Students' Performance Evaluation Using Machine Learning Algorithms

Samah Fakhri Aziz

Department of Computer science
University of AL-Hamdaniya
Mosul, Iraq

Received: 18/5/2020 ; Accepted: 5/7/2020

Abstract :

Predicting student performance is very important to the success of any educational process. Harnessing methods of data mining and machine learning to predict their performance based on data available in schools and student records can explain their behavior, the impact of each factor on the progress of the educational process for students, the relationship of the age stage and follow-up of parents and days of absence. This paper discusses the possibility of harnessing machine learning algorithms to predict student performance and determine the importance of each factor to that performance and Comparing the performance of machine learning algorithms (GBDT-RFDT-Deeplearning) in exploring educational data.

Keywords—Student performance, GBDT, RFDT, Deeplearning, Prediction.

تقييم اداء الطلاب باستخدام خوارزميات التعلم الآلي

سماح فخري عزيز
جامعة الحمدانية /قسم الحاسوب

الملخص :
الخلاصة - إن التنبؤ بأداء الطالب مهم جدا لنجاح أي عملية تعليمية. يمكن لتسخير طرق استخراج البيانات والتعلم الآلي للتنبؤ بأدائها استنادًا إلى البيانات المتاحة في المدارس وسجلات الطلاب أن يشرح سلوكهم وتأثير كل عامل على تقدم العملية التعليمية للطلاب والعلاقة بين المرحلة العمرية والمتابعة حتى الآباء وأيام الغياب. تناقش هذه الورقة إمكانية تسخير خوارزميات التعلم الآلي للتنبؤ بأداء الطالب وتحديد أهمية كل عامل لهذا الأداء ومقارنة أداء خوارزميات التعلم الآلي-GBDT) (RFDT-Deeplearningفي استكشاف البيانات التعليمية.

الكلمات الرئيسية - أداء الطالب ، التنبؤ.

I. INTRODUCTION

The discovery of knowledge from large databases is known as data mining. The purpose of this process is to extract hidden information or repetitive patterns that may be useful in many sciences. Data mining has been used in many fields such as medicine, economics, marketing, and business administration [1]. The educational system data may be useful for extracting information and thus obtaining knowledge that helps in making decisions and improving the educational process [2]. Exploration in student records may be useful in knowing the reasons for the low performance of some students, identifying the causes of certain behaviors, even the causes of some diseases that affect them and the early prediction of the level of students based on their data to help them [3]. Machine learning algorithms are one of the tools used in data mining and

knowledge extraction and have proven their efficiency in many fields. In order to predict student performance, this paper suggests the use of advanced decision tree algorithms (GBDT-RFDT) and comparing their performance with Deep learning.

II. DATA MINING ALGORITHMS

Data mining algorithms help find the factors affecting the performance of students and forecast their performance. The mining algorithms are classified in terms of how they work like Classification, Clustering, Machine Learning, Association rule… etc. In this work, machine learning algorithms are used to predict student performance.

A. *Random Forest (RF)*

Random Forest (RF) is ensemble learning method used for classification, regression that work by creating many decision trees at the time of training. Random forests correct the decision trees habit of over fitting to training set. The training for random forests applies the general technique of bagging, to tree learners [5].

B. *Gradient boosting (GB)*

Gradient boosting (GB) is a method for converting weak learners into strong learners by increasing the weights of the properties that are difficult to classify. After creating the first tree, weights are changed and a sub-tree is created from them. Each tree is a modified version of the original data set, and the results of those trees are taken to achieve more accurate results for the purpose of classification and forecasting[6][7].

C) Deep learning

Deep learning is a group of connected I / O units and each connection has a weight in it. These networks learn by adjusting weights by training them in models that know their results. This network has the ability to handle continuous inputs and outputs value and define patterns. It is suitable for forecasting and classification purposes.

III. SUGGESTED METHODOLOGIES

The algorithms proposed in this research are applied to the data set collected from educational bodies. The number of students to whom the study was applied was 450 instances and 17 attributes. The features taken to predict student performance are as shown in the table below.

TABLE 1: student attribute taken

No	**Attribute**	**Variable**	**Possible Values**
1	Age	numeric	17-28
2	Gender	binary	M-F
3	StageID	nominal	First, Second, Third, Fourth
4	Relation	nominal	Father, Mother
5	Raisedhands	numeric	0-100
6	StudentAbsenceDays	numeric	0-25
7	AcademicachievementofMother	numeric	1-2-3-4
8	Academicachievementof Father	numeric	1-2-3-4
9	Father'sjob	nominal	Governmental, commercial, home
10	Mother'sjob	nominal	Governmental, commercial, home
11	Parents'condition	nominal	Married, Divorced, dead
12	No’sofpastclassfailures	numeric	0-5
13	Traveltime	numeric	1-2-3
14	Studytime	numeric	1-2-3
15	Health status	binary	Good,bad
16	Internettime	numeric	1-2-3
17	Class	nominal	H-M-L (target)

The students' performance levels were divided into three levels (high, medium and low). According to their grades and their levels, they were distributed as shown in Figure (1).

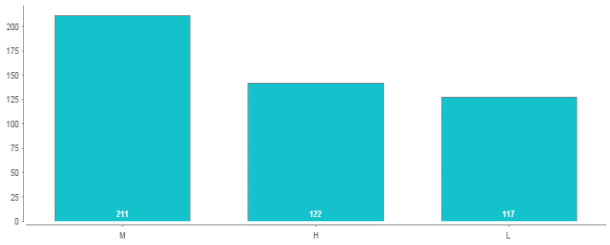

Fig (1): Student performance levels

The data is pre-processed and lost values are checked. Several factors including the performance of the proposed algorithms in terms of accuracy in forecasting, run time of training, forecasting and the impact of each feature on the results were taken into consideration.

The methodology used in this research and its stages is shown below:

1. First, machine learning algorithms gradient boosting decision tree and random forest decision tree and deep neural network were trained on educational dataset and models were obtained.
2. The trained models obtained from the first stage are applied to the test data, as the data set was divided into 80% training group and 20% test group
3. The results of each algorithm are analyzed and compared in terms of accuracy, prediction time, and error rate.

The prediction stages for this research are summarized below

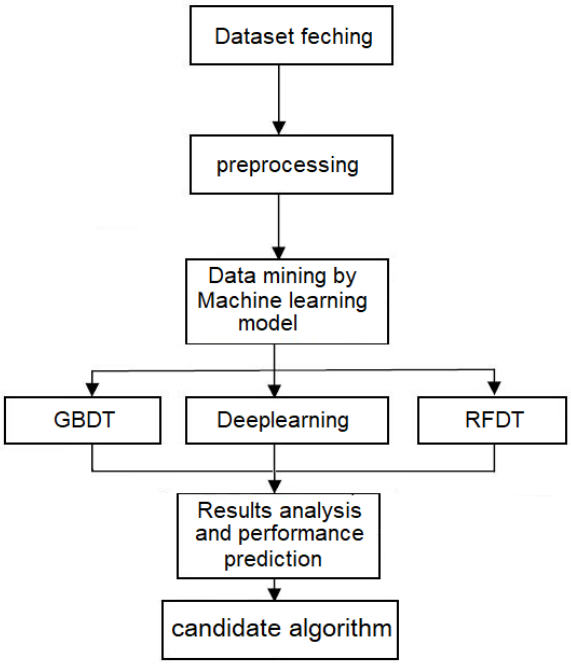

Fig (2): Proposed Frame Work

IV. RESULTS AND DISCUSSIONS

Various classifiers were selected in this paper and a comparative analysis of their performance was performed using the Rapidminer tool. Educational dataset is pre-prepared and later provided for GBDT, RFDT, and Deeplearning. These algorithms were trained and tested and the final results are as shown in the following table:

TABLE 2: prediction accuracy

Model	**Accuracy**	**Standard Deviation**
Deep Learning	0.782	0.0373
Random Forest	0.761	0.035
Gradient Boosted Trees	0.716	0.0454

Table 2 shows the accuracy of the prediction for each algorithm used in this paper.
To show the statistical errors of each model, Confusion Matrix was used and the results were as follows:

TABLE 3: Confusion Matrix of RFDT

	M	**L**	**H**
M	34	2	14
L	4	33	1
H	16	0	32

TABLE 4: Confusion Matrix of GBDT

	M	**L**	**H**
M	39	4	9
L	4	34	1
H	17	0	79

TABLE 5: Confusion Matrix of Deeplearning

	M	**L**	**H**
M	41	7	1
L	6	24	1
H	9	0	21

One important metric for performing machine learning algorithms is a receiver operating characteristic curve, or ROC curve. This scale gives a visual view of the separation ability of classification algorithms. In this paper, each performance of each algorithm is expressed in three curves, where the symbol for the curve that represents the high student level with (H) and the median with (M) and low with (L). This scale gives a visual view of the separation ability of classification algorithms In this research, each performance of each algorithm is expressed in three curves, where the symbol for the curve that represents the

high student level with (H) and the median with (M) and low with (L) as shown below:

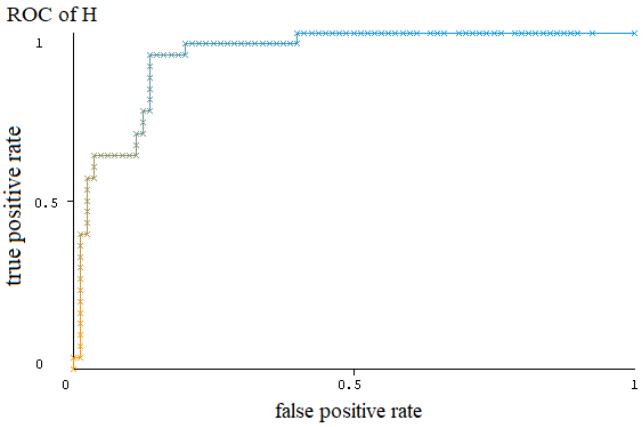

Fig (3): ROC curve of H using Random Forest

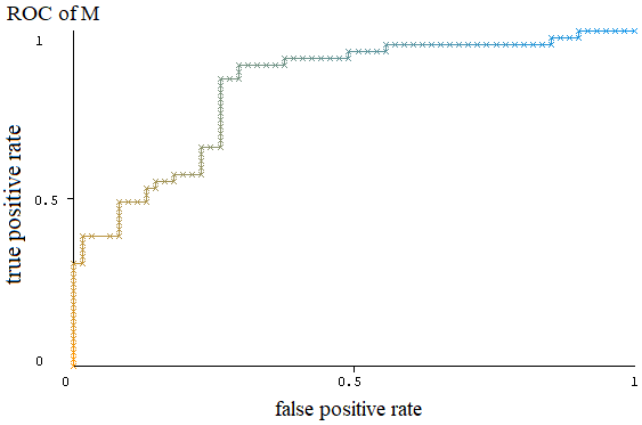

Fig (4): ROC curve of M using Random Forest

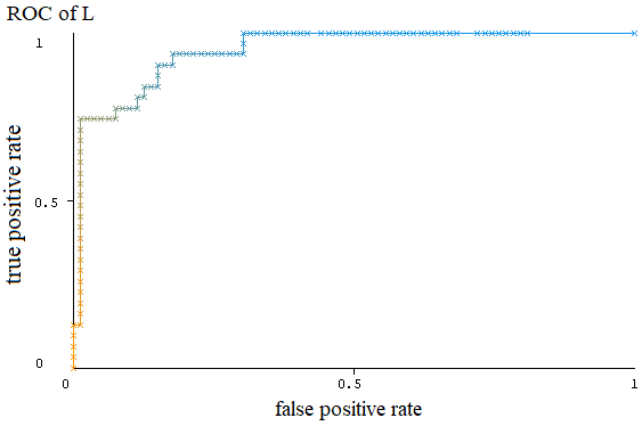

Fig (٥): ROC curve of L using Random Forest

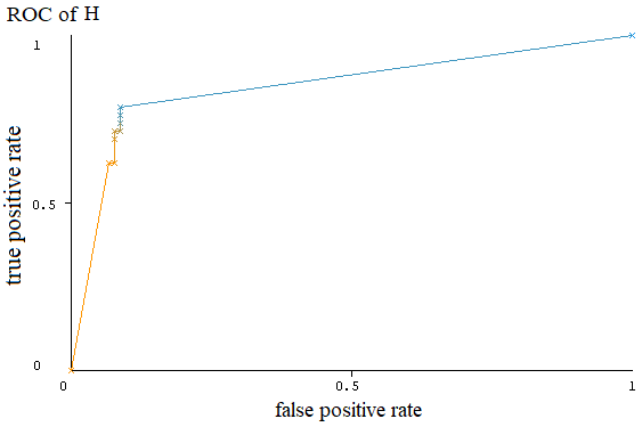

Fig (6): ROC curve of H using Gradient Boosted Trees

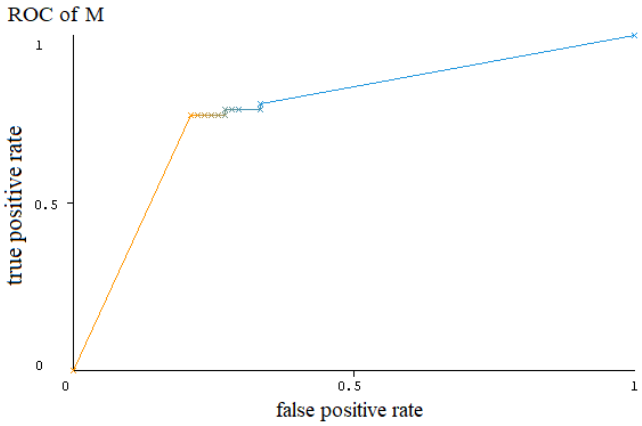

Fig (7): ROC curve of M using Gradient Boosted Trees

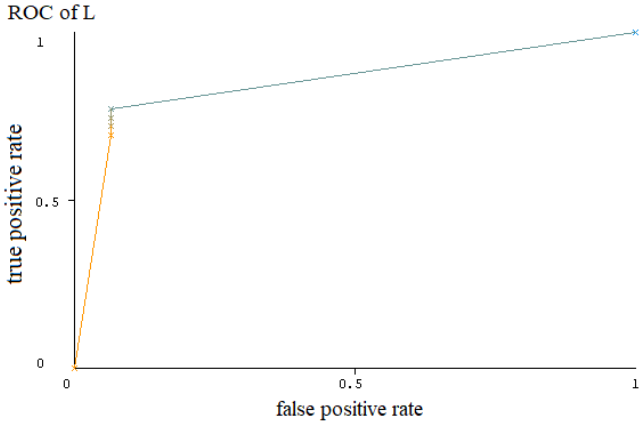

Fig (8): ROC curve of L using Gradient Boosted Trees

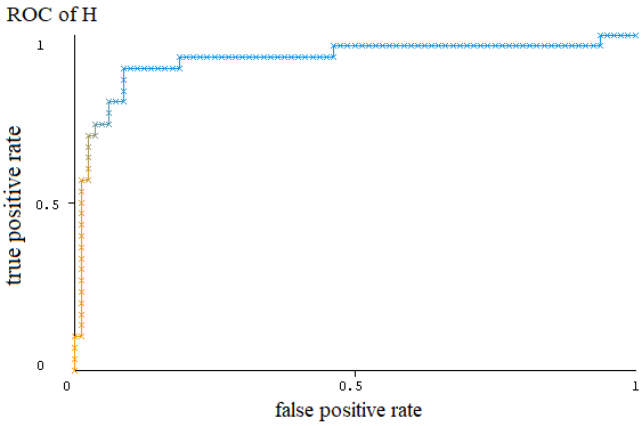

Fig (9): ROC curve of H using Deep Learning

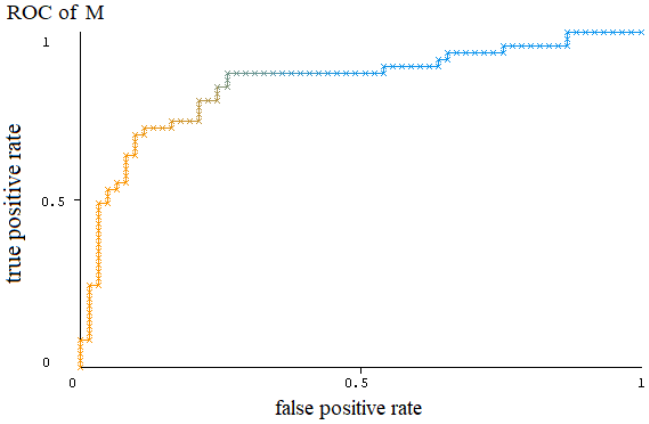

Fig (10): ROC curve of M using Deep Learning

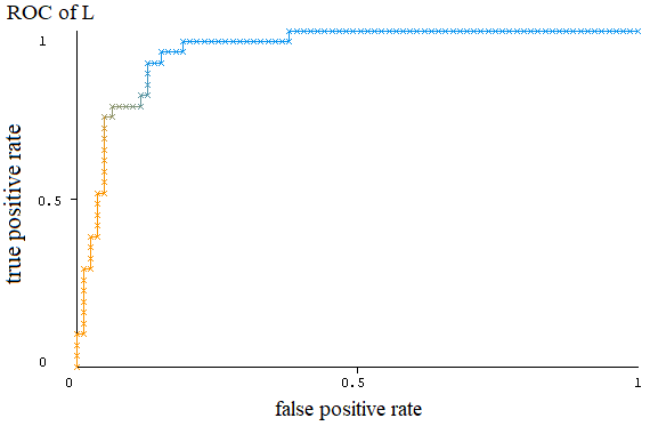

Fig (11): ROC curve of L using Deep Learning

As shown in Figure 3, 5, 9, 10, and 11 it is the one with the highest susceptibility to separation since the curves in Figure 9, 10, 11 belong to Deep Learning and this shows that they are appropriate for a classification of all levels of student performance.

CONCLUSION

Data mining in the education system is extremely important for analyzing and anticipating student performance in academics by looking at different performance factors. This study will provide a solution to pre-assess student performance, which in turn helps to develop student performance and care for it in a timely manner in the right direction. In this study, the performance of machine learning algorithms has been compared for the purpose of data mining educational institutions and it has been shown that these algorithms are promising in the field of student performance forecast and the Deep Learning is an appropriate algorithm for this type of dataset.

REFERENCES

[1] Sultana, J., Rani, M. U., & Farquad, M. A. H. (2019).Student's performance prediction using deep Learning and data mining methods. *International Journal of Recent Technology and Engineering (IJRTE), 8(1S4)*, 2277-3878.

[2] Scheuer, O., & McLaren, B. M. (2012). Educational data mining. *Encyclopedia of the Sciences of Learning*, *Springer* ,1075-1079., doi: 10.1007/978-1-4419-1428-6_618.

[3] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601-618.

[4] Zaker, N. A., Alsaleem, N., & Kashmoola, M. A. (2019). Multi-agent models solution to achieve EMC in wireless telecommunication systems. In Proceedings - 2018 1st Annual International Conference on Information and Sciences, AiCIS 2018 (pp. 311–314). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/AiCIS.2018.00061.

[5] Ho, T. K. (1995). Random decision forests. *In Proceedings of 3rd international conference on document analysis and recognition ,IEEE*, 1, 278-282. doi: 10.1109/ICDAR.1995.598994

Archived from the original (PDF) on 17 April 2016.

[6] Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in medicine*, *22*(9), 1365-1381. doi: 10.1002/sim.1501.

[7] Elith, J., & Leathwick, J. (2017). Boosted Regression Trees for ecological modeling. *R Documentation. Available online: https://cran. r-project. org/web/packages/dismo/vignettes/brt. pdf (accessed on 12 June 2011)*.

[8] *Baradwaj, B. K., & Pal, S. (2012).* Mining educational data to analyze students' performance. *IJACSA, 2(6), 63-69 .arXiv preprint arXiv:1201.3417.*

[9] Asif, R., Merceron, A., & Pathan, M. K. (2014). Predicting student academic performance at degree level: a case study. *International Journal of Intelligent Systems and Applications*, *7*(1), 49. doi: 10.5815/ijisa.2015.01.05.

[10]] Lemeshko, O., Yevdokymenko, M., & Alsaleem, N. Y. A. (2018). Development of the tensor model of multipath QoE-routing in an infocommunication network with providing the required quality rating. Eastern-European Journal of Enterprise Technologies, 5(2), 40–46. https://doi.org/10.15587/1729-4061.2018.141989.

[11] Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, *10*(1), 3-12. ISSN: 1512-8962.

[12] Alsaleem, N. Y. A., Kashmoola, M. A., & Moskalets, M. (2018). Analysis of the efficiency of spacetime access in the mobile communication systems based on an antenna array. Eastern-European Journal of Enterprise Technologies, 6(9–96), 38–47. https://doi.org/10.15587/1729-4061.2018.150921..