International Journal of Innovation and Economic Development
ISSN 1849-7020 (Print)
ISSN 1849-7551 (Online)

URL: http://dx.doi.org/10.18775/ijied.1849-7551-7020.2015.64.2004
DOI: 10.18775/ijied.1849-7551-7020.2015.64.2004

Volume 6
Issue 4
October 2020
Pages 54-61

Predicting Students' Academic Performance Based on Enrolment Data

Alisa Bilal Zorić

Polytechnic Baltazar Zaprešić, Zaprešić, Croatia

Abstract: Efficient education is key to the development and progress of modern society. Identifying factors that affect students' academic performance is a very important step towards efficient education. With fast IT development and lower prices, universities start to collect a huge amount of data. With data mining methods and techniques, universities can use this data, analyze it and get hidden and useful information. This paper presents a model for predicting students' academic performance based on enrolment data using one of the data mining techniques, Neural network. The enrolment data consists of demographic and economic data and information about previous education. Students' academic performance is measured by grade point average in university, and based on that, students are divided into two groups. One group consists of students with a grade point average below 3.5, and the other group consists of students with a grade point average above 3.5. This model may represent the first step for educators to early intervene and reduce the percentage of students leaving universities. They could offer students who are classified below average some additional classes to overcome the more difficult courses because of insufficient prior knowledge, thereby, increasing their likelihood of continuing their studies.

Keywords: Neural networks, Educational Data mining, Student's academic performance

1. Introduction

In Croatia, there is a large number of institutions of higher education and they operate in a very complex and highly competitive environment. Predicting student's academic performance is one of the most important steps towards efficient education and university's profitability, especially for private ones which are fully funded by tuition fees. It affects the modification of the existing programs and the creation of new ones. With accelerated IT development and lower prices, universities start to collect a huge amount of data about their students. These data can be furder analyzed with data mining methods and techniques. A special application of data mining methods and techniques in the educational environment is called Educational data mining. It is an interdisciplinary area that brings together techniques from statistics, artificial intelligence, database systems, machine learning, pattern recognition, data visualization, knowledge acquisition and information theory (Sumathi and Sivanandam, 2006) to find useful patterns and, thus, help understand students' behavior and how they learn.

The application of educational data mining is wide. The most common use is in improving the process of studying (Mardikyan and Badur, 2011), improving course organization (Rashan and Peiris, 2011), supporting students in course selection (Kardan et all., 2013), finding problems leading to dropping out (Bhardway and Pal, 2012), and as a support for decision-making at student admissions (Yadav and Pal, 2012). An empirical study of the potential applications of data mining methods and techniques in higher education was given by Kumar and Chadha (2011). The following applications were pointed out: predicting student performance, organization of syllabus, detecting cheating in the online examination, predicting the registration of students in an educational program and identifying abnormal or erroneous values.

In Croatia, to enroll university, candidates are required to fill in an enrolment form that consists of questions regarding previous education, personal data like gender, date of birth and some questions regarding social status. There are many different factors that influence successful study such as regular class attendance, time spent learning, work during studies, demographic factors, parent education, time since finished high school, finished high school type, time spent learning, scholarship, midterm exams, self-motivation, etc. These factors have been the topic of several research that have shown their impact on successful study. Since some of these data are available on enrolment form, the goal of this paper is to check if there are any indicators available on the enrollment form that can predict students' academic performance. Another goal is to present the opportunities and benefits of educational data mining and to show their application in the specific case using Allyuda Neurointelligence. Allyuda Neurointelligence is a data mining tool which does not require detail knowledge about the complexity and operation of neural networks for using it, but, still, can provide very useful and understandable results.

2. Literature Review

Identifying factors that affect a student's academic performance as well as finding a predictive model is the main topic of many researchers. A systematic literature review on predicting student' performance by using data mining techniques is provided by Shahiri and Husain (2015). The main goal of their paper was to provide an overview of the data mining techniques that have been used to predict students' performance and how the prediction algorithm can be used to identify the most important attributes in a student's data.

Devasia et al. (2016) examine various data mining techniques for the prediction of students' performance. They used all student admission details, course details, subject details, student marks details, attendance details and student's academic history as input. The results of their paper show that Naive Bayesian algorithm is more accurate than other methods like Regression, Decision Tree, Neural networks etc., for comparison and prediction.

Oancea et al. (2013) used the classification power of a neural network to predict the students' results measured by the grade point average in the first year of study. The input variables were: type of the study program (part-time or full-time education), gender, high-school graduation average, age and difference in years from the moment the student graduates high-school until he/she enrolls at university. The goal of their research was to help university management in order to take early action to avoid the phenomenon of leaving education.

Arsad and Buniyamin (2013) presented a study on Artificial Neural Network (ANN) model development in predicting academic performance of engineering students. They used Cumulative Grade Point Average (CGPA) as a dependent variable and grades of fundamental subjects in the first semester as independent variables. Performances of the models were measured by coefficient of Correlation R and Mean Square Error (MSE).

Ramesh et al. (2013) used data mining techniques to predict the performance of the students in final examinations. Another scope of their paper was to identify possible factors that influence student's performance. They found that Multi Layer Perception algorithm was best suited to predict the grades. They concluded that type of school does not influence student performance and that parents' occupation plays a major role in predicting grades

Baradwaj, and Pal (2012.) used the decision tree method for the classification task of evaluating student's performance. The results of their research could help in identifying students who need special attention and dropouts. An early prediction of students at risk of failure may help the management to increase the success rate and student retention.

Osmanbegovic, and Suljic (2012) compared different methods and techniques of data mining for the prediction of students' success. They investigate which variables (socio-demographic variables, achieved results from high school and from the entrance exam, attitudes towards studying) may have an effect on students' success. The goal of their research was to develop a model which can derive the conclusion on students' academic success measured by passing grade at the exam.

Kabakchieva (2012) used various data mining classification algorithms, including a decision tree classifier, a neural network, a rule learner and a Nearest Neighbour classifier to develop a data mining models for predicting student performance, based on their personal, pre-university and university-performance characteristics. The main goal of her research was to demonstrate the great potential of data mining applications for university management.

3. Research Methodology

There are many various data mining methods to build a predictive model for student's performance. The most commonly used method is classification. Among classification algorithms are Neural Networks, Decision Trees, Naïve Bayes, Support Vector Machine and K-Nearest Neighbor (Shahiri and Husain, 2015).

Every data mining process involves six main steps. First, there is a Business Understanding in which a specific problem is translated into a data mining problem. The second step is Data Understanding, which starts with data collected from all applicable data sources. In this step, data load and data integration are done. Data visualization tools are often used in this step to explore the properties of the data. The most important step is Data Preparation, and it can take enormous amounts of time depending on the amount of data analyzed and the number of data sources (Blake and Mangiameli, 2011). Data quality is a major challenge in data mining. The final data set must be cleaned, formatted and constructed into a specific form. In the Modelling and Evaluation step, mathematical models are used to find patterns in the data using sophisticated data tools and parameters are calibrated to optimal values. The last step is Deployment in which the results of data mining are presented (Oracle, 2019).

This research was done on the data collected from 76 students from the University of Applied Sciences Baltazar, Zaprešić. Table 1 presents the number of students distributed by average grade point in high school.

Table 1: Number of students divided by average grade point in high school

Average grade point range	Number of students	Persentage
4.5 to 5.0	12	16%
4.0 to 4.5	31	41%
3.5 to 4.0	27	35%
3.0 to 3.5	6	8%
2.5 to 3.0	0	0%
Total	76	100%

The purpose of this paper is to create a prediction model for students' academic performance, based on enrolment data. We designed a neural network using Alyuda NeuroIntelligence software package and we got a model by which we could determine student's academic performance: good, with an average grade above 3.5, and bad, with an average grade below 3.5. We used university average grade which represents an average value of all final grades individual students earned from the time they first enrolled in University by the end of study because it is a standard measure of success on universities. In Croatia, grading system consists of five grades. Grade 1 means insufficient, student did not pass the exam or satisfy minimal criteria. If student passed the exam, he gets at least 2, which means sufficient. After that, we have grade 3 - good, 4 - very good and 5 – excellent, which represents a top grade. In our

research, we used the following attributes as an input: status of study (full-time students and part-time students), gender (male or female), mother's education (elementary school, secondary education, bachelor's, master's or doctor's degree), father's education (elementary school, secondary education, bachelor's, master's or doctor's degree), average grade during high school (pass, good, very good, excellent), state matura level - Croatian language, mathematics, foreign language, state matura grade - Croatian language, mathematics, foreign language, average monthly income (considerably below average, little below average, average, little above average and considerably above average), accommodation during studies (with parents/relatives, payment of rent or dormitory), scholarship (yes or no) and work during studies (Yes, full time, Occasionally and No). Since the academic year 2009/2010, to enroll in a college in Croatia, candidates are required to pass a state matura exam. State Matura exam is a secondary school leaving examination. The exams are managed and organised by the National centre for external evaluation of education. It consists of 3 compulsory and 1 optional exam. Compulsory subjects from which candidates have to take state matura exams are: the Croatian language, mathematics, foreign language, and they may choose the level they want to take: basic or extended. The points gained in the exams are converted into points for enrollment. Each Croatian higher education institution sets its own criteria of valuing these exams depending on the area of science or art which is taught. Accordingly, the variables of the Level of state matura were defined. State matura grade can have standard values: pass, good, very good, excellent.

Table 2: Students related variables

Varijable Name	Description	Domain
Status	status of study	full-time; part-time
Spol	gender	male or fimale
Obraz_majka	Mother's education	elementary school, secondary education, bachelor's, master's or doctor's degree
Obraz_otac	Father's education	elementary school, secondary education, bachelor's, master's or doctor's degree
Prosjek_sred_skola	average grade during high school	2.5-3.0; 3.0-3.5; 3.5-4.0;4.0-4.5; 4.5-5.0
Hrv_razina	state matura level - Croatian language	basic, high
Mat_razina	state matura level - Mathematics	basic, high
Jezik_razina	state matura level - Foreign language	basic, high
Hrv_ocjena	State matura grade - Croatian language	pass, good, very good, excellent
Mat_ocjena	State matura grade - Mathematics	pass, good, very good, excellent
Jezik_ocjena	State matura grade - Foreign language	pass, good, very good, excellent
Kuc_fin_primanja	average monthly income	considerably below average, little below average, average, little above average and considerably above average
Smjestaj	accomodation during studies	with parents / relatives, payment of rent or dormitory
Stipendija	scholarship	yes, no
Rad_studir	work during studies	Yes, full time, Occasionally,No

We had problems with missing data, because in our dataset, we had students who graduated before 2009, and they did not have state matura exams.

4. Neural Network

A neural network is a mathematical computing model based on the structure and functions of the biological neural system. It is a nonlinear predictive model that learns through training and looks like biological neural networks in structure (Han et al, 2011).

A basic unit of a neural network is a neuron. Every neuron has two parts: the activation function and the net function. The activation function is a mathematical formalism that is used to define the output behavior of a neuron. The output of the activation function is input for the next node and so on until a satisfying solution to the original problem is found. Activation function allows a neural network to learn complicated, non-linear mappings between inputs and response variables (Hinton and Osindero, 2009). The net function determines how the network inputs are combined inside the neuron (Schölkopf et al., 2002). Neural networks have three layers: Input layer, Hidden Layers and Output Layer. The layers consist of nodes interconnected by the activation function. Patterns are presented to the network through the input layer. The input layer communicates with one or more hidden layers. The actual processing is performed in hidden layers through a weighted connection system. The hidden layers connect to an output layer where the response is sent. Each input is sent to each neuron in the hidden layer and then each hidden layer's neuron's output is associated with each neuron in the next layer. We can say that hiddden layer is intermediator between the input layer and the output layer. Information from the environment enters the network through the input layer and is then processed by the second layer (the first hidden layer), which becomes the input to the next hidden layer and so on. This continues until all the hidden layers are processed and the result is transferred to the output layer. The neural network learns from processing many marked examples that are submitted during training. It uses this answer key to learn what characteristics of the input are needed to construct the correct output. After the neural network has processed a sufficient number of learning examples, it can start processing new, unrecognized samples.

The neural network learns from experience, which means the larger the training set is, the more accurate the results are. Neural networks are widely used in different areas such as medicine, computer science, finance, facial recognition, email spam filtering, pattern recognition, self-driving vehicle trajectory prediction, etc..

5. Results

Enrolment data acquired in the process of admission to the university and a neural networks software application designed to assist neural network, data mining, pattern recognition, and predictive modeling experts in solving real-world problems, Alyuda Neurointelligence was used to create the prediction model. After selecting a database, the software goes through five phases: data analysis, data pre-processing, design, training and testing. In the data analysis phase, we are determining the target characteristics that we want to calculate, defining characteristics we will use and characteristics which we will reject like name, surname and personal number. The target characteristic is Average_Ok which has the value of one if university average grade is higher than 3.5 and zero if university average grade is below 3.5. Based on meny research (Ibrahim and Rusli, 2007; Osmanbegovic and Suljic, 2012), we used the university average grade (cumulative grade point average) to evaluate student's performance. University average grade is a average of all university grades based on final exam score consisted of course structure (exercises, seminars, tasks).

In the data analysis step, Alyuda Neurointelligence software package randomly divides Data into three sets: training - blue (68.75 %), validation - green (15.63 %) and testing - red (15.63 %).

In the pre-processing phase, some columns are added if data is marked as Categorical (before pre-processing, there were 17 columns, and after, there were 54).

In the network design phase, a number of hidden layers is selected. The program offers the best topology, which can be changed. In our case, that is a neural network with one hidden layer with 28 neurons.

After the designing phase, there is training in which different parameters like training algorithms, stop training conditions and different training algorithm's parameters can be defined. One of the seven algorithms can be selected: Quick Propagation, Conjugate Gradient Descent, Quasi-Newton, Limited Memory Quasi-Newton, Levenberg-Marquardt, Online Back Propagation and Batch Back Propagation. Different conditions for stop training can be selected: By error value, By error change and By iterations. The best training algorithm does not exist, the choice of algorithm depends on the characteristics of the problem we want to solve. Quick Propagation algorithm was used and for stop training conditions By iterations (500) option was selected.

After training the network, results are presented in Actual vs. Output Table in which variable Match? OK or Wrong is set for every row of input data. There is also a Confusion Matrix as shown in Table 1 and Mean Correct Classification Rate in our case was 93,421053 % (the prediction rate is very high due to insufficient data of lower-grade students (only 8%)).

Table 3: Confusion Matrix

Target output	0	1
0	30	3
1	2	41

In the end, we got a model with which we can check the likelihood of future students' academic performance by entering some parameters obtained during each student's enrolment process.

Different neural network topologies were tested (different number of hidden layers, different algorithms and parameters), and all gave similar results.

6. Conclusion and Recommendations

The analysis of the success of studies is important for attracting new students and the retention of existing ones. Discovering the rules that trigger certain negative occurrences, such as rules that lead to students dropping out, in time, certain actions may be taken to help them remain in the system. Creating a predictive model based on the data that universities already have can be very helpful.

Data mining is one of the most popular techniques to analyze student's performance. The performance of the student is one of the most important aspect of every educational institution (Solomon, Patil, & Agrawal, 2018). Modification of this model can be used for monitoring students' academic progress and to determine important strategies to successful university management as well as for prediction of students who are likely to drop out because of various factors such as socio- economic, academic and psychological.

The final academic status is always a result of the previous semesters, which gives an opportunity for the prediction of a subsequent semester performance based on the previous ones.

These studies revealed the eligibility of data mining techniques to efficiently predict the performance of students at different levels of their studies using various academic performance dependent factors. However, there still exists the need for more researches in conducting and producing an improved framework for SAP and dropout predictions because there are various factors affecting student's performance, and there are special factors that are depending of institution and region in which institution is. There are also special requirements in different countries, even in different universities, for example, some countries have state matura exam which is highly correlated with academic performance, some other countries have other special outputs and factors that affect successfull study. By identifying and analysing these factors, the hole educational process can be better planned and improved.

This research was made using a small database, applying only one method, neural network. With better data and larger quantities of it, the model could easily adapt to the needs of a particular higher education institution and, based on it, they could make decisions about student enrolment or the creation of study programs. By analyzing the data and adapting the model, as well as analyzing the importance of the influence of individual variables on the performance at studies, the model could also be used as a support to decision-making in education, thus contributing to more successful studies and enhancing the quality of education in general. The model created in this paper could be a good starting point for future researchers who could add the specifics of an institution to the model. Another direction for future research would be to use other methods such as association rule, segmentation, decision trees, clustering, outlier detection, etc.

References

- Arsad, P. M., & Buniyamin, N. (2013, November). A neural network students' performance prediction model (NNSPPM). In 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA) (pp. 1-5). IEEE. Crossref
- Bhardwaj, B. K., Pal, S. (2012). „Data Mining: A prediction for performance improvement using classification." arXiv preprint arXiv: pp. 1201.3418.
- Blake, R., Mangiameli, P. (2011), "The effects and interactions of data quality and problem complexity on classification", Data Inform, pp. 160-175. Crossref
- Devasia, T., Vinushree, T. P., Hegde, V. (2016, March). „Prediction of students performance using Educational Data Mining". In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). pp. 91-95. Crossref
- Han, J., Kamber, M., Pie, J., (2011). „Data Mining Concepts and Techniques". Burlington: Elsevier.
- Hinton, G. E., Osindero, S., Teh, Y. W. (2009). „A fast learning algorithm for deep belief nets." Neural computation, Vol. 7 No. 18, pp. 1527-1554. Crossref
- Ibrahim, Z., & Rusli, D. (2007, September). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In 21st Annual SAS Malaysia Forum, 5th September.
- Kabakchieva, D. (2012). „Student performance prediction by using data mining classification algorithms." International Journal of Computer Science and Management Research. Vol. 1 No 4, pp. 686-690.
- Kumar, V., Chadha, A. (2011). „An empirical study of the applications of data mining techniques in higher education". International Journal of Advanced Computer Science and Applications, Vol 2 No.3, pp. 80-84.
- Mardikyan, S., Badur, B. (2011). „Analyzing Teaching Performance of Instructors Using Data Mining Techniques". Informatics in Education, Vol 2 No. 10, pp. 245-257. Crossref
- Oancea, B., Dragoescu, R., Ciucu, S. (2013). „Predicting students' results in higher education using a neural network", available at: at https://mpra.ub.uni-muenchen.de/72041/ (21 December 2019)

- Oracle (2019), "Data Mining Concepts", available at: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON002 (07 November 2019)
- Osmanbegovic, E., & Suljic, M. (2012). „Data mining approach for predicting student performance." Economic Review: Journal of Economics and Business, Vol.1 No.10, pp.3-12.
- Ramesh, V. A. M. A. N. A. N., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. International journal of computer applications, 63(8). Crossref
- Rashan, K. H., Peiris, A. (2011). „Data Mining Applications in the education sector." MSIT, Carnegie Mellon University.
- Schölkopf, B., Smola, A. J., Bach, F. (2002). „Learning with kernels: support vector machines, regularization, optimization, and beyond". MIT press.
- Shahiri, A. M., Husain, W. (2015). „A review on predicting student's performance using data mining techniques." Procedia Computer Science. No. 72, pp. 414-422. Crossref
- Sumathi, S., Sivanandam, S. N. (2006). „Introduction to data mining and its applications", Springer. Vol. 29. Crossref
- Yadav, S. K., Pal, S. (2012). „Data mining: A prediction for performance improvement of engineering students using classification." arXiv preprint arXiv:1203.3832.