

PREDICTING THE ACADEMIC SUCCESS OF STUDENTS USING MACHINE LEARNING ALGORITHMS

Goran Savić¹

Abstract: The aim of this paper is to create a model that will show to students of the Higher Technical School of Professional Studies in Novi Sad if there is a risk of failure after the first year of studies; and also to demonstrate how machine learning models work in a simple and practical example. The model is based on information regarding student grades from courses in the first year of study and the number of attempts to pass the exams. The data was collected from the information system of the Higher Technical School of Professional Studies in Novi Sad, over a period of four academic years (2017/18, 2018/19, 2019/20 and 2020/21). Two algorithms used for modeling: logistic regression and multilayer perceptron. Using a 10-fold cross-validation, both algorithms produced similar accuracy of about 78%.

Key words: classification, machine learning, logistic regression, multilayer perceptron, n-fold cross-validation

1. INTRODUCTION

The success rate of finishing studies represents the ultimate goal for both higher education institutions and students. The aim of this paper is to show students of the Higher Technical School of Professional Studies in Novi Sad if there is a risk of failure after the first year of studies – based on their grades and number of attempts to pass the exams, and also to demonstrate how machine learning models work in a simple and practical example.

This topic was part of numerous research studies; some of them are shown below.

In the research [1], authors created the model that classifies students into one of two categories depending on their success at the end of the first academic year. They used decision trees for predicting. The higher classification rate was produced by REPTree algorithm, but this algorithm was not successful in classifying both classes. A slightly different approach was applied in paper [2], where authors presented a data mining approach to predict the academic success of the first-year students in Portuguese higher institutions. They used 68 features, encompassing socio-demographic, social origin, previous education, special statutes, and educational path dimensions. For research purposes, they developed models based on Decision trees, Support Vector Machines and Multilayer perceptrons. In this study, SVM outperforms other models. In the paper [3], the author proposed a model based on machine learning algorithms to predict the final exam grades of undergraduate students, taking their midterm exam grades as the source data. He compared the performance of random forests, nearest neighbours, support vector machines, logistic regression, Naive Bayes, and k-nearest neighbour algorithms. The proposed model achieved a classification accuracy of 70%-75%. In [4], authors created a model for prediction of students performance and analysis of factors that affect the achieved level of success. The model is based on students socio-demographic data, behavioural data, and personality characteristics. Model performance is measured by student's grade point average achieved over the period of studies. They used three types of machine learning algorithms: logistic regression, decision trees, and neural networks. All three models achieved approximately equal value. The best results were achieved by neural networks (76,4%), then logistic regression (74,8%) and lastly decision trees (71,2%).

2. METHODOLOGY

Given that this task includes a small number of input attributes, and the dataset is relatively small, for the task of predicting student success, i.e. the task of binary classification - will graduate, will not graduate, I chose two simple algorithms: logistic regression and multilayer perceptron. Testing of algorithms will be performed using SciKit Learn and Keras, the high-level API of the Tensorflow platform.

¹ MSc, The Higher Education Technical School of Professional Studies in Novi Sad, Školska 1, Novi Sad, Serbia, e-mail: savic.goran@vtsns.edu.rs

2.1. Logistic regression

Logistic regression is a simple supervised learning algorithm, often used for binary classification tasks. Logistic regression is a mathematical modelling approach that can be used to describe the relationship of several Xs to a dichotomous dependent variable [5]. The core of this method is the logistic function, that is a sigmoid function (an S-shaped curve). Through this function, the logistic regression maps a weighted linear combination of features into real values between 0 and 1. These real values can be interpreted as probabilities: rather than predicting a class, predicting a probability of belonging to a given class of data [6]. According to [7], logistic regression is a factual strategy for breaking down a dataset in which there are one or more autonomous variables that decide a result. It constructs a disuniting hyperplane between the two data sets. Logistic regression gives a functional form f and parameter vector α to express $P(y|x)$ as illustrated in

$$P(y|x) = f(x, \alpha) \quad (1)$$

The parameter alpha is tenacious by the minimum likelihood function. Logistic regression is a parametric method because the contribution of parameters can be interpreted.

2.2. Multilayer perceptron

Multilayer perceptron (MLP) is a widely accepted technique as a promising tool for solving complex engineering problems in real industrial processes. The MLP can include any number of layer with a specified number of neurones in each layer, which transfers information between the first and last layers [8]. **MLP** is a type of neural network technique that consists of an input layer to interpret the signal, an output layer that allows a judgement or assumption about the data, and an infinite number of hidden layers that determine the MLP computational power [9]. A neural network used in machine learning is a system composed of several simple processors (units, neurones), which are connected by communication channels (connections). These connections have corresponding weight coefficients, and each neurone has its own activation function that accepts input signals in order to determine the output signal.

3. DATA

The data collected includes grades from courses and the number of attempts to pass each course in the first year for the Information Technology program, as well as whether those students completed their studies. Each data point represents one input attribute in the algorithm. Considering there are 10 courses in the Information Technology program, the total number of input attributes is 20. The reason why data from students across all programs were not used is that not all programs have the same number of courses, nor the same courses, which would create a problem for the algorithm during training. The data was obtained from the Information System of the Higher Technical School of Professional Studies in Novi Sad, over a period of four academic years (2017/18, 2018/19, 2019/20, and 2020/21). The total number of students is 503.

Numbers of graduates and non-graduate students enrolled by academic year are shown in the following table:

Table 1 – Data

Academic year	Graduate students	Non-graduate students	Total	Success rate
2017/18	64	98	162	39,50%
2018/19	57	55	112	50,89%
2019/20	44	78	122	36,07%

2020/21	19	88	107	17,76%
---------	----	----	-----	--------

I have noticed that students at the Informatics department who enrolled in the first year of study in 2020 have an unusually low success rate. Only 17.76% of them had completed their studies by March 2024 (when the data were collected). One reason for this result probably lies in the fact that in that academic year, due to the pandemic of the COVID-19 virus, lectures were held online, while the other reason is that if some students renewed one of the years, they cannot appear in this table due to the time in which the data were collected. For this reason, I omitted data on students enrolled in the academic year 2020/21 from the analysis.

The total number of students for whom data has been collected is 503. When excluding students enrolled in 2020/21, 396 remain.

If we want the algorithm to be more accurate during classification, it is necessary that the training data be balanced. This means relatively equal presence of all classes. Given that, in our case, there are only two classes, 0 and 1, that is: Not graduated and Graduated, each class should be present in the training data about 50%. In the data set there are 165 instances of Graduated and 231 instances of Not graduated. There are several techniques in machine learning for balancing classes, for example, adding a coefficient to a minority class, synthesising new data, or reducing the number of instances of the majority class. For this work, I thought that synthesising new instances would reduce the accuracy of the algorithm, so I just decided to reduce the number of instances of the majority class, that is, the class - Not graduated. Now the training dataset has a total of 330 instances, of which 165 instances are of the Graduated class and 165 instances are of the Not graduated class.

4. RESULTS

MLP and logistic regression were tested on the balanced dataset. Both algorithms had 20 input attributes – ten for grades and attempts to pass. I also wanted to determine if there is a difference in accuracy on a small dataset if the data is split into train, validation, and test sets or if k-fold cross-validation is used. In machine learning, cross-validation is used to evaluate system performance in cases where we want to avoid overfitting. According to [10], in k-fold cross-validation, the available learning set is partitioned into k disjoint subsets of approximately equal size. Here, “fold” refers to the number of resulting subsets. The model is trained on k-1 subsets, which represent the training set. Then, the model is applied to the remaining subset, which is denoted as the validation set, and the performance is measured. Parameters used for the multilayer perceptron model are: output size=2, hidden layer 1=128, hidden layer 2=16, batch size=64, number of epochs=30, number of folds for cross validation=10, optimiser = ‘adam’, loss function = ‘sparse categorical cross entropy’, metrics = ‘accuracy’. Parameters for the logistic regression model are default parameters of the SciKit Learn library.

Results for tested models are shown in Table 2.

Table 2 –Results

	Algorithm	Accuracy without cross validation	Accuracy with cross validation
1	Logistic regression	75,76%	78,18%
2	Multilayer perceptron	66,67%	78,48%

From the results, it can be seen that better accuracy is achieved if 10-fold cross-validation is used for training the algorithm. When using the dataset split into train, validation, and test sets, MLP stops training after the second epoch due to overfitting. When cross-validation is used, accuracy is significantly higher.

5. CONCLUSION

In Table 2, it can be observed that even a simple machine learning algorithm, such as logistic regression, with a small number of input attributes and a small number of training data, can achieve a result that is very similar to the more complex MLP algorithm. Also, better accuracy can be achieved by using n-fold cross validation on a small training data set.

This model, obtained by applying multilayer perceptron, will be embedded in a desktop application, where Information Technology students can enter their grades from the first year and get the probability of completing their studies. In this way, students of the Higher Technical School of Professional Studies will see a practical example of machine learning algorithms.

For the next research, it is planned to create models for students of other programs at the Higher Technical School of Professional Studies.

6. REFERENCES

- [1] Mesarić, J.; Šebalj, D.: *Decision trees for predicting the academic success of students*, Croatian operational research review, 2016.
- [2] Gil Diniz, P.; Martins da Cruz, S.; Martinho Costa, J.; Moro, S.: *A data-driven approach to predict first-year students academic success in higher education institutions*, Education and Information Technologies, 2020, doi: 10.1007/s10639-020-10346-6
- [3] Yağcı, M.: *Educational data mining: prediction of students academic performance using machine learning algorithm*, Smart Learning Environments, 2022, doi: 10.1186/s40561-022-00192-z.
- [4] Simeunović, V.; Preradović, Lj.: *Using data mining to predict success of studying*, Croatian Journal of education, 2014.
- [5] Kleinbaum, D.; Klein, M.: *Logistic regression, A self learning text*, Springer, ISBN 978-1-4419-1741-6, 2010.
- [6] Choudhary, R.; Gianey, H.K.: Comprehensive review on supervised machine learning algorithms, *International Conference of Machine Learning and Data Science*, 2017, pp. 37–43, <https://doi.org/10.1109/MLDS.2017.11>
- [7] Pathan, S.; Siddalingaswamy, P. C.; Gopalakrishna Prabhu, K.: *Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review*, Biomedical Signal Processing and Control, ISSN: 1746-8094, Vol: 39, pp. 237-262, 2018.
- [8] Ehsan, D.; Behzad, V.: *Applying artificial neural networks for systematic estimation of degree of fouling in heat exchangers*, Computer Aided Chemical Engineering, Volume 52, pp. 685-690, 2023.
- [9] Manavalan, R.: *Automatic identification of diseases in grains crops through computational approaches: A review*, Computers and Electronics in Agriculture, 178, 105802. doi:10.1016/j.compag.2020.105802.
- [10] Berrar, D.: *Cross-validation*, Encyclopedia of Bioinformatics and Computational Biology, 2024.