2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning

Hina Gull[1], Madeeha Saqib[2], Sardar Zafar Iqbal[3], Saqib Saeed[4]
[1,2,3,4]Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box No. 1982, Dammam, Saudi Arabia
[1]hgull@iau.edu.sa,[2]mssaeed@iau.edu.sa,[3]saiqbal@iau.edu.sa,[4]sbsaed@iau.edu.sa

Abstract— **Early indications regarding students' progress help academics to optimise their learning strategies and focus on diverse educational practices to make the learning experience successfully. Machine learning application can help academics to predict the expected weaknesses in learning processes and as a result they can proactively engage such students in better learning experience. We applied logistic regression, linear discriminant analysis, K-nearest neighbors, classification and regression trees, gaussian Naive Bayes and support vector machines on historical data of student grades in one of the undergraduate courses and developed a model to predict the grades of students taking the same course in the next term . Our experiments show Linear discrimination analysis as the most effective approach to correctly predict the students' performance outcome in final exams. Out of total 54 records, 49 were predicted by model as expected giving 90.74% of accuracy.**

Keywords—Machine Learning, Predictive Analysis, Student Performance Prediction, Linear discrimination analysis

I. INTRODUCTION

Academia has seen a significant shift from conventional learning towards a student centric learning approach. Conventional learning approach focuses on delivering lecture and student passively absorb the content and teacher use different assessments to evaluate the performance of students[1]. However, student centric learning approach advocates for fostering a successful learning experience of learners. As a result, it becomes very critical for the instructors to continuously monitor the progress of learners and optimize educational strategies accordingly. Therefore, formative assessments become a vital tool for instructors to assess the effectiveness of learning process.

Machine learning applications have huge potential to help instructors to identify the weak performance of students by enabling an early warning system. As a result, the instructors can focus more on such weak students to make them ready by the time summative assessments are scheduled. In this paper, we have applied different machine learning algorithms on the historic results of a course being taught in bachelor's in computer information systems program to find out the prediction accuracy. These models will be used on formative assessments of future students and if the model predict that students have higher probability of failing a course then alternative educational strategies would be employed to make his/her learning experience improved.

In the next section related work is discussed which is followed by problem statement, discussion of the experiment details and results followed by conclusion at the end.

II. BACKGROUND STUDY

There have been many research contributions in the education context dealing with education quality [2][3][4], teaching reflections [5][6], curriculum design [7] and plagiarism [8]. With the advent of data science and machine learning several research studies have been conducted to predict student performance using machine learning techniques. Iqbal et al. [9] have used advance machine learning techniques to analyse real data of students in private sector university. They argued that early prediction of student grades can help them to perform better in their courses. Similarly, Elbadrawy et al. have used and investigated different recommender systems to correctly predict student grades ahead of time. They argued that Matrix Factorization and Personalized Multi-Linear Regression Models can be successfully used to predict the next term grades of students with low error rate [10]. Several classification models have also been used by Xu Zhang et al [11] to predict grades of student academic performance. By using and comparing classification algorithms Naive Bayes, Decision Tree, Multilayer Perceptron and Support Vector on real student data, they have argued to achieve 65.90% accuracy on the training set and 62.04% accuracy in the test set. Additionally, Rida et al [12] have also used and compare classification algorithms in terms of accuracy, precision, recall and F-factor to predict student grades. Study revealed that among Naïve Bayesian [NB] classifier, Decision Tree [DT], and Multi-Layer Perceptron (MLP), DT is most accurate with 97.69% accuracy and 95.6% of precision on their set of data. Wan Fairos et al. [13] have developed predictive model using supervised learning classification algorithms to predict student performance in one of the universities in Malaysia. Study showed that based on several measures such as accuracy measure, precision, recall and ROC curve, the Naïve Bayes outclass other classification algorithms. Study conducted by Chitra and Rashmi [14] showed the machine learning techniques and mathematical models to aid the instructors in understanding of student's learning styles and behaviours. Another study by Nikola et al. [15] demonstrate a detailed analysis and comparison of several supervised machine learning techniques to output student performance prediction in the final exams, i.e. determining the number of students who are at the verge of dropping out from the course. It also helps them to predict

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

the scores of students in the coming exams. Sotiris Kotsiantis et al [16] in their study have showed efficiency of machine learning methods to forecast student performance in distance education systems. They argued that machine learning methods could enable instructors to predict student grades with substantial accuracy ahead of time i.e. before the final examination that will enable students to perform better. Similarly several other studies [17][18][19][20] have used supervised and unsupervised machine learning algorithms and techniques to predict student performance ahead of time in order to achieve better student performance.

III. Problem Statment

For this study, we would like to early predict (before final exam) student final grade (A, B, C, D, E or F) in one of the undergraduate courses in order to identify weak learners to overcome the difficulties, which they are facing in the learning process. Early grade prediction will not only help instructors to know the students who need academic support, but will also help students to work on their weakness to get good grades in the finals. Furthermore, the findings will help instructors to revisit their educational strategies to foster better learning experience.

IV. Methodology

Real historical data of students taking one of the undergraduate courses is collected across the batch 2016, 2017, 2018 and 2019. The dataset contains 250 students of undergraduate program. The data of each student contains marks obtained by students in assessments (Quiz1=5 Marks, Quiz2=5 Marks, Midterm Exam=20 marks, Project=15 Marks, Lab=15 Marks) as attributes and grade as target. Figure 1 shows the distribution of obtained marks by each student in each attribute/parameter:

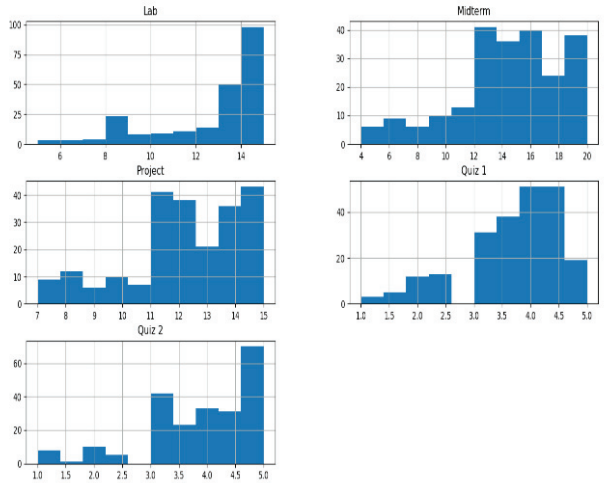

Fig. 1. Marks Distribution

V. Experimentation

Experiments were conducted using python in PyCharm. Data set is divided into ratio of 80% to 20%, having 80% as training data and 20% as testing data. Random student data is also used to predict grades in order to validate the model. Data underwent several steps to get it ready for use.

A. SMOTE for Imbalance Data

SMOTE (Synthetic Minority Over-sampling Technique) was applied for imbalanced classes of data. It is an oversampling method which is used to create synthetic samples of minority classes [21]. In our data F class was minority class, so SMOTE was used to balance data.

B. Parameter Selection

All assessments (Quiz1, Quiz2, Midterm, Project and Lab) were selected as parameter as all of them significantly contribute in the performance of students. We know that the student's final grades are depend on all these assessments in the course.

C. Pattern Identification

This step consists of model training, pattern identification, testing, evaluation results. As mentioned earlier data set was divided into testing and training sets. In the training set, the model is built from the classification techniques. Testing set is used to assess the model. After that results will be evaluated. In order to check which of the algorithm will best suit in prediction, we have tested following six classification algorithms:

Logistic Regression (LR)
Linear Discriminant Analysis (LDA)
K-Nearest Neighbors (KNN).
Classification and Regression Trees (CART).
Gaussian Naive Bayes (NB).
Support Vector Machines (SVM).

D. Best Model Selection

In the experiment, five standard measures Accuracy, Recall, Precision, Kappa and F-measure are used to assess the quality of classification model.

a) Accuracy:

Figures below shows the comparison between different algorithms in terms of accuracy.

Algorithm	Accuracy
LR	0.65
LDA	0.81
KNN	0.78
CART	0.68
NB	0.75
SVM	0.80

Fig. 2. Accuracy

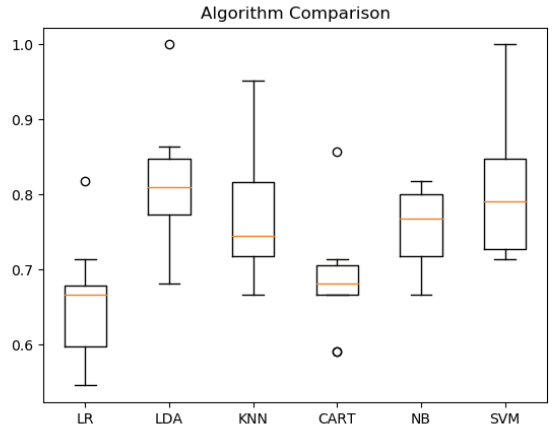

Fig. 3. Algorithm Comparision-Accuracy

It can be clearly seen from the accuracy data that Linear Discrimination Analysis has the highest estimated accuracy value of about 0.81 or 80.1% followed by SVM (Support Vector Machine) which is also 0.80 or 80%.

b) Precision, Recall, F1-Score and Kappa:

TABLE I. MEASURES AND THEIR VALUES

Algorithm	Precision	Recall	F1-Score	Kappa
LR	68.89	68.89	71.11	60.73
LDA	80	80	77.78	74.75
KNN	82.22	82.22	82.22	79.36
CART	71.11	73.33	68.89	77.18
NB	82.22	82.22	82.22	77.12
SVM	86.67	86.67	86.67	86.26

VI. EARLY GRADE PREDICTION

According to the comparative analysis given above, LDA (Linear discrimination analysis) seems to be the most accurate algorithm. So, model based on LDA was selected to predict grades of undergraduate students based on marks obtained in selected assessments (Quiz 1 and 2 (Q1, Q2), Lab (L), Project (P), Midterm Exam (M)) of the course. Table III shows student marks and predicted grades (PG) by the model, expected grades (EG) and status (S). For this case study LDA has correctly predicted 49 records out of 54 total records of students. Through early prediction we have also identified students who may need academic support to better perform in their future assessments.

TABLE II. EARLY GRADE PREDICTION

Sr	Q1	Q2	M	P	L	PG	EG	S
1	1.75	5	16.5	12	14	B	B	
2	5	5	17	11	14.5	A	A	
3	5	4	10.5	10	14.5	C	C	Needs Academic Support
4	5	5	17	12	14.75	A	A	
5	5	5	18	12	14.5	A	A	
6	3.25	4	14.5	13	14.5	B	B	
7	3.75	5	14.5	14	14.5	A	A	
8	1.5	5	7.5	12	14.5	C	C	Needs Academic Support
9	5	5	19.5	11	14.25	A	A	
10	3.75	5	17	10	13.5	A	B	
11	4.25	5	13	9	14.5	C	C	Needs Academic Support
12	2.5	4	14.5	9	14	C	C	Needs Academic Support
13	5	4	16.5	10	13	C	B	
14	3.25	5	11.5	12	12	D	D	Needs Academic Support
15	5	5	18	13	12	A	A	
16	3.25	5	13	14	14	B	B	
17	2.75	3	14	11	13	C	C	Needs Academic Support
18	4	5	17	14	12	B	B	
19	3.25	5	16	13	12	C	C	Needs Academic Support
20	4	4	17.5	14	13.5	B	C	
21	3.75	5	15.5	11	11	C	C	Needs Academic Support
22	5	5	16.5	12	14	B	B	
23	5	5	16	9	12	C	C	Needs Academic Support
24	4.25	5	15.5	13	13	B	B	
25	3.75	5	18.5	13	14.5	A	A	
26	5	5	18.5	13	14.5	A	A	
27	5	5	13	13	14.5	C	D	
28	3.5	5	15	14	14.5	B	B	
29	1.25	5	11	13	14	C	C	Needs Academic Support
30	5	5	19	13	14.5	A	A	
31	3.25	4	16.5	12	14.5	B	B	
32	5	5	20	13	14.5	A	A	
33	3.5	5	15.5	10	12.5	C	C	Needs Academic Support
34	4.5	5	7.5	10	14.5	C	C	Needs Academic Support
35	3.75	5	16.5	12	14.5	B	B	
36	4.25	5	18	12	14.5	B	B	
37	3	5	17.5	12	13.75	B	B	
38	3.75	5	19	12	14.5	A	A	
39	3.75	5	12.5	12	14	C	C	Needs Academic Support
40	3.5	5	11	12	13.5	C	C	Needs Academic Support
41	2.75	4	15.5	12	14.5	B	B	
42	1.75	5	17.5	12	14.75	B	B	
43	4.5	5	18	10	11	C	B	
44	3	3	11	11.5	11	D	D	Needs Academic Support
45	3	4	13	12	9	D	D	Needs Academic Support
46	4	3.5	12	13	8	C	C	Needs Academic Support
47	4	4	13	11	7	D	D	Needs Academic Support
48	4	4	12	14	12	D	D	Needs Academic Support
49	4	4.5	14	11	8	C	C	Needs Academic Support
50	4	5	15	12	10	B	B	
51	4	4	12	14	9	C	C	Needs Academic Support
52	3	4.5	17	11	7	C	C	Needs Academic Support
53	3	3.5	16	12.5	8	C	C	Needs Academic Support
54	4	3.5	16	14.5	8	D	D	Needs Academic Support

VII. Conclusion

Student centric learning has been very instrumental in improving the learning experience of students. Data mining applications can support this process by detecting low performing students early on which will help instructors to improve their educational strategies and keep these students motivated. In our study, we have used the historic results of a course to evaluate the accuracy of students' success. We have tested several classification algorithms and found linear discrimination analysis algorithm to be more accurate. Moreover, model based on LDA was trained to predict student performance outcome in their final exam which predicts records with 90.74% accuracy.

References

[1] S. Saeed, H. Gull, and S. Z. Iqbal, "Web 2.0 Usage by Saudi Female Students for Information Sharing in Public Sector University a Pilot Study," Int. J. Public Adm. Digit. Age, 2017, doi: 10.4018/ijpada.2017070106.

[2] A. Shafi, S. Saeed, Y. A. Bamarouf, S. Z. Iqbal, N. Min-Allah, and M. A. Alqahtani, "Student Outcomes Assessment Methodology for ABET Accreditation: A Case Study of Computer Science and Computer Information Systems Programs," IEEE Access, 2019, doi: 10.1109/ACCESS.2019.2894066.

[3] R. K. Siddiqui, S. Saeed, and F. Wahab, "Understanding Role Of Student Feedback In Quality Assessment: A Case Study," VFAST Trans. Educ. Soc. Sci., vol. 3, no. 2, 2014.

[4] H. Mumtaz, S. Saeed, and F. Wahab, "Quality of University Computing Education : Perception of Pakistani Students," vol. 2, no. 7, pp. 24–30, 2013.

[5] H. Gull, S. Saeed, S. Z. Iqbal, M. Saqib, Y. A. Bamarouf, and M. A. Alqahtani, "Reflections on teaching human computer interaction course to undergraduate students," Proc. - 2018 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2018, pp. 659–662, 2018, doi: 10.1109/CSCI46756.2018.00132.

[6] S. Saeed;, R. Aamir;, and Z. Mahmood, "Reflections on teaching database management systems to undergraduate students," Int. J. Educ. Econ. Dev., vol. 2, no. 4, 2011.

[7] S. Saeed;, M. Saqib;, and A. A. Salam, "Embedding Information Systems Environment Modules in Information System CurriculumNo Title," VFAST Trans. Educ. Soc. Sci., vol. 15, no. 1, 2018.

[8] S. Saeed, R. Aamir, and M. Ramzan, "Plagiarism and its implications on higher education in developing countries," Int. J. Teach. Case Stud., vol. 3, no. 2, 2011.

[9] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran, "Machine Learning Based Student Grade Prediction: A Case Study," pp. 1–22, 2017, [Online]. Available: http://arxiv.org/abs/1708.08744.

[10] A. Elbadrawy and G. Karypis, "Domain-aware grade prediction and top-n course recommendation," in RecSys 2016 - Proceedings of the 10th ACM Conference on Recommender Systems, 2016, doi: 10.1145/2959100.2959133.

[11] X. Zhang, R. Xue, B. Liu, W. Lu, and Y. Zhang, "Grade prediction of student academic performance with multiple classification models," ICNC-FSKD 2018 - 14th Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov., pp. 1086–1090, 2018, doi: 10.1109/FSKD.2018.8687286.

[12] R. M., N. F., and A. A., "Predicting and Analysis of Students' Academic Performance using Data Mining Techniques," Int. J. Comput. Appl., vol. 182, no. 32, pp. 1–6, 2018, doi: 10.5120/ijca2018918250.

[13] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," Indones. J. Electr. Eng. Comput. Sci., 2019, doi: 10.11591/ijeecs.v16.i3.pp1584-1592.

[14] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," in Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019, 2019, doi: 10.1109/COMITCon.2019.8862214.

[15] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," Comput. Educ., 2020, doi: 10.1016/j.compedu.2019.103676.

[16] S. Kotsiantis, C. Pierrakeas, I. D. Zaharakis, and P. E. Pintelas, "EFFICIENCY OF MACHINE LEARNING TECHNIQUES IN PREDICTING STUDE ... EFFICIENCY OF MACHINE LEARNING TECHNIQUES IN PREDICTING STUDENTS ' PERFORMANCE IN DISTANCE LEARNING SYSTEMS," Recent Adv. Mech. Relat. FIELDS Univ. PATRAS 2003 Honour Profr. Constantine L. Goudas, no. July, 2008.

[17] M. Ciolacu, A. F. Tehrani, R. Beer, and H. Popp, "Education 4.0-Fostering student's performance with machine learning methods," in 2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging, SIITME 2017 - Proceedings, 2017, doi: 10.1109/SIITME.2017.8259941.

[18] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," International Journal of Educational Technology in Higher Education. 2020, doi: 10.1186/s41239-020-0177-7.

[19] M. Al luhaybi, A. Tucker, and L. Yousefi, "The Prediction of Student Failure Using Classification Methods : A Case study," 2018, doi: 10.5121/csit.2018.80506.

[20] S. Fedushko and T. Ustyianovych, "Predicting Pupil's Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods," in Advances in Intelligent Systems and Computing, 2020, doi: 10.1007/978-3-030-16621-2_58.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., 2002, doi: 10.1613/jair.953.