Random Forest Regression for Predicting Student Performance

Yuyao Jiang
School of Mathematics and Applied Mathematics
Reading Academy, Nanjing University of Information Science and Technology
yuyaojiang991205@gmail.com

Abstract

It is reported that in some countries, the graduation rate of undergraduates is just over half with high dropout rate. It tends to be late and nothing can be changed when students get unsatisfactory final grades. So predicting student performance in advance is necessary. It allows students to adjust themselves and teachers to guide students better. In predicting student performance, machine learning techniques have a wide range of applications. I will propose a supervised machine learning technique which is random forest regression to construct a student performance prediction model. This application of machine learning is implemented in teaching and learning considering student past grades, demographic, social and school related features. My methodology consist of three steps, the first step aims to build a random forest regression model based on all variables and partition the data set, the second step consists in building another model about the most important variables selected using k-fold cross-validation. The third step aims to compare prediction accuracy of two models. Results show that under 7:3 partition of data set, two models have similar characteristics. However, after replacing it with more accurate k-fold cross-validation, the error of the original model is much smaller than that of the simplified one. The original model is more complex, but with higher accuracy.[1]

1 Introduction

Nowadays, student performance is so crucial that it can affect the next stage of students' education like whether they are able to graduate or enter into ideal school. It is also the main concern of their teachers and parents, who hope that students can learn rich knowledge and receive good education in school, so that they are more likely to find good jobs and better adapt to society. Therefore, this research is to implement machine learning in education. With machine learning becoming more and more popular, some effective algorithms tend to be used to analyze data sets and draw some useful conclusions. Machine learning contains two modes: supervised learning and unsupervised learning [1]. Supervised learning has been focused on in this study considering the need for predictive analysis. Student Performance Data Set [2] from UCI Machine Learning Repository has been chosen and it is about the performance of 395 students in mathematics from two Portuguese schools with 33 relevant attributes such as parents' social status, time spent on study, number of absences. Different attributes represent different importance among them, so it is necessary to identify attributes that influence final grades most. The data set is also used to develop a model to predict students' academic performance. Among various supervised learning algorithms, how random forest regression is implemented in the analysis of data set is what this study describes.

[1] The code for my experiments and methods is at https://github.com/J000y/Paper-Code.

2 Related Work

Various machine learning algorithms have been used to predict student performance. There are classification techniques such as Decision Tree, K-Nearest Neighbor, Naïve Bayes algorithm, together with regression algorithms like Linear Regression, Support Vector Machine.

Decision Tree algorithm. Authors in [3] associated decision tree algorithm with the data mining techniques. J48 decision tree algorithm is found to be the best suitable algorithm to construct the model. Cross-validation method and percentage split method were used to evaluate the efficiency of different algorithms.

K-Nearest Neighbor. In [4], both Support Vector Machine algorithm and K-Nearest Neighbor algorithm were applied on the data set and their accuracy were compared. Authors found that Support Vector Machine achieved slightly better results with correlation coefficient of 0.96, while the K-Nearest Neighbor achieved correlation coefficient of 0.95.

Naïve Bayes algorithm. Classification approach which was Naïve Bayesian classifier was used to predict GPA of graduate student. It was simple probabilistic classifier founded on relating Bayes theorem by naïve impartiality assumptions and was trained extremely expeditiously in supervised education location [5]. Students were clustered into collections using K-Means clustering algorithm.

Linear Regression. In [6], authors observed that linear regression based model is well suited as it predicts the future value rather than a class label. The model is a univariate i.e. it takes only one variable but it can be extended as multivariate model by adding more parameters to get more accurate results.

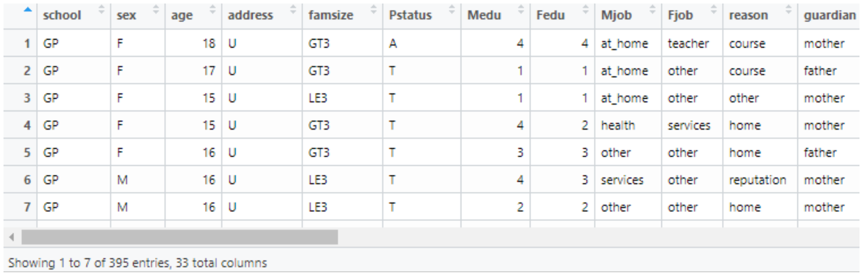
Support Vector Machine. SVM is used as supervised learning model and is a strong classifier which can identify two classes i.e. training and test data for prediction. By utilizing Semantic rules and SVM algorithm, authors analyzed learning of the students and predicted their performance by conducting various tests [7].

In above papers, some models use all attributes as variables, while others use only relatively important ones. In this study, I first extract important variables to build a model, then compare it with the model containing all attributes to observe their differences.

3 Random Forest Regression

3.1 Partition the data set (model 1)

The original data set is complete with no missing values or outliers and is shown partially below.

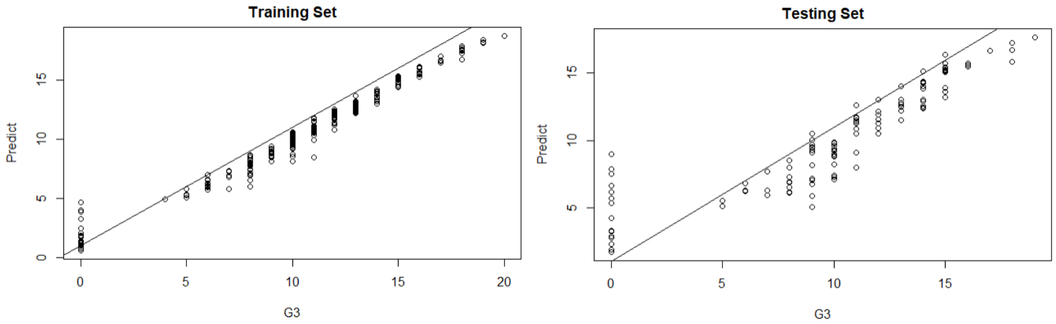


	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother
5	GP	F	16	U	GT3	T	3	3	other	other	home	father
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother

Showing 1 to 7 of 395 entries, 33 total columns

Figure 1: Student Performance Data Set with 395 instances and 33 attributes [2]

Denote the random forest regression model of G3 about 32 other attributes (G3 $\sim$.) as model 1. To facilitate later evaluation of the model's performance, the data set is divided into a training set accounting for 70% and a testing set for 30%. [8] Then package of randomForest is used to analyze with setting a random number seed to 123. Relationship between real grades and predicted ones on training set and testing set is plotted respectively, which illustrates that model 1 has relatively good accuracy.

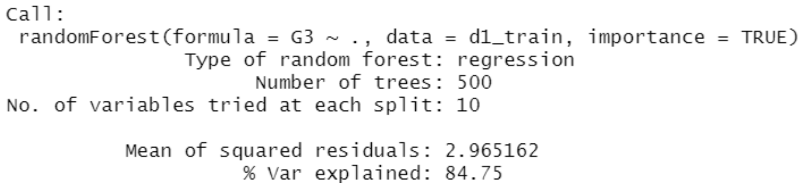


(a) Training Set (model 1)

(b) Testing Set (model 1)

Figure 2: Prediction Accuracy on Training Set and Testing Set (model 1)

"% Var explained" reflects the explanatory variables explained about 84.75% of the total variance and "Mean of squared residuals" is small, indicating that these variables are closely related to students' final grades.

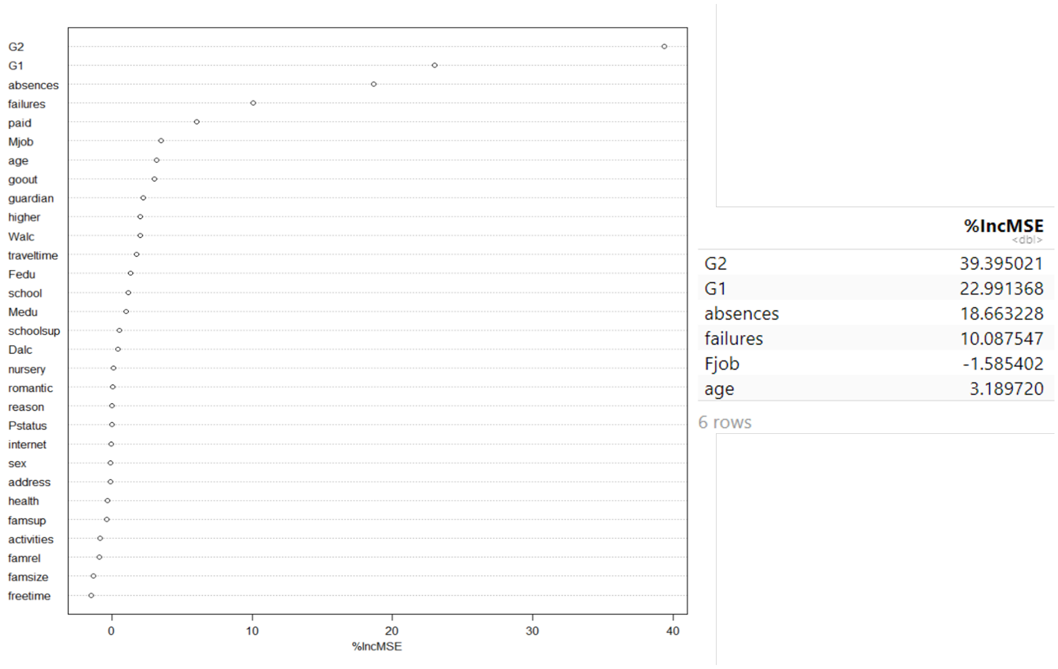


```
Call:
 randomForest(formula = G3 ~ ., data = d1_train, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 10

          Mean of squared residuals: 2.965162
                    % Var explained: 84.75
```

Figure 3: Error Analysis of Training Set (model 1)

3.2 Determine the most important variables (model 2)

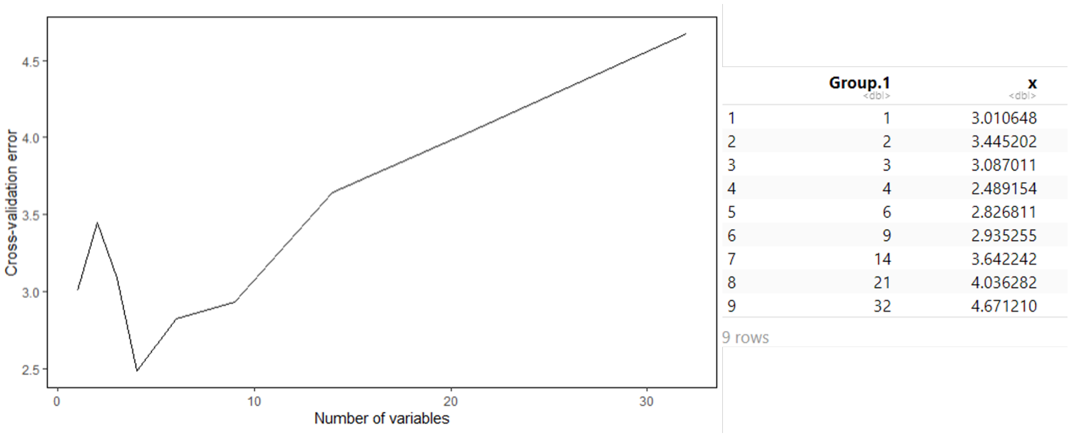


	%IncMSE <dbl>
G2	39.395021
G1	22.991368
absences	18.663228
failures	10.087547
Fjob	-1.585402
age	3.189720

6 rows

Figure 4: Top 30-variable importance

The importance() function is used to calculate the importance of model variables. The larger the value of increase in mean squared error (%IncMSE), the more important the variable. The figure of top 30 important variables from the lowest to the highest is plotted and some specific values are listed above.

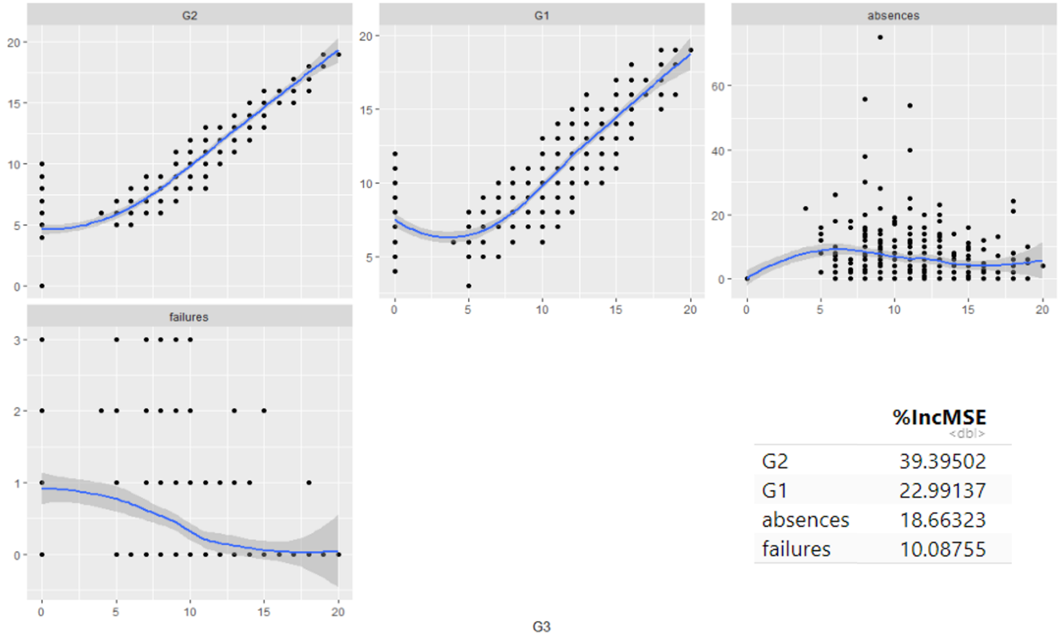


	Group.1 <dbl>	x <dbl>
1	1	3.010648
2	2	3.445202
3	3	3.087011
4	4	2.489154
5	6	2.826811
6	9	2.935255
7	14	3.642242
8	21	4.036282
9	32	4.671210

9 rows

Figure 5: 10-fold Cross-Validation

Then the appropriate number of predictive variables need to be determined. Repeated 10-fold cross-validation in the training set is performed and relationship between the number and cross-validation error is presented using ggplot [9]. From the graph, the error is minimized when the number of variables is about 4.



	%IncMSE <dbl>
G2	39.39502
G1	22.99137
absences	18.66323
failures	10.08755

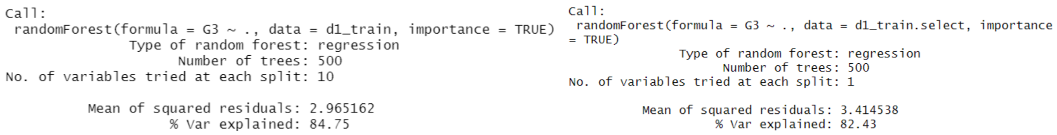
Figure 6: Four most important variables

The first four variables in Figure 4, which are G2, G1, absences and failures, were singled out. The trend of the relationship of them with G3 is obvious, indicating that they are highly correlated with G3. Then build another random forest regression model using four selected predictive variables. Denote it (G3∼G2+G1+absences+failures) as model 2.

4 Evaluation

4.1 Defect of Partition

Compare model 1 and model 2:
"% Var explained" of model 2 is 82.43 compared with 84.75 of model 1. In "Mean of squared residuals", model 2 gets 3.414 and model 1 gets 2.965. Model 2 explains less of the total variance than model 1 and is with larger error, but their differences are little.



```
Call:
 randomForest(formula = G3 ~ ., data = d1_train, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 10

          Mean of squared residuals: 2.965162
                    % Var explained: 84.75
```

(a) Error Analysis of Training Set (model 1)

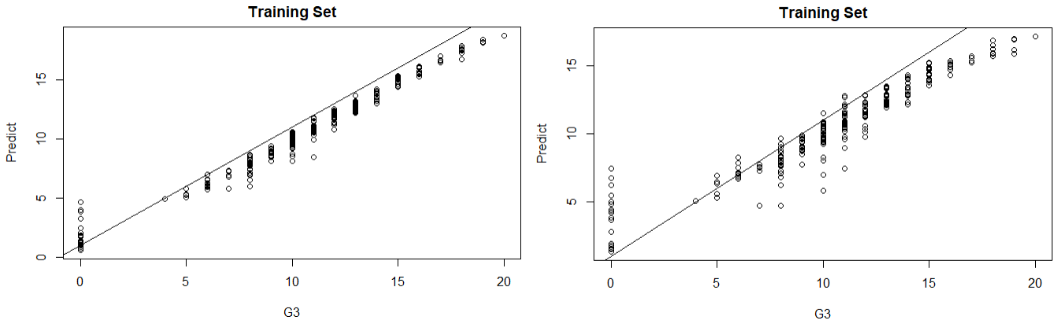
```
Call:
 randomForest(formula = G3 ~ ., data = d1_train.select, importance
= TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 3.414538
                    % Var explained: 82.43
```

(b) Error Analysis of Training Set (model 2)

Figure 7: Comparison about Accuracy [10]

Model 2 performed slightly better than model 1 on training set:

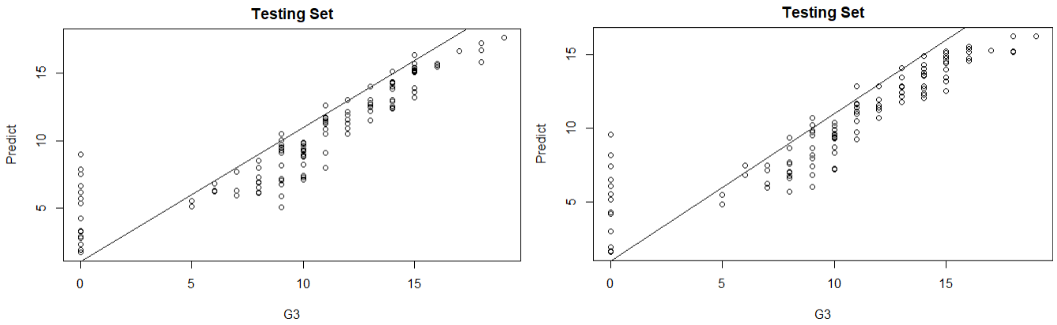


(a) Training Set (model 1)

(b) Training Set (model 2)

Figure 8: Comparison about Prediction Accuracy on Training Set

Two models perform almost the same on the testing set:



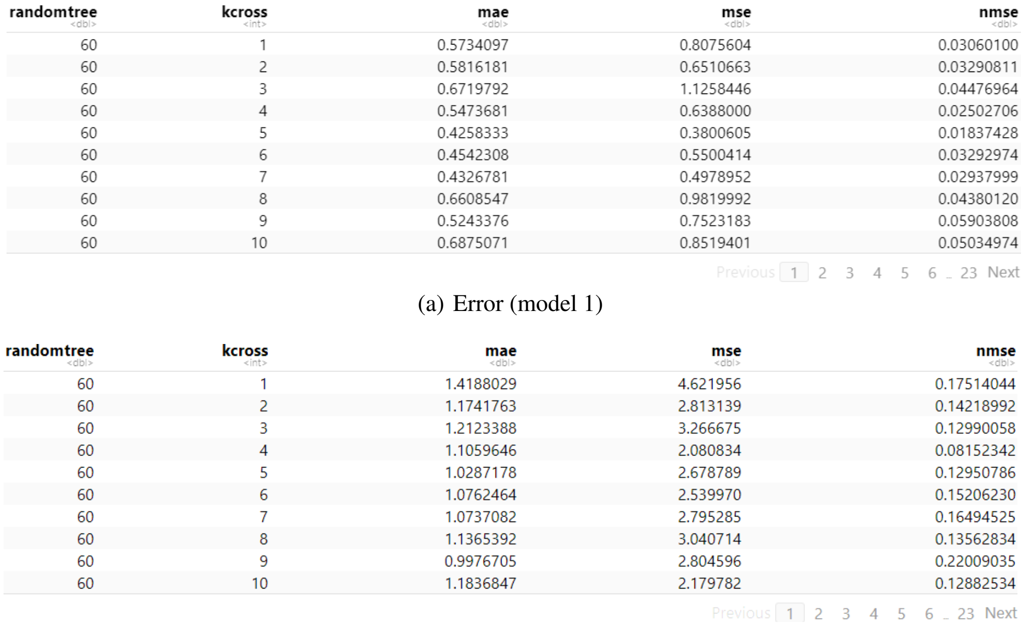
(a) Testing Set (model 1)

(b) Testing Set (model 2)

Figure 9: Comparison about Prediction Accuracy on Testing Set

4.2 Optimization of Partition

In the process of comparison, there is little difference in both results and it is difficult to judge the superiority of two models, indicating that 7:3 partition of data set was not good enough so 10-fold cross-validation is adopted and calculate MAE, MSE, NMSE of two models respectively [11]. The result is that the error between the real value (G3) and the predicted value in model 1 is much less than that in model 2, which means the original model is significantly better than the simplified model.



randomtree <dbl>	kcross <int>	mae <dbl>	mse <dbl>	nmse <dbl>
60	1	0.5734097	0.8075604	0.03060100
60	2	0.5816181	0.6510663	0.03290811
60	3	0.6719792	1.1258446	0.04476964
60	4	0.5473681	0.6388000	0.02502706
60	5	0.4258333	0.3800605	0.01837428
60	6	0.4542308	0.5500414	0.03292974
60	7	0.4326781	0.4978952	0.02937999
60	8	0.6608547	0.9819992	0.04380120
60	9	0.5243376	0.7523183	0.05903808
60	10	0.6875071	0.8519401	0.05034974

(a) Error (model 1)

randomtree <dbl>	kcross <int>	mae <dbl>	mse <dbl>	nmse <dbl>
60	1	1.4188029	4.621956	0.17514044
60	2	1.1741763	2.813139	0.14218992
60	3	1.2123388	3.266675	0.12990058
60	4	1.1059646	2.080834	0.08152342
60	5	1.0287178	2.678789	0.12950786
60	6	1.0762464	2.539970	0.15206230
60	7	1.0737082	2.795285	0.16494525
60	8	1.1365392	3.040714	0.13562834
60	9	0.9976705	2.804596	0.22009035
60	10	1.1836847	2.179782	0.12882534

(b) Error (model 2)

Figure 10: Comparison about Error (partial data)

5 Conclusion

This study used random forest and k-fold cross-validation to analyze the prediction accuracy of different models. The original model is with higher accuracy, but does not seem to be applicable as it contains too many variables. In some cases we may give up a little bit accuracy in order to make the model simpler.

In future work, I would like to use some other criteria to select important variables and make my model can be tested for better analysis and accuracy to be more applicable. Also, I would try some other regression and classification techniques to predict student performance.

References

[1] Bradley C Love. Comparing supervised and unsupervised category learning. *Psychonomic bulletin & review*, 9(4):829–835, 2002.

[2] Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. In *5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, pages 5–12. EUROSIS-ETI, 2008.

[3] Mrinal Pandey and Vivek Kumar Sharma. A decision tree algorithm pertaining to the student performance analysis and prediction. *International Journal of Computer Applications*, 61(13), 2013.

[4] Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, and Sunday O Olatunji. Student performance prediction using support vector machine and k-nearest neighbor. In *2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE)*, pages 1–4. IEEE, 2017.

[5] Fahad Razaque, Nareena Soomro, Shoaib Ahmed Shaikh, Safeeullah Soomro, Javed Ahmed Samo, Natesh Kumar, and Huma Dharejo. Using naïve bayes algorithm to students' bachelor

academic performances analysis. In *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, pages 1–5. IEEE, 2017.

[6] Mahesh Gadhavi and Chirag Patel. Student final grade prediction based on linear regression. *Indian J. Comput. Sci. Eng*, 8(3):274–279, 2017.

[7] Ankita Kadambande, Snehal Thakur, Akshata Mohol, and AM Ingole. Predict student performance by utilizing data mining technique and support vector machine. *International Research Journal of Engineering and Technology*, 4:2818–2821, 2017.

[8] Brijesh Kumar Bhardwaj and Saurabh Pal. Data mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*, 2012.

[9] Boddeti Sravani and Myneni Madhu Bala. Prediction of student performance using linear regression. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–5. IEEE, 2020.

[10] Sadiq Hussain, Zahraa Fadhil Muhsion, Yass Khudheir Salal, Paraskevi Theodorou, Fikriye Kurtoglu, and GC Hazarika. Prediction model on student performance based on internal assessment using deep learning. *iJET*, 14(8):4–22, 2019.

[11] Dorina Kabakchieva. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72, 2013.