Journal of Theoretical and Applied Information Technology
15th October 2024. Vol.102. No. 19

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING ACADEMIC PERFORMANCE PREDICTION FOR AT-RISK STUDENTS: COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS IN EARLY WARNING SYSTEMS

ZAINAB MAHMOUD [1], ABDELRAZEK WAHBA SAYED [2]

[1] Assistant Professor, Jeddah International College, Computer Science and Information Technology Department, Saudi Arabia

[2] Assistant Professor and Head of Law Department, Jeddah International College, Law Department, Saudi Arabia

E-mail: [1] z.rabea@jicollege.edu.sa, [2] a.wahba@jicollege.edu.sa

ID 55515 Submission	Editorial Screening	Conditional Acceptance	Final Revision Acceptance
07-09-2024	08-09-2024	23-09-2024	14-10-2024

ABSTRACT

Academic institutions increasingly leverage technology to enhance student performance, particularly through early warning systems that identify at-risk students. These systems utilize various academic and non-academic factors, including grades and attendance, to forecast performance. This study employs a core dataset from Jeddah International College, consisting of 224 instances and 19 attributes, to evaluate the predictive power of several machine learning algorithms. We conduct a comparative analysis of Gaussian Process, Decision Trees, Linear Regression, Ridge Regression, Gradient Boosting, Random Forest, Support Vector Regression, AdaBoost, and LASSO Regression to rank their performance in identifying at-risk students. Our findings reveal that the Gaussian Process and Decision Trees demonstrate the highest predictive capabilities, achieving the highest R^2 value (0.9657) and the lowest error metrics (RMSE: 0.0424, MSE: 0.0018, MAE: 0.0149). This research outlines the criteria for selecting the most effective models to support academically struggling students.

Keywords: *Machine Learning Algorithms; Academic Metrics; Early Warning System; Students' Performance Prediction.*

1. INTRODUCTION

In today's higher education arena, students enroll in various courses at institutions of higher learning. But for a variety of reasons, some students are unable to finish their courses, which causes many of them to drop out in the middle. Some will not be able to meet the requirements for a Grade Point Average (GPA) and will be expelled from the school. Over the ensuing academic years, this attrition continues, with only 45% of enrolled students eventually graduating[1]. Retention rates are still low despite multiple attempts to raise student success and retention[2].

The core problem addressed in this study is the high dropout rate among students, which is a result of both poor academic performance and a failure to recognize at-risk pupils early in the term. This emphasizes the requirement for efficient predictive models that assist student retention by utilizing both academic and non-academic characteristics. In order to do this, the study raises a number of important research questions: How well can different machine learning algorithms predict which kids are at-risk based on both academic and non-academic factors? Which machine learning models predict student achievement with the best degree of accuracy? Lastly, how can the research results be applied to create a workable early warning system that will enable educators and institutions to provide successful at-risk student support?

Academic success is the most important component in student retention and the best indicator of students' continuity. Therefore, one way to increase retention is to enhance academic success[3]. To boost academic success, identifying at-risk students is imperative. Several of the

student's academic and nonacademic criteria can be used to forecast how well they will perform early in the semester. A predictive model is used as an early warning system to identify at-risk students in the course and notify both teachers and students. After that, instructors can engage with at-risk students in a number of ways to help them improve their performance in the course. Using the early warning system and the course's intervention instructions can help students succeed in the course [4].

To forecast student accomplishment in a course, a general predictive model may yield inaccurate results because learning objectives, activities, and evaluations may vary greatly throughout educational institutions. Additionally, the number of students who fail the course is less than the number of students who pass because there are many students who are disqualified due to absences and other reasons, as well as those who withdraw from the course[5].

In this paper, predictive models are constructed using both academic and non-academic criteria to forecast student performance levels. A number of well-known algorithms were used to investigate their efficacy. LASSO Regression, AdaBoost Regression, Random Forest Regression, Gradient Boosting Regression, Decision Trees Regression, Linear Regression, Ridge Regression, Random Process Regression, and Gaussian Process Regression. These algorithms were chosen because they are widely applied and can identify both linear and nonlinear correlations in the data. We evaluated these algorithms using a dataset that was obtained from the law program at Jeddah International College. It included academic metrics like attendance, assignments, monthly tests, GPA, and so on, along with non-academic characteristics like gender and country. Each student's performance in this course will be assessed out of a possible 100 points; 60 will be given for participation, attendance, assignments, and examinations throughout the semester, and 40 will be set aside for the final exam.

Academic criteria are given more weight than non-academic variables when forecasting a student's academic performance. We performed a thorough analysis to assess each algorithm's performance and compare them using a variety of assessment measures, such as R2, RMSE, MSE, and MAE. The purpose of this study's conclusions is to offer important insights into the best machine learning strategies for precisely projecting exam scores for pupils. All things considered, this work adds to the body of literature by carefully assessing and contrasting the accuracy of various machine learning algorithms in predicting students' academic success. These insights have the potential to significantly improve instructional methodologies and support mechanisms for educators as well as institutions.

The predictive system emails the following parties when it finds evidence of impending academic failure: the student (notifying them of potential hazards and offering suggestions for improvement); the parent or guardian (notifying them of the student's performance to date and the likelihood of failure); the subject lecturer (notifying them of performance details and suggestions for support); and the academic advisor (notifying them of a thorough report and suggestions regarding possible course withdrawal). By doing this, prompt information regarding required interventions and support measures is ensured.

The structure of this paper is as follows: Section 2 reviews various related works on predicting student academic performance. Section 3 introduces a methodology divided into components such as data collection, analysis, preprocessing, machine learning framework, and notification system implementation. Section 4 presents experimental results from different models assessed using standard metrics. Section 5 concludes by summarizing findings, discussing current limitations, and suggesting future research directions in the field.

2. RELATED WORK

Many studies have been carried out to pinpoint the elements linked to and indicative of university student's academic achievement in coursework. Most of these studies have concentrated on forecasting students' final semester grades in courses based on academic information that was available before the semester began (grades in prerequisite courses, cumulative GPA, etc.) and non-academic information that was available at the beginning of the semester (such as socioeconomic status, age, gender, etc.)[6]. Although these studies offer insightful information, they frequently mainly rely on historical data that is static and might not accurately represent the dynamic nature of student achievement. Neither teachers nor students can influence previous performance indicators and non-academic factors. If students are informed that these factors are used in predictive models, it may dampen their motivation because they may believe that their behavior or previous circumstances have predetermined them for failure and that they cannot

do anything to achieve positive outcomes in the future. Therefore, despite their intention to assist students, these models may negatively impact students' performance [4].

Aggarwal et al. [7] compares two models: one constructed solely using academic parameters and the other incorporating both academic and non-academic (demographic) parameters. The dataset comprises information on 6,807 students. Their findings underscore the importance of integrating both types of parameters, suggesting that this dual approach can significantly improve predictive outcomes. Eight classification algorithms are compared to identify the parameters that contribute to creating the most suitable model for classifying students based on their performance. The findings indicate that non-academic parameters cannot be disregarded; relying solely on academic parameters is insufficient. Only when both academic and non-academic criteria are incorporated will the best outcomes be obtained.

Huang et al. [8] studied methods to predict student grades in an Engineering Dynamics course. They used 323 students' amount of data to compare four prediction techniques. According to the study, predicting final grades with midterm exam scores and cumulative GPA had a 64% accuracy rate. This research emphasizes the shortcomings of conventional predictors, which might not fully account for the range of student performance. Accuracy was not appreciably increased by adding grades from necessary courses. 52.5% accuracy was attained by using simply the first midterm exam and a precise procedure; this is comparable to using cumulative GPA or previous grades. The study emphasizes the usefulness of utilizing performance data from the semester for projections, and it suggests that including grades from assignments and quizzes could improve accuracy and enable early predictions.

Marbouti et al . [4] compared predictive methods to identify at-risk students in a course using standards-based grading. They utilized only in-semester performance data available to instructors, aiming to minimize false negatives (type II errors) without significantly increasing false positives (type I errors). This approach aligns with the growing recognition of the need for real-time data in predictive analytics. Employing a feature selection method to reduce variables, the Naive Bayes Classifier and an Ensemble model showed the best results among seven tested methods.

Yılmaz et al . [9]applied artificial intelligence techniques to the results of a questionnaire that included key indicators from three different courses across two faculties. The goal was to classify students' final grade performances and determine the most efficient machine learning algorithm for this task. Several experiments were conducted, and the results suggest that the Radial-Basis Function Neural Network can be effectively used for this purpose, helping to classify student performance with an accuracy of 70%–88%. This highlights how cutting-edge machine learning algorithms have the potential to improve forecast accuracy over more conventional approaches.

It is clear from reviewing several studies on predictive analysis utilizing various educational data mining techniques that prediction models usually consider two categories of parameters:

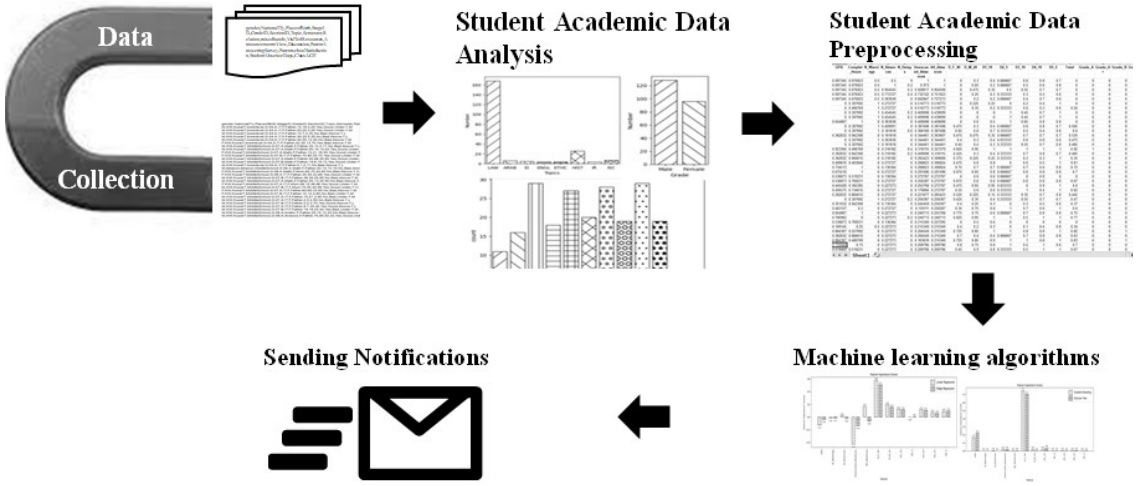

Figure 1: Early Warning Systems Framework.

academic and non-academic. This study, on the other hand, seeks to broaden the scope by adding more dynamic elements that demonstrate continued student involvement and engagement. The study demonstrates how to create an efficient prediction model using machine learning methods including Support Vector Machine [10, 11], Logistic Regression[12], J48 Decision Tree[13], Random Forest [14], and Artificial Neural Networks [15-17]. The effectiveness of these algorithms in combining both academic and non-academic parameters to enhance the accuracy and reliability of student performance predictions will be evaluated against more comprehensive datasets and innovative methodologies to highlight the unique contributions of this research.

3. MATERIAL AND METHOD

This research employs a systematic methodology encompassing key steps aimed at developing an intelligent predictive system. The process involves data collection, rigorous data analysis, meticulous data preprocessing, implementation of diverse machine learning algorithms, and the establishment of a notification mechanism (See Figure. 1). Each phase is meticulously designed to enhance the overall effectiveness and efficiency of the predictive system. Its goal is to ensure precise predictions and timely alerts to both parents and lecturers regarding potential academic challenges students may face.

3.1 Data Collection

The data for this study consists of student details taken from the law program at Jeddah International College. Table 1 show the attributes contained in the dataset. The student details include demographic information such as gender and nationality, in addition to academic details. There are 19 columns and 224 entries in the dataset. The student's grades for the entire semester are represented by (D_F_40, D_M_20, D1_10, D2_5, D3_10, D4_10, and D5_5).

3.2 Data Analysis

Analyzing and comprehending different facets of the dataset is necessary to predict student academic achievement since it provides insights into the variables affecting student results. In order to comprehend distributions, central tendency, and data variability, an examination of the quantitative variables presented in Table 2 comprises statistical analysis using metrics like mean, median, 25th, 50th, and 75th percentiles, as well as standard deviation.

Table. 1: Descriptive Statistics of the Student Academic Dataset.

	Count	Mean	Std	Min	25%	50%	75%	max
GPA	224	3.60	.96	1.6	2.8	3.7	4.5	5
Complet _Hours	224	31.2	13.2	7	20	33	34.3	59
N_Warnings	224	.16	.46	0	0	0	0	2
N_Absences	224	3.90	3.46	0	1	3	6	22
N_Delays	224	.25	.67	0	0	0	0	5
Unexcused_Absences	224	.11	.10	0	.05	.10	.17	.61
All_Absence	224	.13	.11	0	.05	.11	.18	.61
D_F_40	197	27.5	6.94	11	21	29	33	40
D_M_20	224	12.5	4.92	0	8	14	16	20
D1_10	224	6.60	2.68	0	4.5	7	9	10
D2_5	224	4.30	.89	2	4	5	5	5
D3_10	221	6.74	2.11	0	5	6.5	8	10
D4_10	221	7.86	1.85	2	7	8	10	10
D5_5	219	4.45	.71	2	4	5	5	5
Total	206	69.9	17.5	20	60.3	72	82	100

Table. 2: The parameters used in Student Academic DataSet

Attribute Label	Values
Gender	(Male, Female)
Nationality	(SA, Other)
GPA	Numeric (1:5)
Complete_Hours	Integers
N_Warnings	(0, 1, 2)
Topic	(ARAB, EI, ENGL, ETHC, HIST, IR, ISC, LAW)
N_Absences	Integer
N_Delays	Integer
Unexcused_Absences	Percentage
All_Absences	Percentage
D_F_40	Numeric
D_M_20	Numeric
D1_10	Numeric
D2_5	Numeric
D3_10	Numeric
D4_10	Numeric
D5_5	Numeric
Total	Numeric
Grade	(IC (Incomplete), W (Withdrawal), DN (Denial), A+(95:100), A(90:94), B+(85:89), B(80:84), C+(75:79) , C(70:74), D+(65:69), D(60:64), F(0:59)

A number of critical metrics must be computed in order to investigate categorical variables (refer to Table 3), including count (the number of times each category appears in the dataset), unique (number of unique categories in each variable), top (the most frequent category in each variable), and freq (the frequency of the top category in each variable). In order to predict student academic achievement, these metrics supplement the research of quantitative data by offering essential insights into the distribution and properties of categorical variables

Table. 3: Categorical Variables Metrics for the Student Academic Dataset.

	Gender	Nationality	Topic	Grade
Count	224	224	224	224
Unique	2	2	8	12
Top	Male	SA	LAW	F
Freq	128	215	168	29

The correlations between different features are evaluated by correlation analysis, which is then utilized to efficiently choose features for predictive modeling that are used to forecast academic performance for students. The relationships between quantitative factors are represented visually in the heatmap shown in Figure 2. The dataset's features or variables are represented by the x- and y-axes, while correlation values between -1 (white) and 1 (black) are shown by the grayscale color scheme. Stronger positive correlations, closer to 1, are indicated by darker grayscale tones. Through an examination of the features of the student academic performance dataset, important information can be extracted to improve teaching methods, pinpoint successful interventions, and maximize learning objectives.

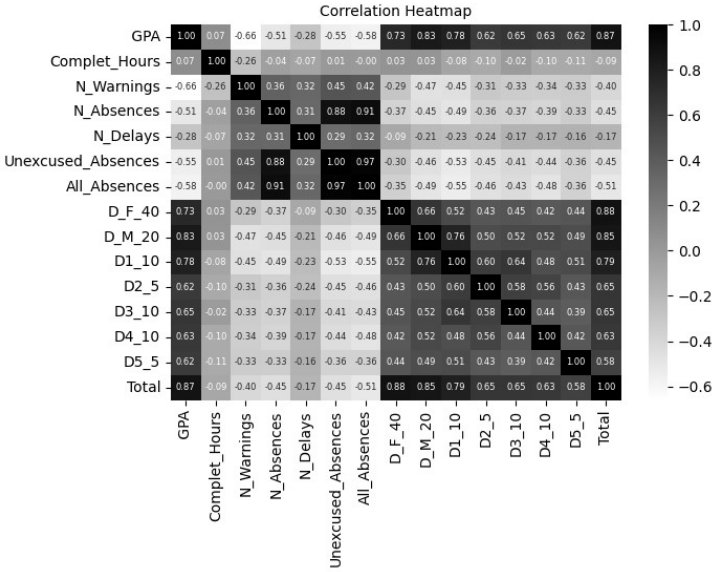

Figure 2: Heatmap illustrating correlations among different Student Academic dataset variables.

3.3 Data Preprocessing

Preparing the Academic Performance Dataset involves a sequential series of steps to preprocess raw data for analysis and modeling. First, the data is examined to make sure it is correct, comprehensive, and doesn't contain any missing numbers. Any missing values are then filled in. After data cleansing, transformations are used to make sure the data aligns with the analysis's or the modeling's underlying assumptions. Normalizing quantitative data to a range, usually between 0 and 1, is a common procedure. Dummy variables are used in order to convert categorical data into numerical values. A distinct dummy variable is used to represent each unique value found in the designated columns of the DataFrame. For example, if the values of a column named "gender" are "Male" and "Female," dummy variables like "gender_Male" and "gender_Female" are created. If and only if the original data point matches that particular value, these new columns will assign a value of 1, and 0 otherwise. Using dummy variables instead of categorical representations improves the performance of statistical and predictive models, enabling more accurate data interpretation. In order to prepare the preprocessed data for additional modeling, analysis, or visualization tasks, it is finally formatted into a suitable structure, such as a structured table. Before data can be efficiently used for forecasting, these first processes in data preparation are necessary to assure data quality and reliability.

3.4 Machine Learning Framework

Gaussian Process Regression, Decision Trees Regression, Linear Regression, Gradient Boosting Regression, Random Forest Regression, Support Vector Regression, AdaBoost Regression, and LASSO Regression are the nine machine learning algorithms used to predict students' academic performance. Based on their nature, these algorithms can be categorized into multiple types:

A linear model is a mathematical model in which the connection between the output variable (Total) and the input variables (GPA, Completed Hours, Number of Warnings, Unexcused Absences, etc.) is linear. It shows how the response variable and the predictor variables are correlated in a straight line. There are two types of linear models: multiple linear regression (which involves more than one predictor variable) and simple linear regression (which involves just one predictor variable). This is furthered by ridge regression, which adds a penalty term to reduce regression coefficients towards zero and minimize the sum of squared residuals [18]. By utilizing the sum of the absolute values of the coefficients, Lasso regression further introduces a penalty component to the goal function. In variable selection, Lasso seeks to minimize squared residuals by precisely pushing some coefficients to zero. Complicated non-linear correlations between the input features and the target variable are not captured by these regression algorithms. Other machine learning algorithms that are capable of capturing non-linear interactions might therefore be more suitable. However, the ease of use, interpretability, and efficiency of linear models in describing linear connections in data make them valuable[19].

Tree-based models are a category of machine learning algorithms that recursively divide the data into subsets based on the values of input features. These models build decision trees in which a feature split is represented by each internal node, the outcome of that split is indicated by each branch, and a prediction is provided by each leaf node[20]. Aiming to minimize impurity at each split, Decision Trees, one of the many variants of tree-based models, divide the dataset into subsets by analyzing feature values[21]. In contrast, Random Forest is an ensemble technique that builds several decision trees and combines their predictions to reduce overfitting and improve generalizability[22]. Tree-based models are particularly good at analyzing feature importance, managing non-linear relationships, and interacting among variables. Their robust performance and capacity to handle complex datasets make them frequently used in a variety of domains[23].

Ensemble models combine a number of different models to improve prediction performance beyond what could be accomplished by a single model. Random Forests, a popular ensemble learning technique, combine predictions from several Decision Trees. In a Random Forest, every tree is trained using a different subset of the data, and the average of all the individual trees' predictions is used to get the final prediction [24, 25]. Gradient Boosting Machines (GBMs), which also use Decision Trees as base learners, are another effective ensemble technique. GBMs use an iterative approach to model building, with each new tree fixing the mistakes made by the preceding ones. A strong predictive model with high accuracy is produced as a result of this iterative process[26]. Another efficient ensemble technique is AdaBoost

Regression, which trains weak learners—usually decision trees—iteratively on weighted versions of the data. To improve overall performance, it modifies these weights to give accurate forecasts for previously mispredicted instances priority [27]. Until a predetermined number of weak learners are trained or the necessary accuracy level is attained, this iterative procedure is continued.

Probabilistic models, such as Gaussian Process Regression, are effective instruments for managing uncertainty and producing forecasts with probabilistic results. Instead of giving a single-point estimate, Gaussian Process Regression models capture the distribution of functions. This enables people to communicate their uncertainty regarding forecasts, which is helpful in situations where decision-making depends on understanding the degree of uncertainty or trust in forecasts.[28].

Machine learning models encompass a variety of algorithms designed to analyze data and make predictions. One example is Support Vector Regression, which finds a hyperplane that minimizes the difference between the predicted and actual points while weighing the trade-offs between complexity and error. Support Vector Machines are extended to regression tasks[29, 30].

3.5 Sending Notifications

The predictive system notifies the appropriate individuals via email when it determines that a student may fail their classes. These parties consist of the following: the student, the academic adviser, the topic lecturer, and the parent or guardian. The goal of this procedure is to make sure that everyone who needs to know is informed in a timely manner so they can support the student and raise their academic performance.

Steps for Sending Notifications:

- Student Notification: The student receives an email warning them that their academic standing is in jeopardy and offering suggestions on how to strengthen their performance.
- Parent/Guardian Notification: The parent or guardian of the student receives an email with details on the student's academic performance as of right now and the likelihood of failing.
- Lecturer Notification: The instructor of the course receives an email with details on the student's performance and suggestions on how to offer the required academic support. With this knowledge, the instructor can provide more classes or modify their approach to better meet the needs of each student.
- Academic Advisor Notification: The academic

Table. 4: Sending Notifications

Case	If	Recipient	Notifications
Predict failure in a subject	GPA is low	Student Parent Guardian	Includes a warning that the academic performance is at risk and the possibility of the student being dismissed from the college if they fail the subject and their cumulative GPA falls below 2.3
		Academic Advisor	Includes a comprehensive report on the student's performance and potential issues and advises the student to withdraw from the subject if the academic workload exceeds the student's level
		Lecturer	Includes information about the student's performance and recommendations on how to provide necessary academic support
Predict failure in a subject	GPA is high	Student Parent Guardian	Includes a warning that the academic performance is at risk and the possibility of failing the subject if the student continues at this level, which will decrease their cumulative GPA
		Lecturer	Includes information about the student's performance and recommendations on how to provide necessary academic support
In case of student absence	Absenteeism rate less than 25%	Student Parent Guardian	Advises providing excuses for absences, if any, to avoid being marked absent and emphasizes the importance of attendance to prevent failing the subject if the absenteeism rate exceeds 25%. It also mentions the impact of absenteeism on the student's performance
In case of student absence	Absenteeism rate exceeds 25%	Student Parent Guardian	Advises providing excuses for absences, if any, to avoid being disqualified from the subject. In case of no excuses, advises the student to withdraw from the subject to avoid decreasing the cumulative GPA
		Academic Advisor	Includes a notification to monitor excuses or withdrawals to prevent affecting the cumulative GPA

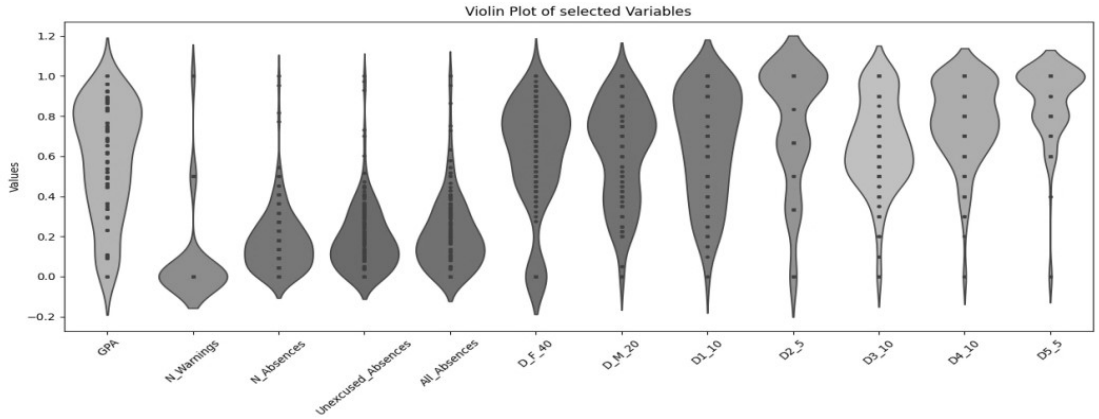

Figure 3: Distribution and summary statistics of factors influencing students' performance visualized by the violin plot.

advisor receives a thorough report regarding the student's performance and any possible problems. If the student's workload is more than they can handle, the adviser should suggest that they drop the course.

Table 4 shows the process of sending notifications to the relevant parties when the predictive system detects the possibility of academic failure, including the type of notification and the associated risks in each case.

The email content is designed to be clear and direct, including the following points:

- Description of the student's current academic status.
- Recommended steps for improvement.
- Additional educational resources and support available to the student.

The predictive method increases the likelihood that a student will succeed in their studies and achieve better academic performance by utilizing an integrated approach to notification delivery.

4. RESULT AND DISCUSSION

We will present the experimental results collected and analyzed to understand how nine machine learning algorithms predict the student's total grade (Total) and assess the final student academic performance outcomes. The important features are: Grade Point Average (GPA), Number of Warnings (N_Warnings), Absences of stedent(N_Absences, Unexcused_Absences, All_Absences), and The student's grades during the semester(D_F_40, D_M_20, D1_10, D2_5, D3_10, D4_10, D5_5). The distribution and summary statistics of 12 features are shown graphically in a violin plot (refer to Figure 3). The data density across different feature values is represented by the width of each violin in this plot; broader sections indicate higher data densities.

Eighty percent of the student academic dataset is made up of training sets, and the remaining twenty percent is made up of testing sets. Machine learning models are structured and their parameters evaluated and adjusted using the training data, while their effectiveness is assessed using the testing dataset. Four evaluation metrics are utilized: Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination, also known as R-Squared (R2) Score. R2 depends on data variance, whereas MSE, RMSE, and MAE measures use the difference between actual and anticipated values of data points to measure error. Scatter graphs demonstrating the accuracy with which machine learning models forecast student achievement in academia are shown in Figure. 4. In this plot, we examine how various machine learning models predict student academic performance by analyzing the relationship between predicted and actual values. These plots visually depict the models' effectiveness in forecasting student outcomes and highlight areas where improvements can be made. The machine learning models' anticipated values for students' academic success are plotted on the x-axis. The real values of student academic performance are represented by the y-axis, which acts as a standard by which the machine learning models' predictions are evaluated.

In studies predicting student performance, we find that machine learning models such as Gaussian Process Regression and Decision Trees Regression are effective and demonstrate a high level of accuracy in forecasting student outcomes. In these models, successful prediction of student performance involves representing data consistently

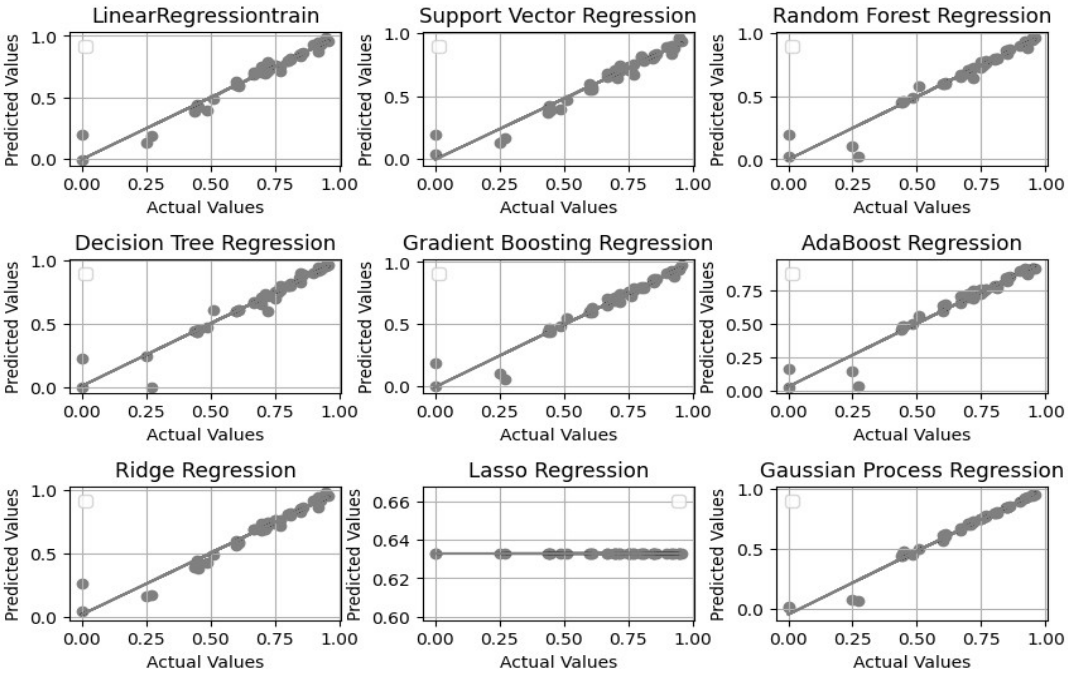

Figure 4: Scatter Plots of machine learning models for the student performance prediction.

with actual values of student performance. A diagonal line connecting the bottom left and upper right corners of the plot is usually followed by data points to show a good correlation between the predicted and actual values. This line may, however, occasionally veer somewhat from the expected, which can be caused by a number of things, including modifications to the training set, errors in the measurement, or other unanticipated effects on student performance. A horizontal line in a Lasso regression indicates that changes in the actual values have little effect on the model's predictions, which are stable.

The performance of machine learning models on the dataset of student performance is shown in Table 5. As per the findings, the Gaussian Process model demonstrated the greatest R-squared value of R2 = 0.9657, hence signifying its exceptional accuracy. Adhering closely, the R2 value for the Decision Trees model was 0.9625, whilst the scores for Linear Regression and Ridge Regression were 0.9588 and 0.9514, respectively. The Gaussian Process model produced the lowest values for RMSE, MSE, and MAE error metrics: 0.0018 for MSE, 0.0424 for RMSE, and 0.0149 for MAE. On the other hand, the regression models exhibited higher values, with LASSO regression showing notable differences. According to the evaluation metrics discussed, the Gaussian Process and Decision Trees models are currently identified as the top-performing models for predicting student performance in this study.

Table 5: Benchmarking of Machine Learning Models for student academic Prediction

Model	MSE	RMSE	MAE	R2
Gaussian Process Regression	0.0018	0.0424	0.0149	0.9657
Decision Trees Regression	0.0020	0.0444	0.0218	0.9625
Linear Regression	0.0022	0.0465	0.0286	0.9588
Ridge Regression	0.0026	0.0505	0.0294	0.9514
Gradient Boosting Regression	0.0027	0.0516	0.0255	0.9494
Random Forest Regression	0.0028	0.0530	0.0258	0.9465
Support Vector Regression	0.0029	0.0540	0.0403	0.9445
AdaBoost Regression	0.0034	0.0579	0.0384	0.9362
LASSO Regression	0.0552	0.2349	0.1916	-0.0513

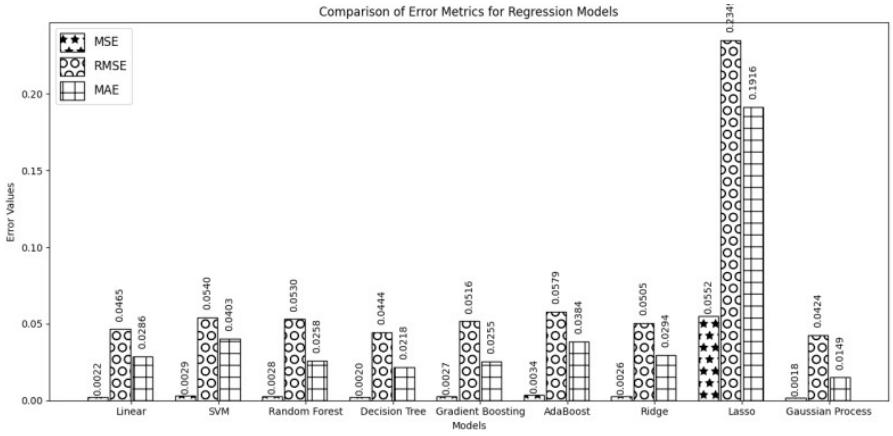

Figure 5: Comparison of Error Metrics for Regression Models

The error metrics (MSE, RMSE and MAE) for various regression models are visualized in bar plots in Figure. 5 that reports that Gaussian model achieved the best results.

Based on the findings of this study, various machine learning algorithms were assessed for their efficacy in predicting the academic performance of at-risk students within university contexts. Here is a synthesized conclusion integrating the performance of each algorithm. The study identified several tiers of performance among the evaluated machine learning models (See Figure. 6).

Excellent Performance: Up to 0.96, the best R-squared values were attained by the Gaussian Process and Decision Tree Regression. This demonstrates their strong capacity for data interpretation and precise prediction-making.

Good Performance: With R-squared values ranging from 0.94 to 0.96, Linear Regression, Ridge Regression, and Gaussian Process Regression likewise showed excellent performance. These models demonstrate consistent ability to predict desired results.

Acceptable Performance: With R-squared values ranging from 0.94 to 0.95, the regression models of gradient boosting, random forest, and support vector demonstrated acceptable performance. Additionally, they showed respectable RMSE values, demonstrating their proficiency in managing data unpredictability.

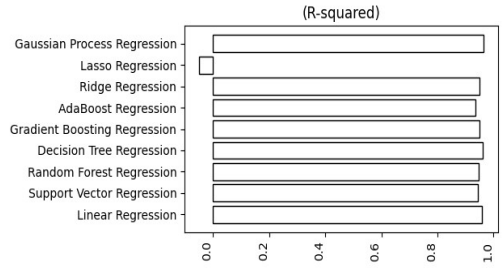

Figure 6: Comparison of R2 for Regression Models

Poor Performance: Lasso Regression showed poor performance, with a negative R-squared value indicating a poor fit between the model and the data. This may be the result of inadequate tuning or the model's incapacity to sufficiently account for the complexity of the dataset.

It is essential to evaluate the results attentively before making any final decisions. Even though the results show that some machine learning algorithms are effective, there are a few points that need to be discussed:

Variance in Accuracy: Although the Gaussian Process model had the greatest R2 value, Decision Trees and other models had results that were quite similar. This implies that certain models might be sensitive to changes in the data, which could have an impact on how accurate they are in various situations.

Diversity in Performance: It is unclear if models such as Lasso Regression are appropriate for the data utilized.

when they perform poorly. The findings might suggest that these models need more fine-tuning or that they are inappropriate for predicting academic success in particular situations.

External Factors: Family support and economic conditions are two examples of potential external factors that could affect student performance but are not taken into account in the results. Future studies should take these aspects into account as they may distort the results.

Sample Size: The information utilized was gathered from a single university's student body. This could restrict how broadly the findings

can be applied. To improve reliability, models must be tested on bigger, more varied datasets.

Interaction Effects: It's possible that certain interactions among the factors under study went unnoticed. For instance, the effect of student absences on grades might be greater than what the models in use now indicate.

5. CONCLUSION

This study addresses the pressing issue of student retention by demonstrating how early warning systems can effectively predict the academic performance of at-risk college students. By focusing on models such as the Gaussian Process and Decision Tree Regression, we found that these approaches provide accurate forecasts and enable timely interventions. This directly responds to the problem of low retention rates and highlights the importance of academic success as a predictor of student continuity.

Our findings suggest that educational institutions can enhance their support programs by implementing customized early warning systems based on these successful models. This proactive approach can create optimized learning environments, addressing the needs of students who are struggling academically.

Furthermore, we identified the necessity for continuous research that explores new datasets and refines machine learning methodologies to enhance the accuracy and applicability of these systems across diverse educational contexts.

Future research should investigate the integration of additional variables, such as psychological and social factors, that may influence student performance. Moreover, exploring the effectiveness of various intervention strategies linked to early warning alerts can provide deeper insights into how to best support at-risk students. Expanding the scope of research to include longitudinal studies could also shed light on the long-term impact of early interventions on student success and retention.

In conclusion, this research contributes valuable insights into how targeted interventions can improve academic performance and retention rates among at-risk students. By answering the challenges identified in the introduction, we emphasize the potential for early warning systems to significantly impact the educational landscape, ultimately fostering a more supportive environment for all students.

REFERENCES:

[1] Lee HB, Villarreal MUJJoEfSPaR. Should Students Falling Behind in School Take Dual Enrollment Courses? 2023;28(4):439-73.

[2] Nieuwoudt JE, Pedler MLJJoCSRR, Theory, Practice. Student retention in higher education: Why students choose to remain at university. 2023;25(2):326-49.

[3] Afzal A, Sami A, Munawar SJIJoH, Society. The role of academic advising and mentoring in promoting student success and retention. 2024;4(1):110-23.

[4] Marbouti F, Diefes-Dux HA, Madhavan KJC, Education. Models for early prediction of at-risk students in a course using standards-based grading. 2016;103:1-15.

[5] Boudjehem R, Lafifi Y. An early warning system to predict dropouts inside e-learning environments. Education and Information Technologies. 2024.

[6] Dawar I, Negi S, Lamba S, Kumar AJSCS. Enhancing Student Academic Performance Forecasting: A Comparative Analysis of Machine Learning Algorithms. 2024;5(6):1-14.

[7] Aggarwal D, Mittal S, Bali VJIJoSDA. Significance of non-academic parameters for predicting student performance using ensemble learning techniques. 2021;10(3):38-49.

[8] Huang S, Fang NJC, Education. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. 2013;61:133-45.

[9] Yılmaz N, Sekeroglu B, editors. Student performance classification using artificial intelligence techniques. International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions; 2019: Springer.

[10] Choudhury S, Ghosh S, Bhattacharya A, Fernandes KJ, Tiwari MK. A real time clustering and SVM based price-volatility prediction for optimal trading strategy. Neurocomputing. 2014;131:419-26.

[11] Remolado AT, Brosas DG, editors. Implementing a Support Vector Classifier for Student Risk Assessment in Colegio De Getafe: A Machine Learning Approach. 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream); 2024: IEEE.

[12] Das A. Logistic regression. Encyclopedia of Quality of Life and Well-Being Research: Springer; 2024. p. 3985-6.

[13] Posonia AM, Vigneshwari S, Rani DJ, editors. Machine learning based diabetes prediction

using decision tree J48. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS); 2020: IEEE.
[14] Pierdzioch C, Risse M. Forecasting precious metal returns with multivariate random forests. Empirical Economics. 2020;58(3):1167-84.
[15] Tealab A. Time series forecasting using artificial neural networks methodologies: A systematic review. Future Computing and Informatics Journal. 2018;3(2):334-40.
[16] Schmidhuber J. Deep learning in neural networks: An overview. Neural networks. 2015;61:85-117.
[17] Suahati A, Nurrahman A, Rukmana O, editors. Academic early warning system: At-risk student prediction using artificial neural network. AIP Conference Proceedings; 2023: AIP Publishing.
[18] Saleh DM, Kadir DH, Jamil DIJQZJ. A comparison between some penalized methods for estimating parameters: simulation study. 2023;8(1):1122-34.
[19] Enwere K, Nduka E, Ogoke UJIIMJ. Comparative Analysis of Ridge, Bridge and Lasso Regression Models In the Presence of Multicollinearity. 2023;3(1):1-8.
[20] Ayulani ID, Yunawan AM, Prihutaminingsih T, Sarwinda D, Ardaneswari G, Handari BDJIJoAS, Engineering, et al. Tree-Based Ensemble Methods and Their Applications for Predicting Students' Academic Performance. 2023;13(3).
[21] Charbuty B, Abdulazeez AJJoAS, Trends T. Classification based on decision tree algorithm for machine learning. 2021;2(01):20-8.
[22] Yadav DC, Pal SJIJoPR. Prediction of heart disease using feature selection and random forest ensemble method. 2020;12(4):56-66.
[23] Quinlan JR. Learning decision tree classifiers. ACM Computing Surveys (CSUR). 1996;28(1):71-2.
[24] Breiman L. Random forests. Machine learning. 2001;45:5-32.
[25] Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2018;8(4):e1249.
[26] Harini K, Rekha KKS, editors. Evaluating Performance Of Identifying At-Risk Students And Learning Achievement Model using Accuracy And F-measure by Comparing Logistic Regression, Generalized Linear Model And Gradient Boost Machine Algorithm. 2022 International Conference for Advancement in Technology (ICONAT); 2022: IEEE.
[27] Wang C, Xu S, Yang JJS. Adaboost algorithm in artificial intelligence for optimizing the IRI prediction accuracy of asphalt concrete pavement. 2021;21(17):5682.
[28] Swiler LP, Gulian M, Frankel AL, Safta C, Jakeman JDJJoMLfM, Computing. A survey of constrained Gaussian process regression: Approaches and implementation challenges. 2020;1(2).
[29] Naicker N, Adeliyi T, Wing JJMPiE. Linear support vector machines for prediction of student performance in school‐based education. 2020;2020(1):4761468.
[30] Chui KT, Fung DCL, Lytras MD, Lam TMJCiHb. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. 2020;107:105584.