Development of a Predictive Decision Support System for Student Graduation using a Decision Tree Algorithm

Ma. Corazon G. Fernando, Marie Luvett I. Goh, Ace C. Lagman, Jay-ar P. Lalata, John Heland Jasper C. Ortega, Maria Vicky S. Solomo

Department of Information Technology, FEU Institute of Technology, Manila, Philippines.

mgfernando@feutech.edu.ph; migoh@feutech.edu.ph; aclagman@feutech.edu.ph; jplalata@feutech.edu.ph; jcortega@feutech.edu.ph; mssolomo@feutech.edu.ph

Abstract - Decision tree algorithm is one of the most popular classification algorithm in supervised learning that has been applied in many areas including mining academic data. One of the main issues under educational data mining is student graduation. In the Philippines According to Philippine Authority of Statistics, there is an imbalance between the student enrolment and student graduation. Students who are full time freshman students do not graduate on time. The study focused on the application of the decision tree algorithm in predicting student graduation by generating rules sets that could early predict and identify students who are prone of not having graduation on time, so proper remediation and retention policies can be formulated and implemented by institutions.

Keywords - Information system, Decision Tree Algorithm, Educational Data Mining, Decision Support

I. INTRODUCTION

At present, Educational Data Mining (EDM) is more particularly focused on developing tools that can be used to discover patterns in academic data. The goals of this area is to improve the services and extract hidden patterns concerning of students’ behavior and academic performances (Baker, 2010).

EDM is more concerned in exploring huge amount of data in order to identify patterns about the microconcepts involved in learning. This area of EDM is often referred to as Learning Analytics – at least as it is commonly compared to more prominent data mining approaches which process data from large repository for better decision-making. One main research focus of educational data mining is student graduation (Ahmed, 2014).

The student graduation rate is the percentage of a school’s first-time, first-year undergraduate students who complete their program successfully (Raju, 2012).

According to National Center for Education Statistics, almost half of the students who are full time freshmen do not graduate on time. The colleges and universities consisting of high leaver rates go through loss of fees and potential alumni contributors.

Addressing this problem is critical as high leaver rates lead to loss of potential alumni contributors and loss of school revenues. The study adapts Seidman’s Formula that Retention can be equated as Early Identification + (Early + Intensive + Continuous) Intervention. The Seidman formula and shows that early identification of students at risk as well as maintain intensive continuous intervention is the key to increase student retention.

With this, our study aims to explore the utility of EDM in addressing the problem of student graduation for a HEI in the Philippines. This research thus encompasses the identification of at-risk students that will enable administration to provide timely help. Due to the availability of the data, the values to be processed are limited only to the following pre-college data, demographic data, entrance examination and college data sets which include first year first term grades.

A. Research Questions

The four specific research questions that this study aims to address are:

1. How feature selection technique can be used to determine significant attributes in predicting student graduation?
2. How effective decision tree algorithm in predicting student graduation?
3. What data model be created that improves the accuracy of predicting student graduation?
4. How effective and usable is the design of the Student Graduation Prediction prototype based on the evaluation of administration confidence?

II. LITERATURE REVIEW

A. Data Mining

Christopher (2010) [6] discussed that Data Mining is application of a specific algorithm in order to extract
patterns from data and transform the information into a comprehensible structure for further use.

Fayyad Et al (1996) stressed out that the Knowledge Discovery in Databases has wide applications in terms business intelligence. The data mining as part of the process of KDD deals with identifying hidden patterns in data. It involves the applications of machine learning algorithms to extract rules or equations commonly used for predictive models.

B. Decision Tree

Decision tree algorithm has been applied in multiple disciplines which include medicines, education and businesses. The greedy search approach has been used as a traditional approach. The purpose of this search is to load a full set of data into memory and the data will be partitioned into a hierarchy of nodes and leaves (Hang Yang,2013).

The CART (Classification and Regression Tree) and CHAID (Chi-square Automatic Interaction Detection are the most common decision tree methods (Nisber, Elder & Miner, 2009) [10]. These methods are used to build decision tree. The recursive partitioning is being used in the data set partitioning. The split search is the first stage of the algorithm. The dataset is partitioned in according to the best split and creates new partition rule. This procedure goes until there are no more splits.

Raju (2012) [9], discussed that student graduation can be improved by early identification of students who are at risk of not having graduation on time. Predictive modeling for early identification of students at risk could be very beneficial in improving student graduation. Research studies show that early identification of leaver students and intervention programs are key aspects that can lead to student graduation.

III. RESEARCH METHODOLOGY

A. Knowledge Discovery in Databases

We used the modified steps of Knowledge Discovery in Databases in the development of the study.

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

1. Problem Understanding. This section entails the researcher to understand the problem and what possible solutions can be proposed. This section determines the research gap that the researcher tries to solve. The goal of the research is to determine if there is a hidden pattern that can be extracted from the educational data that can be used for student graduation prediction.
3. Data Preparation. The data preparation is one of the hardest process in KDD as researchers needed to prepare the data and ready for mining. Data reliability was enhanced in this stage. It includes data cleaning, such as handling missing values. A complex SQL statement was used to extract needed attributes from the data sources. The data was being transformed in to a proper format using data pre-processing technique such as imputation using a data mining tool.

Feature Selection Technique - To determine the statistical significance of a predictor the p value was used. The predictor is statistically significant when a p value is less than the significance level (Agresti, 1990).

4. Data Mining. This stage was the application of machine learning algorithm to generate a data model. A data model can be used for predictive and profiling data sets. To determine the predictive model, decision tree algorithm was used. This entails to produce an if then else rule statements from the datasets. The datasets were divided into two parts, training and testing data. The training data was used to develop a model using machine learning algorithm. To determine the accuracy of the model, it will be then tested to testing data.

The decision tree with a binary target graduation has two outcomes, YES or NO or it can be applied as 1 or 2. Input variables such as demographic student’s data, entrance examination and first year first term grades can be in a form of categorical and binary values. Categorical values can be applied on first year first term student grade and entrance examination results. Binary values can be applied on some of demographic data of student examples are gender, location, scholarship and financial aid.

To extract rule sets the following equations were performed. To compute the entropy the formula was used

4-a. Entropy

Entropy $H(S)$ is a measure of the amount of uncertainty in the (data) set $S$ (i.e. entropy characterizes the (data) set $S$).

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Where,

- $S$ – The current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm)
- $X$ – Set of classes in $S$
- $p(x)$ – The proportion of the number of elements in class $x$ to the number of elements in set $S$. When $H(S) = 0$, the set $S$ is perfectly classified (i.e. all elements in $S$ are of the same class).

4-b. Information gain

Information gain $IG(A)$ is the measure of the difference in entropy from before to after the set $S$ is split on an attribute. To compute the information gain the formula was used.

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

Where,

- $H(S)$ – Entropy of set $S$
- $T$ – The subsets created from splitting set $S$ by attribute $A$ such that
  - $p(t)$ – The proportion of the number of elements in $t$ to the number of elements in set $S$
  - $H(t)$ – Entropy of subset $t$

5. Evaluation of Discovered Knowledge. The most common way to evaluate models is to verify their performances on the test datasets. Evaluation of the models can be easily determined by observing the number of correct predictions to the total number of predictions.

To extract rule sets the following equations were performed. To compute the entropy the formula was used

5-a. Accuracy Computation Formula:

$$Accuracy = \frac{TN+TP}{TP+FP+TN+FN}$$

Equation 1.

The true positive (TP) and true negatives (TN) are correct classifications. The accuracy can be computed by adding the correct predictions of true positive and true negative divided by the total number of items.

TABLE II. CLASSIFICATION RATE TABLE

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Performance Measure of the Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>True Positive</td>
</tr>
<tr>
<td>No</td>
<td>False Negative</td>
</tr>
</tbody>
</table>

The true positive (tp) refers to the number of outcomes that are correctly predicted in terms of graduates on time while true negative (tn) refers to the number of outcomes that are correctly predicted in terms of graduates not on time. This accuracy formula can be found based on the confusion matrix table.
5-b. Error Estimation Formula:

Error Estimation = \frac{1}{100} - \text{Accuracy} \quad \text{Equation 2.}

6. Use of Discovered Knowledge. The discovered classification rules was embedded to the system to determine students who would have learning difficulty in student graduation. A decision support prototype was developed containing the extracted rules sets generated by the decision tree algorithm.

IV. RESULTS AND DISCUSSION

A. Significant Variables

Answering Question No.1 entails to use logistic regression where all predictors are significant, \( p<.05 \) was considered. The resulting significant predictors of status were then processed using Decision Tree Algorithm.

Analysis of the data reveals that eight variables significantly predicts graduation status, namely: gender (B=.888, p<.01, OR=2.44), scholarship (B=.999, p<.01, OR=.36), verbal (B=.307, p<.01), abstract (B=.25, p<.01, OR=1.29), algebra (B=.289, p<.05, OR=1.54), IT fundamentals (B=.43, p<.05, OR=1.33), programming (B=.567, p<.01, OR=1.77) and values (B=.423, p<.01, OR=1.53).

Gender has a positive B coefficient, indicating that female students (coded 2) have higher odds of graduating than male students (coded 1). Female’s odd of graduating is 2.44 times higher than males. On the other hand, the negative B coefficient in scholarship indicates that students without scholarship (coded 2) have lower odds of graduating as compared to those with scholarship (coded 1). The odd of graduating for those with scholarship is almost three \( (\frac{1}{.36}=2.78) \) times higher than those without scholarships.

The B coefficients for verbal analogy and abstract reasoning as components in the entrance examination of the university are positive, indicating that the higher the scores of the students in the verbal analogy and abstract reasoning components, the higher the likelihood that they will graduate to the program that they enrolled. The odd ratio of 1.29 for both verbal analogy and abstract reasoning indicates that for every one (1) point increase in the score in verbal analogy or abstract reasoning, the likelihood of finishing the degree increases by 1.29 times.

Testing plays an important role to the success of the quality of a software. The purpose of the software testing is to discover the defects, improve the quality, reliability and performance of the system and to ensure the software works perfectly based on its functionalities. The proponents used Beta Testing to test the software. In Beta Testing, the software is released to a group of people so that further testing can ensure the product has few faults or bugs. The software was made available to the open public to know the feedback of the users.

The proponents used survey forms to perform the beta testing activity. The researcher used survey questionnaire to evaluate the system. Likert scale was used to address the measurement of the perception of the respondents shown in Table 1. The researchers used random sampling technique. The respondents of the survey consists of two categories: IT and Non-IT respondents.

There were 20 IT respondents and 20 Non-IT respondents. The respondents should be an inhabitant of Manila, regardless of their age and occupation. The proponents used ISO/IEC 9126 software quality model that categorizes software quality into six characteristics (factors) for the questions on the survey form were it is grouped based on the attributes of the software quality factors: Functionality, Reliability, Usability, Efficiency, Maintainability and Portability.

B. Accuracy Results of Decision Tree Algorithm in Prediction

The empirical testing for Decision Tree involves two main parts: (i) accuracy result determination using training data (ii) decision tree data model.

Decision Tree analysis on all the hypothesized predictors were tested using Chie squared Automaton Interaction using data Mining tool Waikato Knowledge Analysis and SPSS. Table IV presents the accuracy results of the decision tree algorithm in predicting student graduation.

Answering Question No.3 entails to determine data models that can be generated by decision tree algorithm.

<table>
<thead>
<tr>
<th>Decision Tree Algorithm</th>
<th>B</th>
<th>S.E.</th>
<th>World</th>
<th>Odds Ratio (OR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (X1)</td>
<td>0.888</td>
<td>0.196</td>
<td>20.613</td>
<td>2.44</td>
</tr>
<tr>
<td>Scholarship (X2)</td>
<td>0.999</td>
<td>0.283</td>
<td>12.243</td>
<td>0.36</td>
</tr>
<tr>
<td>Verbal Equivalence (X3)</td>
<td>0.307</td>
<td>0.081</td>
<td>14.234</td>
<td>1.29</td>
</tr>
<tr>
<td>Abstract Equivalence (X4)</td>
<td>0.25</td>
<td>0.076</td>
<td>10.726</td>
<td>1.29</td>
</tr>
<tr>
<td>Algebra (X5)</td>
<td>0.289</td>
<td>0.138</td>
<td>4.403</td>
<td>1.33</td>
</tr>
<tr>
<td>IT Fundamentals (X6)</td>
<td>0.43</td>
<td>0.131</td>
<td>10.846</td>
<td>1.54</td>
</tr>
<tr>
<td>Programming (X7)</td>
<td>0.567</td>
<td>0.13</td>
<td>19.129</td>
<td>1.77</td>
</tr>
<tr>
<td>Values Education (X8)</td>
<td>0.423</td>
<td>0.133</td>
<td>10.084</td>
<td>1.53</td>
</tr>
<tr>
<td>Constant</td>
<td>5.716</td>
<td>0.842</td>
<td>46.119</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Table IV presents the accuracy results of the decision tree algorithm in predicting student graduation.
C. CHAID Results

The rule sets derived from the decision tree algorithm using CHAID, Chi-square Automatic Interaction Detection, a decision tree technique method which consists of 17 rules for non-graduates on time (coded 0) and for graduates (coded 1).

**TABLE V. RULE SET OF DECISION TREE FOR NON-GRADUATES**

<table>
<thead>
<tr>
<th>Rule</th>
<th>IT Fundamentals</th>
<th>Scholarship</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&gt;2.50 and &lt;=3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&gt;3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>&gt;3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**TABLE VI. RULE SET OF DECISION TREE FOR GRADUATES**

Finally, the fourth research question addresses the issue of measuring the perspectives of the end-users with regard to the software quality characteristics of the developed prototype consisting of the decision tree algorithm. A questionnaire was circulated to guidance officer and head of the Information Technology Department and predictive analytics expert who validated the results asking them to rate the prototype software. Response for the items was measured using five-point Likert scale. The figure below indicates the simulation of input values and the prediction of the algorithm.

**TABLE VII. SUMMARY OF THE WEIGHTED MEAN OF THE FOUR (4) CRITERIA FOR DESCRIPTIVE AND PREDICTIVE ANALYTICS’ OF STUDENT GRADUATION PROTOTYPE**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Expert’s Response Weighted Mean</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functionality</td>
<td>4.55</td>
<td>Very Acceptable</td>
</tr>
<tr>
<td>Design</td>
<td>4.55</td>
<td>Very Acceptable</td>
</tr>
<tr>
<td>Usability</td>
<td>5.00</td>
<td>Excellent</td>
</tr>
<tr>
<td>Reliability</td>
<td>4.60</td>
<td>Very Acceptable</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>4.69</strong></td>
<td><strong>Very Acceptable</strong></td>
</tr>
</tbody>
</table>

Overall the Descriptive and Predictive Analytics’ of Student Graduation Prototype based on the respondents’ response recorder a mean performance of 4.69 with an interpretation of Very Acceptable.
V. CONCLUSION, RECOMMENDATION AND FUTURE WORKS

The study aimed to embed rule sets extracted from educational data using decision tree algorithm. This can be used as a basis in creating a predictive analytics software prototype for student graduation. This will early identify students who are vulnerable of not being able to graduate on time so proper retention policies can be formulated by the administration. There are only few rules for graduates recorded due to limited number of instances.

Decision Tree Algorithm has an accuracy rate of 86.77 in predicting student graduation and the overall acceptability of the Descriptive and Predictive Analytics’ of Student Graduation Prototype based on the respondents’ response recorded an overall mean of 4.69 which has an interpretation of Very Acceptable and concluded that the software can be now used for implementation.

Recommendation and Future Works

The system has plenty of potential for further improvements that future researchers might want to follow through: The continuous study of student graduation rate for new incoming data sets so data it can become voluminous and new patterns can be discovered. Additional attributes can be added as attributes which include financial capability, interest and skills. The study can be applied to other disciplines or courses. The report generation of the prototype can be improved by having archives of reports every year. Possible algorithm combinations can be applied to test sets of data in improving the accuracy of the predictions.

REFERENCES