A Statistical Analysis for Predicting 2nd Year, 3rd Year and 4th Year College Grades Using Least Squares Linear Regression and Logistics Regression while Implementing an Iterative Error-Checking Framework

An Independent Study

Presented to

The Faculty of the Computer Studies Division

Ateneo de Davao University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science in Computer Science

By

Doronila, Mark Jeo

SCHOOL OF ARTS AND SCIENCES

ATENEO DE DAVAO UNIVERSITY

MARCH 2012

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Background of the study

The College is one of the most important phases of a student. This phase, to some, is considered the final milestone in their education. A student in this phase naturally goes to a state of self-assessment from time to time, to check whether they are on the right track. In this stage, self-assessment is one of the helpful ways in ensuring a smooth run in College. It cannot be stressed enough that decision-making in this stage is crucial. Some students are still undecided on what course or field of study they should take. This indecisiveness usually has bad consequences for them. Early dropout, forced to shift at the early years of college and dissatisfaction due to loss of interest within the field of study are among the common consequences they face. Aptitude tests are given to students in order to help them identify their strengths and weaknesses, but these kinds of tests give results on a shallow level. It does not test the knowledge acquired from a school setting. It does not relate to the past academic grades the student acquired. There needs to be another guide or process so that this indecisiveness will somehow alleviate or clear up. The study shows that with the student record of the Computer Science Division students of the Ateneo de Davao University, an accurate model for grades and passing rate is predictable through data mining concepts, statistical regression techniques and an iterative framework.

1.2 Statement of the problem

The study intends to use data mining concepts and statistical analysis to filter through required variables, i.e. major subject grades of 1st year, Computer Science, Information Technology and Information Systems students, as input for the predictive model, wherein the model to be used is the linear least square method and logistics regression. The following processes would result in an accurate grade and a passing percentage based on the filtered data.

The resulting grade from the least square method is comparable to the MSCA grade.

The study will tackle the following problems:

1. Finding the suitable statistical model and data mining concept for acquiring the predicted grades
2. Finding the most effective model and data mining concept for predicting the passing rate of a student in a course
3. Develop a simple iterative error-checking framework for the study.
4. Determine a reasonable margin of error.

1.3 Objective of the Study

For the study to achieve its goals, it needs a suitable data mining concept. The study does not need a data mining technique itself, but rather it only needs a concept to have a basis on how to initiate data gathering and data processing. Having a data mining concept can also be crucial in having an error checking iterative framework The next problem to tackle is finding an effective statistical model for the

passing grade and passing rate prediction. The statistical model should be able to handle and compute the data gathered and give accurate results. After determining the two main components of the study, a simple iterative error-checking framework should be determined for the study. It must be simple and adaptive. It should be adaptive in a sense that it can consider new data and does not rely on a single unchanging model for passing rate prediction. As for all predictive models, the margin of error should be considered. In this study, many factors can affect school performance. As a result, results from the predictive model may not be or close to 100%. As such, a reasonable margin of error should be determined. This is needed so that every percentage can be accounted for, thus resulting in an accurate result.

1.4 Significance of the study

The study sought to use data mining concepts and statistical data modeling to predict the success rate of a 1st year Computer Science, Information Technology or Information Systems student in the succeeding year levels based on the student records, where the basis for passing is the MSCA grade. The study can contribute in cultivation and growth of the student in many ways. Its main importance is that it can help choose the right course of action, given the predicted grade of the student. It can serve as a guide if ever the student feels conflicted or confused on what subject the student should focus on. Furthermore, it can assess the potential of the student in a certain field. It can also gauge where the student excels and where the student is doing poorly; giving the student a layout on where he/she can concentrate. It would be of great help since some students lack self-assessment or make poor decisions that

they later on regret. This study is important to students because one of the things that slow their progress to graduating college is constant shifting of courses. Through this study, the proponent hopes that such consequence will lessen.

1.5 Scope and Limitations of the study

The study considers a straightforward approach on predicting grades that will determine the passing rate of the student. The independent variable in the study is the 1st year level and the dependent variables are the succeeding year levels. It will tackle factors that will directly influence the predictive model. Therefore, it will only consider the student's grades as the main contributor to the passing rate of the student in the succeeding year levels. Other factors such as teacher factor, learning environment, parent's wage, school supplies, etc. will not be included. The other factors are included in the margin of error, so that they are accounted for without it influencing the output of the predictive model. The results of the linear least square method predictive model will encompass the four years of college.