Prediction of Students' Performance in e-Learning Environment using Data Mining/ Machine Learning Techniques

*Brijesh Kumar Verma[1]
Research Scholar,
Amity Institute of Information Technology (AIIT), Amity University, Lucknow, U.P, INDIA
vermamtech05@gmail.com

Hemant Kumar Singh[2]
Associate Professor and Head
Department of Computer Science & Engineering
SMS Institute of Technology, Lucknow, U.P. INDIA
hemantbib@gmail.com

Dr. Nidhi Srivastava[3]
Assistant Professor,
Amity Institute of Information Technology (AIIT), Amity University, Lucknow, U.P, INDIA
nsrivastava2@lko.amity.edu

D.O.I - 10.51201/JUSST/21/05179
http://doi.org/10.51201/JUSST/21/05179

Abstract—The COVID-19 pandemic has drastically changed the way od of learning. During this pandemic the learning has shifted from offline to online. student's performance prediction based on their relevant information has emerged new area for educational institutions for improving teaching learning process, changes in course curriculum. Machine leaning technology can be helpful in predicting the performance of student and accordingly the institutions can make required changes in in their lecture delivery and curriculum.
This paper utilized some machine learning methodologies to predict the students' performance. Educational data of open University(OU) is analysed Based on parameters that are demographic, engagement and performance. In the experimental analysis. In the experimental analysis, the k-NN approach performed best in some cases and ANN performed best in other cases among all compared algorithms on OU dataset.
Keywords: E-Learning Environment (ELE), Educational Data Mining(EDM), Machine Learning (ML), Performance Classification.

1. Introduction

Substantial use of Internet technology has transformed the education system from traditional offline mode to online/blended mode called as E-Learning Environment (ELE). This has emerged as a new area of research for researchers [1]. Jani et al. [2] reasoned that blended learning of face-to-face and using the ELE platform improved the student's understanding and performance as well. During the COVID-19 pandemic all the academic institutions are closed and shifted to online mode that increased the importance of E-leaning Environment [3]. The major challenges for the educational institutions is actual and trustworthy evaluation of student's performance on ELE. It becomes very difficult and complex to e-access the students' performance without cheating by students from Internet, written notes and any other sources [4]. The real students' performance prediction will be helpful to teacher's/course coordinators at the initial phases of the course who needs attention and help [5].

In previous decade Educational Data Mining was an effective tool to find out the useful knowledge and patterns from large educational datasets [6] among all the existing approaches. It includes the use of data mining (DM) methods to data sets [7]. But today Machine Learning (ML) and classification and regression approaches are more effective and accurate in predicting students' performance. Prediction efficiency and accuracy is very much dependent on the data type of features being used, dimension of a dataset, and variety in the dataset.
ML methods that includes k-NN, SVC, ANN, Random Forest (RF), AdaBoost and Decision Tree (DT) are used as key methods for predicting students' performance based on regression and classification analysis on the ELE datasets.

2. Related Work:

This section includes the word done in measuring the student's performance using AI and Machine Learning. The forecast is dependent on teaching style, learning material and access patterns datasets. The review is presented in chronological order that indicates year wise changes in this field over the years.
*In year 2015***,** Elbadrawy et al. [8] applied a class of linear multi-regression techniques to predict the performance of students with the help of the educational data. Models used the data features that includes past performance, interaction with Learning Management System (LMS) and course related activities. The given models were certified on a custom collected dataset of 11,556 student entries, and 832 courses. The result shows the Root Mean Square Error (RMSE) of multi regression model was stated as 0.147, it has increased from the single regression model.
In year 2016, Yee-King et al. [9] proposed a k-NN based model to forecast the student's scores from collaborative social learning. A multivariate classification method was used to tolerate the weak classification. The given method was validated on the custom created dataset in year 2014 from an online course at Coursera. The collected dataset comprised of the total number of User Interface (UI) clicks and mouse-overs created during the course. The achieved classification accuracy was 88%, 77% and 31% for 2, 3 and 10 score bands respectively.
In year 2017, Al-Shehri et al. [10] used k-NN and Support Vector Machine (SVM) ML methods to predict the of students' performance in the final exam. A custom dataset of the University of Minho, Portugal with 395 data samples was used to certify the performance of ML models. Dataset comprised of student family background and individual data attributes.
Through the analysis it was reported that SVM was found slightly better k-NN in terms of accuracy. In the current year, Iqbal et al. [11] compared the three different ML techniques that includes CF, Matrix Factorization (MF) and Restricted Boltzman Machines (RBM) for predicting the score of students used a custom dataset of International Technical University (ITU), Pakistan with 225 student records to validation ML algorithms. Dataset comprised of performance based features including previous academic performance and interview score.
The RBM method shown the best results with an RMSE of 0.3.
In 2018, Hussain et al. [12] made a comparative study to forecast the student engagement and its impact on performance appling numerous learning-based algorithms. They applied DT, Classification, Regression Tree (CART), JRIP Decision Rules, Gradient Boosting Trees (GBT) and Naive Bayes Classifier (NBC) on OU dataset to forecast the student engagement. Dataset of only July 2013 session (384 records) was used with demographic, performance, and learning behaviour features. It was reported that J48 decision tree algorithm surpassed others with maximum accuracy of 88.52% and recall 93.4%. In the same year, Heuer and Breiter [13] In his first assessment find out students at-risk by applying numerous ML methods. They applied standard OULAD Dataset with 32,593 student entries. activity-based and performance topographies were used to forecast the performance. They also applied ML methods SVM, NB,

RF, XGBoost and Logistic Regression (LR) concluded that SVM is best in all applied algorithms with the accuracy of 87.98%.

In 2019, Sekeroglu et al. [14] examined the student performance forecast and classification by applying various ML algorithms. They applied Long-Short Term Memory (LSTM), Backpropagation (BP) and Support Vector Regression (SVR) for forecast while BP, SVM and Gradient Boosting Classifier (GBC) for classification. Student Performance Dataset (SPD) was used for prediction analysis and Students' Academic Performance Dataset (SAPD) used for classification analysis. Datasets consists of student's demographic information, academic background history and behavioural pattern features. Authors concluded that SVR is the best performing algorithm for forecast and BP is best performer for classification.Later, El Fouki et al. [15] proposed an advanced classification model based on deep learning and Principal Component Analysis (PCA) for the prediction of student's performance.

The given multi-dimensional technique reduces the dimensions of data and extract significant information from the data to advance the classification accuracy of the model. A gathered dataset with 496 records consisting of features including student's performance, section information and activity participation. Dataset was pre-processed using PCA for dimensionality reduction and then analysed using deep learning model, Multi-Layer Perceptron (MLP) and Bayes Net. The deep learning was the best performer among these algorithms with classification accuracy of 92.54%. In same year, Hussain et al. [16] given a model based on internal assessment by applying deep learning with Adam optimizer to predict student' performance. In addition to the deep learning model, two other approaches including Artificial Immune Recognition System (AIRS) v2.0 and AdaBoost were also implemented for comparative investigation. The authors used custom dataset of 10,140 records from 3 different colleges in India. Students performance in various tests was the core feature of the used dataset to predict the final scores. The results shown that a deep learning model with binary cross-entropy loss and sigmoid activation was best performer with classification accuracy 95.34%. Later in 2019, Ajibade et al. [17] applied numerous classification algorithms on behavioural learning data of students to forecast the performance. In addition, they used Differential Evolution (DE) for behavioural feature selection. The given methods were validated on the custom dataset with a record of 500 students. Dataset contained demographic, academic, learning process, and behavioural learning characteristics. DT, k-NN and SVM methods were used and DT was the best performer among these with good margin.

Recently in 2020, Tomasevic et al. [18] performed a comparative study to investigate the effect of different features on student's assessment forecast using SVM, k-NN, ANN, DT, Bayesian Linear Regression (BLR), and Regularized Linear Regression (RLR)and statistical approaches. The authors used a part of the OULAD dataset with demographic, engagement and performance features. F1 score and RMSE were used as performance measures for classification and regression models. Authors reported 96.62% F1 score for ANN using engagement and performance features, while 96.04% SVM (RBF kernel) using demographic, engagement and performance features. In the same year, Hooshyar et al. [19] proposed a novel approach PPP based on the delay behaviour of students to predict their performance. The given algorithm focused on student's assignment submission behaviour as the main indicator in forecasting their performance. The proposed method was validated on custom dataset of 242 students from the University of Tartu, Estonia. Common ML approaches including Linear SVM (L-SVM), Radial SVM (R-SVM), DT, Gaussian Process (GP), RF, NN, AdaBoost, and NB were applied on the dataset. In results Neural Network was best for categorical features with 96% accuracy and LSVM best in all aspects classification accuracy of 95%.

In the same year, Waheed et al. [20] proposed the use of Deep Neural Network (DNN) for forecasting the students' performance from the VLE big data.

The authors used the OULAD open-source dataset consisted of 32,593 student records. Dataset features included demographics, clickstream behaviour and assessment performance. The results shown that proposed deep learning-based method surpassed conventional regression and SVM techniques with the accuracy of up to 93%.

3. Data Mining and Machine Learning Techniques overview:

Naive-Bayes: It is based on Bayes theorem which is used for classification. It finds the probability of an object having certain features mapped to a particular class. So it can be said as probabilistic classifier. In this technique occurrence of each feature does not depends on occurrence another feature. For more detail refer to [20].

Random Forest: This is a supervised learning algorithm. Here many decision trees are combined to form a random forest algorithm. That means it is a collection multiple number of classification trees. It is used for classification and regression analysis. Every decision tree comprises of some rule based system. For the taken training dataset having targets as well as features, the decision tree algorithm will have some set of rules. In random forest distinct from decision trees it is not required to calculate information gain to find root node. It uses the rules of every randomly formed decision tree to calculate the outcome and stores the calculated outcome and it calculates the vote for each predicted target. In this way high voted prediction is taken as the final prediction from the random forest algorithm. For more detail refer to [20].

k Nearest Neighbour (k-NN): k-NN is a non-parametric machine learning approach first proposed by Fix et al. [21] in 1951. It classifies input in to one of the target group based on popularity among its neighbours. This algorithm is best suited data distribution is unknown. KNN uses the training dataset directly to make the prediction. So it is simplest and best suited for prediction generation. For more detail refer to [22]

Support Vector Mechanism (SVM)

SVM is very robust supervised ML algorithms proposed by Vapnik in year 1963 and it was extended by Boser et al. [23]. This algorithm creates multiple hyperplanes in the high-dimensional space with the objective to attain good separation among the hyperplanes. The high margin among hyperplanes implies the reduced generalization loss. The test sample is classified in two classes. Every test sample is denoted as an m-dimensional vector and separated by a (m-1) dimensional hyperplane. Test samples may be separated by multiple hyperplanes, however, the best one is selected based on the maximum separation in linear classification case [24,25].

Artificial Neural Network (ANN)

ANN is a machine learning algorithm that simulates the behaviour of human brain. It contains nodes. The network contains layers, nodes and connections. Nodes represents artificial neurons. It has the capability input signal processing and sending it to other neurons.

Usually, ANN consist of input, output layer as well as number of hidden layers with artificial neurons at each layer connected with each other. For more detail refer to [26,27].

4. Experimental Results and discussion:

This section gives the information about the data set and protocols used for performing the experimental analysis and results of implemented Naïve Bayes, Random Forest, k-NN, SVM and ANN Data Mining and Machine Learning algorithms for the student performance analysis based on different combination input features

(i) Dataset: Open University dataset [28] was taken from the Kaggle and used for accomplishment the both experiments. Dataset includes of total 32,593 student entries from 15 diverse countries.

Moreover, dataset also contains courses chosen by students, students' demographics and students' interactions with e-learning Environment. The dataset was cleaned and extracted the desirable features. Dataset cleaning implies dealing with missing values and assigning arithmetic values to phrases for classification analysis. Demographic (D), Engagement (E) and Performance (P) are the input features in the dataset and students' performance measured as pass or fail is the target variable.

(ii) Data Pre-Processing

The final data set is a comma Separated file. that is divided in Demographic, Engagement and Past Performance. Details of dataset features are as follows-

Table-1 Details of Data set features

Demographic features	**Values**	**Description**
Gender	[0,1]	1: Male, 0: Female
Highest Education	0, 0.25,0.50,0.75,1	0: Below high school, 0.25: High school, 0.5: Diploma, 0.75: Bachelor, 1: Post graduate
Age	0,0.5,1	0: <35, 0.5: 35-55, 1: >55

Engagement		
Total clicks	[0-N]	0 - N

Performance		
Score per assessment	[0-1	] 0 - 100
No. of Attempts	[0-1]	0 – N
Final exam score	[0-1]	0 – N

Table-2: Performance comparison of Classification Algorithms to predict final exam result

		D	**E**	**P**	**D+E**	**D+P**	**E+P**	**D+E+P**
Accuracy	CNN	0.7591	0.9620	0.9884	0.9526	0.9934	0.9965	0.9922
	KNN	0.6905	**0.9622**	**0.9992**	**0.9626**	0.9966	0.9966	0.9934
	ANN	**0.7594**	0.9984	0.9990	0.9298	**0.9982**	**0.9984**	**0.9979**
	SVM	0.7594	0.9516	0.9984	0.9340	0.9963	0.9983	0.9958
	NAÏVE BAYES	0.6804	0.9362	0.9893	0.9521	0.9865	0.9964	0.9945
	RANDOM FOREST	0.6805	0.9164	0.9970	0.9533	0.9782	0.9964	0.9947

		D	**E**	**P**	**D+E**	**D+P**	**E+P**	**D+E+P**
F1 Score	CNN	0.8531	0.9460	0.9979	0.9565	0.9886	0.9928	0.9944
	KNN	0.8108	**0.9756**	**0.9995**	**0.9759**	0.9977	0.9977	0.9956
	ANN	**0.8633**	0.9480	0.9989	0.9535	**0.9987**	**0.9988**	**0.9986**
	SVM	0.8632	0.9685	0.9993	0.9564	0.9975	0.9988	0.9972

	NAÏVE BAYES	0.8412	0.9585	0.9983	0.9484	0.9778	0.9988	0.9911
	RANDOM FOREST	0.8436	0.9655	0.9773	0.9425	0.9855	0.9988	0.9862

		D	**E**	**P**	**D+E**	**D+P**	**E+P**	**D+E+P**
J-Index	CNN	0.6799	0.9521	0.9981	0.9500	0.9855	0.9855	0.9923
	KNN	0.6818	**0.9524**	**0.9991**	**0.9530**	0.9955	0.9955	0.9913
	ANN	**0.7596**	0.9012	0.9979	0.9112	**0.9975**	**0.9977**	**0.9973**
	SVM	0.7593	0.9290	0.9987	0.9165	0.9951	0.9977	0.9945
	NAÏVE BAYES	0.7583	0.9270	0.9897	0.9055	0.9981	0.9937	0.9965
	RANDOM FOREST	0.7553	0.9370	0.9917	0.9075	0.9971	0.9957	0.9955

Algorithms were applied to classify the performance of based on demographic, engagement, past performance from the OU education dataset [28]. Results of experiment are shown in Table-2. From the results of Table-2 shows that k-NN performed best for E, P and D+E cases with accuracy of of 0.9622, 0.9992 and 0.9626 respectively. However, for D, D+P, E+P and D+E+P cases ANN performed best with the accuracy of 0.7594, 0.9982, 0.9984 and 0.9979, respectively.

5. Conclusion:

This paper predicted the performance of the dataset by using various data mining and machine learning techniques. In the experiments first data cleaned and prepared in CSV file then some data mining and machine learning algorithms were applied to predict the performance of the students in final examinations and results are presented in tabular form and results of best performing algorithms are highlighted in the table. The experimental results show that K-NN algorithm is performing better than ANN and SVM, Naïve Bayes and Random Forests for various feature variations and in some cases ANN is performing better than other algorithms. In future some more parameters with real dataset will be taken and precision and recall will also be calculated and extended work on missing values.

References:

1] Sweta S. (2021) Educational Data Mining in E-Learning System. In: Modern Approach to Educational Data Mining and Its Applications. SpringerBriefs in Applied Sciences and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-33-4681-9_1

[2] J. Jani, R. Muszali, S. Nathan, and M. S. Abdullah, “Blended learning approach using frog vle platform towards students’ achievement in teaching games for understanding,” Journal of Applied and Fundamental Sciences, vol. 10, pp. 1131–1151, Jan. 2018.

[3] I. Chatziralli, C. V. Ventura, S. Touhami, R. Reynolds, M. Nassisi, T. Weinberg, K. Pakzad-Vaezi, D. Anaya, M. Mustapha, A. Plant, M. Yuan, and A. Loewenstein, “Transforming

ophthalmic education into virtual learning during COVID-19 pandemic: a global perspective," Eye, pp. 1–8, Jul. 2020, publisher: Nature Publishing Group.
[4] H. Mellar, R. Peytcheva-Forsyth, S. Kocdar, A. Karadeniz, and B. Yovkova, "Addressing cheating in e-assessment using student authentication and authorship checking systems: Teachers' perspectives," International Journal for Educational Integrity, vol. 14, p. 2, Feb. 2018.
[5] C. Vegega, P. Pytel, and M. F. Pollo-Cattaneo, "Application of the Requirements Elicitation Process for the Construction of Intelligent System-Based Predictive Models in the Education Area," in Applied Informatics, ser. Communications in Computer and Information Science,H. Florez, M. Leon, J. M. Diaz-Nafria, and S. Belli, Eds. Cham:Springer International Publishing, 2019, pp. 43–58.
[6] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," Procedia Computer Science, vol. 72, pp. 414–422, Jan. 2015
[7] A. S. J. Abu Hammad, "Mining Educational Data to Analyze Students' Performance (A Case with University College of Science and Technology Students)," Central European Researchers Journal, vol. 4, no. 2, 2018.
[8] A. Elbadrawy, R. Studham, and G. Karypis, "Personalized Multi-Regression Models for Predicting Students' Performance in Course Activities," Mar. 2015.
[9] M. Yee-King, A. Grimalt-Reynes, and M. d'Inverno, "Predicting student grades from online, collaborative social learning metrics using K-NN." in EDM, 2016, pp. 654–655.
[10] H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed, J. Alhiyafi, and S. 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), 2017, pp. 1–4, tex.organization: IEEE.
[11] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, "Machine learning based student grade prediction: A case study," arXiv preprint arXiv:1708.08744, 2017.
[12] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores," Oct. 2018, iSSN: 1687-5265 Pages: e6347186 Publisher: Hindawi Volume: 2018. [Online]. Available: https://www.hindawi.com/journals/cin/2018/6347186/
[13] H. Heuer and A. Breiter, "Student success prediction and the tradeoff between big data and data minimization," DeLFI 2018-Die 16. E-Learning Fachtagung Informatik, 2018, publisher: Gesellschaft f¨ur Informatik eV.
[14] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student Performance Prediction and Classification Using Machine Learning Algorithms," Mar. 2019, pp. 7–11.
[15] M. El Fouki, N. Aknin, and K. E. El Kadiri, "Multidimensional approach based on deep learning to improve the prediction performance of DNN models," International Journal of Emerging Technologies in Learning (iJET), vol. 14, no. 02, pp. 30–41, 2019.
[16] S. Hussain, Z. Muhsen, Y. Salal, P. Theodorou, F. Kurto˘glu, and G. Hazarika, "Prediction Model on Student Performance based on Internal Assessment using Deep Learning," International Journal of Emerging Technologies in Learning (iJET), vol. 14, p. 4, Apr. 2019.
[17] S. Ajibade, N. Ahmad, and S. M. Shamsuddin, "Educational Data Mining: Enhancement of Student Performance model using Ensemble Methods," IOP Conference Series: Materials Science and Engineering, vol. 551, p. 012061, Aug. 2019.
[18] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," Computers & Education, vol. 143, p. 103676, Jan. 2020.

[19] D. Hooshyar, M. Pedaste, Y. Yang, L. Malva, G.-J. Hwang, M. Wang, H. Lim, and D. Delev, "From Gaming to Computational Thinking:An Adaptive Educational Computer Game-Based Learning Approach,"Journal of Educational Computing Research, p. 0735633120965919, Oct. 2020, publisher: SAGE Publications

[20] Han, Jiawei, and Micheline Kamber. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2001.

[21] E. Fix, Discriminatory analysis: nonparametric discrimination, consistency properties. USAF School of Aviation Medicine, 1951.

[22] D. Coomans and D. L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-nearest neighbour classification by using alternative voting rules," Analytica Chimica Acta, vol. 136, pp. 15–27, 1982.

[23] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

[24] Y. Pu, D. B. Apel, and H. Xu, "Rockburst prediction in kimberlite with unsupervised learning method and support vector classifier," Tunnelling and Underground Space Technology, vol. 90, pp. 12–18, 2019.

[25] D. Fradkin and I. Muchnik, "Support vector machines for classification," DIMACS series in discrete mathematics and theoretical computer science, vol. 70, pp. 13–20, 2006.

[26] A. Abraham, "Artificial neural networks," Handbook of measuring system design, 2005.

[27] K. Mehrotra, C. K. Mohan, and S. Ranka, Elements of artificial neural networks. MIT press, 1997.

[28] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," Scientific data, vol. 4, p. 170171, 2017.