Student Performance Prediction*

Krisha Aggarwal[1], Raj Aryan[2], Sudhamsu Rawlo[3], and Prof. Suruchi Deshmukh[4]
Department of Computer Science and Engineering
MIT ADT University
Pune, India
krishaaggarwal.mit@gmail.com, Rkrajaryan02@gmail.com,
Rawlosudhamsu@gmail.com, suruchi.deshmukh@mituniversity.edu.in

***Abstract*—The "Student Performance Prediction" project utilizes machine learning to understand and predict factors affecting student performance in higher education. Its main goal is to create a predictive model that helps educators address academic challenges by analyzing student attributes, including historical data, attendance, and assessments. Additionally, the project aims to develop an application that motivates students, provides personalized improvement suggestions, and offers real-time performance monitoring. This initiative strives to enhance education by leveraging technology for a more engaging and successful learning experience.**

***Index Terms*—EDM, Student Performance, Educational Datamining**

I. Introduction

In today's rapidly changing higher education landscape, the pursuit of innovation and excellence remains constant. The evolving needs of students, combined with advancements in educational technology, call for a more insightful and adaptive approach to academic support. This project is conceived to address these challenges and aims to transform the way we monitor and enhance student performance.

At its core, this project leverages the power of machine learning to unravel the intricate factors influencing student success. With a primary focus on predictive analytics, it seeks to develop a precise model for anticipating academic outcomes. By examining various student attributes, including academic history, attendance records, exam results, and other relevant data, this project aims to provide a comprehensive view of student performance.

MIT ADT University.

However, the project's vision extends beyond prediction. It envisions a comprehensive application that motivates, guides, and supports students throughout their educational journey. This application, integrated into the broader educational ecosystem, not only offers tailored insights for academic improvement but also enables real-time monitoring of student performance. It aspires to adapt to individual learning needs, creating an environment conducive to academic growth.

This introduction encapsulates the core essence of a project that strives to be a catalyst for positive change in education. It operates under the belief that technology can not only predict but also actively shape the academic destinies of students, ushering in a new era of personalized and effective learning.

II. Literature Survey

In recent years, the application of machine learning algorithms to predict student performance in higher education has gained significant attention. Several studies have explored the predictive power of various features and algorithms to enhance academic decision-making at educational institutions.

Nieto, García-Díaz, Montenegro, and Crespo (2019) delved into the predictive potential of Support Vector Machines (SVM) and Artificial Neural Networks (ANN) using a dataset of 6,130 students. Their study achieved high accuracy of 84.54

Aggarwal, Mittal, and Bali (2021) emphasized the importance of non-academic parameters in predicting student performance. They considered demographic information and previous marks, employing SVM, AdaBoost, and Random Forest. Their research produced impressive results, with SVM and AdaBoost both reaching an accuracy of 92.4

OuahiMariame (2021) focused on feature engineering and interaction with virtual learning environments. The study applied Artificial Neural Networks (NN) and outperformed several other algorithms, including Naïve Bayes, SVM, RandomForest, and ANN, in evaluating student performance.

Buenaño-Fernández, Gil, and Luján-Mora (2019) explored historical data to predict the performance of computer engineering students. They employed ensemble techniques and achieved an accuracy of 91.5

Ghorbani and Ghousi (2020) compared different resampling methods in predicting student performance, such as Random Over-Sampling, Borderline SMOTE, SMOTE, SVM-SMOTE, SMOTE-Tomek, and SMOTE-ENN. They found that a combination of the Random Forest classifier with the SVM-SMOTE balancing technique provided the best results, with 77.97

These studies collectively illustrate the growing interest in leveraging machine learning to enhance academic decision-making by predicting student performance. Various algorithms, including SVM, ANN, AdaBoost, and ensemble techniques, have demonstrated their effectiveness in this context. Furthermore, the significance of non-academic parameters and the utility of resampling methods underscore the multifaceted approach taken in these endeavors. The outcomes hold promise for optimizing educational strategies and support mechanisms to improve student success in higher education.

III. Methodology

A. Algorithms Used

In this study, we employed various classification algorithms to predict student performance. These algorithms were selected for their effectiveness in classification tasks and their potential to analyze the provided dataset. The following classification algorithms were used:

- **Support Vector Machine (SVM):** SVM is a powerful classification algorithm that seeks to find a hyperplane that best separates data into different classes. It's known for its versatility and ability to handle both linear and non-linear classification.
- **C4.5:** C4.5 is a decision tree algorithm that recursively splits the dataset based on attribute values to create a tree-like model. It's widely used in classification tasks and is known for its interpretability.
- **K-Nearest Neighbors (KNN):** KNN is a simple yet effective algorithm that classifies data points based on the majority class among their k-nearest neighbors. It's a non-parametric and instance-based algorithm.
- **J48 Decision Tree:** J48 is another decision tree algorithm that is based on the C4.5 algorithm. It's used for decision tree induction and classification tasks, similar to C4.5.
- **Multilayer Perceptron (Deep Learning):** The Multilayer Perceptron is a type of artificial neural network with multiple layers of nodes, including input, hidden, and output layers. Deep learning models can learn complex patterns and relationships in data, making them suitable for classification tasks.

These algorithms were employed to analyze the dataset and predict student performance. The choice of algorithms allows for a comprehensive assessment of classification performance.

IV. Results

Acknowledgment

We extend our gratitude to Dr. Ganesh Pathak, Prof. Suruchi Deshmukh, and MIT ADT University for their unwavering support, guidance, and resources. Their contributions have been invaluable in the successful completion of our project, 'Student Performance Prediction.' We are committed to making a positive impact in the field of education with this project.

References

[1] Hernández-Blanco, A.; Herrera-Flores, B.; Tomás, D.; Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, 2019, 1306039. [CrossRef]

[2] Bengio, Y.; Lecun, Y.; Hinton, G. (2021). Deep Learning for AI. *Commun. ACM*, 2021, 64, 58–65. [CrossRef]

TABLE I
SUMMARY OF STUDIES ON PREDICTING STUDENT PERFORMANCE IN HIGHER EDUCATION

Authors	**Features**	**Algorithm**	**Result (Accuracy)**	**Year**
Nieto, et al. (2019)	6130 students' data	SVM, ANN	84.54%	2019
Aggarwal, et al. (2021)	Demographic, previous marks	SVM, AdaBoost, RF	92.4% (SVM), 92.4% (AdaBoost)	2021
OuahiMariame (2021)	Previous Marks	Neural Networks	Outperformed various algorithms	2021
Buenaño-Fernández, et al. (2019)	Historical Data	Ensemble Techniques	91.5%	2019
Ghorbani and Ghousi (2020)	Various resampling methods	Random Forest	77.97%	2020

[3] Almarabeh, H.; Shatnawi, M.; Yasin, M.B.; Hmeidi, I. (2020). Measuring and Enhancing the Performance of Undergraduate Student Using Machine Learning Tools. In *Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS)*, Copenhagen, Denmark, 24–26 August 2020; pp. 261–265.

[4] Wakelam, E.; Jefferies, A.; Davey, N.; Sun, Y. (2020). The potential for student performance prediction in small cohorts with minimal available attributes. *Br. J. Educ. Technol.*, 2020, 51, 347–370. [CrossRef]

[5] Zeineddine, H.; Braendle, U.; Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Comput. Electr. Eng.*, 2021, 89, 106903. [CrossRef]

[6] OuahiMariame, S.K. (2021). Feature Engineering, Mining for Predicting Student Success based on Interaction with the Virtual Learning Environment using Artificial Neural Network. *Ann. Rom. Soc. Cell Biol.*, 2021, 25, 12734–12746.

[7] Reddy, P.; Reddy, R. (2021). Student Performance Analyser Using Supervised Learning Algorithms. Available online: https://easychair.org/publications/preprint/QhZK (accessed on 4 August 2021).

TABLE II
FEATURES USED IN THE STUDY

Attribute	Description
1	School (Binary: ”GP” - Gabriel Pereira or ”MS” - Mousinho da Silveira)
2	Sex (Binary: ”F” - female or ”M” - male)
3	Age (Numeric: from 15 to 22)
4	Address (Binary: ”U” - urban or ”R” - rural)
5	Famsize (Binary: ”LE3” - less or equal to 3 or ”GT3” - greater than 3)
6	Pstatus (Binary: ”T” - living together or ”A” - apart)
7	Medu (Mother’s education: Numeric)
8	Fedu (Father’s education: Numeric)
9	Mjob (Mother’s job: Nominal)
10	Fjob (Father’s job: Nominal)
11	Reason (Reason to choose this school: Nominal)
12	Guardian (Student’s guardian: Nominal)
13	Traveltime (Home to school travel time: Numeric)
14	Studytime (Weekly study time: Numeric)
15	Failures (Number of past class failures: Numeric)
16	Schoolsup (Extra educational support: Binary)
17	Famsup (Family educational support: Binary)
18	Paid (Extra paid classes within the course subject: Binary)
19	Activities (Extra-curricular activities: Binary)
20	Nursery (Attended nursery school: Binary)
21	Higher (Wants to take higher education: Binary)
22	Internet (Internet access at home: Binary)
23	Romantic (With a romantic relationship: Binary)
24	Famrel (Quality of family relationships: Numeric)
25	Freetime (Free time after school: Numeric)
26	Goout (Going out with friends: Numeric)
27	Dalc (Workday alcohol consumption: Numeric)
28	Walc (Weekend alcohol consumption: Numeric)
29	Health (Current health status: Numeric)
30	Absences (Number of school absences: Numeric)
31	G1 (First period grade: Numeric)
32	G2 (Second period grade: Numeric)
33	G3 (Final grade: Numeric)

TABLE III
DATASETS USED IN THE STUDY

Dataset	No. of Traits	Description	Date
1	32	Predict student performance in secondary education. (only math course)	11/2014
2	22	Predict end-semester percentage based on various attributes	9/2018
3	31	Predict students’ end-of-term performances at the Faculty of Engineering	8/2023
4	33	Survey of students’ math course performance in secondary school	-

TABLE IV
MEAN SQUARED ERROR (MSE) FOR DIFFERENT REGRESSION ALGORITHMS

Algorithm	**MSE**	**Algorithm Description**
Random Forest Regressor	0.163	The Random Forest Regressor achieved a remarkably low MSE of 0.163, indicating that it is well-suited for predicting student performance. This algorithm excels at capturing complex relationships within the data and offers highly accurate predictions.
SVM	0.003	Support Vector Machine (SVM) demonstrated an impressively low MSE of 0.003, reflecting its superior performance in regression tasks. Its ability to model both linear and non-linear relationships makes it a robust choice for student performance prediction.
Neural Network	1.847	Neural Network: Our Neural Network model resulted in an MSE of 1.847, which is higher than the other algorithms tested. While it showcases potential, further fine-tuning may be required to harness its full predictive capabilities.
J48 Decision Tree Regressor	46.051	The J48 Decision Tree Regressor exhibited a higher MSE of 46.051, indicating that it may not be the most suitable choice for predicting student performance in this context. Further exploration may be needed to enhance its predictive accuracy.
K-Nearest Neighbor (KNN)	0.381	K-Nearest Neighbor (KNN) achieved an MSE of 0.381, making it a competitive performer in the regression task. Its simplicity and ability to handle non-linear relationships contribute to its effectiveness.
MLP Regressor	4510.892	The MLP Regressor model resulted in a notably high MSE of 4510.892. This suggests that additional optimization and parameter tuning are required to unlock its potential for student performance prediction.