Smart Learning Environments

RESEARCH Open Access

Role of absence in academic success: an analysis using visualization tools

Ronak Etemadpour[1], Yongcheng Zhu[1], Qizhi Zhao[1], Yilun Hu[1], Bohan Chen[1], Mohammed Asif Sharier[1], Shirong Zheng[1] and Jose Gustavo S. Paiva[2*]

*Correspondence: gustavo@ufu.br
[2]Faculty of Computing, Federal University of Uberlandia, UFU, Av. Joao Naves de Avila, 2121, 38408100 Uberlandia, Brazil
Full list of author information is available at the end of the article

Abstract

Understanding the academic performance of students in colleges is an essential topic in Education research field. Educators, program coordinators and professors are interested in understanding how students are learning specific topics, how specific topics may influence the learning of other topics, how students' grades/attendances in each course may represent important indicators to measure their performance, among other tasks. The use of data visualization and analytics is expanding in education institutions to perform a variety of tasks related to data processing and gaining into data-informed insights. In this paper, we present a visual analytic tool that combines data visualization and machine learning techniques to perform some visual analysis of students' data from program courses. Two educational data collections were used to guide the creation of i) predictive models employing a variety of well known machine learning strategies, attempting to predict students' future grade based on grade and attendance previous semesters and ii) a set interactive layouts that highlight the relationship between grades and attendance, also including additional variables such as gender, parents education level, among others. We performed several experiments, also using these data collections, to evaluate the layouts ability of highlighting interesting patterns, and we obtained promising results, demonstrating that such analysis may help the education experts to understand deficiencies on course structures.

Keywords: Education, Information visualization, Machine learning, Multidimensional data, Statistical test

Introduction

The importance of analytics and predictive methods in higher education, as well as the determining factors that contribute to academic performance are discussed in many research studies (Van Barneveld et al. 2012; Mattingly et al. 2012; Gutiérrez et al.) in order to improve the achievement of education goals, offer new modern opportunities for improving education system effectiveness and provide learning personalization.

Although the impact of learning strategies and gender differences on academic performance are addressed by Ruffing et al. (2015), the factors underlying the prediction of academic performance are still of great interest in educational psychology. One reliable task is to ask the school for some anonymized information about previous students to obtain useful information to perform analytic tasks. In this sense,

novel analysis strategies are useful in comprehending educational scenarios involving students performance and related factors, guiding decision making by educational experts.

In this paper we present a computational system for educational data analysis. We intend to demonstrate how predictive analytics and data visualization techniques can provide powerful decision-making aids for educators and school administrators, by providing means to identify and explore trends and patterns on these data, as well as to comprehend the real situation regarding a specific education scenario. We believe that a visual analysis tool employing machine learning and information visualization techniques improves the comprehension, by educational experts, of student's behavior on subjects over the semesters, guiding them in defining effective strategies to mitigate related deficiencies. Our main goal is to investigate the ability of the system's tools in addressing the following research questions:

- **Is it possible to predict the students' grades using grades and attendance in subjects from past semesters? Does students' attendance impact on their grades in the semesters?** Research shows that attendance is one of the most important factor in students' academic performance and achievement (Jones 2006; Kassarnig et al. 2017). Considering students' attendance, we employed a variety of machine learning models to predict students' data trends over several semesters and we compared this prediction to the real data, in order to measure their accuracy.
- **Is there any correlation between students' gender and their performance on different subjects?** We considered gender in a performance statistical evaluation. As this factor as well as race, ethnicity, educational, and psychological factors are also addressed in some studies (Dee 2005; King et al. 2002; Wilson and Shrock 2001).
- **Do external factors, such as parents' education level impact on students' performance in school?** The influence of external factors is also discussed by Gooding (2001). We employed a "Multidimensional Projection" technique to explore the structure of the relationship among students in terms of similarities. The idea is to identify profiles and/or outliers that may explain their performance in the courses over the semesters.

In order to evaluate our system, we considered two educational data collections, one of them containing students records from Exact science programs of the Faculty of Computing of Federal University of Uberlandia, Brazil, and the other one containing students records from two public schools of Portugal.

We present related work on educational analysis in "Related work" section and detail the machine learning techniques and the visual approaches in "Academic visual analysis system" section. The results of our evaluation are presented and discussed in "Results" section, and we present our conclusions in "Conclusion and future works" section.

Related work

Although there are several works addressing the use of data visualization by instructors to enhance learning (Klerkx et al. 2014; Anaya et al. 2016; Thompson et al. 2013), not many studies are found that use visualization approaches to analyze educational data.

The selection, processing, visualization, and analysis of multiple learning and learning environments elements, as well as the links between them have been discussed in Thompson et al. (2013) to provide a comprehension of the learning processes in complex learning environments. Performing limited studies, Lacefield et al. (2018) explore applications of machine learning, predictive analytics, and data visualization to student information available to educational decision makers. They concentrated more on demonstrating individual academic performance histories to identify "at-risk" students in real time for advising, academic coaching, and other support services.

DeCotes (2014) used a Heatmap visualization to represent students' grade performance for different course pairs. In this work three different course pairs are compared and the results have shown that different course pairs yield different behaviors. For instance, the study indicates that students who did poorly in level one course will receive similar poor grades in the level two course. Using this visualization, they could see the transition of grades from one level to another level more clearly.

In terms of predictive analytics, Urrutia-Aguilar et al. (2016) employed also a logistic regression model for the prediction of variables that have impact on the academic performance of first year biomedical students. Soule (2017) also employed multiple logistic regressions to improve prediction techniques regarding the future performance of students in selected university courses and his study showed that in all cases, logistic prediction models matched or exceeded the performance of current prediction methods while using an equal or lesser number of explanatory variables.

Fernandes et al. (2019) performed a descriptive statistical analysis to gain insight from the academic performance of students. Gutierrez et al. () used Random Forest models to predict students' academic performance in different engineering subjects. We employed some of these predictive analytics techniques in our system. Fernandes et al. (2019) used Classification models based on the Gradient Boosting Machine (GBM) to predict academic outcomes of student performance at the end of the school year and they showed that 'grades' and 'absences' attributes were the most relevant for predicting the end of the year academic outcomes of student performance. The analysis of demographic attributes revealed that 'neighborhood', 'school' and 'age' are also potential indicators of a student's academic success or failure. We also investigated 'grades', 'absences' and 'age' factors in our analysis employing visualization and machine learning approaches. We also performed descriptive statistical analysis to gain insight from data and used visualization techniques to show the significant differences.

Academic visual analysis system

We developed a web-based system[1] that provides a set of interactive visualization tools to perform an academic data visual analysis. An overview of our web-page can be found in Fig. 1. The layouts and interaction tools were developed using Javascript D3.js[2] and Python Scikit-learn[3] libraries. Two educational data collection were used to guide the creation of the layouts, and also to evaluate their capability of highlighting interesting patterns. This section describes these collections, and the layouts developed for the analysis.

[1] fall-2018-education.herokuapp.com
[2] https://d3js.org/
[3] https://scikit-learn.org/

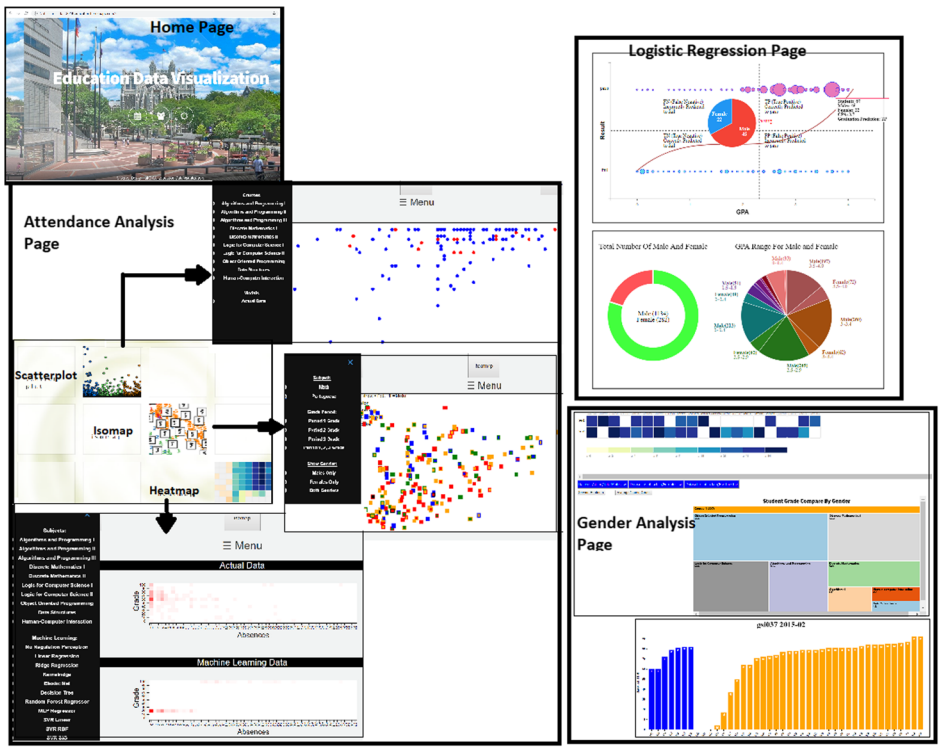

Fig. 1 Academic performance visual analysis web-page available on "http://fall-2018-education.herokuapp.com/". The system provides a set of tools: scatter plots for attendance vs. grades; Isomap projection for role of parents education; Heatmap Matrix visualization for machine learning approaches and predictions; Gender analysis comparing grades for male and female students. Using a drop down menu, users can make selections on different algorithms. Using radio buttons, they can select courses, semesters, periods, and gender

Collection 1: exact science subjects performances

This collection is composed of 1416 records of students from exact science programs of the Faculty of Computing of Federal University of Uberlandia, Brazil. Each record represents the performance of a single student in a specific subject, and provides the following attributes: student ID (anonymity), gender, course semester, subject, grade, and attendance (number of absences). This data collection contains multiple records for a single student, which allows us to follow their progress over the program. The data includes information between 2009 and 2017, and the course subjects are: "Algorithms and Programming I", "Algorithms and Programming II", "Algorithms and Programming III", "Discrete Mathematics I", "Discrete Mathematics II", "Object-Oriented Programming", "Data Structures", "Logic for Computer Science I", "Logic for Computer Science II", and "Human-computer Interaction". The students' grade are ranging from 0 to 100.

Attendance analysis & machine-learning approach

For predictive analysis, we attempt, by using the attributes of each individual student, to predict their grade in future subjects. We investigated several different machine learning models, described as follows.

The perceptron is one of the simplest forms of feeding forward artificial neural networks used for the classification of patterns, and represents a binary linear classifier

(Haykin 2010). The **Multilayer Perceptron** (MLP) (Witten et al. 2016) is a feedforward artificial neural network model that maps multiple input datasets onto a single output dataset via three or more layers of perceptron.

Regression analysis aims at identifying/determining the dependence of a response variable on one or more predictors, including prediction of future values of a response (Weisberg 2005). In this sense, **Ridge regression** (Gruber 2017), also known as Tikhonov regularization, significantly reduces the variance of the model, without substantial increase in its bias as an improvement to least squares estimation methods. **ElasticNet** (Zou and Hastie 2005a) is a linear regression model that is used for sparse models with few non-zero weights. On the other hand, **LASSO** (Tibshirani 2011), also known as least absolute shrinkage and selection operator, is a regression analysis method that simultaneously performs feature selection and regularization to enhance the prediction accuracy and interpret ability of statistical models based on Breiman's nonnegative garrote.

Support Vector Machine (SVM) (Yang et al. 2015) is a machine learning method based on statistical learning theory developed in the mid 1990s and it is a binary classification mode that seeks to maximize the learning machine's generalization ability by minimizing the structural/empirical risk and confidence range. In the case of linear inseparability, SVM first completes the calculation in low-dimensional space, then maps the input space to the high-dimensional feature space through the kernel function, and finally constructs the optimal separation hyper-plane in the high-dimensional feature space, thus better separating non-linear data.

Decision tree (Magerman 1995; Kumar and et al. 2011) is a binary/non-binary tree structure in which the non-leaf nodes represent a test on a feature attribute, the branches represent the output of the feature attribute over a range of values, and the leaf nodes represent a set of predefined categories. The decision process starts from the root node and tests the corresponding feature attributes in the item to be classified, selecting the most adequate output branch according to its values. This process continues until the leaf node is reached. The category associated with the reached leaf nod is then used as the decision result.

Finally, **Random Forest** (Breiman 2001; Osmanbegović and Suljić 2012) is a type of ensemble classifier which employs a bagging of decision trees. Each tree is constructed using a different bootstrap collection sample. During the training process, each tree node is split using the best among a subset of predictors randomly chosen at that node, which makes this strategy accurate and robust against over fitting. It also uses only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), representing a simple tuning technique (Liaw and Wiener 2002).

In this paper, we highlight the results using Multilayer Perceptron (MLP) (Witten et al. 2016), Linear Regression, Decision Tree (Kumar and et al. 2011) and Random Forest Regressor (Osmanbegović and Suljić 2012). In order to visualize the results of the predictive models as well as the original data, we used a Heatmap Matrix visualization. A heatmap is a straightforward visualization strategy that maps a color gradient to a range of measured values, which can be organized in a matrix structure, so that each matrix cell color corresponds to a data value. As shown in Fig. 2, the layout can quickly represents interesting patterns related to the values distribution on the data and highlight peak values. Evolution over time with respect to the other aspects of the data can be represented using this effective analysis strategy.

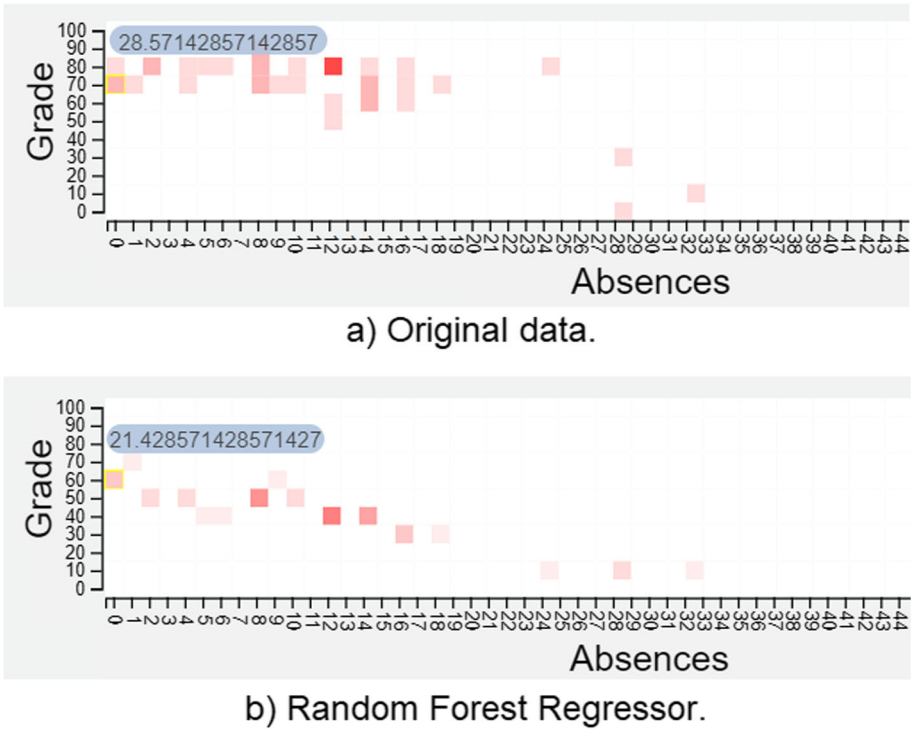

Fig. 2 Heatmap Matrix visualization of Attendance vs. Grade for HCI subject

In this layout, the horizontal axis represents absence amounts, and the vertical axis represents the grades. The percentage of students that meet a particular cell from the matrix is color coded accordingly. For example, in Fig. 2, considering "Human-Computer Interaction" course and the original data, 28.57% of students with zero absence received approx. 70% as their final grade. On the other hand, when analyzing the Random Forest Regressor predictive model Heatmap Matrix, 21.428% of the students are predicted in the same cell, i.e., with zero absence and approx. 70% as final grade. The results of this visualization are described in "Results" section.

Attendance analysis & scatter plots

To show the performances similarity between genders, individual scatter plots per course were created, in which the horizontal-axis represents the grade ($0 < x < 100$), and the vertical-axis represents the number of absences ($0 < y < max - absence$), as shown in Fig. 3. Each circle represents a student: blue circles represent male students and red circles represent female students. The circle distribution on the scatter plot indicates how two variables relate to each other, and how this relationship is observed in each gender.

Gender analysis & statistical heatmap matrix

We also employed Heatmap Matrices for the analysis of the role of genders. One Heatmap Matrix for each gender was constructed, in which the columns represent the courses subjects and the lines represent each semester. The average grades are mapped to the cells colors. For statistically compare the genders grades, we used a non-parametric Wilcoxon statistical test (Fix and Hodges Jr 1955) and implemented z-test to calculate p-values and significance at the 95% level of confidence. A third Heatmap Matrix was then used

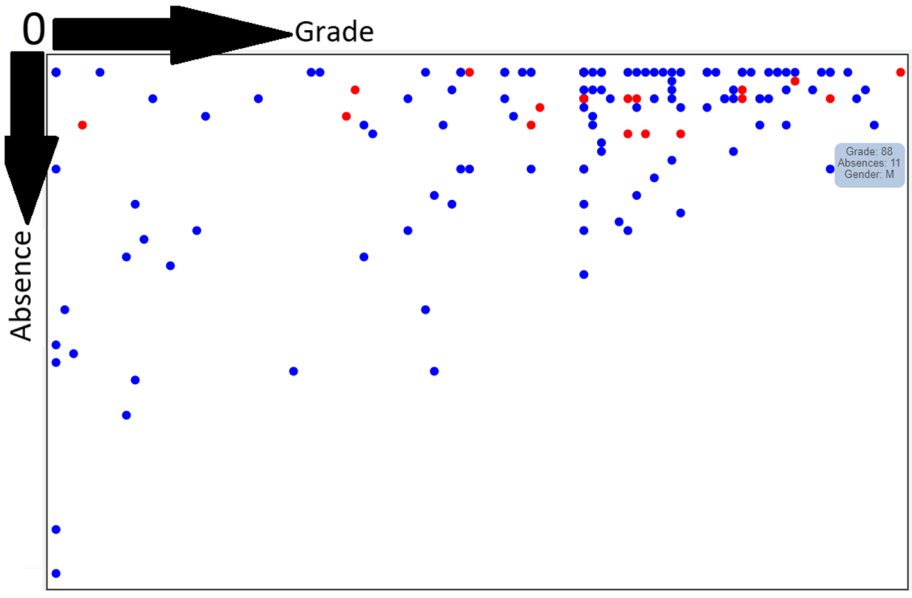

Fig. 3 Scatter plot for "Algorithms and Programming I" subject shows an outlier represented by a student with 11 absences and grade of 88. The pattern shows in general that students with less than two absences received higher grades

in which these p-values comparing grades for different genders are mapped to the cells colors. For example, as shown in Fig. 4, the average of grades in a highlighted course (gsi-2014 in semester 2) for males is 41.8%. The average of grades for the same course is 2.6% for females. In the layout, p-values are color coded if the comparison between females vs. males grades was statistically significant. In Fig. 4, Wilcoxon test shows a significant difference of mean values between males and females for course gsi-2014-semester 2 ($p = 0.0088 < 0.05$)). This visualization can be found in: URL [4].

Gender analysis & course treemap

In addition, we created a Treemap (Johnson and Shneiderman 1991) visualization to show the class population distributions, in order to investigate which subjects concentrated more students over time and to understand the proportion of male and female students in each course over the semesters. Figure 5 is a walkthrough for treemap visualization when "Object-Oriented Programming" is selected.

Logistic regression & pass/fail analysis

Logistic Regression is primarily a data classification technique that separates data points into one of two or more classes. In the present study, we used it to identify students as either "passed" or "failed". We first randomly separated the data into two smaller data sets for the logistic regression: a "training set" containing 80% of the original data and the "test set" containing the remaining 20%. Logistic regression fits the training set to find a best fit probability function, in our case, the probability of passing as a function of GPA or grade. We then assigned a threshold to make predictions. We used the test set to verify whether our prediction function is working fine or not, by calculating the errors. For a student to successfully pass, the probability of passing must be greater than 50%.

[4] http://fall-2018-education.herokuapp.com/gender

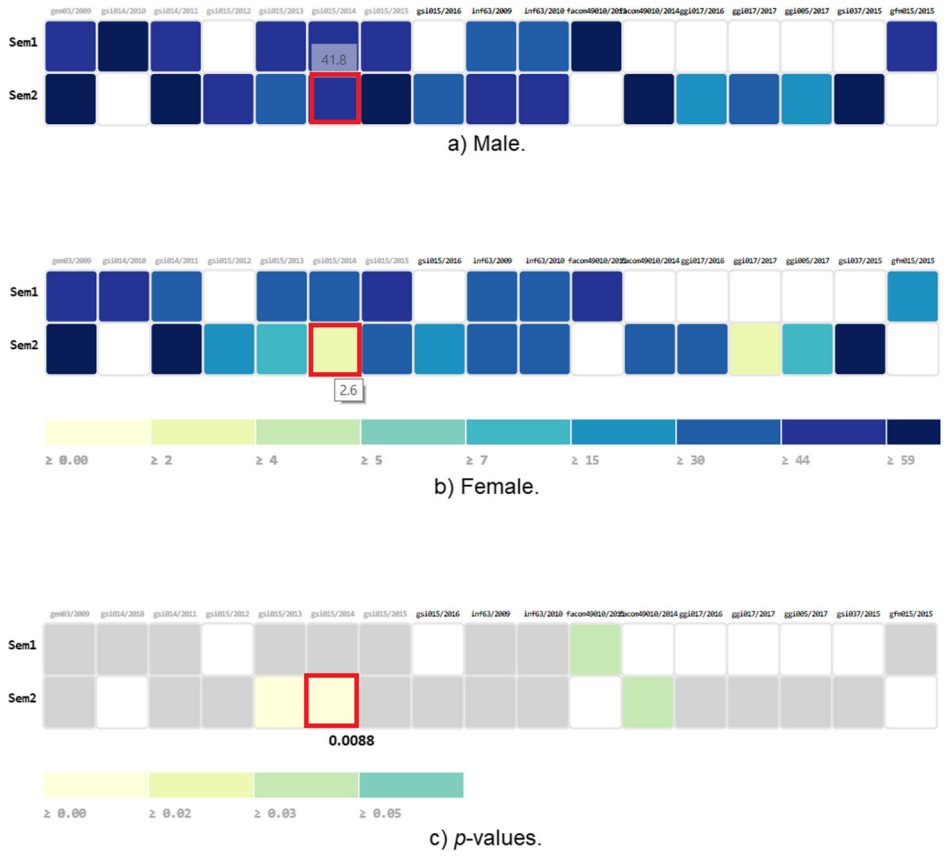

Fig. 4 Heatmap Matrix visualization of Gender vs. Grade for both genders and for associated *p*-values

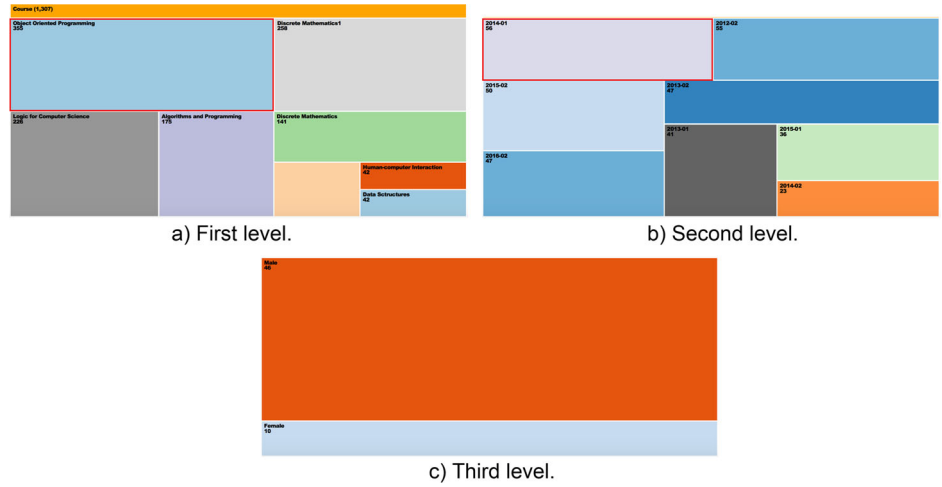

Fig. 5 Treemap visualization for "Object-Oriented Programming" course

To evaluate the predictions, the model generates a confusion matrix based on the test set, using the following indicators:

- *TP*: number of students correctly predicted to pass;
- *TN*: number of students correctly predicted to fail;
- *FP*: number of students incorrectly predicted to pass;
- *FN*: number of students incorrectly predicted to fail.

Using the confusion matrix, we can evaluate the correlation between the explanatory and independent variables, using the following formulas to get the Accuracy (A), Precision (P), Fallout (F), and Recall (R) of the model:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

$$P = \frac{TP}{TP + FP}$$

$$F = \frac{FP}{TN + FP}$$

$$R = \frac{TP}{TP + FN}$$

In this study, a threshold of 2.3 (GPA) was created to make predictions. We used the grade for whole courses to make a better predictor. In order to map predicted values to probabilities, we use a Sigmoid function. The function maps any real value into another value between 0 and 1. Sigmoid functions are commonly used in machine learning to map predictions to probabilities, according to the following equation:

$$S(z) = \frac{1}{1 + e^{-}z}$$

1. S(z) = output between 0 and 1 (probability estimate)
2. z = input to the function (your algorithm's prediction)
3. e = base of natural log

The main purpose of this Logistic Regression Visualization is that it shows the prediction of students' grades based on the performance of students in previous semesters. The resulting layout is shown in Fig. 6.

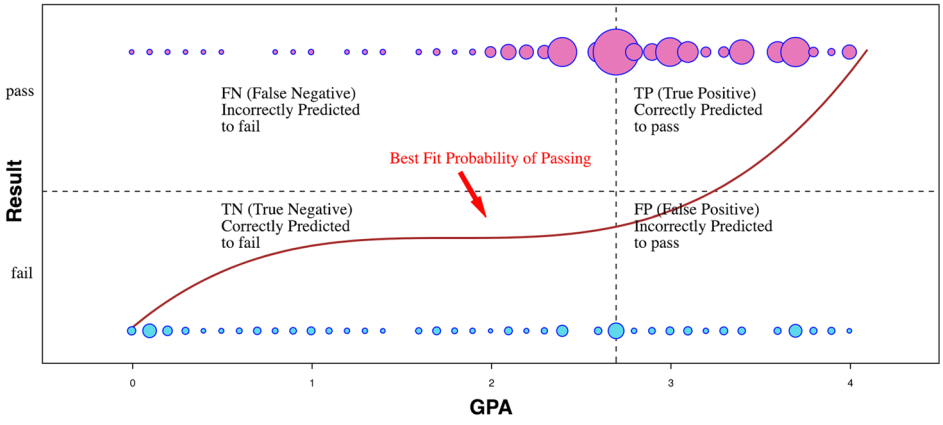

Fig. 6 Logistic Regression Visualization

In this visualization, the Y-axis indicates whether a student passed or failed. The X-axis shows students GPA based on their performance. The visualization is divided into 4 main categories in which confused matrix took place. The situation in which a student is approved in a specific course subject is associated to a false negative, which means that the students were incorrectly predicted to fail, and to a true positive which means the students were correctly predicted to pass. On the other hand, the situation in which a student fails is associated to a true negative, which means the students were correctly predicted to fail, or to a false positive, which means the students were incorrectly predicted to pass. The line curve represents the probability of passing for the students in which the Sigmoid activation method is used. We use bubbles whose size indicates the number of students. In order to make this visualization interactive, a zoom tool is used while hovering over the bubbles. A tooltip associated with the bubbles displays the analysis details, as shown in Fig. 7.

Finally, clicking over a bubble displays a pie chart with the proportion of male/female students as well as their GPA ranges in order to show a clear indication of their performance, as shown in Fig. 8.

Collection 2: public school performances

This collection is composed of 1044 students records collected during the 2005–2006 school years from two public schools of Portugal (Cortez and Silva 2008). Each record contains the performance information of a single student in "Mathematics" (395 records) and "Portuguese" (649 records) subjects. We considered the following attributes, for each record: sex, age, number of past class failures (varying from 0 to 4), number of absences in all three grading periods (varying from 0 to 93), grades in each of the three grading periods (varying from 0 to 20), parents education level, and parents jobs. The parents education level is categorized in four classes: 0:none; 1:primary education (4th grade); 2:5th to 9th grade; 3:secondary education; 4:higher education. The parents jobs can be "teacher", "health care related", "civil services (e.g. administrative or police)", "at home" or "other".

Attendance analysis & Isomap projection

Multidimensional projection (MP) techniques map high dimensional instances from a data collection to individual points on a lower dimensional visualization space, with usually 2 or 3 dimensions. The idea is to preserve on this reduced space the significant

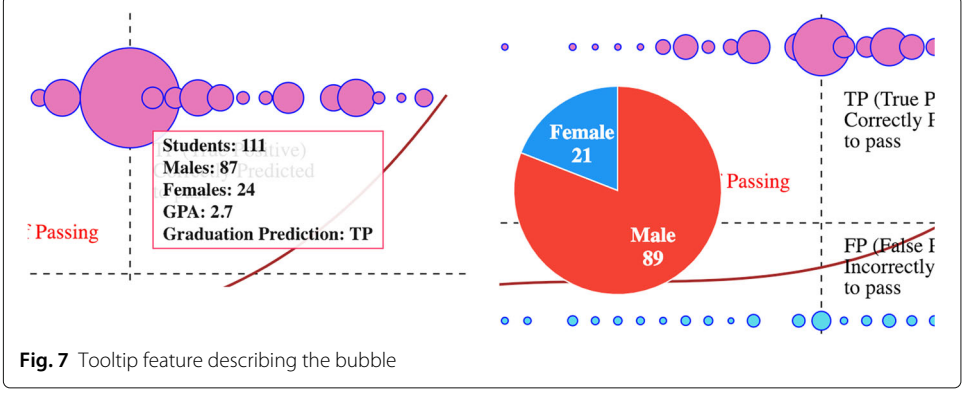

Fig. 7 Tooltip feature describing the bubble

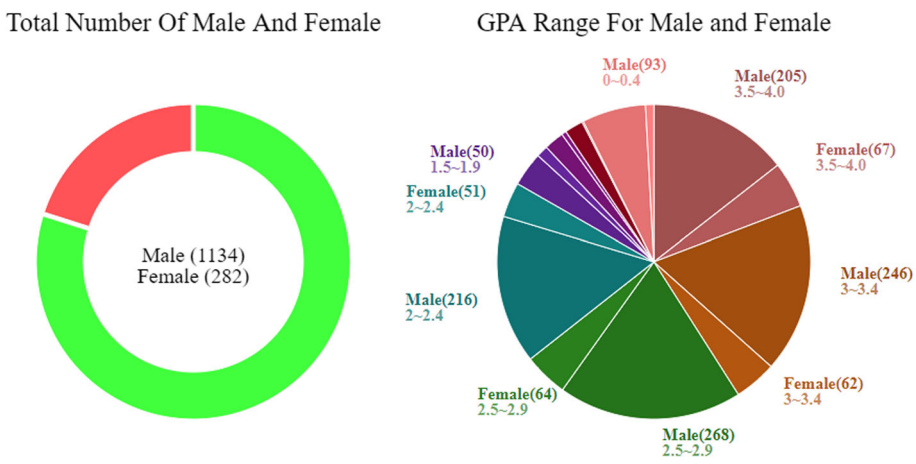

Fig. 8 Pie chart visualization of GPA analysis

relationships observed in the original data, using similarity measures that put similar instances close to each other, and non-similar instances apart. A graphical representation, usually a scatter plot, is then created and displayed to the user, who is able to have an overview of the whole data set, or to concentrate in specific regions of interest. Among these techniques, Isomap visualization (Tenenbaum et al. 2000) is able to precisely maps the data into a 2D scatter plot layout (Etemadpour et al. 2015).

The analysis of educational performance data may significantly benefit from MP techniques, in the sense that it is possible to identify groups of students with similar behavior, or to organize these groups in different categories, as well as to explore the relationship among attributes from instances in a group, in order to iden7tify correlations among them. Hence, we employed Isomap on the quantitative features to map relevant relationships in the original space and to visually investigate the role of additional students information, such as parents' education, on students' performance. The resulting scatter plot layout is shown in Fig. 9.

Each scatter plot point is mapped to a glyph representing a single student. Males are shown by circle glyphs and females by rectangles glyphs. The father's education level is mapped to the glyph stroke color, and the mother's education level to the glyph fill

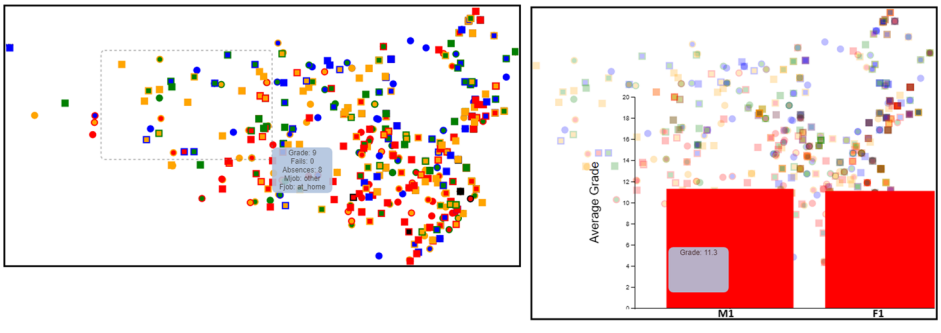

Fig. 9 Isomap visualization of "Portuguese" course. Different levels of education are color coded. Female students are rectangles and male students circles. Mother' education is represented by glyphs colors and father' education is represented by glyphs strokes. A selection on Isomap shows the correspondent average of grades for male and female students

color. These categorical information are mapped as following: Black: No education; Red: Initial education; Orange: Primary education; Green: Secondary education; Blue: Higher education. It is possible to filter data points by gender, and to select specific offered course periods. This visualization is available in URL [5].

We performed several analysis using the resulting layouts. We first focused in analyzing the overall glyphs distribution, in order to identify and quantify groups of students with similar behavior regarding scholar performance. We then investigated these groups to comprehend which attributes "describe" then, and which ones distinguish them from other groups. We also looked for isolated glyphs representing **outliers** students, that is, students whose behavior significantly diverge from the remaining ones, as these students may represent interesting investigation cases. The results of these analysis are discussed in "Isomap: parent education analysis" section.

Results

This section presents the results of applying our proposed analysis strategy in several educational scenarios, using the previously presented data collections. The idea is to show the applicability of our designed tool in analyzing these academic data.

We first performed the analysis of the predictive models using Collection 1 and considering different subjects, comparing the predicted results with the original data, in order to verify their ability to predict students' future grades. We also employ our proposed visualizations to evaluate the capability of the layouts in improving the analysis process, by highlighting the discrepancy between original and predicted values, by evaluating the relationship between students' attendance and grades, as well as by revealing important behavior patterns associated to the students performance.

We also employed the layouts to investigate the role of the students' gender on their academic performance, also taking into account a statistical analysis of the difference between these gender performances. The gender was also considered in an analysis to evaluate the proportion of male and female students in different subjects, as well as the students distribution on a specific subject over time and the proportion of students in several subjects in same and different semesters.

Finally, Isomap was employed in the analysis of Collection 2. We first analyzed the points distribution in the layout, in order to identify groups of students with similar behavior, as well as students with anomalous behavior, represented as outlier points. We then explored these groups, as well as other interesting patterns revealed by the layout, in order to comprehend how each attribute influences the characterization and formation of these groups, as well as to identify correlations among these attributes, the distribution of students among attributes, among other tasks. The following sections present some of our findings.

Heatmap matrix: attendance vs. grade analysis

When investigating the relationship between students' attendance and grades using the Heatmap Matrix visualization, we notice that, for some courses such as "Algorithms and Programming I" & "II" & "III", as well as in "Discrete Mathematics I", students who received higher grades attended either all the sessions or missed only few of them, as shown in Fig. 10.

[5] http://fall-2018-education.herokuapp.com/isomap2

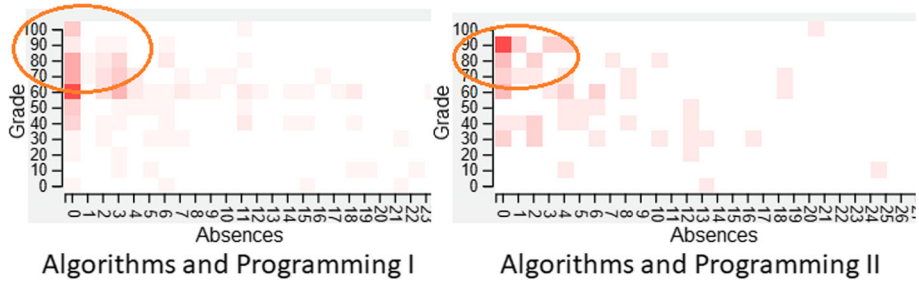

Fig. 10 Attendance vs. grade Heatmap Matrices for "Algorithms and Programming I" and "Algorithms and Programming II" subjects. Students who received high grades attended most of the sessions

In contrary, in courses such as "Data Structures" and "Logic for Computer Science I", sequentially approx. 28.57% and 18.18% of students have high grades with high number of absences, as shown in Fig. 11.

Using different machine learning models, the aim was to develop a predictive model that can derive a conclusion on students' academic success (grade here) ("Academic visual analysis system" section). To make it easier on the eye to compare original data vs. the predictive models, the Heatmap Matrix layouts are placed on top (real data) and bottom (predicted data).

In regard to other findings shown in Fig. 10, the Heatmap Matrix visualization revealed that on some courses such as "Algorithms and Programming", "Discrete Mathematics" and "Logic for Computer Science", that are offered in the first years of Computer Science program, cells with the highest color intensity correspond to approximately 7 or 8 days of absence while their associated grades are around 40% or 50%. It indicates that a student who is absent for at least 7 days will strongly tend to receive a grade as low as 50% or less. These results could be either because of the nature of these courses - they are traditionally known as difficult courses by students from this university, or students' lack of enough knowledge on new Computer Science topics and required study routines. However, for more advanced courses such as "Human-Computer Interaction", that are offered in final semesters, students are more experts and one notice that even less number of students' attendances concluded with higher grades.

We highlight some of the predictive models that show similar patterns to the original data. Figure 13 shows that SVR linear analysis for "Algorithms and Programming I" course correctly predicted some of the grades based on their absences: 1) grade 60 with less than 6 absences; 2) lower grades with higher number of absence. Figures 12 and 13

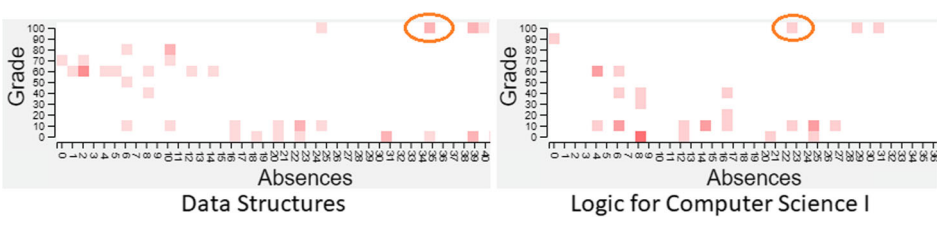

Fig. 11 Attendance vs. grade Heatmap Matrices for "Data Structures" and "Logic for Computer Science I", in which 18.18% and 28.57% of the students received high grades and had high number of absences, respectively

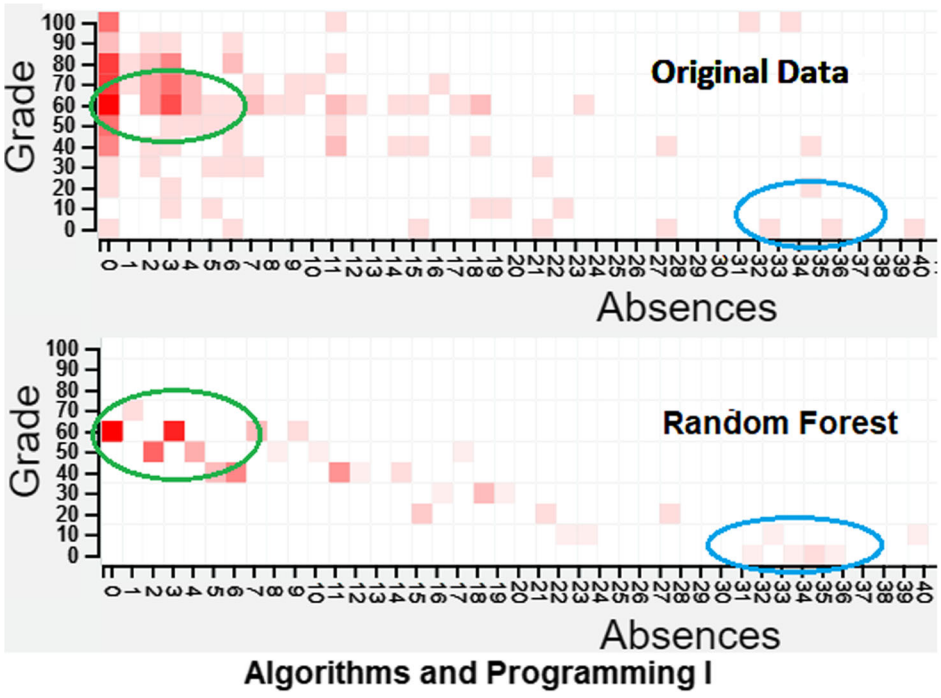

Fig. 12 Algorithms and Programming I: Original data vs. Random Forest prediction

show that SVR and Random forest made more accurate predictions for "Algorithms and Programming I" course compared to all other methods.

As shown in Fig. 14, one notice that the percentage of students with less absence could be approved in "Discrete Mathematics II" with highest grades. This pattern was reproduced by the linear regression. The lowest grades has been correctly

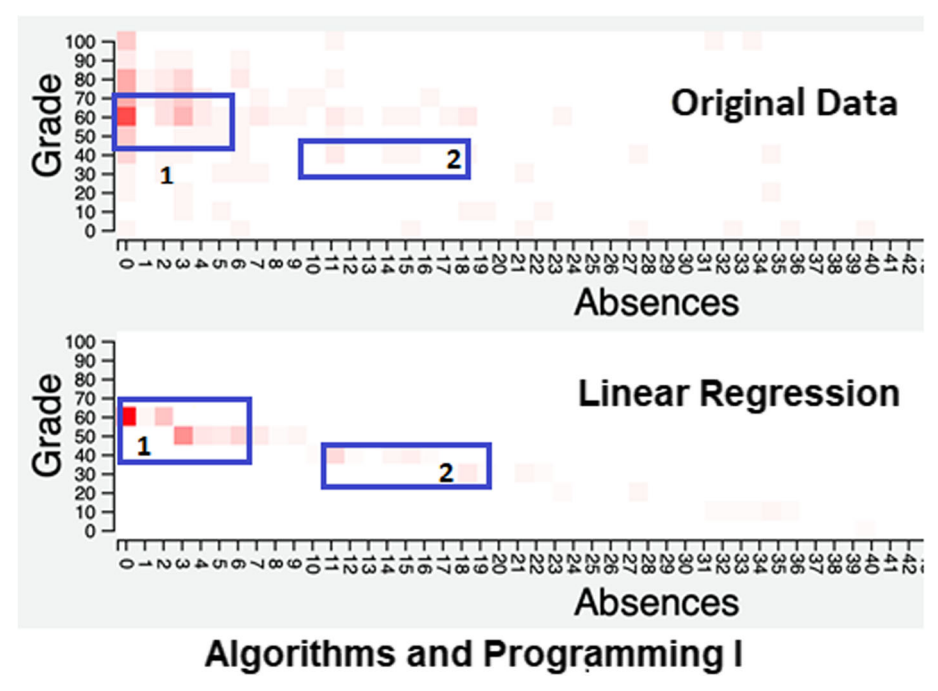

Fig. 13 Algorithms and Programming I: Original data vs. SVR linear prediction

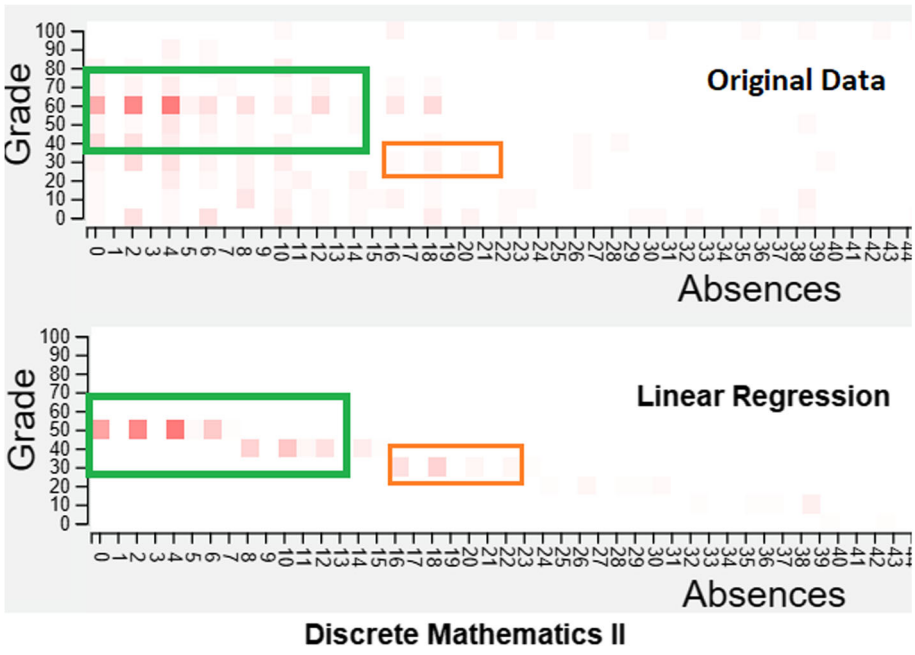

Fig. 14 Discrete Mathematics II: Original data vs. Linear Regression prediction

predicted for highest absences. SVR RBF, Ridge Regression, and Elastic Net presented similar predictions.

Figure 15 shows that the application of the Decision tree predictive model produced similar patterns when compared to the original data. As a result, students with more than 16 absences in "Data structures" present high probability to receive a grade of 10 or less. The scatter plot analysis shown in Fig. 16 confirms these findings.

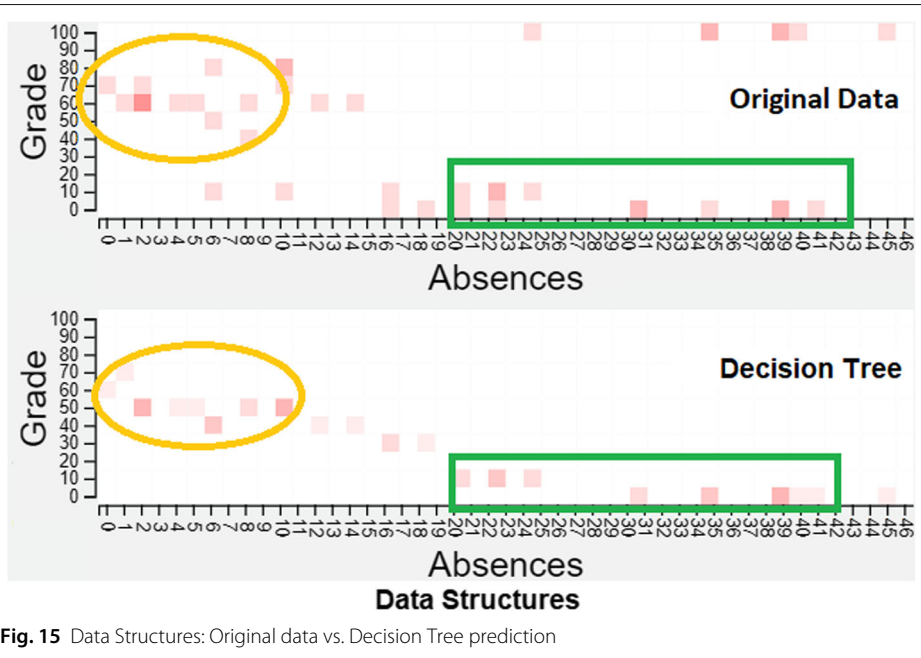

Fig. 15 Data Structures: Original data vs. Decision Tree prediction

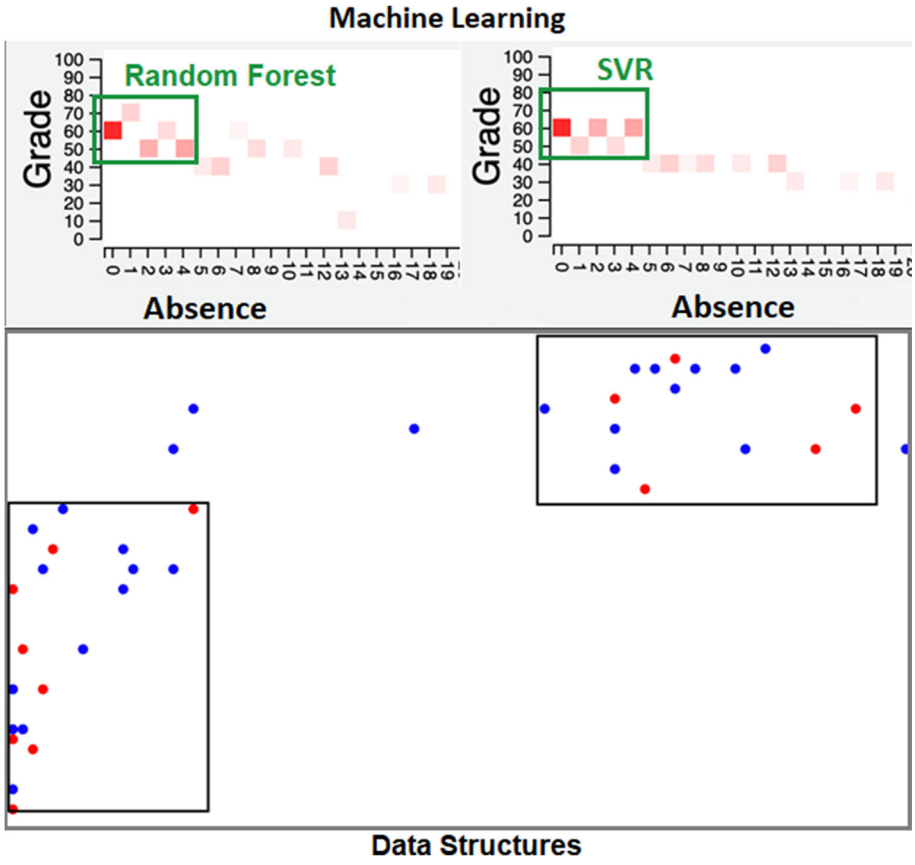

Fig. 16 Two clusters of students: the left one are composed of students with less absence and higher grades; the right one is composed of students with more absences and lower grades. These two patterns are also predicted in the layouts shown in Fig. 15

We were not able to find a single model that fits the best for all the courses as our prediction was only based on the number of days of absence. In future investigations, we intend to identify and evaluate more variables associated with the students' performance and the studying process, including lecturer evaluations, students' previous academic performance on elective or core courses, etc. We believe that a large number of variables will produce a more generalized model.

Scatter plot: attendance vs. grade analysis

The Matrix heatmap for "Algorithms and Programming II", shown in Fig. 10, revealed that students with up to three absences can still receive a high grade, and the predictors were not able to predict this information with accuracy. However, Random Forest and SVR revealed that students with zero or one absence present high probability to receive the highest grades. The highest grade predicted by Random Forest was 60, which suggests that additional factors may be influencing the grades. Moreover, the low number of students that received 90 or above grades in the training set might have contributed to this result.

The scatter plot for "Algorithms and Programming III" revealed that students with 8 or more (up to 12) absences received the lowest grades. This information has been predicted

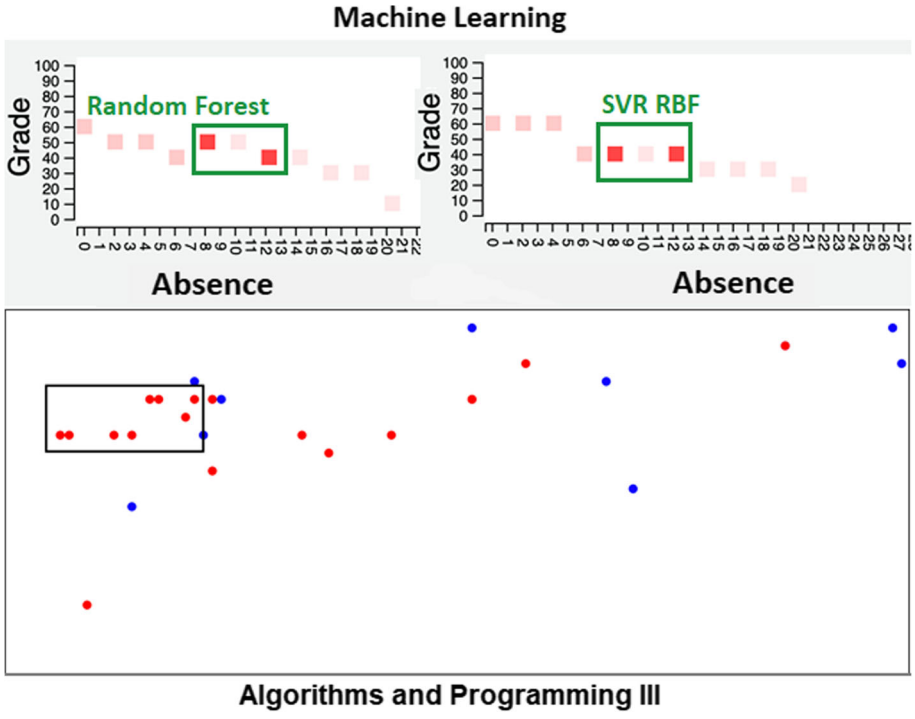

Fig. 17 Machine learning prediction and scatter plot for "Algorithms and Programming II". The scatter plot shows that a lower number of absences lead to better grades and this is predicted with high probability in both Random Forest and SVR

by Random Forest, SVR RBF, and Decision Tree (Fig. 17). The scatter plot also revealed that female students often had the highest number of absences and less grades compared to male students. The region of the layout which presents this pattern is highlighted in Fig. 17.

The scatter plots show that, for "Algorithms and Programming I", if a student is never absent, then he or she will most likely to receive a grade between 65 and 71. For "Algorithms and Programming II", if a student is never absent, then he or she will most likely to receive a grade above 90. For "Discrete Mathematics I", we were not able to conclude anything because the clusters are too scattered. For "Discrete Mathematics II", it seems that the attendance does not matter, and the students will most likely to receive a grade in the 60s. For "Logic for Computer Science II", if a student is never absent, then he or she will receive a grade in the 70s. For "Object Oriented Programming", it seems that if a student has 8 or less absences, then he or she will receive a grade from 60 to 80. Finally, for "Human-Computer Interaction", attendance does not seem to influence on grades, and the student will receive a high grade regardless.

In summary, the scatter plots analysis revealed that students performance is highly dependent of each subject, and that for some subjects there is no correlation between attendance and grade. However, in most cases, the layouts suggest that students with few or no absence tend to receive a grade in at least 65, which in this university if sufficient for approval. The distribution of male and female students on the scatter plots did not produce any significant pattern, in most cases, and we were not able to make any conclusion regarding this information.

Gender vs. grade analysis

The aim of this analysis is to investigate if there is any correlation between students' gender and their grades. Comparing Fig. 4a and b, we notice that the heatmap cells for male students are darker than the associated ones for female students. This occurs for most of the Programming related courses and may suggest gender related differences. However, looking at Fig. 4c, which considers statistical relationships calculated from Wilcoxon *p*-values, we notice that except in four classes, the gender does not show any statistically significant difference in students' performance. A deeper investigation on the data is necessary to identify the reasons for these significant differences. When clicking on a heatmap cell a bar chart is displayed which shows the distribution of students from both genders, as well as their grades. For example, when analyzing the bar charts for "Object-Oriented Programming" in the first semester of 2014 (Fig. 18), one notices that there are less female students compared to male students. Moreover, it is possible to notice that, although a female student received the second highest grade, on average there are more male students with high grades than female students. These layouts quickly communicate these differences as well as their extent over one or several semesters, allowing analyzers to perform a more focused analysis.

In terms of student populations in classes (Fig. 5), Treemap visualization shows that courses that are offered in the first semesters of Computer Science program are highly more populated than the courses that are offered in the last semesters. This can be expected as all students starting these programs have to take some preliminary subjects. The subjects in the next semesters have prerequisite classes that are taken in the earlier semesters. Thus, they tend to have less students as the enrollments are subject to the satisfactory in a prior semester. However, the layout shows some interesting distribution patterns revealing potential abnormal scenarios. For example, we noticed that "Discrete Mathematics" present a larger population in the second semester of 2010 than the first semester. Treemap visualization gives a quick insight to the program coordinators or professors to investigate these specific patterns.

Isomap: parent education analysis

This study aims at investigating if factors other than grades and absence might affect the students' performance. As previously discussed, we employed Isomap (Tenenbaum et al.

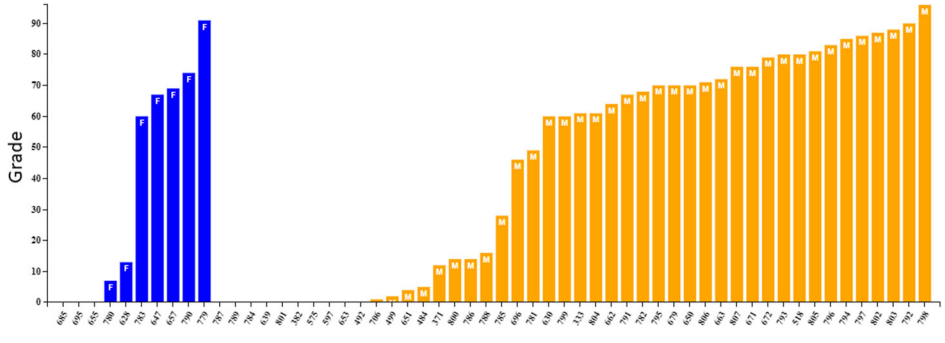

Fig. 18 Bar chart analysis of "Object-Oriented Programming" subject in 2014/01 semester. This Bar chart is presented when users click a cell on Heatmap Matrix visualization of Gender vs. Grade. Female students are shown by blue bars and male students by yellow bars

2000) multidimensional projection on Collection 2 data, resulting on the layouts shown in Figs. 19, 20, 21 and 22, and we found interesting patterns. We highlighted several areas of interest (AOI) to illustrate such patterns. AOI1 shows a cluster of female students who received high grades (*Mean* = 15.23 out of 20) and at least one of their parents has higher or secondary education. In contrary, AOI2 shows a cluster of female students who received low grades (*Mean* = 7.11) and most of their parents has either no education, initial, or primary education. Naturally, these results cannot be generalized to a conclusion that every student whose parent has higher education has a better performance, specially because in AOI3 and AOI4 we notice a cluster of students whose parents are mostly highly educated but their grades are low. Also, there are some exceptional cases that their parents' education level is low, but have very good grades. One of them is indicated by a green horizontal arrow (*Grade* = 19).

However, looking deeply in some students with low grades, indicated with black vertical arrows on Isomap, we notice that their mothers' education level is as low as initial, which may raise some social, behavioral questions about the mothers' role on female students' academic performance. Using Isomap only for male students (Fig. 20), we notice that AOI1 is a cluster of students with high grades (*Mean* = 16.38) and with most of their parents having a higher educational level, while in AOI2 we notice a cluster of students with lower grades (*Mean* = 7.18) and with their parents having a lower education level. Again some exceptions can be noticed in the layout. Figure 21 shows a cluster of students in AOI1 with relatively higher grades than the cluster of students in AOI2. Looking into their parents' education on notice that at least one of the parents in AOI1 has a higher level of education.

These results however shall not be generalized. When analyzing *Portuguese* class, period 2 (Fig. 23), one notices a cluster of students with high grades, but whose parents have initial educational level.

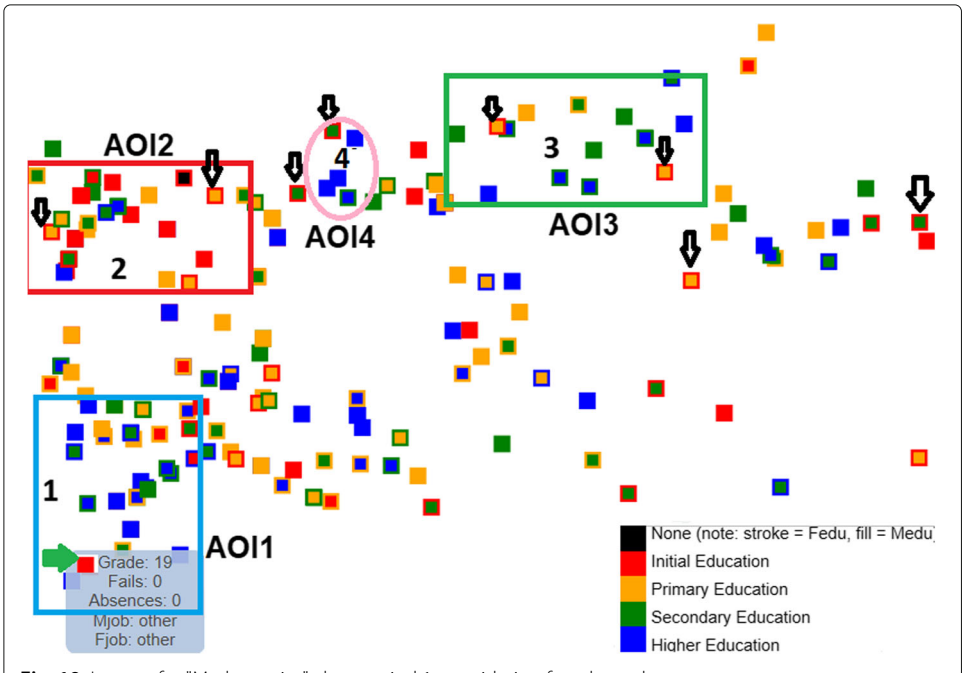

Fig. 19 Isomap for "Mathematics" class, period 1, considering female students

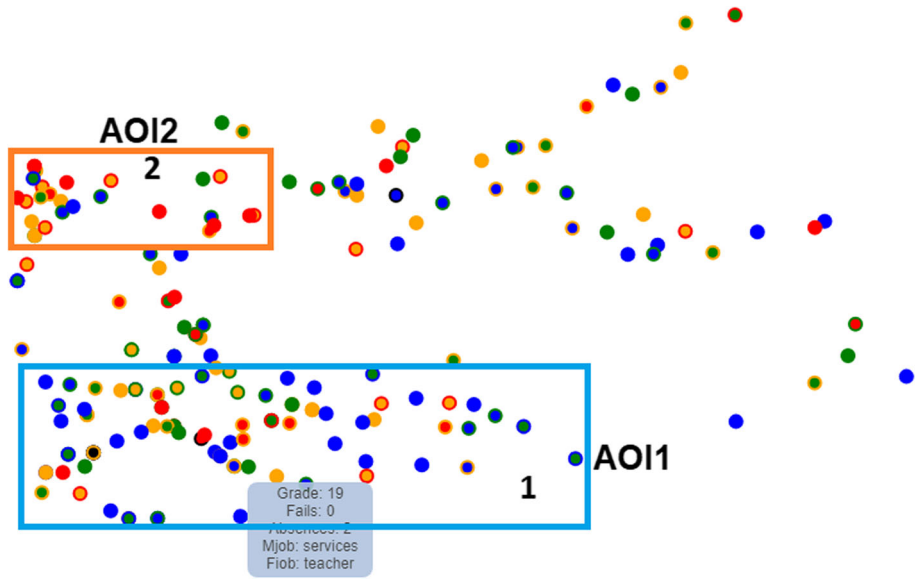

Fig. 20 Isomap for "Mathematics" class, period 1, considering male students

Similarly, other findings were found when analyzing other courses and classes, as shown in Figs. 24 and 25. In summary, the Isomap layouts allowed the identification of groups of students which present similar performance behavior. This visualization utilizes an educational expert to investigate which attributes contribute to forming a group of similar behaviors. This visualization helps to better comprehend specific educational scenarios that could be addressed later by finding possible related issues.

Logistic regression: success prediction

As detailed in "Logistic regression & pass/fail analysis" section, we used a threshold of 2.3 for the GPA analysis. Figure 26 demonstrates a true positive prediction achieved by logistic regression analysis for GPA between 3.5 and 4. The total number of predicted students within this GP range was 198 students, from which 56 are female students, while in the original data the total of number of students within this range of GPA is 278, from which 71 are female students.

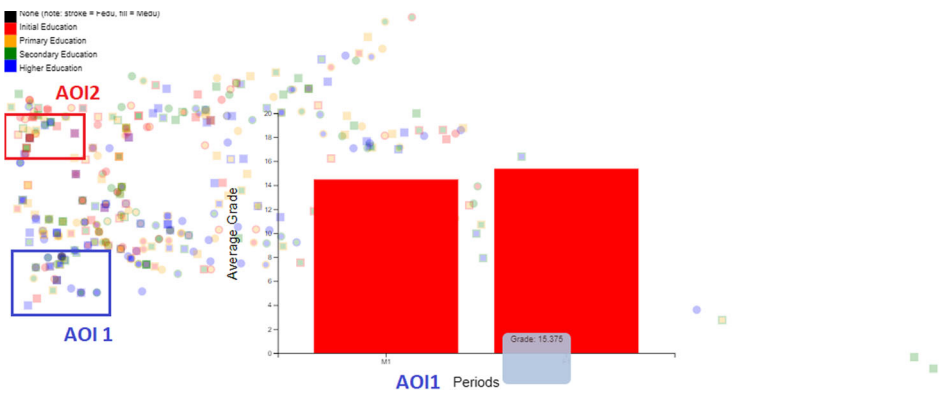

Fig. 21 Isomap for "Mathematics" class, period 1, considering both genders

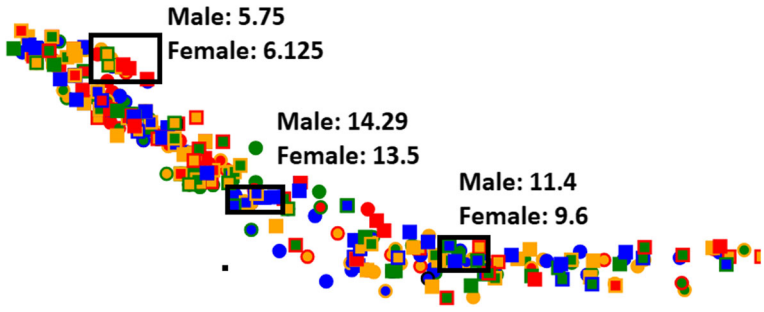

Fig. 22 Isomap for "Mathematics" class, period 2, considering both genders

Furthermore, Fig. 27 shows a true negative prediction achieved by logistic regression for GPA between 0 and 0.4. The total number of predicted students within this GPA range who failed is 94 students, from which 7 are female students, while in the original data the total of number of students within this GPA range is 101, from which 11 are female students.

Conclusion and future works

We developed a web-based visual analytics system that includes several visualization approaches to analyze academic data. Our initial results are reasonable enough to make sense out of this type of multivariate data that helped us to gain insight for better future decision making in the academic environment. We were able to show the relationships between grades and attendance, grades and genders, grades and parents' educational level and students populations and genders. Additionally, we implemented a variety of machine learning models that predict the performance of students based on their attendance (absence rate), and we employed a visualization technique to check the accuracy of different models.

However, to have a deeper insight on the educational data and academic performances, it is necessary to complement the analysis with more social psychological investigations, that are currently beyond our scope. In the future, we intend to collect and analyze a larger volume of data, in order to refine our findings and extract more valuable information

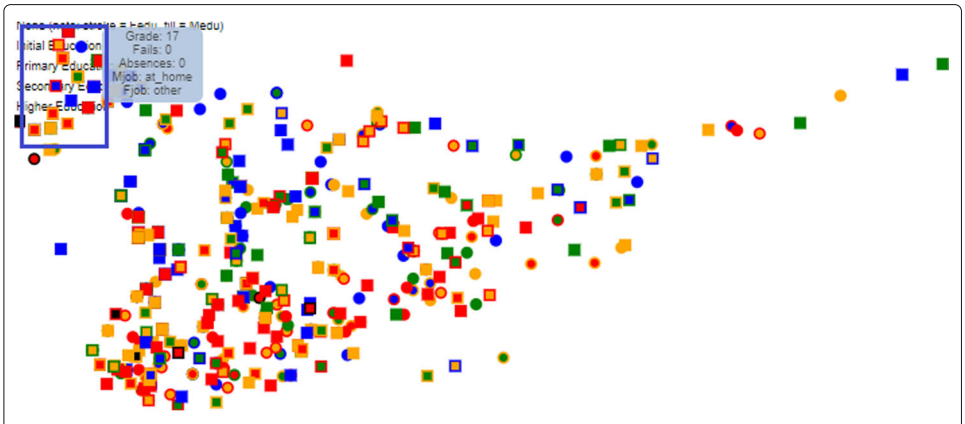

Fig. 23 Isomap for "Portuguese" class, period 2, considering both genders

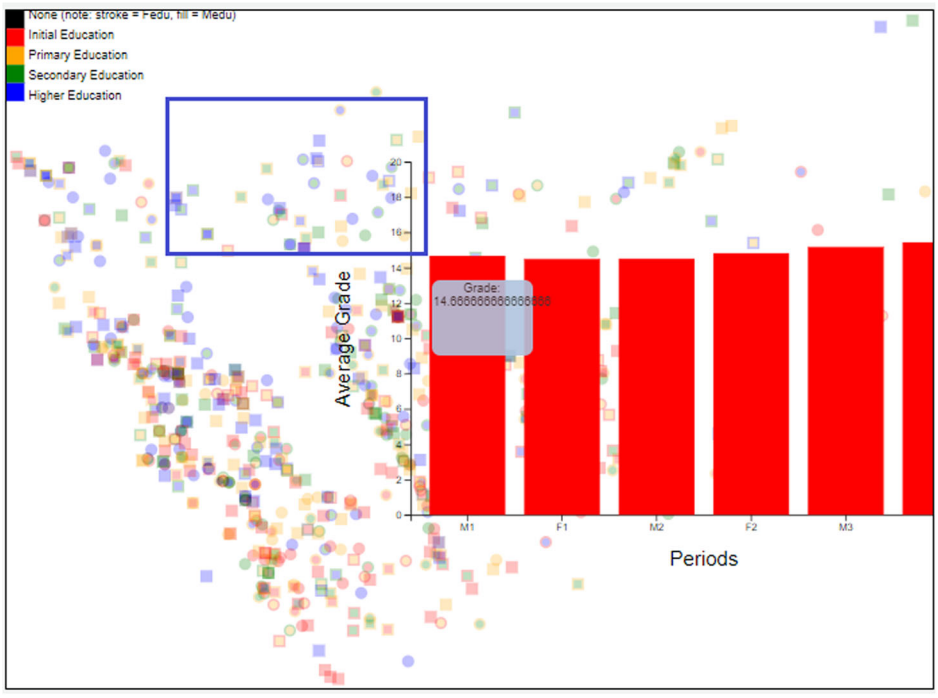

Fig. 24 Isomap for "Portuguese" class, all periods, considering both genders

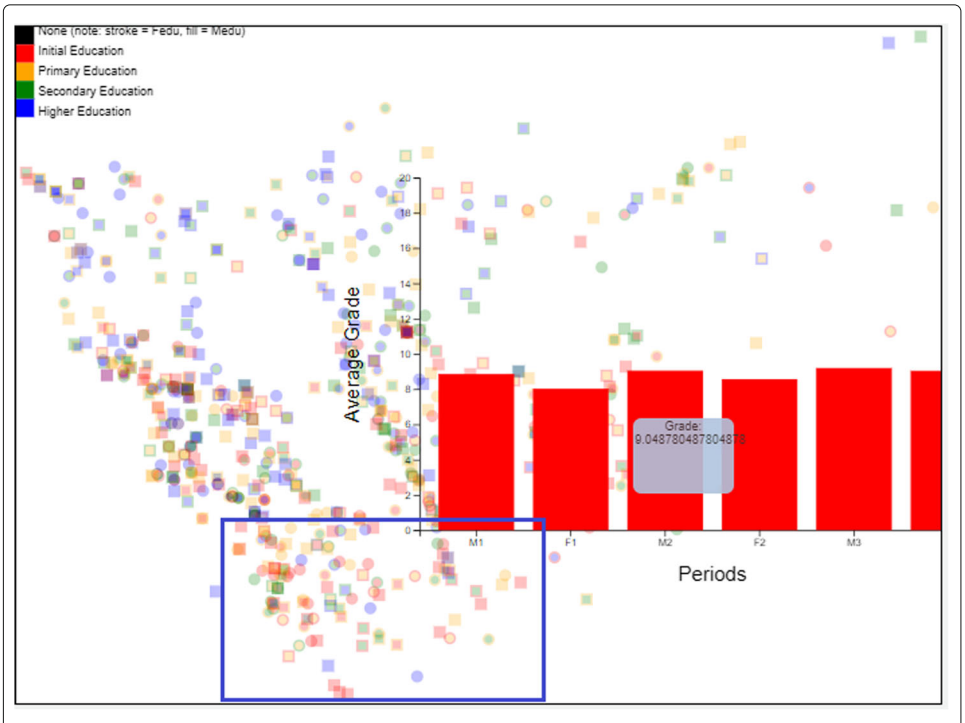

Fig. 25 Isomap for "Portuguese" class, all periods, considering both genders, and a group selection

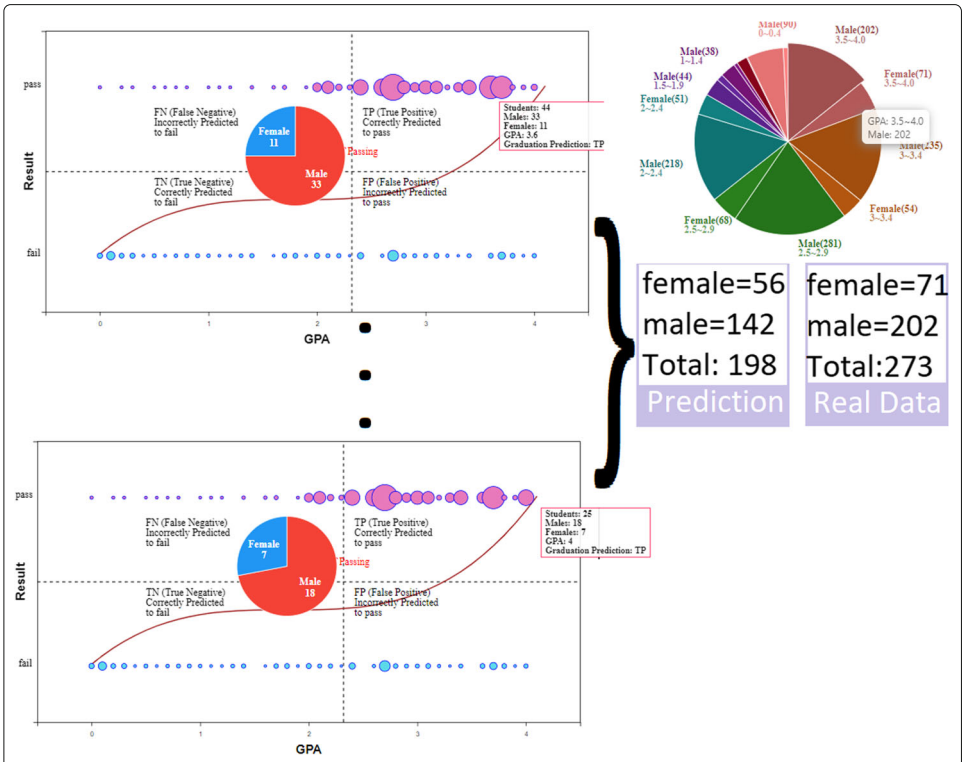

Fig. 26 Logistic Regression analysis for True positive prediction (passed successfully and correctly predicted)

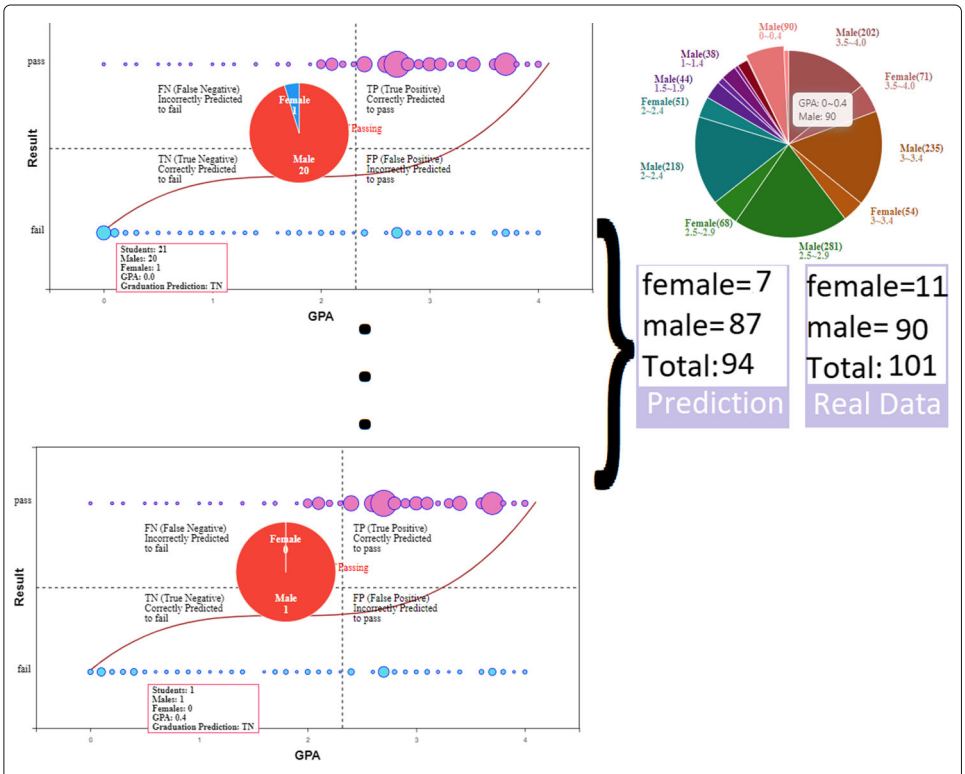

Fig. 27 Logistic Regression analysis for True negative (failed and correctly predicted)

to improve our tool. We also intend to add more visualization and interaction techniques, possibly incorporating more machine learning approaches, in order to provide more analysis tasks, and improve even more the analysis process.

Abbreviations
GPA: Grade point average

Acknowledgements
Not applicable.

Authors' contributions
The authors contributed equally to the effort. All authors read and approved the final manuscript.

Funding
No funding was received for this project.

Availability of data and materials
Two data collections were used in our experiments. For Collection 1 ("Collection 1: exact science subjects performances" section), the authors do not have ethics approval to make the raw student data available to anyone outside the organization in which the experiment was conducted. Going back to ethics committee to request this is not a feasible option at this stage, as the experiment is already completed. The authors have already requested this in the ethics application for future submissions. Collection 2 ("Collection 2: public school performances" section) is publicly available in Cortez and Silva (2008).

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Computer Science, City College of New York, CUNY, 160 Convent Avenue, 10031 New York, USA. [2]Faculty of Computing, Federal University of Uberlandia, UFU, Av. Joao Naves de Avila, 2121, 38408100 Uberlandia, Brazil.

Received: 12 November 2019 Accepted: 5 December 2019
Published online: 07 January 2020

References

A. R. Anaya, M. Luque, M. Peinado, A visual recommender tool in a collaborative learning experience. Expert Systems with Applications. **45**, 248–259 (2016)
L. Breiman, Random forests. Springer. **1**(45), 5–32 (2001)
P. Cortez, A. M. G. Silva, *Using data mining to predict secondary school student performance*. (EUROSIS-ETI, Ostend, 2008)
M. B. DeCotes, Data analytics of university student records (2014). Master's Thesis, University of Tennessee
T. S. Dee, A teacher like me: Does race, ethnicity, or gender matter? American Economic Review. **95**(2), 158–165 (2005)
H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, V. Vapnik, in *Advances in Neural Information Processing Systems*, Support vector regression machines (MIT Press, Cambridge, 1997), pp. 155–161
R. Etemadpour, R. Motta, J. G. de Souza Paiva, R. Minghim, M. C. F. de Oliveira, L. Linsen, Perception-based evaluation of projection methods for multidimensional data visualization. IEEE Transactions on Visualization and Computer Graphics. **21**(1), 81–94 (2015)
E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, G. V. Erven, Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. Journal of Business Research. **94**, 335–343 (2019). https://doi.org/10.1016/j.jbusres.2018.02.012
E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, G. Van Erven, Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. Journal of Business Research. **94**, 335–343 (2019)
E. Fix, J. Hodges Jr, Significance probabilities of the wilcoxon test. The Annals of Mathematical Statistics, 301–312 (1955). https://doi.org/10.1214/aoms/1177728547
Y. Freund, R. E. Schapire, Large margin classification using the perceptron algorithm. Machine Learning. **37**(3), 277–296 (1999)
Y. Gooding, The relationship between parental educational level and academic success of college freshmen (2001). https://doi.org/10.31274/rtd-180813-12012
M. Gruber, *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*. (Routledge, 2017). https://doi.org/10.1201/9780203751220
L. Gutiérrez, V. Flores, B. Keith, A. Quelopana, Using the belbin method and models for predicting the academic performance of engineering students. Computer Applications in Engineering Education. https://doi.org/10.1002/cae.22092
S. Haykin, *Neural Networks and Learning Machines, 3/E*. (Pearson Education India, Chennai, 2010)
B. Johnson, B. Shneiderman, in *Visualization, 1991. Visualization'91, Proceedings., IEEE Conference On*, Tree-maps: A space-filling approach to the visualization of hierarchical information structures (IEEE, 1991), pp. 284–291. https://doi.org/10.1109/visual.1991.175815
D. J. Jones, *The impact of student attendance, socio-economic status and mobility on student achievement of third grade students in title i schools. PhD thesis*. (Virginia Tech, Blacksburg, 2006)

V. Kassarnig, A. Bjerre-Nielsen, E. Mones, S. Lehmann, D. D. Lassen, Class attendance, peer similarity, and academic performance in a large field study. PloS ONE. **12**(11), 0187078 (2017)
J. King, T. Bond, S. Blandford, An investigation of computer anxiety by gender and grade. Computers in Human behavior. **18**(1), 69–84 (2002)
J. Klerkx, K. Verbert, E. Duval, in *Handbook of Research on Educational Communications and Technology*, Enhancing learning with visualization techniques (Springer, 2014), pp. 791–807. https://doi.org/10.1007/978-1-4614-3185-5_64
S. A. Kumar, et al., *Efficiency of decision trees in predicting student's academic performance*. (AIRCC Publishing Corporation, Chennai, 2011)
W. E. Lacefield, E. B. Applegate, in *Annual meeting of the American Educational Research Association*, Data visualization in public education: Longitudinal student-, intervention-, school-, and district-level performance modeling (Poster, New York, 2018)
A. Liaw, M. Wiener, Classification and regression by randomForest. R news. **3**(2), 18–22 (2002)
D. M. Magerman, in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Statistical decision-tree models for parsing (Association for Computational Linguistics, 1995), pp. 276–283. https://doi.org/10.3115/981658
K. D. Mattingly, M. C. Rice, Z. L. Berge, Learning analytics as a tool for closing the assessment loop in higher education. Knowledge Management & E-Learning. **4**(3), 236 (2012)
E. Osmanbegović, M. Suljić, Data mining approach for predicting student performance. Economic Review. **10**(1), 3–12 (2012)
S. Ruffing, F.-S. Wach, F. M. Spinath, R. Brünken, J. Karbach, Learning strategies and general cognitive ability as predictors of gender- specific academic achievement. Frontiers in Psychology. **6**, 1238 (2015). https://doi.org/10.3389/fpsyg.2015.01238
P. Soule, *Predicting student success: A logistic regression analysis of data from multiple siu-c courses*. (OpenSIUC, Carbondale, 2017)
J. B. Tenenbaum, V. De Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction. science. **290**(5500), 2319–2323 (2000)
K. Thompson, D. Ashe, L. Carvalho, P. Goodyear, N. Kelly, M. Parisio, Processing and visualizing data in complex learning environments. American Behavioral Scientist. **57**(10), 1401–1420 (2013)
R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology). **73**(3), 273–282 (2011)
M. E. Urrutia-Aguilar, R. Fuentes-García, V. D. M. Martínez, E. Beck, S. O. León, R. Guevara-Guzmán, Logistic regression model for the academic performance of first-year medical students in the biomedical area. Creative Education. **7**(15), 2202 (2016)
A. Van Barneveld, K. E. Arnold, J. P. Campbell, Analytics in higher education: Establishing a common language. EDUCAUSE Learning Initiative. **1**(1) (2012)
S. Weisberg, *Applied linear regression*. (Wiley, New Jersey, 2005)
B. C. Wilson, S. Shrock, in *ACM SIGCSE Bulletin*, Contributing to success in an introductory computer science course: a study of twelve factors, vol. 33 (ACM, 2001), pp. 184–188. https://doi.org/10.1145/364447.364581
I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. (Morgan Kaufmann, 2016). https://doi.org/10.1016/c2009-0-19715-5
Y. Yang, J. Li, Y. Yang, in *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, The research of the fast svm classifier method, (2015), pp. 121–124. https://doi.org/10.1109/ICCWAMTIP.2015.7493959
H. Zou, T. Hastie, Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). **67**(2), 301–320 (2005)
H. Zou, T. Hastie, Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). **67**(2), 301–320 (2005). Wiley Online Library

Publisher's Note