EXPLORING DIFFERENCES IN PREDICTORS OF ACADEMIC SUCCESS BETWEEN DIFFERENT GENERATIONS OF STUDENTS

Dijana Oreški, Goran Hajdin

University of Zagreb, Faculty of Organization and Informatics (CROATIA)

Abstract

Higher education institutions aim to provide quality education to students. One way to achieve this is by discovering knowledge to predict student’ performance. The knowledge is hidden among the data set and is extractable through the Knowledge Discovery in Data process. This present paper is designed to evaluate the data mining process standard CRISP DM for the purpose of predicting an IT students’ performance. In this paper, we construct a data mining model that tries to predict a students’ academic success. Our data set consists of 405 students from three generations where we gathered information for 16 variables. A specific educational setting is used – university undergraduates and graduates in computer science. This study tries to answer the following research question: What are the differences in predictors of academic success between different generations of students? The main question is further analyzed through two sub-questions: What variables are the best predictors of success? Do student success predictors vary over time? The research results indicated similarities between two generations (B and C) and differences between the “pair” (B and C) and generation A. Among 16 analysed factors three factors had similar results for all three generations, eight factors were similar for the generation B and C, but were different for the generation A. Those eight were: lecture attendance, time management and finding time to learn, admission exam score, conscientiousness, personal learning space, which high school was previously finished, motivation source, and perception of working in teams. Of the 16 factors three were different for the “pair” (B and C): personal responsibility, seminar attendance and gender. Two factors, the GPA (grade point average) in high school and current study year had inconclusive results.

Keywords: academic success, success prediction, higher education, data mining, CRISP DM, neural network.

1 INTRODUCTION

Improvement of an educational system and achievement of conditions for students’ optimal learning requires collecting of performance related data [1]. Educational data mining serves as valuable for data analysis and pattern recognition. There are several recent research papers dealing with this topic from the perspectives of: (i) different academic and non-academic factors included in dataset, (ii) research methodology used in data analysis, which was mostly focused on advanced statistical approaches such as regression analysis, with data mining and machine learning approaches increasingly leaning in recent years, (iii) accuracy of gained predictive models. From methodological point of view various authors emphasized advantages of data mining application in educational domain and argues that predictive accuracy highly depends on the list of variables included, students population and methodology used (e.g. [2], [3]). Data mining approach was also used in the most recent studies of academic performance by [4], [5], [6]. Their results are motivation for data mining application in this paper.

Improvement of an educational system and students’ academic success has been researched as a topic for over two decades. Researchers have analysed different aspects of academic success and possible predictors which might help determine the outcome [7].

Gallagher, Bomba and Crane [8] studied the use of admission exams for prediction of student success. In their study they did not find significant difference between successful and unsuccessful students. Worth noting is that their study was focused on students enrolled in Associates Degree in Nursing programmes which have rather specific requirements.

Ricketts and Rudd [9] studied the relationship between academic performance, leadership and critical thinking. They concluded that GPA, gender, age and innovativeness were the best variables for explanation of critical thinking. Their results are further used by D’Alessio, Avolio and Charles [10] who further researched impact of critical thinking on the academic performance.
According to Cano [11] motivation and time management are significant predictors of students' academic performance. They are part of the affective strategies (along with concentration and attitude), which along with goal strategies (consisted of test strategies and selecting main ideas) are statistically significant predictors of students’ academic performance. On the other hand, research concludes that cognitive monitoring strategies (which consist of information processing, self-testing and study aids) are not a significant factor.

Dollinger, Matyja and Huber [12] concluded in their research that participation and activity during lessons is a significant predictor of students' performance. Their results show that teaching in class does contribute to learning outcomes and that students do benefit from their activity during lessons. They found that there is a positive correlation between attendance, motivation and gender.

Several studies point out conscientiousness as one of the strongest predictors of academic performance. Kappe and van der Flier [13] concluded that conscientiousness correlates with job related performance criteria and that intrinsic motivation is related to GPA and time-to-graduation. Further they conclude that intrinsic motivation and conscientiousness are strongly related. Also, one of the measures of academic achievement was teamwork as part of the team projects which were conducted by students. Kertechian [14] found that students who had higher GPA had also higher level of conscientiousness. Students who were the most motivated and hardworking were the ones with highest probability of passing the exams.

This paper is organized as follows: section 2 presents related work on predictive modelling of academic success, emphasizing the lack of machine learning approaches in such research; then it defines the purpose of our study and research question. Section 3 further describes CRISP DM methodology used in the paper. Section 4 presents the results of this study with regard to the research question. Finally, conclusions and future work are presented.

2 RESEARCH METHODOLOGY

In this study, an analysis was carried out to identify predictors of academic success of information technology (IT) students and to determine the extent of change in predictors among three different generations of students. As seen in the literature review, data mining procedures show promising results, thus, CRISP DM standard for data mining is explained and applied here with a previous description of the data to be used.

2.1 Purpose of the research

The main purpose of this study is to explore predictive power of students' characteristics and students' attitudes on the academic success of the students and identify do those characteristics change over time. Following research question was set up: What are the differences in predictors of academic success between different generations of students? Which was further divided into two sub-questions:

RQ1: Do student success predictors vary over time?
RQ2: What variables are the best predictors of success?

2.2 Data mining process and neural networks

Data analysis was performed by means of Cross Industry Standard Process for Data Mining (CRISP DM) standard process for data mining which consists of six steps: domain understanding, data understanding, data preparation, modelling, evaluation and deployment. With the aim of systematic data analysis, a standard process of CRISP DM is followed here, which was previously used in educational data mining (e.g. [2], [3]). I) First step of CRISP DM is focused on understanding the goals and requirements of domain problem. In this paper we want to investigate academic performance of IT students, explore determinants of student success and identify are there differences in students' behavior and attitudes among generations. II) Second step, data understanding, begins with data collection. In this phase we conducted an initial inspection of data set which is composed of 16 variables related to each student. Descriptive analysis generated distributions of each variable and the results are described in Table 1. III) Data preparation step covers all activities to develop the dataset for modeling. IV) Modelling step is in the heart of each data mining process. Various statistical
and machine learning approaches are developed for modelling ranging from regression to decision trees, Bayesian approach and neural networks. In this paper we applied neural network because of their “ability to handle nonlinear relationships between the output and input variables” [15]. Thus, in modeling phase, various neural network algorithms were selected and applied, and their parameters calibrated to optimal values. The most important parameters are number of hidden layers and number of nodes in the hidden layer. Artificial neurons in artificial nets are connected together into architecture through layers. Each network consists of one input layer, one or more hidden layers and one output layer. Number of hidden nodes is important because hidden layer is used to process data. Hereinafter, number of hidden layers is set to one and number of nodes in hidden layer is set as mean value of input and output nodes (9 nodes). When neural network model is achieved, sensitivity analysis is performed in order to determine how “sensitive” a model is to changes in the value of the parameters of the model [16]. Sensitivity analysis is performed by setting different parameter values to see how a change in the parameter causes a change in the behavior of the output. Last two steps of CRISP DM process, V) evaluation and VI) deployment try to evaluate if the model satisfies goals defined in the first step and decide about implementation of the data mining results into system. RSquare is used here as metric for model reliability and RMSE as metric for model accuracy.

3 RESULTS AND DISCUSSION

3.1 Data understanding

The presented study has been carried out among 405 students (206 female and 199 male students) studying at University of Zagreb, Faculty of Organization and Informatics. In this study students of three different generations (called generation A (n = 113), generation B (n = 92), generation C (n = 200)) filled out questionnaire to collect data. Generation A was consisted of students who enrolled their study programmes before implementation of the Bologna reform, while generations B and C consisted of students who enrolled their study programmes after implementation of the Bologna reform. Questionnaire was previously tested [17] and consisted of 16 variables. Using three samples, we tested whether differences among generations exists and which variables explain differences. Distributions for all variables are reported in Table 1. The majority of the participants are from generation C. Most of them graduated from gymnasium, followed by technical and economic high school. As to their high school GPA distribution, most were very good students. Most of the students were in the middle at the results of admission exam, whereas the remaining were at the bottom of the list and between 10% and 30%, while only a few were in the top 10%. The rest of the Table 1 provides information about the participants’ lecture attendance.

<table>
<thead>
<tr>
<th>Current study year</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>17</td>
</tr>
<tr>
<td>II</td>
<td>73</td>
</tr>
<tr>
<td>III</td>
<td>96</td>
</tr>
<tr>
<td>IV</td>
<td>157</td>
</tr>
<tr>
<td>V</td>
<td>62</td>
</tr>
</tbody>
</table>

Table 1. Variables descriptions and distributions
Which school did you finish

- Economic
- Gymnasium
- Technical
- Other

What was your GPA (grade point average) in high school?

- 1
- 6
- 43
- 217
- 138

What was the first grade you got at the faculty?

- At the bottom
- Between first 10% and 30%
- In the middle
- Top 10%

- 119
- 98
- 126
- 62
- 142
- 114
- 61
- 9

Lecture attendance

- I did, almost all
- I used permissible absences
- Most of the time I did not

- 50
- 218
- 137

Seminar attendance

- I did, almost all
- I used permissible absences
- Most of the time I did not

- 59
- 154
- 189

I manage time well and can find time to learn.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

- 9
- 49
- 126
- 178
- 43

I think I am conscientious.

- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

- 1
- 10
- 58
- 236
- 99
3.2 Predictive model of academic success

Machine learning approach of neural network methodology was used for data analysis in order to answer research questions. Furthermore, sensitivity analysis of the model was performed to identify the most important predictors of success in each of three generations of students. Results are transformed into ranks presented at Table 2. As a result of the neural network modelling, it has been found that GPA in high school is most important predictor of the students’ academic performance in generation A and B, whereas Admission exam score is most important predictor of academic success in generation C. Admission exam score shown to be second most important predictor in generation B. First grade at the Faculty was ranked third predictor in all three generations of students.
The following section provides answer to the first research question. According to the results in the Table 2 all three generations have similar ranking for the following success predictors: first grade you got at the faculty (3), preparation for classes and activity during the lessons (12) and the faculty was at the top of the list of priorities when choosing one (14). Generations B and C have the same or similar ranking for 8 success predictors: lecture attendance (2), time management and finding time to learn (4), admission exam score (5), conscientiousness (6), personal learning space (10), which high school was previously finished (11), motivation source (13), perception of working in teams (15). Generations B and C have different rankings for three success predictors: personal responsibility (8), seminar attendance (9), gender (16), and results are mixed for two success predictors: GPA in high school (1) and current study year (7). One of the results of the analysis points out similarities between two generations (B and C) which were consisted from students who enrolled their study programmes after implementation of the Bologna reform. Based on such results we can conclude that Bologna reform did have a significant impact on the education system. With respect to our research question, to provide a more complete answer about students’ differences among generations, we analyzed the correlations between the three variables: ranks of predictors in generation A, ranks of predictors in generation B and ranks of predictors in generation C. Our results have shown strong positive relationship between predictors of academic success in generations B and C ($r=0.652941176$, $p<0.01$). Correlation in predictors between generations A and B ($r=0.370588235$, $p<0.01$) also between A and C ($r=0.388235294$, $p<0.01$) is lower. Based on the results we can conclude that student success predictors do vary over time but only when there are significant changes within the education system. When comparing results from two post-Bologna reform generations their results are much more similar.

Top three success predictors, based on the success predictor average ranking for the three generations, are GPA in high school (1) with the average of 2.00, admission exam score (5) with the average of 2.67 and first grade you got at the faculty (3) with the average of 3.00. This answers the second research question. It is interesting to note that the top three success predictors are all focused on grades. The results are in line with the previous research from Rountree, Rountree and Robins [18] who point out that grades which students expected to achieve at the beginning of a course were strong indicators of success.
4 CONCLUSIONS

Conducted research was based on 405 students from three different generations (A, B and C) of which two (B and C) were post-Bologna reform generations. Results show similarities between the two post-Bologna reform generations and greater differences between those two generations and pre-Bologna reform one. Further analysis showed that student success predictors do vary over time (RQ1) but only when there is a significant change in the educational system. Among the 16 success predictor variables three that stand out as the best predictors of success (RQ2) are all based on grades: GPA in high school, admission exam score and first grade student got at the faculty.

In anticipation of future research, one should be aware of limitations of our work. Our sample constituted only by students of informatics, so, one must be careful in interpretation of the results. In the future research authors should replicate research among different student populations to generalize results. As from methodology point of view, only one machine learning approach is applied here. In the future research authors will compare different approaches to develop predictive models of higher accuracy and reliability. Furthermore, descriptive models should be included in the analysis to describe characteristics of different groups. Feature selection and feature reduction techniques will be employed in order to select only most relevant variables in modelling phase.

REFERENCES


