GLOBAL ACADEMIC RESEARCH INSTITUTE

COLOMBO, SRI LANKA

GARI International Journal of Multidisciplinary Research

ISSN 2659-2193

Volume: 09 | Issue: 03

On 30th September 2023

http://www.research.lk

Author: Perera DPM, Dinesh Asanka

University of Sri Jayewardenepura, University of Kelaniya, Sri Lanka

GARI Publisher | LMS | Volume: 09 | Issue: 03

Article ID: IN/GARI/ICAS/2023/107AUG | Pages: 150-167 (18)

ISSN 2659-2193 | Edit: GARI Editorial Team

Received: 02.08.2023 | Publish: 30.09.2023

EXPLORING LMS INTERACTION DATA AND PREDICTION OF STUDENTS' GRADES USING STANDARD SUPERVISED CLASSIFICATION ALGORITHMS

[1]Perera DPM, [2]Dinesh Asanka

[1]Department of Information and Communication Technology, Faculty of Humanities and Social Sciences, University of Sri Jayewardenepura, [2]Department of Industrial Management, Faculty of Science, University of Kelaniya, Sri Lanka

ABSTRACT

In the context of higher education financing shifting its dependency from student enrollment to degree completion, the necessity for strategies that elevate educational quality and amplify retention rates has become critical. This study is anchored in the application of predictive modeling techniques to aid in the timely identification of students who may struggle academically, aiming to enhance course retention. Utilizing demographic and behavioral data harvested from Learning Management Systems (LMS), prediction models have been built. These models are specifically engineered to forecast the final grades of students, thus providing a mechanism for the early identification of students at risk of underperformance. Through the employment of preprocessing techniques and regression methods rooted in machine learning, these predictive models are designed to operate at the level of the individual student. The analysis conducted in this study presents compelling evidence suggesting that the degree of a student's engagement with the LMS accurately mirrors their final grades. This research underscores the quantitative value of predictive modeling technology in an educational context and offers an approach for conducting empirical studies of predictive modeling utilizing educational data. Data spanning the years 2020 and 2021 was methodically analyzed to devise early prediction models for students' grades. This analysis involved a comprehensive examination of LMS log files and tracking of students; activity counts. The models were derived from data collected over three academic semesters at the University of Sri Jayewardenepura, Faculty of Humanities and Social Sciences. Eight distinct regression models have been formulated for all students, with algorithms such as Multiple Linear Regression, Random Forest, and K-Nearest Neighbor being thoroughly tested both with and without the use of Principal Component Analysis. The Random Forest algorithm, when applied without Principal Component Analysis, displayed superior performance with the highest R² value (0.3506), indicating its reliability in predicting student outcomes.

Keywords: Student Grade Prediction, Academic Performance, Machine Learning, Linear Regression, Random Forest Classification.

INTRODUCTION

Data has become a pivotal element in numerous sectors, from education and healthcare to engineering and marketing. The rapid growth of data and its ensuing analysis allows departments to discover hidden patterns and valuable facts. Today, data analysis spans diverse fields such as user behavior on websites, patient record tracking, and more. As a crucial asset, data is paramount in delivering insights that can grant businesses a competitive edge.

In healthcare, for instance, data can enable early disease diagnosis, while in finance, it can help predict an organization's future growth (Hooshyar et al., 2019). The field of data science has given rise to various methods for processing and analyzing data. Educational data mining (EDM) has become significant, focusing on extracting meaningful patterns or knowledge from data. EDM involves the thorough exploration of large data sets to uncover correlations and reveal data facts. It's a novel field that uses data mining techniques to analyze hidden knowledge from educational data. The main objective here is to predict student performance based on a large educational database collected from education repositories, web-based education, and surveys (Macfadyen and Dawson, 2010). EDM techniques are applied to mine and analyze vast amounts of student data, resulting in system improvements. These techniques also offer insights into the most influential factors in a student's performance, with the ultimate goal of improving education quality. This enhancement is achieved through predictive models to anticipate a student's performance, particularly for students at risk of dropping out (Al-Musharraf and Alkhattabi, 2016). As the primary stakeholders, students contribute significantly to a country's economic and social development. Digital data, encompassing students' personal details and academic achievements, comprises the majority of universities' data (Kasthuriarachchi and Liyanage, 2017).

Research shows that most students depart educational institutions within the first two years, posing significant concern for both institutions and students. Predicting students' performance quickly is vital for providing early feedback and implementing immediate actions to enhance their academic journey (Khasanah and Harwati, 2017). Scientific research has focused on developing techniques to boost student retention in higher education institutions. The goal is to identify students at risk of dropping out early due to learning difficulties, financial issues, low self-confidence, health concerns, social activities, career responsibilities, and more (Siddique et al., 2021). Data mining has gained substantial attention due to its ability to unveil hidden patterns in extensive educational databases, greatly impacting decision-making across a wide range of industries. This prominence has spearheaded a new era for Artificial Intelligence, Machine Learning, Statistics, and Complex Computation. By exploring various types of data from educational settings, EDM can help understand students and their learning environments, making it a pivotal tool for educational development (Khasanah and Harwati, 2017). This research aims to employ statistical and machine learning techniques to predict students' grades. The dataset includes students' demographic information and academic records from the Faculty of Humanities and Social Sciences, University of Sri Jayewardenepura. The research focuses on the effectiveness of linear regression classifiers and random forest classifiers in predicting student academic success.

Problem Statement and Research Question

Data mining has emerged as a powerful tool in education, enabling examination of various student-generated data types, and unveiling hidden patterns that facilitate learning improvements. Researchers (Hooshyar et al., 2019; 서울교육대학교 컴퓨터교육과 et al., 2021) have successfully deployed data mining methodologies to predict, model, and comprehend frequent patterns of student behavior. However, exploiting Learning Management System (LMS) data in Sri Lanka requires regulatory clearance or direct intervention, posing a significant challenge. Previous research (Macfadyen and Dawson, 2010) has indicated that integrating LMS with data mining techniques can improve students' university performance, emphasizing time as a significant determinant of a student's course performance. This thesis, therefore, aims to evaluate whether a student's interaction with the LMS can be employed to forecast their academic grades. In recent years, academic research has increasingly focused on the participation and completion rates of students pursuing Science, Technology, Engineering, and Mathematics (STEM) disciplines in the United States. High failure rates in these courses and their impact on the country's global market competitiveness underscore this focus (Hellas et al., 2018; Siddique et al., 2021). LMS is a valuable organizational infrastructure that facilitates attendance tracking, assignment grading, and dissemination of crucial information to students (Siddique et al., 2021). Research by Gabrian et al. (2017) asserts that interactivity with an LMS is a significant element of online learning.

Educational Data Mining (EDM) involves the examination of data generated in educational settings, including LMS and traditional classroom interactions, to gain insights into student academic performance. EDM is a burgeoning field that explores, develops, and evaluates methods to enhance education quality (Goriparti et al., 2014). This thesis employs a supervised approach to predict student grades based on LMS interactions, focusing primarily on LMS activity data, such as button clicks. Researchers Dutt and Ismail (2019) successfully utilized various classifiers on the LMS dataset to yield results. The present study builds on the fundamental work by Dahiwal and Joshi (n.d.) on the predictive capabilities of LMS data on online lectures. The core objective of this research is to predict students' grades and identify academically at-risk individuals early, using a high-quality dataset derived from the Faculty of Humanities and Social Sciences at the University of Sri Jayewardenepura. Previous research by Tran and Sato (2017) developed malware detection classifiers using Application Programming Interface (API) sequences and Natural Language Processing techniques, proving the concept of using natural language processing in the field of natural language classification. This methodology for decoding sequence meaning can be applied to domains producing sequences, such as education.

The problem statement for the current study is: To what extent can a classifier predict a student's academic performance during the university period based on their interactions with a Learning Management System (LMS), and their demographic and cognitive data?

This study aims to answer the following research questions:

- RQ1: What is the earliest possible time frame within an academic semester to develop robust classifiers capable of reliably detecting students at risk of failure?
- RQ2: Can process mining features and generic machine learning features be effectively integrated, and to what extent can the newly enhanced features improve classification accuracy?
- RQ3: Can distinct subgroups of courses based on the usage of LMS technologies be identified, which could be utilized to construct a generalized portable model for similar courses in the future?

LITERATURE REVIEW

Learning analytics refers to the process of gathering, analyzing, and reporting data about learners and their environments to comprehend and enhance learning experiences and their settings (Pardo and

Siemens, 2014). A substantial focus of learning analytics research involves utilizing Learning Management System (LMS) data to predict student success by identifying those at risk of failing a course or forecasting their grades (Loewen et al., 2014; Gamo et al., 2017; Steen et al., 1988). This aligns with learning analytics' progressive focus on individualized feedback and interventions. For example, Purdue University developed a Course Management System in 2009 to inform students about their grades and propose action plans based on their academic performance across two semesters (Nguyen et al., 2020). Such a system can reveal insights about the impact of new teaching methods, learning activity fluctuations, and other diverse phenomena. Macfadyen and Dawson (2010) argued that increasing class sizes, sluggish engagement rates, student absenteeism, and challenges in assisting students needing help have complicated instructors' ability to assess student success. Consequently, there is a pressing need for improved tools or strategies to identify at-risk students and provide necessary support. Earlier work (O. and P., 2017) suggested that student online behavior data in an LMS might serve as an early indicator of academic performance. Applying academic analytics to institutional LMS data can enhance understanding of student progress and help identify those at risk of failure (Ferguson, 2012). Incorporating additional variables, such as LMS login information, can significantly boost the predictive accuracy of models originally based solely on student SAT scores.This thesis will explore two strategies for leveraging LMS data to create an early warning system for identifying at-risk students. Approach A will handle LMS activity data in a generic way, while Approach B will focus on a novel method of processing LMS data to extract additional insights.

Students play a pivotal role in any educational institution, contributing significantly to socio-economic advancement through innovation and entrepreneurship. The growing prevalence of learning management systems (LMS) provides extensive data on students' academic performance, allowing for real-time interventions to enhance outcomes rather than relying solely on past performance. This research can support teachers, students, and educational institutions by predicting student success or failure, thereby enhancing overall educational outcomes. Low academic performance remains a substantial challenge within educational systems, with various factors contributing to this. Identifying students at risk at an early stage is critical to enable timely interventions that enhance their future performance. Recent years have seen a growing popularity of predictive methodologies for real-time identification of struggling students to boost retention rates. These models use demographic and behavioral data from students to forecast their performance, thus enabling proactive support for identified at-risk students. Emerging technologies in education are revolutionizing the learning environment. Traditional classrooms are giving way to virtual classrooms due to e-learning or web-based education systems. Learning management systems, an offshoot of e-learning, have become integral to higher education. They provide features that facilitate quizzes/tests, schedule learning activities, enable smooth communication, and offer real-time feedback on student performance. The importance of LMS tools is recognized by both teachers and students alike, as indicated by high percentages in an EDUCAUSE report (Al-Musharraf and Alkhattabi, 2016).

Apart from supporting e-learning, these tools also collect vast digital data about student behavior on LMS. The analysis of this data provides invaluable insights into student performance and learning processes, enabling more effective administrative decisions. Educational Data Mining (EDM) and Learning Analytics (LA) are two fields exploring the potential of historical data for enhancing educational quality (Algarni, 2016). The various applications of EDM include student modeling, classroom behavior modeling, student performance modeling, evaluation, and student and teacher support. It also contributes to the development of tools to facilitate knowledge discovery. The use of these tools enhances the learning experience, addresses course-related problems, and assists in developing cognitive skills for problem-solving. Educational data mining, particularly predicting student performance, is a critical research topic due to its relevance to student retention in higher education. The OECD reports that only 12% of full-time undergraduate students progress to their second year of study, with numbers rising to 20% at the end of a theoretical project period and 24% after three years. The dropout rate after the first year is also significant, varying from 6% in the U.S. to 20% in Slovenia and Belgium's French-speaking regions. These attrition rates pose financial risks for universities. Since the 1970s, access to universities has expanded, but the available spots still fall short of the number of qualified students. In Sri Lanka, for example, only 15.5% of qualified students secured places in universities for the 2010/11 academic year. This under-enrollment problem needs urgent attention for the nation to maximize its developmental potential. Remaining qualified students are left to explore other options such as specialized institutions, private institutions, or studying abroad, most of which incur significant costs. Increased enrollment has been noticed primarily in science courses, including newer areas like computing, food science, and paramedical studies. Private universities have also expanded, offering courses in various disciplines from Business Administration to Sociology. Students who fail to complete a bachelor's degree in public universities often pursue studies abroad. Higher education institutions' funding is based on graduation rates, which puts pressure on them to increase enrollment and retention rates and improve educational quality. Predicting student course grades early and implementing early intervention programs can enhance retention. This can be done through supervised learning approaches, using students' demographic data, and interaction data from educational tools.

A systematic literature review was conducted to understand the commonly used data types for predicting student grades, the machine learning techniques that can help accurately predict student performance, and the grading system used for evaluating outcomes. The search involved related journals such as Science Direct, IEEE Xplore, Google Scholar, and the Journal of Learning Analytics, using keywords related to predictive models, machine learning, and performance predictions, among others. After thorough screening and analysis, several papers revealed feature patterns that can be used for prediction, classifying the features into academic, demographic, preschool, and virtual learning environment-related features. The review also documented the types of machine learning methods used and the methods of evaluation.

METHODOLOGY AND IMPLEMENTATION

This paper employs data analytics methodologies in the pedagogical realm to enhance comprehension of student progression through a course and to prognosticate potential course failures or attrition. Both Learning Management Systems (LMS) and Student Management

Systems (SMS) were utilized as data reservoirs. This section of the paper encapsulates the research methodology implemented in this study.

The primary objectives of this investigation are to:

- Examine the data to identify the most influential variables for achieving accurate predictions.
- Investigate and evaluate the effectiveness of different machine learning and data mining techniques in education to identify the most accurate algorithms for constructing predictive models.
- Collect and assess student activity logs from the Learning Management System (LMS) to analyze their correlation with academic success at university.

This section delineates the universal methodologies employed in the study's experiments, including an overview of datasets, type of features, and their extraction process. Moreover, it outlines the training, evaluation procedures, and the operational context and hardware utilized. Detailed expositions of the procedures employed in particular studies are housed within the pertinent sections.

Figure 1 delineates a series of procedures initiating from raw data and culminating in the prediction of a student's academic performance. The first step involves data preparation, which entails data collection from diverse sources, followed by transformation into a requisite format to form a feature matrix for subsequent training stages. Each stage of this method, beginning with data preparation and feature extraction, is discussed in detail in the following section. In the context of Educational Data Mining, individual student performance expectancy is perceived as the extent to which a student believes that utilizing an LMS will result in academic performance improvement. We hypothesize a positive correlation between these factors and students' grades in this study.

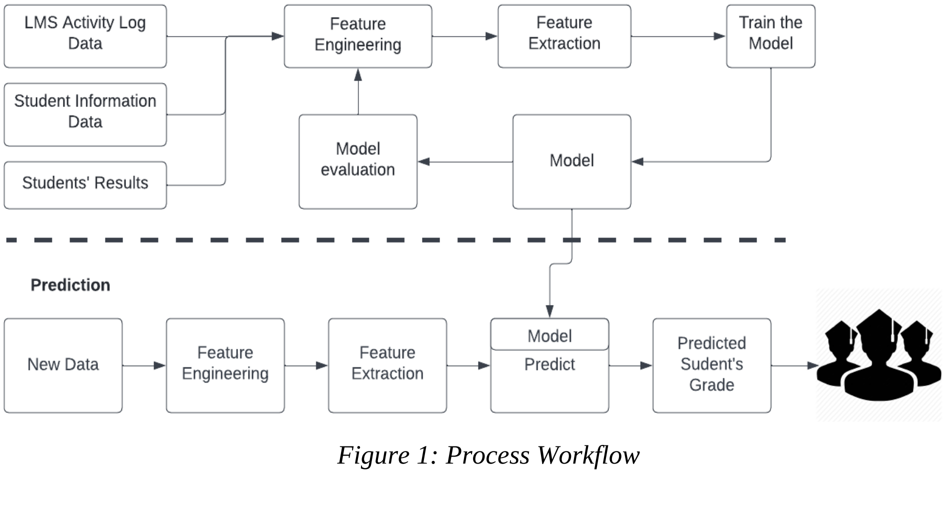

Figure 1: Process Workflow

We state the null (H0) and alternative (H1) hypotheses for each factor as follows:

H0 : Assignment factor has no relationship with Student's grade

H1 : Assignment factor has no relationship with Student's grade

H0 : Chat factor has no relationship with Student's grade

H1 : Chat factor has no relationship with Student's grade

H0 : Feedback factor has no relationship with Student's grade

H1 : Feedback factor has no relationship with Student's grade

H0 : Forum factor has no relationship with Student's grade

H1 : Forum factor has no relationship with Student's grade

H0 : Label factor has no relationship with Student's grade

H1 : Label factor has no relationship with Student's grade

H0 : Lesson factor has no relationship with Student's grade

H1 : Lesson factor has no relationship with Student's grade

H0 : Quiz factor has no relationship with Student's grade

H1 : Quiz factor has no relationship with Student's grade

H0 : Survey factor has no relationship with Student's grade

H1 : Survey factor has no relationship with Student's grade

H0 : Workshop factor has no relationship with Student's grade

H1 : Workshop factor has no relationship with Student's grade

The proposed hypotheses are embodied in the conceptual research model depicted below.

Data for this study was collated from the Faculty of Humanities and Social Sciences at the University of Sri Jayewardenepura. The preliminary data sources included the Learning Management System (LMS), Student Management System (EMS), and Student Grading System (SGS). Given that the structure and formatting of LMS data were not readily amenable to data mining, the initial stage comprised data extraction, followed by reformatting into a suitable structure. Consequently, the data transformation involved pre-processing steps such as data cleaning, outlier detection (Cousineau and Chartier, 2010), handling missing values (Kwak and Kim, 2017), and integration of data from various sources.

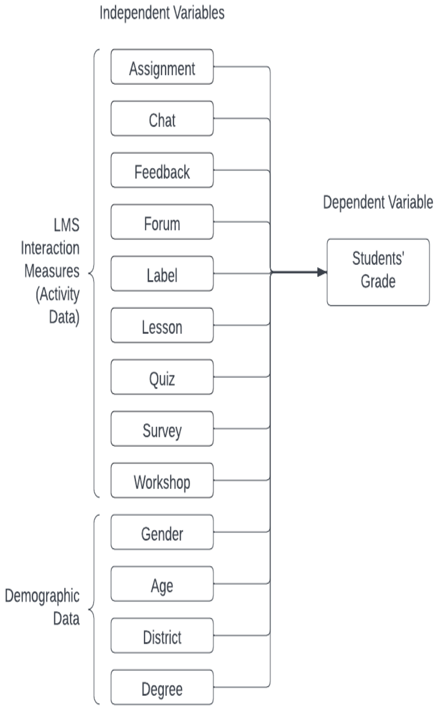

Figure 2: Conceptual Framework

Comprehensive course-related information was procured from three databases: Student Management Systems, Learning Management Systems, and Student Grading Systems. The SGS data encompassed final grades, while the LMS data incorporated the total number of enrolled students, assessment scores, etc. In light of the Covid-19 pandemic, all lectures were conducted online, with all relevant course information disseminated via the LMS. The Faculty of Humanities and Social Sciences at the University of Sri Jayewardenepura utilizes Moodle as their LMS. The study included all undergraduate courses that had at least one student registration. The data considered for this study spans three academic semesters (2020 1st Semester, 2020 2nd Semester, and 2021 1st Semester), inclusive of 2919 students' final results from 530 courses offered by the faculty. However, after aligning with LMS data, the sample size was reduced to 1032 students. All selected courses originated from 16 academic departments of the Faculty of Humanities and Social Sciences at the University of Sri Jayewardenepura.

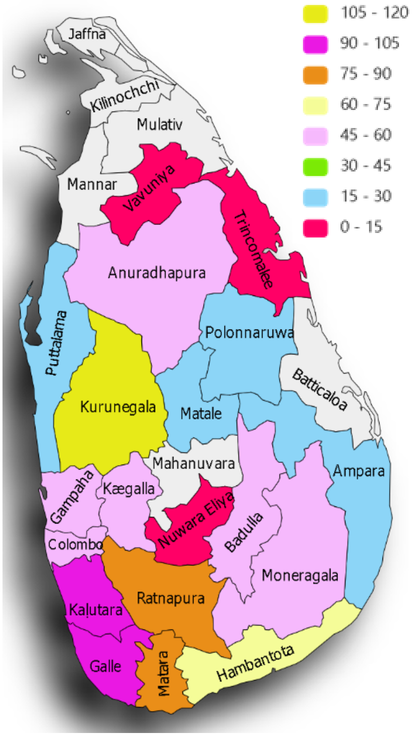

Figure 3: Scattering of Students all over Sri Lanka

Figure 3 portrays the geographic distribution of the 1032 undergraduate participants across the country. (87.69%) of the participants were female, while males constituted (12.31%) which is a proportion reflective of the gender distribution in Humanities and Social Sciences faculties within Sri Lankan universities. The age of the participants ranged from 20 to 30, with the majority of (59.69%) falling within the 21-22 age bracket. The next largest age group, comprising (26.45%) of the sample, was aged 23-24. In terms of degree classification, special degree students made up the largest percentage (56.10%), followed by general degree students (15.6%).

The Learning Management System (LMS) maintains logs of user activities drawn from diverse tools within the system. These logs capture real-time engagements, including mouse clicks, content views, and keyboard strokes, thus providing data on time spent on specific resources such as quizzes, assignments, forums, etc. Moreover, the logs detail content view counts and the frequency of various related activities conducted within the system. These logs, organized by course, participant, and time, provide invaluable temporal information and insights into each module's activity level. These LMS tracking variables serve as independent factors in this study.

Variable	Description	Type
assign_grades	The assignment grading	Discrete
assign_submission_activity_count	The number of assignment submission	Discrete
chat_messages_activity_count	Number of chat messages send	Discrete
comments_activity_count	Number of comments created on LMS	Discrete
forum_created_activity_count	Number of forums created	Discrete
forum_subscription_activity_count	Number of forum subscriptions made	Discrete
forum_posts_activity_count	Number of forum posts created	Discrete
forum_discussion_activity_count	Number of forum discussion created	Discrete
notification_received_activity_count	Number of notifications received	Discrete
messages_activity_count	Number of messages read	Discrete
question_created_activity_count	Number of questions asked via LMS	Discrete
quiz_attempts_activity_count	Number of quizzes attempted	Discrete
quiz_grades	The quiz grading	Discrete

Table 1: List of few independent variables collected from LMS log data

Table 1 presents the tracking variables derived from the LMS. Additional variables were also generated from the initial dataset. This study incorporated data from the Student Management System (SMS) and the Student Grading System (SGS) alongside the LMS variables. The SMS provided detailed student profiles, including demographic information, gender, degree type, age, etc. Grades for assignments, quizzes, projects, and final exams were extracted from the SGS, and all assessments were compiled into a single variable along with their corresponding final score scales.

A predictive model, built using a combination of characteristics from the LMS, SGS, and SMS, was utilized to accurately predict students' academic progress in a course. The course's final grade, established as the dependent variable, was employed to measure academic progress. Final grades, represented in letter format (A+, A, A-, B+, etc.), were extracted from the SGS database, as was the final course score. The datasets were cleaned to exclude records without corresponding final course grades. This included instances such as guest accounts created by faculty, students who did not complete or withdrew from courses, and duplicate values for demographic variables. All the variables from the learning management system were standardized to Z scores to achieve a normal distribution. Outliers were detected using statistical methods like boxplot and Z-score. Data preprocessing was essential in this study. It involved cleaning and structuring the data. Categorical data that cannot be processed directly by machine learning algorithms were transformed using one-hot encoding, a technique that transforms each category value into a binary vector. Principal Component Analysis (PCA) was used to reduce the dimensionality of the dataset while retaining as much information as possible. This process started by standardizing the data, then calculating the covariance matrix to identify relationships between variables. Finally, eigenvectors and eigenvalues of the covariance matrix were computed to identify principal components, which are new variables formed by linear combinations of the original variables. The process of dimension reduction can retain significant information by keeping only the components with the most information, treating these as new variables. These new variables, or principal components, are linear combinations of the original variables, representing the maximum data variance. These components capture most of the data's information. The larger the variance, the more information the data points contain. These principal components provide an optimized view for understanding differences in data observations.

.

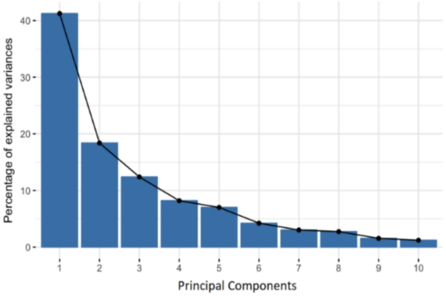

Figure 4: Scxree Plot

Forward selection starts with an empty model and adds the most significant predictors one by one. This process continues until a predefined stopping rule is met, or until all predictors have been evaluated. Determining the most significant variables involves considering criteria such as minimum p-value, maximum increase in R2, and maximum reduction of model's residual sum of squares (RSS). Forward selection halts when all remaining variables have p-values greater than a set threshold. The

The important components can be identified by ordering eigenvectors and eigenvalues. Less significant components (with low eigenvalues) are discarded, resulting in an eigenvector. No changes are made to the data in this step. The last step is to convert the original data set from its original axis to the principal component axis. This is done by multiplying the transpose of the final data set and the feature vector by the transpose of the standardized original data set. Stepwise regression is an automatic model selection process in regression analysis. It either starts with a model with no predictors (forward selection) or with all predictors (backward selection). At each stage, the model searches for the predictors that most improve the model

resulting model contains only variables with p-values below this threshold.

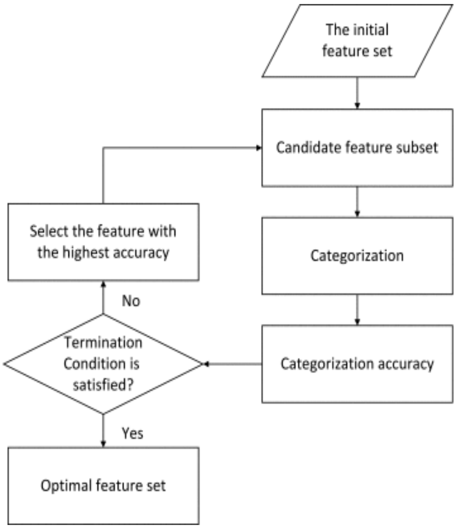

Figure 5: Stepwise Regression Analysis - Forward

Backward selection, or backward elimination, is a process of variable selection that starts with a comprehensive model containing all considered variables. The process then gradually removes the least significant variables. This process continues until a predefined stop condition is met or until there are no variables left in the model. The least significant variable at each step is identified by the highest p-value, and its removal has the least impact on the model's R2 value and the smallest increase in the Residual Sum of Squares (RSS). The stopping condition is when all remaining variables in the model have a p-value less than a previously set threshold. Once this condition is met, the backward elimination stops and returns the model at its current state.

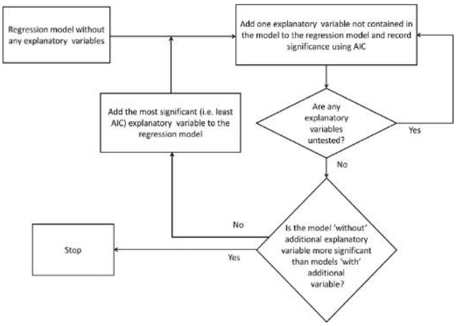
Regression model without any explanatory variables
Add one explanatory variable not contained in the model to the regression model and record significance using AIC
Add the most significant (i.e. least AIC) explanatory variable to the regression model
Are any explanatory variables untested?
Yes
No
No
Is the model 'without' additional explanatory variable more significant than models 'with' additional variable?
Yes
Stop

Figure 6: Stepwise Regression Analysis – Backward

In model evaluation, existing data sets are used to train the model and assess its performance on unseen instances. Given that both the training and test data sets used to fit and test the model are substantial and fundamental, the final performance estimates of the model are neither excessively optimistic nor unduly pessimistic. This task is challenging as it requires a representation of a wider problem. The two most common methods for assessing models are train/test splitting and K-fold cross-validation. While generally effective, these methods can produce misleading results when confronted with severe class imbalances. In such cases, a stratified train/test split or stratified K-fold cross-validation, which adjusts the sample according to class label, is necessary. Predictive models for student performance have been developed using several machine learning techniques. The predicted values range from 0 to 4. The prediction process in machine learning methods is one of the most prevalent approaches in the field today. This process is typically divided into two phases: (1) training the model using historical data (the training set) and (2) predicting unseen data (the test set). Training and testing data sets are subsets of hypothetical universal data sets that encompass all possible combinations of real-world data in machine learning. Models are built by learning the characteristics of a universal data set using the training set. The model's performance is then evaluated using the test data set. It's critical for the model's characteristics to apply to both the training and test data sets; otherwise, the model would not be deemed valid. The following regression classification algorithms were employed in this study:

- Linear Regression
- Random Forest
- K nearest neighbors

Learning Management Systems (LMS) are used by universities to deliver course content. However, their utility extends beyond mere content delivery, providing an environment for online interaction and facilitating improved communication between peers and instructors. LMS also helps identify learner behaviors that may predict future performance. However, the reliability of these predictions hinges on the dataset used. Models built on a wider range of data are likely to be more precise. In addition to using LMS data from various courses and combining it with allocation scores, this study also uses machine learning-based classification algorithms like Random Forest, K-Nearest Neighbor, and multiple regression for analysis. While this method can predict students' grades, it can also integrate more features directly into the learning process. Next, the regression model is trained to predict the performance of the chosen algorithm and recommend the best classification algorithm based on a hidden dataset. The performance of the regression models, which include Multiple regression, Random Forest, and K-Nearest Neighbor, is evaluated using the Spearman's Rank Correlation Test. The Random Forest model has proven to be superior in this context.

Residual Plot - Pre-Model Review

Linear regression aims to find a line that minimizes the difference between predicted and actual values. Residuals, or errors, represent these differences. A residual plot should be analyzed before assessing the model matrix such as R-squared. It shows any potential bias in the model. If the residuals are randomly distributed with no systematic pattern, the model is likely performing well.

$$Residual = Actual\ Value - Predictted\ Value$$

$$e = y - \hat{y}$$

Evaluation Index for Linear Regression Models

Evaluation metrics offer a measure of a model's performance. Metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared, adjusted R-squared, and Root Mean Squared Error (RMSE) are evaluated.

Mean Squared Error (MSE)

This metric, commonly used for regression tasks, calculates the average of the squares of the differences between predicted and actual values. MSE penalizes larger errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Mean Absolute Error (MAE)

This calculates the average absolute difference between target and predicted values. It doesn't penalize large errors as much as MSE does.

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|}$$

R-squared or Coefficient of Determination

This scale represents the variation in the dependent variable as defined by the model's independent variables. It measures how well the model relates to the dependent variable.

Regression Error Calculation

Regression error is the residual difference between the original and predicted values.

Residual Error Calculation

Residual error, or the sum of squares of the differences between each point and the mean of the Y values, is referred to as Total Squares (TSS).

Coefficient of Determination with RSS & TSS

This measures how much of the total change in Y can be explained by the independent variable X.

Root Mean Squared Error (RMSE)

This calculates the square root of the mean of the squared differences between predicted and actual values. RMSE penalizes large errors.

Experimental Results

Model	RMSE	MAE	R2	isPCAUsed
Regression Analysis (With PCA)	0.431	0.311	0.347	Y
Regression Analysis (Without PCA)	0.435	0.361	0.314	N
Regression Stepwise Forward Selection	0.419	0.298	0.337	N
Regression Stepwise Backward Selection	0.428	0.337	0.305	N
Random Forest (With PCA)	0.433	0.293	0.228	Y
Random Forest (Without PCA)	0.395	0.269	0.351	N
kNN (With PCA)	0.467	0.329	0.148	Y
kNN (Without PCA)	0.486	0.339	0.075	N

Table 15. Comparison of Model Performance

Eight distinct regression models were constructed and evaluated in this study. The aim of this task was to determine which models, coupled with varying feature engineering methods, could offer superior performance in predicting student GPAs. These models were elaborated in detail in the preceding section. The four regression algorithms were utilized with all original features and with principal components derived via PCA. PCA was used to decrease the number of dimensions and mitigate the risk of overfitting the models. The problem of "Curse of Dimensionality" becomes especially pronounced when clustering techniques are used to manage high-dimensional data (Zhou et al., 2019). Hence, in one approach of building the regression models, PCA was employed to reduce the feature set. The models were trained using a dataset consisting of randomly selected 75% of the student data. The remaining 25% was used as a test dataset to evaluate each model's performance. This method aims to prevent overfitting. If a model is overfitted, performance measures based on it can lead to distorted results. The performance can significantly alter when the same model is applied to new data.

Performance measurement utilized three metrics: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and R2. These metrics were used to determine the accuracy of the predictions and the extent of deviation from the actual GPA values. RMSE and MAE are error-based measurements, with errors being the difference between predicted and actual GPAs {cite with the paper I sent}. Table 1 presents the obtained values for these measurements. Based on these, the kNN model with PCA features yielded the highest RMSE (0.4856069), while the Random Forest model without PCA had the lowest RMSE (0.394651). In terms of MAE, the Random Forest model without PCA recorded the lowest MAE (0.26881), while Regression Analysis without PCA had the highest MAE (0.3610823). Lower RMSE and MAE values indicate a better fit to the dataset and more accurate results than a model with higher values. Models using PCA generally demonstrated inferior performance to those using all features, with the exception of the kNN regressor, which performed better with reduced dimensionality.

In this study, linear regression analysis was performed with all features as well as with a stepwise approach. This was done to compare results between regressions with selected features and all features. The findings show that the stepwise regression with forward selection outperformed both the linear regression with all features and the stepwise regression with backward selection. Even though the stepwise regression demonstrated relatively good performance compared to linear regression with all features.

R2 is another significant measure used to evaluate model performance. Its definition and interpretation were discussed in section xxx. Among all models, the Random Forest without PCA once again excelled, explaining 35% of the variance in the dependent variable that is accounted for by the independent variables collectively. Conversely, the kNN without PCA only accounted for 7% of the variance, the lowest among all models. Although 35% is at the lower end, the finding that Random Forest is better at explaining variance holds significance for future model refinements. Among the linear regression models, stepwise regression with forward selection had a higher R2 value (34%). To summarize, this section primarily compared model performances in terms of their prediction accuracy using RMSE, MAE, and R2 measures. The Random Forest model using all features outperformed all other models in all performance measurements, while the kNN model had the poorest performance among all tested models. Additionally, the linear regression model performed better when using the stepwise selection method, with stepwise regression using forward selection showing the second-best overall performance.

CONCLUSION

Funding for higher education institutions is determined based on the number of degrees conferred, not the number of students enrolled. Hence, the importance of degree completion is emphasized as incomplete or deferred degrees impact financial stability and efficiency. As a result, emphasis is being placed by educational providers on initiatives designed to elevate the quality of education and increase both enrollment and retention rates. The application of predictive models is increasingly being seen as a beneficial tool for improving student retention. These models, which draw from student demographics and behavioral data, offer real-time identification of students struggling academically. Proactive strategies can then be devised by educators to address these students' learning difficulties. Predictive models thus serve as a safety net for students, providing timely interventions that have the potential to enhance academic performance.

However, it is acknowledged that limited understanding exists regarding student characteristics related to Learning Management System (LMS) usage, as highlighted in a study conducted by the National University of Sri Lanka. Absent this investigation, key student attributes such as attitude towards LMS, self-efficacy, experience, computer literacy, and interaction with educators and peers remain unknown. The study revealed that positive attitudes are generally displayed by students when using an LMS, stimulating engagement in e-learning activities and thus enhancing the effectiveness of LMS in a blended learning environment. It is believed by some researchers that the advent of novel technological applications can affect students' attitudes towards technology, and subsequently their usage decisions. Empirical evidence provided by this study supports the idea of a positive perception towards LMS as a significant driver.

In the realm of academic data mining, a vital role is played by predictive models. They assist in identifying early those students who are at risk due to poor academic performance. This allows the timely implementation of interventions such as personalized recommendations, tutoring, extra classes, or even course suggestions. While these predictive performance models are seen as beneficial to educational institutions, students, and parents alike, the necessity of their accuracy cannot be overstated. Erroneous results and suboptimal decisions can occur if methods are inaccurately applied. The successful prediction of student performance requires the correct utilization of data and machine learning algorithms. For each task, a suitable machine learning approach must be selected, though it is recognized that this alone does not guarantee accurate predictions. Feature engineering, the process of transforming data for machine learning, is seen as crucial in improving prediction accuracy. This research aims to examine how method selection and feature engineering can enhance predictive outcomes by comparing the results from three different machine learning approaches using a rating scale.

Challenges and Limitations

This section briefly discusses the challenges and limitations of employing machine learning for students' grades prediction. The uniqueness of individual courses means that results derived from online courses differ from those of offline courses and cannot be generalized. In online courses, the learning management system (LMS) is the sole platform for student interactions and activity participation. One primary obstacle is the scarcity of accessible data sets with valuable interaction information. Despite the abundance of data, the lack of benchmark data sets for algorithm testing and comparative research (Dahlstrom, 2012) remains a concern. There are no specific guidelines on the types of data to be collected, the motives for data collection, or the compatibility of research findings with educational theory. Inconsistencies and varying sample sizes further complicate matters. There is no stipulation for a minimum sample size, implying the results are not generalizable. More features and larger sample sizes could enhance prediction accuracy. Prediction accuracy hinges on data quality and consistency. Using data collected by teachers to evaluate teaching materials can result in inaccuracies affecting data classification. Changes to performance data, like quizzes and assignment grades, during the semester can lead to inaccurate predictions. LMS logs activities and student information, which are currently the independent variables used for grade predictions. However, the LMS is not the only platform for student engagement. As social networking sites are popular among students, it's worth exploring new data collection methods. The use of monitoring software could provide insights into student activity, but the awareness of being observed could alter student behavior. The inconsistency in the data set, due to fewer failing students than passing ones, can also negatively skew predictions. Handling unbalanced data sets requires a high level of expertise. Accurate predictions are crucial, particularly for at-risk students, to implement interventions that enhance performance (Marbouti et al., 2016).

Future Work

Future research efforts will focus on several enhancements. We aim to develop a basic alert system to monitor student progress and provide feedback to educators. We will also examine the effectiveness of various intervention strategies, considering factors such as email reminders, the accuracy of delay predictions, and the optimal timing for interventions. Additionally, we plan to create an automated system for evaluating the content of forum posts, using text analysis algorithms to score and comment on content-based messages. We aim to develop custom reporting tools or dashboard visuals to track student data according to educational goals, accommodating diverse uses of the Learning Management System (LMS) by instructors and administrators. Further research will also be conducted on the relevance of LMS tracking data and network factors to student performance, with the goal of enhancing instructional design and delivery. We will explore crucial course content, concepts, or learning objectives that significantly impact student performance in the curriculum to assist teacher preparation. Given the increasing use of LMS in post-course modules, we will investigate portability issues for both face-to-face and online programs. We also plan to examine how risk indicators vary between international and domestic students, considering factors such as race, economic status, and language barriers.

REFERENCES

Algarni, A., 2016. Data Mining in Education. Int. J. Adv. Comput. Sci. Appl. 7. https://doi.org/10.14569/IJACSA.2016.070659

Al-Musharraf, A., Alkhattabi, M., 2016. An Educational Data Mining Approach to Explore The Effect of Using Interactive Supporting Features in an LMS for Overall Performance Within an Online Learning Environment 13.

Cousineau, D., Chartier, S., 2010. Outliers detection and treatment: a review. Int. J. Psychol. Res. 3, 11.

Dahiwal, M.B., Joshi, P.A., n.d. Implementation of a Movie Recommendation System for Various Online Streaming Services: A Review 5.

Dahlstrom, E., 2012. ECAR Study of Undergraduate Students and Information Technology, 2012 38.

Dutt, A., Ismail, M.A., 2019. Can We Predict Student Learning Performance from LMS Data? A Classification Approach, in: Proceedings of the 3rd International Conference on Current Issues in Education (ICCIE 2018). Presented at the Proceedings of the 3rd International Conference on Current Issues in Education (ICCIE 2018), Atlantis Press, Yogyakarta, Indonesia. https://doi.org/10.2991/iccie-18.2019.5

European Commission. Joint Research Centre., 2016. Research evidence on the use of learning analytics: implications for education Policy. Publications Office, LU.

Ferguson, R., 2012. Learning analytics: drivers, developments and challenges. Int. J. Technol. Enhanc. Learn. 4, 304. https://doi.org/10.1504/IJTEL.2012.051816

Frey, F., 2017. SPSS (Software), in: Matthes, J., Davis, C.S., Potter, R.F. (Eds.), The International Encyclopedia of Communication Research Methods. Wiley, pp. 1–2. https://doi.org/10.1002/9781118901731.iecrm0237

Gabrian, M., Dutt, A.J., Wahl, H.-W., 2017. Subjective Time Perceptions and Aging Well: A Review of Concepts and Empirical Research - A Mini-Review. Gerontology 63, 350–358. https://doi.org/10.1159/000470906

Gamo, N.J., Birknow, M.R., Sullivan, D., Kondo, M.A., Horiuchi, Y., Sakurai, T., Slusher, B.S., Sawa, A., 2017. Valley of death: A proposal to build a "translational bridge" for the next generation. Neurosci. Res. 115, 1–4. https://doi.org/10.1016/j.neures.2016.11.003

Goriparti, S., Miele, E., De Angelis, F., Di Fabrizio, E., Proietti Zaccaria, R., Capiglia, C., 2014. Review on recent progress of nanostructured anode materials for Li-ion batteries. J. Power Sources 257, 421–443.

https://doi.org/10.1016/j.jpowsour.2013.11.103

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., Liao, S.N., 2018. Predicting academic performance: a systematic literature review, in: Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education. Presented at the ITiCSE '18: 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, ACM, Larnaca Cyprus, pp. 175–199. https://doi.org/10.1145/3293881.3295783

Hooshyar, D., Pedaste, M., Yang, Y., 2019. Mining Educational Data to Predict Students' Performance through Procrastination Behavior. Entropy 22, 12. https://doi.org/10.3390/e22010012

Junco, R., 2014. iSpy : seeing what students really do online. Learn. Media Technol. 39, 75–89. https://doi.org/10.1080/17439884.2013.771782

Junco, R., 2012. Too much face and not enough books: The relationship between multiple indices of Facebook use and academic performance. Comput. Hum. Behav. 28, 187–198. https://doi.org/10.1016/j.chb.2011.08.026

Kasthuriarachchi, K.T.S., Liyanage, S.R., 2017. Knowledge Discovery with Data Mining for Predicting Students' Success Factors in Tertiary Education System in Sri Lanka. Sri Lanka 6.

Khasanah, A.U., Harwati, 2017. A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques. IOP Conf. Ser. Mater. Sci. Eng. 215, 012036. https://doi.org/10.1088/1757-899X/215/1/012036

Krotov, V., 2017. A Quick Introduction to R and RStudio. https://doi.org/10.13140/RG.2.2.10401.92009

Kwak, S.K., Kim, J.H., 2017. Statistical data preparation: management of missing values and outliers. Korean J. Anesthesiol. 70, 407. https://doi.org/10.4097/kjae.2017.70.4.407

Loewen, S., Lavolette, E., Spino, L.A., Papi, M., Schmidtke, J., Sterling, S., Wolff, D., 2014. Statistical Literacy Among Applied Linguists and Second Language Acquisition Researchers. TESOL Q. 48, 360–388. https://doi.org/10.1002/tesq.128

Macfadyen, L.P., Dawson, S., 2010. Mining LMS data to develop an "early warning system" for educators: A proof of concept. Comput. Educ. 54, 588–599. https://doi.org/10.1016/j.compedu.2009.09.008

Marbouti, F., Diefes-Dux, H.A., Madhavan, K., 2016. Models for early prediction of at-risk students in a course using standards-based grading. Comput. Educ. 103, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

Nelson, K., Queensland University of Technology, First Year in Higher Education Centre, 2011. Trends in policies, programs and practices in the Australasian first year experience literature 2000-2010. Queensland University of Technology, Brisbane.

Nguyen, A., Gardner, L., Sheridan, D., 2020. Data Analytics in Higher Education: An Integrated View 31, 13.

O., O., P., C., 2017. Predicting Students' Academic Performances – A Learning Analytics Approach using Multiple Linear Regression. Int. J. Comput. Appl. 157, 37–44. https://doi.org/10.5120/ijca2017912671

OECD, 2021. Education at a Glance 2021: OECD Indicators, Education at a Glance. OECD. https://doi.org/10.1787/b35a14e5-en

OECD, 2019. Education at a Glance 2019: OECD Indicators, Education at a

Glance. OECD. https://doi.org/10.1787/f8d7880d-en

Pardo, A., Siemens, G., 2014. Ethical and privacy principles for learning analytics: Ethical and privacy principles. Br. J. Educ. Technol. 45, 438–450. https://doi.org/10.1111/bjet.12152

Reynoso, J.C.S., 2016. Instalación de WampServer 3. https://doi.org/10.13140/RG.2.2.30976.23042

Siddique, A., Jan, A., Majeed, F., Qahmash, A.I., Quadri, N.N., Wahab, M.O.A., 2021. Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers. Appl. Sci. 11, 11845. https://doi.org/10.3390/app112411845

Şimşek, Ö., Atman, N., İnceoğlu, M.M., Arikan, Y.D., 2010. Diagnosis of Learning Styles Based on Active/Reflective Dimension of Felder and Silverman's Learning Style Model in a Learning Management System, in: Taniar, D., Gervasi, O., Murgante, B., Pardede, E., Apduhan, B.O. (Eds.), Computational Science and Its Applications – ICCSA 2010, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 544–555. https://doi.org/10.1007/978-3-642-12165-4_43

Steen, L.A., National Research Council (U.S.), National Research Council (U.S.), Mathematical Association of America, National Academy of Sciences (U.S.), National Academy of Engineering (Eds.), 1988. Calculus for a new century: a pump, not a filter, a national Colloquium, October 28-29, 1987, MAA notes. Mathematical Association of America, Washington, D.C.

Tran, T.K., Sato, H., 2017. NLP-based approaches for malware classification from API sequences, in: 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES). Presented at the 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES), IEEE, Hanoi, pp. 101–105. https://doi.org/10.1109/IESYS.2017.8233569

Villegas-Ch, W., Palacios-Pacheco, X., Luján-Mora, S., 2020. A Business Intelligence Framework for Analyzing Educational Data. Sustainability 12, 5745. https://doi.org/10.3390/su12145745

Worsley, M., n.d. Multimodal Learning Analytics' Past, Present, and, Potential Futures 16.

서울교육대학교 컴퓨터교육과, Kim, K., Chun, S., Koo, D., Shin, S., 2021. A Trend Analysis of Computer Education based on SNS Data through Data Mining Analysis. J. Korean Assoc. Inf. Educ. 25, 289–300. https://doi.org/10.14352/jkaie.2021.25.2.289