IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 4, March 2025

# Predicting Educational Performance of Students Using Machine Learning

**Kajal Solunke, Guri Sonavne, Anchal Shahane, Anuja Mirajkar, Vishesh Sanap Anushka Shinde**

Guru Gobind Singh Polytechnic Nashik, Maharashtra, India

**Abstract***: The prediction of student performance through automated means is an essential task, particularly given the extensive data present in educational databases. This challenge is being tackled by the field of educational data mining (EDM), which focuses on developing techniques for extracting valuable insights from educational data. These techniques aim to enhance the understanding of students and their learning contexts. Educational institutions often seek to ascertain the number of students likely to pass or fail in order to make necessary preparations. Previous research has indicated that many scholars concentrate on selecting suitable algorithms for classification, often overlooking the challenges encountered during the data mining process, such as high dimensionality, class imbalance, and classification errors. These issues can significantly diminish the model's accuracy. While several established classification algorithms have been utilized in this area, this paper introduces a student performance prediction model that employs a supervised learning decision tree classifier. Furthermore, an ensemble method is incorporated to enhance the classifier's performance, as ensemble approaches are specifically designed to address classification and prediction challenges.*

**Keywords:** Educational Data mining, Educational Data Mining, Predicting Student Performance, Decision Tree, Ensemble

## I. INTRODUCTION

Today, higher education institutions, particularly those specializing in computer science and engineering, encounter significant challenges in their admissions processes. It is essential for each university to develop an admissions framework grounded in valid and reliable criteria that effectively identifies candidates who are likely to excel in their programs. Furthermore, universities should employ the most advanced methods for forecasting applicants' future academic success prior to their admission. This approach would assist decision makers in establishing efficient admissions standards. Educational quality is essential for the advancement of any nation. The volume of data within the education sector is continuously increasing, facilitated by systems such as admissions, academic information management, learning management and e-learning platforms. Typically, the data gathered from students is utilized for generating straightforward queries to aid in decision- making. However, a significant portion of this data remains untapped due to its complexity and the extensive size of the datasets. Consequently, analyzing this vast array of educational data is of considerable interest for predicting student performance. Data mining, which involves extracting valuable insights from large datasets, is also referred to as knowledge discovery in databases (KDD). This technique has been effectively implemented across various fields, including banking, healthcare, and business, and is now being applied to educational contexts, a practice known as Educational Data Mining.

## II. RELATED WORK AND RESEARCH GAP

Dorina et al. introduced a predictive model aimed at assessing student performance by categorizing students into two distinct classes: successful and unsuccessful. This model was developed using the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. The findings indicated that the MLP model achieved the highest accuracy of 73.59% in identifying successful students, while the other three models demonstrated superior performance in identifying unsuccessful students. However, the model faced challenges related to high dimensionality of the data and issues with class balance.

DOI: 10.48175/IJARSCT-24055

D. Filip, F. Pop, C. Serbanescu, and C. Choi, driven by the significant interest in enhancing the use of non- general-purpose devices to achieve computational goals at a lower cost, present a novel model for scheduling microservices across heterogeneous cloud-edge environments. This model employs a specific mathematical formulation to characterize an architecture comprising diverse machines capable of managing various microservices. Recognizing that the introduction of any new model necessitates an initial risk assessment of the solution, we have enhanced the Clouds simulation framework to accommodate experiments involving such systems.

Rebai et al. introduced a model utilizing machine learning techniques to pinpoint the primary elements influencing the academic performance of schools and to analyze the interconnections among these elements. The study revealed that regression trees indicated the most significant factors linked to enhanced performance included school size, competition, class size, parental pressure, and gender ratios. Furthermore, findings from the random forest algorithm indicated that both school size and the proportion of female students significantly affected the model's predictive accuracy.

W. Jin, R. Xu, T. You, Y.-G. Hong, and D. Kim, Edge computing is an emerging computing paradigm that distributes the computational capability to the edge of networks for enabling the computation near to the environment where the sensors and actuators are deployed. Therefore, from the network edge, heterogeneous solutions can be provided to the Internet based on sufficient computing ability. Nevertheless, computing and networking resources are constrained for devices in the network edge. Providing secure services from edge computing is a challenge based on constrained resources.

A. S. Gaur, J. Budakoti, and C.-H. Lung have recently highlighted that the Internet of Things (IoT) has garnered significant attention and is anticipated to play a crucial role across various sectors in the near future, owing to its potential advantages. Nevertheless, the continuous and swift proliferation of IoT devices introduces new challenges, particularly concerning their limited power and resources. One major challenge is ensuring uninterrupted connectivity within mobile IoT environments. Additionally, IoT devices are capable of generating vast quantities of data, necessitating solutions that can effectively minimize the costs associated with data transfer. Lastly, there are obstacles related to the management and deployment of services operating at the mobile IoT Edge Gateway.

## III. EXISTING SYSTEM

Previous predictive models primarily focused on demographic information from students, including factors such as gender, age, marital status, family income, and educational qualifications. They also considered study- related elements like homework completion, study hours, and historical academic performance, including grades. However, these earlier studies were limited to forecasting academic outcomes or potential failures without delving into the underlying reasons for these predictions. In contrast, many recent studies have sought to gather over 40 different attributes during data collection to better assess a student's likelihood of academic success. These attributes typically stem from similar data categories either demographic, study related, or a combination of both leading to a lack of diversity in predictive models. Consequently, the insights into the factors contributing to student dropout rates were not fully captured by these established models. Additionally, prior predictive research from an educational psychology perspective conducted a few experiments to explore the connection between mental health and academic achievement. However, the guidelines provided were overly simplistic and did not adequately illustrate their practical application.

## IV. PROPOSED SYSTEM

The proposed framework begins by merging demographic and academic-related characteristics with aspects of educational psychology, utilizing psychological factors alongside traditionally collected data (such as students' demographic and academic information). We identified the most significant attributes based on their relevance and correlation with academic achievement, following a review of previously utilized variables for forecasting student performance. The aim of this proposal is to examine a student's longitudinal data, academic-related details, and psychological traits to assess their current status whether they are on track, facing challenges, or at risk of failure. Furthermore, we performed a comprehensive evaluation of our proposed model in comparison to earlier similar models.

*User:*

In this module, we will develop the window interfaces for the project. These interfaces facilitate secure login for all users. To establish a connection with the server, users must provide their username and password; only then will they be able to access the server. If a user already exists, they can log in directly; otherwise, they must register their information, including username, password, and email address, with the server. The server will create an account for each user to manage upload and download rates, with the username serving as the user ID. Logging in is typically required to access a specific page.

*Pre Processing:*

This initial module allows the Data User to register and log in. Upon logging in, the Data User has the capability to search for files by their names. Additionally, the Data User can download files, which will display encrypted data. Furthermore, the Data User has the option to submit a trapdoor request to the server. The server may accept this request, after which the Data User can obtain permission from the file owner, enabling the file to be downloaded in plain text.

*Classifier:*

This constitutes the second module of the project. In this module, the Data Owner is required to register and log in. The Data Owner will upload files to the database and may also send requests to the Data User.

*Training Data*:

This is the third module of the project. Within this module, the Cloud Server is able to log in. Upon successful login, it will have access to all information pertaining to data owners. The Cloud Server can also view the information of all users. Additionally, it has the capability to access all stored data files. as shown in Fig. 1

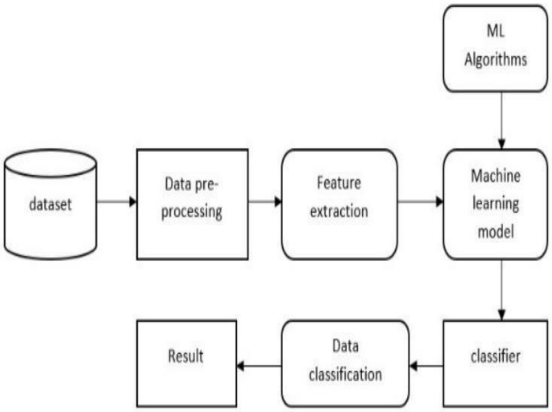

Fig 1. WorkFlow of System

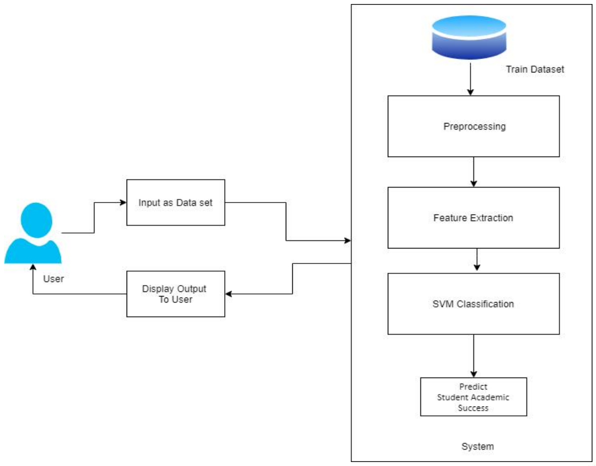

Fig 2. System Architecture

## V. METHODOLOGY

*Artificial Neural Network* (ANN)

ANN is a widely utilized method in Educational Data Mining (EDM) that aims to replicate the human brain's architecture to address intricate issues. It comprises a collection of units that process a weighted array of inputs and generate an output. Numerous studies have employed ANN to forecast student performance. We also selected this technique for its capability to identify all potential interactions among variables and its proficiency in learning from a limited number of examples. Furthermore, previous research indicated that ANN models surpassed other classification methods in accurately categorizing applicants as either accepted or not accepted. In this study, we opted for the Multilayer Perceptron (MLP) architecture for the ANN model, given that the datasets were not sufficiently large to necessitate more intricate topologies.

*Decision Tree*

A Decision Tree consists of a series of nodes organized vertically, with each node symbolizing a feature of an instance and its branches indicating potential values. This method is favored by researchers for its straightforwardness (e.g., [1, 6, 7, 9, 34]). It offers a clear and uncomplicated approach to value prediction. Additionally, it presents several benefits, such as the ability to articulate rules that are easily comprehensible to users, effective performance with both categorical and numerical data, and minimal requirements for intricate data preparation .

*Support Vector Machine*

This classification method constructs a hyperplane that distinguishes objects according to their respective classes. As the distance from the hyperplane to the nearest object increases, the generalization error of the SVM technique decreases. SVM has been utilized in a limited number of studies (e.g., [7, 23, 24]) and is employed in this research due to its effectiveness with small datasets. Moreover, it demonstrates faster performance compared to other techniques .

*Naive Bayes*

Is a straightforward probabilistic method that utilizes Bayes' theorem while assuming independence among variables. It calculates probabilities for each object across all potential classes. In this research, we opted for this method due to its ease of use, strong performance in practical applications, computational efficiency and its widespread recognition in the academic literature.

*Educational Data Mining*

The educational process encompasses various aspects of student performance. To identify students who may face challenges in achieving academic success, it is essential to predict their future performance. When data is effectively transformed into actionable knowledge, it can be utilized for making informed predictions. Consequently, this information has the potential to enhance the quality of education and facilitate students in reaching their academic goals. The field of educational data mining (EDM) employs data mining techniques to analyze information obtained from educational contexts. The application of EDM also supports the development of strategies aimed at improving student performance. Ultimately this will lead to enhanced teaching

## VI. EXPERIMENTAL EVALUATION

*Dataset*

data mining classification techniques: Artificial Neural Network (ANN), Decision Tree, Support Vector Machine (SVM), and Naive Bayes. Each model was developed using a 10-fold cross-validation method, where nine subsets of data were employed for training and one subset was reserved for testing. This procedure was executed ten times, once for each distinct subset, thereby optimizing the total number of observations used for testing. All models.

*Evaluation Metrics*

Table 1 :Dataset Definition

| Attribute Name | Attribute code | Type |
|---|---|---|
| Scode | C1 | Object |
| Gender | C2 | Object |
| Age | C3 | Int |

DOI: 10.48175/IJARSCT-24055

| | | |
|---|---|---|
| Attendance | C4 | Int |
| Performance in class | C5 | Float64 |
| Knowledge for Subject | C6 | Float64 |
| Assignment | C7 | Float64 |
| Performance in Class test | C8 | Float64 |
| Edu_Father | C9 | Int |
| Edu_Mother | C10 | int |
| Job_status_ Mother | C11 | Int |
| Job_Status_Father | C12 | Int |
| Income | C13 | Int |
| Help Chld to learn Lessons | C14 | Int |
| Motivate_chld | C15 | Int |
| PT_interaction | C16 | Int |
| Timespend with SM,MP,Inet | C17 | Int |
| Own a smart device | C18 | int |
| Study pattern at home( PV) | C19 | Int |
| Use of College Library for study | C20 | Int |
| Positive use of mobile phone | C21 | Float64 |
| Study pattern at home(SV) | C22 | Int |
| Percentage of marks obtained | C23 | Int |
| Doing Part time job | C24 | Float64 |

A predictive model is developed utilizing the Linear Regression Algorithm. The model is trained on 70% of the dataset, while the remaining 30% is randomly selected for testing purposes. The implementation is carried out using the Panda tool. The prediction score reflects the anticipated values generated by the linear regression models for various target data. The same features are evaluated against different target classes. For the initial model training, the features from the first semester are employed. When the target class corresponds to the first semester marks(FSM), the predictive model achieves an accuracy of 0.98, with a Mean Absolute Error of 1.85, a Mean Squared Error of 6.27, and an R² score of 0.98, indicating that the model is nearly flawless. However, the primary objective is to forecast the future performance of students. Consequently, the second model is trained using the second semester marks(SSM)while maintaining the same selected features.

| Models | InMode1 | In Mode 2 | In Mode 3 | InMode l4 | In Mode l5 |
|---|---|---|---|---|---|
| | 1st Semester | 2nd Semester | Average of(FSM, SSM) | Highest | Lowest |

Table 2 :Prediction Scores

Accuracy represents the proportion of correctly predicted outcomes and is calculated The formula for calculating the overall score is represented as follows: first semester mark(FSM), To verify the accuracy of the models, the average accuracy of all five models is calculated as.

| Target Class | 1st Marks (FSM) | 2nd Marks (SSM) | FSM and SSM | marks of the FSM and SSM | marks of the FSM and SSM |
|---|---|---|---|---|---|
| Accuracy of in this the Prediction | 0.98 | 0.92 | 0.96 | 0.96 | 0.97 |
| Mean absolute error | 1.85 | 3.92 | 2.52 | 2.89 | 2.24 |
| Mean squared error | 6.27 | 24.5 | 9.69 | 12.98 | 9.01 |
| $R^2$ score | 0.98 | 0.93 | 0.97 | 0.96 | 0.97 |

DOI: 10.48175/IJARSCT-24055

$Average$ Accuracy= 0.98+0.92+0.96+0.96+0.97 = 0.958

To assess the validity of this prediction score, Models 4 and 5 have been developed. Model 4 establishes the Target Class by utilizing the highest scores from both the first and second semesters, while Model 5 determines the Target Class based on the lowest scores from the same semesters. Model 4 achieved an accuracy of 0.96, with a mean absolute error of 2.89, a mean squared error of 12.98, and an $R^2$ score of 0.96. In contrast, Model 5 attained an accuracy of 0.97, with a mean absolute error of 2.24, a mean squared error of 9.01, and an $R^2$ score of 0.97. To further validate the models' accuracy, the average accuracy across all five models is computed.

## VII. CONCLUSION

The performance of students is a critical concern that poses significant challenges. This paper offers an analysis of data mining research results aimed at creating predictive models for student performance. Our study demonstrates the application of machine learning algorithms to enhance the understanding of algorithm efficiency, utilizing data. The analysis incorporates four key factors, including admission scores, first-level courses, and academic achievement tests

## VIII. FUTURE WORK

In the future, we aim to leverage deep learning architectures to enhance prediction capabilities and overall performance. This method facilitates the incorporation of both academic and non- academic elements. Enhancing the early forecasting of student success in higher education necessitates the implementation of several key strategies. First, by adding diverse data sources such as extracurricular activities and personal interests, institutions can develop a more holistic view of each student's profile, resulting in more precise predictions. Second, implementing real- time monitoring and feedback mechanisms allows educators to act swiftly when they detect signs of academic difficulties, thereby preventing problems from worsening and enriching the learning experience for students. Additionally, integrating predictive models with current learning management systems simplifies data analysis and promotes effective communication between predictive analytics and educators. Furthermore, using predictive analytics, personalized learning pathways can be crafted to align with each student's unique strengths and weaknesses, thereby enhancing student engagement and success. Collaboration among educational institutions and researchers is vital for sharing knowledge and advancing the field of predictive analytics in higher education.

## REFERENCES

**[1].** Dorina Kababchieva. (2012). Student Performance Prediction using Data Mining Classification Algorithms. International Journal of Computer Science and Management Research, vol. 1.

**[2].** Analysis of machine learning strategies for prediction of passing undergraduate admission test.

**[3].** Predicting Academic Success of College Students Using Machine Learning Techniques. Jorge Humberto Guanin- Fajardo 1, Javier Guana-Moya 2and Jorge ˜ Casillas 2024, 9, 60.

**[4].** Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems.Hanan Abdullah Mengash DOI 10.1109/ACCESS.2020.2981905, IEEE

**[5].** Development Roadmap for Launching Online Education: A Case Study of an Online Graduate Certificate Course",Kevin Kam Fung Yuen,AMY Ooi Mei Wong

**[6].** M. Al-Saleem, N. Al-Kathiry, S. Al-Osimi, and G. Badr, "Mining Educational Data to Predict Students' Academic Performance," Springer, Cham, 2015, pp. 403–414.

**[7].** Mining Learning Based Predicting Student Academic Success Khalfan Al Mayahi November 03,202

**[8].** Early Prediction of Students' Academic Achievement: Categorical Data from Fully Vol. 48. No. 9. September 2021

**[9].** Student Performance Prediction Using Machine Learning.doi:10.3233/APC210137

**[10].** Educational data mining: prediction of students' academic performance using machine learning algorithms,Mustafa Yagcı.

**[11].** Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning"2020 IEEE International Conference for Innovation in Technology (INOCON) Bengaluru,

**[12].** India. Nov 6-8 2020

**[13].** Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices" International Journal of Educational Technology in Higher Education. 2020, doi: 10.1186/s41239-020-0177-7.

**[14].** M. Al luhaybi, A. Tucker, and L. Yousefi, "The Prediction of Student Failure Using Classification Methods : A Case study," 2018, doi: 10.5121/csit.2018.80506.

**[15].** S. Fedushko and T. Ustyianovych, "Predicting Pupil's Successfulness Factors Using Machine Learning Algorithms and Mathematical Modelling Methods" in Advances in Intelligent Systems and Computing, 2020.

**[16].** Early Prediction of Students Performance in Higher Education Dr. Geeta Tripathi* , Kethati Nandini Reddy , Kotagoda Sahithi, Maddineni Ajay Kumar doi: https://doi.org/10.32628/IJSRST

**[17].** N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., 2002, doi: 10.1613/jair.953.

**[18].** Early Predicting of Students Performance in Higher Education ESSA ALHAZMI AND ABDULLAH SHENEAMER