Using Bayesian Networks and Machine Learning to Predict Computer Science Success

Zachary Nudelman(✉), Deshendran Moodley, and Sonia Berman

Department of Computer Science, University of Cape Town, Cape Town, South Africa
ndlzac001@myuct.ac.za, {deshen,sonia}@cs.uct.ac.za

Abstract. Bayesian Networks and Machine Learning techniques were evaluated and compared for predicting academic performance of Computer Science students at the University of Cape Town. Bayesian Networks performed similarly to other classification models. The causal links inherent in Bayesian Networks allow for understanding of the contributing factors for academic success in this field. The most effective indicators of success in first-year 'core' courses in Computer Science included the student's scores for Mathematics and Physics as well as their aptitude for learning and their work ethos. It was found that unsuccessful students could be identified with ≈91% accuracy. This could help to increase throughput as well as student wellbeing at university.

Keywords: Bayesian Networks · Machine learning
Educational Data Mining · Computer science education

1 Introduction

In the past two decades, the broader field of Educational Data Mining (EDM) has developed into a respected and extensive research field [4]. According to the EDM website,[1] EDM can be defined as *"an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in"*. Using EDM for prediction usually involves developing a model which accepts certain variables (factors affecting the prediction) and outputs an expected result for the predicted variable [3].

EDM, specifically machine learning techniques, can be used to explore factors contributing to the success of high school applicants for the bachelor of computer science curriculum (BSc(CS)) at the University of Cape Town (UCT). While the graduation rate of computer science (CS) majors at the UCT has been reasonably good over the past decade, there is a strong desire to improve throughput and time-to-graduation. Furthermore, in the South African context, UCT needs to consider social redress when admitting applicants. An accurate

[1] http://www.educationaldatamining.org.

S. Kabanda et al. (Eds.): SACLA 2018, CCIS 963, pp. 207–222, 2019.
https://doi.org/10.1007/978-3-030-05813-5_14

model to predict student success could allow UCT to continue along its path of social transformation while maintaining and indeed improving levels of success.

Available literature shows that one of the best predictors of academic success at university relates to the marks attained thus far in a student's academic career [17]. However, when using secondary-school marks to predict university performance in South Africa, this may not be the case, because the basic education system in South Africa is worse than *"(almost) all middle-income countries that participate in cross-national assessments of educational achievement"* as well as *"many low-income African countries"* [18]. The South African Department of Basic Education categorises schools into quintiles according to socio-economic status, where Quintile 1 schools have the least resources and Quintile 5 schools the most. Independent or private schools are usually well resourced; the term may also refer to some schools with a non-government syllabus that are poorly resourced (e.g. some missionary schools). For this paper, 90% of the independent schools considered are private and well-resourced.

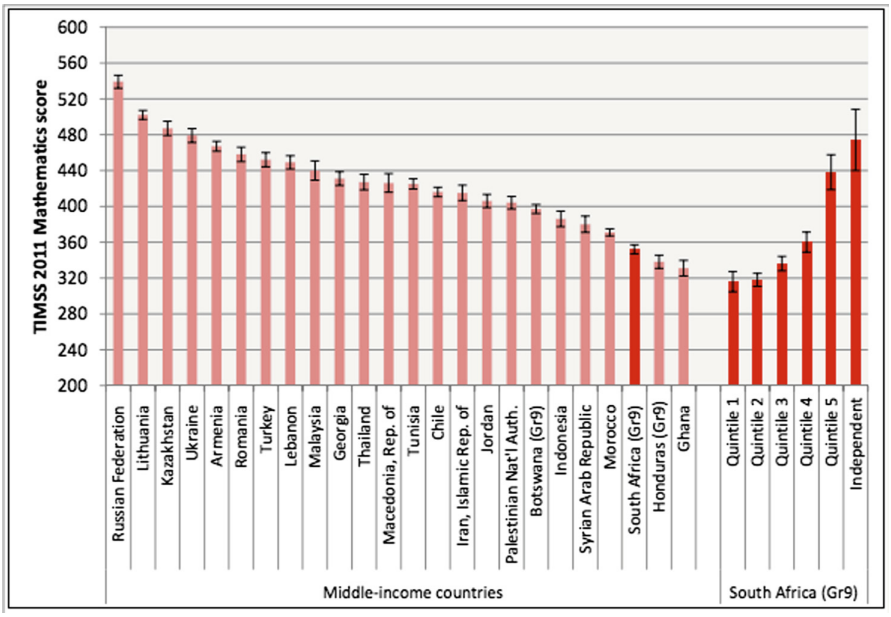

Fig. 1. Average grade-9 test scores for TIMSS middle-income countries, 2011 [18]

Figure 1 shows the disparity in the average mathematics performance of South African school-pupils in different quintiles. A positive correlation exists between quintile and academic performance i.e., in general, the poorer the school, the worse a student's performance in mathematics and science [18]. However, since the quality of education is so varied, a pupil in a lower quintile school may have a greater aptitude for independent learning and/or a better work ethos than a student from a higher quintile school who achieves similar marks.

Another potential factor is the South African province in which a school is located, due to differing standards of education between the provinces in this country. Repeatedly this difference plays a role in the annual *Matric* pass rates, drop-out rates and skills obtained (e.g., literacy) [18].[2] Furthermore, emotional difficulties associated with relocating to a new city may affect academic performance [1]. More than just matric marks, these are potential factors that could also contribute to a more effective prediction model.

This paper aims to explore Bayesian Networks (BN) and Machine Learning (ML) techniques to predict success in the BSc(CS) degree at UCT. Available data included applicants' Matric scores, National Benchmarking Test (NBT) scores, the quintile of their school, and the province in which it is located.

The remainder of this paper is set out as follows: Section 2 reviews related literature on predicting academic performance. Section 3 explains the machine learning methods used in this paper. Section 4 describes how the data was transformed and the model constructed. Section 5 discusses model evaluation and results achieved. Section 6 covers limitations and future work, and Sect. 7 concludes.

2 Related Work

As explained in [16], predicting academic performance of students *"is one of the oldest and most popular applications of DM (Data Mining) in education"* [16]. The most prevalent prediction models include: Decision Trees, Neural Networks, Naïve Bayes, K-Nearest Neighbour, Support Vector Machines, Random Forests and Bayesian Networks [4,17]. Bayesian Networks can handle missing values and have the ability to be queried and give answers that explain their predictions [10]. This allows further enquiry into the causal links between academic success and the predictive variables, i.e., Bayesian Networks are 'white-box' models [20]. This will be crucial for explaining and understanding the predictions of a Bayesian Network model over the typically 'black box' machine learning methods, such as Neural Networks, which do not offer any clear explanations for their predictions. A similar justification is made for using the Decision Tree, Naïve Bayes and Random Forest models as benchmarks for the Bayesian model, i.e., out of the remaining approaches, these model are the easiest to understand [20].

Various papers have compared the predictive performance of these models. The following examples provide a brief overview of the field. Nghe et al. compared the accuracy of Decision Trees and Bayesian Networks in predicting the academic performance of over 21,000 graduate students at two different tertiary institutions [13]. Variables included were grade point average (GPA), prior institution rank, and other factors. They found that Decision Trees were slightly more effective than Bayesian Networks (76.3% vs 71.2% accuracy). Similar results were obtained by [12] for 826 CS students over seven years at the University of the

[2] In South Africa, 'Matric' is name of the formal qualification level of pupils who have passed their secondary school (high school) education after school-year 12 before university—somewhat similar to the Austrian 'Matura' or the German 'Abitur'.

Witwatersrand in South Africa [12]. Their Decision Tree model outperformed the Naïve Bayes algorithm with accuracies of 84.5% and 78.9%, respectively. However, those studies investigated different questions: [13] predicted success of students enrolled in post-graduate programmes as well as 3rd-year students' success from their 2nd-year results, whereas [12] predicted 1st-year final grades based on 1st-year 1st-semester results.

According to [14], the Naïve Bayes classifier (76.7%) outperforms Decision Trees (73.9%) and Neural Networks (71.2%) in predicting 1st-years' academic success in Business Informatics [14]. However, their data set had only 257 student records. More recently, Asif et al. produced similar conclusions with their data set of 200 undergraduate students when predicting future performance [2], however with a bigger difference in the accuracy of the compared models, (Naïve Bayes: 83.7%, Decision Tree: 66.0%).

While there are further papers exploring such topics, their results produced are always similarly varied. It thus appears that the accuracy of prediction models is highly dependent on the variables selected, the question addressed, and the context of the investigation. Consequently it seems necessary to investigate specifically Bayesian Network prediction of CS student performance at UCT, and to compare the accuracy with other techniques.

3 Methods

Classification is the assignment of a label or category to a sample of data based on a number of variables [11]. In particular, *supervised* classification allows a model to learn how to classify some attribute of unknown data samples from labeled data samples [14]. The chosen attribute for classification is known as the 'target' variable. There are various methods to address this problem. The four machine learning methods used in this paper are defined and briefly explained below.

3.1 Bayesian Networks

A Bayesian Network is a directed acyclic graph representing a particular domain. Each node of the graph represents a variable from the domain. The nodes are connected by arcs which represent the dependencies between variables. Each arc is assigned a weight using

$$\textbf{Bayes Theorem } [11]\text{: } P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}.$$

Bayesian Networks can be used for a wide range of applications including reasoning, analysis, diagnosis, risk assessment and evaluation [11]. Bayesian Networks are particularly useful for classification tasks as they provide an explanatory model, in contrast to techniques such as neural networks. Constructing a Bayesian Network requires the assumption of the *Markov property* whereby *"there are no direct dependencies in the system being modeled which are not already explicitly shown via arcs"* [11]. It holds for causal and predictive models.

Naïve Bayes. The Naïve Bayes classification technique uses a Bayesian model where the target variable is the parent of all the predictor nodes, i.e., these models assume independence between every predictor variable [11]. Even with these assumptions it is still an accurate and efficient prediction model for many problems [11].

3.2 Decision Trees

A Decision Tree uses the concept of *information entropy* to divide the classification into subproblems which are simpler to solve. Each node's section of the input space is recursively divided into subsections through its descendants. A node with no descendants indicates a prediction made by the model. Thus the higher up in the tree an attribute appears, the more influential it is in dividing the data. When passed a set of input variables, it can produce an expected classification based on the tree that it has learned. There are a variety of algorithms which use training data to construct these trees such as the C4.5 [15]. Tools exist which automatically invoke this algorithm such as WEKA's J48 classification filter [9].

Random Forests. Random Forests construct multiple decision tree classifiers that are trained on different subsets of the data with independent random subsets of the features of the data [6]. Once numerous trees have been constructed, the model classifies a data input by outputting the most popular decision, i.e., the class chosen by the majority of the trees [6]. This method can reduce the 'overfitting' of data that can occur with Decision Tree classifiers [7].

4 Experimental Design

4.1 Data Pre-Processing and Analysis

The data set used in this paper covered a cohort of 783 students who were admitted into the CS major at UCT between 2007 and 2016. The term 'graduateCS' is used for students qualifying with a BSc(CS); others, who are not 'graduateCS', may graduate with another major instead.

The CS major at UCT is a three-year programme consisting of six 'core' CS courses as well as a mandatory first-year course in mathematics. CS entrant numbers are increasing annually and most courses have consistent pass rates with no clear trends.

UCT offers an option for weak or under-prepared students to complete the same BSc degree over four rather than three years (Extended Degree Program: EDP). CS students are expected to complete Computer Science courses CSC1015F and CSC1016S as well as Mathematics MAM1000W in their first year. EDP students take extended versions of these courses 'stretched' over two years. Consequently, a student who is likely to fail at least one of the three required first year courses should be encouraged to join the EDP. In this context

we define 'at-risk' as 'likely to fail CSC1015F, CSC1016S or MAM1000W'. On average 58.83% of all CS entrants who attempt to complete their first-year in one year fail at least one of CSC1016S and MAM1000W.

Only ≈15% of all CS entrants receive financial aid. The majority of these financial aid receivers come from Quintile 4 schools (30%) whereas only 25% come from Quintile 1, 2, 3 schools (cumulatively).

Table 1. Raw data attributes

Data Context	Attribute (Classifier)	Data Type
Application	Year	Continuous
Application: *Matric results*	Mathematics	Continuous
	Advanced mathematics	Continuous
	English	Continuous
	Physical sciences	Continuous
Application: *NBT results*	Mathematics	Continuous
	Quantitative literacy (QL)	Continuous
	Academic literacy (AL)	Continuous
Progress	Courses Registered	Continuous
	Courses passed	Continuous
	Cumulative GPA	Continuous
	Financial aid	Binary (y/n)
Throughput	Time to graduate	Nominal
	Dropped out (or excluded)	Nominal
	Course marks for each course	Nominal
Secondary school	Province	Nominal
	Quintile	Nominal

Table 1 lists the data elements in our data set. The South African Master Schools list was also used to retrieve the quintile of the applicants' school. We anonymised all data before continuing to use them with new anonymous identifiers. A preliminary analysis of the data lead to selecting the input variables and to determining the state space of each variable. Firstly, 'AP Maths' was reduced to a categorical variable to indicate if a CS entrant had attempted it, since there are too few data points to make a significant finding with respect to marks achieved for it. Additionally, by means of regression analysis, scores with high correlations (above 65%) were averaged and reduced, since these individually would not provide the model with further information. Secondary school physics and maths had a 69.8% correlation, and NBT maths and school maths had a 66.4% correlation. NBT AL and NBT QL had a 75.3% correlation. Consequently, the following reduced variables were introduced: 'AveSciences' for science, maths and NBT maths, and 'ALQL' for NBT AL and NBT QL.

The distribution of the quintile of the schools the 783 CS entrants had attended was analysed. There were only 9 CS entrants from Quintile 1 schools, 22 from Quintile 2 schools, and 54 from Quintile 3 schools. Consequently, these quintiles were combined under the new variable 'Lower': see Table 2. The quintile of the school attended by 49 CS entrants could not be determined because of insufficient information.

Table 2. Computer science intake per Quintile

	Lower	Quint.4	Quint.5	Priv.	Int.	No Data	TOTAL
# Students	85	61	285	187	116	49	**783**
% of total	10.86%	7.79%	36.40%	23.88%	14.81%	6.26%	100%

The distribution of the quintiles of the school attended across provinces was analysed. The majority of students from low quintile schools attended schools in the Eastern Cape and Limpopo, which are among South Africa's poorest provinces. Most students from independent schools came from Gauteng, which is South Africa's economically most productive province. Consequently, the 'Province' variable was split into the following: 'Eastern Cape and Limpopo'; 'Gauteng'; 'Western Cape'; 'International'; and 'Other'.

4.2 Machine Learning Procedure

Before beginning the machine learning procedure, the data had to be split into training and testing sets to avoid over-fitting. Cross-validation was used to train and test the models during the machine learning process, whereby the last year (2016) was kept aside as a hold-out set for evaluation of the final models. The cohort from each year was used as a fold, e.g., 2007 was used as the test set and all other years (2008–2015) as the training set in the first fold, and so on. Thus, there were nine folds in total, one for each of the years between 2007 and 2015.

Measures of Success. CSC1016S is the second-semester course of our first-year CS degree. It is the first course exclusively required by CS entrants and justifiably so as it introduces a more theoretical foundation of modern day programming with a focus on the Java language. The course introduces concepts such as memory referencing, inheritance, and data type abstractions as well as an introduction to data structures with its module on linked lists. These provide a robust foundation for any construction of CS knowledge and consequently, without it, a CS entrant is unlikely to succeed. Succeeding in the course is defined as achieving over 50%. The probability of CS entrants passing this course was found to be 95.53% and their chance of graduating in CS if this course was passed was 55.8%.

MAM1000W is the first-year mathematics course and is the first-year course for Mathematics majors. It introduces *"fundamental ideas in calculus, linear*

algebra, and related topics" [19]. The broad scope of Computer Science requires not only an understanding of these topics, but also fluency with mathematical language and logic. More importantly, MAM1000W requires a far stronger work ethos than what is required at school. Consequently, successfully passing this course allows the CS department to discern whether a student will be able to keep up and succeed in the degree. Succeeding in MAM1000W is defined as achieving over 50% marks for the final course result. The probability of CS entrants passing MAM1000W was found to be 57.39% and their chance of graduating in CS if they passed MAM1000W was 93.64%.

Furthermore, passing all core first year courses (CSC1015F, CSC1016S, and MAM1000W) required for a major in CS on first attempt shows proficient understanding of foundational knowledge for CS as well as an ability to cope with vast jumps in course load and difficulty. Consequently, passing these courses on first attempt provides an even stronger indication of academic success. It was found that the chance of a CS entrant graduating in CS having passed these core first-year courses on first attempt is 97.58%, though their chance of passing them on first attempt is only 41.45%. This consequently shows a student to be 'At-Risk', or not.

After consultation with CS student advisors at UCT, the following target variables for prediction were chosen:

1. Passing CSC1016S, i.e., achieving over 50% on first attempt;
2. Passing MAM1000W, i.e., achieving over 50% on first attempt;
3. At-Risk (*Failing* CSC1015F, CSC1016S or MAM1000W on first attempt);
4. Graduation (eventually, after any amount of time);
5. Graduation (in minimum time as per course-book).

Determining Causal Structure and Parameterization of the Bayesian Network. The network structure was developed over several iterations of development and evaluation with a student advisor.

The variables 'AveSciences', 'AP Maths' and 'ALQL' were included in the structure, since, as explained above, these are often the most powerful predictors of academic success. For reasons explained above, both province and quintile were initially incorporated into the model with links to the Matric score variables. It may seem as though one's school's quintile would be highly correlated with their financial aid state; however, as shown in Sect. 4.1, financial aid is received in fairly equal proportion by CS entrants from all quintiles.

When considering the causal relationships between the variables, it makes little sense to assume that school marks directly affect university results. Rather, there must be some more general (earlier) cause of success for *both* school *and* university. Firstly, a student's success in academic studies is determined by their ability to understand the knowledge that is being provided to them as well as their skills to apply this knowledge. Secondly, the strength of a student's work ethos will affect their academic performance, too. In a Bayesian Network, these variables are known as 'latent' or 'hidden' variables [11]. Similar to any other

variable in the network, there are causal links that point to and from these nodes. If a student has a higher aptitude for learning and understanding knowledge, as well as a strong work ethos, one would expect this to affect the school results (the input variables) as well as the University results (the target variables).

In parameterizing a Bayesian Network, the training data are used to determine prior probabilities of each variable; then the conditional and posterior probabilities are calculated [11]. However, by definition, latent variables are unmeasurable. Consequently, the Expectation Maximization (EM) algorithm was used to approximate the most likely values for these variables. The algorithm chooses a value, e, at random to compute the *"probability distribution over the missing values"* and iteratively uses a maximum likelihood to determine a new value e [11]. The algorithm iterates until the maximum likelihood stabilizes.

Upon comparison of results with the Naïve Bayes model, the Bayesian Network was performing suboptimally. Consequently, an iterative process of relaxing the independence assumptions of the Naïve Bayes was undertaken in order to find the best performing structure. This may provide greater insight into which attributes contributed to effective predictions and which introduced noise in the model.

The Bayesian Network was developed using the Norsys *Netica* software package.[3] The software also allows for learning latent variables with an EM learning algorithm.

Discretising Continuous Variables and Selecting Significant Attributes. Discretising variables for machine learning can *"improve the performance of the algorithm and reduce the computation time considerably"* [8]. In order to determine suitable inflection points for each numeric attribute as well as which variables were significant, the J48 Decision Tree algorithm was used on each fold for each model. The resulting trees were then analysed to determine which attributes were to be included in the Decision Tree as well as what values for numeric attributes were to be used. From this, various possible sets of data were produced for each model. After testing each data set, the best performing data set was chosen. The variables and inflection points found were thus shown to be contributive to the target variable according to the machine learning algorithm.

While this procedure was conducted to improve the performance of the predictive models, its outcome is a result itself. These variables and inflection points are significant as they provide an initial understanding of which variables are indeed contributive to the target variables, and consequently, to success in the CS curriculum. This final state space emanating from the data pre-processing, data analysis and initial experiments is shown in Table 3.

[3] http://www.norsys.com/download.html.

Table 3. Final state space

Attribute	CSC1016S	MAM1000W	At-Risk	Graduation	Time to Gr.
AveSciences	Low (<68) Mid (68...78) High (>78)	Low (<79) High (≥79)	Low (<79) Mid (79...83) High (>83)	Low (<67) Mid (67...83) High (>83)	Low (<79) Mid (79...83) High (>83)
English	Low (<71) Mid (71...83) High (>83)		Low (<72) High (≥72)	Low (<64) Mid (64...76) High (>76)	Low (<77) High (≥77)
ALQL	Low (<68) High (≥68)			Low (<71) High (≥71)	Low (<71) High (≥71)
Adv. Maths attempted	Yes/no	Yes/no	Yes/no	Yes/no	Yes/no
Financ. Aid	Yes/no			Yes/no	Yes/no
Province	Gauteng Western Pr. ECLP International Other			Gauteng Western Pr. ECLP International Other	Gauteng Western Pr. ECLP International Other
Quintile	Low, Q.4, Q.5, Indep.			Low, Q.4, Q.5, Indep.	Low, Q.4, Q.5, Indep.
Target variable	**Fail** (<50) **Pass** (≥50)	**Fail** (<50) **Pass** (≥50)	**No** (passed all at first time)/**yes**	**Yes/no**	**ThreeYears** **Over3years** **NotGradCS**

5 Results and Discussion

5.1 Results

Once the models were constructed and parameterized, each model was tested using the originally withheld testing set. Sensitivity (TPR) and Specificity (TFR) were used to measure the predictive power of each model.

$$\text{Sensitivity} := \frac{TruePredictedPasses}{TruePredictedPasses + FalsePredictedFails}$$

$$\text{Specificity} := \frac{TruePredictedFails}{TruePredictedFails + FalsePredictedPasses}$$

The Matthews Correlation Coefficient (MCC) was also used as a useful measure for model performance on unbalanced data [5]. It measures the correlation between the actual and predicted classifications for all classes. An MCC of 1 indicates a perfect prediction while −1 indicates complete disagreement. A number of experiments were done to predict the five target variables identified earlier, with a focus on predicting At-Risk students. The results are summarized below.

CSC1016S. The Bayesian Network had the highest MCC of 0.38 and the highest TFR of 23%. It also had the joint highest TPR of 98.8%.

MAM1000W. The Bayesian Network and the Naïve Bayes models both scored the highest MCC of 0.43 as well as the highest TPR of 79%. The J48 and Random Forest models attained the best TFR with 70%. The models for this prediction were most effective using only the AveSciences and AP Maths variables as input. This indicates that all other variables did not influence the MAM1000W result.

At-Risk. The variables used for these models were similar to those used for the MAM1000W models, with the addition of Matric English scores. The results for all models for the At-Risk variable are shown in Tables 4, 5, 6 and 7. For this variable, 'Sensitivity' refers to prediction of 'Not At-Risk' students while 'Specificity' refers to prediction of 'At-Risk' students.

Table 4. Performance of the Random Forest for the At-Risk variable

Random Forest	Hold-Out Set	Average	Std.
Specificity	92.42%	90.82%	8.95%
Sensitivity	66.67%	59.70%	13.39%
MCC	61.98%	53.47%	9.30%
F1	84.14%	67.91%	6.71%

Table 5. Performance of the J48 Decision Tree for the At-Risk variable

J48 DT	Hold-Out Set	Average	Std.
Specificity	92.42%	86.92%	13.64%
Sensitivity	66.67%	63.61%	17.58%
MCC	61.98%	53.47%	9.30%
F1	84.14%	67.80%	6.55%

Table 6. Performance of the Naïve Bayes for the At-Risk variable

Naïve Bayes	Hold-Out Set	Average	Std.
Specificity	74.24%	82.40%	18.62%
Sensitivity	70.37%	64.17%	18.71%
MCC	44.54%	49.50%	9.60%
F1	69.72%	65.25%	6.64%

Table 7. Performance of the Bayesian Network for the At-Risk variable

Bayesian Network	Hold-Out Set	Average	Std.
Specificity	92.42%	59.76%	13.77%
Sensitivity	66.67%	90.64%	9.26%
MCC	61.98%	53.34%	8.77%
F1	84.14%	67.91%	6.71%

Table 8. Confusion matrix showing performance of the Naïve Bayes and Bayesian Network models on the 2016 cohort

Naive Bayes Classified as			Bayesian Network Classified as		
	At-Risk	Not At-Risk		At-Risk	Not At-Risk
At-Risk	49	17	At-Risk	61	5
Not At-Risk	16	38	Not At-Risk	18	36
MCC: 0.445	Specificity: 74%	Sensitivity: 70%	MCC: 0.620	Specificity: 92%	Sensitivity: 67%

Table 9. Quintiles of misclassified At-Risk students compared to quintiles of students from full data set. Note the strong %-discrepancy in the 'independent' category!

	Qunit.1	Quint.2	Quint.3	Quint.4	Quint.5	Indep.	TOTAL
# misclassif. stud.	0	1	5	5	16	20	**47**
% of misclassified	0.0%	2.1%	10.6%	10.6%	34.0%	**42.6%**	100%
% of all students	1.1%	2.8%	6.9%	7.8%	36.4%	**23.9%**	

Table 8 shows the predictive performance of the Naïve Bayes and the Bayesian Network for the 2016 cohort. The Naïve Bayes had the lowest TFR of 74.24% on the 2016 cohort while the Bayesian Network had a TFR of 92.42% and an MCC of 0.62. The J48 and Random Forest models produced results similar to the ones of the Bayesian Network. The final Bayesian Network structure is shown in Fig. 2.

Graduating with CS Major. Similar to the CSC1016S models, each variable was seen as contributing to the predictions. The Bayesian Network had the highest specificity of 61% while the Random Forest performed best at predicting graduating with a sensitivity of 91% and the highest MCC of 0.39.

Time-to-Graduation with CS Major. The Bayesian Network attained the greatest Minimum Time True Positive Rate of 70% as well as the best MCC of 0.24. However, the Naïve Bayes had the best True Positive Rate for not graduating and the Random Forest and J48 were able to predict graduating in more than minimum time with a true positive rate of 9%. None of the models was able to achieve satisfactory results in predicting graduating in more than minimum time.

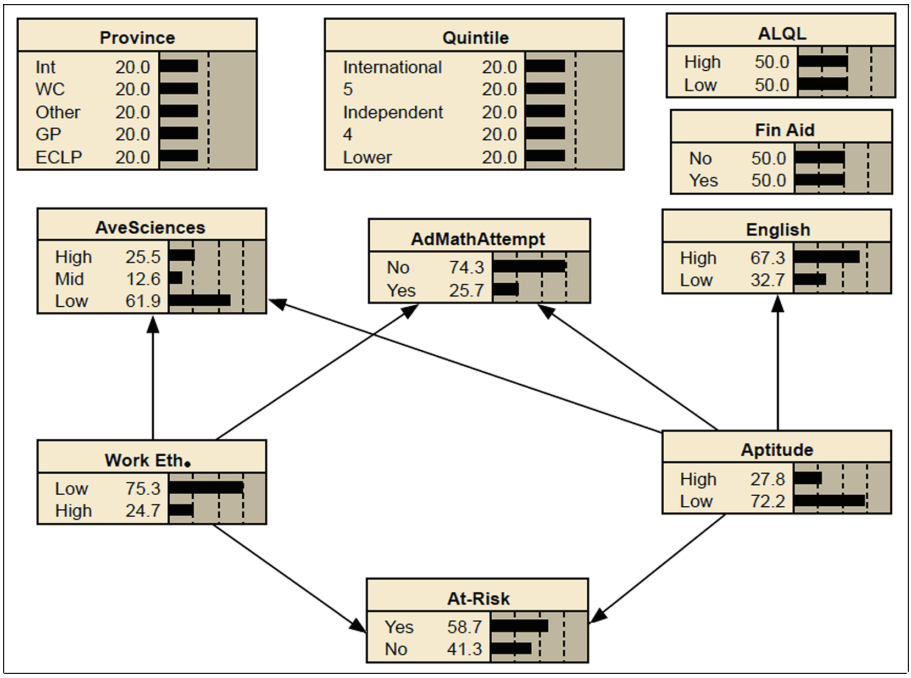

Fig. 2. Bayesian Network for At-Risk students

5.2 Discussion

The *inadequate accuracy for predicting failing* CSC1016S may be a result of the high (86%) CS entrant pass rate of this course—compared to the 61% CS entrant pass rate of MAM1000W. The poor results for Graduating in Minimum Time prediction are less clear; it is possible that there are no signifying features in the given data to indicate how long a CS entrant may take to complete the degree. However, this model performed better than the Graduation model for predicting 'no graduation'. The Bayesian Network achieved 72% accuracy (on average with a standard deviation of 10%) compared to 35% (on average with a standard deviation of 19%) for the Graduation model. Even when using the same data set as the Minimum Time Graduation model (i.e., the same points of inflection and variables included) for the graduation model, the results remained the same. This should be considered in future work.

As far as algorithm performance is concerned, our results were mixed. The Decision Tree and Random Forest models performed similarly to the Bayesian Network for achieving a high specificity (92%) for predicting At-Risk students, while the Naïve Bayes had the highest sensitivity (70%). The findings support the mixed results from the literature which show optimal performance oscillating between the Naïve Bayes, Decision Tree and Random Forest algorithms. In earlier studies, the Naïve Bayes model achieved the highest accuracy of 83.65% [2]

and a lowest accuracy of 76.65% [14]. The best performing Decision Tree was in [12] (84.46%) and the worst in [2] (66.03%). While few studies used Bayesian Network for classification, the one which we found achieved an accuracy of 71.23% [13].

Of particular interest is the At-Risk model. All classification models but the Naïve Bayes performed similarly, with specificities of 92.42%. The prediction errors of this model can be a result of two causes. Firstly, a False Failure Rate (FFR) indicates the model predicted CS entrants At-Risk when they in fact were not. These CS entrants misclassified as being At-Risk would be advised to take the extended courses in vain. However, the second and more pressing source of error is the False Passing Rate (FPR). This figure indicates the number of CS entrants that were classified as Not At-Risk when they in fact were At-Risk, i.e., the model failed to recognise those who are in need of assistance. For the hold-out set, only 5 out of 66 students were misclassified as Not At-Risk. However, it is necessary to try and analyse these misclassified students and determine if there is a prominent reason for their misclassification.

Following such analysis we found that a disproportionate number of CS entrants who were misclassified as not being At-Risk were from independent schools, specifically private schools, (42% compared to 24% for the full data set) as shown in Table 9. Contrary to the initial results found in the Bayesian Network model, attending a private school may be a contributing factor to predicting academic success. Consequently, future research should include this as a variable in prediction models. Furthermore, only 8.33% of misclassified At-Risk students received financial aid compared to 19.8% of students from the full data set. This is another variable that should be explored in subsequent research.

6 Limitations and Future Work

Our data was limited to CS degree applicants over the last ten years. Consequently there were no data for anyone who changed their major to CS or who took CS courses as electives. If larger data sets were used, the models evaluated could have been more accurate or realistic. Additionally, the CS department at UCT has a very limited number of applicants who attend lower quintile schools as well as applicants who are on financial aid. This resulted in the data being biased towards CS entrants in higher quintiles and thus an accurate reflection of success of these categories of CS entrants could not be obtained. Since this study was conducted with UCT students, the situation might perhaps differ at other South African universities.

Bayesian Networks use known data to learn conditional probabilities of the network. Any unknown variable can only be approximated using various algorithmic techniques. Consequently, the variables ‘Aptitude’ and ‘Work Ethos’ in the Bayesian Networks may not be realistic or accurate and their addition should be further explored. Future studies into this specific topic should focus on understanding the effect of different partitionings of the state space for different variables to try and attain a more nuanced understanding of the contributing

factors of performance. As an example, initial experiments show that using a pass mark of 51% can have a substantial impact on predictive performance.

Finally, while financial aid and quintiles initially seemed to have no effect on the predictive performance of the models, analysis of misclassification of At-Risk students suggested that financial aid and quintile indeed affects the models' results. Hence, further studies should explore the inclusion of these factors (possibly with different partitions) in the experiments.

7 Conclusions

Comparatively, Bayesian networks do not outperform other classifiers, but can attain a similar performance to other classifiers. However, the usefulness of a Bayesian Network does not lie solely in its ability to predict classes. Its visual nature provides insight and greater understanding into the causes of success and the contributing factors to success. Bayesian Networks have a high potential to predict students at risk of not passing their core first-year courses in Computer Science. In particular, failing at least one of the first-year mathematics and computer science courses can be predicted with a 90.64% accuracy (on average). This finding justifies the method of identifying At-Risk-students automatically. Once these students are identified, they can be enrolled in the EDP in order to improve students' academic success and graduation throughput.

The key contributing factors were found to be the marks the students received in secondary school for Mathematics, Science and English, whether or not the student had attempted the AP Maths subject at school, and their aptitude and work ethos. It was initially found that in predicting these At-Risk students, the students' province and quintile did not play a discriminatory role, though deeper analysis suggests that more research will be needed to reach a better conclusion in this matter.

While this paper described the effectiveness of Bayesian Networks to predict and analyse academic success in Computer Science at UCT, further research is required to better 'unpack' these results as well as to improve the predictive performance of the underlying models.

References

1. Andrade, M.S.: International students in English-speaking universities: adjustment factors. J. Res. Int. Educ. **5**(2), 131–154 (2006)
2. Asif, R., Merceron, A., Ali, S.A., Haider, N.G.: Analyzing undergraduate students' performance using educational data mining. Comput. Educ. **113**, 177–194 (2017)
3. Baker, R.S.: Data mining for education. Int. Encycl. Educ. **7**(3), 112–118 (2010)
4. Baker, R.S., Yacef, K.: The state of educational data mining in 2009: a review and future visions. J. Educ. Data Min. **1**(1), 3–17 (2009)
5. Boughorbel, S., Jarray, F., El-Anbari, M.: Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS One **12**(6), 1–17 (2017)
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

7. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. SSS. Springer, New York (2001). https://doi.org/10.1007/978-0-387-21606-5
8. Gupta, A., Mehrotra, K.G., Mohan, C.: A clustering-based discretization for supervised learning. Stat. Probab. Lett. **80**(9), 816–824 (2010)
9. Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. **11**(1), 10–18 (2009)
10. Heaton, J.: Bayesian networks for predictive modeling. Forecast. Futur. **7**, 6–10 (2013)
11. Korb, K.B., Nicholson, A.E.: Bayesian Artificial Intelligence, 2nd edn. CRC Press, Boca Raton (2010)
12. Mashiloane, L.: Educational data mining (EDM) in a South African University: a longitudinal study of factors that affect the academic performance of computer science 1 students. Doctoral Dissertations, University of the Witwatersrand (2016)
13. Nghe, N.T., Janecek, P., Haddawy, P.: A comparative analysis of techniques for predicting academic performance. In: Proceedings of the 37th Annual IEEE Frontiers in Education Conference, FIE 2007 (2007)
14. Osmanbegović, E., Suljić, M.: Data mining approach for predicting student performance. Econ. Rev. **10**(1), 3–12 (2012)
15. Quinlan, J.R.: C4.5: Programs for Machine Learning. Elsevier, Amsterdam (2014)
16. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **40**(6), 601–618 (2010)
17. Shahiri, A.M., Husain, W., Abdul, R.: A review on predicting students' performance using data mining techniques. Procedia Comput. Sci. **72**, 414–422 (2015)
18. Spaull, N.: South Africa's education crisis: the quality of education in South Africa 1994–2011. Technical report, Centre for Development and Enterprise, Johannesburg (2013)
19. University of Cape Town: Faculty of Science Handbook (2017). http://www.students.uct.ac.za/usr/apply/handbooks/2017/SCI_2017.pdf
20. Xing, W., Guo, R., Petakovic, E., Goggins, S.: Participation-based student final performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory. Comput. Hum. Behav. **47**, 168–181 (2015)