Towards the Grade's Prediction. A Study of Different Machine Learning Approaches to Predict Grades from Student Interaction Data

Héctor Alonso-Misol Gerlache, Pablo Moreno Ger, Luis de la Fuente Valentín *

Universidad Internacional de La Rioja, Logroño (Spain)

Received 3 January 2021 | Accepted 23 September 2021 | Published 23 November 2021

Abstract

There is currently an open problem within the field of Artificial Intelligence applied to the educational field, which is the prediction of students' grades. This problem aims to predict early school failure and dropout, and to determine the well-founded analysis of student performance for the improvement of educational quality. This document deals the problem of predicting grades of UNIR university master's degree students in the on-line mode, proposing a working model and comparing different technologies to determine which one fits best with the available data set. In order to make the predictions, the dataset was submitted to a cleaning and analysis phases, being prepared for the use of Machine Learning algorithms, such as Naive Bayes, Decision Tree, Random Forest and Neural Networks. A comparison is made that addresses a double prediction on a homogeneous set of input data, predicting the final grade per subject and the final master's degree grade. The results were obtained demonstrate that the use of these techniques makes possible the grade predictions. The data gives some figures in which we can see how Artificial Intelligence is able to predict situations with an accuracy above 96%.

Keywords

Artificial Intelligence, Grade Prediction, Machine Learning, Prediction Technology.

DOI: 10.9781/ijimai.2021.11.007

I. Introduction

With the current change in the digital and business paradigm, society's education has a fundamental role to play. It is not only a question of the anachronistic education systems of the industrial revolution not being valid for a society that is trained for jobs that do not yet exist, but the social and mental models have changed.

Combating failure and early dropout from university is an issue of vital importance, especially because of the economic and social cost it generates [1], becoming an issue that has generated growing concern in recent years. Prevent school failure and increase the quality of teaching is an actual object of the educational. Predicting students' results early enables the university and the teacher to carry out more focused teaching work, as well as allowing students to focus their efforts and plan their studies better. Therefore, predicting student grades will have a direct impact on improving education at all levels, helping to combat school dropout and enabling continuous improvement in the academic process, with a consequent positive impact on society and economy.

Teaching today requires great flexibility to provide useful content to a highly changing and dynamic society. This being so, the work of universities is not merely the transmission of knowledge but must be a focus of innovation to teach students how to face the new challenges and opportunities of society, where flexibility is necessary in both teachers and students, educating in knowledge and skills [2]. In this sense, the online university is presented as a great alternative to face-to-face studies, being increasingly successful and accepted.

Today, the online academic offer is growing considerably, not only as the solution to combine work and training, but the recent COVID19 pandemic has boosted this modality in places where it was previously unthinkable. Online learning is defined as "learning experiences in synchronous or asynchronous environments using different devices (e.g., mobile phones, laptops, etc.) with internet access. In these environments, students can interact with instructors and other students through the different platforms that the market offers [3].

It seems a long time ago, in 1995, when the UOC (Universidad Oberta de Catalunya) appeared in Spain and became the first online university in the world. Since then, the number of students has increased, reaching 900% growth from 2000 to 2018, according to the GAD3 (www.gad3.com), with a forecast of a multiplication of students by 10 in 2026, with growth in both bachelor's and master's students.

There are many advantages to online studies, which has the particularity that students can combine study with other activities, mainly work, and allow them to study anywhere, at a self-controlled rhythm and at any time, depending on the obligations and needs of each student, simply by needing an internet connection [4].

The extensive use of technologies in the current social panorama makes it possible to use technologies that can collect, analyzing and

* Corresponding author.

E-mail addresses: h.gerlache@gmail.com (H. Alonso-Misol Gerlache), pablo.moreno@unir.net (P. Moreno Ger), luis.delafuente@unir.net (L. de la Fuente Valentín).

extracting information from the data generated in all areas. In the educational panorama, this is no exception, and online universities are great generators of data thanks to the use of technological platforms that they use to reach all corners of the world. In fact, the online university studies mode favors the generation of data that allows us to carry out a subsequent study and analysis of the same in order to offer and improve all levels of learning and education in our society [5].

Educational institutions generate and collect huge amount of data. This may include students' academic records, their personal profile, observations of their behavior, their web log activities and faculty profile. This large data set is basically a storehouse of information and must be explored to have a strategic edge among the Educational Organizations [6]. The potential for data analysis in education must focus on developing robust applications that will improve student outcomes, enhance the pedagogy of instructors, improve the curriculum and increase graduation rates for all students, regardless of their background, from kindergarten to university. Today, higher education institutions face the critical challenge of retaining students and ensuring their successful graduation [7].

Today we have enough data to carry out an exhaustive analysis of them, through artificial intelligence techniques, to search for patterns within them that will allow us to improve our knowledge. In addition, the technological platforms integrated into our systems, such as educational ones, allow us to continue generating data that will provide knowledge about future situations. Whether through the processes of knowledge extraction from data (KDD) to the use of Machine Learning or Deep Learning techniques, thanks to the data we are able to make predictions with a high degree of certainty.

The use of this data must be focused on combating the major problems of education, and thus take advantage of the power of artificial intelligence. A problem related to higher education that concerns education authorities worldwide is the high rate of university dropouts. Data from the Spanish Ministry of Education, Culture and Sport (MECD, 2016) indicate that approximately one in five students drop out of university in the first year [8]. Of course, before making predictions, it is essential to find out an algorithm that is best suited for the problem, which requires comparison of algorithms based on certain metrics [6].

Data predictions are possible thanks to the algorithms, their use in educational environments is no exception [9]. As shown in this document, four algorithms with very high success rates are analyzed and compared in order to determine which of them is best suited to the dataset analyzed for grade prediction. These algorithms are Naive Bayes, Decision Tree, Random Forest and Neural Networks.

In this document we are going to try two different approaches in order to check whether they provide promising results. These two approaches are the prediction of the final Master's degree and the final grade of an exam. Both are going to be analyzed with the same dataset and in the conclusions phase, we are going to view the results of each other.

As we said, combating failure and early dropout from university is an issue of vital importance, especially because of the economic and social cost it generates [1], becoming an issue that has generated growing concern in recent years. Currently, many advances are being made, where Artificial Intelligence stands out as a powerful tool to help solve educational problems in the future, forming a scenario for improving educational quality, where technology, focused on the analysis of educational data, can be applied to prevent school failure and increase the quality of teaching, thus improving the student-teacher-university relationship.

The structure of the document is described below. After the introduction of this first chapter, the state of the question is addressed in Chapter II, where the background necessary to address the issue is explored in greater depth. Chapter III presents the main objective and secondary objectives, as well as a description of the methodology used. Chapter IV explains in detail the contribution of the paper and the experimentation carried out. Chapter V analyses the results and, finally, Chapter VI draws a final conclusion to the work and defines future lines of action for the continuation of the work.

II. State of the Art

The International Organization for Standardization stated in 2002 that the ability of educational institutions to manage their students on an individual basis was a key factor in achieving excellence in higher education [10]. This requires that teachers know the characteristics of their students and can guide them adequately to help them achieve their goals and avoid academic failure at the university [11]. This incipient need to improve the quality of education is the reason why many institutions are implementing learning platforms such as Blackboard, Moodle or Sakai, in order to offer their students a complete online platform, where the relationship between students, teachers and academic management is combined and managed. These systems present a learning opportunity that is delocalized and tailored to the interests of each student, and they are major generators of information which, using Artificial Intelligence techniques, can evaluate predictive models for different situations, such as student enrolment or grades [12]. However, this generated data, where a student's past academic history can be reviewed, is not a noise free source of information, which increases the complexity of the already complex problem regardless of the noise data [13], which degrades the quality or performance of the prediction, and it is necessary to discover the underlying correlation between the data and their degree of affectation [14]. In the field of education, Educational Data Mining (EDM) is taking advantage of the large amount of data in the sector, seeks to develop methods that discover the knowledge of data from educational environments [15], with the challenge of making good use of the data to improve the educational process [16]. The analysis of the prediction of grades and dropouts has led to research by different authors, which shows that one model is not better than another in a generalized way, but that the best prediction is given using a combination of models, such as neural networks, support vector machines and ensembles [17]. In the absence of concrete results from research on which algorithms are best suited to this type of problem, there are certain investigations in which the use of decision trees versus Bayesian or neural networks has yielded better results with relatively small data [18]. The analysis of university datasets references a problem with a search space of multiple parameters, of great diversity among them, and while some studies show better figures with the use of decision trees, others show better results using genetic algorithms, which confirms that at present there is no definitive study on the analysis of performance based on qualitative data of students, where it is determined which is the best model to carry out the analysis nor has it been found which of all the parameters of the students is the most influential on their academic performance [19]. Despite this paradigm, surprisingly good figures are being achieved that support the trend in the use of these techniques in the education sector. Thus, studies on prediction of results and university dropouts in the first year of electrical engineering have achieved accuracies of between 75% and 80% through the decision trees [12], even achieving predictions of over 96% to predict student grades before the final exam [15]. Machine learning for education has gained much attention in recent years, with a focus on predicting student performance [20], making clear the usefulness of Artificial Intelligence as a tool for predicting grades in the educational environment. Among the most used supervised learning algorithms in EDM, we find Naive Bayes, k-Nearest Neighbors, Decision Tree based algorithms [21], Random Forest, Support Vector Machines (SVM) and Neural Networks [16].

To the best of our knowledge, grade prediction is possible within the academic environment, but there is no obvious conclusion as either algorithm best suits the conditions of these data sets. Although, depending on the data and the treatment we give them, as well as on the configuration of the algorithms, different data will be obtained, this work provides a new study that compares the algorithms that have given the best results to date, with two clearly differentiated objectives. The first is to find out which of them best fits the data set and, subsequently, to find out which of them can predict students' grades, both for the final exam and for the master's.

III. Objectives and Methodology

Artificial Intelligence and Data Mining bring great possibilities to the field of predicting academic results. However, there are external factors that are not considered in this work, such as the socio-economic data of students, but we have a sufficiently broad set of data to address the problem of grade prediction at university environment, specifically in the University Master's Degree in Computer Security, since the data provided by the university correspond to four courses of this degree.

A. Main Objective

Contribute to the problem of grade prediction by analyzing and comparing different algorithms. The comparison of the algorithms will be done by determining which algorithm predicts more accurately and on which of the two lines of work, the prediction of the final master's degree or the prediction of the final grade of an exam.

In order to achieve objective, other intermediate milestones will need to be achieved, considerate as specifics objectives, as detailed below

B. Specific Objectives

1. Determine the feature extraction model that best fits the data set.
2. Determine the most influential characteristics in the student's academic outcome.
3. Compare the results of the two predictions to be made: final exam grade and final master's grade.
4. Perform training and validation of the selected A.I. algorithms.
5. Perform a comparative analysis of the results and technologies used in the prediction.
6. Predict the student's final exam grade and the master's degree grade.

C. Methodology

As we have seen in the main and specific objectives, the proposal of this work is to determine the best model of exploitation of the dataset for the prediction of students' grades. A methodology is proposed that follows the following five steps: Step 1 – Construction: Construction of a single dataset with the relevant information from each of the eleven files provided by the university. Step 2 – Cleaning: Starting from the single dataset, the data must be cleaned in order to eliminate the possible noise from the data, treating null values, missing data, identification of anomalous values, identification of out-of-range values and elimination of duplicated values. Step 3 – Relationship between the input characteristics: A statistical analysis of the data should be carried out to determine the behavior of the variables, comparing the means, standard deviations and quartiles of all the numerical variables, as well as the correlation between the dataset variables. Step 4 – Implementation of AI algorithms: Division of the dataset in two, so that one is prepared for the prediction of the master's degree grade and the other is prepared for the prediction of student's final exam grade. The following algorithms are considered and compared: Naïve Bayes, DecisionTree, RandomForest and Neural Networks. Step 5 – Analysis of results and conclusions: Finished the implementation of the AI algorithms, this step analyzes the results of each one of them according to the objectives 3 and 6, comparing their use for the two predictions that the work research. The document compares these four algorithms, as they are the best suited to this type of problem, as shown in the state of the art. All the steps could be viewed in the diagram of the Fig.1.

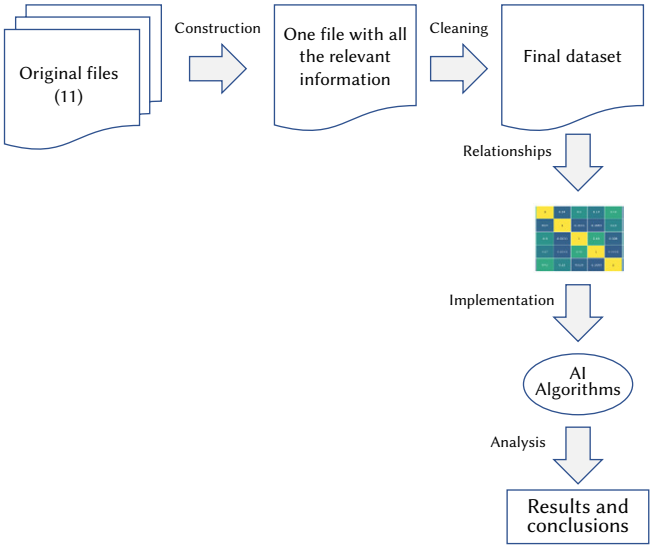


Fig. 1. Diagram of the methodology.

IV. Contribution

A. Data Preparation

The data to be analyzed comes from the LMS (Learning Management System) platform used by the International University of La Rioja (UNIR), corresponding to the University Master's Degree in Computer Security, courses from 2015 to 2018 in on-line mode. This information is provided in 11 different files that need to be unified in a single file, so that it can be analyzed and processed later. These files contain different information about the students and the grade, like the calcifications, evaluate elements, events, forums, users (students and teachers), messages, forums, information about all sessions, the relation of the tasks sent by each student, information related to the tasks and the topics to discuss in the forums. The number of rows and columns is described in Table I.

TABLE I. Number of Rows and Columns in Each File of the Dataset

Id	File	Rows	Columns
1	grades	49513	12
2	evaluate_elements	682	24
3	events	3350557	7
4	forums	197	27
5	users	699	1
6	messages	19230	28
7	rooms	65	4
8	sessions	675064	9
9	task_send	30318	7
10	tasks	315	3
11	topics	530	35

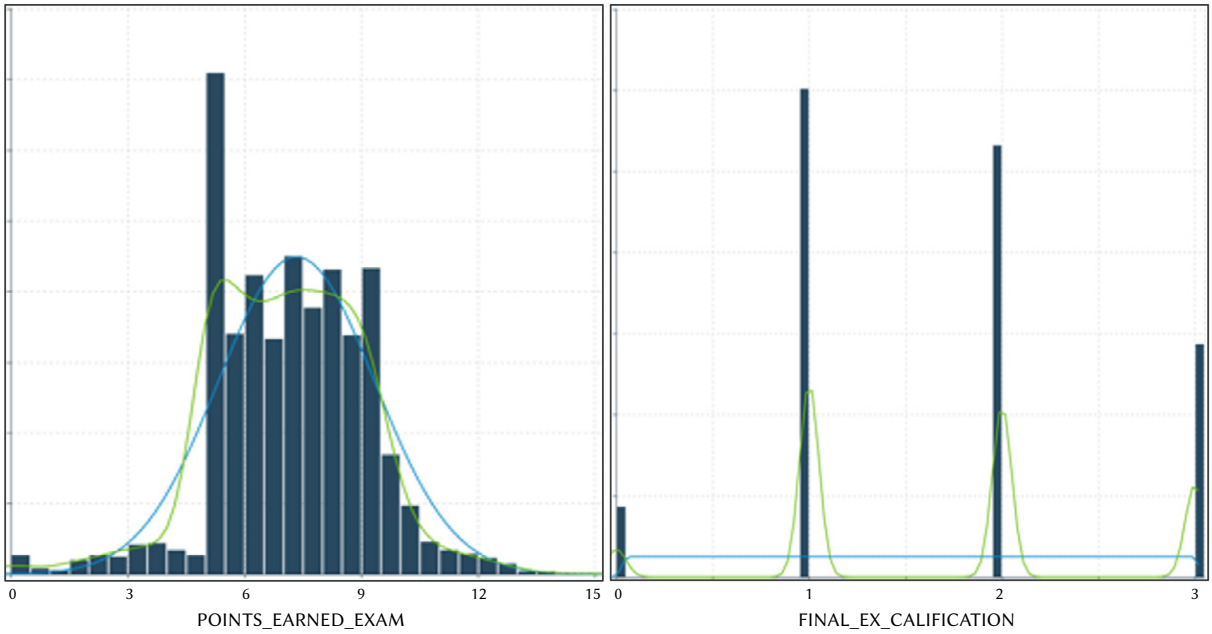


Fig. 2. Data distribution of the final exam grade.

It is important to keep in mind that the data to be worked with is data from student grades that are aseptic in terms of context. In this sense, all students are considered "equal", not according to personal or demographic data of the student, but only those data are collected that the university has as a result of learning under its model, being therefore all data of academic context. In the UNIR evaluation system, which is the origin of the dataset to be processed, there are two clearly differentiated blocks.

On the one hand, there is continuous evaluation based on evaluation activities, data, attendance at virtual classroom sessions and the performance of test-type tests. On the other hand, there is the final exam, which is the most important, and without which the subjects cannot be passed with a mark of more than 5.

The data guarantees the absolute anonymity of the data, not being able to identify any student through the data contained in the files that make up the dataset. With the 11 files, a merge has been realized in order to obtain a unique file with all the relevant data, deleting all those has no relevance in the objective of the scope of this document, like identifiers, versions, external links, etc.

The final dataset consists of 4522 records with student and subject data, and a total of 27 columns, which make up the set of input features for the AI models to be used in the machine learning algorithms. The description of each attribute could be seen at Table II.

B. Analysis Process

It is necessary to identify and understand the behavior of the predictor variables, which according to the proposed objectives 3 and 6. Their behavior can be observed in the Fig. 2 and Fig. 3, where a Gaussian distribution and its categorized correspondence can be seen.

For data analysis, Python 3.7 is used as the main tool for data processing and algorithm generation, in addition to the Watson Studio tool.

Null value analysis (missing values). For each one of the columns of the dataset, the percentage of null values is determined, eliminating directly all those columns that show null data in a percentage equal or superior to 25%. For the rest of the null values, which are less than 25%, the null value is replaced by the average of the column.

TABLE II. Description of the Final Dataset

Attributes description
Student ID.
Subject studied by the student.
Sum of points obtained in continuous assessment.
Possible points in the continuous assessment.
Number of evaluable activities.
Possible points in the course.
Number of the course activities
Number of the course events
Number of the course sessions
Number of the course messages
Number of the read course messages
Number of the evaluable course messages
Number of the read evaluable course messages
Number of the evaluable task and events
Number of the sent evaluable activities.
Points obtained in the continuous assessment.
Possible points of the continuous assessment.
Relation between the earned points of the continuous assessment and the possible points of the continuous assessment.
Average_cont is the 40% of the points earned in the continuous assessment.
Points earned in the exams.
Maximum number of points that could be earned in exams.
Relation between the earned points of the exams and the maximum number of points that could be earned in exams.
Average_exam is the 60% of the points earned in the exams.
Final grade, as the sum of the weighted fields.
Grade description: 4 ⟶ Honourable mention, 3 à ⟶ Merit, 2 ⟶ Notable, 1 ⟶ Pass, 0 ⟶Fail
Final exam grade for each student

Processing of out-of-range values. There are grade values of students with scores above 10, which are the result of having taken the ordinary assessment test and the remedial test. For these values, 187 out of the total of the dataset, which represents 4.1%, the exam grade is determined according to the final grade, putting a 4, 6, 7 or 9 for the final grades of Suspended, Pass, Notable and Merit respectively.

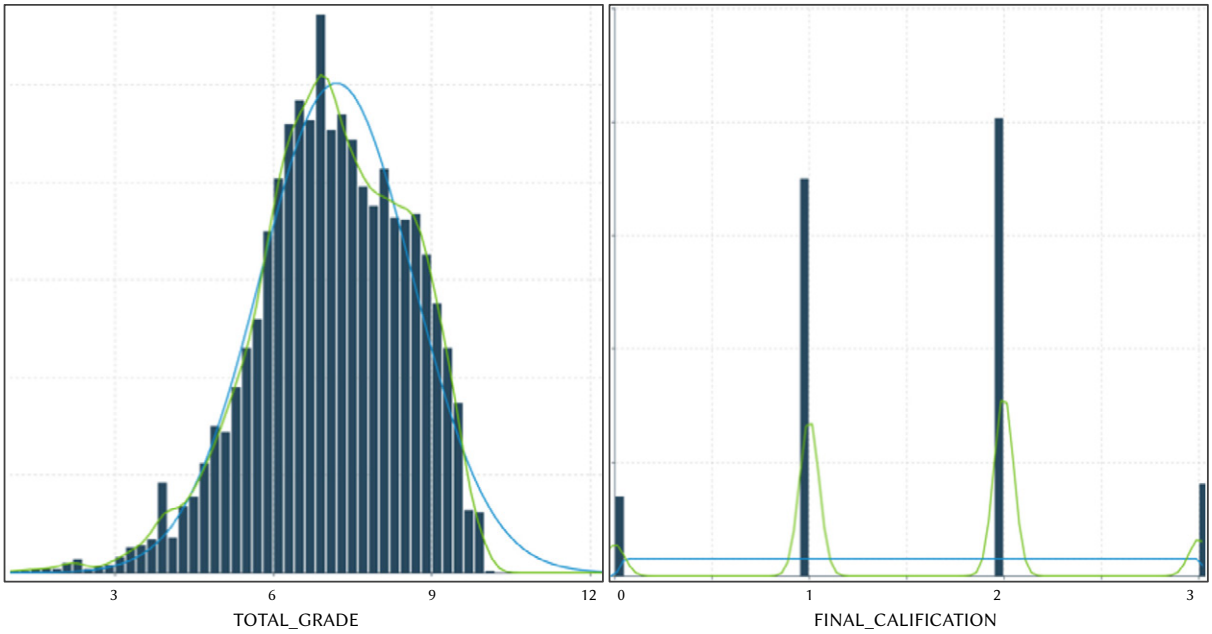


Fig. 3. Data distribution of the master's degree grade.

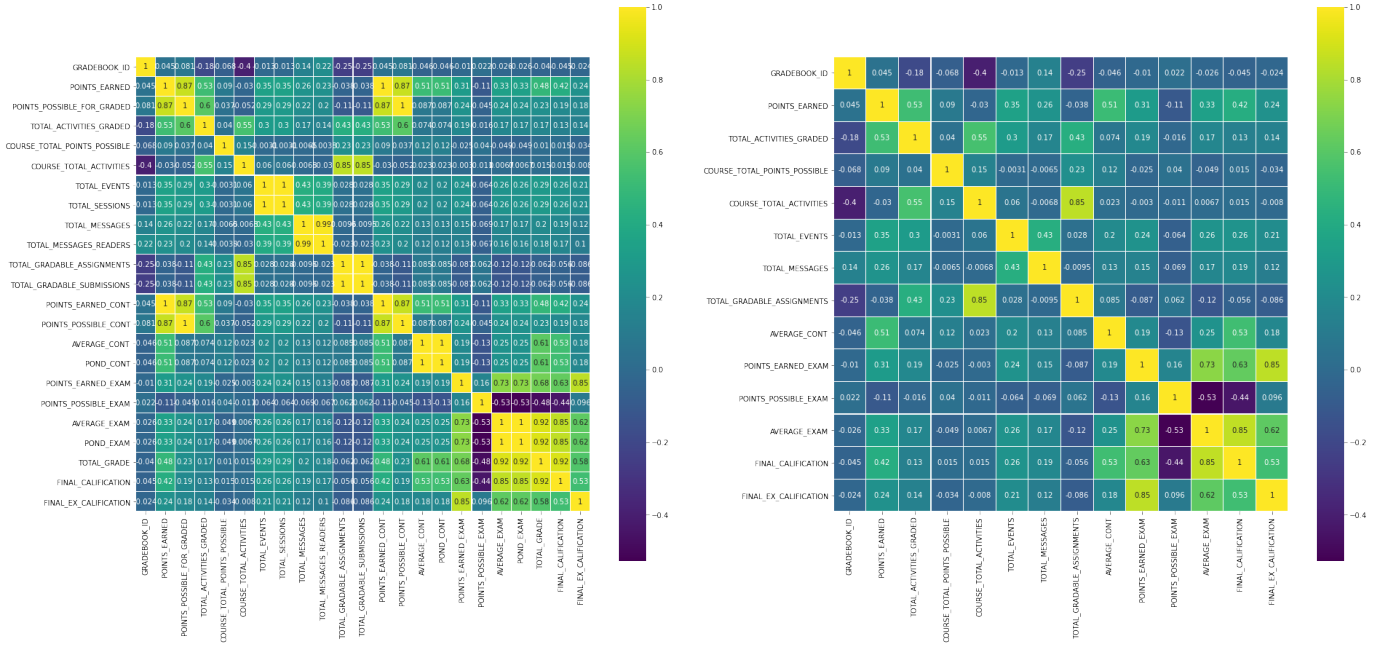


Fig. 4. Correlation of model characteristics, (a) before and (b) after treatment.

Analysis of the correlation of variables. It is important to analyze the degree of correlation between variables, in order to determine which of them do not contribute information to the model, incurring a problem of consumption of unnecessary time and resources. To do this, the correlation matrix of the variables is obtained, as can be seen in Fig. 4.

In this dataset, all those variables that exceed 75% of correlation have been eliminated, having to eliminate a total of 11 input characteristics.

Statistical description of the data. Once the noise has been removed from the data, the data is checked from a statistical perspective, analyzing the mean, standard deviation, minimums, maximums and quartiles, in order to detect possible outliers.

Different behavior of the variables can be observed, as it is the case of Fig. 5, where values out of range are observed and outliers must be solved before proceeding to their use in the Artificial Intelligence algorithms, as can be viewed in Fig. 6.

Duplicate elimination. Once the model's input feature set is clean, it is necessary to ensure that there are no duplicate records. At this point, all duplicate records are checked and removed if necessary.

The objective will always be to provide a solution to a classification problem, where the output of the algorithms will be the probability of obtaining a prediction of the final grade of the exam or of the final qualification of the master, being in both cases a prediction of between four possible values: 0 (fail), 1 (pass), 2 (notable) and 3 (merit). The algorithms Naïve Bayes (BernoulliNB model), Decision Tree (DecisionTreeClassifier model), Random Forest

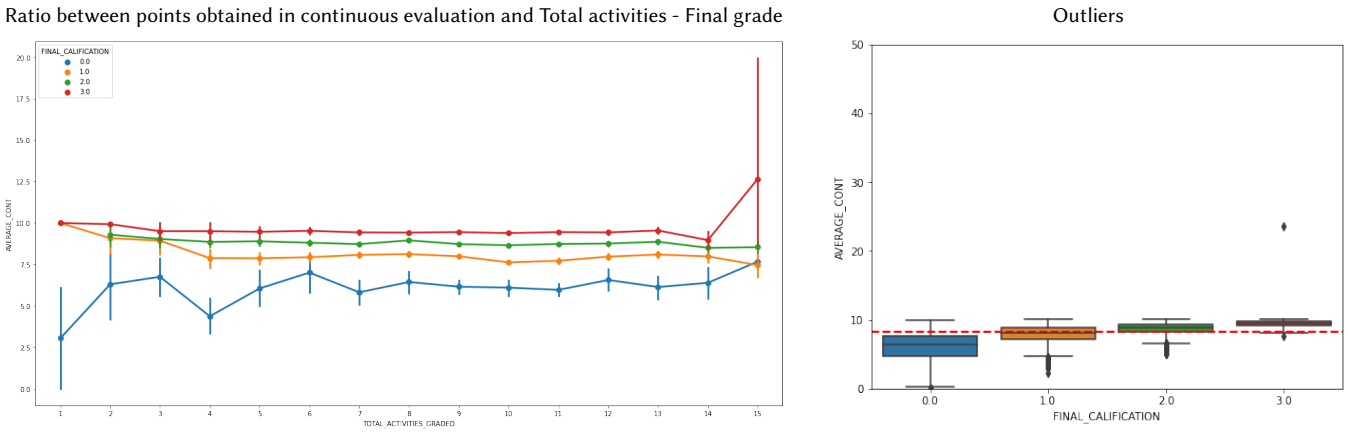


Fig. 5. Behavior of variables where abnormal values are detected.

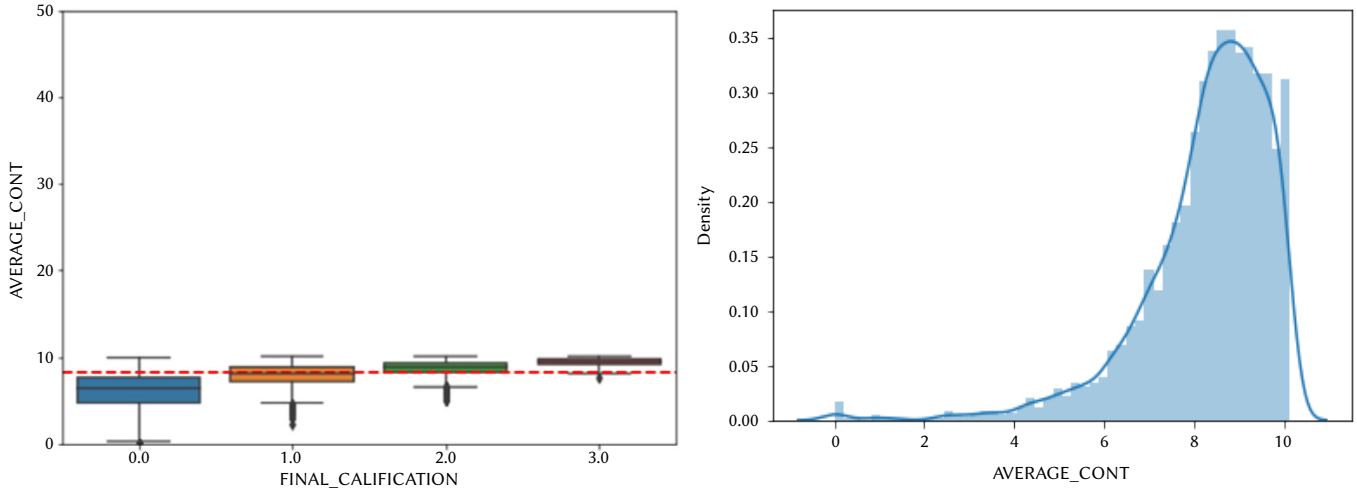


Fig. 6. Abnormal data removal.

(RandomForestClassifier model) and Neural Networks (Sequential model with two hidden layers, Relu activation, Adam optimizer and output activation function for the four Softmax classes) have been parameterized and used.

V. Analysis of Results

Note that under the same input data set, a double classification problem is being addressed. On the one hand, the problem of predicting the final grade of an exam and, on the other hand, the prediction of the final master's degree course. For both predictions, the same configuration of the algorithms will be used, so that the comparison is homogeneous. The comparison of results obtained can be seen in the tables of this section, where the results of each of the algorithms used for each of the predictions made are shown, differentiating between Machine Learning algorithms (ML) and Deep Learning algorithms (DL).

The configuration of the Naïve bayes algorithm configuration is described in Table III.

The configuration of the Decision Tree algorithm configuration is described in Table IV.

The configuration of the Random Forest algorithm configuration is described in Table V.

TABLE III. Naive Bayes Algorithm Configuration

Naive Bayes		
Parameter	Value	Range
Model	BernoulliNB	BernoulliNB
Alpha	1.0	0.5-1.0
Binarize	True	True-False
Fit_prior	False	False
Class_prior	None	None

TABLE IV. Decision Tree Algorithm Configuration

Decision Tree		
Parameter	Value	Range
Model	DecisionTreeClassifier	DecisionTreeClassifier
criterion	entropy	entropy
min_samples_split	20	10-30
min_samples_leaf	4	4-10

TABLE V. Random Forest Algorithm Configuration

Random Forest		
Parameter	Value	Range
Model	RandomForestClassifier	RandomForestClassifier
bootstrap	True	True
criterion	gini	gini
n_estimators	20	10-100

The configuration of the Neural Network algorithm configuration is described in Table VI and Fig. 7.

TABLE VI. Neural Network Configuration

Red Neuronal		
Parameter	Value	Range
Model	Sequential	Sequential
Input	9 dimensions	9 dimensions
Hide layers	2	2-5
Output layer	1	1
Optimizer	Nadam	Nadam, Adam, sgd,
Loss	mean_squared_error	mean_squared_error
Metrics	Accuracy	Accuracy
Activation function. Hidden layers.	Relu	Relu,tanh
Activation function. Output	Softmax	Softmax
Batch_size	20	10-80
Epochs	300	10-500
Num_classes	4	4

```
Model: "sequential_6"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_18 (Dense)             (None, 18)                180
_________________________________________________________________
dropout_13 (Dropout)         (None, 18)                0
_________________________________________________________________
dense_19 (Dense)             (None, 9)                 171
_________________________________________________________________
dropout_14 (Dropout)         (None, 9)                 0
_________________________________________________________________
dense_20 (Dense)             (None, 9)                 40
=================================================================
Total params: 391
Trainable params: 391
Non-trainable params: 0
_________________________________________________________________
```

Fig. 7. Fully-connected Neural Network configuration.

It can be seen that the student's final exam grade is easier to predict than the master's degree grade, which shows that the work of continuous assessment is clearly reflected in the final exam grade. In the case of the final exam grade, results are achieved with an accuracy of 96%, while in the master's degree grade, maximum figures of 70% accuracy are achieved.

In the prediction of the test score, in Table VII, the algorithms that have made the best prediction are Decision Tree and Random Forest, all exceeding a 75% prediction, where the worst result has been Naive Bayes.

TABLE VII. Comparative Table of Results

Algorithm	Type	Master Result	Exam Result
Naïve Bayes	ML	63%	76%
Decision Tree	ML	68%	96%
Random Forest	ML	70%	96%
Neural Networks	DL	62%	81%

It is also important to note that the neural network has a much higher cost of configuration and execution than Random Forest, so, under this configuration, it is convenient to go deeper into Random Forest than into the Neural Network. Regarding the prediction of the final grade of the Master, it is a much more complex prediction, since all the subjects and their results must be taken into account, both in continuous evaluation and in the final exam, but whose relationship is not as direct as in the prediction of the exam grade. In this case, once again the model that has worked best is Random Forest, with 70% correct predictions, while the Neuronal Network has the lowest accuracy, with 62% correct predictions, being significantly worse than Naïve Bayes. The algorithms used give really good figures to be a first approximation, suggesting that the data processing is correct and the methodology appropriate.

VI. Conclusions and Future Lines

A. Preamble

Predicting the grades of students is a powerful tool that helps the student and the university in a remarkable way, and it is a reality today. Artificial Intelligence has put the strings on so that we are able to predict and infer future situations. To make these predictions we need order, data, method and tools that enable us to make them. Thanks to the dataset provided by the university, today we have a set of data from the University master's in computer security, which has allowed us to undertake this project successfully. On the data obtained directly from the LMS (Learning Management System) platform, we have been able to compose a unique set of data with which we have been able to carry out the comparisons of the Artificial Intelligence techniques that best fit, based on a proposed methodology. With the help of tools such as Watson Studio and Python, it has been possible to obtain a multipurpose dataset, which allows its use in algorithms for the prediction of student's final exam grade and the master's degree grade. It's not a simple or fast task, nor is it problem-free, but in the end, a coherent and tangible comparison has been achieved. Through the proposed model, and using Random Forest as a prediction tool, figures of over 95% correct prediction have been obtained and, what is more remarkable in comparison to the objectives of the present work, we have a well-founded comparison of the algorithms used and the proposed methodology, which enable the original dataset to be used for this purpose.

B. Summary of Contributions

The present work has made a comparison of Artificial Intelligence algorithms with the aim of addressing the problem of grade prediction in university environments, seeking to contribute to this problem by providing new information on the treatment of such classification problems in a very defined environment. The following contributions can be identified in Table VIII.

In the State of the Art we have addressed the question of which algorithm is best suited to this type of prediction, demonstrating that there is no one algorithm that is clearly better than another, but that it will depend to a large extent on the data and the treatment that is carried out. With this work, we have been able to verify that with decision tree we have obtained results of 96% accuracy, which gives grounds to continue working with this algorithm, as suggested in future work.

C. Future Work

The results and conclusions obtained in this work present an opportunity to continue working on the prediction of academic grades in UNIR students. Within the future work that can be done, as future lines that can take this work as a basis, the following actions are proposed: i. To deepen in the parameterization of the proposed algorithms, with special focus in Random Forest, in search of higher prediction values than those obtained in this work. Although very high values have been obtained in the student's final exam grade prediction, the same figures are not obtained in the master's degree grade, so there is an exciting field of research in this regard. ii. Increase the dataset

with data from more years and make the model and methodology proposed to see how it behaves. In this sense, it would be interesting to train the model with all the available data and use the current year to make an inference of results, thus validating the model. iii. To increase the dataset with demographic data of the students, which would allow to extend the scope of the study and to be able to conclude in more directions, such as the impact of the family situation on the academic performance, curricular adaptations, new support subjects, etc. iv. Test the same methodology and algorithms in different UNIR studies to see if it can begin to evolve towards generalization at the University.

TABLE VIII. Summary of Contributions

Contribution	Description	Tangible
Dataset collection	Generation of a single dataset from the 11 files provided by the SAKAI platform	The dataset of the SAKAI platform has been obtained and a working methodology has been presented in order to address the problem. Finally a dataset with the 27 most significant characteristics and 4522 records has been obtained
Dataset increase	Generation of the qualification master degree grade from the data of course of each student	Obtaining the characteristic variable, called FINAL_ CALIFICATION
Dataset cleaning	Removal noise from the input variables of the final dataset	The treatmen of the data carried out, before being used by the AI algorithms, is exposed
Implementation of AI algorithms	Implementation of AI algorithms uset to contribute to the problem of university grade prediction (classification problem)	The implementation of the algorithms used in this work can be found and used freely through the following link: https://github.com/HectorAMG/Algoritmos-IA
Algorithms comparision	Comparison of the results obtained	Identification, configuration, use and comparative results of the use of the different algorithms to address the problem of predicting grades of university students

References

[1] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting Student Performance Using Personalized Analytics", *Computer*, vol. 49, no. 4, pp. 61–69, 2016.

[2] L. Gerritsen, "Predicting Student Performance with Neural Networks," Ph.D. dissertation, School of Humanities, Tilburg University, Tilburg, Netherlands, 2017.

[3] Y. Jiang, R. S. Baker, L. Paquette, M. San Pedro, & N. T. Heffernan, "Learning, moment-by-moment and over the long term", in International Conference on Artificial Intelligence in Education, Madrid, Spain, 2015, pp. 654–657, doi: https://doi.org/10.1007/978-3-319-19773-9_84

[4] T. Mishra, D. Kumar, & S. Gupta, "Students' employability prediction model through data mining", International Journal of Applied Engineering Research, vol. 11, no. 4, pp. 2275–2282, 2016.

[5] V. Singh & A. Thurman, "How Many Ways Can We Define Online Learning? A Systematic Literature Review of Definitions of Online Learning (1988-2018)," American Journal of Distance Education, vol. 33, no. 4, pp. 289–306, 2019.

[6] R. Stillwell, & J. Sable, "Public School Graduates and Dropouts from the Common Core of Data: School Year 2009–10", National Center for Education Statistics, US Department of Education, USA, 2013. Accessed: Feb. 15, 2019. [Online]. Available: https://nces.ed.gov/pubs2013/2013309rev.pdf.

[7] R. J. Sternberg. "Teaching College Students that Creativity Is a Decision", Guidance & Counselling, vol. 19, no. 4, pp. 196-200, 2004.

[8] J.M. Tomás & M. Gutiérrez, "Aportaciones de la teoría de la autodeterminación a la predicción de la satisfacción escolar en estudiantes universitarios", Revista de Investigación Educativa, vol. 37, no. 2, pp. 471–485, 2019.

[9] C. J. Villagrá-Arnedo, F. J. Gallego-Durán, F. Llorens-Largo, R. Satorre-Cuerda, P. Compañ-Rosique, & R. Molina-Carmona. "Time-Dependent Performance Prediction System for Early Insight in Learning Trends", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 2, pp. 112-124, 2020, doi: 10.9781/ijimai.2020.05.006.

[10] V.M. Cojocariu, I. Lazar, V. Nedeff, & G. Lazar, "SWOT Analysis of E-learning Educational Services from the Perspective of their Beneficiaries", Procedia-Social and Behavioral Sciences, vol. 116, pp. 1999–2003, 2014, doi: 10.1016/j.sbspro.2014.01.510.

[11] P. Colás Bravo, "El abandono universitario", Revista Fuentes, no. 16, pp. 9–14, 2015, doi: 10.12795/revistafuentes.2015.i16.

[12] S. Regha R. & D. U. Rani, "An Efficient Clustering Based Feature Selection for Predicting Student Performance", International Working Group on Educational Data Mining, vol. 9, no. 2, pp. 524–531, 2017, doi: 10.21817/ijet/2017/v9i2/170902328.

[13] G. W. Dekker, M. Pechenizkiy, & J. M. Vleeshouwers, "Predicting students drop out: A case study", in International Working Group on Educational Data Mining 2009, Córdoba, Spain, 2009, pp. 41–50.

[14] Q. Hu, A. Polyzou, G. Karypis, & H. Rangwala, "Enriching course-Specific regression models with content features for grade prediction", in 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017, pp. 504–513. doi: 10.1109/DSAA.2017.74.

[15] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, & V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques", Computers & Education, vol. 53, no. 3, pp. 950–965, 2009, doi: 10.1016/j.compedu.2009.05.010.

[16] J. Xu, K. H. Moon & M. van der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 5, pp. 742-753, 2017, doi: 10.1109/JSTSP.2017.2692560.

[17] R. Heredia, J. Jobany, H. Rodríguez, G. Aida & J.A. Vilalta, "Predicting performance in a subject using ordinal logistic regression". Estudios pedagógicos (Valdivia), vol. 40. no 1, pp. 145–162, 2014. doi: 10.4067/s0718-07052014000100009.

[18] J. G. Cleary and L. E. Trigg, "K*: An Instance-based Learner Using an Entropic Distance Measure" in Machine Learning Proceedings, M. Kaufmann Publishers, 1995, pp. 108–114. doi:10.1016/b978-1-55860-377-6.50022-0.

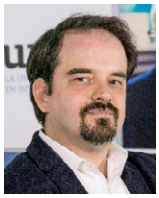
[19] T. Miranda Lakshmi, A. Martin, and V. Prasanna Venkatesan, "An Analysis of Students Performance Using Genetic Algorithm", Journal of Computer Sciences and Applications, vol. 1, no 4, pp. 75-79, 2013, doi: 10.12691/jcsa-1-4-3.

[20] E. Osmanbegovic, M. Suljic, "Data Mining Approach for Predicting Student Performance" in Journal of Economics and Business, University of Tuzla, Faculty of Economics, vol. 10, no. 1, pp. 3-12, 2012. [Online] Available: http://hdl.handle.net/10419/193806.

[21] A. Hamoud, A. S. Hashim, & W. A. Awadh, "Predicting student performance in higher education institutions using decision tree analysis", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 5, no. 2, pp. 26-31, 2018, doi: 10.9781/ijimai.2018.02.004.

Héctor Alonso-Misol Gerlache

Héctor Alonso-Misol Gerlache was born in Madrid in 1976, he has been dedicated to the IT environment for 20 years. Besides his family, he has three passions, technology, education and people, which are reflected in the company he founded in 2020, GradientIA TMC, S.L., providing agile services of SW development with Artificial Intelligence technologies.



Pablo Moreno Ger

Dr. Moreno-Ger was born in Madrid in 1981. He finished his doctorate in Computer Engineering from Universidad Complutense de Madrid (UCM) in 2007 and was an Associate Professor in the Department of Software Engineering and Artificial Intelligence at UCM. Now he is with Universidad Internacional de La Rioja (UNIR), where he is currently the Vice-Rector for Research. Formerly, he was the Director of the School of Engineering and Technology at UNIR, as well as Vice-Dean for Innovation at the School of Computer Engineering at UCM. His main research interests are in technology-assisted teaching, artificial intelligence, learning analytics and serious games. He has published more than 150 academic works in these fields.

Luis de la Fuente Valentín

Luis de la Fuente Valentín is a full-time associate professor at Universidad Internacional de La Rioja, UNIR. He got his PhD at Universidad Carlos III de Madrid, in 2011. He has authored more than 40 papers and participated in several national and European public funded projects, one of them as investigator in charge. His current research interest is on machine learning tools applied to the educational field.