A DATA-DRIVEN APPROACH TO EARLY STUDENT PERFORMANCE PREDICTION USING COLLABORATIVE FILTERING

Y. Suhasini
Research Scholar
University of Technology
Jaipur.

Dr. Suneel Pappala
Guide
University of Technology
Jaipur.

Dr. Shasi Kiran Jangala
Co-Guide
University of Technology
Jaipur.

ABSTRACT

The education sector has emerged as a prominent area for data exploration, leading to an increased interest among companies in educational data. As a result, the demand for data-driven insights in this field has risen significantly. Through the application of data mining and machine learning techniques, I focused on extracting valuable information from educational datasets to develop automated tools aimed at enhancing the education domain. This initiative centers on analyzing student achievement by employing various machine learning and data mining methodologies. Real-world data, including student grades, demographic information, social factors, and school-related characteristics, was gathered from school reports and questionnaires. Four distinct data mining models were utilized: Decision Tree, Random Forest, Neural Networks, and Support Vector Machines. These models incorporated multiple variables that correlate strongly with final grades. The findings demonstrate a commendable accuracy when predicting outcomes based on students' first, second, and third-grade scores. While past evaluations substantially influence students' performance, a comprehensive analysis reveals that several additional factors, such as attendance rates, parents' occupations and education levels, and alcohol consumption, also contribute to academic success. This research opens the door for the development of advanced student prediction tools and provides insights into the challenges faced by special needs students. Ultimately, these findings aim to improve educational quality and optimize school resource management.

INTRODUCTION

The timely prediction of student performance offers a myriad of benefits, including the early identification of students struggling to pass their modules and those at risk of dropping out, course selection pathways, as well as the attributes that influence student retention rates and behaviors. Such intelligent insights empower educational leaders to devise and implement corrective interventions to support academic advising, guide curriculum changes and improvements, and determine the pitfalls of the programs. However, selecting an appropriate machine learning model to estimate student performance accurately remains a complex endeavor. Findings have shown that student academic achievements are typically influenced by a myriad of factors, ranging from academic and non-academic attributes. The variability of these factors necessitates the development of a complex predictive model. Ensemble learning models are already proven to outperform individual learning models concerning the prediction accuracy of student academic performance. Moreover, a recent survey revealed that around 50% of the studies employ supervised learning algorithms, while only 5% of the studies use unsupervised approaches. Indeed, supervised machine learning provides acceptable prediction accuracy however, we believe that augmenting a supervised model with unsupervised learning will

produce even more accurate predictions, with fewer generalization errors. As such, our suggested model integrates the powers of supervised and unsupervised learning. The available models and algorithms focus mainly on improving the prediction accuracy of future student performance. They fall short of generating an explanatory analysis of the exact factors (i.e., variables) that cause the observed student performance. Additionally, relying on a single model, whether this model is linear or non-linear, may be insufficient due to the difficulty of capturing a variety of factors in one predictor model. Factors affecting student performance often differ significantly among students and between academic semesters for the same students. To avoid ambiguity, single linear models would generally suffer from under fitting the data on which they were trained (i.e., consisting of many overlapping student behaviors), leading to high rates of false predictions. Likewise, false predictions would also be high with non-linear models as they are prone to over fitting just parts of the data on which they trained (i.e., the model would only remember some aspects of student behaviors).

Moreover, existing ensemble machine learning solutions do not accommodate for a dynamic weighted contribution of the participating models in predicting student performance. Further limitations concern the disuse of a training set or the use of a single dataset to validate the model, such as. Moreover, some models focus on predicting the achievements of first-year students only. More than 50% of the surveyed studies used SVM and ANN techniques to predict student performance. Moreover, most related approaches that we are aware of are confined to predicting future course grades only without associating them with the key factors that lead to the obtained student performance. In our view, understanding the impact of those enabling and inhibiting factors is quite essential to devise corrective plans to improve student achievements and reduce the risk of dropout. Our approach is distinguished by its clustering feature that helps in understanding the associated factors leading to the predicted future course grades. To address the aforementioned limitations, we contribute a hybrid regression approach along with a semi-supervised learning technique for identifying the enabling factors and inhibitors of student performance in educational programs. The proposed approach seeks to optimize the prediction accuracy of student academic performance based on course grades, and then identify the possible factors that might have caused the observed student achievements. We assume that program strengths and weaknesses are instigated by a set of factors and circumstances, which are believed to have direct or latent effects on student academic results.

LITERATURE REVIEW

Dr. S.K Singh [2024] In contemporary education, pupils' academic performance must be predicted correctly. Logistic Regression and Random Forest are examples of machine learning algorithms that can effectively predict students who are more likely to fail. Logistic Regression is used for the assessment and prediction of such issues as timely interventions and customized care which address multiple factors in a student's life. Conversely, Random Forest may handle big data sets well because its precision is remarkable. With a population of 480 students sampled

from Kalboard360, this study compares logistic regression and random forest about demographic, educational or behavioural features. The findings reveal the effectiveness of both approaches: Logistic regression has an 81% accuracy rate with some space for improvement while Random Forest has a classification accuracy rate of 89% which shows that it can categorize different student outcomes. Such insights from these algorithms have informed tailored interventions aimed at highlighting the significance of employing Machine Learning techniques within the academic arena and understanding their strengths as well as weaknesses for successful strategies.

Venera Nakhipova [2024] This article introduces a novel method that integrates collaborative filtering into the naive Bayes model to enhance predicting student academic performance. The combined approach leverages collaborative user behavior analysis and probabilistic modeling, showing promising results in improved prediction precision. Collaborative Filtering explores user behavior patterns, while Naive Bayes employs Bayes' theorem for probabilistic data classification. Focused on predicting academic success, the integration incorporates collaborative patterns from student data for increased accuracy. The method considers similar students' performance and behavior for nuanced, personalized predictions. Starting with diverse data collection, including collaborative patterns among students, Collaborative Filtering identifies relationships and patterns among those with similar academic histories. These insights enrich the naive Bayes algorithm, creating a holistic approach for more accurate predictions, and contributing to ongoing machine learning initiatives in education.

Dr. Geeta Tripathi [2024] Evaluating students' learning performance is a fundamental aspect of evaluating any educational institution. When addressing challenges related to the learning process, student performance is critical, and it is one of the key factors used to quantify learning outcomes. The topic of research known as educational data mining (EDM) has grown out of the potential to leverage data knowledge to enhance educational systems. EDM is the development of methods to analyze data collected from educational environments, enabling a more complete and precise understanding of students and the enhancement of their educational results. Evaluating the students' learning results is a crucial part of evaluating any educational institution. One of the key variables used to quantify learning outcomes is student performance, which is significant when addressing problems with the learning process. The field of research known as educational data mining, or EDM, was born out of the potential to leverage data knowledge to enhance educational systems. EDM is the process of developing methods for evaluating information obtained from educational environments. This makes it possible to learn more precise and in-depth information about students and enhances their academic achievement. Academic achievement tests (AAT), general aptitude tests (GAT), admission scores, first-level courses, and other early-stage factors are used in the paper's dimensionality reduction mechanism by T-SNE algorithm for the clustering technique. This allows the study to investigate the relationship

between these aspects and GPAs. Regarding the categorization method, the study showcases tests conducted on various machine learning models that forecast student achievement in the initial phases by utilizing diverse attributes such as course grades and entrance exam results. To gauge the models' quality, we employ various evaluation measures. Based on the findings, it appears that early student failure rates can be reduced by educational institutions.

Esmael Ahmed Abdu [2024] Education is crucial for a productive life and providing necessary resources. With the advent of technology like artificial intelligence, higher education institutions are incorporating technology into traditional teaching methods. Predicting academic success has gained interest in education as a strong academic record improves a university's ranking and increases student employment opportunities. Modern learning institutions face challenges in analyzing performance, providing high-quality education, formulating strategies for evaluating students' performance, and identifying future needs. E-learning is a rapidly growing and advanced form of education, where students enroll in online courses. Platforms like Intelligent Tutoring Systems (ITS), learning management systems (LMS), and massive open online courses (MOOC) use educational data mining (EDM) to develop automatic grading systems, recommenders, and adaptative systems. However, e-learning is still considered a challenging learning environment due to the lack of direct interaction between students and course instructors. Machine learning (ML) is used in developing adaptive intelligent systems that can perform complex tasks beyond human abilities. Some areas of applications of ML algorithms include cluster analysis, pattern recognition, image processing, natural language processing, and medical diagnostics. In this research work, K-means, a clustering data mining technique using Davies' Bouldin method, obtains clusters to find important features affecting students' performance. The study found that the SVM algorithm had the best prediction results after parameter adjustment, with a 96% accuracy rate. In this paper, the researchers have examined the functions of the Support Vector Machine, Decision Tree, naive Bayes, and KNN classifiers. The outcomes of parameter adjustment greatly increased the accuracy of the four prediction models. Naïve Bayes model's prediction accuracy is the lowest when compared to other prediction methods, as it assumes a strong independent relationship between features.

Soukaina Hakkal [2024] The huge amount of data generated by an Intelligent Tutoring System becomes useful when analyzed in an appropriate way to provide significant insights about learners, especially his or her performance. Performance data retrieved from historical interactions is the main engine for learner performance prediction, where the likelihood of the learner answering correctly future questions is calculated. Modeling learner performance can provide significant insights into individual students to promote successful learning and maximize educational achievement. This study aims to enhance the learner performance prediction of some logistic regression-based models, namely Item Response Theory, Performance Factor Analysis, and DAS3H using XG Boost, including an empirical comparison of eight

real-world datasets, containing performance log data collected from different online intelligent tutoring systems, involving the first time a new dataset from Moodle Morocco. The results have demonstrated that the XGBoost has enhanced PFA predictive performance on seven datasets with an AUC of up 0.88 and improved the DAS3H AUC on the ASSISTment17 dataset while conserving almost the same predictive results for Item Response Theory on some datasets.

METHODOLOGY

Cluster Based Student Record Classification

The Educational Data Mining (EDM) and learning analytics are effectively used for enhancing the quality of result classification in student performance analysis. The educational institutions are involved in things to reduce the poor results of students. With that concern, many techniques are developed for evaluating the student performances for making the respective faculties to mediate to improve the overall results.

For developing Accurate Student Classification Model, this work comprises of three phases of work, as follows:

i. Cluster based Student Record Classification (CSRC)
ii. Multi-Tier Student Performance Evaluation Model (MTSPEM)
iii. Behaviour based Student Classification System (SCS-B)

WORKING PROCEDURE

The working procedure of the aforementioned models and computations are presented in the following sections. The overall functions involved in the three phases of work are presented in the following Figure 3.1. In EDM, the huge amount of student data is collected through various survey as well as from open-source data repository. For efficient data organization, analysis and classification, K-Means Clustering techniques are used with respect to the obtained academic data. Moreover, using K-Means clustering, the student records are classified under three classes, such as, A. Low B. Average C. Smart

The future prediction of student performances relies on the demographic and academic data. Moreover, the model presents results based on the best, good and average performances of students. The educational models are required to develop the innovative models to preserve students and to make the students to become graduated on right time. For that, efficient study plan should be provided by the tutors based on the result of student classification. Moreover, the model works on deriving the successful academic guide for students. For each student, the classification process is done with the ranking arrangements based on the student score and their performance history.

Results and Discussions

The main objective of the proposed work is to detect the student performances in academics based on their records. For implementation, Waikato Environment for Knowledge Analysis (WEKA), which is an open source environment, is used. The student data files are collected and converted into Attribute Relation File Format (ARFF) for data evaluations. Moreover, the results are evaluated based on the factors such as, classification accuracy, precision value and error rates. The results are compared with the existing models such as, Support Vector Machine (SVM), Multi-layer Perceptron (MLP) and

Artificial Neural Network (ANN). The classifiers, employed for two testing models are divided for training and testing functions. The results of the proposed model are provided in the following figures. The Figure 1 and Figure 2 portray the results obtained for the application of J48 algorithm. Furthermore, the J48 tree visualization is presented in Figure 3.

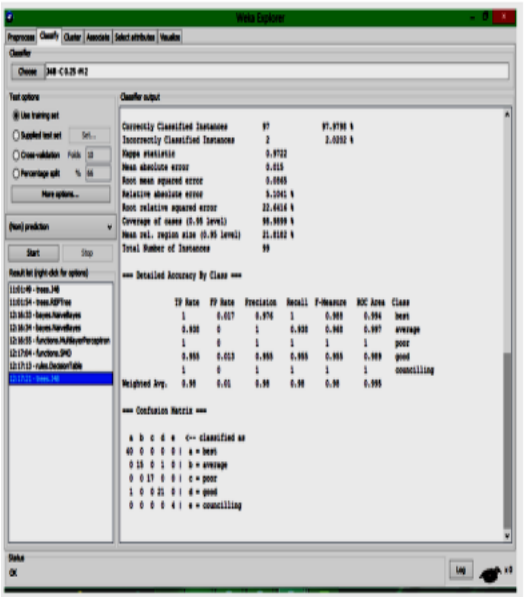

Figure 1: Results obtained for J48 classifier

Figure 2 J48 tree representations for student analysis

Model Design

The model design of the classification operation is explained in this section. Among the compared classification model, the J48 algorithm provides better results than other compared models. The modules in the design model comprises of,

1. Log-in Data
2. Student Record
3. Student Result Analysis

J48 is a tree based classifier model, which is developed based on ID3 algorithm. Moreover, the model uses divide and conquer based node splits, for defining the leaf, child and root nodes. In the given set of 'S' samples, the tree structure is framed as follows

1. If all the samples in 'S' are under the same class or 'S' is having minimal samples, such leaf is considered as the frequent class in 'S'
2. If previous step is not occurred, then selection process is carried out based on single feature with
3. minimal two or more than two possible results. Then, the sample partitions are given as, S1, S2, S3, based on the student cases
4. In recursive manner, the same sets of operations are applied to the other sub nodes.
5. Gain Ratio and Information gain values are ordered based on the heuristic results.

The form results of the proposed model are depicted in the following Figures, in which the log in data are provided in the form in Figure, authentication is in Figure, student data collection is done through the form in Figure, Performance analysis are processed with that. In the results, the student samples are classified under the categories as,

i. Best
ii. Good
iii. Average
iv. .Poor

The results are given for the tutors to design a new learning methodology based on the classification, thereby improving their results and success ratio, effectively. Their corresponding values obtained on the experimentation in WEKA environment are given in Table 1, Table 2 and Table 3.

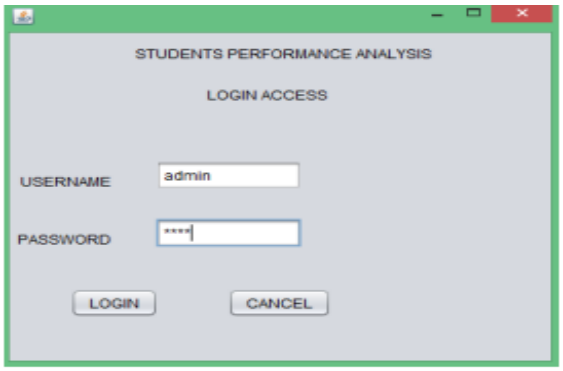

Figure 3: Login credential

Table 1 Values obtained for classification accuracy

Models	Classification Accuracy for Samples				
	40	80	120	160	200
SVM	62.1	69.7	71.0	67.4	68.6
MLP	70.5	81.6	73.3	71.8	72.9
CSRC	94.1	92.4	90.4	94.2	93.2

Table 2 Precision values Vs. Student numbers

Models	Precision Value for Samples				
	40	80	120	160	200
SVM	66.6	75.7	65.5	50.7	44.6
MLP	71.1	76.9	71.9	60.8	54.8
CSRC	94.5	94.3	92.8	81.2	77.2

Table 3 Error rate comparisons

Model	Error rate for Samples				
	40	80	120	160	200
SVM	13.1	14.6	18.2	19.4	21.3
MLP	10.0	11.8	14.6	14.6	17.4
CSRC	3.3	6.1	7.0	6.8	11.1

CONCLUSION

The academic journey of students is often influenced by several cognitive and non-cognitive factors, many of which may not be immediately apparent in traditional evaluation systems. In this research, we explored a hybrid approach for early prediction of student performance using collaborative filtering (CF) enhanced with machine learning algorithms. This approach sought to proactively identify at-risk students by leveraging patterns and similarities in historical academic behavior, engagement metrics, and peer-group performance.

Collaborative filtering, traditionally used in recommender systems, proved to be a powerful paradigm in educational contexts. By treating students as "users" and learning outcomes or academic indicators as "items," we were able to predict a student's likely performance in future assessments based on the observed behavior and outcomes of similar peers. When integrated with supervised learning algorithms such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting, CF showed significantly improved predictive accuracy. These models helped extract underlying patterns in diverse educational datasets while mitigating biases and overfitting.

In conclusion, the use of collaborative filtering coupled with machine learning represents a transformative approach to

student performance prediction. This hybrid model not only identifies at-risk students early but also provides valuable insights into peer dynamics, learning patterns, and institutional factors affecting academic performance. As education systems globally shift toward personalization and data-driven decision-making, such predictive frameworks will play a vital role in enhancing student outcomes, improving retention, and promoting equitable learning environments.

REFERENCES

1. *S.K Singh [2024], "Predicting Academic Performance Using Machine Learning Algorithms", International Journal of Creative Research Thoughts, ISSN: 2320-2882, Volume.12, Issue.3*
2. *Venera Nakhipova [2024], "Integration of Collaborative Filtering Into Naive Bayes Method to Enhance Student Performance Prediction", International Journal of Information and Communication Technology Education (IJICTE),ISSN:1550-1337,vol.20,issue.(1)*
3. *Dr. Geeta Tripathi [2024], "Early Prediction of Students Performance in Higher Education", International Journal of Scientific Research in Science and Technology,ISSN 2395-602X,vol.11,issue.(3),pages.01-10, DOI:10.32628/IJ SRST24112166*
4. *Esmael Ahmed Abdu [2024], "Student Performance Prediction Using Machine Learning Algorithms", Applied Computational Intelligence and Soft Computing,ISSN:1687-9732,vol.(1),DOI:10.1155/2024/4067721*
5. *Soukaina Hakkal [2024], "XGBoost To Enhance Learner Performance Prediction", Computers and Education: Artificial Intelligence, ISSN 2666-920X,Volume.7,https://doi.org/10.1016/j.caeai.2024.100254.*
6. *Ali Cakmak [2017], "Predicting Student Success in Courses via Collaborative Filtering", International Journal of Intelligent Systems and Applications in Engineering,ISSN:2147-67992,vol.5,issue.(1),pages.10–17*
7. *Yupei Zhang [2021], "Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis", Frontiers in Psychology,ISSN 1664-1078,Volume.12 https://doi.org/10.3389/fpsyg.2021.698490*
8. *Christos Sardianos [2019], "Optimizing Parallel Collaborative Filtering Approaches for Improving Recommendation Systems Performance", Information,ISSN:2078-2489,Volume.10 ,Issue.5, https://doi.org/0.3390/info 10050155*
9. *Balqis Albreiki [2021], "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques", Education sciences, ISSN 2227-7102,vol.11,https://doi.org/10.3390/educsci11090552*
10. *Arbër H. [2018] "Student Performance Prediction Using AI and ML: State of the Art", International Conference on Computer and Information Sciences (ICCOINS), ISSN: 2836-9122, Volume.7, Issue.5*
11. *Gabor Tak ´ acs [2009], "Scalable Collaborative Filtering Approaches for Large Recommender Systems", Journal of Machine Learning Research,ISSN:1533-7928,vol.10,pages.623-656*
12. *Rachit Tomar [2015], "User propensity analysis for Movie prediction rating based on Collaborative filtering and Fuzzy system", International Journal of Innovative Science, Engineering & Technology, ISSN:2348 – 7968, Vol.2, Issue.9*