Design of a machine learning based model for academic performance prediction

Moustapha Bikienga[1], Ozias Bombiri[2], Emmanuel Sawadogo
{bmoustaph@yahoo.fr[1], ozibombe@hotmail.com[2], sawemma3@gmail.com}

University Norbert ZONGO, bmoustaph@yahoo.fr[1], University Nazi BONI, ozibombe@hotmail.com[2]

Abstract.

Predicting and analyzing the performance of students is essential to design helpful guidance process that allows good success rates and raises the institution's ranking as one of the criteria for a high-quality university. However the lack of adequate support and personalized guidance increases students failure rate. Nowadays, there are many research findings that propose predictive models based on machine learning methods to do many kinds of tasks. Also, machine learning methods have been applied with success in many domains. The aim of this work is to evaluate the possibility of improving the students guidance system by using machine leaning modeling. We have developed a model with objective of predicting the chance of success of students of the Unit of Training and Research in Science and Technology (UFR-ST) of the University Norbert Zongo (UNZ). The approach used in the design of this model was to estimate students success probability when they make their pathway choice among Mathematics, Physics, Chemistry and Computer science after the semester 3. Several Machine learning algorithms (Adaboost, Random Forest, SVM and KNN) were used to fit model with students of academic years 2017-2018 and 2018-2019 achievements data. The results obtained on the test data reveal a score of above 70% for the best algorithm (Random Forest).

Keywords: academic guidance, predicting chance of success, machine learning, Random Forest

1 Introduction

The use of ICT in education provides exceptional performance, better perception and faster learning. In academic pathway choice, student needs special support from guidance specialists in order to get a comprehensive idea of future academic prospects and job opportunities associated with academic selection according to his abilities. In same educational system, there are constraints on choosing pathways like the number of available study pathways,the number of available places and student' hobby. For systems with limited resources, meaning number of students which apply for

JRI 2022, November 24-26, Ouagadougou, Burkina Faso

DOI 10.4108/eai.24-11-2022.2329809

places is higher than the available places, the process is competitive and based on the abilities. An example of this situation is the Burkina CampusFaso in which best guided students are which had the best achievements in baccalaureate exam.

Several researches have been carried out in order to predict academic performance. Among these studies, the review on the prediction of student performance using data mining techniques was conducted by Amirah Mohamed Shahiri et al in [1]. The idea that led to doing this review is the lack of study on the topic. There is not suitable identified methods for predicting student performance and a lack of investigations of factors affecting students achievements. The literature review shows that features frequently used is cumulative grade point average (CGPA) and internal assessment. The reason is that this has a tangible value for future educational and career mobility. This review also reaches that we can actually improve students achievement and success more effectively in an efficient way using educational data mining techniques.

Pojon, Mura on his thesis [2] examines the application of machine learning algorithms to predict whether a student will be successful or not. After applying different machine leaning algorithms such as Linear regression, Decision Tree and Naive Bayes and features engineering processes on students records datasets, the study concluded that it is possible to successfully predict student performance. When ever, features engineering is the big part of the work and that must be safety conducted.

Paulo Cortez and Alice Silva in [3] also address the problem of student achievement in secondary education using business intelligence and data mining techniques. Data used in this study take into account students grades, demographic, social and school features. Carrying out classification models from Decision Tree, Random Forest, Neural Networks and Support Vector Machines they try to predict the core class to which the student is able to success. According to the results achieved, they conclude that more efficient student prediction tools can be developed, improving the quality of education and enhancing school resource management.

F. Okubo et al [4] proposed a model base on neural network (Recurrent Neural Network) for predicting final grades of students. This study confirms that the Recurrent Neural Network approach performed out the early regression methods.

A similar study was carried out at the Norbert Zongo University by Ozias Bombiri et al in [5], where it is question of the chance of success of new baccalaureate holders. With synthesized data they evaluated some machine learning algorithms behavior when modeling guidance system from students grades. Simulation reached that neural network model is the best adapted to the modeling problem.

Daud et al in [6] carried out a study using scholarship holding students data in order to predict whether a student will be able to complete his degree or not. Using learning analytic, discriminative and generative classification models for experiment, results show that this approach can significantly outperforms the existing methods.

Many others researchers addressed the problem of predicting student's performance using machine learning tools [7, 8, 9, 10, 11, 12, 13, 14].

All these studies show stage of interest for predicting academic performance and the considerations to make when using machine learning tools for that. It can be noticed that a diversity of machine learning algorithms is proposed for modeling guidance process. For those which predict

the student performances regression algorithms are prioritized, while classification algorithms are used for predicting if student will success or not.

A recent review on predicting student's performance using machine learning methods made by Yahia Baashar et al [15] raises that achieving accurate predictions is challenging due to the huge among of educational data. This explain the lack of research on exploring different methods and key attributes that influence the student's academic performance. Results of this review show that artificial neural networks, decision trees, Support vector machine, k-nearest neighbor and naive Bayes are the mostly used methods, while demographic, academic, family/personal and internal assessment are the most frequently used features.

In this work, we address guidance of UNZ MPCI students after completing the first three semesters. This orientation process is intended to help students choose an option that matches their abilities. The aim was to find the machine learning algorithm that fits better the available data and to improve statistically the applying of machine learning technologies to solve the prediction of students performance issues. We are interested to give new orientation on the locally research to treats problems we front in our society. The purpose is to design a machine learning based model that can significantly improve the guidance process. The approach is to build models from the student's records with some algorithms and assess their skills. The remaining of this paper is organized as follows. Section 2 is about the data used and the design of the model. Section 3 presents results and discussion. Section 4 comes up with the conclusion of this study.

2 Material and method

In this section we present the data collected and the process of the modeling used in this work.

2.1 Used Data

Machine learning models are performed using data. The learning process consist on fitting model parameters to capture information embedded in a dataset. That why it is necessary to have data before modeling. For our study we collected data on University Norbert ZONGO students. Data available was those of MPCI training pathway student which have begun their first bachelor level at 2017-2018 and 2018-2019 academic years. We used grades for semester 1 to semester 4, each of these semester having a set of study subjects. Grades of semester 1, 2 and 3 are used as the features of the dataset because the choices are done after semester 3. Each one of students that succeeded all these three semesters have to pick one option among Mathematics, Physics, Chemistry and Computer science. The semester 4 grades as the achievement of student. A full study have to take into account the completion of the bachelor degree. But when doing this work, there were no available data at this level. Achievements are labeled in three levels. Level "A" is those passed the normal session, level "B" is those passed the second session and level "C" is those who don't succeed after the two sessions. The tableau 1 gives numbers of students of each study option and the completion status.

Data analyze process was applied to understand the dataset and best prepare it for machine learning modeling. The aim of this work is to design a predictive model that can help student to

Table 1: Data distribution within options

Study options		Student achievement			sum
		A	B	C	
	Mathematics	18	17	35	70
	Physics	68	19	9	96
	Chemistry	19	9	3	31
	Computer science	20	0	0	20
sum		125	45	47	217

choose a study option in which he is more likely to succeed. Therefore we tried to predict the completion level. The figure 1 shows the correlations of features with the semester 4 completion level.

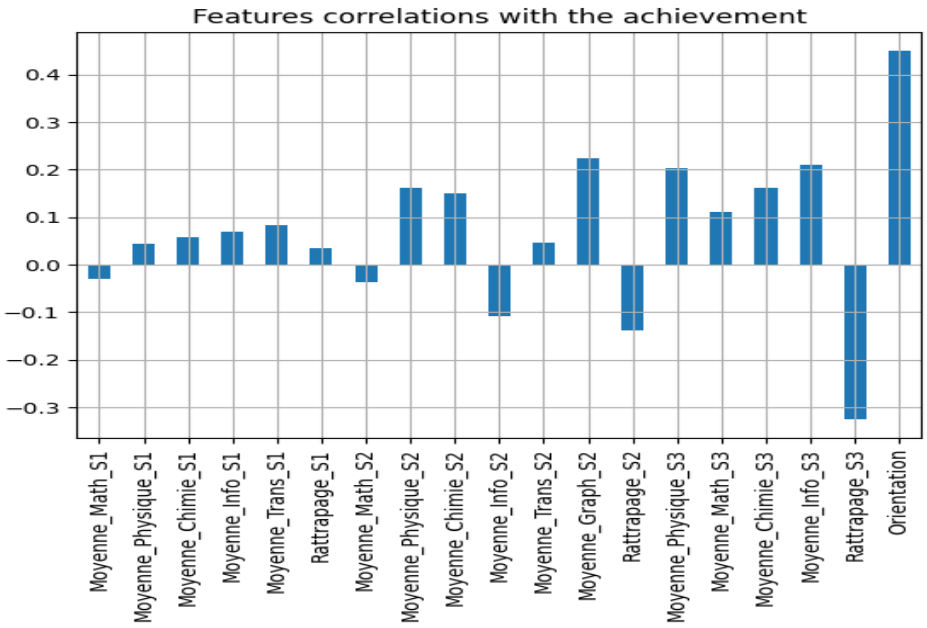

Fig. 1. Features correlations with completion level

As we can see it in the table 1, the dataset is imbalanced. So that we applied data over sampling methods : SMOTE [16] and AdaSyn [17] to balance the distribution on classes within dataset. SMOTE (Synthetic Minority Over-sampling Technique) is a minority class oversampling method that involves creating synthetic examples of the minority class. SMOTE creates synthetic samples for each minority class sample using its k nearest neighbors. ADASYN (adaptive synthetic) is a data

sampling approach for learning from unbalanced datasets. In the sampling technique of ADASYN, the number of synthetic samples to be generated for each minority class sample is determined by the density distribution. Unlike SMOTE, ADASYN automatically calculates the number of synthetic samples. The density distribution is the measure of weights that are assigned to each minority class sample based on their level of learning difficulty. The synthetic sample generation procedure is the same as for SMOTE.

2.2 Design of the guidance system

In this section we present the predictive model of academic guidance based on machine learning. Our study is restricted to the Mathematics-Physics-Chemistry-Computer Science (MPCI) pathway at the University Norbert ZONGO. Based on the student grades for his first three semesters, the system must estimates success rate for each of the four available options (Mathematics, Physic, Chemistry and Computer Science). Therefore, we take into account the marks of all subjects of each semester. The core component of the system machine learning classification model that predicts the label class of achievement. The choice is varied so that at the output we have an achievement level for each available option. Then the student is free to make his orientation regarding the output of the system.

Experimental process was performed by trying multiple algorithms among those which are saying to be adapted to the problem according to the literature review.

For each of these, we have trained and assess models in order to get bests hyper-parameters. Tuning algorithm hyper-parameters was conducted across an implementation of GridSearch. We used some machine learning metrics to evaluate the skill of the model.

Metrics used was accuracy, recall, precision and f1_score. The accuracy is the rate of correct predictions among all predictions. Recall or sensitivity is the rate of true positives, i.e. the proportion of positives that are correctly identified. It is for example the capacity of a model to detect all software defects (in terms of probability it is the probability that a present defect is detected). Precision is the proportion of correct predictions among the points that we have predicted to be positive. It is for example the capacity of a model to indicate a module as defective only when it is really defective (the degree of confidence that one can grant to the predictions of the model).

3 Results and discussion

3.1 Results

After implementing the experimental process, we got results that we describe in this section. The table 2 gives the means accuracy, the recall and the precision after five training and testing for each algorithm we used. As we can see it the best performances are obtained with the Random Forest algorithm.

Table 2: Results of models skill measuring

Algorithm	Accuracy	Recall	Precision	F1_score
AdaBoost	0.669	0.671	0.694	0.676
K-nearest neighbors	0.665	0.664	0.698	0.692
Random Forest	0.715	0.715	0.731	0.723
Support Vector Machines	0.655	0.653	0.677	0.666

3.2 Discussion

Results described in the above section are promising for the design of machine learning model that can do students guidance according to theirs performances. When observing learning curves during the model training process, we found that accuracy gets better from one step to the next and do not stabilize at the end. The figures 2 3 give examples of learning curve of the random forest and AdaBoost algorithm based models respectively. This fact shows that dataset size is not sufficient for the algorithm to learn data structure. So that, more data are needed to perform the study. But there is not other available data at the study moment. Alternatively we applied over-sampling methods to improve results.

In this study we used only the academic achievements as features. These are directly linked to the student performance and easily accessible. There are many others features which can be use to perform academic performance predictive model such as attendance, the age, the social environment parameters, the financial situation, etc. The main difficulty relative on using these is their collecting process.

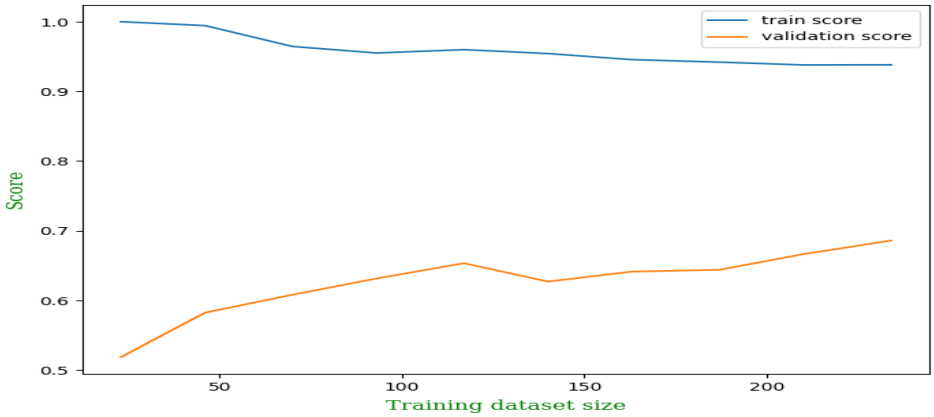

Fig. 2. Learning curve of random forest based model

Fig. 3. Learning curve of AdaBoost based model

4 Conclusion

The use of machine learning is rapidly expanding and is also being extended to the education community for the prediction of academic performance. We have proposed a model of predicting academic performance in order to help Norbert Zongo University MPCI training pathway bachelor students making their training option choice at semester 4. Testing the ability of some machine learning algorithms, we have achieved satisfactory results. This study can lead machine learning system engineering practitioners to design a real model that will be used to help student guidance process. Such a study must be done with more data.

References

[1] Shahiri AM, Husain W, Rashid NA. A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science. 2015 Jan;72:414-22.

[2] Pojon M. Using Machine Learning to Predict Student Performance. 2017. Accepted: 2017-06-26T10:22:44Z.

[3] Cortez P, Silva AMG. Using data mining to predict secondary school student performance. EUROSIS-ETI; 2008. Accepted: 2008-08-20T18:31:05Z.

[4] Okubo F, Yamashita T, Shimada A, Ogata H. A neural network approach for students' performance prediction. In: Proceedings of the Seventh International Learning Analytics & Knowledge Conference. LAK '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 598-9.

[5] Bombiri O, Ouedraogo TF, Some P, Poda P. Towards a Smart Guidance System in CAMPUS-FASO : Simulation Results; 2022. .

[6] Daud A, Aljohani NR, Abbasi RA, Lytras MD, Abbas F, Alowibdi JS. Predicting Student Performance Using Advanced Learning Analytics. In: Proceedings of the 26th International Conference on World Wide Web Companion. WWW '17 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee; 2017. p. 415–421.

[7] Yadav SK, Pal S. Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv:12033832. 2012.

[8] Mueen A, Zafar B, Manzoor U. Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. International journal of modern education & computer science. 2016;8(11).

[9] Angeline DMD, et al. Association rule generation for student performance analysis using apriori algorithm. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA). 2013;1(1):12-6.

[10] Simsek A, Balaban J. Learning strategies of successful and unsuccessful university students. Contemporary Educational Technology. 2010;1(1):36-45.

[11] Abu-Naser SS, Zaqout IS, Abu Ghosh M, Atallah RR, Alajrami E. Predicting student performance using artificial neural network: In the faculty of engineering and information technology. 2015.

[12] Mondal A, Mukherjee J. An Approach to predict a student's academic performance using Recurrent Neural Network (RNN). Int J Comput Appl. 2018;181(6):1-5.

[13] Arunachalam A, Velmurugan T. Analyzing student performance using evolutionary artificial neural network algorithm. International Journal of Engineering & Technology. 2018;7(2.26):67-73.

[14] Aydoğdu Ş. Predicting student final performance using artificial neural networks in online learning environments. Education and Information Technologies. 2020;25(3):1913-27.

[15] Baashar Y, Alkawsi G, Ali N, Alhussian H, Bahbouh HT. Predicting student's performance using machine learning methods: A systematic literature review. In: 2021 International Conference on Computer and Information Sciences (ICCOINS); 2021. p. 357-62.

[16] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. Springer; 2005. p. 878-87.

[17] Barua S, Islam M, Murase K, et al. A novel synthetic minority oversampling technique for imbalanced data set learning. In: International Conference on Neural Information Processing. Springer; 2011. p. 735-44.