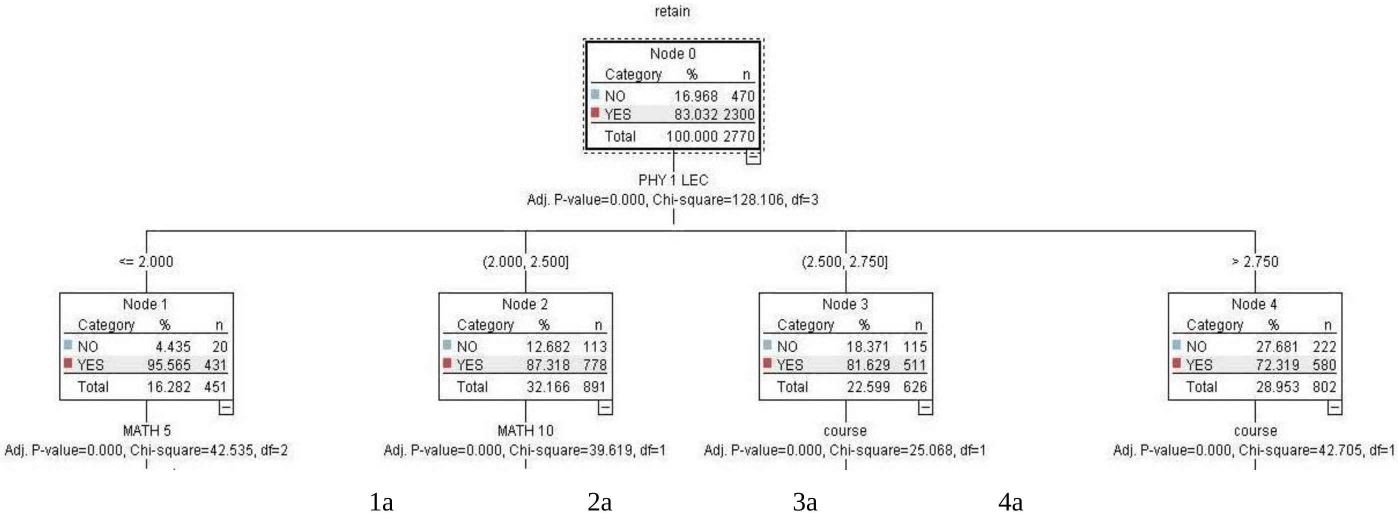


Fig. 4. Overview of the Decision Tree Mapping

Based on figure 4, Node 0 represents a summary for all the records in the dataset. By chi-squared independence test, the first split called parent node divides the category into four nodes. Each parent node splits the category again into child nodes (internal) made of a group of homogeneous values of the selected predictor field. Then the process continues recursively until it reaches the terminal node where all splitting stops.

A. Decision Tree Mapping

Figure 5 illustrated the root node and the first split of the decision tree.   As it exhibited in Figure 5, The root node summarized the total number of students who were retained (Yes) and not retained (No) in the degree program of COE. The first split is PHY 1(LEC) which coincided with the result of predictor importance in figure 3b and the four parent nodes were Node 1, Node 2, Node 3, and Node 4. Each category (in terms of student's numerically value of their grades), indicated the number of students who were retained (Yes) and not retained (No) in the program and their corresponding child nodes, Math 5, Math 10, and course.


retain
PHY 1 LEC
Adj. P-value=0.000, Chi-square=128.106, df=3
<= 2.000
(2.000, 2.500]
(2.500, 2.750]
> 2.750
MATH 5
Adj. P-value=0.000, Chi-square=42.535, df=2
MATH 10
Adj. P-value=0.000, Chi-square=39.619, df=1
course
Adj. P-value=0.000, Chi-square=25.068, df=1
course
Adj. P-value=0.000, Chi-square=42.705, df=1
1a
2a
3a
4a

Node 0

Category	%	n
🟦 NO	16.968	470
🟥 YES	83.032	2300
Total	100.000	2770

Node 1

Category	%	n
🟦 NO	4.435	20
🟥 YES	95.565	431
Total	16.282	451

Node 2

Category	%	n
🟦 NO	12.682	113
🟥 YES	87.318	778
Total	32.166	891

Node 3

Category	%	n
🟦 NO	18.371	115
🟥 YES	81.629	511
Total	22.599	626

Node 4

Category	%	n
🟦 NO	27.681	222
🟥 YES	72.319	580
Total	28.953	802

Figure 5. The root node and the first split

Figure 6 shown the partial part of the CHAID decision tree mapping. In Node 4, if the numerical value (grade) of PHY 1 (LEC) is greater than 2.750, the crosstabulation indicates that 580 out of 802 students were retained and the remaining 222 students were not retained in the program. The second split was COURSE (degree program). The COURSE divided again into CE, ECE, ME and EE. The category under CE, ECE, and ME had seven terminal nodes. It was noted that 56 students under the terminal Node 41 with ECE as their course were not retained in the degree program of COE. While the category EE had two terminal nodes. And all of the students under this category were retained in the EE program.

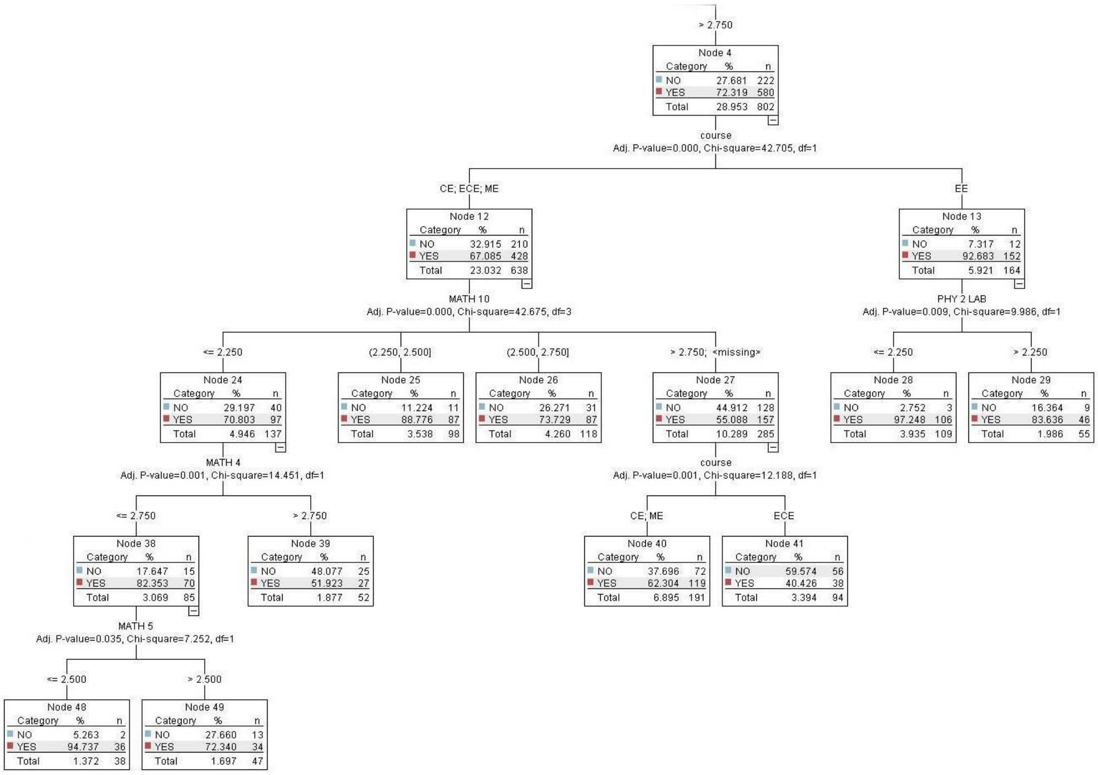


Fig. 6. Node 4 of the Decision Tree Mapping

B. Rules Extracted Based on the CHAID Decision Tree

Table 11 illustrated the rule extracted from Figure 6. The path of the IF-THEN (classification) rules was from the parent node until the terminal node. The classification rules extracted from the decision tree mapping corresponded to the number of terminal nodes tree of CHAID. Since figure 6 had nine terminal nodes, then it had also nine classification rules. R7 indicated that students under this rule were not retained in the ECE program and those students who will be under this conditions will be academically at risk and needs immediate intervention from TUPM.

Table 11. Rules Set Generated by CHAID Tree in the Engineering Program of TUPM

PHY 1 LEC > 2.750
R1. If PHY 1LEC > 2.750 and course = CE, ECE, & ME and MATH 10≤ 2.250 and MATH 4≤ 2.750 and MATH 5≤ 2.500, then **RETAIN**.
R2. If PHY 1LEC > 2.750 and course = CE, ECE, & ME and MATH 10≤ 2.250 and MATH 4≤ 2.750 and MATH 5> 2.500, then **RETAIN**.
R3. If PHY 1LEC > 2.750 and course = CE, ECE, & ME and MATH 10≤ 2.250 and MATH 4 > 2.75, then **RETAIN**
R4. If PHY 1LEC > 2.750 and course = CE, ECE, & ME and MATH 10> 2.250 and MATH 10≤ 2.500,then **RETAIN**.
R5. If PHY 1LEC > 2.750 and course = CE, ECE, & ME and MATH 10> 2.500 and MATH 10≤ 2.750, then **RETAIN**.
R6. If PHY 1LEC > 2.750 and course = CE, ECE, & ME and MATH 10> 2.750 and course = CE & ME, then **RETAIN**.
R7. If PHY 1LEC > 2.750 and course = CE, ECE, & ME and MATH 10> 2.750 and course = ECE, then **NOT RETAIN.**
R8. If PHY 1LEC > 2.750 and course = EE and PHY2 LAB≤ 2.250, then **RETAIN.**
R9. If PHY 1LEC > 2.750 and course = EE and PHY2 LAB> 2.250, then **RETAIN.**

CONCLUSION

- The two top predictive models based on auto classifier were C 5.0 and CHAID based on their overall accuracy. The dataset were divided into training set comprised of 70% for validation and 30% of testing set for evaluation. 10-fold validation were used and to eliminate that the difference of the two models was due to chance, a t-test was conducted. For evaluation, overall accuracy, sensitivity and specificity were calculated. Based on the result, it was concluded that the two models were accurate and valid.
- The recommended predictive model was CHAID based on the calculated result of its overall accuracy, sensitivity and specificity. The predictive model is interpreted through the decision tree by recursive splitting of students' grades in math and physics into smaller subgroups and the rule set was generated based on the decision tree. Thus, CHAID model is the best early warning system for TUPM to detect the students who are academically at risk based on the predictors from math and physics.
- Based on CHAID information, TUPM should prepare a specific intervention program based on the specific needs of the students.
- Establish a close collaboration among TUPM administrators to map out different teaching strategies based on the needs of the students.
- Compensate the faculty members and staff who will be involved in the intervention program so they will be committed to render their services and monitor the performance of each student.

REFERENCES

[1] Abdulsalam,S. O., Babatunde, A. N., Babatunde, R.S. (2015). Comparative Abalysis of Decision Algoritms for Predicting Undergraduate Students' Performance in Computer Programming. *A Multidisciplinary Journal Publication , 2015(2),* 79 – 92.

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

[3] 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. https://doi.org/10.12988/ams.2015.53289 https://doi.org/10.5815/ijisa.2015.01.05

[4] Akinola, O.S., Akinkumni B.O. & Alo, T.S. (2012). A Data Mining Model for Predicting Computer Proficiency of Computer Science Undergraduate Students. *African Journal of Computing And ICT, 5(1),* 43- 52.

[5] Al-Radaideh, E.M, Al-Shawakfa, A. A. Ananbeh, (2011). "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.

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

[7] Baran, B. & Kilic, E. (2015). Applying the CHAID Algorithm to Analyze How Achievement is Influenced by University Students' Demographics, Study Habits, and Technology Familiarity. *Educational Technology &Society, 18(2), 323 – 335.*

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

[9] 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.

[10] Elsayad, A. M.,& Elsalamony, H. A. (2013). Diagnosis of breast cancer using decision tree models and svm. *International Journal of Computer Application,* 0975-8887, 83(5).

[11] Han, J., Kamber, M.,& Pei, J. (2011). *Data Mining Concepts and Techniques* (3rd edition) Morgan Kaufmann Publishers, 225 Wyman Strreet Waltham, MA 02451

[12] IBM SPSS Modelar 18.0 Algorithms Guide. (2016). IBM Corporation.

[13] IBM SPSS Modeler Version 18 Modeling Nodes. (2016). IBM Corporation.

[14] Kabakchieva, D. (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.

[15] Kakavand, S .,Mokfi, T., & Tarokh, M.J. (2014). Predition The Loyal Student Using Decision Tree Algorithms. *International Journal of Information and Communication Technology Research, 4(1), 32 – 37.*

[16] Nurcan, Ö., Metin, K.E., & Eyüp, K. (2015). Predicting Financial Failure Using Decision Tree Algorithms: An Empirical Test on the Manucfacturing Industry at Borsa Istanbul. *International Journal of Economics and Finance, 7(7),* 189 – 206.

[17] Ogar, Emmanuel, N. (2007). Student Academic Performance and Evaluation Using Data Mining Techniques. Fourth Congress of Electronics, Robotics and Automotive Mechanics. DOI 10.1109/CERMA. 2007.78.

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

[19] Zomorodi-Moghadam, M., Das, R., Abdar, M. (2017). Peformance analysis of classification algorithms on early detection of liver disease. *Expert Systems with Application , 67,* 239-251.