Mining Wellness and Performance Data to Identify At-Risk First-Year Engineering Students

S. A. du Plessis and H. L. Botha

***Abstract*—Management information systems, cluster analysis and neural network prediction models were used to mine performance and wellness data of first-year engineering students to identify factors that may cause underperformance. Weighted first-year averages, first and second-year retention rates, as well as throughput rates of the last ten cohorts, were analysed per race, gender, accommodation type and Grade 12 level. Sophisticated management information systems successfully identified groups that were at risk of underperformance and highlighted the possible existence of stereotype threat. Wellness data obtained from a profiling questionnaire and biographical information were combined with Grade 12 and university performance data in neural network models to predict first-year success, first- and second-year retention and success in the minimum period. Correct classification rates of above 80% were obtained and the wellness related variables played a very significant role in these predictions. A cluster analysis confirmed the relationship between wellness and academic performance.**

***Index Terms*—Cluster analysis, management information systems, neural networks, prediction, wellness**

I. Introduction

PREDICTING students' academic performance is very important for Stellenbosch University (SU) because it helps the institution to identify first-year students that are at risk of not succeeding and it informs strategic support programmes that address the needs of the struggling students. Du Plessis & Menkveld [1], [2] have built linear regression models at SU that identified a range of quantitative and qualitative variables (so-called soft factors) that affect academic performance, and have made suggestions on how academic support programmes can help to address the needs identified by these prediction models. These models only focused on the identification of factors that influence the weighted first-year average of students and did not try to predict either retention or graduation rates. It was therefore decided to expand the prediction efforts at SU by investigating whether similar combinations of quantitative and qualitative factors also affect retention and graduation rates. Astin & Oseguera's study [3] was an excellent reference point since the qualitative variables used in SU's model correspond to a large extent with theirs (SU's profiling of first-year students is partially based on Astin's Cooperative Institutional Research Program (CIRP)). It was also decided to expand the range of techniques used to make these predictions since many publications have reported the superior performance of relatively new techniques like cluster analysis and neural networks (for example [4] and [5]).

The first phase of this new approach to prediction concentrated on the engineering faculty and on utilising these new technologies to help identify first-year students at risk. Various research studies have been published that employ these newer techniques to predict retention and throughput rates in engineering programmes (for example [6], [7] and [8]). Most of the more recent studies also use both cognitive and non-cognitive variables and report impressive prediction accuracies of above 70%. Lin, Reid and Imbrie [9], for example, demonstrated an improvement in prediction capability when incorporating nine affective characteristics into an artificial neural network retention model with eleven cognitive factors – they achieved a prediction accuracy above 70%.

A further objective of this study was to utilise SU's management information systems (MIS) as an additional and common sense data mining source to identify students at risk.

The following three data mining techniques were consequently used to identify at-risk first-year engineering students:

1. An in-depth analysis of existing MIS data sets dealing with the academic performance of first-years, as well as retention and throughput rates;
2. A neural network driven cluster analysis to determine the relationship between wellness variables and first-year performance; and
3. Advanced neural networks to build prediction models for first-year performance, retention and throughput, and to identify predictors of performance.

These three investigations and their results will be discussed in more detail below. Before discussing the two neural network approaches, wellness and the way it is measured among first-year students will be explained. The quantitative and qualitative variables that were used in the cluster analysis and the neural network prediction models will also be defined.

Manuscript received March 11, 2012; revised April 05, 2012.

S. A. du Plessis is the head of the Tracking Unit within the Student and Academic Support Division at Stellenbosch University in South Africa (phone: +27-21-808-2598; fax: +27-21-808-2380; e-mail: sadp@sun.ac.za).

H. L. Botha is the Senior Director of the Student and Academic Support Division at Stellenbosch University in South Africa (e-mail: hlb@sun.ac.za).

II. Management Information Systems (MIS)

Sophisticated self-help management information systems were developed to analyse the first-year academic performance of engineering students, to determine first- and second-year retention rates and to calculate throughput rates in the minimum period. The results can be analysed per race, gender, accommodation type and Grade 12 level. Data from the last ten engineering cohorts are used. Many possible combinations of these variables exist and a multitude of scenarios can be generated, but some of the most burning questions these management information systems enable managers to answer include:

1. Are there differences in the performances between men and women?
2. Are there differences between academic achievements of students based on race[1], even if performance per Grade 12 level is taken into consideration?
3. Are there differences between the performances of students living in university residences versus those living in private wards and who are sometimes commuters and also referred to as day-students?

Answers to these questions will be discussed below. Factors causing students to be at risk will also be identified.

A. *Men versus Women in Engineering*

A study of the weighted first-year averages of male and female engineering students clearly indicates that male students outperform the female students in most cohorts, even per Grade 12 symbol. For the lower grades (below 70%), however, this is not necessarily the case – possibly due to the fact that very few women are enrolled in engineering with Grade 12 averages of less than 70% (see figures E5a-E5d in [10] that compare the weighted first-year averages of male and female students, overall, and for specific Grade 12 levels).

This trend that female and male students with similar Grade 12 results perform at different levels at university is in line with Claude Steele's theory of stereotype threat [11]. In "Whistling Vivaldi'' Steele focuses on the phenomenon of stereotype threat as it explains the trend of minority underperformance in higher education. Steele [11] discusses how identity contingencies[2] can have a significant negative effect on a person's functioning, and how these effects can explain racial and gender performance gaps in academic achievement. The above findings seem to confirm the possibility of stereotype threat, at least at first-year level for female students studying Engineering. *First-year female engineering students are therefore at risk of underperformance.*

Previously, Botha and du Plessis [12] have identified self-appraised cognition as another factor that provides some explanation why students with similar final school results perform differently at first-year university level. The underlying hypothesis was: the more students believe they are cognitively capable and equipped to achieve success at university level, the better they perform academically. SU's prediction models confirmed this (see also the findings in later sections about the influence of students' own perceptions of their abilities on their academic performance).

The throughput rates of women are not weaker than those of men in engineering at SU. Although stereotype threat may have surfaced among first-year students there is no reason to believe that it remains a problem.

However, the fact that female students consistently outperform male students at undergraduate level (see figures A1-A4 in [13]) and that there are differences at first-year level for the upper Grade 12 levels, seem to suggest that stereotype threat may indeed be present. The better performance of women in general may cancel out the differences that might have appeared otherwise if men and women did indeed perform on the same level.

B. *White versus black[3] students in Engineering*

A study of the weighted first-year averages of white, coloured and black students clearly indicates that white students outperform the other two groups in most cases; for coloured students even per Grade 12 symbol (due to very small numbers of Indian students, they are not included in this comparison). For black students, however, the results per Grade 12 symbol are mixed and it is not necessarily true that the white students outperform the black students – in fact in a large number of cases the black students actually do better. It is, however, important to remember that the number of black students per Grade 12 level is much lower than for the other two groups – a few good or bad results may therefore have a large influence on the average. This trend, that students from different race groups with similar Grade 12 results perform at different levels at university, is also in line with Steele's findings [11] and seems to confirm the possibility of stereotype threat (see figures E1a-E1d in [14] that compare the weighted first-year averages of white, coloured and black students, overall, and for specific Grade 12 levels). *First-year coloured engineering students are therefore at risk of underperformance.*

Some other evidence, although weaker than the above, for the possibility of stereotype threat for black and coloured students can be found if the first- and second-year retention rates of these groups over the last decade are studied per Grade 12 level (see figures E2a-E3d in [14] that compare the first- and second-year retention rates of white, coloured and black students, overall, and for specific Grade 12 levels). Due to the high retention rate at SU and the relative ease with which our students succeed in achieving the minimum credits to come back after their first year this measure does not come out as strong (and reliable) as the other as an indicator of possible stereotype threat.

[1] In South Africa all universities are required by law to submit annual reports to the Department of Higher Education and Training on the academic performance of their students per race. Four categories of race description are used, namely black (African), coloured, Indian and white. Sometimes black is used as a generic term that includes black, coloured and Indian students.

[2] Identity contingencies are the things you have to deal with in a situation because you have a given social identity, such as being female, male, black, white, emotional sensitive or geeky. Some identity contingencies are more serious than others, but they all carry a sort of stigma.

[3] Black is used here as a generic term that includes black, coloured and Indian students.

Management information on throughput rates per programme and Grade 12 symbols confirms that differences exist between the academic performances of white, coloured and black students per Grade 12 levels. For example, white students with Grade 12 aggregates between 70% and 80% are more successful than coloured and black students with similar Grade 12 aggregates – they graduate faster, more of them graduate and fewer leave without a qualification. Again, this is in line with Steele's research on stereotype threat [11] (what he first observed at the University of Michigan) and a further indication to suspect that stereotype threat is at work. See figures E4a-E4d in [14] that compare the throughput rates (success in the minimum period) of white, coloured and black students, overall, and for specific Grade 12 levels. *Coloured engineering students are therefore at risk of underperformance.*

C. *Residence versus Private*

A comparison was also made between the weighted first-year averages of engineering students living in university residences, students living privately in Stellenbosch and commuting students (students living privately outside of Stellenbosch), overall, and for specific Grade 12 levels – see figures E9a-E9d in [15]. In the majority of cohorts and for most of the Grade 12 levels the students living in residences outperform the other two groups with the students living in town doing slightly better than the commuting students. This finding is in line with similar research done internationally [16] and it suggests that *first-year engineering students living privately*, specifically those commuting from out of town each day, *are also at risk of underperformance*.

The retention data seem to confirm that non-residential students also drop out sooner than their counterparts in residences, also per Grade 12 level. Although the throughput rate data overall, specifically those completing their degrees in the minimum period, also indicate that the residence students fare better, this is not necessarily the case per Grade 12 level – see figures E10a-E12d in [15] to confirm this.

The conclusion: *First-year engineering students living privately, specifically those commuting from out of town each day, are at risk of underperformance.*

III. Wellness

In "Wellness Coaching for Lasting Lifestyle Change" Arloski [17] defines wellness as follows:

- Wellness is a conscious, self-directed and evolving process of achieving full potential.
- Wellness is multi-dimensional and holistic (encompassing such factors as lifestyle, mental and spiritual well-being and the environment).
- Wellness is positive and affirming.

This implies that wellness will help people to achieve their potential, that wellness recognizes and addresses the whole person in all of his/her dimensions and that wellness affirms and mobilizes people's positive qualities and strengths. The implication for higher education is that in order for students to reach their full potential and perform according to their abilities they must be well in all of their dimensions.

SU adopted the wellness model of Bill Hetler (co-founder of the National Wellness Institute) [17]. His model is a comprehensive and inclusive model that looks at wellness in terms of six dimensions, namely physical, emotional, intellectual, occupational, social and spiritual. Since its inception in 1976 this wellness model has served as one of the most common ways to allocate resources for wellness programmes.

In 2002 SU started to build profiles of all her first-year students by means of the Alpha Baseline Questionnaire (ABQ) that was designed around Hetler's wellness model and Astin's CIRP. The ABQ is administered to first-year students at SU at the beginning of the academic year. The ABQ is a web-based questionnaire consisting of 160 items that cover all six wellness dimensions, i.e. the physical (PD), emotional (ED), intellectual (ID), social (SD), occupational (OC) and spiritual (GD). Since it was first launched in 2002 more than 25 000 students have completed the questionnaire. Topics covered in the ABQ include background of the students (parents, accommodation, goals and motivation, financial support), confidence in their own ability to perform, time utilisation patterns in different dimensions, participation or involvement in various wellness activities, reading and writing activities, computer activities, expertise and attitude towards technology, perceptions of various own abilities, help needed on generic skills, special learning needs, values, aspirations and wellness specific items not covered in the rest of the questionnaire, like healthy eating patterns, life satisfaction, the meaning of life, goal setting, and prayer and meditation.

The ABQ is used to build an initial profile of first-year students and to provide and facilitate appropriate support services. ABQ information is also combined with biographical and performance data to build prediction models [1], [2] that identify qualitative and quantitative variables that influence academic performance. Linear regression was used to build these prediction models and they focused only on first-year performance.

A. *Quantitative and categorical variables*

Grade 12 average, race, gender and type of accommodation (university residence or private accommodation) were used in combination with five wellness related variables as input variables in a number of prediction models. The wellness related variables are defined next.

B. *Qualitative or wellness-related variables*

The following five categories of ABQ questions were identified as groups of variables that could possibly influence academic performance:

1. The need of students for additional help with the development of generic skills;
2. Perception of their own abilities in comparison with their peers;
3. Self-confidence with regard to various skills and abilities;
4. Participation in wellness enhancing activities; and
5. Wellness topics not covered elsewhere.

The questions within each of these five groups are outlined below, with the abbreviations of the wellness dimensions to which the questions belong in brackets:

Wellness(W) 1 - Help needed with generic skills: Writing (ID), reading (ID), mathematics (ID), thinking skills (ID), test and examination skills (ID), study skills (ID), subject choices (OD), career choices (OD), financial support (ID), disability (PD), personal development (ED).

Wellness(W) 2 - Perception of own abilities: Academic ability (ID), computer ability (ID), emotional health (ED), leadership (OC), mathematical ability (ID), perseverance (ED), intellectual confidence (ID), social confidence (SD), self-awareness (ED), understanding of other people (SD).

Wellness(W) 3 - Self-confidence: Oral ability (ID), writing ability (ID), problem-solving ability (ID), organization skills (ID), information processing skills (ID), environmental awareness (SD), care for others (SD), seeing the big picture (ID).

Wellness(W) 4 - Participation in wellness enhancing activities: Participated in protests (SD), tutored other students (ID), studied in a group (ID), overwhelmed by everything they had to do (ED), felt depressed (ED), ask teachers for advice (ID), missed classes (ID), discussed politics (ID), socialized with people from other ethnic groups (SD), attended a play/show (SD), visited a museum (ID), communicated via email (ID), did research on the internet (ID).

Wellness(W) 5 – Wellness (not covered elsewhere): Healthy eating habits (PD), in-depth learning (ID), emotional support (ED), knowledge of emotional needs (ED), sense of humour (ED), satisfied with self (ED), positive attitude towards life (GD), goal setting (ID), time for prayer/meditation (GD), life has a purpose (GD).

A score was calculated for each of the above five groups of wellness variables for each student. *W1-HelpSc*, *W2-PerceptionSc*, *W3-ConfidenceSc*, *W4-ActivitySc* and *W5-WellnessSc* were chosen as short names to represent these five scores. They were also used in the cluster analysis and the prediction models as wellness variables.

IV. Cluster Analysis

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. The notion of a cluster varies, depending on which algorithms are used. The authors' choice fell on NeuroXL Clusterizer [18], a neural network clusterization add-in for Microsoft Excel.

NeuroXL Clusterizer was used to investigate the relationship between the wellness related variables *W1-HelpSc*, *W2-PerceptionSc*, *W3-ConfidenceSc*, *W4-ActivitySc* and *W5-WellnessSc*, and the weighted first-year average of first-year engineering students (n = 1510: first-years within engineering at SU from the 2006 to 2009 cohorts who have completed the ABQ). The results of this analysis are depicted in figure 1 and discussed below. Data clustering and neural networks have been used elsewhere in the world to improve the academic performance of students [19], and similar studies as these, serve as a valuable and helpful reference point.

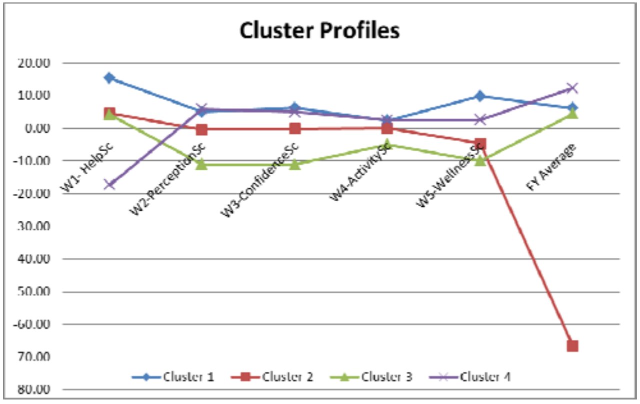

Figure 1: Results of cluster analysis with NeuroXL Clusterizer

Cluster 1 contains students with very high values for *W1-HelpSc*[4], high values for *W2-PerceptionSc* and *W3-ConfidenceSc*, slightly above average values for *W4-ActivitySc* and really high values for *W5-WellnessSc*. The weighted first-year average of students in cluster 1 is the second highest of the four clusters and substantially above average. 26.29% of the students belong to cluster 1.

Cluster 2 contains students with above average values for *W1-HelpSc*, slightly below average values for *W2-PerceptionSc*, almost average values for *W3-ConfidenceSc* and *W4-ActivitySc*, and below average values for *W5-WellnessSc*. The weighted first-year average of students in cluster 2 is the by far the lowest of the four clusters and far below average (66.67% below average). 10.6% of the students belong to cluster 2.

Cluster 3 contains students with above average values for *W1-HelpSc*, well below average values for *W2-PerceptionSc* and *W3-ConfidenceSc*, below average values for *W4-ActivitySc*, and well below average values for *W5-WellnessSc*. The weighted first-year average of students in cluster 3 is above average, but the second lowest of the four clusters. 29.8% of the students belong to cluster 3.

Cluster 4 contains students with well below average values for *W1-HelpSc*, above average values for *W2-PerceptionSc* and *W3-ConfidenceSc*, and slightly above average values for *W4-ActivitySc* and *W5-WellnessSc*. The weighted first-year average of students in cluster 4 is the highest of the four clusters and well above average (12.38% above). 33.31% of the students belong to cluster 4.

Clusters 1 and 4 are the clusters with the highest weighted first-year averages and they are also the clusters with above average values for *W2-PerceptionSc*, *W3-ConfidenceSc*, *W4-ActivitySc* and *W5-WellnessSc*. This represents almost 60% of the students. The top performing students therefore have above average values for the wellness related variables.

Clusters 2 and 3 are the clusters with the lowest weighted first-year averages and their values for *W2-PerceptionSc*, *W3-ConfidenceSc*, *W4-ActivitySc* and *W5-WellnessSc* are below average. They represent the remaining 40% of the students. The weaker performing students therefore have below average values for the wellness related variables.

[4] A high score in *W1-HelpSc* indicates that these students think they need help with generic skills. A lower score is indicative that they feel they are not requiring assistance with their generic skills.

The values of *W1-HelpSc* vary a lot and it does not appear to have a strong relationship with the first-year averages of the students.

It therefore appears as if *students with low scores on the wellness measures underperform*[5].

V. Neural Network Prediction Models / Classification Problems

NeuroIntelligence [20], a neural networks software application designed to assist neural network, data mining, pattern recognition, and predictive modeling experts in solving real-world problems, was chosen to build prediction models at SU. A combination of quantitative, qualitative and categorical variables (described above) are used as inputs and the software is used to predict first-year, retention and graduation outputs. Five different classification problems are defined and solved.

A. *Background*

Classification tasks are tasks connected with the determination of the membership of an object described by means of input data in a definite class. For example, to classification tasks belong bankruptcy forecasting, predicting risk level during crediting, predicting first- and final-year success, etc.

CCR (Correct Classification Rate), input importance and confusion matrix are three terms used extensively in the interpretation of the results of the neural network models and are therefore defined next. CCR stands for Correct Classification Rate and is used in classification tasks as a qualitative characteristic. This rate is calculated by dividing the number of correctly recognized records by the total number of records. CCR is measured in relative units or in percentages. It is used, for example, to indicate what percentage of students are correctly predicted as successful and as unsuccessful at first- and final-year levels.

Input importance is defined as the contribution of the input column/variable to the neural network performance. It is calculated using sensitivity analysis techniques. It indicates for each input variable the magnitude of the role it plays in the predictions.

A confusion matrix [20] is used to analyse the performance of neural network classifications. It displays a square matrix of which the rows and columns represent the target column values and network outputs, respectively. The value in position (i, j) of the matrix (row i, column j of the matrix) is the number of records for which the target column value is in the ith category and whose network output is within the jth category[6]. A neural network that performs perfect classification would have zeroes everywhere except on the diagonal entries.

NeuroIntelligence supports the following seven training algorithms - different combinations of them were used in building the prediction models at SU:

1. Quick propagation,
2. Conjugate Gradient Descent,
3. Quasi-Newton,
4. Limited Memory Quasi-Newton,
5. Levenberg-Marquardt,
6. Incremental back propagation, and
7. Batch back propagation.

B. *Neural Network Models*

The following five neural network prediction models were built with NeuroIntelligence:

1. **<u>Y1 Pass:</u>** *Grade12, Race, Gender, Accommodation, W1-HelpSc, W2-PerceptionSc, W3-ConfidenceSc, W4-ActivitySc,* and *W5-WellnessSc* were used as input variables and *PassFail* as the output variable. Students with a weighted first-year average below 50% failed at least one subject and were assigned the value ***Fail*** for the variable *PassFail* – the others with averages of 50% and above were assigned the value ***Pass*** (although some of them may have failed one or more subjects too) for the variable *PassFail*. The aim of the prediction model was to correctly classify students as either a ***Pass*** or a ***Fail***.
2. **<u>Y1 Retention:</u>** *Grade12, Race, Gender, Accommodation, W1-HelpSc, W2-PerceptionSc, W3-ConfidenceSc, W4-ActivitySc,* and *W5-WellnessSc* were used as input variables and *StayLost* as the output variable. Students who returned after their first year and registered for a second year were assigned the value ***Stay*** for the variable *StayLost* and those who did not return were assigned the value ***Lost*** for the variable *StayLost*. The aim of this prediction model is to correctly classify students as either retained (***Stay***) or ***Lost*** – first-year retention is therefore predicted.
3. **<u>Y2 Retention:</u>** *Grade12, Race, Gender, Accommodation, W1-HelpSc, W2-PerceptionSc, W3-ConfidenceSc, W4-ActivitySc, W5-WellnessSc* and *Y1Ave* (weighted first-year average) were used as input variables and *StayLost* as the output variable. Students who returned after their second year and registered for a third year were assigned the value ***Stay*** for variable *StayLost* and those who did not return after the second year were assigned the value ***Lost*** for the variable *StayLost*. The aim of this prediction model was to correctly classify students as either retained after the second year (***Stay***) or ***Lost*** – second-year retention is therefore predicted.
4. **<u>Min Success 1:</u>** *Grade12, Race, Gender, Accommodation, W1-HelpSc, W2-PerceptionSc, W3-ConfidenceSc, W4-ActivitySc* and *W5-WellnessSc* were used as input variables and *SuccessMin* as the output variable. Students who completed their engineering degrees in the minimum period of four years were assigned the value ***Yes*** for variable *SuccessMin* and those who did not succeed in graduating in four years were assigned the value ***No*** for the variable *SuccessMin*. The aim of this prediction model was to correctly predict which students will complete their degree programmes in the minimum period.

[5] With the exception of *W1-HelpSc*. Also see footnote 4.

[6] Standard mathematical terminology to refer to rows and columns in a matrix, i and j are variables – in a six by eight matrix i = 1,2,…,6 and j = 1,2,…,8.

5. **<u>Min Success 2:</u>** This model is the same as the previous one except that *Y1Ave* (weighted first-year average) is added as an additional input variable. The goal is to determine whether the additional variable will improve the prediction.

Note: For **<u>Y1 Pass</u>** and **<u>Y1 Retention</u>** n = 1510, i.e. the number of first-year engineering students in the 2006-2009 cohorts who have completed the ABQ. For **<u>Y2 Retention</u>** n = 1382, i.e. the number of first-year engineering students who have completed the ABQ in 2006-2009 and who have progressed to their second year in 2007-2010. For **<u>Min Success 1</u>** and **<u>Min Success 2</u>** n = 576, i.e. the number of first-year engineering students from the 2006-2007 cohorts who have completed the ABQ and who therefore had a chance to graduate in the minimum period of four years in 2009 and 2010 respectively. At the time of this study (end of 2011) the final results of 2011 was not known and the latest cohort of students that could have completed their degrees in the minimum period was the first-year cohort of 2007.

C. *The Results: CCR's and Input Importance*

The CCR's (Correct Classification Rates) of the above-described models, as well as their *Input Importance* (weights of the input variables in the prediction), are described and discussed below:

1. **<u>Y1 Pass:</u>** 83.31% of students were correctly classified as either ***Pass*** or ***Fail*** (at the end of their first year**)**. Figure 2 indicates that *Grade 12* was the most important input variable, that the type of accommodation (residence versus private) plays a definite role in predicting academic success and that the wellness related variables (*W1-HelpSc, W2-PerceptionSc, W3-ConfidenceSc, W4-ActivitySc* and *W5-WellnessSc*) all contributed significantly (especially when added together) to this CCR. *Race* and *Gender* also had a role – this was expected due to the differences in academic performance observed between race and gender groups (by studying the corresponding management information systems).

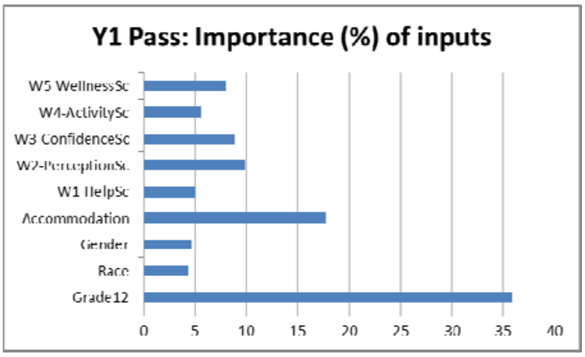

Figure 2: Importance of the input variables in model **<u>Y1 Pass</u>**

2. **<u>Y1 Retention</u>**: 95.76% of students were correctly classified as either ***Stay*** or ***Lost*** (after their first year). Figure 3 indicates that *Grade 12* was the single best predicting variable of first-year retention and the wellness variables also played a part. *Accommodation* had a much smaller role than in **<u>Y1 Pass</u>**. This high CCR value is closely related to the high retention rate of engineering students and to the high quality of students in this programme.

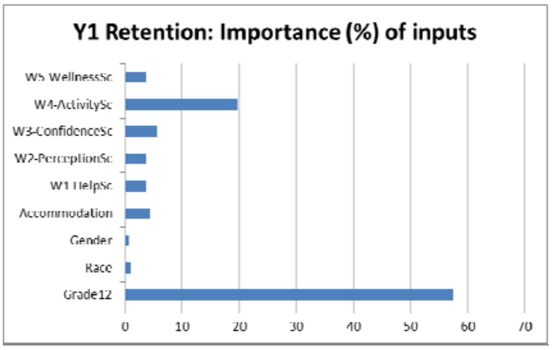

Figure 3: Importance of the input variables in model **<u>Y1 Retention</u>**.

3. **<u>Y2 Retention:</u>** 97.4% of the students were correctly classified as either ***Stay*** or ***Lost*** (after their second year). *Y1Ave*, *Accommodation* and *Grade 12* had the greatest weights (> 10%) while all the other variables (including the wellness related ones) were also significant predictors with weights between 5% and 10% (figure 4). It is interesting to notice that wellness as measured by the ABQ at the beginning of the first year still had such an effect on second year retention.

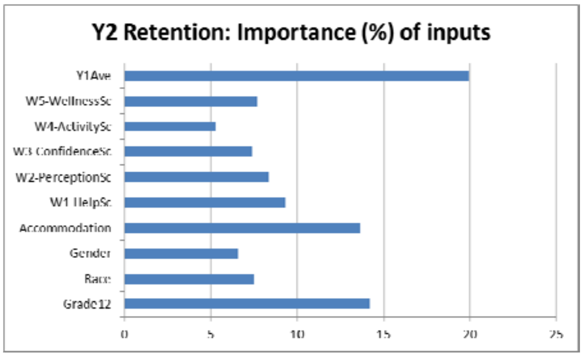

Figure 4: Importance of the input variables in model **<u>Y2 Retention</u>**.

4. **<u>Min Success 1:</u>** 82.99% of the students were correctly classified as either *Yes* or *No* (graduating in the minimum period or not). Figure 5 clearly indicates that the combined weight of the five wellness variables is more than 70%, that the importance of both *Grade 12* and *Accommodation* is approximately 10% and that *Gender* and *Race* have very low importance. It is remarkable that this model can predict with 82.99% accuracy who will complete their engineering studies within the minimum period by only taking the variables in figure 5 into consideration.
5. **<u>Min Success 2:</u>** 86.11% of the students was correctly classified as either ***Yes*** or ***No*** (graduating in the minimum period or not) in this combination of variables. By adding *Y1Ave* to the **<u>Min Success 1</u>** model the CCR is raised to 86.11% and a new combination of weights are applicable. Figure 6 indicates that in his model *Y1Ave* is the variable with the largest weight followed by several wellness variables – their combined weight is almost 50%.

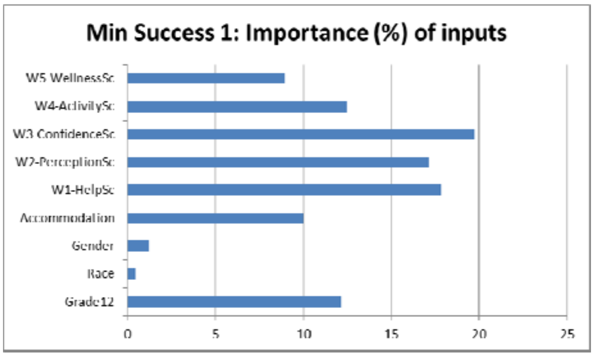

Figure 5: Importance of the input variables in model **Min Success 1**

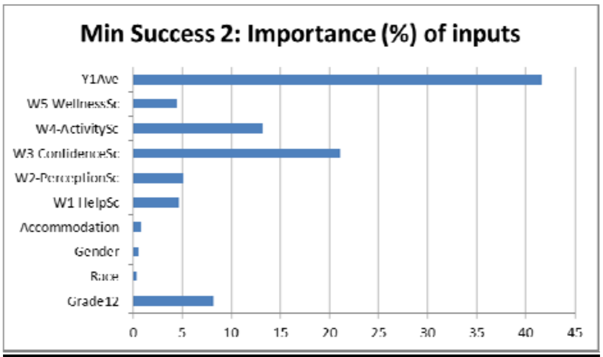

Figure 6: Importance of the input variables in model **Min Success 2**.

D. *Evaluation of the Accuracy of the Models*

1. **Y1 Pass:** This model is successful in predicting a ***Pass*** for 94.22% of the 1160 first-year engineering students that achieved a first-year average of 50% or more (i.e. a pass), but is only successful in predicting 47.14% of the 350 ***Fail*** values. It therefore incorrectly predicts 67 ***Fail*** values for students who passed and 185 ***Pass*** values for students who failed – see table I for the confusion matrix of model **Y1 Pass**. However, an overall CCR value of 83.31% is satisfactory.

Table I: **Y1 Pass** confusion matrix

	Pass	***Fail***
Pass	1093	67
Fail	185	165

2. **Y1 Retention:** This model is successful in predicting a ***Stay*** for 97.51% of the 1404 first-year engineering students who were retained and registered for their second year and a ***Lost*** for 72.64% of the 106 students who left the university after the end of year one. It therefore incorrectly predicted that 35 of those who stayed left and that 29 of those who left stayed. Overall only 64 of the 1510 predictions were wrong (a CCR of 95.76%). Table II represents the confusion matrix of this model.

Table II: **Y1 Retention** confusion matrix

	Stay	***Lost***
Stay	1369	35
Lost	29	77

3. **Y2 Retention:** This model is successful in predicting a ***Stay*** for 99.32% of the 1322 second year engineering students who were retained and registered for their third year and a ***Lost*** for 55% of the 60 students who left the university after the end of year two. It therefore incorrectly predicted that 9 of those who stayed left and that 27 of those who left stayed. Overall only 36 of the 1382 predictions were wrong (a CCR of 97.4%). Table III represents the confusion matrix of this model.

Table III: **Y2 Retention** confusion matrix

	Stay	***Lost***
Stay	1313	9
Lost	27	33

4. **Min Success 1:** 75.88% or 195 of the 257 students who were successful in the minimum period and 88.72% or 283 of the 319 who did not graduate after four years were correctly predicted by this model. It is therefore easier for this model to predict who will not be successful in the minimum period than those who will. 62 of the ***Yes*** predictions and 36 of the ***No*** predictions were wrong (total 98 wrong or 17%). The confusion matrix in table IV summarises this information.

Table IV: **Min Success 1** confusion matrix

	Yes	***No***
Yes	195	62
No	36	283

5. **Min Success 2:** 87.16% or 224 of the 257 students who completed their engineering degrees within the minimum period of four years and 85.27% or 272 of the 319 who did not graduate after four years were correctly predicted (almost equal). 33 wrong predictions were made for the successful group and 47 for the group who was unsuccessful in completing their degrees in the minimum period. Overall only 80 wrong predictions were therefore made – a CCR of 86.11%. See table V for this model's confusion matrix.

Table V: **Min Success 2** confusion matrix

	Yes	***No***
Yes	224	33
No	47	272

E. *Interpretation of Neural Network Results*

The prediction models clearly identified wellness related variables as important predictors of first-year academic performance (who will pass and who will fail), of retention (who will stay and who will not stay) and of success in the minimum period. It also re-affirmed the importance of Grade 12 results as an important predictor and that race, gender and type of accommodation also contribute towards the prediction of who will be successful and who will underperform.

VI. Conclusion

The major findings of this research project can be summarised as follows:

- First-year female engineering students are at risk of underperformance (MIS finding).
- First-year coloured engineering students are at risk of underperformance (MIS finding).
- First-year engineering students living privately, specifically those commuting from out of town each day, are at risk of underperformance (MIS finding).
- Students with low scores on the wellness measures underperform (cluster analysis and neural network predictions).
- The neural network models are surprisingly accurate in predicting academic performance, retention and throughput rates and again demonstrated the importance of both quantitative and qualitative variables (covering all six dimensions of wellness). Academic support programmes designed to develop all six wellness dimensions should therefore be encouraged and enhanced. The causal relationship between academic achievement and wellness should be further investigated.

Interventions to address these reasons for underperformance should be a high priority and should contribute to even better success rates.

This study has also underlined the usefulness of MIS, the potential of cluster analysis and neural network prediction models and has identified the possibility that stereotype threat exist. Future work in 2012 will include testing the neural network prediction models for the weighted first-year academic performance and retention rates with the 2010 and 2011 data – the throughput models (**Min Success 1** and **Min Success 2**) can only be tested once the 2010 cohort has reached the end of their minimum period at the end of 2013.

References

[1] S. A. du Plessis and H. Menkveld, “'n Model vir die voorspelling van eerstejaarsukses ,” Report of the Tracking Unit, Academic Support, Stellenbosch University, 2007.

[2] S. A. du Plessis and H. Menkveld, “Voorspellingsmodelle vir die US (2008-modelle),” Report of the Tracking Unit, Academic Support, Stellenbosch University, 2007.

[3] A. W. Astin and L. Oseguera, “Degree attainment rates at American colleges and universities: Revised edition,” Report of the Higher Education Research Institute, UCLA, Los Angeles, 2005.

[4] J. D. Campbell, “Analysis of institutional data in predicting student retention utilizing knowledge discovery and statistical techniques,” Thesis (Ed.D), Northern Arizona Univ., 2008.

[5] Z. Ibrahim and D. Rusli, “Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression,” presented at the 21st Annual SAS Malaysia Forum, Shangri-La Hotel, Kuala Lumpur, 2007.

[6] R. Alkhasawneh and R. Hobson, “Modeling student retention in science and engineering disciplines using neural networks” in *Conf. Proc. 2011 IEEE Global Engineering Education Conference (EDUCON)*, pp. 660–663.

[7] V. O. Oladokun, A. T. Adebanjo, and O. E. Charles-Owaba, “Predicting students’ academic performance using artificial neural network: A case study of an engineering course,” *The Pacific Journal of Science and Technology*, vol. 9 (1), pp. 72–79, 2008.

[8] P. K. Imbrie, “Use of a neural network model and noncognitive measures to predict student matriculation in engineering,” American Society for Engineering Education, 2007.

[9] J. J. Lin, K. J. Reid, and P. K. Imbrie, “Work in progress - Predicting retention in engineering using an expanded scale of affective characteristics from incoming students,” presented at the 39th ASEE/IEEE Frontiers in Education Conference, San Antonia, M4F1-M4F2, 2009.

[10] *Performance Indicators per Gender within the Faculty of Engineering (internal document).* Available: http://www.sun.ac.za/trackwell/wce2012/menvswomen.docx.

[11] C. M. Steele, *Whistling Vivaldi: And other clues to how stereotypes affect us*: *Issues of our time.* New York: W. W. Norton & Company, 2010, ch. 2.

[12] H. L. Botha and S.A. du Plessis, “An investigation of self-appraised cognition versus measured cognition,” *South African Journal of Higher Education*, vol. 21 (4), pp. 608–627, 2007.

[13] *Throughput Rates of Male and Female Students in Four Year Programmes at Stellenbosch University: Percentages of the 2002-2007 male and female first-year cohorts that successfully completed four year degree programs in the minimum period (internal document).* Available: http://www.sun.ac.za/trackwell/wce2012/menvswomenall4y.docx.

[14] *Performance Indicators per Race within the Faculty of Engineering (internal document).* Available: http://www.sun.ac.za/trackwell/wce2012/whitevsrest.docx.

[15] *Performance Indicators per Accommodation Type within the Faculty of Engineering (internal document).* Available: http://www.sun.ac.za/trackwell/wce2012/resvsprivate.docx.

[16] G. D. Kuh, J. Kinzie, J. H. Schuh, E. J. Whitt, and Associates, *Student success in college, creating conditions that matter.* San Francisco: Jossey-Bass, 2010, ch. 12.

[17] M. Arloski, *Wellness coaching for lasting lifestyle change. Duluth*, Minnesota: Whole Person Associates, 2009, ch. 2.

[18] *Neural network software for clustering and classification in Microsoft Excel (Manual).* Available: http://www.neuroxl.com.

[19] C. E. Moucary, M. Khair, and W. Zakhem, “Improving student's performance using data clustering and neural networks in foreign-language based higher education,” *The Research Bulletin of Jordan ACM*, vol. II(III), pp. 27-34, 2011.

[20] *Artificial neural network software, neural network simulator and classifier (Manual).* Available: http://www.alyuda.com/neural-networks-software.htm.