Implementing Computational Intelligence for Student Classification and Recommendation

Syed Aamer Hashmi[1], Yashpal Singh[1], Harshit Bhardwaj[2]
Department of CSE, Amity University, Jaipur, Rajasthan, India[1]
Department of CSE, ASET, Amity University, Noida, Uttar Pradesh, India[2]
Corresponding author: Syed Aamer Hashmi, Email: amer.syed20@gmail.com

Personalized learning has become very crucial in shaping the future of students, with lot of career options available and lack of knowledge about the best domain that can suite and help individual student, there is a need of effective recommender system which will help them to choose the best courses for them. The idea of this research is to create a classification model which will classify students in fast and slow learner, and this will be used to create a recommender system to suggest the best courses that individual student should opt for. The Professional and personal information of students such as their academic marks, family background is used to create the models. Computational intelligence uses a combination of multiple algorithms of Machine learning and Deep learning to create the final model. The use of Computational intelligence gives very good accuracy compared to traditional methods.

Keywords: Computational Intelligence, Student Classification, Machine Learning.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 567–575. DOI: https://doi.org/10.56155/978-81-955020-7-3-50

1 Introduction

Modern Education System requires smart systems based on AI, which can help in improvement of individual students' performance. Even the Educators need to know about individual students' performance by analyzing the data.

Each student has different level of understanding, learning and grasping the knowledge, also there is difference in subject and area which individual likes and can excel into, it is important to guide students by keeping these factors into account otherwise there can be lack of interest while learning resulting in poor performance in academic and in career. Even the educators and institutions won't be able to provide proper guidance to students without using such systems.

To achieve the goal of creating accurate predictive model, Machine learning algorithms is used, as the model is of binary classification the algorithms like Logistic regression, Naïve bayes, Support Vector Machine with kernel trick and ensemble technique-based algorithm like Random Forest are implemented. To check the accuracy of model accuracy matrices like Accuracy, Recall, Precision and f1 Score are used.

The implemented Model will be accurate and robust, the use of Machine learning ensures that the model will be reliable to use by students, educators and the institutions, the model will also be used in creating the recommendation system, which will recommend students the courses that they should opt for in the coming semesters.

2 Related Work

V. Hegde and H. Sushma Rao [1] explore the assessment of student performance, particularly in programming languages such as C, C++, and Java. They emphasize the profound influence of programming proficiency on students' future career opportunities in today's world. The research goes beyond just academic grades, incorporating attendance data to provide a holistic evaluation of student performance. Hegde and Sushma Rao utilize Educational Data Mining to dissect the landscape of student performance.

P. Rojanavasu [2] explores the use of data mining in education for administration and planning. He analyses admission and student course grade datasets to answer two research questions. The study uses association rule mining to support admission planning, finding useful patterns. For predicting job outcomes, decision tree analysis is employed, showing valuable insights. Overall, the paper illustrates how data mining can benefit educational administration, suggesting educators use these techniques for improved planning and student services.

Arsad et al. [3] explore predicting the academic performance of engineering students using Artificial Neural Networks (ANN). They focus on using the final semester results from student data within the Electrical Engineering program at a college in Malaysia. For their analysis, the authors utilize ANN, a computational tool known for its ability to process complex data. They assess the model's performance using methods such as Mean Square Error (MSE) and the R Coefficient method.

Shahiri, Husain, and Rashid [4] explore the accuracy of different Data Mining Techniques in predicting student performance. They identify key student attributes like demographics and external assessments. Among classification algorithms tested, Artificial Neural Networks (ANN) achieved the highest accuracy (98%), followed by Decision Treses (91%), while SVM and KNN algorithms both scored 83%, and Naive Bayes algorithm achieved the lowest accuracy (76%).

Keisuke [5] used the machine learning for analysing educational data, they highlighted the use of important factors for Machine learning Model creation. They extracted useful information using Data

Mining and have used various algorithms of Machine learning. The study also found challenges like data privacy and data drift and proposes solutions like the implementation of robust data governance policies and educator training initiatives. They concluded by giving valuable resources for educators and policymakers, offering actionable insights on using machine learning to improve educational practices.

V. U. Kumar et al. [6] predicts the student performance using many machine learning algorithms. They used the data of last five years of graduates from K L University, the features like semester scores, participation in extracurricular activities and competitions, and overall attendance are used. Five machine learning algorithms that they used are Decision Tree, K Means, Naive Bayes, SVM, and Hierarchical Clustering.

H. Chen et al. [7] analyses the student behavior during computerized programming tests. they classify students in five distinct types. They found that the efforts put into homework assignments does not consistently correlate with grades, instead the factors like motivation and timely submissions play a more significant role. The paper also suggests restricting one time attempt to test to discourage guessing and underscores the significance of producing high-quality code for academic success.

Ashraf [8] used the Educational Data Mining along with techniques like Structural Equation Modelling (SEM) and Analysis of Variance (ANOVA) to extract patterns from data. The research sorts the students' academic and personal information and conclude that subjects like English, Chemistry, Zoology, and Biology significantly influence overall performance and academic outcomes.

Yang et al. [9] used Artificial Neural Networks (ANN) to predict the students' grades in Massive Open Online Courses (MOOCs), which has become very complex due to the rise of Machine Learning. In today's digital era, online learning platforms like Udemy, edX, and Coursera have become very popular. grades are determined by factors like video views, assessments, and engagement. Instructors can identify struggling and successful students based on these metrics.

Botelho et al. [10] explores the use of online learning platforms and techniques to support student learning processes, the focus is on student persistence which is a critical aspect of learning. They highlighted the challenge of students' hard work without getting any improvement in their results. They used transfer learning techniques in deep learning and traditional modelling, to examine low and unproductive high persistence. By also considering the issues like dropout rates and unproductivity, the study provides insights about when interventions can effectively assist students during their learning journey.

Jalota and Agrawal [11] underscore the significance of data mining for informed decision-making in schools, particularly through machine learning and statistics in Educational Data Mining (EDM). They employ the Kalboard 360 dataset and WEKA software to analyse and forecast student performance. The study also surveys prior research on data mining in education, examines classifier performance metrics, and proposes avenues for future research in educational data mining. The references cited offer additional insights into related studies. Overall, the paper provides valuable insights into predicting student outcomes and enhancing education through data mining.

Ghorbani and Ghousi [12] use Data Mining to predict student performance, addressing imbalanced datasets by comparing resampling methods. They evaluate techniques like SVM-SMOTE and Random Forest classifier, finding that balanced datasets improve algorithm performance, with SVM-SMOTE and Random Forest yielding the best results.

W. Chen et al. [13] developed a model to predict outcomes for short-term online courses using Predictive Learning Analytics. This approach enables instructors to enhance course quality and analyze student dropout rates, quiz scores, and final exam performance. They employed Linear Discriminant

Analysis, Forward Neural Network, Random Forest, KNN, and SVM for analysis and prediction, with Random Forest and SVM yielding the highest accuracy among all algorithms.

Mengash [14] proposes research on predicting students' performance at the onset of graduation. They utilize a dataset comprising 2039 computer science students from a Saudi Public University. High school marks, admission test scores, and aptitude test scores are identified as crucial factors for prediction. Notably, admission test scores exert the most significant influence, warranting higher weighting. The study employs Machine Learning algorithms, including Decision Tree, SVM, Naive Bayes, and ANN, with ANN achieving the highest accuracy of 79% in prediction.

Zaffar and Hashmani [15]propose a methodology for selecting algorithms in Educational Data Mining (EDM) for student datasets. They offer insights for new researchers by evaluating different Feature Selection (FS) algorithms and classifiers. FS aims to boost predictive accuracy by eliminating non-predictive data. The study categorizes techniques into filter, wrapper, and embedded models, emphasizing the importance of combining various FS algorithms and classifiers for improved prediction accuracy. The paper seeks to contribute to the enhancement of education quality and guide researchers in understanding factors influencing student performance for the development of better prediction models.

Yanes et al. [16] Developed Machine Learning (ML)-based suggestions for enhancing student learning in academics. They explore a range of ML algorithms and methods to predict course outcomes, academic performance, and course actions. The document underscores the significance of data preprocessing and feature selection in this process. Experimental results demonstrate the performance of different classification algorithms and the effectiveness of the proposed approach in improving teaching strategies.

Kinnunen et al. [17] investigate student success using the innovative concept of Phonomyography, departing from traditional scoring methods. They adopt an instructor-centric perspective, considering parameters such as the subjects studied, students' intrinsic characteristics, background, behavior, attitude, and the teacher's influence. The study observes students' comprehension of programming languages and tracks their progress over time, offering a fresh approach to evaluating student success.

Tam et al. [18] explore Educational Data Mining (EDM) and learning analytics (LA), aiming to extract insights from vast educational datasets. They focus on identifying prerequisite relationships through online educational platforms, proposing a semi-supervised learning approach that combines a concept-based classifier with explicit semantic analysis (ESA) to formulate prerequisite rules. Their functional prototype shows promising results, particularly in engineering subjects.

Butt et al. [19] used multiple Machine Learning algorithm and proposed a new ensemble system for students performance prediction model, they combined Naïve bayes, Decision tree and ANN algorithm, apart from the normal academic and personal information they have also used Assignment, Quiz and Presentation marks of the students. The accuracy of the proposed model is more than 95%.

3 Data Cleaning and Preprocessing

Any Machine Learning algorithm require quality and correct data to give accurate results, data should be large enough in terms of related features and number of rows as well, Computational intelligence-based algorithms like Random Forest and Artificial Neural Network (ANN) requires considerably large data.

For designing and model Creation, Students relevant data is taken from a private college of Mumbai University, data contains academic details such as marks obtained in10nth ,12th and each semester of

graduation also students Mentoring Session data is taken such as their Attendance, Family background, Gender and how many Backlogs they had. Based on students' overall performance students are classified into Fast and Slow learners by their Mentors. So here the Dependent variable will be Students classification as Fast or Slow learner.

The following is Pearson's correlation Matrix of important independent features.

Figure 1. Pearson's Correlation Matrix of Independent Features

As shown in Figure 1, there are positive correlations between the percentages indicating the consistency in marks obtained by the students, especially 3rd and 4th semester percentages are 60% correlated.

Figure 2 indicates the Boxplot indicating the Inter quartile range of different independent features. The 4th semester percentage has some outliers and rest are in normal range.

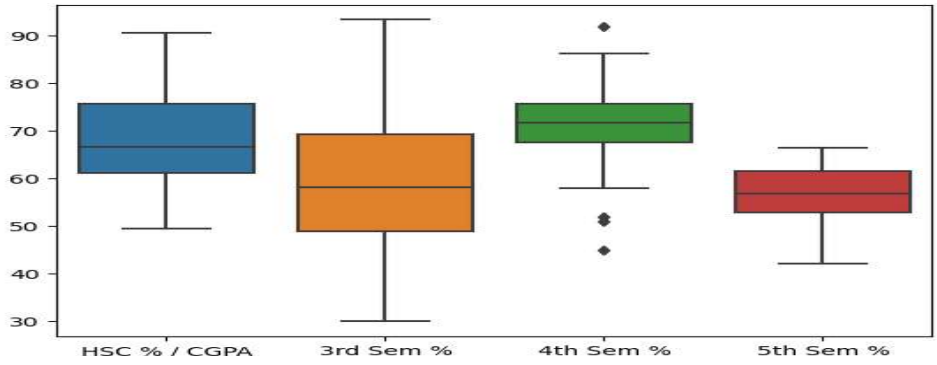

Figure 2. Box Plot indicating Outliers.

4 Classification and Recommendation

Several methods that are utilized for the recommendation and classification are as follows:

4.1 Logistic Regression

Logistic regression uses the sigmoid function, it calculates the value of m and c in similar way of linear regression but uses mx+c value in sigmoid function, the algorithm gives best result for binary classification and linear data.

$$\text{Sigmoid Function} = {}^{1}/_{1 + \bar{e}}\,(mx + c) \qquad (1)$$

For the given dataset Logistic regression has given the accuracy of 87%. The miss classification is mainly of Slow learners being predicted as fast learners. This can be due to some of the features in which logistic regression didn't find the pattern properly.

5 Support Vector Machine (SVM)

The Support Vector Machine can be used for both regression and classification model, it considers all data points as vectors and finds the maximum Margin Hyperplane based on Support Vectors. For calculation purposes it uses the concept of orthogonal projection of one vector on another vector. To handle nonlinear data SVM has kernel tricks such as RBF kernel, Poly kernel and sigmoid kernel.To get the best kernel for the dataset Hyper parameter is implemented using Grid search CV technique and we got the best accuracy of 93%.

6 Naïve Bayes Algorithm

Naïve Bayes algorithm uses the concept of Bayes theorem where we calculate the probability of an event when another event already happened. The problem with naïve bayes is it considers all features are independent of each other's. also, Naïve bayes can be only used for Classification model.

$$P(A / B) = p(B, A) * p(A) / p(B) \qquad (2)$$

For our dataset we got the accuracy of 90% with Naïve bayes algorithm, there is some Nonlinearity in data which the algorithm is not able to capture.

7 Random Forest

Random forest is one of the best Ensemble techniques to get the best accuracy, it uses the concept of Bagging. The algorithm creates multiple Decision tree based on the data and for prediction checks the result of all the trees.

As shown in Figure 3, The Random Forest algorithm in case classification counts the prediction of all the trees and highest number class is declared as prediction class.

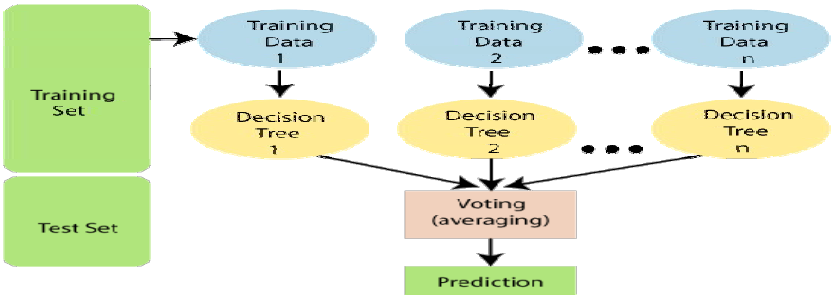

Figure 3. Working of Random Forest

8 Results

As shown in Figure 4, For our dataset Random Forest has given the best accuracy of 97%, indicating that it has learned the data the best among all algorithms.

The data was almost balance so instead of considering Precision, recall or Sensitivity our focus is on Accuracy and F1 score.

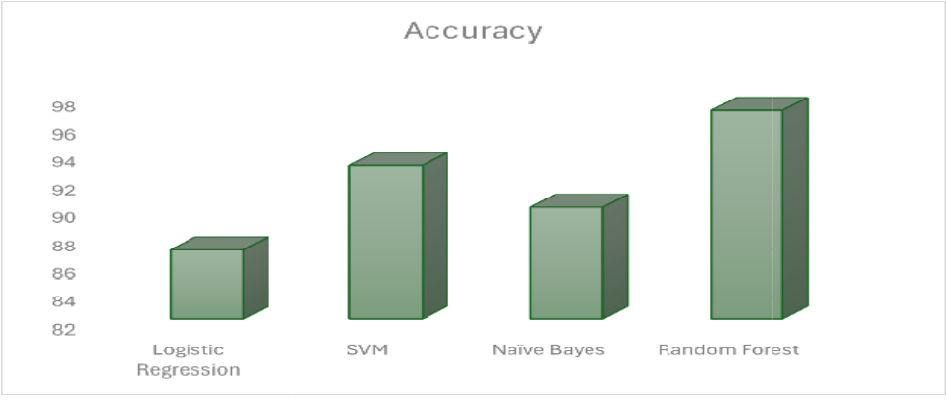

Figure 4. Accuracy matrix of all algorithms

Table 1. Performance Matrix of all algorithms

	Logistic Regression	SVM	Naïve Bayes	Random Forest
Recall	88%	93%	90%	98%
Precision	88%	95%	93%	98%
F1 score	88%	94%	92%	98%
Accuracy	86%	93%	90%	97%

As indicated in the Table 1 the best accuracy is with Random Forest, rest algorithms also gave good accuracies. Also, as data was balance, so all algorithms learn about both the classes well that's why there is not much difference between the accuracy and F1 score for all the algorithms.

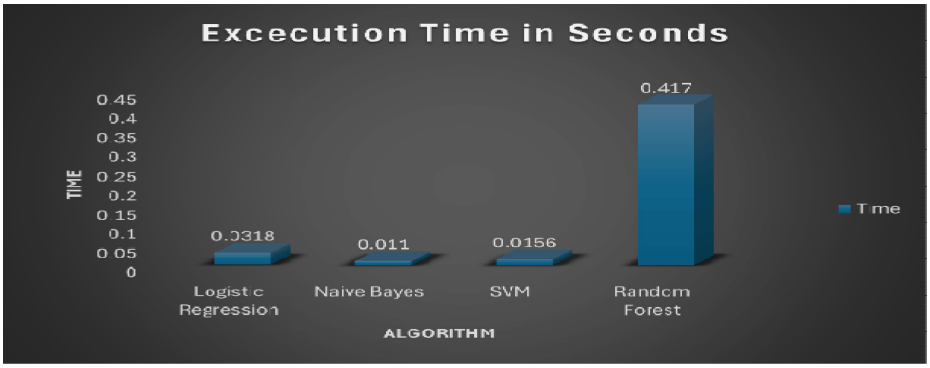

Figure 5. Execution Time taken by each algo.

The Figure 5 is the chart which shows time taken by each algorithm for getting trained, the Naïve bayes takes the least time followed by SVM and Logistic regression, the random forest algorithm as it creates multiple trees while training takes considerably long time, but its accuracy is best, our focus is more on accuracy than the computational time that's why Random Forest is the best algorithm for our research.

Compared to surveyed papers, the major difference in our paper is the features that we have considered and the accuracy of our model. Apart from the data contains information which was carefully filled by the mentors of individual student after interaction and proper analysis of student that's why the data is very reliable. We got the best accuracy of 97% which is better than the surveyed papers.

9 Conclusion and Future Work

The classifier uses the crucial information of students such as academic performances, their attendance, gender and Mentoring information which is collected from the Mentors after keen observation of their mentees, the model classifies data in Fast learner and Slow Learner which can be very beneficial for faculty, college also for the recommender system which will recommend student the courses which they should opt for coming semester. For Future work, the model can use more information such as students' personal information, their sentiment analysis and some psychometric test to understand their personality, technologies like Deep learning, NLP and Gen AI can be used to classify the student. And classifier can be used for various recommendations systems.

References

[1] V. Hegde and H. Sushma Rao, "A Framework to Analyze Performance of Student's in Programming Language Using Educational Data Mining," 2017 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2017, pp. 1–4, 2018, doi: 10.1109/ICCIC.2017.8524244.

[2] P. Rojanavasu, "Educational Data Analytics using Association Rule Mining and Classification," 2019 Jt. Int. Conf. Digit. Arts, Media Technol. with ECTI North. Sect. Conf. Electr. Electron. Comput. Telecommun. Eng. (ECTI DAMT-NCON), pp. 142–145, 2019.

[3] P. M. Arsad, N. Buniyamin, and J. L. A. Manan, "A neural network students' performance prediction model (NNSPPM)," 2013 IEEE Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA 2013, no. November, pp. 26–27, 2013, doi: 10.1109/ICSIMA.2013.6717966.

[4] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," Procedia Comput. Sci., vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.
[5] Keisuke, "Data Mining and Machine Learning Application for educational big data," 2019, pp. 350–355, doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00071.
[6] V. U. Kumar, A. Krishna, P. Neelakanteswara, and C. Z. Basha, "Advanced Prediction of Performance of a Student in an University using Machine Learning Techniques," Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020, no. Icesc, pp. 121–126, 2020, doi: 10.1109/ICESC48915.2020.9155557.
[7] Z. Chen, X. Liu, and L. Shang, "Improved course recommendation algorithm based on collaborative filtering," pp. 466–469, 2020, doi: 10.1109/ICBDIE50010.2020.00115.
[8] M. Ashraf, "Performance Analysis and different subject combination," pp. 287–292, 2018.
[9] T. Y. Yang, C. G. Brinton, C. Joe-Wong, and M. Chiang, "Behavior-Based Grade Prediction for MOOCs Via Time Series Neural Networks," IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 716–728, 2017, doi: 10.1109/JSTSP.2017.2700227.
[10] A. F. Botelho, A. Varatharaj, T. Patikorn, D. Doherty, S. A. Adjei, and J. E. Beck, "Developing Early Detectors of Student Attrition and Wheel Spinning Using Deep Learning," IEEE Trans. Learn. Technol., vol. 12, no. 2, pp. 158–170, 2019, doi: 10.1109/TLT.2019.2912162.
[11] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," 2019 Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput., pp. 243–247, 2019.
[12] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
[13] W. Chen, C. G. Brinton, D. Cao, A. Mason-Singh, C. Lu, and M. Chiang, "Early Detection Prediction of Learning Outcomes in Online Short-Courses via Learning Behaviors," IEEE Trans. Learn. Technol., vol. 12, no. 1, pp. 44–58, 2019, doi: 10.1109/TLT.2018.2793193.
[14] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," IEEE Access, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.
[15] M. Zaffar and M. A. Hashmani, "Performance Analysis of Feature Selection Algorithm for Educational Data Mining," pp. 7–12, 2017.
[16] N. Yanes, A. M. Mostafa, M. Ezz, and S. N. Almuayqil, "A Machine Learning-Based Recommender System for Improving Students Learning Experiences," vol. 8, no. ii, 2020, doi: 10.1109/ACCESS.2020.3036336.
[17] P. Kinnunen, R. McCartney, L. Murphy, and L. Thomas, "Through the eyes of instructors: A phenomenographic investigation of student success," Third Int. Comput. Educ. Res. Work. ICER'07, pp. 61–72, 2007, doi: 10.1145/1288580.1288589.
[18] V. Tam, E. Y. Lam, S. T. Fung, and W. W. T. Fok, "Enhancing Educational Data Mining Techniques on Online Educational Resources with A Semi-Supervised Learning Approach," no. December, pp. 203–206, 2015.
[19] N. A. Butt, Z. Mahmood, K. Shakeel, S. Alfarhood, M. Safran, and I. Ashraf, "Performance Prediction of Students in Higher Education Using Multi-Model Ensemble Approach," IEEE Access, vol. 11, no. December, pp. 136091–136108, 2023, doi: 10.1109/ACCESS.2023.3336987.