Application of Machine Learning Methods to Predict Student Performance: A Systematic Literature Review

Akpofure Avwersuoghene Unghwure1, Mercy Ebitiminipre Ogbise2

1Lecturer, Nigeria Maritime University, Okerenkoko, Delta State, Nigeria
2Nigeria Maritime University, Okerenkoko, Delta State, Nigeria

ABSTRACT: In recent times, the need for the application of machine learning in the educational frontier has become crucial. Most educational administrators and researchers are using various machine-learning methods to improve students' retention, predict students' performance, mitigate students' dropout rate as well as position them to make better decisions in curriculum design and admission policy in their learning citadel.

This study focused on the application of machine learning techniques to predict students' performance in recent times. With a systematic approach, the research identified the existing prediction methods and tools used to predict students' performance, observed the type of variables considered by the researchers in this research area. With the use of Boolean keyword searches in various journal databases and filters, eighty (80) research papers related to students' performance prediction using machine learning methods were selected and considered.

Weka was observed to be the most used tool in the reviewed articles. Most of the datasets used in the reviewed articles were on university students with the authors focusing on the performance of the students' full program. A great number of the authors applied at least one supervised learning methods in their work and reviewed articles from countries with low literacy rate. The results from the various researches will help to improve the academics of students in those countries and monitor the students' performance which would also improve their literacy rate.

Keywords: Machine Learning, Student Performance, Education

1.0 Introduction

In recent times, machine learning (ML) practices have been a big deal in various industries in the world including the educational frontier. The need to automate different tasks such as grading students, improving student retention, testing students, predicting student performance, as well as administrative tasks like material optimizations in the academic facet, has called for the application of machine learning techniques and methods. Arguably, the most important task in every learning institution is to monitor and improve their students' performance. Early prediction of student performance in the right fashion will improve student retention as well as the testing methods used for the students. This practice will also aid the educators and education policymakers by giving them better information about their students' learning ability as well as how best they can help students who are lagging in a given set.

This study focuses on reviewing previous research works on building models to predict student's performance in a learning environment in the last ten (10) years. The authors of this article developed a systematic approach to the review work. This approach is to support the objectives of this study, which are:

1. To identify the existing prediction methods and the tools employed for predicting students performance
2. To study and identify the variable type used for the predictive process.
3. To identify and study the researchers who employed these learning models to analyze students performance

In this paper, the study is structured as follows. In Section 2, we reviewed previous reviewed literature related to students’ performance. Section 3, We performed the review process by using the systematic approached design for this course. Section 4, we presented our results as well as discussed the insights we gather from the results. Section 5, concludes the review as well as state our limitations and future study in this research area.
2.0 Literature Review

The need to predicting student academic performance has become a critical factor in improving the quality of academic curriculum; assist the students while they study as well as providing the tutors more options when training their students. In recent times, there are many works been published related to this subject matter. We found several literature reviews that examined student academic performance modelling in different views. The review of Kavipriya focuses on the students’ performance prediction, analysis, early alert, and evaluation by using different methods of data mining. He reviewed different classifications methods like decision tree, naïve Bayes algorithms etc. He suggested since predicting student’s performance is difficult due to a lot of challenges like data imbalances, there is a need to employ support vector machine that offer the best accuracy in his study [1]. Shahiria A.M. et. al. took a lot at the various data mining techniques implemented in different student’s performance prediction models. They evaluated how these predicting algorithms can be used to identify the most important attributes in students’ database [2]. In this work, they pointed out that cumulative grade point average (CGPA) and internal assessment as the attributes most researchers used to build their model. A review on predictive modelling technique for student academic performance monitoring by A. Mat et. al. focuses on various learning tasks employed when building the predictive models as well as the application of these models [3]. After accessing several researches, they noted that these models can be applied in various fronts like course recommendations, and career path planning. Mishra T. et.al surveyed Students’ performance and employability prediction through Data mining. In this work, they focused mainly on traditional educational set-up [4]. Their findings were most predictors used in traditional educational set-up were demographic and socioeconomic factors. They also mentioned that a few research works have been done on employability prediction. Zaffar M. et. al studied different prediction models for students who enrolled for programming subjects only [5]. Their results were only a few researchers used Filter Feature selection algorithms in the pre-processing stage. The attendance of students, their performance in mathematics, and physics play an important role in influencing their performance.

3.0 Method

3.1 Research questions

The main goal of this systematic literature review is to explore the applications of machine learning methods to educational data to predict/monitor/intervene in student academic performance. In a bid to reach this above goal, we have stated the following main research questions:

RQ1. What machine learning methods and tools were applied to predict student performance?

RQ2. Who are the authors, journals, and the variable type used in this research area?

RQ3. What are the courses/subjects used to access the students’ performance?

To answer the research questions, the required information was retrieved from the chosen literature by both reviewing the contents of the articles and by utilizing the meta-data properties of the article's citation. The authors of this article have classified the chosen articles concerning the machine learning algorithms used, the input data type used, journal article type, stakeholders involved in the studies.

3.2 Literature Collection Process

We follow a standard systematic literature review methodology that involves the use of specific queries search built in a Boolean format on various selected bibliographical database and a set of rules for filtering the result of the queries. The authors of the paper queried seven (7) databases that are related to computer, education, engineering, and general science. To make certain only relevant literature are captured, distinctive search keywords were introduced in this process. Logical operators like “+” and “&” were used to combine these distinctive keywords during the search. The combination of these keywords and logical operators resulted in Boolean search keywords like “Machine Learning” + “Student”, “Predicting Student Performance” which were used to perform the searches on Google Scholar and Science Direct. Date filters were used to filter search results to exclude articles that were published earlier than “2010”.
3.2.1 Location of affiliations

In an attempt to see how scientific contributions have been made towards this front, we classified the research articles according to the location of the corresponding authors. This provided insights on the global publication developments in this research area.

3.2.2 Type of journal source and dataset

The authors of this study attempted to see which journal database published articles within the research area as well as the dataset used in the published work. Given this is an educational research area, did the scholars worked with undergraduate or high school students?

3.2.3 Type of tools/techniques used

The authors of this study examined the tools and techniques employed for this prediction modelling. Currently in the analytics facet, several tools are available to perform various analyses. Tools like WEKA, SPSS, MATLAB, Orange, R, Python are used for data mining, machine learning as well as model deployment.

The machine learning techniques used in the reviewed studies can be categorized into three groups.

- Supervised Algorithms – Regression (Linear and Logistic), Decision Trees (C4.5, ID3, CHAID), support vector machines, naïve Bayes
- Unsupervised Algorithms – Clustering, Correlation
- Semi-Supervised Algorithms- Association Rules and the combination of any supervised and unsupervised algorithms.

4.0 Result and Discussion

This section displays the discoveries from the literature review process that was explained in the earlier sections. The dataset consisted of 98 electronic materials (research articles, review articles, thesis as well as e-books on machine learning and predictive models). The authors of this paper only considered research and review articles on applying machine learning methods to predict academic performance. The selected literature collection consists of 80 articles. The line chart below displayed the reviewed articles in this study by their year of publication.

![Figure 1: A Pie Chart displaying the distribution of reviewed articles by the journal database.](image-url)
Figure 2 presents the distributions of the selected articles from different journal databases for this study. It shows that the research gate leads the collection with 24 articles followed by Science Direct, IEEE Xplore Digital Library, and Google Scholar with 20, 16, and 11 respectively. Semantic Scholar, ACM Digital Library, Hindawi, and Springer were sources for eight (8) articles. One article journal source is unknown [23].

![Figure 2: A Pie Chart displaying the distribution of reviewed articles by the journal database.](image)

Most of the data used for predicting students’ performance in the collection of articles reviewed in this study were University Data. Seventy-one (71) articles used university data with sixty-seven (67) focused on undergraduate students, three (3) articles were on both undergraduate and postgraduate while only one on postgraduate students [24]. Nine (9) articles focused on high school students. This illustrated in figure 3.

![Figure 3: A pie chart of the distribution of dataset type in the reviewed articles.](image)

Out of the selected articles, 49 articles used at least one tool with WEKA as the most-used analytical tool in the collection of reviewed articles. The other 31 articles did not clearly state the tools they employed for their analysis. SPSS, RapidMiner, Matlab were also used in more than three research papers. Figure 4 shows the list of all the tools used with its frequency of use.

![Figure 4: A bar chart displaying the number of articles by tool used](image)
The researchers of the selected papers employed at least one machine learning method in their respective works. There are 61 selected works applied supervised machine-learning algorithms. Seventeen (17) studies applied semi-supervised (a combination of supervised and unsupervised algorithms) and two unsupervised as presented in table 1.

Table 1: A Presentation of the Reviewed Articles by machine learning method used

<table>
<thead>
<tr>
<th>S/N</th>
<th>Machine Learning Method Type</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Supervised</td>
<td>61</td>
</tr>
<tr>
<td>2</td>
<td>Semi-Supervised</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>Unsupervised</td>
<td>2</td>
</tr>
</tbody>
</table>

In an attempt to improve the predictive model, some researchers implemented k-fold cross-validation (CV) in their model. K-fold CV occurs when a dataset is randomly divided into a K number of subsets where each subset is used as a testing data and the remaining subsets are used as the training set. This process is done for K times [12] [14] [18]. The final result is estimated as the average error rate on test examples [21]. Twenty-six (26) articles within the selected studies applied cross-validation to their model at some point [22]. Most of them used 10-fold cross-validation method [13][15][16][17]. C. Lye et.al introduced the cross-validation concept into their study to evaluate the best predictive models in terms of accuracy [19][20]. The Figure below as well as table 2 shows the distribution of the selected papers by their use of cross-validation in their research work.

Figure 5: A Pie Chart showing the distribution of the reviewed articles by the application of cross-validation methods.

Table 2: A Presentation of the Reviewed Articles by the use of cross validation techniques

<table>
<thead>
<tr>
<th>S/N</th>
<th>The use of Cross Validation</th>
<th>Reference Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>[8][12][13][14][15][16][17][20][21][22][28][41][46][47][52][59][60][65] [67][72][77][79][82][83]</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>[6][7][9][10][11][18][19][23][24][25][26][27][29][30][31][32][33][34][35][36][37] [38][39][40][42][43][44][45][48][49][50][51][53][54][57][58][61][62] [63][64][66][68][69][70][71][73][74][75][76][77][78][80][81][84][85]</td>
</tr>
</tbody>
</table>

Looking at the authors of the collected papers in this study, there are some insights the authors of this paper unravel. Using the nationality of the corresponding author and the source of the dataset used to build the models, thirteen (13) of the studies
were carried out in India, Nine (9) in Pakistan, Six (6) in Nigeria and five (5) in China, Malaysia, and the United States [25] as shown below. Given the top three countries who published the highest articles in this study have a relative high illiterate rate (India: 28.80%, Pakistan: 42.10%, Nigeria: 40.40%) to other countries hence, one can deduce that these researchers developed these models in a bid to reduce the illiteracy rate in their countries [86].

Figure 6: A Heat Map showing the number of articles reviewed by the corresponding author’s nationality

M.Zaffar et.al in their study on predicting models for students focused only on students enrolled in programming courses. The authors of this article accessed the selected papers by the courses used for predicting student performance in a bid to see which courses most of the researchers used to build their model. In this study, one hundred and thirty-eight (138) courses were captured and most of the researches used full students’ programs (various courses) to build their predictive models. Courses like “Mathematics”, “English”, “Social Science” were the top three single course used in these studies. A clear picture is shown in the word cloud map below:

Figure 7: A Word Cloud Map showing the courses used in the various reviewed articles.

When considering the performance of any student, one will first evaluate their academic work and their responses to such. All the studies in this field consider academic input of the student dataset they accessed. Features like cumulative grade point average (CGPA), Assignment/Test score, final course mark, activity score, university entry mode, learning style, course attendance are all a function of the student academic involvement. Besides academic involvement factors, other factors also influence students’ performance. K. Alaslani and M. Alandejani (2020) identified some social factors like the use of the social networking sites affect student performance [31] [88]. Hostel Friendship [30], interacting with peers [30], Romantic Relationship [17], Smoking Habits [55]. Socio-economic factors like parents’ earning power and student’s job employment
status [58], number of vehicles [35], student's ability to make money . Other researchers measured students' performance with the teacher's contribution, qualification as well as competence to the course work [70] [16] [77]. The figure below shows the distributions of the reviewed articles in this study by the variable type used in their analysis (Student learning factor, social-economic factor, student's social life factor, and educators inputs).

Figure 8: A Presentation for the reviewed article showing their distribution among the variable type considered in this study.

Conclusion

The art of predicting student performance is highly beneficial to everyone especially to the educational administrators and students. The administrators can evaluate their teaching procedures to ensure their students get the best-fit learning plan from them. Students can also appraise their learning patterns to know the tailor-made learning activities that will meet their needs. The authors of this paper have accessed previous studies on predicting students' performance using machine learning methods. Most of the authors of the reviewed articles are from countries with low literacy rate, hence, this will help academics in those countries to monitor their student performance thereby improving their literacy rate. Weka is the most-used analytical tool used by the researchers when predicting students' performance. When considering the machine learning methods used, the supervised learning method was mostly employed in this research space. Most researchers attempted to predict the performance of students using the students' academic data, social-economic data, and educator competence data as well. This approach proves that a student academic performance does not only depend on his academic input. A full student's academic program is used to predict the performance of students in most of the reviewed work.

References


36. A. Ali, U. Ali, “Predictability of engineering students’ performance at the University of Engineering and Technology, Peshawar from admission test conducted by educational testing and evaluation agency (ETEA), NWFP, Pakistan” World Conference on Educational Sciences 2010 pp 976-982 DOI:10.1016/j.bspro.2010.03.137


