This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

A Platform for Early Class Dropout Prediction of University Students

GUILHERME ANTONIO BORGES[1,2], (Student Member, IEEE), CAROLINA PEDRO[1], JULIO C. S. DOS ANJOS[3,6], ANDRÉ RODRIGUES[1,4], FERNANDO BOAVIDA[5], (Senior Member, IEEE), and JORGE SÁ SILVA[1], (Senior Member, IEEE)

[1]Institute for Systems Engineering and Computers (INESC), Coimbra, University of Coimbra, 3030-790 Portugal
[2]Sul-rio-grandense Federal Institute (IFSUL), Charqueadas, 96745-000 Brazil
[3]Federal University of Ceará Campus Itapajé, Itapaje, 62600-000, Brazil
[4]Polytechnic Institute of Coimbra, Coimbra Business School, Coimbra, 3045-093 Portugal
[5]Center of Informatics and Systems of the University of Coimbra (CISUC), Coimbra, 3030-790 Portugal
[6]Graduate Program in Teleinformatics Engineering (PPGETI/UFC), Campus of Pici, Fortaleza, 60455-970, Brazil

Corresponding author: Guilherme Antonio Borges (e-mail: guilhermeborges@ifsul.edu.br).

"This work was supported in part by the Rectorate of the University of Coimbra under Grant PRR POCH-02-53I2-FSE-000032 and CAPES (Finance Code 001)."

ABSTRACT Higher education faces significant challenges, with student dropout rates remaining a critical issue. While much research has focused on course-level dropout, less attention has been given to the early identification of at-risk students within individual courses. In this context, this study reviews the state-of-the-art and applies Machine Learning (ML) techniques to predict early dropout of students from real-world data. Key demographic, academic, and behavioral factors were analyzed to identify students at risk of dropping subjects, enabling timely intervention and the implementation of preventive strategies. The analysis revealed that student performance, re-enrollment behavior, and attendance patterns are strongly correlated with the risk of early dropout. Models trained on mid-term and final attendance data before final exams showed the best metrics, highlighting the importance of tracking student engagement throughout the semester. Based on these findings, we developed and implemented the Best Assisting Tutor and INteractive Advisor (BATINA) OnBoard Platform, an advanced platform that integrates predictive models in a practical and user-friendly way. This platform provides comprehensive support to students and faculty, enhancing academic performance and enabling early intervention to improve student retention. This study demonstrates how data-driven solutions can enhance educational support systems and help mitigate student dropout.

INDEX TERMS Student Dropout, Machine Learning, Support Platform, Real World-Data, Early intervention, Higher education, Educational Data Mining.

I. INTRODUCTION

THE 21st century is often considered an era of technology [1]. Digital transformation has recently revolutionized society, becoming an integral part of everyday life. Similarly, technology has rapidly transformed learning and teaching processes [2] and is revolutionizing the way students learn.

Along with the evolution of technology, the amount of available and produced data has increased exponentially, including educational data [3] [4]. One main advantage of technology-based learning environments is their ability to easily acquire, gather, and provide detailed data [3]. Such data enable educators to make informed decisions and intervene more effectively, guiding students by adjusting their teaching strategies to meet their specific needs [5]. It also enables the monitoring of student engagement and performance, providing early warnings when students are at risk of dropping or failing a subject by using ML, recommendation systems, and other techniques [6], [7]. Student dropout evaluation can benefit from these approaches.

Dropout is a complex problem in the educational paradigm, affecting students worldwide across various levels, including higher education, high school, and Massive Open Online Courses (MOOC) [7]. Although student dropout comparisons between countries must account for differences in educational systems, a common conclusion emerges in most studies: a substantial percentage of students drop out of their courses before completing them.

To put it in perspective, according to National Center for Education Statistics data [8], between 2020 and 2021, around 24% of first-year full-time undergraduates dropped out in the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

United States of America (USA). Another study concerning the Porto University [9] in Portugal identified a student dropout rate of 32.1% in the first year. In the case of MOOCs, student dropout values can reach 94.37% [10].

A wealth of studies try to understand and generate models capable of predicting the students who dropout (e.g., [7], [10]). However, as it is a dynamic phenomenon influenced by local culture, differences in each educational system, and continuous evolution [11], no single method or system is effective for all scenarios. Nevertheless, to improve the techniques for predicting student dropout, using a real-world dataset is essential to understand and address the phenomenon, increase early detection, and facilitate timely intervention and the implementation of preventive strategies.

Given the above, this study has two goals. First, it evaluates strategies for early detection of student dropout using state-of-the-art ML algorithms and the anonymized real-world data set from University of Coimbra (UC) obtained through the Rectorate of the University of Coimbra project called On-Board. Second, it proposes the BATINA OnBoard Platform to support the teachers with dropout information and their classes. Thus, we organize this study in four steps, as depicted in Fig. 1.

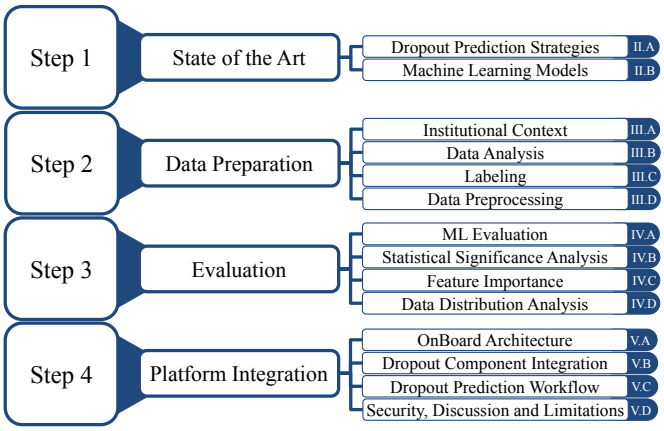


FIGURE 1. Study organization.

Section II pertains to the first step of this study, which reviews state-of-the-art dropout prediction strategies and the baseline ML models used in our evaluations. The second step is conducted in Section III. It presents the UC dataset used in this study, describing its institutional characteristics, data analysis, data preprocessing, and labeling strategies.

Based on the labeled data, the third step (Section IV) addresses data evaluation. The labeled data is evaluated using metrics by several ML models, considering the resulting metrics, statistical analysis, data distribution, and the feature importance technique. The fourth step (Section V) presents the BATINA OnBoard Platform, comprising an overview of its architecture, the integration of an early student dropout prediction component based on the incorporation of the best-performing ML model from our evaluations, the workflow of dropout prediction, and a discussion and limitations of security, ethical and technical. Finally, Section VI summarises the paper's findings and provides guidelines for future research.

II. STATE-OF-THE-ART

A. DROPOUT PREDICTION STRATEGIES

The increasing availability of digital learning data has positioned Educational Data Mining (EDM) as a powerful approach for understanding and optimizing student learning behaviors and outcomes. In this field, predictive modeling has been extensively applied to two related problems: predicting student performance and predicting dropout. Performance prediction models aim to estimate academic outcomes, such as grades [12] or exam success [13], often using classification or regression approaches. In contrast, dropout prediction focuses on identifying students at risk of dropping out of a course before a milestone (e.g., a final exam or semester completion), typically framed as a binary classification problem. While both models rely on supervised learning techniques, dropout prediction poses unique challenges, including early identification, data sparsity, and the need to incorporate behavioral and engagement indicators. These differences require specific modeling strategies and features suited for proactive educational intervention rather than academic outcome forecasting.

Predicting early dropout involves inferring the students' probability of discontinuing their studies. The sooner it is identified that students may be struggling, the greater the chance that interventions designed to protect them from gradually falling behind and eventually interrupting their studies will be effective [14], [15]. To understand how existing solutions achieve this early prediction, we organized Table 1 by considering the following items: reference, study target audience, dropout prediction strategy, and the best-performing ML models. The latter are discussed in subsection II-B.

The dropout prediction strategy determines when an ML model is used to predict early dropout in academic progress. Each work can use one or more strategies classified as admission-based, first-year/semester, course milestone, and class milestone.

The first strategy is to predict dropout, considering the data presented during the admission on the first enrollment ([16], [17], and [18]), which is generally comprised of socioeconomic, demographic and high school grades, administrative information from the enrolled course, and other information. Although these studies demonstrate that such information impacts dropout probability, it alone is insufficient to develop highly accurate predictive models because it disregards the students' experience in the new course.

To illustrate this strategy, the first study in the area with this approach [19] focused on evaluating student dropout rates in distance learning at the Hellenic Open University in Greece. They demonstrated that ML algorithms can empower teachers with the information needed to identify dropout-prone students from the beginning of the school year, using only demographic data from the students. However, the accuracy of the developed model increases as the school year progresses

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

and new curricular data are incorporated. An accuracy of 63% was achieved at the start of the academic term, based solely on demographic data, and exceeded 83% by the middle of the academic term.

As the dropout rate is higher in the first year, the studies presented in [14], [16], [18], [20]–[24], generate ML models to determine the dropout probability considering the data from the first semester and/or the first year. This strategy also enables the identification of factors that can influence student dropout, which can be used to inform the university's intervention or the formulation of strategies to help students stay in the course. The drawback of this strategy is that it needs to consider dropouts that may occur in future semesters.

To address dropout in both early and late semesters, studies [4], [20], [21], and [15] utilize course milestones to predict student dropout. [4] generate models that predict students' dropout in the next semester based on the current semester data. It includes demographic data and general academic data, divided by semester, which features include those based on the students' social network. Similarly, [15] created a model that considers the milestones at the end of each year. They concluded that some characteristics may help predict dropout earlier, while others may become relevant as students progress through their courses.

The study [20] uses data from students' first year to determine two dropout probabilities: First Year Dropout, which gives the dropout probability at the semester right after the end of the first year, and Late Year Dropout, which offers the dropout probability of students from the third semester to the ninth one. As they use the same data from the first year, the performance of the models used by the authors decreases for the late dropout, indicating that there are additional factors influencing academic progress that also impact dropout.

Although the course milestone strategy offers more possibilities for evaluating student dropout during a course, it relies on completing a whole semester or year to obtain data for predicting student dropout. Because of this, the studies [15], [17], [19], [25]–[29], and [14], use the strategy of class milestones. The strategy can be divided into three approaches: fixed structure, dynamic structure, and engagement-based.

The studies [19], [25], and [27] build a model using a fixed class structure. It enables the generation of ML models that can predict a student's dropout rate using evaluations and other events within the structure of classes on a given subject. Although this strategy allows for early intervention before the last class, it limits the professor's freedom to plan and apply different teaching strategies. It also reduces the evaluation flexibility because a fixed assessment structure is required. For example, [19] employs four evaluations, which include two face-to-face meetings and two written assignments, while [27] relies on two exams.

To add flexibility for predicting class dropout, the studies [27] and [15] use a dynamic structure model. This approach combines information that is fixed or takes a prolonged time to change, such as administrative, academic, demographic, and other pre-admission data, with partial class data aggregated by the student's participation. [15] uses aggregated information like Moodle activity count and Moodle activity trend incremented in the middle of the semester. [27] aggregates the Virtual Learning Environment (VLE) information from class interactions at weeks 4 and 8 and before final exams A and B.

TABLE 1. Dropout strategy works for dropout prediction

Article	Target	Prediction Strategy	ML Model
[19]	Higher Education	Class Milestone	NB
[16]	Higher Education	First Year/Semester, Admission Data	DT
[4]	Higher Education	Course Milestone	Rule learner
[28]	MOOC	Class Milestone	LSTM
[17]	Higher Education	Admission Data	GBT
[29]	VLE	Class Milestone	LSTM
[30]	MOOC	Class Milestone	GRU
[25]	Higher Education	Class Milestone	LR
[20]	Higher Education	First Year/Semester, Course Milestone	GLM and RF
[21]	Higher Education	First Year/Semester, Course Milestone	XGBoost
[22]	Higher Education	First Year/Semester	XGBoost, RF
[14]	Higher Education	First Year/Semester, Course Milestone	RF
[26]	MOOC	Class Milestone	CNNAE-LSTM
[18]	Higher Education	Admission Data, First Year/Semester	LR
[15]	Higher Education	Class Milestone, Course Milestone	CAT, NN, and LR
[23]	Higher Education	First Year/Semester	XGBoost
[27]	Higher Education	Class Milestone	NN
[24]	Higher Education	First Year/Semester	RF

Lastly, [15], [17], [26]–[29], and [14] create a model based on the engagement approach by applying the click-stream technique. It is often used on MOOC courses [26], [28], or courses that use a VLE [14], [17], [27], [29], by using metrics like the number of logins at the platform itself, submitted assignments, clicks on links at the forum, and so on. It enables the aggregation of such metrics using a window of N days to predict dropout in the following week or within a defined time frame. Although effective in predicting early class dropouts, it requires a substantial volume of online interaction with students, which may not always be available in traditional university courses.

Although class milestone strategies demonstrate that it is possible to predict early dropout reasonably accurately, there is a need to improve this prediction in traditional higher education courses. With this in mind, we implemented and evaluated a dynamic class strategy to facilitate early intervention, thereby enhancing the possibility of taking actions before the semester ends through an Online Platform.

B. MACHINE LEARNING AND DROPOUT

The dropout prediction problem is typically modeled using ML algorithms as a two-class/binary classification problem. Table 1 summarizes the 'best-performing models' on each paper in column 'Best ML Model' based on what we understand that the authors consider the best ML algorithms through their metrics. These algorithms include Naive Bayes (NB), Deci-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

sion Tree (DT), Logistic Regression (LR), Artificial Neural Network (ANN), Long Short-Term Memory (LSTM), Rule Learning, Gated Recurrent Units (GRU), Generalized Linear Model (GLM), CatBoost (CAT), eXtreme Gradient Boosting (XGBoost), CNNAE-LSTM, Gradient Boosted Trees (GBT), and Random Forest (RF). From all of them, we highlight NB, DT, RF, and XGBoost in our study. We also included K-Nearest Neighbor (KNN) as a baseline model in our evaluations. To provide context for each algorithm's typical performance, we also include examples from recent studies. These examples are not intended as direct benchmarks, as differences in datasets, features, and institutional settings prevent strict comparability. However, they help contextualize our results within existing literature.

Although the table shows that NB, DT, and KNN have not presented the best results in recent studies, they are often used as baseline algorithms to evaluate most datasets. The Naïve Bayes (NB) classifier is based on Bayes's Theorem, known for its simplicity, effectiveness in many classification tasks, fast training, and high scalability [19]. It assumes that attributes are independent of the class, meaning that an attribute's presence or absence does not affect the presence or absence of another attribute within the same class. However, studies that compare the algorithm with others for the same purpose indicate that it can be overcome even in cases where there are substantial dependencies between attributes [6]. To exemplify the performance of NB on recent studies [18] reached an accuracy of 0.88, a recall of 0.3369, and an AUC ROC of 0.6909 using a traditional plain modeling, and [26] reached an accuracy of 0.8107, a recall of 0.9548, and an AUC ROC of 0.8113 on a click stream scenario.

KNN is one of the simplest and most commonly used similarity-based algorithms. It stores all training data and classifies a new data point according to the class, considering the majority of votes of its k nearest neighbors in the given dataset [6], [15]. It can get satisfactory results when several local approximations of low complexity may describe the dataset [31]. An example of recent performance obtained from previous studies is [18], which achieved an accuracy of 0.8081, a recall of 0.3422, and an AUC-ROC of 0.632. Additionally, [15] reported a recall of 0.45 and 0.777 in a click stream scenario.

DT is a tree-based ML algorithm. It organizes the knowledge extracted from data in a recursive hierarchical structure composed of nodes, leaf nodes, and branches. Each internal node represents an attribute associated with a test relevant to data classification. The leaf nodes of the tree correspond to classes. Branches represent each possible result based on the input feature value, allowing the nodes to navigate until they reach a leaf node (classification). One of the biggest drawbacks is overfitting, which is solved by introducing RF [6]. Two examples of results found in recent studies are [20] and [18]. [20] in a first-year approach reached an accuracy of 0.9118, a recall of 0.9004, and an AUC ROC of 0.9473. [18] reached an accuracy of 0.8345, a recall of 0.2406, and an AUC ROC of 0.562 in a first-semester prediction.

The Random Forest (RF) algorithm is a robust and accurate ensemble method comprising several independent DT s, making it less susceptible to overfitting. During model training, each tree is constructed from different subsets of data and features. It ensures that each tree makes distinct decisions, contributing to a diverse final model. Once all trees have made independent predictions, a voting mechanism determines the final class based on the majority vote [7]. In recent studies, [24] reached an accuracy of 0.898, a recall of 0.471, and an AUC ROC of 0.916 on a first-year prediction, and [26] reached an accuracy of 0.836, a recall of 0.932, and an AUC ROC of 0.832 in a click stream scenario.

XGBoost is a ML algorithm based on the boosting method, and uses weak learners as its base classifier. It optimizes the learning process using gradient-boosting techniques on DT. It enables the automatic handling of missing values, efficiently computes large and complex datasets, and is resilient to overfitting due to the integration of regularization methods [23], [32]. [25] reached an accuracy of 0.916, a recall of 0.79, and an AUC ROC of 0.978 on a traditional plain modeling mapped by enrollments. [27] a recall of 0.8, and an AUC ROC of 0.96 on a scenario with class milestones.

The algorithms based on ANN, such as LSTM, GRU, CNNAE-LSTM, and combinations among them, are found in the studies [21], [25]–[30] and [15] with mixed results. In general, traditional ML classifiers perform comparably or better than ANN models in analyzing educational datasets [15], [33]–[35]. Nevertheless, a ANN model typically cannot directly determine a student's profile at risk of dropping out [36], making it harder to explain the predictions to educational stakeholders and practically implement the recommendations.

However, in certain scenarios—particularly those involving click-stream engagement data—ANN models have demonstrated superior performance compared to traditional ML approaches. As shown in Table 1, most of these studies utilize time-series or event-log data collected from learning platforms, which align well with the sequential and high-dimensional nature of neural network architectures. For example, the study [26] achieves an accuracy of 0.876, a recall of 0.9652, and an AUC-ROC of 0.8773 in this kind of scenario using a CNN-LSTM model. This structural compatibility allows models such as LSTM and CNN-LSTM to effectively capture temporal patterns in student behavior [26], [29]. For these reasons, we chose not to include ANNs in our evaluation.

III. DATASET ANALYSIS

A. INSTITUTIONAL CONTEXT AND DATA SELECTION

This work uses a dataset from UC. It is a Portuguese public and autonomous higher education institution created in 1290. It comprises ten academic and research units, along with centers that support cultural and educational initiatives, such as the Common Library, the University Stadium, and the Botanical Garden. Furthermore, it is a well-recognized

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

university that attracts international students from around the world.

The initial database consists of 22 confidential .xlsx files with anonymized student information, covering academic years from 1957/1958 to 2022/2023. However, as Portugal signed the Bologna Agreement in 1999, we selected the data starting from 2000/2001 for our evaluations.

The Bologna Agreement introduced the European Credit Transfer and Accumulation System (ECTS) to register academic credits, which provides a common framework for measuring educational outcomes. It also makes comparing grading systems within and across European countries easier and allows for mobility programs among the signatory university countries.

Records associated with academic years during the COVID-19 pandemic were removed due to that period's exceptional and atypical conditions. The pandemic impacted life in general and, in particular, introduced substantial changes to the educational landscape. In those years, the learning context was completely different from the one experienced until then, with significant changes, including the abrupt shift to online teaching, mobility restrictions, and alterations in teaching methodologies [37]. These factors created variables that could distort comparative analysis. Records related to particular subjects, such as the Master's Dissertation, were also excluded.

B. DATASET ANALYSIS AND FEATURE SELECTION

After analyzing the dataset, relevant information was selected, and data transformation procedures were performed, as detailed in this section. The 35 columns selected as features obtained from the analysis performed in this section are described in Appendix .

Information about students' parents, employment status, and other demographic details was collected from the questionnaires registered in the dataset as part of the university's pre-registration process. This was combined with personal data from the student, including their birth year and any disabilities.

Using each student's birth year and academic year, we calculated their age and created the variable AGE, normalizing the information. Additionally, as the students' professional situation and profession could change with each enrollment, both were updated annually in our datasets.

Information on students entitled to special status was included, with a new variable, "SPECIAL_SITUATION", indicating whether a student had a special situation in a given academic year, which might consist of information like financial difficulties, being a finalist, or serving as an academic association leader. This information was then added to the dataset using the academic year information, which means it could change annually.

The tutoring file contains information about students who received tutoring, including the date they started. From this file, we collected the student ID and the start date of the tutoring. This information was added to our dataset only in the following academic year, as recorded. TUTORING is binary: 0 indicates no tutoring received, and "1" signifies that tutoring was provided.

Students enroll in many subjects at the beginning of each semester, which we use as landmarks to create new variables. The variable RATIO_GRADES was created to reflect the overall performance of students in their subjects/classes. The ultimate goal was to include an indicator of students' course performance at the time of each record in the main dataset. To achieve this, the teaching periods for each course, as recorded in the enrollment file, were compared with the timestamps of each record. The MODE_ENROLLMENT variable indicates whether the student was enrolled in the subject full-time. The NUMBER_ENROLLMENTS variable was created to map the number of subjects that students selected at enrollment.

The mean enrollment records provide information on students' average course grades and the completed ECTS credits. These records allowed us to associate this data with the appropriate academic year and semester (OCCURRENCE_REGIME). The data was added to all student records after these calculations were performed. If multiple records exist for a student, only the most recent record for each case is retained. REG_DESIGNATION maps the purpose of enrollment, which includes "General", re-entry, transfer of institution, transfer of course, and many other minor designations. As most registers are labeled as "General", we aggregated all the others under an "Other" value.

Four variables were mapped for each subject at the time of enrollment in the lecture year. 1) whether it is part of the course study plan (SUBJEC_TYPE). 2) whether the student is enrolled in the subject to improve the grade (IMPROVEMENT). 3) whether or not the student has already been enrolled in the subject (RE-ENROLLMENT). 4) ENROLL_TYPE, which informs if the enrollment in this subject was through normal methods, mobility program, or isolated subjects. As there are a few isolated subjects, we grouped them with a mobility program, creating a group called "Other".

Finally, the attendance file contains detailed records of student attendance. Each entry represents a student's attendance at a specific class in a given month. To determine which subject each class corresponds to, the final output, organized by attendance, consolidates student attendance information by subject and academic year. However, it is essential to standardize the information and account for the relationship between student attendance and the total number of scheduled classes, rather than simply the number of classes attended. The total number of classes for each subject was selected to achieve this.

By revisiting the original attendance file, it was possible to identify the specific classes with attendance records for each student enrolled in each subject for each academic year. The total number of courses with recorded absences was then calculated.

Additionally, by summing up the recorded attendances and absences, it was possible to determine the total number of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

classes held overall and on a monthly basis. These discrepancies may arise from various factors, such as scheduled classes that were not held, students switching classes mid-semester (resulting in records for multiple classes), or first-year students who were not initially placed in the course. Due to unclear reasons for these inconsistencies, such cases were excluded from the analysis.

At this stage, we had the attendance, absences, and justified absences for each month, as well as the number of classes held with attendance records for each month. This data allowed us to calculate different student attendance rates at various times.

Two indices were computed for each student record to analyze the dropout forecast throughout the academic period and achieve one of the pre-established goals. One index pertains to the midterm of the class period, and the other relates to the end of the period, just before the exams. Since the midterm of the class period varies in each case, we estimated the expected number of classes at each midpoint. We calculated the student's attendance rate or justified absence proportionally.

We then created a new column, "RATIO_ATTENDANCE", to represent the ratio of classes attended or justified by the student relative to the total number of classes scheduled at midterm and the end of the period. It is essential to note that attendance only counts for the lecture period. After that, the final exam/evaluation period occurs, during which attendance is not taken into account.

C. LABELING

The labeling strategy is crucial for developing a model that accurately predicts student dropout. This section defines the class milestone strategy based on the RATIO_ATTENDANCE variable. However, the solution must address three issues before labeling: 1) not all subjects have attendance data. 2) there is no standardized quantity of classes per subject. 3) How to define a dropout from a subject.

For some subjects, attendance is recorded only for specific components, such as practical sessions, theoretical classes, or laboratory work, while other components may not have any attendance records at all. In other subjects, attendance records may be maintained for all classes. Additionally, there are subjects where no attendance data is recorded at all. For this reason, although labeling using class milestones cannot be modeled, we decided to include them in a separate dataset for baseline evaluation. We named it Dataset 1, in Table 2.

As class attendance is not standardized for each subject, we modeled the RATIO_ATTENDANCE variable using a ratio approach (ranging from 0 to 1). From a total number of attendances X before the final exam, we defined three milestones for prediction: 1) the beginning of the subject classes (0 of X), which the attendance is not considered; 2) the attendance midpoint ($X/2$); and 3) at the end of the classes before the exam, which considers the whole quantity of attendances from a student considering the total X of scheduled classes. These correspond to three key points in the column milestone on Table 2: the beginning, the middle, and the end, which allows for evaluating the data with a temporal perspective.

From a ML perspective, the problem is framed as a binary classification task, labeling students as either dropouts or non-dropouts. The "0" label in the data denotes students who did not drop out, while "1" denotes students who did. The goal is to develop models that accurately identify students with the label "1". However, defining the label for dropout is challenging due to its varied meanings across contexts [38]. It is relatively straightforward for course-level dropout: a student either completes the course and receives a diploma or does not [16]. However, dropout definitions for specific subjects lack clarity in the literature, necessitating a tailored approach for this study.

Importantly, we empirically examined how well subject-level dropout labels align with actual program-level dropout (i.e., "real" program dropout). A Pearson's Chi-squared test [39] revealed a statistically significant association between these variables, $\chi^2(1) = 166{,}778.18$, p-value < 0.001. Additionally, Fisher's Exact Test [39] was reported to support the strength of association via the odds ratio of 4.82 (p-value < 0.001), indicating that students classified as program dropouts were approximately 4.8 times more likely to exhibit subject-level dropout behaviors, such as missing final assessments. These results provide strong evidence of alignment between course-level and program-level dropout.

While course-level dropout is not a perfect proxy for program dropout, our findings suggest that subject-level disengagement is a strong early indicator of broader risk of dropout. We explicitly acknowledge that these labels are approximations and encourage future work to integrate additional behavioral and longitudinal signals to enhance the fidelity of dropout detection.

Based on course-level dropout information, this study models the labels in subjects in two labels (Table 2): **Approach A** and **B**. The **Approach A** label considers a student to have dropped out if the student fails to attend any scheduled exams or final assessments. The information on student absences from assessments was obtained through the class grades file. This was used to label the three milestones for prediction: the beginning, the middle, and the end, respectively, identified as Datasets 2, 3, and 4 in Table 2. This approach was also applied to Dataset 1, where the subjects' attendance was not recorded.

The **Approach B** is a subset of the Approach A label. It defines a student as having dropped out if the student is not a worker-student, was absent from more than 90% of the total classes, and missed any of the exams or final assessments. This was created to analyze the impact on the dropout rate of students who are fully dedicated to their studies and have low class attendance.

For the dataset of subjects without attendance records (Dataset 1), columns related to attendance were removed, as they were empty, resulting in 34 columns. It contains 1,612,651 records, with 249,090 labeled as "1" using Approach A, representing 15.45% of the total. These records pertain to 13,540 different subjects. Unfortunately, the labeling Approach B can not be applied to this dataset because it has no attendance records.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

TABLE 2. Identification from Datasets to evaluation.

Subject attendance	Milestone	Labeling	Dataset	Dropout
Not recorded	-	Approach A	Dataset 1	15.45%
Recorded	Beginning	Approach A	Dataset 2	12.95%
Recorded	Middle	Approach A	Dataset 3	12.95%
Recorded	End	Approach A	Dataset 4	12.95%
Recorded	Beginning	Approach B	Dataset 5	4.76%
Recorded	Middle	Approach B	Dataset 6	4.76%
Recorded	End	Approach B	Dataset 7	4.76%

One of the purposes of identifying the probability of dropout is to provide an additional support tool for professors in pedagogical planning during the middle of the semester. Therefore, it is necessary to include a temporal variable in the datasets. Therefore, the records in the attendance dataset were tripled, creating three distinct entries for each original entry. Each of these datasets, which includes subjects with attendance records, comprises 327,504 entries and 35 columns. By using the labeling Approach A (Datasets 2, 3, and 4), 42,407 (12.95%) are labeled as "1", whereas Approach B (Datasets 5, 6, and 7) identifies 15,598 (4.76%) student dropouts. These records encompass 3,647 distinct subjects/courses.

D. DATASET PREPROCESSING

The quality of the model largely depends on the quality of the input data [7]. Therefore, the preprocessing phase is indispensable in studies like this one. Before evaluating the ML models, we addressed issues such as missing values, feature engineering, variable encoding, normalization, and class imbalance. It is important to highlight that the columns excluded from our dataset are not included in the 35 selected attributes used in our evaluation.

Two types of missing values were identified: (1) meaningful absences and (2) data loss. For meaningful absences—such as missing values in STUDENT_ORGANIC_DISABILITY—we inferred a negative condition (e.g., "No") based on institutional context. The same was applied to missing values in STUDENT_DISABILITY_OTHER, STUDENT_DISPLACED, and STUDENT_INTERNATIONAL features. Data loss occurs when missing values result from incomplete entries, possibly due to registration issues. For general missing values, numerical features were imputed with the column mean, and categorical features were filled with the most frequent value. Columns with less than 70% completeness were excluded from analysis due to low data coverage.

Feature engineering played a key role in the model's performance. We derived new variables such as AGE (based on academic and birth years), RATIO_GRADES (proportion of passed classes), and RATIO_ATTENDANCE (percentage of attended or justified sessions based on class schedules). Additionally, features with high cardinality were grouped into aggregated categories (e.g., "Other") to avoid sparsity and improve generalization. For example, in the NATIONALITY variable, 90.6% of 1.612.651 records without attendance information were listed as Portugal, and the same percentage was observed in the attendance dataset with 327.504 entries. Thus, we created two groups: "Portugal" and "Others," consolidating less frequent countries into "Others" to reflect the dataset's dominance of Portugal. The same strategy was applied to the features:

- CLASS_TYPE,
- REG_DESIGNATION,
- FCOL_DESIGNATION,
- COUNTRY_USUAL_RESIDENCE,
- COUNTRY_HIGH_SCHOOL,
- MARITAL_STATUS, NATIONALITY,
- NATURALITY,
- STUDENT_ORGANIC_DISABILITY, and
- REGISTRATION_TYPE.

To support algorithms sensitive to feature scales and distance metrics (e.g., KNN and SVM), all numerical features in Appendix A (Table 9) were standardized to have a mean of zero and a variance of one. Categorical variables were one-hot encoded to produce a uniform numerical feature space across all models.

The dataset was also imbalanced, with dropout cases comprising only 4.76% to 15.45% of the data, depending on the dataset (see Table 2). To address this issue, we employed the Synthetic Minority Oversampling Technique (SMOTE) technique on the training datasets, a well-established and effective method [40], which has been used in similar studies [41]. This approach preserved the integrity of the test sets and improved model learning by enhancing the representation of class minorities [40].

IV. RESULTS

A. ML ALGORITHM EVALUATION

This section evaluates all datasets using the following ML algorithms: NB, DT, KNN, RF, and XGBoost. Firstly, all datasets were preprocessed as described in Section III-D. Each dataset was submitted to a grid hyperparameter optimization using the 5-fold cross-validation method. In our context, minimizing False Negative (FN) (missed identification of students who need intervention) is more important than reducing False Positive (FP). Therefore, recall was prioritized in the hyperparameter optimization, as it captures the proportion of true positives among actual positives, effectively measuring the model's ability to identify dropouts. All evaluated hyperparameters for each ML algorithm and the best configuration are presented in Appendix .

The metrics were generated by the 5-fold cross-validation method. To balance evaluation, we also used the AUC-ROC metric in our analysis, which reflects the model's ability to discriminate between the two classes. The SMOTE oversampling technique was applied to the minority class, all training sets (including those used on hyperparameter optimization). All experiments were performed in Python with the scikit-learn library.

Table 3 summarizes the results. Each dataset (DS 1–7) was tested for all five ML models. Recall and AUC-ROC values are shown with standard deviations, considering a 95% confidence interval, and the best values per metric are bolded. To

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

TABLE 3. Metrics obtained for each selected ML model.

Recall					
DS	*NB*	*KNN*	*DT*	*RF*	*XGBoost*
1	0.493 ± 0.026	0.626 ± 0.077	**0.712 ± 0.020**	0.594 ± 0.004	0.668 ± 0.008
2	0.424 ± 0.026	**0.665 ± 0.018**	0.638 ± 0.094	0.643 ± 0.008	0.646 ± 0.006
3	0.456 ± 0.035	0.699 ± 0.007	**0.825 ± 0.012**	0.724 ± 0.010	0.784 ± 0.007
4	0.457 ± 0.030	0.714 ± 0.006	**0.847 ± 0.002**	0.753 ± 0.015	**0.833 ± 0.004**
5	**0.958 ± 0.005**	0.747 ± 0.018	0.728 ± 0.071	0.776 ± 0.006	0.760 ± 0.009
6	0.959 ± 0.008	0.868 ± 0.008	**0.995 ± 0.001**	0.967 ± 0.006	**0.994 ± 0.003**
7	0.959 ± 0.009	0.872 ± 0.009	**0.992 ± 0.005**	0.970 ± 0.002	**0.996 ± 0.002**
AUC ROC					
DS	*NB*	*KNN*	*DT*	*RF*	*XGBoost*
1	0.618 ± 0.003	**0.687 ± 0.002**	0.634 ± 0.002	0.651 ± 0.002	0.659 ± 0.002
2	0.642 ± 0.008	**0.746 ± 0.005**	0.696 ± 0.049	0.728 ± 0.004	0.730 ± 0.003
3	0.660 ± 0.011	0.769 ± 0.003	0.755 ± 0.003	0.767 ± 0.003	**0.777 ± 0.003**
4	0.663 ± 0.010	0.778 ± 0.003	0.783 ± 0.002	0.783 ± 0.005	**0.794 ± 0.003**
5	0.626 ± 0.005	0.815 ± 0.005	0.767 ± 0.047	0.803 ± 0.002	**0.811 ± 0.002**
6	0.871 ± 0.002	0.899 ± 0.003	0.945 ± 0.001	0.934 ± 0.004	**0.947 ± 0.002**
7	0.899 ± 0.006	0.902 ± 0.003	0.951 ± 0.002	0.939 ± 0.005	**0.955 ± 0.001**

TABLE 4. Two-sided Mann–Whitney U test p-values comparing performance between consecutive datasets.

Dataset 1	Dataset 2	p (Recall)	p (AUC-ROC)
DS1	DS2	0.771019	**0.000023**
DS1	DS5	**0.000000**	**0.000039**
DS2	DS3	**0.000122**	**0.000025**
DS3	DS4	0.168326	**0.000211**
DS5	DS6	**0.000000**	**0.000000**
DS6	DS7	0.771013	0.080766

complement this analysis, we used two-sided Mann–Whitney U tests [42] to assess whether the performance distributions differed across dataset versions. Table 4 reports the results for both recall and AUC-ROC metrics, where the statistically significant values ($p < 0.05$) are bolded.

Dataset 1, which does not contain attendance data, represents yearly pre-enrollment scenarios. DT had the best recall (0.712 ± 0.020) and RF the best AUC-ROC (0.634 ± 0.002), showing that tree-based models excel in this early-stage setting.

Almost all the ML evaluations from Datasets with attendance (DS 2 to 7) outperform the metrics obtained in Dataset 1. The only two exceptions on DS 2 were the DT and XGBoost (XGB), whose recall scores were slightly worse, but the AUC-ROC scores were better. Although DS2 slightly improves over DS1 in recall, Table 4 shows that this difference is not statistically significant, indicating that when considering the beginning of the lecture year, there is no significant difference in the predicted values from DS1 and DS2 related to recall.

When the comparison between DS1 and DS5 is made, even though DS5 has a zero value for attendance in its features (i.e., students' attendance is zero at the beginning of the lecture year), Table 4 shows that the p-value among them is significant. It means that the selected data from DS5 enables the identification of patterns that correlate feature values with each output class more effectively than DS1.

For the midterm phase datasets (DS2, DS3, and DS4), we observe that both recall and AUC-ROC improve as more attendance data becomes available. The transition from DS2 to DS3 shows a statistically significant increase in performance, particularly evident in Table 4, where both recall ($p = 0.000122$) and AUC-ROC ($p = 0.000025$) indicate a strong shift. It suggests that partial attendance records (available in DS3) start to reveal reliable patterns that distinguish dropout-prone students from others. However, the difference between DS3 and DS4 is not significant for recall ($p = 0.168326$), although the AUC-ROC still improves significantly ($p = 0.000211$), implying a gain in discrimination capability, even if sensitivity remains stable.

For the end-of-term datasets (DS5, DS6, DS7), the comparison shows a consistent trend toward optimal performance. From DS5 to DS6, both recall and AUC-ROC improvements are statistically significant ($p < 0.0001$), demonstrating that extending attendance tracking into the mid-class period allows ML models to more confidently and accurately identify dropouts. Conversely, the transition from DS6 to DS7 does not show statistical significance in either metric. This result suggests that the predictive value of attendance data plateaus near the end of the term—models trained with DS6 data already capture nearly all distinguishable dropout-related features.

Analyzing the metric values from Table 3, we observe that Decision Trees (DT) often achieved the highest recall scores in earlier datasets (DS1 to DS4), which aligns with their ability to handle sparse and categorical data effectively. However, XGBoost consistently evidenced superior AUC-ROC scores, especially in DS3 through DS7, where attendance-based features became more discriminative. In the final datasets (DS6 and DS7), both DT and XGBoost attained recall values exceeding 0.99. At the same time, XGBoost achieved the highest AUC-ROC scores (0.947 and 0.955, respectively), confirming its ability to identify both dropout and non-dropout cases with high confidence.

In summary, the statistical significance analysis confirms that early-stage models (DS1–DS2) are limited in distinguishing dropouts due to insufficient feature resolution. The intro-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

duction of attendance data in DS3 and beyond results in clear and statistically significant improvements in both recall and AUC-ROC. While models continue to improve through DS6, further gains in DS7 are marginal, suggesting that midterm data suffices for reliable early intervention. XGBoost demonstrated the most balanced and consistently superior performance across all datasets, supporting its use as the default ML algorithm for dropout prediction in the OnBoard platform. These findings not only reinforce the value of incorporating attendance-based milestones in dropout prediction models but also guide the OnBoard platform toward deploying predictive analytics at points where the results are both reliable and actionable.

Finally, the observed superior performance of tree-based models, particularly DT and XGBoost, can be attributed to their compatibility with the structure of our dataset. Many of the most influential features, such as RE-ENROLLMENT, IMPROVEMENT, and categorical student status indicators, are binary or categorical. DTs handle these natively without requiring normalization or complex encoding schemes. Furthermore, DTs tend to prioritize high-information features and generate simple rules, which often leads to high recall scores, especially in imbalanced settings. However, this behavior may lead to overfitting and moderate AUC-ROC values. In contrast, XGBoost builds an ensemble of trees using gradient boosting, which incrementally corrects errors, incorporates regularization, and effectively handles missing or sparse values. These capabilities enable XGBoost to generalize more effectively across heterogeneous educational datasets. This explains why XGBoost consistently achieved top AUC-ROC scores while maintaining competitive recall, making it well-suited for dropout prediction tasks involving complex interactions and behavioral signals.

B. STATISTICAL SIGNIFICANCE ANALYSIS OF MODEL PERFORMANCE

To investigate whether the performance differences among ML models were statistically significant, we applied Wilcoxon signed-rank tests [43] to model scores across all datasets and evaluation folds. The resulting p-values were averaged pairwise and visualized in Fig. 2a–b. These global heatmaps provide an overview of statistical trends across the entire evaluation process. While some model pairs approach significance thresholds ($p < 0.05$), these aggregate values obscure important contextual differences tied to the dataset phases.

To provide a more nuanced view, we repeated the Wilcoxon tests by grouping datasets according to their attendance profile: DS1–DS2 (early stage), DS3–DS4 (midterm), and DS5–DS7 (end-of-term). The analysis of DS1–DS2 was conducted together because the previous section's analysis indicated that there were no significant differences between the results of the DS1 and DS2 groups. The grouped heatmaps in Fig. 3(c–h) demonstrate that statistical differences are minimal at the early stage, consistent with the limited information available before attendance data stabilizes. In contrast, the midterm and especially the end-of-term phases reveal stronger statistical separation, particularly in AUC-ROC, with XGBoost frequently showing significant superiority over other models. These results reinforce the importance of delaying dropout predictions until sufficient attendance data is available, both to improve model reliability and to enable more meaningful comparisons.

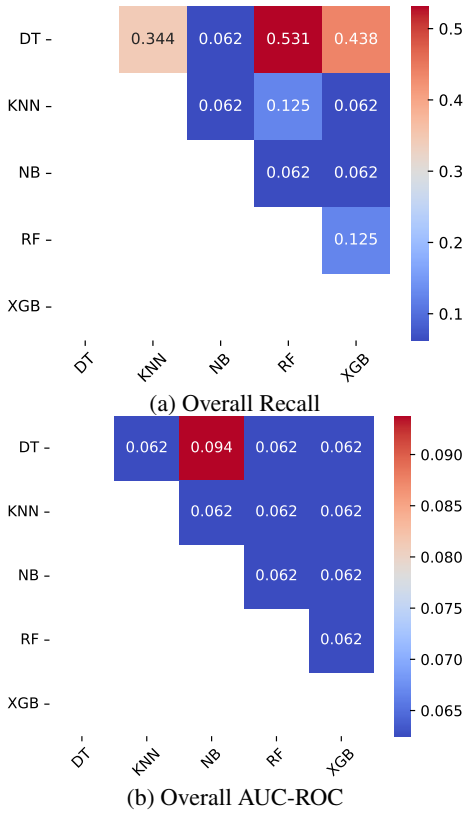


FIGURE 2. Two-sided Wilcoxon signed-rank test p-values comparing ML model performance across all datasets. Lower values (red) indicate statistically significant differences.

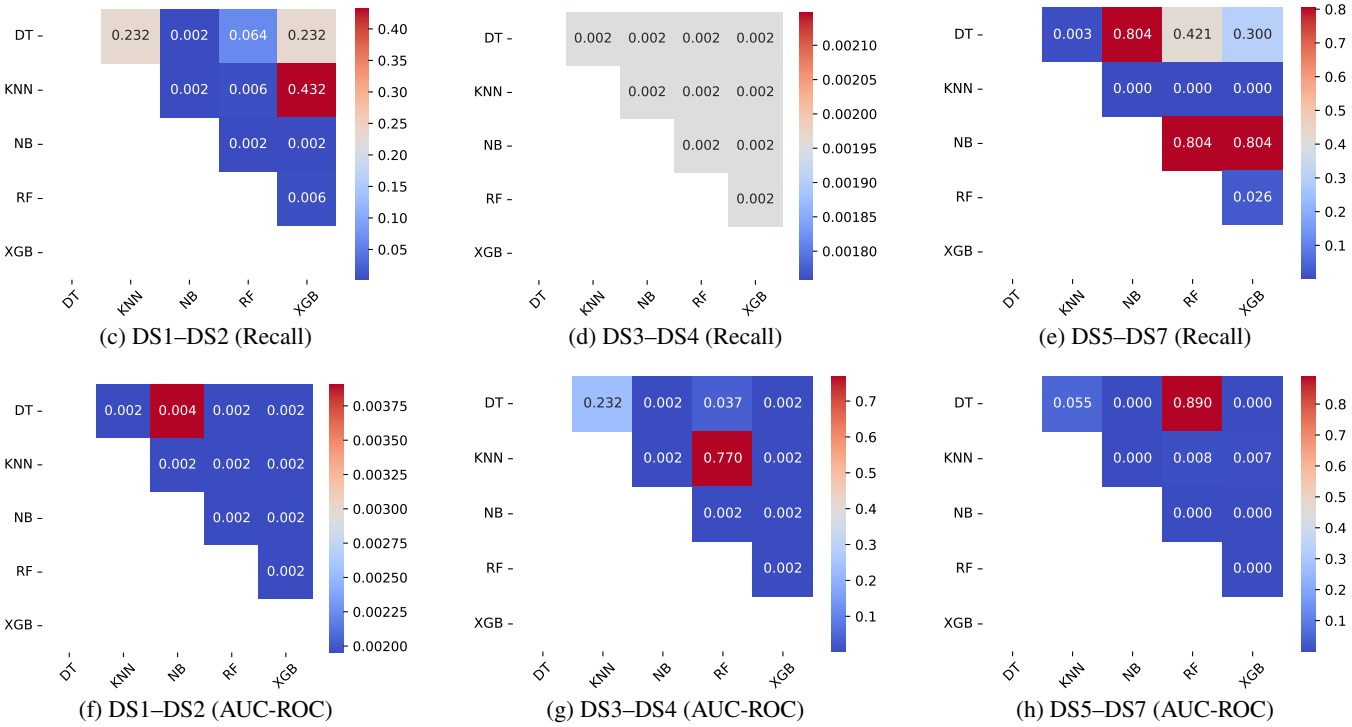
C. FEATURE IMPORTANCE

To better understand the impact of the features in each dataset, we evaluate the Feature Importance (FI) generated by the two best-performing tree-based ML algorithms: XGBoost and DT. The FI ranges from 0.0 to 1.0, where the sum of all importance is 1.0. Based on the analysis, we selected the five most influential features from XGBoost and compared their values with those of the DT features in Table 5.

We chose to analyze from this perspective because the DT algorithm focused the FI scores on fewer features, whereas the XGBoost algorithm tended to spread the FI scores across more features. This occurs because the DT algorithm simplifies the decision-making process by using a single tree model, which can lead to overfitting [23]. In contrast, the XGBoost optimizes the learning process using gradient-boosting techniques on decision trees, which allows it to find more complex relationships between features and the output while avoiding model overfitting [22], [23]. These characteristics

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751



(c) DS1–DS2 (Recall) (d) DS3–DS4 (Recall) (e) DS5–DS7 (Recall)

(f) DS1–DS2 (AUC-ROC) (g) DS3–DS4 (AUC-ROC) (h) DS5–DS7 (AUC-ROC)

FIGURE 3. Two-sided Wilcoxon signed-rank test p-values grouped by dataset phase: early (c, f), midterm (d, g), and end-of-term (e, h). Lower values (red) indicate statistically significant differences between models.

explain why the XBoost results obtained better AUC-ROC values than the DT in Table 3.

Firstly, the RE-ENROLLMENT variable, which indicates whether the student is enrolling in a subject for the first time or not, is the most occurring variable. It points out that the model interprets students in a subject as having a higher likelihood of dropping out of the same subject in the next re-enrollment. The IMPROVEMENT variable was a frequent occurrence in XGBoost results. It indicates that a student is re-enrolling to improve their performance, suggesting that those categorized as 'improvement' are more likely to drop out of a subject in which they were previously enrolled.

The RATIO_GRADES variable reflects the students' performance in assessments throughout their academic journey and was observed in Datasets 1, 3, and 4. It suggests that the student's performance is linked to their decision to drop out. However, the impact on each ML algorithm was different. In the DT case, it was the second most influential feature on Dataset 1, with a score of 0.348 compared to 0.051 from XGBoost. This emphasis on RATIO_GRADES and RE-ENROLLMENT resulted in a higher recall from the DT model, as shown in Table 3. Although it is not a highly accurate predictor, it highlights the importance of this feature when attendance is not taken into account. Conversely, the DT could not find a relationship between RATIO_GRADES and dropout on the remaining datasets.

On Datasets 3 and 4, although the XGBoost model identified a RATIO_GRADES relationship with student dropout, its values of FI decreased when considering class attendance.

TABLE 5. Feature importance from all datasets

Dataset	Feature	Imp. XGB	Imp. DT
Dataset 1	RE-ENROLLMENT	0.660	0.441
	GENDER	0.081	0.075
	RATIO_GRADE	0.051	0.348
	IMPROVEMENT	0.047	0.049
	EDUCATION_FATHER	0.035	-
Dataset 2	RE-ENROLLMENT	0.262	0.527
	GENDER	0.095	-
	SPECIAL_SITUATION	0.058	0.349
	OCCURRENCE_REGIME	0.054	-
	ENROLL_TYPE	0.049	-
Dataset 3	RE-ENROLLMENT	0.331	0.013
	RATIO_ATTENDANCE	0.210	0.856
	GENDER	0.089	-
	IMPROVEMENT	0.062	-
	RATIO_GRADES	0.050	-
Dataset 4	RE-ENROLLMENT	0.317	0.010
	RATIO_ATTENDANCE	0.247	0.867
	GENDER	0.085	-
	IMPROVEMENT	0.057	-
	RATIO_GRADES	0.045	-
Dataset 5	RE-ENROLLMENT	0.217	0.527
	SITUATION_PROF_STUDENT	0.186	0.349
	STUDENT_DISPLACED	0.065	-
	PROFESSION_FATHER	0.053	-
	PROFESSION_MOTHER	0.044	0.051
Dataset 6	RATIO_ATTENDANCE	0.312	0.856
	SITUATION_PROF_STUDENT	0.166	0.130
	RE-ENROLLMENT	0.160	0.013
	REG_DESIGNATION	0.069	-
	IMPROVEMENT	0.052	-
Dataset 7	RATIO_ATTENDANCE	0.431	0.867
	RE-ENROLLMENT	0.159	0.010
	SITUATION_PROF_STUDENT	0.148	0.123
	STUDENT_DISPLACED	0.063	-
	IMPROVEMENT	0.042	-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

The RATIO_ATTENDANCE variable represented students' class attendance on Datasets 3, 4, 6, and 7. It and the RE-ENROLLMENT variable were by far the most significant features in determining student dropout on a subject in both models.

The *GENDER* variable appeared in Datasets 1–4 with similar feature importance values (ranging from 0.075 to 0.095). An analysis of dropout rates by gender reveals that male students consistently exhibited a slightly higher likelihood of dropping out than female students. Specifically, the dropout rates for males in Datasets 1 through 4 were 0.062, 0.045, 0.045, and 0.044, respectively, while the corresponding rates for females were 0.018, 0.022, 0.043, and 0.040. Although these differences are modest, they suggest a consistent pattern of gender-related variation. This trend is consistent with the gender distribution observed during the analyzed period. In Dataset 1, 42.46% of students were male and 57.53% were female; in Datasets 2–4, the distribution was approximately 48.30% male and 51.69% female. However, this gender-related pattern does not persist in Datasets 5–7, which focus on full-time students with low attendance. In these datasets, there is no clear evidence of gender bias in dropout rates.

The SITUATION_PROF_STUDENT variable was present in Datasets 5, 6, and 7 with a high FI score. It maps not only the current profession of a student but also contains information about full-time students as a profession, which biases the models toward this target student profile. As XGBoost is more capable of identifying complex relationships from data, it distributes the FI score across more features than DT on datasets 5, 6, and 7: STUDENT_DISPLACED, REG_DESIGNATION, PROFESSION_MOTHER, and PROFESSION_FATHER.

The STUDENT_DISPLACED variable indicates that the student is not from the Coimbra district or a nearby region. It is present in the FI scores from XGBoost on Datasets 5 and 7. REG_DESIGNATION variable describes the type of entry process into the course (between General and other). It was present in Dataset 6 (midterm milestone), indicating that it is a factor influencing the decision to drop out.

PROFESSION_MOTHER and PROFESSION_FATHER showed up on XGBoost results from Dataset 5 (beginning from classes). The most significant pattern among all professions is "do not know," with FI of 0.023 for the mother and 0.020 for the father. To DT, all the feature importance was concentrated at "do not know" the mother profession with an FI total of 0.051. The OCCURRENCE_REGIME variable indicates whether the student was selected to enroll in the first phase or a subsequent phase. It was identified that this variable influences student dropout only in Dataset 2 FI from XGBoost. The SPECIAL_SITUATION variable on Dataset 2 greatly influenced both ML algorithms' FI results. Such a phenomenon requires independent future research to understand the root causes of its occurrence. However, when attendance is accounted for, this variable loses its FI place on both models.

D. DISTRIBUTION ANALYSIS ON IMPORTANT FEATURES

To better understand the behavioral signals underpinning the models, we analyzed the frequency distributions and dropout rates of key features across the three labeling strategies. Table 6 summarizes the distribution and dropout proportions for the RE-ENROLLMENT feature, which consistently emerged as one of the most important predictors in our models. The first column identifies the datasets analyzed, the second shows the categorical values of RE-ENROLLMENT, the third presents the value distribution across the full dataset, and the fourth reports dropout rates within each group. The final column highlights the relative dropout risk between re-enrolled and first-time enrollees.

TABLE 6. Distribution and dropout rate by RE-ENROLLMENT across labeling strategies

Dataset	Value	All data (%)	Dropout (%)	Rate
DS 1	No	84.63	12.40	—
	Yes	15.37	32.34	2.6x
DS 2-4	No	83.44	7.97	—
	Yes	16.56	37.59	4.7x
DS 5-7	No	83.43	2.26	—
	Yes	16.56	17.41	7.7x

Although re-enrolled students represent only 15.37% of the dataset in DS 1, they exhibited a dropout rate of 32.34%, more than twice the 12.4% rate of first-time enrollees. With attendance-informed labeling, the pattern becomes even more pronounced: under Approach A (DS 2-5), re-enrolled students were nearly five times more likely to drop out; under the stricter criteria of Approach B (DS 5–7), they were more than seven times as likely.

These findings confirm that re-enrollment is a robust and consistent indicator of early academic difficulty. This observation aligns with the Feature Importance scores reported across all datasets in the previous section. Furthermore, the consistency in re-enrollment prevalence (around 16%) across scenarios indicates that the predictive power of this feature is not a statistical artifact of imbalance, but a meaningful behavioral signal.

We also analyzed the distribution of the RATIO_ATTENDANCE variable across LABEL values for Label A Datasets. Figure 4 presents the boxplots for the two evaluation points: midterm and end of classes. Table 7 summarizes key descriptive statistics, including confidence intervals for the mean.

TABLE 7. Descriptive statistics for RATIO_ATTENDANCE by LABEL in Datasets 2–4 (Label A)

Milestone	LABEL	Mean	Median	Q1	Q3
End	0	0.6399	0.7750	0.3571	0.9259
End	1	0.1976	0.0455	0.0000	0.3333
Mid	0	0.6725	0.8125	0.4286	1.0000
Mid	1	0.2522	0.0714	0.0000	0.4825

As shown, RATIO_ATTENDANCE is highly discriminative between dropout and non-dropout cases. At both

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

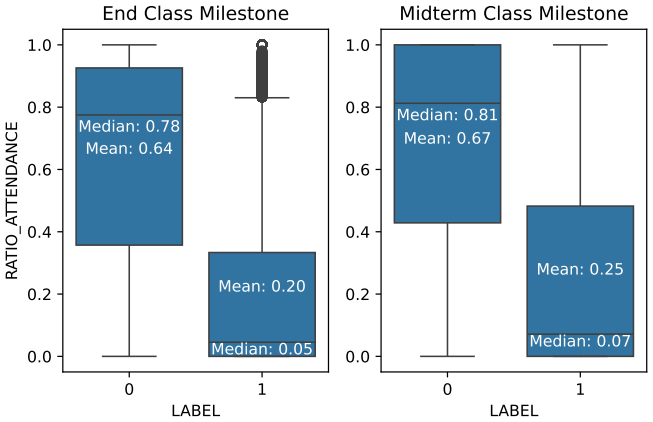


FIGURE 4. Boxplots of RATIO_ATTENDANCE by LABEL.

milestones, non-dropout students (LABEL = 0) demonstrate significantly higher mean and median attendance levels. These distributions suggest that dropout students tend to disengage early, with many exhibiting near-zero attendance throughout the semester. This underscores the value of RATIO_ATTENDANCE as a time-sensitive indicator, supporting the viability of attendance-based labeling for early intervention.

The distribution of RATIO_ATTENDANCE for DS 5–7 is shown in Table 8, where dropout students exhibited systematically minimal attendance, with a median of 0.0 and an interquartile range entirely below 0.20. Non-dropouts showed the expected full-range distribution centered near 0.75.

TABLE 8. Descriptive statistics for RATIO_ATTENDANCE by LABEL in Datasets 5–7 (Label B)

AULAS_DADAS	LABEL	Mean	Median	Q1	Q3
1.0 (End)	0	0.6104	0.7500	0.3000	0.9231
1.0 (End)	1	0.0123	0.0000	0.0000	0.0000
0.5 (Mid)	0	0.6480	0.7819	0.3615	0.9808
0.5 (Mid)	1	0.0200	0.0000	0.0000	0.1951

Due to this extreme skew and binary-like behavior among dropouts, visualizations such as boxplots were uninformative. Instead, we conducted a Mann–Whitney U test [42], which confirmed a statistically significant difference between attendance distributions in dropout and non-dropout groups ($U = 33{,}469{,}357{,}714.50$; $p < 0.00001$). It validates the RATIO_ATTENDANCE influence in the model trained under Label B.

Overall, Label B captures a clearly defined behavioral profile of dropout students — one marked by persistent disengagement. The continued significance of RATIO_ATTENDANCE, despite the imbalanced nature of the dropout group, underscores the feature's robustness and its lack of bias toward dominant values. Together, these two features — one categorical and one continuous — provide complementary and behaviorally grounded signals for early risk identification.

Lastly, the prominence of re-enrollment in courses/subjects reflects a common academic risk pattern: students who retake courses have previously failed or withdrawn, signaling underlying challenges in performance or engagement. Likewise, RATIO_ATTENDANCE is a behavioral proxy for engagement. For instance, in Datasets 5–7, over 90% of students with attendance below 5% failed to complete the course. This strong, monotonic relationship gives the feature high predictive weight. Prior studies, such as [16], [15], and [27], have shown that attendance-related variables are among the most predictive of student dropout and academic risk.

V. ONBOARD PLATFORM

Once we had a suitable dropout ML model, it could be included in an educational system to support professors in their lectures. Non-functional requirements were equally important as they define the system's performance capacities and limitations, ensuring it meets user expectations and required standards. These include:

- Usability - the platform must feature an appealing, user-friendly interface with useful and straightforward features.
- Accessibility - the platform must be simple to navigate and use, allowing it to be accessible to a wide range of users.
- Scalability - the platform architecture must be built to support the future development and integration of new features as needed.
- Privacy - all data must be securely protected by a data privacy regulation such as the General Data Protection Regulation (GDPR) from the European Union (EU).

The strategy for meeting Usability, Accessibility, and Scalability criteria is to utilize a current tool with established frameworks and an interface that integrates with the academic environment. For this proposal, we selected a previous platform developed by our group, called BATINA, as a baseline platform to implement our dropout components, resulting in the OnBoard architecture depicted in Fig. 5.

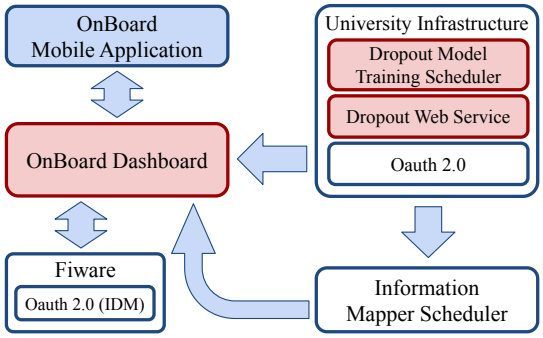


FIGURE 5. OnBoard Architecture to Dropout prediction.

BATINA is a system designed to achieve academic goals and foster connections among students, classes, and teachers. It comprises a web Dashboard and a mobile application based

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

on the IoT Student Advisor and BEst Lifestyle Analyzer (ISABELA) [44] platform. The Dashboard, built using the well-known Dash Plotly Library [45], allows teachers to manage subjects, provide materials, create and monitor questionnaires to students, answer doubts, and view student performance statistics.

The mobile application is specifically designed for students to track their academic progress, interact with a Chatbot, access support materials, view, and complete questionnaires, and submit questions that teachers on the platform will answer. Additionally, students can complete psychological questionnaires, a feature recently added from the ISABELA system. The application is developed in Xamarin Forms and is publicly available for both Android and iOS, excluding the dropout components. The software architecture of the BATINA platform comprises seven main entities: the BATINA Mobile Client, the BATINA Dashboard Server, the FIWARE Service, the UC Services/Infrastructure, the InfoMapper, the ISABELA services, and the ChatBot.

The teachers' Dashboard interacts with three entities: UC, InfoMapper, and FIWARE. FIWARE facilitates OAuth 2.0 login through the IDM service (KeyRock) and provides database access via the Comet Generic Enabler (GE). The UC system provides an OAuth 2.0 interface for teachers to log in using their UC accounts. Once a teacher logs in, InfoMapper, a Java-based service, accesses the UC structure's Web Service to collect class data. The Chatbot available to students is built on DialogFlow and integrated with ChatGPT. Adding ISABELA features requires using additional FIWARE entities, such as Cygnus, Orion, and Intelligence Data Advanced Solution (IDAS), to ensure their proper functioning.

To obtain the OnBoard Platform, the BATINA structure was modified to include a new section dedicated to dropout prediction, focusing on identifying at-risk students by integrating the dropout prediction models into the Dashboard. The components required to allow it are presented in Fig. 5 in red: Dropout Web Services, Dropout Pages, and a Dropout Model Training Scheduler. These components are described below.

A. DROPOUT COMPONENTS

Firstly, the Dropout Web Service enables the generation of dropout probabilities using the best model presented in Section IV-A, which fits the characteristics and time progression of a class without class attendance. This method connects the Dashboard to a Web Service and avoids exposing sensitive data from students as the data is only accessed by the university backend.

The Dropout Pages present the information provided by the Dropout Web Service. The dropout module includes two new pages designed to present dropout-related information. The List Student page, depicted in Fig. 6, provides a comprehensive view of students by subject, showing success rates, questionnaire completion rates, and dropout risk probabilities. Users can click on individual entries for more detailed information, with graphs generated using Plotly displayed below the table.

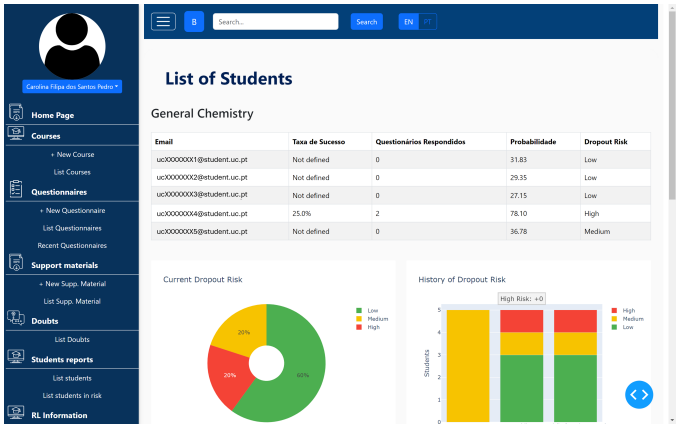


FIGURE 6. Dashboard Page with students' dropout information.

The Web Service was developed to facilitate the communication between the Dashboard and the dropout models. It includes two main routes. The first route retrieves dropout risk data for individual students, performing preprocessing to match the model's training dataset. The second route provides dropout probabilities for all students in a subject/class.

Graphs on the List Students page enable the visualization of dropout risk distribution and its evolution over time. As the dropout probability is a numeric value, the professor can define thresholds to categorize the risk of a student. We set the default threshold value of 66% to identify at-risk students for demonstration purposes. However, it is essential to note that such categories and their relationship with the threshold and the used ML model require further study to determine suitable values in future work.

Course Report Pages include dropout-related tables and graphs. Historical data on high-risk students is displayed for subjects with attendance records, while subjects without attendance records are shown only their current dropout risk. Users can interact with graphs to access detailed lists of at-risk students via the dropout route for all students in a subject or class. This provides professors with all the necessary information related to the class, enabling instructors to make informed pedagogical decisions without disclosing sensitive student information.

Lastly, as the dropout probability is related to multi-factor phenomena, new emergent patterns can emerge from time to time. In this way, the Dropout Model Training Scheduler is required to allow up-to-date models for accurate dropout predictions. As most classes last six months, new data from the university should be collected every six months to update and implement the latest model.

B. DROPOUT PREDICTION WORKFLOW

Fig. 7 illustrates the data flow and inference logic of the dropout prediction module within the OnBoard platform. After passing security validation, the web service receives

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

a request, triggering a data query from the university's academic systems. Each request can be either a student ID for individual predictions or a class ID for batch predictions. For each student, the service checks if attendance data exists. If not, no prediction is made.

When attendance is available, features such as ATTENDANCE_RATIO are extracted. The inference engine evaluates each student in two stages: first, students with low attendance are assessed by Model B. If flagged at risk, their dropout probability is recorded. Remaining students are evaluated by Model A for general dropout risk. The final output is a list of students, each with an associated probability score.

These results are returned to the OnBoard dashboard, enabling instructors to identify at-risk students early and plan preventive interventions based on interpretable prediction scores.

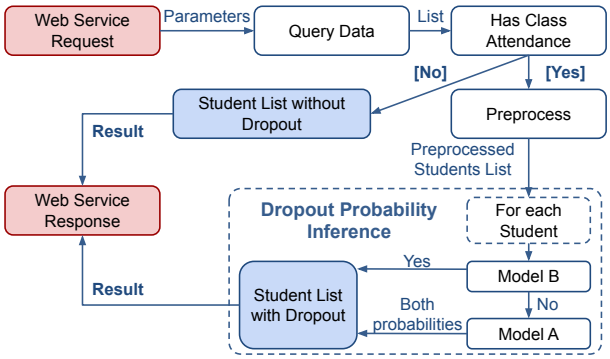


FIGURE 7. Workflow of decision making inside university infrastructure.

C. SECURITY, DISCUSSION AND LIMITATIONS

The OnBoard platform represents a novel educational support tool with significant potential for early intervention in higher education. However, its deployment raises essential concerns related to data privacy, prediction risks, ethical transparency, and human oversight.

The use of student data requires strict adherence to privacy and data protection standards. Since May 25, 2018, the General Data Protection Regulation (GDPR) has been in force in the EU, mandating secure storage, informed consent, and purpose limitation regarding personal data usage [46]. Accordingly, the OnBoard platform is designed to operate only with students' explicit consent and within the scope of the stated educational support objectives. The underlying infrastructure enforces state-of-the-art security protocols, including OAuth 2.0 authentication, role-based access control, and encrypted HTTP communication [47], to safeguard personal data.

To ensure full compliance with GDPR and reinforce data protection, the OnBoard platform implements strong technical safeguards. All data in transit is encrypted using HTTPS with TLS, and data at rest is secured through encrypted storage within the university infrastructure. Access to prediction results is strictly restricted: only authenticated instructors assigned to a specific course can view risk scores for their enrolled students. These predictions are never visible to students or external parties. Furthermore, all access events are logged using an internal audit trail, allowing for accountability and traceability in the handling of sensitive information. These mechanisms ensure that the privacy of at-risk students is protected, even when actionable risk insights are shared with faculty.

To support early intervention, we prioritized recall in our predictive models to maximize the identification of students genuinely at risk of dropping out. This minimizes false negatives—cases where at-risk students are missed—but increases the number of false positives, i.e., students flagged as at risk who may not ultimately drop out. In practical terms, this trade-off is a deliberate design choice intended to avoid missing opportunities for timely support. However, it does carry implications for educators, who may face a higher volume of flagged students. To mitigate undue burden and misinterpretation, the platform emphasizes transparency and clarity of interpretation. Each prediction on the dashboard is accompanied by feature importance for the trained models, helping instructors understand the reasoning behind each risk classification.

Importantly, the platform maintains a human-in-the-loop approach: instructors remain solely responsible for deciding how to respond to flagged students. The risk scores are designed to prompt attention, not prescribe actions. Teachers can use additional context available in the dashboard (e.g., attendance history, performance trends, questionnaire completion) to determine whether follow-up is warranted. This approach ensures that false positives do not result in punitive measures, allowing for supportive and informed outreach. Initial deployments will include feedback mechanisms to assess how instructors perceive and respond to these flags, enabling further refinement of risk thresholds and the interpretability of these tools. These mechanisms will also help calibrate alert sensitivity in alignment with institutional goals and workload considerations.

Bias mitigation is still an ongoing concern. In future interactions, we intend to employ diverse training to address potential systemic biases in the data or model behavior and plan regular bias audits during the platform's iterative deployment. Moreover, we recognize that model predictions can influence future behavior, creating feedback loops that may alter dropout patterns and confound future model performance. Initial deployments will be limited to volunteer instructors, and all interventions will be logged and reviewed to address any issues that arise. Feedback mechanisms will support the refinement of both model and interface, with institutional oversight ensuring alignment with ethical and pedagogical standards.

VI. CONCLUSION

This study investigated early-stage subject-level dropout in higher education using real-world academic data from the University of Coimbra. Through extensive preprocessing, feature engineering, and model evaluation, we developed

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

and validated predictive machine learning models—primarily based on XGBoost and Decision Trees—that demonstrated strong recall and AUC-ROC scores, particularly when leveraging midterm and end-of-semester attendance data. Our findings confirm that attendance, re-enrollment history, and academic performance are key behavioral indicators of dropout risk, aligning with prior literature and institutional expectations.

One of the key insights from our work is that mid-semester attendance metrics provide sufficient predictive power for early intervention, with diminishing returns from additional data collected at the end of the term. This supports the feasibility of timely, actionable predictions. The developed OnBoard platform integrates these predictive insights into a faculty-facing dashboard, providing interpretable risk scores to help student assistance and pedagogical planning.

However, the approach has several limitations. First, the models rely heavily on class attendance data, which may vary in quality and availability across courses and institutions. Second, the prioritization of recall, while useful for flagging at-risk students, increases the likelihood of false positives and necessitates human oversight to prevent misinterpretation. Third, our feature importance analysis suggests potential structural biases, such as higher dropout probabilities associated with re-enrollment status or socioeconomic indicators, which may reflect systemic inequalities. While the platform maintains a human-in-the-loop design and incorporates transparency mechanisms, ongoing monitoring is needed to ensure fair and ethical usage.

Additionally, since all models were trained on data from a single institution, generalizability remains a concern. Local adaptation and validation are essential before deployment in other academic contexts. These challenges underscore the importance of a gradual, feedback-driven implementation approach, supported by ethical governance.

Future work will explore dimensionality reduction techniques such as Principal Discriminant Analysis, Latent Dirichlet Allocation, and embedding methods to address the high dimensionality introduced by one-hot encoding and to improve model efficiency. We also intend to investigate Personalized Federated Learning (PFL) to enable privacy-preserving model training across multiple institutions without sharing raw data. This direction will necessitate a deeper exploration of suitable data encoding strategies and the application of deep learning architectures that can operate effectively within federated environments. Additionally, improving how predictive insights are presented to instructors, through more interpretable visualizations or adaptive feedback, and integrating richer data sources, such as measures of student well-being and engagement, may further enhance the platform's effectiveness and support more personalized interventions.

Ultimately, this work demonstrates that integrating interpretable, data-driven prediction tools into educational support systems holds promise for reducing dropout rates. With careful attention to transparency, fairness, and human-centered design, such tools can proactively help institutions identify and support at-risk students before disengagement.

ACKNOWLEDGMENT

This work was funded by the Recovery and Resilience Plan (PRR), under the terms of the Call for Proposals (AAC) No. 04/C05-i08/2024, 2024.07390.IACDC/2024 - GreenBear, submitted to RE-C05-i08-m04 - 'Supporting the launch of an RD project program aimed at the development and implementation of advanced cybersecurity, artificial intelligence, and data science systems in public administration, as well as a scientific capacity-building program. It was also funded by FCT Pluriannual Funding UID/308: Instituto de Engenharia de Sistemas e Computadores de Coimbra - INESC Coimbra. The authors also acknowledge CAPES (Finance Code 001) and the project OnBoard – "Integrated Tutoring Program for the Prevention of Academic Dropout and Failure," under the Rectorate of the University of Coimbra, PRR POCH-02-53I2-FSE-000032 support.

APPEDIX A: FEATURE DESCRIPTION

Table V describes the selected features.

APPEDIX B: HYPERPARAMETER TUNING

Table VI describes the hyperparameters used in our experiments and the best-performing configuration for each Dataset.

REFERENCES

[1] R. Raja and P. Nagasubramani, "Impact of modern technology in education," *Journal of Applied and Advanced Research*, vol. 3, p. 33, 2018.

[2] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Computers in Human Behavior*, vol. 89, pp. 98–110, 2018.

[3] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021.

[4] J. Bayer, H. Bydžovs, J. Eryk, T. Obšívač, and L. Popelínský, "Predicting drop-out from social behaviour of students," *International Educational Data Mining Society*, 2012.

[5] Z. Chen, J. Zhang, X. Jiang, Z. Hu, X. Han, M. Xu, S. V, and G. Vivekananda, "Education 4.0 using Artificial Intelligence for Students Performance Analysis," *Inteligencia Artificial*, vol. 23, no. 66, pp. 124–137, 2020.

[6] N. Tomasevic, G. N., and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers & Education*, vol. 143, p. 103676, 2020.

[7] Y. Chen and L. Zhai, "A comparative study on student performance prediction using machine learning," *Education and Information Technologies*, vol. 28, pp. 1–19, 2023.

[8] National Center for Education Statistics, "Percentage of Public High School Students Who Graduated in 4 Years and 5 Years, by Selected Characteristics: 2017–18 and 2018–19," 2022. [Online]. Available: https://nces.ed.gov/programs/digest/d22/tables/dt22_326.30.asp (visited on May 7, 2024).

[9] T. H. Nguyen, P. Le, T. T. T. Nguyen, and A. K. Su, "A multivariate analysis of the early dropout using classical machine learning and local interpretable model-agnostic explanations," *CTU Journal of Innovation and Sustainable Development*, vol. 16, no. Special issue: ISDS, pp. 98–106, 2024.

[10] K. Niu, G. Lu, X. Peng, Y. Zhou, J. Zeng, and K. Zhang, "Cnn autoencoders and lstm-based reduced order model for student dropout prediction," *Neural Computing and Applications*, vol. 35, no. 30, pp. 22341–22357, 2023.

[11] O. Lorenzo-Quiles, S. Galdón-López, and A. Lendínez-Turón, "Factors contributing to university dropout: a review," in *Frontiers in education*, vol. 8, p. 1159864, Frontiers Media SA, 2023.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

TABLE 9. Created dataset features, their descriptions, and types.

Mapping time	Feature	Description	Data Type
Pre-entry	GENDER	Student Gender	Binary
Pre-entry	MARITAL_STATUS	Student marital status	Multinominal
Pre-entry	COUNTRY_USUAL_RESIDENCE	Usual country of residence	Multinominal
Pre-entry	COUNTRY_HIGH_SCHOOL	Country where the student completed high school	Multinominal
Pre-entry	NATIONALITY	Student's nationality between "Portugal" and "Other"	Multinominal
Pre-entry	NATURALITY	Country of birth between "Portugal" and "Other"	Multinominal
Pre-entry	EDUCATION_FATHER	Student father's education level at enrollment	Multinominal
Pre-entry	EDUCATION_MOTHER	Student mother's education level	Multinominal
Pre-entry	PROFESSION FATHER	Student father's profession	Multinominal
Pre-entry	PROFESSION MOTHER	Student mother's profession	Multinominal
Pre-entry	SITUATION_PROF_FATHER	Student father's professional situation	Multinominal
Pre-entry	SITUATION_PROF_MOTHER	Student mother's professional situation	Multinominal
Pre-entry	STUDENT_IMPAIRMENT_HEARING	Whether the student has any hearing impairment	Binary
Pre-entry	STUDENT_DIS_MOTOR	Whether the student has any other motor disability not covered in other features	Binary
Pre-entry	STUDENT_DISABILITY_OTHER	Whether the student has any other disability not covered by others	Binary
Pre-entry	STU_IMP_VISUAL	Whether the student has any visual impairment	Binary
Pre-entry	STUDENT_ORGANIC_DISEASE	Organic disability	Multinominal
Pre-entry	STUDENT_DISPLACED	Whether the student is displaced from his homeland	Binary
Pre-entry	STUDENT_INTERNATIONAL	Whether the student is an international student	Binary
Yearly	AGE	Student's age	Numeric
Yearly	SITUATION_PROF_STUDENT	Student's professional situation	Multinominal
Yearly	PROFESSION_STUDENT	Student's profession for students that work during the course	Multinominal
Yearly	SPECIAL_SITUATION	Whether the student has a special status or not	Binary
Yearly	TUTORING	Whether the student was in tutoring	Binary
Yearly	OCCURRENCE_REGIME	Whether the student is enrolled in the first phase or not	Binary
Every Enroll	NUMBER_ENROLLMENTS	Number of subjects at enrollment	Numeric
Every Enroll	MODE_ENROLLMENT	Whether the student is enrolled in the subject full-time or not	Multinominal
Every Enroll	RATIO_GRADES	Proportion of positive grades across all finished classes	Numeric
Every Enroll	REG_VALID_TYPE_APPLICATION	Application record type between "DGES" and "Other"	Multinominal
Every Enroll	REG_DESIGNATION	Record designation between "General" and "Other"	Multinominal
Every subject	SUBJECT_TYPE	whether it is part of the student's study plan of the course or not	Binary
Every subject	IMPROVEMENT	Whether the student is enrolled in the subject to improve the grade	Binary
Every subject	RE-ENROLLMENT	Whether or not the student has already been enrolled in the subject	Binary
Every subject	ENROLL_TYPE	Registration type between "NORMAL" and "Other"	Multinominal
Class ratio	RATIO_ATTENDANCE	The ratio of student attendance in a class period	Numeric
Every subject	LABEL	Whether the student dropped out of the subject	Binary

[12] N. U. R. Junejo, Q. Huang, X. Dong, C. Wang, A. Zeb, M. Humayoo, and G. Zheng, "Sappnet: students' academic performance prediction during covid-19 using neural network," *Scientific Reports*, vol. 14, no. 1, p. 24605, 2024.

[13] E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, p. 4067721, 2024.

[14] S. Matz, C. Bukow, H. Peters, C. Deacons, A. Dinu, and C. Stachl, "Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics," *Scientific Reports*, vol. 13, 2023.

[15] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in finnish higher education," *Technology in Society*, vol. 76, p. 102474, 2024.

[16] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," in *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain*, pp. 41–50, 2009.

[17] M. Nagy and R. Molontay, "Predicting dropout in higher education based on secondary school performance," in *2018 IEEE 22nd international conference on intelligent engineering systems (INES)*, pp. 000389–000394, IEEE, 2018.

[18] J. A. Talamás-Carvajal and H. G. Ceballos, "A stacking ensemble machine learning method for early identification of students at risk of dropout," *Education and Information Technologies*, vol. 28, no. 9, pp. 12169–12189, 2023.

[19] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in *Knowledge-Based Intelligent Information and Engineering Systems* (V. Palade, R. J. Howlett, and L. Jain, eds.), pp. 267–274, Springer Berlin Heidelberg, 2003.

[20] M. Cannistrà, C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni, "Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques," *Studies in Higher Education*, vol. 47, no. 9, pp. 1935–1956, 2022.

[21] L. Blanquet, J. Grilo, P. Strecht, and A. Camanho, "Curbing dropout: Predictive analytics at the university of porto," in *23.ª Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI'2023)*, pp. 143–159, CAPSI, 2023.

[22] M. Vieira Martins, L. Baptista, J. Machado, and V. Realinho, "Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education," *Applied Sciences*, vol. 13, p. 4702, 2023.

[23] T. H. Nguyen, P. Le, T. T. T. Nguyen, and A. K. Su, "A multivariate analysis of the early dropout using classical machine learning and local interpretable model-agnostic explanations," *CTU Journal of Innovation and Sustainable Development*, vol. 16, no. Special issue: ISDS, pp. 98–106, 2024.

[24] M. Delogu, R. Lagravinese, D. Paolini, and G. Resce, "Predicting dropout from higher education: Evidence from italy," *Economic Modelling*, vol. 130, p. 106583, 2024.

[25] J. Niyogisubizo, L. L., E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100066, 2022.

[26] K. Niu, G. Lu, X. Peng, Y. Zhou, J. Zeng, and K. Zhang, "Cnn autoencoders and lstm-based reduced order model for student dropout prediction," *Neural Computing and Applications*, vol. 35, no. 30, pp. 22341–22357, 2023.

[27] O. Goren, L. Cohen, and A. Rubinstein, "Early prediction of student dropout in higher education using machine learning models," in *Proceedings of the 17th International Conference on Educational Data Mining*, pp. 349–359, 2024.

[28] M. Fei and D.-Y. Yeung, "Temporal models for predicting student dropout in massive open online courses," in *2015 IEEE international conference on data mining workshop (ICDMW)*, pp. 256–263, IEEE, 2015.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

TABLE 10. Hyper parameters tuning on each Dataset

ML	Parameters	Values	DS1	DS2	DS3	DS4	DS5	DS6	DS7
NB	Var_smoothing	[1e-12, 1e-11, 1e-10, 1e-9, 1e-8, 1e-7, 1e-6]	1e-12,	1e-12,	1e-12,	1e-12,	1e-12,	1e-12,	1e-12,
DT	criterion	['gini', 'entropy']	entropy	gini	gini	entropy	gini	entropy	entropy
	max_depth	[None, 3, 5, 10]	5	10	3	3	5	3	3
	min_samples_split	[2, 5, 10]	2	2	2	2	2	2	2
	min_samples_leaf	[1, 2, 5]	1	1	1	1	1	1	1
	max_features	[None,'sqrt', 'log2']	NaN	sqrt	NaN	NaN	sqrt	NaN	NaN
XGBoost	subsample	[0.8, 1.0]	0.8	0.8	1	1	1	1	1
	colsample_bytree	[0.6, 0.8, 1.0]	1	1	0.6	0.6	0.8	0.6	0.8
	max_depth	[3, 5, 10]	10	3	3	3	5	3	3
	learning_rate	[0.01, 0.1]	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	n_estimators	[50, 200, 500]	50	500	200	200	500	500	500
	gamma	[1, 5]	1	1	5	1	1	1	1
	min_child_weight	[1, 5, 10]	5	5	1	1	1	1	5
RF	n_estimators	[50, 200]	200	50	200	200	200	50	50
	max_depth	[3, 5, 10]	10	10	10	10	10	10	10
	max_leaf_nodes	[5, 10, 20]	20	20	20	20	20	20	20
	min_samples_split	[2, 5, 10]	2	2	2	2	2	2	2
	min_samples_leaf	[1, 2, 5]	1	1	1	1	1	1	1
	bootstrap	[True, False]	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE
KNN	n_neighbors	[2,3,5,10]	10	10	10	10	10	10	10
	weights	['uniform', 'distance']	uniform	uniform	uniform	uniform	uniform	uniform	uniform
	p	[1, 2]	1	1	1	1	1	1	1
	metric	['euclidean', 'manhattan', 'minkowski']	euclidean	euclidean	euclidean	euclidean	euclidean	euclidean	euclidean

[29] S.-U. Hassan, H. Waheed, N. R. Aljohani, M. Ali, S. Ventura, and F. Herrera, "Virtual learning environment to predict withdrawal by leveraging deep learning," *International Journal of Intelligent Systems*, vol. 34, no. 8, pp. 1935–1952, 2019.

[30] D. Sun, Y. Mao, J. Du, P. Xu, Q. Zheng, and H. Sun, "Deep learning for dropout prediction in moocs," in *2019 eighth international conference on educational innovation through technology (EITT)*, pp. 87–90, IEEE, 2019.

[31] A. C. Lorena, L. F. Jacintho, M. F. Siqueira, R. De Giovanni, L. G. Lohmann, A. C. De Carvalho, and M. Yamamoto, "Comparing machine learning classifiers in potential distribution modelling," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5268–5275, 2011.

[32] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[33] K. H. Wilson, X. Xiong, M. Khajah, R. V. Lindsey, S. Zhao, Y. Karklin, E. G. Van Inwegen, B. Han, C. Ekanadham, J. E. Beck, *et al.*, "Estimating student proficiency: Deep learning is not the panacea," in *In Neural information processing systems, workshop on machine learning for education*, vol. 3, 2016.

[34] R. Basnet, C. Johnson, and T. Doleck, "Dropout prediction in moocs using deep learning and machine learning," *Education and Information Technologies*, vol. 27, pp. 1–15, 2022.

[35] G. Dávila, J. Haro, A. González-Eras, O. R. Vivanco, and D. G. Coronel, "Student Dropout Prediction in High Education, Using Machine Learning and Deep Learning Models: Case of Ecuadorian University," in *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1677–1684, 2023.

[36] I. Sandoval, D. Naranjo-Tovar, J. Vidal, and R. Gilar, "Early Dropout Prediction Model: A Case Study of University Leveling Course Students," *Sustainability*, vol. 12, p. 9314, 2020.

[37] OECD, "OECD Digital Education Outlook 2021," tech. rep., 2021, 2021.

[38] J. Liang, C. Li, and L. Zheng, "Machine learning application in moocs: Dropout prediction," in *2016 11th International Conference on Computer Science & Education (ICCSE)*, pp. 52–57, 2016.

[39] D. Holt, A. Scott, and P. Ewings, "Chi-squared tests with survey data," *Journal of the Royal Statistical Society: Series A (General)*, vol. 143, no. 3, pp. 303–320, 1980.

[40] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res. (JAIR)*, vol. 16, pp. 321–357, 2002.

[41] P. Perchinunno, M. Bilancia, and D. Vitale, "A statistical analysis of factors affecting higher education dropouts," *Social Indicators Research*, vol. 156, 2021.

[42] P. E. McKnight and J. Najab, "Mann-whitney u test," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.

[43] R. F. Woolson, "Wilcoxon signed-rank test," *Encyclopedia of biostatistics*, vol. 8, 2005.

[44] J. Fernandes, D. Raposo, N. Armando, S. Sinche, J. S. Silva, A. Rodrigues, V. Pereira, and F. Boavida, "An integrated approach to human-in-the-loop systems and online social sensing," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 478–483, 2019.

[45] Plotly, "Dash - A Python framework for building analytical web applications." [Online]. Available: https://dash.plotly.com/ (visited on Jun. 1, 2024).

[46] GDPR-info.eu, "GDPR Information." [Online]. Available: https://gdpr-info.eu/ (visited on Jun. 15, 2024).

[47] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, "A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions," *Electronics*, vol. 12, no. 6, p. 1333, 2023.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3581751

GUILHERME ANTONIO BORGES obtained a bachelor's degree in technology of Internet systems from IFSul in 2012 and a master's degree in computer science from UFRGS in 2015. Now, he is a Ph.D. candidate in Electrical Engineering and Intelligent Systems at the University of Coimbra. He is also a full professor of informatics at IFSUL. His research interests include self-adaptive systems, educational data mining, Human-in-the-Loop, artificial intelligence, federated learning, and edge computing. (https://orcid.org/0000-0002-4988-7306)

CAROLINA FILIPA DOS SANTOS PEDRO obtained a bachelor's degree in Biomedical Engineering from the University of Coimbra (UC). She is now a master's student in Biomedical Engineering, specializing in Clinical Informatics and Bioinformatics. (https://orcid.org/0009-0007-9175-2505)

ANDRÉ RODRIGUES has a B.Sc. in Informatics Engineering from the University of Coimbra (Portugal), an M.Sc. in Finance from ISCTE Business School, and a Ph.D. in Informatics Engineering from the University of Coimbra in 2013. He works as a teacher at the Polytechnic Institute of Coimbra, giving classes on networking. His main research interests are Industrial Wireless Sensor Networks, the people-centric Internet of Things, and Network Management. He has authored several papers in international journals and conferences in those areas. He has been a reviewer at top conferences and participated in several European initiatives and projects. He is a researcher at the Centre of Informatics and Systems of the University of Coimbra (CISUC). (https://orcid.org/0000-0003-3849-5515)

FERNANDO BOAVIDA received his PhD in Informatics Engineering in 1990. He is a Full Professor at the Department of Informatics Engineering (DEI) of the Faculty of Sciences and Technology of the University of Coimbra. His main research interests are people-centric Internet of Things, cryptography, and privacy. He is the author/co-author of over 200 international publications (books, book chapters, refereed journals, and conference proceedings) and 50 national publications. He participated in several European projects, such as FP6 E-NEXT, EuQoS (IST-FP6-2004-004503), WEIRD (IST-FP6 Integrated Project 034622), OpenNet (IST-FP6 Specific Support Action 035185), CONTENT (IST-FP6-0384239), GINSENG (ICT-FP7-224282), MICIE (ICT-FP7-225353), and POSEIDON (Grant Agreement no. 786713, H2020-DS-2016-2017/ DS-08-2017). He authored one international book and five Portuguese textbooks, widely used as course books in universities and polytechnic schools of Portuguese-speaking countries, covering computer network engineering, administration, TCP/IP networking, and wireless sensor networks.

JULIO CESAR SANTOS DOS ANJOS is an adjunct professor at the Federal University of Ceara, Campi Itapaje/Brazil, and in the Post Graduate Program in Teleinformatics Engineering (PPGETI/UFC) at the Center of Technology, Fortaleza/Ceara. He obtained a Ph.D. in Computer Science from the Federal University of Rio Grande do Sul (UFRGS/RS) in 2017, a Post-doc in Computer Science on 10/2021 at UFRGS, and by UNICAMP on 05/2022. Master's degree in Computer Science from the Federal University of Rio Grande do Sul in 2012. In addition, he is a bachelor's degree in electrical engineering from the Pontifical Catholic University of Rio Grande do Sul (PUC/RS) in 1991. Research interest: Big Data, Streaming processing, Distributed and Heterogeneous Systems, Hybrid infrastructures, Collaborative Learning, Intelligent Autonomous Systems, and Big Data Analytics with Machine Learning and Deep Learning. (https://orcid.org/0000-0003-3623-2762)

JORGE SÁ SILVA received his PhD in Informatics Engineering in 2001 from the University of Coimbra, where he is an Associate Professor with Habilitation at the Department of Electrical and Computer Engineering (DEEC) and a Researcher of the Institute for Systems Engineering and Computers at Coimbra (INESC Coimbra). His main research interests include the Internet of Things, Network Protocols, and Human-in-the-Loop. He is a senior member of IEEE and a licensed professional engineer. His homepage is at https://home.deec.uc.pt/ sasilva. (https://orcid.org/0000-0002-6273-1285)

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/