Songklanakarin J. Sci. Technol.
47 (6), 406–413, Nov. – Dec. 2025

http://www.sjst.psu.ac.th

Original Article

Statistical forecasting of university academic performance leveraging logistic regression and lasso/ridge regularization on sociodemographic data

Kevin Chamorro*, and Saba Infante

School of Mathematical and Computational Sciences,
Yachay Tech University, Urcuqui, 100119 Ecuador

Received: 20 September 2024; Revised: 11 March 2025; Accepted: 29 July 2025

Abstract

Regression models are crucial in supervised learning and data analysis. Statistical learning methods interpret models and quantify uncertainty, while machine learning techniques handle large-scale predictions. Data modeling serves two main purposes: predicting outcomes and identifying patterns or anomalies within data. This study explores the relationship between sociodemographic factors, including age, gender, socioeconomic status, and educational background, and academic performance in Calculus, Algebra, Biology, and Chemistry among first-semester students at Yachay Tech University, Ecuador (2014–2022). Using a quantitative, correlational methodology within the Knowledge Discovery in Databases (KDD) framework, we developed predictive models through logistic regression enhanced by both Lasso and Ridge regularization. Model performance was assessed with metrics including the confusion matrix, AUC, accuracy, sensitivity, specificity, and Cohen's Kappa. The results show that Lasso consistently outperforms Ridge and baseline logistic regression, achieving accuracies of 88.12% for Calculus, 92.75% for Chemistry, 86.81% for Biology, and 94.49% for Linear Algebra, surpassing the baseline method by up to three percentage points. These models effectively forecast academic outcomes based on sociodemographic data, facilitating early identification of students who may benefit from targeted interventions. This approach not only improves prediction accuracy but also contributes to enhancing educational quality and reducing dropout rates in Ecuadorian higher education.

Keywords: academic performance, regression models, lasso regularization, ridge regularization, KDD

1. Introduction

Academic performance in higher education has long been a subject of global interest, engaging stakeholders ranging from educators and administrators to policymakers and researchers (Honicke & Broadbent, 2016). Previous studies underscore the multifaceted nature of student performance, pointing to factors like teaching quality, learning environments, and institutional support (Schneider & Preckel, 2017). Various predictive techniques have been employed to analyze these factors and forecast academic outcomes. For instance, Support Vector Machines have proven effective in identifying at-risk students in distance learning contexts (Kotsiantis, Pierrakeas, & Pintelas, 2004), while discriminant analysis has been useful in classifying students based on academic achievement (Gutiérrez-Monsalve, Garzón, & Segura-Cardona, 2021). Logistic regression—often used to assess the probability of course completion—has likewise demonstrated robust predictive power in both online and traditional classrooms (Marbouti, Diefes-Dux, & Madhavan, 2016; Yukselturk & Top, 2013).

Despite these findings, most existing research is rooted in contexts such as the United States, Germany, or Australia. Their datasets often focus on a narrower range of predictors (e.g., standardized tests, attendance records, or LMS interactions). Moreover, logistic regression approaches commonly do not incorporate Lasso (L1) or Ridge (L2) regularization, which can address multicollinearity and aid in feature selection. Consequently, there is a lack of evidence on

*Corresponding author
Email address: kevinldu08@gmail.com

how these advanced methods perform when analyzing a broader spectrum of sociodemographic variables in emerging educational contexts, including Ecuador.

In Ecuador, discussions on academic performance have intensified amid challenges related to access, equity, and quality of higher education (Senescyt, 2021). Some local studies do examine socioeconomic and family-related factors (Gutiérrez-Monsalve *et al*., 2021), but few apply regularized logistic regression to comprehensive datasets that encompass variables like age, gender, socioeconomic status, province of birth, re-enrollment data, and type of school. By integrating Lasso and Ridge regularization, our research not only refines the predictive accuracy of logistic regression but also pinpoints the most critical factors influencing academic success in Calculus, Algebra, Biology, and Chemistry at Yachay Tech University. This approach extends beyond prior work by capturing a broader set of predictors and optimizing the detection of at-risk students in an Ecuadorian context.

Hence, our primary contribution lies in developing and validating regularized logistic regression models that leverage an extensive range of sociodemographic and academic features, offering insights into how to improve student retention strategies. By detailing the variable similarities and differences from earlier studies, we underscore the importance of geographically and contextually specific analyses, which can more effectively inform data-driven interventions in higher education.

2. Materials and Methods

Our approach is designed to refine predictive modeling techniques for academic performance, focusing on the identification of sociodemographic data to improve predictive accuracy and operational efficiency in academic settings. Figure 1 shows the KDD methodology employed in our study, which is systematically structured into several phases: selection and preprocessing of the dataset, feature selection through L1 regularization, development of logistic regression models with Lasso and Ridge regularization.

2.1 Data and data sources

This research utilizes data on the academic performance of first-semester students at Universidad Yachay Tech from 2014 to 2022, totaling 2,303 observations. The dataset, managed by the Academic Affairs and Student Welfare Departments, includes grades in four crucial subjects: Calculus I, Linear Algebra, Chemistry I, and Biology I. These numerical variables are essential for evaluating students' academic success in their initial semester and form the basis of the predictive models developed in this study.

In addition to academic grades, the study considers a range of sociodemographic factors as categorical variables, including age, gender, socioeconomic status, and educational background. These variables are detailed in Table 1, providing a comprehensive view of the diverse backgrounds of the student population. To evaluate our models' predictive performances under realistic conditions, we divided the dataset into training and testing subsets using a 70–30 split. Specifically, approximately 70% of the total observations (1,613 students) were allocated to the training set, while the remaining 30% (690 students) served as the test set. By adhering to this ratio, we aimed to ensure that the training set was sufficiently large to capture diverse student characteristics and academic outcomes, while the test set remained substantial enough to offer an unbiased assessment of each model's accuracy, sensitivity, specificity, and other relevant metrics.

The primary objective of this study is to predict whether a student will pass or fail a specific course based on their sociodemographic data, using a comparative analysis of logistic regression with Lasso and Ridge penalization. Initially, we will identify the most important variables using the Lasso model, known for its ability to perform variable selection by penalizing regression coefficients, thereby setting some to zero. After identifying significant variables through Lasso, both Lasso and Ridge models will be implemented to evaluate their predictive performance. By comparing the results of these approaches, we aim to determine which model provides superior predictions and is the most suitable for our dataset.

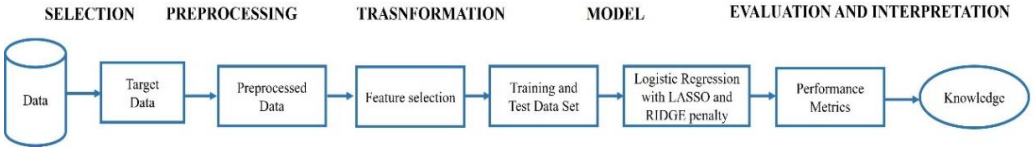

Figure 1. Knowledge Discovery in Databases (KDD) process flow diagram

Table 1. Sociodemographic and academic variables of the student population

Category	Features (Input variables)	Target variable (Output variable)
Sociodemographic factors	Gender, age, ethnicity, marital status, disability, employment, children, country of birth, province of birth	
Academic background	Grade point average, remedial courses, type of school	
Family situation	Homeownership, father's occupation, mother's occupation	
Course enrollment	Enrollment in calculus, algebra, biology, chemistry, remedial courses	
Academic program and courses	Degree program, first semester courses, number of courses passed	
Academic performance		Semester results in calculus, algebra, biology, chemistry

2.2 Data analysis method

2.2.1 From linear to logistic regression

Linear regression is a fundamental statistical model that seeks to describe the relationship between a dependent variable, *Y*, and one or more independent or predictor variables, *X* (Montgomery, Peck, & Vining, 2012),

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \epsilon \quad (1)$$

where:

- Y : Dependent or response variable
- $X_1, X_2, \ldots, X_P$: Predictor or independent variables
- β_0 : The intercept term
- $\beta_1, \beta_2, \ldots, \beta_p$: The coefficients associated with each predictor variable
- ϵ : The random error

To estimate these coefficients, the least squares method is used. This technique aims to minimize the sum of the squares of the differences (residuals) between the observed values of *Y* and the values predicted by the model.

Residual: The residual for a particular observation, *i*, is the difference between the observed value y_i and the predicted value $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i \quad (2)$$

Cost Function: The cost function, J(β), is the sum of the squares of the residuals:

$$J(\beta) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (3)$$

The objective is to find the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ that minimize this cost function.

2.2.2 Matrix form

To generalize the above equation, we rewrite the regression model in matrix form:

$$Y = X\beta + \epsilon \quad (4)$$

where:

- Y : It is response vector.
- X : It is the design matrix, which contains the observations of the predictor variables. It has an additional column of ones for the intercept term
- β : It is the vector of unknown coefficients.
- ϵ : It is the vector of errors.
    - The cost function, which is the sum of the squares of the residuals, is defined as:

$$J(\beta) = (Y - X\beta)^T (Y - X\beta) \quad (5)$$

To minimize $J(\beta)$ with respect to β, we take the derivative of $J(\beta)$with respect to β and set it equal to zero:

$$\frac{\partial J(\beta)}{\partial \beta} = 0$$

This leads to the normal equations:

$$X^T X\beta = X^T Y \quad (6)$$

To solve for β, we simply multiply both sides by the inverse of $X^T X$ (assuming that $X^T X$ is nonsingular and therefore invertible):

$$\beta = (X^T X)^{-1} X^T Y \quad (7)$$

This is the least squares estimator of the coefficients.

To transition from linear regression to logistic regression, we need to consider the logistic function, which can map any input to a value between 0 and 1. This allows us to model binary outcomes or probabilities. In the academic context, the binomial distribution is used to model the number of students who pass (succeed) in a group of n students, where the probability of passing for each student is p. When predicting whether a specific student passes or fails, we are interested in the case where $n = 1$. The probability mass function for a random variable Y, which indicates whether a student passes (1) or fails (0), is:

$$P(Y = y) = p^y (1 - p)^{1-y} \quad (8)$$

To model the probability p of a student passing based on predictor variables, such as sociodemographic variables, we require a function that relates these variables to the probability of passing. This function should transform the range of linear combinations of these variables, which is $(-\infty, \infty)$, to the range of probability, which is (0,1).

Therefore, the logistic regression model posits that the conditional probability of a successful event, denoted by $\boldsymbol{p}$, is given by:

$$p = P(y = 1 \mid x) = \frac{e^{X\beta}}{1 + e^{X\beta}} \quad (9)$$

This probability represents the likelihood of the outcome y=1 given the predictors X. To relate this probability to a linear combination of the predictors, we use the logit link function, defined as:

$$g(p) = \log\left(\frac{p}{1-p}\right) \quad (10)$$

The logit function transforms the odds, defined as the ratio of success (p) to failure ($1-p$), into a linear relationship with the predictors:

$$\log\left(\frac{p}{1-p}\right) = X\beta \quad (11)$$

The inverse of the logit function maps the linear combination of predictors back to the probability of success:

$$p = \frac{e^{X\beta}}{1 + e^{X\beta}} \quad (12)$$

Unlike linear regression, where we can use the normal equations to obtain a closed-form solution for the coefficients, in logistic regression, the coefficients are estimated by maximizing the likelihood function. The

likelihood function indicates how well the model fits the observed data of students who passed and failed. For a set of n students, the likelihood function L is:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad (13)$$

Where, y_i indicates whether student i passed (1) or failed (0), and p_i is the probability predicted by the model that student i passes. In practice, it is more common to work with the logarithm of the likelihood function, called log-likelihood, because it converts the product into a sum and simplifies calculations:

$$\log L(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1-y_i)\log(1-p_i)] \quad (14)$$

The goal is to find the coefficients β that maximize this log-likelihood. This is typically done using iterative numerical methods, such as the Newton-Raphson algorithm.

Regularized regression is a statistical technique that extends linear or logistic regression by adding penalty terms to the cost function, controlling model complexity and preventing overfitting. This method is particularly useful when analyzing academic performance data involving numerous predictor variables, such as test scores, attendance, socioeconomic factors, and study habits, some of which may be redundant or collinear. Techniques like Lasso and Ridge regularized logistic regression specifically address these issues by selecting relevant predictors and reducing multicollinearity, thus improving predictive accuracy on new data (Hastie, Tibshirani, & Friedman, 2009).

Lasso regularized regression (Least Absolute Shrinkage and Selection Operator) is a linear regression technique that incorporates an L1-norm-based penalty term to control model complexity and perform variable selection. This technique is particularly useful when working with high-dimensional datasets or highly correlated variables, as it helps select the most relevant variables and improve model interpretability (Tibshirani, 1996).

$$\min_{\beta_0,\beta} \left\{ -\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + \beta^T x_i) - (1-y_i)\log\big(1 + \exp(\beta_0 + \beta^T x_i)\big) + \lambda \parallel \beta \parallel_1 \right\},$$

where N is the total number of instances in training set, y_i is the real class of the instance i, x_i is the instance feature vector i, λ is the regularization parameter, which controls the trade-off between fitting the data and keeping the coefficients small, β_0 and β are the model coefficients, and $\parallel \beta \parallel_1$ is the L1 norm of the coefficients, which is the sum of the absolute values of the coefficients.

Our study applied L1 regularization to identify and rank the most critical features affecting academic performance in subjects such as Calculus I, Chemistry I, Biology I, and Linear Algebra. Table 2 highlights the features with the greatest impact on model performance, revealing important insights for understanding academic outcomes at Yachay Tech University based on sociodemographic variables. Additionally, Ridge regression is noted for its usefulness across diverse fields-such as social, economic, biological, and health sciences-in analyzing relationships between variables in cases of multicollinearity.

Table 2. Most important features identified by L1 regularization for each subject

Subject	Most important features
Calculus I	Enrolled courses (4), Children (Yes), Ethnicity (Unregistered), Grade score, Employment (Yes)
Chemistry I	Linear algebra (Pass), Third enrollment in chemistry I, Third enrollment in linear algebra, Second enrollment in chemistry I, degree in petrochemical engineering
Linear algebra	Chemistry I (Pass), Third enrollment in linear algebra, Third enrollment in chemistry I, Second enrollment in linear algebra, Second enrollment in chemistry I
Biology I	Calculus I (Pass), Chemistry I (Pass), Ethnicity (Mulatto), Province of residence (Sucumbíos), Employment (Yes)

$$\min_{\beta_0,\beta} \left\{ -\frac{1}{N}\sum_{i=1}^{N} y_i(\beta_0 + \beta^T x_i) - (1-y_i)\log\big(1 + \exp(\beta_0 + \beta^T x_i)\big) + \lambda \parallel \beta \parallel_2^2 \right\},$$

where N is the total number of instances in training set, y_i is the real class of the instance i, x_i is the instance feature vector i, λ is the regularization parameter, which controls the trade-off between fitting the data and keeping the coefficients small, β_0 and β are the model coefficients, and $\parallel \beta \parallel_2^2$ is the squared $L2$ norm of the coefficients, which is the sum of the squares of the coefficients.

The idea behind Ridge is to prevent overfitting and handle multicollinearity, which occurs when the predictor variables are highly correlated. By penalizing the coefficients, Ridge ensures that no individual predictor variable has too much influence, which can be beneficial when the variables are collinear.

2.2.3Machine learning performance metrics

To rigorously assess the performance of our regularized logistic regression model in predicting academic performance on sociodemographic data, we employed a range of established machine learning metrics. Each metric provides distinct insights into the model's predictive abilities, ensuring a comprehensive evaluation.

- Accuracy: Defined as the proportion of true results (both true positives and true negatives) among the total number of cases examined. It is calculated using the formula:
  $$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$
- TP (True Positives): Number of instances correctly predicted as the positive class (e.g., "pass" when the student actually passes).
- TN (True Negatives): Number of instances correctly predicted as the negative class (e.g., "fail" when the student actually fails).
- FP (False Positives): Number of instances incorrectly predicted as positive (predicted "pass" but the student fails).

- FN (False Negatives): Number of instances incorrectly predicted as negative (predicted "fail" but the student actually passes).
- Accuracy indicates the overall proportion of correct predictions the model makes. For additional details on this metric, see Powers (2011).
- Precision: Also known as the positive predictive value, precision measures the ratio of true positive predictions to the total positive predictions made. It is defined as:

$$Pre = \frac{TP}{TP + FP}$$

High precision indicates a low rate of false positive predictions, crucial for medical diagnostics where falsely identifying a condition can led to unnecessary interventions. Refer to Powers (2011) for more insights.

- F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is particularly useful when the class distribution is uneven. The formula is:

$$F1 - Score = 2 \times \frac{Pre \times Recall}{Pre + Recall}$$

where recall (or sensitivity) is the ratio of true positive predictions to the actual positives in the dataset. See Powers (2011).

- AUC (Area Under the ROC Curve): AUC measures the entire two-dimensional area underneath the entire ROC (Receiver Operating Characteristic) curve. It provides an aggregate measure of performance across all possible classification thresholds. The AUC ranges from 0 to 1, where a model whose predictions are 100 % wrong has an AUC of 0.0, and a model whose predictions are 100% correct has an AUC of 1.0. For further reading, see Fawcett (2006).
- Cohen's Kappa: This metric measures the agreement between two raters who each classify items into mutually exclusive categories. To model evaluation, it compares the observed agreement with what might be expected by chance, according to:

$$\text{Cohen's Kappa} = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement, and p_e is the expected agreement under independence. Cohen's Kappa is particularly useful in situations where accuracy may be misleading due to imbalanced class distributions.

These metrics collectively enable a comprehensive evaluation of our model, providing insights into its strengths and areas for improvement in predicting the academic performance on sociodemographic data. For the theoretical background, see Cohen (1960).

3. Results and Discussion

3.1 Exploratory data analysis

Our exploratory data analysis involved examining a dataset of 2,303 first-semester students from Yachay Tech University, covering academic periods from 2014 to 2022. The dataset includes 25 attributes related to courses such as Calculus, Linear Algebra, Chemistry, and Biology, with students representing various Ecuadorian provinces. Figure 2 displays pass and fail rates per subject and includes a choropleth map indicating the geographical distribution of pass percentages.

Figure 2. Heat map of academic performance by province in Ecuador

3.2 Predictive analysis using lasso and ridge models

In our study, we applied L1 regularization (Lasso) to identify and rank the most influential features affecting academic performance in Calculus I, Chemistry I, Biology I, and Linear Algebra. For Calculus I, Table 3 indicates that both Lasso and Ridge logistic regression models performed strongly, with Lasso achieving slightly higher accuracy (0.8927 training, 0.8812 test) compared to Ridge (0.8739 test). Regarding sensitivity—the ability to correctly identify students at risk of failing—Lasso also outperformed Ridge (0.7564 vs. 0.7047 in testing). However, Ridge regression had a marginal edge in specificity, correctly classifying a slightly greater proportion of students who passed. The confusion matrices (Figure 3) show Lasso correctly identifying more failing students, which might be valuable for targeted interventions. The small differences observed suggest both models provide robust predictions, though institutional goals may favor one over the other based on sensitivity or specificity preferences. Additional considerations for Calculus I highlight the importance of strong algebraic skills and pre-calculus foundations, where Lasso's ability to isolate key predictors (like prior math background and sociodemographic variables) may be particularly beneficial.

In Chemistry I, both Lasso and Ridge logistic regression demonstrated effective classification (Table 4). Lasso achieved slightly higher test accuracy (0.9275) and sensitivity (0.7976) compared to Ridge (0.9246 accuracy, 0.7738 sensitivity). Ridge exhibited slightly greater specificity (0.9732 vs. Lasso's 0.9693). The confusion matrices (Figure 4) reveal similar overall capabilities, though Lasso again showed a slight advantage in identifying at-risk students. Chemistry's reliance on theoretical knowledge and practical lab skills suggests that sociodemographic factors and re-enrollment history may be influential predictors, with Lasso effectively eliminating weaker variables and Ridge managing correlated features effectively.

For Biology (Table 5), both models maintained high specificity but had lower sensitivity compared to other subjects. Lasso outperformed Ridge in sensitivity (0.7976 vs. 0.7738),

while Ridge was marginally more specific (0.9665 vs. 0.9594). The slightly lower sensitivity across both models indicates challenges in accurately identifying at-risk students, possibly due to Biology's broad content and reliance on memorization and extensive reading. Nonetheless, Lasso may offer advantages in highlighting key predictors, supporting targeted academic interventions like additional lab or reading support.

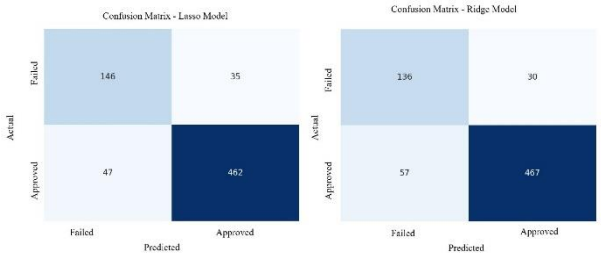

Figure 3. Confusion matrix: Lasso and Ridge models - Academic performance prediction (Calculus I)

Figure 4. Confusion Matrix: Lasso and Ridge models - Academic performance prediction (Chemistry)

For Linear Algebra (Table 6), both models displayed similar performance, with Lasso being slightly better at identifying failing students and Ridge being slightly better at correctly classifying passers. The abstract nature of Linear Algebra, reliant on algebraic reasoning, likely results in fewer critical predictors, explaining the similarity in performance between models. The choice between Lasso and Ridge here would depend largely on the institution's priorities regarding sensitivity and specificity.

Overall, comprehensively evaluating both models and considering their metrics and influential variables, the Lasso model generally appears preferable for future predictions due to its higher sensitivity and capability to identify at-risk students. However, the choice between Lasso and Ridge ultimately depends on institutional priorities at Yachay Tech University: Lasso is ideal for maximizing the detection of students needing intervention, while Ridge better minimizes false positives.

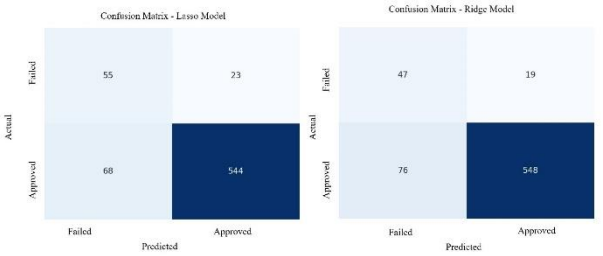

Figure 5. Confusion Matrix: Lasso and Ridge models - Academic performance prediction (Biology)

Table 3. Comparative analysis of metrics between Lasso and Ridge models – Calculus I

Metric	Lasso (Training)	Lasso (Testing)	Ridge (Training)	Ridge (Testing)
Accuracy	0.8927	0.8812	0.8872	0.8739
Sensitivity	0.7417	0.7565	0.7108	0.7047
Specificity	0.9517	0.9296	0.9560	0.9396
Kappa	0.7231	0.6994	0.7048	0.6731
AUC (Area under the curve)	0.941	0.937	0.942	0.935

Table 4. Comparative analysis of metrics between Lasso and Ridge models – Chemistry I

Metric	Lasso (Training)	Lasso (Testing)	Ridge (Training)	Ridge (Testing)
Accuracy	0.9374	0.9275	0.9306	0.9246
Sensitivity	0.8147	0.7976	0.7843	0.7738
Specificity	0.9770	0.9693	0.9779	0.9732
Kappa	0.8236	0.7959	0.8021	0.785
AUC (Area under the curve)	0.978	0.952	0.979	0.951

Table 5. Comparative analysis of metrics between Lasso and Ridge models – Biology I

Metric	Lasso (Training)	Lasso (Testing)	Ridge (Training)	Ridge (Testing)
Accuracy	0.8785	0.8681	0.8795	0.8623
Sensitivity	0.4586	0.4471	0.4276	0.3821
Specificity	0.9705	0.9594	0.9777	0.9665
Kappa	0.8102	0.7746	0.7957	0.7259
AUC (Area under the curve)	0.872	0.849	0.873	0.835

Table 6. Comparative analysis of metrics between Lasso and Ridge models – Linear Algebra

Metric	Lasso (Training)	Lasso (Testing)	Ridge (Training)	Ridge (Testing)
Accuracy	0.933	0.9449	0.9361	0.9333
Sensitivity	0.8701	0.9078	0.8187	0.8156
Specificity	0.9493	0.9545	0.9665	0.9636
Kappa	0.7997	0.8358	0.8004	0.7917
AUC (Area under the curve)	0.972	0.9699	0.973	0.9672

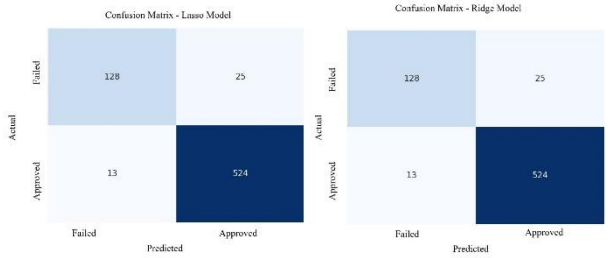

Figure 6. Confusion Matrix: Lasso and Ridge models - Academic performance prediction (Linear Algebra)

4. Conclusions

Our study demonstrates the efficacy of logistic regression models enhanced with Lasso and Ridge regularization in predicting academic performance based on sociodemographic data from Yachay Tech University. By examining four core first-semester courses; Calculus, Chemistry, Biology, and Linear Algebra, our results confirm that Lasso typically outperforms Ridge in correctly identifying students who are likely to fail (i.e., higher sensitivity), without sacrificing specificity. In particular, Lasso achieved test accuracies of 88.12% in Calculus, 92.75% in Chemistry, 86.81% in Biology, and 94.49% in Linear Algebra, surpassing unpenalized logistic regression by up to three percentage points (Tables 3–6). The confusion matrices (Figures 3–6) further illustrate that Lasso provides a more reliable detection of at-risk students, a critical consideration for institutions looking to intervene early.

These performance differences emphasize not only the predictive capabilities of our models but also their immediate potential to inform decisions aimed at improving educational outcomes. By enabling the early identification of students who might benefit from remedial measures, the Lasso-based approach may contribute to reducing dropout rates in a tangible way. This is especially pertinent in courses where conceptual difficulty or lab-based requirements can exacerbate learning challenges, as in Chemistry and Biology, and where sociodemographic indicators, such as re-enrollment history, high-school background, and socioeconomic status, often reveal meaningful patterns that help guide interventions.

Although our dataset stems from Yachay Tech University, these findings have broader implications for other higher education contexts. The flexibility of regularized logistic regression makes it suitable for diverse datasets, provided that the relevant sociodemographic and academic information is collected consistently. Future research might extend this work by integrating additional attributes, such as attendance records, psychoeducational variables, or real-time learning analytics, or by exploring advanced ensemble methods that build upon the strengths of Lasso and Ridge.

In conclusion, our study validates the potential of Lasso and Ridge regularization in improving the predictive power of logistic regression for academic outcomes, particularly when harnessing sociodemographic and academic data. By offering both robust performance and enhanced interpretability, these models represent a valuable tool for educational institutions seeking to elevate academic quality and reduce student dropout rates through data-driven interventions.

Acknowledgements

We extend our sincere gratitude to Yachay Tech University for providing the data essential for this research while ensuring adherence to data privacy standards. This collaboration has been invaluable in facilitating our study and enhancing our understanding of academic performance through advanced statistical methods.

References

Al Husaini, Y., & Ahmad Shukor, N. S. (2023). Factors affecting students' academic performance: A review. *Journal of Educational Research, 12*, 284-294.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measure ment, 20*(*1*), 37–46. doi:10.1177/001316446002000104

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(*8*), 861–874. doi:10.1016/j.patrec.2005.10.010

Gutiérrez-Monsalve, J. A., Garzón, J., & Segura-Cardona, A. M. (2021). Factores asociados al rendimiento académico en estudiantes universitarios. *Formación Universitaria, 14*(*1*), 13-24. doi:10.4067/S0718-50062021000100013

Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review, 17*, 63-84. doi:10.1016/j.edurev.2015.11.002

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence, 18*(*5*), 411-426. doi:10.1080/08839510490442058

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers and Education, 103*, 1-15. doi:10.1016/j.compedu.2016.09.005

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Hoboken, NJ: John Wiley & Sons.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*(*1*), 37–63.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin, 143*(*6*), 565-600. doi:10.1037/bul0000098

Secretaría De Educación Superior, Ciencia, Tecnología E Innovación. (2021). *Plan estratégico institucional 2021—2025*. Retrieved from https://www.educacionsuperior.gob.ec/

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*, 417-453. doi:10.3102/00346543075003417

Villarruel-Meythaler, R. E., Tapia-Morales, K. I., & Cárdenas-García, J. K. (2020). Determinantes del rendimiento académico de la educación media en Ecuador. *Revista Economía y Política, 32*. Retrieved from https://www.redalyc.org/journal/5711/571163421008/html/

Yukselturk, E., & Top, E. (2013). Exploring the link among entry characteristics, participation behaviors and course outcomes of online learners: An examination of learner profile using cluster analysis. *British Journal of Educational Technology, 44*. doi:10.1111/j.1467-8535.2012.01339.