A Prediction Model to Improve Student Placement at a South African Higher Education Institution

Tasneem Abed
School of Computer Science and Applied Mathematics
The University of the Witwatersrand,
Johannesburg, South Africa
tassabed@gmail.com

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand,
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

Ashwini Jadhav
Faculty of Science
The University of the Witwatersrand,
Johannesburg, South Africa
ashwini.jadhav@wits.ac.za

***Abstract*—There is a growing concern over the low pass rates of students in the Science Faculty at a South African Higher Education institution. The Admission Point Score (APS) used to place students into programs may appear to have good discretion in gauging student aptitude, but the reality is that between 2008 and 2015, about 50% of students who met the APS requirements for a Science program failed to meet the requirements to pass. This report attempts to build a recommendation engine that will advise students on their academic trajectory for a chosen program based on features suggested by the Tinto (1975) framework [1]. The results show that classification models from various archetypes of machine learning have good accuracy in predicting the final outcome of a new student. This research argues that a more complex view of student placement will improve the faculties success rates.**

***Keywords*—Admission Point Score, Student placement, Recommendation engine**

I. Introduction

Higher education has become accepted as a vital key for national development in the context of knowledge, community and globalization. In South Africa, this places a big concern on the output of the higher education sector in terms of the number and quality of graduates. Of most concern is the mismatch between the output of the higher education sector and the economic needs of the country that exist [2]. It is important that the performance of this sector is critically and constantly assessed. In this research we will focus on performance of students and the throughput of graduates.

Solely evaluating universities, failure rates in the Science Faculty is quite alarming. In an investigative report on higher education monitoring, it was reported that only 50% of students who entered into a 3-year Mathematical Science degree graduated within 5 years. The statistics for other fields are similar. The highest attrition rate occurs at the end of the first year of study with about 29% of first year students [2].

The academic performance of students is influenced by many factors, however the influence of their performance in high school is heavily weighted in the admission process. Admission boards use the Admission Point Score (APS) as a means of offering students a place in academic programmes. The APS is a weighted calculation based on the symbols received for each Grade 12 subject. National Benchmark Tests (NBTs) may also be looked at to supplement the board's decision on acceptance. The APS may provide an indication of students performance, however, between the years 2008 and 2015, nearly 50% of students who obtained an adequate APS required for a programme failed to complete it [3]. This reflects a downfall of this system in that it lacks qualitative discretion i.e. the content of the school subjects is not taken into account. For example, a student may achieve a high APS by obtaining distinctions for all their subjects and thus qualifying to enrol in a Science degree. However, these subjects may not be Science subjects but Art or Language based subjects instead. Given that there is a positive relationship between the participation in Science related activities in high school and student achievement in Science, these students may not have the adequate skills to succeed in a Science degree despite meeting the APS requirements [4].

At the root of every student's academic trajectory is the decisions they made that led them to register for their courses. These decisions carry a lot of weight for their future. This is where good academic advisory can be the difference between a student failing a course and passing it. Availability of good academic advisory for every student could potentially reduce failure rates as students will be making more informed decisions about their academic trajectories. Students should know, prior to the commencement of their degree, whether they are more likely to pass in minimum time, pass in more than minimum time, or fail to meet the minimum requirements. Of course, students apply for courses that they have an interest in despite the difficulty level. However, if they are advised that passing may be difficult for them, they do not have to abandon the course but rather know that they may have to work harder to pass than expected. From the University's perspective, resources can be allocated to assist the vulnerable student towards completing their academic programme. However, how do we quantify 'good'

978-1-7281-4162-6/20/$31.00 ©2020 IEEE

advisory? Is it enough to look over a student's high school results and develop an idea of their expected success? Are their high school results sufficient indicators of success? Should advisory be more concrete in the form of numbers i.e. probabilities, percentages and statistics?

In this study, a dataset containing the undergraduate enrolment and Matric results from 2008 to 2018 will be modelled using a Bayesian Network in order to predict the success rate of a student given their profile. The conceptual framework proposed by Tinto (1975) [1] is adopted to develop a methodology to predict the success of completing a degree. In particular, six models with different structures and advantages will be implemented and their performances compared. Profiles of students are made up as a combination of features according to the Tinto (1975) three categories: Background attributes, Individual attributes, Pre-Schooling attributes. An analysis of the results will be conducted to see which of the Tinto (1975) categories contribute the most to the predictability of the models. In order to convey the expected success rate of individuals in specific courses meaningfully, a graphical user interface is designed that will display to the student how their features effect their success rate of a chosen programme.

This research will, in general, contribute an indication of how profile features of students influence their rate of success and provide a ranking of importance of features through information gain. Additionally, the need for a more complex mechanism to determine student placement will be argued.

Section II will provide a literature review in order to understand the trends, models, frameworks and results that have been produced which are reverent towards our research goal. Section III will outline the methodology of the pre-processing of the data, the experiment set-up of the models and the evaluation thereof. This is followed by the results and analysis in section IV and finally the concluding remarks.

II. Related Work

In order to have a full understanding of educational data mining and the techniques and processes applied in building such an engine, it is vital to review and understand the context of the research field by delving into the current state of the field. In this section, we will present the background necessary to build a student advisory system for the Science Faculty in a South African context for a research intensive South African University. Section A will discuss how Mathematics and Languages perform as predictors and Section B will discuss the Mathematics behind models used in the educational data mining.

Tinto (1975) proposed a widely cited conceptual framework of the student attrition process. In this model, three groups of characteristics, namely Background, Individual and Pre-College/Schooling, are interrelated and expected to influence a student's determination into achieving the goal of graduating [5]. Higher grade performance and intellectual development is achieved through commitment to the goal which leads to academic integration and reduces the likelihood of dropping out.

A. Mathematics and Languages as Predictors

Universities in South Africa use the APS of students as a criterion for admissions. Some programmes look at percentages of specific subjects to supplement their admission protocols. Computer Science looks at mathematical ability as a primary predictor of success and thus uses Matric Mathematics marks as a criterion for admission. This is backed up by numerous studies that show a clear link between mathematical aptitude and programming [6]. If a student has good results in Matric Mathematics, the chances of them succeeding in a tertiary Mathematics setting is fair. In fact, many universities throughout the world use high school Mathematics results to select their students. However, this may not be the only significant factor to consider [7], [8]. There is also the transition between high school and University as well as the disconnect between the manner in which high school Mathematics and University Mathematics are taught to consider.

The medium of instruction of the University (English) plays an important role in the success of a student based on the student's comfort with the medium. Although it may be overlooked when considering admission into a Mathematical Science program, the ability of a student to comprehend the content that will be provided to them should also be a primary concern, especially given the South African context where many learners are not native English speakers. An investigation into the belief that language ability influences success at University found that achievement in high school language courses is a better predictor of success than mathematics [7].

Perhaps an even more worthwhile category of features to use are "abstract" abilities such as comprehension skills, memorisation skills, programming skills, mathematics skills and inferential thinking skills which are defined as core features or characteristic skills that a student needs to possess in order to succeed in the course [9].

Academic advisory is vital in any student's academic journey. Insufficient advising is not an uncommon practice especially in the world of distance education where students do not have one-on-one interaction with advisers and academic staff. With numerous courses to select from, a student may not be sure of their interest in a course solely based on its title [9]. All the aforementioned concepts provides a solid base for building a system that will help students to understand their academic trajectories and make informed decisions to

optimize their studies to be more feasible, worthwhile and rewarding.

B. Educational data mining

Statistical analysis is a method used for prediction. In more recent years, machine learning has been developed into a field of its own that encompasses a plethora of algorithms that serve the purposes of clustering and prediction. This section will detail the mathematical approaches and machine learning techniques used in building predictive and recommendation engines in the context of education.

Statistical approaches taken to tackle the problem of predicting the most suitable program for a student in order to recommend it to them mainly revolve around supervised machine learning techniques. In a study conducted in Egypt [10], various decision tree algorithms are applied in order to recommend a department with the highest success rate for the student. These algorithms include the Iterative Dichotomiser 3 (ID3), C4.5 and CART algorithms. Of these, C4.5 proved to be the most efficient and robust due to its roots in the ID3 system. Support vector machines (SVMs) have been shown to give good generalization performance on a variety of problems, however, can suffer from slow training and high complexity. The Sequential Minimal Optimization (SMO) algorithm is an advancement of SVMs with better scaling properties and the use of an analytic quadratic programming step [11]. In a study to predict student performance, a number of machine learning algorithms from different paradigms were tested and compared. Results showed a Multilayer Perceptron (MLP) was the most effective algorithm for predicting student performance [12].

As shown, the mathematical approaches used in this field vary considerably and no one approach is proven to be the best. The success of decision tree algorithms, such as the C4.5 algorithm (also called J48), for predicting successful programmes makes it an algorithm to consider for finding the most suitable programme to recommend to a student.

III. Research Methodology

This research aims to develop a recommendation engine to help students make better informed decisions about their academic trajectory. This will be done by modelling trends in success rates based on Matric marks and biographical profiles of previous and current students, as per the Tinto (1975) framework.

A. Data

The dataset is a synthetic dataset generated using a Bayesian Network, where the underlying ground-truth distribution is known. Conditional independence assumptions were used to express the relationships between the risk status and the qualified, year started and various biographical and high school variables. The target variable is the final outcome where conditional independence is assumed between it and the risk status, qualifiers and high school setting.

B. Preprocessing

The target variable contains 3 possible values which are: *QualMin, Qualified* and *Failed*. *QualMin* represents a student who qualified in 3 years, *Qualified* represents a student who qualified in more than 3 years (4-5 years) and *Failed* represents a student who failed to obtain their degree. The school setting represents whether the high school attended by the student was in a rural or urban setting. Three mandatory Grade 12 subject were kept (English First Language (FL) or First Additional Language (FAL), Mathematics and Life Orientation) as well as Computer Studies and Additional Mathematics. The NBT for Academic Literacy (NBTAL), NBT for Mathematics (NBTMA) and NBT for Quantitative Literacy (NBTQL) are present in the data. The variables used in the final dataset are shown in Table I (where Mathematics is shortened to Math) and are put into their respective categories according to the Tinto (1975) framework. Lastly, the dataset was balanced using undersamplimg so that each value of the target variable has an equal number of occurrences.

C. Models

Six classification algorithms are used to predict the target variable of students: a Random Forest, J48, Naïve Bayes (NB), Logistic Regression (LR), Sequential Minimal Optimization, and a Multilayer Perceptron (MLP). The Multilayer Perceptron is a black box model meaning that the structure of the network will not give any insights on the function being approximated. To identify which of the Tinto (1975) categories prove to be more indicative of success, we can look at the information gain ranking and see how they fair against each other. Other models may provide higher accuracies and therefore numerous models from different machine learning paradigms will be implemented and compared. The effectiveness of each model is tested through 10-fold cross validation. This is a re-sampling procedure by which a portion of the training data is not seen by the algorithm during training, but is used for validation. The dataset is first split randomly into a training and testing set. The training set is then split into K partitions (folds) where $K-1$ folds are used for training and the remaining fold is used for validation. This is then repeated until every K fold has served as the validation fold once. The model which gave the best validation accuracy is used and evaluated by the testing set.

TABLE I: Table of selected features placed in their respective category in accordance with Tinto (1975) [1]

Background	Individual	Pre-University
Home Province	NBTAL	Core Math
Age at First Year	NBTMA	Math Literacy
School Quintile	NBTQL	Additional Math
Rural or Urban	Year Started	English FL
International	Plan Description	English FAL
	Target	Computer Studies
		Life Orientation

The evaluation of the models comes by analysing the accuracy of the models predictions as well as the precision and recall values. Precision and recall are evaluation metrics calculated using the resulting confusion matrix, which is a table used to describe the performance of a classifier on test data for which the true values are known. The diagonal elements represents correctly classified instances while off diagonals represent the number of instances incorrectly classified where its true class is given by the column. Precision, recall and accuracy are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where TP are true positives, FP are false positives and FN are false negatives. The accuracy is calculated as

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

In our multiclass case, the confusion matrices will be 3×3 dimensional, so precision and recall will be calculated for each class. Precision represents the proportion of results that are relevant while recall is the proportion of all relevant results that have been correctly classified. A trade off exists between the two in that results will have to be repeatedly generated in order to recall everything thus lowering precision. An information gain ranking tool allows us to see which features played the biggest role in distinguishing which class a student falls into.

IV. Experiments

Table II shows the 6 models and their respective accuracies after 10-Fold cross validation. The Multilayer Perceptron took the longest time to train while the other 5 models were relatively fast. The NB algorithm achieves the top accuracy at 69,18%. The SMO is not far behind. Our next best model, the J48 decision tree, jumps down by a mere 0.72% from the top model. The range of accuracies is 1.71% which shows that all the models performed well relative to each other. The Naïve Bayes model may have performed the best due to the conditional independence assumption of the features. We can see that models that come from the decision tree paradigm outperform the black box MLP model. The success of the J48 model is due to its features such as its ability to handle missing values, the continuous value range and threshold, that it does well in choosing, and its ability to prune the decision tree to remove branches that do not add value to the model [13]. Although not the top model (but very close), this is consistent with [10] in showing the C4.5 algorithm is robust and efficient. A property of trees that make it compatible with the data is that they work well with many features, especially categorical features.

The worst performing model is the MLP, although only 1.71% from the top model. The MLP is limited by the fact that it can not guarantee that the minima it stops at is global. The LR's performance is mid-tier. The multivariate version uses the softmax activation function, which sometimes causes biases, instead of a sigmoid activation function which is typically used for binary classification.

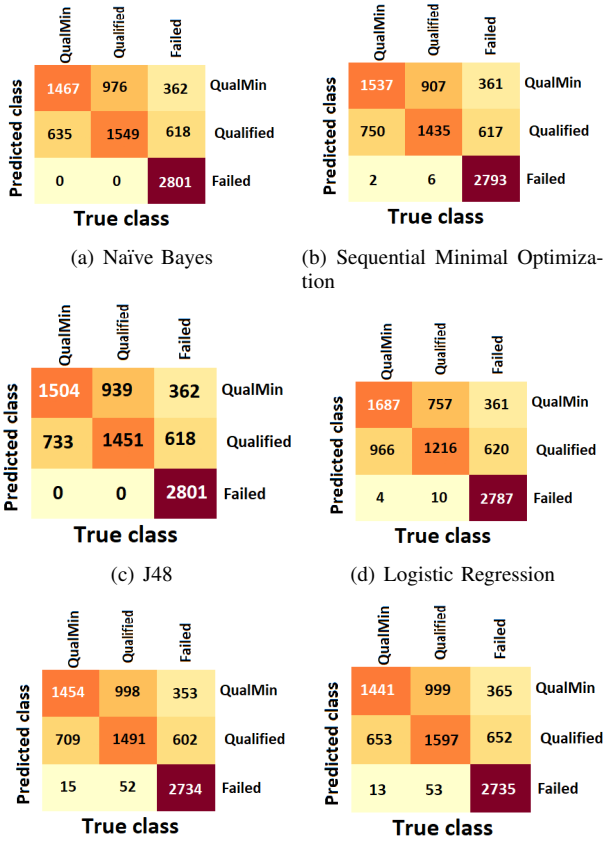
Figure 1 a to f depict confusion matrices for each model. In each case, it can be seen that the *Failed* class value was the most accurately predicted class, *Qualified* was second best and *QualMin* was the least accurately predicted class. Both the NB and J48 models managed to perfectly predict the *Failed* class. Most confusion lies between the *Qualified* and *QualMin* classes. The confusion between these classes comes from the fine line between the number of years it may take a student to graduate. If a student takes 3 years, they fall in the *QualMin* class, but if they take 4 they fall in the Qualifid class. The distinction between students in each class may not be clear from features such as Matric marks, NBT marks or even biographical profiles. External factors may cause a student to fail a year which is not quantified in our data and therefore can not be utilised to improve the predictive accuracy. However, the distinction between students who qualify overall and who fail may well be prevalent in features such as Matric marks and biographical profiles. This leads to the contrast in predictive accuracy between the values and reinforces the argument that all these features play a role in the success of a students academic trajectory.

Figures 2 to 4 are precision and recall graphs pertaining to the NB, J48 and MLP models respectively. The x-axis represents the recall while the y-axis represents the precision. In graph (c) of all three figures we see the *Failed* value takes on a smoother curve than the others. As mentioned before, the *Failed* value has good accuracy amongst all the models with an average accuracy of 87.5%. We can see this performance is reinforced by the high recall each model produces. We are more interested in the recall since it represents the proportion of relevant samples i.e. it represents how many students who failed which are identified by the model. The smoothness of the *Failed* curves in all three cases show that the model is not biased or over-fit. In all three cases we can see the *QualMin* suffers by the jaggedness of the curves. The models may be over-fitting at times.

Table III shows the features ranked according to their information gain. The second column indicates the feature's entropy value *(e)* where $0 \leq e \leq 1$. The colours indicate which of Tinto's categories the feature belongs to, corresponding to

TABLE II: Table of model accuracies to predict the target variable using 10-fold cross validation.

Model	Accuracy
NB	**69.18%**
SMO	68.56%
J48	68.46%
LR	67.67%
RF	67.54%
MLP	67.47%



(a) Naïve Bayes (b) Sequential Minimal Optimization

(c) J48 (d) Logistic Regression

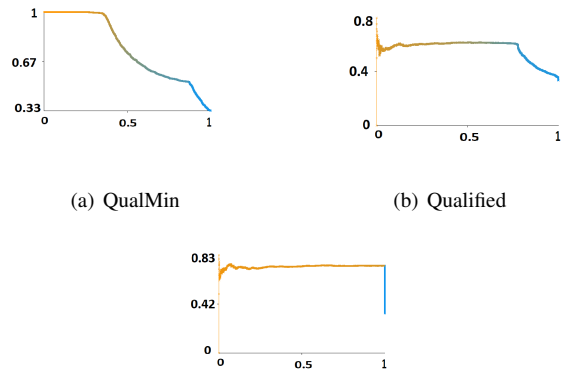
(e) Random Forest (f) Multilayer Perceptron

Fig. 1: Confusion matrices for each model to predict student outcomes in order of descending accuracy

TABLE III: Information Gain

Rank	Information Gain (e)	Feature
1	0.68930	Year Started
2	0.01612	Plan Description
3	0.00127	Home Province
4	0.00121	Rural or Urban
5	0.00035	Quintile
6	0.00001	International
7	< 0.00001	English FAL
8	< 0.00001	NBTAL
9	< 0.00001	NBTMA
10	< 0.00001	Additional Mathematics
11	< 0.00001	Age At First Year
12	< 0.00001	Computer Studies
13	< 0.00001	English FL
14	< 0.00001	NBTQL
15	< 0.00001	Mathematics Literacy
16	< 0.00001	Life Orientation
17	< 0.00001	Core Mathematics

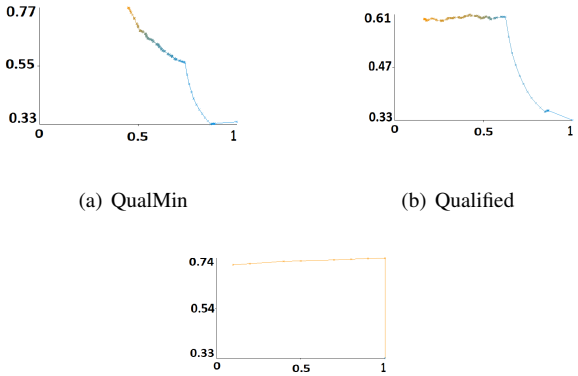
Table I. The top 6 features are the only contributing features to the models, of which none are Pre-University features. Most of the Background features fall in the top 6 but the top 2 features are individual features. These results show the Background and Individual attribute groups of Tinto (1975) have a dominant role in predicting student's performance. This is consistent with the results found by [3]. Year Started is the top contributing feature and this is due to the construction and pedagogy of each module in that year. This feature, however, can not be used for prediction as the structure of future courses



(a) QualMin (b) Qualified

(c) Failed

Fig. 2: Graphs showing precision and recall curves for each outcome class value for the NB model



(a) QualMin (b) Qualified

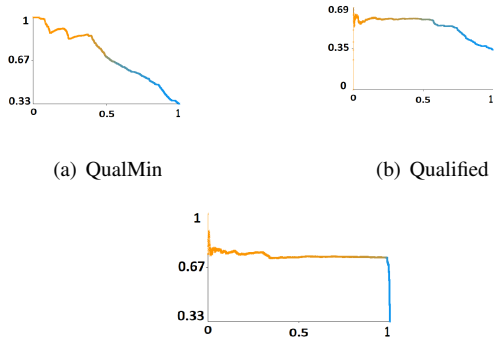
(c) Failed

Fig. 3: Graphs showing precision and recall curves for each outcome class value for the J48 model

can not be known. With Pre-University features being the least contributing feature set, it brings into question why the APS, which we recall is made up of Pre-University subject marks only, is the primary mechanism for offering students places in programs. The Home Province is the 3rd highest contributing feature. This may be an indication of the student's first language. It is proposed that proficiency in the medium of the University is vital to a students expected success and the Home Province as a proxy for Language is consistent with this idea [7].

V. Conclusion

Due to the low pass rates of students in the Sciences in South Africa, there is a need for higher quality academic advisory so students may make better informed decisions about their academic trajectories. More specifically, a recommendation engine that will allow students to gauge their future success and academic affordability of a program



(a) QualMin (b) Qualified

(c) Failed

Fig. 4: Graphs showing precision and recall curves for each outcome class value for the MLP model

is proposed. In addition to this, the mechanism used to select students to be accepted to a program, APS, is debated as being ill-informed. However, discovering the factors that influence a student's future success was not an easy task. Tinto (1975) outlines a framework that identifies three categories of student attributes that influence the attrition rate. In this research, we have employed this framework to identify which category of attributes has the biggest influence on Science students and compare this to the APS mechanism.

Attributes were split into Background, Individual and Pre-University categories. Six classification models from different paradigms were run on the data to predict which of the following three classes a student falls into: *QualMin* (qualify in the minimum time of three years); *Qualified* and *Failed*. The models proved that the combination of these attributes can predict a student's outcome within a 67% - 70% accuracy. The Naïve Bayes model proved to be the most robust with high recall values and mostly the highest precision values as well as the highest accuracy overall. Furthermore, the most influential category of attributes was the Individual attributes which included the Plan description and Year Started. The Background attributes which included the Home Province and Quintile also proved to have high influence. Interestingly, the Pre-University attributes, which are high school grades per subject, had the least influence. APS is calculated as a sum of points pertaining to the marks received per high school subject. Thus we can see a mismatch in the requirements needed to earn a place in a program and the likelihood of succeeding.

The limitations of this work mainly exist in the synthesis of the data. Synthetic data may not capture real world patterns and observations which skews the results, but it does allow us to simulate a theoretical scenario where a proof of concept can be built. This questions the reliability of the data and the use of such models in the real world. Data from the years preceding 2008 are not captured in this data set.

The overall contributions of this work is providing a more complex way of viewing student placement in the Science faculty as opposed to using APS. It is proposed that this is achieved through the implementation of an engine that allows students to input their information and view a distribution of their risk over courses of a similar nature to the one they choose to do. This paper also provides insights into machine learning models that work well with educational data. We further show the importance of each attribute in the predictability of the models by ranking them according to their information gain as seen in Table III.

Future work in this field may include extending the model to accommodate all faculties in the University. This would greatly improve the output of the University as students would be in fields more suited to them or would know the effort they need to put in beforehand. Another future implementation may be to look at first year students who are high risk, at the end of their first year, and propose a change of program for them according to possible pathways suggested by statistical models. As an improvement to the model, data relating to the Peer-Group Interactions and Faculty Interactions of students may be added and ranked as they are part of the framework of Tinto (1975) and may prove useful.

REFERENCES

[1] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of educational research*, vol. 45, no. 1, pp. 89–125, 1975.

[2] I. Scott, N. Yeld, and J. Hendry, "A case for improving teaching and learning in south african higher education," *Higher education monitor*, vol. 6, no. 2, pp. 1–8, 2007.

[3] R. Ajoodha, "Predicting learner attrition for the sciences using background, individual attributes, and schooling at a south african higher educational institute," *Private Communication*, 2019.

[4] D. G. Markowitz, "Evaluation of the long-term impact of a university high school summer science program on students' interest and perceived abilities in science," *Journal of Science Education and Technology*, vol. 13, no. 3, pp. 395–407, 2004.

[5] J. P. Bean, "Conceptual models of student attrition: How theory can help the institutional researcher," *New directions for institutional research*, vol. 1982, no. 36, pp. 17–33, 1982.

[6] P. Byrne and G. Lyons, "The effect of student attributes on success in programming," *ACM SIGCSE Bulletin*, vol. 33, no. 3, pp. 49–52, 2001.

[7] S. Rauchas, B. Rosman, G. Konidaris, and I. Sanders, "Language performance at high school and success in first year computer science," *SIGCSE Bull.*, vol. 38, no. 1, pp. 398–402, Mar. 2006. [Online]. Available: http://doi.acm.org/10.1145/1124706.1121467

[8] P. F. Campbell and G. P. McCabe, "Predicting the success of freshmen in a computer science major," *Communications of the ACM*, vol. 27, no. 11, pp. 1108–1113, 1984.

[9] K. Taha, "Automatic academic advisor," pp. 262–268, 2012.

[10] W. M. Aly, O. F. Hegazy, and H. M. N. Rashad, "Automated student advisory using machine learning," *International Journal of Computer Applications*, vol. 975, p. 8887, 2013.

[11] J. Platt, "Sequential minimal optimization:a fast algorithm for training support vector machines," 1998.

[12] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting student performance: a statistical and data mining approach," *International journal of computer applications*, vol. 63, no. 8, pp. 35–39, 2013.

[13] G. Kaur and A. Chhabra, "Improved j48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, 2014.