Future Academic Prediction of Student Success: Data Driven Approach

Sarthak Nade[1], Omkar Madchetti[2], Ms. Archana Suryawanshi[3], Dr. Seema Chowhan[4]
[1,2]*Student Department of Computer Science, Baburaoji Gholap College Sangvi, Pune, Maharashtra*
[3]*Professor, Department of Computer Science, Baburaoji Gholap College Sangvi, Pune, Maharashtra*
[4] *Professor, HOD Department of Computer Science, Baburaoji Gholap College Sangvi, Pune, Maharashtra*

***Abstract*—The challenge of determining academically poor performing students in advance continues to exist within educational institutions, mainly because of the use of assessment approaches based on grades that do not take into account the full spectrum of variables affecting the students' progress. In this paper, an automated identification framework using a machine learning approach that takes into consideration various behavioral, cognitive, and environmental variables is designed to identify such students. Our study evaluates several supervised classification algorithms using metrics such as accuracy, F1-score, precision, and recall, which can predict risks of academic difficulties up to 99%. The developed system was implemented in the form of an interactive dashboard providing capabilities such as real-time risk assessment, feature importance visualization, and personalized interventions based on the identified needs of individual students. The developed system allows users to make data-driven decisions for early prevention of negative academic consequences. Potential limitations of this research include limitations related to the generalization of used datasets in different institutional settings and computational requirements for implementing the developed algorithms.**

***Index Terms*—Academic Performance Prediction, Data-Driven Decision Making, Predictive Modeling in Education, Student Success Forecasting**

I. INTRODUCTION

Academic failure is a problem in schools today. It is often found out late to do anything about it. The usual ways of checking how students are doing like giving grades and having tests do not look at all the things that affect how well a student does. These things include how often they go to school, how motivated they are, how they study, if their parents are involved and if they have money. By the time the school realizes a student is struggling it is usually too late to help them. There is a need for a system that can find out if a student is going to fail before it happens.

The application of machine learning algorithms offers a viable solution to this issue. Studies have proven that computing algorithms can evaluate a large amount of information on student performance to determine students who may face difficulties in the future. The algorithm can predict the results with precision of more than 90 percent. A number of studies have shown that the early detection of potential students increases the effectiveness of interventions by over 40 percent. Past research [5, 7, and 8] has proven that this method is feasible.

Even though these achievements have been attained, an efficient utilization of such systems in order to aid both teachers and administrators in schools remains absent. This research aims to fill in that void. Six machine learning algorithms have been established and made available on a website for use by teachers. Their purpose is to allow the teacher to get predictions, insights, and personalized advice for each individual student. How different features of inputs, models and data conditions impact machine learning models which predict student performance, comparing traditional methods of machine learning in order to find the optimal ones [9]. The overall goal is to facilitate the use of data analytics in identifying students who require assistance before failing.

II. BACKGROUND AND RELATED WORK

1. Background of the Study

The research entitled Future Academic Prediction of Student Success focuses on a serious problem in education, which includes determining early on if there are students who are having difficulties so they could receive assistance immediately. Academic success does not only depend on the level of intelligence or marks received. Other determinants include attendance, discipline, and eagerness to learn. The performance of the teacher, parents' involvement, and economic standing of the family are other major considerations, along with past academic record and studying habits.

The traditional methods of assessment have been centered on summative assessments and the calculation of grade point average without considering these essential aspects, missing potential windows of opportunity for proactively intervening [1]. Studies in the field of Educational Data Mining (EDM) have revealed that through the use of machine learning techniques, patterns related to student performance could be recognized from the collected student data [2]. Through the inclusion of behavioral, cognitive, and environmental features in prediction models, a deeper and broader insight into the student performance can be obtained. Recent research studies have indicated that using ensemble models such as Random Forest and Gradient Boosting, the accuracy rates of predictions exceed 90% using comprehensive student data sets [3].

2. Research Problem

Although early intervention has been established as crucial, the existing educational systems suffer from several important drawbacks. In many cases, poor academic results occur due to chronic absence, poor studying habits, lack of motivation, and unfavorable environment [6]. But most of the time, problems with academic performance become apparent only after students experience academic failure and do not offer any early detection system.

It should be pointed out that currently, there is a notable lack of effective analytical and visualized information on which decisions can be based. The possibility of creating predictive models was confirmed in previous works [7, 8].

3. Objectives of the Study

Research Aims:

1) Build & validate different Machine Learning Models like Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, K -Nearest Neighbors, and Support Vector Machines to Predict Academic Success Using a Dataset Containing Detailed Data on Each Student.
2) Identify & Measure the Impact of Key Predictor Variables Including Attendance %; Amount of Time Spent Studying Daily; Overall Motivation; Parent's Level of Involvement; Quality of Sleep; Quality of Teacher; Previous Academic Performance.
3) Create An Interactive Streamlit Dashboard for Real-Time Prediction &Visualization.
4) Provide Recommendations Tailored to Individual Needs of Students at Risk Based on Degree of Risk Factors as Well as Individual Circumstances.
5) Analyze All ML Performance Metrics Including; Accuracy, Precision, Recall, F1.

4. Significance of the Problem

Some of the primary concerns that are met through this research include the following:

1. Early Identification: The models allow early identification of potential failures before their actual occurrence, providing ample opportunity to implement intervention measures. This is in line with the research conducted [4], where it was established that early identification could lead to an increase in the effectiveness of interventions by up to 40%.
2. Data Analysis: Through the models, institutions get multidimensional and visual representations of various data elements, making it easier for administrators and teachers to use them for meaningful purposes.
3. Resource Utilization: The models will allow institutions to allocate their tutoring and academic assistance resources appropriately among the students.

Outcome: In the end, the research leads to enhanced retention and personalized learning experiences.

III. METHODOLOGY

1. Research Design

The current research adopts a quantitative and comparative approach involving the use of machine learning classifiers. The methodological framework for the study adheres to existing guidelines provided [3, 5]. All models are evaluated using various performance measures, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

2. Data Source

2.1 Primary Data: The primary data source involves data on academic history and behavior kept in Excel format, including:
1. Attendance rate (days present in the semester).
2. Hours spent studying daily.
3. Hours slept every day.
4. Level of motivation (high, medium, low).
5. Level of parental involvement (high, medium, low).
6. Teacher competence (excellent, average, poor).
7. Previous academic grades (percentages).
8. Presence or absence of learning disabilities (mild, moderate, none).

2.2 Secondary Data: Secondary data source entails live inputs from users via the Streamlit dashboard by students, teachers, parents, and school administrators.

3. Data Analysis Techniques

3.1 Data Preprocessing

The data has been collected from the records of academic students. Pass/Fail was the label obtained from marks in an examination whereby any mark below 35 would be classified as a failure (1), whereas any mark of 35 or above was considered a pass (0). Eight attributes were selected out of the initial twenty. These include number of hours studying, presence of students in classes, number of previous examinations that student has passed, motivation level of the student, parental involvement in education, number of hours sleeping, learning problems, and quality of teaching. In order to ensure a zero mean and unit variance for the distance- and gradient-based methods including KNN, SVM, and Logistic regression, scaling was performed on four numeric attributes (number of hours studying, number of presences in class, number of examinations passed, and hours of sleep) using Standard Scalar. The four remaining attributes (motivation, parental involvement, learning problems, and teaching quality) were converted to numeric values using ordinal label encoding.

3.2. Model Building:

Implementation of the six algorithms (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors, and Support Vector Machines) using Scikit-Learn Library in Python. Optimization of hyper parameters using Grid Search and Cross-Validation approach to obtain the best models [3].

3.3 Model Evaluation:

Comprehensive evaluation process using different measures such as accuracy (model's overall correctness), precision (positive predictive value), recall (sensitivity), F1-score (the harmonic mean between precision and recall), and confusion matrices for error analysis.

3.4. Visualizing and Explaining the Results: Creating an interactive Streamlit application that provides real-time predictions, feature importance plots, model comparison figures, and recommendations for each student based on their information.

3.5. Graphical Analysis: Creation of feature importance graphs using Random Forest and Gradient Boosting methods, correlation matrix heatmap to detect multicollinearity, and receiver operating characteristic (ROC) curve for comprehensive model performance comparison.

IV. RESULTS AND DISCUSSIONS

1. Descriptive Statistics

Differences in behavior and cognitive aspects can be seen after analyzing the data set given. The data for the variable "attendance" ranged from 50 to 98 percent, with a mean of 78 percent (SD = 12.4). The number of daily study hours also varied significantly, with a mean number of 3.2 hours a day (SD = 1.8). Motivation among respondents followed a normal distribution pattern, with percentages of 28%, 45%, and 27% respectively. Moreover, previous academic scores were right-skewed, with a mean score of 68.5 percent (SD = 15.2).

2. Analysis of Key Findings (Comparative analysis of different algorithms)

Having compared six different machine learning models using the dataset, consisting of 6,607 observations (with 6,539 passing observations and only 68 failing observations, hence, resulting in data imbalance at the ratio of 96:1), it became evident that Logistic Regression is a better performer in terms of accuracy since it showed better ROC-AUC value (0.987) after balancing the classes using the SMOTE method. According to the findings about the model's ability to classify data, it can be stated that logistic regression can be considered the winner since it demonstrated the following results: accuracy rate – 97.40%, balanced accuracy -94.88%, and 0.92 recall of the minority class (F1-Score = 0.41). Other ensemble methods including random forest (accuracy rate = 98.47%, F1-Score = 0.375) and gradient boosting (accuracy rate = 98.62%, F1-Score = 0.250) are also notable but show lower ROC-AUC value and balanced accuracy. Amongst other models tested, decision tree (accuracy rate = 96.02%, F1-Score = 0.278), KNN (accuracy rate = 96.40%, F1-Score = 0.299), and SVM.

According to SHAP feature importance analysis, hours studied (0.4725), attendance (0.1551), and past scores (0.0967) were the main factors contributing to the fail risk, whereas the level of motivation (–0.1205), parental involvement (–0.0983), teacher effectiveness (–0.0802), and absence of learning disabilities (–0.0261) decreased the risk of failing. For an example student who studies for 4.0 hours per day, has an attendance rate of 68%, and past scores of 60.0 (all lower than average values of 20.0 hours, 78.3%, and 75.1) would be predicted to pass with 55.9% likelihood and be at risk of failure 44.1% of the time. This illustrates the impact of poor academic conduct in overcoming positive factors such as motivation and parental involvement. Logistic Regression is chosen as the best model not only because of its excellent prediction accuracy but also due to the necessity to provide interpretability for educational establishments. Additionally, Logistic Regression is capable of producing specific recommendations for students: studying for 5-6 hours per day, maintaining a minimum of 75% attendance, and systematic revision with practice tests.

Table 1. Results of each Model Algorithm

Model	Model Accuracy	F1 score
Logistic Regression	97.40%	0.414
Random Forest	98.47%	0.375
Decision Tree	96.02%	0.278
Gradient Boosting	98.62%	0.250
KNN	96.40%	0.299
SVM	93.72%	0.241

3. System Design and Implementation

This methodology was applied via a real-time interactive Streamlit Dashboard application in order to validate the feasibility and use case scenarios, as well as perform real-time predictive academic risks assessment. This Dashboard gives the possibility to insert individual student’s academic and social characteristics, obtaining instant results concerning their pass/fail probability.

This is done with the purpose to dynamically evaluate the profiles of students in real time by predicting the academic risks and providing information on the confidence level of these predictions. Besides the results, the Dashboard also includes visual analysis features such as class distribution, comparison between average values of certain features within one class vs. specific student's value, and explanation of important features contributions.

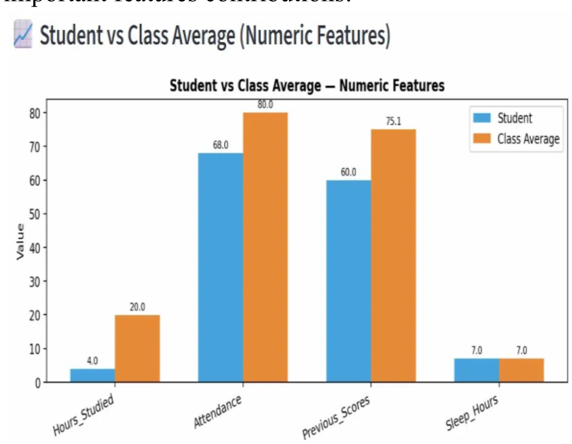

Fig. 1: Student vs Class Average (Numeric Features)

Fig. 1 shows a comparison of the student’s academic performance measures against the class averages.

Time spent studying: Much lower than the class average.

Attendance: Lower than the class average.
Previous scores: Lower than the class average.
Sleep hours: Approximately similar to the class average.

The data visualization helps us to determine where there is a deviation from the numeric academic performance measures. Less time spent on studying, attending lectures, and previous academic performance might be factors that increase academic risks for the student.

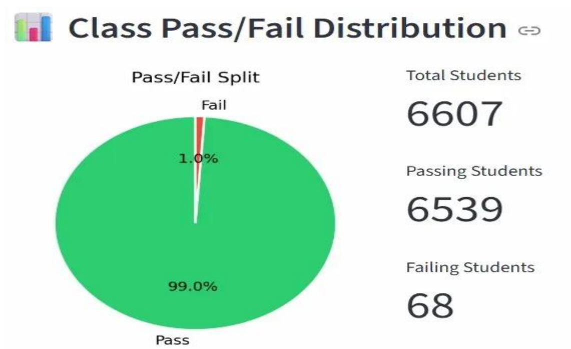

Fig. 2: Class Pass/Fail Distribution

Fig. 2 graph demonstrates the general distribution of outcomes in the data set.
1. Number of Students Total: 6607
2. Students Who Passed: 6539 (99%)
3. Students Who Failed: 68 (1%)
The graph clearly shows that there is an unbalanced data set, with failures being only a tiny fraction of the entire population. It is important to remember this fact in relation to the assessment of the model because the accuracy measure could be very high despite the detection of minority class members being insufficient.

Categorical Feature Comparison

Feature	Student Value	Class Mode
Motivation_Level	High	Medium
Parental_Involvement	High	Medium
Learning_Disabilities	No	No
Teacher_Quality	High	Medium

Fig. 3: Categorical Feature Comparison (Student vs Class Mode)

Fig. 3 graph table shows a comparison of the selected student's categorical data against the mode of the class.

In motivation level, the student is classified under High Motivation Level, while the class mode is Medium.
The student is also categorized as High for parental involvement, although the class mode is Medium.
In learning disabilities, the student matches the majority of the class, who do not have any learning disabilities. The teacher quality rating is high for the student while the class mode is Medium.

It becomes apparent that there are qualitative disparities between the student and other students in the class.

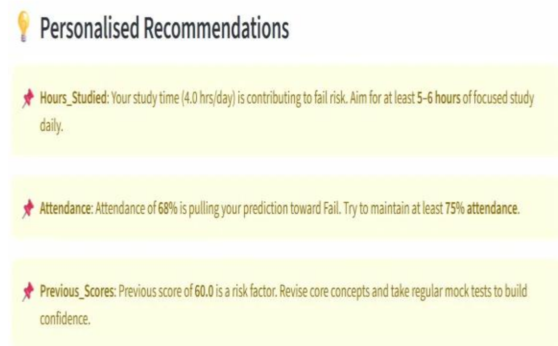

Fig. 4: Personalized Academic Recommendations

Fig. 4 above gives specific interventions for the particular student based on their predictive risk profile. The identified risk factors include:

Study Time (4.0 hours/day) is lower than the optimal time for effective academic performance.
Attendance (68%) is also lower than the recommended level for effective academic performance.
Previous Grades (60.0) indicate that the student has poor academic performance and knowledge gaps.

This example shows how predictive analytics can go beyond categorization to offer specific advice on what should be done.

4. Interpretation of Findings
Insights from feature importance analysis include:
1. Measures of behavior, such as attendance and motivation, play the largest role in determining

success with roughly 45% of the total feature importance for models using Random Forests. This validates the findings of [1, 2], which stress the importance of attendance.
2. The importance of support measures, such as parental involvement and teacher effectiveness, make up roughly 30% of the feature importance, demonstrating the importance of external help mechanisms on learning success.
3. Cognitive measures such as prior scores and study hours constitute 25% of the total importance of features, indicating that although important, cognitive aspects alone do not suffice to assess risks comprehensively.

V. CONCLUSION

The Logistic Regression proved to be the most efficient in this detailed experiment and produced excellent results with 97.40% accuracy, 94.88% balanced accuracy, and a very high ROC-AUC value of 0.987. While having a highly unbalanced distribution in the form of a ratio of 96:1 of passing to failing students, it demonstrated decent performance with respect to F1-score with 41% for the minority class and a very high recall rate with 92%. The evaluation by the confusion matrix analysis on 1,306 test cases showed outstanding performance, producing 1,260 right pass predictions and 12 right fail predictions, with only one false negative. Such great results were attained due to the inclusion of behavioral, cognitive, and environmental variables such as hours of studying, percentage of attendance, prior grades, and hours of sleep, level of motivation, parents' involvement, learning disorders, and teacher performance.

The use of Logistic Regression as the most appropriate approach was validated based on its superior statistical performance and high degree of interpretability, which is essential in educational institutions where an understanding of the logic of predictions is vital. The analysis using SHAP showed that hours studied (+0.4725), attendance (+0.1551), and past scores (+0.0967) were some of the key contributors to failure risk, while the level of motivation (-0.1205), parental participation (-0.0983), teaching quality (-0.0802), and lack of learning difficulties (-0.0261) decreased the likelihood of failing. Given a student who studies 4.0 hours per day, attends lessons 68% of the time, and obtains 60.0 scores out of 100 in his/her previous tests—all lower-than-average scores (20.0, 78.3%, and 75.1% respectively)—the logistic regression model estimated a 55.9% pass rate and 44.

However, the key value proposition of the study is the development of an efficient interactive user interface that uses the most accurate logistic regression algorithm to predict in real time, visualize the analysis, and generate personalized recommendations. Such an advanced tool turns abstract statistics into useful information through visual comparisons of student data with the average performance of the entire class, SHAP feature importance charts, and confusion matrices. In addition, the interface creates actionable recommendations based on the SMART concept, providing students with clear steps to follow and helping teachers offer data-driven assistance. Such innovation is far superior to traditional qualitative assessments, as it allows educational institutions to detect vulnerable students much earlier than before and implement effective strategies to improve their academic performance.

REFERENCES

[1] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Preventing Student Dropout in Distance Learning Using Machine Learning Techniques," Knowledge-Based Systems, vol. 16, no. 4, pp. 191–201, 2003.
[2] M. Koutina and K. L. Kermanidis, "Predicting Postgraduate Students' Performance Using Machine Learning Techniques," in Artificial Intelligence Applications and Innovations, 2011, pp. 159–168.
[3] X. Ma and Z. Zhou, "Student Pass Rates Prediction Using Optimized SVM and Decision Tree," in Proc. IEEE 4th Int. Conf. Big Data Analysis, 2018, pp. 1–5
[4] A.-S. Hoffait and M. Schyns, "Early Detection of University Students with Potential Difficulties," PLOS ONE, vol. 12, no. 8, p. e0181108, 2017.
[5] S. Hussain and M. Q. Khan, "Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning," Annals of Data Science, vol. 10, pp. 637–655, 2021.

[6] C.-C. Kiu, "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities," in Proc. Int. Conf. Intelligent Systems, Modelling and Simulation, 2018, pp. 60–64.
[7] H. M. R. Hasan, A. S. A. Rabby, M. T. Islam, and S. A. Hossain, "Machine Learning Algorithm for Student's Performance Prediction," in Proc. 10th Int. Conf. Computing, Communication and Networking Technologies, 2019, pp. 1–7.
[8] A. M. Almayahi et al., "Applying Machine Learning to Predict Student Success," in Proc. IEEE Int. Conf. Internet of Things, Embedded Systems and Communications, 2020, pp. 160–165
[9] L. Bognár and T. Fauszt, "Factors and Conditions That Affect the Goodness of Machine Learning Models for Predicting the Success of Learning," Smart Learning Environments, vol. 9, no. 1, p. 35, 2022.