Data Mining Approach for Analyzing Graduating Students' Academic Performanceof New Era University – Bachelor Science in Computer Science

Mark R Paraiso, Harry V Torres,
Ralph Chester D Follosco and Jonelle Curioso
College of Computer Studies
New Era University
Quezon City Philippines
{markrio111, torresharry2, rcdfollosco and jonellecurioso}
@gmail.com

Dr. Albert A Vinluan
New Era University
Quezon City Philippines
aavinluan@neu.edu.ph

Abstract— **Data mining has been utilized as a part of diverse studies to recognize and separate new data to enhance a particular objective particularly in huge number of commercial enterprises that still needs to concentrate on different variables which influence the general result of a phenomenon. Utilizing instructive data mining can be an extraordinary thought to create and find new learning in an understudies' execution in some point.**

The fundamental targets of this exploration is utilizing data mining to investigate the understudies' execution of the graduating students of Computer Science course precisely. This examination expects to enhance his/her scholarly execution furthermore to dissect what are the issues that cause in their disappointments, likewise to check whether the understudies enhanced or not in a previous couple of years.

In this study, the researchers utilized different direct relapse and time arrangement in assessing the scholastic execution of the graduating students in Computer Science of New Era University, from academic year of 2011-2015. In different direct relapse, the model has a mean total rate blunder in year 2014 and 2015. It is likewise discovered that the connections between the variables are high. In time arrangement investigation, the scientists figured out that the understudies' scholarly execution in every year from year 2011 to 2012 are in exchange build and diminishing which implies that there is simply

Slight increment ofExecution on the understudies. Utilizing the said procedures, the scientists have the capacity to dissect the scholarly execution of the Computer Science Students.

Keywords- Academic Performance, data mining, educational data mining.

I. INTRODUCTION

Students' academic gain learning execution is influenced by various elements including sex, age, showing stuff, students educating, father/guardian, social monetary status, local location of students, medium in guidelines in schools, educational cost pattern day by day study hour and accommodation as hostelries or day scholar. Many researchers conducted detailed studies about the factors contributing student performance at different study levels a student's instructive achievement unexpected intensely on economic wellbeing of student's parents/guardians in the society. [1]The same that parent's income or social status positively affects the students test score in examination. [2] The higher education performance is depending upon the academic performance of graduate students. [3]

Data Mining (DM) refers to the efficient discovery of valuable, non-obvious information from a large collection of data. One application of DM is Education Data Mining (EDM). Education Data Mining refers to techniques, implements, and research designed for automatically extracting meaning from astronomically immense repositories of data engendered by cognate to people's learning activities in educational settings.

There are increasing research interests in using data mining techniques in education. This emerging field called Educational Data Mining. It can be applied on the data related to the field of education. One of the educational problems that are solved with data mining is the prediction of students' academic performances. Prediction of students' academic performance is more beneficial for identifying the low an indicator of academic performance students. Student's retention is an indicator of academic performance and enrolment management of the university. The ability to predict a student's performance is very important in educational environments. Students' academic performance is International Journal of Data Mining and Knowledge Management Process based upon different factors like social, personal, psychological and other environmental variables. [4] A very promising tool to achieve this objective is the use of Data Mining.

In some cases, students struggle to get out of their own way to achieve academic goals. Poor study habits lack of motivation and poor preparation negatively impact stunt performance. However, students also face more indirect conflicts with high academic achievement form areas like finances and family support.

A. Scope and significance

The study covers Analyzing Graduating Students' Academic performance of New Era University-BSCS Course. The study used Rapidminder in the analysis of the student performance whether the students improved or not. The researchers used the data from year 2011-2015 as the training data year 2012 as the test data.

The proposed study which is about the analysis of student academic performance will benefit the following:

*Professors.*The Study will give them information about the factors that affects the performance of the students of the Computer Science.

Students. To give them knowledge about factors that affects their class performance.

Future Researchers. This study will serve as a guide towards new topic that can be created in future.

B. Statement of the problem

- What are the factors involved that may contribute about Data Mining approach for analyzing graduating Students' Academic Performance of the New Era University – Bachelor of Science in Computer Science Course?
- What are the trends and patterns of the performance of the NEU - BSCS Course?
- What is the level of acceptability of the Data Mining Approach for Analyzing Students' Performance of New Era University – BSCS Course in terms of model accuracy?

II. RELATED WORK

In the past years, there has been a vast interest in using data mining in educational purposes. Data mining gives a more precise understanding when it comes in researches in education which gives specific requirements other than the other. [5]

One of many other problems that data mining can solve is the prediction of students' academic performances with a goal to predict an unidentified variable (result, grades or scores) that describes the students. The estimation of the performance of the students consists of monitoring and guiding them through the teaching process and assessment. Assessment is the measurement of studying outcomes which also indicates the level of students' performance which also concludes that examinations play a significant role in students' life and future.

Examination plays a vital role in any students' life as said earlier. Therefore it becomes essential to predict whether the student will pass or fail in the examination. If the prediction says that a student tends to fail in the examination prior to the examination then extra efforts can be taken to improve his studies and help him to pass the examination. In this connection, the objectives of the present investigation were framed so as to assist the low academic achievers in engineering and they are:

- Generation of data source of predictive variables.
- Identification of different factors, which affects a student's learning behavior and performance during academic career.
- Construction of a prediction model using classification data mining techniques on the basis of identified predictive variables; and
- Validation of the developed model for engineering students studying in Indian Universities or Institutions [6]

Regression is used in classification applications though logical regressions as well as forecasted reports measured using the least square or other methods. Non-linear data can be transformed into useful linear data and analyzed using linear regressions. [7]

Three types of data mining approaches were conducted in study. The first approach is descriptive which is concerned with the nature of the dataset such as the frequency table and the relationship between the attributes obtained using cross tabulation analysis (contingency tables). In addition, feature selection is conducted to determine the importance of the in prediction variables for modeling study outcome. The third type of data mining approach, i. e. predictive data mining is conducted by using four different classification trees. Finally, a comparison between these classification tree models was conducted to determine the vest model for the dataset. [8]

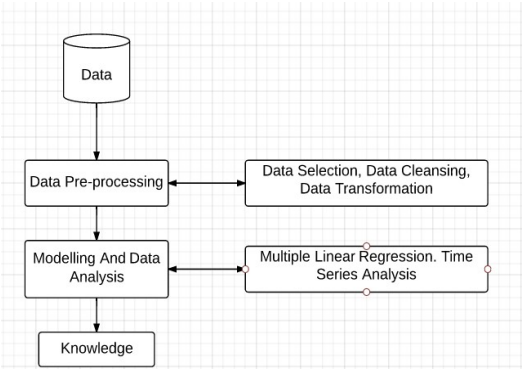

Figure2: Conceptual Framework

The Conceptual Framework above will be used in Analyzing Graduating Students' Academic Performance of New Era University – Bachelor Science in Computer Science. The gathered data are first place in a database for referencing. Using that information, it should go under a data processing which consists of the following methods such as data selection,

data cleansing and data transformation. In this phase of the framework, the data are cleansed meaning, this phase of detecting outliers, data which are not included in building the model. The next phase is the modeling phase in which the techniques and algorithm in data mining are applied. In this case, the researchers use the Regression and Time Series Analysis to build a model out of the processed data. And the last phase is the knowledge or the result of the data mining.

III. METHODOLOGY

This research about Data Mining Approach for Analyzing Students' Academic Performance of New Era University-BSCS Course will use constructive research design. The researcher's will determine the outcome whether the students in CS did increase or decrease their academic performance with the help of data mining techniques that will be used. Using Document analysis for a social research method which is used as a tool for obtaining relevant documentary evidence to support and validate facts stated in a research, especially during the chapter of literature interview. The exercise involves analytic reading and review of lots of review material. This is valuable to help the researcher to extract the relevant portions that can be deemed as statement facts to validate individual research objectives. The sources reading materials have to be acknowledged to prevent plagiarism. Data mining techniques are used to build a model according to which the unknown data will try to identify the new information. Regardless of origin, all data mining techniques show one common failure: "automated discovery of new relationships and dependencies of attributes in the observe data.[9]

The researchers will use constructive research design. The researcher's will determine the outcome whether the students in Bachelor of Science in Computer Science did increase or decrease their academic performance with the help of data mining techniques that will be used.

Regression, this technique used for predicting a continuous numerical outcome such as customer lifetime value, house value, process yield rates. The researchers used this technique to determine the student's performance like learning the percent of improvement.

Linear regression analyzes the relationship between two variables bi fitting a linear equation to observed data. One variable is the independent variable and the other, dependent variable. This model also associates the two variables but it doesn't mean that the one causes the other. Both variables are never the effect of one another. A valuable numerical measure of association between two variables is the correlation coefficient which values between -1 and 1 signifying the strength of association of the observed data for the two variables. The linear regression line has an equation of:

$$Y = a + bX$$

Where:

X = the independent variable

Y = the dependent variable

b = the slope of the line; and

a = the intercept

In finding the correlation r, the Pearson's product moment equation is used.

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

Where:

n is the number of pairs of scores

$\sum xy$ is the sum of the products of n

$\sum x$ is the sum of x scores

$\sum y$ is the sum of y scores

$\sum x^2$ is the sum of squared x scores

$\sum y^2$ is the sum squared y scores

The researchers used correlation to find the relationship of the factors involved in analyzing the students' academic performance. In this way, the researchers can determine if one factor is related to another factor and if two factors affects one

IV. RESULTS AND DISCUSSION

Table 1: Factors involved in Analyzing Graduating Students' Academic Performance of New Era University

Year	Factors and Total average of the factors from year 2011-2015	Type
2011 – 2015	CS 442	Independent
	CS 433	Independent
	CS ELEC 4	Independent
	FREE ELEC 2	Independent
	FREE ELEC 3	Independent
	Rizal	Independent
	OJT	Independent
	Final Grade	Dependent

The table 1 shows the factors involved which are used to analyze the Students' Academic Performance of New Era University – BSCS Course. These factors involved are the course subjects of the students in their course every year. Each factor is equal to the academic subjects of the students. To specify the subjects every year, the researchers labeled the subjects as for example CS442_2011 up to 2013 and so on to know the specific year this subject belongs. The FinalGrade is

the average of the subjects taken by the Computer Science Students. The researcher used FinalGrade_2015 as the dependent factor and the subjects as the factors involved in analyzing. CS442 and CS433 subject focuses on the system of their thesis A and B, refers to the research and the application base on their research. CSELEC4 subject focuses on PC Troubleshooting, Assemble and Disassemble the system unit. Free Elec2 and Elec3 subject focuses on the fundamental Japanese language (Hiragana and Katakana). Rizal subject focus on the history of Dr. Jose Rizal's life, our national hero. On-The-Job Training subject (OJT) refers to the application of all subjects and training of the students.

Table 2: Total average of the factors

Factors	2011	2012	2013	2014	2015
CS442	86.60	87.36	87.74	87.10	86.54
CS433	86.76	86.73	88.19	86.89	84.82
CSELEC4	86.21	88.64	87.97	88.56	86.33
FREEELEC2	86.59	86.86	88.11	87.91	86.69
FREEELEC3	87.19	89.61	88.24	88.34	86.31
Rizal	89.11	88.67	87.93	88.36	86.57
OJT	88.17	87.71	89.38	88.43	85.77
FINALGRADE	87.19	87.90	88.20	87.99	86.08

Table 2 presents the Time Series presented from year 2011 to year 2015. Based on the table, the researchers can be able to analyze the students' academic performance easily by finding their average in each subject per year. In CS442 subjects, it is noticed that there is a decrease of performance because the average of the students are decreasing. From year 2011, the average of the students is 86.60 while 86.54 and there is a 0.06 difference between the two, which means the students are not increasing their performance. In CS433, the average also decreases, though in 2013, the students got the highest average in the historical data with 88.19 and the lowest performing year was the year 2015. In CSELEC4 subject, the highest performing year was year 2012. And the lowest performing year was year 2011. The CSELEC4 subject has an alternate decrease and increase of performance. In FREEELEC2, the highest performing year was in year 2013. And the lowest performing year was in year 2011,which means that the students have an increase in performance and decreased after year 2014. FREEELEC3 most performing year was in year 2012 but decrease their performance after that. In OJT, the most performing year was in year 2011 and the least performing is the year 2015.

Table 3: MAPE and Multiple Linear Regression Accuracy

Mean absolute percentage factor				Year
0.125247	0.001252	0.125247	13%	2011
0.093883	0.000939	0.093883	9%	2012
0.077886	0.000779	0.077886	8%	2013
0.003677	3.68E-05	0.003677	0%	2014
0.002437	2.44E-05	0.002437	0%	2015

To check the validity of the Multiple Linear Regression model, the researchers determine the Mean Absolute Percentage Error (MAPE). In the figure above, the percent error of the model is presented per year since the researchers are analyzing the performance of the students. As seen in the table the accuratemodel are the year 2014and 2015with the percent of 0% while year 2011 is the highest error with 13%.

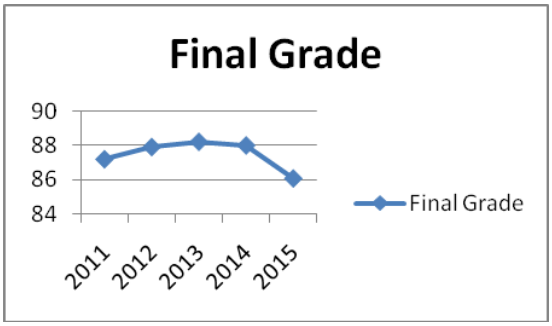

Figure2: Final Grade Graph from 2011 to 2015

On Figure 1, describes the final grade of all the students from year 2011 to 2015. Using the graph, the researchers can say that there is a big decrease of academic performance between 2011 and 2015. The peak of the graph is in year 2013 which means that the grades of the student during that time are higher than the others. The lowest point is in year 2015.

Table 4: Mathematical Model of Regression

Y = 0.926 + (0.013 * OJT_2011) + (0.017 * CS443_2012) + (-0.019 * CSELEC4_2012) + (0.024 + FREEELEC3_2012) + (-0.011 * OJT_2012) + (0.012 * RIZAL_2013) + (0.172 * CS443_2014) + (0.163 * CSELEC4_2014) + (0.176 * FREEELEC2_2014) + (0.162 * FREEELEC3_2014) + (0.178 * RIZAL_2014) + (0.144 * OJT_2014)

Using Multiple Linear Regression predict the relationship of two or more independent variables to be able to predict the outcome. In this case, the dependent factors the FinalGrade_2015 which is the average of the subjects of the Computer Science course such as CS442, CS433, CSELEC4,

FREEELEC2, FREEELEC3, RIZAL and OJT. The table 4 shows the mathematical model generated by Rapidminer.

Table 4 illustrates the applied linear model regression to the data. "Y" is the forecasted and the rest are the factors used in analyzing the students' academic performance/ the model was improved and simplified by excluding the variables with a high p-value which has no effect in building and applying the model.

V. SUMMARY

Base on the results. The following conclusions were drawn.

- The researchers used the Computer Science Students' academic subject such as CS442, CS433, CSELEC4, FREEELEC2, FREEELEC3, Rizal, and OJT. The researchers added the Final Grade factor. They named each subject for each corresponding year to prevent redundant factors and to determine them easily in the data process inside.

  Rapidminer.Inexample: CS442_2011, CS433_2011, CSELEC4_2011, FREEELEC2_2011, FREEELEC3_2011, OJT_2011, FinalGrade_2011, RATING_2011 etc.

- In analyzing the study, the researchers used Regression. In Regression, the researchers used Multi Linear Regressions since they are using two or more factors in the study. Regression is used to predict the relation of two or more independent factors to predict the outcome. The data are presented in graphs form. It is noticed that there's a big improvement in the regression line compared to the previous years of 2012, there is a positive correlation based on the students' final grade in their subjects.

Furthermore, the researchers used the model in Regression because Regression is more precise and the researchers can the improvement through the help of graph. The researchers also used a simple time series to see in simple view the average of the students and analyze it completely. Using time series, the researchers can be able to see the improvements of the students over time and easily compare it to each year. Based on the time series analysis, the students' academic performance did not improve and has an alternate decrease and increase in each year.

- The researchers used the Mean Absolute Percentage Error to know the accuracy of the Linear Model. As seen on table3, the year 2014 and 2015's linear regression has a 0% MAPE in the model which means that model is efficient while year 2011 has the lowest percent accuracy of 87%.

VI. CONCLUSION

In this part of the study, the researchers proposed the study entitled "Data Mining Approach for Analyzing Graduating Students' Academic Performance of New Era University – Bachelor Science in Computer Science". The study aimed to analyze the performance of the students using their academic grades in their subject and use two different data mining approach to find the best model that will analyze the said data.

The researchers used descriptive correlation research which aims at finding the nature, degree, and direction of relationships between variables or using these relationships to make predictions.

Recommendation

The following recommendations are offered for related research in the field of data mining and student performance:

- The researchers recommend the model to be used in other studies related to analyzing students' performance.
- Other data mining techniques can be also used in experimenting. It is advisable to use different techniques to understand and learn each differences and accuracy.
- For future studies, the researchers may encourage utilizing extra data or variables to be dissected, for example, gender, social status, educational attainment etc. in analyzing the students' academic performance and determine if these said factors contribute to their performance.

REFERENCES

[1] Graetz, B. (1995), Socio-economic status in education research and policy in John Ainley et al., Socio-economic Status and School Education DEET/ACER Canberra.

[2] Considine, G. &Zappala, G. (2002). Influence of social and economic disadvantage in the academic performance of school students in Australia. Journal of Sociology, 38, 129-148.

[3] Minnesota Measures (2007) Report on higher education performance. Retrieved on May 24,2008from www.opencongress.org/bill/110.s/642/show-139k

[4] Ahmed, A. and Elaraby, I. (2014) Data Mining: A prediction for student's Performance using Classification Method

[5] Osmanbegovic, E. and Suljic, M. (2012). Data Mining Approach for Predicting student Performance. Economic Review- Journal of Economics and business, Vol. X, Issue 1, May 2012

[6] Yadav, SkK. And Pal, S.(2012). Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification. World of Computer Science and Information Technology Journal (WCSIT) ISNN: 2221-0741 Vol.2, No.2, 51-56,2012

[7] Bara, A., &Lungu, I. (2012). Improving Decision Support Systems With Data Mining Techniques, Retrieved from http://dx.doi.org/10.5772/47788

[8] Kovacic, Z. J. (2010). Early prediction of student success:Mining Students Enrolment Data. Open Polytechnic, Wellington, New Zealand