Student Performance Prediction in Learners Centric Approach with Machine Learning

Harsh Golay[1], *Mudita* Sharan[2], *and Preety* Shoran[3]
[1]Department of Computer Science, Christ University, Bangalore, India
[2]Department of Computer Science, Christ University, Bangalore, India
[3]Department of Computer Science, Christ University, Bangalore, India

ABSTRACT- Predicting student performance helps educators locate students who are at risk and tailor appropriate and timely interventions for those students. This research proposes a learner-centered machine learning framework to integrate demographic, academic and behavioral features in order to predict student grade performance. The dataset consists of 2392 students and 15 attributes including age, gender, parental education, study time, absences, and extracurricular activities. Four supervised learning models - Logistic Regression, Decision Tree, Random Forest and Support Vector Machines (SVM) were trained and measured using 70:30 stratified split. The performance of the model was evaluated using accuracy, precision, recall, and F1-score metrics. Among these, Decision Tree classifier achieved the highest accuracy (92.48%) which was followed by Random Forest (88.31%), SVM (83.51%) and Logistic Regression (75.16%). The results show that such factors as study time, absences, and parental involvement were the most predictive. The proposed learner-centered approach shows that the combination of contextual, behavioral, and academic data can greatly increase the predictive accuracy and the interpretability of the data, facilitating early risk detection and intervention in education.

Keywords: Education data mining; Learner centric methodology; Decision tree; Random forest; SVM; Student performance prediction..

1. Introduction

In recent years, Learning Analytics and Educational Data Mining (EDM) have become important areas of research in enhancing student academic outcomes using data and data-driven insights. Educational institutions gather large amounts of data related to student demographics, behavior, attendance and performance. However, traditional assessment methods -- such as final exams and periodic evaluations -- give us a very limited degree of view in how students are doing, and do not reflect the cognitive behavior of continuous learning and factors that play into the context.

With the advent of Machine Learning (ML) it is now possible to analyze complex educational datasets and find hidden patterns in those data to predict student success or risk of failure in the future. Predictive analytics helps educators to plan out early interventions and personalized learning strategies that will lower dropout rates and aid engagement.
Previous studies have dealt with multiple ML models such as Logistic Regression, Decision Trees, Random Forest, Neural Networks etc for academic prediction. Studies as Waheed et al. (2020) and Li et al. (2020) proved the effectiveness of deep learning for behavior-based prediction while others (Alsariera et al., 2022; Yagci, 2022) put the emphasis on the tree-based and ensemble methods which are superior in interpretability and performance. However, many of the existing works largely focus on accuracy and typically overlook learner-centered and behavioral and contextual aspects, and do not provide an interpretable information that can be used to inform educational interventions.
The research gap therefore exists in the creation of a learner-centric and interpretable ML framework that not only offers high predictive accuracy, but also highlights the academic, behavioral and parental variables that have a high impact on performance. Unlike previous research that focuses on algorithmic optimization alone, this study is focused on the integration of demographic, academic, and behavioral characteristics in a combined predictive model.
The objective of this research is to be able to predict the level of student performance and identify those students who are at risk of poor performance at the early stages using four supervised ML algorithms including Logistic Regression, Decision Tree Algorithm, Random Forest Algorithm and Support Vector Machine (SVM) by applying these techniques to a data set of 2,392 number of students with 15 attributes. The dataset contains variables like age, gender, study time, absences, extracurricular participation and education of parents.
The structure of the proposed work (illustrated in Figure 1) is modeled after the structured ML pipeline in which the overall structure comprises four main phases:
Data Preparation and Preprocessing - cleaning, encoding, and standardizing data;
Exploratory Data Analysis (EDA) - detecting trends, finding correlations;

Model Development and Training - performing four classifiers using ML (using a train-test split of 70 to 30 percent); Performance Evaluation- to judge the performance of the model in terms of accuracy, precision, recall, and F1 score. Among all the models, the Decision Tree classifier showed the best performance with an accuracy of 92.48%, which makes it useful in modeling non-linear relationships and is interpretable.

This work is important because it emphasizes that by incorporating behaviour and contextual characteristics in a learn-centric machine learning paradigm for prediction, interpretability is increased and actionable insights for educators can be provided such as personalized intervention and early on risk detection..

2. Literature Review

New trends in Educational Data Mining (EDM) and learning analytics have been using more models of Machine Learning (ML) and Deep Learning (DL) models for predicting student outcomes. Very Large Datasets, Algorithms, and Learner Centric approach features have been considered in order to improve the prediction accuracy, interpretability and fairness.
The proposed study was written by Waheed et al. who used deep artificial neural networks (ANNs) on Virtual Learning Environment (VLE) click streams data in order to detect

vulnerable students **[1]**.

Their findings indicated that they are more accurate (84.93) than logistics regression and the support vehicle machines which considered the relevance of deep models in depicting learning behavior. On the same note, the Li et al proposed a Sequential Prediction Deep Network (SPDN) that combined the online learning records with patterns of access to campus internet **[2]**.

They found that overall behavior of using the internet was most likely to correlate with performance in school accompanied by their experiment of 505 students demonstrating that multi-source behaviora The field has been viewed in a wider perspective by systematic reviews. A review of the studies in ML-based prediction by Sekeeroglu et al. revealed that studies are at risk of severe gaps including inconsistent validation methodologies, and over reliance on small data sets as well as absence of deep models **[3]**.

The intelligent approaches were also studied by Namoun and Alshanqiti, with the prevailing importance of the supervised ML and the common use of the assessment grades, yet the importance of the assessment outcomes-based prediction and comprehensive learning-learner modeling **[4]**.

A number of empirical studies have made a comparison of classical ML algorithms in the case of algorithm predicting performance. Alsareria et al. compared decision trees, random forests, SVM, logistic regression, and Naivete bayes and it was found that ANNs were the most accurate on several institutional data sets **[5]**.

Similarly, Yağcui stated that tree-based and ensemble models, three of them RF, KNN, and SVM, with midterm and demographic predictor reached high accuracy 70- 75 percent on predicting the final exam grades **[6]**.

Alongside algorithmic comparisons, there are reviews that examine the use of technology of integration in the outcome of learning. In a systematic review of 100 studies on educational technology, Valverde-Berrocoso et al. present mixed evidence concerning ICT as an element that influences performance in mathematics and science, with bigger effects being observed when there is a pedagogical match **[7]**.

The recent studies have shifted on the inclusion of socioeconomic, behavioral, and psychological aspects. In one of the studies on Pennsylvania schools, Chen and Ding discovered that neural networks slightly outgenred classical ML models concerning the forecast of socioeconomic data outcomes **[8]**.

Wen and Juan applied a case of a deep neural sequence model made up of activity logs of online-based students and reached a high prediction accuracy (more than 80 percent) in the early prediction tasks **[9]**.

Aslam et al. also focused on the approach that is human-centered, and added the personality characteristics to academic features, showing better results in terms of accuracy and fairness with gradient boosting and SHAP to explain **[10]**.

At the same time, scientists have designed refined ML models. Ahmed used clustering and classification approaches and hyperparameter optimization where SVM obtained an accuracy that was of approximately 96 percent on institutional data **[11]**.

Mai et al. presented a belief rule base of dual-level progressive classification with explainability limitations (DLBRB-i), todeal with the class imbalance, retain the interpretability **[12]**.

It was suggested by Luo et al. that would use a spatio-temporal feature integration model comprised of demographic, temporal, and regional indicators to improve the predictability of programming courses **[13]**.

Of special [promise] has been time-series and sequential learning strategies. In OULAD, Shou et al. applied multidimensional time- series analysis using multi-head self- attention and LSTMs, getting 74 percent accuracy of the four classes and 99 percent accuracy of the

early-risk analysis **[14]**.
Ying and Ma reformulated prediction as generative adversarial network (GAN)-based synthetic data augmentation as a regression task, and found improvements of up to 21.9 on small datasets **[15]**.

On the whole, the reviewed literature suggests that, in contrast to classical ML, deep learning and user-friendly feature combination (through behavioral logs, demographics, personality, and contextual data) appeal to a much higher predictive accuracy, as well as, to interpretability. Nevertheless, there are still loopholes in unifying assessment, bringing equity and justification of models in varying institutions. This informs the interest of the present study to an ML learner-centric framework that incorporates multidimensional learner data in order to enhance more accurate performance prediction outcomes.

3. Methodology

The research approach chosen in the context of the study is a developed machine learning pipeline that has four primary steps data preparation, preprocessing, exploratory data analysis, and model development.

A. Dataset
The dataset in this research has 2,392 records of students and 15 features. The attributes are dealing with various aspects of student performance which include;
Demographic: Gender, age, ethnicity.
Academic:Parental Education, study time, Absences, Tutoring. Behavior: Sports, Music, Extracurricular, Volunteering.
Results: Grade Class and Grade Point Average (GPA).
The Grade Class was the object of study that fell as a target variable and was divided into five different performance scores coded as numeric classes:
The target variable Grade Class consisted of five classes, namely, Class 0 (severely underperforming), Class 1 (underperforming), Class 2 (average), Class 3 (good/above average), and Class 4 (excellent performance).This system of classification allowed representing the problem as a multi-class classification task.

B. Data Preprocessing
The samples were devoid of missing. Pre processing involved: Converting discrete variables by Gender and Ethnicity to numerical form.
Standardization of continuous variables (e.g., Study Time, Absences) where necessary within the sample to facilitate consistency upon scaling.
Checking on the disseminating of classes as a way of determining possible imbalance especially in lower grade classes (0.0 and 1.0).

C. Dissemination Approaches
Data Journals and Software-Based Sharing.
The descriptive statistics offered the information on the trends in student performance. The mean GPA was 1.90 (on a scale of 0–4). Students were averagely spending 9.7 hours per week studying with an average 14.5-absence per year. Nearly 38 percent of the students reported to engage in extracurricular activities. This was demonstrated in EDA that suggested the existence of correlations between performance levels and academic/behavioral factors.

D. Model Development

The four types of machine learning classifiers were used in order to recognise Grade Class: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine classifiers. To maintain the balance of classes, the dataset sampling was split to divide the dataset into the training (70) and testing (30). Accuracy, Precision, Recall, and F1-Score were used to determine if model performance was good. There was moreover the generation of the confusion matrices to determine the performance each class and a comparative analysis was also performed to determine the best classifier.

Model Evaluation and Performance Metrics

In order to determine how reliable is the prediction models are, the efficiency of each classifier were evaluated with respect to commonly agreed upon evaluation measures Accuracy, Precision, Recall and F1-Score. the dataset was multi-class classification, these measures were calculated on the basis of each of the class and the results were summarized using weighted of averages to handle the imbalance between the classes. Besides that, more information on misclassifications per individual classes was sought through the examination of confusion matrices.

1. Accuracy

Accuracy or correctness of the instances is a measure of the proportion of correctly recognized instances to the total number of instances. Although the most intuitive one, it can also be decidedly misleading when the data are not balanced, as there is a potential that a model can comply with a high level of accuracy, for example, by targeting the most common group..

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy was taken as the major comparison measure among models in this study. The Decision Tree produced the best accuracy rate of 92.48% with random forest generating next price rate of 88.31, the SVM ranked third with accuracy rate standing at 83.51 and Logistic regression ranked last with the accuracy rate of 75.16.

2. Precision

Precision measures the result number of correctly predicted when positive, of the total number of cases that the prediction algorithm believes are positive. It answers the question: “Upon predicting the model to rely on a student as part of a specific grade group, what is the frequency of the correctness of the prediction?

$$Precision = \frac{TP}{TP + FP}$$

Precision is high so that there are fewer false positives. As an example, both Decision Tree and the Random Forest were very precise on all the classes, whilst the other, the Logistic Regression had low precision on the minority classes.

3. Recall

Recall, or sensitivity, or true positive rate is used to clarify the capacity of the classifier not to miss all pertinent cases. It responds to the following inquiry; among all the students who voted in practice in a grade class, what percentage of them was hit when the model was used?

$$Recall = \frac{TP}{TP + FN}$$

Such a metric is especially valuable when it comes to education prediction problems, in which one ought to identify at-risk (low-performing groups) students.
critical. The recall of minority classes was better in both Decision Tree and Random Forest than in in this study.
Logistic Regression and SVM which over-fitted a lot of low- GPA students.

4. F1-Score
F1-Score is the harmonic mean of Precision and Recall; it involves the equal measure where the false positives and the fake negatives are both costly. The F1-Score is more impartial to the majority class as opposed to accuracy.

$$F1\ Score = 2 \times \frac{Precision\ .\ Recall}{Precision + \ Recall}$$

The Decision Tree in this research, however, shows the highest F1- Score (0.92) which denotes a good balance between precision and recall. Random Forest had the highest score of 0.87, and Logistic Regression and the SVM had low scores of
0.73 and 0.65 key since they performed badly across minority classes.

Visualization of Training

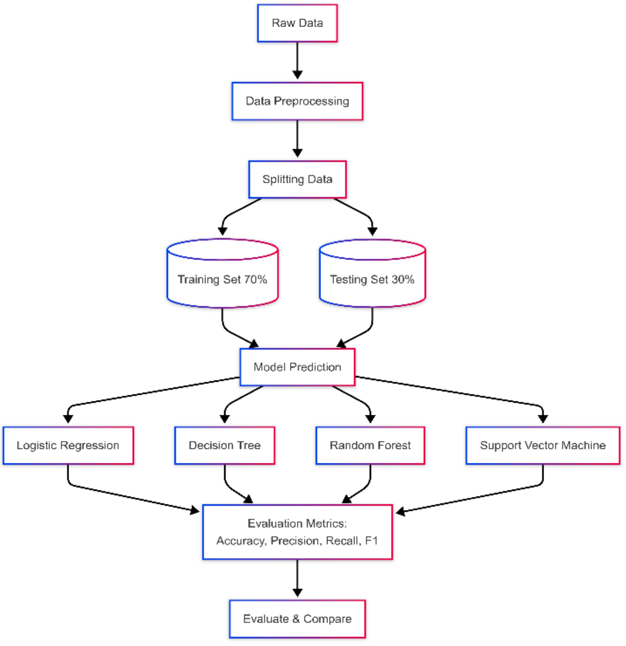

Fig. 1. Visualization of Training

Fig. 1. Visualization of Training shows Raw data forms the basis of the training process, and they are preprocessed to address the containing of categorical encoding and feature preparation. It is then divided into 70 and 30 percent training and testing respectively. The models that are trained on the training set, and then tested on the testing set are instantly called four supervised machine learning models namely, Logistic Regression, Decision Tree, Random Forest and Support Vector machine (SVM) models. Accuracy, Precision,

Recall, and F1-score measures are used to examine model performance and results are compared to find out the best
classifier.

4. Results And Discussion

Table I provides a summary about the four classification models performance. The Decision Tree classifier was best at 92.48 shared with the highest level of accuracy, whereas the low-accuracy was offered by the Logistic Regression (75.16).
The accuracy of the Logistic Regression was found to be 75.16 with the weighted precision, recall, and F1-score of 0.74, 0.75 and 0.73 respectively. gave good performance in most of the classes (Class 4, F1 = 0.94), but weakly in minority classes (Class 0 where the precision of the model is 0.40 and the F1 is 0.15).
Decision Tree had the best overall performance and accuracy of 92.48% and a weighted precision, recall and F1-score of 0.93,
0.92 and 0.92, respectively. It was also found to be equally robust in various classes (macro F1 = 0.87), which is why it is deemed the most appropriate model to use in this study to classify into more than two classes.
Random Forest has achieved the best accuracy of 87.06, weighted precision, recall and F-1-score 0.88, 0.87, and 0.85, respectively. The model had highly predicted the Class 2 (F1 = 0.84), and Class 4 (F1 = 0.96) which meant that it was a highly able predictor up to the point that it could only slightly perform as compared to the Decision Tree.
SVM has been trained and achieved the accuracy of 83.51, weighted precision, recall and F1-score of 0.84, 0.84 and 0.82, respectively. forecasted Class 2 (F 1 = 0.78), Class 3 (F 1 = 0.77), but performed poorly on minority Class 0 (recall = 0.05, F 1 = 0.09).

Table 1. shows the performance of the four machine learning models that were evaluated in this study

Model	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	75.16	0.74	0.75	0.73
Decision Tree	92.48	0.93	0.92	0.92
Random Forest	87.06	0.88	0.87	0.85
Support Vector Machine	83.51	0.84	0.84	0.82

The performance of the four machine learning models that were tested in this study are depicted in table 1. The Decision Tree provided the greatest accuracy of 92.48 average with balanced Precision, Recall, and F1-score, which is why the algorithm is the most effective in predicting student performance. Random Forest classifier was also good, having 88.31% of accuracy, and SVM gave 83.51, offers the middle results, but with less Recall than tree-based classifier. The lowest overall accuracy of 75.16 was obtained with the Logistic Regression, which, however, has proven correct on the majority classes (Class 4.0), it is ineffective on the minority classes since the data was not linear. In this regard, it is worth mentioning that the F1- Scores values that appear in Table I are the mean values of all of four classes (0 - 4) and, as a result, display the general balance between Precision and

Recall of each of the models.

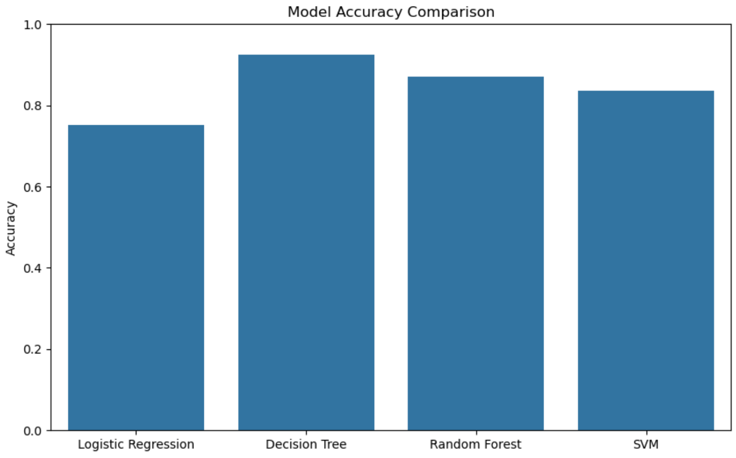

Fig. 2. Comparison of machine learning models (Logistic Regression, Decision Tree, random forest, and SVM) of student performance prediction based on the accuracy of these models

Fig. 2 compares the classification accuracy of four machine learning-based models namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). The most precise of them was the Decision Tree at 92.48 then Random Forest (88.31) and the last one SVM (83.51). The most precise was the Logistic Regression at 75.16.This finding is evidence that tree based algorithms (Decision Tree and Random Forest) are good for modelling the non-linear complexities in the situation with student performance data. Logistic Regression is a linear model, so it did not deal with this level of complexity, and SVM was shown to be not as great as ensemble based.

Confusion Matrix Analysis

Confusion matrices were used to analyze the performance of per-class They highlighted that: Logistic Regression misclassified many students in Class 0 and Class 1, explaining its a weaker recall. SVM performed well but less performed with minority classes. Random Forest and Decision Tree demonstrated were balanced classification, with the Decision Tree achieved nearly perfect predictions Classes 2, 3, and 4.

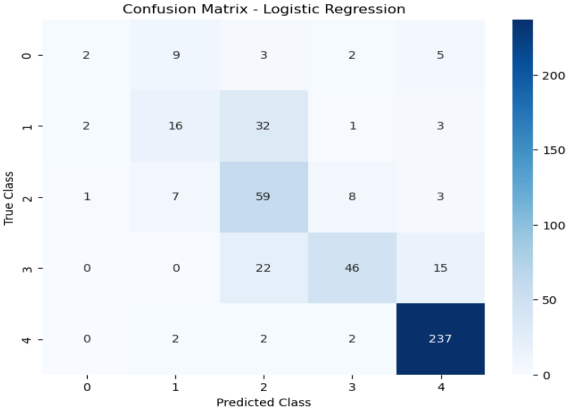

Fig. 3. The confusions of the logistic regression model to show the results of student performance classification.

The confusion of the Logistic Regression model is represented in Fig. 3. The model properly categorized most of them in Class 4.0 (237 instances) which is the highest GPA group. It however poorly matched with lower GPA classes like Class 0.0 and Class 1.0, whereby, in most cases, cross-contamination classified them in other categories.
This is why the overall accuracy rate of the 75.16% in the case of Logistic Regression is comparatively lower. Its linear limits of decision-making were not adequate enough to represent the non-linear dataset relationships with high accuracy and therefore this failed to recall very well with minority classes.

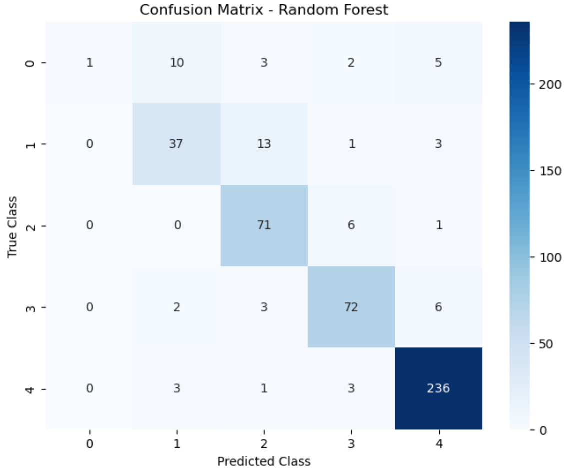

Fig. 4. Student performance prediction after using the Random Forest model confusion matrix..
The confusion matrix of the Random Forest model is presented in Fig. 4. The majority of the examples were rightly accepted by the classifier in all the classes, especially Class 2.0 (74 instances) and Class 4.0 (236 instances) and within these classes only a few misclassifications. Misclassifications were not quite high in cases of intermediate categories with Class 1.0 and Class 3.0 having high recall. This good performance is the reason behind the overall accuracy 88.31 which is second best performing model after the decision tree. Its ensemble methodology also served to mitigate overfitting as well as enhance robustness, but was slightly less precise than the single Decision Tree in this data.

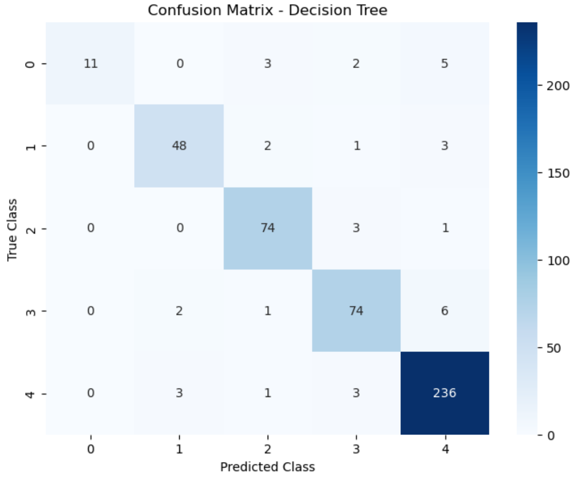

Fig. 5. The decision tree model of student performance prediction has its confusion matrix.

Fig. 5 represents the perplexity table of Decision Tree classifier. The model shows a good predictive accuracy in all grade classes, as Chrases 2.0, 3.0, and 4.0 have almost perfect predictions (74, 74, and 236 accurately predicted respectively). Even the minor classes like 0.0 and 1.0 were categorized rather more effectively as compared to Logistic Regression and SVM. This better use of accuracy justifies the top total accuracy of 92.48, and such makes Decision Tree the best performing model in this research. The capacity to predict student performance in non-linear aspects of interactions with the features was a balanced and interpretable solution.

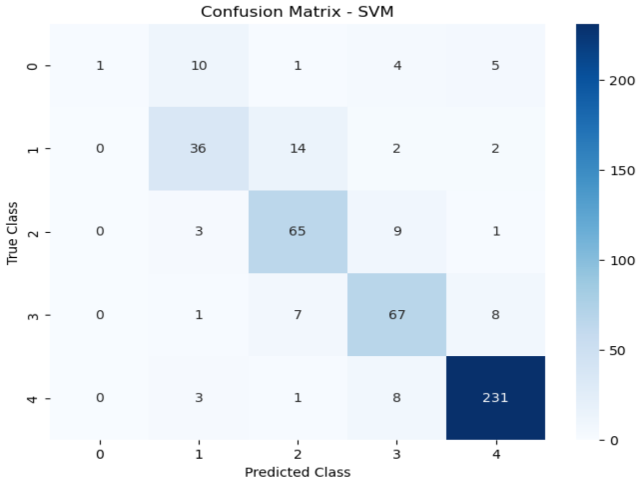

Fig. 6. Confusion double of the Support Vector Machine (SVM) model of student performance prediction.

Fig.6 presents the confusion matrix of SVM classifier. Performing good was the model in higher GPA sets, but mainly Class 4.0 (231 correctly predicted) and Class 3.0 (67 correctlypredicted). It also performed fairly well in terms of Class 2.0 (65 instances).

The model, however, was problematic with minority classes as Class 0.0 and Class 1.0 which exhibited quite a decent amount of misclassified samples into another type. Such imbalance justifies the accuracy of 83.51 in general that is, although that is more precise than that of Logistic Regression, it was worse than that of both Decision Tree and Random Forest.

5. Conclusion

This paper has shown how effectively demographic, academic, and behavioral features of a learner (2392 records and 15 features) can be used by a learner-oriented machine learning model in predicting academic performance of the student. The Decision Tree was the most accurate when compared with other models (92.48%), then there was the Random Forest (88.31%), SVM (83.51%), and lastly, the Logistic Regression(75.16%). From the findings of this study, the following were discovered to be the major predictors of performance and they were: Parental Support, Study Time and Absences that seemed to have a moderate effect moderated by extracurricular activities. These change results also comment on the madness some of which still useful models could have been count on to offer to the student mess in time to ahead to descending to show exhaustion into the earliest indications of the poor academic movement, as well as explanation for tutoring and family interference.And while we have weak-points to these results, they are promising. The experiment was based on one single dataset, used old-school methods of ML and suffered from issues of class imbalance with minorities. To collaborate with them, future studies need to adopt more bigger and heterogeneous data, more powerful software such as deep neural networks and gradient boosting and hybrid assembling, interpretability based on AI models (i.e., SHAP, LIME), etc. The cognitive environment of real-time Learning management System (LMS) is further advanced by the multiple prediction models which produce on one side a more flexible environment and on the other hand the equal and ethical aspect can be understood to stimulate equal result in the didactic progress.

References

1. H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, “Predicting academic performance of students from VLE big data using deep learning models,” Computers in Human Behavior, vol. 104, Mar. 2020, doi: 10.1016/j.chb.2019.106189.
2. X. Li, X. Zhu, X. Zhu, Y. Ji, and X. Tang, “Student Academic Performance Prediction Using Deep Multi-source Behavior Sequential Network,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer, 2020, pp. 567–579. doi: 10.1007/978-3-030-47426-3_44.
3. B. Sekeroglu, R. Abiyev, A. Ilhan, M. Arslan, and J. B. Idoko, “Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies,” Applied Sciences (Switzerland), vol. 11, no. 22. MDPI, Nov. 01, 2021. doi: 10.3390/app112210907.
4. A. Namoun and A. Alshanqiti, “Predicting student performance using data mining and learning analytics techniques: A systematic literature review,” Applied Sciences (Switzerland), vol. 11, no. 1. MDPI AG, pp. 1–28, Jan. 01, 2021. doi: 10.3390/app11010237.
5. Y. A. Alsariera, Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali, “Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance,” Computational Intelligence and Neuroscience, vol. 2022. Hindawi Limited, 2022. doi: 10.1155/2022/4151487.
6. M. Yağcı, “Educational data mining: prediction of students’ academic performance using machine learning algorithms,” Smart Learning Environments, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00192-z.
7. J. Valverde-Berrocoso, J. Acevedo-Borrega, and M. Cerezo-Pizarro, “Educational Technology and Student Performance: A Systematic Review,” Frontiers in Education, vol. 7. Frontiers Media S.A., Jun. 28, 2022. doi: 10.3389/feduc.2022.916502.
8. S. Chen and Y. Ding, “A Machine Learning Approach to Predicting Academic Performance in Pennsylvania’s Schools,” Social Sciences, vol. 12, no. 3, Mar. 2023, doi: 10.3390/socsci12030118.
9. X. Wen and H. Juan, “Early Prediction of Students’ Performance Using a Deep Neural Network Based on Online Learning Activity Sequence,” Applied Sciences (Switzerland), vol. 13, no. 15, Aug. 2023, doi: 10.3390/app13158933.
10. M. A. Aslam, F. Murtaza, M. E. U. Haq, A. Yasin, and M.A.Azam, “A Human-Centered Approach to Academic Performance Prediction Using Personality Factors in Educational AI,” Information (Switzerland), vol. 15, no. 12, Dec. 2024, doi:

10.3390/info15120777.
11. E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," Applied Computational Intelligence and Soft Computing, vol. 2024, 2024, doi: 10.1155/2024/4067721.
12. J. Mai, F. Wei, W. He, H. Huang, and H. Zhu, "An Explainable Student Performance Prediction Method Based on Dual-Level Progressive Classification Belief Rule Base," Electronics (Switzerland), vol. 13, no. 22, Nov. 2024, doi: 10.3390/electronics13224358.
13. Z. Luo et al., "A Method for Prediction and Analysis of Student Performance That Combines Multi-Dimensional Features of Time and Space," Mathematics, vol. 12, no. 22, Nov. 2024, doi: 10.3390/math12223597.
14. Z. Shou, M. Xie, J. Mo, and H. Zhang, "Predicting Student Performance in Online Learning: Multidimensional Time- Series Data Analysis Approach," Applied Sciences (Switzerland), vol. 14, no. 6, Mar. 2024, doi: 10.3390/app14062522.
15. D. Ying and J. Ma, "Student Performance Prediction with Regression Approach and Data Generation," Applied Sciences (Switzerland), vol. 14, no. 3, Feb. 2024, doi: 10.3390/app14031148.