Received October 23, 2017, accepted November 29, 2017, date of publication December 27, 2017, date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2782678

Enabling Adaptability in Web Forms Based on User Characteristics Detection Through A/B Testing and Machine Learning

JUAN CRUZ-BENITO[1], (Member, IEEE), ANDREA VÁZQUEZ-INGELMO[1], JOSÉ CARLOS SÁNCHEZ-PRIETO[2], ROBERTO THERÓN[3], FRANCISCO JOSÉ GARCÍA-PEÑALVO[1], (Member, IEEE), AND MARTÍN MARTÍN-GONZÁLEZ[4]

[1]GRIAL Research Group, Computer Science Department, Research Institute for Educational Sciences, University of Salamanca, 37008 Salamanca, Spain
[2]GRIAL Research Group, Research Institute for Educational Sciences, University of Salamanca, 37008 Salamanca, Spain
[3]GRIAL Research Group, Computer Science Department, and VisUSAL Research Group, University of Salamanca, 37008 Salamanca, Spain
[4]UNESCO Chair in University Management and Policy, Technical University of Madrid, 28003 Madrid, Spain

Corresponding author: Juan Cruz-Benito (juancb@usal.es)

This work was supported in part by the la Caixa Foundation, in part by the Spanish Government Ministry of Economy and Competitiveness throughout the DEFINES Project under Project TIN2016-80172-R, and in part 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. The work of J. Cruz-Benito was supported by the European Social Fund and the Consejería de Educación of the Junta de Castilla y León, Spain.

ABSTRACT This paper presents an original study with the aim of improving users' performance in completing large questionnaires through adaptability in web forms. Such adaptability is based on the application of machine-learning procedures and an A/B testing approach. To detect the user preferences, behavior, and the optimal version of the forms for all kinds of users, researchers built predictive models using machine-learning algorithms (trained with data from more than 3000 users who participated previously in the questionnaires), extracting the most relevant factors that describe the models, and clustering the users based on their similar characteristics and these factors. Based on these groups and their performance in the system, the researchers generated heuristic rules between the different versions of the web forms to guide users to the most adequate version (modifying the user interface and user experience) for them. To validate the approach and confirm the improvements, the authors tested these redirection rules on a group of more than 1000 users. The results with this cohort of users were better than those achieved without redirection rules at the initial stage. Besides these promising results, the paper proposes a future study that would enhance the process (or automate it) as well as push its application to other fields.

INDEX TERMS Adaptability, machine learning, user profiles, web forms, clusters, hierarchical clustering, random forest, A/B testing, human-computer interaction, HCI.

I. INTRODUCTION

Understanding what users do within a system is now a fundamental task in the digital world [1]. Most aspects of modern development workflows include users as a centric part of the design and development process of digital products (i.e., user-centered design [2], [3]). Not only knowing what users do (clicks, workflows, interactions, etc.) within a system is valuable for software developers and designers, but these stakeholders must also pay attention to other related-aspects, like user experience, satisfaction, and trust [4]–[8]. Understanding what users do or feel when they use a system is extremely valuable to validate and improve a system. Analyzing users' interactions or their opinion about what they use makes it possible to ascertain the system's strengths or weaknesses regarding users' experience (mostly user interfaces and parts alike) to improve the system based on evidence.

Besides using the analysis of users' interactions and opinions to improve the worst-perceived parts of a system, developers can use these data to build custom or adaptive solutions for different kinds of users [9]–[11]. Using this idea, software engineers could develop versions of the system in which different version are showed to each kind of user. By knowing user profiles and identifying users' behavior and desires, the system could adapt its components to better match users' expectations and likings, and (probably) boost user performance and satisfaction [8], [9], [12].

For a better understanding of the current paper, the context for this experiment is presented. The research has been conducted using a system that belongs to the Spanish Observatory for University Employability and Employment (OEEU in its Spanish acronym) [13]. 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. To accomplish this mission, the observatory has developed a complex information system [14], [15] that collects and analyzes data to present the insights to the researchers. The system is implemented using the Python language through the Django framework [16] and many other software libraries; it also keeps the information in a MariaDB relational database. To gather data from Spanish universities and students, the OEEU information system has two main tools: one tool is devoted to obtaining students' raw data provided by the university; the other one is a system that generates custom web forms and questionnaires that are to be completed by the graduates after they leave the university. The problem of these web forms is their length, as they typically include between 30 and 70 questions. This second tool for gathering data (the questionnaires) is a centric part in this research.

The goal of this paper is to present a new approach for enabling adaptability in web-based systems using A/B testing methods and user-tracking and machine-learning algorithms that could lead to improving user performance in completing a (large) web form, validating the obtained results through statistical tests. As a secondary goal, the research presented in this paper also aims to produce all machine learning processes in a white-box way, using algorithms and techniques that allow researchers to understand what is happening in every moment. Moreover, to allow readers and other researchers to follow or reproduce the entire process, this paper provide all the code used in the analysis process in Jupyter notebooks available publicly in Github.

The paper has the following structure: section two (Materials and Methods) explains the different algorithms, data, and research framework. Section three (Results) presents the outcomes obtained in the different steps involved in the research: the results regarding the predictive models that provide the most important users' characteristics on completing the web form, those regarding users' profiles found, and those regarding the guidance of users over the different versions of the system to enhance their performance. The fourth section (Discussion) presents different authors' thoughts, proposals, and considerations about this research and its implications, as well as some future works and general conclusions.

II. MATERIALS AND METHODS

This section outlines the materials and methods used for this research. In the case of materials, the data used and the analysis software are described. In the case of the methods, the different steps needed to apply the machine-learning approach to the analysis process as well as the statistics used to prove the validity and significance of the results are presented.

A. MATERIALS

This subsection presents the different materials involved in this research. The materials can be categorized into two main groups: materials related to the experimentation framework and the software tools used to make the proper analysis and support the research process.

The questionnaires and custom web forms included in the OEEU information system gather data from students in two ways: information provided explicitly by the students (the information provided directly) and *paradata* [17]. The paradata from these questionnaires are the auxiliary data that describe the filling process, such as response times, clicks, scrolls, and information about the device used when using the system. All the data used in this research are taken from these two available sources: the raw input tool used by universities and the web forms tool (providing user inputs and their paradata).

Regarding the data used in this research, it is worth noting that to generate the predictive models needed to characterize the main factors that affect users in completing the questionnaires, the authors have chosen only those available before the users began the questionnaire. This is because the research is focused on investigating which factors predetermine participants' success or failure in completing the form, considering all the factors related only to personal context and device and software used to access the web forms. The data about the personal context of the user are provided by the OEEU's system and include information submitted by the university where the user (graduated) studied. All the information that could be used to create the models that predict whether the user will complete the questionnaire (before starting it) is presented in Table 1. Table 1 also explains the data variables used and whether they were valuable for the models. This research has been carried out with a total of 7349 users (all who have some type of experience with the web forms). Of them, the data from 5768 users were considered initially. Finally, data from 3456 users (those resultant after cleaning the data) were used to train and try the machine-learning algorithms (as will be explained in the following section); 1165 users were the cohort introduced in a phase of reinforcement for the questionnaires that validated the rules generated to adapt the web form to users. This number (1165) includes users who did not complete the web form in the first stage as well as users that joined the experiment during the reinforcement and validation phase. Other users (416) only viewed the web forms without starting them. For that reason, were not considered in the experimental report.

TABLE 1. Initial variables gathered from the OEEU information system to build the predictive models of questionnaires' completion.

Name of the variable in the code	Explanation	Type of information that it provides	Was this variable used finally to build the predictive models?
estudiante_id	ID number of student	Personal information	Yes
annoNacimiento	Year of birth	Personal information	No
sexo_id	Gender (male / female)	Personal information	Yes
esEspannol	Is the student Spanish?	Personal information	No
universidad_id	ID of the university where the graduate studied	Personal information	Yes
estudiosPadre_id	Maximum educational level achieved by the graduate's father	Personal information	No
estudiosMadre_id	Maximum educational level achieved by the graduate's mother	Personal information	No
situacionLaboralPadre_id	Current employment status of the graduate's father	Personal information	No
situacionLaboralMadre_id	Current employment status of the graduate's mother	Personal information	No
oficioProfesionPadre_id	Occupation of the graduate's father	Personal information	No
oficioProfesionMadre_id	Occupation of the graduate's mother	Personal information	No
residenciaFamiliar_id	Place of residence of the graduate's family	Personal information	No
residencia_id	Place of residence of the graduate during studies	Personal information	No
idMaster_id	ID number of the master study	Personal information	Yes
especializacionMaster_id	Specialization of the graduate's master	Personal information	No
masterHabilitante	Is an enabling master?	Personal information	No
titularidadMaster_id	Public or not master	Personal information	No
modalidadMaster_id	Modality of the master (online, physical, etc.)	Personal information	No
cursoInicioMaster	Season of the beginning of the master	Personal information	No
cursoFinalizacionMaster	Season of the completion of the master	Personal information	Yes
notaMedia_id	Average grade of the student	Personal information	No
realizacionPracticasMaster	Did the student professionally practice during the master?	Personal information	No
tiempoDuracionPracticasMaster	Time spent by the student in professional practices during the master	Personal information	No
realizacionErasmusMaster	Did the student do an Erasmus stay?	Personal information	No
tiempoDuracionErasmus_id	Time spent by the student in an Erasmus stay	Personal information	No
paisErasmusMaster_id	Country where the student did an Erasmus stay	Personal information	No
viaAccesoMaster_id	Way of accessing the master	Personal information	No
verticalAsignado	Vertical assigned in the A/B testing for the student	Experiment configuration	Yes
cuestionarioFinalizado	Did the student finalize the questionnaire?	Experiment configuration	Yes
numUniversidades	Number of universities involved in the master	Personal information	Yes
numUniversidadesEspannolas	Number of Spanish universities involved in the master	Personal information	Yes
ramaConocimiento_id	Knowledge branch of the master (healthcare, social sciences, engineering, etc.)	Personal information	Yes
realDecreto	Official statement approving of the master studies program	Personal information	Yes
browser_language	Language of the browser used	Device information	Yes
browser_name	Name of the browser used	Device information	Yes
browser_version	Version of the browser used	Device information	Yes
device_pixel_ratio	Device pixel ratio of the browser	Device information	Yes
device_screen_height	Device screen height	Device information	Yes
device_screen_width	Device screen width	Device information	Yes
landscape	Is the device in landscape mode?	Device information	No
os	Operative system of the device	Device information	Yes
os_version	Version of the operative system used	Device information	Yes
portrait	Is the device in portrait mode?	Device information	No
push_notification	Did accept the graduate push notifications for the web form?	Device information	No
push_notification_id	ID number for the push notification subscription	Device information	No
tablet_or_mobile	Is the device tablet or mobile?	Device information	Yes
userAgent	User agent of the device used	Device information	Yes
viewport_height	Height of the window browser	Device information	Yes
viewport_width	Width of the window browser	Device information	Yes

The variables excluded to build the predictive models are those that have more than 10% of their observations with the null value.

The programming language used to conduct all the analyses and calculations was Python. The concrete Python software tools and libraries used to code and execute the different algorithms and statistics were:

- Pandas software library [18]–[20], to manage data structures and support analysis tasks.
- Scikit-learn [21] library, to accomplish the machine learning workflow [22].
- Jupyter notebooks [23]–[25], to develop the Python code used in this research.

All the code developed to analyze user interactions and create machine-learning models, etc. is available at https://github.com/juan-cb/paper-ieeeAccess-2017 [26].

B. METHODS

As found in the bibliography, the concept of A/B testing (also known as bucket testing, controlled experiment, etc.) applied to websites and the Internet could be explained as follows: "show different variations of your website to different people and measure which variation is the most effective at turning them into customers (or people that complete successfully a task in the website, like in this experiment). If each visitor to your website is randomly shown one of these variations and you do this over the same period, then you have created a controlled experiment known as an A/B test" [27]–[29]. In this case, the authors have prepared three different variations, called verticals A, B, and C. In each variation, the authors introduced several changes related to enhancing the users' trustiness, engagement, make the user interface more conversational, etc. All these changes, introduced in the different variations of the web forms (the verticals) used in this research, were proposed by the authors in previous works [30]. These verticals are used as the website variations in which users (students responding to the questionnaires) are meant to test which version is the best regarding the users' performance in the initial stage. To do so, before the experiment, 5768 users were redirected randomly to the different vertical. In the last part of the experiment, the verticals were used to check whether the rules and users' analysis performed during the machine-learning analytics process improve the users' performance in completing the web forms. In this validating phase (which also will be called reinforcement in this paper), 1165 users were redirected to the verticals using the rules generated analyzing the interaction data from the users that acceded randomly to the verticals.

In general, the performed analysis (based on statistics and machine learning) follows common principles in data science regarding data structuration, tidy data approaches, etc. [18], [20], [31]. As stated in the introduction, 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. This is extremely important, from the authors' point of view, in a research project like this, as it allows humans to provide feedback to the algorithmic process.

Moreover, these main principles, the different details for the analysis pipeline, and methods used in this research are presented.

To find the best models and most accurate parameters, researchers have tried the following approaches:

1. Create predictive models using all the data together. In this approach, researchers tried to use different groups of variables to create the model: all the variables collected from the users, using derived variables (like whether the browser or operative system used to access were modern), etc.
2. Create predictive models using the verticals gap. In this case, researchers generated a predictive model per each vertical of the A/B test. In this case, the most relevant configuration regarding the variables to build up the model in the previous step is included.

Using the most accurate models, the researchers applied all the stages that will be described below (as well as the details for building the predictive models) to generate the different clusters and obtain the rules used to redirect users within the system.

The workflow established (available at https://github.com/juan-cb/paper-ieeeAccess-2017 [26]) is as follows:

1. Retrieve datasets about users from OEEU's information system.
2. Filter the desired fields from the datasets and merge datasets in a single 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 users who have only partial information and not all presented in Table 1.
4. Normalize data with the One-hot encoding algorithm for categorical values in columns [22]. To apply the One-hot encoding, researchers used the get_dummies() function from Pandas library, as presented in [26].
5. Considering the data gathered and the kind of variable (labeled) to predict, 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 [32] was selected. In this step, the Random Forest algorithm was executed repeatedly, using a custom method based [26] on *GridSearch* functions from Scikit-learn, to determine the best setup for the dataset given (obtaining the most valuable 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. To obtain these features authors applied *feature_importances_* method from the Random Forest classifier implemented in Scikit-learn library [26].
8. Using the most important features (those that have an importance higher than a custom threshold value of 0.05—the importance score varies between 0 and 1, where 0 is the worst score and 1 the best one), generate clusters applying hierarchical clustering [33]. The reason to use hierarchical clustering is that the algorithm does not require deciding upon the number of clusters to obtain (so, it does not require also to fix previously Euclidean distances and other parameters); it obtains all possible clusters showing the Euclidean distance between them. These clusters represent the groups of users who have participated in the questionnaire according to the most important factors found in the classification.
9. With these clusters, the researchers investigate which clusters exhibit low performance.
10. Using this knowledge about groups of users with low performance and the heuristics observed, software engineers responsible for the OEEU's information system and its web forms could propose changes and

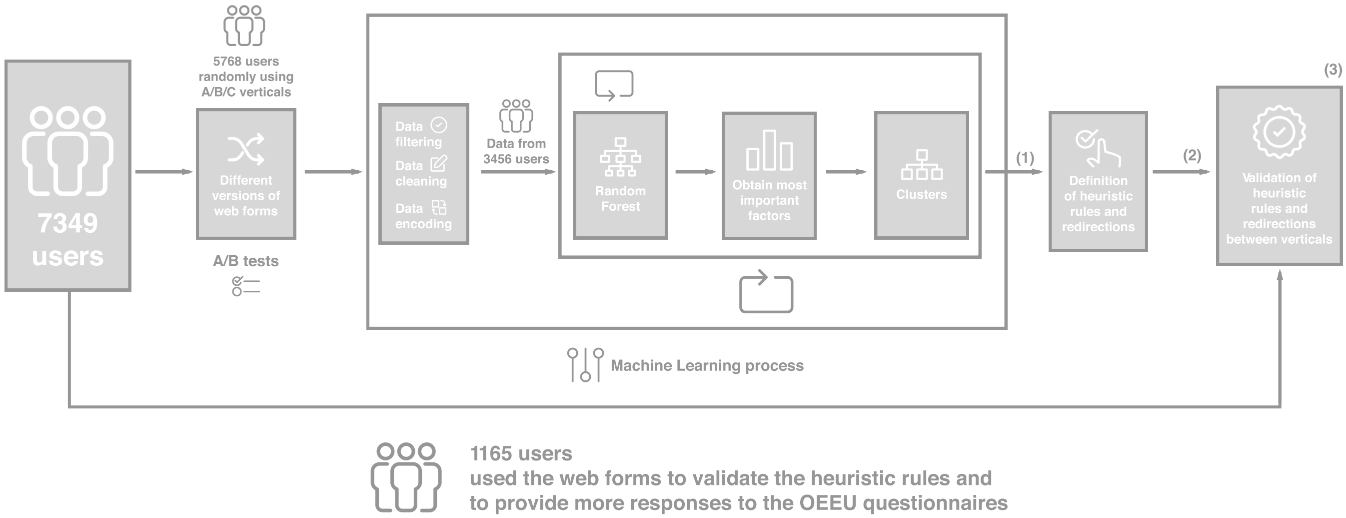

FIGURE 1. **Overview of the process followed. Summary of the materials and methods.**

fixes (rules, redirections, etc.) in the platform that might help users to improve their performance in the future.

11. Once the data-gathering process is finished, the researchers performed a statistical analysis of the finalization rate of the individuals to determine whether the application of the rules had any impact in the improvement of the finalization of the questionnaires. With this purpose, and considering the characteristics of the variables, the authors applied the Chi-squared test given that it is the most convenient alternative for the analysis of the relationship of two nominal variables.

All these steps and a summary of all methods and materials are presented in the Figure 1.

III. RESULTS

This section presents the main results obtained during the research. The outcomes are divided into three subsections: one related to the results obtained during the machine-learning process (best setup, best ways of building predictive models, the predictive models themselves, the most important variables to finalize or the questionnaire, etc.). The second subsection explains the heuristic rules obtained at the end of the machine-learning workflow inferred from the machine-learning results previously explained. These rules were applied to redirect users within the different verticals of the A/B tests. Finally, the results of the redirections are presented, explaining whether they really affected to the users' finalization of the questionnaire.

A. RESULTS REGARDING MACHINE-LEARNING PROCEDURES: PREDICTIVE MODELS AND CLUSTERING

As previously explained, the researchers performed several attempts to find the most accurate predictive models that better explain whether users will finalize the questionnaire. The first attempt was based on using all the data together focusing in primary variables (excluding those that have too many void values); the second one was based on using all variables and derived variables (constructed from primary ones). The third attempt was based on creating separated predictive models depending on the vertical. In this way, the researchers predicted users' behavior regarding the finalization depending on the vertical / interaction features that they experience. In this last approach, the researchers used the best set of variables found previously to build the model.

The results achieved in this phase would correspond to those expected in the (1) mark in Figure 1.

Regarding the first attempt to build the best predictive model, the researchers used all the variables (excluding the cleaned ones applying the rules defined in the methods sections). As presented in https://github.com/juan-cb/paper-ieeeAccess-2017/blob/master/machinelearning-results.ipynb [26], the predictive model generated had an average precision of 0.79 (Table 2 shows the results and explanations of the results metrics) in predicting whether users will finalize the web form before starting it (in fact, this 0.79 is a fairly good

TABLE 2. **Results of the first predictive model built.**

	Precision[a]	Recall[b]	F1-score[c]	Support[d]
False	0.84	0.38	0.52	378
True	0.77	0.97	0.86	815
Avg / total	0.79	0.78	0.75	1193

[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.

precision score). In the case of this research, the authors use the precision score as the main metric to make decisions, as it is focused on penalizing false positives [34].

The crosstab (that expresses the number of good and bad predictions) for this first predictive model can be found in Table 3.

TABLE 3. Crosstab for the first predictive model built.

	False (predictions)	True (predictions)
False (actual)	142	236
True (actual)	27	788

In this first attempt and its 0.79-precision predictive model, the most important factors in the model were (the importance score varies between 0–1, where 1 is the best score and 0 the worst one):

1. *device_screen_width*: 0.297189
2. *viewport_width*: 0.292615
3. *browser_name_Firefox*: 0.100000
4. *device_pixel_ratio*: 0.098356
5. *viewport_height*: 0.096237

In the second attempt, the researchers used the same variables plus two derived variables composed using the primary ones. The derived variables were *modern_browser* and *modern_os*. Those variables were calculated using the versions of operative systems and browsers used by users. In this case, the researchers calculated the median version of the operative system or browser (the midpoint between the oldest version and newest one present) and classified the browser or operative system as modern or not depending on whether its version is equal or superior to the mid version or is lower. These derived variables were prepared because it was impossible to use the literal version of each browser or operative system in the random forest algorithm due their heterogeneous expressions (each browser or OS has its own version's description and format, etc.). In this second attempt, the precision of the predictive model was higher—specifically, a precision of 0.81 (Table 4). The crosstab for this second model is presented in Table 5.

TABLE 4. Results of the second predictive model built.

	Precision	Recall	F1-score	Support
False	0.91	0.34	0.50	378
True	0.76	0.98	0.86	807
Avg / total	0.79	0.78	0.75	1185

In general, this second model performed better than the previous one (at least it was most precise). In this case, the most important factors that define the model were:

1. *tablet_or_mobile*: 0.179032
2. *device_pixel_ratio*: 0.159406
3. *device_screen_height*: 0.097580
4. *device_screen_width*: 0.095784

TABLE 5. Crosstab for the second predictive model built.

	False (predictions)	True (predictions)
False (actual)	129	249
True (actual)	13	794

5. *viewport_height*: 0.089050
6. *os_Android*: 0.063415

Since the variables used to build the predictive model were different from the previous one, it is normal that the factors that define the model could differ.

In the third approach to generate the best predictive model, the researchers generated a predictive model per each vertical in the A/B test applied to the users. In this case, the researchers included all the variables that produced the best predictive model previously: this is, the variables from the second attempt (including the variables *modern_os* and *modern_browser*). In this case, the researchers have trained three different random forest algorithms, found the best setup for each one depending on the data to analyze, and produced a model for each vertical. The results of these predictive models are presented in Tables 6, 7, and 8, and their precision varied between 0.79 and 0.87. The average precision in the three models was of 0.8233, which is higher than the precision achieved in the previous attempts of generating predictive models. Tables 9, 10, and 11 present the crosstabs for each model; they explain how much effective was the prediction depending on the finalization in the web form.

TABLE 6. Results of the predictive model for the vertical A.

	Precision	Recall	F1-score	Support
False	0.92	0.35	0.51	69
True	0.85	0.99	0.92	263
Avg / total	0.87	0.86	0.83	332

TABLE 7. Results of the predictive model for the vertical B.

	Precision	Recall	F1-score	Support
False	0.92	0.37	0.52	161
True	0.74	0.98	0.85	301
Avg / total	0.81	0.77	0.73	462

Regarding the most important factors per each predictive model generated in the third attempt, the results were the following:

Most influential factors for the predictive model for vertical A:

1. *viewport_width:* 0.267931
2. *tablet_or_mobile:* 0.139438
3. *os_iOS:* 0.132425
4. *device_screen_height:* 0.118814
5. *device_screen_width:* 0.067581
6. *device_pixel_ratio:* 0.066577
7. *os_Android:* 0.054088

TABLE 8. Results of the predictive model for the vertical C.

	Precision	Recall	F1-score	Support
False	0.89	0.37	0.52	132
True	0.74	0.97	0.84	238
Avg / total	0.79	0.76	0.73	370

TABLE 9. Crosstab of the predictive model results for vertical A.

	False (predictions)	True (predictions)
False (actual)	24	45
True (actual)	2	261

TABLE 10. Crosstab of the predictive model results for vertical B.

	False (predictions)	True (predictions)
False (actual)	59	102
True (actual)	5	296

TABLE 11. Crosstab of the predictive model results for vertical C.

	False (predictions)	True (predictions)
False (actual)	49	83
True (actual)	6	232

Most influential factors for the predictive model for vertical B:

1. *viewport_height:* 0.294176
2. *viewport_width:* 0.167701
3. *device_screen_height:* 0.102463
4. *device_pixel_ratio:* 0.085122
5. *os_Android:* 0.076196

Most influential factors for the predictive model for vertical C:

1. *device_screen_width:* 0.193903
2. *viewport_height:* 0.143456
3. *device_screen_height:* 0.108721
4. *tablet_or_mobile:* 0.100000
5. *viewport_width:* 0.093584
6. *device_pixel_ratio:* 0.088479
7. *os_Windows:* 0.055153

Analyzing the results, researchers found that the best way, in this case, to obtain the most-precise predictive models for users' interaction, was obtained by splitting the dataset using the vertical criteria. That is, separating the dataset into three datasets, each one including the data from each user cohort that experienced each one of the A/B tests versions. For that reason, the resultant models were selected to generate the clusters and study them to produce the rules to be used in redirecting users among the different visual representations of the web forms. Using these profiles (clusters) and the rules generated, the researchers found what kind of user (and its technological aspects) fits better (is more inclined to finalize) in each version of the questionnaires, forwarding the users using these criteria to each vertical.

After producing the predictive models, the researchers clustered users depending on their finalization ratio and the most important factors extracted in the predictive models. Explaining all clusters generated after producing each predictive model is out of the scope of this paper (but available at https://github.com/juan-cb/paper-ieeeAccess-2017/blob/master/machinelearning-results.ipynb [26]). Thus, only the clusters obtained after finding the best predictive models will be explained (those generated separately per each vertical). As discussed in the methods section, the clusters were generated using hierarchical clustering techniques because these techniques do not require configuring the target number of clusters. This permits all the relevant clusters (relevance due to the Euclidean distance among them) to be obtained regardless of the number. Figures 2, 3, and 4 present the dendrograms corresponding to each set of clusters.

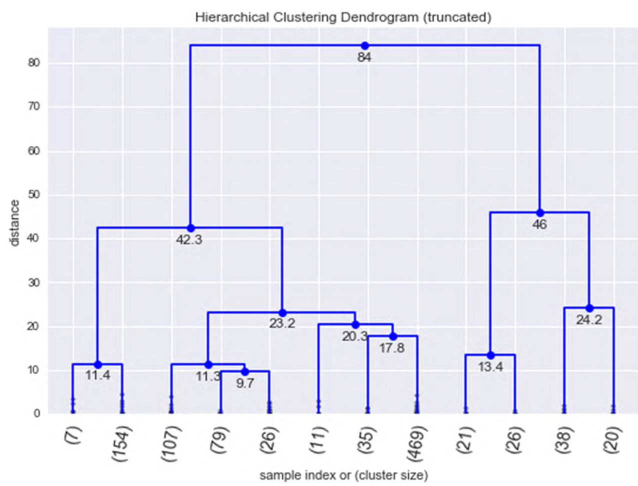

FIGURE 2. Dendrogram representing the clusters found with the predictive model generated using the data from vertical A. Each leaf represents a different cluster obtained (except, in this figure, clusters 8 and 9 that are represented together in the dendrogram due to their closeness in the 9th leaf). 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. Source and full resolution image with all the clusters are available in [26].

After applying the hierarchical clustering algorithm (https://github.com/juan-cb/paper-ieeeAccess-2017/blob/master/machinelearning-results.ipynb [26]) the following numbers of clusters were found: 13 clusters for the vertical A predictive model, 12 clusters for the vertical B model, and 12 clusters for the vertical C.

Analyzing the generated clusters, the researchers found the features that define each cluster and compared them among

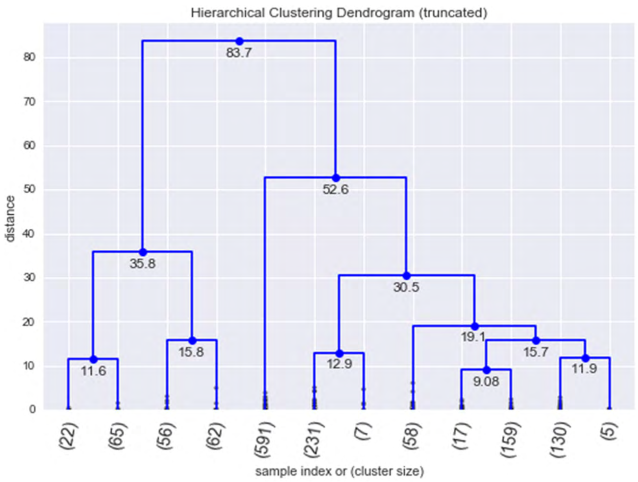

FIGURE 3. **Dendrogram representing the clusters found with the predictive model generated using the data from vertical B. The meaning of the different visual elements is the same than those presented in Fig 2. Source [26].**

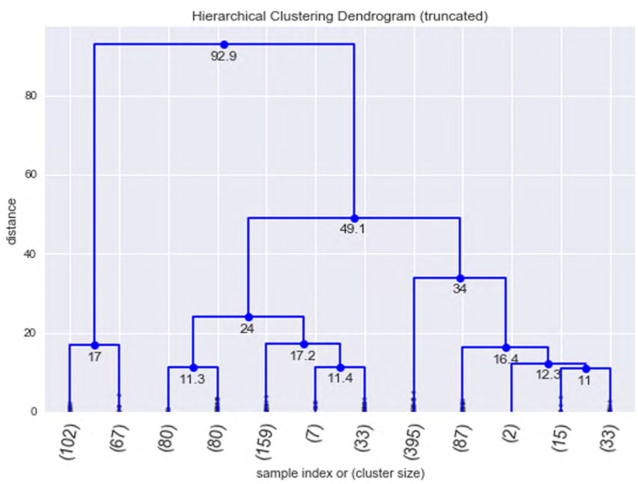

FIGURE 4. **Dendrogram representing the clusters found with the predictive model generated using the data from vertical C. The meaning of the different visual elements is the same as those presented in the previous dendrogram figures. Source [26].**

```
Cluster 8 || feature: os_Windows
count    395.0
mean       1.0
std        0.0
min        1.0
25%        1.0
50%        1.0
75%        1.0
max        1.0
Name: os_Windows, dtype: float64
Mean of feature :os_Windows: 1.0
```

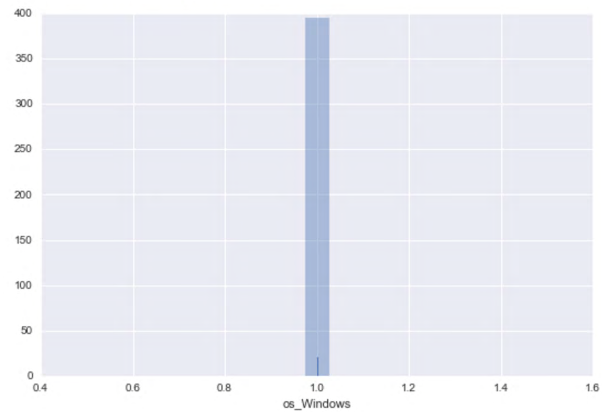

FIGURE 5. **Descriptive statistics and distribution of values for cluster 8 (vertical C), regarding the use of the Windows operating system. Source [26].**

the different models to define the redirection rules. This analysis of clusters and rule generation will be explained in the following subsection.

B. RESULTS REGARDING CRITERIA FOR REDIRECTING USERS WITHIN A/B TESTING VERTICALS

Once the clusters were identified through the produced predictive models, the researchers started to analyze the features of each cluster to establish the proper redirection rules based on the heuristics observed. In the case of this study, these rules were not generated automatically, although using the code and procedures previously presented, it would be possible. The results achieved at this stage correspond to those expected in mark (2) in Figure 1.

First, the most important values of these features were obtained through descriptive statistics and distribution plots (for every identified cluster), as included in [26]. As an example of the features' identification, Figure 5 shows that in vertical C's 8th cluster, the device's operating system of the clustered users is Windows (the most repeated value is 1, i.e., *True*). With this information (and the rest of information obtained through the same process on the rest of features) the researchers could determine the possible devices used by the students in every cluster. In this case, the authors will refer mainly to these factors as *technical features* or *technical info*, as the factors were all related to the technological aspects of the device and software used by users completing the questionnaires.

The descriptive statistics and distribution plots for every technical feature within each cluster are available at https://github.com/juan-cb/paper-ieeeAccess-2017/blob/master/machinelearning-results.ipynb [26].

Once the values (technical specs mainly) of the devices were obtained, the finalization rates of the questionnaires of all clusters were calculated, identifying the performance achieved by users in each of them. This allowed the identification, for example, of the clusters whose finalization rate were smaller than the finalization rate of the whole questionnaire vertical.

In this way, researchers identified the factors (the most relevant features of each vertical's predictive model) linked to the clusters that performed worse than the rest. This information is summarized in Tables 12, 13, and 16 for verticals A, B, and C, respectively.

These tables (12, 13, and 14) helped the researchers to define the redirection rules. For example, Android devices with a 2-pixel ratio (i.e., Android devices with good screen

TABLE 12. Cluster characteristics identification in vertical A. Clusters that performed below the general completion rate of the vertical are marked in red.

Vertical	Vertical completion rate		Total users		Completed questionnaires			Uncompleted questionnaires		
A	76.23% (average)		993		757			236		
Cluster number	**Users count**	**Completion rate**	**Viewport width**	**Tablet or mobile?**	**iOS?**	**Screen height (px)**	**Screen width (px)**	**Pixel ratio**	**Android?**	**Possible device**
1	7	71%	2569	False	False	1440	1560	1 or 2	False	Windows computer
2	154	86.36%	1920	False	False	1080	1920	1	False	Windows computer
3	107	83.17%	1440	False	False	900	1440 or 1600	1	False	Windows computer
4	79	79%	1260	False	False	1024	1280	1	False	Windows computer
5	26	80%	1250	False	False	1080	1800	1	False	Windows computer
6	11	81.81%	896 or 1280	False	False	800 or 1024	896 or 1280	1	True	Convertible device
7	35	82.85%	1290	False	False	800 or 900	1280 or 1440	2	False	Retina Mac computer
8	35	80%	1024	False	False	768	1024	1	False	Windows computer
9	434	82.02%	1366	False	False	768	1366	1	False	Windows computer
10	21	9%	366	True	False	640	360	3 or 4	True	Android mobile (very high resolution)
11	26	15%	360	True	False	600 or 700	360	2	True	Android mobile (good resolution)
12	38	2%	500–400	True	True	600	375	2 or 3	False	iPhone
13	20	84.99%	768 or 1024	False	True	1024	768	1 or 2	False	iPad

TABLE 13. Cluster characteristics identification in vertical B. Clusters that performed below the general completion rate of the vertical are marked in red.

Vertical	Vertical completion rate		Total users		Completed questionnaires		Uncompleted questionnaires	
B	66.5% (average)		1403		933		470	
Cluster number	**Users count**	**Completion rate**	**Viewport height (px)**	**Viewport width (px)**	**Screen height (px)**	**Pixel ratio**	**Pixel ratio**	**Possible device**
1	22	72.72%	628	414	736	3	False	Large iPhone (iPhone 6 Plus, 6s Plus or iPhone 7 Plus)
2	65	1.5%	450–500 or 550–600	320 or 375	480, 568 or 667	2	False	iPhone
3	56	8.9%	550	360	640	2	True	Android mobile (good resolution)
4	62	11.29%	537	360	640	3-4	True	Android mobile (very high resolution)
5	591	75.8%	649	1366	768	1	False	Windows computer
6	231	73.5%	955	1860	1080	1	False	Windows computer
7	7	71.42%	1290	1960	1440	1	False	Non-retina Mac computer
8	58	77.58%	720	1134	800-900 or 1024	2	False	iPad
9	159	76.72%	620	1260	1024, 1080 or 1200	1	False	Windows computer
10	17	76.47%	780	1440 or 1600	900	1	False	Windows computer
11	130	74.61%	894	1260	1024	1	False	Windows computer
12	5	80%	936 or 1144	768-800	1024 or 1080	1	True	Android tablet

resolution), despite their low rate performance, obtain better finalization ratios in vertical A (finalization rate of 15%) than in verticals B and C (finalization rates of 8.9% and 10.7%, respectively), leading to the conclusion that the users with devices that meet these characteristics should be redirected to vertical A.

Repeating this methodology for every device identified, the following rules were obtained:

TABLE 14. Cluster characteristics identification in vertical C. Clusters that performed below the general completion rate of the vertical are marked in red.

Vertical	Vertical completion rate		Total users		Completed questionnaires			Uncompleted questionnaires		
C	65% (average)		1060		689			371		
Cluster number	**Users count**	**Completion rate**	**Screen width (px)**	**Viewport height (px)**	**Screen height (px)**	**Tablet or mobile?**	**Viewport width (px)**	**Pixel ratio**	**Windows?**	**Possible device**
1	102	10.7%	360	550	640	True	360	2	False	Android mobile (good resolution)
2	67	20.84%	360	570	640	True	370	3	False	Android mobile (very high resolution)
3	80	71.25%	1280	895	1024	False	1280	1	True	Windows computer
4	80	77.5%	1440, 1600 or 1920	760	900	False	1440 or 1600	1	True	Windows computer
5	159	72.95%	1920	950	1080	False	1920	1	True	Windows computer
6	7	71.42%	2560	1240	1440	False	1892	1	False	iMac
7	33	72.72%	1920	928	1080	False	1700	1	False	Non-retina Mac computer
8	395	76.96%	1366	645	768	False	1366	1	True	Windows computer
9	87	71.26%	1350	670	800–900	False	1280-1300	1	False	Non-retina Mac computer
10	2	50%	1080	500	1848	True	360	3	False	Android tablet
11	15	66.66%	768	950	1024	False	768	1 or 2	False	Mac computer
12	33	69.69%	1148	1280	800 or 1024	False	1280	2	False	Retina Mac computer

- Redirection to vertical A:
  - Android devices with a 2-pixel ratio.
  - Computers with an operating system different from Android, iOS and Mac OS.
  - Mac OS computers.
  - iPad devices.
  - Convertible devices (those that could be used as tablet or as laptop depending on whether a keyboard or mouse is attached to them).
- Redirection to vertical B:
  - Android devices with a 3- or 4-pixel ratio.
  - Large iPhone devices (iPhone 6 Plus, 6s Plus, or 7 Plus).
  - Android tablets.

If the devices of the users who participate in the reinforcement (validate) phase did not meet any of these characteristics, the redirection was randomly made between verticals A and B (maintaining a 50% distribution).

No users were redirected to vertical C due to the low finalization rates of the clusters in this questionnaire variant. There was only one rule that did not follow this assumption: the case of an Android device with a very high resolution (a pixel ratio of 3 or 4). Despite this case, the researchers decided to close this vertical C, as all the mobile or tablet devices with a very high resolution (like iPhone 6 Plus, 6s Plus, 7 Plus, or Android tablets) work better in vertical B.

The final established heuristic rules were the following (presented as a kind of pseudocode):

1. If the operating system is Android and the device's pixel ratio is 2, the user is redirected to vertical A.
2. If the operating system is Android and the device's pixel ratio is 3 or 4, the user is redirected to vertical B.
3. If the operating system of the device is iOS and its pixel ratio is 3 (iPhone 6 Plus, 6s Plus, or 7 Plus), then the user is redirected to vertical B.
4. If the operating system is neither Android nor Mac OS, iOS, the user is redirected to vertical A.
5. If the operating system of the device is Mac OS, the user is redirected to vertical A.
6. If the operating system is Android and the device's screen height is greater than 1000px, the user is redirected to vertical B.
7. If the operating system is iOS, the device's screen width is 1024px, the device's screen height is 768px, and the device's pixel ratio is 1 or 2 (iPad), the user is redirected to vertical A.
8. If the device's operating system is Android and the device type is neither a mobile nor a tablet (convertible device), the user is redirected to vertical A.
9. If a device does not fit any of the previous conditions, the user is randomly redirected to vertical A or B (with equal probability of being redirected to any of them).

These rules were implemented in the OEEU's ecosystem to apply them whenever a new user enters or resumes the questionnaire.

C. RESULTS REGARDING ADAPTABILITY AND USERS' REDIRECTION WITHIN A/B TEST VERTICALS

After the experiment took place (analyzing the interaction and performance of users who used the system previously), all the users who entered or returned to the questionnaire (and therefore, the target users of the experiment) were sought to obtain the results regarding the application of redirection criteria within the questionnaire verticals. The calculation and validation presented in this phase correspond to the (3) mark in Figure 1.

Before this phase (called reinforcement because the participants are users who access the web forms in a reinforcement made by the OEEU to obtain more responses to the questionnaires) and the application of the redirection rules based on heuristics, 5768 users had started the questionnaire; 4410 of them finished it, leaving a total of 1358 uncompleted questionnaires (and reaching a completion rate of 76.46%). All the data related to this subsection are available at https://github.com/juan-cb/paper-ieeeAccess-2017/blob/master/reinforcement-results.ipynb [26]

In these previous results, the users who *entered* the questionnaire (i.e., reached the welcome page but never started it) were not taken into account. If these users were considered, the results would be as follows:

- Number of students who have *entered* the questionnaire: 6360.
- Number of students who have *not finished* the questionnaire: 1950.
- Number of students who have *finished* the questionnaire: 4410.
- Completion rate *before* reinforcement: 69.34%.

By the time the questionnaires were closed, the final results were the following: 6738 started questionnaires, of which 5214 were completed and 1524 uncompleted. Consequently, the study achieved a questionnaire completion rate of 77.38%, improving the previous rate.

Again, these are the results for the started questionnaires; considering all the users (including the ones who reached the welcome page), the study yields the following results:

- Number of students who have *entered* the questionnaire: 7349.
- Number of students who have *not finished* the questionnaire: 2135.
- Number of students who have *finished* the questionnaire: 5214.
- Completion rate *after* reinforcement: 70.95%.

The total number of target users who entered the questionnaire after the incorporation of the system redirection support was 1165. These 1165 users were classified into three groups:

- Users who *entered* the questionnaire *after* reinforcement (considered as "new users"). There were 1003 new users, becoming the larger group of users who have taken part in the experiment.
- Users who resumed the questionnaire *after* reinforcement and were redirected to a different vertical; 110 users satisfied this criterion.
- Users who resumed the questionnaire *after* reinforcement but were *not* redirected to a different vertical. There were 52 users of this type.

These general results are summarized in Table 15.

TABLE 15. General results in the reinforcement phase.

User type	Total	Results	Completion rate
New users	1003	718 finished questionnaires 285 *not* finished questionnaires	71.59%
Redirected users	110	61 finished questionnaires 49 *not* finished questionnaires	55.45%
Not redirected users	52	25 finished questionnaires 27 *not* finished questionnaires	48.08%

As can be seen in Table 15, the new users' sample reached a completion rate of 71.59%.

This sample includes users who (at least) reached the welcome page of the questionnaire after reinforcement. An improvement in the completion results could be seen when comparing this completion rate (71.59%) with the completion rate before the reinforcement (that includes all the users who entered the questionnaire, 69.34%). Furthermore, it is necessary to consider that these new users are more reluctant in completing the questionnaire, as they have been invited to participate at least twice previously (and they had ignored the invitations), so these results are even more valuable.

Once the participant finalization rates were calculated, the researchers proceeded with the analysis of the impact of the rules formulated to improve the finalization rate, taking as a reference the groups of users who accessed the questionnaire presentation page both before and after the reinforcement phase.

These users were grouped into categories according to the way in which they were assigned to their vertical. To generate these categories, the researchers applied the assignment rules to the group of users who participated prior to the reinforcement and compared the results (ideal vertical assignment) with the vertical to which these individuals were actually sent (actual vertical assigned). Thus, the following three groups of individuals were obtained:

- **Pre-reinforcement users randomly assigned to the wrong vertical (G1, n = 3833):** Composed of users who

TABLE 16. Correlation between the vertical assignment and the finalization rate.

	Finalization rate	Chi-squared	Significance
G1-G2	67.9-74.9	25.927	0.000
G2-G3	74.9-71.6	3.442	0.064
G1-G3	67.9-71.6	5.130	0.024

accessed the questionnaire before the reinforcement and were assigned to a vertical to which they would not have been assigned had the redirection rules been applied.

- **Pre-reinforcement users randomly assigned to the right vertical (G2, n = 1542):** Comprised of users who accessed the questionnaire before the reinforcement and who, despite having been randomly directed, were assigned to the vertical to which they would have belonged to, had the redirection rules been applied.
- **Post-reinforcement users (G3, n=1003):** Users who accessed the questionnaire for the first time after the reinforcement, thus being consequently assigned to the right vertical.

In the case of rule 9, researchers classified all individuals who were randomly directed to vertical C as members of group 1; individuals who were directed to verticals A or B were classified as missing values, as the distribution of those verticals was defined differently from the one defined for the reinforcement phase.

Once the users were classified, the researchers calculated the finalization rate of each group, using the Chi-square statistic to study whether the vertical assignation method influenced the finalization rate. The Chi-square test is the most reliable in this scenario, given that there are two categorical variables (questionnaire finalization and success in the assignment). This statistical test was applied to the three possible combinations of pairs (Table 16).

First, as we can observe in the table, the results of the Chi-square test reflect a significant correlation between the vertical assignation method and the finalization of the questionnaire in pair G1–G2 for a significance level (s.l.) of 0.001. This result is consistent with the methodology employed, given that the clustering process and the later rules of assignment were carried out using the pre-reinforcement users.

Second, for the pair G2–G3, the results indicate no correlation between the assignment method and the finalization rate (s.l. 0.05) which, again, confirms the adequacy of the established rules, as individuals in group 3 were grouped with the same criterion that those in group 2, although the assignment was done in an intentional way rather than randomly.

Finally, it is noticeable that there is also a correlation (s.l. 0.05) between the assignment method and the finalization rate in the case of the pair G1–G3, which confirms that the application of the established rules significantly contributes to the finalization of the questionnaire by the participants.

TABLE 17. Correlation between the application rule and the finalization rate.

	Finalization rate		N		Chi-squared	Significance
	G1	G3	G1	G3		
Rule 1	67.11	72.64	374	106	1.167	0.280
Rule 2	70.99	72.22	362	126	0.069	0.793
Rule 3	62.75	56.52	51	23	0.258	0.612
Rule 4	68.03	76.25	2196	421	11.215	0.001
Rule 5	71.69	73.47	325	49	0.067	0.796
Rule 6	*	*	*	*	*	*
Rule 7	67.12	74.07	73	27	0.445	0.505
Rule 8	72.00	80.00	25	10	0.643**	0.488**
Rule 9	62.53	63.07	427	241	0.019	0.890

*No individuals in group 2. **Fisher's exact test (odds ratio and p-value)

As a final data analysis step, the researchers carried out an in-depth study of the behavior of each of the proposed rules, aiming to delve into the individual effect of each of them on the finalization of the questionnaire.

To this end, a process like the previous analysis was used with each one of the rules, the difference being that only the pair G1–G3 was used (Table 17).

As illustrated in Table 16, although there are differences in all finalization rates, they are significant (s.l. 0.01) only in the case of rule 4. For said rule, the rate of finalization in group 3 is approximately 8% greater than the rate in group 1, which suggests that directing the individuals who access the questionnaire from a non-Mac PC improves their chances of completing the questionnaire.

IV. DISCUSSION

This section presents the discussion of all the issues found in the research, discussing the foundations and effects of some decisions made by the authors. It also includes several future lines of work, suggests a set of recommendations, and closes with a general conclusion.

A. GENERAL DISCUSSION

Regarding the research carried out by the authors, there are several issues to comment on in this paper. To facilitate the comprehension, these issues will be discussed following the same structure of the paper (first, issues related to the methodology; second, those related to the results, and so forth).

First, the authors pose a question: Is it advisable to apply this kind of machine-learning method to this kind of problem? In this case, the researchers were inspired by other authors who have applied these types of processes to a wide range of problems. As an example, this kind of machine-learning algorithmic approach has being used in other fields, such as education [35], with promising results. Beyond the benefits that machine-learning approaches bring to many problems, by also including white-box procedures, the researchers ensure explainable and reproducible results that could be improved or discussed by the scientific community. All these considerations and precedents encouraged the authors to employ this kind of approach to address the problem of improving users' performance within a complex system like that presented. According to the results, the question can be answered positively, as the findings have been valuable and prove the validity of the approach.

Following the discussion, the authors would like to comment that the A/B testing approach used for this research is not a *pure* application of such methodology. While A/B tests are commonly based on singular changes between the different experimentation groups (or verticals), in the presented approach the authors grouped different changes into the same verticals. In this case, this variation of A/B tests does not influence this experimentation, as the researchers attempt to maximize user performance in the questionnaire finalization without a special focus on small changes, but using important differences between the different verticals. Despite that, it is worth noting that this kind of application of A/B tests for the experiment has been previously validated by experts [29].

Regarding the generated predictive models, the cut-off value for their relevant factors to later include in the clusters, the authors stated 0.05 as the minimum value to consider since this is the most common value in classical literature to ensure reliable results. Also in this case, the authors use this cut-off value to generate the clusters using only the most important factors (those that have a specific weight of more than 0.05 in the predictive model), thus excluding less important ones that could introduce noise when building the groups.

Concerning the most important factors that characterize the predictive models and explain the users' profile and preferences while completing the questionnaire, it should be remarked that technical aspects were more important than personal ones. At the beginning of the research and for the predictive models' generation, researchers included personal aspects, such as gender, age, and issues related to education, as part of the dataset. According to the results, such aspects do not have special relevance while modeling the users' behavior in completing the web form. Instead, the present findings indicated that the most important factors for the users were the size of the device screen and the browser window. Moreover, other aspects, like the screen resolution, concrete browser, or operative system, were important, but with a lesser effect. Nevertheless, these are the most important factors for the population of this study and cannot be considered general and valid for other populations. To apply the approach presented in this research in other experiments, the predictive models should be generated again.

Regarding the generation of rules based on heuristics, and as a future study, the researchers would like to automate this process. This will help to reproduce the same process with the same experimental conditions and remove any kind of bias introduced by researchers or administrators. This will be explained in depth in the following subsection.

Related to the reinforcement phase and other conditions of the experiment, with the aim of enhancing users' participation in the questionnaires, the OEEU offered participation in a raffle (the prize would be seven smartwatches) to all graduates completing the web form as a reward. This incentive was used also to promote the reinforcement process where the redirection rules were applied.

Regarding the effectiveness of the use of rules based on cluster analysis during the reinforcement period, cluster analysis was found to be a very useful tool to guide the redirection of users to the version of the questionnaire best suited to the features of the technology with which they completed it.

First, the results of this study confirm that the rules established improved the answer rate by comparing the performance of users who participated after the reinforcement with those who participated before the last reinforcement and were directed to the wrong questionnaire. Additionally, the authors could observe that there are no significant differences between groups G2 and G3, which leads to the understanding that the application of the rules during the reinforcement has maintained the good results regarding to the finalization among the users who would have been randomly assigned to the right vertical.

Second, if the researchers delve into the analysis of the individual behavior of each rule, the results suggest that the improvement in the finalization rate is due to rule four, which redirects users who access the form from non-Mac computers to vertical A, given that the rest of rules have not yielded significant correlations.

Regarding this point, it must be remarked that the users who participated in a reinforcement phase were commonly more reluctant to complete the questionnaire, as they left it in previous stages or were not initially attracted to fulfill it. This also could render even more valuable the results obtained in this research concerning the improvement of users' performance. However, for future studies it would be interesting to apply a research design that includes an experimental and a control group from the beginning to be able to assess the effect of the rules under the same conditions.

Another interesting future line of research would be an analysis of the threshold cut-off to perform the factor selection, given that a higher minimum value may simplify the number of rules and make more efficient the redirection process. As a first step, the authors intend to analyze rule four to gain a better understanding of the predictive importance of the elements behind its formulation.

Finally, the authors believe that the approach and procedures presented in this research are transferable to other application fields. The process presented in this paper follows some *traditional* approaches and methods within the machine-learning research field, and the prediction challenge is present in many other problems beyond web form completion. The proposed methodology may also help to transfer this experience to other problems with the additional value of providing a white-box approach for the algorithms used. In the future, the authors would like to attempt to apply such methodology to predict the employability of Spanish graduates. This will also validate the genericity of the methodology, which will only require some minor changes depending on the dataset.

B. HOW TO APPLY THIS RESEARCH IN PRODUCTION IN THE REAL WORLD

One of the main concerns related to this research could be stated as follows: Is it possible to use this contribution in a real industry setting? Is it possible to integrate this kind of approach in production systems and enable an automated process? From the point of view of the researchers, the answer is yes to both questions. There are many examples in the industry on how data sciences processes can be transformed from Jupyter notebooks to enterprise-ready systems put in production. In this case, the researchers outline the approach proposed by the Airbnb engineering team on how their ML Automator [36] tool helped in translating a Jupyter notebook into an Airflow machine learning pipeline [37] and use this kind of analytics process in production systems. This automating effort must include—apart from the machine-learning algorithms and process—rule generation or the identification of the proper Euclidean distance to separate the clusters generated. To automate the rule generation, probably researchers would have to employ artificial intelligence techniques such as neuronal networks, that could learn to generate these rules as done by humans in this paper.

C. GENERAL CONCLUSION

This paper presents a novel study in the field of Human-Computer Interaction. The main results achieved have been quite promising and encourage authors to continue the labor of improving users' performance in completing large web forms. Adaptability can be achieved by detecting users' behaviors, preferences, and profiles using machine-learning techniques and offering the best user interface and user experience to each kind of user detected. Based on the results, the authors also propose several future works that could push this research to be adopted in the industry and other application fields.

ACKNOWLEDGMENTS

The authors thank Felicidad García-Sánchez for designing Figure 1.

REFERENCES

[1] R. Colomo-Palacios, F. J. García-Peñalvo, V. Stantchev, and S. Misra, "Towards a social and context-aware mobile recommendation system for tourism," *Pervasive Mobile Comput.*, vol. 38, pp. 505–515, Jul. 2017.
[2] D. A. Norman and S. W. Draper, Eds., *User Centered System Design: New Perspectives on Human-Computer Interaction*. Hillsdale, NJ, USA: L. Erlbaum Assoc. Inc., 1986.
[3] D. A. Norman, *The Design of Everyday Things: Revised and Expanded Edition*. New York, NY, USA: Basic Books, 2013.
[4] J. A. Bargas-Avila, O. Brenzikofer, S. P. Roth, A. N. Tuch, S. Orsini, and K. Opwis, "Simple but crucial user interfaces in the World Wide Web: Introducing 20 guidelines for usable Web form design," in *User Interfaces*, R. Matrai, Ed. Rijeka, Croatia: InTech, 2010. [Online]. Available: https://www.intechopen.com/books/user-interfaces/simple-but-crucial-user-interfaces-in-the-world-wide-web-introducing-20-guidelines-for-usable-web-fo, doi: 10.5772/9500.
[5] B. Shneiderman and C. Plaisant, *Designing the User Interface*, 4th ed. Reading, MA, USA: Addison-Wesley, 2005.
[6] M. Seckler, S. Heinz, S. Forde, A. N. Tuch, and K. Opwis, "Trust and distrust on the Web: User experiences and website characteristics," *Comput. Human Behavior*, vol. 45, pp. 39–50, Apr. 2015.
[7] B. Shneiderman, "Designing trust into online experiences," *Commun. ACM*, vol. 43, no. 12, pp. 57–59, Dec. 2000.
[8] S. P. Anderson, *Seductive Interaction Design: Creating Playful, Fun, and Effective User Experiences, Portable Document*. London, U.K.: Pearson Education, 2011.
[9] S. Krug, *Don't Make Me Think: A Common Sense Approach to Web Usability*. Bengaluru, India: Pearson Education, 2000.
[10] D. Malandrino, F. Mazzoni, D. Riboni, C. Bettini, M. Colajanni, and V. Scarano, "MIMOSA: context-aware adaptation for ubiquitous Web access," *Pers. Ubiquitous Comput.*, vol. 14, no. 4, pp. 301–320, May 2010.
[11] G. Pruvost, T. Heinroth, Y. Bellik, and W. Minker, "User interaction adaptation within ambient environments," in *Next Generation Intelligent Environments*. New York, NY, USA: Springer, 2016, pp. 221–263.
[12] C. Flavián, R. Gurrea, and C. Orús, "The effect of product presentation mode on the perceived content and continent quality of Web sites," *Online Inf. Rev.*, vol. 33, no. 6, pp. 1103–1128, 2009.
[13] 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)," Universidad Politécnica Madrid, Madrid, Spain, 2016.
[14] 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 3rd Int. Conf. Technol. Ecosyst. Enhancing Multiculturality (TEEM), Porto, Portugal, 2015.
[15] A. Vázquez-Ingelmo, J. Cruz-Benito, and F. J. García-Peñalvo, "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, USA: IGI Global, 2018.
[16] Django Software Foundation. (Mar. 15, 2015). *Django Web Framework*. [Online]. Available: https://www.djangoproject.com/
[17] S. Stieger and U.-D. Reips, "What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study," *Comput. Human Behavior*, vol. 26, no. 6, pp. 1488–1495, Nov. 2010.
[18] W. McKinney, *Python for Data Analysis: Data Wrangling With Pandas, NumPy, and IPython*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2012.
[19] W. McKinney. (2017). *Pandas, Python Data Analysis Library*. Accessed: 2017. [Online]. Available: http://pandas.pydata.org/
[20] W. McKinney, "Data structures for statistical computing in python," in *Proc. 9th Python Sci. Conf.*, vol. 445. Austin, TX, USA, 2010, pp. 51–56.
[21] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[22] S. Raschka, *Python Machine Learning*. Birmingham, U.K.: Packt Publishing Ltd., 2015.
[23] T. Kluyver *et al.*, "Jupyter notebooks—A publishing format for reproducible computational workflows," in *Proc. ELPUB*, 2016, pp. 87–90.
[24] M. Ragan-Kelley *et al.*, "The Jupyter/IPython architecture: A unified view of computational research, from interactive exploration to communication and publication," in *Proc. AGU Fall Meeting Abstracts*, 2014.
[25] F. Perez and B. E. Granger, "Project Jupyter: Computational narratives as the engine of collaborative data science," Tech. Rep., 2015. [Online]. Available: https://blog.jupyter.org/project-jupyter-computational-narratives-as-the-engine-of-collaborative-data-science-2b5fb94c3c58
[26] J. Cruz-Benito, A. Vázquez-Ingelmo, and J. C. Sánchez-Prieto. (2017). *Enabling Adaptability in Web Forms Based on User Characteristics Detection Through A/B Testing and Machine Learning*, Code Repository That Supports the Research Presented in the Paper. [Online]. Available: https://github.com/juan-cb/paper-ieeeAccess-2017
[27] Y. Xu, N. Chen, A. Fernandez, O. Sinno, and A. Bhasin, "From infrastructure to culture: A/B testing challenges in large scale social networks," presented at the 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Sydney, NSW, Australia, 2015.
[28] E. Dixon, E. Enos, and S. Brodmerkle, "A/b testing of a webpage," U.S. Patent 7 975 000, Jul. 5, 2011.
[29] D. Siroker and P. Koomen, *A/B Testing: The Most Powerful Way to Turn Clicks Into Customers*. Hoboken, NJ, USA: Wiley, 2013.

[30] J. Cruz-Benito *et al.*, "Improving success/completion ratio in large surveys: A proposal based on usability and engagement," in *Proc. 4th Int. Conf Learn. Collaboration Technol. Technol. Edu. (LCT)*, 2017, pp. 352–370.
[31] H. Wickham, "Tidy data," *J. Statist. Softw.*, vol. 59, no. 10, pp. 1–23, 2014.
[32] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
[33] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep. 1967.
[34] Scikit-Learn. (Sep. 4, 2017). *API Reference—Scikit-Learn Documentation: Metrics*. [Online]. Available: http://scikit-learn.org/stable/modules/classes.html
[35] 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*, vol. 5, pp. 23792–23802, 2017.
[36] R. Chang. (Sep. 4, 2017). *Using Machine Learning to Predict Value of Homes On Airbnb*. [Online]. Available: https://medium.com/airbnb-engineering/using-machine-learning-to-predict-value-of-homes-on-airbnb-9272d3d4739d
[37] M. Beauchemin. (Sep. 4, 2017). *Airflow: A Workflow Management Platform*. [Online]. Available: https://medium.com/airbnb-engineering/airflow-a-workflow-management-platform-46318b977fd8

JUAN CRUZ-BENITO received the M.Sc. degree in intelligent systems from the University of Salamanca, Spain, in 2013, where he is currently pursuing the Ph.D. degree 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 is involved in human–computer interaction, educational virtual worlds and technologies for educational purposes, and disciplines that he has developed in many innovation and research projects. He has participated in many European and national research and development projects, such as TRAILER, VALS, USALSIM Virtual Campus, and the Spanish Observatory for University Employability and Employment, as a Software Engineer, Researcher, and Developer.

ANDREA VÁZQUEZ-INGELMO was born in Salamanca, Castilla y León, Spain, in 1994. She received the bachelor's degree in computer engineering from the University of Salamanca, Salamanca, in 2016, where she is currently pursuing the master's degree in computer engineering. She is a member of the Research Group of Interaction and eLearning. Since 2016, she has been a part of the National Project Spanish Observatory for University Employability and Employment as a Developer and a Researcher.

JOSÉ CARLOS SÁNCHEZ-PRIETO received the bachelor's degree in pedagogy and the master's degree in ICT applied in education from the University of Salamanca, Spain. He is currently pursuing the Ph.D. degree with the Faculty of Education, Said University, within the Program on Education in the Knowledge Society. His area of research is the assessment of attitudes among in-service and pre-service teachers.

ROBERTO THERÓN received the Diploma degree in computer science from the University of Salamanca, the B.A. degree from the University of A Coruña, the bachelor's degree in communication studies and the bachelor's degree in humanities from the University of Salamanca, and the Ph.D. degree from the Research Group Robotics, University of Salamanca. His Ph.D. thesis was on parallel calculation configuration space for redundant robots. He is currently the Manager of the VisUsal Group (within the Recognized Research Group GRIAL), University of Salamanca, 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. He has authored of over 100 articles in international journals and conferences. In recent years, he has been involved in 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 received the Extraordinary Doctoral Award for his Ph.D. thesis.

FRANCISCO JOSÉ GARCÍA-PEÑALVO received the degrees in computing science from the University of Salamanca and the University of Valladolid, and the Ph.D. degree from the University of Salamanca. He is currently the Head of the Research Group Interaction and eLearning. He was the Vice Chancellor of Innovation with the University of Salamanca from 2007 to 2009. He has led and participated in over 50 research and innovation projects. 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 is a member of the Program Committee of several international conferences and a reviewer of several international journals. 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 currently the Editor-in-Chief of the *International Journal of Information Technology Research* and the *Journal of Education in the Knowledge Society*. He is also the Coordinator of the multidisciplinary Ph.D. Program on education in the Knowledge Society.

MARTÍN MARTÍN-GONZÁLEZ received the master's degree in economic development and public policy and the Ph.D. degree in economics from the Autonomous University of Madrid (UAM), and the degree in economics from the University of La Laguna. He was a Researcher with the Faculty of Economics, UAM. He is currently a Researcher with the UNESCO Chair in University Management and Policy, Technical University of Madrid. 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.