Research Article

Linear Support Vector Machines for Prediction of Student Performance in School-Based Education

Nalindren Naicker ¹, Timothy Adeliyi,² and Jeanette Wing²

¹ICT and Society Research Group, Department of Information Systems, Durban University of Technology, Durban 4001, South Africa

²ICT and Society Research Group, Department of Information Technology, Durban University of Technology, Durban 4001, South Africa

Correspondence should be addressed to Nalindren Naicker; nalindrenn@dut.ac.za

Received 6 June 2020; Revised 30 August 2020; Accepted 16 September 2020; Published 5 October 2020

Academic Editor: Mahmoud Mesbah

Copyright © 2020 Nalindren Naicker et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Educational Data Mining (EDM) is a rich research field in computer science. Tools and techniques in EDM are useful to predict student performance which gives practitioners useful insights to develop appropriate intervention strategies to improve pass rates and increase retention. The performance of the state-of-the-art machine learning classifiers is very much dependent on the task at hand. Investigating support vector machines has been used extensively in classification problems; however, the extant of literature shows a gap in the application of linear support vector machines as a predictor of student performance. The aim of this study was to compare the performance of linear support vector machines with the performance of the state-of-the-art classical machine learning algorithms in order to determine the algorithm that would improve prediction of student performance. In this quantitative study, an experimental research design was used. Experiments were set up using feature selection on a publicly available dataset of 1000 alpha-numeric student records. Linear support vector machines benchmarked with ten categorical machine learning algorithms showed superior performance in predicting student performance. The results of this research showed that features like race, gender, and lunch influence performance in mathematics whilst access to lunch was the primary factor which influences reading and writing performance.

1. Introduction

Predicting the factors impacting student performance early in the academic programme can assist in combating the high dropout rate experienced by higher education institutions. Predicting factors can guide lecturers adjusting their lessons in order to assist students who are at risk of failing. Several studies have been conducted on using machine learning algorithms for early prediction of student performance. These studies fall into the field of Educational Data Mining (EDM) which provides great value to educational institutions [1]. EDM is an emerging field in data mining and lies in the intersection between education, computer science, and statistics [1]. Not only is Educational Data Mining limited to education but also it covers fields like transportation, sports, marketing, and sales [2].

There are several data mining tools used in EDM to analyse and predict student performance to the benefit of educational institutions. These interventions improved pass rates, curbed dropout rates, and increased retention rates [3]. There are several tools that are used in EDM. Data manipulation and feature engineering tools include Microsoft Excel, EDM Workbench, Python and Jupyter notebook, and Structured Query Language (SQL). No one tool can be used for EDM as different tools suit different tasks [4].

A wide range of classification algorithms can be used to predict processes and performance, namely, random forest, support vector machines, AdaBoost, decision tree, Naïve Bayes, and K-nearest neighbour [3]. Kumar et al. [2] used EDM effectively in research to improve retention rates by predicting slow learners' in a high school class and providing

them with intervention to improve. They found Naïve Bayes, Multilayer Perceptron, SMO, J48, and REPTree to be the most widely used techniques in EDM research. This paper will explore other techniques in the field of EDM.

This research paper will attempt to answer the following research question in the field of EDM:

What are the strong impacting factors for school based learners' performance in reading, writing, and mathematics?

The paper is organized as follows:

In the next section, a review of related literature is presented, Section 3 introduces the materials and methods used in this research, followed by Section 4 that provides the applied data mining algorithms on the selected dataset and presentation of experimental results and, finally, the paper ends with the section that concludes the work.

2. Related Works

Acharya and Sinha in [5] showed, in a study of students majoring in computer science, the best results for early prediction was obtained with the decision tree class of algorithms. In this study, the training consisted of 309 instances whilst the testing comprised 104 instances. The study found that caste and religion played a vital role in student performance in India. Other strong factors related to academic performance were the family size and family income.

In another study by Koutina and Kermanidis in [6], early prediction of postgraduate master's students was useful to assist tutors. Data on a total of 117 instances were collected from three courses, namely, Advanced Language Technology, Computer Networks, and Information Systems Management. Naïve Bayes and 1-NN achieved the best prediction results compared to well-known classification algorithms. The results show that, in small datasets, Naïve Bayes and 1-NN can perform better than decision trees [6]. Support for better performance on Naïve Bayes on smaller datasets was also shown in a study by Belachew and Gobena in [7]. The study found that the size of the dataset and the imbalance and distribution of class values are the main challenges in work obtaining better accuracies. In this study, predictors of student performance were the presence in class and in-term performance. It was also found that student's occupation, the type of degree, and their possession of another master's degree were not good indicators for early prediction of student performance.

In a study by Kotsiantis et al. [8] on the performance of distance learning, students showed that Naïve Bayes was able to produce more than satisfactory accuracy when compared to other state-of-the-art algorithms. Two experiments were conducted using datasets of 354 and 28 instances. The results showed that 28 instances were too few and the number of instances needed for better accuracy was 70 instances.

El Aissaoui et al. [9] implemented an adaptive e-learning system to take into account learning styles so convenient content can be provided to enhance learning. This study

automatically detected learning styles using machine learning. K-means clustering was used to extract learning styles from login sequences. Thereafter, using a supervised learning algorithm, namely, Naïve Bayes, the learning style for a new learner could be predicted. A real dataset extracted from the e-learning system's log file was used in the experiments.

In a study, Xu et al. [10] used machine learning for tracking student performance in a degree programme. Students' future performance in degree programmes based on ongoing information could be predicted. The following predictors were cited in the study, namely, backgrounds of the students, the courses that were selected by students, the information provided by the courses to predict future performance, and the progress made by the students. This study asserts that predicting students' performance in degree programmes is not a one-off task but it is a progressive process of tracking and updating. They developed a novel algorithm to predict students' performance based on a bilayered structure comprising a base predictor layer and an ensemble predictor layer. Base predictors are trained using offline data (past performance) and ensemble predictors are trained using online student data (current performance). The dataset had 1169 anonymous undergraduate students. The data contained student's precollege information (high school GPA and SAT), courses that students take each quarter, the course credits, and the obtained grades. The study concluded that the proposed method achieved superior performance over benchmarked approaches. Mao et al. in [11] used Bayesian Knowledge Tracing and Long Short-Term Memory (LSTM) to predict posttest scores and learning gains prediction. Two training datasets namely, Pyrenees and Cordillera were employed. The findings showed that the learning environment can predict students' performance and learning gains.

Cui et al. in [12] investigated student performance prediction by assessing the possibility of a common set of student activity variables that predict performance, the machine learning classifiers that perform well across different courses model that could be used to predict performance based on LMS data. There were 18 features extracted from data processing on the LMS activity. The recursive feature elimination (RFE) algorithm was employed to select the most important features. The features were ranked and the top 5 features were selected. In the experiment, 8 classical machine learning algorithms were used, namely, logistic regression (LR), Naïve Bayes (NB), neural network (NN), support vector machine (SVM), decision tree (DT), k -nearest neighbours (kNN), random forest (RF), Gradient boosting machine (GBM), and 1 ensemble classifier, namely, ensemble model (EM). The results showed that the usability of the LMS could predict student performance and can provide feedback to support students and higher education institutions.

Gray and Perkins in [13] showed that students "At Risk" can be identified as early as week three of the semester with approximately 95% accuracy. The dataset consisted of 4970 students, 32 features, and 5 classes, namely, pass, supplementary, repeat year, fail, and repeat semester. Zohair in [14] explored identifying key features using a small dataset using

clustering algorithms. Records of 50 students were broken up into 2 small datasets. The classifiers K -nearest neighbour (KNN), Naïve Bayes (NB), LDA, support vector machines (SVM), and MLP-ANN were used for the training. A comparative analysis showed which algorithm was superior when dealing with prediction using small datasets. The study concluded that support vector machines produced the most acceptable results on small datasets.

Linear support vector machine (LSVM) is a variant of support vector machine (SVM) which is one of the most popular supervised machine learning methods. It has been proven in the literature that LSVM can guarantee global optimization for regression or classification problems in small-to-large datasets [15, 16] and deals with predictive binary classification, that is, the assignment of class labels to unlabeled data [17]. Based on the benefits LSVM proffers, this paper adopted LSVM for the statistical analysis of student performance.

The aim of this study is to compare the performance of linear support vector machines with state-of-the-art classical machine learning algorithms in order to determine the algorithm that would improve prediction of student performance.

3. Materials and Methods

The experiments were conducted on a computer running Windows 10 operating system with the configuration of Intel® Core™ i5-8250U CPU @ 1.60 GHz (8 CPUs), 1.8 GHz, 8 GB RAM memory, and 500 gigabytes hard disk drive. The dataset used in this study was selected from a set of categorical data obtained from a public domain referred to as the Kaggle data and available online (<https://www.kaggle.com/spscientist/students-performance-in-exams/activity>). The database consists of 1000 high school learner records based on student performance in mathematics, reading, and writing. No preprocessing of the data was required. The features of the dataset include the following: (1) gender, (2) race/ethnicity, (3) parental level of education, (4) access to lunch, (5) test preparation, (6) mathematics score, (7) reading score, and (8) writing score.

A machine learning approach was adopted to establish the contributing factors that impact student performance. The training dataset comprised of all 1000 student data and thereafter the testing dataset comprised of 1000 student data for the Kaggle data. In the training stage, the classification rules were adopted. The testing stage is used to test the accuracy of the classification rules [5]. All results are provided for the testing of the dataset. The machine learning library in MATLAB R2020a was used to implement algorithms and obtain statistical results.

Figure 1 shows the architecture of the student performance testing used in this study.

3.1. Classification. A popular machine learning algorithm, the linear support vector machine (LSVM) classifier, was used as a supervised learning model to analyse data for classification. SVM is one of the most actively developed

FIGURE 1: The architecture of the student performance testing model using learning algorithms.

classifications and regression methodologies in data mining and machine learning. It provides salient properties such as the margin maximization and nonlinear classification via kernel tricks and has proven to be effective in many real-world applications [18].

According to Foody and Mathur in [19] a linear support vector machine is often able to classify a dataset to a higher accuracy than conventional statistical classifiers. Based on the advantage that the linear support vector machine classifier algorithm provides, we used the algorithm to classify and predict factors that impact student performance [20, 21]. The linear SVM classifier was benchmarked with ten other algorithms such as coarse decision tree, medium decision tree, fine decision tree, logistic regression, Gaussian Naive Bayes, Kernel Naive Bayes, quadratic SVM, cubic SVM, fine Gaussian SVM, and medium Gaussian SVM.

3.2. Performance Measures. The standard performance metrics, namely, accuracy, total misclassification cost (TMC), prediction speed (PS), training time (TT), and area under Receiver Operating Characteristics (ROC) were calculated in the experimental comparison of classifiers [22]. A parallel coordinate plot was used to visualise the multivariate data. A parallel coordinate plot maps each row in the data table as a line, or profile, and allows for the visualisation of data points across many dimensions.

4. Results and Discussion

The LSVM was used to analyse the performance by extracting useful knowledge from the student dataset. The usefulness of the algorithm was to interpret relationships among variables and to determine factors that affect student performances in their mathematics, reading, and writing score. Table 1 presents the frequency and percentage of all variables for a student population of $n = 1000$.

The summary data from Table 1 was trained, using linear SVM 5-fold cross-validation, a preprocessing operation for classification. The trained data was analysed using linear SVM (LSVM) and benchmarked with fine decision tree (FDT), coarse decision tree (CDT), medium decision tree (MDT), logistic regression (LR), Gaussian Naive Bayes (GNB), Kernel Naive Bayes (KNB), quadratic SVM (QSVM), cubic SVM (CSVM), and fine Gaussian SVM (FGSVM) and medium Gaussian SVM (MGSVM). Table 2 shows the results for all the benchmarked algorithms for the area under ROC curve (AUC), accuracy, total misclassification cost (TMC), prediction speed (PS), and training time (TT).

Table 2 shows that LSVM has the highest accuracy of 90.1% and appositely the lowest total misclassification cost of 99. According to time metrics, LSVM was outperformed by other classifiers. GNB boasts the best performance for prediction speed while CDT has the fastest training time.

TABLE 1: Description of student performance dataset ($n = 1000$).

Factors	Categories	Frequency	Percentage (%)
Gender	Female	518	51.8
	Male	482	48.2
Race/ethnicity	Group A	89	8.9
	Group B	190	19.2
	Group C	319	31.9
	Group D	262	26.2
	Group E	140	14.0
Parental level of education	Some college	226	22.6
	Some high school	179	17.9
	High school	196	19.6
	Bachelor's degree	118	11.8
	Master's degree	59	5.9
	Associate's degree	222	22.2
Lunch	Standard	645	64.5
	Free/reduced	355	35.5
Test preparation	None	642	64.2
	Completed	358	35.8
Math score	0–10	2	0.2
	11–20	2	0.2
	21–30	12	1.2
	31–40	34	3.5
	41–50	100	10.0
	51–60	189	18.9
	61–70	270	27.0
	71–80	215	21.5
	81–90	126	12.6
	91–100	50	5.0
Reading score	0–10	0	0
	11–20	1	0.1
	21–30	7	0.7
	31–40	19	1.9
	41–50	70	7.0
	51–60	178	17.8
	61–70	238	23.8
	71–80	252	25.2
	81–90	173	17.3
	91–100	62	6.2
Writing score	0–10	1	0.1
	11–20	2	0.2
	21–30	7	0.7
	31–40	25	2.5
	41–50	89	8.9
	51–60	177	17.7
	61–70	243	24.3
	71–80	248	24.8
	81–90	140	14.0
	91–100	68	6.8

TABLE 2: The linear SVM results versus 10 benchmarked algorithms.

	FDT	CDT	MDT	LR	GNB	KNB	QSVM	CSVM	FGSVN	MG SVM	LSVM
Accuracy (%)	83.7	75.2	82.4	89.4	68.4	67.8	90	85.7	70.2	89.5	90.1
TMC	163	248	176	—	316	322	100	143	298	105	99
PS (obs/sec)	49000	52000	54000	21000	41000	11000	54000	46000	42000	47000	54000
TT (sec)	0.936	0.213	0.220	1.465	0.385	0.735	0.598	1.336	0.400	0.306	0.509

4.1. *Strong Predictors for Mathematics.* Figure 2 illustrates the parallel coordinate plot with five dimensions represented by $N = 5$ (race/ethnicity, parental level of education, lunch,

test preparation, and math score) such that every vertical line, that is, factors impacting the mathematics score, appears exactly once. The standard deviation is 65% shown as

point 0 on the parallel coordinate plot. The standard deviation from 0 to 2 gives scores greater than or equal to 65% while the standard deviation between -1 and 0 gives scores greater than or equal to 52% but less than 65%; furthermore, the standard deviation between -4 and -1 gives scores between 0% and 49%.

In Figure 2, the topmost plot for mathematics score shows that a standard deviation greater than 0 represents scores greater than or equal to 65% and enables us to predict factors that impacted student performance for the mathematics score. The five-dimension vertical lines predict that the high level of education of parents does not necessarily impact student performance as a student with a parent with some college or high school degree performed better than a student whose parent has an associate's or master's degree.

The parallel coordinate plot in Figure 2 shows that another important factor that affected student performance in math was the type of lunch. It was observed that students with a standard lunch performed better than students with a free or reduced form of lunch. It was also observed that the test preparation did not affect student performance for a student who scored more than 65% in math. The bottom row in Figure 2 with a standard deviation of less than -1 shows a math score lower than 50%. It was identified that a student with free/reduced lunch had a bad performance in math and it was further observed that students that fall under race group c and are females also had a poor performance in math while the level of parents' education was not an important factor.

4.2. Strong Predictors for Reading. In Figure 3, the topmost plot for reading score shows a standard deviation greater than 0 represents reading scores greater than or equal to 65%, which enables us to predict factors that impact student performance for reading score.

In Figure 3, the five-dimension vertical lines predict that the high level of education of parents still does not impact student performance as a student with a parent with some college or high school degree performed better than a student whose parent has an associate's or master's degree. However, contrary to the observation of the math score, the reading observation gave a different result. Females in group C race/ethnicity performed better in reading than math; furthermore, the parent level of education is not an important factor in student performance. The bottom row in Figure 3 with a standard deviation of less than -1 shows a reading score lower than 50%. It was identified that the free/reduced lunch factor has a significant impact in affecting the performance of students. Furthermore, females in group E race/ethnicity recorded the lowest reading scores.

4.3. Strong Predictors for Writing. Figure 4 shows the parallel coordinates with five dimensions represented by $N=5$ vertical lines impacting the writing score.

In Figure 4, the topmost plot for writing score further reiterates that parental level of education does not impact student performance as a student with a parent with some college or high school degree performed better than a student whose parent has an associate or masters degree. Females in group C race/ethnicity performed better in writing than math;

FIGURE 2: Parallel coordinates with five dimensions represented by $N=5$ vertical lines impacting the mathematics score.

FIGURE 3: Parallel coordinates with five dimensions represented by $N=5$ vertical lines impacting the reading score.

furthermore, the parent level of education is not an important factor in student performance. The bottom row in Figure 4 with a standard deviation of less than -1 shows a reading score lower than 50%. It was identified that the free/reduced lunch factor has a significant impact on affecting the writing score of students. Furthermore, females in group E race/ethnicity recorded the lowest reading scores.

4.4. Receiver Operator Characteristic (ROC). Receiver Operator Characteristic (ROC) curve is a graphical plot used to

TABLE 3: Summary of AUC values.

	FDT	CDT	MDT	LR	GNB	KNB	QSVM	CSVM	FGSVN	MG SVM	LSVM
AUC	0.88	0.80	0.89	0.97	0.74	0.73	0.96	0.94	0.78	0.96	0.97

FIGURE 4: Parallel coordinates with five dimensions represented by $N = 5$ vertical lines impacting the writing score.

FIGURE 5: ROC for the linear SVM versus benchmarked algorithms.

show the diagnostic ability of binary classifiers. The AUC provides a trade-off between true positives and false positives [3]. Table 3 shows the summary of results of AUC values for the linear SVM versus the ten benchmarked algorithms.

Figure 5 shows the ROC curve for the linear SVM compared to the ten benchmarked algorithms used in this study.

The results depicted in Figure 5 show that linear SVM outperformed other algorithms iterating the accuracy of the linear SVM in classifying and predicting factors that impact student performance.

5. Conclusions

The results show that, in predicting students' performance, linear support vector machines showed superior performance when applied to student data. The algorithm predicted that the parental level of education does not influence students' performance whilst impactors such as race, gender, and lunch have a bearing on student performance. In future work, we will consider using an ensemble of classical machine learning algorithms to boost accuracy in the prediction of student performance.

Data Availability

The data is available publicly at <https://www.kaggle.com/spscientist/students-performance-in-exams/activity>.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The authors acknowledge the Durban University of Technology for making funding opportunities and materials for experiments available for this research project.

References

- [1] A. V. Manjarres, L. G. M. Sandoval, and M. S. Suárez, "Data mining techniques applied in educational environments: literature review," *Digital Education Review*, no. 33, pp. 235–266, 2018.
- [2] M. Kumar, S. Shambhu, and P. Aggarwal, "Recognition of slow learners using classification data mining techniques," *Imperial Journal of Interdisciplinary Research*, vol. 2, no. 12, pp. 741–747, 2016.
- [3] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and R. Najoua, "Educational data mining and analysis of students' academic performance using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, pp. 447–459, 2018.
- [4] S. Slater, S. Joksimović, V. Kovanovic, R. Baker, and D. Gasevic, "Tools for educational data mining: a review," *Journal of Educational and Behavioral Statistics*, vol. 42, no. 1, pp. 85–106, 2017.
- [5] A. Acharya and D. Sinha, "Early prediction of students' performance using machine learning techniques," *International Journal of Computer Applications*, vol. 107, no. 1, pp. 37–43, 2014.

- [6] M. Koutina and K. L. Kermanidis, "Predicting postgraduate students' performance using machine learning techniques," in *Proceedings of the International Conference on Artificial Intelligence Applications and Innovations*, pp. 159–168, Springer, Corfu, Greece, September 2011.
- [7] E. B. Belachew and F. A. Gobena, "Student performance prediction model using machine learning approach: the case of Wolkite university," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 2, pp. 46–50, 2017.
- [8] S. Kotsiantis, C. Pierrakeas, and P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.
- [9] O. El Aissaoui, Y. E. M. El Alami, L. Oughdir, and Y. El Alloui, "A hybrid machine learning approach to predict learning styles in adaptive E-learning system," in *Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development*, vol. 5, pp. 772–786, Springer, Tetouan, Morocco, January 2019.
- [10] J. Xu, K. H. Moon, and M. Van Der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742–753, 2017.
- [11] Y. Mao, C. Lin, and M. Chi, "Deep learning vs. Bayesian knowledge tracing: student models for interventions," *Journal of Educational Data Mining*, vol. 10, no. 2, 2018.
- [12] Y. Cui, F. Chen, and A. Shiri, "Scale up predictive models for early detection of at-risk students: a feasibility study," *Information and Learning Sciences*, vol. 121, no. 3-4, pp. 97–116, 2020.
- [13] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," *Computers & Education*, vol. 131, pp. 22–32, 2019.
- [14] L. M. A. Zohair, "Prediction of student's performance by modelling small dataset size," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 27, 2019.
- [15] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *European Journal of Operational Research*, vol. 265, no. 3, pp. 993–1004, 2018.
- [16] W.-J. Niu, Z.-K. Feng, B.-F. Feng, Y.-W. Min, C.-T. Cheng, and J.-Z. Zhou, "Comparison of multiple linear regression, artificial neural network, extreme learning machine, and support vector machine in deriving operation rule of hydropower reservoir," *Water*, vol. 11, no. 1, p. 88, 2019.
- [17] M. Somvanshi and P. Chavan, "A review of machine learning techniques using decision tree and support vector machine," in *Proceedings of the 2016 International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1–7, IEEE, Pune, India, August 2016.
- [18] J. Vaidya, H. Yu, and X. Jiang, "Privacy-preserving SVM classification," *Knowledge and Information Systems*, vol. 14, no. 2, pp. 161–178, 2007.
- [19] G. M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1335–1343, 2004.
- [20] J. Huang, J. Lu, and C. X. Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy," in *Proceedings of the Third IEEE International Conference on Data Mining*, pp. 553–556, IEEE, Melbourne, FL, USA, November 2003.
- [21] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [22] H. Ferreira, "Machine learning metrics," 2019, <https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-othermetrics-in-machine-learning-894688cb1c0a>.