A Review: Predicting Student Success at Various Levels of their Learning Journey in a Science Programme

Judith Goodness Khanyisa Mabunda
School of Computer Science
and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
1489219@students.wits.ac.za

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

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

***Abstract*—This paper examines how features affect student persistence or dropout at South African higher education institutions, based on three previous studies. In the previous studies, high school grades were used as a valid predictor of student success. The quality of a high school's learning environment has an effect on almost every aspect of higher education success. Students who are better prepared coming out of high school are ideally suited to do well in higher education institutions, who they are, how much money they have, and where they go don't matter. This review aims to identify effective features that warrant student success from high school grades and choice of academic courses during registration in higher education. The following questions are used to guide this review: How can we define student success? Which features should we focus on? Which models work? Based on data mining techniques such as machine learning models that the previous studies have used to predict student success, it has been revealed that the most important features that influence student success in a Computer Science programme are Prior Computer experience, Mathematics, English from High school and the choice of a course.**

***Index Terms*—Student success, Prediction, Higher Education, Machine Learning**

I. Introduction

Student success is significant in South African higher education because it is commonly used as a metric for the institution's performance. At-risk students would have a much greater chance of thriving if they are detected early and preventative measures are taken. High school grades have long been a strong predictor of student success in postsecondary institutions [11]. Students who complete four years of mathematics, science, and English in high school have an 87 percent chance of graduating from college, compared to a 62 percent chance for those who do not complete the coursework [4], [14]. Machine learning algorithms have been widely used for prediction in recent years.

Admission to science programmes in South African universities, requires that students perform well in both pure mathematics and English as confirmed by [12], [15]. However, studies that applied machine learning techniques show that pure mathematics and English cannot be the only predictors of student success in higher educational institutions [1]. The effective and efficient application of machine learning techniques entail many decisions, ranging from how to define student success, through which student features to focus on, up to which machine learning method is more appropriate to the given problem.

II. How can we define success?

The authors in each paper have outlined what drives student success, in simple words, they have defined what student success is. [2] defined student success as dependent on biographical and enrolment observations as features that ensures efficient student performance at a South African higher educational institution. This study argues that there exists characteristics, attributes, and features in a student profile that can accurately predict the student's performance from the first year of registration until qualifying, providing a contribution that suggests a support and supplementary mechanism to the current university Admission Point Score (APS) system, which has evidently been struggling, generating between 25.7% and 32.2% minimum time (3-year) graduates in South Africa for the academic years 2000 to 2017 [2].

[3] believes that student enrolment and biographical data are rich sources of information that can help universities and staff solve a number of problems, such as identifying at-risk students, restricting student intake, and changing course content to understand and assist disadvantaged students.

[1] predicted student success for first year computer science students based on three features and believes that prior computer experience is required for students to perform well in a computer science course.

III. Which Features Should we focus on?

The authors considered different features and looked closely at the extent to which these features impact student success.

978-1-6654-4067-7/21/$31.00 ©2021 IEEE

[1] considered two groups when investigating the effect that prior computer experience had on the final first year computer science results. Students in group 1 had no prior computer experience. Students in group 2 had prior computer experience. The paper further outlined all three features that were looked at when predicting success for first year computer science students. The features include: Final grade 12 results such as Pure Mathematics as a predictor (this is a known predictor of performance for computer science as confirmed by [15]), Past computer experience as a predictor and Language performance as a predictor. The features are shown in TABLE 1.

On the other hand, paper [3] considered school quintile, high school grades, Naional Banchmark Test scores as well as the major a student chose when registering at university as predictors when predicting the Success of First Year University Students. From the results it was concluded that a student's undergraduate major, school quintile, Life Sciences and Mathematics marks in matric have a significant effect on their chance of completing their studies. The features of this paper are outlined in Fig. 1.

Some features used by [2] are the same as the ones used by [3]. But some are different. TABLE I listed all the features that [2] used to predict student performance at each year of study until qualifying, for students at a South African higher education institution. Fig. 2 represents features used throughout in [2].

TABLE I: A table presenting the various features used for prediction of student success in [1].

Features
Mathematics as a predictor
Past computer experience as a predictor
Language performance as a predictor

Attribute Name	***Description***
Qualified	Whether a student qualified or not (Class Values)
SchoolQuintile	Quintile 1 is the group of schools in each province catering for the poorest 20% of learners, while Quintile 5 is the group of schools in each province catering for the least poor 20%.
LifeOrientation	Life orientation grades
MathematicsMatricMajor	Grades of a student if they took pure maths
MathematicsMatricLit	Grades of a student if they took maths literacy
AdditionalMathematics	Grades of a student if they took advanced maths
EnglishFirstLang	English grade if taken as a first language
EnglishFirstAdditional	English grade if taken as an additional language
NBTAL	Grade in the National Benchmark Test Academic Literacy section
NBTMA	Grade in the National Benchmark Test Mathematics paper
NBTQL	Grade in the National Benchmark Test Quantitative Literacy section
AdditionalLanguage	Grade in additional language, if taken
PhysicsChem	Grade in Physics and Chemistry
Geography	Grade in Geography
LifeSciences	Grade in Life Sciences

Fig. 1: A figure presenting the various features used for prediction in [3].

IV. Which Models work?

To achieve their objectives and answer research problem statement and questions that are set out for their papers to answer, the authors considered different models that are suitable and best possible approaches for their given problems.

[1] used four models. Linear regression was used to predict the true results of students, the outcome produced from the model is a real number. The classification models such as Logistic regression, Naive Bayes and Decision tree were used to predict a student's results as either a PASS or FAIL. The four models are used to identify which features are important in higher educational institution success to first year computer science students.

The three classification models performed very well in group 2 as they resulted in highest accuracy than in group 1. Although the classification models performed better in group 2, the difference in accuracies between group 1 and group 2 was not much when using the logistic regression model and Naive Bayes model. The differences for these two models between group 1 and group 2 are 2.65 and 0.37 percents respectively. The difference in accuracies when using the decision tree classifier model was greater between the two groups as compared to using the logistic regression model and Naive Bayes model. The difference was 8.51 percent when using the gini index and 13.2 percent when using entropy in decision tree model.

Results from the hypothesis test and confidence interval test showed that the students in group 2 outperformed the students in group 1. Based on these results, it is worth considering past computer experience as an additional criterion to studying computer science. However, we should not ignore students without prior computer experience considering the fact that the accuracies produced from both group 1 and group 2 differed by a small percentage when using both the logistic regression model (difference of 2.65 percent) and Naive Bayes model (difference of 0.37 percent).

On the other hand, through six machine learning models like Random Forest, Logistic Model Trees (LMT), Decision Trees (J48), Sequential Minimal Optimization (SMO), Multinomial Logistic Regression and Naive Bayes, the authors predicted students' first, second, and final year outcomes based on synthetic dataset. In the 1st, 2nd and final years of student enrollment, all of these models performed outstandingly in predicting student success, however, amongst these models, Random Forests performed better with accuracy of 94.40%, 93.70% and 95.45% respectively [2].

#	Feature	1st Year	2nd Year	Final Year
1	English Home Language			
2	Plan Description			
3	Quintile			
4	Home Province			
5	Year Started			
6	Language			
7	Progress Outcome YOS1			
8	Home country			
9	Aggregate YOS2			
10	Rural or Urban			
11	Second Year Outcome			
12	Age at Third Year			
13	Mathematics Literacy			
14	NBTAL			
15	Age at First Year			
16	Computers			
17	NBTQL			
18	Age at Second Year			
19	Life Orientation			
20	NBTMA			
21	Plan Code			
22	English FAL			
23	Additional Mathematics			
24	Mathematics Major			

Fig. 2: A figure presenting the various features used for classification in [2]. The table sorts the features according to whether they were used for the prediction of the students' 1st Year Outcome, 2nd Year Outcome, or Final Year Outcome.

Another study aimed at investigating which features of a student best predict whether they will graduate so as to identify vulnerable students and offer them crucial assistance applied six machine models such as Bootstrap Aggregating (Bagging), Bayesian Network, Logistic Regression, Multilayer Perceptron (MLP), K-Nearest Neighbours (KNN) and Random Forests. Bagging was found to be the most effective model for these features, correctly classifying 75.97% of the data. Random Forests followed closely, correctly classifying 75.57%, while KNN came in last with 64.83% [3].

These models are briefly described below.

- Bayesian Network - A probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).
- Logistic Regression - used to model the probability of a certain class or event existing such as pass/fail, win/lose.
- Multilayer Perceptron (MLP) - is a class of feedforward artificial neural network. The Perceptron consists of an input layer and an output layer which are fully connected.
- K-Nearest Neighbours (KNN) - K-Nearest Neighbour is a non-parametric classification algorithm. This algorithm works by finding the distances between a new data point and all the labelled datasets in the data, it reads through the whole dataset to find out the k nearest neighbours closest to the new data point, then the votes for the most frequent label in classification will be the class for the new data point [10].
- Bootstrap Aggregating (Bagging) – a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.
- Random Forests - Random forest algorithm is a supervised classification model. This model creates the forest with number of trees [6]. Random forest classifier works well with missing values. It does not overfit the model when there are more trees in the forest.
- Naïve Bayes (NB) - has been widely used in text classification. Given a set of labelled data, NB often uses a parameter learning method called Frequency Estimate (FE), which estimates word probabilities by computing appropriate frequencies from data [13].
- Linear Regression - models a relation between a dependent variable and one or more independent variable.

Random Forest followed by Logistic model trees, Multinomial Logistic Regression, Sequential Minimal Optimization and Decesion tree are suitable and best possible approaches for predicting student success.

Modules (features)	**Pearson Correlation Coefficient**	**n (sample size)**
Pure mathematics grade 12 results	0.346117	214
Computer studies grade 12 results	0.320650	119
English First Language grade 12 results	0.164604	214

Fig. 3: This figure shows the result of the linear regression from [1].

TABLE II summarises and compares the accuracies of each model that the authors obtained when conducting their studies.

V. Discussion

All of these indicators of student success have been studied in the literature to varying degrees, and there is widespread consensus on their significance. While other studies have concentrated on Mathematics and English as the best predictors of student success in a Computer Science program, the findings in [1] indicate that previous computer experience is also important for student success in higher educational institutions.

On the other hand, [3] demonstrate that school quintile, NBT grades, a course or major a student registered for, life sciences and mathematics from high school matters. [2] has as well focused on high school and intra-university grades and individual attributes such as home language, home province, age, etc.

In other studies, a handful of additional elements of student success have emerged, representing new dimensions, and variations on common indicators. Examples of such indicators are theoretical perspectives (such as Sociological Perspectives,

TABLE II: The accuracy of each model.

Models	Paper1 [1]		Paper2 [2]			Paper3 [3]
	Group 1	Group 2	Year1	Year2	Year3	After
Logistic Regression	59.85%	62.5%	-	-	-	75.48%
Decision Tree	60.24%	79.10%	-	-	-	-
Naive Bayes	60.95%	61.32%	83.95%	83.40%	84.40%	-
Decision Tree (J48)	-	-	87.55%	86.20%	91.45%	-
Random Forest	-	-	94.40%	93.70%	95.45%	75.57%
Sequential Minimal Optimization (SMO)	-	-	87.25%	84.5%	89.20%	-
Multinomial Logistic Regression	-	-	87.80%	86.20%	90.70%	-
Logistic Model Trees (LMT)	-	-	91.90%	91.75%	93.15%	-
Bayesian Network	-	-	-	-	-	74.12%
Multilayer Perceptron (MLP)	-	-	-	-	-	75.09%
K-Nearest Neighbours (KNN)	-	-	-	-	-	64.83%
Bootstrap Aggregating (Bagging)	-	-	-	-	-	75.97%

Organizational Perspectives, Psychological Perspectives, Cultural Perspectives and Economic Perspectives). The indicators also includes student background characteristics, precollege experiences, and enrolment patterns. Student engagement such as Student behaviors, activities, and experiences in post secondary education [8].

The importance of student engagement can be divided into two categories. The first is the amount of time and effort students put into their studies and other educational pursuits. "A person's level of commitment to the learning process has a huge effect on learning." [5]. The second category of student engagement is how the school allocates resources and organizes the curriculum, other learning opportunities, and support services to enable students to participate in behaviors that contribute to positive interactions and results such as persistence, satisfaction, learning, and graduation [7]. As [9] concluded, individual effort and participation in academic, interpersonal, and extracurricular offerings on a campus decide the effect of higher eductaional institution.

VI. Conclusion

From this review, it's concluded that there are more influential features on student success other than high school grades and choice of course during registration. I suggest everyone that tries to predict student success look closely at various features that might just make a very huge difference in ensuring that the results are correct and efficient to make decisions. The best prediction models should be considered when predicting student success. Different students perform differently based on different features, that is why it is very vital that various features be considered when predicting student success.

Acknowledgment

This work is based on the research supported in part by the National Research foundation of South Africa (Grant number: 121835).

References

[1] Different models relating prior computer experience with performance in first year computer science.

[2] Educational data-mining to determine student success at higher education institutions. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–8. IEEE, 2020.

[3] Using machine learning techniques and matric grades to predict the success of first year university students. In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–5. IEEE, 2020.

[4] Clifford Adelman. *Answers in the tool box: Academic intensity, attendance patterns, and bachelor's degree attainment*. US Department of Education, Office of Educational Research and Improvement, 1999.

[5] PA Alexander and PK Murphy. The research base for apa's learner-centered psychological principles', paper presented at the. In *Annual Meeting of the American Educational Research Association*, 1994.

[6] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[7] George D Kuh. Assessing what really matters to student learning inside the national survey of student engagement. *Change: The magazine of higher learning*, 33(3):10–17, 2001.

[8] George D Kuh, Jillian L Kinzie, Jennifer A Buckley, Brian K Bridges, and John C Hayek. *What matters to student success: A review of the literature*, volume 8. National Postsecondary Education Cooperative Washington, DC, 2006.

[9] Ernest T Pascarella and Patrick T Terenzini. *How College Affects Students: A Third Decade of Research. Volume 2*. ERIC, 2005.

[10] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[11] Gary R Pike and Joseph L Saupe. Does high school matter? an analysis of three methods of predicting first-year grades. *Research in higher education*, 43(2):187–207, 2002.

[12] Sarah Rauchas, Benjamin Rosman, George Konidaris, and Ian Sanders. Language performance at high school and success in first year computer science. *ACM SIGCSE Bulletin*, 38(1):398–402, 2006.

[13] Jiang Su, Jelber S Shirab, and Stan Matwin. Large scale text classification using semi-supervised multinomial naive bayes. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 97–104. Citeseer, 2011.

[14] Edward C Warburton, Rosio Bugarin, and Anne-Marie Nunez. Bridging the gap: Academic preparation and postsecondary success of first-generation students. statistical analysis report. postsecondary education descriptive analysis reports. 2001.

[15] Laurie Honour Werth. Predicting student performance in a beginning computer science class. *ACM SIGCSE Bulletin*, 18(1):138–143, 1986.