Heliyon 6 (2020) e04081

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country

Frederico Cruz-Jesus *, Mauro Castelli, Tiago Oliveira, Ricardo Mendes, Catarina Nunes, Mafalda Sa-Velho, Ana Rosa-Louro

NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, Campus de Campolide, 1070-312, Lisboa, Portugal

ARTICLE INFO

Keywords:
Education
Applied computing
Information systems
Data analysis
Evaluation in education
Teaching research
Achievement
Education reform
Quantitative research
Artificial intelligence
Data science

ABSTRACT

Understanding academic achievement (AA) is one of the most global challenges, as there is evidence that it is deeply intertwined with economic development, employment, and countries' wellbeing. However, the research conducted on this topic grounds in traditional (statistical) methods employed in survey (sample) data. This paper presents a novel approach, using state-of-the-art artificial intelligence (AI) techniques to predict the academic achievement of virtually every public high school student in Portugal, i.e., 110,627 students in the academic year of 2014/2015. Different AI and non-AI methods are developed and compared in terms of performance. Moreover, important insights to policymakers are addressed.

1. Introduction

The Europe 2020 Strategy aims at tackling "the problem of early school leavers by reducing the dropout rate to 10% from the current 15%, whilst increasing the share of the population aged 30–34 having completed tertiary education from 31% to at least 40% in 2020". As the European Commission recently stated, "early school leaving is an obstacle to economic growth and employment. It hampers productivity and competitiveness, and fuels poverty and social exclusion" (European Commission, 2017). Presently, school dropout is one of the most common forms of school failure. The European average stands at 10.7%.

Understanding the drivers behind academic achievement (AA) is an everlasting global challenge that concerns students, their families, and teachers, but also public decision-makers, and everyone concerned about development and wellbeing at a global level (Noell et al., 2019; Valli Jayanthi, Balakrishnan, Lim Siok Ching, Aaqilah Abdul Latiff and Nasirudeen, 2014). The importance of AA on the general development of regions, countries, and civilization, in general, is a material that has long concerned policymakers and researchers (Hattie, 2009). AA is associated with human capital, benefiting individuals and organizations as it promotes the spread and transmission of information, culture, and knowledge. AA is proven to be positively associated with economic growth and salary increases, engendering overall development. It is also an inhibitor of social exclusion as it promotes social progress, especially for those that are socioeconomically disadvantaged (Dronkers et al., 2012).

Despite the importance of AA, most research conducted on this stream has employed traditional (statistical) methods on sample data (Hattie, 2009). Thus, despite its positive potential on AA (Dunn et al., 2013; van der Scheer and Visscher, 2018), recent developments and interest in artificial intelligence (AI) data science or big data fields (Sivarajah et al., 2017) have not been harnessed in this truly global challenge, so crucial in terms of economic development, equality, and wellbeing. Once these (new) streams of methods are employed in the context of AA, policymakers, parents, and teachers are better equipped to engender education and reduce school dropout rates. Education is indisputably the best tool society has for development and equality. We expect this paper to make a significant contribution in this regard as, to the best of our knowledge, this is one of the first studies that does so. Thus, we intend to answer the following research questions:

1. How do AI methods compare against each other, and traditional ones, in predicting high-school AA?
2. What are the most important high-school AA drivers on a national scale?

* Corresponding author.
E-mail address: fjesus@novaims.unl.pt (F. Cruz-Jesus).

https://doi.org/10.1016/j.heliyon.2020.e04081
Received 21 January 2020; Received in revised form 4 May 2020; Accepted 22 May 2020
2405-8440/© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

In answering these questions, this paper is organized as follows: Section two presents a literature review on AA; Section three details the methodology employed; Section four presents the results, whereas Section five the discussion and implications. Finally, Section six provides conclusions and future work.

2. Literature review

In recent years, big data and artificial intelligence methods have been given great attention due to their potential to engender development and wellbeing at individual-, firm-, and societal levels. In a world where data is widely available, new, and more efficient ways of analyzing it are of paramount importance (Delen and Zolbanin, 2018). There is growing evidence that big data and artificial intelligence methods may yield several benefits such as firms' performance (Côrte-Real et al., 2019), innovativeness (Ghasemaghaei and Calic, 2019), marketing efficiency (Erevelles et al., 2016), and education (Hattie, 2009; van der Scheer and Visscher, 2018).

Accordingly, assessing AA is an important topic as engendering individuals' education, in general, is amongst the most important global challenges nowadays, both in developed and developing countries. As with other global challenges (Choi et al., 2018), despite the (limitation in terms of) employed (traditional) methods, AA antecedents are a fairly explored research subject. One of the first studies to address the drivers of AA was the "The Coleman Report," which argued that amongst the key antecedents were students' family characteristics, but also the other students in school and their respective backgrounds, i.e., the students' environment (Coleman and Hopkins, 1966). However, as time went by, new findings were added, notably the consideration of students' teachers. Hence, Greenwald et al. (1996) posited that aspects such as teachers' education, experience, along with smaller classes, were positively associated with AA. From the literature, it is noticeable that, overall, AA antecedents may be comprised of three natures - students', parents' and, schools' characteristics.

Students' characteristics have continually been identified as the key antecedent of AA (please see Table 1), which seems plausible since students themselves should be the main stewards of their AA. As examples of students' characteristics, there is evidence that gender is important as females usually perform better than males in academic results (Mensah and Kiernan, 2010; Torrecilla Sánchez, Olmos Miguélañez and Martínez Abad, 2019), depending, nevertheless, upon the specific scientific areas of studies. Ethnicity and cultural background have also been marked as characteristics that influence AA (Ahmed et al., 2019; Avery and Walker, 1993; OECD, 2012). Although with opposing results, technology adoption is also mentioned as a relevant driver of AA, as some authors argue that personal computer (PC) and Internet access positively affects AA while others argue the reverse (Kubey et al., 2001). In a recent study, Huang (2018) conducted a meta-analysis on the effect that social network site (SNS) use yields on academic achievement, arguably one of the most debatable aspects in AA literature (Wakefield and Frawley, 2020). Their findings suggest that, in fact, SNS may yield a negative effect on AA, but it is minimal, especially in the case of Facebook.

Parents' characteristics and their involvement have also been long identified as key drivers of AA (see, e.g., Fan, 2001; Torrecilla Sánchez et al., 2019), with some recent studies indicating that the effect of mothers and fathers differ (see, e.g., Otani, 2019). One of the most

Table 1. Previous studies addressing academic achievement.

References	Methods	St	Pa	Sc
(Hanushek and Kimko, 2000)	Regression models	x		x
(Hoxby, 2000)	Regression models	x		x
(Fan and Chen, 2001)	General linear model	x	x	
(Barnett et al., 2002)	Linear Programming techniques			x
(Driessen et al., 2005)	Frequency, Variance, and Structural models	x	x	x
(Rivkin et al., 2005)	Regression models			x
(Archibald, 2006)	Hierarchical linear models	x		x
(Jackson et al., 2006)	Internet recorded	x		
(Lee and Bowen, 2006)	Hierarchical linear model	x	x	
(Marks et al., 2006)	Item Response Theory; Regressions models	x	x	x
(Jeynes, 2007)	Regression models		x	
(Codjoe, 2007)	Interviews	x		
(Croninger et al., 2007)	Hierarchical linear models	x		
(Lee, 2007)	Hierarchical linear models; Regression models	x	x	x
(Lei and Zhao, 2007)	Hierarchical linear models; ANOVA tests	x		
(Steinmayr and Spinath, 2008)	Regression models	x		
(Caro et al., 2009)	Hierarchical linear models; Panel data models	x		
(Mensah and Kiernan, 2010)	Tobit regression models; Univariate and Multivariate analyses	x	x	
(Hartas, 2011)	Univariate analyses of variance; Chi-square tests		x	
(Patterson and Pahlke, 2011)	Regression models	x	x	
(Hanushek and Woessmann, 2012)	Regression models	x		x
(S. Huang and Fang, 2013)	Regression model, Artificial Neural Networks, Radial Basis Function, and Support Vector Machines.	x		
(Brunner et al., 2013)	Multiple group factor analytic models; Full maximum likelihood	x		
(Wally-Dima and Mbekomize, 2013)	Descriptive statistics T-tests	x		
(Bosworth, 2014)	Regression models	x		x
(Krassel and Heinesen, 2014)	Regression discontinuity design; Control for school fixed effects; Regression models	x	x	x
(Vigdor et al., 2014)	Probit regression; Regression models	x		
(Hodis et al., 2015)	Hierarchical linear models	x		
(Lee and Mallik, 2015)	Ordinary least squares	x		
(Miguéis et al., 2018)	Random Forests, decision trees, support vector machines and naïve Bayes	x	x	x
(Yağci and Çevik, 2019)	Artificial neural networks	x	x	x

popular characteristics related to parents is their participation in students' academic life (Wilder, 2014). Another relevant factor is the parents' socioeconomic status (Caro et al., 2009; Scherer and Siddiq, 2019), which depends upon parents' income, occupation, and education level (Sirin, 2005; Steinmayr et al., 2010). Few studies have addressed the increasing importance that this factor may have in times of economic distress, such as those some countries are experiencing presently.

Regarding school aspects, the first focus is mainly on the class and school size – measured in the number of students – although contradictory results are frequently reported (Leithwood and Jantzi, 2009). Class size is one of the topics that has elicited more debates as a result of dubious and different conclusions. Nonetheless, it is a topic of special interest to policymakers due to the financial impact that different class sizes have. On one side, smaller class sizes have been pointed to increase AA (see, e.g., Bosworth, 2014), whereas other studies point out that class size reduction is not directly linked to better AA (see, e.g., Wößmann and West, 2006). Other school aspects, e.g., regarding infrastructure and IT, have also been identified as potential drivers of AA. Note that its importance is heavily dependent on the context of the study, e.g., the case of the digital divide. Teachers are also usually recognized as an important influencing factor of AA (Kutaka et al., 2017; Noell et al., 2019).

One of the most interesting facts, when one assesses prior studies focusing on the drivers of AA, is that it appears data science and more sophisticated data analysis techniques have not yet been fully (if barely) used. Despite the growing interest in this phenomenon and its unquestionable importance, to the best of the authors' knowledge, researchers have been essentially using a wide variety of traditional methods to shed light on AA drivers. Research on AA has traditionally been survey-driven (e.g., surveying a student cohort and following them for a specified period to determine their success) (Caison, 2007). As reported in Delen (2010), these survey-based research studies have been criticized for their lack of generalized applicability to other institutions (and contexts) and the difficulty and costliness of administering such large-scale survey instruments. An alternative approach to traditional survey-based research is a data-driven strategy that leverages the vast amount of data commonly available in institutional databases, using machine learning techniques to extract insights from the data, something that the present study intends to start. Even though existing literature demonstrates the superiority of data-driven approaches based on AI techniques concerning survey-based methods (Caison, 2007), the use of AI in the field of education is still in its infancy. There are, however, some exceptions. Huang and Fang (2013), in what they argue was the first study to compare four mathematical models to predict AA – multiple regression, multilayer perception network, radial basis function model, and support vector machines. However, this study was limited to 2.907 students for whom the only variables available pertain to past academic results. Moreover, the dependent variable is the AA in a specific area/course – engineering dynamics. Miguéis, Freitas, Garcia, and Silva (2018), used a two-stage approach to segment and predict the AA of 2.459 of a European Engineering School. They found that random forests (RF) surpassed other methods in terms of performance. However, the authors used the previous year's grades to predict the ones in the following year, thus excluding other potentially important factors. It should be noted that if the (true) determinants of the first year's AA – arguably the same as the second – were still not assessed. Another example of a study employing AI methods to predict AA is given by Yağci and Çevik (2019). These authors used artificial neural networks (ANN) to predict AA in science courses, using a sample of 1.972 students from Malaysia and Turkey. The authors included variables pertaining to the students, their parents, and their schools. Note that these studies are limited in terms of sample size, together with the independent and dependent variables used.

Table 1 presents a list of studies focused on AA, describing the techniques used as well as the classification, in terms of the three dimensions presented earlier, of the independent variables tested.

3. Methodology

3.1. Artificial intelligence techniques

Compared with the standard methodology followed by statisticians, AI techniques rely on a different approach for finding a solution to a given problem. The standard procedure employed by statisticians is to build a model that, based on the available input data, can successfully predict the output values. On the other hand, machine learning exploits a different strategy. More in detail, given a supervised optimization problem (i.e., a problem in which data consists of a set of training examples, where each example is a pair consisting of the values of the input variables and the expected output value), an AI technique automatically builds a model that matches input data into the expected target values. In other words, the task of building the model is demanded of the AI technique, and the domain expert is only responsible for collecting the input-output pairs used to train the model.

In the context of the considered academic achievement application, given a training set containing where each input consists of a vector of variables representing a student, and the output indicates whether the student was promoted to the following year. We considered AI techniques to automatically build a model that, given the variables' values associated with students that were not considered in the training phase, can produce as output the expected outcome (promoted or not promoted) that the student will obtain at the end of the school year. A description of the variables used is reported in Section 3.3. The following subsections describe the AI methods considered in the experimental phase. Different AI techniques were considered, to cover tree-based classifiers, instance-based classifiers, and thus obtaining a clear understanding of the performance of different learning algorithms. The reader is referred to Bishop (2006) for a more in-depth explanation of the techniques considered.

3.1.1. Artificial neural networks

Artificial neural networks (ANN) are one of the best-known and widely used AI techniques. They are biologically inspired, and they mimic the structure of the human brain (Haykin, 1994; "History of neural networks," 2015).

A simple ANN consists of many simple and connected units called neurons, each producing a sequence of real-valued activations. Input neurons are activated through the input data provided to the ANN, while other neurons are activated through weighted connections from previously active neurons (Schmidhuber, 2015). Some neurons are responsible for providing the output value(s) that is the prediction of the ANN for a given input. The training process of an ANN consists of finding weights that make the ANN produce the desired output for the input data. In this study, we considered acyclic ANNs.

The main problem when using an ANN is that the final model is a sort of black box consisting only of the set of weights on the connections among the neurons. Thus, the readability of the model is very limited.

3.1.2. Decision trees

Decision Trees (DTs) are a supervised machine learning technique that can address both regression and classification problems (Breiman et al., 1984). A decision tree builds a classification model that has the form of a tree structure, where the internal nodes of the tree contain the independent variables, and the leaves correspond to the possible target classes. Each internal node has several branches, corresponding to the possible values that the variable can assume. The creation of a DT is based on an iterative process in which, at each iteration, a variable is selected to enter the tree based on the homogeneity of data (calculated with the Gini index or with the entropy). Thus, the variable at the root of the tree corresponds to the best predictor. One of the main advantages of DTs relies on the fact that it produces simple if-then-else rules that can be

interpreted. Additionally, the position of the independent variables in the tree indicates their importance for addressing the classification task.

3.1.3. Extremely Randomized Trees

Extremely Randomized Trees (ERTs), also known as Extra Trees (ETs) (Geurts et al., 2006), differ from classic decision trees in the way they are built. In particular, to determine the best split for separating the training samples of a node, random splits are drawn for each of the k randomly selected features, and the best split among those is chosen. When k is set 1, the ERTs correspond to a randomly created decision tree.

3.1.4. Random forest

Random forest (RFs) belongs to the family of ensemble methods (Zhang and Ma, 2012). The idea exploited by RFs is to build different decision trees, each one considering a randomly selected subset of the independent variables of the problem. This point is a fundamental aspect because it leads to decision trees with different structures that can model different aspects of a given problem. Increasing the number of decision trees is usually beneficial for reaching a better prediction of the target variable, without negatively affecting the ability of the final ensemble model to deal with unseen data (Kleinberg, 1996). Thus, one of the main advantages of RF with respect to other techniques is its ability to counteract, or at least limit, overfitting.

3.1.5. Support vector machines

Support vector machines (SVMs) (Cortes and Vapnik, 1995) are a popular ML method for addressing classification and regression problems. Focusing on a classification problem where each training observation belongs to one of the possible two classes, the main idea of SVMs is to determine the best hyperplane that separates instances of one class from the instances of the second class. The best separating hyperplane is the one that maximizes the margin between the observations that are closer to the decision boundary (called support vectors). Considering that the best separating hyperplane maximizes the margin between the two classes, SVMs are usually able to produce classifiers characterized by a good generalization ability (that is, they can produce satisfactory performance over unseen observations) (Hastie et al., 2017). The final model produced by an SVM is difficult to understand and interpret by a human being, thus limiting its use domains.

3.1.6. K-Nearest Neighbors

K-Nearest Neighbors (KNNs) (Cover and Hart, 1967) is a machine learning technique that can be used for addressing both classification and regression problems. In the case of a regression problem, the algorithm assumes that closer (for instance, with respect to the Euclidean distance) data points in the search space should have a high probability of belonging to the same target class. The KNN algorithm calculates the distance between each pair of points in the training set and, subsequently, it classifies a new data point p by using a majority vote: from the training set, it considers the K points that are closer to p and assigns p to the class to which the majority of the K neighbors belong. The algorithm is very simple to implement, and its performance only depends on the choice of the K parameter.

3.2. Logistic regression

Logistic regression (LR) is a well-known technique that is commonly applied in the field of statistics. LR is also used in machine learning as a baseline for testing the performance of more advanced techniques. Given a binary classification problem, LR aims at modeling the posterior probabilities of the two classes via linear functions, while ensuring that they sum to one and remain in [0, 1] (Hastie et al., 2017).

Due to the relatively large number of possible independent variables (16), we used the SAS® Stepwise method in the LR. This method consists of a combination between forward and backward selection techniques in which effects (independent variables) initially specified in the model do not necessarily stay there. In forward selection, there are no initial effects on the LR as they are added sequentially and only stay if they are statistically significant. In backward selection, however, every possible effect is initially present, and each is removed if no statistical significance is found. Hence, in the stepwise method, "the same entry and removal approach for the forward selection and backward elimination methods is used to assess contributions of effects as they are added to or removed from a model" (SAS). The significance level of 0.05 was used in this method.

3.3. Experimental phase

This section presents the experimental settings and the data used in the experimental phase. The results are discussed in the following section. All the experimental phase was performed using the scikit-learn machine learning package, version 0.20.3 (Pedregosa et al., 2011).

The dataset used in this study is an anonymized dataset provided by the Directorate-General of Statistics for Education and Science (DGEEC) of the Portuguese Ministry of Education. Each observation contains data associated with a student, and the target is a binary variable indicating whether the student was promoted to the following year. The variables associated with each observation are reported in Table 2, and one-hot encoding was applied to categorical variables. One-hot encoding is a representation of categorical variables as binary vectors. The process for obtaining the one-hot encoding of a categorical variable first requires that the categorical values are mapped into integer values. Subsequently, each integer value is represented as a binary vector that contains all zero values, except the index of the integer which contains a one. This transformation is necessary, when there is no ordinal relationship between the categories, to remove any bias associated with the integer representation of the categories.

Data pertains to the academic year of 2014/2015. It comprises virtually every student from public high schools (secondary level), corresponding to 110,627 students in the 10th, 11th and 12th grades. The dataset was divided into training and test observations, with the training set containing 60% of the observations and the test set containing the remaining 40%. Subsequently, considering that for the posed classification problem, the dataset is unbalanced, a preprocessing phase aimed at creating a balanced training set was performed. A dataset is unbalanced when at least one class is represented by only a small number of training examples (called the minority class), whereas the other class makes up the majority. In this study, the dataset contains instances belonging to two classes, where one contains most of the instances (majority class). Ignoring this feature of the training set would result in a biased classifier, with all the instances classified in the class corresponding to the majority class. In order to avoid this potential issue, the training set was created by randomly selecting instances from the original dataset, maintaining the original distribution of the data. After this step, the SMOTE oversampling technique (Chawla et al., 2002) with Tomek links (Batista et al., 2004; Tomek, 1976) was used to obtain a new training set with the same number of instances for each one of the two classes. The main idea of SMOTE is to form new minority class examples through the interpolation of several minority class examples that lie together. While the use of SMOTE allows obtaining balanced class distributions, it does not overcome all the issues characterizing data sets with skewed class distributions. In particular, one common problem relies on the fact that some majority class examples might be invading the minority class space. Additionally, the use of SMOTE can expand the minority class cluster introducing artificial minority class examples in the majority class space. This situation is one of the main causes leading to overfitting because a machine learning model must create specific rules for handling the points that are invading the space of the class to which they do not belong.

Tomek links were removed from the data set to avoid this problem. Given two points a and b belonging to different classes, and denoting with $d(a, b)$ the distance between a and b, a Tomek link is defined as follows. The pair (a,b) is a Tomek link if there is not a point c such that

Table 2. Independent variables of the considered dataset. Variables represent demographic information, financial information of students' families, and information about the school and the area in which the school is located.

Variable	Description
x0	Year of the study cycle
x1	Portuguese citizenship (1 = Yes)
x2	Portuguese naturality (1 = Yes)
x3	Gender (1 = Female)
x4	Student's age (years)
x5	Number of enrolled years in high school
x6	Number of failures in the educational career
x7	Scholarship
x8	Level of financial support received by government
x9	Availability of a Personal Computer (PC) at home (1 = Yes)
x10	Internet access (1 = Yes)
x11	Class size (# students)
x12	School size (# students)
x13	Economic level of residence area
x14	Population density of residence area
x15	Rural residence area (1 = Rural)
x16	Number of unit courses attended in the present academic year

$d(a, c) < d(a, b)$ or $d(b, c) < d(b, a)$. If two points form a Tomek link, they are considered noise or borderline points, and they are removed (Batista et al., 2004).

After the application of SMOTE with Tomek links, the training set contained 107,338 observations.

For all the considered techniques (ANNs, RF, DT, ETs, SVMs, and KNN), a parameter tuning phase was executed to determine the ideal values of the parameters. The tuning of the parameters was performed through the random search functionality provided by scikit-learn. All the techniques considered underwent an extensive tuning phase, where 100 randomly generated configurations of the parameters were considered. For each configuration, 30 independent runs were executed, and in each run, 3-fold cross-validation was performed. This procedure is a fundamental step for ensuring the statistical significance of the results. At the end of this process, the best configuration (the one producing the best average AUROC - area under the receiver operating characteristics - value on the validation folds) was selected.

4. Results

4.1. Main results

In this section, results achieved with ML techniques are compared against the ones achieved by an LR, a technique commonly used in "traditional" data analysis. It is important to highlight the fact that LR was applied in a "naive" fashion, without the application of SMOTE and Tomek links. That is, the objective is to compare a machine learning methodology against a traditional data analysis procedure commonly used in the literature. The first part of the analysis is focused on the comparison among ML techniques.

The results achieved by ML techniques are summarized in the boxplots reported in Figure 1. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, and the whiskers extend to the most extreme data points that are not considered outliers. Accuracy, recall, the area under the ROC, and the lift score are the metrics reported. Focusing on the accuracy, one can see that KNNs, RF, and DTs are the best performers, producing comparable results among training and test instances. Thus, these three techniques can produce robust models that are able to handle unseen data. ANNs and ETs also achieved an accuracy greater than 0.75 on both training and test observations. SVMs are the poorest performer among the considered techniques, also presenting a significant difference between training and test accuracy values.

Considering the recall values, SVM is the best performer, producing similar results on both training and test sets. The performance achieved by RF is comparable to the one obtained by ANNs and ETs, with the recall on the test set ranging from 0.69 (RF) to 0.73 (ANNs). The three techniques present some overfitting (i.e., recall on unseen data lower than the one on training data). Finally, DTs and KNNs achieved a recall greater than 0.8 on the training set, but they are characterized by severe overfitting, which leads to a recall on the test set that is lower than 0.65.

The most important metric when considering a classification problem is the area under the ROC curve (AUROC), which summarizes in one single value the precision and the recall of a classifier when varying the threshold. The higher the AUROC, the better the model is at predicting the correct class of each observation. Thus, the AUROC curve is the most robust metric for evaluating the global performance of a classifier (Castelli et al., 2019). For this reason, the ML techniques and the LR model are ranked based on the AUROC values. According to the boxplots of Figure 1, RF is the best performer (on the test set) in terms of AUROC, followed by ANNs and ETs. DTs and KNNs presented an AUROC value greater than 0.8 on the training set but, due to the presence of overfitting, their performance on the test set is lower with respect to the three aforementioned competitors. SVMs performed poorly on the task under examination, presenting AUROC values smaller than 0.65 on both training and test instances. This result strengthens the fact that ensemble techniques (like RF), where a single final model is built by combining different weak learners (decision trees in the case of RF), are able to produce robust results and can outperform the results achieved by a single-learner-based model (Aggarwal, 2014).

To summarize, results obtained by the considered ML techniques show that RFs perform the best in terms of AUROC over the test set, with ANNs and ETs able to provide comparable results. DTs and KNNs are the best performers on the training instances, but they suffer a severe amount of overfitting. A set of statistical tests was performed to validate the AUROC results reported in Figure 1, statistically. Preliminary analysis using the Kolmogorov-Smirnov test showed that the data were not normally distributed, and hence a rank-based statistic was used. The Wilcoxon rank-sum test for pairwise data comparison was used (with $\alpha = 0.1$ and a Bonferroni correction) with the alternative hypothesis that the samples do not have equal medians of AUROC. The results of the statistical test are summarized in Table 3.

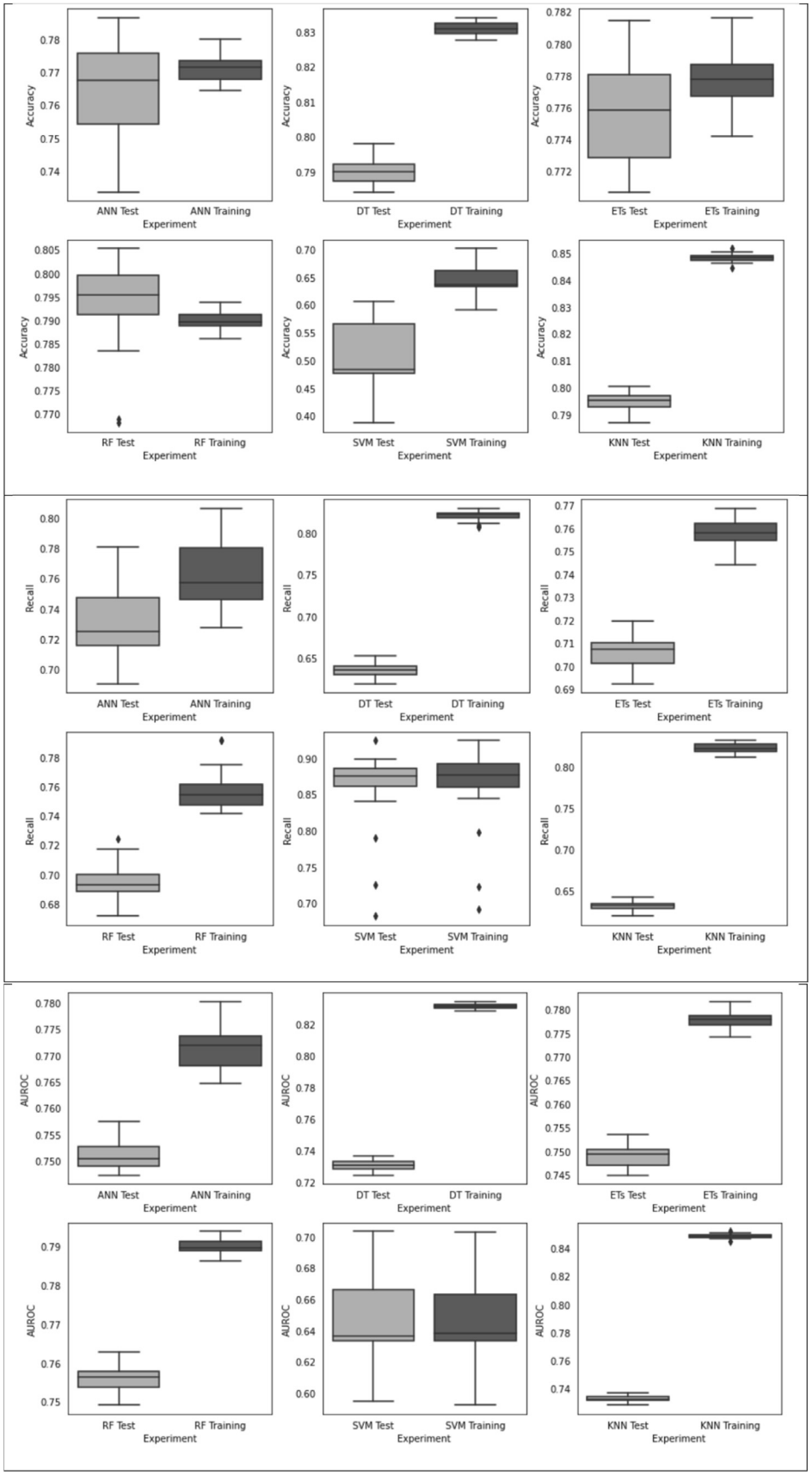

Figure 1. Boxplots of accuracy, recall, and AUROC for training and test instances for the ML techniques considered. Higher values correspond to a better performance of the models.

Extending the analysis to include the LR, one can notice from Table 4 that the AUROC value achieved by the LR-based model is the worst among the considered techniques, thus strengthening the suitability of AI techniques in addressing the problem at hand. For all the techniques considered, this table summarizes the values of accuracy, recall, and AUROC achieved by the best model on unseen observations. The fact that SMOTE was not applied before the use of LR produced a model that always returns the same prediction (corresponding to the over-represented class). Nevertheless, this was a conscious choice, as our purpose, more than just comparing AI vs. non-AI techniques, is to compare the respective approaches. Hence, we decided not to use a traditional technique (LR) with an advanced sampling technique (SMOTE).

All in all, the analysis corroborates the initial hypothesis that AI techniques can provide a competitive advantage compared to "traditional" statistical techniques when dealing with a complex problem characterized by a vast amount of data. More in detail, among the different competitors, RF is the technique that produced the best results: this result is coherent with AI literature that shows the competitiveness of ensemble techniques in addressing classification problems (Aggarwal, 2014; Sagi and Rokach, 2018).

The last step of this analysis is to simulate the impact of employing each of the four considered methods at the beginning of each academic year, assessing to what extent each could successfully identify students that essentially fail at the end. For the education field, providing a tool that could preventively rather than reactively "mark" students with a higher probability of failing the year by warning the teachers at the beginning of each academic year, would bring astounding benefits. For this purpose, we used the scores (estimated probability by the model of failing the year) of each method in the test set to sort them by decreasing order of failing. After that, we split each test set into 20 equally sized groups (i.e., ventiles or vigintiles). For each ventile, we computed the real/effective number of failures and its rate, the cumulative failure rate until that ventile, the lift (i.e., the failure rate of that ventile over the average failure rate), the cumulative lift (cumulative failure rate over the average) and the captured failures (i.e., how many failures were identified over the total number of failures). Note that all these parameters pertain to the actual situation of the students, as the estimated situation, i.e., the one predicted by the models, were only used to sort the students in ventiles using these scores. The RF and DT presented the best performance in predicting, with one academic year in advance, those students that effectively failed (as reported in Table 5). As it is natural, the first ventile contains the students with a higher failure probability. Hence, the first ventiles are the ones in which the results are more reliable in predicting failure. As expected by the performance assessment (please see Table 3), the RF and DT clearly outperform both the traditional method (LR) and the SVM. In fact, it is interesting to note that the difference in terms of performance between these two latter techniques is much less pronounced than the difference among the first two methods (LR and SVM versus RF and DT). The detailed results are reported in Tables A1 to A7, inAppendix.

4.2. Feature importance

This subsection reports an analysis that involves some of the AI methods considered. The objective is to understand which features are deemed as important by the different AI methods for addressing the classification task at hand. This analysis involves RFs, DTs, and ETs. For the other considered AI methods, it is not possible to perform a similar study because they are "black-box" models. Thus, they make it impossible to extract useful information that may allow understanding the process used for producing a particular output for a given observation.

The importance of a feature (that is a node in the considered AI methods) is computed as the (normalized) total reduction of the node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value, the more important the feature. For RFs, the feature importance is calculated by averaging over all the trees of the ensemble.

Table 6 reports the feature importance values extracted for DTs, RFs, and ETs. According to these values, it is interesting to point out that all the considered methods deem the variable "Number of unit courses attended in the present academic year" as the most relevant for the problem under exam. This aspect is coherent as regards the analysis discussed in the previous section and summarized in Table 6. Additionally, gender and age are relevant for all the methods, even though their importance is significantly lower than one of the previously mentioned variables ("Number of unit courses attended in the present academic year").

5. Discussion and implications

5.1. Discussion

We used AI to predict the academic achievement of virtually every public high school student in Portugal in a specific academic year. Every model was estimated using data (independent variables) pertaining to the beginning of each academic year, whereas the dependent variable is in respect of the end of the year. Hence, the estimated models can act as an effective prevention tool for failing academic years. Our results clearly demonstrate that an AI approach manifestly outperforms a more traditional one -the ubiquitous approach in the literature until this point. We posit that implementing an AI stratagem could substantially improve the prediction performance of AA. Regarding the ANN results, despite meaningful, they are, to some extent, less pronounced. The failure rate in the first ventile is 80%, 4.26 higher than the average (lift), which corresponds to 21% of the overall failures. In DT, the failure rate of the first ventile is 86%, a lift of 4.59. The model with the best performance in this regard is RF. The failure rate in the first ventile is 87%, 4.65 higher than the average. The students failing in this 5% group account to 23% of the total number of students failing that academic year (cumulative captured response). Had these students' teachers and parents known *a priori* (at the beginning of the year)

Table 3. P-values returned by the Wilcoxon test.

	ANN	DT	ET	RF	SVM	KNN
ANN	-	$<10^{-8}$	$3.49*10^{-3}$	$3.31*10^{-6}$	$<10^{-8}$	$<10^{-8}$
DT		-	$<10^{-8}$	$<10^{-8}$	$<10^{-8}$	$2.28*10^{-2}$
ET			-	$<10^{-8}$	$<10^{-8}$	$<10^{-8}$
RF				-	$<10^{-8}$	$<10^{-8}$
SVM					-	$<10^{-8}$
KNN						-

Table 4. Accuracy, recall, and AUROC values on the test instances for the best model produced by the considered techniques.

Test Set	ANN	DT	ET	RF	SVM	KNN	LR
Accuracy	76.5%	79.0%	77.6%	79.4%	51.2%	79.5%	81.1%
Recall	73.0%	63.4%	70.6%	69.4%	86.3%	63.2%	48.7%
AUROC	0.75	0.73	0.75	0.76	0.65	0.55	0.55

Note that for this problem, accuracy is biased.

Table 5. Lift and captured response of the models.

Test Set	ANN	DT	ET	RF	SVM	KNN	LR
Cumulative Lift at 5%	4.26	4.59	3.54	4.65	2.45	3.42	2.59
Cumulative Lift at 15%	3.13	3.10	2.82	3.28	1.41	2.96	2.66
Cumulative Captured Response 5%	21%	23%	18%	23%	12%	17%	13%
Cumulative Captured Response 15%	47%	46%	42%	49%	33%	44%	40%
Threshold 15%	0.758	0.703	0.647	0.722	0.659	0.727	0.349

Table 6. Feature importance values for Random Forest, Decision Trees, and Extra Trees.

Random Forest (RFs)		Decision Trees (DTs)		Extra Trees (ETs)	
Feature	Importance	Feature	Importance	Feature	Importance
Number of unit courses attended in the present academic year	0.5300	Number of unit courses attended in the present academic year	0.5539	Number of unit courses attended in the present academic year	0.6770
Student's age (years)	0.1429	School size (# students)	0.1368	Gender	0.1167
School size (# students)	0.0924	Economic level of residence area	0.0770	Student's age (years)	0.0484
Gender	0.0737	Gender	0.0524	Economic level of residence area	0.0330
Economic level of residence area	0.0363	Student's age (years)	0.0506	School size (# students)	0.0277

Note: Feature importance is only possible to be computed in tree-based models.

that the learners were almost five times more likely to fail at year-end, many of the failures may have been prevented. For the sake of readability, please see the Appendix for the detailed results of every model. By observing them, it is noticeable that RF and DT irrefutably outperform both traditional LR and SVM. A striking fact is that the difference between the last two methods is much less noticeable, although LR still outperforms SVM. Hence, we demonstrate that AI methods do not always lead to better results. Another interesting point is that, in many cases, ANN are tough to perform better than DT, although at the expense of interpretability. In this case, that is not true, as DT tend to perform better, at least for the students identified as being very likely to fail.

As marking "just" 5% of the students would limit the total failures captured, we also highlight the performance parameters for the first three ventiles, i.e., up to the 15% of students with a higher estimated likelihood of failing. In the first three ventiles, the RF, ANN, and DT identified students with an effective failure rate of 62%, 59%, and 58%, respectively, corresponding to 3.28, 3.13, and 3.10 times the average. Moreover, the percentages of captured failures are 49%, 47%, and 46%. These results are astounding, as "just" by "marking" 15% of students at the beginning of each academic year, almost 50% of failures are identified and could thus, to some extent, be prevented. Even for the false positives, i.e., students that would pass in any event, it is reasonable to assume that their grades could improve substantially. Note that the opposite approach can be drawn, i.e., one could focus on the less likely identified students to fail. In this case, using the RF, more than one-third of students (35%) – the seven ventiles with less probability of failing - have a failure rate of 2.6% (!), corresponding to a failure rate that is seven times smaller in respect of the global average. Moreover, this group of students, i.e., the 35% identified with less probability of failing, effectively fail less than 23 times than those identified by the RF as being in the top-15% in terms of failure.

Apropos the most important AA drivers, important conclusions are drawn from our results. The two most important variables are related to the academic record of the student. Those enrolled in fewer courses, meaning that they are repeating some, and the number of failed academic years since the beginning of their academic paths, are the most important antecedents of failing the academic year. This ramification is a problematic finding as it indicates that once a student fails, it is very likely that he or she will fail again in the future. One might argue that failing a course works almost as having a criminal record, as far as AA is concerned. In education, especially at the secondary level, this should ideally not happen. On the other hand, failing an academic year is deeply related to the internal characteristics of students that should be constant across years. Another conclusion, this one more expected, has to do with the fact that, on average, females present better AA than males. It would be interesting to assess whether this happens regardless of the unit course or not. Finally, it appears that class size does not present a significant impact on one's AA. Only for (very) large and uncommon classes, with more than 30 students, is AA marginally affected. This factor is significant because some policymakers argue that in smaller classes, students would perform better. As having smaller courses implies a higher financial burden, finding that it does not affect AA, is an important fact for policymakers (only classes with more than 30 students should be avoided). This finding is a good example of the benefits that using AI methods in large datasets may yield for the public sector, especially in times of economic meltdown (Weerakkody et al., 2017). The area of residence also appears not to have an impact on AA, which is a good sign, as Portugal is an extremely unequal country, where most resources and people are found in the coastal regions, especially Lisbon and Oporto. It appears that centralization does not affect education, at least not in Portugal. Finally, our results present meaningful theoretical and practical implications, both in the field of AI and education.

5.2. Theoretical implications

The two main theoretical implications are: first, this paper is among the first initiatives to use AI techniques for a large-scale AA study. As the results demonstrate that AI techniques have a better performance in general in terms of prediction than traditional ones, we suggest the use of AI methods in this context; Secondly, we were able to shed some light on AA antecedents, using state-of-the-art methods that, to the best of our knowledge, have not yet been employed in this context.

5.3. Practical implications

First, by implementing the approach used in this paper, EU member-states to achieve the Europe 2020 goals for AA. If it were possible to safeguard students' privacy and data, employing artificial intelligence methods like the ones we used in this project at the beginning of each academic year, could provide teachers with valuable information to engender their students' academic achievement and, therefore, reduce school dropout. The idea is to replicate our approach to virtually every student, classifying each in terms of the likelihood of passing, or failing, the year. We acknowledge that this is not a consensual nor straightforward approach, as it needs to be implemented with other measures to prevent a self-fulfilling prophecy, i.e., a student that could fail because the teacher thought in advance that that is likely to happen when, actually, it was not.

Second, our results allow engendering AA by providing decision-makers, schools, and teachers a better understanding of its drivers, as well as individual (student-level) prediction of AA. Our results provide valuable information towards the most critical drivers of academic achievement. Thus, policies could be targeted at the most influential antecedents of academic achievement.

Thirdly class size does not present a significant impact on one's AA. This element is also extremely important because an ongoing common debate in some European countries is whether class sizes should be smaller, as some argue that this would yield higher academic achievement and, consequently, lower school dropouts (particularly meaningful in high school). Although some policymakers argue that in smaller classes, students would perform better, the truth is that this also adds a higher financial burden, which is an even more relevant constraint in times of financial difficulties. Our results seem to indicate that class size does not affect AA, at least not in a meaningful way, thus shedding some light on this controversial argument.

5.4. AI and model interpretability

The scientific literature demonstrated a rising interest concerning the interpretability (or explain ability) of AI models (Dosilovic et al., 2018; Preece, 2018). The growing interest in this topic is accompanied by the popularity of AI-based models. Despite their ability to produce human-competitive results on an increasing number of complex tasks, these models are essentially black boxes: no information is provided about how they achieved their predictions. In other words, as a final AI user, we know the prediction that an AI model produced, but we do not know why this particular prediction was made. This dimension could be an important limitation to the wide-scale adoption of AI, as users showed more willingness to use a particular model if they can understand why particular decisions are being made (Ribeiro et al., 2016).

In the context of the considered application, the possibility of obtaining explainable models could allow the Portuguese Ministry of Education to make better-informed decisions for fostering academic achievement. Additionally, at the European level, the importance of reaching an explainable AI is also motivated by the recent introduction of the general data protection regulation (GDPR), which forbids the use of solely automated decisions. Thus, while some of the techniques used in this study allow for a partial interpretation of the models (i.e., it is possible to understand which variables have the more considerable influence on the output of the model), future efforts should be dedicated to the implementation of a fully interpretable AI model.

Explainable AI will give to human users the ability to not only understand an AI model but also to identify and correct the errors of the model. This assessment could lead to a long-life learning process, where the AI model continuously improves based on the feedback of human experts.

5.5. Conclusions and future work

To the best of our knowledge, this was one of the first studies (if not the first) that used artificial intelligence AI and traditional techniques to predict academic achievement (AA) at a national level, i.e., including virtually every (high school) student. Based on a sample of 110,627 students from all public high schools in Portugal in a specific academic year, we can conclude that: (1) in general, the AI methods reveal a better performance compared to traditional ones. For example, to the first ventile (5% of the students with a high likelihood estimated by the model to fail), 87%, 80%, 46%, and 49% are well-classified, respectively, for RF, ANN, SVM, and LR. The AUROC also reveals that RF is the best one. (2) *the most critical drivers are the number of unit courses attended in the present academic year*, the *number of failures in the education career*, and student *gender*. Females present better AA than males. It would be interesting to assess whether this happens regardless of the unit course or not. As further work, we suggest using unit course data instead of annual input to figure out if there are differences among unit course (disciplines).

Declarations

Author contribution statement

F. Cruz-Jesus: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

M. Castelli: Conceived and designed the experiments; Performed the experiments; Wrote the paper.

T. Oliveira: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data. R. Mendes, A. Rosa-Louro: Performed the experiments.

C. Nunes: Analyzed and interpreted the data.

M. Sa-Velho: Performed the experiments; Analyzed and interpreted the data.

Funding statement

This work was partially supported by national funds through FCT (Fundação para a Ciência e a Tecnologia) under project DSAIPA/DS/0032/2018 (DS4AA).

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e0408110.

Acknowledgements

This work was partially supported by national funds through FCT (Fundação para a Ciência e a Tecnologia) under project DSAIPA/DS/0032/2018 (DS4AA).

References

Aggarwal, C.C., 2014. Data classification: algorithms and applications. In: Data Classification: Algorithms and Applications.

Ahmed, Z., Asim, M., Pellitteri, J., 2019. Emotional intelligence predicts academic achievement in Pakistani management students. Int. J. Manag. Educ.

Archibald, S., 2006. Narrowing in on educational resources that do affect student achievement. Peabody J. Educ. 81 (4), 23–42.

Avery, P.G., Walker, C., 1993. Prospective teachers' perceptions of ethnic and gender differences in academic achievement. J. Teach. Educ. 44 (1), 27–37.

Barnett, R., Glass, J.C., Snowdon, R., Stringer, K., 2002. Size, performance and effectiveness: cost-constrained measures of best-practice performance and secondary-school size. Educ. Econ. 10 (3), 291–311.

Batista, G.E.A.P.A., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. In: ACM SIGKDD Explorations Newsletter.

Bishop, C.M., 2006. Machine learning and pattern recogniton. In: Information Science and Statistics.

Bosworth, R., 2014. Class size, class composition, and the distribution of student achievement. Educ. Econ. 22 (2), 141–165.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Retrieved from. https://books.google.pt/books?id=JwQx-WOmSyQC.

Brunner, M., Gogol, K., Sonnleitner, P., Keller, U., Krauss, S., Preckel, F., 2013. Gender differences in the mean level, variability, and profile shape of student achievement: results from 41 countries. Intelligence 41 (5), 378–395.

Caison, A.L., 2007. Analysis of institutionally specific retention research: a comparison between survey and institutional database methods. Res. High. Educ. 48 (4), 435–451.

Caro, D.H., McDonald, J.T., Douglas Willms, J., 2009. Socio-economic status and academic achievement trajectories from childhood to adolescence. Can. J. Educ. 32 (3), 558–590.

Castelli, M., Vannesch, L., Rubio Largo, Á., 2019. Supervised learning: classification. In: Guenther, R., Steel, D. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Elsevier, Oxford, pp. 342–349.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res.

Choi, Y., Lee, H., Irani, Z., 2018. Big data-driven fuzzy cognitive map for prioritising IT service procurement in the public sector. Ann. Oper. Res.

Codjoe, H., 2007. The importance of home environment and parental encouragement in the academic achievement of African-Canadian youth. Can. J. Educ. 30 (1), 137–156.

Coleman, J., Hopkins, J., 1966. Equality of Educational Opportunity. U.S. Department of Health, Education and Welfare, pp. 666–675.

Côrte-Real, N., Ruivo, P., Oliveira, T., Popovič, A., 2019. Unlocking the drivers of big data analytics value in firms. J. Bus. Res.

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Machine Learning.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. 13 (1), 21–27.

Croninger, R., Rice, J., Rathbun, A., Nishio, M., 2007. Teacher qualifications and early learning: effects of certification, degree, and experience on first-grade student achievement. Econ. Educ. Rev. 26 (3), 312–324.

Delen, D., 2010. A comparative analysis of machine learning techniques for student retention management. Decis. Support Syst. 49 (4), 498–506.

Delen, D., Zolbanin, H.M., 2018. The analytics paradigm in business research. J. Bus. Res.

Dosilovic, F.K., Brcic, M., Hlupic, N., 2018. Explainable Artificial Intelligence: A Survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings.

Driessen, G., Smit, F., Sleegers, P., 2005. Parental involvement and educational achievement. Br. Educ. Res. J. 31 (4), 509–532.

Dronkers, J., Van Der Velden, R., Dunne, A., 2012. Why are migrant students better off in certain types of educational systems or schools than in others? Eur. Educ. Res. J. 11 (1), 11–44.

Dunn, K., Airola, D., Lo, W.-J., Garrison, M., 2013. Becoming data-driven: exploring teacher efficacy and concerns related to data driven decision making. J. Experim. Educa. 81, 222–241.

Erevelles, S., Fukawa, N., Swayne, L., 2016. Big Data consumer analytics and the transformation of marketing. J. Bus. Res.

European Commission, 2017. Early School Leaving. Retrieved. https://ec.europa.eu/education/policies/school/early-school-leaving_en. (Accessed 20 October 2019).

Fan, X., 2001. Parental involvement and students' academic achievement: a growth modeling analysis. J. Experim. Educ. 70 (1), 27–61.

Fan, X., Chen, M., 2001. Parental involvement and students' academic achievement: a meta-analysis. Educ. Psychol. Rev. 13 (1), 1–22.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42.

Ghasemaghaei, M., Calic, G., 2019. Does big data enhance firm innovation competency? The mediating role of data-driven insights. J. Bus. Res.

Greenwald, R., Hedges, L.V., Laine, R.D., 1996. The effect of school resources on student achievement. Rev. Educ. Res. 66 (3), 361–396.

Hanushek, E., Kimko, D., 2000. Schooling, labor force quality, and the growth of nations. Am. Econ. Rev. 90 (5), 1184–1208.

Hanushek, E., Woessmann, L., 2012. Schooling, educational achievement, and the Latin American growth puzzle. J. Dev. Econ. 99 (2), 497–512.

Hartas, D., 2011. Families' social backgrounds matter : socio-economic factors, home learning and young children's language, literacy and social outcomes. Br. Educ. Res. J. 37 (6), 893–914. https://doi.org/10.1080/01411926.2010.506945.

Hastie, T., Tibshirani, R., Friedman, J., 2017. In: second ed. The Elements of Statistical Learning. Math. Intell.

Hattie, J., 2009. Visible learning: a synthesis of over 800 meta-analyses relating to achievement. In: Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement.

Haykin, S., 1994. A Comprehensive Foundation, Neural Networks.

History of neural networks, 2015. SpringerBriefs in Applied Sciences and Technology.

Hodis, F., Johnston, M., Meyer, L., McClure, J., Hodis, G., Starkey, L., 2015. Maximal levels of aspiration, minimal boundary goals, and their relationships with academic achievement: the case of secondary-school students. Br. Educ. Res. J. 41 (6), 1125–1141.

Hoxby, C., 2000. The effects of class Size on student achievement: new evidence from population variation. Q. J. Econ. 115 (4), 1239–1285. Retrieved from. http://www.jstor.org.proxy.library.ucsb.edu:2048/stable/info/2586924.

Huang, C., 2018. Social network site use and academic achievement: a meta-analysis. Comput. Educ. 119, 76–83.

Huang, S., Fang, N., 2013. Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models. Comput. Educ. 61, 133–145.

Jackson, L., von Eye, A., Biocca, F., Barbatsis, G., Zhao, Y., Fitzgerald, H., 2006. Does home internet use influence the academic performance of low-income children? Dev. Psychol. 42 (3), 429–435.

Jeynes, W.H., 2007. The Relationship between Parental Involvement and Urban Secondary, 42. School Student Academic Achievement, pp. 82–110.

Kleinberg, E.M., 1996. An overtraining-resistant stochastic modeling method for pattern recognition. Ann. Stat.

Krassel, K., Heinesen, E., 2014. Class-size effects in secondary school. Educ. Econ. 1–15.

Kubey, R.W., Lavin, M.J., Barrows, J.R., 2001. Internet use and collegiate academic performance decrements: early findings. J. Commun. 51 (2), 366–382.

Kutaka, T.S., Smith, W.M., Albano, A.D., Edwards, C.P., Ren, L., Beattie, H.L., Stroup, W.W., 2017. Connecting teacher professional development and student mathematics achievement: a 4-year study of an elementary mathematics specialist program. J. Teach. Educ. 68 (2), 140–154.

Lee, C.L., Mallik, G., 2015. The impact of student characteristics on academic achievement: findings from an online undergraduate property program. Pacific Rim Property Research Journal 21 (1), 3–14.

Lee, H., 2007. The effects of school racial and ethnic composition on academic achievement during adolescence. J. Negro Educ. 76 (2), 154–172.

Lee, J.-S., Bowen, N., 2006. Parent involvement, cultural capital, and the achievement gap among elementary school children. Am. Educ. Res. J. 43 (2), 193–218.

Lei, J., Zhao, Y., 2007. Technology uses and student achievement: a longitudinal study. Comput. Educ. 49 (2), 284–296.

Leithwood, K., Jantzi, D., 2009. A review of empirical evidence about school size effects: a policy perspective. Rev. Educ. Res. 79 (1), 464–490.

Marks, G., Cresswell, J., Ainley, J., 2006. Explaining socioeconomic inequalities in student achievement: the role of home and school factors. Educ. Res. Eval. 12 (2), 105–128.

Mensah, F.K., Kiernan, K.E., 2010. Gender differences in educational attainment: influences of the family environment. Br. Educ. Res. J. 36 (2), 239–260.

Miguéis, V.L., Freitas, A., Garcia, P.J.V., Silva, A., 2018. Early segmentation of students according to their academic performance: a predictive modelling approach. Decis. Support Syst. 115, 36–51.

Noell, G.H., Burns, J.M., Gansle, K.A., 2019. Linking student achievement to teacher preparation: emergent challenges in implementing value added assessment. J. Teach. Educ. 70 (2), 128–138.

OECD, 2012. Equity and Quality in Education - Supporting Disadvantaged Students and Schools.

Otani, M., 2019. Relationships between parental involvement and adolescents' academic achievement and aspiration. Int. J. Educ. Res.

Patterson, M., Pahlke, E., 2011. Student characteristics associated with girls' success in a single-sex school. Sex. Roles 65 (9–10), 737–750.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res.

Preece, A., 2018. Asking 'Why' in AI: explainability of intelligent systems – perspectives and challenges. Intell. Syst. Account. Finance Manag. 25 (2), 63–72.

Ribeiro, M., Singh, S., Guestrin, C., 2016. Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Rivkin, S., Hanushek, E., Kain, J., 2005. Teachers, schools, and academic achievement. Econometrica 73 (No. 2), 417–458.

Sagi, O., Rokach, L., 2018. Ensemble Learning: A Survey. *Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery*.

Scherer, R., Siddiq, F., 2019. The relation between students' socioeconomic status and ICT literacy: findings from a meta-analysis. Comput. Educ. 1 (138), 13–32.

Schmidhuber, J., 2015. Deep Learning in neural networks: an overview. Neural Network. 1 (6), 85–117.

Sirin, S.R., 2005. Socioeconomic status and academic achievement: a meta-analytic review of research. Rev. Educ. Res. 75 (3), 417–453.

Sivarajah, U., Kamal, M.M., Irani, Z., Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. J. Bus. Res. 70, 263–286.

Steinmayr, R., Dinger, F., Spinath, B., 2010. Parents' education and children's achievement: the role of personality. Eur. J. Pers. 24 (6), 535–550.

Steinmayr, R., Spinath, B., 2008. Sex differences in school achievement: what are the roles of personality and achievement motivation? Eur. J. Pers. 22 (3), 185–209.

Tomek, I., 1976. TWO MODIFICATIONS of CNN. IEEE Transactions on Systems, Man and Cybernetics.

Torrecilla Sánchez, E.M., Olmos Miguélañez, S., Martínez Abad, F., 2019. Explanatory factors as predictors of academic achievement in PISA tests. An analysis of the moderating effect of gender. Int. J. Educ. Res. 1 (96), 111–119.

Valli Jayanthi, S., Balakrishnan, S., Lim Siok Ching, A., Aaqilah Abdul Latiff, N., Nasirudeen, A.M.A., 2014. Factors contributing to academic performance of students in a tertiary institution in Singapore. Am. J. Educ. Res. 2 (9), 752–758.

van der Scheer, E.A., Visscher, A.J., 2018. Effects of a data-based decision-making intervention for teachers on students' mathematical achievement. J. Teach. Educ. 69 (3), 307–320.

Vigdor, J.L., Ladd, H.F., Martinez, E., 2014. Scaling the digital divide: home computer technology and student achievement. Econ. Inq. 52 (3), 1103–1119.

Wakefield, J., Frawley, J.K., 2020. How does students' general academic achievement moderate the implications of social networking on specific levels of learning performance? Comput. Educ. 144, 103694.

Wally-Dima, L., Mbekomize, C., 2013. Causes of gender differences in accounting performance: students' perspective. Int. Educ. Stud. 6 (10), 13–26.

Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., Dwivedi, Y.K., 2017. Open data and its usability: an empirical view from the Citizen's perspective. Inf. Syst. Front 19 (2), 285–300.

Wilder, S., 2014. Effects of parental involvement on academic achievement: a meta-synthesis. Educ. Rev. 66 (3), 377–397.

Wößmann, L., West, M., 2006. Class-size effects in school systems around the world: evidence from between-grade variation in TIMSS. Eur. Econ. Rev. 50 (3), 695–736.

Yağci, A., Çevik, M., 2019. Prediction of academic achievements of vocational and technical high school (VTS) students in science courses through artificial neural networks (comparison of Turkey and Malaysia). Educ. Inf. Technol. 24 (5), 2741–2761.

Zhang, C., Ma, Y., 2012. Ensemble machine learning: methods and applications. In: Ensemble Machine Learning: Methods and Applications.