7th International Conference on Higher Education Advances (HEAd'21)
Universitat Politècnica de València, València, 2021
DOI: http://dx.doi.org/10.4995/HEAd21.2021.13196

Recent Advances in Academic Performance Analysis

Linlin Zhang, Kin Fun Li, Imen Bourguiba
Department of Electrical and Computer Engineering, University of Victoria, Canada.

Abstract

Academic performance analysis has gained popularity in the past decade. Using various prediction and classification methods, researchers aim to provide clues to help students to improve their performance, and to assist educational institutions to improve quality and make better administrative decisions. This work provides a brief survey of 56 papers related to academic performance prediction, published in 2019 and 2020. Statistics and analysis on the prediction target categories, the target population size, prediction and classification methodologies used, and evaluation metrics are presented. It is found that the most commonly used techniques are decision tree, ensemble methods, and neural networks. Futhermore, these techniques also give the highest accuracy in their target prediction.

Keywords: *Academic performance; prediction; classification; student population; performance evaluation.*

1. Introduction and Motivation

Academic performance analysis has become an important research area, to predict student performance and to identify at-risk students. Proper analysis not only helps students to improve their academic performance and to prevent failure, but also helps educational institutions to improve quality and make better administrative decisions. In this work, a brief survey on academic performance related literature published in 2019 and 2020 is given.

From IEEE Xplore, 56 papers related to student performance prediction and educational data mining are collected (available here). Seven papers are published in journals while the remaining 49 are presented at conferences. Figure 1 shows the countries where the institutions are located. The top three referenced countries have the highest world population, and logically associated with a large number of universities (India has 4004, USA has 3281 and China has 2310). In addition to the ones shown, a single instutition is referenced in 18 individual papers.

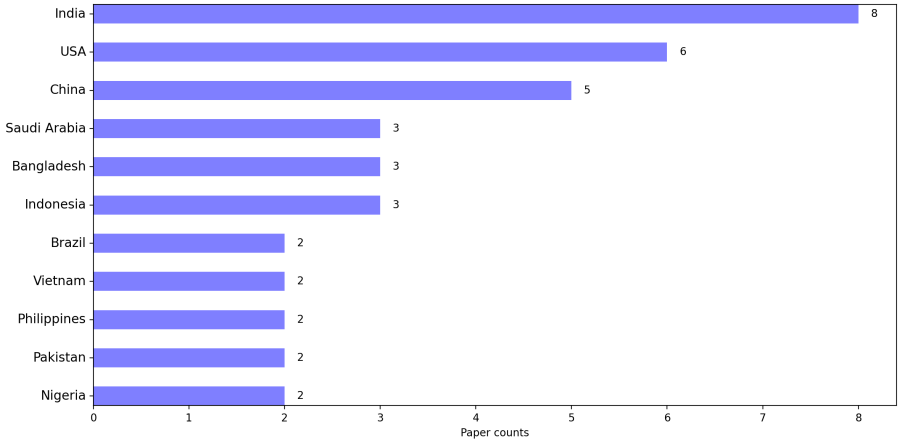

Figure 1. Paper count in countries of institution location.

For the 56 papers, the analysis targets are discussed in Section 2, while the prediction methodologies used and performance evaluation are presented in Section 3.

2. Analysis targets

2.1. Target population

The analysis target statistics are shown in Figure 2. The "information technology related" group includes computer science, electrical and computer engineering, information system and technology, and software. There are two papers, Nakagawa *et al.* (2019) and Zhang *et*

al. (2020) use the same online open-source datasets, the ASSISTments 2009-2010 "skill-builder" provided by the online educational service ASSISTments, and Bridge to Algebra 2006-2007 used in the Educational Data Mining Challenge of KDDCup.

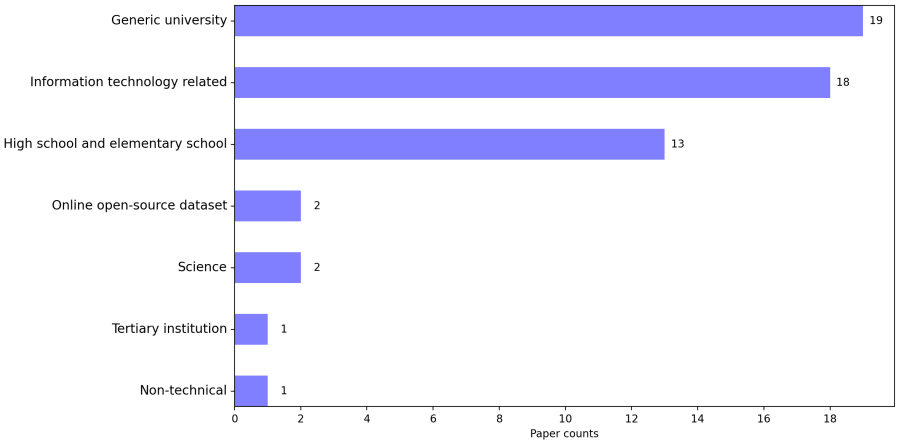

Figure 2. Analysis target and paper count.

About 73% of the papers target higher education, while 23% focus on high school and elementary school students. Many papers investigate students from information technology related departments. This seems natural as authors from these departments probably are researchers in data mining and analytics, and therefore are applying their expertise to analyze students within their department.

2.2. Target population size

Table 1 shows the target population size of the 56 papers. There are two papers that collect information from over 50,000 students. These investigations are conducted by the same authors, Mai *et al.* (2019a, 2019b), and the data are collected from 2009 till 2019 from the entire university.

Typically, population size from a department is much smaller, with 62.5% of the papers having a population size under 1000 (35 out of 56). Three papers have target population size less than 100. Khan *et al.* (2019) study 50 students from the department of system and networking and only focus on predicting a student's final grade in a course. Lai *et al.* (2019) investigate 55 students and present a model to identify students who are likely to fail in reading. This is an interesting project as the students are provided with a science text to read. All participants wear a mobile headset to record electrocardiogram-based attention, an eye tracker to track eye movements, and a mobile webcam to capture facial behavioral cues. Olalekan *et al.* (2020) gather statistics on 44 students from a department in a tertiary institution. The number of papers in each target population category and the corresponding

average population size are shown in Table 2. The impact of target population size on the prediction model's performance is addressed in Section 3.2.

Table 1. Target population size and corresponding count of papers.

Target population size	Paper counts
unknown	10
size <= 200	9
200 < size <= 400	9
400 < size <= 600	10
600 < size <= 1000	7
1000 < size <= 2,000	5
2,000 < size <= 5,000	3
5,000 < size <= 10,000	1
10,000 < size <= 50,000	0
50,000 < size <= 100,000	2

Table 2. Paper count and average size in each target population category.

Target population category	Paper count	Average population size
Generic university	19	9,133
Tertiary institution	1	44
High school and elementary school	18	1,379
Information technology related	13	569
Non-technical	1	145
Science	2	269
Online open-source dataset	2	2,000

3. Prediction methodologies and evaluation

3.1. Prediction and classification techniques

In academic analysis, the major goal is to predict a certain outcome or to classify an instance as pass/failure, grade point average (GPA), graduated or not, etc. Figure 3 shows the prediction and classification techniques used in the surveyed papers. Most works employ more than one techniques with decision tree, ensemble methods, and neural networks being the most frequently used.

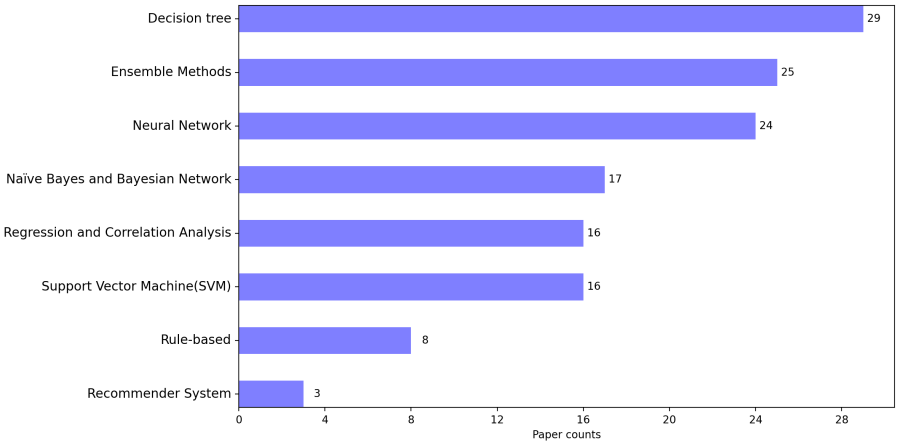

Figure 3. Prediction and classification techniques used in surveyed papers.

3.2. Prediction performance analysis

The most commonly used metric to evaluate prediction performance is accuracy found in 40 papers, followed by ROC (receiver operating characteristic) curve and MSE (mean squared error) each in 3 papers, and RMSE (root MSE), MAE (mean absolute error), and f-score each in one paper. Seven papers have not made any evaluation. The highest accuracy attained by their prediction of the 49 papers is shown in Figure 4. Four papers, Alsalman *et al*. (2019), Akram *et al.* (2019), Jayaprakash *et al.* (2019) and El-Rady (2020), obtain over 95% accuracy, while Islam *et al.* (2019) achieve only 64%.

From the correlation line in Figure 4, it can be observed that as the target population size increases, the prediction accuracy improves too. Another reason for some of the high accuracy obtained is the prediction and classification methods used.

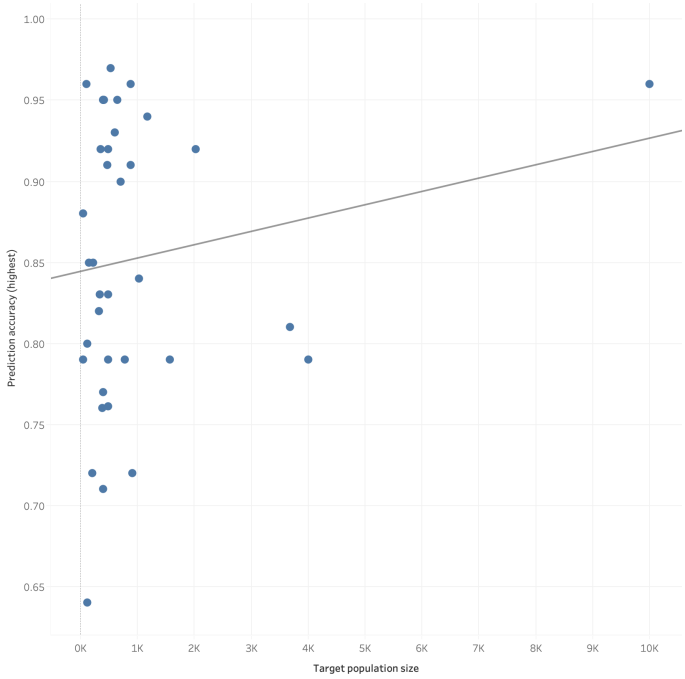

Figure 4. Accuracy measure and target population size.

On the other hand, small population size seems to be the reason for low accuracy. For instance, the paper by Islam et al. (2019) with 64% accuracy in predicting grade, has only 120 students as samples. Small dataset might have bias or easily lead to overfitting. In addition, the prediction methods deployed in this paper is logistic regression, which is one of the least accurate among all prediction methods, as illustrated in Table 3, which shows the accuracy of the prediction methods.

Table 3. Accuracy of prediction methods.

Prediction method	Mean of accuracy	Standard deviation
Decision Tree	89%	6%
Ensemble Methods	86%	9%
Neural Network	84%	8%
SVM	83%	6%
Naïve Bayes and Bayesian Network	83%	17%
Regression and Correlation Analysis	71%	9%

As shown in Table 3, the most accurate prediction method is decision tree, while it is also the most frequently used prediction method (see Figure 3). Interestingly, ensemble methods

(random forest) and neural network (multi-layer percentron) are the second and third most commonly deployed prediction methods, and also the second and third most accurate, respectively, as noted in Section 3.1.

4. Conclusion and Future works

This brief survey provides a glimpse of the academic prediction work in the literature in 2019 and 2020. Decision tree, ensemble methods and neural networks are the most commonly implemented prediction techniques, while also achieving top performance in accuracy.

We are currently working on a comprehensive survey that covers academic performance literature in the 2010s. Investigation includes data preprocessing, feature selection, and other influencial factors on prediction performance.

References

Akram, A., Fu, C., Li, Y., Javed, M. Y., Lin, R., Jiang, Y., & Tang, Y. (2019). Predicting Students' Academic Procrastination in Blended Learning Course Using Homework Submission Data. IEEE Access, 7, 102487–102498. doi:10.1109/access.2019.2930867

Alsalman, Y. S., Khamees Abu Halemah, N., AlNagi, E. S., & Salameh, W. (2019). Using Decision Tree & Artificial Neural Network to Predict Students Academic Performance. 2019 *10th Int Conf on Information & Communication Systems* (ICICS). doi:10.1109/iacs.2019.8809106

El-Rady, A. A. (2020). An Ontological Model to Predict Dropout Students Using Machine Learning Techniques. 2020 *3rd Int Conf on Computer Applications & Information Security* (ICCAIS). doi:10.1109/iccais48893.2020.9096743

Islam, R., Sazid, M. T., Mahmud, S. R., Ferdous, C. N., Reza, R., & Hossain, S. A. (2019). Parametric Study of Student Learning in IT Using Data Mining to Improve Academic Performance. 2019 Joint 8th Int Conf on Informatics, Electronics & Vision (ICIEV) & 2019 *3rd Int Conf on Imaging, Vision & Pattern Recognition* (icIVPR). doi:10.1109/iciev.2019.8858551

Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020). Predicting Students Academic Performance using an Improved R&om Forest Classifier. 2020 *Int Conf on Emerging Smart Computing & Informatics* (ESCI). doi:10.1109/esci48226.2020.9167547

Khan, I., Al Sadiri, A., Ahmad, A. R., & Jabeur, N. (2019). Tracking Student Performance in Introductory Programming by Means of Machine Learning. 2019 *4th MEC Int Conf on Big Data & Smart City* (ICBDSC). doi:10.1109/icbdsc.2019.8645608

Lai, S., Liu, J., Niu, B., Tian, H., & Wu, F. (2019). Combining Facial Behavioral Cues, Eye Movements & EEG-Based Attention to Improve Prediction of Reading Failure. 2019 *Int Joint Conf on Information, Media & Engineering* (IJCIME). doi:10.1109/ijcime49369.2019.00103

Mai, T. L., Do, P. T., Chung, M. T., & Thoai, N. (2019a). An Apache Spark-Based Platform for Predicting the Performance of Undergraduate Students. 2019 IEEE *21st International*

Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE *5th International Conference on Data Science and Systems* (HPCC/SmartCity/DSS). doi:10.1109/hpcc/smartcity/dss.2019.00041

Mai, T. L., Do, P. T., Chung, M. T., Le, V. T., & Thoai, N. (2019b). Adapting The Score Prediction to Characteristics of Undergraduate Student Data. 2019 *International Conference on Advanced Computing and Applications* (ACOMP). doi:10.1109/acomp.2019.00018

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019). Graph-based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network. 2019 *International Conference on Electrical and Computing Technologies and Applications* (ICECTA).

Olalekan, A. M., Egwuche, O. S., & Olatunji, S. O. (2020). Performance Evaluation Of Machine Learning Techniques For Prediction Of Graduating Students In Tertiary Institution. 2020 *Int Conf in Mathematics, Computer Engineering & Computer Science* (ICMCECS). doi:10.1109/icmcecs47690.2020.240888

Zhang, J., Mo, Y., Chen, C., & He, X. (2020). Neural Attentive Knowledge Tracing Model for Student Performance Prediction. 2020 IEEE *International Conference on Knowledge Graph* (ICKG). doi:10.1109/icbk50248.2020.00096