Article

Identifying At-Risk Students for Early Intervention—A Probabilistic Machine Learning Approach

Eli Nimy 1, Moeketsi Mosia 2,* and Colin Chibaya 3

1 School of Natural and Applied Sciences, Sol Plaatje University, Kimberley 8300, South Africa
2 Centre for Teaching, Learning, and Programme Development, Sol Plaatje University, Kimberley 8300, South Africa
3 Department of Computer Science and Information Technology, Sol Plaatje University, Kimberley 8300, South Africa
* Correspondence: moeketsi.mosia@spu.ac.za

Abstract: The utilization of learning analytics to identify at-risk students for early intervention has exhibited promising results. However, most predictive models utilized to address this issue have been based on non-probabilistic machine learning models. In response, this study incorporated probabilistic machine learning for two reasons: (1) to facilitate the inclusion of domain knowledge, and (2) to enable the quantification of uncertainty in model parameters and predictions. The study developed a five-stage, probabilistic logistic regression model to identify at-risk students at different stages throughout the academic calendar. Rather than predicting a student’s final or exam mark, the model was focused on predicting the at-risk probabilities for subsequent assessments—specifically, the probability of a student failing an upcoming assessment. The model incorporated student engagement data from Moodle, as well as demographic and student performance data. The study’s findings indicate that the significance and certainty of student engagement and demographic variables decreased after incorporating student-performance variables, such as assignments and tests. The most effective week for identifying at-risk students was found to be week 6, when the accuracy was 92.81%. Furthermore, the average level of uncertainty exhibited by the models decreased by 60% from stage 3 to 5, indicating more reliable predictions at later than earlier stages. The study highlights the potential of a probabilistic machine learning model to aid instructors and practitioners in identifying at-risk students, and thereby to enhance academic outcomes.

Keywords: machine learning; probabilistic machine learning; predictive analytics; at-risk students; early intervention

1. Introduction

How to enhance student learning is an open question, and one that is an intellectual project for many scholars. Questions relating to student learning are so crucial that many theories have been developed to help scholars and practitioners understand student learning. Recently, there have been demands that higher-education institutions improve student learning [1] through data analytics, commonly known as learning analytics (LA). While scholars are not unified in one definition of LA, it is widely accepted that LA refers to “the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environment in which it occurs” [2] (p. 34). Learning analytics primarily relies on data extracted from learning management systems (LMS) and employs data analysis and visualization tools for descriptive analytics, along with machine learning algorithms for predictive learning analytics. In some ways, the use of data analytics to improve student learning is related to developments in the notions of big data. While what is meant by big data differs for different contexts, its general meaning may be characterized by what is called the 3Vs, namely, (i) volume—the amount of data, (ii) velocity—the pace at which data are generated, (iii) and variety—the
different types of data generated [3]. Primarily, higher education provides greater variety of data than volume and velocity; thus, much of learning analytics research leverages these different types of data to optimize student learning and the environment in which learning occurs. This paper focuses on predictive learning analytics, which relies on algorithms for predicting at-risk students. In this paper, machine learning is understood as a modelling process of combining data with algorithms to form models that can uncover patterns in data to predict the future [4]. Using machine learning in the context of educational data mining has seen an upsurge in leveraging data, with a primary focus on understanding at-risk students and predicting at-risk students [5–8].

Identifying at-risk students remains difficult, despite numerous studies using learning analytics to identify these students, so that institutions can improve their remedial actions [7–10]. Most studies applied machine learning techniques based on regression and classification tasks to improve accuracy and performance in predicting at-risk students [7]. These studies primarily adopted non-probabilistic machine learning approaches, and generally used decision trees, random forests, neural networks, support vector machines, and naïve Bayes as algorithms [7,10].

Despite the increase in documented evidence of research on the use of machine learning in learning analytics, including the studies referred to above, and others, little is known about building models that account for uncertainty in predicting at-risk students. As stated above, big data in the education context relates more to the various types (variety) of data than to volume and velocity; thus, predicting at-risk students has a high level of uncertainty, which is not accounted for when using non-probabilistic machine learning. Even more importantly, the ability to predict at-risk students enables higher-education institutions to create opportunities for students to receive early interventions. The accuracy of predicting at-risk students is insufficient for higher-education institutions due to, amongst others, a shortage of resources and because both academic and psychosocial support programs are not available on time, because of high student enrolments. Thus, modelling uncertainty helps provide support where it is needed the most. This paper seeks to apply a probabilistic machine learning (PML) approach to quantify uncertainty in predicting at-risk students, to optimize the impact of support programs. In so doing, the objectives of this study were as follows.

- To determine features that can serve as predictors of at-risk students.
- To build a probabilistic machine learning model and evaluate its performance.
- To identify the best academic calendar week for identifying at-risk students for early intervention.

The present study proposes a novel approach to addressing the problem of identifying at-risk students for early intervention by using a different combination of indicators in a geographical area (South Africa) that has received little scholarly attention in learning analytics. By exploring the relationship between the identified indicators and at-risk students, we also aim to highlight the potential for geographic variations in the significant indicators for predicting at-risk students. Moreover, the study introduced a novel application of probabilistic machine learning (PML) to the at-risk student identification problem, which is a departure from traditional machine learning models that only provide point estimates. This approach allows for the quantification of uncertainty in model predictions, providing a more reliable assessment of the probability of students being identified as at-risk. This innovative use of PML represents a valuable contribution to the field of learning analytics, particularly for addressing the challenges of predicting at-risk students. Overall, this study offers a significant contribution to the existing body of knowledge on at-risk student identification by introducing a unique approach and providing novel insights.

2. Literature Review

2.1. Probabilistic Machine Learning

The demand for machine learning approaches that are grounded in principles has been increasing among non-specialists in diverse fields, including education, healthcare, and
finance. This trend has led to greater support for probabilistic modeling, driven by a need for transparent models that provide students with insight into the factors that underpin the model’s prediction, as well as measurements of uncertainty. In other words, there is a desire for models that know when they do not know [4,11]. Probabilistic machine learning (PML) models are particularly useful because they enable us to understand why a specific prediction was made and with what level of certainty that prediction was made. As [12] notes, PML is focused on decision making, which is the ultimate goal of identifying at-risk students—namely, determining whether a student is at-risk. However, this process inherently involves uncertainty, and accurately describing a predicted outcome is not always possible [13].

Probabilistic models offer the advantage of transparency, and they can also integrate domain knowledge through prior probability distributions [14]. Prior probability distributions are informed by the experimenter’s understanding of the data-generating process or the system that produced the data [15]. Specifically, they reflect the experimenter’s knowledge about the value of a parameter of the model prior to observing the data [16]. In this study’s context, this could entail what is known about students’ grades before examining their performances throughout the semester.

2.2. Predictive Analytics for at-Risk Student Identification

In higher education, identifying at-risk students for timely intervention is a significant problem [5,6,8]. Studies on at-risk student identification provide evidence on the usefulness of utilizing various machine learning algorithms to identify low-performing or at-risk students as a course progresses.

A time-series clustering approach to identifying at-risk online students was proposed by [8], who investigated if time-series clustering can outperform traditional methods in terms of accuracy, using dynamic data that consisted of Moodle logs, and static data representing students’ demographics and academic performances observed over 16 weeks. The analysis used six predictive algorithms: logistic regression (forward, backward, and stepwise), decision trees, rule induction, and boosting. The study concludes that the time-series clustering approach, in accuracy and feasibility, outperforms aggregated models built on machine learning and time-series algorithms alone. The decision tree was the best-performing algorithm, achieving an accuracy of 84%, and it had the ability to accurately capture at-risk students from week 10.

Er [6] investigated the effects of time-variant (dynamic) data and time-invariant (static) data on the accurate prediction of at-risk students in an online course. Er [6] suggests combining multiple algorithms to achieve more accurate results. This was presented in three decision schemes: Scheme 1 requires at least one algorithm to classify a student as being at-risk for the student to be considered at-risk; and Schemes 2 and 3 needed at least two and three algorithms, respectively. The three algorithms used to support the three schemes for 10 weeks were the instance-based learning classifier, decision trees, and naïve Bayes. The study concludes that excluding time-invariant data has no significant influence on the overall results of classifying at-risk students.

In another study, low-cost learning analytics practices were used to identify at-risk students in a quantitative methods in business course taken by undergraduate students [5]. The study used student demographic data, summative assessments, and clicker responses as formative assessments. Using data generated from the learning management system, the researchers used Google Sheets and Google Forms to collect and analyze clicker data. The study used linear and logistic regression to identify at-risk students in five stages over 12 weeks, with three weeks intervals per stage. The study found that linear regression could be more acceptable, due to its more refined recall performance (the ability of predictive models to identify at-risk students) with prediction intervals. It is less frequently used than logistic regression in previous studies.

Reading data from Student’s eBook was used to develop an early identification system for students at-risk of academic failure in the context of eBook-based teaching and...
learning [17]. This study used data from 90 undergraduate students in an elementary informatics course over 16 weeks. The researchers created prediction models using 13 algorithms. The algorithms were evaluated using Cohen’s kappa to determine the degree to which the model performed better than chance, and accuracy. The study found that the algorithms achieved their best performances from week 15, and the predictive models classified at-risk students with an accuracy of over 79% from week three. The researchers noted that predictive models with transformed data yielded poor performances.

Various machine learning techniques have been applied to identify at-risk students [5,6,8,17], the most common ones being decision trees, the random forest, neural networks, the support vector machine, and naïve Bayes. The machine learning techniques are implemented and compared using accuracy, F-score, recall, and precision [1,7,10]. However, these techniques were, in most cases, implemented to improve predictive accuracy, suitability, and performance, depending on the type, nature, and availability of data [7]. The studies were contextualized by identifying significant indicators that enhance the predictive power, including clicks, eBook interactions, quizzes, and attendance [5,17,18]. Furthermore, researchers applied advanced techniques, such as k-means and time-series clustering, to improve the accuracy of the predictive models [8,19]. Additionally, multiple algorithms were employed either to select the best algorithm or to use all algorithms collectively to classify a student as being at-risk [6,17]. Lastly, automated machine learning tools, such as Auto-Weka and Auto-Sklearn, were utilized to enhance the prediction of student success [1].

3. Materials and Methods

This study followed the five-stage learning analytics methodological approach proposed by [20]. The five-stage process consists of the steps capture, report, predict, act, and refine (Figure 1).

![Figure 1. Five-stage learning analytics approach.](image)

3.1. Data Description

The Moodle and student management information systems served as sources of student engagement and academic performance data. The data were based on a compulsory second-semester module for all first-year students registered at a South African university for the 2020 academic year. The student data from the two systems were merged and anonymized for analysis.
The combined student data contained 517 Moodle and academic performance observations for 517 students. Table 1 describes the student data obtained over 11 weeks (14 September–30 November 2020).

Table 1. Student data description.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics and performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TITLE</td>
<td>Nominal</td>
<td>Student’s Title (Mr, Ms)</td>
</tr>
<tr>
<td>QUAL CODE</td>
<td>Nominal</td>
<td>Student qualification code</td>
</tr>
<tr>
<td>CLASS GROUP</td>
<td>Nominal</td>
<td>Student’s class group (Class A-F)</td>
</tr>
<tr>
<td>AS 1</td>
<td>Numerical</td>
<td>Assessment 1 mark</td>
</tr>
<tr>
<td>TM 1</td>
<td>Numerical</td>
<td>Test 1 mark</td>
</tr>
<tr>
<td>TM 2</td>
<td>Numerical</td>
<td>Test 2 mark</td>
</tr>
<tr>
<td>CA 1</td>
<td>Numerical</td>
<td>Continuous assessment</td>
</tr>
<tr>
<td>Moodle Data (Student engagement)</td>
<td></td>
<td></td>
</tr>
<tr>
<td># OF COURSES</td>
<td>Numerical</td>
<td>Number of courses taken by the student in 2020</td>
</tr>
<tr>
<td>TIME ON SITE</td>
<td>Numerical</td>
<td>Time student spent on Moodle (cumulative)</td>
</tr>
<tr>
<td>TIME ON COURSES</td>
<td>Numerical</td>
<td>Time student spent on the course (cumulative)</td>
</tr>
<tr>
<td>TIME ON ACTIVITIES</td>
<td>Numerical</td>
<td>Time student spent on course activities (cumulative)</td>
</tr>
</tbody>
</table>

3.2. Data Preprocessing

This section describes the data pre-processing procedure conducted before the probabilistic machine learning models were trained.

3.2.1. Label Encoding

In label encoding, nominal variables are replaced with numerical values between 0 and the number of classes in the nominal variable minus 1 ($0 - n - 1$) [21]. All the nominal variables in student data were encoded using label encoding; the encoding was done in an alphanumeric order.

3.2.2. Feature Scaling

Standardization was used as a feature-scaling method to deal with the high variability in measurements—the following formula was used:

$$x_{std} = \frac{x - u_x}{\sigma_x}$$  \hspace{1cm} (1)

where $x$ is a feature (numerical variable); $\sigma_x$ and $u_x$ are the sample standard deviation and mean, respectively.

3.2.3. Feature Selection

In this study, feature selection was performed through correlation analysis and feature importance (using extra trees classifier). This was a crucial step, as features used for training PML or machine learning models significantly influence performance; thus, having irrelevant or partially irrelevant features can negatively impact a model’s performance [22]. Two methods were used to identify consistent features that can be used for modeling.

3.2.4. Correlation Analysis

Correlation analysis was used as a statistical method to evaluate the strengths of relationships between all the features and the target feature (subsequent assessments). Pearson’s correlation coefficient was used, which is calculated as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$  \hspace{1cm} (2)
where \( r \) is the correlation coefficient; \( x_i \) and \( y_i \) are the values of the \( x \) variable and \( y \) variable, respectively; and \( \bar{x} \) and \( \bar{y} \) are the mean values of the \( x \) variable and \( y \) variable, respectively.

3.2.5. Feature Importance

The extra tree classifier (ETC) algorithm was used to select important features for predicting subsequent assessments for identifying at-risk students. The ETC is an ensemble algorithm that seeds numerous tree models built randomly from the student data and sorts out the most-voted-for features [23]. The ETC was constructed with 100 trees using the default criterion (Gini–Gini impurity) for identifying essential features.

3.2.6. Handling Class Imbalance

The problem of identifying at-risk students was approached as a binary classification task, wherein students deemed at-risk were assigned a classification of 0, and those not at-risk were assigned a classification of 1. However, the data exhibited a significant class imbalance—a minimum of 76% of students passed each subsequent assessment. As it is well known, class imbalance occurs when one class has significantly more samples than the other [24]. To mitigate this issue, the synthetic minority oversampling technique (SMOTE) was employed in this study. SMOTE was used to oversample the minority class by creating synthetic examples, which in turn enabled matching the number of samples in the majority class. The SMOTE technique can be likened to a form of data augmentation [24].

3.3. Model Stages

The study aimed to build a PML model to identify at-risk students for early intervention. The early intervention part of the study’s aim was crucial. To incorporate the early intervention aspect of the aim, a five-stage PML model was built by incorporating the ongoing assessment and student engagement data, where \( s \) represents stages 1 to 5.

**Stage 1 (2nd week):** Predicting Assessment 1 (AS_1):

\[
p_1 = \text{Sigmoid}(\beta_0 + \beta_1 X_1 + \beta_2 X_2)
\]

(3)

**Stage 2 (4th week):** Predicting Assessment 1 (AS_1):

\[
p_2 = \text{Sigmoid}(\beta_0 + \beta_1 X_1 + \beta_2 X_2)
\]

(4)

**Stage 3 (6th week):** Predicting Test 1 (TM_1):

\[
p_3 = \text{Sigmoid}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)
\]

(5)

**Stage 4 (7th week):** Predicting Test 2 (TM_2):

\[
p_4 = \text{Sigmoid}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)
\]

(6)

**Stage 5 (10th week):** Predicting Continuous Assessment (CA_1):

\[
p_5 = \text{Sigmoid}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5)
\]

(7)

In each stage, the variables in Table 1 served as input in the training data to predict the at-risk probability (i.e., the probability of a student failing an assessment) of an assessment at a particular stage \( s \). The features are \( X_1 \) for gender and \( X_2 \) for time spent on Moodle before taking an assessment (\( X_2 \) is accumulative, and is thus different for each model). \( X_3, X_4, \) and \( X_5 \) represent the assessments taken by students.

3.4. Formulating a Logistic Regression Model

A logistic regression model was implemented as a base model to compare model formulation and performance to the probabilistic logistic regression model. In logistic regression, the linear relationship between student data variables and binary outcome
(at-risk or not at-risk) estimates were mediated by a sigmoid function to ensure the logistic regression model produces probabilities [16]. The logistic regression model was formulated as:

\[ \logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \]  

\[ p = \frac{1}{1 + e^{-\logit(p)}} \]  

The logistic regression models the probability \( p \) that student \( i \) is at-risk based on \( k \) features at stage \( s \), where \( \logit(p) \) is the logistic regression model for stage \( s \), represented as \( \logit(p) \).

3.5. Formulating a Probabilistic Machine Learning Model

Like the logistic regression model above, a binary-classification PML model was formulated to classify a student as at-risk or not-at-risk for subsequent assessments.

3.5.1. Probabilistic Machine Learning Model Prior

Uninformative priors were used to express objective information due to limited information on student engagement and performance for the year 2020. These priors assumed a normal distribution with a mean of 0 and a standard deviation of \( 10^6 \).

3.5.2. Probabilistic Machine Learning Model Likelihood

Since the outcome is binary (risk or not at-risk), a Bernoulli distribution was used to model the probability of the data given the parameter \( p \), as follows:

\[ P(\text{Not at-risk}|p) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \]  

where \( y_i = 0 \) if a student is at-risk and \( y_i = 1 \) if the student is not at-risk, and \( p_i \) is the probability of a student being at-risk.

3.6. Parameter Uncertainty and Model Evaluation

A forest and a posterior plot were used at every stage, \( s \), with a 94% highest density interval (HDI) for parameter uncertainty. The HDI is one of the ways of defining a credible interval. The HDI credible interval was used to indicate which distribution points are most credible and which cover most of the distribution. Credible intervals are the uncertainty levels in the model’s parameters.

The evaluation of the predictive models’ performance was conducted utilizing several established metrics, namely, accuracy, F1-score, precision, and recall. These metrics were derived from the false negative (FN), false positive (FP), true negative (TN), and true positive (TP) values obtained from the confusion matrix. Such a comprehensive evaluation ensures a thorough and objective assessment of the models’ predictive capabilities in identifying at-risk students for early intervention.

\[ \text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \]  

Accuracy is a measure of how correctly the model identifies the at-risk students. It is defined as the ratio of the number of correctly identified at-risk students to the total number of students in the dataset. High accuracy means that the model can accurately identify at-risk students.

\[ \text{Precision} = \frac{TP}{TP + FP} \]  

Precision is the ratio of true positives (students who are identified as at-risk and are actually at-risk) to the total number of students identified as at-risk (true positives plus
false positives). A high precision indicates that when the model identifies a student as at-risk, it is likely to be correct.

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{13}
\]

Recall is the ratio of true positives to the total number of at-risk students in the dataset (true positives plus false negatives). A high recall indicates that the model can identify most of the at-risk students in the dataset.

\[
F1 - \text{Score} = \frac{2TP}{2TP + FP + FN} \tag{14}
\]

F1-score is a measure of the model’s accuracy that considers both precision and recall. It is the harmonic mean of precision and recall. A high F1-score indicates that the model performs well in identifying at-risk students.

The proposed study presents several significant societal benefits. Firstly, it can improve academic outcomes by identifying at-risk students earlier, allowing for timely interventions and support to improve academic achievement. This can ultimately lead to higher student retention rates and reduce costs associated with student attrition. Secondly, it can enable more efficient use of resources by allowing targeted interventions, reducing the need for resources to be allocated to students who may not need them. Thirdly, this study highlights the potential of probabilistic machine learning for incorporating domain knowledge and uncertainty, thereby improving the accuracy and usefulness of predictive models in various fields beyond education. Fourthly, it can enhance teachers’ effectiveness by enabling them to tailor their teaching approach and resources to meet the needs of each student. Fifthly, it can promote equity by ensuring that all students receive appropriate support, regardless of background or demographic characteristics. Overall, this study offers far-reaching benefits for society, showcasing the potential of probabilistic machine learning for improving academic outcomes, reducing educational disparities, and enhancing the overall efficiency of the education system.

4. Results

This section starts by responding to the first objective, which was to determine features that can serve as predictors of at-risk students. Therefore, to accomplish the first objective, the study used Pearson’s correlation and an extra trees classifier to select features that were useful for predicting at-risk students in each stage. The Pearson’s correlation and ETC results are shown in Tables 2 and 3, respectively. The variables with asterisks (*) are assessments predicted at each stage—assessment 1 (in stages 1 and 2), test 1 (in stage 3), test 2 (in stage 4), and continuous assessment 1 (in stage 5). Thus, correlation and feature importance were observed between variables with and without an asterisk (*).

Table 2. Pearson’s correlation between variables with and without an asterisk.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
<th>Stage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>0.1429</td>
<td>0.1631</td>
<td>0.3422</td>
<td>0.3412</td>
<td>0.2272</td>
</tr>
<tr>
<td>QUAL CODE</td>
<td>0.0581</td>
<td>0.0276</td>
<td>−0.0597</td>
<td>−0.0543</td>
<td>0.0403</td>
</tr>
<tr>
<td>CLASS GROUP</td>
<td>−0.0612</td>
<td>−0.0597</td>
<td>0.1509</td>
<td>0.0252</td>
<td>−0.0039</td>
</tr>
<tr>
<td># OF COURSES</td>
<td>0.0581</td>
<td>0.0616</td>
<td>0.1488</td>
<td>0.0160</td>
<td>0.1588</td>
</tr>
<tr>
<td>TIME ON SITE</td>
<td>0.3003</td>
<td>0.3839</td>
<td>0.7380</td>
<td>0.6277</td>
<td>0.5053</td>
</tr>
<tr>
<td>TIME ON COURSES</td>
<td>0.2927</td>
<td>0.3775</td>
<td>0.7497</td>
<td>0.6107</td>
<td>0.5167</td>
</tr>
<tr>
<td>TIME ON ACTIVITIES</td>
<td>0.2426</td>
<td>0.2798</td>
<td>0.7427</td>
<td>0.4853</td>
<td>0.4856</td>
</tr>
<tr>
<td>AS 1 (S1 &amp; S2) *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM 1 (S3) *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TM 2 (S4) *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA 1 (S5) *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The variables with asterisks (*) are assessments predicted at each stage.
Table 3. Extra trees classifier: feature importance between variables with and without an asterisk.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
<th>Stage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>0.0359</td>
<td>0.0349</td>
<td>0.0517</td>
<td>0.0500</td>
<td>0.0380</td>
</tr>
<tr>
<td>QUAL CODE</td>
<td>0.1153</td>
<td>0.1040</td>
<td>0.0347</td>
<td>0.0746</td>
<td>0.0598</td>
</tr>
<tr>
<td>CLASS GROUP</td>
<td>0.0980</td>
<td>0.0929</td>
<td>0.0402</td>
<td>0.0540</td>
<td>0.0564</td>
</tr>
<tr>
<td># OF COURSES</td>
<td>0.1768</td>
<td>0.1690</td>
<td>0.0412</td>
<td>0.0560</td>
<td>0.0723</td>
</tr>
<tr>
<td>TIME ON SITE</td>
<td>0.1945</td>
<td>0.2077</td>
<td>0.1577</td>
<td>0.1207</td>
<td>0.0896</td>
</tr>
<tr>
<td>TIME ON COURSES</td>
<td>0.1985</td>
<td>0.2061</td>
<td>0.1352</td>
<td>0.0991</td>
<td>0.0878</td>
</tr>
<tr>
<td>TIME ON ACTIVITIES</td>
<td>0.1809</td>
<td>0.1852</td>
<td>0.1716</td>
<td>0.0787</td>
<td>0.0975</td>
</tr>
<tr>
<td>AS 1 (S1 &amp; S2) *</td>
<td></td>
<td></td>
<td>0.3676</td>
<td></td>
<td>0.1958</td>
</tr>
<tr>
<td>TM 1 (S3) *</td>
<td></td>
<td></td>
<td>0.2562</td>
<td></td>
<td>0.1093</td>
</tr>
<tr>
<td>TM 2 (S4) *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.1958</td>
</tr>
<tr>
<td>CA 1 (S5) *</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The variables with asterisks (*) are assessments predicted at each stage.

In stage 1, the student engagement variables and the TITLE variable showed stronger positive correlation values than QUAL CODE, CLASS GROUP, and # OF COURSES. Higher correlation values for these variables were also seen in stage 2. From stages 3 to 5, strong and moderately positive correlations were observed, including for assessment variables. However, correlation values were at their highest in stage 3, week 6. Weak positive and negative relationships and lower feature importance values were observed among QUAL CODE, CLASS GROUP, # OF COURSES, and the asterisk variables throughout the five stages. Thus, these variables were not considered for predicting at-risk students. A strong positive correlation was observed among the student engagement variables, and thus, including them all would have resulted in a multicollinearity problem. Since TIME ON COURSES had a stronger correlation and a higher feature importance value for most stages than the other student engagement variables, it was considered for predicting at-risk students, and the other variables were omitted.

While student engagement variables had higher feature importance values for stage 1 and stage 2, a significant drop in feature importance was noted after including performance variables (assessment 1, test 1, and test 2) in stages 3 to 5. After identifying critical variables for predicting at-risk students, a five-stage PML model was constructed by incorporating the ongoing assessments.

The study’s second objective was to build a PML model to predict at-risk students and to evaluate its performance. The performance of the PML model was assessed and compared with that of a standard logistic regression model, as presented in Table 4. While both models demonstrated similar performance levels in all five stages, the PML model offers the additional advantage of enabling the quantification of uncertainty pertaining to both model parameters and predictions. In this context, the uncertainty quantification is useful because it provides a measure of reliability in the model’s predictions. This is important when identifying at-risk students, where decisions will often be made based on a model’s predictions. By quantifying the uncertainty, decision makers can better understand the limitations and potential risks associated with a particular decision or action. The evaluation of the models’ performances in the five stages revealed slight discrepancies between the LR and PLR models. Specifically, the metrics of accuracy, F1-score, precision, and recall exhibited minor variations in these stages. The LR and PLR models demonstrated higher or lower performance in the various stages. These differences are illustrated in the form of bold values—the better-performing model being denoted by bold text. Notably, the LR model was found to outperform the PLR model in terms of F1-score in stages 2 and 5, and the PLR model demonstrated superior precision in stages 3 and 5.
Table 4. Probabilistic logistic regression (PLR) and standard logistic regression (LR) model’s performance.

<table>
<thead>
<tr>
<th>Stages</th>
<th>Weeks</th>
<th>Accuracy</th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LR</td>
<td>PLR</td>
<td>LR</td>
<td>PLR</td>
</tr>
<tr>
<td>Stage 1</td>
<td>Week 2</td>
<td>62.98</td>
<td>62.98</td>
<td>60.73</td>
<td>60.73</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Week 4</td>
<td>67.43</td>
<td>67.30</td>
<td>66.67</td>
<td>66.49</td>
</tr>
<tr>
<td>Stage 3</td>
<td>Week 6</td>
<td>92.60</td>
<td>92.81</td>
<td>92.63</td>
<td>92.83</td>
</tr>
<tr>
<td>Stage 4</td>
<td>Week 7</td>
<td>82.77</td>
<td>82.77</td>
<td>83.83</td>
<td>83.83</td>
</tr>
<tr>
<td>Stage 5</td>
<td>Week 10</td>
<td>78.29</td>
<td>78.29</td>
<td>79.79</td>
<td>79.75</td>
</tr>
</tbody>
</table>

Numbers are in bold to denote instances where one model outperformed the other.

Moreover, a significant increase in model performance was observed from stages 3 to 5. The highest accuracy was achieved at stage 3, in the sixth week of the 2020 second-semester academic calendar. This finding responds to the study’s third objective, which sought to determine the optimal week in which to predict at-risk students. The observed increases and decreases in the model’s performance mean that more or less predictive power was obtained over time, thereby suggesting an optimal week for predicting at-risk students.

A 94% HDI for the model’s parameters was observed throughout the five stages, as shown in Table 5. The $(x_1, x_2)$ represents the 94% HDI values, where $x_1$ is the lower HDI value and $x_2$ is the upper HDI value, and $\bar{x}$ is the mean value of $x_1$ and $x_2$. $|x_d|$ is the absolute difference between $x_2$ and $x_1$. The model is more certain about a parameter if the $|x_d|$ is close to 0.

Table 5. Posterior plot values: 94% HDI values for the model’s parameters.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
<th>Stage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>0.47 (0.58)</td>
<td>0.54 (0.60)</td>
<td>1.10 (1.12)</td>
<td>1.20 (0.74)</td>
<td>0.62 (0.68)</td>
</tr>
<tr>
<td>TIME ON COURSES</td>
<td>0.70 (0.35)</td>
<td>0.95 (0.36)</td>
<td>3.5 (1.30)</td>
<td>0.80 (0.59)</td>
<td>0.71 (0.48)</td>
</tr>
<tr>
<td>AS 1</td>
<td>1.00 (0.60)</td>
<td>1.00 (0.60)</td>
<td>0.63 (0.41)</td>
<td>0.71 (0.39)</td>
<td></td>
</tr>
<tr>
<td>TM 1</td>
<td></td>
<td></td>
<td>0.88 (0.47)</td>
<td></td>
<td>0.17 (0.50)</td>
</tr>
<tr>
<td>TM 2</td>
<td></td>
<td></td>
<td></td>
<td>1.1 (0.52)</td>
<td></td>
</tr>
</tbody>
</table>

Throughout the five PML models, the uncertainty of TITLE and TIME ON COURSES increased after including student performance variables. As the semester progressed, the PML models demonstrated more certainty for assessment 1 (AS_1). Furthermore, the PML models demonstrated 43%, 80%, and 56% decreases in model uncertainty for TITLE, TIME ON COURSES, and AS_1, respectively, from stage 3 to stage 5. This is an implication of great model reliability from stage 3 to 5, as model predictions can be performed with greater certainty.

To showcase the capabilities of a probabilistic machine learning (PML) model, a student’s at-risk status was predicted for all five stages using a probabilistic logistic regression (PLR) model. The PLR model operates by making predictions through sampling from the posterior distribution. In details, 1000 samples used to make predictions with a 95% credible interval. The 95% credible interval indicates the range of values for which the PML model predicts the student’s risk status with 95% certainty. The results of these predictions are presented in Table 6. The student was predicted as being at-risk throughout stages 1 to 4, and higher levels of certainty were reported for stages 3 and 4, where the differences between the upper and lower limits were minimal. The student was identified as not being at-risk at stage 5. The 61% probability exceeded the 50% threshold; however, this prediction still indicates that there is a possibility that the student could be at-risk. This prediction demonstrates how the PLR model can incorporate the possibility of a student being either at-risk or not-at-risk, rather than simply providing a point estimate of 61%, as a standard logistic regression (LR) model would do without a credible interval. Therefore, the PLR...
model provides a more comprehensive view of the prediction’s uncertainty, making it a valuable tool in decision-making scenarios.

Table 6. At-risk student prediction with a probabilistic machine learning model.

<table>
<thead>
<tr>
<th>Stages</th>
<th>Probability Not At-Risk</th>
<th>95% Credibility Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Low Limit</td>
</tr>
<tr>
<td>Stage 1</td>
<td>38%</td>
<td>33%</td>
</tr>
<tr>
<td>State 2</td>
<td>24%</td>
<td>20%</td>
</tr>
<tr>
<td>State 3</td>
<td>2%</td>
<td>1%</td>
</tr>
<tr>
<td>State 4</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>State 5</td>
<td>61%</td>
<td>42%</td>
</tr>
</tbody>
</table>

5. Discussion

This paper has presented a multistage probabilistic machine learning (PML) model designed to identify at-risk students for early intervention. This section discusses the accomplishment of the objectives outlined in Section 1.

Regarding the first objective, the study employed Pearson’s correlation and an extra tree classifier feature importance score to identify effective predictors for identifying at-risk students throughout the semester. Results showed that TIME ON COURSES, TITLE (gender), ASSESSMENT 1, TEST 1, and TEST 2 were significant predictors. This finding aligns with the most predictive and useful features found in previous studies under the categories of demographics, student engagement, and performance [5,9,25]. The study also identified TIME ON COURSES as a useful feature for predictive models in learning analytics, in contrast to dynamic student engagement features emphasized by other studies [5,8].

In relation to the second objective, the study proposed a probabilistic logistic regression (PLR) model that demonstrated comparable accuracy results to a regular logistic regression (LR) model, along with the added benefit of providing high-density interval (HDI) values and credible intervals to quantify uncertainty in model parameters and predictions. In contrast, regular LR models only provide point-estimate predictions and model parameters, without accounting for the possibility of a student being both at-risk and not-at-risk due to the ambiguities present in learning analytics data.

The third objective was achieved by identifying the sixth week of the academic calendar as the optimal time to identify at-risk students. This finding is consistent with the optimal week found in previous studies [6,8,17].

6. Conclusions

This study presented the development of a probabilistic machine learning (PML) model to identify at-risk students throughout multiple stages based on their demographics, engagement, and performance data. Such identification of at-risk students in different stages can enable instructors to intervene and support students at the optimal times to prevent academic failure.

The present study aimed to develop a logistic regression model design within the framework of probabilistic machine learning for identifying at-risk students. However, it is recommended that a comparative investigation be conducted in the future to assess different probabilistic machine learning model designs, with the goal of identifying the optimal model for this specific task. Additionally, while the clustering of student data is a well-studied topic, prospective research endeavors may compare the performances of probabilistic clustering methods with those of traditional clustering techniques. Moreover, it is suggested that future studies explore the integration of data assimilation techniques into probabilistic machine learning models, in the context of identifying at-risk students. Such a combination is hypothesized to yield more accurate and reliable estimates, particularly in scenarios where there exists a significant degree of uncertainty or incomplete information.

From a macro-level perspective, this study may serve as a reference point for future studies that aim to adopt probabilistic machine learning approaches for modeling student
data to solve different learning analytics problems that are targeted at improving student and university success.

**Author Contributions:** Conceptualization, E.N. and M.M.; methodology, E.N.; software, E.N.; validation, E.N., M.M. and C.C.; formal analysis, E.N.; investigation, E.N.; resources, E.N., M.M. and C.C.; data curation, E.N.; writing—original draft preparation, E.N.; writing—review and editing, E.N., M.M. and C.C.; visualization, E.N.; supervision, M.M. and C.C.; project administration, M.M. and C.C.; funding acquisition, M.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable for studies not involving humans or animals.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing is not applicable to this article.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**


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