ISSN 1991-9468 (Print), 2308-1058 (Online) http://edumag.mrsu.ru

УДК 004.85-057.875

doi: 10.15507/1991-9468.107.026.202202.216-228

Original article

Predicting and Analyzing Student Absenteeism Using Machine Learning Algorithm

L. Mukli✉*, A. Rista*
University Aleksandër Moisiu Durres, Durres, Albania
✉ *linditamukli@uamd.edu.al*

Abstract

Introduction. In a developed society, the state should invest in the education of the younger generation. In less developed countries, Albania included, there are no nation-wide studies to show the factors that affect the lack of students in classrooms. The purpose of this study is to predict, analyze, and evaluate the possible causes of student absenteeism using machine learning algorithms. The attributes taken into account in this study are related to the family, demographic, social, university, and personal aspects according to academic criteria.

Materials and Methods. Student absenteeism covers any student that has not attended class, irrespective of the reason. The data set consists of 26 attributes and 210,000 records corresponding to the teaching hours of 500 students during an academic year at Faculty of Information Technology. The students participating in the survey range from 18 to 25 years of age of both genders. The compilation of the student questionnaire was based on reviewing the literature and analyzing 26 attributes that we categorized into 5 groups included in the questionnaire.

Results. This paper provides knowledge in the analysis and evaluation of factors that lead students to miss lectures using machine learning. It is important to note that this study was conducted on students of this faculty, and as such, the results may not be generalized to all universities. That's why, researchers are encouraged to test the results achieved in this paper on other clusters.

Discussion and Conclusion. The paper provides recommendations based on the findings by offering different problem-solving strategies. The questionnaire used only for 500 Faculty of Information Technology students can be widely applied in any educational institution in the region. However, the results of this study cannot be generalized for the student and youth population of other regions or other countries. This paper provides an original and easily usable questionnaire suitable to various study programs and universities.

Keywords: student absenteeism, family, demographic, social, university, personal aspects, data mining, machine learning

Conflict of interests: The authors declare no conflict of interest.

For citation: Mukli L., Rista A. Predicting and Analyzing Student Absenteeism Using Machine Learning Algorithm. *Integration of Education*. 2022;26(2):216–228. doi: https://doi.org/10.15507/1991-9468.107.026.202202.216-228

Научная статья

Прогнозирование и анализ прогулов студентов с использованием алгоритма машинного обучения

Л. Мукли✉*, А. Риста*
Университет Александра Моисиу Дурреса, г. Дуррес, Албания
✉ *linditamukli@uamd.edu.al*

Аннотация

Введение. Целью данного исследования является прогнозирование, анализ и оценка возможных причин прогулов студентов с использованием алгоритмов машинного обучения. Применяемые алгоритмы эффективны при анализе данных, полученных в результате опросов студентов.

Материалы и методы. Для изучения проблемы было проведено анкетирование, в котором приняли участие 500 студентов в возрасте 18–25 лет. Исследование проводилось на основе количественного метода сбора данных, при котором были получены числовые и стандартизированные значения, что привело к установлению взаимосвязей и тенденций на основе научных материалов. На втором этапе такой алгоритм использовался для анализа полученных результатов.

Результаты исследования. Определены факторы, влияющие на пропуск занятий: занятость студентов, территориальная удаленность от вуза, проблемы со здоровьем. Для решения указанных проблем авторами даны рекомендации руководителям вузов. Результаты исследования подтвердили, что использование методов классификации и конкретных анализируемых алгоритмов служит хорошим инструментом для анализа поставленных задач.

Обсуждение и заключение. Представленная в статье анкета может широко применяться в любом учебном заведении. Однако результаты данного исследования нельзя обобщать на студенческое и молодежное население других регионов или стран. Материалы статьи будут полезны для совершенствования учебно-воспитательного процесса в вузе.

Ключевые слова: прогул студентов, семья, демографический, социальный, вуз, личностные аспекты, анализ данных, машинное обучение

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Мукли Л., Риста А. Прогнозирование и анализ прогулов студентов с использованием алгоритма машинного обучения // Интеграция образования. 2022. Т. 26, № 2. С. 216–228. doi: https://doi.org/10.15507/1991-9468.107.026.202202.216-228

Introduction

Many developed countries use assessments tools and national surveys to assess the quality of teaching, as well as the determination of indicators that affect the motivation of students to achieve the best possible results in their studies.

S. Larabi-Marie-Sainte, R. Jan, A. Al-Matouq, S. Alabduhadi have pointed that student's academic performance can be affected by several factors and one of them is student absences [1]. Marsh, Paulsen, and Richardson suggest that "student ratings demonstrate acceptable psychometric properties which can provide important evidence for educational research" [2–4]. Despite being aware of the harmful effects that absenteeism holds on academic performance, the absenteeism level remains high. J. Childs, R. Lofton have showed that the root causes of chronic absenteeism are complex [5]. M.H. Bahadori, A. Salari, I. Alizadeh, F. Moaddab, L. Rouhi have recommended that educational planners and policymakers pay more attention to the factors mentioned by students as the most important causes of absenteeism [6]. FTI part of UAMD (*Aleksandër Moisiu University of Durrës*, 2021) "the second-largest public academic institution of the Republic of Albania which enrolls about 500 students each year" is experiencing high rates of absences. If not addressed accordingly, the problem of absenteeism may reduce academic performance and have an impact on many social issues. Many factors influence student absenteeism, thus predicting it many a time proves to be very challenging. Özcan, found that "poor academic outcomes, parental involvement, school management, and school schedules, as well as health issues and a lack of social activities, are the main factors influencing student absenteeism" [7]. Additionally, Balkis et al. observed that the major reason that was given by students for non-attendance, related to attitudes towards teacher and school, lack of motivation, level of parents education [8]. Based on the work of I. Dey and Kassarnig et al., it is understood that attendance is amongst the most crucial elements in determining a student's academic performance and success [9; 10]. Wadesango and Machingambi found that auditor condition, socio-economic factors, and relations between students and lecturers are the main factor leading students toward non-attendance [11]. In their work, B.N. Young, W.O. Benka-Coker, Z.D. Weller, S. Oliver, J. W. Schaeffer, S. Magzamen, have shown the connection between student absenteeism and the test scores [12]. Referring to complex factors that influencing high student absenteeism, the usage of Data Mining (DM) and Machine Learning (ML) algorithms is a good method to analyze and predict student absenteeism. Helm et al. refer to ML as "an application of artificial intelligence (AI) that provides to build a model based on training data to make predictions or decisions without being programmed" [13]. DM is a process that extracts and discovers patterns with intelligent methods from a large dataset [14]. Based on the nature of the study and dataset organization, we chose classification methods to evaluate the data. Kantardzic states that "classification techniques are part of predictive methods and categorize a given dataset into classes" [15]. By using these methods, we can predict the unknown values by utilizing the known ones [15]. This dataset consists of 26 attributes and 210,000 records corresponding to the teaching hours of 500 students ranging from 18 to 25 years of age during an academic year at FTI. The attributes analyzed refer to demographic, family, university, and personal factors according to academic performance. This study is most helpful to UAMD and can be easily utilized by other universities in Albania. It is also a valuable tool for all universities in the world, serving to guide the management and provide a sense of understanding of the factors that make students not attend. The rest of the paper is structured as follows: Section 2 – an overview of the data mining classification techniques; Section 3 – the methodology; Section 4 is geared towards the findings of the study; Section 5 relates the discussion to the overall results observed and gives some recommendations.

Literature Review

Classification is one of the methods in data mining that analyzes a large amount of data to predict group membership for data instances [15]. The main goal of classification is to identify the category or class under which new data will fall. This section briefly explains an overview of the analyzed algorithms in this study.

Bayes Net algorithm. Bayes Net falls under the category of probabilistic graphical modeling that is used to compute unknown values by using concepts of probability [16]. It can be represented by using a directed acyclic graph that is used to represent a Bayesian Network, which contains a set of links and nodes. The nodes represent the variables and the links denote the relationships between these variables. A directed acyclic model will value the unknown value of an event occurring based on the conditional probability distribution of each random variable. A Conditional probability table is used to represent the distribution of each variable in the network.

Naive Bayes algorithm. "Naive Bayes is a classification technique based on Bayes' theorem with an assumption of independence among predictors" [17]. This independence quality comes by assuming that the presence of one feature in a class does not impede the presence of any other feature on that same class. This assumption holds true even if these features are dependent on one another. Due to the easy nature of building the Naive Bayes model, it can be effectively used for vast datasets. Utilizing Bayes' theorem and its task of describing the probability of an occurrence generated by previous related conditions, we can understand conditional probability. Conditional probability as explained by Bramer is "the probability of an event happening given that it has some relationship to one or more distinct events" [17].

Logistic Regression algorithm. As displayed in the study by M. Maalouf, "Logistic Regression is a predictive modeling technique that uses one or more independent variables to determine one outcome. The outcome is measured with a dichotomous variable meaning it will have only two possible outcomes" [18]. Logistic Regression can provide probabilities and classify different types of data using continuous and discrete datasets.

Random Forest algorithm. "Random Forest is a supervised algorithm used for both classification and regression" [19]. It is trained with the "bagging" method, which means a blend of learning models. In order to get an accurate prediction, random forest builds and merges multiple decision trees.

Hoeffding Tree algorithm. "The Hoeffding Tree is an incremental decision tree learner for a large dataset assuming that the distribution generating examples does not change over time" [20]. This model utilizes the Hoeffding bound by categorizing into a prescribed precision and estimating statistic values calculated from the necessary observations. The algorithm is responsible for saving the statistics needed for splitting an attribute. It does so once it finds sufficient statistical evidence for an optimal splitting feature, which results in the expansion of the node [20].

Random Tree algorithm. Random Trees algorithms are based on ensemble classification, which means a method that makes predictions by averaging over the predictions of several independent base models [21]. It takes as input the features vector, classifies it with tall trees in the forest, and outputs the largest class. This algorithm can be used as well for regression issues. In this case, the average response of all the trees in the forest is the classifier response.

J48 algorithm. Based on Mathuria's work we determine that "the J48 classifier is an implementation of the C4.5 decision tree algorithm" [22]. First, the attribute values of the dataset are used to build a decision tree that serves to classify the instances. When the classifier is able to recognize the attribute that accurately categorizes numerous instances, it means that it encountered the training set. The branch of this decision tree is then terminated by assigning probable feature values to it [22].

REPTTree algorithm. REP Tree is based on C4. 5 algorithms described by Mohamed et al. that "outcome a decision tree using information gain and prunes it using reduced-error pruning (with back-fitting)" [23].

Materials and Methods

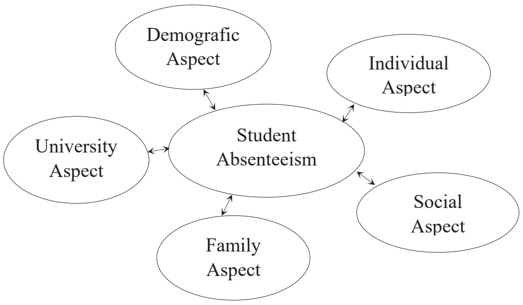
This section identifies the methods to fulfill the goal of the study by providing hypothesis, sampling, data collection, analysis, and assessment criteria. The study is composed of a quantitative method of data collection, where numerical and standardized data were collected, leading to generation of relationships and trends based on academics. These relationships were established using machine learning algorithms. First step is to determine the most effective algorithm for processing the data obtained from the questionnaires. Second step, this algorithm will be used to analyze the data and generate the results needed to test our hypotheses.

Target Population and sampling. The study population is targeted on students of FTI, from the bachelor and master categories. In total, 500 students have been selected from five study programs that the faculty offers on a percentage balance. Respondents were randomly selected from each study program, belonging to both genders and age groups ranging from 18 to 25 years old. Each respondent has equal rights and conditions. All respondents were informed of their participation in the study.

Data Collection Methods. Both primary and secondary data have been considered as useful for the study to fulfill the aim of our study:

– Primary Data. The primary data were collected through a survey questionnaire for the consolidation of our aims and objectives.

– Questionnaire Design. The data collection tool for the research was developed based on literature from scholars on the reasons that lead students to be absent from class. Figure 1, shows the aspects that have been analyzed in this study related to the factors that directly affect student absenteeism according to academics.



F i g. 1. The aspect of student absenteeism

The design of survey questionnaires was divided into five parts. The first part assesses the attributes of the demographic data such as gender (GE) and residence (RS). The second part assesses the attributes related to the family aspect such as monthly family income (IFM), parent's educational level (LPE), a family member with a chronic condition (CCF), legal issues in family (LF), a family member suffering from alcoholism (AMF), parent's marital status (MSP), and if family takes interest in the student's achievements (AIF). The third part assesses the attributes related to university aspects. Categories as shown in the questionnaire are: enjoying field of study (SPE), satisfied with auditorium conditions (CAS), dedicated lecturers (LD), campus safety (SC), satisfied with academic leaders (LAS), and study program (PS). The fourth part relates to individual aspects divided into the following categories: employment status (SE), who do you live with (LEFO), distance from the university (UD), chronic health conditions (CHC), and motivation to attend class (CAM). The last part investigates social aspect's impact on absenteeism, where the following attributes are assessed: frequent bars and night clubs (CBF), satisfied with student activities (ASS), relationship with other students (SR), relationship with lecturers (LR), friends facing issues with alcohol, drugs or legal issues (LADF), and sufficient study space (SS).

– Hypotheses. To construct our hypotheses, we focused on one of the 5 aspects presented and analyzed in our work: the family aspect. In Albania as well as in many other countries, family is considered the foundation of a healthy individual and a stable society as a whole. Based on this elevated role that family holds in the Albanian society, we constructed the following hypotheses:

H1: The attribute that has the greatest effect on student absenteeism is part of the family aspect category.

H2: Students with higher absenteeism rates come from lower-income families.

– Secondary data. Secondary data was used to understand the background of the distribution of absences during the 2019–2020 academic year. Secondary data were collected from the FTI database. Lecturers record the absences of each student in every lesson.

Dataset Design. The dataset is built from 210,000 records and 26 attributes. At UAMD during an academic year, a student attends 10 courses, over the span of 28 academic weeks. Every student attends 15 academic hours weekly, 3 hours per course. The number of records (210,000) corresponds to the number of academic hours of 500 students taken as a sample during an academic year. While the number of attributes (26) corresponds to the number of questions raised in the questionnaire design.

Assessment Criteria. Analysis and evaluation of data are done in the Weka tool [24]. The first step is to evaluate the different algorithms and to analyze the data. The algorithm that performs best and is simpler in interpreting the data is selected. D. M. Powers comes to the same conclusion as S. Arlot, A. Celisse, by stating that "the evaluation and comparison of algorithms are done in terms of accuracy, recall, and precision" [25]. Arlot & Celisse defines cross-validation as "a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data" [26]. We conducted the experiments similarly to the method described in Arlot & Celisse, separating the dataset in 10 subsets or folds. However, we utilized different subset values. We modeled the DM algorithm by using 1 subset as a test set and 9 subsets as a training set. The test set subset processes the performance of the algorithm. We attained an average by repeating this process 10 times. Similar to Arlot & Celisse we based our selection of k = 10 on the dataset size.

Choosing an Algorithm. In this phase, we analyzed the data obtained from the preprocessing phase by comparing different algorithms according to accuracy, recall and F-measure, classifying them as secondary findings of this study, and we chose the algorithm with the best performance for interpreting the data, which we have classified as the primary findings of the study.

In our study, we analyzed different algorithms, and their performance measurements were compared to one other. The algorithms are: BayesNet, Naive Bayes, Logistic, HoeffdingTree, Decision Stup, J48, Random Tree, Random Forest, REPTTree, OneR, LogitBoost and MultiClasClassifier. Table 1 shows the results of algorithms.

As can be seen from Table 1, all algorithms present acceptable results, and they display a high accuracy and classification of more than 80% of instances.

To analyze the main findings of this study, the Bayes Net algorithm is chosen, due to the simplicity of interpretation and the high accuracy that it shows. Figure 2, shows the visual graph of the Bayes Net algorithm for the Residence attribute. This attribute was chosen randomly as an example to explain how Bayes Net Algorithm generates results. The algorithm classifies the frequented hours as well as the non-frequented hours according to all the attributes taken in the analysis.

For the Residence attribute the average number of absences for students living in the village is calculated:

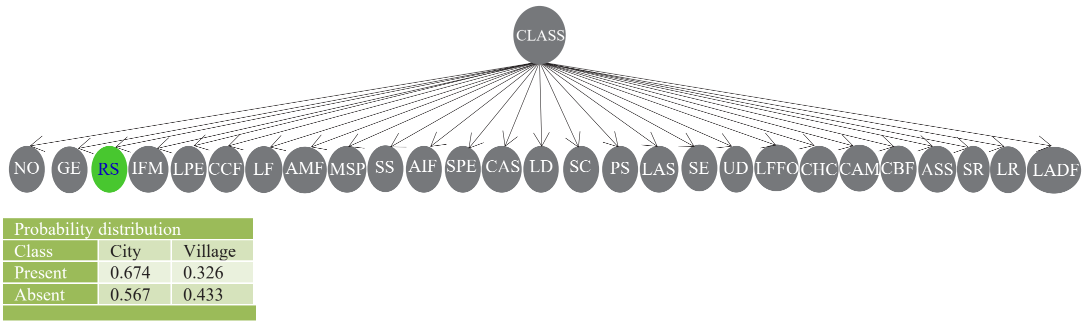
$$Nrav = 0.433 \times Nt/Nrv$$

Where Nrav represents the average number of absences committed by students living in villages throughout the academic year; Nt represents the total number of absences throughout the academic year; *Nrv* represents the number of students living in the village.

Table 1. **Results of algorithms**

Algorithms	Correctly Classified Instances, %	Incorrectly Classified Instances, %	Recall	Precision	F-measure
Bayes Net	97.50	2.50	0.975	0.975	0.975
Naive Bayes	83.98	16.02	0.840	0.834	0.836
Logistic	91.53	8.47	0.915	0.914	0.914
HoeffdingTree	95.66	4.34	0.957	0.957	0.957
J48	97.49	2.51	0.975	0.975	0.975
Random Forest	97.43	2.57	0.974	0.974	0.974
Random Tree	97.41	2.59	0.974	0.974	0.974
REPTTree	97.49	2.51	0.975	0.975	0.975

Source: Authors' calculation.



Probability distribution		
Class	City	Village
Present	0.674	0.326
Absent	0.567	0.433

Fig. 2. Visual graph of Bayes Net algorithm

Results

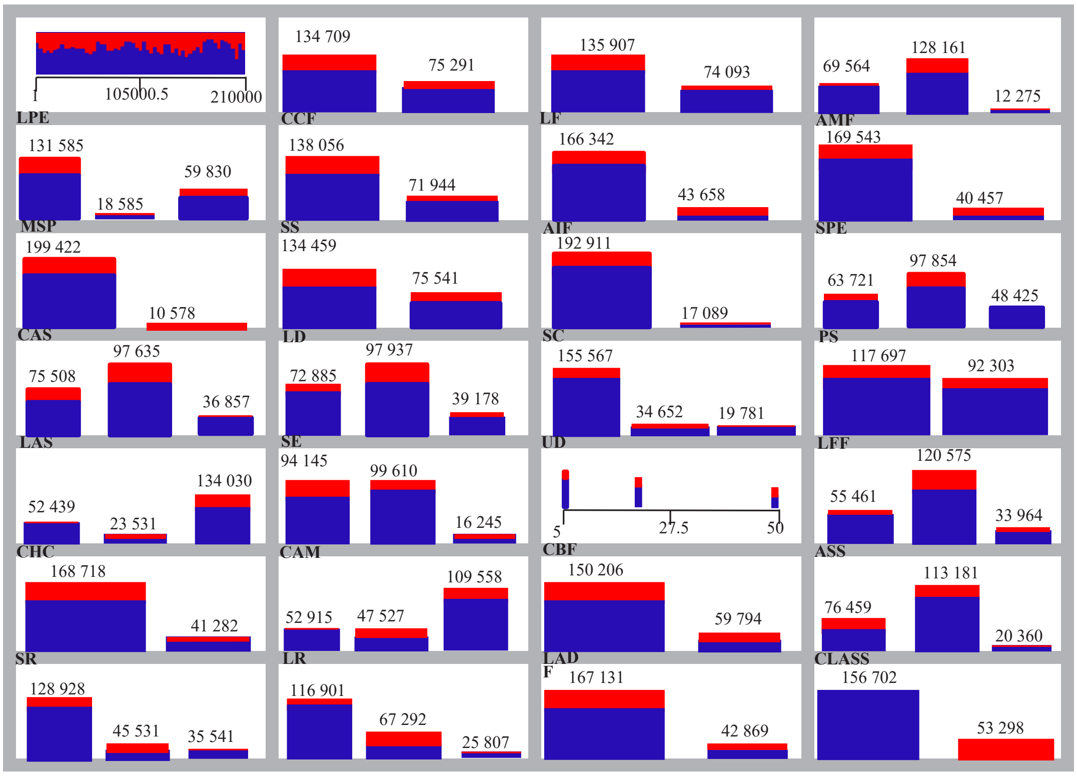
The first step in the data mining process is data preprocessing, which involves cleaning, preparation, normalization, and transformation of data. Based on the fact that the database was built on the basis of primary data (data obtained from the questionnaires), as well as secondary data (data obtained from the FTI database) it was not necessary to perform their cleaning, normalization, or transformation. Figure 3, shows the distribution of absences according to all the attributes considered. The blue color indicates the number of teaching hours that are attended by students during an academic year. The red color indicates the number of hours missed by students in this academic year. Visually, it shows a significant percentage of absences committed by students for the academic year we analyzed.

In total, we see that we have 53,298 unattended hours out of 210,000 total annual teaching hours of 500 students taken as samples. Translated into percentages, the total number of unattended hours by all students taken as a sample is 25.38%. As can be seen from the presentation of data during preprocessing phase, the shortcomings constitute a significant percentage.

Using the Bayes Net algorithm, we are tabulating below the results for the data obtained from the questionnaires according to the attributes: Demographics, Family, University, Individual, and Social Aspect.

Referring to the attributes categorized in terms of demographics, we see that 64.8% of the number of the sample students live in urban areas. Additionally, 64.2% of the surveyed students were male and 35.8% were female. Male students tend to be absent more compared to females by a margin of 2.7%. On the other hand, students living in the countryside tend to record more absences compared to students living in the city by a margin of 9% (Table 2).

Referring to the attributes classified in terms of family, we found that 61% of the samples are expressed with low family income and 62.6% of students have parents with secondary education.



F i g. 3. Visualization results from the preprocessing phase

Table 2. **Distribution of questionnaire answers (QA) and average absences according to demographic aspect,** %

Attributes	Description	Distribution of QA	Average absences
Gender (GE)	Male	64.2	26.3
	Female	35.8	23.6
Residence (RS)	Village	35.2	31.2
	City	64.8	22.2

Source: Authors' calculation.

Table 3. **Distribution of QA and average absences according to family aspect,** %

Attributes	Description	Distribution of QA	Average absences
Monthly family income (MFI)	Low	61.0	31.2
	Medium	33.2	15.8
	High	5.8	19.3
Parent's educational level (LPE)	Middle School	8.8	43.8
	High School	62.6	21.9
	Bachelor's	28.6	3.9
Family member with a chronic condition (CCF)	Yes	34.2	42.1
	No	65.8	16.7
Legal issues in family (LF)	Yes	20.8	72.8
	No	79.2	12.9
Family member suffering from alcoholism (AMF)	Yes	19.2	73.4
	No	80.8	13.9
Parent's marital status (MSP)	Married	95.0	21.8
	Divorced	5.0	92.8
Family takes interest in your achievements (AIF)	Yes	91.8	24.1
	No	8.2	39.3

Source: Authors' calculation.

Students who have at least one family member with legal and alcohol problems tend to have a very large number of absences, leading to failure of the academic year. We also see that a very important attribute is the marital status of the parents. Students who have divorced parents (despite being only 5%), miss on average about 92.8% of the hours, by which they miss the full academic year. Referring to other attributes categorized in terms of family, it is worth noting that students with parents who have low education, low monthly income, and do not express interest in the achievements of their children, tend to make more absences from school (Table 3).

Based on the results displayed in the table above hypothesis H2: "Students with higher absenteeism rates come from lower-income families", is false. The attribute that affects student absenteeism most is parent's marital status (divorced, 92.8%) (Table 4).

According to the results obtained, we see that only 23% of students really like the field of study they are pursuing.

Table 4. **Distribution of QA and average absences according to university aspect, %**

Attributes	Description	Distribution of QA	Average absences
Enjoying field of study (SPE)	Not at all	46.6	35.2
	Adequate	30.4	26.5
	Very much	23.0	3.8
Satisfied with auditorium conditions (CAS)	Not at all	38.6	28.4
	Adequate	35.8	28.5
	Very much	17.6	11.1
Dedicated lecturers (LD)	Not at all	18.6	33.8
	Adequate	55.5	30.2
	Very much	34.8	14.4
Campus safety (SC)	Not at all	16.6	63.3
	Adequate	74.2	18.1
	Very much	9.2	11.3
Satisfied with academic leaders (LAS)	Not at all	9.4	74.3
	Adequate	65.6	24.3
	Very much	25.0	9.9
Study program (PS)	Bachelor	56.0	23.6
	Master	44.0	27.6

Source: Authors' calculation.

Additionally, 17.6% of students are satisfied with the conditions of the auditorium, while the rest expressed insufficient satisfaction or none at all. Students who have expressed dissatisfaction with university leaders as well as those who do not feel safe on campus tend to have a relatively large number of absences. We also emphasize that students who do not like their field of study, are not satisfied with the conditions of the auditors, and are dissatisfied with the dedication of the lecturers, have a tendency to perform more absences compared to students who have expressed otherwise. Regarding the study program, students who attend master studies have a relatively higher percentage of absences compared to students attending bachelor studies (Table 5).

Referring to the attributes categorized in the individual aspect, only 25.2% of students are very motivated to go to university, which presents a very concerning factor. Students who have expressed that they work full-time, suffer from a chronic illness, and those who do not feel motivated to go to university, tend to have a very large number of absences. Regarding the distance from the university, we see that students living at a relatively long distance (20–50 km) have a higher percentage of absences compared to students living closer to the university. Also, referring to the roommate attribute, we see that students who live alone tend to commit more absences (Table 6).

From the results, it is evident that for student-student relations and the student-lecturer relations, a small percentage of students have expressed that they have very good ones. It is noteworthy that students who have expressed their association with friends who consume alcohol, drugs, or have legal problems, and students who do not have good relations with other students and lecturers, have a large percentage of absences. Also, we see that students who regularly frequent bars and nightclubs have a much higher percentage of absences compared to students who do not.

Table 5. **Distribution of QA and average absences according to individual aspect, %**

Attributes	Description	Distribution of QA	Average absences
Employment status (SE)	Part-time	44.8	28.9
	Full-time	7.8	59.5
	No	47.4	16.38
Who do you live with (LEFO)	Alone	16.2	34.7
	Family	26.4	9.9
	Friends	57.4	29.8
Distance from the university (UD)	0–5 km	43.6	18.9
	5–20 km	30.8	20.6
	20–50 km	25.6	42.2
Chronic health conditions (CHC)	Yes	19.6	66.9
	No	80.4	15.2
Motivation to attend class (CAM)	Not at all	22.6	64.4
	Adequate	52.2	15.2
	Very much	25.2	11.3

Source: Authors' calculation.

Table 6. **Distribution of QA and average absences according to social aspect, %**

Attributes	Description	Distribution of QA	Average absences
Frequent bars, nightclubs (CBF)	Yes	28.4	56.5
	No	71.6	13.1
Satisfied with student activities (ASS)	Not at all	36.4	44.9
	Adequate	54.0	15.3
	Very much	9.6	5.3
Relationship with other students (SR)	Good	61.4	13.4
	Adequate	21.6	72.0
	Very good	17.0	8.1
Relationship with lecturers (LR)	Good	55.6	10.8
	Adequate	32.0	56.8
	Very good	12.4	9.4
Friends facing issues with alcohol, drugs, or legal issues (LADF)	Yes	20.4	60.6
	No	78.6	15.1
Sufficient study space (SS)	Yes	64.0	15.2
	No	36.0	43.5

Source: Authors' calculation.

Finally, we see that students who are dissatisfied with student activities as well as do not have the necessary study space, tend to be absent more compared to students who have expressed otherwise.

After analyzing the impact of all 5 relevant aspects and their pertaining attributes, it is evident that our first hypothesis H1: "The attribute that has the greatest effect on student absenteeism is part of the family aspect category", is true. The family aspect constitutes the attribute that has the highest effect on student absenteeism (parents marital status, divorced, 92.8%).

Discussion and Conclusion

This study focused on analyzing and evaluating the factors that lead students to drop out of the class through machine learning algorithms. Based on the findings of the study, conducted in FTI, this section offers recommendations to solve such a problem. Referring to the findings of the study, it is noted that the use of classification methods and specific analyzed algorithms serve as a very good tool to analyze the objectives of our study, but also beyond. Based on the literature studies, we selected to analyze 26 attributes which we categorized into 5 groups. In the first study group that includes demographic aspects, we recommend that a more detailed study should be conducted in the future for these attributes to determine what are the real factors driving students who live in rural areas to have more absences compared to students who live in the city. Similarly, deeper social investigation beyond our scope is required to analyze why men tend to be absent more compared to women.

In the second group that includes the family aspect, according to the study, students who have family problems, low income, divorced, or sick parents tend to make more absences than the other category of students. As mentioned earlier when studying the validity of our hypotheses, the family aspect is crucial in every society. When observing individual attributes, lower income of the family proved to have a lesser negative effect on student absenteeism than parent's marital status. This may result from the lack of societal and governmental resources needed to help parents and children navigate the financial and emotional outcomes of a divorce. Despite the government efforts to help students, there is more to be done. The financial assistance of students under the family aspect category should be considered a priority. Higher education institutions should set up student counseling offices with specialists to help students cope with family problems as easily as possible. Also, in order to minimize the shortcomings, it is necessary to work with the parents so that they are very vigilant towards the achievements of their children in the university, finding different forms of simulation to attend regularly and to achieve high results.

In the third group categorized in this study, we noticed that students who tend to make more absences are those who are not satisfied with the conditions of auditoriums, university leaders, the commitment of lecturers and do not feel safe on campus. It is the duty of the university to improve the conditions of the auditorium with elements that help with the safety of the students and to facilitate the interaction of the students with the lecturers. The university should make constant investments in improving the conditions of the auditoriums by providing more study space, more comfortable facilities, and better technology. Additionally, the university should support and encourage the research work of lecturers so that they can keep up with the latest developments in their field of teaching. This way, the study program for the students becomes not only more enjoyable, but also closer to the practical knowledge that the student needs for his employment.

Referring to the attributes categorized in the individual aspect, we noticed that the high number of student absences is influenced by factors such as: student employment; distance from university; school motivation, and student health problems. To solve these problems, we recommend that university leaders com e to terms with businesses to enable students a flexible schedule so that they can attend classes and employment. Regarding University of Durres, we recommend that university leaders cooperate with the government to provide funding to build affordable dormitories for students at or near the university. This solution also addresses the absenteeism trend in the residence attribute discussed above. In addition, we recommend the academic staff of the university to be more flexible and to be closer to the students, by motivating, and encouraging students to attend the lesson.

Finally, we offer some recommendations about the attributes classified in the social aspect group. Alcohol and drug abuse is a disturbing phenomenon to every society. Even in our study, it was observed that students who frequent bars, nightclubs, or associate with people who consume alcohol and drugs tend to be more absent compared to students who have expressed otherwise. The academic staff of our university constantly works in this direction, but we still need continuous work and an appropriate policy to limit and control the businesses that students attend. Additionally, having student counseling offices near the university with specialized staff can help address this issue. In order to reduce the absences that students make because they do not have good relations with other students, with lecturers, and are dissatisfied with student activities, the university should create the opportunity to organize as many activities and set up as many student clubs as possible, because students often are their main actors and feel themselves quite involved in them.

REFERENCES

1. Larabi-Marie-Sainte S., Jan R., Al-Matouq A., Alabduhadi S. The Impact of Timetable on Student's Absences and Performance. *Plos one*. 2021;16(6):e0253256. doi: https://doi.org/10.1371/journal.pone.0253256

2. Marsh H.W. Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research. *International Journal of Educational Research*. 1987;11(3):253–388. doi: https://doi.org/10.1016/0883-0355(87)90001-2

3. Paulsen M.B. Evaluating Teaching Performance. New Directions for Institutional Research. *Special Issue: Evaluating Faculty Performance*. 2002;(114):5–18. doi: https://doi.org/10.1002/ir.42

4. Richardson J.T. Instruments for Obtaining Student Feedback: A Review of the Literature. *Assessment & Evaluation in Higher Education*. 2005;30(4):387–415. doi: https://doi.org/10.1080/02602930500099193

5. Childs J., Lofton R. Masking Attendance: How Education Policy Distracts from the Wicked Problem (s) of Chronic Absenteeism. *Educational Policy.* 2021;35(2):213–234. doi: https://doi.org/10.1177/0895904820986771

6. Bahadori M.H., Salari A., Alizadeh I., Moaddab F., Rouhi Balasi L., et al. The Root Causes of Absenteeism in Medical Students: Challenges and Strategies Ahead. *Educational Research in Medical Sciences*. 2020;9(2):e107120. doi: http://dx.doi.org/10.5812/erms.107120

7. Özcan M. Student Absenteeism in High Schools: Factors to Consider. *Journal of Psychologists and Counsellors in Schools*. 2020. p. 1–17. doi: https://doi.org/10.1017/jgc.2020.22

8. Balkis M., Arslan G., Duru E. The School Absenteeism among High School Students: Contributing Factors. *Educational Sciences: Theory and Practice*. 2016;16(6):1819–1831. doi: https://doi.org/10.12738/estp.2016.6.0125

9. Dey I. Class Attendance and Academic Performance: A Subgroup Analysis. *International Review of Economics Education*. 2018;28:29–40. doi: https://doi.org/10.1016/j.iree.2018.03.003

10. Kassarnig V., Bjerre-Nielsen A., Mones E., Lehmann S., Lassen D.D. Class Attendance, Peer Similarity, and Academic Performance in a Large Field Study. *PloS ONE*. 2017;12(11):0187078. doi: https://doi.org/10.1371/journal.pone.0187078

11. Wadesango N., Machingambi S. Causes and Structural Effects of Student Absenteeism: A Case Study of Three South African Universities. *Journal of Social Sciences*. 2011;26(2):89–97. doi: https://doi.org/10.1080/09718923.2011.11892885

12. Young B.N., Benka-Coker W.O., Weller Z.D., Oliver S., Schaeffer J.W., Magzamen S. How Does Absenteeism Impact the Link between School's Indoor Environmental Quality and Student Performance? *Building and Environment*. 2021;203:108053. doi: https://doi.org/10.1016/j.buildenv.2021.108053

13. Helm J.M., Swiergosz A.M., Haeberle H.S., Karnuta J.M., Schaffer J.L., Krebs V.E., Ramkumar P.N. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Current Reviews in Musculoskeletal Medicine*. 2020;13(1):69–76. doi: https://doi.org/10.1007/s12178-020-09600-8

14. Schuh G., Reinhart G., Prote J.P., Sauermann F., Horsthofer J., Oppolzer F., Knoll D. Data Mining Definitions and Applications for the Management of Production Complexity. *Procedia CIRP*. 2019;81:874–879. doi: https://doi.org/10.1016/j.procir.2019.03.217

15. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. 3rd ed. John Wiley & Sons; 2019. doi: https://doi.org/10.1002/9781119516057

16. Niedermayer D. An Introduction to Bayesian Networks and Their Contemporary Applications. In: Holmes D.E., Jain L.C. (eds.) Innovations in Bayesian Networks. Studies in Computational Intelligence. Springer, Berlin, Heidelberg; 2008. Vol. 156. p. 117–130. doi: https://doi.org/10.1007/978-3-540-85066-3_5

17. Bramer M. Principles of Data Mining. 3rd ed. London; 2016. doi: https://doi.org/10.1007/978-1-4471-7307-6

18. Maalouf M. Logistic Regression in Data Analysis: An Overview. *International Journal of Data Analysis Techniques and Strategies*. 2011;3(3):281–299. doi: https://doi.org/10.1504/IJDATS.2011.041335

19. Biau G., Scornet E. Rejoinder on: A Random Forest Guided Tour. *TEST.* 2016;25(2):264–268. doi: https://doi.org/10.1007/s11749-016-0488-0

20. Pfahringer B., Holmes G., Kirkby R. New Options for Hoeffding Trees. In: Orgun M.A., Thornton J. (eds.) AI 2007: Advances in Artificial Intelligence. AI 2007. Lecture Notes in Computer Science. Vol. 4830. Berlin, Heidelberg: Springer; 2007. doi: https://doi.org/10.1007/978-3-540-76928-6_11

21. Kalmegh S. Analysis of Weka Data Mining Algorithm Reptree, Simple Cart and Randomtree for Classification of Indian News. *International Journal of Innovative Science, Engineering & Technology*. 2015;2(2):438–446. Available at: http://ijiset.com/vol2/v2s2/IJISET_V2_I2_63.pdf (accessed 21.12.2021).

22. Mathuria M. Decision Tree Analysis on J48 Algorithm for Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2013;3(6). Available at: https://www.academia.edu/4375403/Decision_Tree_Analysis_on_J48_Algorithm_for_Data_Mining (accessed 21.12.2021).

23. Mohamed W.N.H.W., Salleh M.N.M., Omar A.H. A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms. In: 2012 IEEE International Conference on Control System, Computing and Engineering. 2012. p. 392–397. doi: https://doi.org/10.1109/ICCSCE.2012.6487177

24. Srivastava S. Weka: A Tool for Data Preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. *International Journal of Computer Applications*. 2014;88(10):26–29. Available at: https://research.ijcaonline.org/volume88/number10/pxc3893809.pdf (accessed 21.12.2021).

25. Powers D.M. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *arXiv preprint arXiv*. 2020;2010:16061. doi: https://doi.org/10.48550/arXiv.2010.16061

26. Arlot S., Celisse A. A Survey of Cross-Validation Procedures for Model Selection. *Statistics Surveys*. 2010;4:40–79. doi: https://doi.org/10.1214/09-SS054

Submitted 25.08.2021; approved after reviewing 10.01.2022; accepted for publication 17.01.2022.
Поступила 25.08.2021; одобрена после рецензирования 10.01.2022; принята к публикации 17.01.2022.

About the authors:

Lindita Mukli, Dean of Faculty of Information Technology, Lecturer at Mathematics Department, University Aleksandër Moisiu Durres (L1 Rruga Curila, Durres 2001, Albania), Ph.D., Associated Professor, **ORCID: https://orcid.org/0000-0003-4472-0053**, linditamukli@uamd.edu.al

Amarildo Rista, Lecturer at Information Technology Department, University Aleksandër Moisiu Durres (L1 Rruga Curila, Durres 2001, Albania), **ORCID: https://orcid.org/0000-0001-9471-4749**, amrildorista@gmail.com

Contribution of the authors:

L. Mukli – contributed to the conception of absenteeism measurement for the faculty and the overall perception of the survey results.

A. Rista – contributed to the statistical analysis and dataset construction.

All authors have read and approved the final manuscript.

Об авторах:

Мукли Линдита, декан факультета информационных технологий, преподаватель кафедры математики Университета Александра Моисиу Дурреса (2001, Албания, г. Дуррес, Рруга Куррила, L. 1), доктор философии, доцент**, ORCID: https://orcid.org/0000-0003-4472-0053**, linditamukli@uamd.edu.al

Риста Амарилдо, преподаватель кафедры информационных технологий Университета Александра Моисиу (2001, Албания, г. Дуррес, Рруга Куррила, L. 1), **ORCID: https://orcid.org/0000-0001-9471-4749**, amrildorista@gmail.com

Заявленный вклад авторов:

Л. Мукли – разработка концепции измерения невыходов на работу для профессорско-преподавательского состава и общего восприятия результатов опроса.

А. Риста – статистический анализ; построение набора данных.

Все авторы прочитали и одобрили окончательный вариант рукописи.