Using Unsupervised Learning for Mining Behavioural Patterns from Data. A Case Study for the Baccalaureate Exam in Romania

Mariana-Ioana MAIER*, Gabriela CZIBULA, Lavinia-Ruth DELEAN

Babeș-Bolyai University, 1 M. Kogălniceanu Street, 400084, Cluj-Napoca, Romania
mariana.maier@ubbcluj.ro (**Corresponding author*), gabriela.czibula@ubbcluj.ro, lavinia.sugar@stud.ubbcluj.ro

Abstract: *Behavioural data mining* is an interesting paradigm in the field of knowledge discovery focused on uncovering meaningful patterns that describe behavioural characteristics. With the broader goal of improving decision-making in various areas, behaviour mining has proven to be useful in several application domains ranging from information systems to social science studies. This paper addresses the topic of behaviour mining in the field of educational data mining and analyses, as a proof of concept, the application of unsupervised learning-based models (self-organising maps and association rules) for identifying certain patterns in the behaviour of the high school students from the Real Sciences specializations when choosing the optional exam item for the Romanian baccalaureate. The experiments conducted for real data sets collected from Romanian high school students have shown that features like class specialization, gender, motivational patterns, or the average score obtained at a certain school subject influence the students' choosing or rejecting that subject as a baccalaureate exam item. The uncovered behavioural patterns are useful in outlining the profile of the present-day high school student and may be integrated in a recommender system for assisting students and teachers in the educational processes.

Keywords: Data mining, Behaviour mining, Unsupervised learning, Students' profile, Association rules, Self-organising map.

1. Introduction

Data mining (DM) techniques are being widely used nowadays for extracting relevant knowledge from data and for improving decision-making processes in various domains. Unsupervised machine learning models offer a wide range of methods for uncovering hidden patterns in data from various practical domains, such as bioinformatics, medicine, software engineering, educational data mining. In order to produce more tangible results, the mining process should be more goal-oriented (Chen, 2006) and this can be achieved by incorporating the semantics of the data into the mining task. DM already deals with this challenge, due to the growing interest in mining complex data like graph mining, text mining, academic data mining, etc (Chen, 2006).

The challenge of mining complex data and extracting behavioural characteristics underlying the raw data leads to a new paradigm in the field of knowledge discovery, namely that of *behaviour mining* (Chen, 2006). Behavioural patterns mined from various type of data lead to a better comprehension of the relationships between the entities from a data set, by uncovering those entities exhibiting a similar behaviour, even though they have different characteristics (Maiti & Subramanyam, 2019). Behavioural pattern mining has already been applied in the DM literature for various real-world scenarios such as detecting behavioural patterns of smartphone users (Sarker et al., 2019), mining behavioural patterns from spatial data (Maiti & Subramanyam, 2019) or mining malicious behavioural patterns (Seifi & Parsa, 2018).

As an application domain for highlighting the usefulness of employing *unsupervised learning* (UL) models for mining behavioural patterns from data, the present study is focused on *educational data mining* (EDM). The target of EDM is to mine data collected from educational environments and is of great interest in the data mining field. The case study targeted in this paper is focused on finding relevant patterns with regard to choosing or rejecting a subject at the national baccalaureate exam.

The baccalaureate is a national exam in many countries around the world which can be passed by high school graduates. It represents an important moment in the life of every person because the score obtained in this exam is relevant in the admission process to many faculties in the country and abroad. In the Romanian educational system (Romanian Ministry of National Education, 2022), the baccalaureate includes five modules, each of them being graded separately. The first four modules (A-D) are Competency Exams. They are marked with levels of competencies and their scores don't have any influence on the final score.

https://doi.org/10.24846/v32i2y202307

The fifth module (E) is the written examination, and the marks obtained in its tests are averaged to get the final score.

The module E consists of the following written exams: Romanian Language and Literature (*Ea*), Maternal Language and Literature (*Eb*) only for those studying in their mother tongue, a compulsory subject specific to the class profile (*Ec*), and an optional exam item (*Ed*) specific to the profile of the respective class.

The present study is dedicated to high school students from Real Sciences classes and their attitude and preferences with regard to choosing a certain subject for the *Ed* exam. The other specializations will be addressed in additional studies in the future. For the *Ed* exam, students from Real Sciences classes can choose between the following subjects: Physics (*P*), Chemistry (*C*), Biology (*B*), and Computer Science (*CS*).

In this paper, a study is conducted with the aim of highlighting the relevance of UL-based behaviour mining applied to educational data. Statistical and UL-based analyses are applied with the goal of identifying some patterns and behavioural characteristics of Romanian high school students from Real Sciences classes in relation to choosing a subject for the *Ed* test. *Association rule* (AR) mining and *self-organizing maps* (SOMs) are used as UL models in the considered educational-related behaviour mining case study. SOMs were selected as UL tools for visualizing clusters of students with a similar behaviour in what concerns the choice of the subject for the baccalaureate exam. The experimental evaluation conducted on real data sets collected from Romanian high school students and the AR and SOM models are able to detect behavioural patterns of high school students in selecting a subject for the baccalaureate. Additionally, the results of the UL-based analysis were proven to be highly correlated with those of the statistical analysis when studying the students' profile. Based on the available information, there hasn't been a study like the present one published so far, in the EDM literature.

To summarize, the present paper aims to answer the following research questions: **RQ1:** *Which are the preferences of students from Real Sciences classes when choosing the subject for the Ed exam and how are the students' features correlated with their choices?* **RQ2:** *To what extent are ARs able to identify some subsets of features influencing students' choices for the Ed exam?* **RQ3:** *To what extent are SOMs able to detect similar behavioural patterns for students, when using different features for characterizing them?*

The remainder of this paper is structured as follows. Section 2 presents the related work, while Section 3 sets forth the methodology for the conducted study. The analysed case study is described in Section 4. Section 5 discusses the results obtained and Section 6 summarizes the conclusions of this paper and proposes several future research directions.

2. Related Work

Students who choose science subjects for the baccalaureate exam perceive these fields as requiring certain traits and a certain personality. To attract students to subjects in the field of exact sciences, effective methods must be found to overcome the barriers for students with a less science-oriented identity (Taconis & Kessels, 2009).

Since students can choose a subject at an exam, they have the opportunity to choose that subject in a strategic way, namely the subject that offers them the highest chance to obtain a higher grade. This enables less prepared students to achieve an optimal result, but it also disadvantages students who want to achieve good results in more competitive subjects. This aspect leads to a situation where the subjects for which higher marks are perceived as being easier to obtain, are increasingly chosen by students, the rest representing an option only for those very well trained in that field (Kupiainen, et al., 2016).

Relative advantage is an important factor in predicting the subjects chosen by students for study and assessment. Thus, the fact that a school achieves a level of performance in a certain subject of study also influences the choice of students regarding the subject they will study (Davies et al., 2009).

(Carter, 2006) aimed to identify the reasons why students from USA who seem to have a natural

ability for *CS* don't choose to continue their studies for pursuing a career in this domain. The author used a questionnaire for the study. The main conclusion of the research is that students have an incorrect or no understanding of what the field of *CS* entails, and this is the reason for their decision to avoid the deepening of *CS* study.

(Kumar & Chadha, 2012) investigated the existence of *association rules* (ARs) in data collected from the students' assessment. Their mined ARs highlight a number of variables, including student engagement, curricula, instructional approaches, and evaluation processes that may have an impact on students who have not achieved a suitable degree of success at the post-graduate level.

A machine learning algorithm framework had been introduced so far to predict the students' performance at the baccalaureate exam in Morocco (Qazdar et al., 2019).

To analyse students` performance, and to what extent one could uncover some patterns in educational data sets from the academic level, unsupervised machine learning methods have been used, like: *self organising maps* (SOMs) (Oneț-Marian et al., 2021), *autoencoders*, or *t-distributed stochastic neighbor embedding* (Maier et al., 2021).

One can note that for the baccalaureate exam in Romania no studies have been carried out to find the students' preferences or some patterns for these preferences at the *Ed* exam. This paper aims to bring about analysis methods together with UL methods, with the target to uncover some patterns that should influence the students' choice of a subject at the baccalaureate exam.

3. Methodology

The proposed methodology for carrying out the current analysis is presented in this section.

3.1 Research Instrument

A questionnaire was used for collecting information about the Romanian high school students' and their preferences related to the *Ed* exam. The questionnaire contains 14 items, of three different types: *multiple choice* (MC) - items with one possible answer, *Likert scale* (LS), and *open answer* (OA). The 14 items are presented below.

Item I1 (MC) is related to the class specialization of the respondents. For Real Sciences classes, there are three possible specializations: *Mathematics-CS intensive CS* (I), *Mathematics-CS* (M), and *Natural Sciences* (S). The specializations (I) and (M) differ by the number of CS hours/week. The curriculum for specialization (S) differs from that for (M) only from the third year of high school by the number of *P*, *C*, *B* hours/week; *CS* subject is missing starting with the third year of high school at specialization S.

Item I2 (MC) is dedicated to the grade of the surveyed students. In Romania, high school education lasts four years for the specializations mentioned in this paper, so the 9th grade is the first year of high school (students with the age of 14-15 years), and the 12th grade is the last year of high school (the age of 18-19 years).

Item I3 (MC) is a demographic one, regarding the gender of the respondents: *male, female*, or *not specified*. There is no gender imbalance in Romanian schools between students, so the classes are heterogeneous from this point of view. This paper aims to verify if there are gender differences related to the students` preference for a subject.

Item I4 (MC) is the first item related to the respondents' preference for an *Ed* exam subject. Here, the respondents are asked which is the subject they want to choose for the *Ed* exam. The options are the four possible subjects for this exam and *Undecided* (U) option, for the students who haven't made their mind yet in this direction.

Item I5 (MC) is aimed, on the one hand, to verify the answers of those who have decided on a subject, and on the other hand, to identify the tendencies for the undecided students. For the following items, the chosen subject refers to the answer given for this item.

Items I6 and I9 are of the OA type, to give respondents the possibility to share their reasons for choosing or rejecting a subject.

Item I8 (MC) is aimed to identify the subjects that students are sure that they won't chose for the *Ed*

exam. They have to pick an answer from the set of four possible subjects.

Items I7 and I10 (MC) represent the average score obtained so far at the chosen, and at the rejected subject, respectively. In the Romanian educational system, the average scores for passing a subject belong to 5.00-10.00 interval. The average scores lower than 5.00 don't allow students to pass a year of study. Thus, the possible answers at these items are the following intervals of average scores: 5.00 - 5.99, 6.00 - 6.99, 7.00 - 7.99, 8.00 - 8.99, and 9.00 - 10.00.

Items I11-I14 (LS) are designed to identify the students' attitude towards choosing a subject for the *Ed* exam (I11 is for *P*, I12 - for *C*, I13 - for *B*, and I14 - for *CS*). These items are LS, and the answers are coded from 1 to 5, as follows: (1) *Definitely I won't choose it*, (2) *Unlikely that I will choose it*, (3) *Undecided*, (4) *Almost sure I will choose it*, and (5) *Definitely I will choose it*.

3.2 Formalisation

The analysed problem is formalized as follows. $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ shall denote a set of data with *n* instances, where any instance *si* characterizes the answers of a student at the questionnaire. Each instance has a unique collection of features that define it, i.e. $\mathcal{F} = \{ft_1, ft_2, ..., ft_k\}$, features which were identified as relevant in the proposed questionnaire.

In the present paper, the students' answers in the questionnaire are considered the feature values. As a consequence, every student *si* is seen as a vector with the dimension *k:* $s_i = (si_1, si_2, ..., sik)$, where the value of the feature ft_j for the student *si* is represented by s_{ij}.

3.3 Analysis Methods

This subsection presents the methods employed for analyzing the collected answers given by respondents. Firstly, a method for statistical analysis are presented. Then, two UL methods used in this paper are introduced: AR and SOM.

3.3.1. Statistical Analysis Method

The Chi-Square test is used in this study to determine the presence or absence of correlations between two or more features. In this respect, there is a computed Pearson Chi-Square test value (*p-value*) for each hypothesis. A *p-value* lower than a chosen significance level (common choice: 0.05) rejects the null hypotheses (Wasserstein & Lazar, 2016). According to Bearden et al. (1982), the Chi-Square test is sensitive to the sample size, and relevant results are obtained for samples with a minimum size of 100 instances.

3.3.2 Unsupervised Analysis Methods

Association Rule (AR) is the expression of an implication of the form X→Y, where X and Y are sets of distinct items. An association rule's *confidence* and *support* levels can be used for assessing its strength. Confidence establishes how frequently items in Y appear in rules that contain X, and support indicates the percentage of instances that comprise all of the items listed in that association rule (Agrawal et al., 1993). The formal definition for *confidence* is expressed by equation (1) and for *support* - by equation (2), where N represents the cardinal of the data set.

$$c(X \rightarrow Y) = \frac{fr(X,Y)}{fr(X)} \quad (1)$$

$$s(X \rightarrow Y) = \frac{fr(X,Y)}{N} \quad (2)$$

For mining ARs from a data set, the *Apriori* algorithm is used, which was implemented in *apyori* library from *Python* language, enforced by *Google Colab*. This algorithm works based on the *Apriori Principle*, which states that if a set of items is frequent, then all its subsets are frequent as well. Thus, the algorithm tries to search for those item sets which are frequent in a data set and to enlarge them on condition that those item sets are found frequently enough in the analysed database.

Self-organizing maps (SOMs) (Somervuo & Kohonen, 1999) are artificial neural networks that are trained via UL to generate a low-dimensional map of the input space. The number of neurons from the input layer of the generated map is equal to the dimensionality of the input data (i.e. number of data features), while the output layer consists of neurons generally distributed in a 2D space. Every input instance is plotted during training to a neuron on the output layer in such a way that topological relationships from the input space are maintained after training, thereby ensuring

that instances which are close to each other in the input space are plotted to neurons that are near one another on the map. The input space (i.e. the collection of numerical vectors with the dimension *k*, as it is illustrated in subsection 3.2) is encoded into a two-dimensional space using SOMs. Thus, by maintaining the input data's structure, a mapping $f : \mathbb{R}^k \to \mathbb{R}^2$ is encoded by the UL models. For measuring the accuracy and quality of the learned SOMs, the *average quantization error* (AQE) is employed (Kohonen et al., 2009). After each training epoch, the AQE is computed as the mean of the Euclidian distances between the input vectors and their *best matching units* (BMUs). Lower values for AQE suggest better (more accurate) maps, i.e. a better mapping of the input space into the 2D output space.

The proposed implementation of the SOM algorithm was carried out by using a torus topology and the U-matrix approach for visualizing the trained map (Lötsch & Ultsch, 2014): whiter areas indicate the boundaries separating the clusters, whereas darker areas express clusters of comparable instances. The proposed SOM implementation employed the following parameters: an adaptive learning rate (0.01 for initial value), 50x50 map dimensions, and 400 training epochs. The SOMs were constructed using the unnormalized data sets.

In this paper, ARs and SOMs are used for detecting some hidden patterns which could be useful in analysing the clusters of students grouped based on different criteria.

4. Case Study

This section presents the case study chosen for this paper. It comprises the description of the data set utilized for this study, and the analysis of the data set.

4.1 Data Set

To obtain the data set used for this study, the questionnaire presented in the subsection 3.1 was applied in 2022 to Romanian high school students from Real Sciences classes. The obtained data set is available in (Maier et al., 2023a). It consists of 301 instances, i.e. the answers of respondents from 9 counties in Romania. The main characteristics of the samples and their related percentages are presented below.

4.2 Data Analysis

In relation to the specialization of the students, the sample contains 48.8% of answers from I, 22.3% of answers from M, and 28.9% of answers from S.

The respondents are from all levels of study in high school, as follows: 30.9% from 9th grade, 32.23% from 10th grade, 18.6% from 11th grade, and 18.27% from 12th grade.

54.49% are male, 41.2% are female, and 4.32% of respondents didn't specify their gender.

As the chosen subject for the *Ed* exam (I4) is concerned, *P* is preferred by 4% of respondents, *C* – by 1.7%, *B* – by 29.2%, *CS* – by 37.9% of students, and 27.2% of the surveyed students haven't decided yet.

The ranking of chosen subjects is preserved when those who are undecided have to decide which would be their favorite subject at I5: *CS* is on the first place with 49.17% of answers, on the second place is *B* with 39.53%, then *P* with 7.64% and *C* with 3.65% of answers. A reason for *CS* being on the first place could be the high percentage of students from *M* and *I* specializations. Even if the percentage of students from *S* specialization is smaller, a great interest for *B* can be noticed.

I6 is an OA item and, for facilitating the study, its related answers have been analyzed and categorized into a set of motivational patterns: Mc = {*the specialization of my class, useful in the future, I like it, easy, by elimination, I don't know*}. On the first places are the easy (35.9%), and *I like it* (32.6%) patterns. Then, there are the following patterns: *useful in the future* (18.6%), *the specialization* of my class (8%), I don't know (3.3%), and *by elimination* (1.7%).

I7 includes the average scores of the respondents for the subjects chosen at I5. On the first place is the range between *9* and *10* (63.1%), then there are: *8-8.99* range (27.2%), *7-7.99* (7.6%), *6-6.99* (1.3%), and *5-5.99* range (0.7%).

From I8 answers, a reversed ranking results as against I5, i.e. on the first two places are *C*

(35.22%), and *P* (34.55%), then there is *CS* (16%) and then *B* (14%).

I9 is an OA item and, analogous to I6, its related answers have been analyzed and categorized into a set of motivational patterns: Mr = {*not the specialization of my class, not useful in the future, difficult, I don't like it, I don't know*}. On the first places are *difficult* (58%) and *I don't like it* (32%) patterns. Then, there are *not useful in the future* (5.6%), *not the specialization of my class* (2.7%), and *I don't know* (1.7%) patterns.

I10 includes the average scores of the respondents for the subjects rejected at I8. On the first place is the range between *9* and *10* (33%), then there are: *8-8.99* range (24%), *7-7.99* (21%), *6-6.99* (13%), and *5-5.99* range (9%).

Items I11-I14 express the students' attitude towards choosing a certain subject. The answers' summary is plotted in the Figure 1 and it illustrates the answers to the previous items.

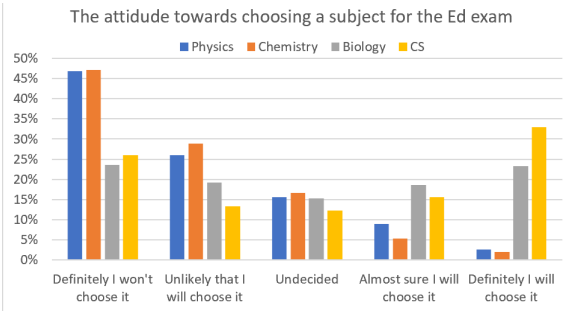

Figure 1. The students' attitude towards chosing a certain subject for the *Ed* exam

5. Results and Discussion

This section presents the experimental findings for the data set presented in the subsection 4.1, using the suggested methodology, together with an analysis of the obtained results.

5.1 Statistical Analysis Results

The statistical analysis of the input data set is conducted using the *IBM SPSS Statistics 29.0.0.0* software, after the coding of the data set. The following analysis is meant to find some relations between the students' characteristics using Chi-Square test. The obtained results are discussed below.

The study starts with the analysis of *the relation between the class specialization* (I1), *the grade* (I2), *and the chosen subject* (I5). For this analysis, the null hypotheses is stated: *There is no relation between the class specialization, grade and the chosen subject*. After applying Chi-Square test, the *p-value<0.001* was obtained, so the null hypothesis was rejected. Figure 2 plots the relation between the three features from the hypothesis above. One can notice that *CS* is in the top of students' preferences from I and M specializations and *B* is the favorite of students from S specialization irrespective of their grade. Another observation is that students from I are more determined with regard to *CS* then their mates from M. Students of specialization M from 11th grade are the most determined as regards *CS*, which can be explained by the number of *CS* hours/week: students from I have 4 hours/week in the 9th and 10th grades and 7 hours per week in the next two years, while those from M have 1 hour/week in the 9th and 10th grades and 4 hours/week from the 11th grade on; however, the Ed exam at *CS* is identical for both M and I students.

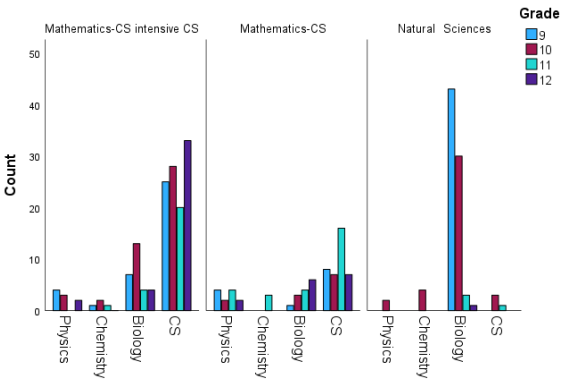

Figure 2. The distribution of class specialization and grade in relation to the chosen subject

Gender is another feature of interest in the present study, so the next null hypothesis is: *There is no relation between the specialization, gender, and chosen subject*. The *p-value=0.005* obtained at the Chi-Square test rejects the null hypothesis, so there are some relations between the specialization, gender, and the chosen subject that can be observed in Figure 3. It can be seen that *CS* has the highest percentage among the boys from I and M specializations. For the classes with specialization S, boys and girls make up a majority in choosing *B*. However, there is considerable interest in *B* on the part of girls from the I specialization.

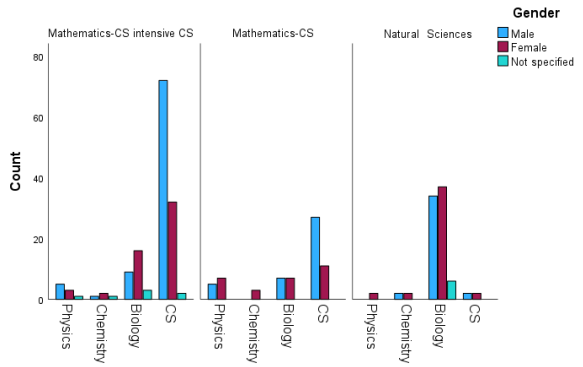

Figure 3. The distribution of genders on the chosen subject

Another aspect is the *motivation for choosing a subject* (I6). The study continues with the null hypothesis: *There is no relation between the chosen subject and the motivation to choose it.* At the Chi-Square test, the *p-value<0.001* was obtained. This means that the null hypothesis was rejected, so there is a relation between the chosen subject and the motivation for the respective choice.

The results are plotted in Figure 4. Looking at the plot, it can be noticed that, regardless of the chosen subject, an important reason is that respondents perceive a subject as being easy, because they feel well-prepared, and this could be seen as the students` level of training. Another reason, almost as important, is the fact that students like the chosen subject. In the third place, it matters that the students continue their studies in the chosen field (they consider it *useful in the future*). These three motivational patterns are connected, given that the preference for a subject (*I like it*) determines the manifestation of interest for the students who chose it and, implicitly, a good training in the field (*easy*). The implication is also valid the other way around: even if a student doesn't have a preference for a certain subject, when he/she studies it, he/she may discover things that he/she is passionate about, which could increase his/her interest in what he/she is learning. Also, the interest in a subject and a good training in its field determine the desire to continue studying that subject (*useful in the future*). The connection between these patterns is also shown by Stirling (2013).

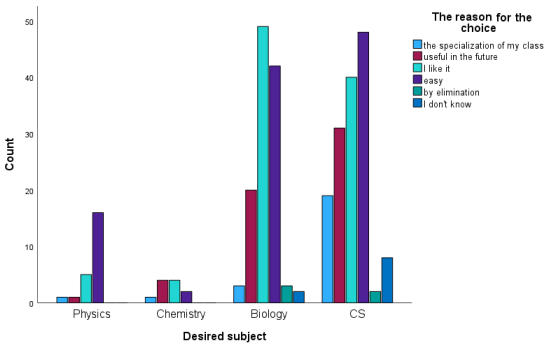

Figure 4. The distribution of the motivational patterns for the chosen subjects

To verify if the average score obtained for a subject influences the students' decision, the following null hypothesis is stated: *There is no relation between chosen subject and the related average score*. The Chi-Square test gave the *p-value=0.822*, so the null hypotheses was accepted. This means that grades don't influence the students' preference for a subject.

The next aspect is the *motivation for rejecting a subject* (I9). In this respect, the null hypothesis is stated: *There is no relation between the rejected subject and the motivation for rejecting it*. The Chi-Square test gave the *p-value=0.067*; this implies that the null hypotheses was accepted. However, based on Figure 5 it can be noticed that the main reason for rejecting a subject is because it's perceived as being *difficult* by the students. The second reason for the rejection is that students *don't like* that subject. These two major reasons are the opposite of the two major reasons for choosing a subject: easy, and *I like it*, which had been discussed previously.

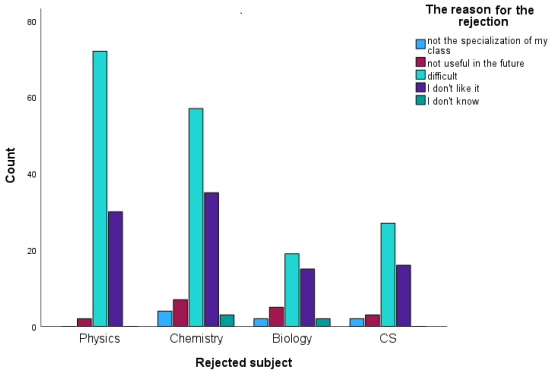

Figure 5. The distribution of the motivational patterns for the rejected subjects

Another interesting aspect is linked to the existence of a relation between *the rejected subject and the average score obtained at that subject.* The null hypotheses is: *there is no relation between the rejected subject and the average score obtained at the rejected subject*. For the Chi-Square test, the *p-value<0.001* was obtained. It can be concluded based on this value that there are some connections between poor school results at a subject and its rejection.

5.2 Results of the Unsupervised Larning-based Analysis

Further on, this paper focuses on the presentation and discussion of the outcomes produced by the ARs and SOMs-based UL models.

5.2.1 Association Rules

After applying the *Apriori* algorithm on the data set described in subsection 4.1, some association rules were identified and the most relevant are presented below. In order to obtain them, the value for the minimum confidence was set at 0.8, the value of minimum support at 0.08, the value of minimum length for a rule at 2 and the maximum length at 14. For the complete data set and these settings, the *Apriori* algorithm provided a set of 219 rules. In this set, the obtained rules were strongly influenced by the relation between items I4 and I5, which was presented in subsection 3.1, so it is obvious that there are rules with a confidence value of 1 which have the chosen subject in the set X and the same as a desired subject in the set Y. An *Undecided* answer appears only in a single rule, i.e. the students who chose the value 1 for the item I12 (the attitude towards choosing *C*) and are undecided at I4, reject *C* as a subject at *Ed* exam. Taking into account these observations, it results that I5 has almost the same influence as I4 during the mining process.

For a better comprehension of these rules, item I4 was removed from the initial data set. After applying the same algorithm to the new data set, 158 rules were obtained, and they are available in (Maier et al., 2023b). For many of these rules, the connection between the chosen subject (I5) and the attitude towards this subject on LS is checked, i.e. when a student chooses a subject, the attitude on LS is 5 for that subject. Analogously, this fact is available for the rejected subject (I7), too: if a student rejects a subject, then the value on LS for the attitude regarding that subject is 1.

With the aim of detecting some patterns in the input data set, some significant rules that were obtained are presented below. The minimum support for the following rules was set at 0.08, which means that each itemset of a rule is found in at least 8% from the entire dataset (i.e. 24 rows where an itemset is found).

Rules with confidence 1:

- 9th grade students from S choose *B;*
- Students who are sure that they won't choose either *C* or *B* and have the average score at the chosen subject between *9* and *10*, will choose *CS*;
- 9th grade students who are sure that they won't choose *CS*, will choose *B*;
- Students who are sure that they won't choose either *P* or *B*, will choose *CS*.

Rules with confidence between 0.8 and 0.99:

- Students from specialization S who are sure they won't choose *CS* and have the average score between *9* and *10* at the selected subject will choose *B*;
- Boys who are sure that they will choose *CS* and have the average score at the chosen subject between *9* and *10*, are from specialization I;
- Students who are sure they won't choose either *CS* or *P* and for whom the rejected subject is difficult, will choose *B*;
- Students who are sure they will choose *CS* and reject any other subject because they don't like it, are from I;
- Boys from specialization S choose *B*;
- Students from S who chose a subject because they like it, will choose *B*;
- Students from I who reject *B*, will choose *CS*, in case they have an average score between *9* and *10*;
- Students from S who choose a subject because it's easy, will choose *B*;

- Students from I who are sure they won't choose either *B* or *C*, will choose *CS*;

- 12th grade students who are sure they will choose *CS*, are from specialization I.

5.2.2 Self-Organising Maps

The following figures depict the U-matrix visualization of the SOMs unsupervisedly trained on the set of data described in subsection 4.1. Whiter regions signify the boundaries separating the clusters, and darker regions indicate clusters of related instances. The obtained maps unfold some patterns unsupervisedly learned on the analysed data set, which are related to different features. These features were established according to the results obtained for the statistical analysis and AR.

Figure 6 is the representation of the students' features in relation to their specialization. Two clusters can be noticed: (s1) students from specializations M and I, and (s2) students from specialization S. A reason for grouping different instances in (s1) is that the only difference between M and I lies in the number of *CS* hours/week and otherwise, they are identical. Some outliers can be noticed, i.e. instances with M and I specialization in the cluster (s2). The reason is that there are more students from M and I who choose subjects more specific to the specialization S (*P*, *C*, or *B*) than there are students from S choosing *CS*, which is specific to M and I.

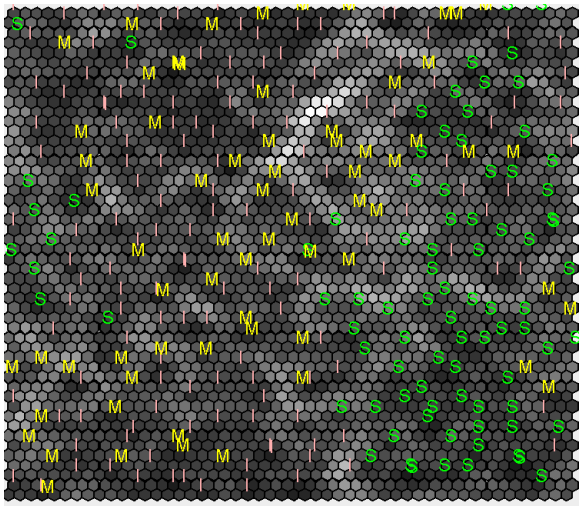

Figure 6. SOM visualisation of the class specialization analysis (SOM1)

Figure 7 depicts the SOM visualization when grouping students' features from the perspective of the answers for item I4. For readability reasons, *CS* was denoted by *I*. It can be noticed that there are three important clusters, bounded by the instances which are undecided (*U*) regarding the subject preferred for *Ed* exam: (c1) the cluster of students who chose *CS* (I on the plot), (c2) the cluster of students who chose *B*, and (c3) a small cluster of students who chose *P*. Those who chose *C* and some instances with *P* and *U* belong to the cluster (c2).

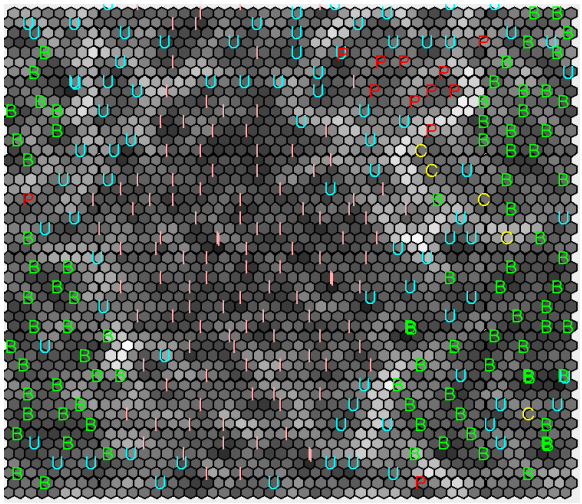

Figure 7. SOM visualization of the item I4 analysis (SOM2)

Figure 8 is a visualization of the analysed data set from the perspective of item I5. This representation preserves the clusters which resulted from the item I4, but it adds two small clusters with *C* and some instances with *P* in clusters (c1) and (c2).

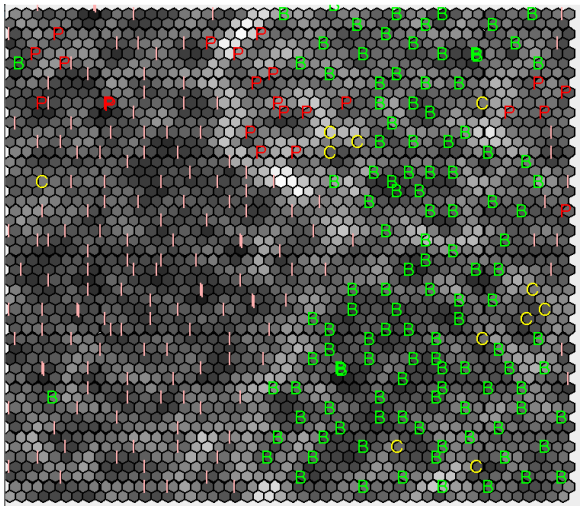

Figure 8. SOM visualization of the item I5 analysis (SOM3)

Figure 9 illustrates the SOM visualization for the rejected subject, with two major clusters: (r1) students who reject *C* and *B*, and (r2) students who reject *P* and *CS* (I).

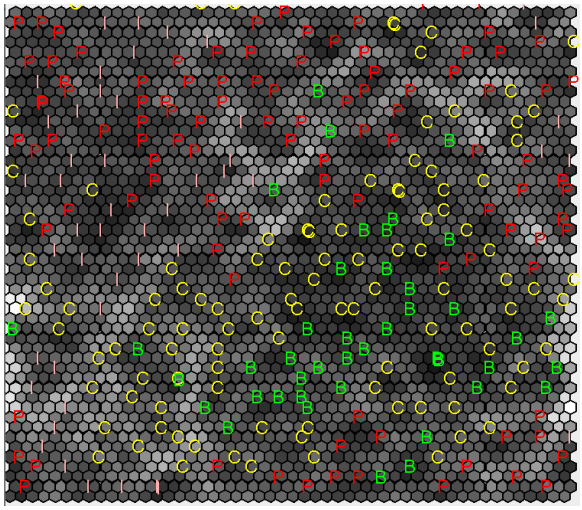

Figure 9. SOM visualization of the rejected subject analysis (SOM4)

A fairly good distinction between clusters can be observed in the Figures 6 to 9. The quality of the SOMs illustrated in Figures 6 to 9 may be visualised in Figure 10, which depicts the evolution of the AQE during the training epochs. A convergence of the AQE during the training process is observed in Figure 10. The final values obtained for the AQE (between 0.74 and 0.79) at the end of the training process are close to zero, thereby confirming the accuracy of the trained maps.

Inside of the bigger areas that can be noticed in the SOM visualisations, there is typically no obvious distinction between classes. This may be a consequence of the curricular similarities concerning the discussed specializations. Almost every year, in the high school admission process, the students' scores for these three specializations are very similar. During the four years of high school, any student can switch to another specialization if he/she wants, and if there are enough places in the classes from the targeted specialization. At the end of high school, students can attend any faculty, so they are not constrained by a certain specialization related to their present or future interests.

5.3 Discussion

After the statistical analysis, the *RQ1* was answered. Based on the obtained results, it can be deduced that the students from the Real Sciences classes overwhelmingly chose *CS* or *B* for the *Ed* exam regardless of their general school performance. Moreover, in the case of the *P* and *C* subjects, the students declared that they didn't choose them because they were not of interest for them and they were not prepared to take such exams. Students chose their subject for the *Ed* exam being motivated by two major aspects: they perceived that subject being easy, and they liked that subject. Therefore, students motivated their choices not in terms of the results they hoped to achieve, but in terms of their previous learning experiences and the level of preparation they believed they had in one subject or another.

The UL-based analysis answered *RQ2* and *RQ3*. Thus, AR identified some subsets of rules which are in accordance with the statistical analysis. SOMs prove that there are some patterns related to choosing or rejecting a subject.

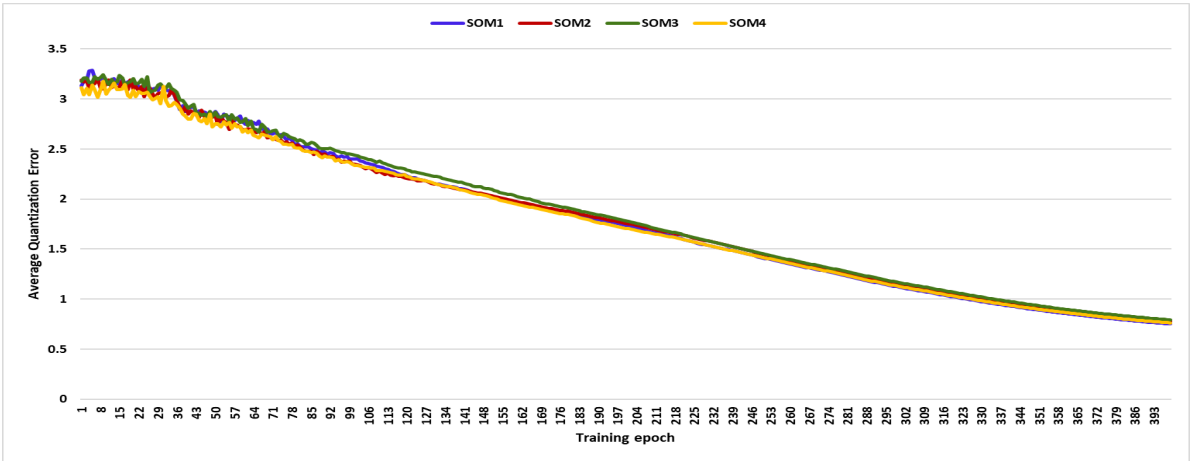

Figure 10. The average quantization error computed for each SOM clustering, as it can be visualized in Figures 6-9

6. Conclusion and Future Work

This paper presented, as a proof of concept, a study on applying unsupervised learning for mining behavioural patterns for high school students in relation to their subject prefferences for the baccalaureate exam. The proposed proof of concept used a survey based on small-sized data collected from high school students from 9 counties in Romania. The survey was carried out in order to identify the tendencies of students over the four years of high school.

Unsupervised learning techniques were applied as descriptive models in order to learn patterns and gain insights into the analysed data. The descriptive UL models employed in this paper (both ARs and SOMs) revealed behavioural aspects related to students (the choice of a subject at the baccalaureate exam). Moreover, both models may be further employed in a supervised learning scenario. For instance, by using the learned SOM, a new student who answers the survey may be classified according to the class (cluster) in which the BMU of the instance is mapped.

Research questions that served as the basis for carring out this analysis had been answered. The experiments provided empirical support for the idea that there are hidden patterns in the characteristics of Romanian high school students regarding their preferences for the *Ed* exam, patterns that could be relevant in identifying the behavioural tendencies of students. The AR and SOM models are able to unsupervisedly detect these patterns. Furthermore, it was underlined that, when analysing the profile of the students, there is a strong connection between the outcomes of the UL-based analysis and those of the statistical analysis.

The study presented in this paper has a major practical relevance, as it can be used for developing a recommender system for educational environments, to help students and teachers in the educational process. Through such a recommender system, teachers could anticipate the students' preferences or dislikes for certain subjects. Thus, teachers could design differentiated activities for students, in order to avoid the rejection of certain subjects or guide students so they can improve their performance in their preffered subjects. Even if the approach presented in this paper was applied to a case study from the EDM domain, it has a general character and it may be applied for mining behavioural patterns from any type of data. In order to better highlight the generality of the proposed approach, the aim is to further apply the methodology described in this paper in other application domains such as software defect detection, marketing and advertising, or sports.

Future work could focus on expanding this study for data collected from other high school specializations, or from high school graduates, and on including students' results obtained at the *Ed* exam. Further extensions and additional analyses may be carried out for detecting the clusters from the learned SOMs (by applying clustering algorithms to the SOM neurons) or for directly partitioning the data by applying hard or fuzzy clustering techniques (Ruspini et al., 2019). Also, the intention is to include other UL models in the presented analysis, such as *Uniform Manifold Approximation and Projection* (McInnes et al., 2018).

Acknowledgements

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/ CCCDI – UEFISCDI, project number PN-III-P4-ID-PCE-2020-0800, within PNCDI III.

REFERENCES

Agrawal, R., Imielinski, T. & Swami, A. (1993) Mining association rules between sets of items in large databases. *ACM SIGMOD Record.* 22(2), 207-216.

Bearden, W. O., Sharma, S. & Teel, J. E. (1982) Sample Size Effects on Chi Square and Other Statistics Used in Evaluating Causal Models. *Journal of Marketing Research.* 19(4), 425-430. doi: 10.2307/3151716.

Carter, L. (2006) Why Students with an Apparent Aptitude for Computer Science Don't Choose to Major in Computer Science. *ACM SIGCSE Bulletin.* 38(1), 27-31.

Chen, Z. (2006) From Data Mining to Behavior Mining. *International Journal of Information Technology & Decision Making.* 5(4), 703-711. doi: 10.1142/S0219622006002271.

Davies, P., Davies, N., Hutton, D., Adnett, N. & Coe, R. (2009) Choosing in schools: locating the benefits of specialisation. *Oxford Review of Education.* 35(2), 147-167.

Kohonen, T., Nieminen, I. T. & Honkela, T. (2009) On the quantization error in SOM vs. VQ: A critical and systematic study. In: Principe, J. C. & Miikkulainen (eds.) *Advances in Self-Organizing Maps.* Berlin, Heidelberg, Springer, pp. 133-144.

Kumar, V. & Chadha, A. (2012) Mining Association Rules in Student's Assessment Data. *IJCSI International Journal of Computer Science Issues.* 9(5), 211-216.

Kupiainen, S., Marjanen, J. & Hautamäki, J. (2016) The problem posed by exam choice on the comparability of results in the Finnish matriculation examination. *Journal for Educational Research Online.* 8(2), 87-106.

Lötsch, J. & Ultsch, A. (2014) Exploiting the Structures of the U-Matrix. In: Villmann, T, Schleif, F.-M., Kaden, M. & Lange, M. (eds.) *Advances in Self-Organizing Maps and Learning Vector Quantization.* Springer International Publishing, pp. 249-257.

Maier, M.I., Czibula, G. & Oneţ-Marian, Z.E. (2021) Towards Using Unsupervised Learning for Comparing Traditional and Synchronous Online Learning in Assessing Students' Academic Performance. *Mathematics.* 9(22), 2870. doi: 10.3390/math9222870.

Maier, M.I., Czibula, G. & Delean, L.R. (2023a) *High School Students' Answers from Real Sciences Classes.* https://docs.google.com/spreadsheets/d/1-YfyjoCIMssdAcN9BtLMvICBunSnhcUEM2XWKXmSsHs/edit?usp=sharing [Accessed 20th January 2023].

Maier, M.I., Czibula, G. & Delean, L.R. (2023b) *The set of rules for high school students preferences.* https://docs.google.com/spreadsheets/d/1g2GBcBkJC8Q-4_0Dco8O26weRJ_y3iKasijQwvHzPe4/edit?usp=sharing [Accessed 20th January 2023].

Maiti, S. & Subramanyam, R. (2019) Mining behavioural patterns from spatial data. *Engineering Science and Technology, an International Journal.* 22(2), 618-628. doi: 10.1016/j.jestch.2018.10.007.

McInnes, L., Healy, J. & Melville, J. (2018) UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv: 1802.03426.* https://arxiv.org/abs/1802.03426 [Accessed 19th January 2023].

Oneț-Marian, Z., Czibula, G. & Maier, M. (2021) Using Self-Organizing Maps for Comparing Students' Academic Performance in Online and Traditional Learning Environments. *Studies in Informatics and Control.* 30(4), 31-42. doi: 10.24846/v30i4y202103.

Qazdar, A., Er-Raha, B., Cherkaoui, C. & Mammass, D. (2019) A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Education and Information Technologies.* 24, 3577–3589. doi: 10.1007/s10639-019-09946-8.

Romanian Ministry of National Education (2022) *Baccalaureate | Ministry of Education.* https://edu.ro/bacalaureat [Accessed 20th January 2023].

Ruspini, E. H., Bezdek, J. C. & Keller, J. M. (2019) Fuzzy Clustering: A Historical Perspective. *IEEE Computational Intelligence Magazine*. 14(1), 45-55. doi: 10.1109/MCI.2018.2881643.

Sarker, I. H., Colman, A. & Han, J. (2019) RecencyMiner: mining recency-based personalized behavior from contextual smartphone data. *Journal of Big Data.* 6(1), 1-21. doi: 10.1186/s40537-019-0211-6.

Seifi, H. & Parsa, S. (2018) Mining malicious behavioural patterns. *IET Information Security.* 12(1), 60-70. doi: 10.1049/iet-ifs.2017.0079.

Somervuo, P. & Kohonen, T. (1999) Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters.* 10, 151-159. doi: 10.1023/A:1018741720065.

Stirling, D. (2013) Motivation in Education. *Aichi Universities English Education Research Journal.* 29, 51-72.

Taconis, R. & Kessels, U. (2009) How Choosing Science depends on Students' Individual Fit to 'Science Culture'. *International Journal of Science Education.* 31(8), 1115–1132. doi: 10.1080/09500690802050876.

Wasserstein, R. L. & Lazar, N. A. (2016) The ASA statement on p-values: context, process, and purpose. *The American Statistician.* 70(2), 129-133. doi: 10.1080/00031305.2016.1154108.