www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(1), Jan-April 2025, 177-189. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2025.v16i1.551

A similarity Based Algorithm for Predicting Academic Success in First-year Undergraduates

Cinthia Rodríguez Maya[1], *Carlos Gershenson García*[2], *Helena Montserrat Gómez Adorno*[3]

[1] Posgrado en Ciencia e Ingeniería de la Computación, Universidad Nacional Autónoma de México
[2] Department of Systems Science and Industrial Engineering, Binghamton University
[3] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México
E-mail: cinthia.rodriguez@ciencias.unam.mx, cgg@binghamton.edu, helena.gomez@iimas.unam.mx

Abstract. According to the European Commission, "Early school leaving is linked to unemployment, social exclusion, poverty, and poor health" (European Commission, 2024). Due to the importance of reducing school drop-out rates, several authors have analyzed this phenomenon. Before the boom of Artificial Intelligence, instead of using or implementing programming methods, researchers applied written formulas and rudimentary graphical visualizations to predict academic completion and the main factors behind it; besides, the development of Machine Learning (ML) algorithms has enhanced the precision and performance of an ample variety of investigations including the educational field. In this paper, we use a dataset of undergraduate-level students at the Universidad Nacional Autónoma de Mexico (UNAM) to predict timely academic completion. We use seven common ML algorithms and propose a novel algorithm based on students' similarities according to the most relevant features in common. This algorithm shows a higher precision than some traditional categorical ML algorithms. This innovative way to predict academic success can support educators, pedagogues, and policymakers make better decisions at UNAM.

Keywords: Machine Learning, Academic Achievement, Scholar Dropout, Learning Analytics

Article Info
Received November 13, 2024
Accepted March 12, 2025

1 Introduction

According to the Dirección General de Evaluación Institucional (DGEI, General Directorate of Institutional Evaluation), an organization that collects and reports statistical data about students and professors at the Universidad Nacional Autónoma de Mexico (UNAM, National Autonomous University of Mexico)[1] , the cost of a student in 2023 was around MXN 74,342[2] (3,913 USD). High dropout rates and delayed completion of higher education are associated with considerable personal and social costs. Dropping out from higher education represents a cost for the government and society, an unnecessary expense for the family, and an experience of failure for the student (Latif et. al., 2015).

Educational Data Mining (EDM) is the use of Data Mining methods on educational data such as student information, educational records, exam results, student participation in class, and the frequency of students' asking questions. In recent years, EDM has become an effective tool for identifying hidden patterns in educational data, predicting academic achievement, and improving the learning/teaching environment (Yagci, 2022). EDM has used different ML techniques with high accuracy to predict students' future performance (Romero et. al., 2010).

[1] https://www.dgei.unam.mx/hwp/

[2] https://web.siia.unam.mx/siia-publico/?tabla indicadores basicos=&entidades=entidadesTodaUNAM

This paper analyzes a group of first-year bachelor students at UNAM. This group of students started university at the end of 2016. We will refer to them as Generation 2017[3] . This study only considers degrees with a duration of 5 years and people who studied in a school attached to UNAM and answered a vocational test called PROUNAM in the second year of high school.

The features of the students come from two databases: *Academic records* and the *PROUNAM* test. The first database contains information about grades, failed subjects, and school or campus at UNAM, among other attributes related to academic performance, whereas the second database comprises the results about cognitive skills, intellectual aptitudes, and vocational interests.

The main contribution of this paper is a new classification ML algorithm to predict academic completion on time. This algorithm is based on student similarities. It can be argued that students are most alike when all their characteristics are exactly the same. However, this behavior could be unyielding and infrequent in some cases. Furthermore, some student's features are more relevant than others. Thus, our algorithm applies a formula to measure the difference or distance between the values of every attribute, looking for the minimum distance. Besides, it considers the attributes' relevance through a calculated relation of feature weights. The weight of a feature is the relevance associated with it. This prediction method is the main difference between our proposed algorithm and other traditional ML algorithms.

In the context of our research, we choose to evaluate and compare the algorithms using the precision metric. It is preferable to reduce the number of false positive values as much as possible because it is preferable to say a student will not finish their major on time and suggest extra help rather than ignore a student whose optimistic forecast is wrong. The results show our proposed algorithm has a higher precision than some traditional ML algorithms.

The remainder of this paper is organized as follows: Section 2 presents related work on predicting academic achievement using machine learning algorithms or statistical techniques. Section 3 shows the main steps around the process of data exploration, feature selection, and feature engineering. Besides, it explains our final dataset and its attributes, meanings, and ranges. Section 3.4 describes our proposed similarity-based algorithm and how it works. Section 3.4.1 shows a mathematical proof of our algorithm's correctness and complexity. Section 4 presents the precision of executing several traditional classification ML algorithms over our final dataset; we compare these results against our proposed algorithm. Finally, Section 5 presents our conclusions and outlines potential directions for future research.

2 Related work

Understanding the drivers behind academic achievement (AA) is an everlasting global challenge that concerns students, their families, teachers, public decision-makers, and everyone concerned about development and well-being at a global level (Noell et. al., 2019).

AA reflects the progress toward acquiring educational skills, materials, and knowledge, usually spanning various disciplines. It refers to achievement in academic settings rather than the general acquisition of knowledge in non-academic settings (Bolt et. al., 2011).

Some authors have made studies applying ML algorithms to datasets from different universities worldwide, trying to identify factors that might predict AA.

In Cambodia, there is evidence that high levels of mathematical skills are related to better academic performance in national exams (Penh, 2018). To obtain their conclusions, the researchers applied surveys to students, calculated percentages, and visualized different graphs. The data set from which these results were obtained was collected from 22 high schools and contained 1,204 samples.

In (Roy et. al., 2017), the authors analyzed 395 students in two secondary schools in Portugal and found that parent's education, alcohol consumption, and romantic relationships are correlated with academic performance.

[3] In UNAM, a Generation is named according to the year of admission + 1. In the United States, Generation 2017 would be the 'class of 2021' (since they should finish in 2021)

The authors in (Lecompte et. al., 1983) analyzed a sample of 874 first-year college students in Belgium. They calculated and examined correlation graphs and found that financial difficulties are the most common feature of students withdrawing from university.

In the same way, in (Betts et.al., 1999), the authors analyzed the Grade Point Average (GPA) of more than 5,000 undergraduates at the University of California, San Diego, and found that GPA is strongly correlated with family income. The authors applied statistical formulas to their data. In contrast, (Cruz-Jesus et.al., 2020) applied a study over a sample of 110,627 students in Portugal and found that economic factors were not correlated with academic performance.

The authors in (Vallmur et.al., 2001) did a study with data from Queensland University, Australia. The researchers selected a sample of 197 first-year university students from the Faculties of Science (n = 149) and Information Technology (n = 48); this sample included 103 males and 94 females, with a mean age of 21.24 years. The techniques used in this research were standard regression and ANOVA. The authors found that economic status is not correlated with academic completion. Nevertheless, students who have full-time work have higher GPAs than students with part-time employment responsibilities.

In public schools of the Federal District of Brazil, the authors in (Fernandes et.al., 2019) collected 238,575 records from 2015 and 247,297 records from 2016, and their study revealed that attributes like grades, absences, neighborhood, school, and age are potential indicators of a student's academic success or failure. The authors used descriptive statistics to obtain their conclusions.

At UNAM, the Coordinación de la Universidad Abierta Innovación Educativa y Educación a Distancia (CUAIEED, Coordination of Open University, Educational Innovation, and Distance Learning) conducted a study on academic data related to the medicine major at the Facultad de Medicina (Medicine Faculty) in Ciudad Universitaria (C.U., UNAM's central campus). The main factors that are negatively correlated with the academic achievement of these students were older age, insufficient general knowledge before university, a precarious financial situation, family violence, and poor English knowledge (Monteverde-Suárez et.al., 2024).

Table 1 shows a more specific description of the techniques and outcomes obtained by the authors previously mentioned. The authors of (Roy et. al., 2017) applied Naïve Bayes, ID3/C4.5, and Multi-layer Perceptron algorithms to predict a final grade (A, B, C, D, Fail). The highest obtained precision was 73.92%, achieved by the ID3/C4.5 algorithm.

The authors of (Cruz-Jesus et.al., 2020) and (Monteverde-Suárez et.al., 2024) made a binary classification. The goal of (Cruz-Jesus et.al., 2020) is to predict if a student will be approved for the following year. The most accurate algorithm used support vector machines and obtained a precision of 87%. The authors of (Monteverde-Suárez et.al., 2024) applied Naïve Bayes and artificial neural networks to predict if a student will fail a subject at the end of the first year. The highest precision was 74% obtained by using artificial neural networks.

This paper proposes a novel algorithm explicitly inspired by the student's features and behaviors. We hypothesize that students who present similar conditions or share the same feature values will also have similar academic outcomes.

3 Methodology

This section is about the process of data analysis and information processing.

3.1 Databases

To analyze the students of the Generation 2017, we used two databases: *Academic records* and *PROUNAM* test.

Table 1. Results of some authors using different ML algorithms

Authors	Algorithm	Type of classification	Precision
Roy et. al., 2017	Naïve Bayes	Multiclass: final grade (A, B, C, D, Fail)	68.60%
	ID3/C4.5		73.72%
	MLP (Multi-layer Perceptron)		51.13%
Cruz-Jesus et. al., 2020	Artificial Neural Networks	Binary: The student was or was not promoted to the following year	77.00%
	Decision Trees		79.00%
	Extremely Randomized Trees		77.00%
	Random Forest		80.00%
	Support vector machines		87.00%
	K-nearest neighbors		79.00%
Monteverde-Suárez et. al., 2024	Artificial Neural Networks	Binary: Regularity at the end of the first year	71.00%
	Naïve Bayes		74.00%

3.2 Academic records

The original database comprises 560,282 records from June 2008 to August 2020 and has 42 features. This dataset is handled and maintained by the Dirección General de Administración Escolar (DGAE, General Directorate of Academic Administration); this institution manages all the official academic information about students at UNAM at all levels and entities. DGAE registers and carefully examines all data about enrollments, grades, subject registrations, and graduation of students as valid information.

The database was composed initially of 42 features for each student, but only 10 remained; the reason for deleting 32 features was because they contained irrelevant or repetitive information, while other candidate features, like the correlation matrix, were dropped after a statistical analysis. The 10 remaining features are the student's grade obtained in the first year, the number of subjects that the student has not approved in the first year, the grade obtained at high school, the current major and campus or school, the major's knowledge area, in person major modality, remote major modality, hybrid major modality, and the student's age.

3.3 PROUNAM test

This test was responded by students whose high school is incorporated into UNAM.
This psychometric exam aims at students in their professional career selection, comparing their results in several aptitudes with the recommended scores for every major. PROUNAM's results advise students regarding what skills improve in case of having poor results in a desired major whose high scores in specific skills could increase the success of accomplishment. The percentage of students of the generation 2017 who answered the PROUNAM test is 39.83%

The measured abilities of this test are abstract reasoning, numerical aptitude, mechanical aptitude, shape assembly, discrimination of figures, cryptograms, word recognition, verbal skill, and language use. As to professional interests, PROUNAM considers physical sciences, mechanics, mathematics, biological and health sciences, ecology and environment, altruism/social service, political sciences, social sciences, administrative/financial, organizational/persuasive, artistic visual plastic, musical expression, oral expression, and written expression. The provided database has 27,190 student records with 1,024 features each.

The features that we selected from this database were 9: the academic potential of the student, the student's father's highest academic level, the number of people that work at the student's house, the number of rooms at the student's house, student's mother highest academic level, the quantity of money that enters student's house approximately every month, student's aptitudes, student's interests result, major's knowledge area that the student wanted at high school. The rest of the features were considered irrelevant; they were either repetitions of other features but in different formats or variables containing serial numbers, dates, and auxiliary strings to print exams on paper.

3.4 Feature Selection and Feature Engineering

The final dataset results from merging academic records and the PROUNAM test. It contains 16,887 students with 19 features each. All students correspond to Generation 2017, with 3 to 5 - year majors, and answered the PROUNAM test. The target is a binary variable indicating whether the students completed their majors on time or not.

Feature engineering involves creating, transforming, extracting, and selecting features to build an accurate ML algorithm. We decided to reduce the range values of some features and combine others to optimize and improve the performance of traditional machine learning algorithms and our proposed algorithm. Table 2 shows four columns: the first and second correspond to the names and descriptions of our features, and the third column contains the ranges or set of possible values of the selected features in their original version. The fourth column contains the reduced version of the features. Below, we explain the process for the feature reduction.

The variables academicPotential, aptitudesResult, and interestsResult were reduced from a percentage (0% to 100%) into three categories. These come from the PROUNAM test and help students classify their results hastily.

The eight different values of motherDegree and fatherDegree were reduced to a single feature called parentsEducations; the possible values of this feature are three: the parents have the same academic level, the father studied more than the mother, or the mother studied more than the father.

The variables roomsInHouse and workersInFamily were combined in roomsWorkers with three possible values: the student lives in a house with a proportional quantity of people and rooms, or there are more rooms than people. Otherwise, there are more people than rooms.

Concerning the grades features, all were transformed from the original range of 0.00 to 10.00 to a classification given by a national educative institution in Mexico called Secretaría de Educación Pública (SEP, Ministry of Public Education); this organization has proposed an official grade scale; this scale considers only four groups. [4]

The student's age, from the integer original value, was delimited in a set of four variables, according to a study named "Ages of the brain" [5]

The campus, schools, and majors are number codes that represent relevant information, but the formats are serial numbers; we decided to analyze the majors and schools to get the percentages of egress and classify the schools and majors into four groups according to the Likert difficulty scale [6].

3.5 Traditional Machine Learning Algorithms

We address the problem of academic success prediction as a supervised learning problem since we have a target variable in our database indicating if the students will accomplish their majors on time. It is a binary classification task because we must predict a binary value. Table 3 shows our selected Machine Learning algorithms. All the algorithms were implemented using Python's scikit-learn machine learning library, version 1.2.0, and Keras, version 2.2.1. The experiments were performed using the 10-fold cross-validation technique. The optimal values were selected using the GridSearchCV function, which helped us find the best combination of hyperparameters.

[4] https://es.wikipedia.org/wiki/Calificacin escolar

[5] https://www.gaceta.unam.mx/las-edades-del-cerebro/

[6] https://www.marquette.edu/student-affairs/assessment-likert-scales.php

Table 1. Features of the final dataset

Attribute	Description	Normal value	Reduced values
academicPotential	Capability of a student to accomplish a major	(0% to 100%)	1: under 35% 2: 35% - 65% 3: 66% - 100%
aptitudesResult	Avg. of skills	(0% to 100%)	same previous
interestsResult	Avg. of interests	(0% to 100%)	same previous
desiredArea	The knowledge area the student was interested in, during high school	1: Sci. & Engineering 2: Biological sciences 3: Social sciences 4: Humanities & arts 5: Not decided yet	Stayed the same
fatherDegree	Highest academic level	1: No instruction 2: Primary school 3: Secondary school 4: Teacher Education 5: Technical major 6: High school 7: Bachelor 8: Postgraduate	Combined with motherDegree 1: father=mother 2: father>mother 3: father<mother
motherDegree	Highest academic level	same fatherDegree	
roomsInHouse	Total rooms in the student's house	Integer (1-10)	Combined with workersInFamily 1: rooms<workers 2: rooms=workers 3: rooms>workers
workersInFamily	Workers at house	1, 2, 3, 4, 5, 5+	
monthlyIncome	Monthly household income (Mexican pesos)	1: ≤ \$1,500 2: [\$1,501, \$3,000) 3: [\$3,001, \$4,500) 4: [\$4,501, \$6,000) 5: [\$6,001, \$8,500) 6: [\$8,501, \$11,500) 7: [\$11,501, \$14,000) 8: ≥ \$14,000	
failedSubjects	Failed subject 1st year	Integer	0,1 (1 means 1 or more)
firstYearGrade	Grade at 1st year	Float (0.00 to 10.00)	1: Failed 2: Enough 3: Satisfying 4: Outstanding
highschoolGrade	Grade at high school	Float (0.00 to 10.00)	= firstYearGrade
studentAge	Student's age	Integer	1:9-20, 2:21-29, 3:30-40
major	The major's id	Integer	Integer (1, 2, 3, 4)
knowledgeArea	The major's area	Same desiredArea's list except no. 5	
school	The school's id	Integer	
hybridClasses	The major is hybrid	Boolean (1,0)	Stayed the same
inPersonClasses	The major is onsite	Boolean (1,0)	Stayed the same
remoteClasses	The major is online	Boolean (1,0)	Stayed the same

Table 2. Optimal hyperparameters settings selected for each algorithm based on grid-search and cross-validation

Algorithm	Hyperparameters tested	Optimal Value
Artificial Neural Networks	Activation: ['relu', 'sigmoid', 'tanh'] Epochs: [10, 100, 1000] Layers: from 3 to 7	Activation: 'relu' Epochs: 100 Layers: 3
Decision Trees	Criterion: ['gini', 'entropy'] max depth: from 5 to 12	Criterion: 'gini' max depth: 10
K-nearest Neighbors	n neighbors: from 5 to 22 algorithm: ['ball tree', 'kd tree']	n neighbors:19 algorithm: kd-tree
Logistic Regression	solver: ['lbfgs', 'liblinear', 'adam']	solver: 'liblinear'
Naïve Bayes	algorithm:['Gaussian','Categorical', 'Bernoulli']	Algorithm: 'Categorical'
Random Forest	n estimators: from 50 to 150 max depth: from 5 to 15	n estimators: 100 max depth: 14
Support Vector Machines	C : [0.1, 1, 10, 100], kernel : ['sigmoid', 'rbf', 'poly']	C: 1 kernel: 'rbf'

3.6 Proposed similarity-based algorithm

Let n be a student in our training dataset. We know if n accomplished their major on time or not. To determine if a student m will accomplish their major on time, it seeks how similar m and n are according to their feature values. To determine whether m is similar to n, they must exhibit the highest degree of similarity in the most relevant features, even if m and n do not have the same exact value in each feature.

The weight of the features in our dataset was calculated using the χ^2 function.

The χ^2 statistic is calculated between each feature and the target variable. Features that are highly correlated with the target variable will have higher scores. The feature weights are given by the variable $\text{wfeature}_{i.}$

Our algorithm calculates how similar are m and n using Equation 1 and 2:

$$\text{d} = |feature(student(n))_i - feature(student(m))_i | \tag{1}$$

$$relfeature_i = wfeature_i - d * \left(\frac{wfeature_i}{maxfeature_i - minfeature_i}\right) \tag{2}$$

Where:
student (n): is any student from the training dataset
student (m): is a student to predict their result
$\text{feature(student(n))}_i$: is the corresponding value of the feature i for the student n
$\text{feature(student(m))}_i$: is the corresponding value of the feature i for the student m
wfeature_i : is the importance or weight of the feature i
maxfeature_i : is the highest possible value of the feature i
minfeature_i : is the lowest possible value of the feature i
relfeature_i : is the calculated value of the similarity between the students m and n for the feature i

For every feature i our algorithm accumulates the relfeature_i as shown in Equation 3.

$$totalScore = \sum_{i=1}^{F} relfeature_i \tag{3}$$

Where F is the number of features in our dataset
The totalScore in equation 3 represents the similarity score between the students m and n. If a student z in the training data set with totalScore is higher, we can conclude that m is more similar to z and n is discarded. To predict if student m will accomplish their majors in curricular time, our algorithm checks the target variable of student z and will assign the same value for student m.

In the case that several students u, v, w, ... in the training dataset share the same totalScore highest value, they are considered as a group G, G is composed of students who finished their major in curricular time and students who did not. Let x be the number of students who accomplished their majors and y the number of students who did not accomplish their majors in curricular time. The result of our algorithm is given by the comparison x > y or x < y.

If x > y, our algorithm returns 1, which means that the student m will accomplish their major on time; otherwise, our algorithm returns 0, which means that the student m will not accomplish their major on time.

In the case that x = y, our algorithm generates a random value (1,0). It is important to mention that the case x = y never occurred in all our experiments.

Figure 1 shows a representation of how our proposed algorithm works. We have the features from 1 to F on the x axis and their possible values on the y axis.

Every colored line S i is a student or group of students whose value of totalScore is the same. For every S_i, there are several students who approved or did not approve in curricular time; x is the number of students who finished their major on time, and y is the number of students who did not.

When we want to know if a new student or new data will finish on time or won't, our algorithm takes the new student and extracts each feature F_i from it and applies equation 1 and 2 between the new student and every S_i until calculating the maximum totalScore described in equation 3.

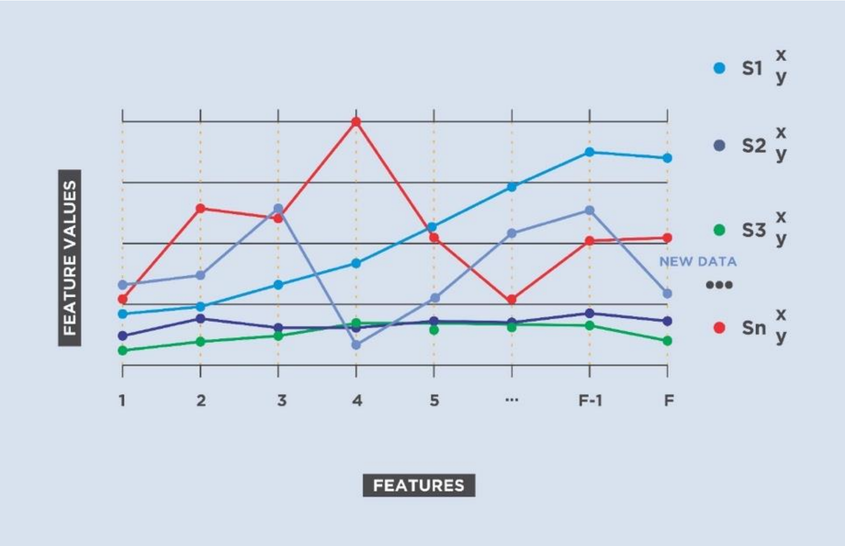

Fig. 1. Graphical representation of how our algorithm works.

Algorithm 1 presents the pseudocode of our proposed algorithm. The input data is the student's result we want to predict (line 1). Therefore, the output will be 0 if the student's prediction is negative. Otherwise, it is 1 (line 2). Other data are the training set, the weights of the features, and the minimum and maximum values of these attributes (lines 3 to 6). Lines 7 to 39 calculate the process of comparing and looking for the highest similarity among the new student and a student or group of students in the training set.

Algorithm 1: Proposed algorithm based on similarities

```
input: new student   ▷ the student to predict their result
output: {1, 0}    ▷ 1 ← new student will finish on time, 0 ← otherwise
data: training = {student 1, student 2, ..., student m}
f1w, f2w, ..., fFw                ▷ weights of the features
f1min, f2min, ..., fFmin          ▷ minima values of every feature
f1max, f2max, ..., fFmax          ▷ maximums values of every feature
procedure prediction by similarities(new student)
final group = [ ], similar = 0
for i = 0 to training.length - 1 do
    sum v = 0
    for k = 1 to F do
        distk = abs(training[i].featurek - new student.featurek)
        difk = featurekmax - featurekmin
        var = featurekw - (distk * (featurekw/(difk)))
        sum v += var
    end for
    if sum v >= similar then
        similar = sum v
        if sum v > similar then
            final group = []
            final group.append(training[i].target)
        else
            final group.append(training[i].target)
        end if
    end if
end for
j = 0, answer = 0, aprove = 0, noaprove = 0
while j < f inal group.length do
    if final group[j] == 1 then
        aprove + 1
    else
        noaprove + 1
    end if
end while
if aprove > noaprove then
    answer = 1
end if
if aprove == noaprove then
    answer = random(0, 1)
end if
return answer
```

3.7 Proposed Algorithm's complexity analysis

Let the first for which iterates over the training set T from 0 to T.length-1

T is composed of m students S, ie:
T = {S 1 , S 2 , ..., S m }

Every S has F features: feature1, feature2, ..., feature
T has length m; and for every student S in T, we get every feature from 1 to F in
constant time, so, we have $O(m) * O(F) \Rightarrow O(m * F)$

Because for m students S in T we extract every F feature.

Our algorithm contains a last loop to iterate over the set final group of size n that contains the students whose similarity is the highest with the new student. This loop takes O(n) because it evaluates if each student in final group list completed the major in curricular time. This process is performed in only a single iteration of n students.

In conclusion, the complexity of the algorithm is:
⇒ O(m * F) + O(n)
⇒ O((m * F) + n)

4 Experimental results

In this section, we present and compare the precision obtained by traditional ML algorithms (mentioned in Section 3.3) and our proposed algorithm. All the experiments were executed over the students' database without preprocessing the features (normal value) and with the combined values of features (reduced value). We also present the results of the classification models when the algorithms are trained only with the PROUNAM features or the academic records features.

The PROUNAM database contains the exam results applied in the second year of high school as a guide to support students' election majors. The number of features considered in this set is nine students' attributes. These features are: academicPotential, aptitudesResult, desiredArea, fatherDegree, interestsResult, monthlyIncome, motherDegree, roomsInHouse, and workersInFamily.

The analysis that only considers the PROUNAM database considers the previous nine features, but applying feature engineering more precisely reduced and combined feature techniques: motherDegree and fatherDegree were combined into parentsEducation and roomsInHouse and workersInFamily are combined into roomsWorkers (see Table 2).

Table 4 shows the results of executing the algorithms over the PROUNAM database. We can observe a general poor precision; every algorithm is under 60%. Furthermore, the data set with the normal values performs better than the reduced version. The highest precision is obtained using the Naïve Bayes algorithm, overcoming our proposed algorithm by 2.54%.

Table 3. Precision of known ML algorithms and our algorithm using only PROUNAM database

Algorithm	Precision - normal value	Precision – reduced value
Artificial Neural Networks	57.74%	56.80%
Decision Trees	56.00%	55.61%
K-nearest Neighbors	53.04%	53.94%
Logistic Regression	22.34%	54.74%
Naïve Bayes	58.95%	56.25%
Random Forest	57.50%	55.00%
Support Vector Machines	55.16%	54.61%
Proposed algorithm	56.41%	54.71%

The academic records database contains grades and the number of failed subjects in the first year of a Bachelor's degree. Ten students' features are considered. in this set failedSubjectsFirstYear, firstYearGrade, highschoolGrade, hybridClasses, inPersonClasses, remoteClasses, knowledgeArea, major, school, and studentAge.

Table 5 presents the classification precision when training the ML learning algorithms and our proposed algorithm only with the features from the academic records database. The column normal value shows the precision of the classification models when trained without preprocessing the features. The column reduced value shows the precision of the classification models when trained on the features with reduced values (as explained in Section 3.2).

We can observe a significant improvement using academic records instead of only applying the PROUNAM database. Besides, the precision is better when we use the reduced values in most algorithms. The highest score is obtained by Random Forest but with the features in their normal form. Our proposed algorithm has a precision of 77.47% under Random Forest by 2.64%

Table 5. Precision of known ML algorithms and our algorithm using only academic records

Algorithm	Precision - normal value	Precision – reduced value
Artificial Neural Networks	77.32%	78.52%
Decision Trees	78.94%	78.31%
K-nearest Neighbors	76.59%	77.69%
Logistic Regression	75.96%	77.12%
Naïve Bayes	77.74%	76.65%
Random Forest	80.11%	78.38%
Support Vector Machines	76.96%	78.56%
Proposed algorithm	77.47%	78.30%

Finally, we used all the features presented in Table 2 to train the traditional ML algorithms and our proposed algorithm. Table 6 shows the precision obtained by each algorithm. Our algorithm yielded better results than K-nearest Neighbors, Logistic Regression, Naïve Bayes, and Support Vector Machines. When data is combined, our proposed algorithm achieves better results than K-nearest Neighbors, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machines. When data is combined, all the algorithms improve their precision except for Decision Trees, Naïve Bayes, and Random Forest. Even though our algorithm is 2.71% under the best score precision, the context and nature of our problem require a detailed description of the behavior and characteristics of the students we want to detect who are at risk of dropping out of college. Therefore, our algorithm, which constructs a set of students to get a prediction, allows us to examine the students inside that set who are the most similar to the new student and analyze their features in common or discover if a particular pattern exists. This information allows for insights about majors or other features that could be discussed with pedagogues and policymakers and help them make better decisions to enhance the academic success of higher education students at UNAM.

Table 6. Comparison between known ML algorithms and our algorithm using both academic records and PROUNAM test

Algorithm	Precision - normal value	Precision – reduced value
Artificial Neural Networks	77.67%	78.56%
Decision Trees	78.59%	78.27%
K-nearest Neighbors	74.10%	76.70%
Logistic Regression	75.44%	76.76%
Naïve Bayes	76.88%	75.88%
Random Forest	80.30%	77.74%
Support Vector Machines	76.07%	77.19%
Proposed algorithm	77.59%	78.13%

4.1 Features Relevance Analysis

Feature importance can provide a way to rank the features based on their contribution to the final prediction; it allows practitioners to understand which features in a dataset contribute most to the final prediction in a Machine Learning model. Table 7 shows the feature relevance of our proposed algorithm and Random Forest model, which had the highest precision.

Table 4. Feature relevance of our model and Random Forest model

Our Model	Random Forest
failedSubjectsFirstYear	school
school	firstYearGrade
major	major
knowledgeArea	failedSubjectsFirstYear
firstYearGrade	knowledgeArea
hybridClasses	highschoolGrade

highschoolGrade	studentAge
remoteClasses	inPersonClasses
studentAge	hybridClasses
inPersonClasses	remoteClasses

We can appreciate that the first and last five features appear in a different order but in the same block of importance. Our similarity algorithm and Random Forest consider almost the same features to predict academic success. The five more important features are school, the number of failed subjects and grades in the first year, the studied major, and the knowledge major.

5 Conclusions and future work

Our proposed algorithm is more precise when it only considers the academic records database in its reduced form with a value of 78.30%.

After analyzing several works that propose success classification models, academics found that they obtain different performances depending on the context and the nature of the problem they face. Therefore, it is not possible to identify only a classification algorithm that stands out from the others.

Regarding the analysis of relevant characteristics to identify academic success, we observe that the number of failed subjects in the first year is one of the most significant factors in determining whether a student at UNAM will finish their major on time. Student's monthly income does not impact academic performance, similar to the outcomes obtained at Queensland University (Vallmur et.al., 2001) and Portugal (Roy et. al., 2017). Similar to Brazil (Fernandes et.al., 2019), schools and majors affect students' success significantly. In other words, there are some majors where most students succeed or fail.

We also observed in our experiments that, not surprisingly, the grade in the first year is more important than the grade obtained in high school.

Considering the results of taking both databases academic records and PROUNAM test, our proposed algorithm has shown better results than several traditional categorical Machine Learning algorithms. Our proposal improved an average score of 1.02% on five of those algorithms.

We will discard the PROUNAM test in future experiments to consider more students. The algorithms applied only using the PROUNAM database showed the lowest precision, with an average score of 52.14%. To improve student similarity, we will also test other kinds of distances in Equation 1, like Manhattan, Minkowski, and Chebyshev.

Acknowledgments

Thanks to CONAHCYT (Consejo Nacional de Humanidades, Ciencias y Tecnologías) for supporting Cinthia Rodríguez Maya with a scholarship for her Ph.D. studies. This work has been carried out with partial support from the DGAPA UNAM-PAPIIT project number IN104424 and CONAHCYT project number CF-2023-G-64.

References

Betts, J. R., & Morell, D. (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of Human Resources*, 34, 268–293.

Bolt, N. (2011). In S. Goldstein & J. A. Naglieri (Eds.), *Academic achievement* (pp. 8–9). Springer. https://doi.org/10.1007/978-0-387-79061-9

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. *Heliyon*, 6(6), e04081. https://doi.org/10.1016/j.heliyon.2020.e04081

European Commission. (2024, 10, 01). Quality education and training for all. from https://education.ec.europa.eu/education-levels/school-education/early-school leaving

Latif, A., Choudhary, A., & Hammayun, A. (2015). Economic effects of students' dropouts: A comparative study of the causes of students' dropouts globally. *International Journal of Economics, Commerce and Management*, 3(1) https://doi.org/10.4172/2375-4389.1000137

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., Erven, G. V. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, 94, 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Lecompte, D., Kaufman, L., Rousseeuw, P., & Tassin, A. (1983). Search for the relationship between academic performance and some psychosocial factors: The use of a structured interview. *Acta Psychiatrica Belgica*, 83, 598–608.

Monteverde-Suárez, D., Gonzalez-Flores, P., Santos, R., Manuel, G. M., Zavala-Sierra, I., Luz, V., & Sánchez-Mendiola, M. (2024). Predicting students' academic progress and related attributes in first-year medical students: An analysis with artificial neural networks and Naïve Bayes. *BMC Medical Education*, 24. https://doi.org/10.1186/s12909-023-04918-6

Noell, G. H., Burns, J. M., & Gansle, K. A. (2019). Linking student achievement to teacher preparation: Emergent challenges in implementing value added assessment. *Journal of Teacher Education*, 70(2), 128–138. https://doi.org/10.1177/0022487118800708

Penh, P. (2018). *Education in Cambodia: Findings from Cambodia's experience in PISA for development*.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Roy, S., & Garg, A. (2017). Predicting academic performance of students using classification techniques. In *2017 International Conference on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)* (pp. 568–572). https://doi.org/10.1109/UPCON.2017.8251112

Vallmuur, K., & Schweitzer, R. (2001). Who succeeds at university? Factors predicting academic performance in first year Australian university students. *Higher Education Research & Development*, 20, 21–33. https://doi.org/10.1080/07924360120043621

Yagci, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9. https://doi.org/10.1186/s40561-022-00192-z