EduInsight: Harnessing Data Science and Machine Learning to Predict and Enhance Student Success

Nur Yasmin Yusrina1, a) and Neoh Hoo Thye1,2, b)

*1Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia*

*2Centre for Big Data and Blockchain Technologies, CoE for Advanced Cloud, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia*

a) Corresponding author: 1221302828@student.mmu.edu.my  
b) htneoh@mmu.edu.my

**Abstract.** Student performance prediction is an important area in educational data mining that helps identify students who may need academic support. The research goal involves creating an adaptable machine learning system named EduInsight which predicts student results throughout multiple educational datasets. Three datasets were used: Cyprus (145 students), Portuguese (648 students), and Mathematics (394 students). Although the Portuguese and Mathematics datasets come from the same school and questionnaire, they differ by subject, while the Cyprus dataset comes from a different region and context. This study used common features from different datasets to maintain equal comparison.The datasets went through encoding procedures followed by normalization process and binary grade conversion (Pass/Fail) and reversed Likert scaling for variables including parental education and study time.Recursive Feature Elimination (RFE) is used for feature selection. The datasets are trained and evaluated by four machine learning models: Random Forest, Logistic Regression, XGBoost and Support Vector Machine. The research evaluates how well a standardized machine learning framework works when applied to different educational environments. Test accuracy results from datasets show either better or similar performance where Centre for Big Data and Blockchain Technologies Bayes improved test accuracy from 34.5% (validation) to 51.7% (test) on Cyprus dataset and maintained 84.6% on Portuguese dataset. The study demonstrates the scalability and effectiveness of EduInsight as an approach for predicting student results. The research findings enable educators to use data for better decision-making and academic planning as well as early student support.

# Introduction

Predicting student academic achievement has become an increasingly popular research subject in educational data mining, giving accurate predictions for timely intervention, resource planning, and individualized support. Cortez and Silva [1] described one of the early applications of data mining for the prediction of performance in Portugal's secondary school, demonstrating how grades in the classroom as well as socio-demographic characteristics influence educational achievement. Developing on this basis, Peña-Ayala [2] surveyed extensively diverse educational data mining studies, highlighting the value of predictive as well as descriptive models for student success. Recent research has highlighted the influence of parental education, family context, study skills, and extracurricular activities in determining student achievement, thus underscoring the need for a more comprehensive, data-driven approach.

This paper presents EduInsight, a machine learning model for assessing student performance on three educational datasets: Cyprus, Portuguese, and Mathematics. Whereas the Portuguese data sets are drawn from the same school and both have the same questionnaire designs, they differ on topics and thereby allow for comparison of how struggles in academic influence prediction measures. For them to be comparable, a combination of the common features within all three data sets was maintained for analysis.

One of the notable observations here is the consistent increase from validation to test accuracy across various models and datasets, which points towards the pipeline being extremely generalizable across academic and cultural contexts. For example, Naïve Bayes on the Cyprus dataset increased from 0. 517, and Naïve Bayes on the Portuguese dataset remained strong at 0.846. The current work is based on prior research in educational data mining. Marbouti et al. [3] and Al-Shabandar et al. [4] examined early academic predictors for dropout prediction, while Aulck et al. [5] found the importance of non-academic characteristics. The study by Sulaiman et al. [6] analyzed the problems that institutions face during the COVID-19 pandemic regarding online learning policies and student engagement at Multimedia University. The research conducted by Hussain et al. [7] introduced a deep learning framework which utilized the Levenberg-Marquardt algorithm to forecast academic outcomes through continuous assessment data. Other technology-based interventions, such as LearnwithEmma [8], reported online learning platforms with integrated teaching agents improved performance in Discrete Mathematics during the pandemic because over 75% of users reported improved performance and understanding on the tests.

According to their research results, the chatbot demonstrated useful benefits for student engagement but student satisfaction depended on emotional connections together with how well the tool matched learning preferences. EduInsight introduces a feature-matched, cross-cultural, and generalizable pipeline to predict student performance enabling scalable, data-driven educational interventions.

# DATA AND METHODOLOGY

EduInsight, a cross-cultural machine learning model that predicts student academic attainment from a standard set of socio-demographic, behavioral, and academic characteristics. The study design is structured in four main stages: data gathering, data preparation, model deployment, and feature exploration.

## Data collection

Three datasets were employed in this study. The first dataset is from a Cyprus higher education institution and comprises 145 student records [9]. The other two datasets were collected from the same Portuguese secondary school, but on two different topics: Portuguese language (648 students) and Mathematics (394 students) [10]. Although the Portuguese and Mathematics datasets relied upon one questionnaire and one school, their subject-specific nature introduces variation in academic challenge and student performance. The Cyprus dataset, however, was developed from a unique questionnaire with unique item structures. In comparative analysis, only a subset of common attributes from all three datasets—such as gender, age, parental education, study time, and extracurricular activities—were used for modeling.

## Preprocessing of datasets

To ensure comparability and consistency between all three datasets Cyprus, Portuguese, and Mathematics a basic preprocessing pipeline was established. While both the Portuguese and Mathematics datasets share the same structure of questionnaires and are from the same school, the Cyprus dataset was constructed around a different survey design. For example, the Portuguese dataset's sex, Medu, Fedu, and studytime were renamed to GENDER, MOTHER\_EDU, FATHER\_EDU, and STUDY\_HRS to be aligned with the Cyprus dataset.

Similarly, columns such as activities and romantic were renamed to ACTIVITY and PARTNER. For the Mathematics and Portuguese datasets, only students aged 15 to 22 were retained and those whose study hours were neither 1 nor 4 were removed. For the Cyprus dataset, where age and study hours were encoded categorically, values above 3 and 5 were removed to prevent inconsistent or noisy inputs. Binary responses (e.g., Yes or No, Male or Female) were represented as 0 and 1, whereas multi-categories were label-encoded in the same way for all datasets.

One of the critical preprocessing measures involved applying reverse Likert scaling to select ordinal variables in this case, MOTHER\_EDU, FATHER\_EDU, and STUDY\_HRS. These were reversed to enable consistency in interpreting throughout the set so that reduced values currently allude to favorable study situations. Grades lower than 10 in Portuguese and Mathematics datasets were labeled as "Fail" and grades equal to or above 10 as "Pass".

## Machine Learning Models

The overall aim here was to construct up models that could classify student academic performance categorized into Pass and Fail-performance levels based on only the common features defined in the three datasets. To be consistent and equal in comparison, the same steps of modeling were performed separately on each of the datasets with the same preprocessing rules, evaluation metrics, and the same algorithms. Four standard classification algorithms were used to do this: Naïve Bayes, Random Forest, Support Vector Machine**,** and Extreme Gradient Boosting. Every model is different and was selected to determine the other learning methodologies on learning data. Na ve Bayes classifier, particularly the Gaussian one, was attempted first as a baseline since it is computationally cheap and has an assumption of feature independence. Random Forest classifier was used to capture non-linear interactions and relationships among features. Finally, the XGBoost gradient boosting algorithm, a highly efficient algorithm, was used because of its ability to handle tabular data and scalability. Training and validation were conducted with the default hyperparameters so that the baseline performance could be measured in isolation of the impact of hyperparameter tuning. By employing identical models and training processes for the three datasets, this study allows for a fair comparison of model performance across two distinct education systems and determines the extent to which predictive variables are culture generalizable.

## Feature Selection

Good feature selection is a critical task in building stable and interpretable machine learning models. There were two techniques employed in this study to rank and refine the input features for student performance prediction: Spearman correlation analysis and Recursive Feature Elimination (RFE). These were independently applied to each of the datasets—Cyprus, Portuguese, and Mathematics—after encoding all the categorical variables and standardizing data.

### Spearman Correlation Analysis

To quantify the association strength between each feature and the target variable (GRADE), Spearman's rank correlation was computed. As a non-parametric statistic, it is appropriate to quantify monotonic relationships in educational data where most features are ordinal or nominal in nature. In each data set, a correlation matrix of all pairs was generated, and ranked absolute association strength with the target variable.

An initial filtering step was performed based on Spearman correlation values. The features with an absolute correlation value ≥ 0.05 or ≤ -0.05 with the target variable (GRADE) were retained for subsequent analysis. While this cutoff is very low, it was used intentionally so as not to eliminate features with smaller—potentially still useful—associations, particularly in smaller datasets where effect sizes tend to be subtle.

### Recursive Feature Elimination (RFE)

To further pre-select the best-performing features, RFE used a base estimator in the form of a Random Forest classifier. It successively eliminated the least important features and ranked the remaining features based on their contributions to model performance. RFE ranked the first 8 top features for every dataset, and these were further used to train and test the final machine learning models.

The combination of Spearman correlation filtering and RFE guaranteed that the resultant feature sets were not only statistically significant but also optimized in predictive performance. The features, specifically selected for each dataset, were next used in modeling and cross-dataset comparison. A table of the top RFE-selected features for each dataset is presented in Table 1 and a bar chart representing the top RFE-selected features are shown in Figure 1.

|  |  |  |
| --- | --- | --- |
| **TABLE 1.** Final selection of variables/ features in the datasets | | |
| **Variables** | **Description** | **Selected In** |
| AGE | Student age (Cyprus: ordinal 1–3; Portugal: 15–22, numeric) | Cyprus, Portuguese, Math |
| MOTHER\_EDU | Mother's education level (Cyprus: 1–6; Portugal: 0–4) | Cyprus, Portuguese, Math |
| FATHER\_EDU | Father's education level (Cyprus: 1–6; Portugal: 0–4) | Cyprus, Portuguese, Math |
| MOTHER\_JOB | Mother's job type (encoded categorical) | Cyprus, Portuguese, Math |
| FATHER\_JOB | Father's job type (encoded categorical) | Cyprus, Portuguese, Math |
| #\_SIBLINGS | Family size (Portugal: LE3 or GT3; Cyprus: number of siblings) | Cyprus |
| STUDY\_HRS | Weekly study time (Cyprus: ordinal 1–5; Portugal: 1–4) | Cyprus, Portuguese, Math |
| ACTIVITY | Participation in extracurricular activities (1 = Yes, 0 = No) | Portuguese, Math |
| PARTNER | In a romantic relationship (1 = Yes, 0 = No) | Cyprus, Portuguese, Math |

A graph with different colored lines

AI-generated content may be incorrect.**FIGURE 1**. RFE-Selected features in bar chart

# Results and discussion

This section presents the performance of four machine learning models—Random Forest, Naïve Bayes using Logistic Regression, XGBoost, and Support Vector Machine (SVM)—on three education datasets: Cyprus, Portuguese, and Mathematics. Each model was trained on the top eight features chosen through Recursive Feature Elimination (RFE), and tested through validation and test metrics like accuracy, precision, recall, and F1-score. Special focus lies in consistency and changes between validation and test accuracy as a metric of the pipeline's generalizability.

## Model Performance: Cyprus Dataset

Table 2 shows the performance metrics of all models on Cyprus dataset. For Cyprus dataset, XGBoost emerged as the winner with validation accuracy of 0.586 and test accuracy of 0.621. The 6% boost shows a good-generalized model with stable feature importance. The F1-score increased from 0.582 to 0.589 in favor of its robustness. Random Forest generalizes too (Val: 0.552 → Test: 0.586), while Naïve Bayes performed far behind on both splits. SVM caught up a bit with validation to train, tying XGBoost's test accuracy but lagging in F1.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TABLE 2.** Performance metrics of all models on Cyprus dataset | | | | | | | | |
| **Algorithm** | **Validation Data** | | | | **Testing Data** | | | |
| **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Naïve Byes | 0.345 | 0.340 | 0.331 | 0.332 | 0.517 | 0.438 | 0.452 | 0.436 |
| Random Forest | 0.552 | 0.548 | 0.551 | 0.543 | 0.586 | 0.547 | 0.543 | 0.542 |
| SVM | 0.517 | 0.438 | 0.452 | 0.436 | 0.621 | 0.570 | 0.535 | 0.505 |
| XGBoost | 0.586 | 0.590 | 0.596 | 0.582 | 0.621 | 0.592 | 0.588 | 0.589 |

## Model Performance: Portuguese Dataset

Table 3 shows the performance metrics of all models on Portuguese dataset. For the Portuguese dataset, both Naïve Bayes and SVM achieved the highest validation and test accuracy of 0.846, with excellent consistency and generalization. Naïve Bayes performed better in terms of F1-score (0.503) compared to SVM (0.458). XGBoost and Random Forest came close second but still had acceptable performance. The result shows that if there is a large and well-structured dataset, even simpler models like Naïve Bayes can deliver good results.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TABLE 3.** Performance metrics of all models on Portuguese dataset | | | | | | | | |
| **Algorithm** | **Validation Data** | | | | **Testing Data** | | | |
| **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Naïve Byes | 0.846 | 0.676 | 0.520 | 0.503 | 0.846 | 0.676 | 0.520 | 0.503 |
| Random Forest | 0.823 | 0.570 | 0.527 | 0.525 | 0.769 | 0.417 | 0.455 | 0.435 |
| SVM | 0.846 | 0.423 | 0.500 | 0.458 | 0.846 | 0.423 | 0.500 | 0.458 |
| XGBoost | 0.808 | 0.565 | 0.539 | 0.542 | 0.762 | 0.416 | 0.450 | 0.432 |

## Model Performance: Mathematics Dataset

In the Mathematics dataset, Naïve Bayes once again outperformed other models, with a validation accuracy of 0.671 and an improved test accuracy of 0.696. It also achieved the highest test F1-score (0.530), indicating balanced performance between precision and recall. SVM performed consistently as well, with matching validation and test accuracy at 0.658. XGBoost underperformed on the test set, possibly due to overfitting the validation data. Table 4 shows the performance metrics of all models on Mathematics dataset.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TABLE 4.** Performance metrics of all models on Mathematics dataset | | | | | | | | |
| **Algorithm** | **Validation Data** | | | | **Testing Data** | | | |
| **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Naïve Byes | 0.671 | 0.590 | 0.520 | 0.465 | 0.696 | 0.683 | 0.558 | 0.530 |
| Random Forest | 0.532 | 0.438 | 0.445 | 0.440 | 0.595 | 0.453 | 0.473 | 0.446 |
| SVM | 0.658 | 0.555 | 0.520 | 0.483 | 0.658 | 0.555 | 0.520 | 0.483 |
| XGBoost | 0.633 | 0.575 | 0.570 | 0.571 | 0.557 | 0.457 | 0.464 | 0.456 |

## Overall Model Performance

Figure 2 shows comparison of test accuracy across models and datasets. SVM and Naïve Bayes demonstrate strong performance on the Portuguese and Mathematics datasets according to the chart while SVM maintains competitive results. The Cyprus dataset shows Random Forest together with XGBoost achieve better accuracy results than Naïve Bayes. The findings demonstrate that different models excel on different datasets yet Naïve Bayes together with SVM show consistent strong performance in student performance prediction tasks.

A graph of different colored bars

AI-generated content may be incorrect.

**FIGURE 2.** Comparison of test accuracy across models and datasets

## Key Predictors of Student Performance

On each of the three datasets that were analyzed Cyprus, Portuguese, and Mathematics the same set of predictors was identified using feature selection and model interpretation. The most significant and common predictors were age, parent education, study time, and romantic relationships.

These variables were selected by RFE in all the datasets and exhibited high relevance in distinguishing between students who would pass and those likely to fail. Age was the most important variable across all datasets. In the Cyprus dataset, in which age is used as an ordinal category, older students were allocated more frequently to lower performance classes.

Similarly, in the Portuguese and Mathematics datasets, age continued to have predictive power, affirming the general finding that path and rate of academic study are significant, regardless of cultural or institutional setting. Specifically, parental levels of education historically the conventional gold standard for academic assistance and socio-economic stability proxies had inconsistent and relatively weaker correlations with end performance, particularly in the Cyprus dataset.

This suggests that while structural advantages are significant, behavioral and developmental predictors may have greater direct effects on academic performance in real student populations. Study hours per week, reversed in direction to put interpretive flow in the anticipated direction, was a consistent variable across all databases. Their consistency in culturally diverse samples also lends credence to their generalizability and usefulness for the construction of early-warning systems and support interventions based on a holistic portrait of student success.

## Practical Implications of Results

The results of this study highlight the growing potential of using machine learning and predictive analytics on education systems in a scalable and feasible way. The continued robustness of models such as XGBoost and Naïve Bayes after being trained on small, interpretable input features suggest that complex student outcomes might be predictably modelled without having to gather highly sensitive or obtrusive data.

This implies that schools can begin to implement predictive systems based on basic academic, behavior, and demographic data making early intervention possible and ethical. The most profound finding, perhaps, is the persistent effect of developmental and behavior traits such as age, marital status, and study hours across the three datasets. The recurring influence of PARTNER and AGE on RFE shows the role of pacing- and social-emotional variables in student success.

Additionally, the cross-subject, cross-country design of the study demonstrates that generalizable models can be applied successfully in a variety of curricula and cultural educational contexts. Despite subject matter and dataset structure differences, performance of the models was strong when standardized preprocessing and feature alignment were employed. Briefly, this study presents a model not only for the forecasting of student performance but also for operational educational planning, providing institutions with an explicit, adaptive, and socially accountable approach to academic success prediction.

## Limitations and Future Work

The most evident limitation is the imbalance between the size of the dataset among the three data sources. While the Cyprus dataset only had 145 records, the Portuguese and Mathematics datasets were comprised of 648 and 394 records, respectively. The imbalance would have contributed to the variation in the performance of classifications, particularly the very high rates observed in certain models such as Naive Bayes and XGBoost. Secondly, the scope of features used is limited. As a result, while the models are effective at predicting students into Pass or Fail, they do not have good interpretability of the causality of student underperformance.

Another limitation is the absence of external or real-world testing. To address these constraints, future research must strive to expand the collection of datasets from more diverse educational institutions, geographies, and student groups. The incorporation of longitudinal data would allow for monitoring of academic achievement over time, improving causal inference and prediction stability. Furthermore, the incorporation of psychosocial, motivational, and institutional variables would enhance the interpretability and equity of models, yielding more actionable outcomes for educators. Pilot studies in schools would assess the feasibility, ethical acceptability, and sufficiency of the model as a decision-support tool to guide its development into an effective and ethical scale-up as a student success prediction framework.

# Conclusion

The study presented EduInsight, a predictive modeling framework that was developed to forecast student academic performance in three real-world educational datasets in Portugal, Cyprus, and a sub-set of Mathematics. With a view to shared socio-demographic and behavioral factors, the study demonstrated that consistent preprocessing through reverse Likert scaling and recursive feature elimination enables precise prediction of performance using interpretable variables of low complexity. The developmental and behavioral traits were found to consistently appear across datasets, highlighting the necessity for schools to adopt a more holistic method of student success incorporating emotional and social aspects alongside cognitive achievement. With further validation, extension, and real-world testing, the framework may be shaped into an ethical and effective decision-support system to enhance student performance in culturally diverse learning environments.

# References

1. P. Cortez, and A.M.G. Silva, “Using data mining to predict secondary school student performance,” (2008).
2. A. Peña-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works,” Expert Systems with Applications **41**(4, Part 1), 1432–1462 (2014).
3. R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, and M. Radi, “Machine learning approaches to predict learning outcomes in Massive Open Online Courses,” *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, USA, 2017, pp. 713–720.
4. F. Marbouti, H.A. Diefes-Dux, and K. Madhavan, “Models for early prediction of at-risk students in a course using standards-based grading,” Computers & Education **103**, 1–15 (2016).
5. L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, “Predicting Student Dropout in Higher Education,” (2016).
6. R. Idris, and J. Bacotang, “Exploring STEM Education Trends in Malaysia: Building a Talent Pool for Industrial Revolution 4.0 and Society 5.0,” International Journal of Academic Research in Progressive Education and Development **12**(2), 358–369 (2023).
7. M.M. Hussain, S. Akbar, S.A. Hassan, M.W. Aziz, and F. Urooj, “Prediction of Student’s Academic Performance through Data Mining Approach,” Journal of Informatics and Web Engineering **3**(1), 241–251 (2024).
8. C. Susaie, C.-K. Tan, and P.-Y. Goh, “Learning Experience with LearnwithEmma,” Journal of Informatics and Web Engineering **1**(2), 30–44 (2022).
9. B.Ş. Nevriye Yilmaz, “Higher Education Students Performance Evaluation,” (2019). https://www.kaggle.com/datasets/csafrit2/higher-education-students-performance-evaluation
10. P. Cortez, “Student Performance,” (2008).https://doi.org/10.24432/C5TG7T.