# Literature Review

# A Programme Recommendation Engine: Predicting student performance in Degrees


Tasneem Abed (1408535)
University of the Witwatersrand
Supervised by Dr Ritesh Ajoodha

March 31, 2019

## Introduction

This research project will attempt to build a recommendation engine that utilises Matric marks and biographical information of students of Mathematical Science degrees in order to determine the most suitable academic trajectories for new incoming students. It aims to help lower failure rates as well reduce course changes after first year. 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. This literature review will present the background necessary to build a student advisory system for Mathematical Science programs in a South African context for the University of the Witwatersrand. Section 1 will look at issues surrounding performance and admission of students in universities and relate it back to high school performance as well as biographical features and advise given to students. Results about these topics found in research papers in this field will be discussed. Section 2 will look at the mathematics and techniques used to model similar problems and how these will impact the approach taken to build a student advisory system in this research. As contributions to this field, this research will discover trends between high school and university performance of students and propose a solution to alleviate the negative impacts that stems from poor program guidance.

## 1 Educational data mining

The investigations into trends and patterns in various levels of education is a necessary part of building a strong education sector where each student is aware of their options. There is a wealth of discoveries in the field of education data mining and this section will elaborate on some of these and show how they will contribute to optimizing a recommendation engine.

The final 3 years of high school in South Africa (grades 10 - 12), requires learners to take 4 mandatory subjects, namely English (first or second language), Mathematics (pure or literacy), a second approved language (first or second additional language) and Life Orientation. Additionally, a minimum of 3 other subjects must be chosen. These include Geography, Biology, Physical sciences, History and Music amongst others. In the final year of high school, Matric, learners write national examinations that are set by the Department of Basic Education. A combination of the marks of these examinations and the marks earned during the final school year are used to calculate an Admission Point Score (APS), with the highest score for a single subject being 7. Life Orientation is excluded from this calculation [Van der Westhuizen and Barlow-Jones 2015].

Universities in South Africa use the APS of students as a criterion for admissions amongst others. Some programs look at percentages of certain 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 [Rauchas *et al.* 2006]. This is backed up by numerous studies that show a clear link between mathematical aptitude and programming [Byrne and Lyons 2001]. In any Mathematical Science degree, mathematical aptitude would naturally be the primary concern of an admission board as it is directly correlated to the subject matter of the degree content. If a student has poor results in Matric Mathematics, the chances of them succeeding in a tertiary setting is minimal. In fact, according to Campbell and McCabe [1984], 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. Success in a range or subset of high school subjects should be considered .

The medium 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 might 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. English performance in high school is a quantified measure of comfort with the English language. The term 'comfort' is used here to describe the appreciation a student has for precise language use as opposed to the basic understanding of the language at face value. An investigation into the belief that language ability influences success in university [Rauchas *et al.* 2006] found that achievement in high school language courses is a better predictor of success than mathematics. Although this finding was for Computer Science specifically, it is a good indication of the effect language aptitude has on Mathematical Science success as a whole. Of more concern, English as a first language proved to be a much better predictor. Of course, those who are not native English speakers would not be penalised, but this does raise the question of weather these students should be advised on taking supplementary English courses either before or at the commencement of their degree in an effort to curb failure rates. At this point, it is necessary to look at the bigger picture. If a student's home language is not English, could the home province of a student provide any insights into their academic trajectory?

Biographical profiling of students at university is practiced very early into the student-university relationship. Before a student is offered a place, the university already has a wealth of information about them, including their home language, race, gender and home province. These features are considered in an investigation into at-risk undergraduate profiles [Ajoodha and Jadhav 2019]. Gender is a factor that has shown to provide some insight into persistence. Females are less likely to persist in scientific majors, however this is not indicative of academic achievement or potential

[Campbell and McCabe 1984]. Expected value of performance of males in a Mathematical Science degree at a university in South Africa has increased by approximately 7% where females remained constant [Ajoodha and Jadhav 2019]. A distribution of home languages of students in Mathematical Science degrees shows that 38% of students have English as a home language. Although this was the modal home language, it is very low when considering the implications of English ability as discussed previously [Ajoodha and Jadhav 2019]. The home province of a student and their home language are directly linked. It is also difficult to consider home province as there may be less students who come from locations further from the university thus creating data limitations. Race is a sensitive topic, especially in a South African context. Nevertheless, a study on the influence of race on student performance in Mathematical Science degrees showed that 63% of students had an associated Black race description and, alarmingly, 71% of at-risk profiles also had an associated Black race description. These studies have produced significant results which shows that it is worth taking into account biographical features when looking at success rates and not just high school marks. Perhaps an even more worthwhile category of features to use is abstract abilities. Abilities such as comprehension skills, memorization skills, programming skills, math skills and inferential thinking skills are defined as course features or characteristic skills that a student needs to possess in order to succeed in the course [Taha 2012]. No insights are given as to why these features are used instead of high school marks, however the study can be used to compare the effect of using different types of features for the same purpose. Unfortunately, there is no data on abstract abilities like the aforementioned available from the university this research is based on, so high school marks and biographical information will be used.

Academic advisory is a key step 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 advisors and academic staff. With numerous courses to select from, a student may not be able to know their interest in a course solely based on its title [Taha 2012]. All the above mentioned feature engineering* and importance 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.

## 2 Mathematics and techniques used in educational data mining

The mathematics behind discovering patterns and correlations in data between variables is vast and many approaches can be taken. These techniques are necessary to delve into as they will substantiate the approach used in this research project. Statistical analysis is a method used for prediction and correlations. 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 used in the discovery of results related to education as well as machine learning techniques used in building predictive and recommendation engines in the context of education.

Statistical analysis is an approach taken towards the collection and interpretation of data and has been vastly used for many years. In a study done in 1984 that sought to predict the success of first years in a Computer Science major, a statistical analysis approach was taken to determine if there was a significant difference between students of Computer Science, Engineering, Other Sciences and Others and any entrance variables, and to determine which combinations of entrance variables could predict the group of a student [Campbell and McCabe 1984]. Common statisti-

cal measures such as the mean, variance and standard variations are calculated for each group and entrance variables. The Chi-square statistic is used to compare the percentages of males in each group where a large Chi-square value represents noticeable differences. F-statistics indicate noticeable differences among means where p-values assume equality among means and represents the chance of observing data as extreme or more extreme than that observed in the sample. One sided t-tests were used to compare the Science group versus the Other group to see if the Science group had significantly higher mean values. Statistical analysis was not sufficient to justify that the difference found between the groups were useful or important, so discriminant analysis was done instead. Discriminant analysis is a statistical classification technique that is used to assign a student a group. Wilks' lambda discriminates between 2 groups on the basis of multivariate t-test and gave a 68.4% accuracy which proved to be the best classification method. A Jackkniffing procedure was then applied to verify the results. This involved removing a student and performing discriminant analysis on the remaining data then adding the student and letting the model classify them then repeating this for each student. In the paper that focused on language ability as mentioned above, statistical analysis was again the approach taken, this time in the form of Pearson's product-moment correlation. This statistic measured the strength of association that existed between 2 variables [Rauchas *et al.* 2006]. In this case, 4 Matric marks of which 3 were languages and the remainder was Mathematics. Small values (i.e closer to 0) showed weak positive correlation between Mathematics and Computer Science as well as between first languages and Computer Science but stronger correlations were found between English first language and Computer Science. No prediction techniques were needed for this study.

Mathematical approaches taken to tackle the problem of predicting the most suitable program for a student in order to recommend it to them revolve around both supervised and unsupervised machine learning algorithms. In a study from Cairo, Egypt, the K-means clustering algorithm is applied to divide student records into clusters based on similarities in marks [Aly *et al.* 2013]. This algorithm uses the distance between data points and predefined centroids to find the closest centroid to a query point and classifies it according to the centroids class. The K-means algorithm will converge to a local optimum, however it is not guaranteed that an optimum clustering will be obtained [Aly *et al.* 2013]. After the cluster to which a student belongs has been identified, various decision tree algorithms are applied in order to recommend a department with the highest success rate for the student. These algorithms include the ID3, C4.5 and CART algorithms. Of these, C4.5 proved to be the most efficient and robust. In a similar fashion, a study from Abu Dhabi also clustered students into biclusters based on similar academic skills and interests, however the xMotif algorithm was used instead of K-means as it is designed for biclustering specifically [Taha 2012]. No accuracy of the algorithm was given however it is evident that it was sufficient as the closeness between the lists ranked by the recommendation system increased consistently as the number of students increase, so it is worth looking into. A successful recommendation tool known as Collaborative Filtering is used to filter for information using the opinion of other people.

Another approach to modelling education data to find patterns in high school marks and biographical information with regards to university success is Bayesian networks. A Bayesian network representation is a directed, acyclic graph where each node represents a random variable and the edges correspond to the influence of one node on another [Koller *et al.* 2009]. This approach is able to model causal relationships i.e. the occurrence of an event causes or has a strong influence on the other. In the context of education, Matric marks and biographical factors are events that can be considered as causes or influences on the success a student has in their degree. This was the primary rationale for a recent study conducted at a South African university. Bayesian estimates

were calculated for students who failed with their profiles i.e. the probability of a biographical profile given that the student did not meet the minimum requirements to obtain their degree [Ajoodha and Jadhav 2019]. This approach allows for individual features to be grouped into one feature making it easy and efficient to see and compare results and to tweak calculations on as granular a level as a value of a single feature. Another advantage of this approach is that prior probabilities are accounted for.

As shown, the mathematical approaches used in this field vary considerably and no one approach is proven to be the best. Although clustering is not a concern in this research, the success of decision tree algorithms such as the C4.5 algorithm for predicting successful programs makes it an algorithm to consider for finding the most suitable program to recommend to a student. A Bayesian network approach also provides a robust framework for dealing with all the features that will have different types of values. It is a probabilistic approach and will model conditional probabilities with ease which, in this context, is extremely useful and informative.

## Conclusion

The literature in this field has proven to produce informative and necessary observations that will influence the methodology in this study. For example, the weight placed on certain features such as gender and English ability will vary from others due to the results found in the papers mentioned above. The success of mathematical approaches taken such as the C4.5 algorithm and Bayesian networks provides both reasoning and choice which gives this research flexibility. The South African context of this research makes for an interesting viewpoint in the global scheme of this field as certain observations may solely be linked to the history of this country. Thus, this research will also demonstrate how to deal with sensitive issues regarding biographical data.

## References


[Ajoodha and Jadhav 2019] Ritesh Ajoodha and Ashwini Jadhav. Identifying at-risk students using biographical and enrollment observations for mathematical science degrees at a south african university. *Personal communication*, 2019.

[Aly *et al.* 2013] Walid Mohamed Aly, Osama Fathy Hegazy, and Heba Mohmmed Nagy Rashad. Automated student advisory using machine learning. *International Journal of Computer Applications*, 975:8887, 2013.

[Byrne and Lyons 2001] Pat Byrne and Gerry Lyons. The effect of student attributes on success in programming. *ACM SIGCSE Bulletin*, 33(3):49–52, 2001.

[Campbell and McCabe 1984] Patricia F Campbell and George P McCabe. Predicting the success of freshmen in a computer science major. *Communications of the ACM*, 27(11):1108–1113, 1984.

[Koller *et al.* 2009] Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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

[Taha 2012] Kamal Taha. Automatic academic advisor. pages 262–268, 2012.

[Van der Westhuizen and Barlow-Jones 2015] Duan Van der Westhuizen and Glenda Barlow-Jones. High school mathematics marks as an admission criterion for entry into programming courses at a south african university. *The Independent Journal of Teaching and Learning*, 10(1):37–50, 2015.