Journal of Natural Sciences and Engineering, Vol. 1, (2020)
DOI number: 12.34567/JONSAE2020123

Student Attendance Pattern Detection and Prediction

Ibrahim Muzaferija[1], Zerina Mašetić[2], Samed Jukić[3], Dino Kečo[4]
[1]International Burch University, Sarajevo, Bosnia and Herzegovina
ibrahim.muzaferija@stu.ibu.edu.ba
zerina.masetic@ibu.edu.ba
samed.jukic@ibu.edu.ba
dino.keco@ibu.edu.ba

***Abstract* – Since the early beginnings of education systems, attendance has always played a crucial role in student success, as well as in the overall interest of the matter. The most productive way of increasing the student attendance rate is to understand why it decreases, try to predict when it is going to happen, and act on causing factors in order to prevent it. Many benefits of predicted and increased attendance rate can be achieved, including better lecture organization (i.e. lecture time and duration, lecture class choice, etc). This paper describes the steps in the extraction of knowledge from the university's student database and making a model that predicts whether the student will attend the class or not. Results show that the attendance patterns are best reflected when employing a decision tree algorithm, a C4.5 model that is interpretable and able to predict the attendance with 0.81 AUC performance measure.**

Keywords - Data Mining, Educational Data Mining, Machine Learning

1. Introduction

Data mining (DM) is an approach to discover useful information in data. It uses statistical and machine learning (ML) techniques to operate on large volumes of data to discover hidden patterns and relationships that describe the behaviors of systems that produced the data. Relationships and patterns discovered provide helpful insight into decision making, as well as making predictions, thus solving numerous problems.

In recent years, there has been an increase in the use of ML techniques in many fields, such as education, economics, business, statistics, medicine, and sport. The main objective of this paper is to apply ML techniques in the educational field to analyze student behaviors and to predict whether the student will attend the class.

Traditionally, educational institutions are collecting large volumes of data related to students, classes, faculty members, and educational processes. However, collected data is often not analyzed enough to provide significant results. In general, collected data is used for producing simple reports that are not highly significant in contributing to the decision making process in the institutions.

Currently, educational systems aim to enhance the teaching and learning process by carefully analyzing collected data, and discovering patterns related to student behavior and their final outcome. Reasons are to identify which students will perform well, so that they can be awarded scholarships and more importantly, to identify the students who may fail so that some form of help and assistance may be offered to them.

Besides identifying students by their performance, it's also important to discover which aspects of teaching and learning systems facilitate student learning and success. One of the aspects that are closely related to student performance is student attendance, meaning that students who have a higher attendance rate also have a higher success rate in the end [1].

The paper is structured in seven sections: 1. Introduction section; 2. The previous work section describes the previous efforts for the topic; 3. The methods and materials section describes data cleaning and processing steps; 4. The model creation section describes model selection and creation methodology; 5. The results section provides model results and evaluation; 6. In the discussion section, a comparison between this study and previous studies is made; 7. The conclusion section provides recommendations for future work in the area of educational data mining.

2. Previous Work

Gurmeet Kaur and Williamjit Singh [2] applied machine learning methods from the WEKA tool in order to predict students' performance from the College of Science and Technology – Khan Younis. Thir work was concluded with two classification algorithms, Naive Bayes and J48, which provided an accuracy of 63.59% and 63,53% respectively.

C. Anuradha and T. Velmurugan [3] conducted a comparative analysis of the evaluation of classification algorithms in the prediction of students' performance. The dataset was obtained from the college database, containing 19 attributes that describe the student, his family, and the living environment, as well as previous performances. Their goal was to compare algorithms in predicting students’ performance in end semester examinations. The results show that Bayesian classifiers, as well as JRip and J48, had the highest accuracy which is very close to 70%.

Abeer Badr El-Din Ahmed and Ibrahim Sayed Elaraby [4] describe the importance of Educational Data Mining (EDM) and Knowledge Discovery in Databases (KDD) in achieving the main goal of higher education institutions, that is, providing quality education to students. They used classification algorithms to identify those students who needed special attention in order to reduce the failing ratio and taking appropriate action at the right time, resulting in a decrease of the falling ratio by more than 15%.
Anal Acharya and Devadatta Sinha [5] used a dataset that contains a huge number of features that describe a student, by applying feature selection algorithms like Correlation-Based Feature Selection (CBFS) and Information Gain Attribute Evaluation (IGATE), they reduced the number of features and performed cross

modeling with five machine learning algorithms: Decision Trees (DT), Bayesian Networks (BN), Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP). Features related to gender, university, time, and family are the ones having the highest information gain, as well as the models created using decision tree algorithms, provide 10-15% more reliable performance in comparison to other classification algorithms.

The study conducted by Havan Agrawal and Harshil Mavani [6] confirms that past performances have indeed got a significant influence over current performances. Further, they used neural network algorithms and confirmed that the accuracy of the algorithms is proportional to dataset size, meaning that with the increase of dataset size, the algorithms generalize the problem better.

In this paper, we'll address the problem with a selection of best-performing machine learning algorithms for EDA, as proposed by Anal Acharya and Devadatta Sinha [5] and Gurmeet Kaur and Williamjit Singh [2], such as Logistic Regression, Decision Tree, Rule-based, k-NN, etc. Moreover, an increased number of data samples is obtained in order to improve the algorithms generalizing ability, in contrast to the number of data samples used in the previous study conducted by Gurmeet Kaur and Williamjit Singh [2].

3. Methods and Materials

The research is based on CRISP-DM [7] methodology as it describes common approaches used by data mining experts, while the paper contains a simplified version of the processing model shown below.

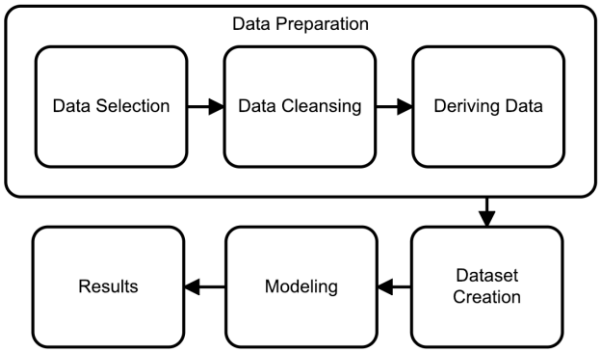

Figure 1. Data processing workflow

A. Data selection

Initial data was obtained from International Burch University's Student Academic System [8] and contains 2nd-year student attendance data from the years 2016/2017 and 2017/2018. Although the dataset doesn't contain all the details about the students and their classes (such as day of the week in which the class was held, exact start and end time of classes, professor ID, etc.), it's enough to extract the patterns of student attendance behavior and create a model that predicts it.

The data was obtained as an SQL file, and after importing the file to the local database, RapidMiner [9] was used to fetch the tables and store them in CSV format. Every further operation is done using the RapidMiner, as it has the Weka [10] extension.

The following table displays whether or not an attribute of the original dataset was copied over to the data mining dataset. All the selected attributes were considered relevant to the task of predicting student attendance to classes.

Table 1. Initial dataset attribute selection

Table	***Attribute***	***Accepted***	***Notes***
students	student_id	x	No need for additional IDs
	student_number	x	No need for additional IDs
student_courses	student_id	✓	Student ID
	course_code	✓	Course ID
	branch	x	Same values in other tables
	year	x	Same values in other tables
	semester	x	Same values in other tables
student_attendance	student_id	✓	Student ID
	attendance_id	✓	Class attendance ID
course_attendance	attendance_id	✓	Class attendance ID
	course_code	✓	Course ID
	branch	✓	Branch
	year	✓	Year
	semester	✓	Semester number
	course_date	✓	Starting date of the week in which class was held
	type	✓	Type of the class

	topic	x	Not relevant / High cardinality
	duration	✓	Duration of the class

B. Data Cleansing

In order to get an insight into data quality, graphical and statistical methods were used to detect anomalies, faults, outliers, missing values, etc. First, the dataset was divided into four parts: 1st semester of 2016, 2nd semester of 2016, 1st semester of 2017, and 2nd semester of 2017.

After examination, data related to both semesters of the year 2016 contained no anomalies and were consistent, thus were labeled as clean data. Furthermore, 2nd semester of the year 2017 contained incomplete data due to university system failure (class attendance from the last 2 weeks is missing), and 1st-semester data were not consistent (having a huge number of recorded attendances in the 14th week and almost none in 15th week).

The dataset contained automatic attendance values that were irrelevant for creating a model and those samples were removed. Some attendance samples recorded before and after the semester were marked as outliers. Samples related to midterm and final exams showed the decrease of recorded attendances due to the nature of exam weeks, as instead of multiple lectures in those weeks, only one was held - the exam. Those samples were not relevant in predicting the lecture attendance and were discarded.

C. Deriving Data

From the course_date attribute, containing the date of the week in which the class was held, week attribute was derived, containing week number in the semester.

The attribute attended is added to the table student_attendances and contains the value 1, which reflects that the student attended the class. Later when joining tables, this attribute will have missing values which indicate that students didn't attend the class.

The dataset contains only the records of students that attended the class and no records of students that didn't attend. In order to populate the attribute attended with reflection did the student attend the class, joining the tables is necessary.

First, by performing an inner join of student_courses and course_attendance tables, matching course_code from one table with course_code from another, a new table is created containing a matched list of students per course attendance IDs.

Next, by performing a left join of the previously created table and student_attendance table, matching both attendance_id and student_id from one table with attendance_id and student_id from another table, a new table is created containing attendance values where the student attended the class and missing values where the student was absent. Finally, missing values were replaced with 0, indicating that the student was absent.

D. Dataset Creation

During the data preparation phase, attributes considered most relevant were selected to shape the model's prediction capabilities. Then, using the RapidMiner tool, all data was cleaned and exported as a CSV dataset that will be used in training and testing the model. The final dataset contains about 58,000 attendance samples from the 2nd semester of the year 2016, and the following table displays qualitative and quantitative aspects of all the attributes present on the final dataset. The goal attribute (or prediction class) is "**attended"** which indicates did the student attend the class (marked as 1) or not (marked as 0).

Table 2 - Final dataset attribute description

Attribute	***Data type***	***Range***	***Missing values***	***Distinct values***	***Unique values***	***Statistics***
id	integer	[1,58019]	0	58019	58019	—
attended	integer	0,1	0	2	0	Least: 1 (21327) Most: 0 (36692)
course_code	nominal	MAN 201, (...)	0	85	0	Least: IRES 305 (5) Most: MAN 201 (6784)
branch	nominal	A,B,C,D,E, F	0	6	0	Least: D (1628) Most: A (37368)
type	nominal	Recitation, lecture, lab	0	3	0	Least: recitation (1954) Most: lecture (46511)
duration	integer	[1,4]	0	4	0	Min: 1 Max: 4 Average: 1.684
week	integer	[1,15]	0	15	0	Min: 1 Max: 15 Average: 7.861

4. Model Creation

This machine learning problem belongs to the classification types [11]. In order to reach the business goal, the complete understanding of data is required to generate the model. Currently, there are several modeling algorithms for classification types of problems, and they are shown in the table below.

In order to correctly create, evaluate and validate the model, one of the key steps is the separation of the data into training, testing, and validation.

Table 3. Machine Learning algorithms

Type	***Name***
Functions	Logistic Regression
Trees	ID3 (Decision Tree) C4.5 (J48) Random Forest
Rules	One-Rule PRISM
Memory-Based	k-NN

The most convenient method for training and testing separation is called Cross-Validation [12], as it splits the data into folds, and crosses the results of training and testing with different folds. The cross-validation is conducted using five folds of training data. Validation data will not be used in cross-validation in order to provide reliable testing results at the end.

5. Results

All the decision tree algorithms had the minimal gain set to "0.01" in order to prevent premature pruning of the tree branches, and pruning confidence threshold to "0.25". Other model settings have been kept on the default values because they are preselected for optimal model performance. After applying manifold training and testing methods known as cross-validation [13], building the models with different algorithms yielded promising results, as shown using the metrics such as accuracy, the area under the curve (AUC), precision, recall, fallout, and f-measure [14]. Moreover, models have been evaluated with validation data holdout and the results match with the cross-validation testing results presented below.

Table 4. Evaluations of created models

Algorithm	***Accuracy***	***AUC***	***Precision***	***Recall***	***Fallout***	***F-Measure***

Logistic Regression	75.37%	0.803	71.09%	55.63%	13.16%	62.41%
ID3	68.38%	0.697	56.20%	63.31%	28.68%	59.54%
C4.5	77.41%	0.812	73.04%	61.12%	13.12%	66.55%
Random Forest	66.48%	0.700	56.41%	38.73%	17.39%	45.92%
One-Rule	74.60%	0.500	69.25%	55.65%	14.39%	61.69%
PRISM	64.15%	0.500	71.90%	4.07%	0.93%	7.70%
K-NN	70.42%	0.672	58.13%	69.86%	29.25%	63.45%

The machine learning algorithm that creates the most accurate model is a decision tree algorithm known as C4.5. The reason is the enhanced method of tree pruning that reduces misclassification errors due to noise and too many details in the training data set, as described in the study conducted by Anuja Priyam et al [15]. The accuracy of the model is fairly satisfying, taking into consideration that previous works provided an accuracy of less than 70%. As opposed to previously mentioned studies, our data set contains more examples thus produces a more accurate prediction model. This process allows the extraction of relevant information from the model and helps draw the lines of action for this business problem.

Table 5. Confusion matrix for C4.5 model

	true 0	***true 1***	***class precision***
predicted 0	31878	8291	79.36%
predicted 1	4814	13036	73.03%
class recall	86.88%	61.12%	

In regards to interpretability, the decision tree generated by the C4.5 algorithm is easy to interpret as the size of the tree generated is 357 and the number of leaves is 230. The most important attribute on the dataset, as taken from the model, is the course code.

Furthermore, it's wrong to assume that one student attending classes has the same cost, from a business perspective, as one that never goes to class. That means that students that attend classes are beneficial and students that miss classes have a cost. With that in mind, the model needs to help in finding the solutions that decrease the overall cost. There are four possibilities:

1. We predicted the student would attend class and he did;
2. We predicted the student would not attend class and he did not;
3. We predicted the student would attend class, but he did not;
4. We predicted the student would not attend class, but he did.

Journal of Natural Sciences and Engineering, Vol. 1, (2020)
DOI number: 12.34567/JONSAE2020123

Point 1 is the best scenario, so it needs to have a negative cost (to be a benefit). Point 2 is the worst case, so it needs to have the highest cost. Point 3 is also negative, but not as negative as the previous one. Point 4 is positive, but not as good as the first point. With that information, it is possible to build a cost matrix for the class "Attended":

Table 6. Cost matrix for the model

		Actual	
		T	F
Prediction	T	-15	15
	F	-5	5

Building the cost-matrix doesn't affect the model's performance but aids in the final outcome of prediction by introducing the business bias and targeting to increase the business value.

6. Discussion

The possible issue with the study conducted by Gurmeet Kaur and Williamjit Singh [2] is the small number of instances (as low as 52) contained in the dataset and used to build the model. In order to make a model more accurate and more prone to generalization, Havan Agrawal and Harshil Mavani [6] propose using a higher number of instances, which made the model described in this paper more accurate. Moreover, cross-validation, as one of the extra steps that are taken in model construction, increased the model's overall ability to generalize and provide higher accuracy than models in previous studies.

While conducting the research, it was noticed that the quantity and quality of data plays a crucial role in the final outcome and performance. We highly devise to use a high number of instances in future studies, and continuum stream of attendance data in deployed models to continuously train the model as the trends responsible for student attendance dynamic behavior progresses over time.

The feature engineering task in the data preparation step has yielded significant model improvement as compared to the models from previous studies that are built without deriving new attributes. Moreover, the induction of external data has also improved the performance of the model as outliers were removed.

7. Conclusion

This study has shown that patterns for student attendance exist and can predict whether the student will attend the class. The importance of student data quantity and quality is presented, as well as the methods for cleaning and transforming the data. The creation of a machine learning model should include cross-

validation as one of the key steps, and we devise using multiple algorithms for achieving the best results. When there is a business value to achieve, it's recommended to use a cost-matrix to further adjust the model and increase the business value. The model for predicting student attendance can be used to improve in the area of causing factors and increase the attendance ratio, which will subsequently increase the passing ratio, i.e., the number of students that graduate. Future works can include an increase in data set examples, as well as dimensionality increase by adding attributes for external factors of students' attendance, such as a professor who held the lecture and weather information of the day.

REFERENCES

[1] A. S. N. Kim, S. Shakory, A. Arman, C. Popovic, and L. Park, "Understanding the impact of attendance and participation on academic achievement," 2019. [Online]. Available: https://doi.org/10.1037/stl0000151. [Accessed: 14-Feb-2020].

[2] "Prediction Of Student Performance Using Weka Tool," Vidya Publications. [Online]. Available: http://ijoes.vidyapublications.com/paper/Vol17/02-Vol17.pdf. [Accessed: 26-Nov-2018].

[3] "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance." [Online]. Available: http://www.indjst.org/index.php/indjst/article/view/74555/58051. [Accessed: 26-Nov-2018].

[4] A. B. El-Din Ahmed and Ibrahim Sayed Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method," HR PUB. [Online]. Available: http://www.hrpub.org/download/20140105/WJCAT3-13701793.pdf. [Accessed: 26-Nov-2018].

[5] "Early Prediction of Students Performance using Machine Learning Techniques," Semantics Scholar. [Online]. Available: https://pdfs.semanticscholar.org/6447/4a9172a97cdf5d39c6fdcc21fc0c61fc7df3.pdf. [Accessed: 26-Nov-2018].

[6] "Student Performance Prediction using Machine Learning." [Online]. Available: http://www.ece.uvic.ca/~rexlei86/SPP/otherswork/V4I3-IJERTV4IS030127.pdf. [Accessed: 26-Nov-2018].

[7] "IBM Knowledge Center." [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm. [Accessed: 19-Dec-2018].

[8] International Burch University, "Home," International Burch University. [Online]. Available: https://www.ibu.edu.ba/. [Accessed: 19-Dec-2018].

[9] "Lightning Fast Data Science Platform for Teams | RapidMiner©," RapidMiner, 19-Jan-2016. [Online]. Available: https://rapidminer.com/. [Accessed: 19-Dec-2018].

[10] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 19-Dec-2018].

[11] "[No title]." [Online]. Available: https://www.cs.princeton.edu/~schapire/talks/picasso-minicourse.pdf. [Accessed: 10-Nov-2019].

[12] "[No title]." [Online]. Available: https://www.cs.princeton.edu/~schapire/talks/picasso-minicourse.pdf. [Accessed: 10-Nov-2019].

[13] "3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.21.3 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html. [Accessed: 10-Nov-2019].

[14] L. Egghe, "The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations," Inf. Process. Manag., vol. 44, no. 2, pp. 856–876, Mar. 2008.

[15] "Comparative Analysis of Decision Tree Classification Algorithms" [Online]. Available: https://inpressco.com/wp-content/uploads/2013/03/Paper17334-3371.pdf. [Accessed: 05-July-2020].