Article

Check for updates OPEN ACCESS

Journal of International Students
Volume 16, Issue 21 (2026), pp. 1-15
ISSN: 2162-3104 (Print), 2166-3750 (Online)
jistudents.org
DOI: https://doi.org/10.32674/admcxq87

An Explainable Hybrid Artificial Intelligence Framework for Predicting Academic Success Among International Students

Preeti Nashier [1], Purushottam Kumar [2], Prakash Kuppuswamy [3*], Ajay Chakravarty [4], Santhi Muttipoll Dharmarajlu [5], Phaneendra Kanakamedala [6], Saroj Kumar Gupta[7], and Vijaya Ramineni [8]

[1]Research Scholar, Kalinga University, Chhattisgarh, India; kptvtv@gmail.com
[2]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, Uttar Pradesh, India; dr.purushottam.kmr@gmail.com
[3]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India; prakashcnet@gmail.com
[4]College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India; ajay.chakravarty1@gmail.com
[5]Department of Nursing, College of Nursing and Health Sciences, Jazan University, Jazan, Saudi Arabia; srajalu@jazanu.edu.sa
[6]Department of Computer Science and Information Technology, Lakireddy Bali Reddy College of Engineering, Andhra Pradesh, India; phanikanakamedala@gmail.com
[7]Computer Science Engineering (AI & ML) Department, Parul University, Vadodara, Gujarat, India; sarojkumargupta@gmail.com
[8]Department of Information Technology, DVR & Dr. HS MIC College of Technology, Vijayawada, Andhra Pradesh, India; vijayaramineni5@gmail.com

* Correspondence: kptvtv@gmail.com, prakashcnet@gmail.com

ABSTRACT

Universities increasingly use artificial intelligence to identify students at risk of not completing their studies, but most predictive models are opaque and are rarely checked for international students. This study developed an explainable hybrid framework that combines XGBoost, LightGBM, and CatBoost through a logistic regression meta-learner and explains its predictions with SHapley Additive exPlanations (SHAP). Using a public dataset of 4,424 undergraduates at a Portuguese polytechnic, including 110 international students, we compared the hybrid model with seven tuned single learners under 10-fold cross-validation and a held-out test set, tested predictions at three points in the first year, and audited performance and explanations by international status. The hybrid model predicted dropout, continued enrolment, and graduation with 77% accuracy (macro-F1 = .70), and was not significantly better than strong single learners or logistic regression. Prediction at enrolment was weak, and first-semester results raised macro-F1 by .12. Explanations were consistent across methods and identified approved course units and tuition-

fee status as the main drivers of risk. Accuracy for international students was similar to that for domestic students (.82 vs. .78), but fee arrears and debt weighed more heavily in their predictions, and nationality influenced their risk estimates. Institutions should prefer simple, validated models, rerun predictions after each semester, audit results for international students, exclude nationality, and pair alerts with financial support.

Keywords: academic success, early warning, explainable artificial intelligence, international students, learning analytics, secondary data analysis, SHAP, student dropout

© Author(s), 2026. Published by Star Scholars Press.
This article is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. https://creativecommons.org/licenses/by/4.0/

INTRODUCTION

Universities increasingly use institutional data to identify students who may not complete their programs and to offer support before problems become irreversible (Romero & Ventura, 2013). Machine learning (ML) has become the main tool for this work, and systematic reviews report steady growth in models that predict grades, dropout, and graduation from demographic, socioeconomic, and academic records (Ab Rahman et al., 2025; Albreiki et al., 2021). For international students, the stakes of such systems are distinctive. They often face financial, linguistic, and social pressures that domestic peers do not, their persistence patterns can differ from institutional averages (Fass-Holmes, 2016), and non-academic support can matter as much as academic help for their progress (Gilbert et al., 2026). Universities are also beginning to use artificial intelligence (AI) to recruit and serve international students, and students' trust in these systems depends on transparency and on perceived fairness (Al-Dmour et al., 2025).

Two problems limit the use of predictive models in advising. First, the most accurate models, such as gradient-boosted trees and ensembles, are difficult to interpret, and advisers cannot act responsibly on a risk score they cannot explain (Barredo Arrieta et al., 2020; Rudin, 2019). Explainable artificial intelligence (XAI) methods, especially SHapley Additive exPlanations (SHAP), address this by attributing each prediction to the input variables (Lundberg & Lee, 2017). Second, models are usually evaluated on the whole student population. Accuracy for a small subgroup can differ from overall accuracy without anyone noticing, and international students are often exactly such a subgroup (Baker & Hawn, 2022; Gardner et al., 2019). A model that performs well on average may therefore still be unreliable for the students whom international offices most need to reach.

Recent studies report very high accuracy for hybrid and deep models, often above 93% (Chavez et al., 2023; Paul et al., 2025). These figures are rarely accompanied by a comparison with strong single learners under the same validation protocol, by an analysis of how early in a program the prediction is available, or by a check of performance for international students. This study addresses these gaps. We develop an explainable hybrid framework that combines three gradient-boosting learners through a logistic regression meta-learner, explain its predictions with SHAP, and audit its performance and explanations for international and domestic students in a public higher education dataset. The study asks four research questions (RQs):

- RQ1: How accurately do single learners and a hybrid stacking ensemble predict dropout, continued enrolment, and graduation, and does the hybrid outperform the best single learner?
- RQ2: How does predictive performance change as academic information accumulates from enrolment to the end of the second semester?

• RQ3: Which factors drive the predictions, and are the global explanations consistent across explanation methods?

• RQ4: Do predictive performance and the main explanatory factors differ between international and domestic students?

The study contributes a fully reproducible and transparent evaluation protocol, an honest estimate of what a hybrid ensemble adds over simpler models, and a subgroup audit that international student offices can repeat on their own data before adopting an early-warning system.

LITERATURE REVIEW

Machine Learning for Predicting Academic Success and Dropout

Educational data mining has long framed student attrition as a prediction problem. Early case studies predicted dropout from pre-university and first-year records with decision trees and related classifiers (Dekker et al., 2009; Márquez-Vera et al., 2013), and later work showed that course-level performance data allow earlier and more accurate identification of failing students (Asif et al., 2017; Costa et al., 2017). This line of work is consistent with Tinto's (1975) view that persistence reflects a student's academic and social integration over time, which is why records from the first semesters carry so much predictive information.

Recent studies have moved toward more complex models. Jiao et al. (2022) combined learning-process and summative assessment data through evolutionary computation and found knowledge acquisition and class participation to be the strongest predictors. Haron et al. (2025) compared RepTree, k-nearest neighbors, and naïve Bayes for an academic monitoring system and found RepTree most accurate. Chavez et al. (2023) trained a neural network on about 32,000 Open University records without personal attributes and reported 93.81% accuracy for a pass–fail outcome. Paul et al. (2025) reported 98.43% accuracy for identifying at-risk students among 33,627 Indian college students. Reviews nevertheless caution that reported accuracies depend heavily on the outcome definition, class balance, and validation design, and that overfitting and weak generalization are common (Ab Rahman et al., 2025; Albreiki et al., 2021).

The dataset used in the present study illustrates this dependence. It was published by Realinho et al. (2022) to support a three-class prediction of dropout, continued enrolment, and graduation. Bettahi et al. (2025) replicated their institutional pipeline on this dataset and obtained an F1-score of 0.792 for the easier binary dropout task, well below the 0.911 they reached on their own institutional data. Three-class prediction is harder still, mainly because students who are still enrolled resemble both graduates and dropouts. Comparisons of "hybrid" models are informative only when every model is tuned and validated in the same way.

Explainability, Trust, and Responsible Use

Explainability is a precondition for using predictions in advising. Barredo Arrieta et al. (2020) and Belle and Papantonis (2021) distinguish models that are interpretable by design from post hoc explanations of complex models. Rudin (2019) argued that high-stakes decisions should rely on interpretable models wherever these are competitive, because post hoc explanations can be incomplete. SHAP offers a principled post hoc approach: it decomposes each prediction into additive feature contributions based on Shapley values (Lundberg & Lee, 2017), and TreeSHAP computes these contributions exactly for tree ensembles (Lundberg et al., 2020). Educational studies increasingly use SHAP to identify risk and protective factors (Bettahi et al., 2025; Paul et al., 2025), but they rarely test whether the explanations are stable across explanation methods or across student groups.

Explainability also bears on trust. In a study of AI tools in international student recruitment, Al-Dmour et al. (2025) found that concerns about privacy and algorithmic bias reduced prospective students' trust and willingness to engage. Nelson (2026) showed that structured AI workflows in an English-medium

university improved performance while raising questions of equity and academic integrity. Both studies suggest that AI systems for international students need to be transparent and to be checked for unequal effects.

International Students, Subgroup Validity, and Fairness

Research on international student success shows that institutional averages can mislead. Fass-Holmes (2016) found that international undergraduates at a U.S. public university persisted and graduated at higher rates than expected, and that most attrition reflected reasons other than academic failure. Gilbert et al. (2026) found that international postgraduate researchers depended on housing, community, pastoral care, and paid work to progress, factors that institutional records capture poorly. In learning analytics, Gardner et al. (2019) proposed slicing analysis, which evaluates a model separately for each student group, and Baker and Hawn (2022) reviewed evidence that predictive models can be less accurate for groups that are small or under-represented in training data. International students are typically such a group. Yet studies that predict academic success seldom report results for international students separately, and the present study responds to this gap. The review leads directly to the four RQs: whether a hybrid ensemble improves prediction (RQ1), when predictions become useful (RQ2), what drives them (RQ3), and whether they hold for international students (RQ4).

RESEARCH METHOD

Research Design

The study used a quantitative, predictive design based on secondary analysis of an existing, anonymized institutional dataset. A predictive design suits the purpose of the study, which is to estimate how well a student's later academic outcome can be forecast from information available at enrolment and during the first year, and to explain those forecasts. The study tests no causal hypotheses, and the explanatory results describe associations learned by the models.

Data Source and Participants

The data came from the publicly available dataset "Predict Students' Dropout and Academic Success" (Realinho et al., 2021), described in detail by Realinho et al. (2022). It contains records of 4,424 students enrolled in 17 undergraduate degree programs at the Polytechnic Institute of Portalegre, Portugal, between the 2008/2009 and 2018/2019 academic years. Each record includes 36 predictors and an outcome measured at the end of the normal program duration: dropout (n = 1,421; 32.1%), still enrolled (n = 794; 17.9%), or graduated (n = 2,209; 49.9%). The dataset has no missing values, and our integrity check found no duplicate records, so all 4,424 students were included.

We define an *international student* as a student whose nationality is not Portuguese, which is the definition encoded in the dataset's binary International variable; in these data the variable agrees exactly with the Nationality code. The sample includes 110 international students (2.5%) of 20 nationalities. Most came from Portuguese-speaking countries: 38 were Brazilian, 14 Santomean, 13 Cape Verdean, and 5 Guinean; 13 were Spanish, and 15 other nationalities had three or fewer students each. Among international students, 32 (29.1%) dropped out, 24 (21.8%) were still enrolled, and 54 (49.1%) graduated, compared with 32.2%, 17.8%, and 50.0% among domestic students.

Variables

The predictors fall into five groups: demographic (e.g., age at enrolment, gender, marital status, nationality, international status, displaced status), socioeconomic (parents' qualifications and occupations, scholarship holder, debtor, tuition fees up to date), pre-enrolment academic (application mode and order, previous qualification and its grade, admission grade, course, daytime or evening attendance), academic progress in the first and second semesters (curricular units credited, enrolled, evaluated, approved, and without evaluation, and the semester grade average), and macroeconomic conditions at enrolment

(unemployment rate, inflation rate, and gross domestic product growth). Nine nominal variables were one-hot encoded for the linear and kernel models, and continuous variables were standardized for these models. Tree-based models used the original codes, and CatBoost treated the nominal variables as categorical.

Proposed Framework

Figure 1 summarizes the seven-step framework. Each student *i* is described by a predictor vector $\mathbf{x}_i$ and an outcome $y_i \in \{0 = \text{dropout}, 1 = \text{enrolled}, 2 = \text{graduate}\}$. Three base learners, XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018), each produce a vector of class probabilities $\mathbf{p}_i^{(m)}$ for $m = 1, 2, 3$. To avoid leakage, these probabilities are generated out of fold by five-fold cross-validation within the training data. Following stacked generalization (Wolpert, 1992), the nine probabilities form the input $\mathbf{z}_i$ of a multinomial logistic regression meta-learner:

$$P(y_i = k \mid \mathbf{z}_i) = \frac{\exp(\mathbf{w}_k^{\top}\mathbf{z}_i + b_k)}{\sum_{l=0}^{2}\exp(\mathbf{w}_l^{\top}\mathbf{z}_i + b_l)} \tag{1}$$

The meta-learner parameters minimize the L2-regularized cross-entropy

$$\mathcal{L}(W, b) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{2}\mathbb{1}(y_i = k)\log P(y_i = k \mid \mathbf{z}_i) + \lambda \parallel W \parallel_2^2 \tag{2}$$

where $\mathbb{1}(\cdot)$ is the indicator function and λ controls the strength of regularization. The predicted outcome is the class with the highest probability.

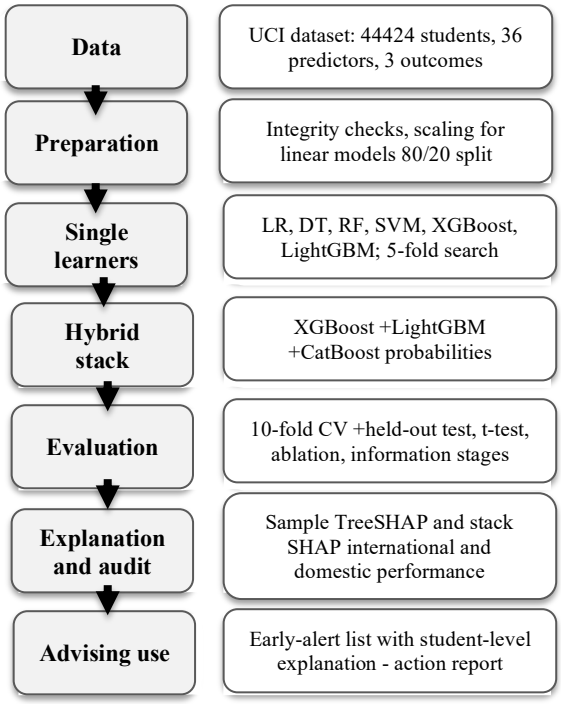
To explain a model *f*, SHAP assigns to each predictor *j* the Shapley value

$$\phi_j(f, \mathbf{x}) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(d - |S| - 1)!}{d!}\left[f_{S\cup\{j\}}(\mathbf{x}) - f_S(\mathbf{x})\right] \tag{3}$$

where *F* is the set of *d* predictors and f_S is the model's expected output when only the predictors in *S* are known. The contributions are additive, so that $f(\mathbf{x}) = \varphi_0 + \Sigma_j\, \varphi_j$, where φ_0 is the average prediction. We explain the predicted probability of dropout, the outcome of most concern for early warning.

Figure 1

Framework for Explainable Prediction of Academic Outcomes



Note. LR = logistic regression; DT = decision tree; RF = random forest; SVM = support vector machine; CV = cross-validation; CI = confidence interval; SHAP = SHapley Additive exPlanations.

Model Training and Evaluation

We compared the hybrid stack with seven single learners: multinomial logistic regression, decision tree, random forest (Breiman, 2001), support vector machine with a radial basis kernel (Cortes & Vapnik, 1995), XGBoost, LightGBM, and CatBoost. The data were split once into a stratified training set (80%; n = 3,539) and a held-out test set (20%; n = 885). Hyperparameters for every single learner were selected by grid search with stratified five-fold cross-validation on the training set, using the macro-averaged F1-score as the criterion because the three classes are unbalanced. The stack reused the tuned settings of its base learners.

Each model was then evaluated in two ways: by stratified 10-fold cross-validation within the training set, reported as mean and standard deviation, and once on the untouched test set. We report accuracy, macro-averaged precision, recall, and F1-score, and the macro-averaged one-versus-rest area under the receiver operating characteristic curve (AUC). For class k, precision is $TP_k / (TP_k + FP_k)$, recall is $TP_k / (TP_k + FN_k)$, and $F1_k$ is their harmonic mean; the macro-F1 is the unweighted mean of $F1_k$ over the three classes. Test-set 95% confidence intervals (CIs) were obtained from 2,000 bootstrap resamples. We compared the stack with each base learner using the corrected resampled t test over the 10 folds (Nadeau & Bengio, 2003) and an exact McNemar test on the test set.

An ablation study compared the full stack with three two-learner stacks and with a simple soft-voting average of the three base learners. To address RQ2, the stack was retrained and tested at three information stages: at enrolment (no curricular unit variables), after Semester 1 (first-semester variables added), and after Semester 2 (all variables).

Explanation and Subgroup Audit

Two explanation methods were used. First, the stack's predicted probability of dropout was explained on the test set with model-agnostic permutation SHAP, using 100 training students as the background distribution. Second, to obtain out-of-sample explanations for every student, including all 110 international students, CatBoost was refitted in each fold of a 10-fold cross-validation over the full sample and explained with exact TreeSHAP on the held-out fold. Agreement between the two explanations was measured by Spearman's rank correlation (ρ) of the mean absolute SHAP values. For RQ4, out-of-fold predictions from the same 10-fold procedure were compared between international and domestic students (a slicing analysis; Gardner et al., 2019), with bootstrap CIs and a permutation test of the difference in accuracy (5,000 permutations).

All analyses used Python 3.13 with scikit-learn 1.9.1 (Pedregosa et al., 2011), XGBoost 3.4.1, LightGBM 4.7.0, CatBoost 1.2.10, and SHAP 0.54.1, with a fixed random seed of 42.

Ethical Considerations

The study is a secondary analysis of a public, fully anonymized dataset released under a Creative Commons Attribution 4.0 license. The dataset contains no names or identifiers; it was compiled and anonymized by the data owners from institutional records (Realinho et al., 2022). The authors had no contact with students, and no attempt was made to identify individuals. Institutional review board approval was therefore not required. Because the models could influence how students are treated, we treat the predictions as decision support for advisers and report results by international status.

RESULTS

Predictive Performance of Single and Hybrid Models (RQ1)

Table 1 reports the performance of all models. In cross-validation, macro-F1 ranged from .668 for the decision tree to .717 for the hybrid stack. Six of the eight models, including logistic regression, lay between .708 and .717; only the decision tree and the random forest (.698) were lower. On the held-out test set, the stack reached an accuracy of .766 (95% CI [.738, .794]) and a macro-F1 of .697 [.663, .729]. It had the highest AUC, .893 [.874, .910], although the differences in AUC were small. Logistic regression reached an accuracy of .774 [.747, .801] and a macro-F1 of .703 [.670, .737]. The best single learner on the test set by macro-F1 was XGBoost (.718).

The stack did not outperform its base learners by a meaningful margin. In cross-validation, its macro-F1 exceeded that of CatBoost by .008 (corrected $t(9) = 0.86$, $p = .410$), that of LightGBM by .006 ($p = .378$), and that of XGBoost by .007 ($p = .326$). On the test set, the stack and CatBoost disagreed on 28 students. The stack was right on 12 of these and CatBoost on 16 (exact McNemar $p = .572$). The ablation (Table 1, Panel B) showed that every two-learner stack and the simple soft-voting average performed as well as the three-learner stack; the soft-voting average matched its cross-validated macro-F1 (difference < .001, $p = .971$) and needed one fifth of the training time.

All models found the *enrolled* class hardest. On the test set, the stack recalled .754 of dropouts and .900 of graduates but only .415 of students who were still enrolled; it classified most of the others as graduates or dropouts.

Table 1

Predictive Performance of Single Learners, the Hybrid Stack, and Ablated Ensembles

Model	CV accuracy	CV macro-F1	Accuracy	Precision	Recall	Macro-F1	AUC
Panel A: Single learners and the hybrid stack							
Logistic regression	.781 (.016)	.708 (.023)	.774	.723	.694	.703	.882
Decision tree	.748 (.019)	.668 (.029)	.747	.685	.668	.673	.855
Random forest	.776 (.015)	.698 (.021)	.775	.733	.691	.704	.887
Support vector machine	.784 (.016)	.712 (.025)	.765	.710	.685	.693	.881
XGBoost	.780 (.022)	.710 (.031)	.776	.729	.710	.718	.889
LightGBM	.781 (.021)	.711 (.031)	.765	.709	.699	.703	.883
CatBoost	.781 (.017)	.708 (.031)	.771	.710	.682	.690	.892
Hybrid stack (all three learners)	.786 (.024)	.717 (.037)	.766	.709	.690	.697	.893
Panel B: Ablation of the hybrid stack							
Stack: XGBoost + LightGBM	.783 (.019)	.709 (.029)	.774	.723	.705	.712	.886
Stack: XGBoost + CatBoost	.786 (.024)	.717 (.037)	.768	.713	.694	.701	.893
Stack: LightGBM + CatBoost	.787 (.021)	.717 (.036)	.768	.713	.693	.700	.892
Soft vote (all three learners)	.787 (.021)	.717 (.032)	.774	.725	.704	.712	.892

Note. CV columns show the mean (*SD*) over stratified 10-fold cross-validation within the training set (n = 3,539). The remaining columns refer to the held-out test set (n = 885). Precision, recall, F1, and the area under the receiver operating characteristic curve (AUC, one-versus-rest) are macro-averaged over dropout, enrolled, and graduate. For the hybrid stack, 95% bootstrap CIs on the test set were [.738, .794] for accuracy, [.663, .729] for macro-F1, and [.874, .910] for AUC. Stacks combine the listed gradient-boosting learners (XGBoost, LightGBM, CatBoost) through a logistic regression meta-learner; the soft vote averages their probabilities.

Prediction at Three Information Stages (RQ2)

Table 2, Panel A shows how performance grew as students progressed. At enrolment, using demographic, socioeconomic, and pre-enrolment variables only, the stack reached an accuracy of .633, a macro-F1 of .530, and an AUC of .767, and it identified .627 of the students who later dropped out. Adding

first-semester results raised macro-F1 to .650 and dropout recall to .725. Second-semester results added a smaller gain, to a macro-F1 of .697 and dropout recall of .754.

International and Domestic Students (RQ4)

Table 2, Panel B reports out-of-fold performance for the 110 international and 4,314 domestic students. The stack's accuracy was .818 (95% CI [.745, .882]) for international students and .783 [.771, .795] for domestic students. The difference of 3.6 percentage points was not statistically significant (permutation p = .415). The model identified 26 of the 32 international students who dropped out (.813), compared with .774 of domestic students who dropped out. The confidence intervals for international students are about six times wider than those for domestic students because the group is small. Removing the Nationality and International variables did not change overall performance (CatBoost out-of-fold accuracy .781 with and .782 without them); for international students, accuracy was .818 with and .809 without them, a difference of one student.

Table 2

Performance by Information Stage and by International Status

Stage or group	n	Accuracy	Macro-F1	AUC	Dropout recall	Enrolled recall
Panel A: Information stage (hybrid stack, held-out test set, n = 885)						
At enrolment (24 predictors)	885	.633	.530	.767	.627	.170
After Semester 1 (30 predictors)	885	.732	.650	.872	.725	.327
After Semester 2 (36 predictors)	885	.766	.697	.893	.754	.415
Panel B: Student group (hybrid stack, out-of-fold predictions, all 36 predictors)						
International students	110	.818 [.745, .882]	.805 [.721, .876]	.914	.813	—
Domestic students	4,314	.783 [.771, .795]	.714 [.699, .729]	.897	.774	—
All students	4,424	.783 [.771, .796]	.717 [.702, .732]	.898	.775	—

Note. Panel A: at enrolment, all curricular unit variables were excluded; after Semester 1, only first-semester curricular unit variables were added. Panel B: predictions for each student come from the fold in which that student was held out (10-fold cross-validation over all 4,424 students); 95% bootstrap CIs in brackets. The accuracy difference between international and domestic students was 3.6 percentage points (permutation p = .415). Enrolled recall is not reported by group because only 24 international students were still enrolled. AUC = area under the receiver operating characteristic curve (macro-averaged, one-versus-rest).

Explanatory Factors (RQ3 and RQ4)

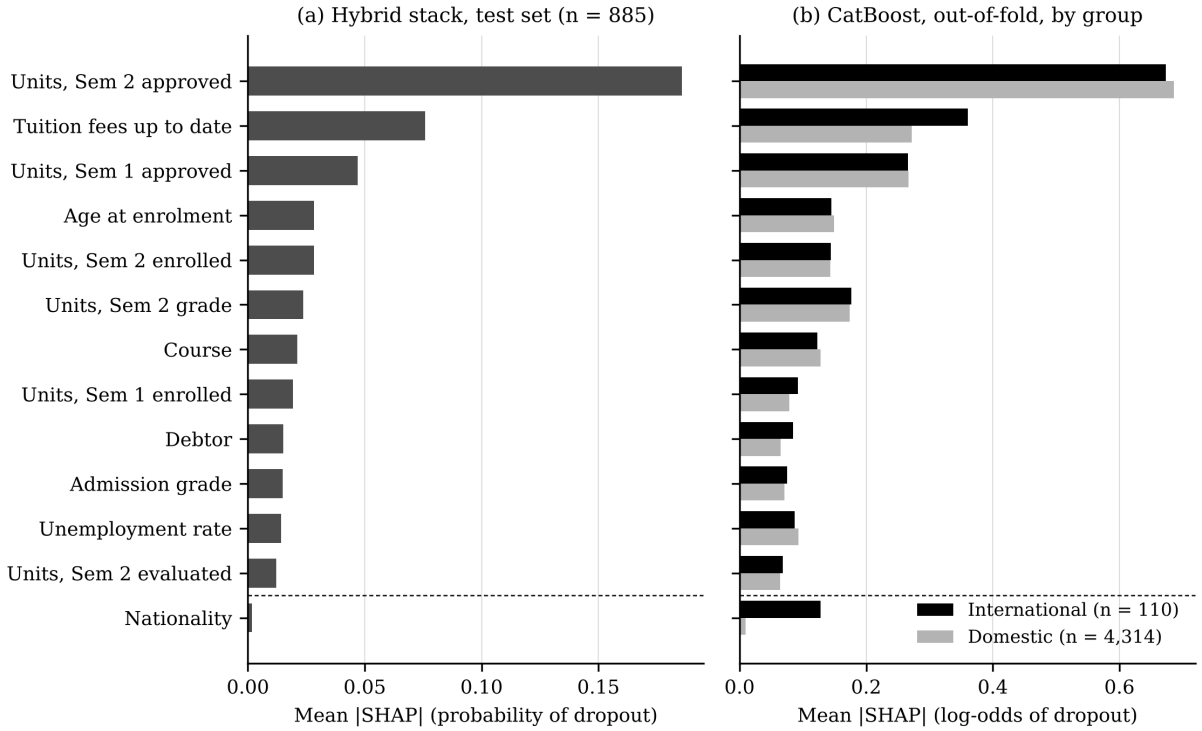
Figure 2a shows the main drivers of the stack's predicted probability of dropout on the test set. The number of curricular units approved in Semester 2 dominated: on average it shifted a student's predicted probability of dropout by 18.6 percentage points, more than twice as much as the next factor, whether tuition fees were up to date (7.6 points). Units approved in Semester 1 (4.7 points), age at enrolment, units enrolled and the grade average in Semester 2, and course followed. More approved units, fees paid on time, and holding a scholarship lowered the predicted risk, whereas older age at enrolment raised it. Macroeconomic variables played a minor role. The SHAP values summed exactly to each prediction (maximum error < 10^{-14}). The two explanation methods agreed closely: the rank correlation between the stack's permutation SHAP and CatBoost's out-of-sample TreeSHAP importance across all 36 predictors was ρ = .94, and eight of their top-10 predictors were the same.

Figure 2b compares the drivers for international and domestic students. The overall rankings were similar (ρ = .91), and academic progress dominated in both groups. Three differences stand out. First, whether tuition fees were up to date contributed about a third more to the predictions for international students (mean |SHAP| = 0.36, 95% CI [0.29, 0.44]) than for domestic students (0.27; permutation p = .005). Second, being a debtor also contributed more for international students (0.084 vs. 0.065; p = .023). These differences reflect the students' circumstances: 20.9% of international students had tuition fees in arrears, compared with 11.7% of domestic students (Fisher's exact p = .007), and 26.4% were debtors, compared with 11.0% (p < .001). Fewer international students held scholarships (17.3% vs. 25.0%, p =

.073). Third, nationality contributed almost nothing to predictions for domestic students (mean |SHAP| ≈ 0.010) but about 0.13 for international students, comparable to the contribution of course. These subgroup comparisons are exploratory and were not adjusted for multiple testing.

Figure 2

Main Drivers of the Predicted Probability of Dropout



Note. Bars show mean absolute SHAP values. Panel (a) explains the hybrid stack's probability of dropout on the held-out test set with permutation SHAP. Panel (b) explains CatBoost's log-odds of dropout with out-of-sample TreeSHAP for all 4,424 students, separately by group. The 12 predictors shown above the dashed line are the most important in Panel (a); nationality is added below it for comparison. Units = curricular units; Sem = semester.

Local explanations show how the framework can support an adviser. For one international student in the test set, the stack predicted a dropout probability of .94, against a baseline of .34. The explanation attributed most of the increase to having no curricular units approved in Semester 2 (+.44) or Semester 1 (+.07) and a Semester 2 grade average of zero (+.05). Age 30 at enrolment added a little (+.02), while having tuition fees up to date lowered the risk slightly (−.02). The student did drop out. An adviser who saw this explanation at the end of Semester 1 would know that the risk came from academic disengagement rather than from finances.

DISCUSSION

This study set out to build an explainable hybrid framework for predicting academic outcomes and to check whether it works as well for international students as for domestic students. Four findings stand out.

First, the hybrid stack predicted outcomes as well as any model tested, but not better than strong single learners or a simple average of them (RQ1). Its test accuracy of about 77% and macro-F1 of about .70 are consistent with the difficulty of the three-class task on this dataset and with the binary F1 of .792 reported by Bettahi et al. (2025), and they are far below the 93–98% often reported for other datasets and

tasks (Chavez et al., 2023; Paul et al., 2025). These high figures usually come from binary outcomes, different populations, or less strict validation, and our results show that tuned single models under the same protocol close most of the gap that "hybrid" architectures are said to fill. The finding supports the cautions of Albreiki et al. (2021) and Ab Rahman et al. (2025) about overfitting and reporting practices, and it echoes Rudin's (2019) argument: when a regularized logistic regression is within one percentage point of the most complex model, as it was here, institutions have good reason to prefer the simpler, more transparent model or a simple average of a few learners.

Second, the timing of a prediction matters more than the choice of algorithm (RQ2). Predictions made at enrolment identified only about 63% of eventual dropouts and 17% of the students who would still be enrolled, whereas first-semester results raised the macro-F1 by .12. This agrees with earlier work showing that course-level performance is the strongest signal of risk (Asif et al., 2017; Costa et al., 2017; Jiao et al., 2022), and with Tinto's (1975) view of persistence as a process of academic integration. For international offices, the practical implication is that an early-warning system should be run again after each assessment period rather than once at entry, and that predictions at entry are too weak to justify any action beyond a general welcome and support offer.

Third, the explanations were stable and plausible (RQ3). Two different SHAP methods applied to two different models produced nearly the same ranking of predictors, which increases confidence that the explanations reflect the data rather than the quirks of one model (Lundberg et al., 2020). Academic progress dominated, and the next most important factor was financial: whether tuition fees were paid on time. All five predictors that Realinho et al. (2022) found important across four algorithms appear among our top seven.

Fourth, the model performed at least as well for international students as for domestic students, but it relied on partly different information (RQ4). The accuracy for international students was slightly higher, although the difference was not significant and the estimate is uncertain because only 110 international students were available. This pattern mirrors Fass-Holmes (2016), who found that international undergraduates persisted better than expected. The subgroup explanations add something that overall accuracy cannot show: financial status weighed more heavily in predictions for international students, who were nearly twice as likely as domestic students to be behind with fees and more than twice as likely to be debtors. This is consistent with Gilbert et al. (2026), who found that non-academic support, including help with paid work and living costs, was central to international researchers' progress. It suggests that the first response to an alert for an international student should often be a conversation about finances, emergency funds, or payment plans rather than academic remediation alone.

The nationality finding calls for caution. For international students, the model's risk estimates depended partly on their country of origin, a pattern learned from very few students per country. Such estimates are unreliable and could lead to treating students differently because of where they come from, which is the kind of algorithmic bias that Baker and Hawn (2022) warned about and that undermines the trust international students place in institutional AI (Al-Dmour et al., 2025). Because removing nationality did not reduce accuracy, we recommend leaving nationality out of operational models and monitoring performance by international status instead.

Implications for Practice and Policy

The results suggest four practical steps for universities that serve international students. First, institutions should validate any vendor or in-house model locally, compare it against a regularized logistic regression, and prefer the simpler model when the difference is small. Second, they should rerun predictions after each semester and route alerts to advisers rather than acting automatically. Third, they should report model performance separately for international students before deployment, as a slicing analysis (Gardner et al., 2019), and repeat this each year. Fourth, they should pair academic alerts with financial support pathways for international students, whose fee and debt status weighs heavily in their risk. At the policy

level, institutional AI policies should require subgroup audits and exclude protected or proxy attributes such as nationality unless their use is justified and reviewed.

Limitations and Future Research

The study has several limitations. The data come from one Portuguese polytechnic over ten academic years, so the results may not generalize to other countries, to research universities, or to recent cohorts. International students made up only 2.5% of the sample and were mostly from Portuguese-speaking countries, so the subgroup estimates are imprecise and may not apply to international students who study in a language other than their own. The dataset does not record language proficiency, visa status, housing, employment, or sense of belonging, factors that matter for international students (Gilbert et al., 2026). The model uses semester results, so its most accurate predictions arrive after academic difficulty has already appeared. Finally, SHAP values describe associations learned by the model, not causal effects, and no intervention was tested. Future research should replicate the audit in institutions with larger international populations, add data on language, finances, and belonging, test whether explanation-based alerts improve advising decisions and retention in a controlled evaluation, and examine students' and advisers' trust in explained predictions.

CONCLUSION

This study developed and evaluated an explainable hybrid framework that predicts dropout, continued enrolment, and graduation from institutional records and explains each prediction with SHAP. Under a strict and reproducible protocol, the hybrid stack predicted outcomes with about 77% accuracy, as well as but not better than strong single learners, and predictions became useful only after the first semester. Explanations were consistent across methods and showed that academic progress and tuition-fee status drive risk. An audit by international status found no loss of accuracy for international students but showed that financial factors and nationality weigh differently in their predictions. The study contributes a transparent evaluation and audit protocol that institutions can apply before adopting AI early-warning systems. We recommend that universities pair such systems with subgroup audits, adviser review, and financial support pathways, so that AI helps international students rather than labelling them.

Author Contributions: Conceptualization, Preeti Nashier and Prakash Kuppuswamy; methodology, Preeti Nashier, Prakash Kuppuswamy, and Vijaya Ramineni; formal analysis, Preeti Nashier; data curation, Preeti Nashier and Vijaya Ramineni; validation, Purushottam Kumar, Ajay Chakravarty, Santhi Muttipoll Dharmarajlu, Phaneendra Kanakamedala, and Saroj Kumar Gupta; writing original draft preparation, Preeti Nashier and Prakash Kuppuswamy; writing review and editing, all authors; supervision, Prakash Kuppuswamy. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study is a secondary analysis of a publicly available, fully anonymized dataset (Realinho et al., 2021) and did not involve contact with human participants.

Acknowledgments: The authors sincerely thank their respective institutions for providing the necessary academic environment, computational resources, and technical support that facilitated the completion of this research.

AI Acknowledgment: No generative artificial intelligence (AI) tools were used in the preparation, analysis, writing, or revision of this manuscript. All research activities, analyses, interpretations, and conclusions were carried out by the authors, who take full responsibility for the content and integrity of the article.

Data Availability Statement: The data analyzed in this study are openly available in the UCI Machine Learning Repository at https://doi.org/10.24432/C5MC89 under a Creative Commons Attribution 4.0 license. The analysis code and full numerical results are available from the corresponding author upon reasonable request.

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

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 the publisher and/or the editor(s). The publisher 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.

REFERENCES

Ab Rahman, N. F., Wang, S. L., Ng, T. F., & Ghoneim, A. S. (2025). Artificial intelligence in education: A systematic review of machine learning for predicting student performance. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 54*(1), 198–221. https://doi.org/10.37934/araset.54.1.198221

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences, 11*(9), Article 552. https://doi.org/10.3390/educsci11090552

Al-Dmour, H., Al-Dmour, R., Al-Dmour, Y., & Al-Dmour, A. (2025). Transforming international student recruitment: The role of AI, personalization, and trust in Jordanian higher education. *Journal of International Students, 15*(8), 25–52. https://doi.org/10.32674/m2fmc286

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data, 4*, Article 688969. https://doi.org/10.3389/fdata.2021.688969

Bettahi, A., Belouadha, F.-Z., & Harroud, H. (2025). A modular and explainable machine learning pipeline for student dropout prediction in higher education. *Algorithms, 18*(10), Article 662. https://doi.org/10.3390/a18100662

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chavez, H., Chavez-Arias, B., Contreras-Rosas, S., Alvarez-Rodríguez, J. M., & Raymundo, C. (2023). Artificial neural network model to predict student performance using nonpersonal information. *Frontiers in Education, 8*, Article 1106679. https://doi.org/10.3389/feduc.2023.1106679

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior, 73*, 247–256. https://doi.org/10.1016/j.chb.2017.01.047

Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Proceedings of the 2nd International Conference on Educational Data Mining* (pp. 41–50). International Working Group on Educational Data Mining.

Fass-Holmes, B. (2016). International undergraduates' retention, graduation, and time to degree. *Journal of International Students, 6*(4), 933–955. https://doi.org/10.32674/jis.v6i4.327

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 225–234). Association for Computing Machinery. https://doi.org/10.1145/3303772.3303791

Gilbert, D. J., Poliah, K., Helm, H., Milburn, M., & Yates, K. (2026). Non-academic support helps international postgraduate researchers succeed in their studies. *Journal of International Students, 16*(15), 55–80. https://doi.org/10.32674/51qbn076

Haron, N. H., Mahmood, R., Md Amin, N., Ahmad, A., & Jantan, S. R. (2025). An artificial intelligence approach to monitor and predict student academic performance. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 44*(1), 105–119. https://doi.org/10.37934/araset.44.1.105119

Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. *Artificial Intelligence Review, 55*(8), 6321–6344. https://doi.org/10.1007/s10462-022-10155-y

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 3146–3154). Curran Associates.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates.

Márquez-Vera, C., Romero, C., & Ventura, S. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje, 8*(1), 7–14. https://doi.org/10.1109/RITA.2013.2244695

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning, 52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

Nelson, E. C. W. (2026). Structured AI workflows in Saudi EMI research writing: Performance gains, equity, and academic integrity across three semesters. *Journal of International Students, 16*(15), 81–102. https://doi.org/10.32674/e7jdhc67

Paul, R., Sarker, S., El Aouifi, H., Hussain, S., Baruah, A. K., & Gaftandzhieva, S. (2025). Analyzing dropout of students and an explainable prediction of academic performance utilizing artificial intelligence techniques. *Frontiers in Education, 10*, Article 1698505. https://doi.org/10.3389/feduc.2025.1698505

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In S. Bengio et al. (Eds.), *Advances in Neural Information Processing Systems 31* (pp. 6638–6648). Curran Associates.

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data, 7*(11), Article 146. https://doi.org/10.3390/data7110146

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). *Predict students' dropout and academic success* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12–27. https://doi.org/10.1002/widm.1075

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125. https://doi.org/10.3102/00346543045001089

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Author Bios

Preeti Nashier is a doctoral student at Kalinga University, Chhattisgarh, India. Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Data Science, Internet of Things, and Intelligent Computing Systems. She is particularly interested in exploring AI-driven solutions for real-world challenges and developing innovative computational approaches. Email: kptvtv@gmail.com.

ORCID: https://orcid.org/0009-0006-5715-4555

Purushottam Kumar, PhD, is a faculty member in the School of Computer Science, Engineering and Technology, Bennett University, Greater Noida, India. His research interests include Artificial Intelligence, Machine Learning, Cloud Computing, Big Data Analytics, Cybersecurity, and Distributed Computing. His academic interests focus on developing intelligent computational frameworks and scalable technology-driven solutions. Email: dr.purushottam.kmr@gmail.com.

ORCID: https://orcid.org/0000-0001-6575-8232

Prakash Kuppuswamy, PhD, is a faculty member in the Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India. His research interests include Cybersecurity, Network Security, Post-Quantum Cryptography, Blockchain Technology, Artificial

Intelligence, and Internet of Things. His research focuses on developing secure, intelligent, and resilient computational frameworks for emerging technological applications. Email: prakashcnet@gmail.com.

ORCID: https://orcid.org/0000-0001-9804-8835

Ajay Chakravarty, PhD, is a faculty member in the College of Computing Sciences and Information Technology, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh, India. His research interests include Machine Learning, Artificial Intelligence, Data Mining, Cloud Computing, Software Engineering, and Natural Language Processing. His academic work focuses on intelligent algorithms, computational optimization, and innovative approaches to emerging information technology challenges. Email: ajay.chakravarty1@gmail.com.

ORCID: https://orcid.org/0000-0001-9454-9739

Santhi Muttipoll Dharmarajlu, PhD, is affiliated with the Department of Nursing, College of Nursing and Health Sciences, Jazan University, Jazan, Saudi Arabia. Her research interests include Nursing Education, Clinical Nursing, Community Health, Patient Safety, Healthcare Quality Management, and Evidence-Based Nursing Practice. Her academic interests focus on improving healthcare delivery, patient outcomes, and innovative approaches to nursing education and practice. Email: srajalu@jazanu.edu.sa.

ORCID: https://orcid.org/0000-0003-0281-1731

Phaneendra Kanakamedala is a faculty member in the Department of Computer Science and Information Technology, Lakireddy Bali Reddy College of Engineering, India. His research interests include Artificial Intelligence, Machine Learning, Deep Learning, Data Analytics, Computer Networks, and Internet of Things. His research focuses on developing intelligent algorithms and efficient computational models for practical engineering and technological applications. Email: phanikanakamedala@gmail.com.

ORCID: https://orcid.org/0000-0002-7869-6058

Saroj Kumar Gupta, PhD, is affiliated with the Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Parul University, Vadodara, Gujarat, India. His research interests include Artificial Intelligence, Machine Learning, Deep Learning, Computer Vision, Predictive Analytics, and Intelligent Automation. His research focuses on developing advanced learning algorithms and data-driven computational solutions for emerging real-world applications. Email: sarojkumargupta@gmail.com

ORCID: https://orcid.org/0000-0003-1304-8858

Vijaya Ramineni is a faculty member in the Department of Information Technology, DVR and Dr. HS MIC College of Technology, Vijayawada, India. Her research interests include Artificial Intelligence, Machine Learning, Data Mining, Internet of Things, Cloud Computing, and Network Security. Her academic interests focus on intelligent information systems, emerging computing technologies, and innovative solutions to contemporary technological challenges. Email: vijayaramineni5@gmail.com.

ORCID: https://orcid.org/0009-0009-8460-0233