Human and Artificial Intelligence for the Society of the Future
European Distance and E-Learning Network (EDEN) Proceedings
2020 Annual Conference | Timisoara, 22-24 June, 2020
ISSN 2707-2819
doi: 10.38069/edenconf-2020-ac0023

USE OF ARTIFICIAL INTELLIGENCE TO PREDICT UNIVERSITY DROPOUT: A QUANTITATIVE RESEARCH

Francesco Agrusti, Gianmarco Bonavolontà, Mauro Mezzini, Roma Tre University, Italy

Abstract

The main aim of the research is to predict, as early as possible, which student will drop out in the Higher Education (HE) context. Artificial Intelligence (AI) is used for replacing repetitive human activities, e.g. in the field of for autonomous driving or for the task of classification pictures. In these areas IA competes with the man with fairly satisfactory results and, in the case of college dropout, it is extremely unlikely that an experienced teacher can "predict" the student's academic success based on only on data provided by administrative offices. In this study used administrative data of about 6,000 students enrolled in the Department of Education of the University of Roma Tre to train convolutive neural nets (RNC). The trained network provides a probabilistic indicating, for each student, the probability of abandonment. Then, the trained network provides a predictive model that predicts whether the student will dropout. The accuracy of the obtained deep learning models ranged from 67.1% for the first-year students up to 94.3% for the third-year students.

Introduction

In the comparative study on dropping out of higher education in Europe conducted by Vossensteyn and other researchers (2015), it was found that successful studies are seen as a crucial factor for personal success in 28 of the 35 participating countries. Early recognition of dropout is a key prerequisite for reducing dropout rates: several studies highlight the importance of monitoring individual and social characteristics of students as they have a strong impact on the probability of success of students in higher education. A key objective of the Europe 2020 strategy is in fact to reduce drop-out rates by seeking to achieve at least 40% of 30-34 year olds completing higher education (Vossensteyn et al., 2015). As reported in the literature, students generally leave during their first year of university (Larsen et al., 2013), immediately after upper secondary school: in this period, they must develop their sense of responsibility and self-regulation (Pintrich & Zusho, 2002). Individual skills and dispositions are investigated in different psychological and pedagogical models in relation to the phenomenon of early abandonment in terms of

personality characteristics (Pincus, 1980). Numerous studies have explored the impact of the economic and social status of students (e.g. race or income) and the organisational services provided to students by the university (e.g. faculty-student relationship) on the drop-out rate (Pincus, 1980). For decades, one of the most used and discussed models has been Tinto's "student integration" model, which underlines the importance of the academic and social integration of students in predicting the phenomenon of early school leaving (Tinto, 2010). One of the other main models is the one proposed by Bean (1988), the "student attrition" model, based on the attitude-behaviour of the student, which measures individual and institutional factors and evaluates their interactions in order to predict university dropout. Another interesting model of student/institution integration is the Pascarella model (Pascarella & Terenzini, 2005), which emphasizes the cruciality for student success of having informal contacts with teachers. In other words, in this model, background characteristics interact with institutional factors influencing student satisfaction with the university. Numerous studies have demonstrated the positive effects of student-university interaction on persistence (Pascarella & Terenzini, 2005). Event history modelling is another model much discussed in literature: proposed by Des Jardins, Albourg, and Mccallan (1999), this model takes into account the role of the succession of different events in the different stages of the student's educational career, changing the importance of factors from year to year, depending on the time period. In all these models, the relationship between students and institutions is crucial to reduce drop-out rates and several variables have been identified to improve student retention (Siri, 2015). In Italy, due to the very high drop-out rates of university students (ANVUR, 2018), several specific studies were conducted (Burgalassi et al., 2016) which confirmed the value of the baccalaureate vote (and of the entry skills of students more generally) together with the socio-demographic traits of the students (mostly the socio-economic context) as valid indicators of university drop-out compared to the outcome of the first year of study. Many of the models and studies conducted, both national and international, have presented different analyses from the psychological point of view, building psychological-motivational models focused on expectation, reasons for involvement, personal value and motivation in general. These models and surveys all involve the collection of data by interviewing students directly, through the use of tools (usually questionnaires) specially administered. The study presented in this article, however, aims to use only the data available in any university statistical office, without therefore, at least at this stage of research, interviewing students directly. In this regard, it was decided to proceed to the analysis of these data through the use of Artificial Intelligence (AI). Today, AI is used to replace human activities that are repetitive, for example, in the field of autonomous driving or for the task of classifying images. In these areas, IA competes with man with quite satisfactory results and, in the case of abandonment of the educational system, it is extremely unlikely that an experienced teacher will be able to "predict" the educational

success of the student on the basis of data provided by the administrative offices. These recent advances on neural networks have shown that AI may be able to compete (or even exceed) with human capabilities in the tasks of classification and recognition. Here below are then first shown some of the most important studies obtained thanks to the IA, on the prediction of university dropout. Then the metrics for the evaluation of these models and then the methodology used and the results obtained in our research are presented. Preliminary conclusions on the study are therefore briefly drawn.

State of the Art

Many research projects have used data mining techniques to study the Dropout phenomenon. Specifically, in this section we will discuss work that has investigated university dropout by developing predictive models through EDM (Educational Data Mining), or the use of data mining in education, applying computer methods to analyse large data collections. From the analysis of the literature it emerged that the decision tree algorithm (DT) is the one most used for the development of predictive models aimed at identifying university dropout. A research project funded by the Colombian Ministry of Education tried to identify predictive models of early school leaving by analysing 62 attributes belonging to socio-economic, academic and institutional data. Also in this case a decision tree has been implemented (algorithm J48) and for the validation of the model the cross-validation folder has been used with an accuracy of more than 80% (Pereira et al., 2013). Similarly, research was conducted in India to develop a DT based on the ID3 algorithm that could predict students dropping out of university. The study is based on the analysis of 32 variables on a sample of 240 students selected through a survey. Model performance was evaluated using the accuracy index, accuracy, recall and F-measure (Sivakumar et al., 2016). In 2018, research presented a classification based on the DT algorithm. The study analyses 5288 cases of students belonging to the Chilean public university (cohorts of students belonging to 44 university courses in the fields of humanities, arts, education, engineering and health). The attributes selected for the analysis are related to the student's demographic variables, economic situation, and data on previous academic performance prior to his or her admission to university. The accuracy index of the best model developed was 87.2% (Ramírez & Grandón, 2018). In addition to the DT, other classification methods have been used in order to implement models for the prediction of university dropout. Some researchers have used specific methodologies such as CRISP-DM (Cross Industry Standard Process for Data Mining), to predict at the end of the first semester students at risk of dropping out. The dataset consists of over 25 thousand students and 39 variables for each student and the algorithms used are: DT, artificial neural networks (ANN) and logit model (LR). The results show an accuracy of 81.2% for the model developed with ANN (Delen, 2011). Similarly, a research conducted at the University of Genoa, employed the ANNs to detect students at risk of

dropping out. The study refers to a population of 810 students enrolled for the first time in a degree course in medicine in the academic year 2008-2009 and the data come from administrative sources, an ad hoc survey and telephone interviews (Siri, 2015). Another example is the work done at the College of Technology in Mato Grosso. The research presents a model developed with the Fuzzy-ARTMAP neural network using only the registration data collected for a period of seven years from 2004 to 2011. The results show an accuracy rate of more than 85% (Martinho et al., 2013). In Brazil at Universidade Federal do Rio de Janeiro, a research project compared different algorithms (DT, SimpleCart, Support Vector Machine, Naïve Bayes and ANN) analysing data from 14,000 students (Manhães et al., 2014). Similarly, at the University of Technology and Economics in Budapest, using data from 15,285 university students regarding their secondary and university education, 6 types of algorithms were employed and evaluated to identify students at risk of dropouting. Accuracy, recall, precision and the ROC curve are the metrics used for the evaluation and the results showed the best model developed by the Deep Learning algorithm with an accuracy rate of 73.5% (Nagy & Molontay, 2018). A similar research has employed five classification algorithms (LR, Gaussian Naive Bayes, SVM, Random Forest and Adaptive Boosting) analysing 4432 data from the students of the degree courses in Law, Computer Science and Mathematics of the University of Barcelona in the years 2009 and 2014. The research found that all machine learning algorithms reached an accuracy of around 90% (Rovira et al., 2017). The Instituto Tecnológico de Costa Rica implemented a model derived from the algorithms of Random Forest, Support Vector Machine, ANN and LR. The reference sample is composed of 16,807 students enrolled between the years 2011 and 2016 and the best model is that resulting from the algorithm of Random Forest (Solis et al., 2018). The above studies highlight a heterogeneous use of datasets, algorithms, metrics and performance methodologies. Therefore, it is unlikely to be possible to define with certainty which model is better than the other, but research confirms the effectiveness of the EDM approach to the study of university dropout. The main difference that characterizes this work from those present in the literature is given by the introduction of convolutive neural networks to analyse data belonging to the educational field.

A quantitative research at Roma Tre University

One of the most important problems in the field of IA is the problem of classification (LeCun et al., 2015). In this problem you have an object, which can be an image, a sound or a sentence and you want to associate to this object a class taken within a finite set K of classes. A neural network (RN) can be seen as a function φ that takes an input from a vector n-dimensional x and produces a value, called the prediction of x. The prediction is correct when $\varphi(x) = f(x)$ and otherwise incorrect. Contrary to the classic programming paradigm, where the programmer to design an algorithm must have a deep and complete knowledge

of the problem of interest such as in (Malvestuto, Mezzini, & Moscarini, 2011; Mezzini, 2010; 2011; 2012; 2016; 2018; Mezzini & Moscarini, 2015; 2016), to implement an RN the programmer may also be completely unaware of the mechanism or semantics of classification.

We collected, from the administration office of Roma Tre University, a dataset of students enrolled in the Department of Education (DE). The years of enrolment ranges from 2009 up to 2014 comprising a total of 6078 students. We found that 649 of all students were still active at the time when we acquired the dataset (August 2018), while the remaining 5429 closed the course of their studies either because they graduated or because they dropped out or by other reasons, explained later. We refer to this set of students as the no active students. Note that in the following when we will refer to the enrolment year (or simply the year) of a student we mean the number of years passed since her/his first enrolment to university, that is, we refer to an integer value between 0 and 9 since no student is enrolled for more than 9 years. In general, each of the no active student is classified in two different classes: Graduated and Dropout. We excluded later all students which do not classified in these two classes, like for example students who changed faculty within the R3U or went to another university. The number of such students is 118. The number of graduated students is 2833 while the number of who dropped out is 2478. We obtained, from the R3U's administrative office, most of the (out of what were available) administrative fields of all students. The attributes relative to the student's academic career are the following: Exam name, Score of the exam, Maximum score of the exam, ECTS of the exam, Exam date (month/day), Academic year, Type of validation. They represent the attributes relative to each test or exam given by the student. Note that the field "ECTS of the exam" refer to the European Credit Transfer and Accumulation System. In order to construct the training set all the domains of the dataset are converted, using an arbitrary bijective function, to a non-negative integer domain. For example, the domain of the attribute GENDER, was converted to the domain {0,1} where 0 correspond to "male" and 1 to "female". We created a table STUDENT, whose schema S contains all the attributes provided by administrative offices. We limited our tests only to the students that are still active at the year 3 because after that year the number of those students dropping out to university is very small and not significant from statistical and/or practical purposes. If a student ends his/her career in the year z, $0 \leq z < y$, then f_y will take the value δ for every year $z < y \leq 3$. The value of δ, which was arbitrarily chosen to be equal to -1, can be considered as a NULL value and it does not appear in the original domain of any field on the scheme S. Furthermore, for each year of enrolment $y \in \{1, 2, 3\}$ an integer m is set to represent the maximum number of exams sustained by any student on the year of enrolment y. We found that $m_1 = 24, m_2 = 19$ and $m_2 = 23$. Thus, for any field in List 3, for each year y and for each z, $0 \leq z \leq m_y$, we added a field denoted as $g_{y,z}$. If a student in the year $y > 0$ of her/his academic career

completes successfully no more than j exams, then the value of the field $g_{y,z,}$ is set to δ for each $j < z \leq m_y$. Overall the table STUDENT has 530 fields (although we collected data up to year 5 totalling 897 fields).

Table 1: Here we report the confusion matrix for the epochs with the best F_1 measure on the validation set. The confusion matrix for the test set was computed using the very same model that achieved the best F_1 measure on the validation set. Column T stands for table type (A, B or C).

			Validation								Test							
Year	T.	Arch.	Dropout		Degree		Acc.	Prec.	Recall	F1	Dropout		Degree		Acc.	Prec.	Recall	F1
			True	False	True	False					True	False	True	False				
0	B	RNV2	166	105	111	50	64.12%	61.25%	76.85%	68.17%	144	138	121	38	60.09%	51.06%	79.12%	62.07%
0	B	INCRV4	183	120	96	33	64.58%	60.40%	84.72%	70.52%	151	155	104	31	57.82%	49.35%	82.97%	61.89%
0	B	DFSV1	160	96	132	43	67.75%	62.50%	78.82%	69.72%	159	116	113	54	61.54%	57.82%	74.65%	65.16%
1	A	RNV2	65	10	228	20	90.71%	86.67%	76.47%	81.25%	44	16	199	37	82.09%	73.33%	54.32%	62.41%
1	A	INCRV4	67	17	221	18	89.16%	79.76%	78.82%	79.29%	50	23	192	31	81.76%	68.49%	61.73%	64.94%
1	A	DFSV1	65	22	216	20	87.00%	74.71%	76.47%	75.58%	52	26	189	29	81.42%	66.67%	64.20%	65.41%
1	B	RNV2	47	43	186	40	73.73%	52.22%	54.02%	53.11%	33	33	205	52	73.68%	50.00%	38.82%	43.71%
1	B	INCRV4	60	74	155	27	68.04%	44.78%	68.97%	54.30%	52	76	162	33	66.25%	40.63%	61.18%	48.83%
1	B	DFSV1	62	94	143	24	63.47%	39.74%	72.09%	51.24%	51	87	141	24	63.37%	36.96%	68.00%	47.89%
1	C	RNV2	61	23	215	24	85.45%	72.62%	71.76%	72.19%	44	25	190	37	79.05%	63.77%	54.32%	58.67%
1	C	INCRV4	54	24	233	17	87.50%	69.23%	76.06%	72.48%	42	34	195	43	75.48%	55.26%	49.41%	52.17%
1	C	DFSV1	54	28	229	17	86.28%	65.85%	76.06%	70.59%	45	30	199	40	77.71%	60.00%	52.94%	56.25%
2	A	RNV2	35	6	228	15	92.61%	85.37%	70.00%	76.92%	15	6	221	21	89.73%	71.43%	41.67%	52.63%
2	A	INCRV4	38	13	221	12	91.20%	74.51%	76.00%	75.25%	17	8	219	19	89.73%	68.00%	47.22%	55.74%
2	A	DFSV1	33	6	228	17	91.90%	84.62%	66.00%	74.16%	14	4	223	22	90.11%	77.78%	38.89%	51.85%
2	B	RNV2	16	9	243	15	91.52%	64.00%	51.61%	57.14%	11	14	213	41	80.29%	44.00%	21.15%	28.57%
2	B	INCRV4	15	5	247	16	92.58%	75.00%	48.39%	58.82%	12	13	214	40	81.00%	48.00%	23.08%	31.17%
2	B	DFSV1	17	10	242	14	91.52%	62.96%	54.84%	58.62%	15	21	206	37	79.21%	41.67%	28.85%	34.09%
2	C	RNV2	29	7	211	16	91.25%	80.56%	64.44%	71.60%	17	15	237	14	89.75%	53.13%	54.84%	53.97%
2	C	INCRV4	32	14	201	13	89.62%	69.57%	71.11%	70.33%	22	23	235	18	86.24%	48.89%	55.00%	51.76%
2	C	DFSV1	30	10	205	15	90.38%	75.00%	66.67%	70.59%	25	18	240	15	88.93%	58.14%	62.50%	60.24%
3	A	RNV2	19	3	94	6	92.62%	86.36%	76.00%	80.85%	13	11	91	10	83.20%	54.17%	56.52%	55.32%
3	A	INCRV4	19	1	96	6	94.26%	95.00%	76.00%	84.44%	13	4	98	10	88.80%	76.47%	56.52%	65.00%
3	A	DFSV1	20	3	94	5	93.44%	86.96%	80.00%	83.33%	12	3	99	11	88.80%	80.00%	52.17%	63.16%
3	B	RNV2	14	6	91	11	86.07%	70.00%	56.00%	62.22%	7	8	94	16	80.80%	46.67%	30.43%	36.84%
3	B	INCRV4	14	4	93	11	87.70%	77.78%	56.00%	65.12%	1	4	98	22	79.20%	20.00%	4.35%	7.14%
3	B	DFSV1	16	8	89	9	86.07%	66.67%	64.00%	65.31%	5	10	92	18	77.60%	33.33%	21.74%	26.32%
3	C	RNV2	17	3	94	8	90.98%	85.00%	68.00%	75.56%	11	5	97	12	86.40%	68.75%	47.83%	56.41%
3	C	INCRV4	18	6	106	4	92.54%	75.00%	81.82%	78.26%	19	8	105	14	84.93%	70.37%	57.58%	63.33%
3	C	DFSV1	17	2	95	8	91.80%	89.47%	68.00%	77.27%	12	5	97	11	87.20%	70.59%	52.17%	60.00%

We build a table called Y_LABEL containing two attributes: STUDENTID and DROPOUT, where the last represents the label of each student. It has a numerical domain with the following meanings: 0, if the student graduated, 1 if the student dropped out. From the table student described above, we derived three type of tables denoted as STUDENT_A_x, STUDENT_B_x and STUDENT_C_x for $0 \leq x \leq 3$ where x is the number of years from the first enrolment.

In the schema of tables STUDENT_A_x we added all the attributes in List 1 and all the attributes in List 2 (of the type f_y), and all the attributes of List 3 (of type $g_{y,z}$) for all $y = 0, \ldots, x$.

The tables denoted as STUDENT_B_x, $x = 0, \ldots, 3$, contain only the attributes of List 1 and List 2. That is, we considered in these tables only administrative fields and we excluded the fields related to the academic careers of the students (the ones of type $g_{y,z}$).

The tables STUDENT_C_x, $x = 0, \ldots, 3$, have been constructed in the following way. We computed, for each student, the following aggregate statistics: DIFFYEAR and for each year $x > 0$, NUMBEREXAMS$_x$, AVGSCORE$_x$ and SUMETCS$_x$. The first statistic contains the value YEAR OF BIRTH − YEAR OF BEGINNING OF STUDIES − 19 that is, the difference in years between the age of the student (at the date of the enrolment) and 19. The other statistics contains, for each student and for each year $x = 1,2,3$ respectively, the number of exams successfully passed, the average score of the exams successfully passed and the sum of the ECTS gained. We thus obtained the schema of STUDENT_C_x by adding to the schema of each table STUDENT_B_x, all the above four fields. The idea we want to test here is whether it is better and effective to use only some significant aggregate statistics or, instead, it is better and effective to use all the attributes relative to the academic career (like in tables STUDENT_A_x). For the tests of both CNN and BN we choose a random permutation of all no active students. Next, we partitioned all students in twelve different mutually disjoint groups containing approximately 450 students each thus obtaining a partition $\mathcal{P} = \{P_0, P_1, \ldots, P_{11}\}$. For all $0 \leq i \leq 11$ the group P_i is used as a validation set V_i and the group $P_{i+1 \bmod 12}$ as a test set T_i and the students in the remaining groups, as the training set A_i. In the validation set for the year x we put only the students who, at that year of enrolment, were still active. We trained three models based on the CNN architectures mentioned above by taking from each of the table above (A or B or C) the training, validation and test sets from the partition $\mathcal{P}$. We got data from a total of 43200 epochs. For each epoch the confusion matrix of both the validation and the test sets were produced. We found that the F_1 measure, was the better indicator for the selection of the best model. We calculated the accuracy for the validation set, for the training set and the F_1 measure. Training and validation data were taken from the table STUDENT_B_0. Furthermore, we computed the value of the F_1 measure for the year 1 and for the years 2 and 3 for the three different tables STUDENT_A_x, STUDENT_B_x and STUDENT_C_x. We observe that in all three cases the value of the F_1 measure relative at the table STUDENT_B_x is always worse in every year. This clearly shows that using only administrative data gives very poor performance in predicting the dropout of a student. In Table 1 we report the data of the confusion matrix, for both validation and test sets, in which the validation set, among the twelve possible different sets of the partition $\mathcal{P}$, achieved the best score on the F_1 measure.

Conclusions

We explored the effectiveness of predicting the dropout from university using three different sets of features. The first one, containing all the academic and administrative features (tables STUDENT_A_x). The second one, containing only administrative features (tables STUDENT_B_x) and the third (tables STUDENT_C_x) containing the administrative features and 3 aggregate statistics about the academic career of the students. The experiment showed that using only administrative features does not give good results and the models using only them are always outperformed by models using also the academic career features or aggregate statistics. Furthermore, the models using, besides administrative features, also aggregate statistics perform slightly worse than the models using only and all the academic careers features. From all the above discussion we clearly conclude that the more accurate data we have the more precise and effective the model's predictions could be. Since it is not required that the prediction process is made in real time, we can train hundreds of models and make multiple prediction in order to reduce the random variation found in the early phase of training. Clearly the system can be made finer by introducing a prediction model every semester or even every trimester or it can be extended to other faculty or other types of students.

References

ANVUR. (2018). *Report on the state of the university system and research 2018.*

Bean, J. P. (1988). *Leaving college: Rethinking the causes and cures of student attrition.* Taylor & Francis.

Burgalassi, M., Biasi, V., Capobianco, R., & Moretti, G. (2016). The phenomenon of early school leaving. A case study on the degree courses of the Department of Education Sciences of the University "Roma Tre". *Italian Journal of Educational Research, 17*, 131-152.

Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice, 13*(1), 17–35.

Des Jardins, S. L., Albourg, D. A., & Mccallan, P. B. (1999). An event history model of student departure. *Economics of Education Review, 18*, 375–90.

Larsen, M. R., Sommersel, H. B., & Larsen, M. S. (2013). *Evidence on dropout phenomena at universities.* Danish Clearinghouse for educational research Copenhagen.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

Malvestuto, F. M., Mezzini, M., & Moscarini, M. (2011). Computing simple-path convex hulls in hypergraphs. *Information Processing Letters, 111*(5), 231–234. https://doi.org/10.1016/j.ipl.2010.11.026

Manhães, L. M. B., da Cruz, S. M. S., & Zimbrão, G. (2014). The Impact of High Dropout Rates in a Large Public Brazilian University–A Quantitative Approach Using Educational Data Mining. *CSEDU, 3*, 124–129.

Martinho, V. R. D. C., Nunes, C., & Minussi, C. R. (2013). An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. *Proceedings of the 2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, 159–166.

Mezzini, M. (2010). On the complexity of finding chordless paths in bipartite graphs and some interval operators in graphs and hypergraphs. *Theoretical Computer Science, 411*(7), 1212–1220. https://doi.org/10.1016/j.tcs.2009.12.017

Mezzini, M. (2011). Fast minimal triangulation algorithm using minimum degree criterion. *Theoretical Computer Science, 412*(29), 3775–3787. https://doi.org/10.1016/j.tcs.2011.04.022

Mezzini, M. (2012). Fully dynamic algorithm for chordal graphs with O(1) query-time and O(n2) update-time. *Theoretical Computer Science, 445*, 82–92. https://doi.org/10.1016/j.tcs.2012.05.002

Mezzini, M. (2016). On the geodetic iteration number of the contour of a graph. *Discrete Applied Mathematics, 206*, 211–214. https://doi.org/10.1016/j.dam.2016.02.012

Mezzini, M. (2018). Polynomial time algorithm for computing a minimum geodetic set in outerplanar graphs. *Theoretical Computer Science, 745*, 63–74. https://doi.org/10.1016/j.tcs.2018.05.032

Mezzini, M., & Moscarini, M. (2015). On the geodeticity of the contour of a graph. *Discrete Applied Mathematics, 181*, 209–220. https://doi.org/10.1016/j.dam.2014.08.028

Mezzini, M., & Moscarini, M. (2016). The contour of a bridged graph is geodetic. *Discrete Applied Mathematics, 204*, 213–215. https://doi.org/10.1016/j.dam.2015.10.007

Mezzini, M., Bonavolontà, G., & Agrusti, F. (2019). Predicting university dropout by using convolutional neural networks. In INTED2019.

Nagy, M., & Molontay, R. (2018). Predicting Dropout in Higher Education Based on Secondary School Performance. *Proceedings of the 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, 389–394.

Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research*. Jossey– Bass Higher & Adult Education.

Pereira, R. T., Romero, A. C., & Toledo, J. J. (2013). Extraction Student Dropout Patterns with Data mining Techniques in Undergraduate Programs. In *KDIR/KMIS*, 136–142.

Pincus, F. (1980). The false promises of community colleges: Class conflict and vocational education. *Harvard Educational Review, 50*(3), 332–361.

Pintrich, P. R., & Zusho, A. (2002). The development of academic self–regulation: The role of cognitive and motivational factors. In Development of achievement motivation. Elsevier. 249–284.

Ramírez, P. E. & Grandón, E. E. (2018). Prediction of Academic Dropout in a Chilean Public University through the Classification based on Decision Trees with Optimized Parameters. *University Education, 11*(3), 3-10.

Rovira, S., Puertas, E., & Igual, L. (2017). Data–driven system to predict academic grades and dropout. *PLoS one, 12*(2), e0171207.

Siri, A. (2015). Predicting students' dropout at university using artificial neural networks. *Italian Journal of Sociology of Education, 7*(2).

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology, 9*(4), 1–5.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. *Proceedings of the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 1–6.

Tinto, V. (2010). From theory to action: Exploring the institutional conditions for student retention. In *Higher education: Handbook of theory and research* (pp. 51–89.). Springer.

Vossensteyn, J. J., Kottmann, A., Jongbloed, B. W. A., Kaiser, F., Cremonini, L., Stensaker, B., Hovdhaugen, E., & Wollscheid, S. (2015). *Dropout and completion in higher education in Europe: main report.* https://doi.org/10.2766/826962