Comparison of Machine Learning algorithms for the Burnout projection

Comparación de algoritmos de Machine Learning para la proyección de Burnout

LARA-GONZALEZ, Luis Rey†, DELGADO-LUNA, Martha Angélica, DE LEON-GALVAN, Beatriz Elena and VENEGAS-GUERRERO, José Carlos

Universidad Politécnica de Juventino Rosas

ID 1st Author: *Luis Rey, Lara-González* / **ORC ID**: 0000-0002-0902-5955, **CVU CONACYT-ID**: 607818

ID 1st Coauthor: *Martha Angélica, Delgado-Luna* / **ORC ID**: 0000-0002-3749-6973

ID 2nd Coauthor: *Beatriz Elena, De León-Galván* / **ORC ID**: 0000-0002-8974-4561

ID 3rd Coauthor: *José Carlos, Venegas-Guerrero* / **ORC ID**: 0000-0003-0725-6730

Abstract

The present study aims to carry out a projection of student burnout risk detection in young university students using Machine Learning technics (Neuronal Networks, KNN, SVM, Random Forest). A descriptive method was proposed, with a cross-sectional and stratified design in which a sample of 791 students from 4 different universities. This study opens up an innovative field of research by integrating resources from psychological evaluation and virtual resources, in addition, it would allow the generation of preventive actions to treat various implications of Burnout in school dropout and low academic performance through the analysis of information and the generation of algorithms that allow the projection of burnout risk. Due to the combination of experience of professionals in psychology, education and engineering, as well as the contribution to the projection of a syndrome that affects students, makes this article an innovative proposal.

Resumen

El presente estudio tiene por objeto realizar una proyección de detección de riesgo de burnout estudiantil en jóvenes universitarios utilizando algoritmos de Machine Learning (Neuronal Networks, KNN, SVM, Random Forest). Para esto se planteó una metodología de tipo descriptivo, con un diseño de corte transversal y estratificado en la que se toma una muestra de 791 estudiantes de 4 diferentes universidades. Este estudio abre un campo de investigación innovador al integrar recursos de la evaluación psicológica y de recursos virtuales, además de que permitiría generar acciones preventivas para tratar diversas implicaciones del Burnout en la deserción escolar y el bajo rendimiento académico a través del análisis de información y la generación de algoritmos que permitan realizar la proyección de riesgo de burnout. Por la conjunción de experiencia de profesionales de la psicología, de la educación y de ingeniería, así como el aporte a la proyección de un síndrome que afecta a los y las estudiantes, hace de este artículo una propuesta innovadora.

Academic Burnout, Machine Learning, Students

Burnout académico, Máquinas de aprendizaje, Estudiantes

† Researcher contributing first author.

Introduction

The 2005 World Report of the United Nations Educational, Scientific and Cultural Organization (UNESCO), entitled "Towards knowledge societies" indicates that the new information and communication technologies have created the conditions for the emergence of knowledge societies (Forero de Moreno, 2009). In the current context, technology provides different possibilities to solve problems regarding the collection and analysis of information in the development of research, which range from google forms to the most sophisticated tools in quantitative analysis such as SPSS, R, Mintab, Stats and Excel, among others.

We see that technological advances can be a good support in the development of research, from the structuring of instruments, gathering information, to the systematization and analysis of data. It has been reported by several studies in different countries and in Mexico a syndrome that affects students, Burnout has been defined as a "loss of energy sources of the subject, so it can be defined as a combination of physical fatigue, emotional fatigue and cognitive fatigue (Rosales Ricardo & Rosales Paneque, 2013, p. 340).

In the academic field, young university students present difficulties that affect the phenomenon of academic abandonment and that are associated with high stress loads during their training process and that, in addition to the presence of avoidance or escape behaviors, in addition to feelings of not being able to give more of themselves, among others, make up the so-called academic burnout syndrome (Rosales Ricardo & Rosales Paneque, 2013).

The measurement of Burnout has been given through various studies and instruments, including scales, questionnaires, inventories, and others. Among these we can find the Maslach Burnout Inventory Student Survey or MBI-SS (Schaufeli, et al, 2002;) adapted from the Maslach Burnout Inventory General Survey (Maslach & Jackson, 1981); the Emotional Tiredness Scale or ECE (Ramos Campos, et al, 2005); the Engagement Scale (Caballero Domínguez, et al, 2006); and The Unidimensional Scale of Student Burnout or EUBE (Barraza Macías, 2011) among others.

Academic burnout, in education, is reported by Caballero (Caballero Domínguez, et al, 2012) who proposes that it be considered as a good predictor of the concurrence of anxiety symptoms, and to a lesser degree, depression, and that it is feasible constitutes a risk factor for the development of major conditions of mental disorder, especially clinically significant anxiety.

Therefore, Burnout is a variable that impacts the educational system, its study and deepening provide useful elements to institutions and in general to the educational system to generate strategies, programs and actions that support its reduction and have a favorable impact on permanence and education. academic performance of the student population.

This article presents the conjunction of the knowledge that intelligent systems contribute to the sciences of education, to project possibilities of suffering Burnout in the student population, which will allow to anticipate and generate preventive actions that avoid affectations in the academic trajectory.

Machine learning, translated into Spanish as "machine learning", is a field derived from artificial intelligence that broadly consists of developing processes that allow machines to learn autonomously. Machines learn on their own from a set of data that an instructor first enters and then corrects manually. It is throughout this process that the computer extracts information that it transforms into knowledge based on experience (Siegel, 2013).

In the present work, an analysis of different algorithms is generated that, based on the data obtained from the application of the EUBE instrument (Barraza, 2011) in four universities, allows us to analyze the predictive power that will define the most appropriate algorithm for the projection of Burnout in the higher education student population.

The following algorithms are applied: neuronal networks, random forest, support vector machine and KNN (k-Nearby neighbors).

LARA-GONZALEZ, Luis Rey, DELGADO-LUNA, Martha Angélica, DE LEON-GALVAN, Beatriz Elena and VENEGAS-GUERRERO, José Carlos. Comparison of Machine Learning algorithms for the Burnout projection. ECORFAN Journal-Democratic Republic of Congo. 2021

Software used

Rstudio

RStudio is an integrated development environment (IDE) for the R programming language, dedicated to statistical computing and graphics. It includes a console, syntax editor that supports code execution, as well as tools for plotting, debugging, and workspace management.

R is a programming language used as software for statistical computing, which does not have a graphical interface, that is, when pressing a button in the RStudio environment which was developed for programming with this R language, it does not display a menu or a pop-up window. What makes this statistical analysis program so widely used is that it is free software and its codes are open, that is, they are available to any user (Celis, 2020).

For this analysis, machine learning tools were used using the R language libraries and algorithms, these algorithms are explained below.

The Neuronal Networks algorithm

Neural networks or also called annexation systems are an algorithmic model that is based on a large set of simple neural units, called artificial neurons, in a way that is very similar to the behavior of the axons of neurons that any biological organism has. Each neuronal unit connects with many others and the links between them can increase or inhibit the state of the neurons that are adjacent to it.

Each individual neuronal unit operates using addition operations. There may be a limiting function, a threshold function at each of the connections in the unit itself, so that the signal can pass a limit before propagating to another neuron.

Random Forest Algorithm

A Random Forest model is made up of a set (ensemble) of individual decision trees, each trained with a random sample extracted from the original training data by bootstrapping. This implies that each tree is trained with slightly different data.

In each individual tree, the observations are distributed by bifurcations (nodes) generating the structure of the tree until reaching a terminal node. The prediction of a new observation is obtained by adding the predictions of all the individual trees that make up the model.

The Support Vector Machine (SVM) algorithm

The classification-regression method Vector Support Machines (SVMs) was developed in the 90's, within the field of computational science. Although it was originally developed as a binary classification method, its application has been extended to multiple classification and regression problems. SVMs have turned out to be one of the best classifiers for a wide range of situations, which is why it is considered one of the benchmarks in the field of statistical learning and machine learning.

The KNN algorithm:

KNN (K-Nearest Neighbors, K-Neighbors) algorithm is one of the most popular neighborhood-based classifiers in machine learning (Kataria and Singh, 2013), given its simplicity and efficiency in detecting and classifying elements into categories. The parameter k in KNN refers to the number of neighbors with which the belonging to a category is defined, this parameter is usually determined empirically, depending on the problem it is tested with different values of K, choosing the parameter with the best performance in precision. The operation of the algorithm is based on calculating a distance matrix between all the points of the training dataset.

It is a supervised learning algorithm, that is, from an initial data set its objective will be to correctly classify all new instances. The typical data set for this type of algorithm is made up of several descriptive attributes and a single objective attribute (also called a class) (Ad Tech & Analytics, 2017). The objective of this study is to determine the most powerful algorithm to project Burnout in university students. Therefore, the central hypothesis that guides this research is that there is an algorithm that predicts the projection of presenting Burnout in university students above 95%.

In the presentation of the article, the methodology is developed where the steps that were followed for the development of the present study are specified, results where the predictive percentages of each of the algorithms are described, the discussion of the results, where the point out approaches derived from the results and the confirmation or not of the hypothesis. Finally, in the conclusions, the contribution of the present investigation is planned as well as some lines of investigation.

Methodology

All projects have a life cycle; that is, its realization comprises a series of logical stages that begins with the formulation of the original idea and the development of an activity plan, continues with the identification and verification of the capacities and needs of the interested parties and, finally, It comprises the design, implementation and evaluation stages. (Rice and Salinas, 1997).
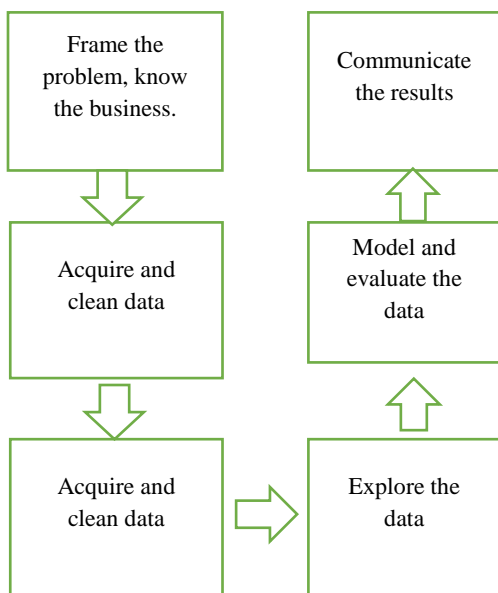
Proposed model

Figure 1

A scale created and validated by Dr. Arturo Barraza and collaborators (Barraza, 2008) of the Pedagogical University of Durango was applied, it consists of 15 items and reports a reliability level of 91 in Cronbach's alpha and 90 in reliability by halves according to the Spearman-Brown formula. This instrument measures the level of physical, emotional and cognitive exhaustion (Barraza Macías, 2011).

Due to its distinctive characteristics, the scale can be defined as self-descriptive (Anastasi & Urbina, 1998, in Barraza Macías, 2011) and domain-specific (Hogan, 2004, in Barraza Macías, 2011). It is an instrument that can be applied in any type of population, which makes it currently one of the most widely used instruments in Mexico and recovers the one-dimensional perspective of burnout (Rosales Ricardo & Rosales Paneque, 2013).

For the systematization and analysis of the information, a data structure was taken that consisted of 791 samples obtained from the application of the instrument by Arturo Barraza et al. (2008), a total of 15 questions; each question is answered with these possible options "never, sometimes, almost always and always. The following figure shows an example of questions 12, 13, 14, and 15; The column "conclusion_b1" indicates the Burnout level that the respondent presents. In other words, the rows represent the respondents, and the columns represent each of the questions (b1_1 .. b1_15) questions asked with their respective answer and the last column (Conclusion.b1) shows the level of Burnout presented by the respondent ("never , sometimes, almost always and always).

b1_12	b1_13	b1_14	b1_15	Conclusion.b1
Never	Never	Usually	Never	Does not present
Never	Never	Never	Sometimes	Does not present
Never	Never	Never	Never	Does not present
Never	Never	Never	Never	Does not present

Results

Below is the application of each of the algorithms.

Random Forest algorithm

It implements the Breiman Random Forest algorithm (based on the original Fortran code from Breiman and Cutler) for classification and regression.

data = 791 = total sample
Training = 70% of the data
Ntree = Number of trees to grow. This should not be set to too small a number, to ensure that each input row is predicted at least a few times.
ntree = 500
target = conclusion.b1
target = conclusion.b1

LARA-GONZALEZ, Luis Rey, DELGADO-LUNA, Martha Angélica, DE LEON-GALVAN, Beatriz Elena and VENEGAS-GUERRERO, José Carlos. Comparison of Machine Learning algorithms for the Burnout projection. ECORFAN Journal-Democratic Republic of Congo. 2021

model = randomforest (Training, independent_variables, target, ntree)
Pred = generic function for predictions from the results of various model fitting functions.
Pred = (model, 30% remaining data, ”class”); class is the type of algorithm applied.

The result is as follows:

Projection			
Current	Mild	Moderate	Does not present
Mild	66	0	6
Moderate	0	0	0
Does not present	2	0	162

Table 2

The algorithm predicts 66 cases correctly in the "SLIGHT" value, having a total of 6 non-projected errors. Also in the value "NOT PRESENT" it manages to project 162 cases and a total of 2 errors.

Projection percentage = 95.9%

Support Vector Machine algorithm

SVM is used to train a support vector machine. It can be used to perform general classifications and regressions (nu and epsilon), as well as to estimate density. A formula interface is provided.

Training = 70% of the data
target = conclusion.b1
target = conclusion.b1
model = svm (independent_variables, target)
Pred = (model, 30% remaining data,)
The result is as follows:

Projection			
Current	Mild	Moderate	Does not present
Mild	63	0	9
Moderate	0	0	0
Does not present	11	0	153

Table 3

The algorithm predicts 63 cases correctly in the value "SLIGHT", having a total of 9 non-projected errors. Also in the value "NOT PRESENT" it manages to project 153 cases and a total of 11 errors.
Projection percentage = 91.5%

K-Nearest Neighbor Classification algorithm

k-nearest neighbor classification for the test set of the training set. For each row of the test set, the k closest training set vectors are found (in Euclidean distance) and the ranking is decided by majority vote, with randomly broken ties. If there are ties for the k-th closest vector, all candidates are included in the vote.

KNN requires that the independent or predictor variables be numeric. And the dependent variables or what I want to classify is a category.

Training = 50% of the data
Temporary = remaining 50% of Training data
Val = 50% of Temporal
Test = 50% of the rest of Temporal
Pred = knn (Training (independent_variables), Val, Training (category_variable, k = 1)

k = 1

Projection			
Current	Mild	Moderate	Does not present
Mild	51	0	10
Moderate	0	0	1
Does not present	10	0	127

Table 4

k=2

Projection			
Current	Mild	Moderate	Does not present
Mild	49	0	12
Moderate	0	0	1
Does not present	10	0	127

Table 5

k=3

Projection			
Current	Mild	Moderate	Does not present
Mild	49	0	12
Moderate	0	0	1
Does not present	7	0	130

Table 6

Looking for the suitable value of k for the algorithm, we find that:

k=4

Projection			
Current	Mild	Moderate	Does not present
Mild	54	0	7
Moderate	0	0	1
Does not present	5	0	132

Table 7

The algorithm predicts 54 cases correctly in the "SLIGHT" value, having a total of 7 non-projected errors. Also in the value "NOT PRESENT" it manages to project 132 cases and a total of 5 errors.

Projection percentage = 93.9.%

Neural Networks Algorithm

Training = 70% of the data
size = relation to input variables = 15
maxit = maximum convergences = 10000
decay = weight drop parameter = 0.001
range = range of initial weights close to 1 = 0.25
model = nnet (independent_variables, Training, size, maxit, decat, range)
Pred = (model, 30% of the remaining data)

It should be noted that this algorithm uses more computational processing time compared to other algorithms seen.

The result is as follows:

Projection			
Current	Mild	Moderate	Does not present
Mild	60	1	11
Moderate	0	0	0
Does not present	20	0	144

Table 8

The algorithm predicts 60 cases correctly in the value "SLIGHT", having a total of 12 non-projected errors. Also in the value "NOT PRESENT" it manages to project 144 cases and a total of 20 errors.

Projection percentage = 86.8%

Results and conclusions

The study of Burnout becomes relevant in higher education institutions, since these are usually highly stressful places. In these academic spaces, students face a series of demands on a daily basis that, under their own assessment, constitute stress-generating situations, which, being prolonged and frequent, can lead to burnout (Rosales Ricardo and Rosales Paneque, 2013).

Along with academic burnout, it is a good predictor of the concurrence of symptoms of anxiety, and to a lesser degree, depression, so it can become a risk factor for the development of major conditions of mental disorder, especially of clinically significant anxiety (Caballero (2012). Therefore, young university students may present difficulties that affect the phenomenon of academic abandonment (dropout) that is associated with high stress loads or as a consequence of presenting exacerbated symptoms of anxiety or depression during their training process In this sense, this research has an impact on the educational quality of Higher Level Institutions, particularly on the indicators of educational lag and terminal efficiency.

Although machine learning algorithms have advanced for years, it has only been used recently for behavioral sciences (DelPozo-Banos et al., 2018). For example, these algorithms are used in computational psychiatry to improve the diagnosis of mood disorders: stress (Silva et al., 2020), depression (Webb et al., 2020), and suicidal tendencies (Kessler et al., 2015), allow possibilities of obtaining detailed knowledge about diseases, disorders. These methods are important because they allow you to analyze data from different sources. For example, Kaczor and his colleagues used machine learning techniques to detect stressful situations using digital sensors used by emergency medicine physicians and a self-assessment questionnaire (Kaczor 2020).

The application of the algorithms allows us to project the Burnout results based on a percentage of the data (70%) which is taken as training data and the rest of the data (30%) are taken to make the prediction by applying the machine learning algorithms.

Collecting the results of each of the algorithms used, we have:

Algorithm	% from Projection	Chosen one
Random Forest	95.9 %	x
KNN	93.9 %	
SVM	91.5 %	
NNET	86.8 %	

Table 9

It is observed that the algorithm with the highest percentage of predictive result is that of Random Forest with a 95.0% projection percentage with respect to the data provided based on the survey carried out.

Future work, since the type of input variable for all algorithms is categorical, remains to adjust the input parameters (size, maxit, decat, range) of the Neural Networks (NNET) algorithm, because Only settings that are suggested in the parameters of said algorithm were used, with the aim of analyzing whether it is possible to improve the prediction percentage.

There are other data from the respondents that were not taken into account in this research, such as age, gender, the study period in which data collection was applied, so future work would be to integrate these data as independent variables to provide greater elements for decision-making and characterizations of the academic burnout.

This study, then, could allow a projection of student burnout at the university level to be carried out and in this way generate strategies and actions focused on its prevention and / or intervention, for example, through Tutoring programs, which represent an element important innovator of Educational Quality and that contribute to "reduce dropouts and lag in higher education" (Romo López, 2011, p. 76)

References

Ad Tech & Analytics. (2017). El algoritmo K-NN y su importancia en el modelado de datos |Analítica web. Recuperado 13 de abril de 2021, de https://www.analiticaweb.es/algoritmo-knn-modelado-datos/

Anastasi, A. and Urbina, S. and Salinas, M.E.O.,isbn=9789701701867,url=https://books.google.com.mx/books?id=FV01zgFuk0cC,1998.Prentice-Hall

Barraza Macías, A. (2008). Escala Unidimensional del Burnout Estudiantil. Investigación Educativa Duranguense. 9. https://www.researchgate.net/publication/28237109_Escala_Unidimensional_del_Burnout_Estudiantil

Barraza Macías, A. (2011). Validación psicométrica de la escala unidimensional del burnout estudiantil. *Revista Intercontinental de Psicología y Educación. 13*(2). 51-74. http://www.redalyc.org/pdf/802/80220774004.pdf

Caballero Domínguez, C. C., Abello Llanos, R. y J. Palacio Sañudo. J. (2006). BURNOUT, ENGAGEMENT Y RENDIMIENTO ACADÉMICO ENTRE ESTUDIANTES UNIVERSITARIOS QUE TRABAJAN Y AQUELLOS QUE NO TRABAJAN. *Revista Psicogente, 9*(16), 11-27. https://www.researchgate.net/profile/Jorge_Palacio_Sanudo/publication/284701793_Burnout_engagement_y_rendimiento_academico_entre_estudiantes_universitarios_que_trabajan_y_aquellos_que_no_trabajan/links/568a75ec08ae1975839d84b2.pdf

Caballero Domínguez., C. C., Abello Llanos, R. y Palacio Sañudo, J. (2007). Relación del burnout y el rendimiento académico con la satisfacción frente a los estudios en estudiantes universitarios. *Avances en Psicología Latinoamericana, 25*(2), 98-111. https://www.redalyc.org/articulo.oa?id=799/79925207

Caballero, C. (2012). El burnout académico: Pre-valencia y factores asociados en estudiantes universitarios del área de la salud de la ciudad de Barranquilla. Doctoral dissertation. Universidad del Norte, Barranquilla (Colombia).

Celis Mario A, 2020 Introducción a Rstudio, Recuperado 4/Mayo/2021, https://rpubs.com/Mariocelis25/603283.

DelPozo-Banos, M., John, A., Petkov, N., Berridge, D. M., Southern, K., Loyd, K., et al. (2018). Using neural networks with routine health records to identify suicide risk: Feasibility Study. JMIR Mental Health, 5(2), e10144

Forero de Moreno, I. (2009). LA SOCIEDAD DEL CONOCIMIENTO. *Revista Científica General José María Córdova, 5*(7), 40-44. https://www.redalyc.org/articulo.oa?id=476248 849007

Kaczor, E. E., Carreiro, S., Stapp, J., Chapman, B., & Indic, P. (2020). Objective measurement of physician stress in the Emergency Department using a wearable sensor. In Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020 (pp. 3729–3738). Hawaii: University of Hawai'i at Mānoa.

Kessler, R. C., Warner, C. H., Ivany, C., Petukhova, M. V., Rose, S., Bromet, E. J., et al. (2015). Predicting suicides after psychiatric hospitalization in US Army soldiers: The Army Study to assess risk and resilience in servicemembers (Army STARRS). JAMA Psychiatry, 72(1), 49–57. https://doi.org/10.1001/jamapsychi atry.2014.1754.

Maslach, C. y Jackson, S. E. (1981) The measurement of experienced burnout. *Journal of Occupational Behaviour, 2*, 99-113. https://onlinelibrary.wiley.com/doi/abs/10.1002 /job.4030020205

Ramos Campos, F. Manga Rodríguez, D. y Morán Astorga, C. (2005). ESCALA DE CANSANCIO EMOCIONAL (ECE) PARA ESTUDIANTES UNIVERSITARIOS: PROPIEDADES PSICOMÉTRICAS Y ASOCIACIÓN. *6 Congreso Virtual de Psiquiatría.* Interpsiquis. https://www.researchgate.net/publication/27184 6180_Escala_de_cansancio_emocional_ECE_p ara_estudiantes_universitarios

Rice and Salinas, 1997] Rice, F. P. and Salinas, M. E. O. (1997). Desarrollo humano: estudio del ciclo vital. Pearson Educación.

Romo López, A. (2011). *La tutoría: una estrategia innovadora en el marco de los programas de atención a estudiantes*. ANUIES.

Rosales Ricardo, Y. y Rosales Paneque, F. R. (2013). Burnout estudiantil universitario. Conceptualización y estudio. *Salud Mental, 36*(4). 337-345. http://www.redalyc.org/articulo.oa?id=5822896 9009

Silva, E., Aguiar, J., Reis, L. P., Sá, J. O. E., Gonçalves, J., & Carvalho, V. (2020). Stress among Portuguese Medical Students: The EuStress Solution. Journal of Medical Systems, 44(2), 1–6. https://doi. org/10.1007/s10916-019-1520-1.

Schaufeli, W. B., Martínez, I. M., Marques Pinto, A., Salanova, M. y Bakker, A. B. (2002). Burnout an engagement in univertity students. *Journal of Cross-cultural Psychology, 33*(5), 464-481. https://journals.sagepub.com/doi/10.1177/0022 022102033005003

Siegel E., John Wiley & Sons. 2013 Predictive analytics : the power to predict who will click, buy, lie, or die. Hoboken : John Wiley & Sons, 2013.

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. Journal of Consulting and Clinical Psychology, 88(1), 25–38. https://doi.org/10.1037/ccp0000451.