The Relationship between Secondary Education Outputs and University Education Inputs: A Data Mining Approach

Dr. Mohammed Alghobiri[۞]

Abstract

The outputs of secondary education are the main inputs of university education. Therefore, high-quality inputs into university education can be assumed to yield excellent outputs, which is a major goal of any higher-education institution.

The present study compared the outputs of secondary education, as measured by students' grades, to those same students' grades from the first year of university. Grades for the first academic year of university are considered the most important in that they significantly affect students' subsequent academic careers by setting the tone for their pursuit of excellence [12].

To measure the relationship between student's grades at the secondary-school level and at the university level, the present study used the Pearson's correlation coefficient (r) and the Expectation Maximization (EM) algorithm to cluster students into two groups. Then study measured the characteristics of each group using the open-source software programs WEKA and SPSS. The study used the academic records of King Khalid University (KKU) students, specifically those students in the Faculty of Administrative and Financial Sciences (FAFS).

Although it is widely believed that high grades in secondary school predict high grades during the first year of university study, the present study found that there was a very weak correlation between

[۞]MIS Department, King Khalid University, maalghobiri@kku.edu.sa

grades during these two periods. The results showed that secondary-school grades do not predict grades during the first year of university. However, the results showed that the average rate indicates some variations in the characteristics of the two groups during the first year of university study. Finally, the study identified a number of recommendations for improving the quality of higher education inputs.

Keywords: Academic performance, Secondary output, Higher education output, Data mining

1. Introduction

The labour market of the Arabian Gulf countries in general and of Saudi Arabia in particular is attractive to labour from countries worldwide. At the same time, many experts agree that the Kingdom's education system at all levels has failed to meet the needs of its labour market [1]. Furthermore, the unemployment rate for Saudi Arabians is increasing, yet expatriates and foreign labourers continue to move into the Kingdom.

This increased unemployment among Saudis is linked to a numbor of reasons but mainly to the mismatch between the global market needs and the Saudi curriculum [2]. In the past, a university degree has increased a student's chances of getting the job he or she wants. However, the current increasing unemployment rate among university graduates should motivate universities to examine why this not be so true at present and to revamp their curricula with an eye toward restoring the status quo.

Since education is a circular process, each stage affects the other in a way or another. The higher-education system must improve its

science programmes, in particular, and provide more incentives for students to pursue science degrees.

2. Previous Studies

Many studies have addressed the issue of students' academic performance, some focusing on midterm exams and their effects on yearly grades. Other studies have focused on yearly grades and their correlation with those from previous academic years. Still others have investigated the effects of certain subjects on students' yearly grades [10, 13, 3].

The present study differed from previous ones in that it compared students' performance during two stages in their educational journey.

3. Study Problem

There is an apparent gap between higher-education output and labour-market needs. This may be caused by the poor academic performance of a large number of graduates. There are other possible reasons for this such as the degrees being earned do not match up with the degrees needed in the present job market. This cannot be excluded as a possible reason for graduates not being hired.

4. Study Objectives

The present study had the following aims.

1. To measure and assess the overall educational outputs of the main focuses of Saudi higher education.
2. To propose guidelines for improving the quality of the Kingdom's educational system.
3. To contribute to reducing the gap between higher-education outputs and labour-market needs.

5. Methodology

The present study used the Cross Industry Standard Process for Data Mining (CRISP-DM) approach to perform the following steps.

1. Collect data from the Deanship of Admission and the Registration Department.
2. Pre-process the data, which included:
    a. Selecting data
    b. Addressing missing data values
    c. Creating files in CSV format
3. Apply the Expectation Maximization (EM) algorithm and the Pearson's correlation coefficient.
4. Analyse the results.

5.1 Data mining - what is it?

Data mining is a process used to uncover useful data in large databases [10]. The main techniques of data mining include classifying and predicting, clustering, detecting outliers, associating rules, analysing sequences and time series and mining text. In addition, there are a few newer techniques, including social-network analysis and sentiment analysis. Some textbooks, including Han and Kamber (2000), Hand et al. (2001) and Witten and Frank (2005), include detailed introductions to data-mining techniques. In real-world applications, the data-mining process can be broken into six major phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment, as defined by the CRISP-DM [6, 10].

Data mining is primarily used by companies that are strongly consumer focused, including retail, financial, communications and marketing organizations. Data mining enables them to 'drill down' into their transactional data and determine pricing, customer preferences and product positioning with the aim of increasing sales, customer satisfaction and corporate profits. Using data mining, a retailer can

employ point-of-sale records of customer purchases to develop products and promotions that satisfy specific customer segments [4, 6].

5.2 Educational data mining

A newly emerging field, educational data mining (EDM), develops methods to uncover data from databases in educational environments [3]. The goals of EDM involve predicting students' future learning behaviours, studying the effects of educational support and advancing scientific knowledge about learning. EDM can be used by an institution to take accurate decisions and predict students' results. EDM results can help institutions improve educational processes and focus on what to teach and how to teach. Students' learning patterns can be captured and used to develop techniques for better teaching [13, 8].

5.3 Applying EDM

5.3.1 Data description

The present study used the 2012–2015 academic data of students from King Khalid University (KKU), specifically students from all departments in the Faculty of Administrative and Financial Sciences (FAFS). The particular data used were secondary-school grades and first-year university grades.

Figure 1 shows the distribution of the sample according to GPA and grade variables.

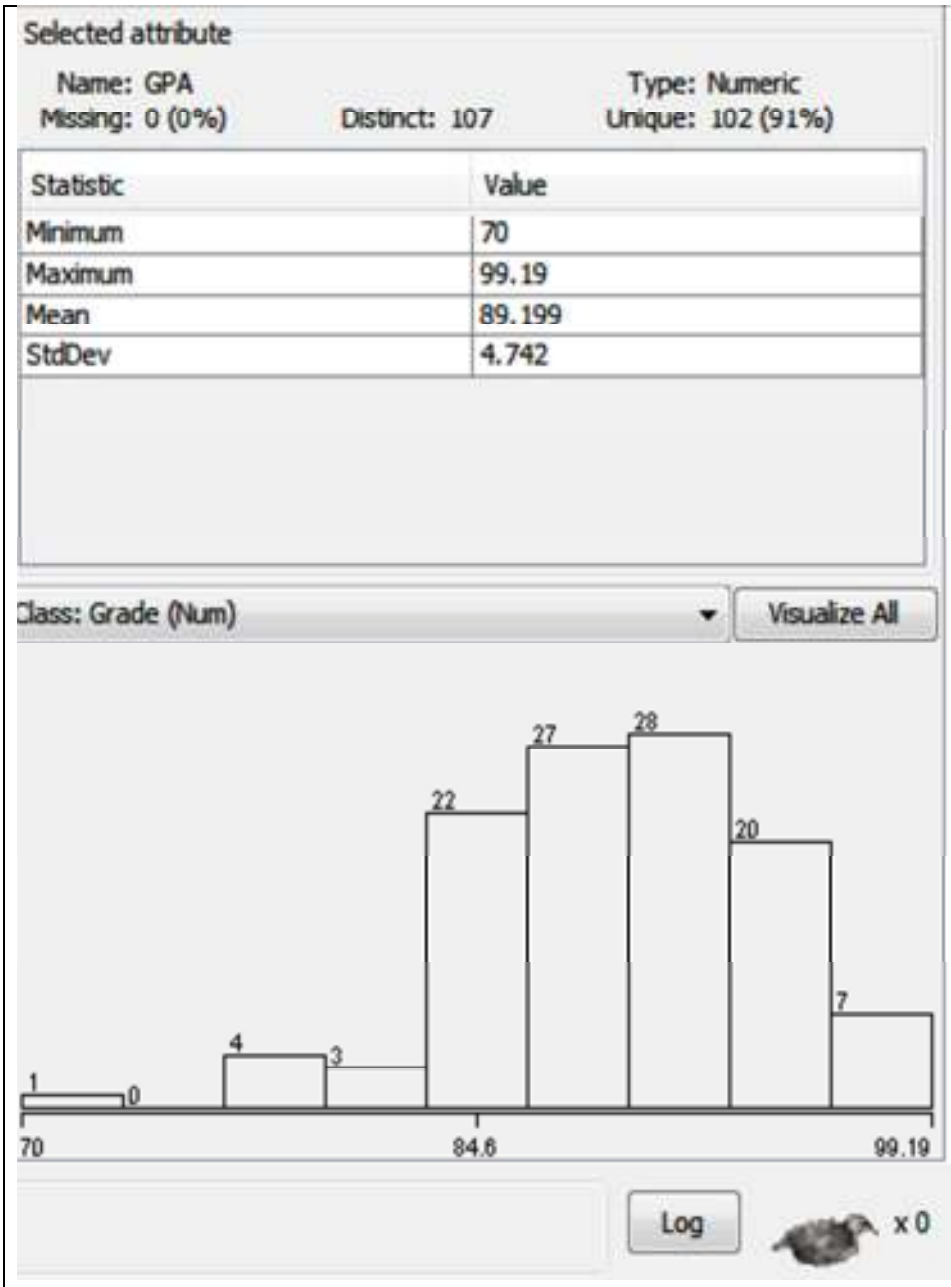

Fig 2a:distribution of GPA

Figure 1. GPA and grade variables of the study sample.

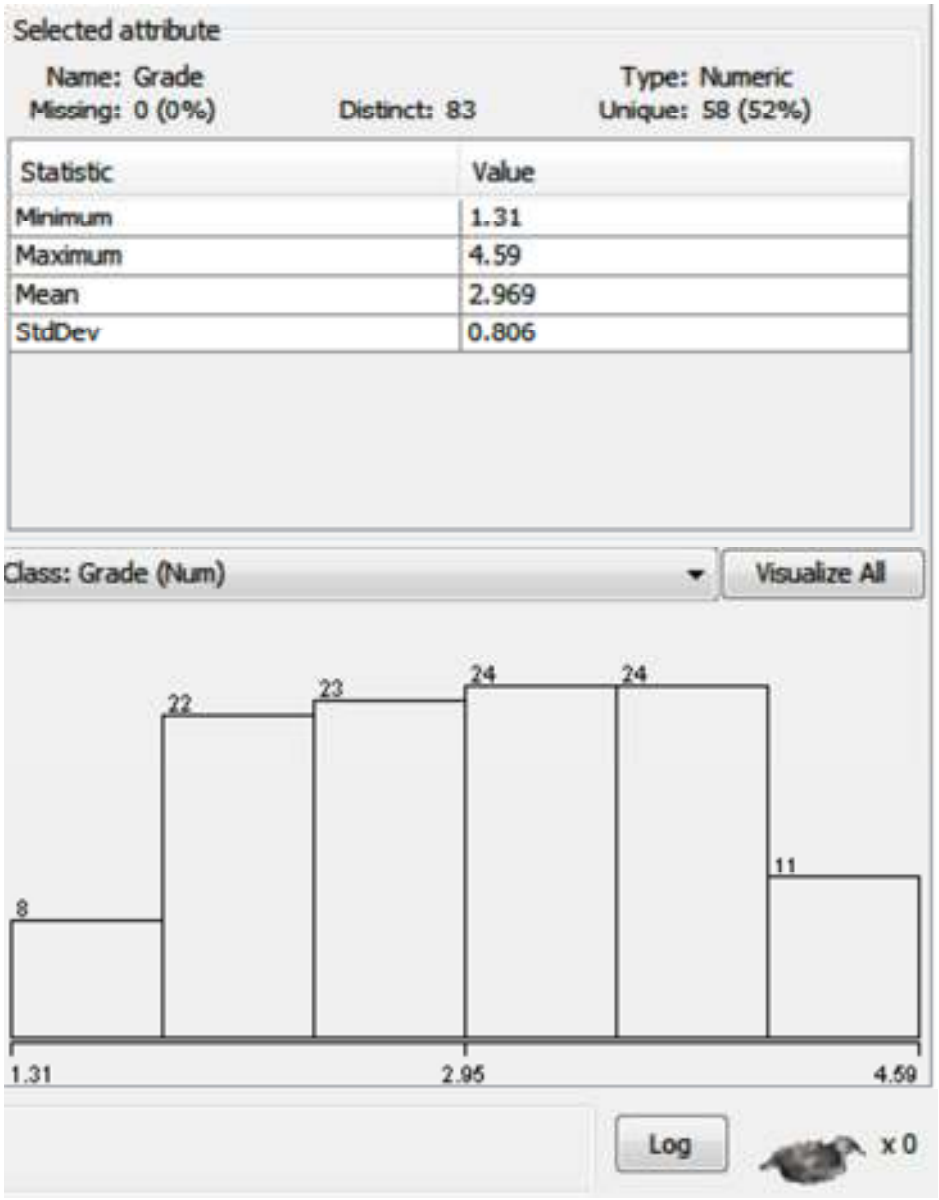

Fig 2b:distribution of Grade

5.3.2 EM algorithm

The EM algorithm is used to find the local maximum likelihood parameters of a statistical model when the equations cannot be solved directly. Typically, these models involve latent variables in addition to unknown parameters and known data observations. That is, either there are missing values among the data, or the model can be formulated more simply by assuming the existence of additional unobserved data points. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component to which each data point belongs.

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values, that is, the parameters and the latent variables, and solving the resulting equations simultaneously. In statistical models with latent variables, this usually is not possible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation [13].

The basic structure of the EM algorithm is as follows.

1. Init-step: Assign random values to the model's parameters.
2. E-step: Assign each point to the model that it best fits (these assignments are continuous, not binary).
3. M-step: Update the parameters using the points assigned in the previous step.
4. Iterate until parameter values converge [11, 13].

5.3.3 Pearson's correlation coefficient

The Pearson's correlation coefficient is used to measure the strength of a linear association between two variables. In the coefficient, a value of r = 1 means a perfect positive correlation and a value of r = -1

means a perfect negative correlation. So, this test could be used, for example, to find out whether people's height and weight were correlated, meaning that the tallerFVCCC they are, the heavier they are likely to be.

Requirements for using the Pearson's correlation coefficient are as follows.

- The scale of measurement should be interval or ratio.
- The variables should be approximately normally distributed.
- The association should be linear.
- There should be no outliers in the data.

5.3.3.1 Equation

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

4. Results and Discussion

The researcher used the Pearson's correlation coefficient to measure the relationship between students' grades in secondary school and in university. The two variables used were secondary-school grades and university grades. The relationship between the two was 0.0158, indicating a very weak relationship.

Table 1. Correlation between GPA and Grades

	GPA	Grade
GPA Pearson's Correlation	1	0.0158 *
Sig. (2-tailed)	0	0.014
N	112	112
Grade Pearson's Correlation	0.0158 *	1
Sig. (2-tailed)	0.014	0
N	112	112

* = Correlation is significant at the 0.05 level (1-tailed).

Table 1 shows the values of the specified correlation test, which is Pearson's r. Each table row corresponds to one of the variables, and each column corresponds to one of the variables. The correlation coefficient was 0.0158. The significance of this correlation was .014. Since 0.0158 is more than .014, it indicates an insignificant relationship between the two variables and, furthermore, that there was no relationship between students' performance at the secondary-school level and the university level.

The researcher also applied the EM algorithm to cluster the sample, which resulted in two groups, as shown in Figure 2.

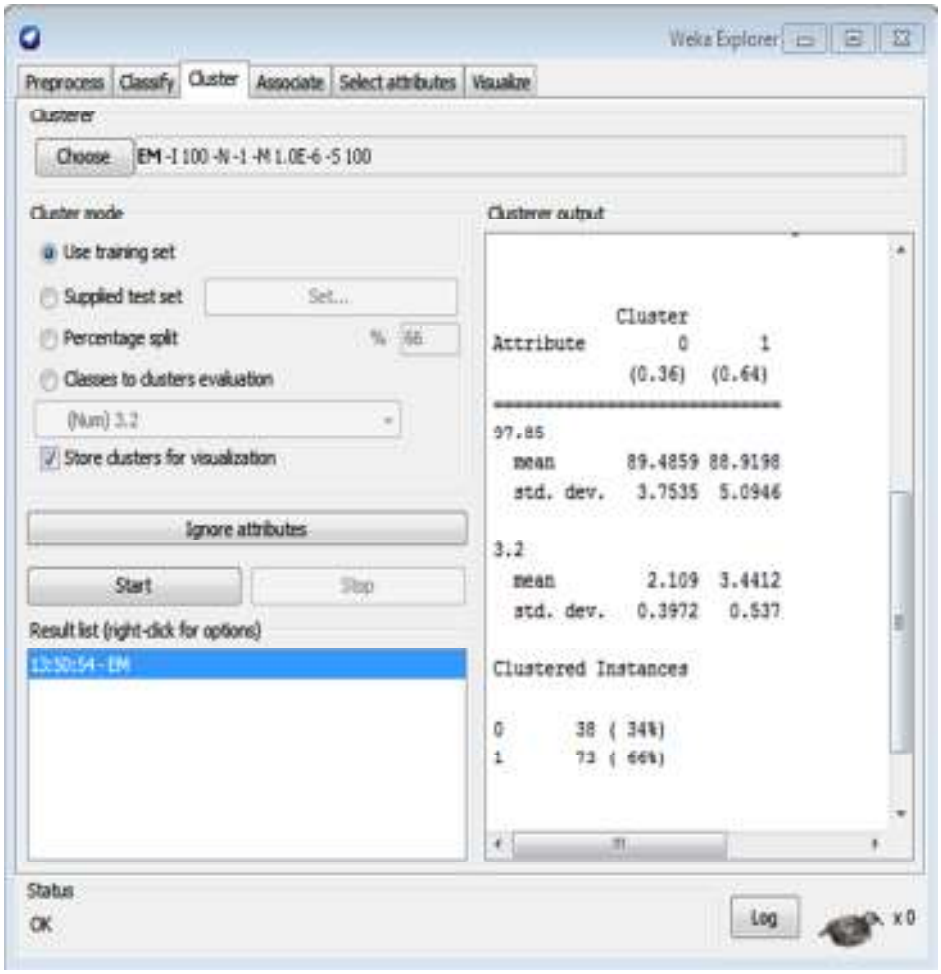

Fig 2. EM algorithm results.

As Figure 2 shows, the average grades of students in the two groups at the secondary-school level were equally matched, being 89.4 and 88.9, and there was a large standard deviation between the two groups.

The average grades of students in the two groups at the university level were not equally matched. The average of the first group

was 2.1, which is weakly equivalent, and the average of the second group was 3.4, which differs to a good degree.

The results of both the Pearson's correlation coefficient and the EM algorithm showed that students' grades at the secondary-school level were weakly correlated with their grades at the university level.

5. Conclusions

The educational systems in many countries, including Saudi Arabia, are undergoing noticeable declines in students' academic performance, necessitating thorough investigation and redress. The present study showed that most students admitted to the FAFS at KKU with excellent secondary-school grades did not attain university grades of the same distinction. Since academic performance is widely considered a major predictor for obtaining a good job and contributing to meeting labour-market needs, this decline in performance is cause for concern.

The results of the present study show that most secondary-school students pass with more than 90%, yet their failure rate during the first year of university was high. This contradiction suggests an apparent gap between the two levels of study. Either secondary-school study is too easy or the first year of university is very difficult, or both. Therefore, it is vitally important to decide which circumstance is the priority to address and plan and implement means to do so. As secondary education contributes to success in higher education, it is essential to verify the source of the problem identified so that solutions can be effected.

References

[1] Al-Asmari M.G.H. (2008), "Saudi Labor Force: Challenges and Ambitions" JKAU: Arts & Humanities, Vol. 16 No. 2, pp: 19-59 (2008 A.D. / 1429 A.H.)

[2] Alfawaz, H. & A. (2014), Would the Educational Programs help in Solving Saudi Arabia's Employment Challenges?. *International Journal of Academic Research in Economics and Management Sciences January 2014, Vol. 3, No. 1 ISSN: 2226-3624*

[3] Barros, B. & Verdejo, M. F. (2000). Analysing Student Interaction Processes in Order to Improve Collaboration: The Degree Approach. *International Journal of Artificial Intelligence in Education, 11,* pp. 221-241.

[4] Dhillon, I. S. & M. (2001). Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning, vol. 42*, pp. 143-175.

[5] Florin, G. (2011). Data Mining: Concepts, Models and Techniques. Springer-Verlag: Berlin and Heidelberg, Germany.

[6] Hand, D. M. & S. (2001), Principles of data mining. Adaptive Computation and Machine Learning Series. MIT Press.

[7] Hand, D. J., M. H. & S. (2001). Principles of Data Mining (Adaptive Computation and Machine Learning). The MIT Press: Cambridge, MA.

[8] Han, J. & K. (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc.: San Francisco, CA.

[9] International Conference on Knowledge Discovery and Data Mining, 1999, pp. 32, 41. ACM Press: New York, NY.

[10] Jiawei, H. M. & J. P. (2012). Data Mining Concepts and Techniques, 3rd Ed. Elsevier Inc.: Philadelphia, PA.

[11] Suchita, B. & R. (2013). Predicting Students' Academic Performance Using Education Data Mining. *International Journal of Computer Science and Mobile Computing, 2,* pp. 273-279.

[12] Siri A, R. B. L. & S. From School to University: A Case Study of the Nursing Students. *Hellenic Journal of Nursing Science. 2010;3:22–30*

[13] Tan, P.N. K. V. & S. (2002). Selecting the Right Interestingness Measure for Association Patterns. In *KDD '02: Proceedings of the 8th ACM SIGKDD.*

[14] Witten, I. & F. (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd Ed. Morgan Kaufmann: San Francisco, CA.