Predictive Models for Early Detection of Engineering Students at Risk of a Course Failure

Andres Gonzalez-Nucamendi
School of Engineering and Science
Tecnologico de Monterrey
Cd. de Mexico, Mexico
anucamen@tec.mx

Julieta Noguez
School of Engineering and Science
Tecnologico de Monterrey
Ave. Eugenio Garza Sada 2501, Monterrey 64849, NL, Mexico.
jnoguez@tec.mx

Luis Neri
School of Engineering and Science
Tecnologico de Monterrey
Cd. de Mexico, Mexico
neri@tec.mx

Víctor Robledo-Rella
School of Engineering and Science
Tecnologico de Monterrey
Cd. de Mexico, Mexico
vrobledo@tec.

Rosa Maria Guadalupe García-Castelán
School of Engineering and Science
Tecnologico de Monterrey
Cd. de Mexico, Mexico
rmggarci@tec.mx

***Abstract*— In this study, the creation of predictive models for the detection of students at risk of failure is presented. A sample of 618 student profiles , 19% of which failed a given course, were used to detect students at risk of failing. The student profiles were determined from the constructs of Self-Regulation Learning and Affective Strategies (SRLAS) and Multiple Intelligences (MI). The first part of this work, includes an Exploratory Factor Analysis of the data. The predictive phase of the study uses nine Machine Learning classification techniques (KNN, SVM, LDA, QDA, Decision Trees, Random Forest, ADA_Boosting, XGBoosting, and Bayes) to classify students that passed or failed a given course. The *Log-Math* and the *Anxiety* dimensions turned out to be relevant variables regarding the success of the student. Decision Trees and Random forests provided the best predicting power. Our results are encouraging in the sense that the methodology followed proves successful at identifying the main factors related to student failure. This may help instructors to timely identify students at risk of failure and their possible causes, to implement appropriate strategies to mitigate this undesirable outcome.**

Keywords*— *Course failure, Educational innovation, Engineering students, Higher education, Predictive models, Machine learning

I. Introduction

The Tecnologico de Monterrey cares for each enrolled student and tiresly works to successfully lead all of them to the completion of their academic path. For this reason, studies are being carried out to determine the possible reasons that cause academic failure, which may give rise to delays in graduation and even school dropout. Course failure may be due to various factors, including academic, economic, familiar, and cognitive reasons [1]. In this study, we focus on the creation of predictive models for the detection of students at risk of failure.

In previous works, the authors focused on the determination of *student profiles* based on the constructs of Self-Regulation Learning and Affective Strategies (SRLAS) [2] and Gardner's multiple intelligences (MI) [3]. The relation between the dimensions of these two constructs and the course grades was analyzed using Machine Learning tools for building models aimed to predict student performance [4]. Also, we have validated and applied surveys to know the cognitive profile of students adapting Gargallo et al.'s SRLAS construct [2,4,5] as well as Gardner's MI construct [3,6].

Next, our team carried out a study applying basic descriptive and inferential statistics, as well as regression and correlation analyses to compare various *measures* that could be used to estimate MI and SRLAS cognitive levels [7]. This study also tackled the problem of the optimism-pessimism bias present in students' self-diagnosing. Surveys based on the SRLAS and MI constructs were applied to a total sample of 618 students (19% of which failed a given course), to obtain their student profiles. Knowing the final grade of the students in a given course on a 1–100 scale, the characteristics common to students with relatively good or poor academic performance were identified. For this task, basic statistical techniques, including principal component analysis, clustering, correlation, and multiple regression analysis were used. It was discovered that logical-mathematical intelligence and the level of self-regulation learning are important for a good general academic performance. On the contrary, the level of anxiety and the need for extrinsic motivation is an obstacle to a suitable academic performance [7]. Additionally, these results allowed us to assess the potential of these constructs in terms of generating early warnings from the student profiles to be used by the teacher at the beginning of the course [4].

Going beyond the general prediction of academic performance, a more specific study aimed to detect students at risk of failing a particular course is presented in this paper.

Consequently, the variable to be predicted is no longer continuous but a dichotomous qualitative variable (pass/fail) for which classification methods usually give better results than regression methods.

The hypotheses of this study are:

1. It is possible to identify the key factors leading to course failure by engineering undergraduate students.
2. Through the application of different classification techniques, it is possible to obtain a good prediction of student failure.

In this work, an Exploratory Factor Analysis of the data is included. In the predictive part, Machine Learning classification techniques are used to classify students that passed or failed a given course. The specific techniques considered are KNN, SVM, LDA, QDA, Decision Trees, ADA_Boosting, XGBoosting, Random Forest, and Bayes. The results of this research may help instructors to timely identify students at risk of failure and their possible causes, to implement early appropriate strategies to mitigate this undesirable outcome.

The organization of this work is as follows. Related work is presented in Section II, while the methodology is presented in Section III. Results and analysis are shown in Section IV. Discussion is presented in Section V. Finally, Section VI presents the main conclusions and future work.

II. RELATED WORK

Previous research aimed at predicting academic achievement employed a variety of machine learning algorithms, including multiple, probit, logistic regression, neural networks, and C4.5 and J48 decision trees. However, the incorporation of learning analytics techniques that involve simultaneous analysis of student performance data has improved the accuracy of predictive models in recent years (e.g., [8]).

Student data obtained during their registration, as well as environmental factors to determine students at risk of failure, were used by [9]. The study adapted three data mining methods, namely, random forest, logistic regression, and artificial neural network algorithms to improve the estimation accuracy, including the possibility of classifying students by risk levels.

Graduate and undergraduate students' learning abilities, study habits, and academic interaction characteristics were used by [10] applying a Multilayer Perception Neural Network Model to determine if students were at risk regarding their academic performance. This study reports that classification accuracy of 85% was achieved. The researchers concluded that the proposed model could be used to determine whether the student could be academically successful.

Other authors, such as [11], developed a model that uses student grades obtained from the inter-semester period, as well as their demographic characteristics. In that study, classification models based on the Gradient Boosting Machine (GBM) were used to predict academic performance. The results showed that the factors that most influenced performance estimation were the performance scores of the previous year and the number of absences. The authors found that demographic characteristics, such as neighborhood school and age could also affect student success or failure.

Recently, efforts have been increased to apply machine learning techniques to improve predictive models' accuracy in academic performance. Random forests [12], genetic programming [13], and Naïve Bayes algorithms were used in recent studies [14]. The prediction accuracy of these models reaches very high levels. Nevertheless, the prediction accuracy of student academic performance requires a deep understanding of the factors and features that impact student results and student achievements [15].

Curricular analytics techniques were used by [16] to understand the educational trajectories of students. In particular, their study defines an educational trajectory as a process, i.e., a set of events that occur in a certain order during the permanence of the students in a given program. They used Process Mining to study the dynamic behavior of educational trajectories to identify which educational trajectory patterns lead to late dropout and which ones do not.

A new model based on machine learning algorithms (random forests, support vector machines, logistic regression, Naïve Bayes, and *k*-nearest neighbor) is proposed by [17] to predict the final exam grades of undergraduate students, taking their midterm exam grades as the source data. The results showed that the proposed model achieved a classification accuracy of 70–75%. This study presents a contribution to the early prediction of students at high risk of failure and determines one of the most effective machine learning methods.

The following section describes the methodology used in this study to obtain predictive models to identify students at risk of failing at an early stage. It is based on the determination of student profiles according to the constructs of Self-Regulation Learning and Affective Strategies (SRLAS), and multiple intelligences (MI) and considering students' previous grades.

III. METHODOLOGY

This section describes the details of the dataset, pre-processing techniques, and machine learning algorithms employed in this study.

A. Data identification and collection

As reported in the first part of this investigation [7], data were obtained through the application of MI and SRLAS questionnaires to 1,693 undergraduate engineering students enrolled mainly in mathematics, physics, and software engineering courses. The students answered online the adapted MI and SRLAS questionnaires, where they also provided personal information for statistical purposes (gender, age, major, etc.). After the completion of the questionnaires, the system assigned values for each of the dimensions of the MI and SRLAS constructs (each comprising eight dimensions or variables) used to define their *student profile,* as it is shown in Tables 1 and 2 below. Nevertheless, not all the students in the initial sample provided complete answers, and all incomplete entries were excluded from the data analysis. Therefore, the final

database was reduced to $N = 618$ students. This database was compared with the final grades provided by our institution's Student Services Department.

TABLE 1. GARDNER'S MULTIPLE INTELLIGENCES (MI) [7]

VARIABLE	DESCRIPTION
Lin	Sensitivity to spoken and written language, ability to learn languages, and capacity to use language to accomplish certain goals. Includes the ability to use language effectively to express oneself rhetorically or poetically. Language as a mean to remember information.
LogMath	Capacity to analyze problems logically, carry out mathematical operations and investigate issues scientifically. Ability to detect patterns, reason deductively, and think logically. Associated with scientific and mathematical thinking.
Mus	Involves skill in the performance, composition, and appreciation of musical patterns. Capacity to recognize and compose musical pitches, tones, and rhythms.
BodKin	Potential of using one's whole body or parts of the body to solve problems. Capacity to use mental abilities to coordinate bodily movements.
Spa	Potential to recognize and use the patterns of open space and more confined areas.
Inter	Capacity to understand the intentions, motivations, and desires of other people, enabling people to work effectively with others.
Intra	Capacity to understand oneself, to appreciate one's feelings, fears, and motivations. It involves having an effective working model of oneself and being able to use such information to regulate one's own life.
Nat	Enables human beings to recognize, categorize, and draw upon certain features of the environment; combines a description of the core naturalistic ability with the role characterization that many cultures value.

TABLE 2. SELF-REGULATION LEARNING AND AFFECTIVE STRATEGIES (SRLAS) [7]

Variable	Description
IntMot	The degree to which students are intrinsically motivated and have the self-confidence to study autonomously and with enthusiasm. It includes those factors that depend on each individual and can be controlled by the student.
ExtMot	The degree to which students depend on external factors to focus on their learning.
FitMood	The degree to which students seek to maintain optimal physical and mental states to obtain better learning outcomes.
Anx	Degree of anxiety that students display while working on their academic activities.
SelfReg	The degree to which students know their academic strengths and limitations, and make the best decisions to improve their performance. It includes taking control of their learning process, time and space, and adapting to course conditions.
SocInt	The degree to which students value collaborative work and seek the support of colleagues and teachers to face academic difficulties.
InfSearch	The degree to which students have adequate strategies to search, discriminate and select relevant information for their courses.
InfProc	The degree to which students have adequate strategies to organize and process course information and to transfer this information to other courses or contexts.

B. Reliability analysis of student profile questionnaires

The reliability of the MI and SRLAS questionnaires was assessed by estimating Cronbach's α-value for each dimension. This value is a measure of the degree of correlation among the questions for a specific dimension.

Reliability tests of the adapted questionnaires were also performed with a sub-sample of 248 students before applying them to the entire student sample. The average Cronbach's α-value was 0.950 for the eight MI dimensions and 0.792 for the eight SRLAS dimensions, which are satisfactory values, and are presented by [4], [6], and [5], respectively.

C. Data balancing

As shown in Fig. 1, the dataset on which the predictive study was carried out is moderately unbalanced because it contains 81% of approved students and only 19% of failed students. The latter is the target of our prediction.

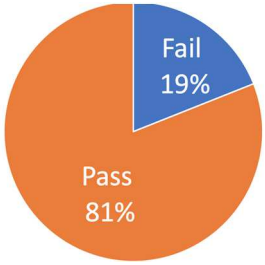


FIGURE 1. THE IMBALANCE BETWEEN THE NUMBER OF *PASS* AND *FAIL* FLAGS WITHIN OUR STUDENT SAMPLE

D. Establishing Data Mining (DM) model and algorithm implementation

For this study, different data mining techniques were used, in three main processes:

(a) Factor analysis

Factor analysis is a statistical method used to describe the variability between observed correlated variables in terms of a potentially smaller number of unobserved variables called *factors*. To discover the implicit factors in the database with 16 variables (8 in each construct), *Minitab* was used to perform a factorial analysis using the Varimax rotation method, which minimizes the number of variables with high values and factor loadings, facilitating thus the interpretation of the results. The algorithm was run for two, three, and four factors (see below).

(b) Predictive power of each variable

Density graphs of each of the MI or SRLAS variables were used for those students who passed and failed the course. An analysis of the predictive power of each of the 16 variables included in the two constructs was performed. To facilitate the comparison, the superimposition of these density graphs of some of the most significant variables for those who passed and failed the course are shown in Section IV below. The programming was done with Python.

(c) Classification Algorithms.

Nine classification algorithms were applied using Python's *Scikit-Learn* library.

1. For the SVM classifier, the *rbf kernel* was applied.

2. For the KNN classifier, the *KNeighborsClassifier* was applied, which is a more widely used technique than its alternative *RadiusNeighborsClassifier*.

3. For the Tree classifier, the *DecisionTreeClassifier* method was combined with the default parameters, while the visualization was carried out with the *graphviz* software imported from the Python code programmed for this task.

4. The execution of the Random Forest classifier was carried out through the *RandomForestClassifier* class by manuvering default parameters.

Classification using classical discriminant analysis was performed through:

5. *LinearDiscriminantAnalysis* (LDA) method.

6. *QuadraticDiscriminantAnalysis* (QDA) method.

7. The ADA_Boosting classifier was executed using the *AdaBoostClassifier* class using 90 estimators of the Tree class.

8. The XGBoosting classifier was executed using the *GradientBoostingClassifier* class with the parameters n_estimators = 150, min_samples_split = 2, and random_state = 0, and finally,

9. Bayesian classifier was executed using the *sklearn.naive_bayes.GaussianNB* class.

IV. Results and Analysis

The main results of the different techniques applied in the three main processes are shown below:

A. Factorial analysis.

As indicated above, a factor can be seen as a list of variables with their factor loadings that weigh the importance of its dimensions. The factorial loading of each variable is its projection on an axis of a new coordinate system obtained from a certain rotation of the original reference system. Only the absolute value of the factorial loadings matters to deduce the importance of each variable in the factor. In Table 3, the most representative variables of each factor are marked with bold and asterisks. Tables 3 and 4 focuses only on the four outstanding variables of each factor, instead of the complete list of factor loads.

The factor analysis showed interesting results. When using two factors, it is observed in Table 1 that the two constructs are conveniently separated to form two factors: the MI Factor, where the highest factor loads belong to multiple intelligences dimensions, and the SRLAS factor, where the dimensions of that construct have the highest factor loadings.

Analyzing three factors, as shown in Table 3, results in a *mental structure* factor, composed of the *LogMath*, *Intra*, *Lin*, and *BodKin* dimensions. Factor 2, which we can call *motivational discipline* or *motivational self-regulation*, is represented by the dimensions *SelfReg*, *InfProc*, *InfMot*, and *InfSearc*, all belonging to the SRLAS construct. Finally, factor 3, which we could call the *student's concerns* factor, is made up of the *ExtMot* and *Anx* dimensions of the SRLAS construct and the *Inter* and *Spa* dimensions of the MI construct.

TABLE 3. Two factors: Full factor loadings

Variable	Factor1	Factor2
Intra	0.686	-0.447
Inter	0.314	**-0.556***
Lin	0.616	**-0.523***
Spa	0.277	**-0.577***
LogMath	0.653	**-0.582***
Mus	0.447	-0.248
BodKin	0.512	-0.5
Nat	0.472	-0.391
IntMot	**0.701***	0.408
ExtMot	0.187	0.227
FitMood	0.607	0.274
Anx	0.102	0.23
SelfReg	**0.804***	0.454
SocInt	0.554	0.266
InfSearch	**0.715***	0.328
InfProc	**0.704***	0.29

TABLE 4. Formation of three factors. Only the 4 most important variables are shown

Variable	Factor1		Variable	Factor2		Variable	Factor3
LogMath	0.829		*SelfReg*	-0.868		*ExtMot*	-0.678
Intra	0.812		*InfProc*	-0.839		*Inter*	0.640
Lin	0.778		*IntMot*	-0.831		*Spa*	0.478
BodKin	0.697		*InfSearch*	-0.730		*Anx*	-0.478

When the algorithm was applied to form four factors, as shown in Table 5 below, the result was similar to that of three factors, with the difference that factors 3 and 4 were a split in the variables of MI and SRLAS of factor 3 that had resulted when the formation of three factors was sought.

B. Predictive Power.

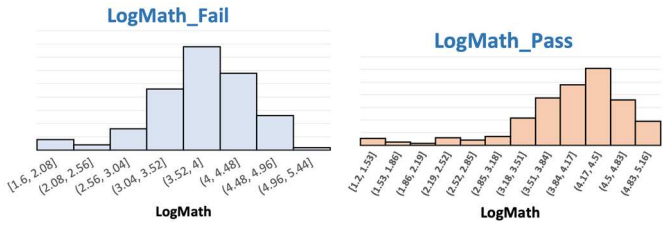
To facilitate the visualization of the pass/fail predictive power of each specific MI or SRLAS variable, the histograms of the *fail* vs. *pass* categories can be compared. These diagrams allow a better visualization through the density functions, which are an analogy to discrete probability distributions, but instead of being bound to the interval [0,1], they can take values in the interval $[0, \infty)$. Density estimates are closely related to

histograms, but they are endowed with properties such as smoothness and continuity that facilitate comparison. An example of the *LogMath* dimension is shown in Fig. 2, where the *x*-axis corresponds to the intensity of this dimension (from 1 to 5) in the student profile defined above.

TABLE 5. FORMATION OF 4 FACTORS. ONLY THE 4 MOST IMPORTANT VARIABLES ARE SHOWN

Var	Factor 1	Var	Factor 2	Var	Factor 3	Var	Factor 4
LogMath	0.850	*SelfReg*	-0.880	*ExtMot*	0.695	*Inter*	0.860
Intra	0.807	*IntMot*	-0.840	*Anx*	0.681	*Spa*	0.414
Lin	0.762	*InfProc*	-0.820	*Inter*	-0.270	*ExtMot*	-0.260
BodKin	0.684	*InfSearch*	-0.740	*Spa*	-0.250	*Lin*	0.206

Through the superposition of density functions, a more detailed description of the differences in fail/pass performance is achieved than that shown by correlations or regressions. For example, from Figure 2 it can be affirmed that if a student has a *LogMath* intelligence above 4, it is very likely that she will pass because the "pass" curve is above the "fail" curve. Moreover, if her *LogMath* level is between 2.5 and 4, the probability that she will fail is greater, because the "fail" curve now lies above the "pass" curve. Finally, if her *LogMath* intelligence is less than 2.5, the probabilities of passing or failing are similar, because both curves are quite similar in this range of values. These last students can be considered atypical because they are too few and too far from the previous behavior.



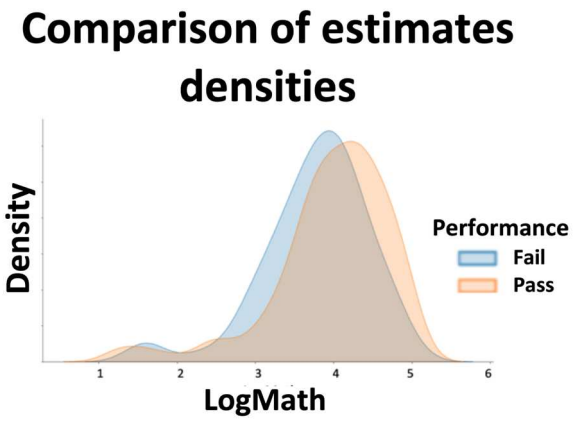


FIG 2. EXAMPLE OF HISTOGRAMS OF OBSERVED DATA AND COMPARISON OF THEIR ESTIMATED DENSITIES, FOR THE *LOGMATH* VARIABLE.

Fig. 3 (above) shows the distribution density for the linguistic variable (*Lin*). It is observed that in the central part, with values between 3.5 and 4.5, the probabilities of passing are greater than the probabilities of failing, while with values between 1.5 and 3.0, the probability of failing is greater. Fig. 3 (below) shows the distribution density for the Anxiety variable (*Anx*). It is interesting to observe that the more anxiety a student presents, the greater the probability that she will fail, and vice versa.

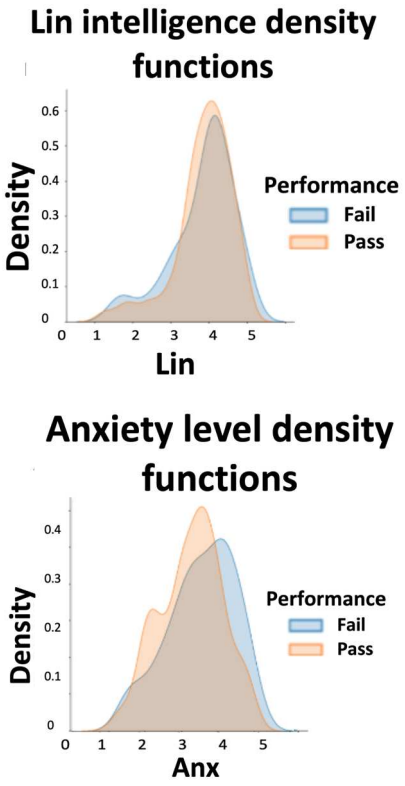


FIGURE 3. COMPARISON OF ESTIMATED DENSITIES OF LIN Y ANX

In this way, the predictive power of each of the 16 variables used in this study was analyzed. Tables 6 and 7 present the dimension values ranges according to Pass/Fail probabilities obtained observed from the curves corresponding to all variables of the MI and SRLAS constructs, respectively.

TABLE 6. MI VARIABLES VALUES

Variable	Prob(Pass) > Prob(fail)	Prob(Fail) > Prob(Pass)	Prob(Pass) ≈ Prob(fail)
LogMath	4 to 5	2.5 to 4	1 to 2.5
Lin	3.2 to 4.2	1.5 to 3.2	1 to 1.5; 4.2 to 5
BodKin	1.5 to 2.5	3.8 to 4.8	1 to 1.5; 4.8 to 5
Intra	2.5 to 3.7	3.7 to 5	---
Inter	3.8 to 5	1 to 3.8	---
Spa	1.5 to 2.5; 3.5 to 5	2.5 to 3.5	---
Mus	---	---	1 to 5
Nat	3.2 to 4.5	4.5 to 5	1 to 3.2

C. Classification Algorithms.

To obtain an individualized prediction of the students' academic performance, different classification techniques were applied, as mentioned above: SVM, KNN, Decision Trees, Random Forest, ADA_Boosting, SX_Boosting, Bayesian classification, and Discriminant Analysis (LDA and QDA). In Table 8 below a comparative study of the derived results is shown.

TABLE 7. SRALS VARIABLES VALUES

Variable	Prob(Pass) > Prob(fail)	Prob(Fail) > Prob(Pass)	Prob(Pass) ≈ Prob(fail)
IntMot	4.1 to 5	1 to 4.1	---
ExtMot	1 to 3.6	3.6 to 5	---
Anx	1 to 4	4 to 5	---
SelfReg	3.5 to 4.7	1 to 3.5	4.7 to 5
FitMot	---	---	1 to 5
SocInt	---	---	1 to 5
InfSearch	3.2 to 4.2	---	1 to 3.2, and 4.2 to 5
InfProc	4 to 5	2.5 to 3.5	1 to 2.5, and 3.5 to 4

Table 8 was built from the confusion matrix of each algorithm. The Global Precision refers to the total proportion of well-classified cases and the Global Error is its complement, that is, the total proportion of classification errors. *Pass Precision* refers to the proportion of students who passed the course and were classified correctly out of the total number of students who passed. Finally, *Fail Precision* refers to the proportion of failed students who were correctly classified, concerning the total number of failed students.

TABLE 8. PREDICTORS EVALUATION

	Global Precision (%)	Global Error (%)	Pass Precision (%)	Fail Precision (%)
SVM_rbf	80.65	19.35	100	0
KNN	68.39	33.2	75.20	40.00
Decision Trees	74.19	25.8	79.20	**53.33**
Random Forest	**83.87**	16.13	96.00	33.33
ADA_Boosting	80.65	19.35	95.16	22.58
XGBoosting	81.94	18.06	92.00	40.00
Bayes	72.90	27.10	0.896	3.30
LDA	80.00	20.00	99.20	0
QDA	78.71	21.29	96.60	16.70

V. DISCUSSION

It was possible to corroborate hypothesis 1 through the identification of the relevant variables using the factor analysis technique, as well as by visualizing the predictive power of each variable, and obtaining density functions.

Regarding the prediction variables used by other authors we can observe that most of the studies use variables that inform about scholarships, internet access, number of students in the courses, previous academic history, mathematics level, socioeconomic level, location of the school, number of units enrolled, the experience of teachers, nationality, political context, gender, and age, among others [17]. Studies were also found that use fewer variables on which there may be early intervention, for example, variables that describe study habits and learning abilities [14] or learning strategies, coping strategies, and cognitive factors [18]. However, our study uses student profiles based on the constructs of Self-Regulation Learning, and Affective Strategies (SRLAS), and multiple intelligences (MI), and considering their grades in previous courses.

To corroborate hypothesis 2, it is important to mention that most studies on the academic results of students have focused on various objectives such as the prediction of the final grade, and the prediction of failure or dropout. In this regard, [17] presents an interesting table with a list of authors indicating the objectives pursued by each of them, the variables they use, the school level of the sampled population, the sample sizes, the algorithms used, and their respective precisions. Therefore, it was possible to apply and compare various classification techniques to obtain a good measure of the prediction of student failure.

It is relevant to highlight that in classification problems, global precision is not always the best measure to evaluate the results of the classification algorithms, especially when the categories to be predicted are very unbalanced. The above generates a useless rule, since it assigns everyone to the majority class, giving a global error equal to the proportion of the majority class, and providing a false perception of the generosity of the algorithm.

Therefore, in the present study, we have chosen instead to present the four precision discriminators shown in Table 8. It should be clarified for the proposed objective, that the main discriminator for failure prediction is "*Fail precision*". However, if one is not careful enough, one could still get a useless rule that indicates an accuracy at 100%, at the cost of a large overall inaccuracy and a 0% prediction on the Pass prediction.

Taking into account the above considerations, the choice of the best classification algorithm requires a thorough study. Based on our objective, it can be stated that *Decision Trees* was the best classification algorithm because it is the best Fail predictor with an accuracy of 53%, as shown in Table 8. However, its global accuracy is not the best (74.2%), and it is still incorrectly predicting 25.8% of the students who fail because its' Pass accuracy is 79.2%.

On the contrary, considering only the global accuracy, the best algorithm is Random Forest, with 83.9%, which incorrectly classifies only 4.0% of the students who passed because its Pass accuracy is 96.0%. However, Random Forest is only correctly classifying 33.3% of the students who failed. Is this what it is sought? We study this issue further in a forthcoming paper.

The final decision corresponds to the institution and department, considering that a bad prediction of failed students leaves students at risk of failing without the timely opportunity to receive special attention. On the other hand, false classification of possible failed students, who are not actually at risk of failing, could use institutional resources for a task that may be unnecessary.

Finally, it is important to mention that the next stage of this research is expected to show the possible causal relationships between the classification of each student and their main characteristics. That is, to show the variables that most influence a given student to be classified as approved or failed, which is essential for the design of interventions that help prevent his/her failure.

Some of the algorithms used in this work allow the identification of the predictor variables and their values. *Decision Trees* can show specific classification rules in terms of the levels of the predictor variables, and *Random Forest* can provide the set of the most relevant variables ordered according to their importance when doing the classification. In future work, we will be addressing these tasks in the next stage of the research so that the results obtained so far have practical use.

VI. Conclusions and Future Work

In this paper, we have presented new advances in the research agenda that seek to improve student learning. This is the continuation of several studies carried out previously, which started with the design, validation, and implementation of instruments to estimate the dimensions of the constructs of multiple intelligences (MI) and self-regulation learning and affective strategies (SRLAS) of our students. The new contributions presented in this paper consist of (a) a different exploratory analysis where the Factorial Analysis is used and (b) the analysis of the predictive power of the variables through probability density functions before the use of classification algorithms.

One of the advantages of the classification approach outlined in this work compared to our previous approach, which consisted of Principal Component Analysis (PCA) and Clustering, is that we advance in the customization of the methodology, because through the *predicting methods* a more accurate individualized prediction for student success or failing a course is possible.

Future work will require the need to program a better balance between Pass and Fail accuracies, which could be achieved if the probability threshold for assigning the Failed or Pass status is modified from its default 50%-value. This would help in the decision-making of educational institutions, which can use the methodology outlined in this research to provide timely academic support to students at risk of failing a course.

For the next stage of this research, it is also planned to carry out studies on the efficiency of the interventions to avoid failures.

Acknowledgment

The authors acknowledge financial support from the *Writing Lab, and the Institute for the Future of Education*, from the Tecnológico de Monterrey, México, in the production and presentation of this work.

References

[1] Cagliero L., Canale L., Farinetti L., Baralis E., & Venuto E. (2021) Predicting Student Academic Performance by Means of Associative Classification. *MDPI, Appl. Sci.* 2021, 11, 1420, https://doi.org/10.3390/

[2] Gargallo López B, Suárez J, Pérez-Pérez C. The CEVEAPEU questionnaire. An instrument to assess the learning strategies of university students. *RELIEVE-Rev Electrón de Investig Eval Educ* 2009; *15*(2):1–31.

[3] Gardner H. Frames of mind: The theory of multiple intelligences. 3rd ed. *Hachette* UK; 2011.

[4] Noguez J, Neri L, González-Nucamendi A, Robledo-Rella V. (2016). Characteristics of self-regulation of engineering students to predict and improve their academic performance. In: 2016 IEEE Frontiers in education conference (FIE). 2016, p. 1–8. http://dx.doi.org/10.1109/FIE.2016.7757479.

[5] Neri Vitela L, Noguez Monroy J, & Alanis Funes G. (2015). Validación de instrumento para determinar habilidades de autorregulación de los alumnos. In: *Memorias CIIE*. Tecnologico de Monterrey, 2015, p. 994–949.

[6] Noguez Monroy J, Escárcega Centeno D, Escobar Castillejos D. (2015). Validación de instrumento para inteligencias múltiples y estrategias de aprendizajes. *In: Memorias CIIE. Tecnologico de Monterrey* 2015, p. 1000–1005.

[7] Gonzalez-Nucamendi A., Noguez J., Neri L., Robledo-Rella V. García Castelan R.M. Escobar-Castillejos D.E. (2021). The Prediction of Academic Performance using Engineering Student's Profiles. *Computers and Electrical Engineering*. Special Issue VSI-tei. ISSN: 0045-7906. https://doi.org/10.1016/j.compeleceng.2021.107288

[8] Amare M.Y. & Simonova, S. (2021). Global challenges of students' dropout: A prediction model development using machine learning algorithms on higher education datasets, *The 21st International Scientific Conference Globalization and its Socio-Economic Consequences,* SHS Web Conf., 2021, Volume 129

[9] Hoffait, A., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*(2017), 1–11. https://doi.org/10.1016/j.

[10] Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. *Bulletin of Education and Research, 40*(3), 157–164.

[11] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, 94 (2018), p.p. 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

[12] Zabriskie, C., Yang, J., DeVore, S., & Stewart, J. (2019). Using machine learning to predict physics course outcomes. Phys.. Rev. Phys. Educ. Res., 15(2), 020120. https://doi.org/10.1103/PhysRevPhysEducRes.15.020120

[13] Banzhaf, W., Nordin, P., Keller, R. E., & Francone, F. D. (1998). Genetic programming: an introduction: on the automatic evolution of computer programs and their applications. *Morgan Kaufmann Publishers Inc*. ISBN: 1-55860-510-X

[14] Ornelas, F., & Ordonez, C. (2017). Predicting student success: A naïve Bayesian application to community college data. *Tech., Knowledge and Learning,* 22(3), 299–315. https://doi.org/10.1007/s10758-017-9334-z

[15] Alshanqiti, A., & Namoun, A. (2020). Predicting student performance and its influential factors using hybrid regression and multi-label classification. *IEEE Access*, 8, 203827–203844. https://doi.org/10.1109/access.2020.3036572

[16] Salazar-Fernandez J.P., Sepúlveda M., Munoz-Gama J., Nussbaum M. (2021). Curricular Analytics to Characterize Educational Trajectories in High-Failure Rate Courses That Lead to Late Dropout. *Applied Sciences*. 2021, 11, 1436. https://doi.org/10.3390/app11041436

[17] Yağcı M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments* (2022) 9:11

[18] Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: A machine-learning approach. Higher Education, 80(5), 875–894. https://doi.org/10.1007/s10734-020-00520-7