# Use educational data mining to enhance student achievement and performance of physical subjects

Nihal BOURAS*, Laila AYAICHI, Aziz AMAAZ, Abderrahman MOURADI, Abderrahman EL KHARIM

Laboratory of Applied and Didactic Sciences, Abdelmalek Essaadi University, Tetouan, Morocco

*nihal.bouras@etu.uae.ac.ma

## Abstract :

Every year, educational institutions analyze the success rate of students in order to develop their academic standard and avoid failure. To determine this rate, there are several types of techniques, such as statistics, physical examination, and in-course data mining techniques.

In this study, we will follow one of the educational data mining techniques to explore massive data, in order to extract useful elements to improve the performance and performances of physics students using information on their academic results, their behaviour in class, their participation in activities, etc.

The technique of extracting data from KNIME and other algorithms was used in this study.

## Keyword:

Educational Data Mining, Database, Student Performance, KNIME.

## 1. Introduction:

Institutions of higher education are designed to deliver education of high quality that can promote the development of consciousness and understanding (Adekitan & Salau, 2019).

The school record of students in these institutions is of considerable concern, attracting the attention of many researchers and parents (Ramachandran et al., 2023). It is crucial for educational institutions to oversee the academic achievements of their students and take the necessary steps to improve it (Alwarthan et al., 2022)

Educators should assess the performance of university or college students in achieving their goals and fostering a setting focused on ongoing enhancement (Mansson, 2016).

Various factors come into consideration when assessing the performance of an educational institution (Dabhade et al., 2021; Prodanova & Kocarev, 2023). Based on these criteria,


M. Khaldi et al. (eds.), *Proceedings of the E-Learning and Smart Engineering Systems (ELSES 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 14,
https://doi.org/10.2991/978-94-6463-360-3_49

ranking adjustments are required for the institution (Saleh et al., 2021). The commitment of educational institutions to the provision of quality education remains essential to achieve exceptional (Adejo & Connolly, 2017).

Currently, many researchers are working on pedagogical datasets to analyze factors that aim to create student performance problems in order to solve them (Enaro & Chakraborty, 2019)

This study exposes the deployment methodology of techniques of mining data to anticipate assess performance and generate the required outcomes. It focuses on the implementation of techniques predicting performance and providing insights into the substantial potential of data mining methods.

## 2. The Problem:

This study aims to comprehend the diverse factors influencing student performance and forecast upcoming semester outcomes through data analysis and past performance. The major aim of this study is to extract valuable information from students' databases to obtain the characteristics and discover the attributes that disrupt their behavior in order to solve them.

## 3. Research Methodology:

For this study, the evaluation of the performance of undergraduate students in physics-chemistry teaching at the higher education institution ENS of Tétouan integrates the analysis of their results during the previous semesters as well as their characteristics.

Attributes like student regularity, mean notes, and pertinent lessons details, student classroom participation level, etc. Utilized in a data exploration design to anticipate the academic efficiency of 520 pupils. The application of implemented a decision tree model to forecast the likelihood of failure among 520 students was undertaken to gain relevant knowledge. This information is intended to enable the management team to initiate appropriate enabiling timely and preemptive intervention. The research categorezed student grades into five classes : excellent, vey good, good, acceptable and failure. Eight entry characteristics, such as the student’s department, grades from the previous year, level of class participation, elements like labortory reports, homework grades, seminar grades, completion by manipulation along with overall grades, were incorporated into developed decision tree model

The data collected The information gathered from the survey document and the division educational of the institution were combined into an Excel chart.

| N° d'ordre | CNE | Leisure | family income | Time spent of social media | Entrance exam score | GPA of the previous semestres | Career objective | Circle of friends effect | Students participation in extracurricular activites |
|---|---|---|---|---|---|---|---|---|---|
| 1 | S132246573 | Judo | mean | 3 | 14,43 | 15,76 | Master new skills | Positive | Oui |
| 2 | P131123418 | Race | mean | 4 | 12,62 | 11,84 | Evolution of the soul | Positive | Non |
| 3 | P100005312 | Gymnastique | mean | 5 | 13,76 | 13,48 | Acquire the necssary confidence | Négative | Oui |
| 4 | P135246706 | Music | mean | 2 | 15,45 | 14,39 | Being confident in one's self | Négative | Oui |
| 5 | P131101549 | Race | mean | 2 | 16,43 | 15,22 | Master new skills | Positive | Non |
| 6 | S131067652 | Traveling | mean | 3 | 11,57 | 15,89 | Acquire the necssary confidence | Positive | Non |
| 7 | S137174780 | Reading | mean | 1 | 13,56 | 10,19 | To learn a new language | Positive | Oui |
| 8 | S139033305 | Race | mean | 3 | 14,62 | 12 ,74 | Master new skills | Positive | Oui |
| 9 | S130066778 | Judo | mean | 2 | 13,11 | 13,71 | Being confident in one's self | Positive | Oui |
| 10 | P130335666 | Painting | mean | 1 | 12,89 | 12,72 | Evolution of the soul | Positive | Non |
| 11 | N120064208 | Reading | mean | 4 | 10,21 | 16 | Acquire the necssary confidence | Positive | Oui |
| 12 | P131041339 | Writing | mean | 1 | 14,33 | 13,35 | Evolution of the soul | Négative | Oui |
| 13 | L141057130 | Tennis | mean | 1 | 11,22 | 12,68 | Evolution of the soul | Négative | Non |
| 14 | N138355093 | Music | mean | 2 | 11,63 | 12,81 | To learn a new language | Positive | Oui |
| 15 | M130258303 | Traveling | mean | 4 | 12,65 | 10,05 | Evolution of the soul | négative | Oui |

**Fig 1 :** Specifications Excel Table Extract

A data set of 520 students was collected. In addition, the collection of specific academic details, such as cumulative rank point average (GPA) and student acceptance criteria, was conducted with the institution's faculty department to improve the details of the information collected.

On the Konstanz Information Miner (KNIME) analysis platform, a predictive model was developed to identify the underlying relationship between the characteristics of the dataset. This model aims to provide reasonably predict the final year class of major students in computer education, based on their cumulative average in the initial two of the three full years of study for experimental physics courses.

In the KNIME workflow, data were loaded into the platform using the Excel driver, and the statistical properties of the dataset were obtained via the statistical node. The data underwent prior processing through a stratified division of samples; 70% were allocated to and 30% for evaluations. Color coding of the grade class was implemented to allow visual evaluation through the use of point clouds. Standardization was applied to the samples, and a Principal component analysis was used to reduce the size of the variables.

## 4. Results and Discussion:

The outcomes of predictives models, employing both KNANE and regression-based approaches, are presented in this section.

### 4.1 Results obtained from the model based on knime :

The forecast capacity of each of the five input variables was examined. through the front freature select metanode of the KNIME sever. The results revealed that third-year APM was the variable with the highest reliability is closely trailed by second-year APM to improve precision. The first annual grade point average, program and entry year were identified as the variables with the least impact on classification experience.

The algorithms' performance, regarding prediction confusion matrices, is displayed Table 1 for the decision tree forecast and Figure 2 for the logistic regression forecast. These matrices depict the algorithms predictive performance for each classification class, identifying True Positives, False Pasitives, True Negatives and False Negatives predictions. (Fahmy Amin, 2022; Heydarian et al., 2022) for classes 1, 2/1 and 2/2. Tables 3 then provide a comparison of the performance of the two models.

**Table 1 :** Confusion matrix for the decision tree predictor.

| | 2/2 | 3rd | 2/1 | 1st |
|---|---|---|---|---|
| 2/2 | 176 | 9 | 21 | 0 |
| 2/1 | 16 | 0 | 237 | 3 |
| 1er | 0 | 0 | 5 | 44 |

**Table 2 :** Confusion matrix logistic Regression predictor

| | 2/2 | 3rd | 2/1 | 1st |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| **2/2** | **184** | **7** | **20** | **0** |
| **2/1** | **13** | **0** | **246** | **2** |
| **1er** | **0** | **0** | **7** | **42** |

**Table 3 :** Comparison of model performance

| | Tree Ensemble | Logistic Regression |
|---|---|---|
| **Correct Classified** | **486** | **493** |
| **Accuracy** | **87.884%** | **89.15%** |
| **Cohen's Kappa (k)** | **0.803** | **0.823** |
| **Wrong Classified** | **67** | **60** |
| **Error** | **12.116%** | **10.85%** |

### 4.2 Regression-based model results :

As an element of a comparative assessment and to strengthen the affirmation of the findings of the data exploration model, which disclosed a minimal precision of 85.89%, linear and quadratic regression models were established to scrutinize the collection of details regarding academic achievement. The regression analysis was not executed on KNIME with the intention of constructing a model that consolidates all datasets, rather than dividing them into two for training and testing, as mandated by KNIME's data exploration modules.

### 4.3 The linear regime :

The formula deduced from the linear regression framework through decomposition results is described in the following formula:

$$y = 0.4865 - 0.0057x_1 - 0.0016x_2 + 0.1811x_3 + 0.2788x_4$$

Considering the standardized coefficient (β) shown in Figure 4, we see that second-year cumulative GPA has the greatest impact on anticipated cumulative GPA, followed by first-year cumulative GPA and academic program, while the year of registration has the slightest impact on the dependent factor.

**Table 4 :** Quadratic results of the regression analysis

| | Estimate | Standard Error (SE) | Beta (β) |
|---|---|---|---|
| **(Intercept)** | **0.4865** | **0.0197** | **-** |
| **Program ($X_1$)** | **-0.0057** | **0.0017** | **-0.0164** |
| **Year of Entry ($X_2$)** | **-0.0016** | **0.0018** | **-0.0053** |
| **First Year GPA ($X_3$)** | **0.1811** | **0.0090** | **0.1809** |
| **Second Year GPA ($X_4$)** | **0.2788** | **0.0089** | **0.3141** |

According to the information in Figure 5, the model displays a coefficient of determination R2 of 0.955, indicating that the ultimate cumulative average of LE PC students may be within reason anticipated from their results (GPA) during the earliest two years of their three-year course.

**Table 5 :** Performance measures for regression

| Performance parameters for ml algorithms | Linear Regression |
|---|---|
| **MAE** | **0.44079** |
| **MSE** | **0.23040** |
| **RMSE** | **0.48000** |
| **$R^2$ score** | **0.955** |

## 5. Conclusion :

Over the years, the evolution of the management of higher education systems has shifted from reactive decision-making to a proactive approach involving proactive analysis of System efficiency. Ensuring sufficient quality within the education system is essential for the success of learners and the comprehensive worth of the knowledge conveyed. Early detection of student learning problems and difficulties has many advantages, as it provides a unique opportunity to take rapid action on causal factors, thus preventing trends towards school failure and dropout.

For this research, a A data exploration method was employed to evaluate the soundness of this assumption, conducting a foretelling examination to infer the ultimate grade and classification of learners in their ultimate year, derived from their cumulative GPA over the initial three years of the program. Program and year of registration were employed as forecasting elements in a KNIME process, with the application of six independent data mining algorithms, run individually to allow comparative The performance of the six dedicated algorithmic was examined. A peak precision of 89.15% was reached, and, when employing regression techniques structures for yield confirmation, coefficients of determination R2 of 0.955 were obtained through the application of pure linear and quadratic regression models. These results suggest that the performance of PC LE students in their third and the last year of study can certainly be a reasoned one. anticipated on the basis of their performance in the first three academic terms.

## 6) References:

**1)** Adejo, O., & Connolly, T. (2017). An Integrated System Framework for Predicting Students' Academic Performance in Higher Educational Institutions. *International Journal of Computer Science and Information Technology*, *9*(3), 149-157. https://doi.org/10.5121/ijcsit.2017.93013

2) Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon*, *5*(2), e01250. https://doi.org/10.1016/j.heliyon.2019.e01250

3) Alwarthan, S. A., Aslam, N., & Khan, I. U. (2022). Predicting Student Academic Performance at Higher Education Using Data Mining : A Systematic Review. *Applied Computational Intelligence and Soft Computing*, *2022*, 1-26. https://doi.org/10.1155/2022/8924028

4) Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. *Materials Today: Proceedings*, *47*, 5260-5267. https://doi.org/10.1016/j.matpr.2021.05.646

5) Enaro, A. O., & Chakraborty, S. (2019). A Review on the Academic performance analysis of higher Education Students using Data Mining Techniques. *2019 International Conference on contemporary Computing and Informatics (IC3I)*, 256-260. https://doi.org/10.1109/IC3I46837.2019.9055632

6) Fahmy Amin, M. (2022). Confusion Matrix in Binary Classification Problems : A Step-by-Step Tutorial. *Journal of Engineering Research*, *6*(5), 0-0. https://doi.org/10.21608/erjeng.2022.274526

7) Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM : Multi-Label Confusion Matrix. *IEEE Access*, *10*, 19083-19095. https://doi.org/10.1109/ACCESS.2022.3151048

8) Mansson, D. (2016). Exploring College Students' Expressed Concern About Their Academic Performance. *College Student Journal*, *50*(1), 121-129.

9) Prodanova, J., & Kocarev, L. (2023). Universities' and Academics' Resources Shaping Satisfaction and Engagement : An Empirical Investigation of the Higher Education System. *Education Sciences*, *13*(4), Article 4. https://doi.org/10.3390/educsci13040390

10) Ramachandran, K. K., Ravichand, M., Joshi, K., Vekariya, V., Saini, D., & Ponnusamy, R. (2023). Investigation of the Educational Performance on the Revolutionary Philosophical Electoral Online Learning Platform Centred on Deep Learning. *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, 639-642. https://doi.org/10.1109/AISC56616.2023.10085556

11) Saleh, M. A., Palaniappan, S., Abdallah, N. A. A., & Baroud, M. M. B. (2021). Predicting student performance using data mining and learning analysis technique in Libyan Higher Education. *Periodicals of Engineering and Natural Sciences (PEN)*, *9*(3), 509. https://doi.org/10.21533/pen.v9i3.2212