2021 20th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Machine Learning Techniques to Predict Academic Performance of Health Sciences Students

Hana Alharthi, PhD.
Dept. of Health Information Management & Technology. College of Public Health
Imam Abdulrahman Bin Faisal University (IAU)
Dammam, Saudi Arabia
e-mail: Halharthi@iau.edu.sa

***Abstract*—Prediction of academic performance of health sciences students prior to being fully engaged in academic studies will identify those students who may need early intervention. Machine learning (ML), a branch of artificial intelligence, can be used to predict the academic performance of such students and the factors that continue to impact their academic performance. *Objective*: To use a best fit model in ML to predict the academic performance of health science students and rank the most important factors affecting their performance. *Method*: The academic records of 3468 students were extracted from the student information system (SIS), which included preparatory year great point average (GPA), high school GPA, Achievement Test (AT), General Aptitude Test (GAT), and cumulative GPA upon graduation. Multiple machine learning algorithms were used to develop the best fit model to predict students' performance GPA and identify factors that contributed to GPA. *Results*: The best performing classifier based on area under the curve (AUC) is random forest (.773) followed by naïve bayes (.758), Support Vector Machine (.686), k-nearest neighbors (.684) and decision tree (.658), the three scoring methods showed preparatory year GPA, gender, and high school GPA were the top variables predicating student cumulative GPAs. *Conclusion*: Random forest model can assist college administrators and faculty in health colleges to predict which students are more likely to underperform during their undergraduate studies.**

Keywords: Machine learning; algorithms; Classifiers; GPA ; ML.

I. INTRODUCTION

Anxiety profoundly affects quality of life of those affected by it. Universities continue to invest in resources to support the academic performance of their students. Yet the students that may need it the most are less likely to seek this support. As such, it is helpful to identify students that may struggle and underperform, especially as it relates to students in the biomedical fields who require strong academic performance to navigate the next step of their career whether further training or entering the workforce. Harnessing information technology can be a powerful tool to identify such students. Institutions have a breadth of student databases that can be easily saved, extracted, examined, and analyzed. Educational data mining (EDM) is a sub-field of data mining that recognize patterns that are not known before by utilizing smart algorithms [1]. As a result, educators can gain insights into their own academic environment that leads to better understanding on how to improve student academic performance, decrease number of failing students and decrease percentage of dropouts [2].

Data mining is utilized in many sectors such as in economics, business, health, retail, to name a few. The education sector has also used data mining [3]. To predict students' performance or retention rates among other issues. For example, current cumulative GPA for engineering students was used to predict their final semester GPA [4]. Another study used data mining to predict students at risk of dropping out [5].

Although numerous studies have targeted failing students to reveal correlates of failing and/or predicting which student will fail, we aimed to focus on students who are not failing yet are low performers, those with a "C" or " D" GPA, to provide educators with helpful information to establish the needed resources to maximize the performance of such students. Specifically, we evaluated student preparatory year GPA (first year), high school GPA , Achievement Test (AT) and the General Aptitude Test (GAT) and demographics to predict those students who are likely to graduate with a 'C" or "D" GPA from a health field (medicine, dentistry, nursing, applied medical sciences, pharmacy, and public health). We employed the orange data mining software to develop the best fit model to predict student GPA upon graduation.

II. LITERATURE REVIEW

Although admissions to health sciences colleges is highly competitive and is based on high standards of previous academic performance (e.g., high school GPA; high scores in standardized tests, etc.), some students struggle, underperform, and graduate with lower cumulative GPA. Therefore, capturing these low performers at early stage and providing them with appropriate support may make a difference to elevate their academic performance and aid in their future success, whether to continue to graduate school in some cases or enter their respective career sector.

Measuring student performance would enable decision makers and stockholders (management, administrative stuff, and faculty) to have the right information to enable them to make the best decisions for their students. This would rely on solid data that can predict student performance and reliance

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00015

on data mining. Numerous studies have addressed this issue, some focused on overall performance over long period of time (>1 year) while others evaluated shorter periods (one course or one semester) [6]. Course participation and perceptions examined for 533 students in their first year of study, classified them into three categories of academic performance , low-risk, medium-risk, and high-risk who are more likely to fail or drop out [7].

Multiple research work has used cumulative GPA as the main predictive attribute to assess student performance and most found that high school GPA, gender, ethnicity, quantitative SAT scores, verbal SAT scores, significantly impacted students' graduation [8],[9]. Further age, gender, parent's marital status, parent's qualification, cumulative GPA of 7500 students was evaluated to predict students' academic performance in their first year of college [10]. Several studies predicted students' final GPA using students grades information [11],[12]. Educational data mining (EDM) is a sub- branch of data mining that is based on statistics and machine learning to explore and investigate education data. It is applied in different aspects of the branch of education such as academic performance, admission, graduation, retention, gifted students among others [13].

EDM is utilized mostly to measure student performance [14]. Further, machine learning algorithms are used to analyze education data. These algorithms would provide models to identify students at risk and provide early warning alerts for the university staff to take the necessary actions. For example, decision trees (DT), neural networks (NN), and support vector machine (SVM) produced models that predicted student dropouts [15]. Another study that predicted student's dropout based on cumulative GPA was modeled using k-nearest neighbors (KNN) with 87% accuracy [16]. To predict students time of graduation naïve bayes algorithm was used which produced a model with 70.83% accuracy [17]. Decision tree classifier predicted student success or drop out with 60.5% accuracy using socio-demographic variables in addition to major and courses being studied [18]. The same algorithm was used to predict student first year performance in a business informatics by using their high school state exam marks and first year success , the model produced accuracy of 76.65% [19] .

Another study used neural network algorithm to predict student performance in a specific course and was 92.3% accurate [20]. Support vector regression algorithm outperformed multiple linear regression using socioeconomic and university academic information to predict student' academic performance [21].

III. METHODOLOGY

A. Data collection

Imam Abdulrahman Bin Faisal University, in Dammam Saudi Arabia (IAU) has six health colleges: College of Medicine, College of Dentistry, College of Nursing, College of Applied Medical Sciences, College of Public Health, College of Applied Medical Sciences – Jubail. Admission requires high school students to take two standardized tests. One is the Achievement Test (AT) which test their knowledge in Biology, Chemistry, Physics and Math. The other is the General Aptitude Test (GAT) which measures analytical and deductive skills. The equation that determines their admission is 30% high school GPA, 40% AT, and 30% GAT. We collected data from 3468 student records across these colleges using institutional student information system (SIS) from 2012-2019 and data placed in excel sheet. The dataset has seven features and one target variable with four values as shown in table 1.

B. Data preprocessing

One column in excel sheet was created to further formulate the research question. A formula was used to create grade codes for the cumulative GPA based on university policy as shown in table 2. The grade code becomes the class variable with values "A", "B" ,"C" or "D". If class values is "C" or "D" then student need counselling otherwise, no need for counselling.

Table 1. Students related variables

Attributes	Description
Gender	M, F
PREP_GPA	GPA after first year of study
High School_ GPA	GPA upon high school graduation
GAT	General Aptitude Test
AT	Achievement Test
College	A health college
Year	Admission year
Grade Code (target)	Class (A,B,C,D)

Table 2. Grade point average (GPA) scale

Code	GPA out of 5
A	5.00 – 4.75
B	4.74 – 4.00
C	3.99 – 3.00
D	2.99 – 2.00

C. *Classification models*

Several supervised classifiers imbedded in the orange data mining software were used in this study. Each classifier was used to generate a models based on the features and the class value to answer the research question. The classifiers that were used are : KNN, Tree, SVM, Random Forest, Naïve Bays. Training and testing were conducted on the dataset available for this work. The dataset was divided into 66:34 for training and testing, respectively. To avoid overfitting and underfitting, 10-Fold cross validation was used. Fig.1 shows the workflow for academic performance of health science students' dataset.

D. *Model evaluaiton*

To evaluate the best fit model for this study, several metrics were used that include area under the curve (AUC), accuracy, F1, precision and recall. Students who are in need for consultation to enhance their academic performance can be assured by the best model evaluated by these measurements.

IV. RESULTS AND DISCUSSIONS

We evaluated 3468 student data records that covers a seven-year span between 2012-2019 to predict which students are likely to graduate with a "A" , "B", "C" or "D" cumulative GPA. We used five machine learning classifiers, KNN, Decision Tree, SVM, Random Forest, Naïve Bays to create the models. We show that the best performance based on AUC is random forest (.773) followed by naïve bayes (.758), SVM(.686), KNN(.684) and decision tree (.658) as shown in table 3. The confusion matrix for the random forest is another indicator of the usefulness of this algorithm (Fig. 2).

Further, to evaluate which features has more effect on predicting CGPA, three scoring methods were selected. Info. Gain, Gain ration and ReliefF. Preparatory year GPA ,Gender, and high school GPA were the top variables predicating student performance. Interestingly, standardized test did not predict graduating GPA, this is consistent with numerous studies world-wise that are revealing that standardized high school tests do not predict performance at college level [22].

Different research wok in machine learning in predicting students' performance has reported different results. Three studied have shown random forest classifier to outperform other classifiers. One study predicts students' academic performance based on demographics, student previous performance information, course and instructor information, and student general information, random forest was the most appropriate model [23]. A study to predict the right path for engineering students attending preparty year found random forest algorithm as the best fit model for the Mechanical engineering department [24]. Finally, random forest has outperformed other classifiers (decision trees, support vector machines, naive Bayes, bagged trees and boosted trees) in predicting students' final year performance based on information available at the end of the first year with 96% accuracy [25]. These results are consistent with our findings. However, two studies to predict students' GPA showed extreme gradient boosting classifier outperformed random forest [26]. And Naive Bayes scored 83.65%/ accuracy over random forest with 71.15% accuracy to show students' academic achievement at the end of the final year [27].

Information gained through this modeling can be used by universities for early intervention measures to support talented students who are good enough to gain admission yet struggle at the college level. This model is ready to be tested on batch of 2019-2022 once their information is available on the student's information system. Also, the data set can be expanded to include other tracks such Engineering, Sciences and Management, Arts and Education to be part of further studies to test the model.

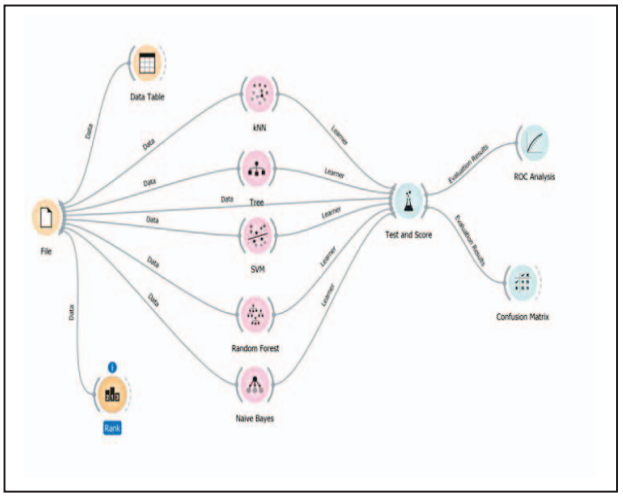

Figure 1. Orange workflow for health science students database

Table 3. Classifier's evaluation matrix

Model	AUC	CA	F1	Precision	Recall
kNN	0.684	0.662	0.639	0.630	0.662
Tree	0.658	0.659	0.654	0.652	0.659
SVM	0.686	0.645	0.604	0.611	0.645
Random Forest	0.773	0.705	0.693	0.690	0.705
Naive Bayes	0.758	0.660	0.661	0.663	0.660

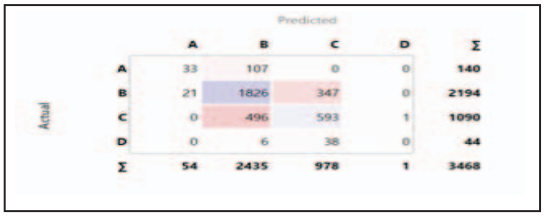

Figure 2. Random forest confusion matrix

V. CONCLUSION

The focus of this research was to apply multiple machine learning classifiers to predict students' performance in health colleges. Their performance is measured upon graduation with a "B", "C", or "D" GPA. Health college students should be motived to graduate with a B or higher GPA as they are excellent students who are good enough to get admission to health colleges but struggling at their university studies. Therefore, the outcome of this research is to serve a guidance to university staff to provide these students with the at most support they deserve.

ACKNOWLEDGMENT

I thank Mr. Yousef Salahat (IAU) for assistance with data collection and Ms. Wadhwa Aldossary for manuscript formatting .

REFERENCES

[1] K. Umamaheswari, S. Niraimathi,"A Study on Student Data Analysis Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 8. August 2013.

[2] E. Osmanbegovic, M. Suljic, "Data Mining Approach for Predicting Student Performance", Journal of Economics and Business, Vol. X, Issue 1. May 2011.

[3] S. Rawat, S. Sreenatha et al., "Ascertaining the Factors Influencing Students' Performance for Engineering Pedagogy" Vol.28,pp. 30-33. January 2015.

[4] M. Suhaimi, N. Abdul-Rahman et al., " Review on Predicting Students' Graduation Time Using Machine Learning Algorithms", International journal of modern education and computer science, Vol. 11,issue 7. July 2019.

[5] G. Dekker, M. Pechenizkiy, M., " Predicting students drop out: A case study", Proceedings of the 2nd International Conference on Educational Data Mining, pp. 41-50, 2009.

[6] Tatar, S. Dilek, " Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average?", Applied sciences., Vol. 10, Issue 14, July 2020.

[7] J. Superby, F. Meskens, " Determination of factors influencing the achievement of the first-year university students using data mining methods", the 8th international conference on intelligent tutoring systems, P.32: 234, 2006

[8] A. Shahiri, H. Wahidah, "A review on predicting student's performance using data mining techniques", Procedia mput Sci.,Vol.72, pp. 414–422, 2015.

[9] G. Zhang, M. Anderson et al., "Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study", Journal of Engineering Education, Vol. 93 (4), pp. 313–320, 2004.

[10] M. Goga, S. Kuyoro et al., "A recommender for improving the student academic performance", Social and Behavioural Sciences, Vol. 180 , pp. 1481 – 1488, 2015.

[11] M. Al-Barrak, M. Al-Razgan, "Predicting students final gpa using decision trees: a case study", International Journal of Information and Education Technology, vol. 6, no. 7, pp. 528, 2016 .

[12] Paris, L. Affecndy et al., "Improving Performance Prediction using Voting technique in data Mining", World Academy of Science, Engineering and Technology, Vol. 38, 2010

[13] A. Nandeshwar, S. Chaudhar, " Enrollment prediction models using data mining. Retrieved July 25, 2021, from http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf, (2009).

[14] D.Kabakchieva, "Predicting Student Performance by using Data Mining methods for classification." ,Cybernetics and Information Technologies, Vol. 13, 2013.

[15] R. Pereira, J. Zambrano, "Application of decision trees for detection of student dropout profiles In Machine Learning and Applications", IEEE International Conference, pp.528-531, IEEE. Decemebr , 2017.

[16] M. Quadri, N. V. Kalyankar. "Drop out feature of student data for academic performance using decision tree techniques." Global Journal of Computer Science and Technology, Vol.10, 2010.

[17] A. Purwinarko, W. Hardyanto, et al., " Academic achievement analysis of Universitas Negeri Semarang students using the naïve bayes classifier algorithm, Journal of Physics: Conference Series, Volume 1918, Mathematics and Its Application,2021.

[18] Z. Kovačić, "Early Prediction of Student Success:Mining Students Enrolment Data", Proceedings of Informing Science & IT Education Conference , 2010.

[19] Edin, and Mirza Suljić. "Data mining approach for predicting student performance." Economic Review 10.1 (2012): 3-12.

[20] N. Stanković1, M. Dragovan et al., "Artificial Neural Network Model for Prediction of Students' Success in Learning Programming", J. Sci.d, Res., Vol 80, March, 2021.

[21] D. Pranav, A. Ravina, et al., "Educational data mining for predicting students' academic performance using machine Learning", Materials today, proceedings, Vol 47, part 15, pp. 5260-5267, June 2021.

[22] E. Allensworth, k. Clark, "Are GPAs an Inconsistent Measure of College Readiness across High Schools? Examining Assumptions about Grades versus Standardized Test Scores", University of Chicago Consortium on School Research, April 2018.

[23] S. Amjad, M al-Emran, K,Shaalan, "Mining Student Information System Records to Predict Students' Academic Performance, March 2019,The International Conference on Advanced Machine Learning Technologies and Applications

[24] M. Ezz, A. Elshenawy, "Adaptive recommendation system using machine learning algorithms for predicting student's best academic program", 0202, Education and Information Technologies, 25:2733–2746

[25] V. Miguéis, A Freitas, P. Garcia, " Early segmentation of students according to their academic performance: A predictive modelling approach ", 2018, Decision Support Systems V. 115, Pages 36 - 51

[26] K. Jaiswal, P. Pathak, V. Pawar," Prediction of degree student achievement analysis",2021, IOP Conf. Ser.: Mater. Sci. Eng. 1070 012057

[27] Asif, R., Merceron, A., Abbas, S., & Ghani, N. (2017). Analyzing undergraduate students ' performance using educational data mining. Computers in Education, 113, 177–194.