A Practical Model for the Evaluation of High School Student Performance Based on Machine Learning

Mostafa Zafari [1], Abolghasem Sadeghi-Niaraki [2,3,*], Soo-Mi Choi [3] and Ali Esmaeily [1]

1   Department of Surveying Engineering, Faculty of Civil and Surveying Engineering,
    Graduate University of Advanced Technology, Kerman 7631885356, Iran; m_zafari1@yahoo.com (M.Z.);
    aliesmaeily@kgut.ac.ir (A.E.)
2   Geoinformation Technology Center of Excellence, Faculty of Geodesy& Geomatics Engineering,
    K. N. Toosi University of Technology, Tehran 1969764499, Iran
3   Department of Computer Science and Engineering and Convergence Engineering for Intelligent Drone,
    Sejong University, Seoul 143-747, Korea; smchoi@sejong.ac.kr
*   Correspondence: a.sadeghi.ni@gmail.com

Abstract: The objective of this research is to develop an machine learning (ML) -based system that evaluates the performance of high school students during the semester and identify the most significant factors affecting student performance. It also specifies how the performance of models is affected when models run on data that only include the most important features. Classifiers employed for the system include random forest (RF), support vector machines (SVM), logistic regression (LR) and artificial neural network (ANN) techniques. Moreover, the Boruta algorithm was used to calculate the importance of features. The dataset includes behavioral information, individual information and the scores of students that were collected from teachers and a one-by-one survey through an online questionnaire. As a result, the effective features of the database were identified, and the least important features were eliminated from the dataset. The ANN accuracy, which was the best accuracy in the original dataset, was reduced in the decreased dataset. On the contrary, SVM performance was improved, which had the highest accuracy among other models, with 0.78. Moreover, the LR and RF models could provide the same performance in the decreased dataset. The results showed that ML models are influential for evaluating students, and stakeholders can use the identified effective factors to improve education.

Keywords: student performance; machine learning (ML); educational data mining (EDM); education; artificial intelligence (AI)

1. Introduction

Artificial intelligence (AI) has always been a hot topic for discussion because, in the twenty-first century, the world is run by this technology in almost all spheres of life [1]. Machine learning (ML) is a branch of artificial intelligence that systematically applies algorithms to find the combination of basic relationships among data and produces information. ML's main purpose is to predict future events or future scenarios that are unknown for computers [2,3]. Thanks to data mining and machine learning, the combination of these methods can process data, patterns and models for reasoning, understanding, planning, problem solving, predictions and the manipulation of objects. One of the main advantages of ML is that it can complete complex and time-consuming tasks, and time spent on this work is freed up to use in other matters [4]. For instance, this time can be used by teachers to work on progress, interact with students and prepare for classes [5]. Artificial intelligence and ML have been used in various ways in educational institutions, including automating processes and administrative tasks, curriculum and content development, instruction and the learning processes of students. Instructors are able to assess without bias by using ML. Although ML may not certainly replace human evaluation, it is close to

doing such work [6]. MLs are currently highly advanced and can do more than scoring exams with the answer key. They can provide information about student performance and even perform more conceptual assessments such as scoring the essays [7–9] or students' engagements [10]. MLs can collect information about how students acted and can evaluate this performance [1]. Machine learning and data mining are powerful tools for instructors and institutions to explore the educational database; this has increased with the assistance of ML, and enabled decision-makers to extract information from data for decisions and policies. The application that ML uses in the educational database is called Educational Data Mining (EDM) [11,12]. The educational database is available at different levels of detail for different tasks and often expands in several software systems over some time [13]. EDM tries to use features and patterns in a database to make an effective analysis for predicting student performance [14]. This produced information can be used by data scientists, instructors, administrators, a student's parents, etc. [15]. The evaluation of student performance refers to how much a student approaches the educational goals and specifies how well a student learned, how motivated to learn a student is or how good the teaching method was [16]. The information obtained from the evaluation gives teachers this insight so they can make the right decisions to improve the learning of the student and give appropriate feedback. The individual differences of each student, such as personality, motivation, self-efficacy, intelligence and self-control, have a close relationship to his/her performance, so in this research, all these differences were covered by choosing the proper features.

The educational system in Iran is divided into two main levels, K-12 education and higher education, consisting of 6 years of primary education and 6 years of high school education. Students spend 24 h in class each week. The curriculum contains mathematics, science, foreign languages and so on. High school students aged 12 to 18 years old are divided into four fields (streams): humanism, science, and technical and vocational streams. Choosing their stream is based on his/her grades and the results of his/her examination, not based on their interest. Furthermore, grades are determined on a scale between 0 and 20 in all levels of education and students are assessed during the semester and end of the semester as well. The lowest score to pass a lesson is 10. Some of the most highlighted features of the Iranian educational system include a teaching strategy that is teacher-centered in all schools and the assessment of students only using grades in each lesson [17].

The rest of the paper is as follows. Studies conducted to evaluate student performance with the machine learning approach are referred to in the next section. In Section 3, the selection and collection of data, and the case study are described in detail. After that, the model selection process, and the design and development of models are discussed, and the overall definition of each algorithm is presented. Furthermore, in the results section, the performance of each model on datasets is shown and is discussed. In the final section, the results are discussed, and a summary of the paper is presented.

2. Related Works

Many research works have been focused on the integration of AI and ML in different parts of education, and various methods and tools have been used to carry out such tasks. One of these parts is the assessment of student performance. The performance of students can be evaluated from various aspects. Several studies evaluated students in general in terms of student performance [18–22] and some other studies evaluated students for a specific purpose such as academic achievement [23,24], reading ability [25,26], grading [27–29], dropout prediction [30–33], etc. Below, some state-of-the-art research studies have been discussed for each of the mentioned tasks that evaluated the student performance from different aspects.

2.1. Evaluation of Student Performance

To evaluate the performance of students in geographic lessons and classify students, data which consisted of 15 attributes such as grades in homework assignments, oral assessments, short tests, semester exams, etc., were obtained from 307 students and fed into Naive Bayesian (NB), Random Forest (RF), Sequential minimal optimization (SMO), multilayer perceptron (MLP) and Bayes net (BN) algorithms. The accuracy of the mentioned models was 68.99, 67.04, 65.38, 59.81 and 65.67, respectively. Data have been collected concerning the "geography" module [34]. Similar to the previous reference in ref [15], the authors employed linear regression (accuracy = 64.25), decision tree (accuracy = 59.8), and naive Bayes (accuracy = 71) algorithms in data analysis with 303 input features including student demographics, the salaries of educators, and assessment data. Moreover, in ref [12], logistic regression (accuracy = 0.839), elastic net (accuracy = 0.839), decision tree (accuracy = 0.767), random forest (accuracy = 0.828) and neural network (accuracy = 0.833) algorithms were utilized to evaluate student performance in English lessons. Their database was NA-PLAN which is a set of standardized literacy and numeracy tests sat by all students in Australia and the tests cover five learning areas known as "test domains" (reading, writing, spelling, grammar and punctuation, and numeracy), alongside student background information which is collected by schools from students' parents via enrolment forms. Overall, 2,235,804 rows of student information were used in the analysis. For monitoring student performance at a Brazilian technical high school, authors in ref [35] developed an ML system based on NB, SVM, tree-based method (Simple CART), and rule-based method (OneR) algorithms. Their dataset contained information about course name, age, sex, birthplace, course duration, the identification of each discipline studied, number of faults in the first to fourth bimester, and student status at the end of the course. In this research, the usefulness of ML was highlighted for monitoring student performance. In ref [18], student performance was predicted using multi-class support vector classification. Real data from 395 secondary students in Portugal were employed, which was collected by a questionnaire method and school reports. The data contained grades assigned and student info, and each instance is categorized in five levels as A to F. This research indicates that by predicting the performance of the students, they can provide appropriate extra tasks to improve students' education. In ref [36], the answers of 702 students in the 10th grade were analyzed in chemistry lab classes. The researchers used the k-mean clustering method for segment answers, then a DL algorithm was employed for the development of explanatory models of the segmentation. Their results highlight that one key factor for chemistry learning is the attention and involvement of students in classes.

2.2. Evaluation of Student Performance for Reading Ability

Ref [37] automatically predicts the students' ability to read high-level literature. In this process, they discover mispronunciation during the tasks of reading sentences and words. In addition, the children should read a list of English words during voice recording. Therefore, the data contain native English and Spanish-speaking children's recorded voices from kindergarten to the second grade. In this research, first, an expert rated the recorded sounds with the correction of pronunciation, speaking rate and fluency criteria, and then a supervised ML predicted students' reading ability with extracted features from voices. A linear regression model was used with 0.946 accuracy and a 0.828 Pearson correlation coefficient.

2.3. Evaluation of Student Performance for Grading

In ref [38], the performance of 246 high school students in open-ended physics questions was graded by ML. The study was conducted on four physics questions, and they used the correlation coefficient method to feature selection and SMOT (synthetic minority over-sampling) to eliminate the imbalanced data problem. Bagging and AdaBoost.M1 (adaptive boosting) algorithms were applied as ensemble learning techniques (bagging and boosting) to define the grade (class) of the exam papers. The accuracy of the mentioned

algorithm ranged between 0.84–0.99 for Q1 (528 features), 0.65–0.84 for Q2 (1037 features), 0.85–0.91 for Q3 (724 features) and 0.87–0.98 for Q4 (316 features) for ten iterations. In ref [39], the scores of Portuguese high school students were predicted by a multilinear regression model (accuracy = 65.87), a random forest (accuracy = 94.42), a support vector machine (accuracy = 87.74), an artificial neural network (accuracy = 73.24), and an extreme gradient boosting machine (accuracy = 97.12). Optimal hyperparameters were found by the grid search and employed the lasso method to the features selection. Their datasets comprise the information of 362,261 high school students such as urban, income, age, employment, cultural level and grades. In the training process, the XGB has a better performance compared to the other algorithms. On the other hand, the SVR has a better performance on test data, which indicates XGB has an over-fit. After calculating the feature importance, the school size has a great impact on academic achievement. On the contrary, socioeconomic factors do not have a significant impact. The authors in this study [40] made a model for predicting middle school programming talent using ANN. The ANN is a satisfactory method for forecasting participants' skills, such as analytical thinking, problem-solving, and programming aptitude. They have used surveys to collect information that consists of four sections: demographic information, paper-folding problems, map sketching, and analytical thinking questions. The demographic information questions consist of gender, age, parents' educational levels, parents' jobs, grades in mathematics, homework time, the possession of an electronic device and time spent with the device, coding knowledge, and computer screen time. After the completion of the survey, the participants then took the 20-level Classic Maze course (CMC) on code.org. The participants' final scores in the CMC were calculated based on the level they completed and the lines of codes they wrote. Overall, 13 input features were provided as demographic information (five), grades in mathematics (one), paper-folding problems (three), map sketching (one), and analytical thinking questions (three). For the prediction phase, they employed ANN models consisting of one hidden layer, 13 inputs, and one target with accuracy = 91.26.

2.4. Evaluation of Student Performance for Dropout Prediction

In ref [41], an anonymized dataset with 3,575,724 instances was used to predict students' dropout probability. Features were employed to include free/reduced lunch eligibility and school demographic makeup and are drawn from a range of school districts, educational organizations, and agencies. The generalized algorithms on this data were decision trees, support vector machines, XGBoost, logistic regression, and random forests, of which the best were random forest classifiers (accuracy = 0.88) with n = 15 estimators and a max depth of 10. The authors have used the Gini importance algorithm to calculate the importance of each feature.

2.5. Evaluation of Student Performance for Academic Achievement

In ref [24], the authors analyzed and determined student achievement classes (KKM) with ANN, and the accuracy was 99.26. KKM is the level of achievement of basic competencies that must be achieved by students and is made as a teacher's reference in assessing the competencies of their students. A total of 131 records with 24 attributes (student information: number of student data, names of students, grades in all subjects, student attendance value, student behavior score, total number of subjects, and information on student grade promotion) were imported as the dataset.

3. Materials and Methods

The main objective of this study is to design and develop an ML system for classifying the performance of high school students in four classes as very good, good, medium and bad. In particular, the study aims to seek answers to the following research questions:
- To make the right decisions and appropriate policies, what are the main factors and most significant features for evaluating student performance? Additionally, which

of the three types of data, including demographics, behavior and grades, have the most influence?

- What is the best and most effective ML model for evaluating and classifying student performance that can determine a good boundary in the data?
- In order to make a more efficient and convenient model for predicting new instances, when we run models on data that only include the most important features, what impact on their performance occurs?

In order to answer the research questions, in this applied research, we are seeking to use ML models that can evaluate the performance of students with high precision to resolve this issue in education. In order to do that, as illustrated in Figure 1, the required data is collected from students by using an online questionnaire and this data is individually labeled by evaluators and then stored in the original database. ML models are fitted on the data and efficient models with high performance are identified. Then, the most important features for models and the assessment of students are identified and these features are separated from the original database and restored in the decreased database. Moreover, the ML models are trained again with the data in the decreased database. By eliminating low-importance features, the models become more practical for implementation. Furthermore, the best models are used to predict labels for new input features. Therefore, the human evaluators' part is no longer required in the process of predicting student performance and students are evaluated with high accuracy by ML models.
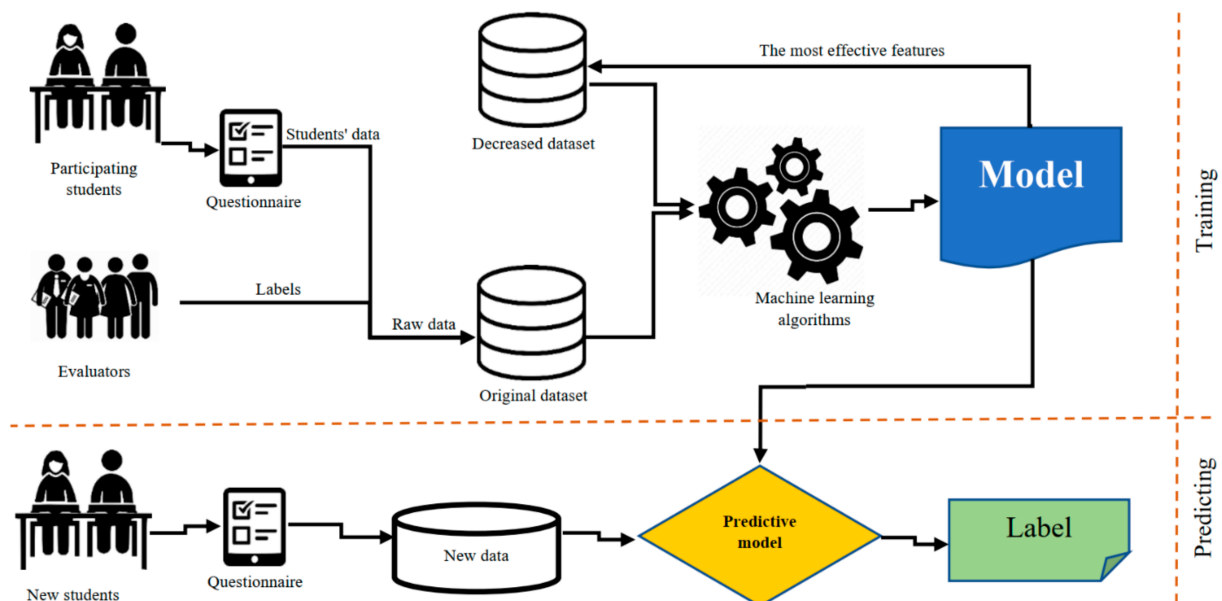
Figure 1. An overview of the research approach.

3.1. Dataset

The dataset used in this study was collected from 459 high school students who have been studying in different levels and different fields such as humanism, science, and technical studies in 2020–2021. One of the high schools is a boys' school and another is a girls' school. Both high schools are located in Arian Shahr, South Khorasan Province, Iran.

Various aspects can be considered for evaluating the performance of students, and different conclusions can be obtained from the perspective of each one. Hence, we must have extensive features for an accurate assessment. Through a scrutiny of the literature regarding the integration of AI in education and evaluating student performance with ML, the most important and influential characteristics that expressed student performance in these studies were identified. These studies are mentioned in the related work section. Many features can be considered for student evaluation. In general, students' actions are

very decisive and determinative because the performance of a student depends on his actions and the manner they were conducted. In this study, students' actions are categorized into two main categories: a behavioral category that represents the usual and individual behavior of the student and grades that show the student's efficiency in different fields of study. Moreover, demographic features of students and their families were included to consider environmental factors affecting the students. These selected features give general and significant information about student efficiency, and students are comprehensively monitored. A total of 65 features were identified as suitable features. Too many questions cause fatigue in students and their focus is decreased, so the quality of the answers is reduced. To optimize data collection, we reduced the number of characteristics to 35 by eliminating less appropriate options which were not too important for the evaluation and were too time-consuming and costly for collecting. We kept only features that had a direct relationship with the performance of the students. The 35 selected features cover all aspects of a student and are a proper measure for performance evaluation. After that, according to the required features, we designed questions that students responded to by selecting existing options. One-to-one surveys were conducted with the students.

Three major types of data that were collected from each student included demographic, behavioral, and grades data. Behavioral and demographic data were collected through the online questionnaire and grades were obtained from teachers. Table 1 indicates the attributes of the dataset that elaborate the 35 features used in this research and each value of the feature. In detail, demographic data describe the personal information of each student. In this part of the data, families with more than three members and less than three members are labeled by L and S, respectively. The quality of family relationships is categorized into four levels. The student's guardian, parents' education and parents' jobs were demanded. To determine parents' jobs, five overall categories were determined as follows: (1) education and social occupations such as teachers, the press, psychologists, etc.; (2) health occupations such as doctors and nurses, etc.; (3) service occupations such as in sales, sports, tourism, etc.; (4) agriculture and environmental occupations such as botany, livestock, etc.; and (5) engineering. The behavioral data of the students include information such as routine personal and educational life. The level of socialization of the student is characterized by friendly relationships and go-out rate. How students use non-school time was determined with questions about the amount of internet, television and video game usage. Other features, such as extra classes, identify informal classes outside of school hours. Two important features, the number of absences and attention in class, were added to the dataset with the help of teachers. Due to the fact that grades play a major role in student assessment in the educational system of Iran, the scores of student exams in different classes were added to the data as a variable for evaluating his/her performance. The wide range of class subjects, including math, science, foreign language, sports and art, was used two times in the semester for a comprehensive evaluation. These grades were obtained from the assessment of students during the semester by teachers of each class which were specified in the dataset with "first class name midterm" and "second class name midterm". Overall, 35 features were selected to establish the research based on these and develop an ML system that evaluates students based on these features. Moreover, the most important features that determine the students' performance will be identified and the performance of the models examined on these identified features. After collecting the information, each student was evaluated by six teachers, two principals and two authors of this paper and the performance of each of them were labeled with A (very good), B (good), C (medium) and D (bad) according to the dataset features collected.

3.2. Model Selection

Machine learning models were used to classify the performance of high school students. ML is a subset of computer science that is applied in mathematics and statistics. The main objective of supervised ML is to build an algorithm that can receive input data and use statistical analysis to predict outputs while being updated by adding new data. In the

past decade, artificial intelligence and machine learning have been drawn toward education tasks, including the extraction of useful data and knowledge production to support decision making for students' progress. In this study, student performance was evaluated by supervised ML models. Studies that had similar data have used random forest [42–44], gradient boosting [41], support vector machines (SVM) [45–47], elastic net [12], naive Bayes [15,34,48], logistic regression [12,41,49], decision tree [42,50,51], and ANN [52–54] algorithms. Therefore, the preprocessed data executed on previously mentioned models and best performance with the default parameters were obtained by RF, SVM and LR models; Thus, these models were selected for more tuning and analysis.

Table 1. Attributes of the database with value details.

Demographics Data			Behavioral Data			Grades	
Variable	**Group**	**Count**	**Variable**	**Group**	**Count**	**Variable**	**Group**
Gender	Male	221		Less t 1	23	First math midterm	0–20
	Female	237		1–2	127		
Age	15–18	-	Study time	2–3	177	Second math midterm	0–20
Family members	L	301		3–4	82		
	S	157		More t 4	49	First science midterm	0–20
	Excellent	118	Supervision by school	Yes	78		
	Good	253		No	380	Second science midterm	0–20
Family relation	Normal	56	Supervision by family	Yes	256		
	Bad	21		No	202	First language midterm	0–20
	Worse	10	Higher edu	Yes	404		
	Father	319		No	54	Second language midterm	0–20
Guardian	mother	107	Internet	Yes	374		
	other	32		No	84	First sports midterm	0–20
	No formal	9	Extra class	Yes	137		
Mother Educational Background	Primary. s	54		No	285	Second sports midterm	0–20
	High. s	140	Work	Yes	176		
	Bachelor	201		No	282	First art midterm	0–20
	Graduate	55		Less t 1			
	No formal	15	Usage of each internet, TV, game	1–2		Second art midterm	0–20
Father Educational Background	Primary. s	86		2–3			
	High. s	99		3–4			
	Bachelor	197		More t 4			
	Graduate	62		1	59		
	Education and social	152		2	91		
	Health	146	Free time	3	169		
Mother-job	Services	61		4	104		
	Agriculture and env	55		5	35		
	Engineering	44		1	27		
	Education and social	229		2	119		
	Health	153	Go-out rate	3	146		
Father-job	Services	34		4	90		
	Agriculture and env	25		5	76		
	Engineering	17		1	40		
	0	300		2	75		
Travel duration	1	120	Friendship	3	82		
	2	33		4	102		
	3	5		5	160		
			Absences	0–10			
			Attention	0–100			

3.2.1. Random Forest (RF)

Random forests or random decision forests are a well-known method for classification and regression tasks and is an ensemble model made of many decision trees. A random subset of data and features are imported to each tree, and trees are fitted on this subset for

the training of decision trees in the model. Moreover, RF uses the average votes of trees to make predictions [55].

3.2.2. Support Vector Machines (SVM)

SVM is a powerful and versatile model of machine learning that is capable of conducting classification and regression tasks in linear and nonlinear space. SVM is a suitable model for medium and small dataset classification and has good performance in this kind of dataset [55]. By tuning the parameters, such as C, the model's hardness in the violation data into the margins can be controlled, and the narrower bell-shaped curve can be determined with the gamma parameter which all these hyperparameters are going to be tuned into the models' training phase.

3.2.3. Logistic Regression (LR)

LR is a common model of statistics, which is used in ML and calculates the outputs' probability. It is also called the sigmoid function. LR is a linear model, but predictions are deformed using the logistic function. LR itself is a binary classifier; however, multi-classification can be done with multi-binary LR. The algorithm coefficients should be estimated from the training data which, for this matter, the algorithm used maximum likelihood estimation.

3.2.4. Artificial Neural Network (ANN)

ANNs are derived from neurons from the brains of living organisms that perform the calculations. ANNs are based on several neurons or nodes along with each other, which form a layer called the hidden layer. By placing these hidden layers sequentially, the neural network forms. All neurons of each layer are connected to the front and back layers; those connections are called weight and such networks are called fully connected layers. Each neuron has an activation function. The activation function is a function that takes the sum of data as input and generates output data [56]. Another important component of ANNs is the loss function that calculates the gradient (prediction error) of the model. The loss function is used for optimizing the weight and improves the performance of the model. The ANN network has many parameters and is capable of discovering complex relationships between data. Therefore, an ANN model is fitted on the existing dataset. Different numbers of neurons, hidden layers and various types of loss function were employed in the current study. Finally, the most accurate and optimum performance was provided by our designed network. The designed network has two hidden layers plus input and output layers. The arrangement of layers consists of 35 neurons with a linear activation function, and in the next layer, there are 105 neurons with a linear activation function that forms the first hidden layer. Then, a dropout exists between this layer and the next layer that eliminates 40 percent of non-necessary connections. In the second hidden layer, there are 70 neurons with linear activation functions that are connected to the four neurons with a SoftMax activation function in the output layer. The most optimal loss function was the mean squared error and the most optimal optimizer was the Adam optimizer with a 0.0005 learning rate. The network was run with a batch size equal to 47 in 200 epochs.

3.3. Feature Selection

Due to the cost and time of collecting data and importing new data into the existing model to receive predictions by the model, the existence of many features may be problematic, and the utilization of an operational model or framework will be time-consuming and difficult. Therefore, the required features of the model are lower; firstly, the cost of collecting and training the model is reduced. Secondly, student performance prediction is done with less information which increases the utilization and comfort of the model. This may or may not affect the accuracy and performance of the model. To evaluate and calculate the effect of any of the features and their importance for the model, the Boruta algorithm was employed. The "Boruta algorithm is a wrapper built around the random

forest classification algorithm" [57], and works similarly to the random forest in that it is an assembly approach based on several classifier votes. In this algorithm, a copy of all features adds to the randomness dataset and then a random forest model is fitted to the dataset and examines the importance of each feature. In each epoch, the features that have more importance are determined and features with low importance are eliminated.

To design and develop the above-mentioned models, the Python programming language was used with popular libraries such as pandas, NumPy, scikit-learn, TensorFlow and Keras. Each of the models has several hyperparameters that specify how the model calculates. The models were trained using cross-validation techniques and the best hyperparameters were identified during the training phase. Cross-validation is a useful method for evaluating and comparing the learning of an algorithm, which divides the input data into several parts and each of these sectors is called a fold. Then, the model is trained on the fold-1 part of input data and evaluated by one remaining fold that was set aside.

4. Results

Some attributes of the dataset contain categorical and string values such as parents' education or job attributes. Hence, this data should be encoded to numerical values for the training process. Due to the different range of feature values, data were normalized between 1 and 2 because some models such as SVM are sensitive to these different scales. Data were run on the mentioned models after pre-processing and preparing, and models were evaluated by the k-fold cross-validation technique in which the K variable was considered 10. In Figure 2, the models' learning curves are illustrated. In these curves, the accuracy rates of training and cross-validation models are shown during the addition of more instances (or more epochs).
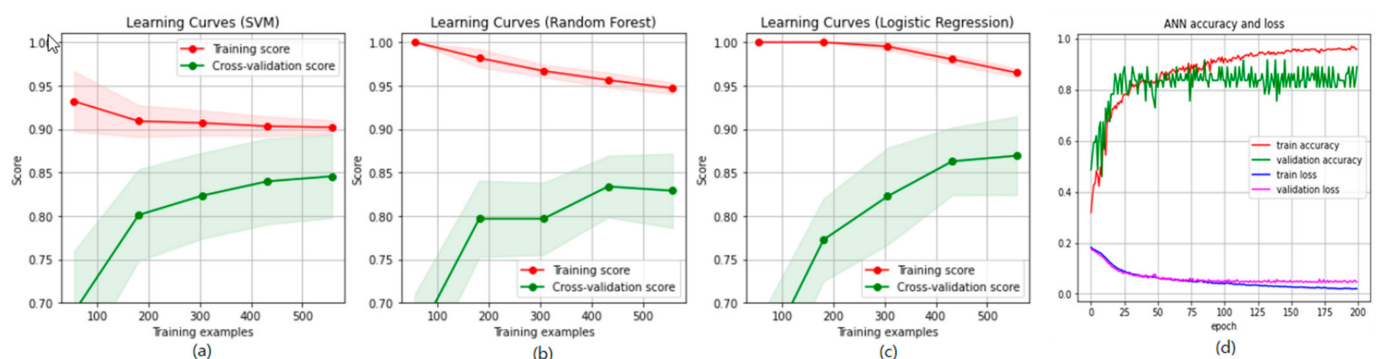

Figure 2. Learning curve of models on the train set and cross-validation. (**a**) is associated with the learning curve of SVM, (**b**) is associated with RF, (**c**) is associated with LR and (**d**) is associated with ANN.

As demonstrated in Figure 2, the models have low bias so that selective models are suitable for our data and have good performance; however, the RF and LR models have more bias than SVM. Nevertheless, they are not overfitted and this bias is negligible. In the case where more training data is available, the models would have more certainty and the value of the bias would decrease. In Figure 3, the accuracy of the models obtained on training data and test data of the original dataset is shown on the box chart.

The ANN model has the best performance on the training data. Moreover, this model has more certainty in the accuracy results than other models. In contrast, the RF model has a varied range of results, and the maximum and minimum lines are much different. Besides, RF has the least accuracy (0.82) on training data. On the right-hand side of Figure 3, box charts show the performance of the models on the test data. The accuracies of the models are diverse. The best performance of the models belongs to ANN and then, in order from highest to lowest, SVM, LR and RF models have lower accuracy in terms of the mean value. Notice that, the yellow line inside the boxes shows the median number and the white number inside the boxes shows the mean value in cross-validation results.

Therefore, there is a difference between mean and median values in the box related to the LR model. If the median value was considered to compare the performance of models, the LR model has the worst performance on test data but the upper side of the LR box is larger than other boxes and this indicates that the model also has high accuracy in some iterations. Figure 4 illustrates the confusion matrix of classifiers that shows the model's ability to distinguish class boundaries. All models have had a good ability to distinguish classes. Although models have been misclassified on the boundary between each class, in class 0 and 1, which represent the A and B classes, all models have had great separation.
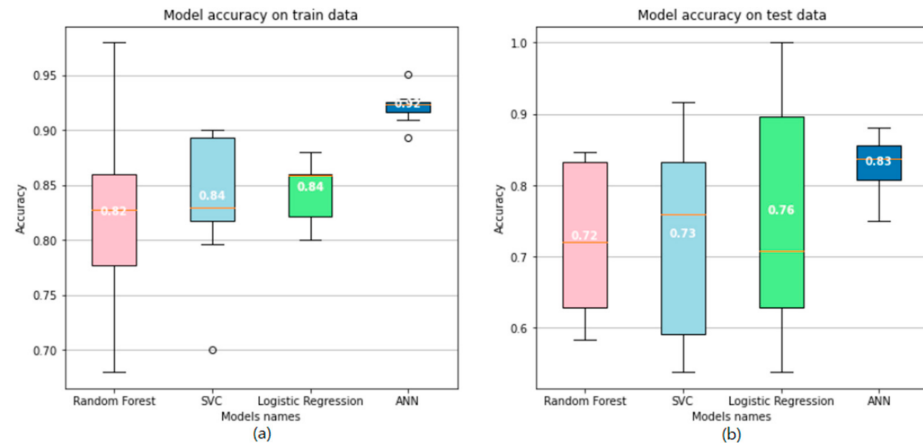
Figure 3. Accuracy of models on the original dataset. (**a**) demonstrates the accuracy of the models on train data and (**b**) demonstrates the accuracy of the models on test data.
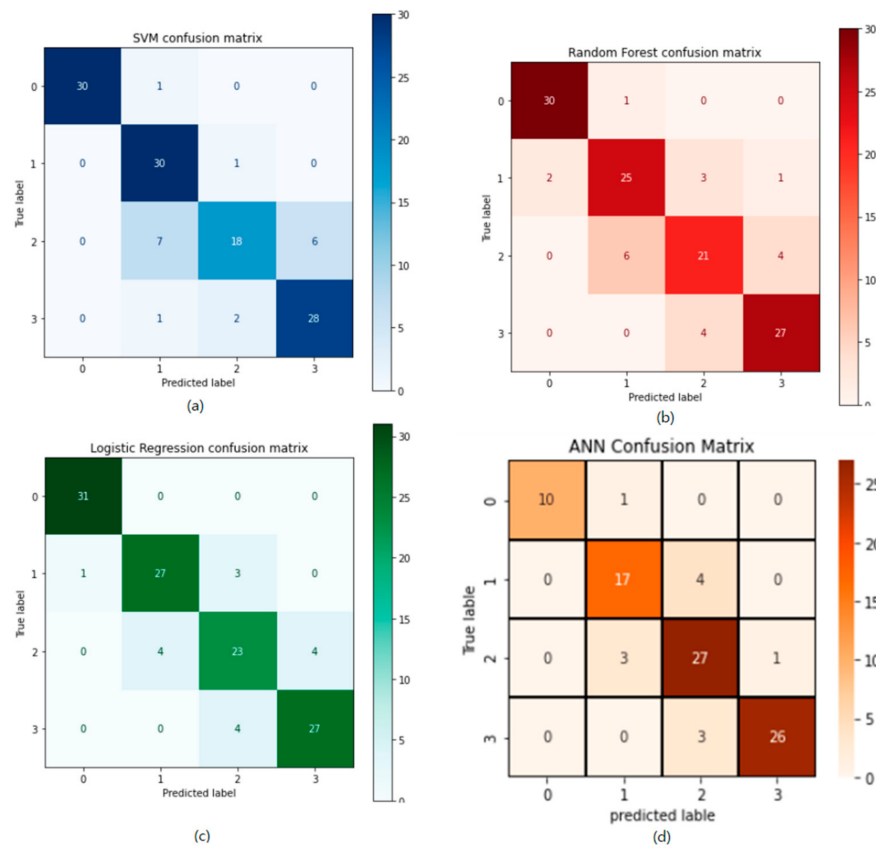
Figure 4. (**a**) is confusion matrix of SVM, (**b**) is confusion matrix of RF, (**c**) is confusion matrix of LR and (**d**) is confusion matrix of ANN.

The Boruta algorithm was used to obtain the most important feature of the dataset to evaluate the performance of students. This algorithm was implemented on the preprocessed original dataset using ten random forest estimators. The results of the Boruta algorithm are in Table 2. In this table, the importance for each feature, which is called ranking, is specified for each feature. The number of features that the model decides to be maintained can be specified by adjusting the algorithm threshold. As the table indicates, the number of high-quality features which have a number 1 ranking are selected for keeping. Thus, the number of features that are important for the model is 17.

Table 2. The results of the Boruta algorithm.

Feature Names	Ranking	Support	Feature Names	Ranking	Support	Feature Names	Ranking	Support
Gender	14	FALSE	Sup-family	18	FALSE	First math midterm	1	TRUE
Age	11	FALSE	Higher-edu	17	FALSE	Second math midterm	1	TRUE
Family-members	12	FALSE	Internet	19	FALSE	First science midterm	1	TRUE
Family-relation	10	FALSE	Extra class	4	FALSE	Second science midterm	1	TRUE
Mother-edu	6	FALSE	Work	3	FALSE	First language midterm	1	TRUE
Father-edu	8	FALSE	Internet use	1	TRUE	Second language midterm	1	TRUE
Mother-job	6	FALSE	TV use	1	TRUE	First sports midterm	1	TRUE
Father-job	9	FALSE	Do Game	1	TRUE	Second art midterm	1	TRUE
Guardian	15	FALSE	Free time	5	FALSE	First art midterm	1	TRUE
Travel-duration	12	FALSE	Go-out rate	2	FALSE	Second art midterm	1	TRUE
Study time	1	TRUE	Absences	1	TRUE	Friendship	1	TRUE
Sup-school	16	FALSE	Attention	1	TRUE			

To make the model more convenient and efficient, low importance data which ranked lower than 1, were eliminated from the dataset by the Boruta algorithm and the models were retrained again on the new dataset by a cross-validation technique with 10-fold. The performance of the models on the train and test instance of the decreased dataset are shown in Figure 5.
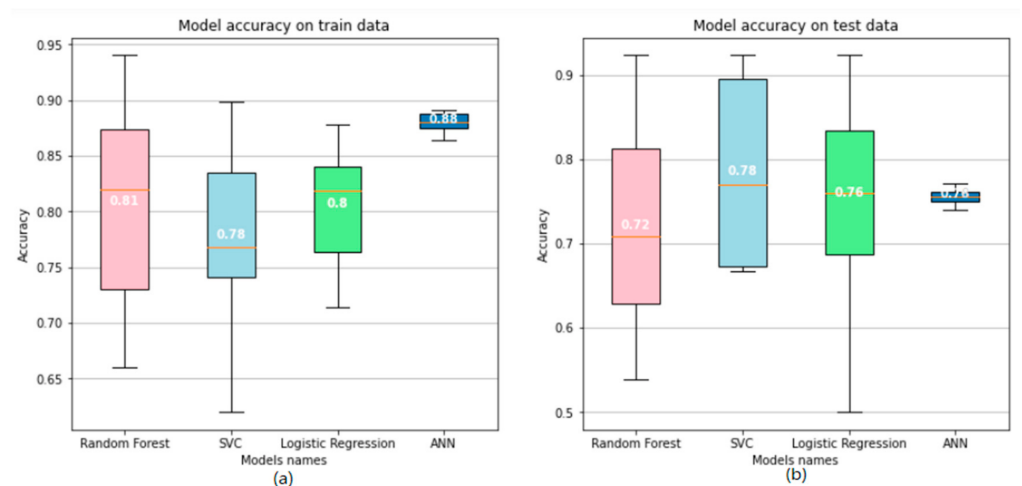

Figure 5. Accuracy of the models on the decreased dataset. (**a**) demonstrates the accuracy of the models on train data and (**b**) demonstrates the accuracy of the models on test data.

On the right-hand side of Figure 5, the accuracy of the models on training data with Ranking 1 is shown. The length of the boxes is long and the results of the model's performance on the decreased dataset have a more diverse range than the results on the original dataset. The average accuracy value of the SVM model has the most change and

the elimination of low important features reduced the performance of this model. The ANN and LR model and then the RF model had less impact. The left side of Figure 5 shows the performance of the models on the test set of decreased datasets. The elimination of low-importance features has reduced the accuracy of the ANN model. Thus, this model has the highest impact among the remaining models. RF and LR models have maintained their accuracy and have similar performance such as their performance on the original dataset. Although the accuracy of the SVM model was reduced in the training data, the accuracy of this model has been better in the test set of the decreased dataset. Moreover, this model has the highest accuracy among the rest of the models. In general, by reducing the low importance features, models have been more optimized and the difference between the accuracy of the train and the test instances is lower; thus, the models are more fit than models on the original dataset.

5. Discussion

The classification that was conducted in this research is based on information about students' personal lifestyles, parent information, and scores in each discipline (lesson). The type of data features used in this study was selected based on studies on student performance evaluation using ML and the data features it had used [15,24,39,40,42,52]. These features include all aspects of students' home life, school life, how they use their time, interests, work, etc. Therefore, an evaluation of a student with these features provides a comprehensive result of his/her performance. Overall, 35 attributes were considered for each instance. Next, the dataset was imported into selected models including RF, ANN, LR and SVM, which were trained by a cross-validation technique where k = 10.

Among the models used, the ANN model with two hidden layers containing 105 and 70 neurons had the best performance (i.e., 0.83) on the original dataset. Moreover, the LR model was able to provide good performance in some iterations. The highest accuracy performed by this model was on the upper quartile of the chart in Figure 3 that was equal to 0.90. Besides, the performance of the SVM model had the most changes on the decreased training data, and the accuracy of this model decreased. However, the SVM accuracy on the decreased test data was much better; thus, it had the best performance compared to the other models. Moreover, after reducing the dataset from the low-importance features, the RF and LR model still had the same accuracy as in the original dataset. Although the models are able to provide the same performance, they became more efficient and more convenient to use for predicting a new instance. In the case where models are more efficient, non-expert users can use them more conveniently with high precision. Consequently, this will make this system more promising, and the trend of replacing artificial intelligence for educational evaluations will accelerate. Furthermore, models have a few variances that do not improve with hyperparameter tuning in the original dataset. To resolve this problem, more data should be collected and entered into the models to reduce the variance. Our findings are consistent with [44,51] that have a similar dataset to ours. In terms of the accuracy of ANN, RF and SVM, our models have better accuracy than [39], [34,39], respectively. In other words, when the models were implemented in the decreased dataset, which only includes the most important and significant features, first, the difference in accuracy in train and test instances was more adjacent, so the variance of models decreased, and the models were more optimized in our data. This demonstrates that some of the unnecessary features in the dataset increase noise and make the algorithms confused. Thus, the algorithms are not able to fit properly on the data. Second, the accuracy of the ANN and SVM models was changed after eliminating features with a lower rank than 1. On the contrary, the ANN model had a noticeable fall among other models, and the SVM model accuracy increased moderately to 0.78 and reached the highest accuracy among other models. The efficiency of the LR and RF remained unchanged, and they were able to offer the same accuracy.

To calculate the most important features, the Borate algorithm was employed. This algorithm calculates the importance of each feature and removes the features with low

importance. The most important features for evaluating student performance are, in order of importance, absences, second math midterm, attention, second science midterm, first sports midterm, second art midterm, first science midterm, second language midterm, first language midterm, first art midterm, first math midterm, internet use, second sports midterm and study time. In contrast, low-importance features were supervision by the school, internet, higher education, guardian, supervision by family, gender, travel duration, father's and mother's educational background and family relations. These low-importance features are randomly valued and do not have significant value for the models. However, some features, in order of the amount of its importance, including age, free time, father's job, mother's job, extra classes, work, go-out rate, friendship, usage of TV and games are more important than previous features, and help models to have better classification. Indeed, attention in the class is very important, and whenever a student pays more attention in class, they will have more focus and more understanding of the class content [36]. Thus, to improve the student's performance, attractive classroom activities must be created in order for students to have more enthusiasm to pay attention. One of the disadvantages in the education system of Iran is the teacher-centered classroom. By changing this method to student-centered classrooms, more attraction will be created, and the attention of students will be attracted. Moreover, in cases where the student becomes absent, he/she should not be left behind from the curriculum and it should be compensated by arrangements such as recording the class for them, additional classes, additional support and so on. Another important factor is the students' grades in the lessons, which have much more weight compared to the rest of the variables. To evaluate student performance, we should not only consider these criteria; other influential criteria such as problem-solving, critical thinking, creativity, etc., should somehow be included in calculations. Excessive usage of the internet and social media reduce the attention of youth and distract them [58], so less use of the internet increases student's focus and performance as well. In general, for evaluating student performance, the most influential type of data is grades, and all features related to this category remained in the decreased dataset. Additionally, 50 percent of behavioral features were identified as important features by the Borate algorithm.

Due to the lack of an appropriate educational database and insufficient connections between the available databases, the process of data collection in this way was difficult. Although the questions in the questionnaire were explained for students during an online justified session, the questionnaires were filled out by students online, and there was no supervision during the process of filling out the questionnaires. Consequently, some information was unreliable due to a lack of student's proper understanding of the questions, so these data were removed. Furthermore, some further features were not included in the data, such as teacher quality and parent salary, because of the absence of resources. Adding these features could provide a more accurate assessment.

The developed models in this research evaluate the overall performance of students promptly, and this evaluation is without bias and intention which makes the system impressive. For a more accurate assessment of the student, all the features of the original dataset can be used in analysis. Otherwise, this can be done in a faster way with 17 of the features mentioned earlier. Moreover, our results allow decision-makers, schools and teachers to better understand the factors affecting students' performance at the individual level. This valuable information can be used by politicians to adopt correct decisions and choose appropriate strategies to improve education. The information obtained can be used to understand the improvement of student academic achievement by teachers. Efficient models and explainable AI could allow the Iran Ministry of Education to make an informed decision for better student achievement and the policies of employing AI in schools.

6. Conclusions

In the present study, an ML-based system was designed and developed to evaluate the performance of high school students in the overall state. This system evaluates students from different aspects and classifies each student into four classes: very well, good, medium

and bad. Due to the lack of appropriate educational databases, the best information needed was collected individually with the help of online questionnaires and the collaboration of teachers. The most influential features were determined and analyzed, and the most efficient and optimal ML models were identified as well. Moreover, the models were trained again with the decreased dataset that contained data with the most important features. Therefore, more efficient models were made, and students were evaluated without human intervention.

For further research, the collected data can be used for other work such as dropout prediction and academic achievement by adding related labels. Additionally, the performance of models can be improved by adding more new data. As mentioned above, the student's attention is an influential criterion for student evaluation. Sohlberg and Mateer developed a model for attention sub-components based on clinical cases of experimental neuropsychology. Students' attention in class can be divided into three parts based on this model. These parts include focused attention, which refers to their ability to focus on a stimulus, sustained attention, the ability to keep this focus over a long period, and selective attention, the ability to focus on a particular stimulus when there are other distracting stimuli [59]. For further research, to consider the percentage of students' attention, each of these sub-components can be included in the calculation, or all sub-components can be used to obtain a precise and general attention criterion. This task can easily be performed by using intelligence systems. By developing an algorithm such as deep learning, students' reactions in class can be monitored, and by tracking the eye angles of the students, their response to class stimuli can be detected. By employing this system, if there is a lack of attention from students, the system warns teachers or students, and students will be able to maintain their attention in class. Additionally, the precise percentage of the focused attention, sustained attention and even selective attention will be obtained. The obtained data from the system are reliable and without human error or bias. This will make the attention criteria more accurate for evaluation.

Author Contributions: Conceptualization, M.Z. and A.S.-N.; methodology, M.Z., and A.S.-N.; software, M.Z.; validation, M.Z., A.S.-N., S.-M.C., and A.E; formal analysis, M.Z.; investigation, A.S.-N., S.-M.C., and A.E; resources, M.Z., and A.E; data curation, M.Z. and A.S.-N.; writing—original draft preparation, M.Z..; review and editing, M.Z., A.S.-N., S.-M.C., and A.E; supervision, A.S.-N. and A.E; project administration, S.-M.C.; funding acquisition, S.-M.C. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Ethics approval is not required for this type of study, according to the local legislation.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available on request due to restrictions e.g., privacy or ethical.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mondal, K. A Synergy of Artificial Intelligence and Education in the 21 st Century Classrooms. In Proceedings of the 2019 International Conference on Digitization (ICD), Sharjah, United Arab Emirates, 18–19 November 2019; pp. 68–70.
2. Awad, M.; Khanna, R. Machine learning in action: Examples. In *Efficient Learning Machines*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 209–240.
3. Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Choi, S.-M. Asthma-prone areas modeling using a machine learning model. *Sci. Rep.* **2021**, *11*, 1912. [CrossRef] [PubMed]
4. Sadeghi-Niaraki, A.; Mirshafiei, P.; Shakeri, M.; Choi, S.-M. Short-Term Traffic Flow Prediction Using the Modified Elman Recurrent Neural Network Optimized Through a Genetic Algorithm. *IEEE Access* **2020**, *8*, 217526–217540. [CrossRef]

5. Karsenti, T. Artificial intelligence in education: The urgent need to prepare teachers for tomorrow's schools. *Form. Prof.* **2019**, *27*, 112–116. [CrossRef]

6. Fahimirad, M.; Kotamjani, S.S. A review on application of artificial intelligence in teaching and learning in educational contexts. *Int. J. Learn. Dev.* **2018**, *8*, 106–118. [CrossRef]

7. Li, H.; Dai, T. Explore Deep Learning for Chinese Essay Automated Scoring. *J. Phys.* **2020**, *1631*, 012036. [CrossRef]

8. Arianti, N.D.; Irfan, M.; Syaripudin, U.; Mariana, D.; Rosmawarni, N.; Maylawati, D.S. Porter Stemmer and Cosine Similarity for Automated Essay Assessment. In Proceedings of the 2019 5th International Conference on Computing Engineering and Design (ICCED), Singapore, 11–13 April 2019.

9. Filho, A.H.; Do Prado, H.A.; Ferneda, E.; Nau, J. An Approach to Evaluate Adherence to the Theme and the Argumentative Structure of Essays. *Procedia Comput. Sci.* **2018**, *126*, 788–797. [CrossRef]

10. Yun, W.H.; Lee, D.; Park, C.; Kim, J.; Kim, J. Automatic Recognition of Children Engagement from Facial Video Using Convolutional Neural Networks. *IEEE Trans. Affect. Comput.* **2020**, *11*, 696–707. [CrossRef]

11. Romero, C.; Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* **2007**, *33*, 135–146. [CrossRef]

12. Cornell-Farrow, S.; Garrard, R. Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia. *Commun. Stat. Case Stud. Data Anal. Appl.* **2020**, *6*, 228–246. [CrossRef]

13. Silva, C.; Fonseca, J. *Educational Data Mining: A Literature Review*; Europe and MENA Cooperation Advances in Information and Communication Technologies; Springer: Cham, Switzerland, 2017; pp. 87–94.

14. Romero, C.; Ventura, S. Data Mining in Education. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 12–27. [CrossRef]

15. Harvey, J.L.; Kumar, S.A.P. A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning. In Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019; pp. 3004–3011.

16. Cassano, R.; Costa, V.; Fornasari, T. An Effective National Evaluation System of Schools for Sustainable Development: A Comparative European Analysis. *Sustainability* **2019**, *11*, 195. [CrossRef]

17. Clark, N. Education in Iran. World Education News Reviews. Available online: https://wenr.wes.org/2017/02/education-in-iran (accessed on 7 February 2017).

18. Athani, S.S.; Kodli, S.A.; Banavasi, M.N.; Hiremath, P.G.S. Student performance predictor using multiclass support vector classification algorithm. In Proceedings of the 2017 International Conference on Signal Processing and Communication (ICSPC), Coimbatore, India, 28–29 July 2017; pp. 341–346.

19. Chen, J.F.; Do, Q.H. Training neural networks to predict student academic performance: A comparison of cuckoo search and gravitational search algorithms. *Int. J. Comput. Intell. Appl.* **2014**, *13*, 1450005. [CrossRef]

20. Costa, E.J.F.; Campelo, C.E.C.; Campos, L.M.R.S. Automatic Classification of Computational Thinking Skills in Elementary School Math Questions. In Proceedings of the 2019 IEEE Frontiers in Education Conference (FIE), Covington, KY, USA, 16–19 October 2019.

21. Du, Y.; Yang, L. What affects the difficulty of Chinese syntax? In Proceedings of the 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 15–17 November 2019; pp. 71–74.

22. Cai, C.Y.; Yan, K.; Lu, H.; Ye, M. Intelligent Placement Model Based on Decision Tree. In Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 19–21 October 2018; pp. 837–841.

23. Yagci, A.; Cevik, M. 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.* **2019**, *24*, 2741–2761. [CrossRef]

24. Saiful, M. Implementation of the Neural Network (NN) Algorithm in Analysis of Student Class Increment Data Based on Report Card Value. *J. Phys.* **2020**, *1539*, 012034.

25. Chen, C.M.; Wang, J.Y.; Chen, Y.T.; Wu, J.H. Forecasting reading anxiety for promoting English-language reading performance based on reading annotation behavior. *Interact. Learn. Environ.* **2016**, *24*, 681–705. [CrossRef]

26. Amir, O.; Gal, K.; Yaron, D.; Karabinos, M.; Belford, R. Plan recognition and visualization in exploratory learning environments. In *Studies in Computational Intelligence*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 524, pp. 289–327.

27. Ma, K.; Yang, L. Automatic Extraction and Quantitative Evaluation of the Character Relationship Networks from Children's Literature works. In Proceedings of the 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 15–17 November 2019; pp. 188–193.

28. Salim, Y.; Stevanus, V.; Barlian, E.; Sari, A.C.; Suhartono, D. Automated English Digital Essay Grader Using Machine Learning. In Proceedings of the 2019 IEEE International Conference on Engineering, Technology and Education (TALE), Yogyakarta, Indonesia, 10–13 December 2019.

29. Saha, S.K.; Rao Ch, D. Development of a practical system for computerized evaluation of descriptive answers of middle school level students. *Interact. Learn. Environ.* **2019**, 1–14. [CrossRef]

30. Gil, J.S.; Delima, A.J.P.; Vilchez, R.N. Predicting students' dropout indicators in public school using data mining approaches. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 774–778. [CrossRef]

31. Nangia, S.; Anurag, J.; Gambhir, I. A Machine Learning Approach to Identify the Students at the Risk of Dropping Out of Secondary Education in India. In *International Conference on Soft Computing and Signal Processing*; Springer: Singapore, 2020; Volume 1118, pp. 557–569.

32. Sansone, D. Beyond Early Warning Indicators: High School Dropout and Machine Learning. *Oxf. Bull. Econ. Stat.* **2019**, *81*, 456–485. [CrossRef]

33. Şara, N.B.; Halland, R.; Igel, C.; Alstrup, S. High-school dropout prediction using machine learning: A Danish large-scale study. In *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015-Proceedings*; Louvain-la-Neuve Ciaco: Bruges, Belgium, 2015; pp. 319–324.

34. Kostopoulos, G.; Kotsiantis, S.; Verykios, V.S. A prognosis of junior high school students' performance based on active learning methods. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2017; pp. 67–76.

35. De Melo, G.; Vasconcelos Filho, E.P.; Oliveira, S.M.; Calixto, W.P.; Ferreira, C.C.; Furriel, G.P. Evaluation techniques of machine learning in task of reprovation prediction of technical high school students. In Proceedings of the 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Pucon, Chile, 18–20 October 2017; pp. 1–7.

36. Figueiredo, M.; Lurdes Esteves, M.; Neves, J.; Vicente, H. Lab classes in chemistry learning an artificial intelligence view. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2014; Volume 299, pp. 565–575.

37. Black, M.P.; Tepperman, J.; Narayanan, S.S. Automatic prediction of children's reading ability for high-level literacy assessment. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 1015–1028. [CrossRef]

38. Çınar, A.; Ince, E.; Gezer, M.; Yılmaz, Ö. Machine learning algorithm for grading open-ended physics questions in Turkish. *Educ. Inf. Technol.* **2020**, *25*, 3821–3844. [CrossRef]

39. Costa-Mendes, R.; Oliveira, T.; Castelli, M.; Cruz-Jesus, F. A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. *Educ. Inf. Technol.* **2021**, *26*, 1527–1547. [CrossRef]

40. Çetinkaya, A.; Baykan, Ö.K. Prediction of middle school students' programming talent using artificial neural networks. *Eng. Sci. Technol. Int. J.* **2020**, *23*, 1301–1307. [CrossRef]

41. Coleman, C.; Baker, R.S.; Stephenson, S. A better cold-start for early prediction of student at-risk status in new school districts. In Proceedings of the International Conference on Educational Data Mining (EDM), Montreal, QC, Canada, 2–5 July 2019; pp. 732–737.

42. Lee, S.; Chung, J.Y. The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Appl. Sci.* **2019**, *9*, 3093. [CrossRef]

43. Rebai, S.; Ben Yahia, F.; Essid, H. A graphically based machine learning approach to predict secondary schools performance in Tunisia. *Socio-Econ. Plan. Sci.* **2020**, *70*, 100724. [CrossRef]

44. Aguiar, E.; Lakkaraju, H.; Bhanpuri, N.; Miller, D.; Yuhas, B.; Addison, K.L. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, Poughkeepsie, NY, USA, 16–20 March 2015; pp. 93–102.

45. Xiao, Y.; Hu, J. Assessment of Optimal Pedagogical Factors for Canadian ESL Learner's Reading Literacy Through Artificial Intelligence Algorithms. *Int. J. Engl. Linguist.* **2019**, *9*, 1–14. [CrossRef]

46. Taga, M.; Onishi, T.; Hirokawa, S. Automated Evaluation of Students Comments Regarding Correct Concepts and Misconceptions of Convex Lenses. In Proceedings of the 2018 7th International Congress on Advanced Applied Informatics (IIAI-AAI), Yonago, Japan, 8–13 July 2018; pp. 273–277.

47. Tomkins, S.; Ramesh, A.; Getoor, L. Predicting Post-Test Performance from Online Student Behavior: A High School MOOC Case Study. In Proceedings of the International Conference on Educational Data Mining (EDM), Raleigh, NC, USA, 29 June–2 July 2016; pp. 239–246.

48. Aslan, S.; Cataltepe, Z.; Diner, I.; Dundar, O.; Esme, A.A.; Ferens, R.; Kamhi, G.; Oktay, E.; Soysal, C.; Yener, M. Learner Engagement Measurement and Classification in 1:1 Learning. In Proceedings of the 2014 13th International Conference on Machine Learning and Applications, Detroit, MI, USA, 3–6 December 2014; pp. 545–552.

49. Hu, X. Automated recognition of thinking orders in secondary school student writings. *Learn. Res. Pract.* **2017**, *3*, 30–41. [CrossRef]

50. Yousafzai, B.K.; Hayat, M.; Afzal, S. Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. *Educ. Inf. Technol.* **2020**, *25*, 4677–4697. [CrossRef]

51. Cruz-Jesus, F.; Castelli, M.; Oliveira, T.; Mendes, R.; Nunes, C.; Sa-Velho, M.; Rosa-Louro, A. Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon* **2020**, *6*, e04081. [CrossRef]

52. Hung, J.L.; Rice, K.; Kepka, J.; Yang, J. Improving predictive power through deep learning analysis of K-12 online student behaviors and discussion board content. *Inf. Discov. Deliv.* **2020**, *48*, 199–212. [CrossRef]

53. Sokkhey, P.; Okazaki, T. Comparative Study of Prediction Models for High School Student Performance in Mathematics. *IEIE Trans. Smart Process. Comput.* **2019**, *8*, 394–404. [CrossRef]

54. Luis-Rico, I.; Escolar-Llamazares, M.C.; De la Torre-Cruz, T.; Jimenez, A.; Herrero, A.; Palmero-Camara, C.; Jimenez-Eguizabal, A. Entrepreneurial Interest and Entrepreneurial Competence Among Spanish Youth: An Analysis with Artificial Neural Networks. *Sustainability* **2020**, *12*, 1351. [CrossRef]

55. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media: Sebastopol, CA, USA, 2019.

56. Naser, J.A. Neural Networks; A Brief Introduction. In Proceedings of the American Power Conference, Chicago, IL, USA, 29 April–1 May 1991; Volume 53.

57. Kursa, M.B.; Rudnicki, W.R. Feature selection with the Boruta package. *J. Stat. Softw.* **2010**, *36*, 1–13. [CrossRef]

58. Abbas, J.; Aman, J.; Nurunnabi, M.; Bano, S. The impact of social media on learning behavior for sustainable education: Evidence of students from selected universities in Pakistan. *Sustainability* **2019**, *11*, 1683. [CrossRef]

59. Sohlberg, M.M.; Mateer, C.A. Effectiveness of an attention-training program. *J. Clin. Exp. Neuropsychol.* **1987**, *9*, 117–130. [CrossRef]