DATA VISUALIZATION AND EXPLORATION OF STUDENTS DATABASE

Miroslava Jordovic Pavlovic[1], MSc; Dragana Knezevic[1], MSc; Slobodan Petrovic[1], Mr
[1]Business and Technical College of Applied Sciences, Uzice, SERBIA,
miroslava.jordovic-pavlovic@vpts.edu.rs, dknezevic28@gmail.com, slobodan.petrovic@vpts.edu.rs

Abstract: *The main goal of this paper is the application of efficient analytic tools for educational data processing, which will lead to improving the quality of education and enhancing school resource management. Data from the students' database of the Business and Technical College of Applied Sciences Uzice were used for the research. Although student achievement is highly influenced by past evaluations, an explanatory analysis has shown that there is a need for more relevant features in order to develop interesting automated tools that can aid education domain.*

Keywords: *Learning analytics, educational data processing, Cytoscape, Octave.*

1. INTRODUCTION

We are living in the time of Big Data. The amount of data that any organization is dealing with is massive. School possesses considerable amount of structured data like grades, enrollment data, progression rates as well as unstructured data like students' opinions expressed through surveys, web blogs, etc.Those are very powerful resources, containing information about learners, their learning and learning environment. These databases contain key answers for: improvement of students' performance, identification of different factors which effects students' learning behavior, drop-out rates, etc. The institution can improve decision making and optimize success due to valuable information hidden in the database, such as trends and patterns.
Learning analytics (LA) is a field of data analytics, whose main goals are optimizing teaching and learning. Learning analytics refers to the measurement, collection, analysis and reporting of data about the progress of learners and the contexts in which learning takes place. Using the increased availability of big datasets around learner activity and digital footprints left by student activity in learning environments, learning analytics take us further than data currently available can [1]. Campbell, DeBlouis, and Oblinger explain that "Analytics marries large data sets, statistical techniques, and predictive modeling" and say that analytics "could be thought of as the practice of mining institutional data to produce 'actionable intelligence'" [2]. Learning analytics could make significant contributions in the following areas: as a tool for quality assurance and quality improvement, as a tool for boosting retention rates, as a tool for assessing and acting upon differential outcomes among the student population, as an enabler for the development and introduction of adaptive learning [1]. A variety of tools and approaches is used in Learning analytics to provide educators with quantitative intelligence to make informed decisions about students' learning.
Student achievement is highly influenced by past evaluations, personal, social, psychological, financial and other environmental variables. What are the factors that mostly affect student achievement? Is it possible to identify at-risk students? Is it possible to predict student performance? This paper will focus on these questions with the aim to emphasizes the importance of getting started with adaptive learning in college courses.

2. LITERATURE REVIEW

G.Dimic at al. in their paper described the application of classification methods (Decision Trees and Naive Bayes)for prediction of student's success on the last exam. Data from the on-line course created on the Moodle LMS system were used for research. Standard measurements of model accuracy were considered (the number of correctly and incorrectly classified samples, TP, FP, Precision, Response, F-measure, ROC curve area). The same authors discussed the

importance of good choice of classifiers for an extremely small set of training data in a paper [4]. They found that the Naive Bayes and J48 algorithms generate precise classification models with a prediction accuracy greater than 70%.
In paper [5] Bayesian classification method is used on student database to predict the students division on the basis of previous year database, as well as to identify those students who need special attention in order to reduce failing ratioand to take appropriate actions at right time.
Paper [6] describes three predictive models obtained from the student data set by three machine learning algorithms: the C4.5 decision tree algorithm, the CART algorithm, and ID3 decision tree algorithm. Decision trees model is successfully identifying the students who are likely to fail. These students can be considered for proper counseling so as to improve their result.
The aim of case study Identifying at-risk students at New York Institute of Technology [1] was to increase retention of students in the first year of their studies by creating an at-risk model to identify students most in need of support and to provide information about each student's situation that would assist support counsellors in their work. Recall of the model is 74%; in other words, approximately three out of every four students who do not return to their studies the following year had been predicted as at-risk by the model. This high recall factor is due to the choice of model as well as the inclusion of a wider range of data than other similar models. Financial and student survey data were included in the model as well as pre-enrolment data.
A. Desai at al. [7] performed k-means clustering on the students data set. The students of a particular class are classified as Cluster 1 (Below Average), Cluster 2 (Average) and Cluster 3 (Above Average) based on their academic performances in order to enhance teaching-learning process. The algorithm has ability to change the number of clusters and to use various of non-academic parameters. A web-based tool was developed in order to enable students to fill in the tests conveniently and for the teachers to view the student profiles and generate clusters based on their requirement. The result is the teachers have a better insight of the student's performance and accordingly can conduct remedial classes or advance test programs for different groups of the students to improve their academic scores.

3. METHODOLOGY

The students' database of the Business and Technical College of Applied Sciences Uzice (BTC Uzice) contains the following data: the school year of enrollment for each semester of studies, the success of the student from the secondary school (the maximum number of points is 40), the success on the preliminary exam - the number of points at the math test plus the number of points at the general knowledge test (the maximum number of points is 60), educational profile, type, duration and location of secondary school. In this paper, only a database of students from the Information Technology study program was analyzed. Data were collected in the period 2007-2013. The total number of analyzed records was 453. We used Cytoscape and Octave. In the data preprocessing phase several activities were undertaken: cleaning irrelevant data, selecting variablesand transforming the original data into a form that is more convenient for the application of the selected processing tools.

4. RESULTS AND DISCUSSION

One of the important issues to deal with is the question of which type of high schools are completed by students enroll ed in study program Information Technology at BTC. Regard to this, two categories of secondary schools, a four-year and three-year secondary school are first to be expressed. Four-year high schools of different profiles in relation to the three-year schools have a much larger part. Four-year high schools completed 79.8% (or 341 students) and a three-year 20.1% (or 86 students) of the observed sample. Figure 1 shows the graphic representation of relationship of these type of schools with concrete profiles completed by students. It can be seen from the picture that Information Technology is enrolled by students mostly after technical high school, whether it's profiles lasting three or four years.

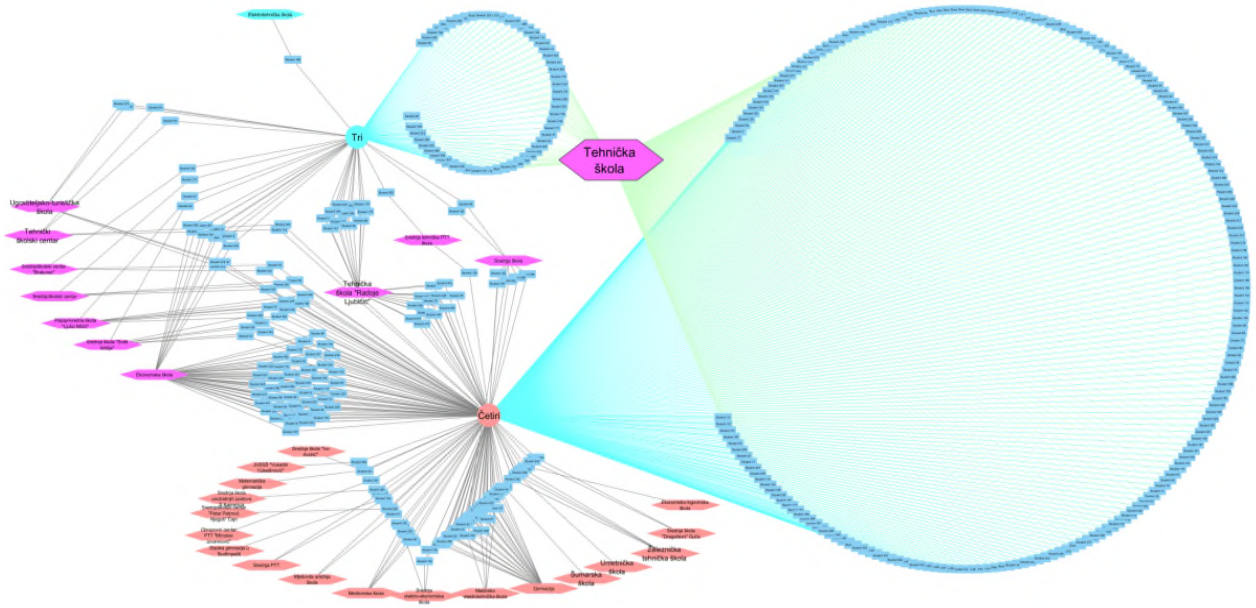

Figure 1: Three-year and four-year high schools finished by students BTC

Based on the same problem (whether students complete three-year or four-year high schools), we can make a comparison per enrollment year in one or individual diagrams. The consolidated diagram of all enrollment years (2007 / 08-2012 / 13) is given in the following figure (Figure 2).

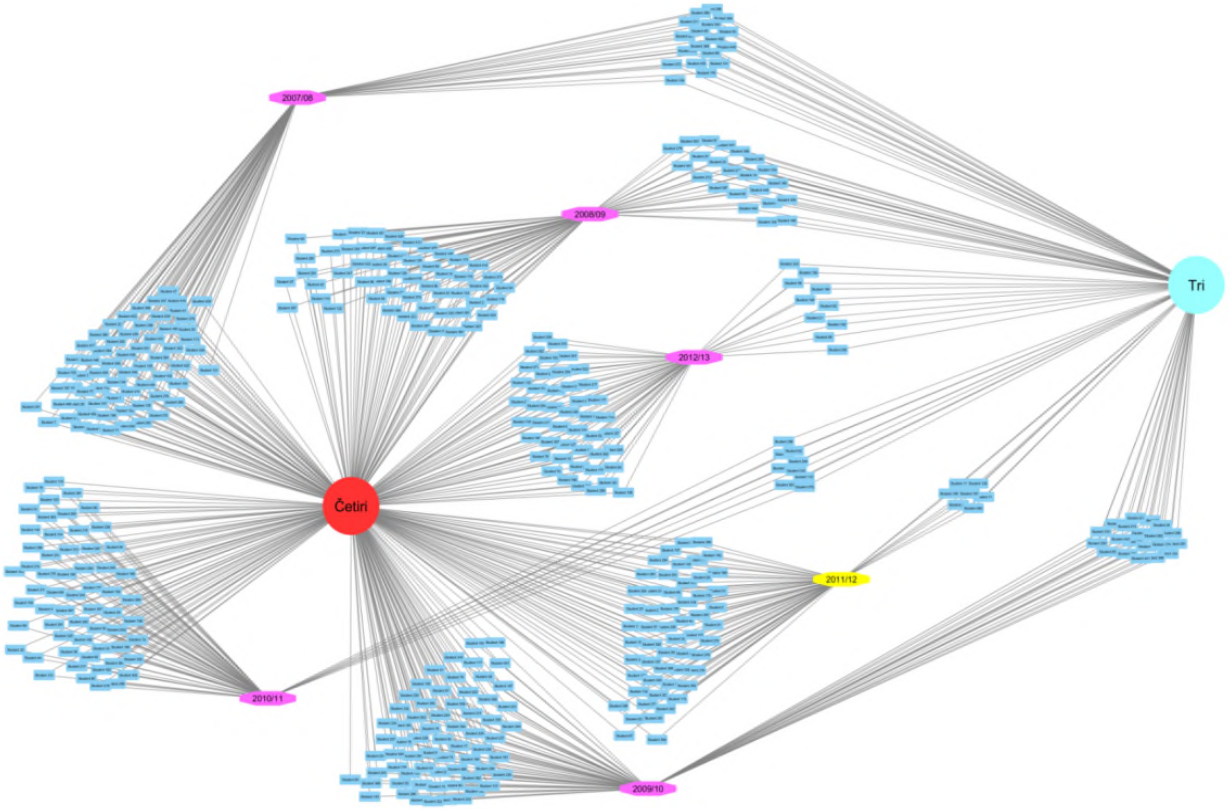

Figure 2: The overview three-year and four-year high schools finished by students BTC per enrollment year

It can be seen from Figure 1 thatin school years 2010/11 and 2011/12 was the smallest number of those who completed three-years high school (only 9 or 7 students), so in these generations the highest level of student success can be expected during the education at BTC. We could individually display such results as shown at following pictures.

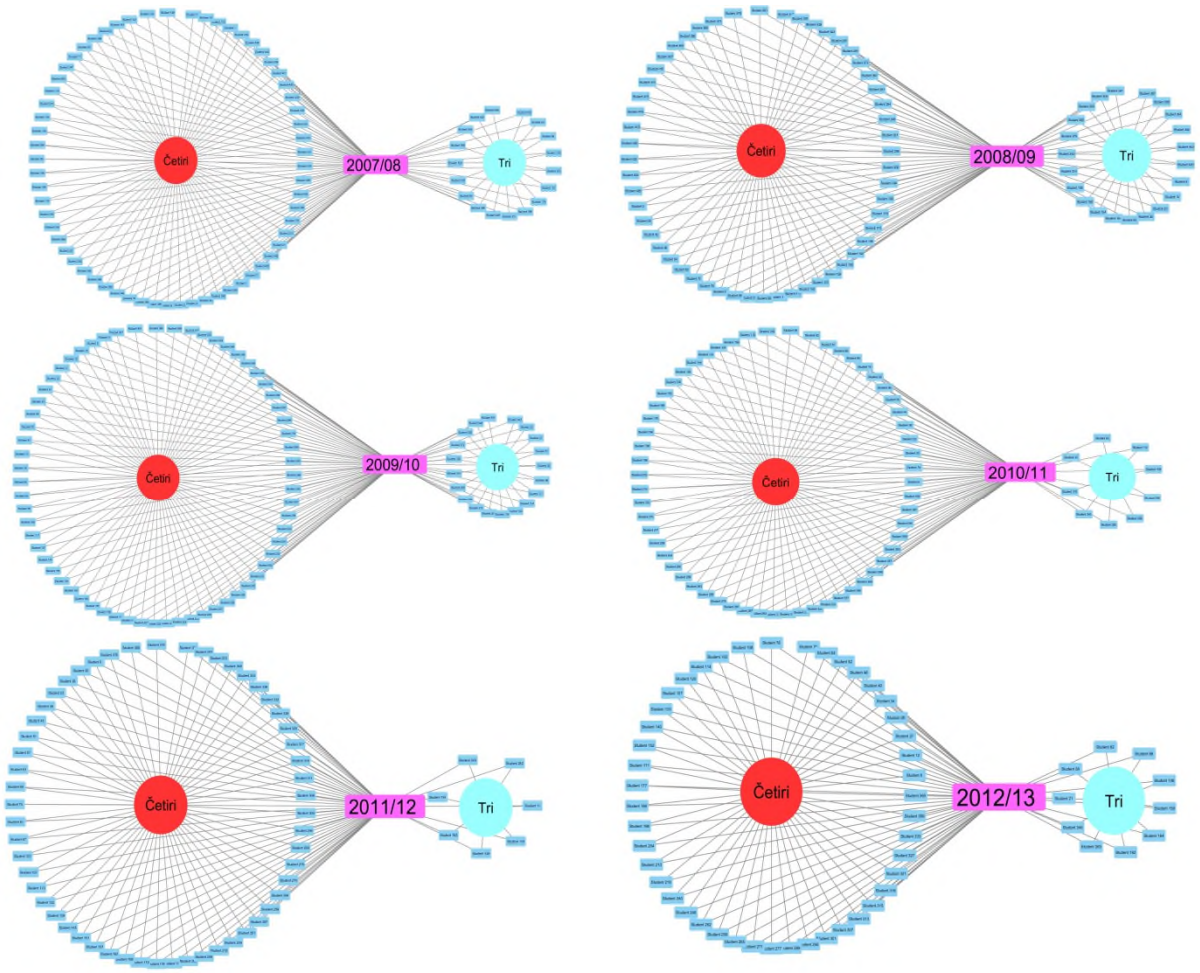

Figure 3: The overview of three-year and four-year high schools completed by students who enrolled BTC in period 2007-2012

In terms of students' passing through the year, we can separately observe each of the enrollment years and then compare the performance of students for each generation. Figure 3 showsdistribution of students enrolled from 2007/08, 2008/09 ... until 2012/13 school year. Symbols for students who have finished studying without the renewal of academic year are circularly placed, colored in dark green color, that is, students have been regularly enrolled in all six semesters. Also, with the same color but out of the circle are presented students who completed their studies but have renewal one academic year. Rest of the students dropped out of study at some time and have not continued until now. The number of such students varies year by year.

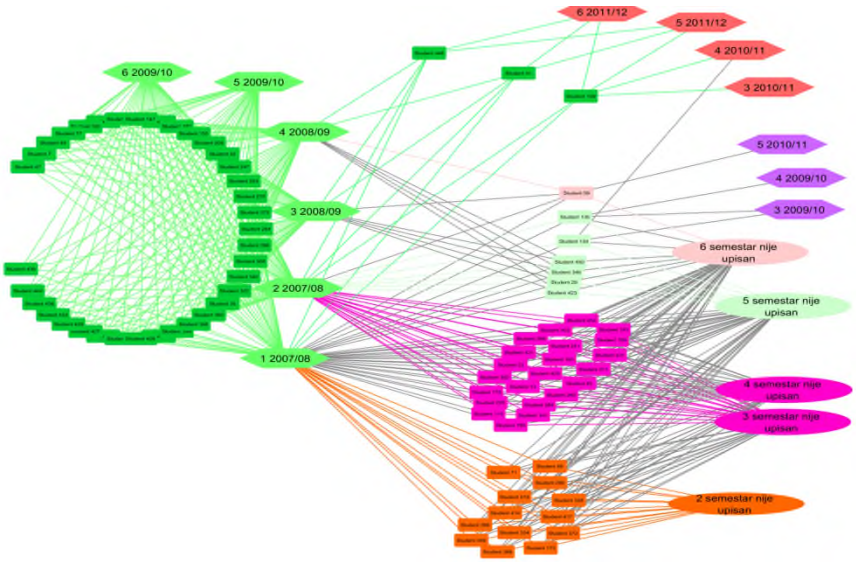

Figure 4: Performance of students enrolled in the first semester 2007/08

Generation enrolled in 2007/08, from total of 81 students, only 36 students or 44%, finished regularly plus three students who completed their studies with renewal of one academic year. Other students, even 42, gave up at some time, which makes up more than half of enrolled students.

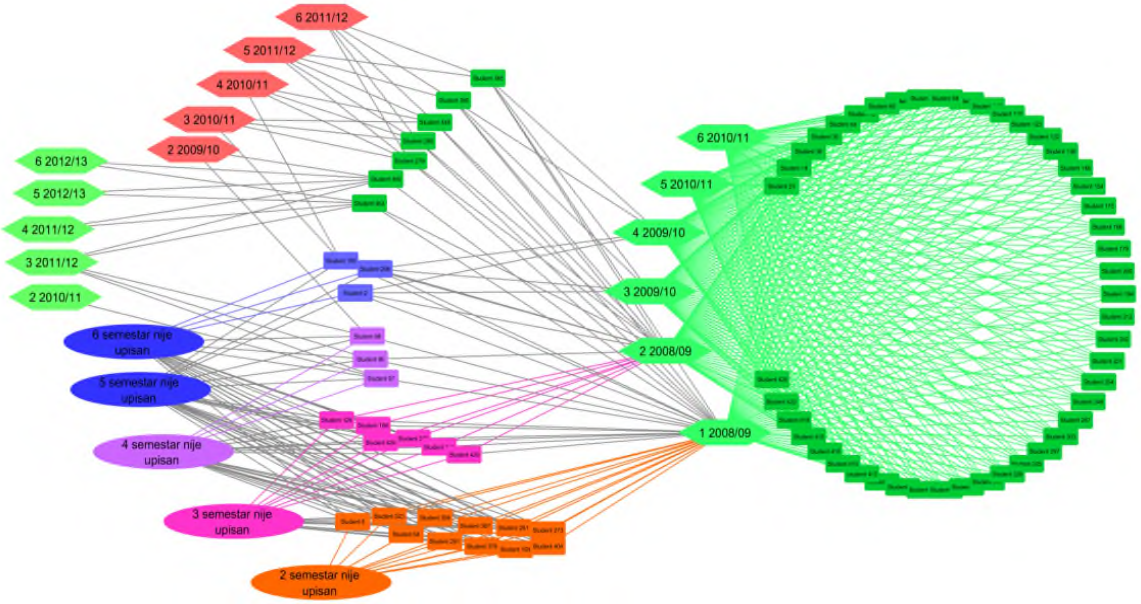

Figure 5: Performance of students enrolled in the first semester 2008/09

Generation enrolled in 2008/09, from total of 76 students, only 46 students or 61% finished within the deadline, without renewal of the year, plus seven other students who renewed one year. Other students, even 30, gave up at some time of their study.

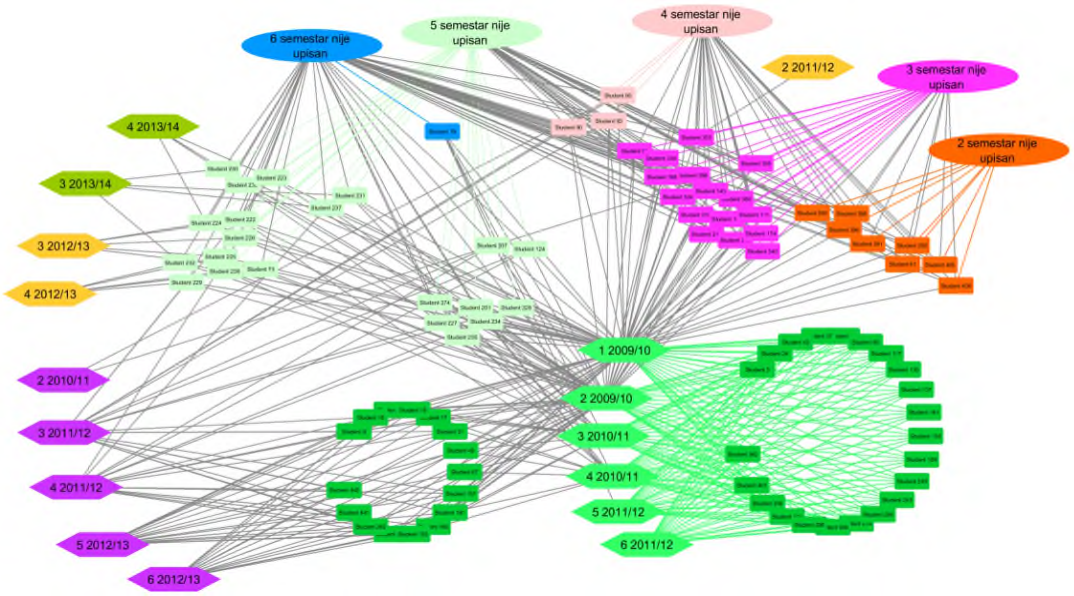

Figure 6: Performance of students enrolled in the first semester 2009/10

Generation enrolled in 2009/10, out of 87 students, only 22 students or 25% finished within the deadline, without renewal of the year, plus 16 students who renewed one year. Other students, even 49 of them, gave up at some time of their study.

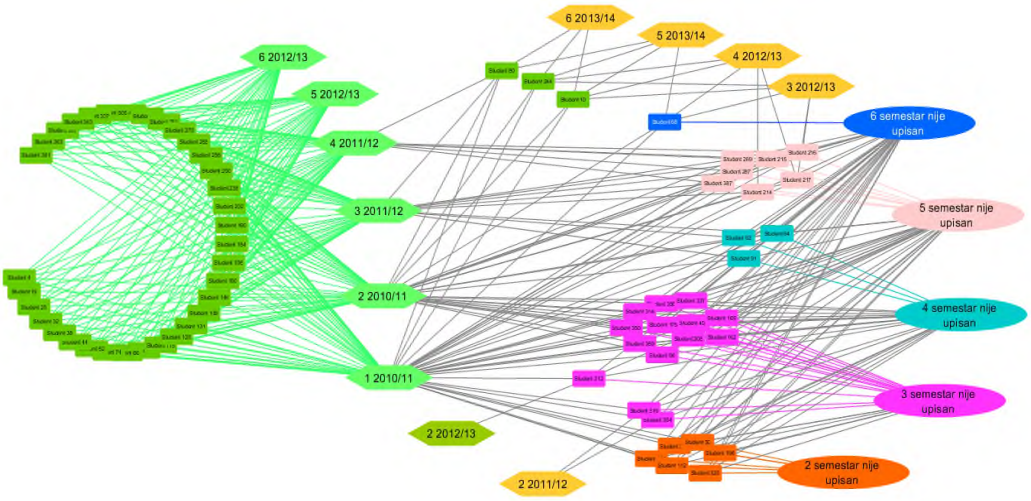

Figure 7: Performance of students enrolled in the first semester 2010/11

Generation enrolled in 2010/11, from total of 68 students, 34 students or 50% finished within the deadline, without renewal of the year. Other students, even 31, gave up at some timeof their study.

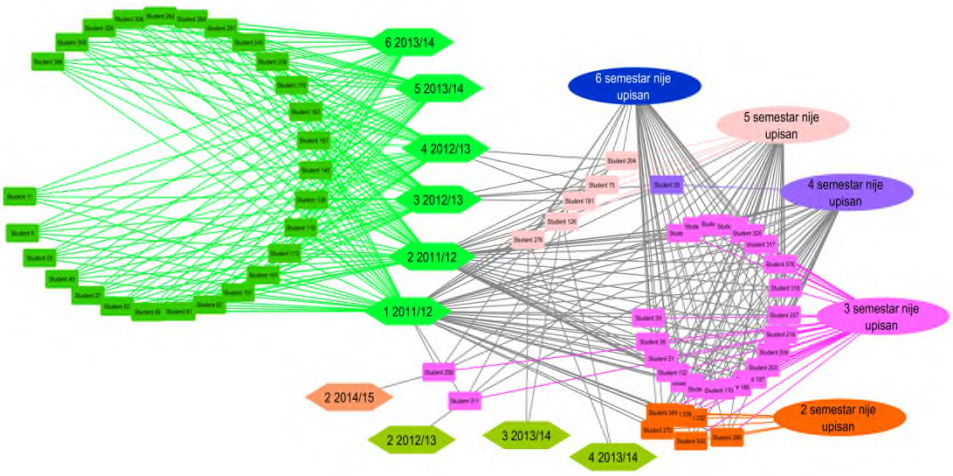

Figure 8: Performance of students enrolled in the first semester 2011/12

Generation enrolled in 2011/12, from total of 62 students, 27 students or 44% finished within the deadline, without renewal of the year. Other students, even 35 of them, gave up at some point in the study.

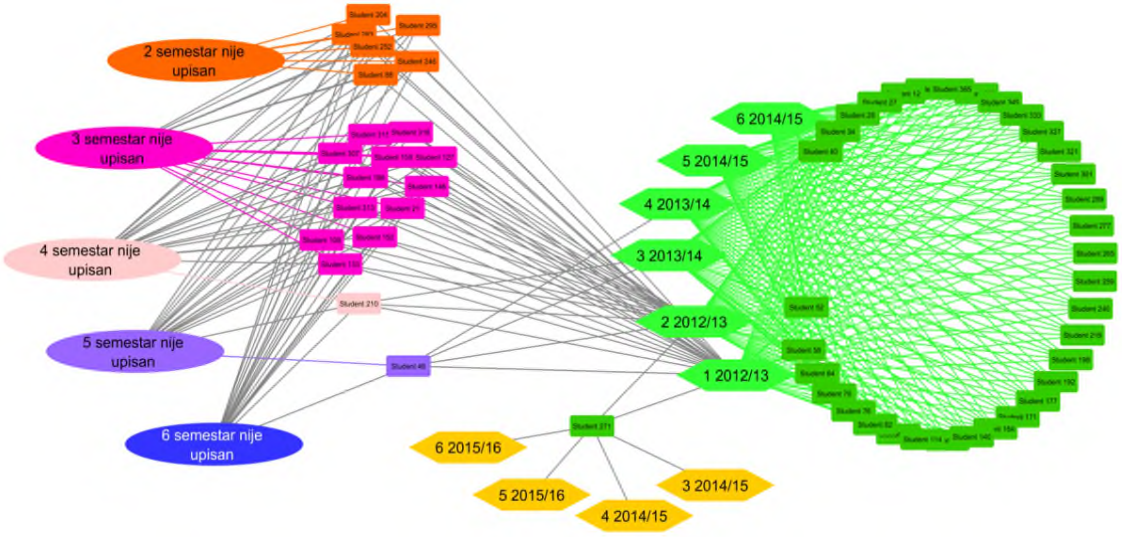

Figure 9: Performance of students enrolled in the first semester 2012/13

Generation enrolled in 2012/13, from total of 55 students, 34 students or 62% finished within the deadline, without renewal of the year plus one student who renewed one year. Other students, even 20, gave up at some time of their study.

According to the results, we can conclude that the largest percentage is percentageof students who actually completed the studies regularly. The highest percentage have generation enrolled in 2012/13 (62%), and the smallest generation 2009/10 (only 25%).

The reason for this could be found in the success of students from secondary schools, or from the success of students on preliminary exam. In order to demonstrate the existence of a relationship between these categories and the performance of students we carried on our further research in that direction.

Figure 10 shows the distribution of regular students by two variables, the success of the student from the high school and the success on the preliminary exam with the definition of students through theircollege performance. The green circles represent the students who have completed their studies regularly without renewal of academic year,the turquoisestars represent the students who completed the studies with the renewal of one year,the red ax color are the students who renewed two years. This distribution shows that among the regular students, the guarantees of high-level passing can be: if points from a high school are above 30, the points earned on the preliminaryexam must be at least 25 or more, and if the points achieved at the preliminary exam areover 40,points from high school must be at least 25 and more. There is an extremely small number of students that regularly studied and have renewed two years the number is practically statistically negligible.

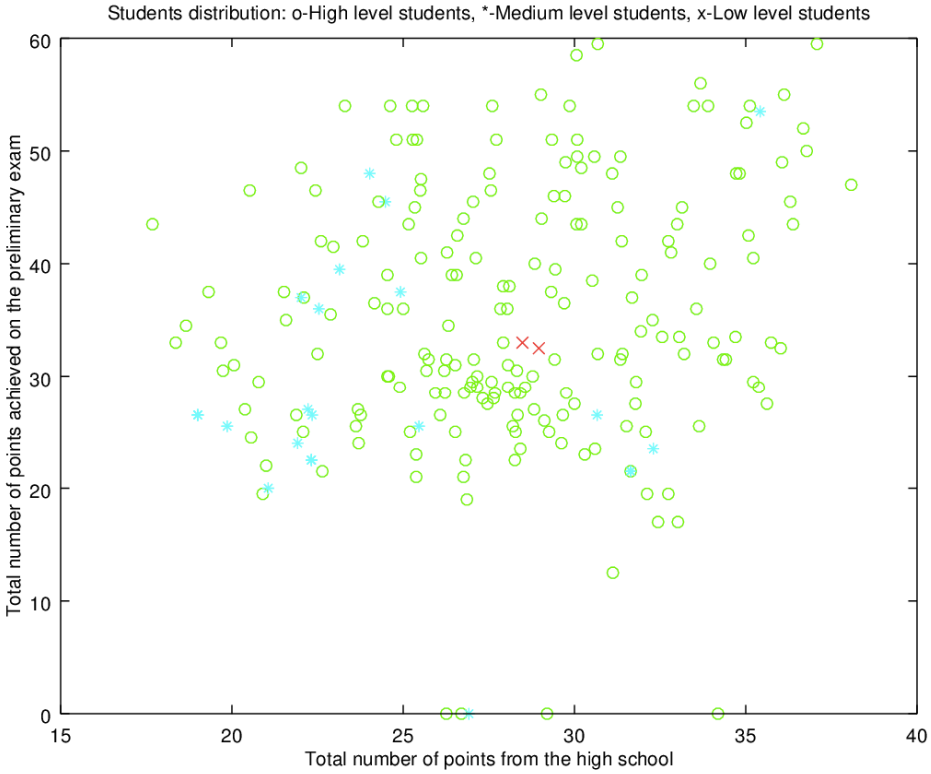

Figure 10: Distribution of regular students by two variables, the success of the student from the high school and the success on the preliminary exam with the definition of students through theircollege performance

The distribution of all enrolled students, including students who dropped out of their studies at some time (violet triangles), looks like at Figure 11. Previous analysis showed that the percentage of students who gave up their studies is quite large, and further this distribution shows that the previous conclusion for the distribution of regular students that a certain number of points achieved on the preliminary exam and a certain number of points that student brings from a high school can be a guarantee of a high level pass, which is not the case with regard to giving up. Even those students are at-risk students. At this point of analysis the alarm should be turned on. It is necessary for teachers and school management to take appropriate measures adapted to profile of at-risk students, using Learning analytics techniques. Of course, measures can have a positive effect if they are applied at the right moment. So a prediction of potential at-risk students is needed. We conclude, based on the diagram on Figure 11, that for such a prediction not only parameters of success on the preliminary exam and success in high school are relevant. According to the authors application of the Data Mining algorithm-Decision Trees over these data (415 records for analysis) for the prediction of at-risk students did not give satisfactory accuracy. The prediction accuracy on the test data set was 60%. The used parameters were: the number of points on the preliminary exam, the number of points from the secondary school, the four-year/three-year education and the technical/non-technical education profile in the secondary school. The algorithm nevertheless showed the important fact that the parameters of the greatest importance to the success of the students, from the listed parameters that were input at the analysis, are the number of points on the preliminary exam and the number of points from the high school. We conclude that it is necessary to complete the database with additional data. Additional sources can be LMS, surveys, results of preliminary tests, etc., that is all data that would indicate the flow and progress of student's activities.

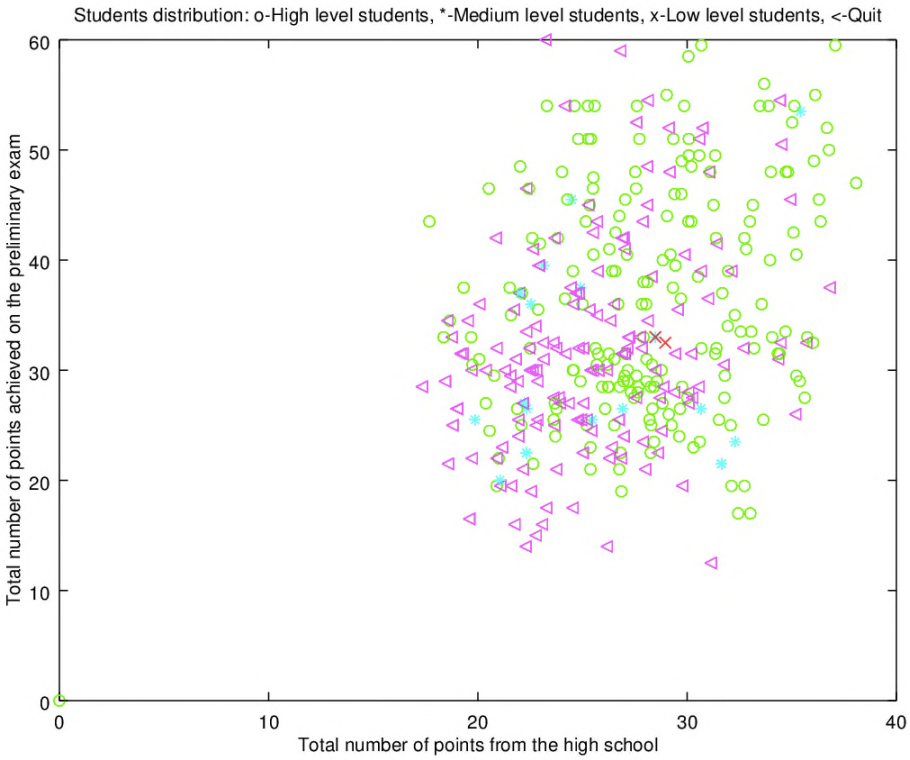

Figure 11: The distribution of all enrolled students in the period 2007-2013. by two variables, the success of the student from the high school and the success on the preliminary exam with the definition of students through theircollege performance

5. CONCLUSION

The visualization and explorationof data from students' database showed students' performance during the period of 2007-2013. Some interesting conclusions were made as well as some new ideas of managing teaching-learning process. Learning analytics is a new and rapidly developing field, it combines expertise from different academic disciplines such as educational data processing and predictive modeling. Students who are having great difficulty with their studies could beindentified so they could be offered timely support. In the future work authors will focus in such predictive tools but the institution must take more efforts in gathering more detailed information about the students in order to achieve suitable prediction accuracy.

REFERENCES

1. https://www.jisc.ac.uk/reports/learning-analytics-in-higher-education, visited at August 2017.
2. John P. Campbell, Peter B. DeBlois, Diana G. Oblinger, "*Academic Analytics: A New Tool for a New Era*", EDUCAUSE Review, vol. 42, no. 4 (July/August 2007): 40–57, Aveilable at: http://er.educause.edu/articles/2007/7/academic-analytics-a-new-tool-for-a-new-era
3. GabrijelaDimić, DraganaProkin, Kristijan Kuk, "*Primena Decision Trees i Naive Bayes klasifikatoranaskuppodatakaizdvojeniz Moodle kursa"*, INFOTEH-JAHORINA Vol. 11, March 2012.
4. GabrijelaDimić, DraganaProkin, Kristijan Kuk, BoškoBogojević, "I*zborklasifikatoraza mali obučavajućiskupobrazovnihpodataka"*, INFOTEH-JAHORINA Vol. 12, March 2013
5. Brijesh Kumar Bhardwaj, SaurabhPal,Data Mining: "*A prediction for performance improvement using classification"*, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.
6. K. Bunkar, U. K. Singh, B. Pandya and R. Bunkar, "*Data mining: Prediction for performance improvement of graduate students using classification,"2012* Ninth International Conference on Wireless and Optical Communications Networks (WOCN), Indore, 2012, pp. 1-5.
7. A. Desai, N. Shah and M. Dhodi, "*Student profiling to improve teaching and learning: A data mining approach,*" 2016International Conference on Data Science and Engineering (ICDSE), Cochin, 2016, pp. 1-6.