

Article

Predicting Student Dropout from Pre-Enrollment Data in Mexican Higher Education: A Theoretically Grounded and Calibrated Machine Learning Approach

Blanca Carballo-Mendívil ^{*}, Adrián Jesús Pérez-Morales, Alejandro Arellano-González ,
María del Pilar Lizardi-Duarte and Nidia Josefina Ríos-Vázquez 

Departamento de Ingeniería Industrial, Instituto Tecnológico de Sonora, Ciudad Obregón 85000, SO, Mexico; adrian.perez236023@potros.itson.edu.mx (A.J.P.-M.); alejandro.arellano@itson.edu.mx (A.A.-G.); maria.lizardi@itson.edu.mx (M.d.P.L.-D.); nidia.rios@itson.edu.mx (N.J.R.-V.)

* Correspondence: blanca.carballo@itson.edu.mx

Featured Application

The proposed model can be integrated into an institutional dashboard to report dropout risk by student, cohort, academic program, and campus, and to facilitate the prioritization of mentoring, academic support, and early intervention strategies.

Abstract

Student dropout remains a persistent challenge in higher education, with significant academic, social and institutional implications. Although machine learning (ML) models for dropout prediction have proliferated, most select input variables based on data availability rather than theory and rarely report psychometric validation of the constructs used. This study addresses both gaps by developing and validating an early warning system at a Mexican university using exclusively pre-enrollment data from more than 46,000 student records from 2014 to 2025. Following the CRISP-DM methodology, eight theoretically grounded constructs, anchored in Tinto's integration model, Bean's attrition model, and Cabrera et al.'s persistence model, were operationalized from institutional intake questionnaires and assessed for internal consistency using Cronbach's alpha prior to model training. Eight supervised ML algorithms were benchmarked across distance-based (Logistic Regression, SVM, AdaBoost, ANN) and tree-based (Random Forest, XGBoost, LightGBM, CatBoost) families. A tuned and isotonically calibrated Random Forest achieved the best overall performance (recall = 0.743, F1 = 0.524, ROC-AUC = 0.734, PR-AUC = 0.478) on a strictly held-out test set that was not used at any stage of model development. The 2023–2025 cohorts, whose dropout labels were not yet observable under the institutional definition, were scored prospectively to generate operational risk profiles. SHAP analysis identified high-school GPA, parental education, and household asset indices as dominant predictors, which mapped directly onto the three theoretical frameworks. These findings demonstrate that psychometrically grounded pre-enrollment data alone can support an operationally deployable dropout detection system, enabling proactive, evidence-based retention interventions from the first day of enrollment.



Academic Editors: Chuang Wang and Yuyang Cai

Received: 4 June 2026

Revised: 22 July 2026

Accepted: 28 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: student dropout; dropout prediction; early warning system; machine learning; ensemble methods; pre-enrollment data; theoretical grounding; educational data mining

1. Introduction

Student dropout remains one of the most persistent challenges in Higher Education Institutions (HEIs) globally, as it affects both the individuals involved and the institutions responsible for their education, as well as the efficiency of government and private investment, and undermines national equity objectives (OECD, 2019; UNESCO, 2021).

In this context, the application of machine learning (ML) as a predictive tool holds promise for addressing student retention, as classical statistical modeling methods are limited in their ability to capture complex nonlinear relationships among predictor variables and do not enable timely generation of predictive probability estimates. In such cases, ML provides a potentially effective alternative (Delen, 2010; Sghir et al., 2023) because HEIs collect a substantial amount of data through enrollment records, intake surveys, and online learning platforms, and can apply this method to identify high-risk students early in their academic careers (Baneres et al., 2019; Von Hippel & Hofflinger, 2021). Random Forest, XGBoost, LightGBM, and CatBoost have been shown to be highly effective in identifying students who are likely to drop out of their courses early. These tree-based ensemble techniques were significantly better than traditional classification models previously used to identify student dropouts in large student datasets (Villar & de Andrade, 2024).

Nevertheless, a structural gap remains in ML: models rarely base their variable selection on established educational theory, and psychometric validation of input constructs is rarely reported. Reviews of predictive literature confirm that most ML dropout studies select variables based on data availability rather than theory (Reina Marín et al., 2025). Among those who engage with theoretical frameworks, none report criterion validity or measurement invariance for their psychological constructs (Jud et al., 2025). When reliability is reported, such as Cronbach's alpha coefficients for scale-based inputs, it is the exception rather than the rule (Musso et al., 2020). Furthermore, no study using pre-enrolment constructs in Latin American higher education has achieved a sufficiently large sample size to enable robust generalization, as a recently conducted research project achieved: an AUC score of 0.69 and a sensitivity of 0.88 using the XGBoost method with pre-enrolment data of more than 40,000 students (Carballo-Mendivil et al., 2025). This study demonstrated the effective implementation of the model but also revealed an opportunity to improve its accuracy and probabilistic calibration.

At a theoretical level, Tinto's (1993) integration model remains the most cited conceptual model for integrating students into college life; however, it appears to be defined differently in practice across studies. While academic integration is usually assessed using a scale or instrument with demonstrated psychometric properties, such as reliability and validity, some research relies on a single administrative measure, such as a student's cumulative grade point average (CGPA), to assess academic integration. Additionally, social integration is often not a significant predictor of retention in settings where students do not live on campus or commute to classes (Shmeleva & Froumin, 2020). On the other hand, while there are other models cited in the literature, such as Bean's (1980) attrition model and Cabrera et al.'s (1993) persistence model, they are rarely cited as the primary framework of study. Bean's model is usually used in combination with Tinto's model rather than on its own, and only one study uses Cabrera as the primary framework (Vilela Braga et al., 2018). Furthermore, few studies have assessed the psychometric properties of instruments developed to assess constructs within either model, and those reports are outside North America and Western Europe. The only study outside this context that makes a formal psychometric effort is Castro-Montoya et al. (2025) in Colombia, but it reports that the validation is incomplete. Furthermore, studies that use dropout intentions as a proxy for behavior face serious validity issues because prior research indicates that intentions account for less than 10% of actual withdrawals (Lykkegaard & Qvortrup, 2024).

This study aims to address these gaps. It used over 46,000 pre-enrollments from Mexican university cohort groups from 2014 through 2025: (1) grounding all predictors in [Tinto \(1993\)](#), [Bean \(1980\)](#), or [Cabrera et al. \(1993\)](#); (2) checking the internal reliability of Cronbach's alpha scores on the training dataset before developing any models; (3) comparing the performance of eight machine learning algorithms developed with hyperparameter tuning, isotonic probability calibration, and threshold optimization; and (4) interpreting how different models make decisions based on SHAP value output linked to eight theoretical constructs. To the best of our knowledge, the combined use of psychometrically grounded constructs, systematic model comparison with probability calibration, and SHAP-based interpretability has not been previously reported at this scale in Latin American HEIs, and each element of this claim is supported by recent evidence. A recent systematic review of ML-based dropout prediction concluded that models rarely operationalize educational theory and that psychosocial constructs remain underrepresented, with predictors typically retained because of their availability and predictive power rather than their theoretical meaning ([Delen et al., 2020](#); [Kemper et al., 2020](#); [Lopez-Muñoz et al., 2026](#)). Reliability and validity evidence for input constructs are generally confined to instrument development studies, whereas predictive ML studies report model performance metrics ([Hernández-Fernández et al., 2023](#); [Vaarma & Li, 2024](#)). Probability calibration likewise remains an emerging rather than routine practice in dropout prediction, in which discrimination metrics still dominate ([Kwon, 2026](#); [Nagy & Molontay, 2024](#)). Latin American studies follow the same pattern, prioritizing predictive performance over construct measurement and calibrated probability estimation ([Gonzalez-Nucamendi et al., 2023](#); [Gutierrez-Pachas et al., 2023](#); [Palacios et al., 2021](#)).

In line with these frameworks, the descriptive objectives listed above were complemented with directional statistical hypotheses that link each of the eight theoretical constructs to dropout risk: (H1) a weaker prior academic trajectory (lower high school GPA) is associated with a higher probability of dropout ([Tinto, 1993](#)); (H2) lower socioeconomic status and cultural capital (e.g., lower parental educational attainment, fewer household assets) are associated with a higher probability of dropout ([Bourdieu, 1998](#); [Cabrera et al., 1993](#)); (H3) health conditions and health-risk behaviors (e.g., an illness affecting academic performance, more frequent alcohol consumption) are associated with a higher probability of dropout; (H4) employment at entry, especially when unrelated to institutional engagement, is associated with a higher probability of dropout ([Bean, 1980](#)); (H5) longer commute times and distances are associated with a higher probability of dropout ([Cabrera et al., 1993](#)); (H6) greater family responsibilities and weaker family support are associated with a higher probability of dropout ([Bean, 1980](#)); (H7) lower motivation and expectations (e.g., lower career-choice satisfaction and weaker persistence dispositions) are associated with a higher probability of dropout ([Tinto, 1993](#)); and (H8) demographic characteristics associated with non-traditional student profiles (e.g., older age at first enrollment) are associated with a higher probability of dropout ([Bean, 1980](#)).

Given the predictive, non-experimental design of this study, these hypotheses were formulated and evaluated in terms of predictive associations, rather than causal effects. A hypothesis was considered empirically supported when the corresponding construct exhibited a substantial SHAP contribution in the hypothesized direction (see Section 4.3).

2. Literature Review

2.1. Conceptual Frameworks for Student Dropout

Three theories were utilized in conjunction to establish the study's theoretical foundation. As posited by [Tinto's \(1993\)](#) Academic and Social Integration Theory, students' likelihood of withdrawing from their studies is contingent on the extent of their affiliation

with the college community, as evidenced by academic pursuits and social endeavors. Students are at risk of dropping out if they experience a sense of disconnection in either area of their lives. Bean's (1980) Nontraditional Student Theory builds upon Tinto's theory by incorporating additional external factors, including employment, familial obligations, and financial pressures. In addition to internal forces, external factors must be considered when examining the factors that contribute to students' decisions to discontinue their education, much like employees' decisions to remain in their current roles or pursue further education. In this context, employees must assess the potential costs and benefits associated with continuing employment or pursuing academic studies in the field. Cabrera et al.'s (1993) Persistence Theory builds upon the foundations laid by both earlier theories by incorporating two components that must be considered when predicting whether a student will persist in their academic journey: the financial support provided by the institution and external organizations and the level of institutional support extended to the student. Adequate resources for students have been shown to facilitate their connection to institutions, regardless of their level of academic or social integration.

The decision to use all three models (Tinto, Bean, and Cabrera) rather than a single model was based on theory. Many empirical studies have shown a significant overlap between Tinto and Bean's concepts. Cabrera et al. (1992) and Cabrera et al. (1993) had suggested that combining these two models into a single conceptual framework helps explain why students drop out of college. Recent research indicates that multi-model approaches can include a wider variety of predictors of dropping out than single-model approaches, including academic performance, behavioral factors, and external pressures such as job requirements (Müller & Klein, 2023; Nikolaidis et al., 2022). Literature reviews of dropout studies indicate that no single model includes all facets of student attrition. Therefore, an additional or combined conceptual framework is needed for predictive purposes and institutional applications (Pinheiro et al., 2023).

These theoretical models were also used to organize the eight categories in this study that group clusters of variables: (1) previous academic success/tracking, (2) socioeconomic status/cultural capital, (3) health/well-being, (4) employment linkages to academics, (5) travel/logistics, (6) family obligations/social support, (7) motivation/expectations, and (8) demographics.

Although these theoretical constructs are primarily based on the work of Tinto (1993), Bean (1980), and Cabrera et al. (1993), other theories have also been used to provide a comprehensive overview of the student experience, extending the operational definitions of certain concepts. The health and well-being construct draws on Jessor and Jessor's (1977) Problem Behavior Theory, which frames alcohol consumption as one indicator within a broader syndrome of risk behaviors that compromise adolescent development, a framework extended here to academic persistence in early university years. The extracurricular participation variable draws on Astin's (1984) Student Involvement Theory, which posits that the amount of physical and psychological energy a student devotes to the academic experience, including extracurricular activities, is positively associated with persistence and development. The motivation and vocational adjustment construct draws on Kerby's (2015) integrated retention model, which identifies students' perceived fit with their chosen program and vocational alignment as key predictors of persistence. Spady's (1970) sociological model, recognized as a foundational precursor to Tinto's work, established that students' social origins and family backgrounds, including parental characteristics, shape their capacity for academic and social integration. Building on this, parental education is operationalized as an indicator of cultural capital, following Bourdieu's (1998) framework. Therefore, these additional theoretical frameworks did not replace the core models; rather, they expanded the measurement of each core model to include aspects that Tinto

(1993), Bean (1980), and Cabrera et al. (1993) recognized but could not fully define using items alone.

2.2. Machine Learning Applications in Dropout Prediction

Recently, the use of machine learning (ML) to predict student dropouts in educational institutions has significantly increased (Delen, 2010; Sghir et al., 2023). As previously mentioned, most existing research using supervised learning methods has focused on tree-based ensemble methods, such as Random Forest, XGBoost, LightGBM, and CatBoost, as well as distance-based classifiers, such as logistic regression, support vector machines, and neural networks. Both families of algorithms have been applied to student dropout prediction, although their relative performance varies considerably depending on dataset characteristics, class imbalance, and the optimization criteria prioritized (Villar & de Andrade, 2024).

There is no consensus on the relative performance of each classifier type. Random Forest is highly competitive, often performing among the best when demographic, socioeconomic, and academic variables are used as input variables (Kabathova & Drlik, 2021). However, Random Forest models do not always outperform other classification models. Other boosting models, such as XGBoost, LightGBM, and CatBoost, tend to outperform Random Forest when optimized for imbalanced datasets or when the primary objective is to maximize the early recall of at-risk students (Carballo-Mendivil et al., 2025; Villar & de Andrade, 2024). Systematically comparing multiple models avoids the assumption that one “best” model exists a priori. Within Latin America, Gutierrez-Pachas et al. (2023) combined ML classifiers with survival analysis methods over 13 years of demographic, socioeconomic, and academic records from a Peruvian university, modeling not only who drops out but also when, and highlighting the value of long-run institutional datasets in the region.

Recently, an XGBoost model was created using only pre-enrollment data, resulting in an area under the curve (AUC) score of 0.69. The XGBoost model was selected because it achieved higher recall rates for at-risk students than the Random Forest model, despite their similar AUC scores (Carballo-Mendivil et al., 2025). Additionally, studies in Latin American higher education report ROC-AUC values ranging from 0.78 to 0.96, with the highest values associated with models that incorporate in-semester academic performance data alongside administrative records (Arévalo-Cordovilla & Peña, 2025; Palacios et al., 2021). Models relying exclusively on pre-enrollment data, as in the present study, tend to have more moderate AUC values, reflecting the inherent constraint of predicting outcomes before any university-level academic behavior is observed.

3. Materials and Methods

The present study adhered to the CRISP-DM process, as delineated by Chapman et al. (2000). The preliminary step involved establishing a business understanding of the issue, defining dropout as an institutional priority requiring probabilistic risk scoring, and subsequently interpreting the data through an exploratory analysis of variable distributions, missing values, and class imbalance. The second step was to address construct operationalization, translate it into clusters of measurable variables, and evaluate their internal consistency using Cronbach’s alpha (α). The next step in the data preparation process involved cleaning, imputation, feature engineering, and encoding. The fifth step in the process entailed constructing and comparing the performance of eight supervised algorithms on the training sets, and then evaluating them using a range of metrics. The final step involved generating risk scores and establishing a reproducible framework for evaluating the subsequent cohorts.

3.1. Study Context and Data Source

This study was conducted at the Sonora Institute of Technology (ITSON), Cd. Obregón, Sonora, in northwestern Mexico. At this university, a student is considered a dropout if they are enrolled in an academic program and fail to re-enroll for at least four consecutive semesters after a break.

The study's dataset included 47,000 records of new, first-time undergraduates who began attending classes at ITSON between 2014 and 2025 across its four campuses: Guaymas, Empalme, Navojoa, and Ciudad Obregón (Figure 1). Due to two cohort anomalies that occurred during this timeframe, the data decreased significantly: one occurred when surveys were transitioned to remote delivery in 2021 due to the pandemic, and the other resulted from a database failure in 2024. However, neither event was severe enough to produce a sufficient distributional bias to exclude these observations from the study.

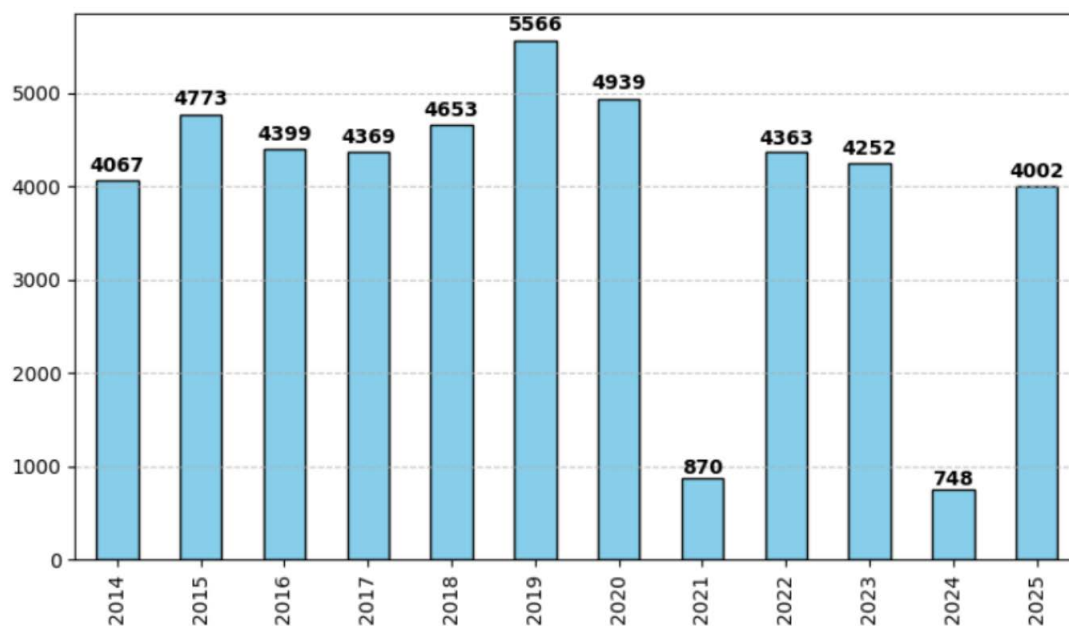


Figure 1. Number of incoming undergraduate students per enrollment cohort at ITSON (2014–2025).

These data were obtained from two sources: the institutional record of the dropout and data collected through a questionnaire administered during the admission process. This database provided access to demographic information on each student, including their socioeconomic status, academic performance, and personal traits of interest. Although the questionnaire collected broad information about the admitted class, the data generated by the questionnaire had only been used for descriptive purposes, with no utilization in predictive models or decision support tools, until the previous study was conducted (Carballo-Mendivil et al., 2025), which established a proof of concept, but fell short of achieving adequate precision and calibrated probabilities.

3.2. Variable Framework and Construct Operationalization

Eight theoretically grounded constructs were operationalized using the institutional intake questionnaire (Table 1). Each construct is tied to at least one of the three theoretical frameworks that guide this study: the Academic/Social Integration Model (Tinto, 1993), the Non-Traditional Student Model (Bean, 1980), and the Persistence Model (Cabrera et al., 1993). Therefore, each predictor was entered into the model based on a theory and was not simply included for convenience. The constructs developed by Tinto represent students' academic preparedness and fit upon entering their institution. Those developed

by Bean illustrate extrinsic factors, such as work and family responsibilities, that can interfere with students' ability to engage academically. The constructs developed by Cabrera et al. (1993) expanded the model to include financial and institutional support, which may influence the connection between students and institutions, regardless of the level of student–institution integration. Where multiple frameworks converge on the same construct, all theoretical bases are retained in Table 1, as this overlap is itself theoretically meaningful, as Cabrera et al. (1992) demonstrated by merging Tinto's and Bean's models into a single predictive framework.

Table 1. Theoretical constructs, variable operationalization, and internal consistency of composite indices.

Construct	Theoretical Basis	Key Variables	Internal Consistency
Prior academic trajectory	Tinto (1993)	GPA, failed subjects, years in high school	Extracurricular Index: $\alpha = 0.895$
Socioeconomic status & cultural capital	Cabrera et al. (1993); Spady (1970); Bourdieu (1998)	Parental education, household assets, travel	Basic Assets Index: $\alpha = 0.845$; Secondary Assets Index: $\alpha = 0.711$; Travel Index: $\alpha = 0.529$ (items retained individually)
Health and well-being	Tinto (1993); Jessor and Jessor (1977); Astin (1984)	Illness, special needs, alcohol consumption	Special Needs Index: $\alpha = 0.011$ (items retained individually)
Employment & academic linkage	Bean (1980); Kerby (2015)	Employment status, job–field relatedness	Single items
Transportation & logistics	Cabrera et al. (1993)	Distance to campus, commute mode, commute time	Single items
Family responsibilities & social support	Tinto (1993); Bean (1980)	Dependents, marital status, parental support; career-choice conversations (family, friends, teachers)	Single items & Family Factor Index: $\alpha = 0.564$ (items retained individually)
Motivation & expectations	Tinto (1993); Kerby (2015)	Career satisfaction, choice factors, persistence factors	Career Satisfaction Index: $\alpha = 0.779$; Choice Index: $\alpha = 0.572$ (items retained individually); Persistence Factor: $\alpha = 0.291$ (items retained individually)
Demographic characteristics	Tinto (1993); Spady (1970)	Age, gender	Single items

Note: Indices with $\alpha < 0.70$ were not created; constituent items were retained individually in the feature set.

Most questionnaire items served as single predictors, such as high school GPA, age, and commute time, and were maintained individually. However, nine indices were created from multiple items conceptualized to measure related constructs. In these cases, Cronbach's alpha was used to assess the internal consistency computed exclusively from the training partition. An alpha value of 0.70 was selected based on the guidelines for assessing the quality of scales and tests described by George and Mallery (2019). Indices that met the criterion of $\alpha \geq 0.70$ were combined into composite indices via row means, and the individual items comprising each index were deleted to reduce the dimensionality. In contrast, indices that failed to meet this criterion were not created, and all items contained within them remained part of the original feature set. This process ensured that dimensionality reductions would only be made when it could be demonstrated that aggregating items at the scale level was statistically valid and reliable. In addition, selecting 0.70 as the

cutoff point occurred before training the models, so there was no risk of overfitting due to data leakage.

3.3. Data Preparation Pipeline

The data preparation steps were as follows: first, the temporal split strategy was applied, obtaining (1) 30,398 records for training (2014–2022), (2) 7600 records for internal validation (20% of the training set was randomly selected and stratified for division), and (3) 9002 entries for testing (2023–2025).

Second, the missing-value imputation process entailed replacing missing values in the numeric variables with values derived from the training set. For categorical variables, missing values were either imputed from the training set, designated as ‘Unknown,’ or intentionally left blank and retained, then encoded as binary flag features for tree-based modeling. However, in the case of institutional variables (such as program and campus), the “unknown” treatment was replaced by the training mode.

Third, as previously mentioned, several items were grouped to create nine indices, but only four exceeded Cronbach’s alpha (>0.7): The Extracurricular Index ($\alpha = 0.895$), Basic Assets Index ($\alpha = 0.845$), Career Satisfaction Index ($\alpha = 0.779$), and Secondary Assets Index ($\alpha = 0.711$). Columns with records of individual items were removed. The dataset included 76 columns.

The nine candidate indices were constructed from a total of 49 item-level variables. Several questions on the instrument were multi-option, so each response option was scored as a separate dichotomous item. The indices were the Extracurricular Index (five items), Basic Assets Index (four items), Secondary Assets Index (nine items), Career Satisfaction Index (five items), Choice Index (nine items), Special Needs Index (five items), Family Factor Index (four items), Travel Index (two items), and Persistence Factor (six items). The full wording of each item, translated from the original Spanish instrument, together with its mean, standard deviation, and standardized factor loading estimated on the training partition, is reported in Appendix A (Table A1). This allows the operationalization and measurement of each construct to be reproduced and verified.

At this point, the preparation pipeline for the tree model was complete, and the dataset for training was saved for further analysis. However, for distance-based models, three additional steps were applied: (1) Z-score normalization with Winsorization (1st–99th percentile), (2) One-Hot Encoding with VIF-based Collinearity Removal, and (3) SMOTE class balancing (the `scale_pos_weight` or `class_weight = ‘balanced’` parameter was applicable to tree-based models).

3.4. Training Models and Evaluation

Two families of algorithms were tested separately for each family: (1) tree-based models, including Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), and CatBoost (CatB). And (2) Distance-based models included Logistic Regression (LR), Support Vector Machine (SVM), AdaBoost (AdaB), and Artificial Neural Network (ANN).

Each model received the appropriately pre-processed data, as specified above. An initial evaluation of the default 0.50 threshold was performed for each model separately. A subsequent evaluation was conducted in which the decision thresholds were adjusted to optimize the F1 score for each model over the grid [0.10, 0.90]. For the model that demonstrated the highest level of performance (Random Forest), the following additional evaluations were conducted: (1) hyperparameter tuning via a randomized search with cross-validation (25 iterations, 3-fold stratified); (2) isotonic probability calibration using `CalibratedClassifierCV`; and (3) the decision threshold was re-optimized on the internal validation set.

The model performance was assessed using Accuracy, Precision, Recall (sensitivity), F1-Score, ROC-AUC, and PR-AUC. The Brier Score was additionally reported as a measure of probabilistic calibration quality, alongside the reliability diagram. The reliability diagram, also known as the calibration curve, was used to evaluate the accuracy of the probability calibration by demonstrating the relationship between the predicted and actual values. In addition, the Variable Importance was analyzed using the Mean Decrease in Impurity (MDI) and SHAP summary plots.

In addition, to assess whether the observed differences between the best-performing models could be attributed to sampling variability, 95% confidence intervals (CIs) for the ROC-AUC, F1-score, and recall were estimated using a stratified bootstrap procedure ($B = 1000$ resamples) on the internal validation set. Pairwise differences between Random Forest and the other tree-based ensembles were evaluated through the bootstrap distribution of each metric difference; a difference was considered statistically meaningful only when the 95% CI of the difference excluded zero.

Finally, the optimized model was evaluated using the held-out test partition (7600 records from the 2014–2022 cohorts), which was not used at any stage of training, tuning, calibration, or threshold selection, to approximate real-world applicability. The 2023–2025 cohorts (9002 records) were reserved for prospective risk scoring only (Section 4.5), because under the institutional definition, a student must remain unenrolled for four consecutive semesters to be classified as a dropout, so their outcome labels were not yet observable at the time of this study.

4. Results

4.1. Comparative Model Performance

A comprehensive overview of the outcomes of the eight models is presented in Table 2. The Random Forest model demonstrated an optimal F1-Score of 0.531 and a nearly equivalent ROC-AUC Score of 0.743. In contrast, the CatBoost and XGBoost models exhibited F1-Scores of 0.503 and 0.496, respectively. In general, distance-based models demonstrated substandard performance compared with the remaining seven models on both the F1-score and ROC-AUC. However, the AdaBoost model achieved higher PR-AUC scores, indicating greater sensitivity compared to the other models. Brier Scores followed the same pattern as ROC-AUC, with tree-based models achieving lower values (RF = 0.169; CatBoost = 0.206) than distance-based classifiers (ANN = 0.298; AdaBoost = 0.239), indicating better probability calibration in ensemble methods even before the isotonic calibration step.

Table 2. Optimized performance metrics for all evaluated models (internal validation set).

Model	F1	Precision	Recall	Accuracy	ROC-AUC	PR-AUC	Brier
Random Forest	0.531	0.425	0.708	0.665	0.743	0.520	0.169
CatBoost	0.503	0.391	0.706	0.627	0.704	0.473	0.206
XGBoost	0.496	0.414	0.619	0.664	0.703	0.472	0.195
LightGBM	0.493	0.383	0.691	0.664	0.693	0.460	0.197
AdaBoost	0.481	0.351	0.763	0.560	0.678	0.684	0.239
SVM	0.468	0.336	0.771	0.531	0.661	0.656	0.203
Logistic Regression	0.458	0.332	0.739	0.532	0.645	0.644	0.228
ANN	0.404	0.383	0.506	0.650	0.604	0.624	0.298

Bootstrap analysis indicated that Random Forest achieved an ROC-AUC of 0.743 (95% CI [0.730–0.756]), compared with 0.704 (95% CI [0.689–0.718]) for CatBoost and 0.703 (95% CI [0.688–0.718]) for XGBoost. The 95% CIs of the pairwise differences between the Random Forest and gradient-boosting ensembles excluded zero for ROC-AUC ($\Delta\text{RF-CatBoost} = 0.039$, 95% CI [0.031–0.046]; $\Delta\text{RF-XGBoost} = 0.040$, 95% CI [0.031–0.050]) and F1-score ($\Delta\text{RF-CatBoost}$, 95% CI [0.017–0.039]; $\Delta\text{RF-XGBoost}$, 95% CI [0.021–0.048]). For recall, the difference excluded zero with respect to XGBoost (95% CI [0.068–0.109]) but included zero with respect to CatBoost (95% CI [−0.014–0.019]), indicating that both models had comparable sensitivity. Consequently, the advantages of Random Forest in discrimination (ROC-AUC) and precision–recall balance (F1) are statistically supported, whereas its recall is statistically indistinguishable from that of CatBoost.

4.2. Random Forest Optimization and Probability Calibration

Table 3 presents a comparison between the base Random Forest and the tuned-and-calibrated version. The primary outcome of hyperparameter tuning and isotonic calibration was a shift in the optimal threshold from 0.31 to 0.25, with a concomitant prioritization of recall (0.709 vs. 0.736) at a modest-precision cost (0.425 vs. 0.413). F1 demonstrated negligible change (0.531 vs. 0.529), whereas the ROC-AUC showed stability (0.743 vs. 0.738). The calibration curve demonstrated that the estimated probabilities of the tuned model exhibited a substantial improvement in alignment with the observed event rates compared with the base model, which exhibited a systematic overestimation of the dropout risk.

Table 3. Performance comparison between the base and tuned-and-calibrated Random Forest (internal validation set).

Model	Threshold	F1	Precision	Recall	Accuracy	ROC-AUC	PR-AUC	Brier
RF (Base)	0.31	0.531	0.425	0.709	0.665	0.743	0.520	0.1694
RF (Tuned + Calibrated)	0.25	0.529	0.413	0.736	0.649	0.738	0.490	0.1676

Aggregate metrics showed a notable improvement in recall (0.709 to 0.736), with only slight decreases in precision, F1 score, and ROC-AUC. This trade-off was intentional, based on the institution’s desire to minimize false negatives. Additionally, although there was a marginal improvement in the Brier score, it confirms that the isotonic calibration process reduced the probabilistic error. This is evident from the improved reliability diagram alignment while maintaining stability in the discrimination metrics.

The impact of using an isotonic calibration approach for probability estimates is illustrated in Figure 2. The base model tended to consistently overestimate the likelihood of dropouts at all points along the probability scale; thus, it was far from the ideal diagonal line that would represent perfect accuracy, especially in the middle range (0.20–0.50) of the probability scale, which contains the bulk of students. Calibration clearly aligned the actual estimated probabilities with observed dropout rates, indicating that the calibrated model could be used to estimate true dropout probabilities versus simply providing relative rankings. The institutional use of this type of model requires the capability to identify and act upon differences, such as a student who is estimated to have a 30% chance of dropping out versus a student who is estimated to have a 70% chance of dropping out.

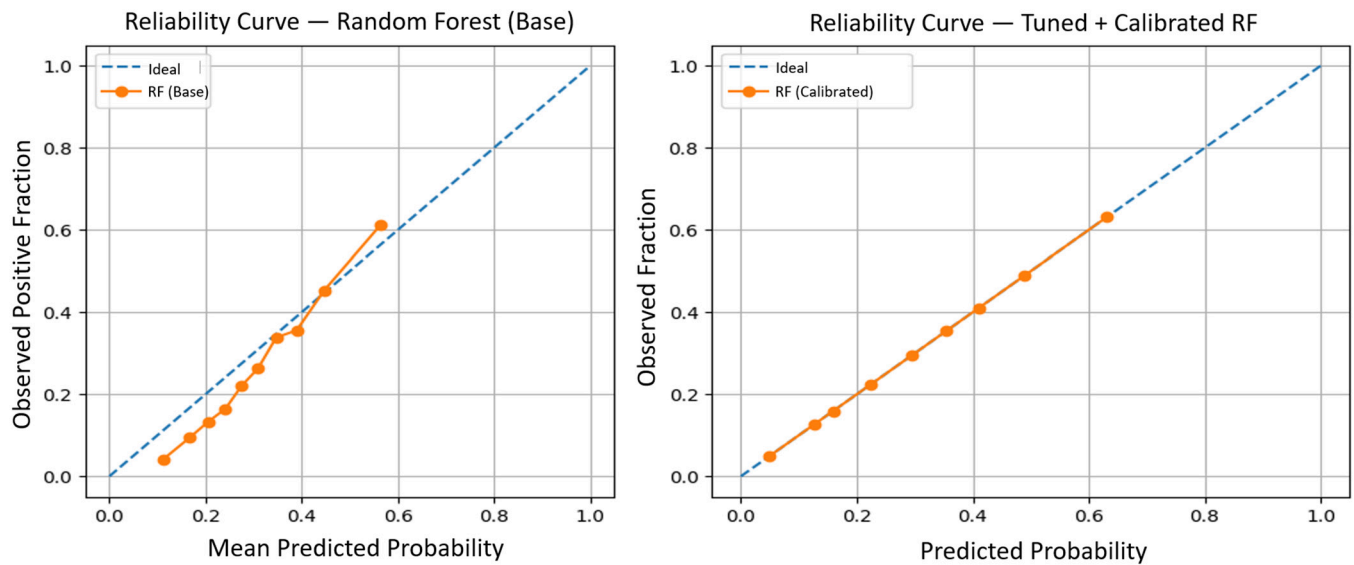


Figure 2. Reliability curves comparing the base Random Forest and the tuned-and-calibrated Random Forest. The calibrated model showed a substantially closer alignment with the ideal diagonal, indicating improved probability calibration accuracy.

4.3. SHAP-Based Variable Importance

To transition from the prediction of dropouts as a group to the comprehension of the student characteristics most strongly associated with each student's individualized dropout risk prediction, two different methodologies were utilized to assess the relative importance of each predictor (student characteristic): (1) Mean Decrease in Impurity (MDI), shown in Figure 3, quantified the contribution of each feature to the decision regarding the allocation of data during model training, and (2) Shapley Additive Explanations (SHAP) values ascertained the extent to which each feature contributed to an individual student's predicted probability of dropping out and the direction of the effect, whether positive or negative (Figure 4).

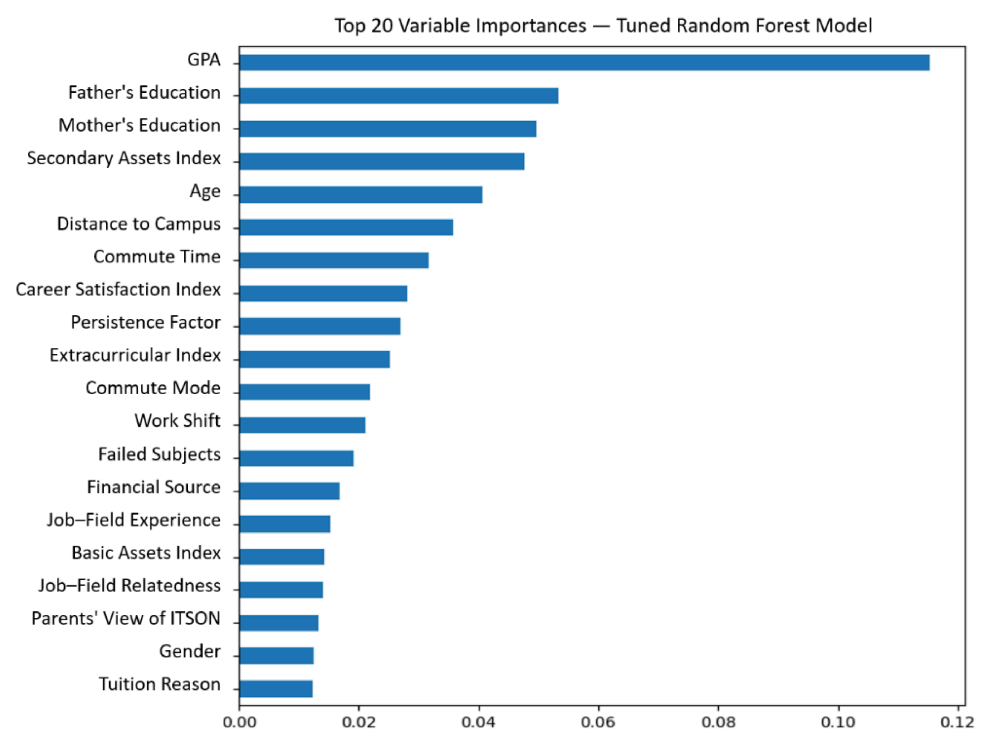


Figure 3. Top 20 Variable Importances—Tuned Random Forest Model.

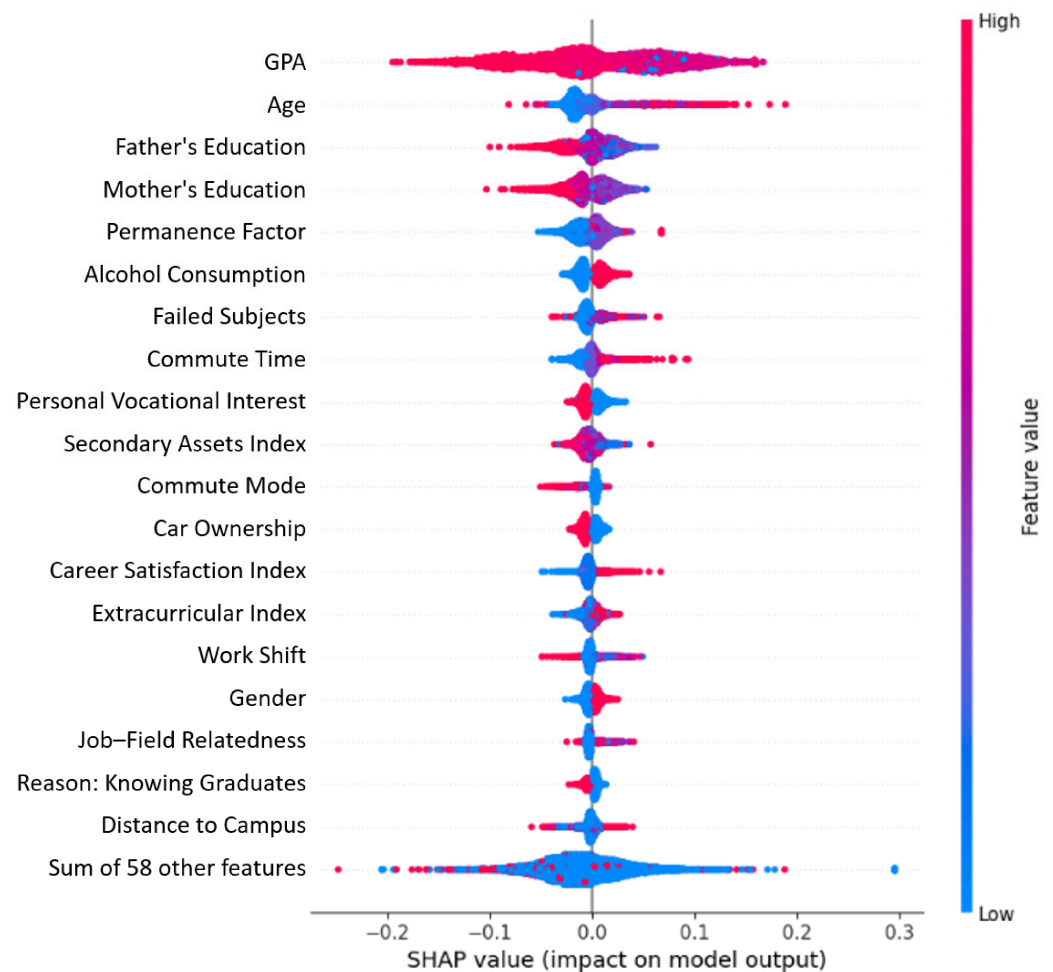


Figure 4. SHAP summary plot for the tuned Random Forest model. Each dot represents an individual student. The color indicates the feature value (red = high, blue = low), and the horizontal position indicates the direction and magnitude of the impact on the dropout probability.

Both the MDI ranking (Figure 3) and SHAP summary plot (Figure 4) converged on the same top-ranking predictors. High School GPA was the most important predictor of student dropout; students with low GPAs were assigned large positive SHAP Values, which increased their predicted probability of dropping out, whereas students with a higher GPA were assigned larger negative SHAP values, which contributed to a lower predicted dropout probability. Parental Education Level ranked second, providing additional associative evidence consistent with Bourdieu's (1998) theory of Cultural Capital. Next in line were the Permanence Factor (a single questionnaire item asking which factor could hinder the student's permanence at the university), how often students consumed alcohol, how many subjects a student failed in high school, how old they were when they first enrolled in college, and how long it took them to commute to class. These findings are coherent with Tinto's (1993) integration model and Bean's (1980) external-pressures framework.

Viewed against the directional hypotheses formulated in the Introduction, these results provided associative support, in the hypothesized direction, for H1 (prior academic trajectory), H2 (socioeconomic status and cultural capital), H3 (health and well-being), H5 (transportation and logistics), H7 (motivation and expectations), and H8 (demographic characteristics). In contrast, the constructs linked to H4 (employment and academic linkage) and H6 (family responsibilities and social support) did not rank among the top predictors, indicating weaker predictive associations with dropout in the pre-enrollment context. It

should be stressed that these results reflect the predictive associations learned by the model and do not constitute causal tests of the underlying theories.

4.4. Validation on the Held-Out Test Set

Table 4 summarizes the results for the held-out test dataset (7600 records, 2014–2022). The optimized and calibrated Random Forest yielded a recall of 0.743, an F1 score of 0.524, an ROC-AUC of 0.734, a PR-AUC of 0.478, and a Brier Score of 0.167. A minor decline was observed in each metric relative to the internal validation values, as expected when transitioning from optimized conditions to fully unseen data, which represented an acceptable generalization gap. The confusion matrix revealed that the model correctly identified most students who would eventually drop out, while producing a moderate number of false positives, a justifiable trade-off in retention contexts, where failing to flag a true at-risk student carries substantially higher institutional costs than unnecessary outreach.

Table 4. Performance metrics for the held-out test set (2014–2022 cohorts).

Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Brier
RF (Tuned + Calibrated)	0.639	0.404	0.743	0.524	0.734	0.478	0.167

The ROC Curve for the held-out test dataset (see Figure 5) illustrates the model's continued ability to distinguish between students at risk and those who are not at risk on new, unseen data, with an Area Under the Curve (AUC) of 0.734. The orange dashed diagonal line represents the performance of a random classifier and serves as the baseline for comparison, corresponding to an AUC of 0.50. This AUC was slightly lower than that obtained in the internal validation (0.738); however, it demonstrated consistent separation across all thresholds tested and far exceeded the expected value.

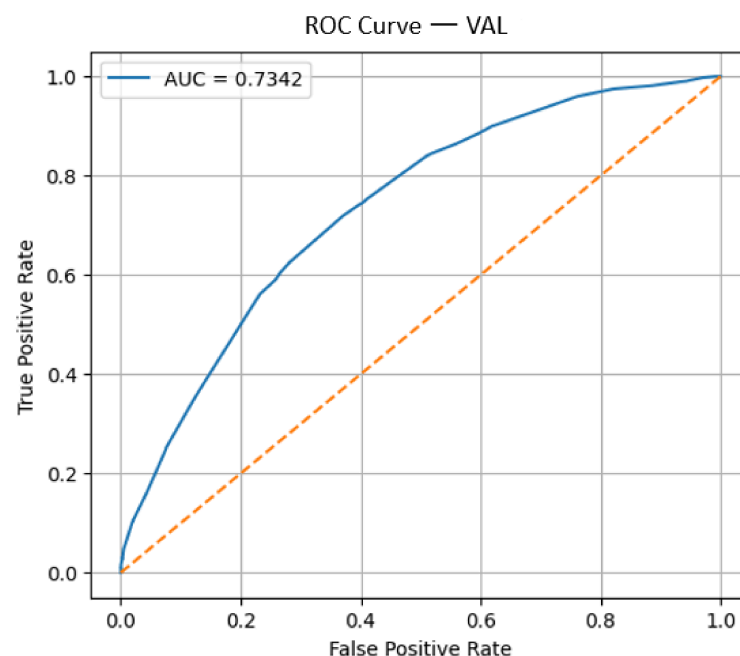


Figure 5. ROC curve for the tuned and calibrated Random Forest on the held-out test set. AUC = 0.734.

The precision-recall curve (Figure 6) provides an alternative view of the model and is particularly useful for problems with an extreme class imbalance. As shown, Model PR-AUC = 0.478 is significantly better than the Baseline AUC = 0.268 (red dashed line).

This demonstrates that the model can provide a signal value when probabilities are used to detect the minority class. The model maintained a reasonable level of precision across the range of possible recall values. However, it eventually began degrading. This shows that the model focuses on correctly detecting true dropouts and does not collapse at higher-recall levels.

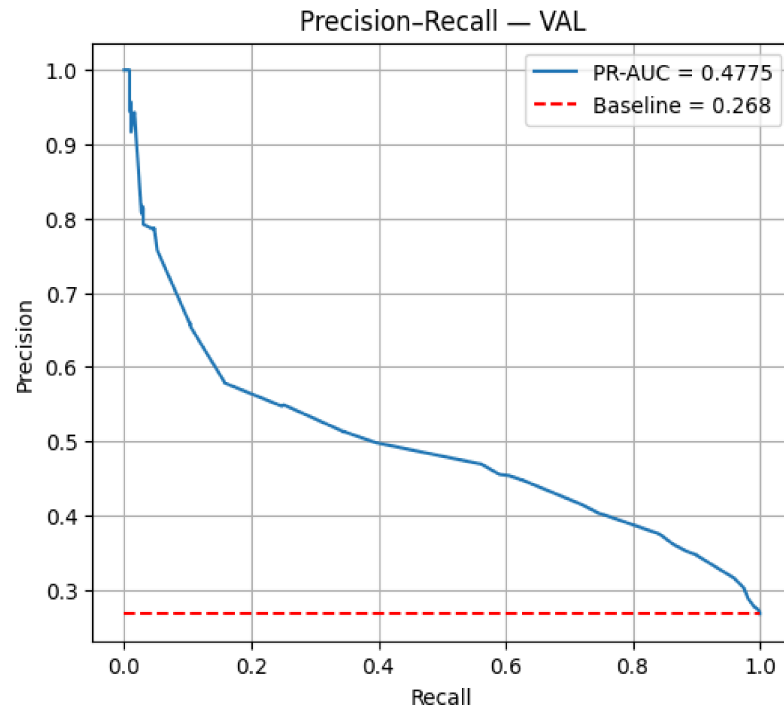


Figure 6. Precision–recall curve for the tuned and calibrated Random Forest on the held-out test set. PR-AUC = 0.478; baseline = 0.268.

The results from the reliability curve constructed using the validation dataset (Figure 7) also show that the model maintains a calibrated probability estimation for new, or unseen, data. The model’s estimated probabilities tend to follow the diagonal line across most of the range of possible values (0.00–1.00). However, there was some evidence of slightly higher estimated probabilities above 0.60, with observed dropout rates of approximately 0.58 and estimated probabilities of approximately 0.63. While this represents a minor deviation from ideal calibration, it has no negative impact on the functional use of the risk scores; they can be interpreted as true probability estimates and therefore retain their operational value as measures of the relative risk.

Finally, Figure 8 presents the confusion matrix for the held-out test data, which shows that the classifier had 3343 true negatives (non-dropouts) and 1510 true positives (dropouts). In addition, the classifier produced 2224 false positives and 523 false negatives. More importantly, the number of false negatives was significantly lower than that of false positives, which reflects the prior optimization approach of selecting a recall-based threshold. Thus, from an institutional perspective, the most important output of the system is the 1510 students whose risk of dropping out was correctly identified.

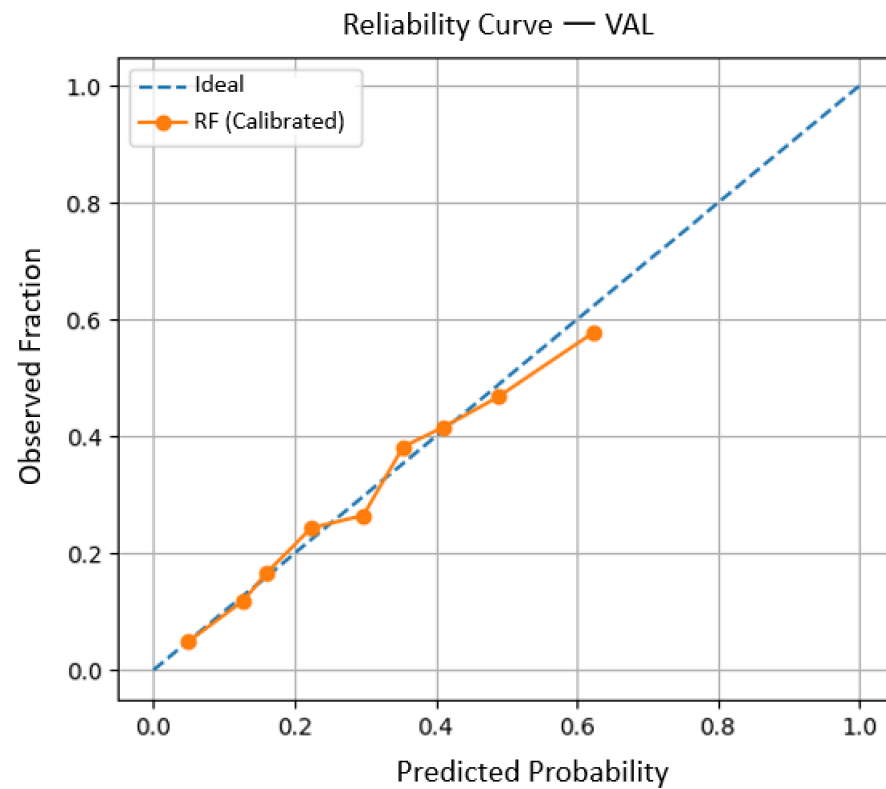


Figure 7. Reliability curve for the tuned and calibrated Random Forest on the held-out test set. The estimated probabilities were closely aligned with the observed dropout rates, confirming that the calibration generalizes beyond the training partition.

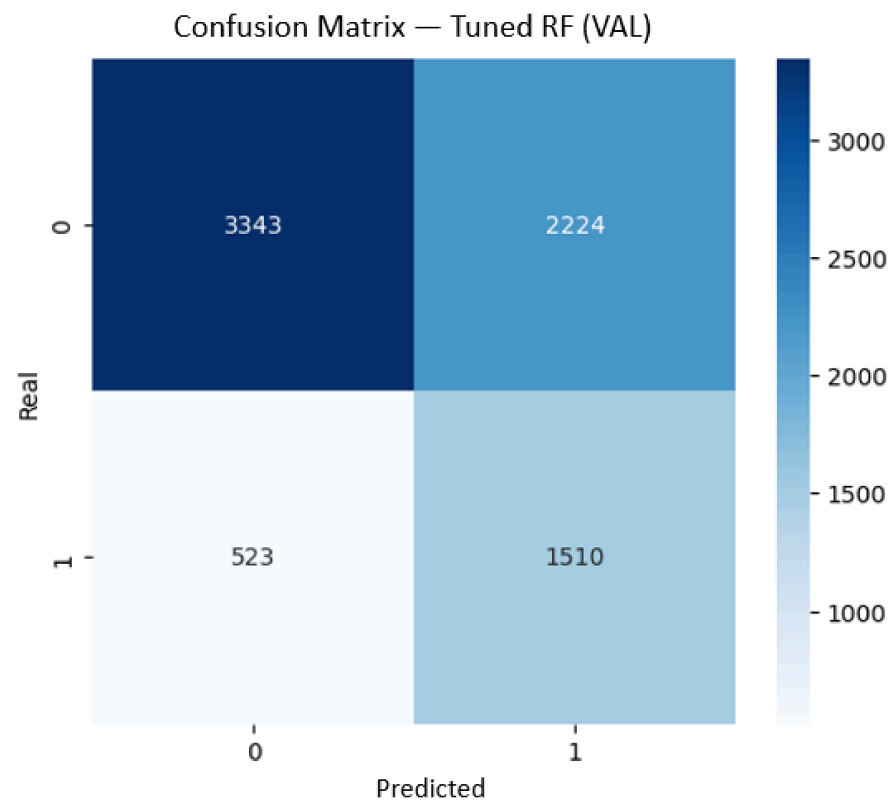


Figure 8. Confusion matrix for the tuned and calibrated Random Forest on the held-out test set. True negatives: 3343; false positives: 2224; false negatives: 523; true positives: 1510.

4.5. Operational Risk Profiling

The final model was applied to the 2023–2025 cohorts, resulting in the identification of several students with an elevated probability of dropping out. For illustrative purposes, the 20 students with the highest predicted risk are listed in Table A2 (Appendix B), with dropout probabilities ranging from 59.5% to 73.7%. In addition to being male, most cohort members in the highest-risk category were between 20 and 32 years old. Several of these students were also confirmed to have not been enrolled for the term spanning August to December 2025. Because the institutional definition of a dropout requires four consecutive semesters without enrollment, outcome labels for the 2023–2025 cohorts were not observable at the time of writing. Therefore, these cohorts are used exclusively for prospective risk scoring, and the enrollment-status column in Table A2 reflects a preliminary manual verification for the 20 highest-risk students rather than a formal outcome label.

At the program level (Figure 9; restricted to programs with at least 30 incoming students in the 2023–2025 cohorts, with conditional-admission codes merged into their parent programs), the Bachelor's in Physical Activity and Sport Management (LDCFD; mean predicted probability = 0.346, $n = 75$), Manufacturing Engineering (IMAN; 0.291, $n = 81$), and Bachelor's in Exercise and Physical Sciences (LCIEF; 0.278, $n = 252$) exhibited the highest average dropout probabilities, suggesting structural curricular or contextual factors shared across student populations. Notably, Biosystems Engineering (IBIOS) also ranked among the high-risk engineering programs, likely reflecting its demanding quantitative curriculum.

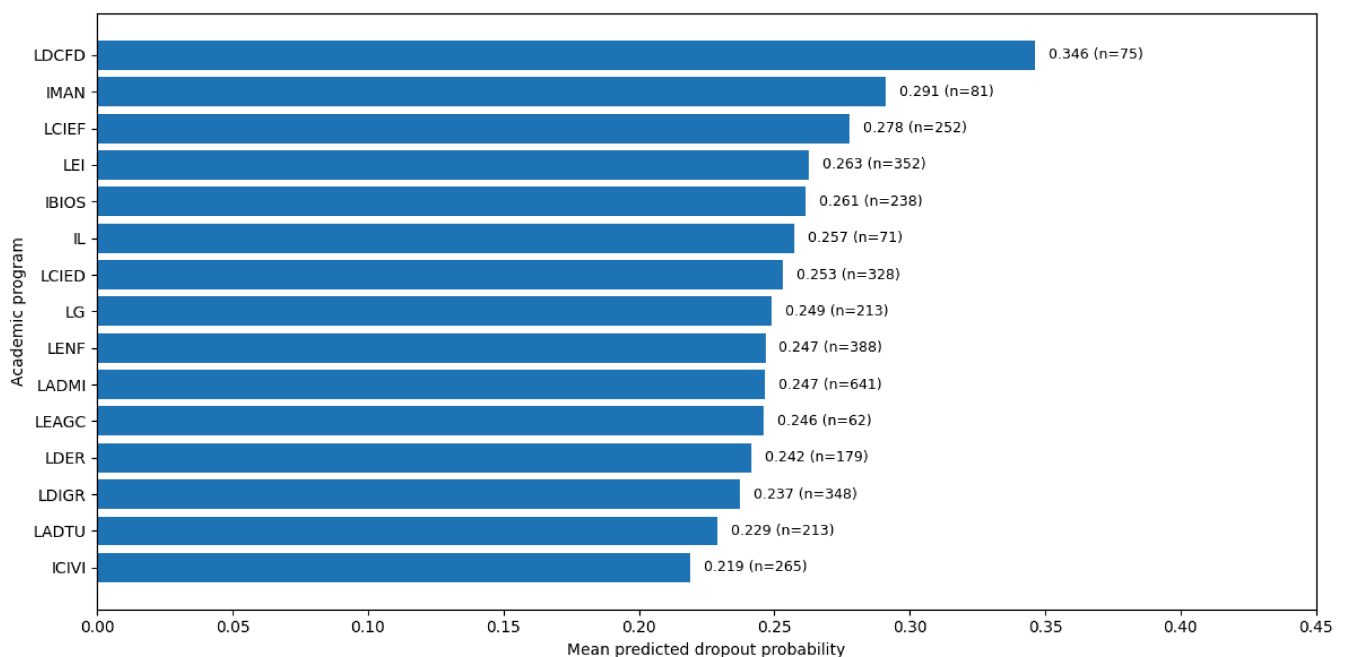


Figure 9. Top 15 Academic Programs by average predicted dropout risk (2023–2025 cohorts; programs with at least 30 incoming students; calibrated probabilities from the model trained exclusively on the labeled 2014–2022 cohorts).

5. Discussion

5.1. Model Performance and Benchmarking

As mentioned before, The Random Forest model trained achieved an area under the curve (AUC) of 0.734 and a recall of 0.743 on the held-out test set (2014–2022 cohorts). These values are considerably better than the previous benchmark (Carballo-Mendivil et al., 2025), where an XGBoost model with an AUC of 0.69 was obtained using a subset of features

available during the pre-enrollment period. This performance improvement is consistent with evidence indicating that strengthening the theoretical foundation and technical implementation of pre-enrollment features, probability calibrations, and thresholds yields quantifiable gains without requiring new data sources (Baneres et al., 2019; Von Hippel & Hofflinger, 2021).

However, this performance must be considered relative to the type of data used for training the model. Machine learning models that rely solely on pre-enrollment or admission data, as in the current study, generally yield lower AUCs than those that use first-semester academic performance records. Some of these models produced AUCs greater than 0.90.

The improved AUC over that of the previous study can be attributed to three specific design decisions made during the development of this model. First, the feature engineering process was theory-driven rather than data-driven. Each feature cluster was based on a theoretical construct (Bean, 1980; Cabrera et al., 1993; Tinto, 1993), and internal reliability was assessed using Cronbach's alpha before developing any models to ensure that the composite scores reflected reliable, sound theoretical constructs rather than random grouping of survey items (George & Mallery, 2019). Second, isotonic probability calibration was performed after hyperparameter optimization. This calibration greatly improved the alignment between the predicted probabilities and the actual dropout rates. Most similar studies did not perform post-hyperparameter optimization calibrations. Third, the optimal classification threshold was determined to optimize recall rather than accuracy, as the institution's goal was to minimize false negatives in a retention setting.

Among the eight algorithms evaluated, tree-based ensemble models consistently outperformed their distance-based counterparts, consistent with prior comparative evaluations (Kabathova & Drlik, 2021; Villar & de Andrade, 2024). The Artificial Neural Network achieved the lowest F1 score (0.404), likely because the high proportion of categorical and binary questionnaire items in the dataset does not lend itself well to neural architectures without embedding layers, a structural limitation that tree-based models do not share. CatBoost and XGBoost ranked second and third, respectively, reinforcing the idea that gradient boosting models are a competitive alternative to Random Forest when maximizing recall is the primary optimization criterion. However, Random Forest achieved the highest F1 score after threshold optimization and was therefore selected as the final model.

The precision of 0.404 in the held-out test set indicates that six out of ten students identified by the RF model as potentially at risk of dropping out would not experience adverse outcomes. The literature demonstrates that this trade-off exists across multiple domains in educational research. While it is true that the overly broad application of these types of models could result in high levels of alert fatigue among counseling professionals and stigmatizing labeling of students as at-risk, there appears to be no reason why institutions should not be willing to incur the relatively low costs associated with contacting students incorrectly designated as at-risk if doing so would prevent them from dropping out of school.

Finally, two methodological caveats deserve consideration. First, the modeling strategy adopted in this study is one of several defensible alternatives. Survival analysis techniques can model the timing of attrition rather than only its occurrence, as illustrated in the Latin American context by Gutierrez-Pachas et al. (2023). Simpler interpretable baselines, such as logistic regression, although outperformed in this study, remain valuable for institutional transparency. Second, the SHAP values quantify predictive associations conditional on the trained model and available data; they do not estimate causal effects. Accordingly, the variable-importance results reported in this study should be used to prioritize outreach and

motivate subsequent (quasi-)experimental evaluation of interventions, rather than being read as causal levers whose direct manipulation would necessarily reduce dropout.

5.2. Theoretical Implications

SHAP analysis provided an empirical assessment of the predictive relevance of the three theoretical frameworks on which the variable selection was based. The identification of High School GPA as the strongest predictor was consistent with Tinto's (1993) Academic Integration Theory. This supports the idea that students with weaker prior qualifications may struggle to adapt to higher-level studies. Consequently, they are more likely to drop out because of the gap between their existing skills and the demands of university courses, even before any semester data are collected. The results of this study were clear-cut: a low high school GPA was associated with a high likelihood of dropping out of college; conversely, high GPA levels were associated with strong negative values and a lower predicted risk. These findings are consistent with evidence from Latin American higher education showing that pre-admission academic trajectory is among the strongest institutional predictors of dropout, independent of other socioeconomic factors (Tapasco Alzate et al., 2019). These results can be used to counsel prospective students on admissions and the potential need for early tutoring to improve their performance in the future.

Parent education, in terms of both fathers' and mothers' educational attainment, ranked second and third in importance in the SHAP summary plot, respectively. This supports both Bourdieu's (1998) cultural capital theory and the socioeconomic dimension of Cabrera et al.'s (1993) persistence model. University students from families in which neither parent had a high level of education had limited cultural capital to help them navigate their university life. They may also have fewer role models and less knowledge of university operations. They may also have needed to rely on others to help them seek assistance within institutions of higher learning. These findings are supported by research indicating that first-generation college students face a higher risk of dropping out of college or university in various educational settings (Cabrera et al., 1993; Vilela Braga et al., 2018).

Both the MDI and SHAP importance ranks placed the secondary asset indices, commute distance, and commute time in middle tiers. These rankings reflect the structural barriers anticipated by Bean's (1980) non-traditional student model. Students who commute long distances spend considerable time and money, which competes with their academic engagement. This is particularly pertinent for ITSON students, given the campus locations, which attract students from several cities throughout Sonora.

The Career Satisfaction Index, the Permanence Factor item, and the Extracurricular Index, all of which were developed using questions about motivation and belonging, appeared in both importance measures, supporting Tinto's (1993) Social Integration Theory and Kerby's (2015) Vocational Adjustment Theory during the pre-enrolment period.

Overall, these results indicate that the theoretical frameworks not only provided conceptual justification for selecting variables but also anticipated the empirical hierarchy of their predictive importance. Additionally, the top three predictors (High School GPA, Parent Education, and Household Assets) corresponded exactly to the top three theoretical frameworks (Tinto, Cabrera/Bourdieu, and Bean), suggesting that the Integrated Multi-Model Approach captured a more complete and accurate representation of dropout risk than would have been possible using each framework individually (Cabrera et al., 1992; Müller & Klein, 2023). The correspondence between predictions based on theoretical frameworks and empirical hierarchies of importance is particularly important in Latin America, where it is relatively uncommon to validate theoretically based constructs in ML models using psychometrics (Castro-Montoya et al., 2025).

5.3. Practical Implications for Institutions

The model provides three outputs at the individual level. The first output allows tutoring offices, scholarship units, and academic counselors to identify which students to contact for outreach, based on their individual calibrated dropout probability scores. Ideally, this should be done immediately after the first day of the attendance. As these dropout probability scores are calibrated, a student estimated to have a 70% chance of dropping out is considered to be in substantially greater need of early intervention than one estimated to have a 30% chance of dropping out. Therefore, resources can be allocated according to risk intensity rather than being flagged with binary code. However, this significant benefit of interpretability was possible only by including an isotonic calibration step in the model.

The second output (program) relates to the aggregate risk profile of each program. It has been demonstrated that dropout risk is not uniformly distributed among ITSON's various academic programs. For instance, the dropout risk appears to be notably higher for programs such as the LDCFD, IMAN, and LCIEF. This suggests that structural elements, such as curricular intensity, vocational alignment between student expectations and program realities, and the availability of student support resources, interact with individual-level risk in ways that cannot be fully counteracted by a single university-wide strategy. Therefore, the program-level data provided by the model are useful for examining whether curriculum design, instructional methods, and/or student support service delivery systems need to be redesigned.

Finally, the high recall value (0.743) obtained by the model during testing indicates that approximately seven out of every ten students who would ultimately drop out were correctly identified at the time of enrolment. Consequently, 1510 of the 2033 dropouts in the held-out test partition were correctly identified. Students with comparable risk profiles in incoming cohorts can now receive targeted support from their first day at university, such as mentoring programs, emergency financial assistance, academic skills training workshops, and transport vouchers.

5.4. Ethical Considerations

The use of predictive models in schools raises ethical issues that must be acknowledged explicitly. Using demographic and socioeconomic indicators (e.g., parental education and household possessions) alongside geographic distance to campus to identify students who may need additional support is risky. If these predictions are used to deny students access to the services they need, rather than providing them access, the model can exacerbate existing educational inequalities. Therefore, the model's outputs can only trigger proactive interventions that are optional for both students and institutions. The outputs cannot determine whether a student is admitted to a course or program or whether any form of academic sanction is justified. In other words, the difference between identifying at-risk students and labelling them as such is not merely procedural. Rather, this is the ethical condition under which warning systems can be fairly implemented.

Additionally, algorithmic fairness requires ongoing evaluations to monitor potential differential errors based on gender, program type and campus. Although this study did not audit the model for differential error rates across the aforementioned subgroups, future implementations of this model should undergo systematic subgroup-performance analyses. As previously stated, false positives for certain demographic groups could lead to additional burdensome interventions for disadvantaged students, thereby creating a situation in direct opposition to the equitable goals that retention programmes were developed to achieve. Finally, students must be made aware that their institutions utilize predictive analytics in institutional retention programs. Furthermore, institutions must provide students with the

option to opt out of predictive analytics for retention purposes and clearly communicate how these analytics are utilized in their efforts. A human-in-the-loop governance structure would be beneficial for maintaining the human element in the decision-making process regarding retention strategies. Specifically, a well-trained counselor should review the risk flags generated by predictive modeling systems and make contact-related decisions using this information.

5.5. Limitations

This study has several limitations. First, the model was developed using data from a single university with a specific demographic, geographic, and curricular profile of the students. Therefore, it is difficult to ascertain whether the model's external validity extends to all Mexican universities or to other South American countries.

While the ITSON intake questionnaire (the source of the data) has been administered to every incoming cohort at ITSON since 2014, providing a substantial longitudinal sample size, there is no guarantee that other universities will collect similar data in the future. Therefore, cross-institutional validation, possibly through federated modeling, where individual university data remain private to their respective institutions, should be a high priority for demonstrating the generalizability of the findings.

Second, the model exclusively uses pre-enrolment data. This makes the current study unique but also limits its potential applicability. Consequently, factors that arise after enrolment, such as first exam failure, personal or financial crises, and family disruptions, are essentially invisible to the model at the time of the initial prediction. Implementing a longitudinal update mechanism, whereby the model can be continuously retrained with data becoming available as the semester progresses, would greatly increase predictive accuracy while maintaining the primary goal of providing timely dropout warnings to the university. Improvements in predictive accuracy have been demonstrated in studies that include first-semester data (for example: [Baneres et al., 2019](#); [Von Hippel & Hofflinger, 2021](#)), suggesting that this is probably the best way to increase the accuracy of future versions of the model.

Third, the dropout outcome was defined. The model predicts two outcomes: persistence versus abandonment. However, dropping out can have several consequences depending on the timing and circumstances. For instance, a student may voluntarily withdraw, be academically dismissed, transfer programs, or temporarily discontinue their studies. Each of these outcomes presumably has a different risk profile and responds differently to various interventions. Defining dropout as a multi-class problem could produce more actionable institutional intelligence than the current model, and this approach is recommended for consideration when developing subsequent models.

Fourth, the model only considers quantitative factors that influence decision-making. Although many non-quantitative influences exist (e.g., motivations and relationships), these do not appear in the quantitative model. Therefore, combining predictive modelling with qualitative data from student experiences (i.e., a mixed-methods approach) could provide a better explanation of why students identified as being at risk leave school and/or determine which intervention might be most effective for each risk profile.

Fifth, because dropout labels for the 2023–2025 cohorts will become observable only once those students accumulate four consecutive semesters without enrollment, the model's temporal generalization to genuinely future cohorts could not be evaluated. The held-out test performance reported here reflects within-period generalization across the 2014–2022 cohorts. A prospective evaluation of the 2023 cohort, whose labels will mature first, is an immediate next step and will provide a true out-of-time assessment of the early warning system.

5.6. Directions for Future Research and Institutional Implementation

While improving the predictive model's ability to accurately predict dropouts is crucial, it is not sufficient to lower dropout rates alone. While improving the model's accuracy in predicting dropouts can improve retention rates in real-world decision-making settings, research has repeatedly shown that this does not necessarily happen. This is because translating model-predicted risk scores into meaningful actions in an academic setting requires significant additional work from academic administrators.

A key next step for ITSON, as well as for other universities planning to adopt similar frameworks, is to integrate model-based risk scores into the regular management practices of academic programs. This would avoid relying solely on the potential benefits of using predictive models in centralized institutional initiatives, which often become less specific and less urgent once they are sent from administrative offices to students who require the most support.

More specifically, we suggest developing an interactive Business Intelligence Dashboard to present program-level and student-level risk profiles to the Academic Coordinator of each program. Such a dashboard could utilize platform tools such as Microsoft Power BI or Tableau to display the data produced by the dropout prediction pipeline. This type of visualization tool would enable program coordinators to view the risk distribution of their incoming cohort from day one, identify which students have been contacted, document the results of any interventions undertaken, and continue to do so throughout their degree progress.

Importantly, presenting the model's outputs at the program level, rather than directing all alerts through a central student services office, reduces the gap between the prediction and action. It places risk information in the hands of those responsible for supporting students in their academic pursuits.

Additionally, an approach based on a program-level dashboard addresses an important structural limitation associated with institution-wide alert systems. In these systems, the intensity of follow-up decreases as alerts move up the institutional hierarchy (from department to college) because of the diffusion effect. Therefore, by placing the responsibility for managing risk-identified students at the program level (i.e., where those students are likely to receive support) and making accountability explicit and local, a predictive model becomes a continuous management tool rather than simply a classification exercise.

6. Conclusions

In this study, we developed and demonstrated the validity of an operational early warning system for student dropout at one of Mexico's largest public universities. This system utilizes only data collected at the time of enrollment, that is, all the data available to any institution about students when they enroll before any academic performance is observed at the university level. It was used Tinto's (1993) integration model, Bean's (1980) framework for nontraditional college students, and Cabrera et al.'s (1993) persistence model to anchor each of our predictors. We also ran Cronbach's alpha internally on our predictors before training the models. Finally, we selected the best machine learning algorithm from eight supervised algorithms in two categories, applied isotonic probability calibration to optimize the thresholds, and trained a tuned Random Forest. Our tuned Random Forest achieved a test set recall of 74.3% and a test set ROC-AUC of 73.4% on a held-out test set drawn from the labeled 2014–2022 cohorts. This represents a meaningful improvement over the previous standard for dropout prediction established by ITSON. This is also consistent with the upper bound demonstrated for the predictive power of pre-enrollment data in Latin American higher education contexts.

This study uniquely integrates psychological measurements with statistical modeling. Thus, it addresses an important methodological gap in the literature. Most machine learning (ML)-based student dropout predictions do little, if anything, to confirm whether the input variables used to train them reflect the intended constructs. Conversely, virtually all theory-based student retention studies rarely build such models.

In addition to addressing this methodological gap, this study achieved a distinctive methodological integration by using a single institutional database of over 46,000 records covering nearly a decade of admissions. This study also includes theoretical foundations and the predictive models. To the best of our knowledge, no such study has been conducted in the context of Mexican higher education.

SHAP analysis revealed that the theoretical frameworks provided a justification for selecting the input variables and explained why some were statistically more important than others. Specifically, high school grades, parental education levels, and family wealth levels, corresponding to Tinto, Bourdieu/Cabrera, and Bean, respectively, were the top three predictors. These findings demonstrate that a theoretically informed subset of features yields understandable and empirically valid results. Additionally, the output probabilities produced by the final model can be clearly differentiated at multiple risk levels, enabling resource allocation based on risk intensity rather than simply producing a yes/no flag.

This study has four avenues for future research. First, incorporating the semester one academic performance into a longitudinal update would increase the model precision without compromising the ability to identify potential dropouts. Second, the impact of interventions guided by this model on actual dropout rates should be evaluated in a quasi-experimental manner. The current lack of causal evidence in this area of research is well-documented. Third, the transferability of the model to other institutions should be tested using federated modeling techniques that protect the student data. Finally, fairness-aware modeling should be employed to audit the error rates per demographic subgroup and ensure that the system reduces, rather than exacerbates, socioeconomic inequalities.

The above-mentioned represent clear avenues for future studies. However, the simplest way to translate predictive accuracy into actual institutional retention impacts is to translate model outputs into business intelligence dashboards accessible to academic department coordinators at the program level, rather than having to go through central administrative offices.

Author Contributions: Conceptualization, B.C.-M. and A.A.-G.; Methodology, B.C.-M., A.J.P.-M. and N.J.R.-V.; Formal analysis, B.C.-M. and A.J.P.-M.; Data curation, B.C.-M. and A.J.P.-M.; Writing—original draft, B.C.-M. and A.J.P.-M.; Writing—review & editing, A.A.-G., M.d.P.L.-D. and N.J.R.-V.; Supervision, M.d.P.L.-D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ITSON, Mexico. The APC was funded by the Program for the Promotion and Support of Research Projects (PROFAP), project reference number 2026-0581.

Institutional Review Board Statement: This study was conducted in accordance with the principles of the Declaration of Helsinki. Data were provided by ITSON in anonymized form for academic research purposes, in compliance with Mexican federal data-protection legislation (Ley Federal de Protección de Datos Personales en Posesión de los Particulares). Formal ethics committee approval was not required because there was no direct interaction with human subjects.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is not publicly available because of the institutional data privacy policy. Aggregated results and anonymized outputs are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript.

LARQ	Bachelor's in Architecture
IBIOS	Biosystems Engineering
ICIVI	Civil Engineering
IELME	Electrical and Electronics Engineering
IL	Logistics Engineering
IMAN	Manufacturing Engineering
IMECA	Mecatronics Engineering
IISIS	Industrial and Systems Engineering
LADMI	Bachelor's in Business Administration
LADTU	Bachelor's in Tourism Business Administration
LCIED	Bachelor's in Education Sciences
LCIEF	Bachelor's in Exercise and Physical Sciences
LCP	Bachelor's in Public Accounting
LDCFD	Bachelor's in Physical Activity and Sport Management
LDER	Bachelor's in Law
LDIGR	Bachelor's in Graphic Design
LEAGC	Bachelor's in Arts Education and Cultural Management
LEI	Bachelor's in Early Childhood Education
LENF	Bachelor's in Nursing
LG	Bachelor's in Gastronomy
LPSIC	Bachelor's in Psychology

Appendix A

Table A1 reports, for each composite index described in Sections 3.2 and 3.3, the constituent *item-level variables* of the institutional intake instrument, their wordings translated from the original Spanish, their descriptive statistics, and their standardized factor loadings computed exclusively on the training partition (2014–2022). Because several questions of the instrument are multi-option, each response option was scored as a separate dichotomous item; the source question of each item is indicated in parentheses. This information is provided so that readers and future researchers can see exactly how each construct was operationalized and measured and can reproduce and verify the data collection and index-construction process.

Table A1. Constituent item-level variables, item wordings, descriptive statistics, and standardized factor loadings for the nine candidate indices (training partition, 2014–2022).

Construct/Index	Item Code	Item Wording (English Translation)	M	SD	λ
Extracurricular Index ($\alpha = 0.895$; index created)	Group_ArtCultural (Q41)	During high school, participated in artistic-cultural groups	0.51	0.50	0.68
	Group_Sports (Q41)	During high school, participated in sports groups	0.61	0.49	0.55
	Group_Religious (Q41)	During high school, participated in religious groups	0.33	0.47	0.95
	Group_Political (Q41)	During high school, participated in political groups	0.32	0.47	0.96
	Group_StudentRep (Q41)	During high school, participated in student-representation groups	0.38	0.49	0.86

Table A1. Cont.

Construct/Index	Item Code	Item Wording (English Translation)	M	SD	λ
Basic Assets Index ($\alpha = 0.845$; index created)	Asset_Sewer (Q27)	Assets available at home: drainage	0.86	0.35	0.65
	Asset_Electricity (Q27)	Assets available at home: electricity	0.93	0.25	0.88
	Asset_GasStove (Q27)	Assets available at home: gas stove	0.90	0.30	0.76
	Asset_Water (Q27)	Assets available at home: drinking water	0.90	0.30	0.79
Secondary Assets Index ($\alpha = 0.711$; index created)	Asset_TvOpen (Q27)	Assets available at home: open-air television	0.51	0.50	0.00
	Asset_TvCable (Q27)	Assets available at home: cable television	0.47	0.50	0.59
	Asset_DomesticHelp (Q27)	Assets available at home: domestic staff	0.07	0.25	0.22
	Asset_WaterHeater (Q27)	Assets available at home: water heater	0.33	0.47	0.59
	Asset_WashingMachine (Q27)	Assets available at home: washing machine	0.71	0.45	0.66
	Asset_Landline (Q27)	Assets available at home: landline telephone	0.35	0.48	0.47
	Asset_Microwave (Q27)	Assets available at home: microwave oven	0.51	0.50	0.64
	Asset_AC (Q27)	Assets available at home: air conditioning	0.81	0.39	0.57
	Asset_Internet (Q27)	Assets available at home: internet connection	0.54	0.50	0.50
Career Satisfaction Index ($\alpha = 0.779$; index created)	Sat_WantedToStudy (Q51)	The chosen program is the one you want to study	0.95	0.25	0.00
	Sat_FirstChoice (Q51)	The chosen program was your first-choice option	0.84	0.40	0.15
	Sat_MajorChange (Q51)	You are planning to change programs (reverse-keyed)	0.32	0.66	0.94
	Sat_UniversityChange (Q51)	You are planning to change universities (reverse-keyed)	0.26	0.65	0.94
	Sat_MajorConfusion (Q51)	You feel confused about this program (reverse-keyed)	0.34	0.65	0.90
Choice Index ($\alpha = 0.572$; items retained individually)	Reason_Academic (Q56)	Reason for choosing ITSON: academic prestige	0.79	0.40	0.02
	Reason_Teachers (Q56)	Reason for choosing ITSON: high-level professors	0.52	0.50	0.41
	Reason_BestInstitution (Q56)	Reason for choosing ITSON: better than other institutions	0.40	0.49	0.49
	Reason_KnowAlumni (Q56)	Reason for choosing ITSON: knowing ITSON graduates	0.34	0.47	0.38
	Reason_MajorLevel (Q56)	Reason for choosing ITSON: high level of the chosen program	0.43	0.50	0.40
	Reason_FriendsEnrolled (Q56)	Reason for choosing ITSON: friends enrolled at ITSON	0.18	0.39	0.16
	Reason_EasyAdmission (Q56)	Reason for choosing ITSON: ease of admission	0.26	0.44	0.28
	Reason_StudentVibe (Q56)	Reason for choosing ITSON: student atmosphere	0.44	0.50	0.49
	Reason_Facilities (Q56)	Reason for choosing ITSON: adequate facilities	0.70	0.46	0.54

Table A1. Cont.

Construct/Index	Item Code	Item Wording (English Translation)	M	SD	λ
Special Needs Index ($\alpha = 0.011$; items retained individually)	SpecialNeed_Motor (Q6)	Special educational need: motor problems	0.79	0.40	0.02
	SpecialNeed_Hearing (Q6)	Special educational need: hearing impairment with hearing aids	0.52	0.50	0.41
	SpecialNeed_HearingCom (Q6)	Special educational need: complete or near-complete hearing impairment	0.40	0.49	0.49
	SpecialNeed_Visual (Q6)	Special educational need: visual impairment corrected with glasses	0.34	0.47	0.38
	SpecialNeed_VisualCom (Q6)	Special educational need: complete or near-complete visual impairment	0.43	0.50	0.40
Family Factor Index ($\alpha = 0.564$; items retained individually)	Factor_Friends (Q50)	Career choice influenced by conversations with friends	0.33	0.47	0.47
	Factor_Parents (Q50)	Career choice influenced by conversations with parents or guardians	0.53	0.50	0.56
	Factor_Relatives (Q50)	Career choice influenced by conversations with other relatives	0.28	0.45	0.54
	Factor_Teachers (Q50)	Career choice influenced by conversations with high school teachers	0.27	0.44	0.41
Travel Index ($\alpha = 0.529$; items retained individually)	Travel_Domestic (Q42)	Trips taken within the country	0.69	0.46	0.60
	Travel_International (Q43)	Trips taken outside the country	0.29	0.45	0.60
Persistence Factor ($\alpha = 0.291$; items retained individually)	Factor_Vocational (Q50)	Career choice influenced by vocational guidance in high school	0.42	0.49	0.23
	Factor_MajorInfo (Q50)	Career choice influenced by program information obtained at ITSON	0.61	0.49	0.54
	Factor_JobOpportunity (Q50)	Career choice influenced by future employment opportunities	0.53	0.50	0.34
	Factor_PersonalTaste (Q50)	Career choice influenced by personal vocation or liking	0.54	0.50	0.18
	Factor_Alumni (Q50)	Career choice influenced by knowing an ITSON graduate	0.09	0.29	0.20
	Factor_ExamScore (Q50)	Career choice influenced by admission-exam results leaving no other option	0.02	0.14	−0.05

Note: M = mean; SD = standard deviation; λ = standardized factor loading from a one-factor solution estimated exclusively for the training partition (2014–2022). The variable names correspond to the recoded analysis dataset; the original questionnaire item (QXX) is given in parentheses. Items marked “reverse-keyed” are worded in the direction opposite to the construct and require reverse scoring prior to aggregation.

Appendix B

Table A2 presents an individual-level illustration of the model output referenced in Section 4.5. It is provided in this appendix because its purpose is illustrative; the substantive analyses in the main text are conducted at the aggregate (program) level.

Table A2. The 20 students with the highest predicted dropout risks (2023–2025 cohorts).

N	Year	Gender	Age	Program	% Dropout Prob	Enroll AD25 (Manual Check)
1	2023	M	24	ICIVI	73.7	No
3	2023	M	20	IISIS	73.7	Yes
4	2023	M	20	LCIEF	73.7	Yes
2	2025	M	20	LADMI	73.7	Yes
5	2025	M	21	IELME	72.3	No
6	2023	M	21	IMECA	67.6	No
7	2023	M	29	LDCFD	67.6	No
8	2023	M	26	LG	67.6	No
9	2023	M	24	IISIS	67.6	No
10	2023	M	21	LDIGR	67.6	No
11	2025	M	32	LDCFD	67.6	Yes
12	2025	M	25	LCP	67.6	Yes
13	2025	M	24	IISIS	67.6	Yes
14	2025	M	25	LARQ	67.6	Yes
15	2023	M	23	LCIEF	66.7	No
16	2025	M	24	IBIOS	66.7	Yes
17	2025	F	21	LENF	66.7	Yes
18	2025	M	21	LPSIC	59.5	Yes
19	2023	M	28	IISIS	59.5	No
20	2023	F	23	LPSIC	59.5	No

Note: Probabilities were generated with the tuned and isotonically calibrated Random Forest trained exclusively on the labeled 2014–2022 cohorts. “Enrolled AD25” reflects the preliminary manual verification of enrollment in the August–December 2025 term.

References

- Arévalo-Cordovilla, F. E., & Peña, M. (2025). Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data. *Scientific Reports*, 15(1), 29420. [\[CrossRef\]](#) [\[PubMed\]](#)
- Astin, A. (1984). Student involvement: A development theory for higher education. *Journal of College Student Development*, 40, 518–529.
- Baneres, D., Rodriguez, M. E., & Serra, M. (2019). An early feedback prediction system for learners at-risk within a first-year higher education course. *IEEE Transactions on Learning Technologies*, 12(2), 249–263. [\[CrossRef\]](#)
- Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12(2), 155–187. [\[CrossRef\]](#)
- Bourdieu, P. (1998). *The state nobility: Elite schools in the field of power*. Stanford University Press.
- Cabrera, A. F., Castañeda, M. B., Nora, A., & Hengstler, D. (1992). The convergence between two theories of college persistence. *The Journal of Higher Education*, 63(2), 143–164. [\[CrossRef\]](#)
- Cabrera, A. F., Nora, A., & Castañeda, M. B. (1993). College persistence. *The Journal of Higher Education*, 64(2), 123–139. [\[CrossRef\]](#)
- Carballo-Mendivil, B., Arellano-González, A., Ríos-Vázquez, N. J., & Lizardi-Duarte, M. d. P. (2025). Predicting student dropout from day one: XGBoost-based early warning system using pre-enrollment data. *Applied Sciences*, 15(16), 9202. [\[CrossRef\]](#)
- Castro-Montoya, B., Vélez-Gómez, P., Segura-Cardona, A., & French, B. F. (2025). A cultural adaptation of Tinto’s student integration theory in undergraduate students of a private university in Colombia. *Cogent Education*, 12(1), 2479384. [\[CrossRef\]](#)
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Step-by-step data mining guide*. SPSS.
- Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498–506. [\[CrossRef\]](#)
- Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a Bayesian belief network-based DSS for predicting and understanding freshmen student attrition. *European Journal of Operational Research*, 281(3), 575–587. [\[CrossRef\]](#)
- George, D., & Mallery, P. (2019). *IBM SPSS statistics 26 Step by Step*. Routledge. [\[CrossRef\]](#)
- Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., & García-Castelán, R. M. G. (2023). Predictive analytics study to determine undergraduate students at risk of dropout. *Frontiers in Education*, 8, 1244686. [\[CrossRef\]](#)
- Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G., & Gomez-Nieto, E. (2023). Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the Latin American context. *Education Sciences*, 13(2), 154. [\[CrossRef\]](#)

- Hernández-Fernández, A., Ávalos-Ruiz, I., & Pantoja-Vallejo, A. (2023). Psychometric properties of a dropout prediction tool for students in Andalusia. *Frontiers in Education*, 8, 1121985. [CrossRef]
- Jessor, R., & Jessor, S. L. (1977). *Problem behavior and psychosocial development: A longitudinal study of youth*. Academic Press.
- Jud, M., Weissenbacher, B., & Neubauer, A. C. (2025). AI insights into initial teacher education: Predictive models for student dropout. *European Journal of Teacher Education*, 48(5), 1123–1142. [CrossRef]
- Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences*, 11(7), 3130. [CrossRef]
- Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, 10(1), 28–47. [CrossRef]
- Kerby, M. B. (2015). Toward a new predictive model of student retention in higher education. *Journal of College Student Retention: Research, Theory & Practice*, 17(2), 138–161. [CrossRef]
- Kwon, J. B. (2026). A portable, generalizable machine learning framework for long-term student dropout prediction. *IEEE Access*, 14, 30830–30843. [CrossRef]
- Lopez-Muñoz, G. L., Serrano, C. G., & Sanchez-Ferreira, C. (2026). Student dropout prediction in higher education: A systematic review of machine learning methods and risk factors. *Journal of Information Technology Education: Research*, 25, 20. [CrossRef] [PubMed]
- Lykkegaard, E., & Qvortrup, A. (2024). Longitudinal mixed-methods analysis of tertiary students' dropout considerations. *Frontiers in Education*, 9, 1399714. [CrossRef]
- Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. *Higher Education*, 80(5), 875–894. [CrossRef]
- Müller, L., & Klein, D. (2023). Social inequality in dropout from higher education in Germany. Towards combining the student integration model and rational choice theory. *Research in Higher Education*, 64(2), 300–330. [CrossRef]
- Nagy, M., & Molontay, R. (2024). Interpretable dropout prediction: Towards XAI-based personalized intervention. *International Journal of Artificial Intelligence in Education*, 34(2), 274–300. [CrossRef]
- Nikolaidis, P., Ismail, M., Shuib, L., Khan, S., & Dhiman, G. (2022). Predicting student attrition in higher education through the determinants of learning progress: A structural equation modelling approach. *Sustainability*, 14(20), 13584. [CrossRef]
- OECD. (2019). *Education at a glance 2019: OECD indicators*. OECD. [CrossRef]
- Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher Education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy*, 23(4), 485. [CrossRef] [PubMed]
- Pinheiro, C. B., Ribeiro, J. L. L. d. S., & Fernandes, S. A. F. (2023). Theoretical models of dropout in higher education and notes on the national context. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 28, e023015. [CrossRef]
- Reina Marín, Y., Quiñones Huatangari, L., Cruz Caro, O., Maicelo Guevara, J. L., Alva Tuesta, J. N., Sánchez Bardales, E., & Chávez Santos, R. (2025). Data mining to identify university student dropout factors. *Applied Sciences*, 15(22), 11911. [CrossRef]
- Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, 28(7), 8299–8333. [CrossRef] [PubMed]
- Shmeleva, E., & Froumin, I. (2020). Factors of attrition among computer science and engineering undergraduates in Russia. *Voprosy Obrazovaniya/Educational Studies Moscow*, 3(1), 110–136. [CrossRef]
- Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1), 64–85. [CrossRef]
- Tapasco Alzate, O. A., Ruiz Ortega, F. J., Osorio García, D., & Ramírez Ramírez, D. (2019). Deserción estudiantil: Incidencia de factores institucionales relacionados con los procesos de admisión. *Educación y Educadores*, 22(1), 81–100. [CrossRef]
- Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). University of Chicago Press.
- UNESCO. (2021). *Reimagining our futures together: A new social contract for education*. UNESCO. Available online: <https://unesdoc.unesco.org/ark:/48223/pf0000379381> (accessed on 15 January 2026).
- Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, 76, 102474. [CrossRef]
- Vilela Braga, L. F. H., Santos, N. M. d. S. B. F., Toledo, J. C., Albuquerque, A. A., & Gonzales, J. F. B. (2018). Influential factors in student retention: A study involving undergraduate students in accountancy. *Revista de Educação e Pesquisa em Contabilidade (REPeC)*, 12(3), 303–322. [CrossRef]
- Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, 4(1), 2. [CrossRef]
- Von Hippel, P. T., & Hofflinger, A. (2021). The data revolution comes to higher education: Identifying students at risk of dropout in Chile. *Journal of Higher Education Policy and Management*, 43(1), 2–23. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.