Open Access

ARTÍCULO RESULTADO DE INVESTIGACIÓN
Copyright ©
Educación y Humanismo

EDUCACIÓN Y HUMANISMO

ISSN: 0124-2121
E-ISSN: 2665-2420
http://revistas.unisimon.edu.co/index.php/educacion
Educación y Humanismo 26(47): pp. 1-19. Julio-Diciembre, 2024
https://doi.org/10.17081/eduhum.26.47.6379

Predictive model for the classification of university students at risk of academic loss

Modelo predictivo para la clasificación de estudiantes universitarios en riesgo de pérdida académica

Recibido: 04-07-2023 Aceptado: 27-11-2023 Publicado: 10-02-2024

María Gamboa-Mora
Universidad Nacional Abierta y a Distancia, Bogotá, Colombia
Correspondence author: maria.gamboa@unad.edu.co

Felix Vivián Mohr
Universidad La Sabana, Chía, Colombia

Vicky Ahumada-De La Rosa
Universidad Nacional Abierta y a Distancia, Bogotá, Colombia

Sulma Vera-Monroy
Universidad La Sabana, Chía, Colombia

Alexander Mejía-Camacho
Universidad de Cundinamarca, Fusagasugá, Colombia

Abstract

For higher education institutions, predicting the risk of academic loss is a priority issue due to the resources invested by institutions, students and the academic community in general. **Objective**: the objective of this research was to propose a suitable model that allows predicting students who are at risk of academic loss in a chemistry course. **Methodology**: the quasi-experimental, predictive, longitudinal research was developed with data from 103 students from four Colombian universities. To build the model, a comparison of five algorithms was implemented. Data was processed with Jupyter-Python. **Results**: the logistic regression model (LR) was built based on the students' results on the Saber 11 test (Colombian nation-wide university admission exam), in which the penalty of false positives with different weights from the false negatives improved the performance of the model. **Conclusions**: it is concluded that LR is substantially better than grasping or a guessing approach, furthermore, it was shown to perform better than a neural network model.

Keywords: Logistic regression model, academic loss, comparison, chemistry course, higher education.

How to cite this article (APA): Gamboa-Mora, M., Mohr, F.V., Ahumada-De La Rosa, V., Vera-Monroy, S. & Mejía-Camacho, A. (2024). Predictive model for the classification of university students at risk of academic loss. *Educación y Humanismo*, 26(47), pp. 1-19. https://doi.org/10.17081/eduhum.26.47.6379

Resumen

Para las instituciones de educación superior, predecir el riesgo de pérdida académica es un tema prioritario debido a los recursos invertidos por las instituciones, los estudiantes y la comunidad académica en general. **Objetivo:** el objetivo de esta investigación fue proponer un modelo adecuado que permita predecir a los estudiantes que están en riesgo de pérdida académica en un curso de química. **Metodología**: la investigación cuasi-experimental, predictiva y longitudinal se desarrolló con los datos de 103 estudiantes de cuatro universidades colombianas. Para construir el modelo se implementó una comparación de cinco algoritmos. Los datos se procesaron con Jupyter-Python. **Resultados:** el modelo de regresión logística (LR) se construyó con base en los resultados de los estudiantes en la prueba Saber 11 (examen nacional colombiano de admisión a la universidad), en el que la penalización de falsos positivos con pesos diferentes a los falsos negativos mejoró el rendimiento del modelo. **Conclusiones:** se concluye que LR es sustancialmente mejor que un enfoque codicioso o de adivinanzas, además, se demostró que funciona mejor que un modelo de red neuronal.

Palabras clave: Modelo de regresión logística, pérdida académica, comparación, curso de química, educación superior.

Introduction

Context and motivation

Predicting the risk of academic loss of students has become a priority for all higher-education institutions (Cheema, 2014; Ndirika and Njoku, 2012; Hall and Mechan, 2000), the purpose is to be able to identify students who are in this situation in their different courses to implement support strategies that serve to facilitate learning. To this regard, Gamboa et al. (2020) showed that Colombian universities implement tutorials, monitoring, social work, Olympics, educational websites, study groups and collaborative-work forums, as innovative and successful strategies which favor training processes.

Prediction processes are complex, in recent years different studies have been reported in which students have been adequately classified around various aspects, for example: Lee and Chung (2019) used a decision tree model to predict students who are at risk of dropping out of school; on the other hand, Son and Fujita (2019) applied a multi-adaptive neuro-fuzzy inference system with representative sets (MANFIS-S) to predict academic performance in university courses, involving socioeconomic and academics variables, like Yang et al. (2018) who applied a multiple linear regression model (MLR) combined with a principal components analysis (PCA), in the same way as Waheed et al. (2020) used the artificial neural network model (ANN).

In engineering faculties, predictive studies of early approval classification have been reported, such as the work of Miguéiset et al. (2018), who used decision trees with 95% precision. On the other hand, Ranjeeth et al. (2021) used data mining and machine learning to predict academic performance and make corrective decisions to assist students, which resulted in a classification close to 80% precision.

In chemistry, Alhadabi and Karpinski (2020) worked on a model to predict academic performance in relation to self-efficacy in learning chemistry during the first semester of university courses, proving the increase in the positive effect on the goals achievement and obtaining good evaluated the factors associated with the loss of a general chemistry university course, using logistic regression, some significant predictors were used such as: mathematics scores in the massive SAT tests - Scholastic Aptitude Test, which are requirements to be accepted into a university, in addition to the grade point average in the secondary education.

In this sense, in Colombia, there is a challenge to consolidate a predictive model to classify university students at risk of academic loss (Gamboa et al., 2020; Vargas and Ardila, 2019; Rodríguez et al., 2018; Gamboa, 2014) making use of the information collected by ICFES (Colombian Institute for Education Evaluation) whose mission is to evaluate through standardized tests the competencies achieved by students at different educational levels, as stipulated by the Ministry of National Education (2004). In the transition from high school to higher education, students must take the Saber-11 tests, which enables higher education institutions to have information on the competencies of admitted applicants (ICFES, 2018).

As that the Colombian educational system constitutes a responsibility of universities to assume the training and graduation of students, it becomes necessary to make the decision to choose which students should benefit from institutional support, which is a sensitive matter. It is generally not desirable that all the students receive support service. Not only would this imply substantial fixed cost for the institutions, but also it would also be an unnecessary burden in terms of time for students who in fact don't need to be supported (Suárez-Montes and Díaz-Subieta, 2015). Due to this, it is worth proposing a suitable regression model that allows predicting which university students of first year are at risk of academic loss in any course, based on the results of the standardized Saber 11 tests in order to rationalize institutional resources, those of the students and those of the educational community (Peña and González, 2022; Ávila et al., 2021; Junca, 2019).

Prediction model

The task of predicting whether a student is at risk can be considered as a binary classification problem (Tai Chui et al., 2020). Importantly, the problem is not to predict a concrete grade (which would be regression) but a risk property that has previously been well-defined by the institution. For example, a student may be (a-posteriori) at risk if his achieved grade is only slightly above the threshold to pass the course.

Unfortunately, it is often not possible in practice to reliably predict whether a student is at risk (Beaulac and Rosenthal, 2019). The general problem is that the predictor variables that are typically available do not allow for a (even close to) separability between the two classes.

Method

This product is derived from the project entitled Variables associated with academic performance in the subject of Chemistry in four Colombian universities, was approved in the call for projects funded by National Open and Distance University of Colombia – UNAD under Code PG15-2019, developed during the period 2020-2021 with the cooperation of the Universidad de La Sabana, the Universidad de Cundinamarca (UdeC), the Universidad Distrital Francisco José de Caldas (UDFJC) and the Universidad Nacional de la Patagonia Austral in Argentina (UNPA).

A total of 246 participants from UNAD, La Sabana, UdeC, and UDFJC universities signed the F-11-5-1 consent form, which was approved by the ethics committee of UNAD, in accordance with Resolution 8430 of 1993, Article 11, and the provisions of Colombian Law 1581 on Data Protection of 2012.

Data collection

This paper corresponds to a case study developed with 103 students enrolled in the general chemistry course in four Colombian universities, in the 2020-2 period, framed in a predictive and longitudinal research.

The data of the participants from the PG15-2019 project, who voluntarily shared their results from the Saber 11 test, were included in the model. The sample used in this study was of the intentional type, employing a quasi-experimental method. Information was collected by compiling an ex post facto database that included data on students' Saber 11 test scores and their performance in the chemistry course.

Tools

To predict the mode a comparison of five algorithms was implemented. Data was processed with Jupyter-Python.

Results

This work uses an approach based on logistic regression (LR) to predict the student risk (Coussement et al., 2020). LR is arguably a good choice for the problem at hand compared to other classification schemes (Heredia et al., 2014). First, since it would be desirable to specify the risk of a student in terms of probabilities, support vector machines (SVM), which do not natively deliver probabilities, can be dismissed. Besides, LR can be extended much better to several classes (if different risk stages shall be distinguished) compared to SVM, which are by default only binary classifiers. Second, since a roughly monotone behavior between the relevant predictor variables and the grade (and hence the risk) can be expected, a linear approach seems less prone to over-fitting compared to decision trees. As well, decision tree algorithms do not so directly cater for asymmetrical losses since, in general, they don’t directly optimize the loss function but rather purity-related metrics such as information gain or the Gini index. Third and in contrast to recent approaches relying on neural networks, LR (in fact being a trivial neural network) requires orders of magnitude less training data to build a solid model. Since LR optimizes the cross-entropy, the asymmetric loss can be directly integrated by assigning different weights to the instances of the two classes (Salmerón-Pérez et al., 2010; Ramos et al., 2018).

Risk prediction with asymmetrical loss - asymmetric error rate and cross-entropy

The error rate (or equivalently, the accuracy) is, by a large margin, the most common metric for binary classification on balanced datasets. The error rate estimates the generalization performance of a learner by computing the loss

$$L = \frac{1}{n}\sum_{i=1}^{n} \quad [y_i \neq \hat{y}_i] \tag{1}$$

where y_i it's true and $\hat{y}_i$ is the predicted label of the i instance of some validation set of size *n*.

Preparing the dataset

It is assumed that the data basis consists of a table with one column specifying the final grade (number) and all other attributes being predictor variables. In this work, it is additionally assumed that all attributes are scaled into the unit interval via range normalization. Range normalization does not affect the correlation between the attributes and the grade but is specifically relevant if the original data contains categorical attributes, which need to be converted to a 0-1-valued Bernoulli encoding; without the scaling, the categorical attributes might be ignored by the learner.

The target column is then transformed into a binary column by evaluating each value against some threshold τ. The threshold τ is the grade that is considered to separate students that were at risk from those that were not. For instance, if grades are between 0 (worst) and 5 (best) and students pass with a 3, it might be sensible to set τ to 3.5.

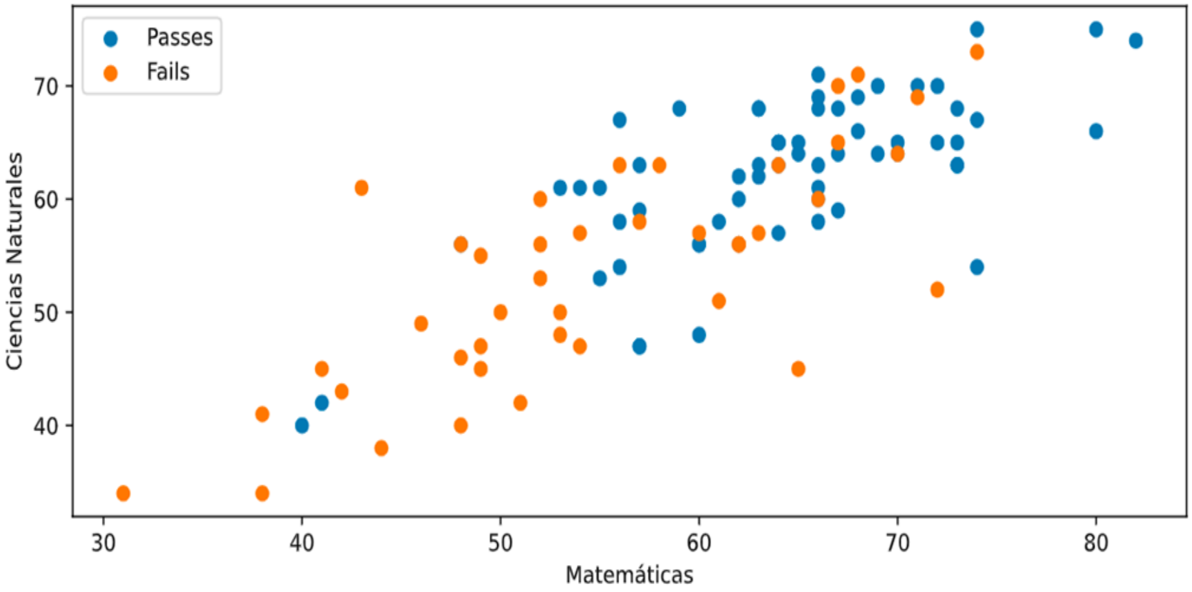
Since LR has native support for multi-class classification, it is conceivable to introduce a series of thresholds and define different levels of risk characterized by different (monotonic) threshold values $\tau_1, .., \tau_n$. If respective costs $c_1, .., c_{n+1}$ for misclassifying the classes are available, the cross entropy can be just extended in the common way. However, since the case study in this work only considers the case of two groups, this extension is not considered or explained in more detail.

The following presents the results for the analyzed dataset. The dataset consists of 103 instances, each of which described through five numerical attributes and associated with a grade between 0 (worst) and 5 (best). The numerical values have been range-normalized and hence reside in [0, 1]. Choosing a risk threshold τ = 3.5 yields a slightly imbalanced dataset with 60% of the instances being negative and the rest being positive. The considered values of c are between 0.25 (false positives 4 times worse than false negatives) and 4 (false negatives 4 times worse than false positives, which are consistent with the results on Harada (2020).

Fig. 1 shows the case of a core module in chemistry taught in four universities in Colombia. The figure gives an overview based on the two attributes with the highest linear correlation with the grade; orange points reflect students at risk. As can be clearly seen, it is not possible to linearly separate the two groups even only close. Of course, it may occasionally happen that the two groups can be, for specific courses, highly separable.

However, despite some "clean" results, the much more realistic situation is that the two groups have a substantial intersection in the input space.

Given the fact that it is not typically possible to predict with certainty whether a student is at risk, it is important to clarify that there is also a typical asymmetry in the cost associated with inaccurate predictions (Bai et al., 2020). Arguably, it is more severe to predict that a student will pass the course while he will in fact fail it than if failure is predicted and he would pass it (even without institutional support). On the other hand, being overly conservative and predict a risk even for unlikely cases will result in unreasonable support costs (Soo et al., 2021). Hence, a middle ground prediction solution is required. While this asymmetric assessment should lead to an asymmetric loss function used to evaluate a predictor, existing work has apparently not covered this aspect.



Note. Data processed in Jupyter-Python.

Figure 1
Prediction model for chemistry courses in four Colombian universities. *Non-Separability*.

Comparison of learning algorithms

The comparisons between the different classifiers is shown in Table 1 and Fig. 2. Here, we report only mean values, because the standard deviations are fairly small; so standard deviations are not shown to maintain readability. Table 1 gives high level results of the asymmetric cross entropy loss for c = 0.25, c = 1, and c = 4. In Fig. 2, the left plot shows the scores in terms of the original loss function in Eq. (2) for the different values of c. The right plot shows the scores in terms of the (asymmetric) cross entropy loss in Eq. (4). Note that the results in the plot assume that LR, DT and SVM were configured to use class weights according to c, which is not the case for the ANNs, which do not have support for this in scikit-learn (even though this is not a conceptual limitation of ANNs).

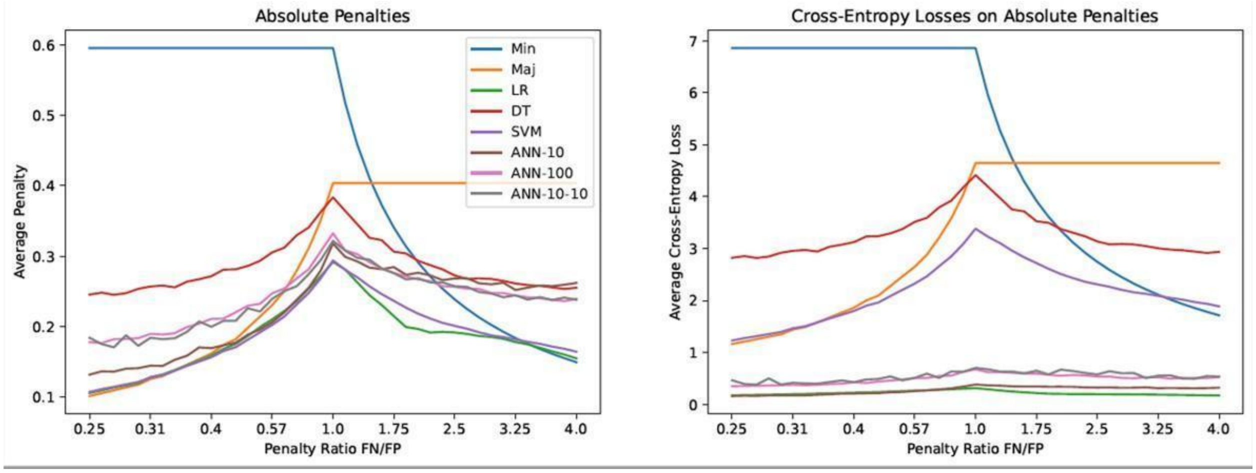
Table 1
Comparison of learning algorithms. Values are mean asymmetric cross entropy losses as in Eq. (4). Lower values are better.

algorithm	c = 0.25	c = 1	c = 4
Min	6.86	6.86	1.72
Maj	1.16	4.65	4.65
LR	0.18	0.31	0.17
DT	2.79	4.38	2.9
SVM	1.23	3.38	1.89
ANN-10	0.16	0.38	0.31
ANN-100	0.35	0.68	0.5
ANN-1010	0.35	0.73	0.53

The first observation is that the majority and minority classifiers behave exactly as expected. In the case of the evaluation in this paper, the negative class is the majority class. That means that adopting the majority classifier will predict the negative class, and this curve will increase with increasing c, because, for small values of c, little importance is given to false negatives, and hence a very low error is observed. As c tends to 1, the error for the majority class attains its maximum, because from that point on, the penalty for a false positive is always 1 (and 1/ c for false positives). Since the majority classifier does not create false positives, its performance is stuck at the fraction of the minority class. The same interpretation holds for the minority classifier, which starts off from the majority class fraction and starts dropping from 1 onward as false negatives get punished more.

An important even though not surprising observation in the left plot is that the minority and majority approach are in fact performing best as the value of c diverges from 1. That is, as c tends to 0 from the right, the majority classifier becomes the globally best model whereas the minority classifier becomes the best as c tends to infinity (and in fact already for quite small values of c like 4. This has a natural explanation: as c approaches 0, false negatives are penalized less, so making those errors is every time less problematic. In other words, if the condition tells that those false negatives are essentially irrelevant, then no prediction model is even needed anymore, but all students can be simply treated as negative cases. Once again, this is essentially the same as pretending that there are (almost) only instances of one of the classes. Similarly, as c grows, this essentially means that false positives are irrelevant compared to false negatives, and in this case one can simply treat all students as positive cases. So, what this means is that there is a need for a discriminating model only if the value of c is rather close to 1, i.e., if false positives and false negatives should both be avoided.

Interestingly, the dominance of the trivial classifiers is not reflected on the right plots in the cross-entropy. The reason for this is simply that there are some models, like LR and neural networks, that make the same mistakes as the trivial classifier but with much more uncertainty. This uncertainty plays out well for those classifiers because they still assign a substantial (even though insufficient) probability to the right class and are penalized much less than the trivial classifier, which assigns minimal probability to the right class. However, this is a matter of scale, and for more extreme values of c, this difference ultimately disappears.



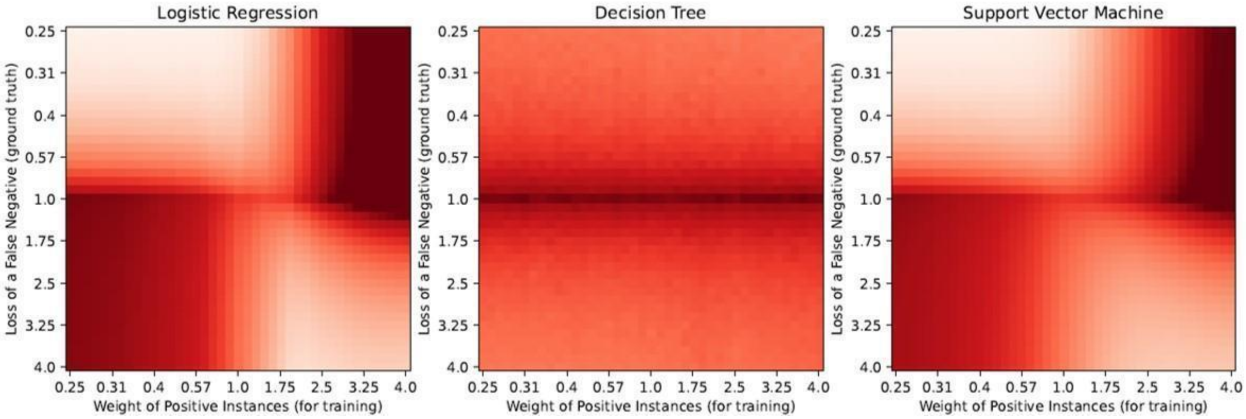
Note. Data processed in Jupyter-Python.

Figure 2

Comparison of learning algorithms.

One main result that can be observed in the figure is that there is no algorithm that is able to out-perform logistic regression in any of the two metrics and any value of c under consideration. On small values of c, the majority classifier obtains a marginal advantage, but LR is essentially on par. In terms of cross-entropy, a 10-unit neural network obtains almost the same performance as LR but does never outperform it. In this sense, the results confirm the initial motivation to choose logistic regression being a simple model that well considers the asymmetric loss and is tailored towards the generally monotone behavior of the data.

Next, the neural networks (ANN) and the support vector machine (SVM) show very interesting behavior in the results figures. While the SVM performs well in terms of the average penalty and substantially outperforms the ANNs in this metric, it is exactly the other way around for the asymmetric cross- entropy loss. It is easy to explain why the SVM performs bad in terms of cross- entropy, because, as for the trivial algorithms, no probabilities are produced, so the SVM is heavily punished for being very sure of false predictions. However, it is not so clear why neural networks are performing so well in terms of cross- entropy. One possible explanation for this is that, while making more severe factual mistakes on average, they are often just a glimpse away from making the correct prediction and hence are not so strongly penalized by the cross-entropy. The final observation is that decision trees are in no sense a competitive model here. This might be somewhat surprising, but there are at least two possible explanations. First, one could adscribe it to the fact that decision boundaries are needed that are not parallel to the axes, which will imply that the decision tree creates too fine granular areas. Second, the purity in the middle of the data is often poor, which might lead to a significant overfitting of the tree to regional patterns in order to achieve high purity on them. In practice, this purity cannot be achieved due to a real overlap of the classes, so the decision tree learns regions that are just by chance pure in the training data but not in reality.



Note. *Data processed in Jupyter-Python.*

Figure 3
Comparison of penalties for different true and anticipated values of C.

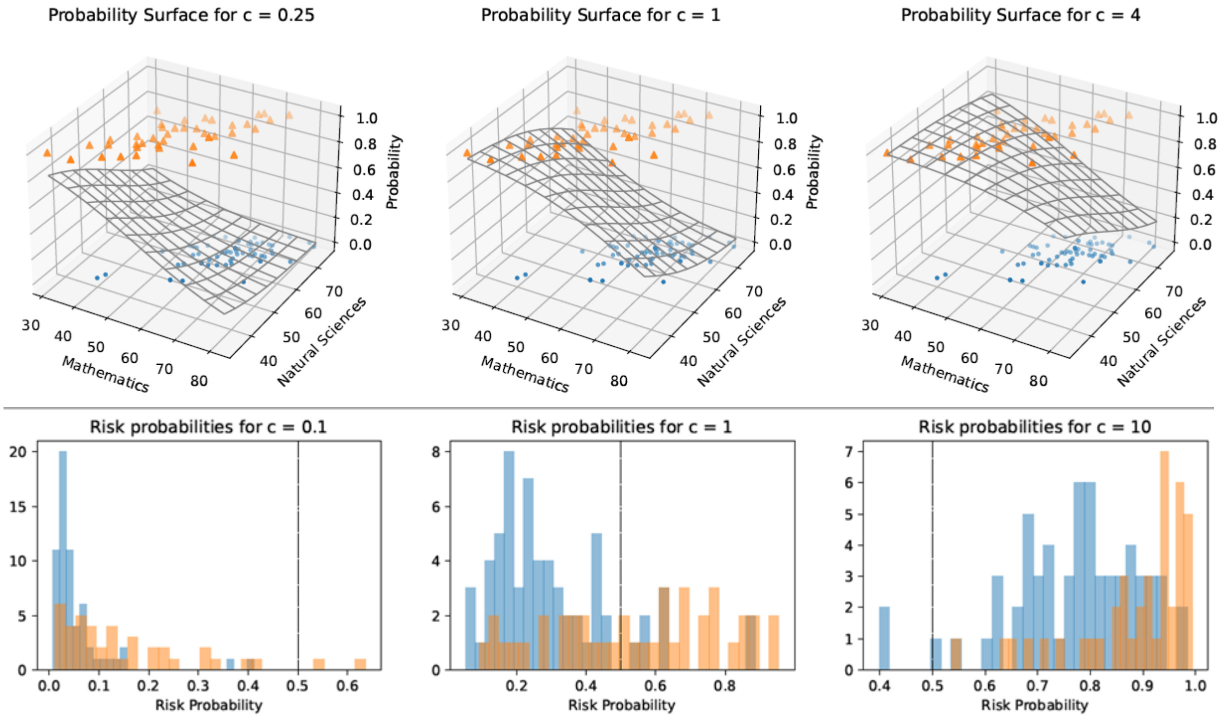
In order to assess the ability of LR to react to the asymmetrical loss conditions, the penalties are shown for different combinations of true and anticipated penalties. These results are shown in Fig. 3. In this figure, the darker the color, the higher the prediction error. For completeness the behavior is reported for all three learners that are configurable in this sense.

Two main observations can be made here. First and most obviously, the DT algorithm does not perform well no matter how the parameters are adjusted. It only starts to perform slightly less bad for values of c deviating from 1, but this is only because the problem then gets simpler and predicting the majority/minority class is less severe. Second, both LR and SVM are clearly able to take the asymmetric loss into consideration well. This can be recognized from the light areas in the diagonal quadrants. Likewise, both can be strongly misleading if asymmetry is assumed but in the wrong sense. Third, not taking asymmetry into account, i.e., pretending that false positives and false negatives are equally bad, is in fact not even too bad on the examined dataset. In the plot, this corresponds to the vertical line in the center, which still has rather reasonable values for both LR and SVM. In fact, looking at the upper half of that vertical line and comparing it to the values in the upper left quadrant, there seems to be almost no difference between adapting or not adapting to the true value of c if the true value is smaller than 1. However, the message is clearly not that adaptation does not matter; it does matter, but the degree to which it matters might also depend on the concrete data and circumstances.

Creating a belief model

While the left of Fig. 2 suggests that SVMs are a very competitive model for this type of predictions, the big advantage of logistic regression is that it can create probabilistic models of whether students are at risk. This is a huge practical advantage, because these probabilities can be used to rank students based on these probabilities and, in this way, to create a prioritization among them.

Fig. 4 shows example probability distributions in the input for our data for three different values of c together with the respective histograms for the probabilities. The orange triangles are the students at risk, and the blue points are the students not at risk. The grids in the upper row show the probability distribution learned by the LR for different values of c. For a value of $c = 0.25$, false positives are considered four times as bad as false negatives, which results in a higher inclination of the model to predict the negative class. On the other hand, a value of $c = 4$ implies that the model must be well certain to predict the negative label, because false negatives are penalized heavily.



Note. Data processed in Jupyter-Python.

Figure 4
Probability distributions and histograms.

The histograms in the bottom row show, per class, how many students are associated with which probabilities of being at risk; the vertical line is a visual aid to recognize the 0.5 decision boundary. For example, in the case of $c = 0.25$, only few students are classified positively (are on the right side of the vertical line) and almost all of them are indeed positive. There are a couple of false negative cases, but since $c < 1$, these are not weighted so strongly. On the other hand, for a case of $c = 4$, there are few false negatives, only five. There is a high number of false positives here, but since $c = 4$, these are much less important than potential false negatives, which here have been tried to avoid if “reasonably” possible. The case of c = 1 shows a middle-ground compromise, which gives a balanced picture of false positives and false negatives.

Preliminary experiments showed that for values of c = 0.1 or c = 10, the model essentially makes no mistake on the heavily penalized side anymore but in fact does also not make almost any prediction for that class anymore. In other words, for values of c close to 0 or much larger than 1, the LR model effectively degenerates to a constant predictor, which is exactly what is expect based on the discussions in the introduction.

Discussions and conclusions

Risk probability vs. Risk threshold

The above definition of the dataset implies that the probabilities produced by the probabilistic predictor for an unseen student are not probabilities of failing a course but probabilities of being at risk (Gladshiya and Sharmila, 2021; Gazdula and Farr, 2020; Olaleye and Vincent, 2020). This difference is very important for the interpretation of predictions. Depending on the configuration of the model, a probability of 95% of being at risk is maybe not at all a certain criterion of indeed failing; this probability could be for example only 50%. In other words, a high probability produced by a probabilistic model simply means that a student falls with high probability into the risk group, and it might be advisable to examine the case of that student a bit further (Robinson et al., 2019; Planinic et al., 2019; Ene and Ackerson, 2018).

The above observation is particularly crucial if the approach is applied in a rather conservative regime. In the above example with grades between 0 and 5, a conservative choice of τ would maybe be 4.0. In that case, a high predicted probability only means that there is a high probability that the student will achieve a performance of less than 4.0, but since there is no estimate on the particular grade, no immediate conclusions about the risk of indeed failing the exam can be drawn.

At this point, two important advantages of probabilistic classifiers over non-probabilistic classifiers like SVMs should be highlighted. First, the user gets a notion of uncertainty of the model. If most students have a high or a low probability, then the model was able to separate well the two cases. Otherwise, the two groups are harder to separate. Second, the probabilities can be used to decide in a fine granular way which cases should be examined further. Without probabilities, it is not possible to distinguish between two students at risk at the prediction level. In order to achieve a priorisation of the students, the analyst then must manually look at the risk group and create one (Fay and Negangard, 2017). In contrast, LR produces probabilities, which directly allows to sort the students by their probability of being at risk (Deri et al., 2018). Note that while, in principle, it would be possible to include all students within the list and even consider students with a risk of less than 50% as students at risk, there is no theoretical justification for doing so: the LR model has, by definition, found the 50% threshold that minimizes costs.

With respect to asymmetric loss risk prediction, the error rate can be inappropriate for several reasons. First, if there is one highly over-represented class, then learners have little incentive to ever predict any of the other classes. For binary classification problems, this has led to metrics such as the F-measure or the ROC measure [Faw04], which aim to demand high rates of correct predictions for both classes. Specifically, the latter has also been used in the context of student drop-out prediction [BST12, TIS^{+}14]. For multi-class classification, the cross-entropy has been considered a meaningful alternative for the error rate in the case of unbalanced datasets; the cross-entropy is discussed below in more depth (Wang et al., 2020).

However, simply treating the two classes in a more democratic way is maybe not even the intention of the decision maker who is more concerned about different costs of different prediction errors. And in fact, if the true concerns about, say, a false positive, are of small importance, then there is no point in assigning too much weight to it. In the simplest case, the decision maker may assign a different penalty to different mistakes. Formalizing this through a generalization of Eq. (1) leads to

$$L_C = \frac{1}{n}\sum_{i=1}^{n} C_{y_1}[y_i \neq \hat{y}_i] \tag{2}$$

where c_j is the cost of predicting a class other than j when j would be the correct answer and $c = (c_1, \cdots, c_k)$ is a vector that fully describes this customization of the error rate. In the standard case, it is just that $c_j = 1$ for all classes j, but making this factor explicit allows controlling the importance of different classes by the user.

Clearly, introducing the factor c_j does nothing else than simulating a change in the class distribution. Assuming i.i.d. data, the standard loss just considers each class with its "natural" frequency while the weighted loss can be seen to compute the standard loss on an "adjusted" class distribution.

An important advantage of the asymmetric error rate over the other metrics is that it is easier to adjust learners to optimize for it than for, say, the F-measure or ROC (Zois et al., 2019). In other words, since c is under the control of the decision maker, the algorithms can be easily adapted to work under the conditions. A simple and crude mechanic would be to simply over-sample the classes until the distribution in the data reflects c. However, many learning techniques optimize the maximum log-likelihood of the data, which is equivalent to minimizing the cross-entropy, into which the vector c can be incorporated directly. It is now explained in detail how to achieve this codification.

The use case of risk predictions calls not only for binary predictions but more importantly for probabilities of a student of being at risk. That is, rather than only an extreme prediction $\hat{y}_i \in \{0,1\}$, a probabilistic *belief* $\hat{y}_i \in [0,1]$ of the model that the student is at risk is demanded. In those cases, the error rate seems inappropriate, and a typical measure used for this is the cross-entropy:

$$L = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} log\, p(y_j|x_i) \tag{3}$$

in which (x_i, y_i) are the pairs in the database used for validation, y_i has been expanded to a Bernoulli vector e_l where the true label of x_i is the l-the one, and $p(y_j|x_i)$ is the estimated probability of class y_j for instance x_i.

The cross-entropy can be easily extended to cope with asymmetric cost, because wrong predictions of one class can be weighted higher than those of other classes. In the binary case, the situation is particularly simple, because this can be just encoded using some constant c that specifies the factor by which, say, false positives are worse than false negatives. Incorporating this into the above equation yields an adjusted loss function

$$L_c = -\frac{1}{n}\sum_{i=1}^{n} \sum_{j=1}^{k} c_j y_{ij} \, log \, p(y_j|x_i) \tag{4}$$

Just like for the asymmetric error rate, it is easy to see that the weighted cross-entropy is nothing else than the simulation of standard cross-entropy on a different distribution. As such, basically any learner can be used to optimize for the asymmetric cross-entropy if a sampling procedure is applied previously to obtain the desired distribution (Vargas et al., 2022).

On the suitability of logistic regression

Logistic Regression (LR) is arguably a very natural candidate to address the above problem and as matter of fact, the optimization problem of LR is precisely formulated as the problem to minimize Eq. (3), and every implementation can be easily adapted to minimize for Eq. (4); in fact, this is already implemented in standard machine learning libraries like scikit-learn. So LR seems like a canonical algorithm to address the above problem (Niu, 2020).

If it is believed that a non-linear decision boundary is required, neural networks are a natural option to consider next. It is common practice to use a softmax activation function in the output layer and to compute errors through the (point-wise) cross-entropy. This is essentially equivalent to logistic regression with the only difference that hidden layers can be used to automatically learn an advantageous non-linear feature mapping (Tsiakmaki et al., 2020). The experiments of this paper could not identify this necessity for the used data, but depending on the available features, such a mapping could prove useful. However, in the light of the bias-variance trade-off, it seems better to stick with a simple model, which is robust and also with few data points, unless there is evidence that a better model can be learned with a neural network.

Other models like decision trees, a Bayesian approach or support vector machines seem to be less suitable for this task. The only of these models that also optimizes the cross-entropy is the Bayesian approach, but it does not take any advantage of a monotonic behavior of the target variable in the inputs, which is precisely what is exploited with LR and it also needs to make assumptions about the (joint) distribution of predictor variables. While often a Gaussian is used for this purpose, the data often does not justify such an assumption. Decision trees and support vector machines do not optimize cross-entropy, and the latter do not even produce native probabilities (Selwyn et al., 2021; Joshi, 2020). In other words, while basically every learning algorithm could be used to learn a model on the type of the considered data, LR seems to be the best fit. It basically has no significant disadvantages over the other approaches but a number of advantages such as robustness (few examples required to obtain reasonable models), native production of probabilities, and immediate optimization of the (asymmetric) cross-entropy loss.

The discussion of suitable models is particularly important in the light of the recent rise of Automated Machine Learning (AutoML). AutoML has recently been used to predict course failures (He et al., 2021; Tsiakmaki et al., 2021; Mohr et al., 2021; Mohr et al., 2023). However, one problem of AutoML tools is that they typically use a nested cross-validation to evaluate candidates and assess the final performance. This is clearly a problem if little data is available, which is often the case in the context of student risk prediction, in which it is not uncommon that datasets have less than 100 instances. Using AutoML tools in such situations is problematic from a methodological viewpoint. In order to avoid problems of data snooping, it is, in this particular case, better to previously argue for a suitable model and stick to that one. Experiments can then *confirm* in the aftermath that such choice was good, but without having these results any effect on the choice of the model. This is exactly what this paper does.

To improve comparability between different values of c, the penalties are normalized in all cases so that always the highest penalty is 1. For example, in the last case, instead of penalising false positives with 1 and false negatives with 4, false positives are penalized with 0.25 and false negatives with 1. In this way the average overall penalty is also always between 0 and 1 and not, say, sometimes between 0 and 1 and sometimes between 0 and 4, which would hurt the interpretability of the result figures.

The LR approach is compared against five other algorithms. First, most simplistic, the majority (MAJ) and minority (MIN) classifier are considered, respectively, as demonstrated by Ashraf et al. (2020) in his writing. Next, three neural networks (ANN) (once with one hidden layer of 10 units, once with two hidden layers of 10 units each, and once with one hidden layer of 100 units) according to the works of Lau et al. (2019), a decision tree (DT) like the one used by Park and Dooris (2020), and a support vector machine (SVM) (Burman and Son, 2019) are considered. The DT and SVM can be configured in scikit-learn with specific instance weights, so that the same logic as for LR is applied.

All approaches are evaluated based on a 100-times repeated hold-out cross- validation with 50% training data. That is, of the 103 data points, 52 are used for training and 51 for validation of predictions. All of the 100 splits are drawn based on stratified sampling to preserve the class distribution in the folds.

In each validation run, two metrics are being measured. The first measure is the penalty, i.e., the asymmetric predictive loss as per Eq. (2). The second measure is the asymmetric cross-entropy measure (Gill at al., 2019) as per Eq. (4). The cross-entropy loss allows looking not only at the correctness of final decisions but also takes into account the certainty the model has about those assessments. Notably, some models don't produce probabilities, like the SVM or the majority/minority classifier, and others may assign a 0 probability to the correct label, e.g., decision trees. In such cases, the cross-entropy would degenerate to infinity, so the probability of the classes is "corrected" to assign a probability of at least 10^{-5}. This gives rise to cross-entropy scores of up to about 11.5 per instance.

Recommendations

Due to the nature of the data and the complexity of the analyzed context, there are limitations that make it difficult to propose an absolutist model; instead, this work builds a risk prediction model that allows to associate, with relative ease, a student profile with the probability of needing some academic support from a higher education institution, in such a way as to reduce academic loss in the chemistry course.

Also, it was possible to build a logistic regression model that adequately classified risk groups from the asymmetry cross-entropy values obtained from the scores achieved by the students in the natural sciences and mathematics components applied by the ICFES in Colombia. Finally, in the risk model, it is important to establish penalties that allow to adequately differentiate the real scenarios on which it can be implemented in a higher institution, this conception will allow a better articulation between the results and the institutional decision-making.

It is recommended to continue with the validation of the model, implementing tests with different samples to corroborate the efficiency.

References

Alhadabi, A., & Karpinski, A.C. (2020). Grit, self-efficacy, achievement orientation goals, and academic performance in university students. *International Journal of Adolescence and Youth, 25*(1), 519-535. https://doi.org/10.1080/02673843.2019.1679202

Ashraf, S., Saleem, S., Ahmed, T., Aslam, Z. and Muhammad, D. (2020). Conversion of adverse data corpus to shrewd output using sampling metrics. *Visual Computing for Industry, Biomedicine and Art, 3*(1), 1-13. https://doi.org/10.1186/s42492-020-00055-9

Ávila, L. K., Ospino, E., & Páez, A. J. (2021). *Análisis de resultados de las pruebas saber 11 implementando técnicas de minería de datos [Analysis of Saber 11 test results by implementing data mining techniques].* Universidad del Norte. http://hdl.handle.net/10584/9877

Bai, R., Zhang, C., Wang, L., Yao, C., Ge, J., & Duan, H. (2020). Transfer Learning: Making Retrosynthetic Predictions Based on a Small Chemical Reaction Dataset Scale to a New Level. *Molecules, 25*(10), 2357. https://doi.org/10.3390/molecules25102357

Beaulac, C., & Rosenthal, J. S. (2019). Predicting University Students' Academic Success and Major Using Random Forests. *Research in Higher Education, 60,* 1048–1064. https://doi.org/10.1007/s11162-019-09546-y

Burman, I., & Som, S. (2019). Predicting students academic performance using support vector machine. *Amity international conference on artificial intelligence (AICAI)*: 756-759. IEEE. https://doi.org/10.1109/AICAI.2019.8701260

Cheema, J. R. (2014). The Migrant Effect: An Evaluation of Native Academic Performance in Qatar. *Research in Education, 91*(1), 65-77. https://doi.org/10.7227/RIE.91.1.6

Coussement, K., Phan, M., De Caigny, A., Benoit, D., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems,* 135, 1-13. https://doi.org/10.1016/j.dss.2020.113325

Deri, M., Mills, P., & McGregor, D. (2018). Structure and Evaluation of a Flipped General Chemistry Course as a Model for Small and Large Gateway Science Courses at an Urban Public Institution. *Journal of College Science Teaching, 47*(3), 68–77. https://doi.org/10.2505/4/jcst18_047_03_68

Ene, E., & Ackerson, B. (2018) Assessing learning in small sized physics courses. *Physical Review Physics Education Research, 14*(010102), 1-21. https://doi.org/10.1103/PhysRevPhysEducRes.14.010102

Fay, R., & Negangard, E. (2017). Educational Case. Manual journal entry testing: Data analytics and the risk of fraud. *Journal of Accounting Education, 38*, 37-49. https://doi.org/10.1016/j.jaccedu.2016.12.004

Gamboa, M. (2014). *La evaluación externa en el área de Ciencias a través de las pruebas masivas a gran escala TIMMS y PISA. Análisis del desempeño de los estudiantes colombianos y españoles*. Universidad Distrital Francisco José de Caldas y Universidad Nacional Abierta y a Distancia. https://repository.unad.edu.co/bitstream/handle/10596/2792/9789588832692.pdf?sequence=4&isAllowed=y

Gamboa, M., Ahumada, V., Vera-Monroy, S., Mejía-Camacho, A., & Romero, J. C. (2020). Estudio de las variables asociables al rendimiento académico en la asignatura de Química en cuatro universidades colombianas. Universidad Nacional Abierta y a Distancia. https://doi.org/10.22490/9789586517454

Gazdula, J., & Farr, R. (2020). Teaching Risk and Probability: Building the Monopoly Board Game In to a Probability Simulator. *Management Teaching Review, 5*(2), 133-143. https://doi.org/10.1177/2379298119845090

Gill, H. S., Khehra, B. S., Singh, A., & Kaur, L. (2019). Teaching-learning-based optimization algorithm to minimize cross entropy for Selecting multilevel threshold values. *Egyptian Informatics Journal*, *20*(1), 11-25. https://doi.org/10.1016/j.eij.2018.03.006

Gladshiya, V., & Sharmila, K. (2021). A HML-EVC Model for Analyzing the Risk of the Students to Predict the Success Probability in the Field of Education. In: *10th International Conference on System Modeling & Advancement in Research Trends (SMART)*, 341-344. https://doi.org/10.1109/SMART52563.2021.9676327

Goyal, M., & Vohra, R. (2012). Applications of Data Mining in Higher Education. *International Journal of Computer Science Issues*, *9*(2). https//10.17148/IJARCCE.2020.9124

Hall, K., & Marchan, P. (2000). Predictors of the Academic Performance of Teacher Education Students. *Research in Education, 63*(1), 89-99. https://doi.org/10.7227/RIE.63.9

Harada, T., (2020). Learning from success or failure?–Positivity biases revisited. *Frontiers in Psychology, 11*, 1627. https://doi.org/10.3389/fpsyg.2020.01627

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems,* 212, 1-27. https://doi.org/10.1016/j.knosys.2020.106622

Heredia, J. J., Rodríguez, A. G., & Vilalta, J. A. (2014). Predicción del rendimiento en una asignatura empleando la regresión logística ordinal [Predicting Performance in a Subject Using Ordinal Logistic Regression]. *Estudios Pedagógicos, XL*(1), 145-162. http://dx.doi.org/10.4067/S0718-07052014000100009

Instituto Colombiano para Evaluación de la Educación – ICFES. (2018). *Guía de orientación Saber 11. 2019-1* [Colombian Institute for Educational Evaluation - ICFES. (2018). Orientation Guide, Saber-11 Test. 2019-1. ICFES publishing]. ICFES https://www.icfes.gov.co/documents/20143/193560/Guia+de+orientacion+saber+11+de+2019.pdf/13d64150-fa02-9062-8bb8-dcee660607c5

Joshi, A. V. (2020). Decision Trees. In: Machine Learning and Artificial Intelligence. *Springer, Cham*. 53-63. https://doi.org/10.1007/978-3-030-26622-6_6

Junca, J. A. (2019). Desempeño académico en las Pruebas Saber 11 [Academic performance in the Saber 11 tests]. *Panorama Económico, 27*(1), 8-38. https://doi.org/10.19053/01211129.v30.n58.2021.13823

Lau, E. T., Sun, L., & Yang, Q. (2019). Modeling, prediction and classification of student academic performance using artificial neural networks. *SN Applied Sciences, 1*(9), 1-10. https://doi.org/10.1007/s42452-019-0884-7

Lee, S., & Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction. *Applied Sciences, 9*(15), 3093. https://doi.org/10.3390/app9153093

Miguéis, V. L., Freitas, A., García, P., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115*, 36-51. https://doi.org/10.1016/j.dss.2018.09.001

Ministerio de Educación Nacional (MEN). (2004). *Estándares básicos de competencias en Ciencias Naturales y Sociales. Formar en ciencias: ¡El desafío! Lo que necesitamos saber y saber hacer* [Basic standards of competencies in Natural and Social Sciences. Science Education, the challenge! What we need to know and know how to do]. MEN. https://www.mineducacion.gov.co/1759/articles-81033_archivo_pdf.pdf

Mohr, F., Wever, M., Tornede, A., & Hüllermeier, E. (2021). "Predicting Machine Learning Pipeline Runtimes in the Context of Automated Machine Learning," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9), 3055-3066. https://doi.org/10.1109/TPAMI.2021.3056950

Mohr, F., Wever, M. (2023). Naive automated machine learning. *Machine Learning, 112*(4), 1131-1170. https://doi.org/10.1007/s10994-022-06200-0

Ndirika, M. C. and Njoku, U. J. (2012). Home Influences on the Academic Performance of Agricultural Science Students in Ikwuano Local Government Area of Abia State, Nigeria. *Research in Education, 88*(1), 75-84. https://doi.org/10.7227/RIE.88.1.7

Niu, L. (2020). A review of the application of logistic regression in educational research: common issues, implications, and suggestions, *Educational Review, 72*(1), 41-67. https://doi.org/10.1080/00131911.2018.1483892

Olaleye, T., & Vincent, O. (2020). A Predictive Model for Students Performance and Risk Level Indicators Using Machine Learning. In: *2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS 2020)*, 1-7. https://doi.org/10.1109/ICMCECS47690.2020.240897

Park, E., & Dooris, J. (2020). Predicting student evaluations of teaching using decision tree analysis. *Assessment & Evaluation in Higher Education*, *45*(5), 776-793. https://doi.org/10.1080/02602938.2019.1697798

Peña, Y., & González, J.J.F. (2022). Modelo de predicción de los resultados de la prueba ICFES Saber 11 en el área de matemáticas a partir de variables socioeconómicas [Prediction model of the results of the ICFES Saber 11 test in the area of mathematics based on socio-economic variables. *Studies in Engineering and Exac*t] *Sciences*, *Curitiba*, *3*(1), 31-37. https://doi.org/10.54021/seesv3n1-006

Planinic, M., Boone, W., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research, 15*(020111), 1-14. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Ramos, D., Pedroso, J., Lozano, A., & González, J. (2018). Deconstructing Cross-Entropy for Probabilistic Binary Classifiers. *Entropy, 20,* 208. https://doi.org/10.3390/e20030208

Robinson, K., Perez, T., Carmel, J., & Linnenbrink, L. (2019). Science identity development trajectories in a gateway college chemistry course: Predictors and relations to achievement and STEM pursuit. *Contemporary Educational Psychology*, *56*, 180-192. https://doi.org/10.1016/j.cedpsych.2019.01.004

Rodríguez, F. J., Benavides, H., & Riascos, A.J. (2018*). Predicción del desempeño académico usando técnicas de aprendizaje de máquinas* [Prediction of academic performance using machine learning techniques]. Universidad de los Andes. ICFES.

Salmerón-Pérez, H., Gutierrez-Braojos, C., Fernández-Cano, A., & Salmeron-Vilchez, P. (2010). Self-regulated learning, self-efficacy beliefs and performance during the late childhood. *RELIEVE, 16*(2), 1-18. https://doi.org/10.7203/relieve.16.2.4136

Selwyn, N., Pangrazio, L., & Cumbo, B. (2021). Attending to data: Exploring the use of attendance data within the datafied school. *Research in Education, 109*(1), 72–89. https://doi.org/10.1177/0034523720984200

Son, L. H., & Fujita, H. (2019). Neural-fuzzy with representative sets for prediction of student performance. *Applied Intelligence*, *49*, 172–187. https://doi.org/10.1007/s10489-018-1262-7

Soo, J., Lok, V., Bong, K., Wha, Y., & Ook, B. (2021). Quantitative risk-based inspection approach for high-energy piping using a probability distribution function and modification factor. *International Journal of Pressure Vessels and Piping,* 189, 1-14. https://doi.org/10.1016/j.ijpvp.2020.104281

Suárez-Montes, N., & Díaz-Subieta, L. B. (2015). Estrés académico, deserción y estrategias de retención de estudiantes en la educación superior [Academic stress, desertion, and retention strategies for students in higher education]. *Revista de Salud Pública*, *17*(2), 300–313. https://doi.org/10.15446/rsap.v17n2.52891

Tai-Chui, K., Chun, D., Lytras, M., & Miu-Lam, T. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior, 107*. https://doi.org/10.1016/j.chb.2018.06.032

Tsiakmaki, M., Kostopoulos, G., & Kotsiantis, S. (2021). Fuzzy-based active learning for predicting student academic performance using autoML: a step-wise approach. *Journal of Computing in Higher Education, 33*, 635–667. https://doi.org/10.1007/s12528-021-09279-x

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Transfer Learning from Deep Neural Networks for Predicting Student Performance. *Applied Sciences*, *10*(6), 2145. https://doi.org/10.3390/app10062145

Vargas, V., & Ardila, L. F. (2019). *Predicción del desempeño en las pruebas Saber 11 utilizando variables del contexto socio-económico de los aplicantes mediante un análisis estadístico con técnicas de machine learning* [Performance prediction on Saber 11 Tests using socio-economic variables of the applicants through a statistical analysis with machine learning techniques]. Universidad Nacional de Colombia.

Vargas, V., Gutiérrez, P., & Hervás, C. (2022). Unimodal regularisation based on beta distribution for deep ordinal regression. *Pattern Recognition, 122*. https://doi.org/10.1016/j.patcog.2021.108310

Waheed, R., Sarwar, S., Sarwar, S., & Khan, M. K. (2020). The impact of COVID-19 on Karachi stock exchange: Quantile-on-quantile approach using secondary and predicted data. *Journal of Public Affairs, 20*(4), e2290. https://doi.org/10.1002/pa.2290

Wang, Y., Pan, Z., Yuan, X., Yang, C., & Gui, W. (2020). A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network. *ISA Transactions, 96*, 457-467. https://doi.org/10.1016/j.isatra.2019.07.001

Yang, S., Lu, O., Huang, A., Huang, J., Ogata, H., & Lin, A. (2018). Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis. *Journal of Information Processing,* 26, 170–176. https://doi.org/10.2197/ipsjjip.26.170

Zois, E., Alexandridis, A., & Economou, G. (2019). Writer independent offline signature verification based on asymmetric pixel relations and unrelated training-testing datasets. *Expert Systems with Applications, 125*, 14-32. https://doi.org/10.1016/j.eswa.2019.01.058