Annotated Bibliography

A Programme recommendation engine: Predicting student performance in Degrees

Tasneem Abed (1408535)
University of the Witwatersrand

March 6, 2019

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.

Aim: To identify at-risk students biographical profiles by contextualizing the student by use of their biographical profiles and enrollment information.

Style/Type: Private communication

Cross references: The approach taken in this paper puts to use the Bayesian network techniques and concepts as proposed by Koller *et al* [2009], specifically Bayesian networks. The paper also places an importance on home languages spoken by students as an indicator for performance. This agrees with Rauches *et al* [2006] where language performance as an indicator of success was investigated.

Summary: This paper explores the relationship between biographical characteristics of a student with respect to their university performance. Firstly, biographical characteristics that were either successful or unsuccessful in obtaining a degree were explored. These included characteristics such as gender, home language, home province and race. Secondly, the correlations between the student's performance and these characteristics were presented. Lastly, the posterior probability of a biographical profile failing to complete a degree was calculated using Bayesian analysis. A ranking of the outcomes of these probabilities was provided in the form of a table. Most notable observations were then summarized. For example, the home language with the highest Bayesian estimate of not completing the minimum requirements to obtain a degree in a

Mathematical Sciences field, the gender with the highest Bayesian estimate for the same etc. Joint probabilities of at-risk biographical profiles were also discussed, such as Black and English speaking students or Black students whose home province is the Free state. The report also gives an appendix with numerous visuals that describe some results that the study has found.

[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.

Aim: To use educational data mining to build an automated student advisory framework to help guide students to a more suitable education track.

Style/Type: Journal article

Cross references: This paper and Taha [2012] use different algorithms to perform clustering.

Summary: The paper begins by defining what educational data mining is and how it is applied. The advisory framework proposed aims to improve the students performance and quality of education. One of the main reasons for high failure rates is the incorrect choice of the student's department. The system will predict the department the student is most likely to succeed in. The paper goes on to talk about the machine learning techniques that were adopted. These include the C4.5, CART and ID3 decision tree algorithms which were tested for predicting the department a student is most likely to choose. C4.5 proved to be most accurate with an accuracy of 98%. The K-means cluster algorithm was then employed to divide the students into a number of clusters and then determined the rate of success in each cluster. The entire framework was then tested on a case study from an institute in Cairo.

[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.

Aim: To identify the factors that influence success in the first year of a Computer Science major, not just success in a single programming course.

Style/Type: Journal article

Cross references: This paper finds that higher SAT verbal scores played a role in distinguishing Science and Engineering students from those in other fields. This notion agrees with Rauches *et al* [2006] which finds language ability to be an important indicator of university success. However, it also challenges the paper in that it finds that females have lower persistence rates yet have generally higher high school English marks. This finding also correlates with Ajoodha *et al* [2019] in that sex was a indicative variable and in the result itself.

Summary: This research was conducted in 1984. The paper identifies the need for success prediction in universities due to the increasing number of enrolments. It first manages to find that a useful indicator to weather a student is successful in completing their degree or not is if they managed to complete their freshman year. The paper tries to see if students who left Computer Science(CS) to go to

an Engineering or other science field are distinguishable from those who left to go to any other non-scientific field. A figure showed that an equal percentage of students remained in CS after first year and switched to a non-scientific field. Some results found were that those who persisted in CS or switched to an Engineering or scientific field had significantly higher SAT math and verbal scores. Also, men were more likely to persist than women. However, men had, on average, lower high school grades in math and English. In terms of techniques, statistical and discriminant analysis was used as well as binary classification and the jackknifing procedure. The highest classification accuracy obtained was 68.4%, however, high error rates were obtained for students who were classified as Other when they were actually CS. The paper concluded that students who persisted in CS, Engineering or any science field differed from those in other fields by their SAT math and verbal scores, high school ranking and background in high school maths and science.

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

Aim: This extract of the textbook aims to introduce the reader to the Bayesian Network representation and the mathematical and logical concepts related to it.

Style/Type: Expository, book chapter.

Cross references: Bayesian networks as a model for determining causal probabilistic relationships among random variables is the adopted strategy by Ajoodha *et al* [2019] due to its suitability in the research topic.

Summary: The extract starts out by exploring independence properties of a probability distribution and how it is used to compactly represent high dimensional distributions. The Naïve Bayes model is described in detail from the independence assumptions it makes to the applications of it and its drawbacks. Factorized joint distributions as a product of conditional probability distributions(CPDs) are described with examples. Bayesian networks are not restricted to representing distributions that satisfy the Naïve Bayes independence assumption. They are directed acyclic graphs whose nodes represent the random variables of the domain and the edges represent direct influence on a node by another. To illustrate all the concepts, one example of a student Bayesian network was continuously used. This uses random variables of course difficulty, intelligence of a student, grade, SAT scores and a recommendation letter. Local probabilities represent the nature of dependence of each variable to its parents. A fundamental concept that is defined towards the end of the extract is that of the Chain Rule for probabilities in Bayesian networks.

[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.

Aim: The research paper aims to investigate the theory that language ability and performance at a high school level influences success in Computer Science at a first year university level.

Style/Type: Conference proceeding

Cross references: In Ajoodha *et al* [2019], the correlation between home language and completion of a degree is looked at and it was found that English speaking students had the highest rate of success. This result correlates with the results of this paper in that English ability plays a significant role in success at a university. Contrary to Campbell *et al* [1984] which states that using a single high school subject grade is insufficient in determining success rates, this paper focuses on high school English marks.

Summary: The report starts by introducing the notion that although high school performance in mathematics is the main indicator used to determine acceptance into Computer Science at a university level, it might not be the most important factor. It suggests that language ability and comfort (specifically in English a it is the medium used in the university) might provide more insight into the expected success in a Computer Science degree. Of the 4 modules taken in first year, only one has more English oriented tasks as opposed to mathematical tasks. Based on this it is easy to assume that those who lack in English ability would struggle in the module that involves essay writing, however, the modules proved to be indistinguishable. Factors that influenced the research was that performance in Computer Science needed to be distinguished from more general computing courses, most students are not native English speakers and that high school performance is not the only factor influencing university success. Pearson's correlation figures were obtained for each of the relationships between either BCO or FAC (university modules) and English FL, English SL, mathematics and other FL languages in school. It was found that the language competence is at least as good a predictor as Mathematics. However, English FL proves to be a strong indicator. The paper goes on to discuss the impact of this finding and also defines what it means to be 'comfortable' with the English language.

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

Aim: To introduce a collaborative filtering system called Automatic Academic Advisory(AAA) to overcome the problem of poor course guidance for students in distance education.

Style/Type: Conference prceeding

Cross references: Different to all other paper, especially Mohamed *et al* [2013], this paper looks at more abstract variables as opposed to high school marks, such as inferential thinking skills and memorization skills. This would agree with Campbell *et al* [1984] that high school marks are not the only factors to consider when trying to calculate success rates.

Summary: Distance Education is a newly popular education tool for learning off site of an institution, primarily over the internet. The paper focuses its attention on overcoming the poor advise given to distance education students about course choices with regards to their skills and interests. Students are categorized based on their similarity of course features which will then be used to calculate success rates in order to recommend a list of ranked courses to a new student based on which cluster they fall into. An assumption that those who

have similar academic performance and interests on prior courses tend to have the same interests for future courses. The xMotif algorithm was applied for biclustering and a scoring system was used to identify course features per bicluster which involved complex similarity equations. The system is re-optimized with every new student user. It was evaluated by an experiment done on university students in Texas and the UAE. Results showed there were only small differences between the courses ranked by students based on prior academic performance and the lists ranked by AAA.