Comparative Analysis of Machine Learning Based Predictive Models of Student Success: Survey versus Learning Management System Data

Alen Kišić[1]

[1]VERN University, Palmotićeva 82/1, Zagreb, 10000, Croatia

Abstract

Traditionally, the most common and accurate method of measuring opinion has been sample surveys, which ask carefully defined questions on precisely define samples of the population. Such an approach also comes at a high price: large investments of time, effort, and money for the researchers who design the research and collect the data, but also for the respondents who volunteer their answers. The problem with surveys is the honesty of the respondents, as well as the sample itself. Recently, an alternative to such an approach has emerged with the potential to supplement or even completely replace previously used research methods that would reduce costs for researchers and eliminate effort for respondents. Researchers started using data from social networks. In the domain of education, this potential is extremely large because students and teachers use learning management systems (LMS) for their teaching and learning.

The research conducted here applies machine learning algorithms to develop predictive models of student success based on: (i) students' activity data on LMS Moodle, (ii) students' satisfaction with the course measured by surveys. The main goal of the research is : (i) to compare the performances of predictive models based on LMS data with predictive models based on survey data, (ii) to identify predictors of student success. Results indicate that LMS data-based predictive models give models of higher accuracy and reliability in comparison to survey based predictive models.

Keywords

Machine learning, artificial intelligence, LMS data, survey satisfaction data.

1. Introduction

The digital transformation era has affected all aspects of society and education is not an exception. In higher education, predicting the academic success of students is important to serve as a basis for early intervention and optimization of educational resources. Traditionally, surveys have been the primary tool for gathering data on student habits, attitudes, and perceptions, serving as the foundation for developing predictive models. However, with the adoption of learning management systems (LMS), a new, potentially more effective method for predicting student success based on LMS data, came up. This paper explores and compares two approaches for developing predictive models of student success. The first approach is based on data obtained by student satisfaction surveys, whereas the second approach is based on the LMS data of students' interaction with the system. By applying a machine learning algorithm to both datasets, this study aims to determine which approach gives a more accurate and reliable predictive model.

While surveys provide valuable insights into the subjective experiences of students, collecting data through LMS offers several significant advantages. First, this approach is faster because it eliminates the need for the time-consuming process of survey design, distribution, and analysis. Second, it is cheaper since reduces the costs associated with conducting surveys and processing data. Third, it is simpler to implement, with automated data collection that minimizes human error and bias.

Data from LMS provide objective and continuous insight into student activities, including time spent on the platform, interactions with course materials, and learning patterns. This data, unlike periodic surveys, allows the development of more dynamic models that can track student progress in real-time and predict potential challenges before they become critical.

Proceedings for the 15th International Conference on e-Learning 2024, September 26-27, 2024, Belgrade, Serbia

kisic.alen@gmail.com (A. Kišić)

0000-0002-2196-1092 (A. Kišić)

This paper aims to: (i) compare the performances of predictive models obtained by LMS data and survey data developed by machine learning algorithms in both cases, (ii) identify predictors of students' academic success by using their subjective opinions and usage of LMS.

This paper is organized as follows. The second section gives an overview of recent previous research related to the research topic. The third section explains two datasets used in the research as well as machine learning algorithms applied here: artificial neural networks and decision trees. The fourth section provides research results and discusses their implications. The fifth section concludes the paper and gives guidelines for further research.

2. Related work

In the literature review, focus was on the: (i) investigation of statistical and machine learning approaches which previously have been used for student success prediction, (ii) exploration of data used for student success prediction.

Predicting student success is one of the goals for higher education institutions, as it can inform admissions decisions, guide interventions, and ultimately improve student outcomes [1]. Numerous studies have explored the use of statistical and machine learning algorithms to model student success, with a focus on identifying the most important factors and developing accurate predictive models [2].

Various studies proposed an artificial neural network model (ANN), e.g. [2] trained ANN on 121 features extracted from the records of over 60 000 students [2]. The model aimed to identify students likely to graduate, transfer to a different school, or drop out. Another study examined five commonly used machine learning models for predicting short-term and long-term academic success, with a focus on the trade-off between model interpretability and accuracy [3].

While machine learning methods have shown promise in improving predictive accuracy, some studies have found their interpretability as a drawback when comparing it with traditional statistical models, such as linear and logistic regression, in the context of academic achievement prediction [4].

Various types of data have been utilized for predicting student success in academic settings. These include demographic information, success grades in specific courses, class test scores, attendance records, assignment scores, midterm scores, and student-related data such as gender, parental education, test preparation, and lunch type [5].

Hussain study used demographic data and success grades in courses to predict student success [6]. Alfanaz study used data from students' learning outcomes in the basic control systems course to predict student performance through decision tree, KNN, SVM, and Naive Bayes algorithms [7].

The *'Students Performance in Exams'* dataset from Kaggle including attributes like ethnicity, gender, parental education, test preparation, and lunch type was utilized for predicting student success in Fahmida research [8].

Additionally, data mining techniques have been employed to extract useful information from student datasets, focusing on factors like student participation in discussion forums, accessing learning materials, and academic performance in online learning environments. Data from intelligent computer tutors, online classes, academic records, and standard assessments are used for predicting student success in online learning using machine learning techniques [9].

Labeled student education data was utilized for predicting student success in academic performance using ANN classifiers, support vector classifiers, random forests, and decision tree in the study of Partha [10].

Data on students` interactions with technology is commonly used for predictive modeling of student success, focusing on accurate outcome predictions [11].

The study pf Eleyan et.al. used machine learning algorithms on data from two secondary schools in Portugal to predict student final grades [12]. The data used for predicting student success includes personal information, academic evaluation, VLE activities, psychological factors, student environment, practical work, homework, mini projects, and student absences in the study of Ouatik [13].

Nyamane et. al. utilized student data from a LMS to predict academic success in blended learning environments [14].

The literature review showed that diverse datasets were used previously for students' success prediction by using different machine learning and statistical learning algorithms. However, we did

not found paper which would compare performances of predictive models based on different data sources. This is motivation for the work presented here.

3. Methodology

The goal of this research is to compare the performances of machine learning algorithms on data sets from learning management systems with data sets obtained from surveys to find out which one gives better predictive models of student success.

To do so, methodology is focused on: (i) the data to be used, and (ii) the machine learning algorithms of the artificial neural network and decision tree which will be applied to both data sets.

This section first gives a brief explanation of artificial neural networks followed by a description of the data to be used here.

3.1. Machine learning algorithms

Artificial Neural Networks (ANN) are computer models inspired by the structure and functioning of biological neural networks in the human brain. These complex machine learning systems consist of interconnected nodes or "neurons" organized into layers, capable of processing and transforming input data through a series of mathematical operations to produce a desired output.

The structure of an ANN consists of three main parts: the input layer, which receives the initial data; one or more hidden layers, where processing takes place; and the output layer, which provides the final result. Each neuron in the network has an associated weight and activation function, which together determine the strength and nature of the connection between neurons.

The learning process in ANN takes place through iterative adjustment of the weights of connections between neurons, to minimize the difference between the predicted and actual output. This process, known as network training, is usually carried out using algorithms such as backpropagation, which allow the network to "learn" from the examples presented, gradually improving its ability to generalize and predict.

A decision tree is a machine learning algorithm that uses a tree-like structure. It starts from the root node and branches into possible outcomes, where each node represents a test on an attribute, each branch an outcome of the test, and each leaf node a final decision or classification.

The algorithm builds the tree from top to bottom, choosing at each step the attribute that best divides the data set according to a certain metric (e.g. information gain or gini index). This process is repeated recursively for each branch, creating subtrees, until a stopping criterion is met.

Decision trees are popular for their interpretability - it's easy to follow the path from root to leaf and understand how the model makes decisions. However, they can be prone to overfitting the data, especially if allowed to become too complex.

Literature review have shown good performance of artificial neural networks and decision trees in educational setting. Furthermore, previous research on the comparison of survey and social network data in politics explored four machine-learning algorithms and identified ANN as the most accurate [15].

In the context of educational data mining and learning analytics, artificial neural networks show great potential due to their ability to detect complex, non-linear patterns in data sets. This characteristic makes them suitable for the analysis of educational data, where interactions between different variables can be subtle and difficult to detect with traditional statistical methods.

3.2. Data description

Both datasets are collected among third-year students of information technology at the University of Zagreb, Croatia. The sample consists of 76 students who took the course and fulfilled the survey.
List of variables both from survey data is enlisted in table 1.

Table 1.
Survey variables used in research

Survey variables
COURSE ORGANIZATION AND COMMUNICATION
All my obligations and deadlines in the course are clearly defined.
In the LMS, I manage well by chapters, topics and tasks.
I can satisfactorily monitor my progress in the LMS course.
I am satisfied with the possibilities of communication with the teacher.
COURSE MATERIALS AT LMS
The teaching materials are clear to me and help me to learn.
KNOWLEDGE AND SKILLS EXAMINATION
Knowledge tests refer to the contents of available teaching materials.
The method of implementation of the knowledge test is useful for mastering the material.
The feedback after the test was useful to me.
TEACHING IN GENERAL
My interest in this course.
Classes are held regularly.
I regularly attended lectures.
The teacher presented the teaching content clearly and comprehensibly in the lectures.
Methods, examples, and tasks facilitate the achievement of learning outcomes.
I regularly attended seminars/exercises.
The teacher clearly and comprehensibly presents the teaching content at the seminars/exercises.
Methods, examples and tasks in seminars/exercises facilitate the
The teachers know IT tools and techniques.
Lectures and other forms of teaching were coordinated.

Survey variables were grouped into four categories: course organization and communication, course materials at LMS, knowledge and skills examination, and teaching in general. Students were asked to express their level of agreement with the statements within each group. All variables had values on a Likert scale from 1 to 5 where 1 indicates “strongly disagree”, 2 indicates “disagree”, 3 indicates “neither agree or agree”, 4 indicates “agree” and 5 indicates “strongly agree”. The only exception is variable *My interest in this course* where students should rate it on a scale from 1 to 3.

LMS data consists of variables referring to a number of students log into each activity and resource on LMS Moodle. The following resources and activities were taken into account: File, Forum, Student report, Folder, Choice, File submission, Overview report, Page, System, Test, and Assignment. An overall number of points achieved during the course was the output variable in both predictive models. The maximum number of points that a student could achieve in the course is 100. As such, this is a regression machine learning problem. In the data preparation phase, various activities were performed to prepare data for predictive modeling. First, min-max normalization was applied to each data set. Secondly, outlier detection was performed. Third, missing values were explored. On the prepared data, modeling was performed.

4. Research results and discussion

In the training phase of model development, different artificial neural network architectures were explored to find the best one. In the end, ANN with three layers was employed for both data sets. In the case of survey data, there were 17 neurons in the input layer, 10 neurons in the hidden layer and 1 neuron in the output layer. In the case of LMS data, there were 11 neurons in the input layer, 6 neurons in the hidden layer and 1 neuron in the output layer. Feedforward artificial neural network multilayer perceptron (MLP) was used.

Decision tree was post-pruned. The best model was selected based on trade off between model quality and explainability.

Model quality was measured by RSquare and RASE. Values for both ANN models are presented in Table 2. Survey-based predictive model achieved lower reliability (Rsquare of 0.671) and a higher error rate (RASE of 0.342). The LMS-based predictive model outperformed the survey-based model by both criteria, reliability, and accuracy.

Table 2.
Neural network predictive model evaluation

Model	RSquare	RASE
Survey based predictive model	0.671	0.342
LMS based predictive model	0.814	0.232

Values for both decision tree models are presented in Table 3. Survey-based predictive model achieved lower reliability (Rsquare of 0.546) and a higher error rate (RASE of 0.351). The LMS-based predictive model outperformed the survey-based model by both criteria, reliability and accuracy.

Table 3.
Decision tree predictive model evaluation

Model	RSquare	RASE
Survey based predictive model	0.546	0.351
LMS based predictive model	0.743	0.249

However, ANN predictive model outperformed DT predictive model.

Sensitivity analysis was performed on both predictive models with the aim to identify most important predictors of student success. Results are presented at figures 1 and 2.

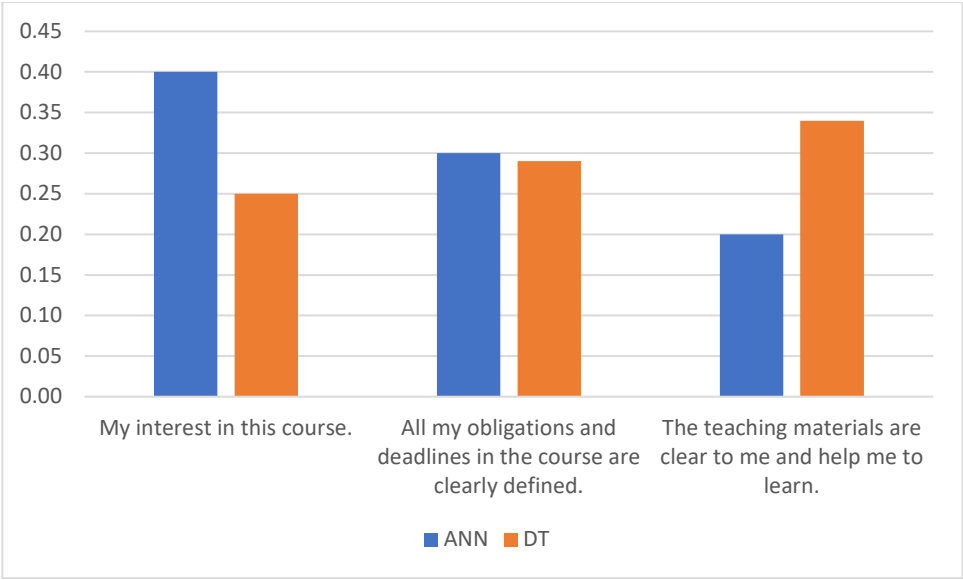

Figure 1: Sensitivity analysis for survey based predictive models

Sensitivity analysis of survey based predictive models based on ANN and DT identified top three predictors: *My interest in this course*, *All my obligations and deadlines in the course are clearly defined* and *The teaching materials are clear to me and help me to learn.* According to the results, student motivation is the best drive for success based on ANN model. Teaching materials and their usefulness is the best predictor in case of DT model.

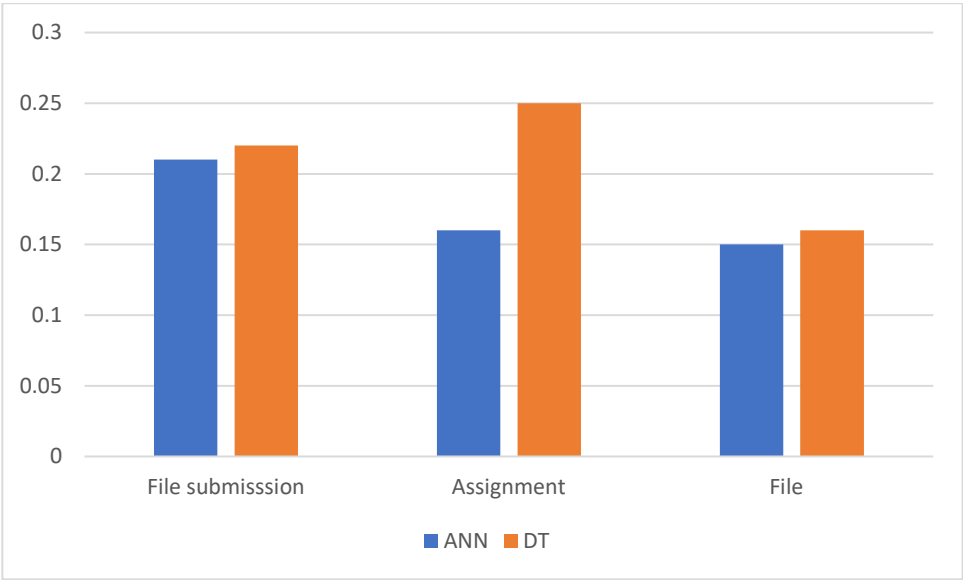

Figure 2: Sensitivity analysis for LMS based predictive models

Sensitivity analysis of LMS based predictive models identified top three predictors: Number of logs to file submission, number of logs to assignment and number of logs to file resource. Whereas ANN model identified File submission as the best predictor, DT model yielded Assignment as the most important predictor.

Comparison of two machine learning models shown neural network model as more accurate and reliable then decision tree model

5. Conclusion

This paper aims to contribute to the growing field of educational data mining and provides empirical evidence that LMS data combined with machine learning algorithms outperforms survey data with machine learning algorithms. Such an approach comes with a lower price of research: smaller investments of time, effort, and money for the researchers who design the research and collect the data.

Machine learning algorithms have proven to be powerful when analyzing LMS data. Artificial neural network provide better accuracy whereas decision tree gives high level of interpretability.

Research results can serve as input for educational decision-making based on LMS data and lead to future strategies for monitoring and supporting student success in a digital educational environment.

Research results contribute to further digitalization of higher education and support applications of artificial intelligence and machine learning for decision-making.

In future research, a larger sample of students will be employed along with students of different study programs. IT students investigated here have specific characteristics and that should be taken into account when generalizing results.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. Gándara, H. Anahideh, M. P. Ison, and L. Picchiarini, "Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction," AERA Open, vol. 10, p. 23328584241258741, 2023.

[2] S. Voghoei, J. M. Byars, S. J. King, S. Shapouri, H. Yaghoobian, K. M. Rasheed, and H. R. Arabnia, "Students Success Modeling: Most Important Factors," arXiv preprint arXiv:2309.13052, 2023.

[3] C. Kung and R. Yu, "Interpretable models do not compromise accuracy or fairness in predicting college success," in Proceedings of the seventh acm conference on learning@ scale, 2020, pp. 413-416.

[4] S. Cornell-Farrow and R. Garrard, "Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia," Communications in Statistics: Case Studies, Data Analysis and Applications, vol. 6, no. 2, pp. 228-246, 2020.

[5] S. Tosun and D. B. Kalaycıoğlu, "Data Mining Approach for Prediction of Academic Success in Open and Distance Education," Journal of educational technology and online learning, 2024. doi: 10.31681/jetol.1334687

[6] M. Hussain, S. Akbar, S. A. Hassan, M. W. Aziz, and F. Urooj, "Prediction of Student's Academic Performance through Data Mining Approach," Journal of informatics and web engineering, 2024. doi: 10.33093/jiwe.2024.3.1.16

[7] R. Alfanz, R. K. Hendrianto, and A. H. A. M. Siagian, "Predicting Student Performance Through Data Mining: A Case Study in Sultan Ageng Tirtayasa University," Journal of Advanced Computational Intelligence and Intelligent Informatics, 2023. doi: 10.20965/jaciii.2023.p1159

[8] F. F. Ananna, R. Nowreen, S. S. R. Al Jahwari, E. Costa, L. Angeline, and S. R. Sindiramutty, "Analysing Influential Factors in Student Academic Achievement: Prediction Modelling and Insight," International Journal of Emerging Multidisciplinaries Computer Science & Artificial Intelligence, 2023. doi: 10.54938/ijemdcsai.2023.02.1.254

[9] S. A. A. Kharis, G. F. Hertono, E. Wahyuningrum, and Y. Yumiati, "Design of student success prediction application in online learning using fuzzy-knn," Barekeng, 2023. doi: 10.30598/barekengvol17iss2pp0969-0978

[10] P. S. Ghosh, R. Roy, S. Mandal, M. Chowdhary, and S. Bokshi, "Data mining approach to predict academic performance of students," 2023. doi: 10.54646/bijcs.2023.21

[11] C. Brooks, V. Kovanović, and Q. Nguyen, "Predictive modeling of student success," in Handbook of Artificial Intelligence in Education, Edward Elgar Publishing, 2023, pp. 350-369.
[12] N. Eleyan, M. Al Akasheh, E. F. Malik, and O. Hujran, "Predicting Student Performance Using Educational Data Mining," in 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2022, pp. 1-7. IEEE.
[13] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Predicting student success using big data and machine learning algorithms," International Journal of Emerging Technologies in Learning (iJET), vol. 17, no. 12, pp. 236-251, 2022.
[14] S. Nyamane, A. Jadhav, and R. Ajoodha, "Predicting Academic Success in Blended Learning Environments: A Probabilistic Bayesian Approach Leveraging Student Trajectory Data," Available at SSRN 4663410, 2023.
[15] A. Kišić, "Modeli za predikciju ishoda političkih izbora korištenjem društvene mreže Facebook i algoritama strojnog učenja," Ph.D. dissertation, University of Zagreb, Faculty of Organization and Informatics, 2023.