 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at *Sustainability* **2018**, *11*, 105; doi:10.3390/su11010105

Article

Prediction of confusion attempting algebra homework in an intelligent tutoring system through machine learning techniques for educational sustainable development

Syed Muhammad Raza Abidi [1,*], Mushtaq Hussain [2], Yonglin Xu [3] and Wu Zhang [4]

[1] Shanghai University; razaabdi@live.com
[2] Shanghai University; mushi_pcr@yahoo.com
[3] Shanghai University; wffzxyl@gmail.com
[4] Shanghai University; wzhang@shu.edu.cn

* Correspondence: razaabdi@live.com; Tel.: +86-185-016-16009

Abstract: Incorporating substantial sustainable development issues into teaching and learning is the ultimate task of Education for Sustainable Development (ESD). The purpose of our study is to identify the confused students who have failed to master the skill(s) given by the tutors as a homework using Intelligent Tutoring System (ITS). We have focused ASSISTments, an ITS in this study and scrutinized the skill-builder data using machine learning techniques and methods. We used seven candidate models that include: Naïve Bayes (NB), Generalized Linear Model (GLM), Logistic Regression (LR), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), and Gradient Boosted Trees (XGBoost). We trained, validated and tested learning algorithms, performed stratified cross-validation and measured the performance of the models through various performance metrics i.e., ROC (Receiver Operating Characteristic), Accuracy, Precision, Recall, F-Measure, Sensitivity & Specificity. We found GLM, DT & RF are high accuracies achieving classifiers. However, other perceptions such as detection of unexplored features that might be related to the forecasting of outputs can also boost the accuracy of the prediction model. Through machine learning methods, we identified the group of students which were confused attempting the homework exercise and can help students foster their knowledge, and talent to play a vital role in environmental development.

Keywords: education for sustainable development, confusion, intelligent tutoring system (ITS), ASSISTments, machine learning, computer-based homework, algebra mathematics technology education, sustainable development.

---

1. Introduction

The Intelligent tutoring systems (ITSs) and MOOCs both have corresponding educational approaches. ITS differs with MOOCs in many aspects, for instance; ITS facilitates instant feedback, scaffolding practice in solving the pedagogical problems. Students while learning through web-based interfaces have opportunities to take the hint, watch the related topic videos, and guidance to practice the concept and attempt the right answer. MOOC, on the other hand, provides much interactive learning through learning management system (LMS) in which various forms of instruction like video lectures, moderated discussion boards (MDBs), and forums available for learning with peer feedback. [1].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1
Peer-reviewed version available at *Sustainability* **2018**, *11*, 105; doi:10.3390/su11010105

Conversely, students' enhancement in learning is due to the most powerful intelligent tutoring systems (ITSs) and these have got commercial effectiveness. Although the availability of intelligent tutors is limited, their substantial cost subsidizes construction of a content [2,3].

Many researchers generated numerous ITS for various students. For instance: Intelligent Tutoring Tools (ITTs) project of Byzantium [4], AutoTutor, Atlas and Why2 [5], Andes [6], ASSISTments [7].

In this study, mainly our focus will be on ASSISTments (ITS), as we collected data from Skill-builder data 2009-2010 [8].

1.1. ASSISTments (ITS)

The ASSISTments (https://www.assistments.org) is an ITS that provides a free web-based platform that facilitates school students and teachers to assess their students learning [9]. They also performed an experiment and found that students who received technology assisted feedback had higher scores than the students without receiving hint or scaffolding.

ASSISTments comprises of problems related to mathematics with hints, immediate feedback, and answers. These problems are grouped into problem sets and teachers allocate to their students [10].

Essentially, ASSISTments came into being when authors were coaching in middle school especially math pupils. The objective was pursuing the students who have basic knowledge and to recognize what they want to do. In previous years each student masters in the skill(s) of practicing strictly by hand, not by computer. ASSISTments provides teachers to keep an eye on the comprehension and trail the skill(s) they have grasped [7].

It is an environment produced by Worcester Polytechnic Institute, USA, specially designed for middle schools. More than 50,000 students registered and use ASSISTments for doing homework and master the skill(s) without taking response the next day in the class because in-time instant feedback facility of ASSISTments for particular problem associated with the skill(s) from the various subjects that include (Mathematics, Science, English and Statistics etc.). Similarly, teachers give two options for students' assignments: 1. Teachers compile questions with remedies, hints, and problem-related videos to master the skill(s) chosen by the student for a particular subject. 2. Teachers can also use built-in problem set and broadcast the homework for the students.

In the word ASSISTments, assist associated with the teachers while assessment related to students. Moreover, a student attempts the wrong answer to given problem, the prompt feedback pings student to rectify by taking a hint or to try another option, while teachers can log the results in accurate instantaneous output and later they can use this evidence to make a strategy for the next lesson [7].

Two types of educational contents are available in ASSISTments. One is related to the mathematics textbook homework or the problems that teachers write themselves for their students and the second type is specifically designed for skill learning practice and mastery called "skill builders". In ASSISTments, current skill-builder data consists of more than 300 topics in related to middle school mathematics. The purpose of skill builder is to master the skill by practicing the problem assigned by the teacher defined standard or principle [9].

Skill-builder assures that student must have an expert in the topic or skill before going to move forward to grasp other tasks. It is one of the best kind of content in ASSISTments to test and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

acknowledge what student has learned and it is mandatory for each student to correct 3 questions in a row until gets the preliminary ability on the chosen topic area [7].

Many peer-reviewed journals published articles in the context of prediction. ASSISTments was broadly and widely used as data mining exploration. e.g., [11], by means of Bayesian networks, e.g., [12], or using the platform to make classifiers of students' demonstrative state. e.g., [13].

The ultimate goal of our research is to identify and predict the confusion of students while solving the mathematics homework using mastery skill-builder learning after attempting the teacher defined criterion (e.g., correcting three consecutive answers on similar math problems). Even though by taking instant feedback from ITS, students get confused. So, by using machine learning classifiers we would be able to categorize students, who are confused and who do not. As machine learning algorithms work on the principle of statistics and for this objective, we used statistical programming language R and RapidMiner 8.1 to analyze and predict the confused students in an ITS Skill-builder session and showed the results.

To accomplish this task following are our research question which we will consider to bridge the gap.

- Can we categorize which machine learning algorithms are the best fit to classify mastery skill learning confusion among the students using skill-builder in an intelligent tutoring system on the basis of chosen skills?

Further, the structure of the paper is as follows: In Section 2, we present a short overview of related works and research on the particular subject. In Section 3, we define the related methods used and proposed predictive methods. In Section 4, we interpret the results and discuss prediction performance. In Section 5, we leave the reader with concluding thoughts, shortcomings & future recommendations.

2. Related Works

Material regarding the course of mathematics in ASSISTments comprises of difficulties with solutions and in-time suggestions. Furthermore, substantial assistance readily available over the internet to resolve the issue that students solve online. Another type of material was precisely designed for mastery focused skill training named as "Skill-builders" as discussed above. At the moment, ASSISTments covers more than 300 matters related to mathematics for middle school and the capability given to teachers to allocate skill-builders to pupils to rehearsal those problems that emphasis on the desired skill(s) until unless they get the pre-defined standards for accuracy [9].

Many types of research corroborate the significance of ITS while using in a class of students in school [14].

Still, very limited researches discovered the importance of ITS used as homework [15].

Hence, it was very inspiring when [16] communicated auspicious outcomes when ANDES and ITS used in this manner.

ASSISTments used by massive students of the middle and high school at present for their evening homework and due to instant advice regarding homework, students feel comfortable and tutors become able to monitor the reports specifying students achievements [15].

So far, for the evening homework, multilayered tutoring systems are not suitable as on the other hand, technology-supported instructions which disseminate same questions with a fast response about the problem is more appropriate [17].

According to Singh, homework on the web-based tool using the tutoring system is more authentic and robust in learning and mastering the skill(s) of a student compared to previous old-fashioned paper based traditional style. Also, this research focuses on instantaneous response with the tutoring system against the feedback received by students from the tutor the next working day, which is obviously time-consuming and reduces the learning ability as a whole. He further investigated that 8th-grade math students who were indulged in both scenarios observed that they expanded pointedly with an effect size of 0.40 by using technology-assisted homework [14].

As per Fyfe, around the globe, ITS and technology-assisted homework achieved fame and pervasiveness in schools [18].

Conferring to Ma, Adesope and colleagues, the handiness of personalized and well before advice is the solid foundation of intelligent tutoring systems [19].

The objective of the study by Fyfe, reveals an investigational assessment of algebra class of middle school students who have variable preceding knowledge affected by the technology-based response [18].

Generally, numerous researches support that by using the in-time response from ITSs, as usual, has constructive properties on learning outputs as opposed to no response from ITSs [20,21].

Lee & colleagues, Baker & colleagues and Gupta & Rose, they all classify that confusion and both its roots and penalties can easily be recognized through the performance and actions of students [22–24].

Confusion affects students to halt, reproduce and start problem-solving to rectify own confusion. The only way to cope with confusion is that every student must have bottomless knowledge of complicated matters, as fought with confusion is, of course, an intellectual action [25,26].

On the other hand, if a healthy learning atmosphere offers an adequate platform and timely assistance to the students and they themselves efficiently normalize their confusion could achieve positive outcomes [25–27].

Moreover, many scholars used different methods like "classification or knowledge engineering" to detect the disturbance changes in students, particularly confusion [28].

Likewise, Conati & MacLaren established a detector built on logged data and grouping of survey questions to forecast self-described student disturbance. Although this model was healthier to recognize attentive and inquisitive students but ineffective at classifying confused students [29].

Baker and colleagues conducted substantial research especial focus on computer software designed for education e.g. ITSs to automatically identify confusion through affect detection and they collected this information through semantic actions of students and labeled the existing PSLC DataShop log files. In this research, they defined confusion as the slower patterns of students' actions while attempting the pre-defined teacher criterion before the starting of mastery skill-builder assignment or homework. Authors focused the preliminary step and observed the percentage of clip actions [24].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at *Sustainability* **2018**, *11*, 105; doi:10.3390/su11010105

3. Methods

This section clarifies and illustrates the effectiveness of raw data to classification via machine learning methods. **Figure 1** depicts the visualization of raw data to classification workflow:

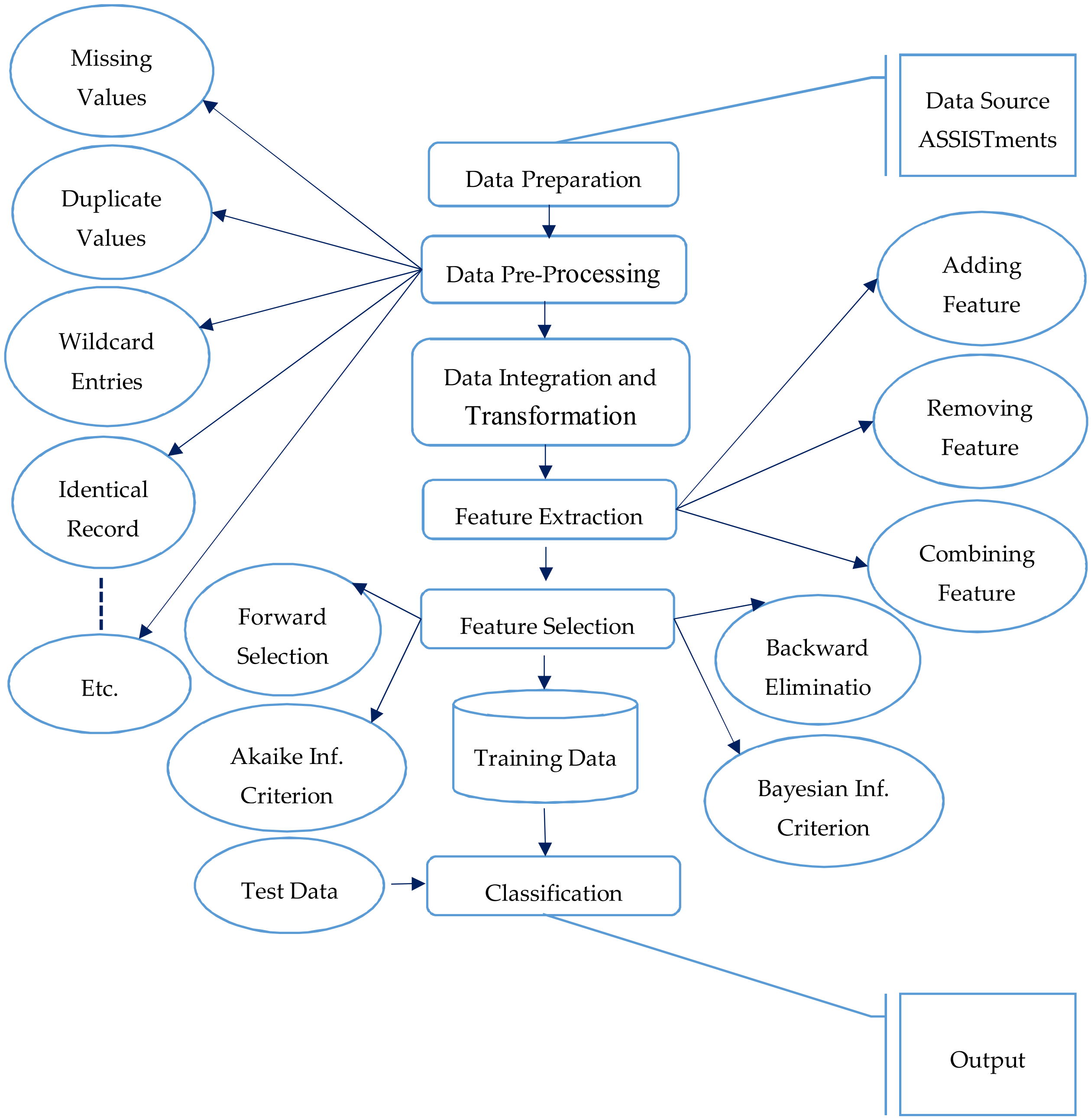


Figure 1. A pictorial view of raw data to classification workflow

3.1. Preparation of Data

In this study, we used dataset collected from ASSISTments, Skill-builder data 2009-2010 [8].

Skill-builder problem sets have the following features:

- Questions are based on one specific skill; a question can have multiple skill tagging's.
- Students must answer three questions correctly in a row to complete the assignment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

- If a student uses the tutoring ("Hint" or "Break this Problem into Steps"), the question will be marked incorrect.
- Students will know immediately if they answered the question correctly.
- If a student is unable to figure out the problem on his or her own, the last hint will give an answer.
- Currently, this feature is only available for math problem sets.

In this whole data set, various features are available related to the mastery skill-builder learning. There were almost 72 schools participated with 93 mastery skills in algebra mathematics and about 28 features. We targeted the school ID-73 because it has maximum records availability amongst the other school IDs and selected 10 mastery skills i.e., (Absolute Value, Addition and Subtraction Positive Decimals, Box and Whisker, Circle Graph, Multiplication Fractions, Ordering Fractions, Percent Of, Subtraction Whole Numbers, Venn Diagram, and Write Linear Equation from Graph) as the maximum students selected these chosen skills and after removing duplicate values, we got total 166 distinct student IDs remained.

3.1.1. Measurements and Covariates

We have selected the predictors (original, attempt_count, ms_first_response, correct, hint_total, overlap_time, opportunity) from the list of features available in the dataset and measured ROC, accuracy, precision, recall, F-measure, sensitivity & specificity as performance indicators used by machine learning algorithms.

3.1.2. Discretization of Predicted Variable

After precise selection of predictors, we are interested in learning what features apprise the status of the confused/not confused student. So, determining this, we used a feature extraction technique to select the predicted variable. We chose and combined three variables with concern to form new feature called “student state”, and on the basis of that, we categorized the status of the confused/not confused student, ‘1’ designates confused and ‘0’ for not confused.

3.1.3. Experimental Manipulations or Interventions

We have used cross-validation technique to divide our dataset into a standard (80% – 20%) of training and test datasets respectively with stratified sampling, as our response variable is dichotomous.

3.1.4. Statistical Analysis

For statistical analysis, we used statistical programming language R (https://cran.r-project.org/) and used RStudio (https://www.rstudio.com/) to perform basic descriptive and regression analysis. Also checked the correlation between explanatory and response variables and identified which variables are significant, bring information to the model and which variables do not.

3.2. Pre-processing of Data

Data extracted either from databases, log files or Microsoft Excel files need to be cleaned. Although, it was in good shape the cleaning of data before moving ahead is an utmost part of the pre-processing. Data could be noisy, missing or uneven. Machine learning algorithms itself perform

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1
Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

pre-processing of data up-to some extent but these algorithms will be more robust if we manually accomplish this step.

3.3. Integration and Transformation of Data

For better statistical analysis and classification, data must be integrated and transformed. For this objective, **Table 1** illustrates ten mastery skills and each skill has four attributes (stated above in *3.1. Preparation of Data*) for each student.

Table 1. Mastery skills and corresponding attributes

Skill Name	Attribute-1	Attribute-2	Attribute-3	Attribute-4
Mastery skill-1	ATT-1	ATT-2	ATT-3	ATT-4
Mastery skill-2	ATT-1	ATT-2	ATT-3	ATT-4
Mastery skill-3	ATT-1	ATT-2	ATT-3	ATT-4
Mastery skill-4	ATT-1	ATT-2	ATT-3	ATT-4
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Mastery skill-10	ATT-1	ATT-2	ATT-3	ATT-4

3.4. Feature Extraction

Feature extraction is a procedure for creating new attributes amongst the existing features. Figure 1, shows a snapshot of the feature extraction step. In classification, it is considered to be an important step as the performance measure of learning process dependent on significant explanatory variables. In many real-world cases, we cumulatively extract features, alter if needed, combine them and produce one variable and same procedure might be adopted for the selection of response variable. In this study, we selected 10 mastery skills and also nominated the associated explanatory variables and club them to form 40 explanatory variables for each student in each mastery skill.

3.5. Feature Selection

As revealed in Figure 1 above, there are many criterions available for feature selection. For Instance: Backward Elimination, Forward Selection, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), DIC (Deviance Information Criterion), Bayes factor and Mallow's Cp etc. We used Backward Elimination using the adjusted R^2 method with the cutoff P-value of 0.05 to construct our model because it is a common way [30]. We started with the full model and eliminate one variable at a time until the parsimonious model is reached [31].

3.6. Training of Model

Beforehand, the prediction of confusion amongst the students attempting algebra mastery skill homework in ITS, we essentially trained machine learning algorithms to curtail the difference between the actual and predicted values. For this objective, we split the data into (80% – 20%) ratio with stratified sampling, as our response variable is nominal.

In this study, we designated seven learning classifiers: Naïve Bayes (NB), Generalized Linear Model (GLM), Logistic Regression (LR), Deep Learning (DL), Decision Tree (DT), Random Forest

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

(RF) & Gradient Boosted Trees (XGBoost) to classify dichotomous response variable with the value "Confuse" or "Not Confuse".

3.6.1. Description of Machine Learning Algorithms for Constituent Models

As mentioned above, seven learning algorithms were used to figure out candidate models. We have cautiously and broadly reviewed prior research that implements machine learning methods and techniques for the classification problems. Our research has employed learning algorithms for the evaluation of performance that includes accuracy and precision. The comprehensive description of the algorithms used is presented in the following:

- Naïve Bayes (NB)

It is fast and efficient probabilistic classifier with an extensive past record of research. Due to its robustness, precision, and competence, this method is usually referenced.

One of the most important aspects of NB is, it has the property of scalability, meaning that adding more predictors (input) variables do not cause drastic changes in performance. Moreover, NB has verified over many years in the diversity of domains of academic research [32–34].

- Generalized Linear Model (GLM)

It is eventually assessed by the famous statistical principle of maximum likelihood estimation (MLE) and it helps to minimize the supposition that difference between observed and the predicted value of response variable which is called residual and is Gaussian distributed [35].

GLMs are actually the enhancement of old-style linear models and the series of instructions inside these models turn to data by using the MLE technique. These models give tremendous, really fast and parallel computation with a small number of explanatory variables with non-zero constants [36].

- Logistic Regression (LR)

In the study of Peng, Lee & Ingersoll, they revealed when the response variable has two branches, LR prevails logical method. They also highlighted the usefulness of logistic model that was exposed to be braced by the statistical significance test of each explanatory variable, the conclusive and expressive goodness of fit, and probabilities related to prediction [37].

It is broadly used statistical technique for the classification of binary output. When predicting the output of the response variable of nominal in nature, usually the logistic regression algorithm used. It uses the statistical logistic function to classify items between "0 and 1", or it can also handle the nominal variables which have a limited number of categories. For example; Range from (0 - 9), or (A - Z), etc. LR essentially establishes the relationship between a categorical response variable and commonly a continuous explanatory variable(s) by adapting the response variable to likelihood (probability) scores [36].

- Deep Learning (DL)

As per Li & colleague, DL works on the basis of a neural network that takes information that offers information about other data as input and produces the outcome by using many layers [38].

On the other hand, the old-style neural network can only consider a single hidden layer, DL initiates the process by using extensive hidden layers which comprise of nodes to produce the outcome. DL has the ability to tune & select the model at an optimal level by itself and it also achieves

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

mining of features instinctively without involvement and interaction of individuals or humans which spectacularly saves a plenty of determination and time [39].

- Decision Tree (DT)

It depicts a tree like building, where it has nodes (internal and leaf). It is made by training data which consists of data rows or records. Each record is formed by a set of features and outcome label. Features contain either distinct (integer) or continuous (non-integer) values. Primarily, data whose outcome label is un-identified, DTs are employed to classify them and according to the feature values of the data record, route from root to leaf must be trailed [40].

- Random Forest (RF)

Random forest by Breiman associates multiple tree input variables in a group. New occurrences being classified are broken down the trees, and each tree states a classification [41].

The “forest” then chooses which label to allocate to this new occurrence built on the cumulative number of polls specified by the set of trees [42].

RF generates a number of arbitrary trees on various subsets of a data and the subsequent model builds on polling of these trees. Because of this variance, it is less likely to overtraining. In a classification task, the minimal leaf size is 2 and 5 for regression [36].

- Gradient Boosted Trees (XGBoost)

It is correlated to Gradient Boosting Machine (GBM) which is another boosting algorithm. It produces good accuracy due to the competences of parallel computing and the effective linear model solver. It also creates decision trees which are individual understandable models [43].

Due to the groups of DTs, XGBoost is more authoritative and compound model. It trains the model by repeating the process, again and again, refining a single tree model. Instances are assigned new weights according to their earlier prediction. Ultimately, the final model is a weighted sum of all established models. It is a group of either classification or regression tree models and both obtain predictive outcomes through steadily enhanced approximation [36].

3.7. Testing (evaluation) of the Model

Model evaluation is an important part of the implementation of machine learning techniques. When a machine becomes train on the known data then we evaluate the model on unseen data to verify that the model is good enough, learned and classified correctly.

3.7.1. Performance Metrics

In this study, we adopted the most common and widely used performance metrics of [44]. They have used the ROC Area under the curve (AUC) to calculate the performance of prediction models.

- ROC Curve or AUC

It contains several thresholds and each threshold produces a 2 x 2 contingency table, **Table 2** which comprises 4 inside central entries. ROC curve demonstrates the association between true positive rate and false positive rate [45].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

Table 2. Contingency table (confusion matrix)

		Predicted Class (Y)	
		Y	N
Actual Condition (X)	Y	TP	FN
	N	FP	TN

(Y: Yes, N: No, TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative)

We also used accuracy, precision, recall, F-Measure, sensitivity, and specificity performance metrics.

- Accuracy

Symbolizes a predicted value approves with a real value [46].

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + False\ Positive + True\ Negative + False\ Negative)}$$

- Precision

Recognizes the likelihood of a positive test outcome. High values specify that the likelihood of the test dataset being perfectly classified [47].

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

- Recall

Assesses the number of true positives of the real class forecasted by the models. High recall shows improved classifier performance [47,48].

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

- F-Measure

Indicates which algorithm performed well. We can take swift choices about the algorithms accuracies through this performance metric [49].

$$F1\ Score = \frac{2PR}{(P + R)}$$

Where, P: Precision and R: Recall.

- Sensitivity

Is the capability that accurately classifies with the "Confuse" in this study of students' mastery skill using ITS [44]. Also called True Positive Rate (TPR) [50].

$$Sensitivity = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

- Specificity

Is the skill that precisely classifies with the "Not Confuse" students [44]. Also known as True Negative Rate [50].

$$Specificity = \frac{True\ Negative}{(True\ Negative + False\ Positive)}$$

Detailed results for this present study are publicized in Section 4 (Results).

3.8. Classification

Plenty of machine learning/data mining tools are available. We have used RapidMiner 8.1 for our investigation & testing. RapidMiner studio is well equipped with data mining/machine learning tasks with state-of-the-art sufficient collection of machine learning algorithms along with data access, pre-processing, blending, cleansing, modeling, visualization & validation operators which give high-tech advanced platform to perform machine learning/data mining task in the most efficient and well-organized manner [51,52]. As we are executing supervised learning succeeding some linear & non-linear classifiers are used for classification, applied and verified.

Following are the list of classifiers we have selected for our experiment are shown in **Table 3**.

Table 3. List of machine learning classifiers used in this study

Machine Learning Classifiers Used
• Naïve Bayes (NB)
• Generalized Linear Model (GLM)
• Logistic Regression (LR)
• Deep Learning (DL)
• Decision Tree (DT)
• Random Forest (RF)
• Gradient Boosted Trees (XGBoost)

3.9. Statistical Analysis and Parameters

In order to discover the significance of explanatory variables for the prediction of students' confusion in algebra mastery skill in ITS, it is imperative to explore the predictors (explanatory variables) and its impact on response variable statistically. Although machine learning algorithms intrinsically perform statistical test and analysis of variables, nevertheless, it is always good practice to check manually before applying any machine learning method and technique.

Figure 2 is the weights (ranks) of the attributes which show the universal significance of each attribute for the value of the target attribute, independent of the modeling algorithm.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

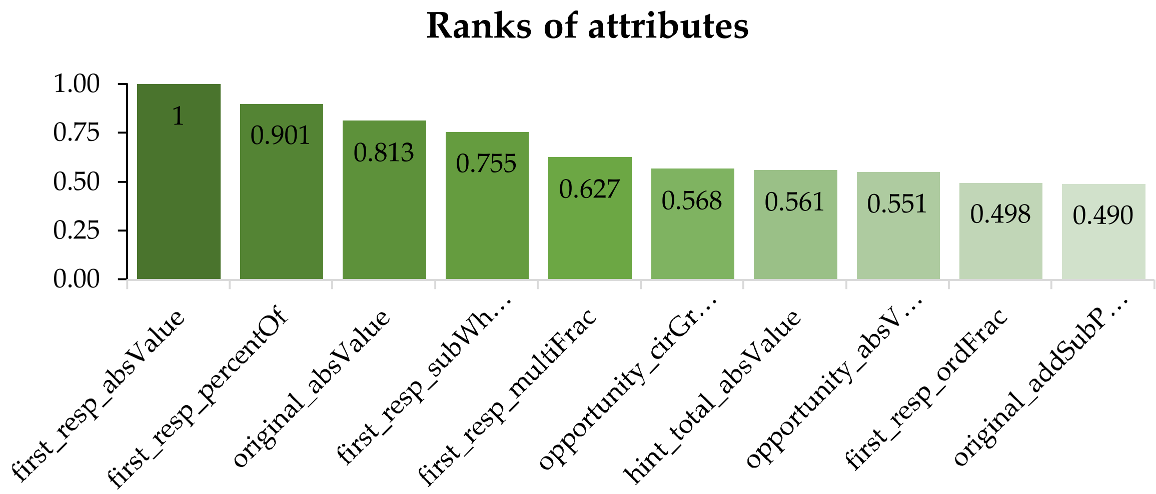


Figure 2. Representation of ranks of attributes

We have used statistical programming language R with the standard cut-off level of probability value (P-value 0.05). **Figure 3** shows a graphical display of a correlation matrix P-values using R-Language package (ggplot2).

In this statistical summary of correlation, we found 9 predictor variables are most significant, i.e., their values are ($P < 0.05$) related to the dichotomous response variable. Correlation is used to measure how strong a linear association between two numeric variables and there are many types of correlation coefficient exist. i.e., (Pearson, Kendall, Spearman). We used Pearson's correlation coefficient as it is commonly used in linear regression. It is denoted by (r or R) and its value always in the range from -1 to +1, where +1 specifies strong positive correlation and -1 the strong negative correlation.

 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at *Sustainability* **2018**, *11*, 105; doi:10.3390/su11010105

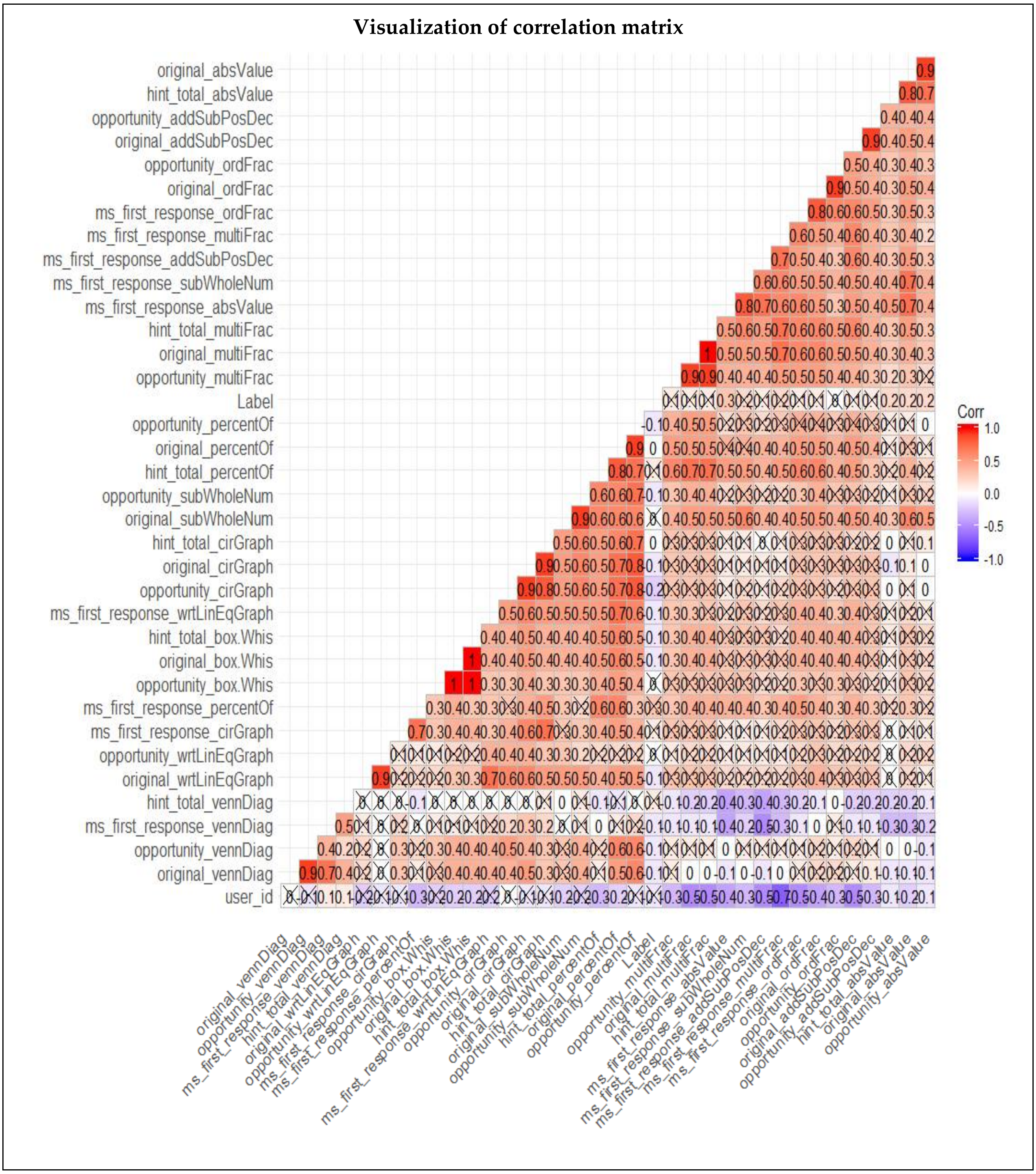


Figure 3. A graphical display of a correlation matrix P-values

Statistically, after using backward elimination technique, we end up with the final model which validates the significance of 9 explanatory variables which are shown in **Table 4**, which portrays descriptive statistics, **Table 5** shows regression analysis including predictor's coefficients, standard errors, P-values etc., and **Table 6** reveals regression summary.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at *Sustainability* **2018**, *11*, 105; doi:10.3390/su11010105

Table 4. Descriptive statistics

Min	1Q	Median	3Q	Max
-0.94696	-0.04588	0.05360	0.13200	0.44114

Table 5. Regression analysis of predictors (explanatory variables)

Coefficients	Estimate	Std. Error	t value	pr (> I t I)	
(Intercept)	0.5588550	0.0475093	11.763	< 2e-16	***
ms_first_response_absValue	0.0247446	0.0043088	5.743	4.73e-08	***
original_addSubPosDec	0.0104006	0.0058002	1.793	0.074887	.
original_box.whis	-0.1641421	0.0685545	-2.394	0.017838	*
opportunity_box.whis	0.0317376	0.0196939	1.612	0.109082	
original_cirGraph	0.0530879	0.0123338	4.304	2.95e-05	***
opportunity_cirGraph	-0.0041942	0.0010725	-3.911	0.000137	***
hint_total_vennDiag	0.0273546	0.0059493	4.598	8.77e-06	***
original_wrtLinEqGraph	-0.0602548	0.0218751	-2.754	0.006577	**
opportunity_wrtLinEqGraph	0.0005436	0.0002309	2.355	0.019787	*

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6. Regression statistics summary

Residual standard error	Degrees of freedom	Multiple R-squared	Adjusted R-squared	F-statistic	P-value
0.2936	156	0.3213	0.2821	8.205	6.123e-10

Furthermore, **Figure** 4, reveals the efficient feature of R language which graphically shows what maximum adjusted R^2 value could be achieved through given explanatory variables.

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

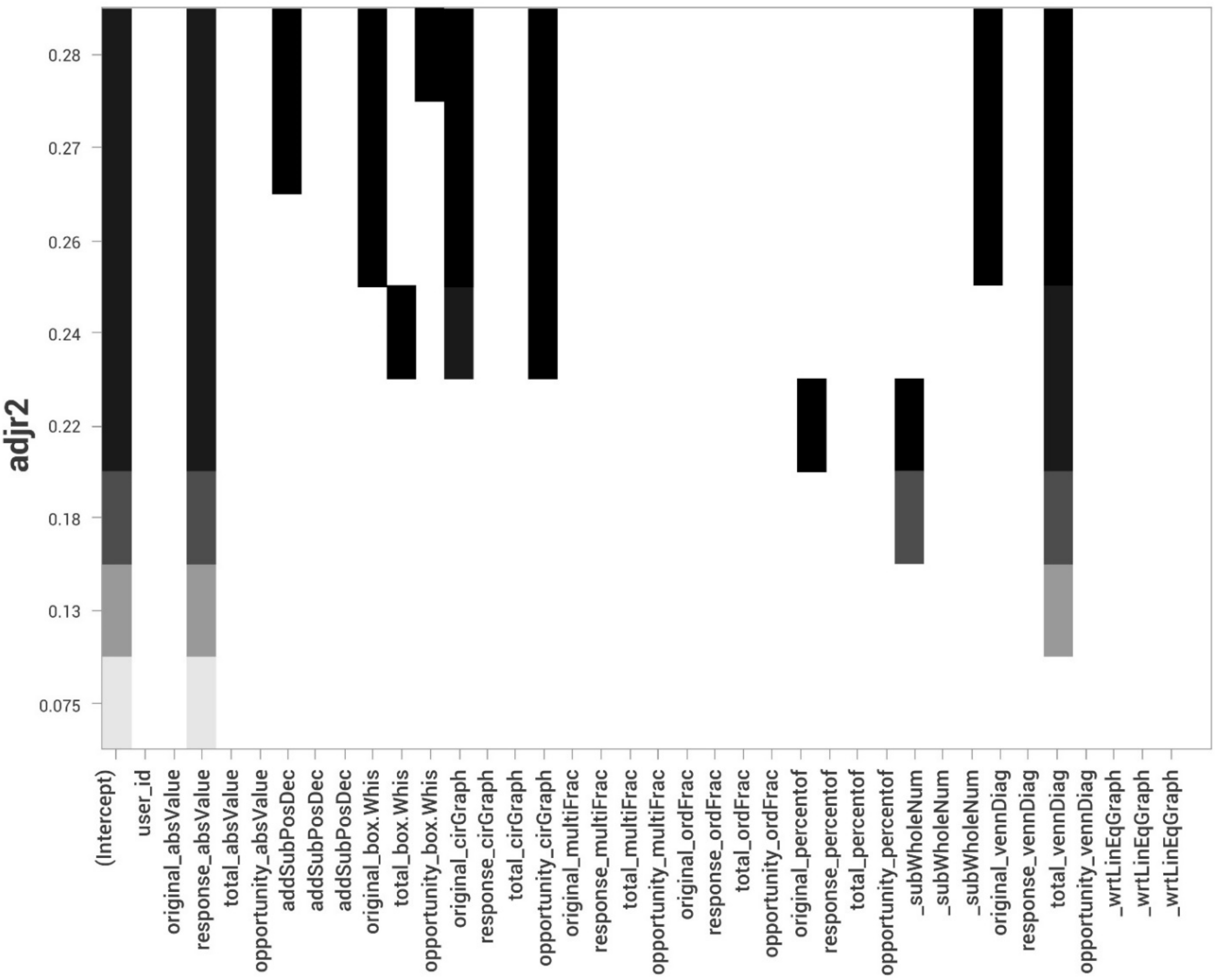


Figure 4. Maximum adjusted R^2 representation

4. Results

As discussed in Section 3, Methods, ROC is Receiver Operating Characteristic and also recognized as ROC AUC or just simply ROC curve. It demonstrates the relationship between the true positive rate (TPR) and false positive rate (FPR). It also determines cooperation between sensitivity and specificity as both are contradictory i.e. when the sensitivity rises specificity declines. The accuracy can be monitored if the curve is nearer to top left corner and could be considered finest results but if curve comes closer to the diagonal angle (45°) the result would not be accurate. Moreover, ROC AUC value is > 0.9, portrays excellent results, value is in between 0.8 – 0.9 considers good, between 0.7 – 0.8 reflects fair, and < 0.6 illustrates poor [53].

Graphical representation of ROC AUC is shown in **Figure 5** and **Figure 6** depicts the AUC values graphs for seven machine learning algorithms which correctly predicted the confusion amongst the students attempting algebra mastery skill in ITS.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

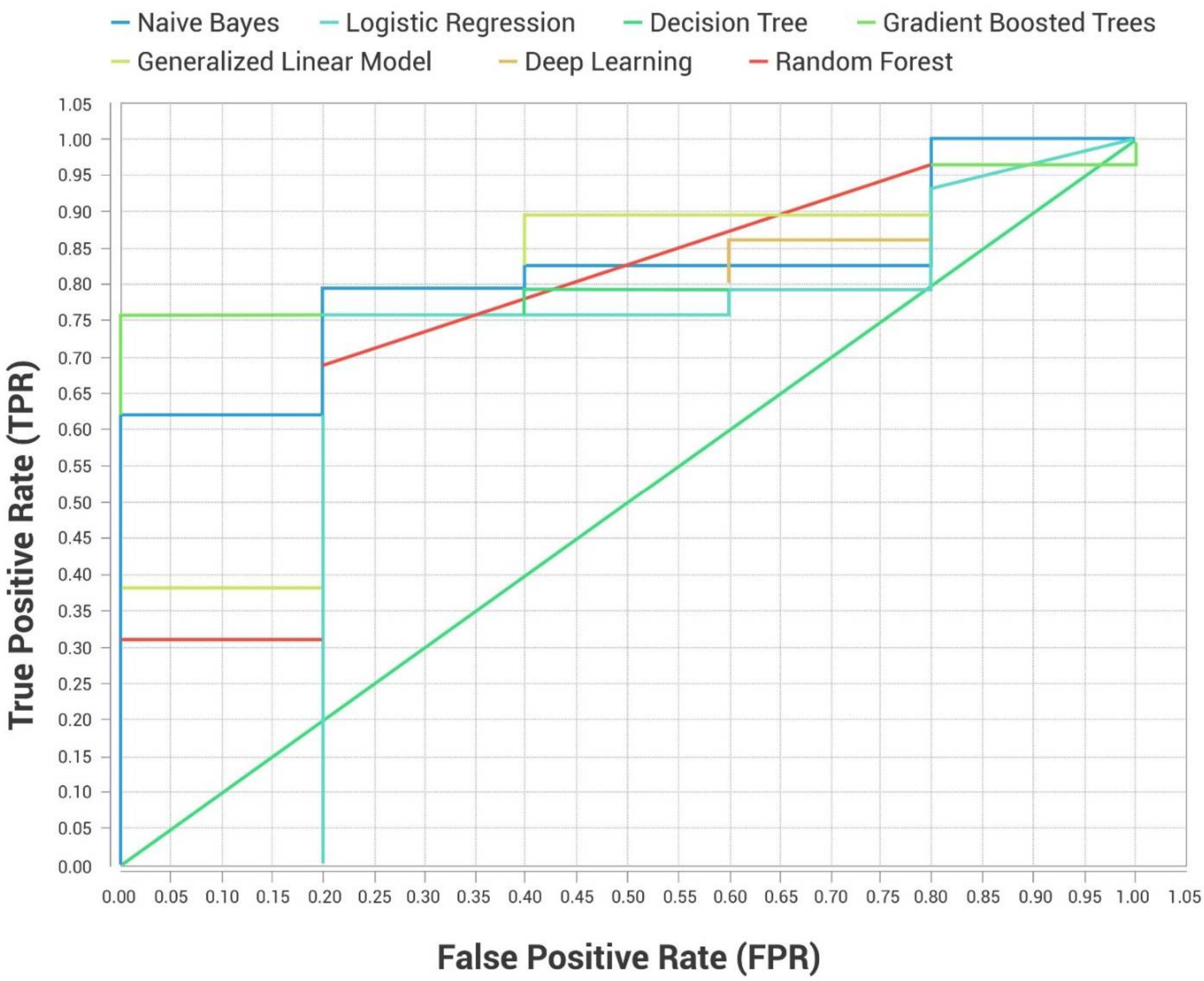


Figure 5. A symbolic view of ROC AUC of 7 selected machine learning algorithms

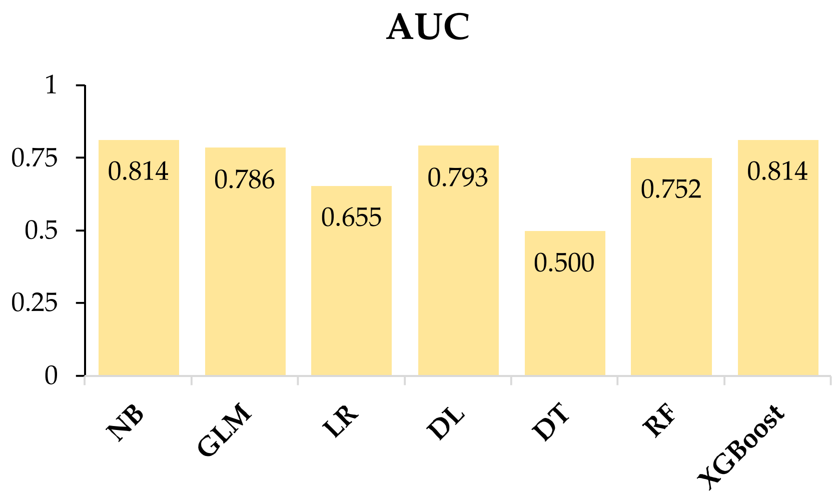


Figure 6. Graphical representation of ROC AUC of nominated machine learning models

We have constructed seven candidate models built on various machine learning methods. The performance achieved by each classifier is shown in **Figure** 7, which reveals the accuracy performance metric of each model by repetitive sampling validation technique, in which it randomly replicates division of training and test data.

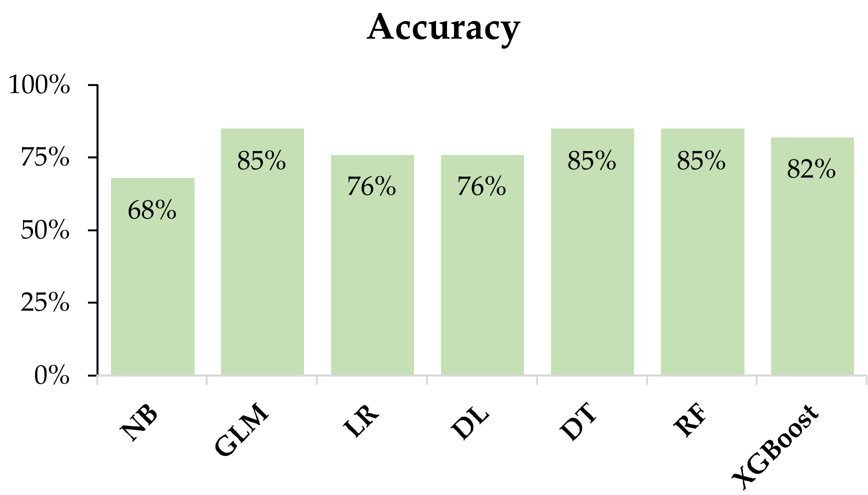
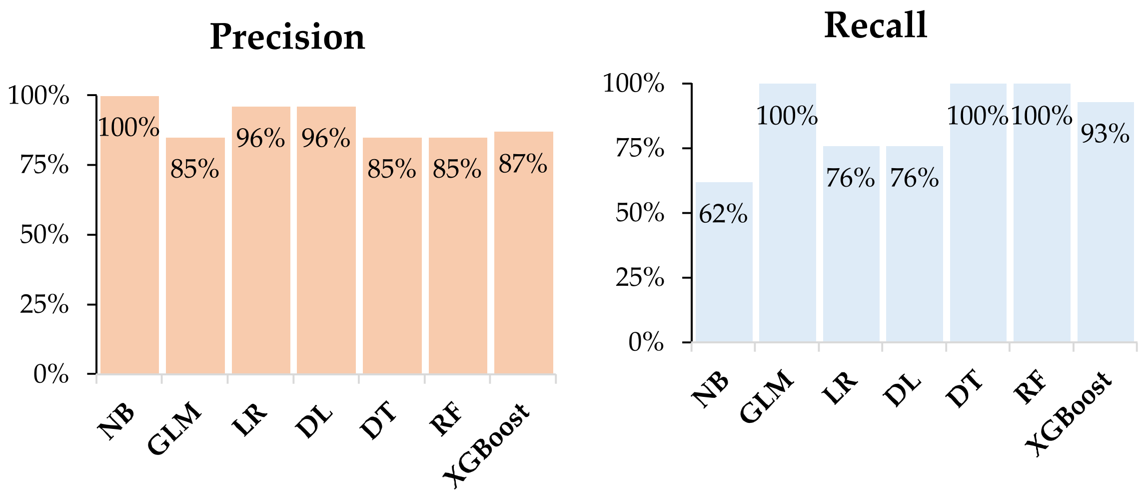


Figure 7. Candidate models' summary with respect to accuracy

These results illustrate the ratio of time we are able to acceptably predict the cases. We attained maximum accuracy with GLM, DT, and RF, 85.3% each respectively. We also employed other classifiers, i.e., NB: 67.6%, LR: 76.5%, DL: 76.5%, and XGBoost: 82.4%.

We have checked the other performance metrics which we discussed in Section 3. **Figure 8** displays the performance of seven machine learning algorithms regarding precision, recall, F-measure, sensitivity, and specificity.



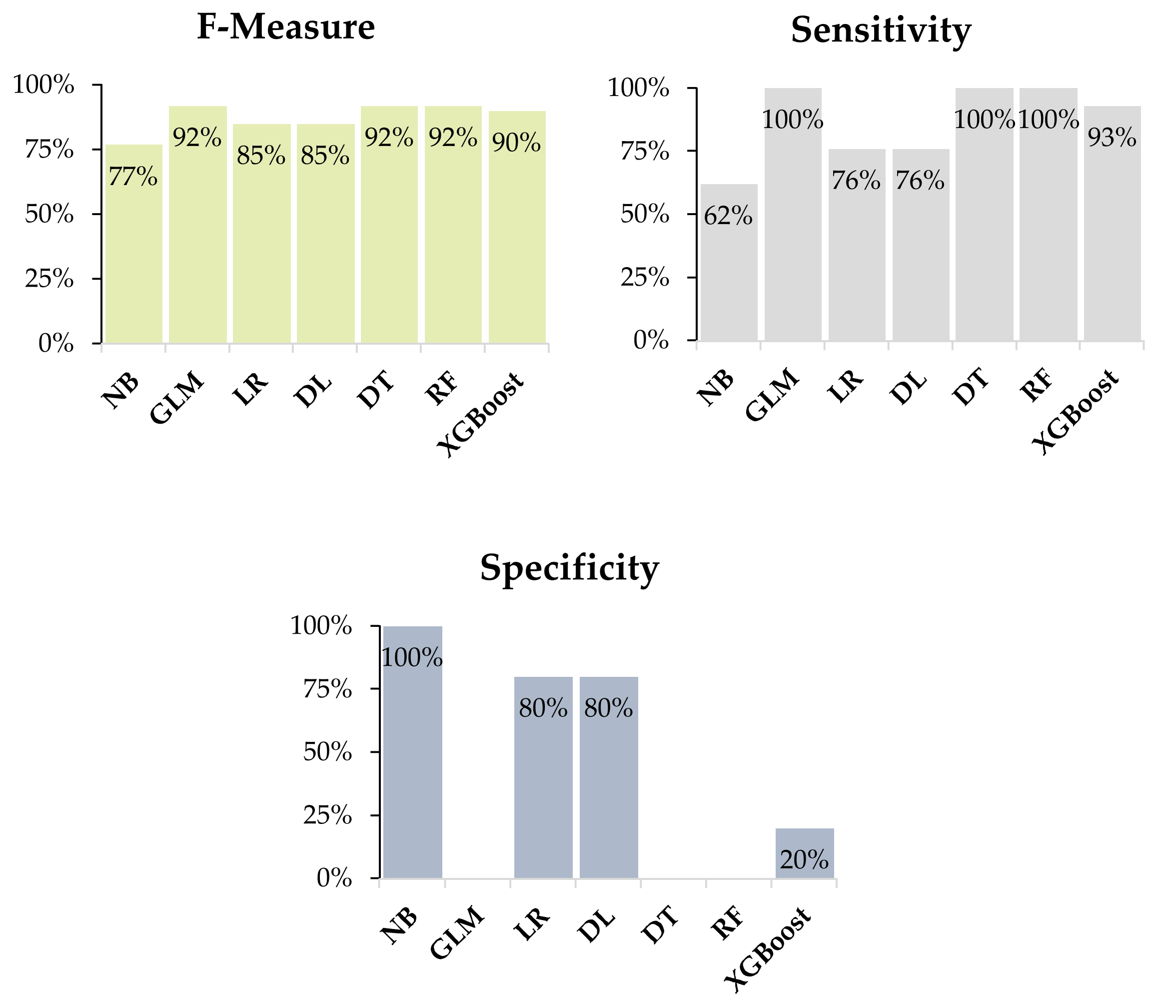


Figure 8. Performance of machine learning algorithms relating to performance measures

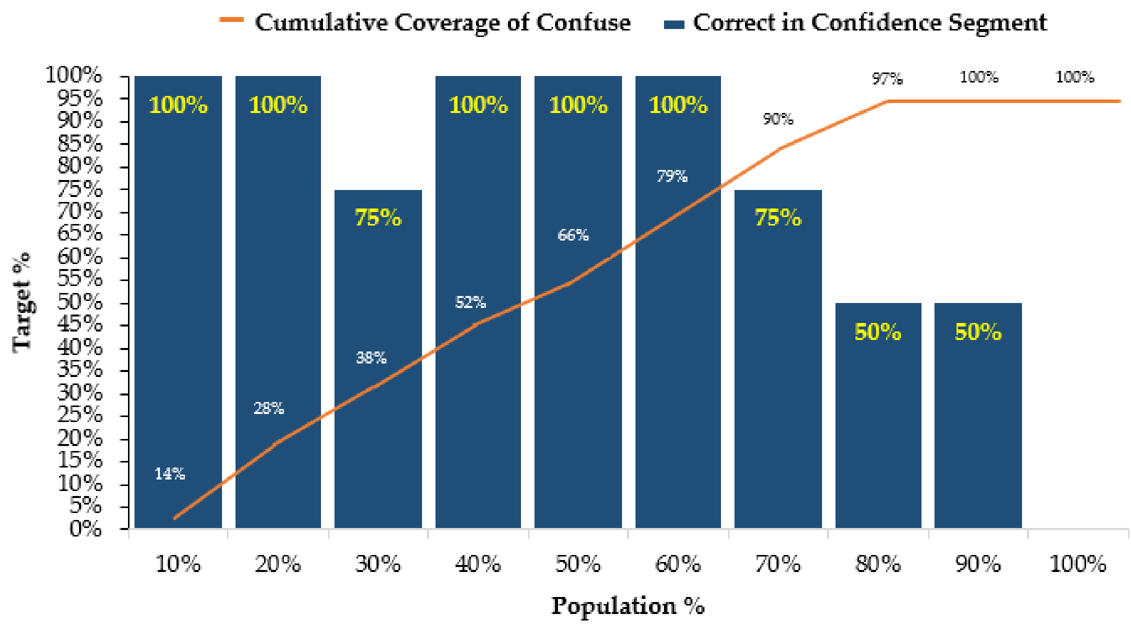
Table 7. Complete detail of learning algorithms along with runtime

Model	Accuracy	Precision	Recall	F-Measure	Sensitivity	Specificity	Runtime
NB	67.6%	100.0%	62.1%	76.6%	62.1%	100.0%	87 ms
GLM	85.3%	85.3%	100.0%	92.1%	100.0%	0.0%	5 s
LR	76.5%	95.7%	75.9%	84.6%	75.9%	80.0%	772 ms
DL	76.5%	95.7%	75.9%	84.6%	75.9%	80.0%	1 s
DT	85.3%	85.3%	100.0%	92.1%	100.0%	0.0%	527 ms
RF	85.3%	85.3%	100.0%	92.1%	100.0%	0.0%	3 s
XGBoost	82.4%	87.1%	93.1%	90.0%	93.1%	20.0%	1 min 33 s

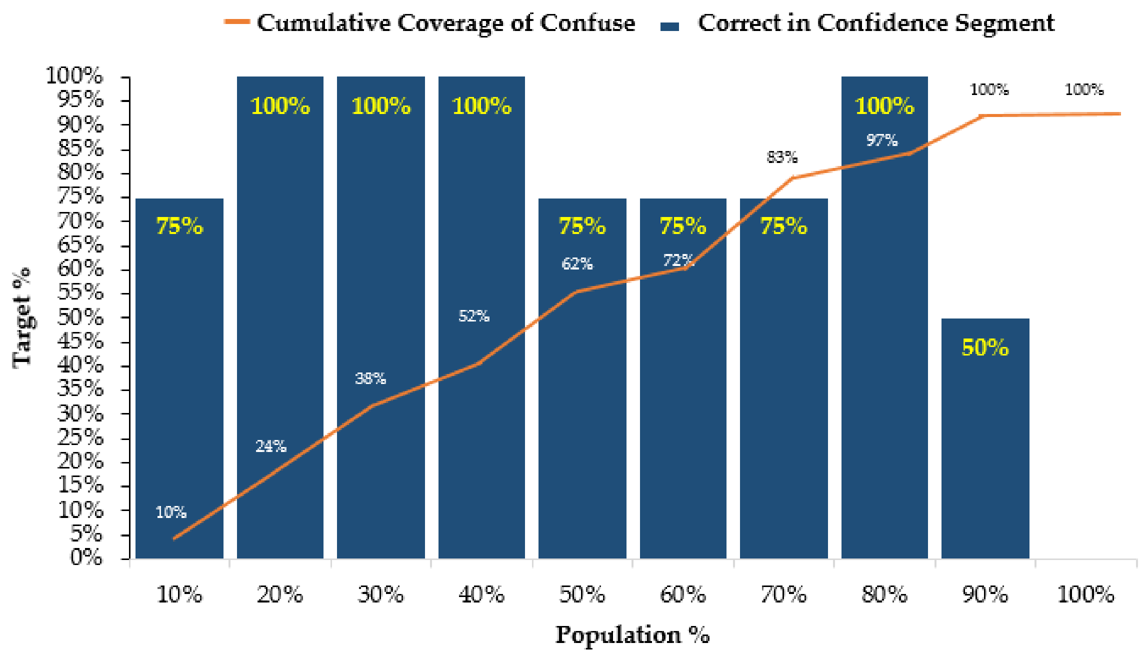
ms: millisecond, s: second, min: minute

Figure 9 displays the lift charts of high achieving accuracy machine learning models. A lift chart is a graphical illustration of the enhancement that a model delivers when related against a random guess [54]. It shows the efficiency of the model by measuring the ratio between the outcome obtained "with the model and without a model" [36].

Generalized linear model (GLM) Lift chart



Decision tree (DT) Lift chart



Random forest (RF) Lift chart

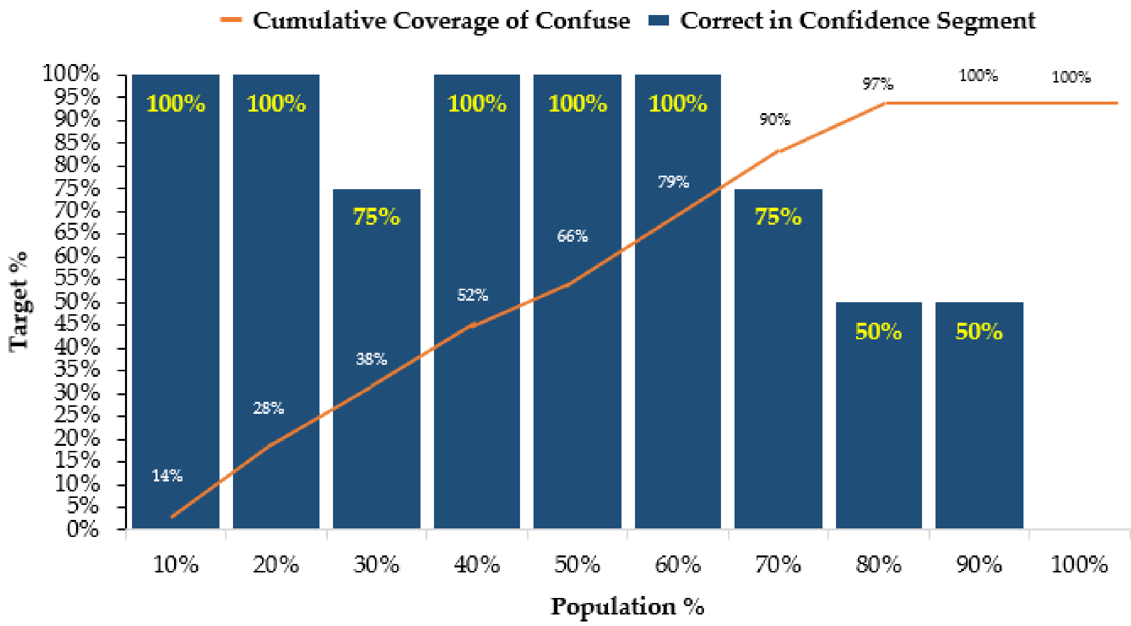


Figure 9. High achieving accuracy models' lift chart

5. Discussion and Conclusions

This research investigates models for the prediction of confused students attempting homework using skill-builder in ITS. Analyzing confusion is a task of classification and machine learning has plenty of robust classification algorithms. So, in this study, we used machine learning methods for the experiment. Performing techniques of data mining on ITS is a tough task because of the many features related are in various extents with heaps of noisy data and missing fields. We then extracted explanatory variables (input features) and targeted (output) response variable from ITS. Then, we applied machine learning models NB, GLM, LR, DL, DT, RF, and XGBoost respectively. The results demonstrate that GLM, DT, and RF models attained a high accuracy of 85.3% in predicting the students' confusion in the algebra mastery skill in ITS.

Such a result can provide assistance to tutors of the schools in the next day of the class and identifying group of students which were confused attempting the homework exercise in mastery skill builder and will also highlight which skill(s) need(s) more attention to practice. Furthermore, tutors can also govern learning behaviors and performances of each student during various mastery skill(s) and could be able to focus only problematic skill(s) in the next day of the class which will save a lot of time and effort of both tutors and students.

Our study has many decent inferences both educationally and practically. Firstly, to the best of our information and facts, our research, amongst the previous studies for predicting confusion by using machine learning methods for educational sustainable development, is one of the rare studies that have focused. ITS contributes sustainable development in education as the development focuses the necessities of the present-day without compromising the future needs. The objective of sustainable development is to stable our environmental, economic, and social needs [55]. Sustainable development in education is an interdisciplinary learning approach covers the combined environmental, social, and economic aspects of the formal and informal curriculum. This educational approach can assist students to develop their aptitudes, knowledge, and experience to show a significant role in the ecological development and become liable members of a society. Participation and sharing teaching and learning techniques and methods are also required to encourage and empower learners to change and alter their performances and take corrective actions for sustainable development. Critical thinking, visualizing future, and decisions making are the skills and abilities that ESD promotes [56].

5.1. Shortcomings

The shortcoming in this study is, we have used a limited number of variables as there are more attributes available which can be used for further investigation and could be statistically stronger. Another shortcoming is, by doing rigorous optimization techniques like changing criterion, pruning, selecting a threshold of machine learning models (algorithms) could achieve better results.

5.2. Future Recommendations

In future work, we will design to apply some strategies to augment our model further. First, a more decent optimization parameter can be used for building the more accurate model. For instance: In DT, we can change the criterion i.e.; gain_ratio, Information_gain, gini_index, accuracy, maximal depth parameter etc., in RF, we can set the same criterion, number of trees and maximal depth etc. and in XGBoost, we can alter maximal depth, min rows, min split improvement, number of bins etc. to optimize performance. Secondly, other kinds of classification methods, techniques can be measured. Though the machine learning techniques used and applied in this study are fairly comprehensive but still, there are various unexplored methods/techniques can be applied to the prediction problem in the domain of students in intelligent tutoring system. Thirdly, other structures and features in the data may enhance the prediction correctness and accuracy can be added. Furthermore, as per the tutors' perspective, we can identify the benefits associated while detecting the confusion in a group of students solving mathematics homework using skill-builder in ITS.

Author Contributions: Conceptualization, Mushtaq Hussain and Yonglin Xu; methodology, Mushtaq Hussain, Syed Muhammad Raza Abidi; software, Syed Muhammad Raza Abidi; validation, Mushtaq Hussain, Yonglin Xu, and Syed Muhammad Raza Abidi; formal analysis, Syed Muhammad Raza Abidi; resources, Syed Muhammad Raza Abidi; writing—original draft preparation, Syed Muhammad Raza Abidi; writing—review and editing, Mushtaq Hussain; visualization, Syed Muhammad Raza Abidi, Yonglin Xu; supervision, Prof. Dr. Wu Zhang; funding acquisition, Prof. Dr. Wu Zhang.

Funding: This research was funded by "NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA, grant number 61572434, 91630206, and 61303097" and "THE NATIONAL KEY R&D PROGRAM OF CHINA, grant number 2017YFB0701501".

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

References

[1] V. Aleven *et al.*, "Integrating MOOCs and Intelligent Tutoring Systems: edX, GIFT, and CTAT," *Proc. 5th Annu. Gen. Intell. Framew. Tutoring Users Symp.*, p. 11, 2017.

[2] K. R. Koedinger, J. R. Anderson, W. H. Hadley, M. A. Mark, and Others, "Intelligent tutoring goes to school in the big city," *Int. J. Artif. Intell. Educ.*, vol. 8, pp. 30–43, 1997.

[3] V. Aleven, J. Sewall, B. M. McLaren, and K. R. Koedinger, "Rapid Authoring of Intelligent Tutors for Real-World and Experimental Use," *Sixth IEEE Int. Conf. Adv. Learn. Technol.*, no.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at *Sustainability* **2018**, *11*, 105; doi:10.3390/su11010105

Icalt, pp. 1–5, 2006.

[4] Kinshuk, “Computer Aided Learning for Entry Level Accountancy Students,” 1996.

[5] R. Freedman, “Atlas: A plan manager for mixed-initiative, multimodal dialogue,” *AAAI-99 Work. Mix. Intell.*, pp. 1–8, 1999.

[6] A. Gertner, C. Conati, and K. VanLehn, “Procedural help in Andes: Generating hints using a Bayesian network student model,” *Aaai/Iaai*, no. Pearl 1988, pp. 106–111, 1998.

[7] N. T. Heffernan and C. L. Heffernan, “The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching,” *Int. J. Artif. Intell. Educ.*, vol. 24, no. 4, pp. 470–497, 2014.

[8] N. Heffernan, “ASSISTmentsData,” 2012. [Online]. Available: https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data/skill-builder-data-2009-2010. [Accessed: 02-Feb-2018].

[9] J. Roschelle, M. Feng, R. F. Murphy, and C. A. Mason, “Online Mathematics Homework Increases Student Achievement,” *AERA Open*, vol. 2, no. 4, p. 233285841667396, 2016.

[10] J. Roschelle, R. Murphy, M. Feng, S. R. I. International, C. Mason, and J. Fairman, “Rigor and Relevance in an Efficacy Study of an Online Mathematics Homework Intervention,” 2014.

[11] M. Feng, N. Heffernan, and K. Koedinger, “Addressing the assessment challenge with an online system that tutors as it assesses,” *User Model. User-adapt. Interact.*, vol. 19, no. 3, pp. 243–266, 2009.

[12] Z. Pardos and N. Heffernan, “Tutor Modeling vs. Student Modeling,” 2012.

[13] M. O. Z. San Pedro, R. S. J. D. Baker, S. M. Gowda, and N. T. Heffernan, “Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7926 LNAI, pp. 41–50, 2013.

[14] R. Singh *et al.*, “Feedback during web-based homework: The role of hints,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6738 LNAI, pp. 328–336, 2011.

[15] K. Kelly, N. Heffernan, C. Heffernan, S. Goldman, J. Pellegrino, and D. Soffer-goldstein, “WEB-BASED HOMEWORK,” vol. 3, no. 2011, pp. 417–424, 2014.

[16] K. Vanlehn, C. Lynch, and K. Schulze, “The Andes physics tutoring system: Lessons learned,” *Int. J. …*, vol. 15, no. 3, pp. 1–51, 2005.

[17] K. Kelly, N. Heffernan, C. Heffernan, S. Goldman, J. Pellegrino, and D. Soffer Goldstein, “Estimating the effect of web-based homework,” *Lect. Notes Comput. Sci. (including Subser. Lect.*

Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7926 LNAI, pp. 824–827, 2013.

[18] E. R. Fyfe, "Providing feedback on computer-based algebra homework in middle-school classrooms," *Comput. Human Behav.*, vol. 63, pp. 568–574, 2016.

[19] W. Ma, O. Adesope, J. C. Nesbit, and Q. Liu, "Intelligent tutoring systems and learning outcomes : A meta-analysis," vol. 106, no. 4, pp. 901–918, 2014.

[20] J. Hattie and H. Timperley, "The Power of feedback. Review of Educational Research," *Rev. Educ. Res.*, vol. 77, no. 1, pp. 81–112, 2007.

[21] L. Alfieri, P. J. Brooks, N. J. Aldrich, and H. R. Tenenbaum, "Does Discovery-Based Instruction Enhance Learning?," *J. Educ. Psychol.*, vol. 103, no. 1, pp. 1–18, 2011.

[22] N. K. Gupta and C. P. Rose, "Understanding Instructional Support Needs of Emerging Internet Users for Web-based Information Seeking," *JEDM - J. Educ. Data Min. (ISSN 2157-2100)*, vol. 2, no. 1, pp. 38–82, 2010.

[23] D. M. C. Lee, M. M. T. Rodrigo, R. S. J. D. Baker, J. O. Sugay, and A. Coronel, "Exploring the relationship between novice programmer confusion and achievement," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6974 LNCS, no. PART 1, pp. 175–184, 2011.

[24] R. S. J. d. Baker *et al.*, "Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra," *Proc. 5th Int. Conf. Educ. Data Min.*, pp. 126–133, 2012.

[25] B. Lehman, S. D'Mello, and A. Graesser, "Interventions to regulate confusion during learning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7315 LNCS, pp. 576–578, 2012.

[26] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser, "Confusion can be beneficial for learning," *Learn. Instr.*, vol. 29, pp. 153–170, 2014.

[27] Z. A. Pardos, R. S. J. D. Baker, M. O. C. Z. San Pedro, S. M. Gowda, and S. M. Gowda, "Affective States and State Tests: Investigating How Affect Throughout the School Year Predicts End of Year Learning Outcomes," *Proc. Third Int. Conf. Learn. Anal. Knowl. - LAK '13*, vol. 1, pp. 117–124, 2013.

[28] R. Pekrun, T. Goetz, W. Titz, and R. P. Perry, "Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research," *Educ. Psychol.*, vol. 37, no. 2, pp. 91–105, 2002.

[29] C. Conati and H. MacLaren, "Empirically building and evaluating a probabilistic model of user affect," *User Model. User-adapt. Interact.*, vol. 19, no. 3, pp. 267–303, 2009.

[30] D. H. Vu, K. M. Muttaqi, and A. P. Agalgaonkar, "A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using

climatic variables," *Appl. Energy*, vol. 140, pp. 385–394, 2015.

[31] M. C. Rundel, "Linear Regression and Modeling," 2018. [Online]. Available: https://www.coursera.org/learn/linear-regression-model. [Accessed: 18-Jan-2018].

[32] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Los," *Mach. Learn.*, vol. 29, no. 1, pp. 103–130, 1997.

[33] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," pp. 4–15, 1998.

[34] V. C. Smith, A. Lange, and D. R. Huston, "Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses," *J. Asynchronous Learn. Netw.*, vol. 16, no. 3, pp. 51–61, 2012.

[35] V. K. Y. Ng and R. A. Cribbie, "The gamma generalized linear model, log transformation, and the robust Yuen-Welch test for analyzing group means with skewed and heteroscedastic data," *Commun. Stat. Simul. Comput.*, vol. 0918, pp. 1–18, 2018.

[36] RapidMiner, "RapidMiner Documentation," 2016. [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/. [Accessed: 29-Apr-2018].

[37] C. Peng, K. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, pp. 3–14, 2002.

[38] W. Li, M. Gao, H. Li, Q. Xiong, J. Wen, and Z. Wu, "Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2016–Octob, pp. 3130–3137, 2016.

[39] W. Xing and D. Du, "Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention," *J. Educ. Comput. Res.*, 2018.

[40] R. R. Kabra and R. S. Bichkar, "Performance prediction of engineering students using decision trees," *Int. J. Comput. Appl.*, vol. 36, no. 11, pp. 8–12, 2011.

[41] L. Breiman, "Random Forests, Machine Learning," vol. 45, no. 1, pp. 5–32, 2001.

[42] E. Aguiar, N. V. Chawla, J. Brockman, G. A. Ambrose, and V. Goodrich, "Engagement vs performance," *Proceedins Fourth Int. Conf. Learn. Anal. Knowl. - LAK '14*, pp. 103–112, 2014.

[43] R. Cobos, A. Wilde, and E. Zaluska, "Predicting attrition from massive open online courses in FutureLearn and edX," *CEUR Workshop Proc.*, vol. 1967, no. March 2017, pp. 74–93, 2017.

[44] R. G. Pontius and K. Si, "The total operating characteristic to measure diagnostic ability for multiple thresholds," *Int. J. Geogr. Inf. Sci.*, vol. 28, no. 3, pp. 570–583, 2014.

[45] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 November 2018 doi:10.20944/preprints201811.0460.v1

Peer-reviewed version available at Sustainability 2018, 11, 105; doi:10.3390/su11010105

student difficulties from learning session data," *Artif. Intell. Rev.*, pp. 1–27, 2018.

[46] T. Devasia, T. P. Vinushree, and V. Hegde, "Prediction of students performance using Educational Data Mining," *2016 Int. Conf. Data Min. Adv. Comput.*, pp. 91–95, 2016.

[47] M. Sweeney, H. Rangwala, J. Lester, and A. Johri, "Next-Term Student Performance Prediction: A Recommender Systems Approach," 2016.

[48] X. Ge, J. Liu, Q. Qi, and Z. Chen, "A new prediction approach based on linear regression for collaborative filtering," *2011 Eighth Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD*, vol. 4, no. 10871091, pp. 2586–2590, 2011.

[49] S. Rovira, E. Puertas, and L. Igual, "Data-driven system to predict academic grades and dropout," *PLoS One*, vol. 12, no. 2, 2017.

[50] Wikipedia, "Sensitivity and specificity," 2018. [Online]. Available: https://en.wikipedia.org/wiki/Sensitivity_and_specificity. [Accessed: 03-May-2018].

[51] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, vol. 2006, pp. 935–940.

[52] K. Godwin, M. Almeda, and M. Petroccia, "Classroom activities and off-task behavior in elementary school children," *Proc. 35th Annu. Meet. Cogn. Sci. Soc.*, no. 2001, pp. 2428–2433, 2013.

[53] C. E. METZ, "Basic principles of ROC analysis," 2018. [Online]. Available: http://gim.unmc.edu/dxtests/ROC1.htm. [Accessed: 04-May-2018].

[54] Microsoft, "Lift Chart (Analysis Services - Data Mining)," 2018. [Online]. Available: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=sql-analysis-services-2017. [Accessed: 07-May-2018].

[55] "Education for sustainable development | Higher Education Academy." [Online]. Available: https://www.heacademy.ac.uk/knowledge-hub/education-sustainable-development-0. [Accessed: 16-Nov-2018].

[56] "The Brundtland Commission." [Online]. Available: https://www.sustainabledevelopment2015.org/AdvocacyToolkit/index.php/earth-summit-history/past-earth-summits/58-the-brundtland-commission.