Proposing a Machine Learning Approach to Analyze and Predict Employment and its Factors

Francisco J. García-Peñalvo[1,3,4]*, Juan Cruz-Benito[1,3,4], Martín Martín-González[5], Andrea Vázquez-Ingelmo[1,3,4], José Carlos Sánchez-Prieto[1,4], Roberto Therón[1,2,3]

[1] GRIAL Research Group. University of Salamanca (Spain)
[2] VisUSAL Research Group. University of Salamanca (Spain)
[3] Department of Computer Science. University of Salamanca (Spain)
[4] Research Institute for Educational Sciences. University of Salamanca (Spain)
[5] UNESCO Chair in University Management and Policy. Technical University of Madrid (Spain)

Received 3 November 2017 | Accepted 22 January 2018 | Published 2 February 2018

Abstract

This paper presents an original study with the aim of propose and test a machine learning approach to research about employability and employment. To understand how the graduates get employed, researchers propose to build predictive models using machine learning algorithms, extracting after that the most relevant factors that describe the model and employing further analysis techniques like clustering to get deeper insights. To test the proposal, is presented a case study that involves data from the Spanish Observatory for Employability and Employment (OEEU). Using data from this project (information about 3000 students), has been built predictive models that define how these students get a job after finalizing their degrees. The results obtained in this case study are very promising, and encourage authors to refine the process and validate it in further research.

Keywords

Employability, Employment, Artificial Intelligence, Machine Learning, Random Forest, Academic Analytics, OEEU.

DOI: 10.9781/ijimai.2018.02.002

I. Introduction

THE concept of employability has steadily gained importance in recent years, becoming one of the pillars of the European educational strategy within the European Higher Education Area (EHEA) framework. However, the empirical research is still insufficient to build a strong theoretical foundation. It is worth noting that applied research on employability has, nowadays, an exploratory approach. This is because this research area presents difficulties regarding to have adequate, reliable and updated data, as well as this early status not only prevents agreement on research results are reached, but also poses many questions as to which methodologies and approaches are most appropriate to address these issues. For these reasons, the area is still growing and need to push the outcomes to further research levels.

Several research projects have been developed in recent years to provide more information on the employability of graduates, many of which have been driven by the OECD and the European Commission. These projects have faced at least two problems: first, the lack of a single, consensual definition of employability and, secondly, the difficulty of obtaining summary indicators to assess it.

Indeed, employability is a theoretical construct whose definition varies according to academic discipline and the perspective used, as well as the socioeconomic context to which it refers. There is no clear consensus on the factors that compose or determine it, nor on the employment outcomes to which it leads. Therefore, evaluating employability is a tough task. In any case, given the complexity of the notion of employability, it would be worth using several variables and indicators that assess different labor, educational and sociodemographic issues, rather than a summary indicator.

Most of the studies on employability that have been developed since the 1990s have focuses on identifying the competencies that graduates will need throughout their career path. Some have gone further, introducing other variables related to the training and education that offer the academic institutions, the sociodemographic context, the institutional and normative framework and the productive structure ("broad" approach as presented in [1]).

This kind of research, despite its initial status, is focused on develop successful strategies and outcomes that could help policymakers and institutions to enhance and promote those detected factors that contribute to get more chances of employment and better employments. For that reasons and its application in the society, it is possible to affirm that the project is in the scope of emergent areas like the Academic Analytics [2-6] or Institutional Intelligence [7, 8].

This paper aims to present a new method to analyze employability factors and to analyze how people gets employed. To achieve that, this paper proposes a machine-learning-based approach that produce predictive models on employment, providing the main factors that affect the predictive model and finding the most relevant ones. This approach contrasts to the previous state of the art in this research area. As will be explained in the Background section, previous approaches are based on basic statistical processes and tries to accomplish the problem of employment and its factors as a whole, instead of weighting the relevance of each factor to build more complex models. To illustrate these considerations, the paper provides a case of study where has been

* Corresponding author.
E-mail address: fgarcia@usal.es

applied the approach and shows some promising results.

The research presented in this paper is developed under the scope of the Spanish Observatory for University Employability and Employment (OEEU in its Spanish acronym) [9]. This observatory gathers data about employment and employability parameters among the Spanish graduates (after they leave the university) to analyze the information they provide and understand what the employment trends and most important employability factors are for this population [10]. To accomplish this mission, the observatory has developed a complex information system [6, 11-13] that collects and analyzes data to present the insights to the researchers [14, 15]. These data collected are used as a dataset to test the machine learning approach that will be presented in the following sections.

This paper has the following structure: second section (Background) presents the state of the art in the case of research applied on employability and employment analysis. Third section (Proposal) describes the machine learning approach and the methods and materials used in this research. Section fourth (Case Study: OEEU) shows the case study developed and the initial results achieved. Fifth section (Discussion) discusses about the implications of the research presented and the results achieved. Sixth section (Conclusions) finalizes the paper with some final remarks and introduces some future work.

II. Background

One of the main competencies studies promoted by the OECD was the "Definition and Selection of Key Competences" (DeSeCo). Since the first editions of the Programme for International Student Assessment (PISA) it had become clear that job success depended on a much greater range of competencies than those considered in the project. The DeSeCo project was created to identify these key competencies, aiming to serve as a framework to guide and complement two international programs to evaluate competencies: the aforementioned PISA and the Adult Literacy and Lifeskills (ALL). The DeSeCo project began in 1997 and ended in 2003, when it published the final report entitled "Key Competencies for a Successful Life and Well-functioning Society" [16]. In this project worked academics and experts from different fields (sociologists, philosophers, psychologists, economists, anthropologists, historians, statisticians, educators, etcetera) and social institutions (political parties, unions, employers, associations, etcetera) to define and figure out key competencies based on previous research. The final list of competencies was discussed in depth in two international symposia until achieve an agreement on the most important. This project is one of the foundational approaches and projects for the employment and employability analytics knowledge area.

One of the most influential projects in the EHEA when defining, identifying and classifying competences was the Tuning Educational Structures in Europe (usually called Tuning project). This project was funded by the European Commission within the Socrates framework. The project was divided into two phases, the first of which was more significant. It was active between 2000 and 2002. Its main objective was to figure out and classify the competences that the graduates require on their career path. Experts from different fields of knowledge (Business Administration, Geology, History, Mathematics, Physics, Education and Chemistry), from several European countries (Germany, Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Iceland, Italy, Norway, Netherlands, Portugal, Spain and Sweden) worked on the project. A total of 5803 graduates, 944 employers, and 998 scholars were surveyed; with the participation of more than 100 educational institutions of the European Union. The research finally divided the competencies into two main groups — specific and generic— and, in turn, the generic ones into three groups: instrumental, interpersonal and systemic.

Driven by the creation of the EHEA, different employability and competency projects have been developed that have enabled comparisons between European countries and universities using a common methodology. One of these projects was the "Higher Education and Graduate Employment in Europe" (usually called CHEERS project — "Careers after Higher Graduation. A European Research Study"). The project was promoted and financed by the European Commission within the "Targeted Socio-Economic Research Programme" (TSER). It began in 1997 and ended in 2000. It was led by The International Centre for Higher Education Research at the University of Kassel (Germany) and included other countries like Germany, Austria, Spain, Finland, France, Iceland, Italy, Japan, Norway, Netherlands, United Kingdom, Czech Republic and Sweden. Between 1998 and 2000 the different research groups sent a standardized questionnaire to the graduates who had completed their studies in the academic year 1994/1995. 37000 subjects were interviewed (about 3000 from each university). The information collected was based on their studies and the career path to analyze the relationship between higher education and employment (job position, mismatch in the labor market, etcetera). An important part of the questionnaire, in which the project had put special emphasis, was the assessment of the level of graduates' competencies and the level that they required by employers.

The CHEERS project was the starting point for the development, in 2006, of "The Flexible Professional in the Knowledge Society: New Demands on Higher Education in Europe" project, usually called REFLEX project. This project aimed to answer, among others, the following questions: what competences do graduates require fulfilling the demands from the modern knowledge society? To what extent has higher education provided these competencies? How can the mismatches between acquired and required competencies be solved?; To what extent are the graduates' expectations met? [17]. It was funded by European Union within the 6th Framework Programme (FP) for Research and Technological Development. It was led by the Research Center for Education and the Labor Market of the University of Maastricht. 14 European countries (Germany, Austria, Spain, Finland, France, Italy, Norway, Netherlands, United Kingdom, Belgium, Czech Republic, Portugal, Switzerland, Estonia) and Japan participated. The methodology and the questionnaire were like that adopted in the CHEERS project. A total of 40787 graduates were surveyed in 1999/2000, 5500 of which corresponded to graduates in Spanish universities (National Agency for Quality Assessment and Accreditation of Spain, 2008). Once again, information about the competencies that the project researchers considered relevant for the promotion of employability was compiled.

In Spain, the study carried out by the Catalan University Quality Assurance Agency (AQU) is one on the most important at the regional level. AQU conducts a telephone survey periodically (every three years) since 2001. The sample is made up of university graduates who finished their studies in any Catalan university three years before the date on which they were surveyed. Among other aspects the outcomes of this study are reports that analyze information related to the quality of employment, job stability, earnings, education-job skills match, job satisfaction, the process of finding a job, mobility, students' satisfaction with their studies, etc. In relation to competences, the questionnaire incorporates a section about skills acquired and their usefulness in the workplace.

Among the latest initiatives to evaluate the competencies of graduates in Spain highlight the *Libro Verde sobre la empleabilidad de los egresados de la Comunidad Valenciana* (Green book on the graduates' employability of Valencian Community). This book was published in 2013 because of a project promoted and funded by the *Generalitat Valenciana* and developed by the Valencian Agency

of Assessment and Prospective (AVAP) and the universities of the region. The report presents the results of a study aimed to clarify the employability and employment situation of graduates at the *Comunidad Valenciana (*Valencian Community) and provided a series of conclusions and recommendations to reinforce and improve their employability. In this project was carried out a survey for those who had completed their short and long cycle studies during the years 2008 and 2009, comprising 2099 graduates in the study. This survey gathered information about the level of mastery of 21 generic competences that the graduates should have, as well as the level required by employers and the level acquired in universities.

Regarding to the methodologies applied in these studies, it is worth noting that they mainly use descriptive statistics, basically applying basic measures of central tendency and frequency distribution. They also rely mainly on basic charts to present the information in a simple way, basically using of histograms and bar charts. As an example, it is common to find in this type of studies the percentage distribution of graduates according to the strategy used in the search for employment, the time (months) taken for graduates to find their first job, the average earnings, the percentage of graduates satisfied with their jobs or their studies, the distribution of the graduates according to the education and job match, the average level of competences required by employers and acquired at the universities, etc. The number of variables involved in these studies varies notably depending on the scope of the project and its objectives. Among the largest is the REFLEX project, which almost included 500 variables distributed in 11 categories: study program, other educational and related experiences, transition from study to work, first job after graduation, employment history and current situation, current work, work organization, competencies, evaluation of study program, values and orientations, about yourself, plus the variables of country and study identification.

This kind of studies have attracted the attention from numerous researchers from different disciplines who, given the increasing availability of data, have been able to carry out some empirical papers. Thus, much of the research has focused on identifying the competencies required by employers and evaluating their impact on employment [18-21] as well as assessing the mismatches between acquired and required competencies [22, 23]. In addition, these research works allowed deepening the relationship education-labor market, as well as the mismatches and their effects [23-28]. They have applied all kinds of analytical methodologies. Although there is no common practice, they use more sophisticated techniques and tools than those applied by general studies, linked to econometrics, psychometry and other quantitative and qualitative methods from research in social sciences.

III. Proposal

In the case of this paper, researchers followed this kind of new approaches. Using novel methods and techniques like machine learning could open new possibilities and ways to work in assessing employability and employment. Also, these methods could unleash new ways for manage huge amounts of information as a whole but considering each factor in a weighted way within the predictive models to be built. Previously some projects use basic statistics due the difficulty of handling big datasets in a non-automated mode, but in our way, the same procedure can be used to crunch all data related jointly.

In general, the proposal for the analysis (based on machine learning) follows common principles in data science regarding data structuration, tidy data approaches, etc. [29-31]. Because it is a proposal and it is on its initial stage, the approach should be explainable to assess its appropriateness. For this reason, the machine-learning process has been implemented in a white-box way; thus, the researchers have selected algorithms and methods to make the workflow explainable.

Moreover, these main principles, the different details for the analysis pipeline, and methods used in this research are presented below. All of these details are explained in the following workflow (available at [32]):

1. Retrieve dataset about students from OEEU's information system.
2. Filter the desired fields from the datasets and enclosed them in data frame (a data structure like a table).
3. Data cleaning: remove noise data, remove columns (variables) with too many null (NaN) values, and remove all students who have only partial.
4. Normalize data with the One-hot encoding algorithm for categorical values in columns [33].
5. Considering the data gathered and the kind of variable (labeled) to predict (student gets employed or not), the algorithm to use must be related to supervised learning. This is because this kind of algorithm makes predictions based on a set of examples (that consist of a labeled training data set and the desired output variable). Moreover, regarding the dichotomous (categorical) character of the variable to predict, the supervised learning algorithm to apply must be based on classification (binary classification, as we have a label of finalization equal to true or false). According to the authors' previous experience, the possibility of explaining results and the accuracy desired for the classification, a Random Forest classifier algorithm [34] was selected. In this step, the Random Forest algorithm was executed repeatedly to determine the best setup for the dataset given (obtaining the most adequate parameters for the execution).
6. With the best configuration found, train the random forest algorithm (with 33.33% of the dataset) and obtain the predictive model.
7. Using the predictive model, obtain the most important features for the predictive model.

At this point, researchers have built a model that could predict if a person will get employed or not, showing also what are the factors that affect more the result. After that, researchers could use these factors to filter information, generalize knowledge across the dataset, extract what values on these factors lead to get employed or not, etc. As an example, using these most relevant factors, researchers could clusterize students to gain deeper knowledge about what are the main characteristics between those who get an employment and those who do not.

In general, in this paper will be demonstrated how this kind of approach could be suitable for the goal of modelling employment and its factors. To do that, the algorithms and the code used is available publicly at [32]. Unfortunately, due the privacy restrictions that affect the OEEU will not be revealed some kind of data involving individual graduates, universities or other sensitive information, it will be only displayed aggregated (and anonymized) information. For that reason, regarding the processes related to analyze the factors that define the model, only will be shown some generalist figures that could illustrate the procedure and give some clues about a real implementation.

The programming language used to conduct all the analyses and calculations was Python. The Python software tools and libraries used to code and test the approach were:

- Pandas software library [29, 35], to manage data structures and support analysis tasks.
- Scikit-learn [36] library, to accomplish the machine learning workflow [33].
- Jupyter notebooks [37-39], to develop the Python code used in this research.

IV. Case Study: OEEU

As presented in the previous sections, the projects that investigated employability and employment disciplines varied in number and type of variables available, scope of the project, etc. For this paper, researchers used the information gathered by the Spanish Observatory for Employability and Employment (OEEU). This project keeps information about 182000 students graduated from degree and master studies. It includes about 400-500 variables per each edition of the study (one edition about degree graduates – 134129 students involved – and other edition about master studies –47822 students–).

The data used in this case study correspond to the information available from graduated students that finalized a degree in the course 2009-2010. This is, information about 134129 students, with 493 variables per student [9, 32]. Despite of the dimensions of the dataset, it is worth noting that not all the students have information for all the possible variables. The Observatory gathers the information by using two input methods: the raw records from the Spanish universities and the information provided by the students through fulfilling questionnaires. These two main sources of information have only part of the variables marked as required, for that reason some of them appear with empty values. Also, the fact of ingesting information via web forms (like in this case) make extremely difficult to get all the information, because the graduate can quit the web form in any moment. For that reason, are required methods that clean and wrangle the information like those presented in the previous section.

Apart of cleaning and wrangling the data properly, researchers have excluded some variables included in the OEEU's dataset, since they are related exclusively to some universities (the universities could add some questions to the OEEU questionnaire), and using only the common variables to all students in their test to create the predictive model related to employment. This reduced the dataset from 493 total variables to 383 possible variables per student. These 383 variables per student can be observed in the 8th cell at the provided notebook [32].

Following the workflow outlined in the previous section, after filtering the desired variables, researchers cleaned all those variables which presents to much *NaN* (empty) values. In this case, the threshold used to remove all the weak variables was 10%. This is, all the variables with more than 10% of empty values were discarded for the model construction. This threshold is strict to obtain a stronger model. There are other common procedures to deal with void variables or measures (fill the empty values with others from the dataset, with the mean of the column, etc.), but in this case, researchers preferred to avoid any kind of artificial data that could contaminate the result. After removing all these non-valuable variables, researchers dropped all the students that had any empty value in their information (completing by this way the data-cleaning stage). After this hard-cleaning process, researchers counted with 26 data variables from 9744 graduates. As previously commented, following other conservative methods to deal with empty values would lead to use more variables (columns) and observations (rows), but this is not the focus in this initial test of the approach presented.

After all this work in data preparation, began the machine learning phase. In this case, the third part of the dataset (third part of graduates) was marked as the portion to train the random forest algorithm. Also, researchers selected the variable to predict using the others. In this case, the variable to predict was *'haEstadoDesempleado'* (Have the student been unemployed?) which contains two possible values: 0 (false) value for those students that got a job after finalizing the degree, and 1 (true) for those students that were not employed.

After that, and with this 33% of the observations (3305 students) and the variable to predict, researchers tested programmatically the best setup for the random forest algorithm. This is, the best configuration values for the parameters *randomforestclassifier__max_features* and *randomforestclassifier__min_samples_leaf*. Using the best values found for the parameters, researchers executed (trained) the algorithm to get the corresponding predictive value.

Table I presents the quality metrics [40] of the predictive model built to predict the graduates' employment. As displayed, the precision of the predictive model classifying and predicting the employment or not was of 0.71 (where 0 is the worst precision and 1 the best).

TABLE I. Results of the Predictive Model Built

	Precision[a]	Recall[b]	F1-score[c]	Suport[d]
False	0.73	0.12	0.20	1066
True	0.70	0.98	0.82	2239
Avg / total	0.71	0.70	0.62	3305

[a]The precision is the ratio *tp / (tp + fp)* where *tp* is the number of true positives and *fp* the number of false positives. The precision is intuitively the classifier's ability of not labeling as positive a sample that is negative. This score reaches its best value at 1 and worst score at 0.

[b]The recall is the ratio *tp / (tp + fn)* where *tp* is the number of true positives and *fn* the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. This score reaches its best value at 1 and its worst score at 0.

[c]The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and its worst score at 0. The relative contribution of precision and recall to the F1 score is equal. This score reaches its best value at 1 and its worst score at 0.

[d]The support is the number of occurrences of each class in each predicted label.

On the other hand, the crosstab that expresses the number of good and bad predictions for the predictive model can be found in Table II.

TABLE II. Crosstab for the Predictive Model Built

	False (predictions)	True (predictions)
False (actual)	125	941
True (actual)	46	2193

Following the process, the important factors found for the model are presented in the Table III. Despite the random forest provide a importance score for all the variables involved in the predictive model, researchers stated 0.05 as the minimum value to consider the factor as relevant. This is because 0.05 is a common value to ensure reliable results in several analytical processes.

TABLE III. Most Relevant Factors for the Predictive Model Built

Name of the variable	Explanation	Importance score
universidad_id	The university where studied the graduate	0.437347
otrosCriterios SeleccionarPuesto Trabajo_circunstancias Personal	The graduate's opinion about the relevance of choose a job depending on the conditions related to personal context: family conciliation, etc.	0.150274
sexo_id	Graduate's gender	0.107880
residenciaExtranjero Motivos_cod	Graduate's reasons to live abroad during the degree	0.097754
otrosCriterios SeleccionarPuesto Trabajo_prestigio	The graduate's opinion about the relevance of choose a job depending on the conditions related to the prestige of the employer, the tasks to be done or to the position within the company.	0.075377
titulacion_id	The degree studied by the graduate.	0.054354

The importance score varies between 0–1, where 1 is the best score and 0 the worst one

As previously presented, after obtaining the predictive model, researchers can use the most relevant factors to analyze in deep what are the specific situations (values of factors) that lead students to get or not a job. For example, using the most relevant factors, could be generated clusters that group graduates using their similar characteristics. As an example, Fig. 1 presents the different clusters obtained after applying a hierarchical clustering algorithm to the clean dataset using the factors as variables to group the users. The representation is truncated to show the more related clusters jointly (showing only 12), but in fact, applying the hierarchical clustering were obtained 55 different groups of students.

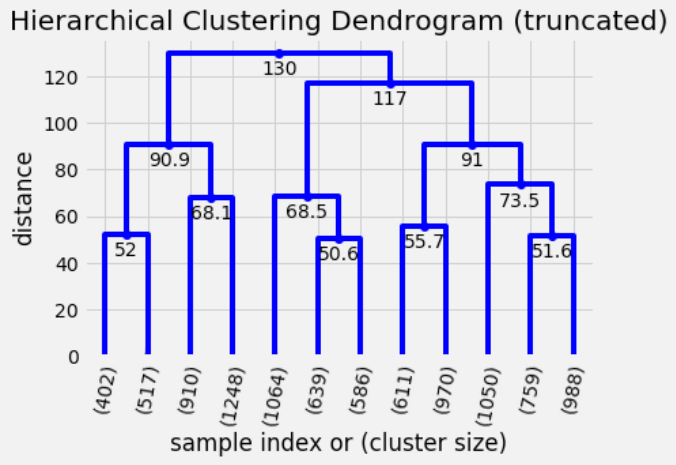

Fig. 1. Dendrogram that represents the clusters of graduates regarding to the most relevant factors detected in the random forest. Each leaf represents a different cluster obtained. The different values that appear near the claves display the Euclidean distance that explains the separation between the different clusters. Finally, the numbers below the leaves (at the bottom of the figure) present the number of users included in the corresponding cluster.

For example, if researchers choose one of the 55 resulting clusters, could observe that (cluster 22):

"229 students from 14 different universities, all of them women who studied one of 20 selected degrees, who do not consider the prestige of employer as a key factor to select a job, who lived abroad during the degree mainly because they work abroad, and do not consider the conditions related to personal context as a key factor to select a job have a chance of get job of 86,90%."

As previously explained, the full information about clusters and results obtained after applying the process is not fully provided in the paper or in an external notebook because of the Observatory's legal and privacy restrictions. Also, this kind of information is out of the scope of this paper, since it is focused on explaining how machine learning methods could be applied to this problem.

V. Discussion

As outlined in the background, the research on employability and employment is a knowledge area in development. The main projects developed in the previous years have pursued to define the main factors that define the employability and employment. The research methods used previously were related to basic statistics and simple analysis, despite some independent researchers went deeper in the methods, applying other related to econometrics, psychometry and other quantitative and qualitative methods of social research. One of the problems observed by the researchers when working with this kind of projects (i.e. in the case of the OEEU), is how to manage to handle large amounts of data and make more specific analysis. The approach presented in this paper regarding the application of machine learning methods, allows to automate some part of the analysis while it allows to gain general knowledge and deal with the problem of analyze how people gets employed from a broader perspective, understanding all the data as a whole.

The machine learning approach has been applied successfully previously by the authors and other researchers in fields like Human-Computer Interaction [41], education [42], etc. For that reason, was tried this approach in a complex research like employment and employability.

Considering the case study presented, the results are quite promising. Despite of the complexity of the data, and the different issues with the dataset, researchers have been able to get a predictive model with a 0.71 precision score (0 the worst score, 1 the best one). This result opens the possibility of keep working in this approach to enhance the results and try deeper analysis. Regarding other scores achieved like, F1, recall, etc., should be outlined that the model built performs poorly in detecting the real "False" values (not employed students), so it could lead to bad predictions regarding students without employment.

In this case, researchers are confident in that managing better the empty values and applying other kind of strategies in data cleaning and wrangling will allow to get better predictive models and outcomes. Following with the results and the case study, it is worth noting that the predictive model involved a support of 3305 observations (graduates), which is a high number in the case of a test like this. It is because researchers want to try a real case to validate the seminal idea of applying machine learning. Despite the model and procedures should be validated in a better way and tested more, the results achieved in a real case like this are considerable.

About the predictive model built, it is not surprising that the some of the most important factors that define if a student gets employment or not after graduating are those related to the university or the degree. Currently there are severe gaps regarding the employment ratio depending on the studies and the students' knowledge area. Also, the predictive model highlights other factor sadly known nowadays: the gender. There are many international studies that deals with the fact that the gender (specially for women) is a handicap to apply for some jobs and positions. The predictive model built found also this variable as a fundamental factor that define how graduates achieve an employment. On the other hand, the predictive model also presents other factors that are not as well-known as the previous ones: the reasons to live abroad during the studies, the importance of employer prestige to choose a job, or the facilities provided by the employer to adjust and balance personal context and life with the work. In general, it is clear that these factors are truly relevant in the model. These 5 factors sum a score of more than 0.8 out of the maximum 1 score that could be achieved by all the 383 factors included in the predictive model. Also, it is possible to think that the least relevant factor of these 5 (the degree obtained by the student) has a very low score (0. 054354), but this score is the weight of a solely factor in a model with 383 different factors, so achieve a 0.05 score out of 1 maximum score it is not low with this amount of different variables observed.

This kind of algorithmic approaches that include all the factors as part of possible models, shed light over some aspects avoiding previous bias and prejudices. In contrast to the background, where the international research community define the factors to study in each project, in this case, authors propose to use all the factors letting the machine learning algorithms to select on their own those truly relevant factors. This switch the *traditional* approach, making the exploratory process to depend only on the dataset available and adapting the focus to the facts and metrics obtained previously.

Also, are provided within the case study some examples on how to obtain deeper information and insights about the concrete metrics that

affect the employment. This is, considering the most relevant factors, how to get common characteristics between those students who get an employment and those who do not. There are other many approaches to make this process possible, but the proposed using clustering analysis is evident and easy to understand and apply.

Regarding the implications of this research, authors agree to point out that this kind of approaches could open new possibilities on using data to enhance the education and students' opportunities in labor market. Following with the idea of Academic Analytics, this kind of employment analytics could be included in academic intelligence processes in universities and other higher education institutions. Also, policymakers could follow this kind of data-driven employability and employment analytics to design and propose new ways of preparing students to their professional future.

VI. Conclusion and Future Work

This paper presents a novel study in the field of employability and employment analytics. The main results achieved have been quite promising and encourage authors to continue the labor of improving the generation of predictive models for employability and employment. The nature of this kind of problems is extremely complex and varies on the time, but with this kind of algorithmic and automated processes could address it better than the traditional approaches. Based on the results, the authors are committed to continue developing the approach to get better results and improve the process until it could be applied successfully in further research works.

Acknowledgment

The research leading to these results has received funding from "la Caixa" Foundation. The author Juan Cruz-Benito would like to thank the European Social Fund and the *Consejería de Educación* of the *Junta de Castilla y León* (Spain) for funding his predoctoral fellow contract. This research has been cofounded by the University of Salamanca through the program of financial aid for the predoctoral contract of José Carlos Sánchez-Prieto (*Programa III: Ayudas para contratos Predoctorales*) cofounded by Banco Santander. This work has been partially funded by the Spanish Government Ministry of Economy and Competitiveness throughout the DEFINES project (Ref. TIN2016-80172-R).

References

[1] R. W. McQuaid and C. Lindsay, "The concept of employability," *Urban studies,* vol. 42, no. 2, pp. 197-219, 2005.

[2] P. Baepler and C. J. Murdoch, "Academic analytics and data mining in higher education," *International Journal for the Scholarship of Teaching and Learning,* vol. 4, no. 2, p. 17, 2010.

[3] J. Bichsel, *Analytics in higher education: Benefits, barriers, progress, and recommendations*. EDUCAUSE Center for Applied Research, 2012.

[4] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," *EDUCAUSE review,* vol. 42, no. 4, p. 40, 2007.

[5] D. A. Gómez Aguilar, F. J. García-Peñalvo, and R. Therón, "Analítica Visual en eLearning," *El Profesional de la Información,* vol. 23, no. 3, pp. 233-242, 2014.

[6] F. Michavila, M. Martín-González, J. M. Martínez, F. J. García-Peñalvo, and J. Cruz-Benito, "Analyzing the employability and employment factors of graduate students in Spain: The OEEU Information System," presented at the Third International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'15), Porto, Portugal, 2015.

[7] Oficina de Cooperación Universitaria. OCU, "Libro Blanco Inteligencia Institucional en Universidades," ed: Madrid: OCU, 2013.

[8] F. J. García-Peñalvo, "Inteligencia Institucional para la Mejora de los Procesos de Enseñanza-Aprendizaje," *GRIAL Research Group*.

[9] F. Michavila, J. M. Martínez, M. Martín-González, F. J. García-Peñalvo, and J. Cruz-Benito, "Barómetro de Empleabilidad y Empleo de los Universitarios en España, 2015 (Primer informe de resultados)," 2016.

[10] W. Greller and H. Drachsler, "Translating learning into numbers: A generic framework for learning analytics," *Journal of Educational Technology & Society,* vol. 15, no. 3, pp. 42-57, 2012.

[11] A. Vázquez-Ingelmo, J. Cruz-Benito, and F. J. García-Peñalvo, "Improving the OEEU's data-driven technological ecosystem's interoperability with GraphQL," presented at the Fifth International Conference Technological Ecosystems for Enhancing Multiculturality 2017 (TEEM'17), Cádiz, Spain, October 18-20, 2017, 2017.

[12] A. Vázquez-Ingelmo, J. Cruz-Benito, F. J. García-Peñalvo, and M. Martín-González, "Scaffolding the OEEU's Data-Driven Ecosystem to Analyze the Employability of Spanish Graduates," in *Global Implications of Emerging Technology Trends*, F. J. García-Peñalvo, Ed. Hershey, PA: IGI Global, 2018, pp. 236-255.

[13] A. García-Holgado, J. Cruz-Benito, and F. J. García-Peñalvo, "Analysis of Knowledge Management Experiences in Spanish Public Administration," presented at the Third International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'15), Porto, Portugal, 7-9, October, 2015.

[14] J. Alcolea Picazo and S. Pavón de Paula, "Los datos como recurso estratégico Libro Blanco Inteligencia Institucional en Universidades (pp. 17-43)," *Madrid, Spain: OCU (Oficina de Cooperación Universitaria),* 2013.

[15] M. Zeleny, "Management support systems: Towards integrated knowledge," *Human systems management,* vol. 7, pp. 59-70, 1987.

[16] D. S. Rychen and L. H. Salganik, *Key competencies for a successful life and well-functioning society*. Hogrefe Publishing, 2003.

[17] National Agency for Quality Assessment and Accreditation of Spain, *Titulados universitarios y mercado laboral, Proyecto REFLEX*. Madrid: ANECA, 2008.

[18] R. G. Biesma, M. Pavlova, G. Van Merode, and W. Groot, "Using conjoint analysis to estimate employers preferences for key competencies of master level Dutch graduates entering the public health field," *Economics of Education Review,* vol. 26, no. 3, pp. 375-386, 2007.

[19] A. García-Aracil and R. Van der Velden, "Competencies for young European higher education graduates: labor market mismatches and their payoffs," *Higher Education,* vol. 55, no. 2, pp. 219-239, 2008.

[20] H. Heijke, C. Meng, and G. Ramaekers, "An investigation into the role of human capital competences and their pay-off," *International Journal of Manpower,* vol. 24, no. 7, pp. 750-773, 2003.

[21] E. Kelly, P. J. O'Connell, and E. Smyth, "The economic returns to field of study and competencies among higher education graduates in Ireland," *Economics of Education Review,* vol. 29, no. 4, pp. 650-657, 2010.

[22] M. J. Freire Seoane, M. Teijeiro Álvarez, and C. Pais Montes, "La adecuación entre las competencias adquiridas por los graduados y las requeridas por los empresarios," 2013.

[23] P. Kellermann, "Acquired and Required Competencies Of Graduates," in *Careers of university graduates: Views and experiences in comparative perspectives*, vol. 17, U. Teichler, Ed. Dordrecht: Springer Science & Business Media, 2007, pp. 115-131.

[24] P. Kellermann and G. Sagmeister, "Higher education and graduate employment in Austria," *European Journal of education,* vol. 35, no. 2, pp. 157-164, 2000.

[25] H. Schomburg and U. Teichler, *Higher education and graduate employment in Europe: results from graduates surveys from twelve countries*. Springer Science & Business Media, 2007.

[26] U. Teichler, "Research on the relationships between higher education and the world of work: Past achievements, problems and new challenges," *Higher Education,* vol. 38, no. 2, pp. 169-190, 1999.

[27] U. Teichler, "Graduate employment and work in selected European countries," *European Journal of Education,* vol. 35, no. 2, pp. 141-156, 2000.

[28] U. Teichler, "New perspectives of the relationships between higher education and employment," *Tertiary Education & Management,* vol. 6, no. 2, pp. 79-92, 2000.

[29] W. McKinney, *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[30] W. McKinney, "Data structures for statistical computing in python," in

Proceedings of the 9th Python in Science Conference, 2010, vol. 445, pp. 51-56: SciPy Austin, TX.

[31] H. Wickham, "Tidy data," *Journal of Statistical Software,* vol. 59, no. 10, pp. 1-23, 2014.

[32] J. Cruz-Benito. (2017). *Jupyter notebook developed to support the research presented in the paper "Proposing a machine learning approach to analyze and predict employment and its factors"* Available: https://github.com/juan-cb/paper-ieeeAccess-2017

[33] S. Raschka, *Python machine learning*. Packt Publishing Ltd, 2015.

[34] L. Breiman, "Random Forests," *Machine Learning,* journal article vol. 45, no. 1, pp. 5-32, October 01 2001.

[35] W. McKinney. (2017). *Pandas, Python Data Analysis Library. 2017*. Available: http://pandas.pydata.org/

[36] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* vol. 12, no. Oct, pp. 2825-2830, 2011.

[37] M. Ragan-Kelley *et al.*, "The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication," in *AGU Fall Meeting Abstracts*, 2014.

[38] F. Perez and B. E. Granger, "Project Jupyter: Computational narratives as the engine of collaborative data science," Technical Report. Technical report, Project Jupyter2015.

[39] T. Kluyver *et al.*, "Jupyter Notebooks-a publishing format for reproducible computational workflows," in *ELPUB*, 2016, pp. 87-90.

[40] Scikit-learn. (2017, 4/09/2017). *API Reference - scikit-learn documentation: Metrics*. Available: http://scikit-learn.org/stable/modules/classes.html - module-sklearn.metrics

[41] J. Cruz-Benito, A. Vázquez-Ingelmo, J. C. Sánchez-Prieto, R. Therón, F. J. García-Peñalvo, and M. Martín-González, "Enabling Adaptability in Web Forms Based on User Characteristics Detection Through A/B Testing and Machine Learning," *IEEE Access,* vol. 5, 2017.

[42] A. Zollanvari, R. C. Kizilirmak, Y. H. Kho, and D. Hernández-Torrano, "Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors," *IEEE Access,* 2017.

Francisco J. García-Peñalvo

Francisco J. García-Peñalvo received the degrees in computing from the University of Salamanca and the University of Valladolid, and a Ph.D. from the University of Salamanca. He is currently the Head of the Research Group Interaction and eLearning. He has led and participated in over 50 research and innovation projects. He was the Vice Chancellor of Innovation with the University of Salamanca from 2007 to 2009. He has authored over 300 articles in international journals and conferences. His main research interests focus on eLearning, computers and education, adaptive systems, Web engineering, semantic Web, and software reuse. He was a Guest Editor of several special issues of international journals, such as Online Information Review, Computers in Human Behavior, and Interactive Learning Environments. He is also a member of the Program Committee of several international conferences and a Reviewer of several international journals. He is currently the Editor-in- Chief of the International Journal of Information Technology Research and the Education in the Knowledge Society Journal. He is also the Coordinator of the multidisciplinary Ph.D. Programme on Education in the Knowledge Society.

Juan Cruz-Benito

Juan Cruz-Benito received an M.Sc. degree in intelligent systems from the University of Salamanca, Spain, in 2013, where he is currently pursuing a Ph.D. in computer sciences. He is one of the youngest members of the Research Group Interaction and eLearning, where he specializes in software solutions based on technology ecosystems and open source software. He works in Human-Computer Interaction, educational virtual worlds and technologies for educational purposes, disciplines that he has developed in many innovation and research projects. He has participated in many European and national R&D projects, such as TRAILER, VALS, USALSIM Virtual Campus, and the Spanish Observatory for University Employability and Employment (OEEU), where he participated as a software engineer, researcher, and developer.

Martín Martín-González

Martín Martín-González is a researcher at the UNESCO Chair in University Management and Policy of the Technical University of Madrid (UPM). Prior to his work at the UPM he worked as a researcher in the Faculty of Economics in the Autonomous University of Madrid (UAM). He holds a PhD in Economics and a master's degree in Economic Development and Public Policy from the UAM. He majored in economics at the University of La Laguna (ULL). His areas of research are the Economics of Education, the Economics of Higher Education, Employability, the Evaluation of Educational Policies, Public Economics, Applied Economics, Economic Development, Higher Education and Vocational Training.

Andrea Vázquez-Ingelmo

Andrea Vázquez-Ingelmo was born in Salamanca, Castilla y León, Spain in 1994. She received a bachelor's degree in Computer Engineering from the University of Salamanca, Salamanca, in 2016, and she is currently completing her master's degree in computer engineering from the same university. She is a member of the Research Group of Interaction and eLearning (GRIAL). Since 2016, she has been part of the national project "Spanish Observatory for University Employability and Employment" as a developer and researcher.

José Carlos Sánchez-Prieto

José Carlos Sánchez-Prieto received a bachelor's degree in Pedagogy and a master's degree in ICT applied in education from the University of Salamanca (Spain). Currently he is a Researcher in Training at the Faculty of Education of said university, where he conducts his PhD research within the Programme on Education in the Knowledge Society. His area of research is the assessment of attitudes among in-service and pre-service teachers.

Roberto Therón

Roberto Therón studied Computer Science at the University of Salamanca (Diploma) and the University of La Coruña (BA). After joining the Research Group Robotics at the University of Salamanca, he presented his thesis work, "Parallel calculation configuration space for redundant robots," receiving the Extraordinary Doctoral Award. He subsequently obtained a Bachelor in Communication Studies (University of Salamanca) and a Bachelor in Humanities (University of Salamanca). At the same university, he continues to carry out his research work as the manager of the VisUsal group (within the Recognized Research Group GRIAL), which focuses on the combination of approaches from computer science, statistics, Graphic Design and Information Visualization to obtain an adequate understanding of complex data sets. In recent years, he has been dedicated to developing advanced visualization tools for multidimensional data, such as genetics or paleo-climate data. In the field of Visual Analytics, he develops productive collaborations with groups and institutions internationally recognized as the Laboratory of Climate Sciences and the Environment (France) or the Austrian Academy of Sciences (Austria). He is the author of over 100 articles in international journals and conferences.