# Using Machine Learning Techniques and Matric Grades to Predict the Success of First Year University Students

***Abstract*— Student enrolment and biographical data are rich sources of information that could help universities and staff tackle a diverse range of problems, such as identifying at-risk students, student intake limitations, and course content adjustments. South Africa faces a unique economic and political history which creates new sets of challenges in the determination of which students are at risk of failing their degrees. This paper investigates which attributes of a student best predict whether they will graduate so as to identify vulnerable students and offer them crucial assistance. Different machine learning algorithms are applied to the data and the results are compared. The data was synthetically generated using a Bayesian network with features such as the major a student chooses, their school quintile, high school grades as well as NBT scores. Bagging produced the best results, correctly classifying 75.97% of the data.**

***Keywords— education; learning analytics; grades prediction; machine learning.***

## I. INTRODUCTION

South Africa faces a university crisis with an extremely high dropout rate. Even though enrolment has doubled from approximately 500 000 in 1993 to 938 201 in 2011 largely due to apartheid ending, [19] universities have been unprepared to accommodate this increase with a first-year dropout rate of around 40%. [1]

Looking at a 2010 cohort of students, only 22% of them completed their three-year degree within three years. [2] Moodley [2] estimates that the number of dropouts during the period 2000 to 2005 cost the National Treasury R4.5 billion in grants and subsidies to higher education institutions without a return on their investment. Historical inequality, a lack of support and funding plus a worsening in the quality of government education makes South African students especially vulnerable and increases the importance of analysing student data to extract meaningful information that can be used to improve the amount and quality of support to students.

The aim of this study is to identify if a student's high school marks and their undergraduate major are good indicators of whether they are at-risk of not completing their degree. This is vital because the earlier an at-risk student can be identified, the earlier university staff can intervene and provide the student with the support and resources they require in order to graduate.

## II. RELATED WORK

Prediction of student performance has been studied for many years, with one of the earliest papers, Campbell [3] in 1984, applying linear regression and finding Scholastic Aptitude Test (SAT) scores and high school grades as the best predictors of success. A paper published in 1985 by Butcher [4] found similar features, American College Testing (ACT) and Grade Point Average (GPA), to be the best features, but received much greater predictive success than Campbell's [3] study. In 1992, Danko [5] also found that GPA and exam scores were the best predictors of success in a group of accounting students.

Building on these studies by adding more descriptive variables like course preference, Goold [6] conducted a study in 2000 using a range of variables such as age, gender, high school grades and schooling type. This study had moderate success of between 42% and 65%. In South Africa, Spark [7] looked at high school maths grades and level of maths and found a high correlation between them and a student's success in a first-year computer course.

In more recent years, researchers started applying Artificial Intelligence algorithms to the problem. In 2006 Superby [8] used decision trees, random forests, neural networks, and linear discriminant analysis on the data, with 375 dependent variables, a significant increase from previous studies. Results were mixed, with the best results using linear discriminant analysis at 57.35%. This study supported the findings of previous studies by finding high school grades and maths grades as important variables, but additionally found attendance of courses and study skills to be important.

Most recently, Daud [9] in 2017 and Tsiakmaki [10] in 2018 have both applied a wide range of machine learning algorithms on large groups of student data. Daud [9] used more unique variables such as family expenditure, income, and personal information, and found good results with the support vector machine algorithm, with an F1 score 0.867. Significant variables were electricity expenditure, location, and employment status.

Tsiakmaki [10] used subject grades like previous studies, but variables such as how the student entered the department and number of times a student failed a course were new. Random forests gave a fair accuracy and the best of all the algorithms in this study.

## III. Data

This study was performed on synthetic data which was generated using a Bayesian network. A Bayesian network is a type of probabilistic model that uses a directed acyclic graph (DAG) to represent a group of variables and their relationships. Each node represents an attribute and each edge joining them shows the dependencies between them. Nodes that are not connected represent attributes that are conditionally independent from one another. The node that the edge originates from is the parent node and the node being pointed to is the child node, with the edge forming a conditional causal link from the parent to the child node. This causal relationship uses Bayes' Theorem which is defined in terms of conditional probability:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

where A and B represent two events,

P(A|B) is the probability of A given B is true

P(B|A) is the probability of B given A is true

P(A), P(B) are the independent probabilities of A and B.

The network that was used to generate the data in this study can be seen in Figure I. The variable Qualified was used as the measurement of success and could take on one of two values, either F (Failed) or Q (Qualified).

Most undergraduate degrees last 3 to 4 years and are based on a credit system. In order to receive a qualification, a student is required to accrue a certain amount of credits in total, with more credits given for more advanced courses. In certain faculties, if a student fails a certain amount of times or receives low grades, they can be excluded from re-entering that degree and repeating any courses. Exclusion from a course as well as dropping out of the university fall under the Failed category.

To perform pre-processing on the data, non-sensical values were removed like marks above 100 as well as rows with no high school data. Since 6178 students failed and 4431 students qualified, which would have caused the algorithms to overfit the data, the function ClassBalancer as implemented by Eibe [11] was used. This function "reweights the instances in the data so that each class has the same total weight."

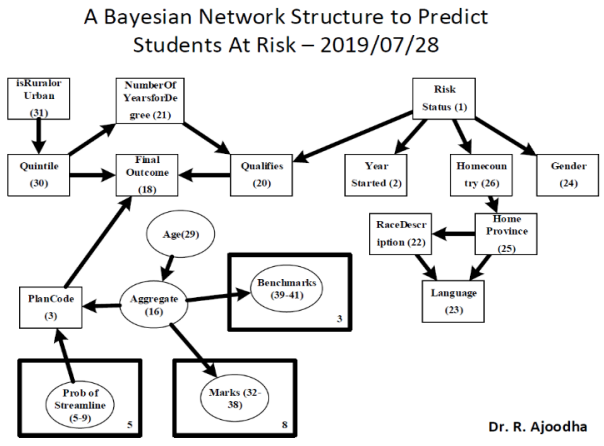


**Figure 1**

61 attributes were collected in total, which was narrowed down to the 18 most relevant. These variables are summarised in Table I.

TABLE I. Student Attributes Selected

| *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 |

## IV. Methodology

Attribution selection was performed first, followed by the application of various models. Initially, Information Gain [12] with 10-fold cross-validation was used in order to find the most useful attributes, using a cut-off of 0.001 average merit. The results are shown in Table II.

TABLE II. Information Gain Results

| *Attribute* | *Average Merit* | *Average Rank* | *Used* |
|---|---|---|---|
| Majors | 0.233 +- 0.003 | 1 +- 0 | Yes |
| SchoolQuintile | 0.014 +- 0.001 | 4 +- 0 | Yes |
| LifeSciences | 0.007 +- 0.001 | 5.6 +- 0.66 | Yes |
| MathematicsMatric Major | 0.007 +- 0. 001 | 5.7 +- 0.78 | Yes |
| NBTAL | 0.006 +- 0.001 | 7 +- 0.63 | Yes |
| LifeOrientation | 0.006 +- 0.001 | 7.7 +- 0.64 | Yes |
| EnglishFirstLang | 0.003 +- 0.001 | 9.9 +- 1.3 | Yes |
| NBTQL | 0.002 +- 0 | 10.4 +- 0.92 | Yes |
| Geography | 0.002 +- 0 | 10.5 +- 0.92 | Yes |
| PhysicsChem | 0.002 +- 0 | 12.3 +- 0.64 | Yes |
| AdditionalMathemat ics | 0.002 +- 0 | 12.8 +- 0.87 | Yes |
| AdditionalLanguage | 0.001 +- 0.001 | 13.5 +- 2.06 | No |

| *Attribute* | *Average Merit* | *Average Rank* | *Used* |
|---|---|---|---|
| MathematicsMatricLit | 0 +- 0 | 15.3 +- 0.64 | No |
| EnglishFirstAdditional | 0 +- 0 | 15.6 +- 0.92 | No |
| NBTMA | 0 +- 0 | 16.7 +- 0.9 | No |

The following models were used to learn the predictive functions, along with 10-fold cross validation:

- Bayesian Network – A probabilistic model that uses a directed acyclic graph (DAG) to represent variables and their conditional dependencies. These models work well when predicting the probability that one of several known causes is a contributing factor in a given event, such as a student qualifying. [13]
- Logistic Regression - A statistical model that uses regression to estimate the parameters of a logistic function with a binary dependent variable. It explains the relationship between one dependent binary variable and one or more independent variables. This model is ideal for this study as the dependent variable can only take one of two values, F or Q. [14]
- Multilayer Perceptron (MLP) – A deep learning model that comprises multiple layers of input nodes connected as a directed graph between the input and output layers. It is a feedforward artificial neural network that generates a set of outputs from a set of inputs and uses backpropagation to train the network. [15]
- K-Nearest Neighbours (KNN) – A simple algorithm that classifies a data point based on the classification of the k closest points around it, assuming similar data points are close to each other. The success of this algorithm depends on the k value chosen as well as the distance metric, and is very sensitive to outliers, so multiple k values are tested to find the one with the best predictive results. [16]
- Bootstrap Aggregating (Bagging) – An ensemble machine learning model which is composed of two parts: aggregation and bootstrapping. It selects samples with replacement and runs the learning algorithms on the selected samples. It aggregates the predictions of multiple weak machine learning models and chooses the best result, reducing variance and overfitting. [17]
- Random Forests – Another ensemble machine learning model. It creates multiple different decision trees which each produce their own class prediction and the class that is the mode or mean of the individual trees becomes the predictive model used in testing. It follows the idea that the whole is greater than the sum of its parts. With Random Forests, overfitting is less likely, and it is a good model when there are many missing values, which the case for the dataset used in this study, shown in Table III. [18]

TABLE III. NUMBER OF MISSING VALUES PER VARIABLE

| *Attribute* | *Not Missing* | *Missing* |
|---|---|---|
| SchoolQuintile | 10075 | 535 |
| LifeOrientation | 9312 | 1298 |
| MathematicsMatricMajor | 10231 | 379 |
| MathematicsMatricLit | 115 | 10495 |
| MathsLevel | 10610 | 0 |
| AdditionalMathematics | 2252 | 8358 |
| EnglishFirstLang | 5517 | 5093 |
| EnglishFirstAdditional | 4929 | 5681 |
| NBTAL | 6912 | 3698 |
| NBTMA | 6817 | 3793 |
| NBTQL | 6876 | 3734 |
| AdditionalLanguage | 10258 | 352 |
| PhysicsChem | 9995 | 615 |
| Geography | 4717 | 5893 |
| LifeSciences | 8729 | 1881 |

To evaluate their performance, the following evaluation metrics were used:

$$Recall = \frac{TP}{TP+FP} \quad (2)$$

$$Precision = \frac{TP}{TP+FN} \quad (3)$$

$$F-score = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}_i| \quad (4)$$

TP: True positives

FP: False positives

FN: False negatives

Precision is defined as the fraction of relevant instances among the retrieved instances, while recall is defined as the fraction of the total amount of relevant instances that were actually retrieved.

For example, if a model identifies 10 qualified students in a data set that comprises 14 qualified students and some failed students. 8 of the 10 identified as qualified did actually qualify (TP), while the other 2 did not (FP). Therefore, the precision is 8/10 while its recall is 8/14.

The F-score measures a model's accuracy. It is defined as the weighted harmonic mean of a test's precision and recall and provides a more realistic measure of a test's performance than precision or recall individually as it balances the use of both.

Mean absolute error (MAE) measures the size of the errors in a model. It takes an actual data point and subtracts the predicted value, which is the error for that point. It does this for every data point in the data set, sums up the absolute values of these errors and takes the average.

TABLE IV. CONFUSION MATRIX EXAMPLE

| | ***Positive*** | ***Negative*** |
|---|---|---|
| ***Positive*** | TP | FP |
| ***Negative*** | FN | TN |

Confusion matrices are used to measure the success of machine learning algorithms. It shows which proportion of a binary variable were correctly and incorrectly classified. An example matrix is shown in Table IV. In the context of this study, the top left cell of the matrix (TP) shows the number of data points that were Qualified and correctly classified , while the top right cell (FP) shows the number of data points that were Qualified but labelled as Failed. The bottom left cell (FN) shows the number of data points that were F but incorrectly labelled as Qualified, and the bottom right cell (TN) shows the data points correctly labelled as Failed.

## V. RESULTS

The results corresponding to the abovementioned machine learning models are presented in Table V.

Overall, the performance of the models was positive. Bagging produced the best results, classifying 75.97% of the data correctly. Random Forests came in a close second with 75.57% correct classifications. This success may be due to Random Forest's ability to perform well with many missing values. KNN performed the poorest, classifying 64.83% of the data correctly and with an F-score of 0.706 for F and 0.563 for Q. The algorithm suffered the most with the prediction of Q, predicting nearly as many correctly as incorrectly. The optimal value for k was found to be 1, which may explain why the predictive results were so low.

The remaining algorithms performed very similarly, with correctly classified instances ranging from 74.12% to 75.48%. Bayesian network correctly classified 74.12% of the data, with a mean absolute error of 0.3363. It performed better predicting F compared to Q, with 1240 incorrect classifications and 4749 correct for F, and 3115 correct and 1506 incorrect for Q.

TABLE V. RESULTS OF MACHINE LEARNING MODELS

| ***Algorithm*** | ***Qualified*** | ***TP Rate*** | ***FP Rate*** | ***Precision*** | ***Recall*** | ***F-score*** | ***Correctly Classified Instances*** | ***Mean absolute error*** | ***Confusion Matrix*** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | ***a*** | ***b*** |
| Bagging | a = F | 0.791 | 0.281 | 0.785 | 0.791 | 0.788 | 75.97 % | 0.3209 | 4738.84 | 1249.35 |
| | b = Q | 0.719 | 0.209 | 0.727 | 0.719 | 0.723 | | | 1300.45 | 3320.36 |
| Random Forests | a = F | 0.793 | 0.293 | 0.778 | 0.793 | 0.786 | 75.57 % | 0.3223 | 4748.39 | 1239.79 |
| | b = Q | 0.707 | 0.207 | 0.725 | 0.707 | 0.716 | | | 1352.13 | 3268.69 |
| Logistic Regression | a = F | 0.787 | 0.287 | 0.781 | 0.787 | 0.784 | 75.48 % | 0.3384 | 4711.15 | 1277.04 |
| | b = Q | 0.713 | 0.213 | 0.721 | 0.713 | 0.717 | | | 1324.56 | 3296.25 |
| Multilayer Perceptron | a = F | 0.814 | 0.331 | 0.761 | 0.814 | 0.787 | 75.09 % | 0.3026 | 4874.26 | 1113.93 |
| | b = Q | 0.669 | 0.186 | 0.735 | 0.669 | 0.701 | | | 1528.43 | 3092.39 |
| Bayesian Network | a = F | 0.793 | 0.326 | 0.759 | 0.793 | 0.776 | 74.12 % | 0.3363 | 4749 | 1240 |
| | b = Q | 0.674 | 0.207 | 0.715 | 0.674 | 0.694 | | | 1506 | 3115 |
| K-Nearest Neighbours | a = F | 0,748 | 0.481 | 0.668 | 0.748 | 0.706 | 64.83 % | 0.3517 | 4479.1 | 1509.09 |
| | b = Q | 0.519 | 0.252 | 0.614 | 0.519 | 0.563 | | | 2221.86 | 2398.96 |

The Multilayer perceptron model correctly classified 75.09% of the data and had a MAE of 0.3026. Logistic regression ranked third, correctly classifying 75.48 % of the data.

Bagging just edged out the other algorithms in terms of performance, perhaps due to the fact that it aggregates the results of multiple models. A student's major was found to be the most significant attribute which was surprising. School quintile was second, probably because it is highly correlated with the income level and quality of high school education the student received. Grades in subjects like Life Sciences, Pure Mathematics, Life Orientation and English were also good predictors of success, but unexpectedly, Maths Lit, AP Maths and NBT maths marks were not.

## VI. CONCLUSION

In this research the aim was to discover, through the use of a wide range of machine learning models, if the features of a student's high school career are good predictors of their success at university in South Africa, and if so, which features. The feature space used in this study consisted of the most commonly taken high school subjects, NBT marks, the quintile of the school the student attended, as well as the major a student chose when registering at university.

Using information gain to determine how much each attribute contributed towards the model a threshold was determined, and four features were removed; Additional Language, Mathematics Matric Lit, English First Additional and NBTMA, leaving 11 features to use for prediction.

Bayesian Networks, Logistic Regression, Multilayer Perceptron, K-Nearest Neighbours, Bootstrap Aggregating, and Random Forests were applied to the dataset, using 10-fold cross validation to evaluate the accuracy of the classifiers. 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%. It can be concluded from the results that a student's undergraduate major, school quintile, Life Sciences and Mathematics marks in matric have a significant effect on their chance of

graduating.

Learning analytics is a rich and exciting area of research and there is a great amount of knowledge to be discovered that could significantly improve the lives and quality of education in South Africa for both students and educators. Due to South Africa's high university dropout rate, attention needs to be given to identifying at-risk students as early on as possible so that the necessary support and guidance can be provided, ensuring every student has the highest chance of graduating, regardless of their background or disadvantages.

Going forward, research in this area should focus on gathering real data on students in South Africa and testing whether these results hold up and which attributes are significant. The findings of this study, when applied to real data, could help educators make more informed decisions when trying to identify vulnerable students by providing them with reliable indicators that are readily available and easily accessible.

## ACKNOWLEDGEMENT

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

## REFERENCES

[1] "South Africa's university pass rate shocker," BusinessTech, 13-Nov-2019. [Online]. Available: https://businesstech.co.za/news/government/353575/south-africas-university-pass-rate-shocker/.

[2] P. Moodley, R. J. Singh, "Addressing student dropout rates at South African universities," Alternation (Durban), DHET, 2015.

[3] 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, May 1984.

[4] D. F. Butcher and W. A. Muth, "Predicting performance in an introductory computer science course," Communications of the ACM, vol. 28, no. 3, pp. 263–268, Jan. 1985.

[5] K. Danko, J. C. Duke, and D. P. Franz, "Predicting Student Performance in Accounting Classes," Journal of Education for Business, vol. 67, no. 5, pp. 270–274, 1992.

[6] A. Goold and R. Rimmer, "Factors affecting performance in first-year computing," ACM SIGCSE Bulletin, vol. 32, no. 2, pp. 39–43, 2000.

[7] L. Spark, "Matric maths as a predictor of success in Information Systems---a study of Information Systems students at the University of the Witwatersrand", Proceedings of the 35th Conference of SACLA, pp 268-271, 2005.

[8] J. Superby, J.P. Vandamme, N. Meskens, "Determination of factors influencing the achievement of the first-year university students using data mining methods", Workshop on educational data mining, vol. 32, pp. 234, 2006.

[9] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, "Predicting Student Performance using Advanced Learning Analytics," Proceedings of the 26th International Conference on World Wide Web Companion - WWW 17 Companion, pp. 415–421, 2017.

[10] M. Tsiakmaki, G. Kostopoulos, G. Koutsonikos, C. Pierrakeas, S. Kotsiantis, and O. Ragos, "Predicting University Students Grades Based on Previous Academic Achievements," 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1–6, 2018.

[11] F. Eibe, ClassBalancer. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/filters/supervised/instance/ClassBalancer.html.

[12] M. Hall, InfoGainAttributeEval. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html.

[13] R. Bouckaert, BayesNet. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/BayesNet.html.

[14] X. Xu, Logistic. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/functions/Logistic.html.

[15] M. Ware, MultilayerPerceptron. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/functions/MultilayerPerceptron.html.

[16] S. Inglis, IBk. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/lazy/IBk.html

[17] F. Eibe, Bagging. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/meta/Bagging.html.

[18] R. Kirkby, RandomForest. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/RandomForest.html.

[19] Statistics on Post-School Education and Training in South Africa. Department of Higher Education and Training, 2018, pp. 11,23

[20] Ritesh Ajoodha, Ashwini Jadhav, and Shalini Dukhan. 2020. Forecasting Learner Attrition for Student Success at a South African University. In Conference of the South African Institute of Computer Scientists and Information Technologists 2020 (SAICSIT '20), September 14–16, 2020, Cape Town, South Africa. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3410886.3410973