Influence of personality and modality on peer assessment evaluation perceptions using Machine Learning techniques

Cristina Cachero, Juan Ramón Rico-Juan, Hermenegilda Macià

PII: S0957-4174(22)02168-6
DOI: https://doi.org/10.1016/j.eswa.2022.119150
Reference: ESWA 119150

To appear in: *Expert Systems With Applications*

Received date : 17 February 2022
Revised date : 14 October 2022
Accepted date : 24 October 2022

Cristina Cachero: Conceptualization, Methodology, Funding acquisition, Writing - Review & Editing, Validation, Supervision

Juan Ramón Rico-Juan: Methodology, Data curation, Software, Writing- Original draft preparation, Formal analysis

Hermenegilda Macià: Resources, Investigation, Writing- Original draft preparation, Funding acquisition, Project administration

- Effect of personality traits on the acceptance of peer assessment (PA) in students
- Evaluation if it changes depending on the PA modality (individual, pairs or threes)
- Data Analysis using ML techniques, the best predictions are with the RF algorithm
- Application of eXplainable AI techniques showing best predictors
- These predictors are true regardless of PA modality and practical considerations

Influence of personality and modality on peer assessment evaluation perceptions using Machine Learning techniques

Cristina Cachero[c,], Juan Ramón Rico-Juan[c,], Hermenegilda Macià[d,*],,

[c] *Departament of Software and Computing Systems, University of Alicante, 03690, Alicante Spain*

[d] *Departamento de Matemáticas, Escuela Superior de Ingeniería Informática, Universidad de Castilla-La Mancha, 02071 Albacete, Spain*

Abstract

The successful instructional design of self and peer assessment in higher education poses several challenges that instructors need to be aware of. One of these is the influence of students' personalities on their intention to adopt peer assessment. This paper presents a quasi-experiment in which 85 participants, enrolled in the first-year of a Computer Engineering programme, were assessed regarding their personality and their acceptance of three modalities of peer assessment (individual, pairs, in threes). Following a within-subjects design, the students applied the three modalities, in a different order, with three different activities. An analysis of the resulting 1195 observations using ML techniques shows how the Random Forest algorithm yields significantly better predictions for three out of the four adoption variables included in the study. Additionally, the application of a set of eXplainable Artificial Intelligence (XAI) techniques shows that Agreeableness is the best predictor of Usefulness and Ease of Use, while Extraversion is the best predictor of Compatibility, and Neuroticism has the greatest impact on global Intention to Use. The discussion highlights how, as it happens with other innovations in educational processes, low levels of Consciousness is the most consistent predictor of resistance to the introduction of

[*] Corresponding author: Tel.: +34-967599200; Fax: +349-967599224

Email addresses: ccachero@dlsi.ua.es (Cristina Cachero), juanramonrico@ua.es (Juan Ramón Rico-Juan), hermenegilda.macia@uclm.es (Hermenegilda Macià)

Preprint submitted to Expert Systems with Applications *October 14, 2022*

peer assessment processes in the classroom. Also, it stresses the value of peer assessment to augment the positive feelings of students scoring high on Neuroticism, which could lead to better performance. Finally, the low impact of the peer assessment modality on student perceptions compared to personality variables is debated.

Key words: Peer Assessment (PA), Personality, Quasi-experiment, Use Behaviour, eXplainable Artificial Intelligence (XAI), Machine Learning (ML)

1. Introduction

The educational community has, in recent years, regularly advocated involving peer students in assessment practices at all educational levels (Li et al., 2020, 2016), since, if carefully designed, peer assessment (PA) can provide several advantages in higher education for both teachers –as instructors– and students –as evaluators–. From the instructor perspective, peer assessment increases teachers' efficiency in grading and facilitates a formative evaluation in which students receive rapid feedback on their work with a reasonable investment of time (Falchikov & Goldfinch, 2000; Topping, 2003). Students, meanwhile, benefit from a pedagogical strategy that facilitates their learning (Adachi et al., 2018; Double et al., 2020; Sanchez et al., 2017; Wang et al., 2012) and promotes their social-affective development (Li et al., 2020). The change of role from student to evaluator helps them to develop a habit of reflection and a constructive critical spirit (Panadero et al., 2013), which, in turn, fosters autonomy (Li et al., 2020; Shen et al., 2020) and improves self-regulation capabilities (Nicol et al., 2014; Reinholz, 2016). Even when faced with complex tasks, students, with the proper use of domain-specific scaffolding, may improve their performance, while their perceived mental effort is reduced (Könings et al., 2019). All these advantages have been backed by recent meta-analyses focused on the effect of peer assessment on learning across multiple educational settings (Li et al., 2020; Zheng et al., 2020), and also by qualitative studies in which the perceptions of both students and teachers have been analysed (Adachi et al., 2018; To &

Panadero, 2019).

However, to attain such gains, it is important to take into account the hindrances and challenges specific to implementing a successful peer assessment environment (Adachi et al., 2018). Some of these inhibitors to learning and achievement have been widely studied in the context of peer assessment. This is the case of time and resource constraints, power relations, perceived expertise and reliability and accuracy of students' judgement skills (Liu & Carless, 2006; Panadero et al., 2019). However, others, such as individual differences, are still under-researched (Chang et al., 2021; Rivers, 2021), despite being widely recognised in educational psychology as affecting both learning and achievement (An & Carr, 2017). This paper aims to help fill this gap by examining the effect of students' personality on their subjective assessment of the merits of the peer assessment technique (perceived usefulness, ease of use, and compatibility), under the assumption that different personality profiles may have different subjective perceptions of these merits, which will affect (a) their behavioural intention (BI) towards the use of the technique and (b) the degree to which they benefit from using it. This aspect is analysed in first research question addressed in this paper.

Furthermore, peer assessment can be applied individually (each evaluator individually reviews the work of their peers) or in collaboration with others (evaluators meet in groups of 2 or more members to jointly assess the work of their peers). On the basis of Vygotsky's social development theory (Vygotsky, 1980), which emphasises the key role of social interaction in learning, it is possible to argue that increasing the number of assessors (and therefore making the assessment process more social) should benefit the outcomes of the process and augment the BI towards the use of a peer assessment evaluation process. However, in Rico-Juan et al. (2022), it was shown that the impact of the peer assessment modality (PAM) on this BI is lower than expected. In this paper, we explore the hypothesis that this may be because the individual students' personality profiles may be mediating this effect. This aspect is analysed in the second research question addressed in this paper.

To study these issues, we have used ML-based techniques rather than purely statistical ones, as the former are able to learn both linear and non-linear relationships between variables, and are therefore able to accommodate a wider variety of data distributions and lead to more accurate predictions (Bates & Watts, 2007). Furthermore, we have applied explainability (eXplainable Artificial Intelligence, XAI-ML) (Arrieta et al., 2020) to these more accurate ML models to be able to quantify the influence of the input (independent) variables on the target variable at a general, group or individual level (with post-hoc techniques). This has allowed us to design a more effective action plan than what would have been possible had we used other data analysis approaches. The potential advantages associated with using ML techniques to analyse data have led to these techniques being increasingly used in a variety of fields (Dalzochio et al., 2020; Dhini et al., 2021; Liu, 2020), including education where we can find a general review of ML applied (Menon & Janardhan, 2021) and some examples of concrete uses (Embarak, 2021; Rico-Juan et al., 2019; Wu, 2021).

This research might be useful for instructors aiming to maximise students' satisfaction and learning outcomes by (a) helping them to predict which student profiles are going to benefit the most of this assessment technique when they introduce it in their teaching contexts, and (b) helping them to choose the most appropriate peer assessment modality to maximize those benefits.

The remainder of this paper is organised as follows: Section 2 presents the conceptual model that provides definitions for the main experimental constructs. Section 3 presents the related empirical work regarding the impact of personality on the different components of use behaviour. The experimental design used in this study is explained in Section 4, while the details of its execution can be found in Section 5. The ML-based analysis of the data is presented in Section 6 and the implications to be drawn from the data in relation to the current educational landscape are discussed in Section 7. Finally, in Section 8, the main conclusions and some future lines of research are outlined.

2. Conceptual Model

In order to empirically assess how the inclusion of a pedagogical innovation affects both students' productivity and attitudes towards that innovation, we draw on two existing theoretical models.

On the one hand, the set of individual differences that may impact work productivity can be organised, following the Model of Human Performance (Blumberg & Pringle, 1982), into two dimensions: Capacity (including variables such as level of education, cognitive abilities and work experience, to name a few) and Willingness (psychological and emotional characteristics, such as motivation or personality). The model adds a third dimension, unrelated to individual differences, called Opportunity, which refers to the context in which the innovation is deployed (tools, materials, working conditions, etc.). In the context of peer assessment, a number of capacity and opportunity variables have been extensively studied with respect to grade accuracy and impact on learning outcomes (e.g. types of tasks, or the effect of implementing a paper-based or a computer-mediated peer assessment process (Zheng et al., 2020)). However, there is an empirical gap regarding the impact of Willingness variables in general, and personality in particular, on these same variables (Chang et al., 2021; Rivers, 2021).

On the other hand, the attitude towards the use of pedagogical innovation refers to the need for a pedagogical innovation not only to improve learning outcomes for students, but also to foster a positive attitude towards its adoption and the learning process in general. The behavioural intention towards innovations has been widely studied in the literature, where several models have been proposed for its operationalisation (Diéguez et al., 2012; Lai, 2017). Most of these models have been derived from the Technology Acceptance Model (TAM) (Davis, 1989), and have been applied to different domains, including education (Rakoczy et al., 2019). This paper uses the adaptation presented in (Martínez et al., 2013), called Unified Method Adoption Model (UMAM), which centres on the behavioural intention towards new working methods, rather than technologies. It decomposes the Behavioural Intention construct into six main sub-components: Usefulness, Ease of Use, Compatibility, Subjective Norm, Voluntariness and Intention to Use. Again, there is an empirical gap regarding how personality affects behavioural intentions towards new evaluation methods in education (Chang et al., 2021; Rivers, 2021).

The hypothesis underlying this research is that the same three factors which, according to the model of work performance, impact productivity, together with productivity itself, may impact the Behavioural Intention to adopt new evaluation methods in education. This hypothesis is reflected in the conceptual model presented in Fig. 1. This same hypothesis has been partially confirmed in the context of education in previous research works. For example, in Mailizar et al. (2021), the authors study the impact of certain capacity variables on the intention to adopt e-learning, while in Abu-Al-Aish & Love (2013) various capacity and opportunity variables, together with performance, proved to be important to explain behavioural intentions to use m-learning.

Given the previously mentioned research gap regarding the Willingness dimension, this study focuses on the relationship between the five Personality sub-components (Neuroticism, Extraversion, Agreeableness, Conscientiousness and Openness to Experience) and four of the six behavioural intention sub-components (Intention to Use, Usefulness, Ease of Use and Compatibility). These constructs, together with their relationships, are marked in bold in Fig. 1. We have intentionally omitted the Subjective Norm and the Voluntariness sub-components of the BI construct from the study, since they are not applicable to our experimental context. Also, we have left out of the scope of the study the internal relationships that may exist among the different sub-components of each construct.

Next, we further delve into each of the theoretical constructs involved in this experimental study.

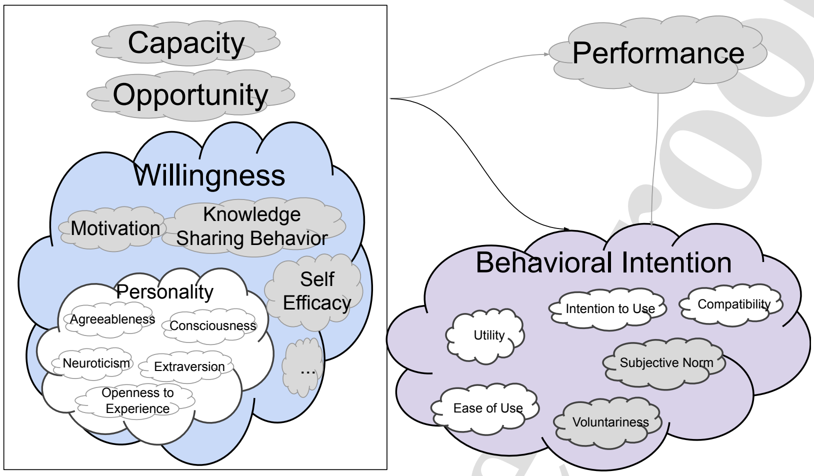

Figure 1: Conceptual Model: Overview

2.1. Personality

According to the American Psychological Association (APA), personality can be defined as *individual differences in characteristic patterns of thinking, feeling, and behaving*. Personality matters because it predicts and explains intention, behaviour and productivity at work (Barrick, 2005; Curtis et al., 1988).

Personality psychologists consider the Big Five (BF) personality model (John et al., 2008) and Eysenck's Hierarchical Three Factor model (PEN) (Eysenck, 1994) to be the two theories that best represent the personality structure (Avia et al., 1995; Feldt et al., 2010). Additionally, a comparison between models has demonstrated the benefits of the BF model in terms of both completeness and measurement reliability (Balijepally et al., 2006). For these reasons, BF currently dominates the personality research landscape (Cruz et al., 2015; De Raad & Schouwenburg, 1996; Terzis et al., 2012), including education (Bergold & Steinmayr, 2018; Caprara et al., 2011; Terzis et al., 2012). It is widely accepted among the research community that personality traits are useful to predict both achievement and behaviours.

The BF model intended to classify all major sources of individual personality differences. For this purpose, it includes five factors: Extraversion, Openness to experience, Agreeableness, Conscientiousness and Neuroticism. The definitions of these factors are as follows (McCrae & John, 1992):

- Neuroticism (N) represents individual differences in the tendency to experience distress, and in the cognitive and behavioural styles that follow from this tendency. High N scorers experience greater levels of nervous tension, depression, frustration, guilt, self-consciousness and ineffective coping. At the opposite end of the spectrum, individuals scoring low in N may be defined as calm and relaxed.

- Extraversion (E) represents individual differences in terms of positive emotionality. High E scorers tend to be cheerful, enthusiastic, optimistic, energetic, dominant, talkative, sociable, and warm, while low E scorers can be described as quiet, reserved, retiring, shy, silent, and withdrawn.

- Agreeableness (A) represents individual differences in friendliness. Characteristics such as altruism, nurturance, caring, trust, modesty, and emotional support are typical of high A scorers, while hostility, indifference to others, self-centeredness, spitefulness, and jealousy are traits usually associated with low A scorers.

- Conscientiousness (C) represents individual differences in impulsiveness and will to achieve. High C scorers are thorough, neat, well organised, diligent, achievement-oriented and usually governed by conscience, while low C scorers are not.

- Openness to Experience (O) refers to individual differences in intellect traits. High O scorers show intellectual interest, aesthetic sensitivity, need for variety, and unconventional values. They are curious, imaginative and perceptive, and are open to fantasies, feelings, sensations, and values. Individuals rated low on the O factor tend to judge in conventional terms, favour conservative values, and repress anxiety.

2.2. Behavioural Intention

Regarding behavioural intention, both existing theories from social psychology and a number of empirical studies show that future behaviour can be predicted by intentions to engage in the behaviour (Agarwal & Prasad, 2000). The specific perception attributes that contribute to accepting /adopting a specific innovation vary depending on the nature of the innovation. The acceptance of new products is usually best predicted through the Technology Acceptance Model (TAM), which defines two intention drivers: usefulness and ease of use (Davis, 1989). This model has been used and adapted in the field of education since the very introduction of computers in the learning process (Bagozzi et al., 1992; Koufaris, 2002; Lazar et al., 2020; Lowther et al., 1998; Šumak et al., 2011).

However, when the acceptance involves not only using new technologies but also changing established processes, as is the focus of this paper, it has been proven that patterns of intention determinants differ (Hardgrave et al., 2003), and new variables, such as compatibility with the current way of working or pressures from the environment (social norm) may also play a key role (Diéguez et al., 2012). To the best of our knowledge, in the particular case of pedagogical innovations, the potential explanatory value of these additional variables is still unknown (see e.g. Casey et al. (2021); Rakoczy et al. (2019)), although the TAM has been expanded in other directions, e.g. by adding new axes to the model (phases of use, users, and components) and considering usefulness and ease of use for each of the 27 combinations of these aspects (Persico et al., 2014). In this paper, we recognise the potential value of all these variables (see Fig. 1). Given the context of this study, our data analysis centres on the following set of behavioural intention sub-components:

- Usefulness (U): a.k.a. Perceived Usefulness. Degree to which a person believes that using a particular method will enhance his/her job performance.
- Ease of Use (EoU): a.k.a. Perceived Ease of Use. Degree to which a person

believes that using a particular method will be free of effort.

- Compatibility (C): Degree to which a method is perceived as being consistent with existing values, principles, practices and the past experience of the potential adopter.
- Intention to Use (I2U): Overall intention to adopt the method in the future if given the opportunity.

We next present the main empirical results published thus far regarding the relationship between these two sets of variables.

3. Related Work: Impact of Personality on I2A

As mentioned, the Big Five is the most parsimonious and comprehensive framework of personality. On the other hand, regarding BI, the TAM model - and its different adaptations - is also well known both inside and outside the educational community.

Given the popularity of the two models, several papers have addressed in the past the relationship between personality and the intention to use different technologies, from ERPs (Benlian & Hess, 2010) and green information technology (Dalvi-Esfahani et al., 2020) to self-driving cars (Qu et al., 2021), or mobile banking adoption (Agyei et al., 2020). In all of these, it was found that personality dimensions can be valuable determinants of users' intentions and perceptions. There is also a considerable body of literature on the relationship between personality and adoption of new processes in areas other than education. In Jarillo-Nieto et al. (2015), the authors use the BF model to show how higher levels of neuroticism are related to lower attachment to structured processes, while higher levels of agreeableness, openness to experience and conscientiousness are related to higher satisfaction with the adoption of new processes; no relationship has been found between extroversion and process attachment or satisfaction. In Computer Science, we can also find a number of papers that address the relationship between personality and adoption of new technologies/technical processes: such is the case of developers' intention to use Agile Methods (Gandomani et al., 2014), IT professionals' ability to adapt to a technological innovations (Gallivan, 2004), analysts' intention to use Model Driven Engineering techniques (Toala et al., 2018), or adoption/rejection of innovative Software Engineering processes and practices in industrial settings (O'Connor & Yilmaz, 2015).

However, in the field of education, the number of studies on the relationship between personality and adoption of pedagogical process innovations is relatively low. In Bhagat et al. (2019), the authors conclude that online learning solutions are known to appeal to students with higher Conscientiousness and Intellect/Imagination, and lower Neuroticism. In Devaraj et al. (2008), the authors centre on the interaction between TAM and personality in the context of the use of collaborative technologies for education. Their main findings are: (1) Conscientiousness significantly moderates the relationship between Perceived Usefulness (PU) and Intention to Use; (2) Extraversion moderates the relationship between Subjective Norm and Intention to Use; (3) Neuroticism is negatively associated with Perceived Usefulness; (4) Openness is positively associated with Perceived Usefulness and (5) Agreeableness is positively associated with Perceived Usefulness and moderates the relationship between Subjective Norm and Intention to Use. In Lazar et al. (2020), the authors extend the TAM model with a personality trait (Anxiety) and a compatibility trait (Familiarity) in the context of blended learning in higher education. In this model, Anxiety is considered a mediator factor that negatively affects Usefulness and, to a lesser extent, Ease of Use and Intention to Use. Furthermore, in Rivers (2021), the authors include the cultural dimension when they examine the role of personality traits and academic self-efficacy in acceptance, actual use, and achievement. They conclude that, in the context of a socially distanced asynchronous university course held in Japan and supported by Moodle, Agreeableness and Conscientiousness have positive direct effects on online academic self-efficacy, in addition to positive indirect effects on the acceptance of Moodle. Moreover, Agreeableness and Conscientiousness have an indirect effect on course achievement while none of the five-factor model personality traits has an influence on actual Moodle use.

Lastly, to the best of our knowledge, only Terzis et al. (2012) addresses the relationship between personality and adoption of new ways of assessment, namely a Computer Based Assessment (CBA). In their study, the authors report that Neuroticism has a negative effect on Usefulness of computer-based assessments, while Agreeableness and Conscientiousness have a positive effect on Ease of Use.

It is also noteworthy that none of the quantitative related articles apply ML techniques for the data analysis of the empirical data, and therefore fail to study possible non-linear relationships in the data set.

4. Research Method

As argued in Section 1, PA requires careful design and implementation for it to be an effective tool for formative assessment processes (Wanner & Palmer, 2018). This section presents the objectives and defines the context, the research questions, the variables and measurement instruments, and the experimental design.

4.1. Objectives and context definition

The objective of this study was to systematically assess the impact of the personality and the peer assessment modality (individual, pairs and in-threes) on students' behavioural intention towards the use of peer assessment, while maintaining other well-known influential factors constant.

In order to fulfil this objective, and study the effect that different modalities of PA may have on the behavioural intention towards PA, depending on the particular personality profile, the chosen experimental design was that of a quasi-experiment. A quasi-experiment is a type of controlled experiment in which the subjects are not randomly assigned to treatments, but rather preexisting groups are assigned to each treatment -in our case, the peer assessment modality-. It is thus possible to study cause-effect relationships in scenarios such as education, in which students are pre-organised in groups, usually with different instructors and timetables, and a truly random assignment and management of participants across the different experiment modalities is, therefore, highly complicated (Kampenes et al., 2007).

Our study was conducted with first-year students at universities. This age group is of particular significance, since the school-to-university transition poses a major challenge for students in pursuing learner independence, self-assessment autonomy and academic performance demands (Webster & Yang, 2012; Yucel et al., 2014). To & Panadero (2019) showed how the use of peer assessment in this population can support the development of this self-assessment autonomy, provided that students are guided by an appropriate scaffolding process.

4.2. Research questions

The research questions addressed in this study were designed to be answered using quantitative data. The questions are the following:

- RQ1: What is the relationship between the different personality traits and the BI towards a peer assessment evaluation process?
- RQ2: How does the relationship between personality traits and BI change depending on the peer assessment modality being evaluated?

4.3. Variables and Measurement Instruments

The variables considered in this study were already defined in Section 2 (see model constructs marked in bold in Fig.1). In order to operationalise these constructs, the following two questionnaires, both validated by the research community, were selected:

- Big Five Inventory (BFI-44) (Li et al., 2015): Spanish version. This questionnaire includes items for the five personality traits: E (8 items), A (15 items), C (11 items), N (5 items), and O (5 items).

- UMAM-Q (Diéguez et al., 2012): This questionnaire was developed to specifically measure method adoption. It includes 42 items divided into six scales: U (7 items), EoU (7 items), C (7 items), SN (7 items), V (7 items) and I2U (7 items). For this study, only the U, EoU, C and I2U components were included.

Additionally, an Evaluation Preference variable was defined and measured through the question 'Which method of assessment do you prefer?', with two possible answers (Peer assessment/Teacher assessment). Also, a Modality Preference variable was defined and measured through the question 'Which PAM do you prefer?', with three possible answers (Individual/In pairs/In threes). These questions were added to a final questionnaire, together with a set of open questions to capture the perceived advantages/disadvantages of each modality of the peer assessment evaluation method.

4.4. Experimental planning

In this study, we planned to gather data from 85 students (8 female), who were enrolled on the Calculus and Numerical Methods (CNM) course[1] in the first year of the Computer Engineering programme at the University of of Castilla-La Mancha, Albacete Campus (Spain).

The university divides all the students into groups of similar size at the beginning of the course. For first-year students of Computer Engineering, there were two pre-existing groups (A, B), which were maintained throughout the course. The course lasted 15 weeks.

The two groups were scheduled to carry out the same three open activities (Activity 1, Activity 2 and Activity 3). Each activity was associated with a unit of content as follows:

- Activity 1: Mathematical induction. Unit 1: Numbers, sequences and series. One exercise.

[1] https://www.esiiab.uclm.es/plan.php?que=grado&curso=2020-21&idmenup=planestudios

Activity	Week	#Total Enrolled	#Total Participants	Group	PAM	#Enrolled	#Participants
Activity 1	Week 4	85	70	A	Individual	44	36
				B	Pairs	41	34
Activity 2	Week 9	85	64	A	Pairs	44	30
				B	In Threes	41	34
Activity 3	Week 14	85	59	A	In Threes	44	31
				B	Individual	41	28

Table 1: Execution data: date, PA Modality assignment and participants in each part of the study

- Activity 2: Optimisation problem and Taylor's formula. Unit 2: Differential Calculus. Two exercises.
- Activity 3: Area enclosed by two curves and area bounded by function and x-axis. Unit 3: Integral Calculus. Two exercises.

The activities were designed to be solved on paper. The students were required to upload a digital copy (photo) of their work into a Moodle workshop activity, which was also configured to be used to support the peer assessment process.

All the students carried out all the activities, following a within-subjects design. In this design, each student also completed in the UMAM questionnaire (behavioural intention construct) three times, that is, after participating in each of the assessment modalities. The order of PAMs was changed between groups in order to mitigate the risk of a treatment order effect. The final design is shown in Table 1.

For the peer assessment process, two documents were prepared in advance for each activity: a detailed solution of the activity and the rubric for its assessment. These documents were scheduled to be handed in to the students immediately before the beginning of the peer assessment. A time slot of 90 min was reserved for the PA of each activity. During that time slot, the researchers' previous experience suggested 9 evaluations as the maximum number to be carried out without losing quality, and 6 as the maximum number of peer reviews that students working on their own could perform without getting bored. It was thus decided that students participating in the individual modality would receive six assignments, students participating in the pairs modality would receive four assignments each (eight in total) and students participating in the in-threes modality would receive three assignments each (nine in total). Within each group (A and B), for the pairs and in-threes modality, the students were randomly divided into sub-groups. The activities that each individual/group were required to evaluate were also assigned randomly by the Moodle system.

The final structure of the empirical study was as follows:

- Week 2: Students complete a Consent Form to allow their data to be used as part of the study, and a Big Five personality questionnaire. A pilot task is proposed and both groups go through the peer evaluation process together with the instructor, who explains each phase of the process. In this way, the students become familiar with both the capabilities of the Moodle workshop tool and the use of rubrics for the PA.

- Week 4: Students carry out Activity 1 under exam conditions. In the following class, the students receive the solution and the rubric corresponding to Activity 1. Group A applies the individual PAM, while Group B applies the pairs PAM. Finally, all the groups complete the UMAM questionnaire for the assigned PAM.

- Week 9: Students carry out Activity 2 under exam conditions. In the following class, the students receive the solution and the rubric corresponding to Activity 2. Group A applies the Pairs PAM, and Group B applies the in-threes PAM. Finally, all the groups fill in the UMAM questionnaire for the assigned PAM.

- Week 14: Students carry out Activity 3 under exam conditions. In the following class, the students receive the solution and the rubric corresponding to Activity 3. Group A applies the in-threes PAM, and Group B applies the individual PAM. Finally, all the groups fill in the UMAM questionnaire for the assigned PAM.
- Week 15: All students fill in a questionnaire with a closed question where they choose between a peer assessment evaluation and an expert evaluation, and a set of open questions regarding their qualitative perceptions about the peer assessment evaluation method and the different PAMs.

4.4.1. Pandemic Contingency Plan

Due to the pandemic situation, the conditions under which the study was to take place were unclear. In order to allow the study to be implemented in case students were totally or partially self-isolating, an online system based on Microsoft Teams was put in place. In this way, for the collaborative peer assessment, each group was able to talk and share their screens to conduct the collaborative evaluation task, regardless of their physical location.

5. Execution of the study

This study took place in the first term of the 2020/2021 course (from September to December), in accordance with the planning. During the sessions of interest for the study, none of the students were self-isolating due to COVID-19.

Table 1 shows, for each activity, the week in which it took place (Week), the number of students enrolled on the course at that time (#Total Enrolled), and the number of students that performed the activity and uploaded it to Moodle (#Total Participants). Moreover, for each group, the table includes the peer assessment modality assigned to that group for that activity (PAM), the expected number of subjects in that group (#Enrolled) and the actual number of students that performed the activity (#Participants).

The PA data for each activity were gathered independently in each group 2 days after the execution of the activity, in the students' group class session.

For every activity upload, (a) between 3 and 6 peer assessments and (b) an expert assessment (performed by one of the researchers) were collected. A total of 1388 assessments were obtained (1195 from the students).

The same instructor/expert supervised all the peer assessment sessions. She was in charge of solving doubts and incidents, and preventing any kind of interaction between the different peer assessment individuals/teams. Therefore any instructor bias, if present, can be assumed to have equally affected the two groups.

6. Data analysis

The data were analysed using a variety of open source software tools and libraries[2]. A standard normalisation (S_n) has been applied to the input instances (Eq. 1).

$$S_n(y) = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=0}^{n}(y_i - \bar{y})}} \tag{1}$$

The mean age of participants was 18.

6.1. Distribution of variables

Figures 2 and 3 present the combined violin and box plot graphics showing the distribution of all the variables included in the study.

6.2. Machine Learning Analysis

To explore the relationship between Personality and BI, we follow the methodology showed in Fig. 4 which consists of testing a set of selected ML algorithms

[2] Several Python routines were developed by one of the authors. The main Python libraries used include scikit-learn v0.24.0, xgboost v0.90 and catboost 0.24.0 to implement ML algorithms, SHAP library v0.37 to analyse XAI-ML aspects, and plotnine v0.7.1 and ggplot2 v3.3.2 (statistical language R) to build plots.

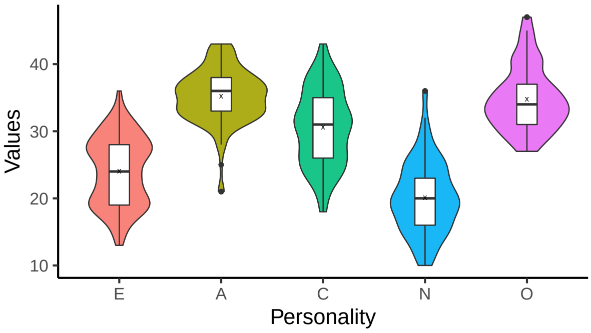

Figure 2: Combined violin and box plot showing the distribution of the Personality variables.

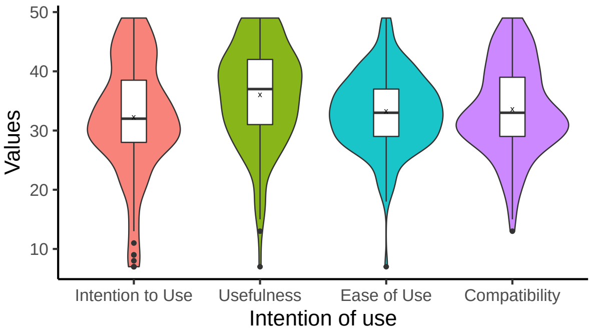

Figure 3: Combined violin and box plot showing the distribution of the Behavioural Intention variables.

belonging to different ML learning styles using cross-validation to choose the best one, and applying post-hoc explainability tools to obtain information regarding the impact of the model's characteristics, both at the global and individual level. The set of algorithms tested is:

455

- *Baseline*: This algorithm computes the average of the outcome without taking into account the predictors (independent variables).
- *Decision tree* (Breiman, 2017): It predicts the value of a sample based on simple learning rules in a hierarchical manner. The tree is built from the

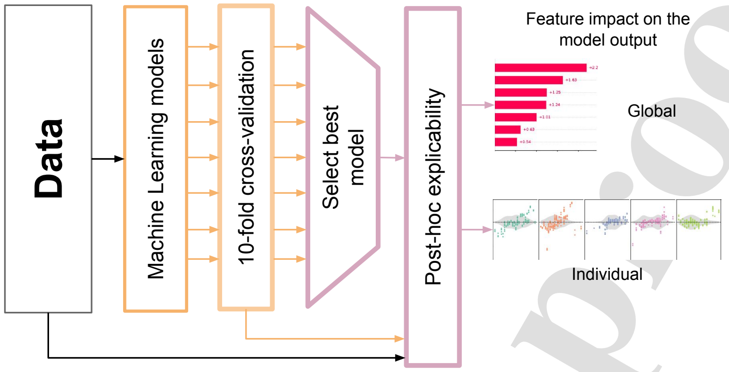

Figure 4: Diagram of the ML-based methodology applied to extract relevant information.

training samples, considering only one feature per rule.

- *Random Forest* (Breiman, 2001): Build multiple decision trees and combine all the individual predictions to obtain a final, more robust behavior.

- *AdaBoost (Adaptive Boosting)* (Freund & Schapire, 1997): The algorithm constructs several linear regressors in the training phase. When it predicts a new sample it does so by considering all linear predictions weighted with a confidence value learned during the training process.

- *XGBboost* (Chen & Guestrin, 2016) and *CatBoost* (Dorogush et al., 2018): These algorithms are based on boosting applied on decision trees where the optimization is performed using derivable cost functions together with gradient descent as is done in neural networks. Each algorithm implements the optimizations differently. They have obtained good results in open challenges.

- *Support vector machine* (Cortes & Vapnik, 1995): It is based on two steps. First, the original data space is mapped into another, usually higher dimensional, data space. Secondly, it tries to find a linear separation in the resulting space.

- *Neural Network (Multilayer Perceptron)* (Hinton, 1990): The traditional neural network architecture is used where all layers are fully connected.

- *Nearest Neighbours* (Cover & Hart, 1967): It calculates a prediction value based on the k (parameter) nearest samples of the training set and calculates a final prediction based on the proximity of the neighbours according to the Euclidean distance. We have set k to values 1, 3, 5, 7 and 9.

All the algorithms described have been used with the default parameters provided by the software.

6.2.1. Selection of the best model

To validate the models trained by the ML algorithms, we applied, as is usual in this field, the technique of k-fold cross-validation. It consists of creating a k fixed number of partitions - the most common value being 10 -, and using one as a test and the rest for training. The process is repeated in order to use each partition once as a test.

Then, to compare the results of the trained models, we chose the square root-mean-square error (RMSE) metric (Pentreath, 2015), whose formula is defined in equation 2. In this formula, y and $\hat{y}$ are vectors of size n, y contains the true values and $\hat{y}$ contains the prediction values. This metric is widely used in regression problems, and two of its characteristics are that (a) it more heavily penalises larger differences between the real and the predicted value (residuals), and (b) when applying the square root, it maintains the same units of measure.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \tag{2}$$

Figure 5 shows the average RMSE of the 10-fold cross-validation results for the five personality predictors under study. We obtained the best average results for the Random Forest and XGBoost algorithms; the worst results were those of the nearest neighbour with a neighbour, the decision tree and artificial neural networks[3].

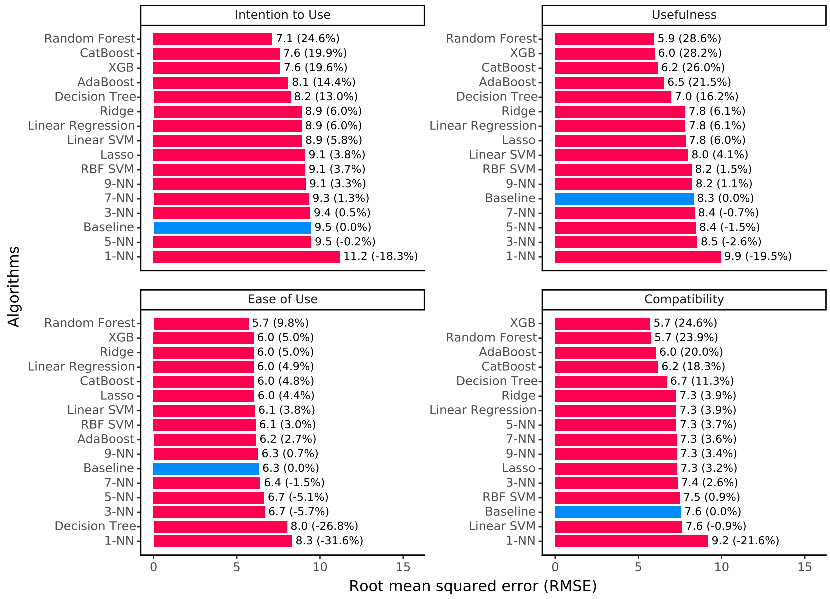

Figure 5: Average RMSE of the 10-fold cross-validation results for the five personality predictors using several machine learning algorithms. The lower the error values, the better. The numbers at the end of the bars indicate the RMSE error, and the percentage of the relative improvement from the baseline is shown between brackets, $(error(baseline) - error(algorithm))/error(baseline)$. The baseline results are highlighted in blue.

Figure 6 graphically represents a pairwise comparison of significance between the errors (RMSE) made by the algorithms considered for prediction after applying the 10-fold cross-validation technique. The method used is the Wilcoxon signed-rank test (Stapor et al., 2021; Wilcoxon, 1945) which is non-parametric and does not require the normal distribution of the data. It can be seen that the Random Forest algorithm significantly outperforms the other algorithms (green

[3] We decided not to show the results of the neural networks in the figure as they were the worst, far below the baseline. This phenomenon is arguably explained by the lack of training samples, which leads to underfitting.

bullets) for the four BI outcomes (Intention to Use, Usefulness, Ease of Use and 510 Compatibility). This means it is a good algorithm to build a model to explain the behaviour of the BI variables based on the personality variables, and was thus chosen for the explainability tests.

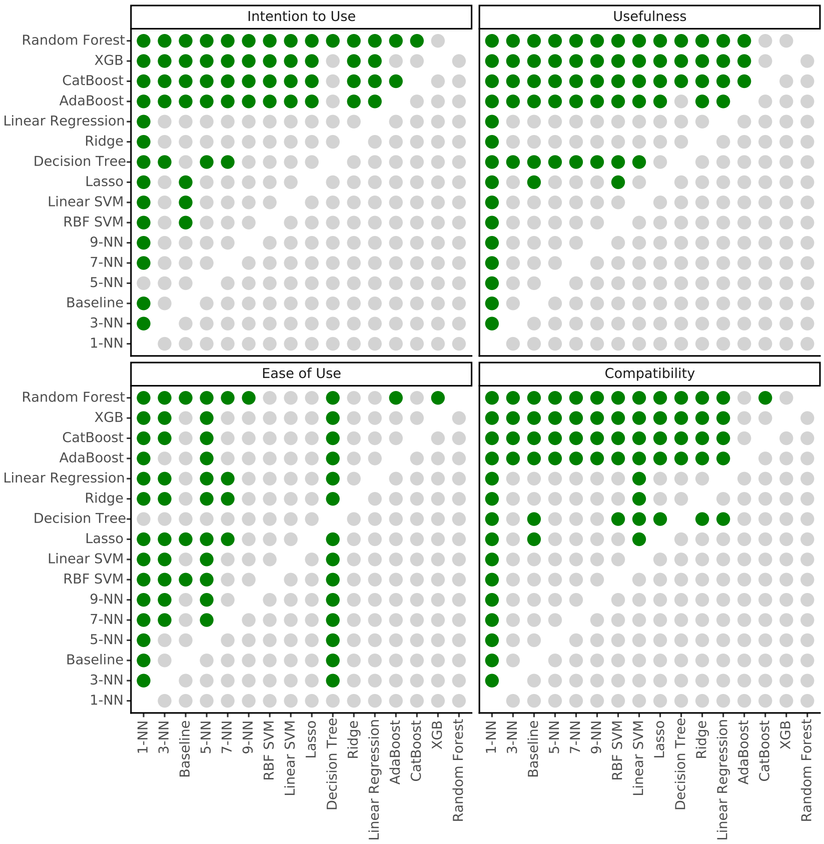

Figure 6: Pairwise comparison of significance between the RMSE values obtained from the 10-fold cross-validation applying the Wilcoxon signed-rank test. Green bullets indicate that the row algorithm is significantly better than the column algorithm.

6.2.2. Explanation of the Machine Learning model

Among the variety of ML algorithms used, there are some that are intrinsically explainable (Arrieta et al., 2020) such as those related to linear regression, neighbourhood (k nearest neighbours), or those based on a single decision tree, but there are others that are difficult to explain and some publications even refer to them as "black boxes" (Doornenbal et al., 2021; Fung et al., 2021). The latter algorithms are those that use multiple decision trees, support vector machines or neural networks. The Random Forest algorithm falls under this latter category. Fortunately, there are two main approaches that can be used to coherently explain these "black boxes" model predictions. The first is based on making permutations (Breiman, 2001) on the values of each predictor (independent variable) to compare their variability and that of their predictions (dependent variable). In this way, we can estimate the importance of the predictors of an already trained model. The second approach, which is the one that has been followed in this paper, is based on the construction of a new linear model that explains the complex model already trained. If Shapley's values (Roth, 1988) are used for this second approach, it is more accurate than the first. In essence, a Shapley value represents the average of a participant's expected marginal contribution after considering all possible combinations with the rest. These values are based on game theory and provide a solution that equitably distributes benefits and costs among participants. This approach is often used in situations where each participant contributes unequally. In addition, this method ensures local accuracy, missingness and consistency. Recent advances in this approach are explained by Lundberg & Lee (2017a,b), who present an extension of Shapley's values. This extension allows for a unified approach to explaining the predictions made by any trained model, and also allows individual weights to be calculated for each sample to extract individual, group or global explanations. The SHAP (SHapley Additive exPlanations) (Lundberg, 2019) tool allows these advanced techniques to be utilised.

Fig. 7 summarises the complete process. On the one hand, the data is used to train an ML model to make predictions, and on the other hand, the data and the trained ML model are used to build a new explainer model (using a locally applied linear approach) that explains why the predictions are made.

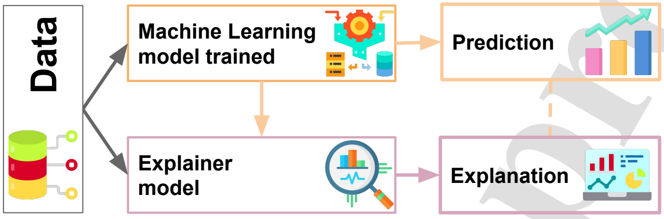

Figure 7: General scheme of the post-hoc explicability of the machine learning models.

We next present the answers to the research questions according to the explainer model generated.

6.2.3. RQ1: Impact of personality variables on Behavioural Intention towards the use of a peer assessment evaluation process

Fig. 8 shows the average impact of the five personality variables (Neuroticism, Conscientiousness, Extraversion, Openness to Experience and Agreeableness) on the four BI sub-components (Intention to Use, Usefulness, Ease of Use and Compatibility).

This figure shows how Agreeableness is the best predictor of Usefulness and Ease of Use (the higher the Agreableness score, the higher the Usefulness and Ease of Use perceptions). However, for Compatibility, Extraversion is the best predictor (the higher the student's Extraversion score, the higher the perceived compatibility of the student with the technique), while Neuroticism is the best predictor of the Intention to Use in the peer assessment evaluation process (the higher the Neuroticism score, the higher the intention expressed by the student to use the technique in the future if given the chance).

In order to further explore the individual relationships between personality and BI variables, Fig. 9 shows a grid plot. For each combination, the shape of the relationship is depicted. In this figure, we can see that some personality

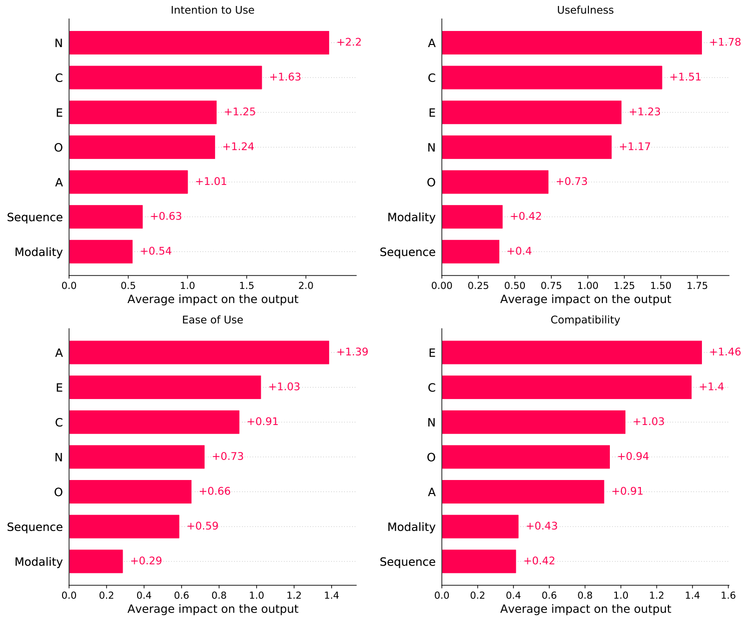

Figure 8: Average impact of Personality variables on BI variables.

variables have a more or less linear relationship with respect to the BI variables. This is the case of the Consciousness variable, which shows a more or less linear relationship with respect to Intention to Use, Usefulness and Compatibility. Others, on the contrary, show an inverse linear relationship. For example, Neuroticism shows a slightly inverse linear relationship with respect to Ease of Use. Finally, for other combinations there appears to be no relationship (see e.g. graph in the intersection of Openness to Experience and Usefulness).

Cohort analysis. Our next analysis goes one step further and checks whether the importance of the predictors varies when groups (cohorts) are formed. This additional analysis allows better decisions to be made and measures to be taken to empower students with certain personality profiles. To perform this search for alterations in general patterns in certain groups, we launched an automatic process to form the cohorts, using a maximum number of 3. This

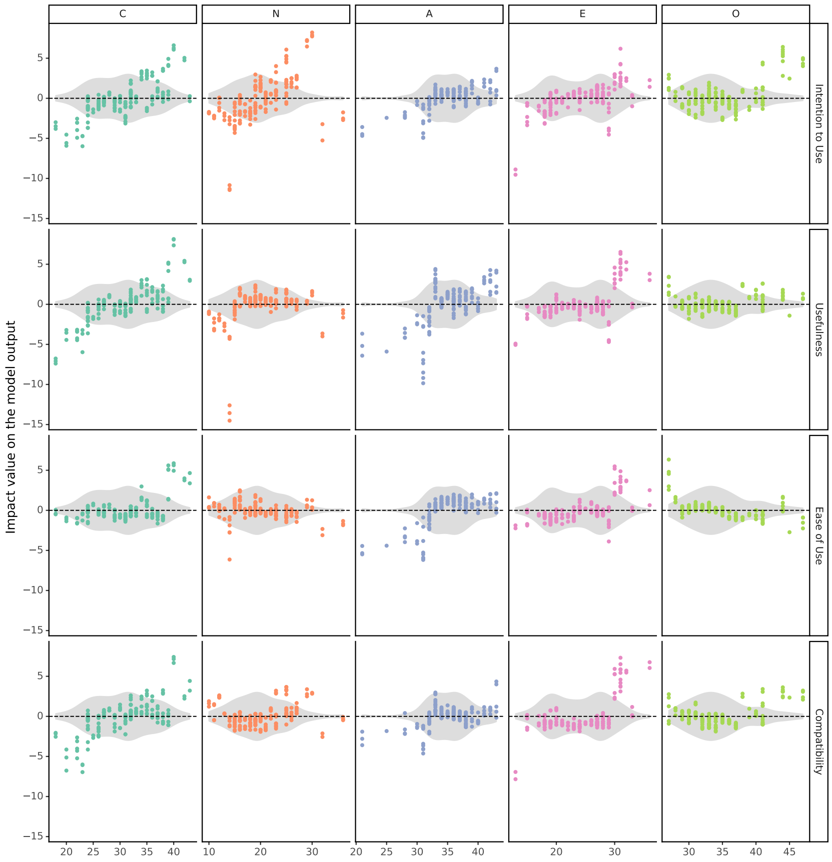

Figure 9: Individual impact according to the output model of Personality variables on BI variables. The density of samples in each area is plotted on the horizontal axis. The grey shading next to the points represents the overlapping violin plot. The average impact value, together with the ranking position of the Intention to Use variable, are indicated at the bottom of each sub-figure.

number corresponds to the traditional high, medium and low levels for each variable, and it ensures that subsets are not generated with too small a number of samples. The main advantage of the automatic approach is that it searches 580 for optimal cohorts according to logical criteria, which in turn allows for the discovery of rules that, a priori, would be difficult to find.

Output	cohort	M	SD	count	(1)	(2)	(3)
Intention to Use	N < 18.5 (1)	29.3	9.9	61	-		*
Intention to Use	N >= 18.5 & C < 31.5 (2)	30.7	7.9	52		-	*
Intention to Use	N >= 18.5 & C >= 31.5 (3)	37.8	8.7	46	*	*	-
Usefulness	A < 32.5 (1)	27.9	6.5	36	-	*	*
Usefulness	A >= 32.5 & N < 14.5 (2)	36.1	8.0	10	*	-	*
Usefulness	A >= 32.5 & N >= 14.5 (3)	42.2	5.4	113	*	*	-
Ease of Use	A < 31.5 (1)	27.7	5.8	20	-	*	*
Ease of Use	A >= 31.5 & C < 38.5 (2)	33.5	6.0	126	*	-	*
Ease of Use	A >= 31.5 & C >= 38.5 (3)	39.1	4.6	13	*	*	-
Compatibility	E < 29.5 & C < 26.5 (1)	28.6	6.6	36	-	*	*
Compatibility	E < 29.5 & C >= 26.5 (2)	33.8	7.0	99	*	-	*
Compatibility	E >= 29.5 (3)	40.1	6.0	24	*	*	-

Table 2: Average and number of students on the output predictor value using cohort rules. Differences significant at 95% between the different rules (1), (2) or (3) according to the output variable are marked with an *.

The problem of automatic formation of cohorts is similar to the construction of a decision tree, where it is necessary to find, in an ordered manner, subsets of samples whose size is increasingly smaller and which are organised to correctly predict the output variable. For this purpose, we used the utility of the SHAP tool based on the Gini impurity (eq. 3) of the data in a given set.

$$Gini(y) = 1 - \sum_{i=0}^{n} P(y_i)^2 \tag{3}$$

where y is a vector with n elements and $P(y_i)$ denotes the probability of the element y_i being misclassified. The process of dividing a set into subsets stops when the set maximum -3 in our case- is reached.

Table 2 shows how people scoring below 18.5 in Neuroticism exhibit a lower Intention to Use a peer assessment approach, while the highest Intention to use is provided by participants scoring both high in Neuroticism and in Consciousness. Regarding Usefulness, students scoring high in Agreeableness and Neuroticism are more likely to regard peer assessment as a useful evaluation strategy. Additionally, peer assessment is regarded as easier to apply by stu-

Test	Preference	Median E	Median A	Median C	Median N	Median O
Evaluation	expert	24	36**	31	19**	35
	peers	24.5	33.5**	32	21.5**	33.5
Peer Modality	individual	24	35	29	16**	36.5**
	pairs	21	34	31	21**	33**
	in-threes	26	36	31	20**	34**

Table 3: Median values of personality predictors according to their *evaluation preference* and their *peer modality preference*. The range for all the variables is [7,49]. Significant differences at 95% are indicated by *, and at 99% by **.

dents scoring high in Agreeableness and Consciousness. Lastly, compatibility is higher for participants whose level of extraversion is also higher.

Statistical analysis of Evaluation Preference. Finally, the Evaluation Preference variable, measured through the question Which method of assessment do you prefer?, with two possible answers (Peer assessment/Teacher assessment), and given the fact that the distribution of the E A, C, N, and O variables was not normal, was analysed with five independent U Mann-Whitney tests (Mann & Whitney, 1947). Table 3, first row (Evaluation) shows both the descriptive statistics (medians) of the five variables (E, A, C, N, O) for the two groups (students preferring an expert evaluation and students preferring a peer evaluation), and whether their differences are significant. Distributions for the scores for the two groups were similar, as assessed by visual inspection. Only A (U = 3601.05, p = 0.001) and N (U = 3627.5, p = 0.0009) are significantly different between the two groups, with students scoring higher on Agreeableness and lower on Neuroticism preferring expert assessment.

6.2.4. RQ2: Impact of PAM on BI regarding the use of a peer assessment evaluation process

Additionally, to test whether the PAM being applied or the order in which it was being applied had an effect on the BI sub-components, these were also included in the ML model as predictors (see Sequence and Modality bars, in

Fig. 8). This figure shows that all the personality variables have a far greater effect than either Sequence or Modality on the BI sub-components.

Statistical analysis of Modality Preference. Lastly, the Modality Preference variable, measured through the question 'Which PAM do you prefer?', with three possible answers (Individual/In pairs/In threes), due to the lack of normality of the distribution of the personality variables, was also analysed using non-parametric tests.

A Kruskal-Wallis H test (Kruskal & Wallis, 1952) was conducted to determine if there were differences among the three modality preference groups ("individual", "pairs", and "in-threes") regarding their level of E, A, C, N, and O . Distributions of all the personality variables were similar for all groups, as assessed by visual inspection of a boxplot.

As can be seen in Table 3 (Peer Modality row), only N ($\chi^2(2) = 15.54$, $p = 0.0004$) and O ($\chi^2(2) = 10.27$, $p = 0.006$) were statistically significantly different between the different levels of the modality preference group. On the one hand, students scoring higher on N prefer collaborative modalities, with the subsequent post-hoc analysis showing significant differences occurring between individual (Mdn=16) and pairs (Mdn=21) (p=0.006), and individual (Mdn=16) and in-threes (Mdn=20) (p<0.0001). On the other hand, students scoring higher on O prefer the individual modality, with the subsequent post-hoc analysis showing significant differences occurring between individual (Mdn=36.5) and pairs (Mdn=33) (p=0.0009), and between individual (Mdn=36.5) and in threes (Mdn=34) (p=0.0009). In both cases, the differences between the pairs and in-threes modalities were non-significant.

7. Discussion

Personality traits can be useful determinants of learners' perceptions and beliefs. The current study focuses on the impact of personality on the behavioural intention to use an evaluation strategy that shifts the focus from the instructor to the student. Not only is this evaluation strategy more cost-and time-efficient, but it also helps to overcome some of the challenges posed by traditional forms of evaluation in online and blended education environments - as has been the case during the current pandemic situation -. The outcomes reported are therefore relevant to a wide range of educators and technologists interested in implementing a shift in focus in their classrooms. Our findings show how personality impacts a student's behavioural intention to use a peer assessment strategy, above and beyond the particular peer assessment modality chosen by the instructor (see Fig. 8 and Table 2):

- Agreeableness and Conscientiousness are the best positive predictors for Usefulness
- Agreeableness and Extraversion are dominant positive predictors for Ease of Use
- Compatibility is best predicted by Extraversion and Conscientiousness (also with a positive impact)
- Intention to Use is best predicted by Neuroticism and Conscientiousness (positive relationship)

It was earlier stated that Agreeableness concerns the control of frustration, meaning that agreeable students are perhaps more likely to recognise the usefulness and ease of use of the peer assessment evaluation technique in their study environment. In this sense, our results are consistent with the conclusions of the related literature (Benlian & Hess, 2010; Devaraj et al., 2008; Rivers, 2021; Terzis et al., 2012).

Regarding Neuroticism, it seems that more insecure, nervous persons, who, at the same time, are highly achievement-oriented (highly conscious), are much more willing to adopt, if given the option, a peer assessment evaluation process, perhaps under the impression that they will be judged more benevolently - as was stated by some students in the open questions at the end of the experiment -. Moreover, the highly structured evaluation process involved in peer assessment may play a role in its higher regard by students with higher Neuroticism scores.

These two facts counteract in the peer assessment context the inherent negative emotions that people scoring high in Neuroticism are expected to feel when they come up against changes (Terzis et al., 2012).

The significant role of Extraversion in Compatibility draws attention to the specific behavioral demands of evaluating their peers as opposed to letting the expert evaluate everybody. Extraversion represents a tendency to be energetic, sociable, and assertive, and to experience positive thoughts and emotions. Within learning situations that demand open interactions and exchanges through the sharing of opinions with teachers or peers - as is the case with collaborative peer assessment modalities -, extraverted students are at a natural advantage due to their being inherently social, outgoing, and cooperative (Lounsbury et al., 2003). In addition, peer assessment gives cues to these students indicating that they are involved in a learning community, and this sense of belonging is particularly appealing for highly extraverted people. The fact that Extraversion is not significantly different across the three PAMs (see Table 3) suggests that what matters to extraverts is participating in the PA, regardless of the specific PAM.

Furthermore, we coincide with Rivers (2021) in that the strong positive effect of Conscientiousness on all the BI variables (see Fig. 8) points toward its dominant position as the most consistent predictor not only of academic achievement (Kappe & Van Der Flier, 2012) but also of less resistance to innovations in educational processes. Conscientious students are careful and responsible, and they exhibit a high level of performance and a strong sense of purpose and will. They are achievement-oriented, and are usually good students, and so their confidence in their own ability and the belief that they may obtain better grades when assessed by peers may be a strong motivation for their positive perceptions regarding peer assessment.

The differences found between our results and those of other studies also suggest that the personality dimensions play a different role on behavioural intention depending on the educational context and the characteristics of the evaluation method being considered, and therefore it is important to replicate the study whenever either one of them vary. In Terzis et al. (2012), where the object of study was a Computer Based self-Assessment tool (CBA), Agreeableness and Consciousness had a positive impact on perceived Ease of Use. Meanwhile, in our study, which features peer assessment evaluation in a face-to-face educational context, Extraversion has a bigger effect than Consciousness on this variable. One explanation may lie in the social interaction involved, which is lacking in CBA. Additionally, Terzis et al. (2012) found Neuroticism to have negative effects on Usefulness, while in our case the effect is positive. Some possible reasons, which we have already mentioned, are (a) the belief that they will be judged more benevolently by peers than by an instructor, and (b) the highly structured nature of this type of evaluation. Our study also detected a key role of Extraversion on Compatibility. This relationship was not considered in the original model used by Terzis et al. (2012).

Regarding the low (although positive) impact of Openness To Experience on the BI variables, our results differ from other findings in the related literature, which mark Openness as a significant correlate of both behavioural intentions (Terzis et al., 2012) and academic achievement (Asendorpf & Van Aken, 2003). In other fields, Openness To Experience also correlates with a higher ability of IT professionals to adapt to a technological innovation (Gallivan, 2004) and a higher preference to take responsibility for a whole process and not individual parts (Feldt et al., 2010). Given these antecedents, we would have expected it to also play a key role during the adoption of new evaluation methods. One possible explanation to this lack of impact is that the social and highly structured characteristics of the peer assessment process make other personality traits stand out, thus shadowing the effect of Openness To Experience.

Lastly, our results are not consistent with the literature highlighting a negative association between Neuroticism and Usefulness (Devaraj et al., 2008). According to our data, in the context of peer assessment, Neuroticism positively impacts on Usefulness.

This paper has also addressed the impact of the different peer assessment modalities on students' intention to adopt a peer assessment strategy. In this regard, our analysis shows how the particular modality (individual, pairs, in threes) has a lower impact on the behavioural variables than any of the personality traits considered in the study (see Fig. 8). This low impact of PAM is consistent with previous results Rico-Juan et al. (2022). One possible explanation for this result is that the students' perception of social interaction during the peer assessment process is achieved regardless of whether they perform the assessment individually or in groups. Also, the statistical analysis of the students' evaluation preference (see Table 3) shows that a higher neuroticism is associated with preferring an evaluation that is made in groups (pairs or in-threes) rather than individually. This may be explained by the perception that the evaluation responsibility is shared among the evaluators when a group modality is used. The other significant result shows how students scoring higher in Openness to Experience prefer an individual PAM. Again, this may be explained by their higher preference to take responsibility for a whole process and not individual parts (Feldt et al., 2010).

7.1. Practical considerations and concerns

The results of this study have implications for educators willing to successfully integrate peer assessment practices in their classroom. Not only do they need to become technically competent with the systems that support this type of evaluation - which, incidentally, also prepares them for situations such as the current pandemic -, but they also need to be prepared to handle potential adoption resistance and to understand how by including peer assessment practices they are impacting individuals with differing personality profiles. Regarding resistance, as it happens with other innovations in educational processes, low levels of Consciousness is the most consistent predictor of resistance to the introduction of peer assessment processes in the classroom. Regarding impact, the main finding is how introducing peer assessment can help students scoring high in Neuroticism. It is well known how these students tend to feel anxious when faced to an evaluation process, which may hamper their performance. Our results suggest that introducing peer assessment evaluation processes to complement other existing types of evaluation (tests, essays, oral expositions, etc.) may palliate this problem and help augment their positive feelings towards the evaluation process.

Additionally, our data shows how peer assessment is highly compatible with extroverted students, while introverts tend to reject it. To improve the introverts' feeling of compatibility with this technique, it would be advisable to devise instructional interventions that focus on introverted students and that are aimed at promoting their participation and interaction during the course. These interventions and the inclusion of peer assessment techniques are even more important in the case of online learning environments, where learner-to-learner interactions have been found to be one of the most significant indicators of achievement (Macfadyen & Dawson, 2012). Also, instructors should be aware of how the traditional modes of assessment used in these on-line courses (i.e. static forums, reports, readings, multiple choice quizzes, etc.) hamper extroverts, and how they should be oriented toward more creative real-time formats if their aim is to allow for sociable, lively and outgoing individuals to use their innate traits to work towards greater achievement outcomes (Rivers, 2021).

In summary, variety in terms of assessment methods is vital for inclusion. For this endeavour, the existing Learning Management Systems provide a good scaffolding to personalise the instructional design according to different parameters, students' personality being one of them.

As far as the PAM is concerned, we have detected a minor impact of the PAM applied on the students' BI. However, instructors should aim at a group evaluation modality (pairs, in-threes) if their objective is to further benefit students scoring high in Neuroticism.

Last, regarding the research approach applied, in this paper we have shown how, thanks to recent advances in explainability in Machine Learning models, many of the drawbacks associated with the application of ML for these types of problems have been overcome. It is now possible to explain why a model makes a certain prediction, and thus analyse the importance of the predictor variables involved in it. Therefore, the educational research community can now select the ML algorithm that best fits their data, while maintaining the capacity to explain the reason for the algorithm predictions, which is necessary to inform the design of educational interventions.

7.2. Limitations of the study

While the current study provides valuable data from an understudied context and includes several risk mitigation strategies, there remain some limitations which must be acknowledged. These mitigation strategies and limitations have been classified into four categories: internal, external, construct and conclusion (Cook et al., 1979), as shown below.

Threats to internal validity are concerned with the possibility of hidden factors that may provide alternative explanations for the result. All the students agreed to participate in the study, and it was confirmed that any absences on the day of the peer assessment (experimental mortality) were not due to the experiment itself. The interaction between subjects during the study was controlled. The same structure and the same instructors were used for the different treatments, and the order of treatments was randomised. The artefacts were carefully designed, the assignments were realistic and the entire design was carefully explained and tested through the use of a pilot study. However, the particular artefacts and evaluation guides used may have had an impact on the results. This risk is unfortunately unavoidable without new replications using different activities and evaluation guides.

Also, the sample size (85) is slightly inferior to that of the other related empirical study found in the literature (117 subjects in Terzis et al. (2012)), although in range with other peer assessment experimental designs that can be found in literature (see e.g. Chang & Lin (2020), with 60 subjects, Fang et al. (2022), with 100 subjects, or Panadero & Jonsson (2020) with 64 subjects) . In order to palliate this risk, a within-subjects design was devised. In within-subject designs, every participant provides repeated measures, making the study more cost effective.

Threats to external validity are concerned with the generalisation of the results. The participants within the current study cannot be used to generalise beyond the immediate context, which undermines the potential applicability of the results to other courses, universities or countries. Additionally, the study relies on self-reported measures. Both cultural factors and cultural etiquette have a significant role in self-reported behaviours, which also hampers generalisability. In order to mitigate this risk, new replications in different courses, universities and cultural contexts are needed.

Threats to construct validity are related to the relationship between theory and observation. We mitigated this type of risk by using measurement instruments (UMAM-Q and BF) previously validated in the literature.

Finally, *threats to conclusion validity* are related to the relationship between the treatment and the outcome. The scheme used for the analysis applies a selection of the best algorithms belonging to different ML families together with a cross-validation technique to finally select the algorithm that minimises the error in its prediction. This algorithm poses no a priori restrictions to the distribution of the variables, and serves as a basis for applying explainability techniques to post-hoc models based on game theory, that guarantees local accuracy, missingness and consistency.

8. Conclusions and future work

Given the educational shift in process, where students are increasingly taking centre stage, this article examines the effect of personality traits on the acceptance of three different modalities of peer assessment in first-year students enrolled in a Computer Engineering programme in Spain. The study was devised to fill an empirical gap in the existing literature.

Peer assessment evaluation minimises the role of instructors during its execution, relies on peer to peer interaction, and computer systems tend to play a more important role than in traditional forms of assessment. Neurotic people tend to avoid face-to-face interactions, and tend to feel more anxious when they are evaluated, while extroverted people thrive on interactions, Agreeable people favour altruism, nurturance and trust, and conscious people feel academically secure and are high achievers. This may explain why Neuroticism, Consciousness, Agreeableness and Extraversion are the main drivers of different aspects of Behavioural Intention. It is hoped that the data provided can serve as a basis for future replications within other affected educational contexts.

The scheme followed for data analysis, based on explainable ML, is one of the most recent and advanced. Three advantages of this approach over traditional statistical analysis are: (a) it detects nonlinearities between variables, (b) it does not restrict the distribution of the data (e.g. normal, log-normal, etc.) and (c) it allows us to obtain patterns or regularities in the form of rules, groups or criteria. In our case, the Random Forest algorithm made the best predictions and was therefore chosen to apply the explainability techniques.

As future work, we plan to replicate the study in different courses and universities. We also plan to include self-efficacy beliefs and course achievement as additional variables of the model, in order to assess how they interact with personality and academic performance. Finally, we need to dive into how taking into account the students' profiles during the instructional design (personality, motivation, cultural context, etc.) can maximize the social value of education, and how it helps to promote a 'sustainable educational system'.

Acknowledgements

We would like to thank the students who kindly agreed to participate in our study. This work has been partially funded by the the University of Alicante's Redes-I^{3}CE de investigación en docencia universitaria del Instituto de Ciencias de la Educación (REDES-I3CE-2020-5069), by the EU Erasmus+ Programme (EduTech (609785-EPP-1-2019-1-ES-EPPKA2-CBHE-JP) and SkoPS (2020-1-DE01-KA226HE-005772) projects), by the Spanish Ministry of Science and Innovation (Access@IoT (PID2019-111196RB-I00) project), by the GVA (AICO/2020/143) project, and by the UCLM group cofinanced with ERDF funds (research grant 2021-GRIN-30993). This article reflects the views only of the authors, and the funding bodies cannot be held responsible for any use which may be made of the information contained therein.

References

Abu-Al-Aish, A., & Love, S. (2013). Factors influencing students acceptance of m-learning: An investigation in higher education. *International Review of Research in Open and Distributed Learning*, *14*, 82–107. doi:10.19173/irrodl.v14i5.1631.

Adachi, C., Hong-Meng Tai, J., & Dawson, P. (2018). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment and Evaluation in Higher Education*, *43*, 294–306. doi:10.1080/02602938.2017.1339775.

Agarwal, R., & Prasad, J. (2000). A field study of the adoption of software process innovations by information systems professionals. *IEEE Transactions on Engineering Management*, *47*, 295–308. doi:10.1109/17.865899.

Agyei, J., Sun, S., Abrokwah, E., Penney, E. K., & Ofori-Boafo, R. (2020). Mobile banking adoption: Examining the role of personality traits. *SAGE Open*, *10*, 2158244020932918. doi:10.1177/2158244020932918.

An, D., & Carr, M. (2017). Learning styles theory fails to explain learning and achievement: Recommendations for alternative approaches. *Personality and Individual Differences*, *116*, 410–416. doi:https://doi.org/10.1016/j.paid.2017.04.050.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

Asendorpf, J. B., & Van Aken, M. A. (2003). Personality–relationship transaction in adolescence: Core versus surface personality characteristics. *Journal of personality*, *71*, 629–666. doi:https://doi.org/10.1111/1467-6494.7104005.

Avia, M., Sanz, J., Sánchez-Bernardos, M., Martínez-Arias, M., Silva, F., & Graña, J. (1995). The five-factor modelII. Relations of the NEO-PI with other personality variables. *Personality and Individual Differences*, *19*, 81–97. doi:https://doi.org/10.1016/0191-8869(95)00007-S.

Bagozzi, R. P., Davis, F. D., & Warshaw, P. R. (1992). Development and test of a theory of technological learning and usage. *Human relations*, *45*, 659–686. doi:https://doi.org/10.1177/001872679204500702.

Balijepally, V., Mahapatra, R., & Nerur, S. P. (2006). Assessing personality profiles of software developers in agile development teams. *Communications of the Association for Information Systems*, *18*, 4. doi:10.17705/1CAIS.01804.

Barrick, M. R. (2005). Yes, personality matters: Moving on to more important matters. *Human performance*, *18*, 359–372. doi:https://doi.org/10.1207/s15327043hup1804_3.

Bates, D., & Watts, D. (2007). *Nonlinear regression analysis and its applications*. Wiley.

Benlian, A., & Hess, T. (2010). Does personality matter in the evaluation of ERP systems? Findings from a conjoint study. In *18th European Conference on Information Systems, ECIS 2010*.

Bergold, S., & Steinmayr, R. (2018). Personality and intelligence interact in the prediction of academic achievement. *Journal of Intelligence*, *6*, 27. doi:10.3390/jintelligence6020027.

Bhagat, K. K., Wu, L. Y., & Chang, C.-Y. (2019). The impact of personality on students' perceptions towards online learning. *Australasian Journal of Educational Technology*, *35*. doi:https://doi.org/10.14742/ajet.4162.

Blumberg, M., & Pringle, C. D. (1982). The missing opportunity in organizational research: Some implications for a theory of work performance. *Academy of management Review*, *7*, 560–569. doi:https://doi.org/10.5465/amr.1982.4285240.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32. doi:https://doi.org/10.1023/A:1010933404324.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Caprara, G. V., Vecchione, M., Alessandri, G., Gerbino, M., & Barbaranelli, C. (2011). The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology*, *81*, 78–96. doi:https://doi.org/10.1348/2044-8279.002004.

Casey, J. E., Pennington, L. K., & Mireles, S. V. (2021). Technology acceptance model: Assessing preservice teachers acceptance of floor-robots as a useful pedagogical tool. *Technology, Knowledge and Learning*, *26*, 499–514. doi:https://doi.org/10.1007/s10758-020-09452-8.

Chang, C., & Lin, H.-C. K. (2020). Effects of a mobile-based peer-assessment approach on enhancing language-learners oral proficiency. *Innovations in Education and Teaching International*, *57*, 668 679. doi:10.1080/14703297.2019.1612264.

Chang, C.-Y., Lee, D.-C., Tang, K.-Y., & Hwang, G.-J. (2021). Effect sizes and research directions of peer assessments: From an integrated perspective of meta-analysis and co-citation network. *Computers & Education*, *164*, 104123. doi:https://doi.org/10.1016/j.compedu.2020.104123.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *CoRR*, *abs/1603.02754*.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* volume 351. Houghton Mifflin Boston.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, *20*, 273–297. URL: https://doi.org/10.1023/A:1022627411411. doi:10.1023/A:1022627411411.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27. doi:http://dx.doi.org/10.1109/TIT.1967.1053964.

Cruz, S., da Silva, F. Q., & Capretz, L. F. (2015). Forty years of research on personality in software engineering: A mapping study. *Computers in Human Behavior*, *46*, 94–113. doi:10.1016/j.chb.2014.12.008.

Curtis, B., Krasner, H., & Iscoe, N. (1988). A field study of the software design process for large systems. *Communications of the ACM*, *31*, 1268–1287. doi:https://doi.org/10.1145/50087.50089.

Dalvi-Esfahani, M., Alaedini, Z., Nilashi, M., Samad, S., Asadi, S., & Mohammadi, M. (2020). Students green information technology behavior: Beliefs and personality traits. *Journal of cleaner production*, *257*, 120406. doi:https://doi.org/10.1016/j.jclepro.2020.120406.

Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., & Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in Industry*, *123*, 103298. doi:10.1016/j.compind.2020.103298.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, (pp. 319–340). doi:https://doi.org/10.2307/249008.

De Raad, B., & Schouwenburg, H. C. (1996). Personality in learning and education: A review. *European Journal of personality*, *10*, 303–336. doi:https://doi.org/10.1002/(SICI)1099-0984(199612)10:5<303::AID-PER262>3.0.CO;2-2.

Devaraj, S., Easley, R. F., & Crant, J. M. (2008). Research NoteHow Does Personality Matter? Relating the Five-Factor Model to Technology Acceptance and Use. *Information Systems Research*, *19*, 93–105. doi:https://doi.org/10.1287/isre.1070.0153.

Dhini, A., Surjandari, I., Kusumoputro, B., & Kusiak, A. (2021). Extreme learning machine–radial basis function (ELM-RBF) networks for diagnosing faults in a steam turbine. *Journal of Industrial and Production Engineering*, (pp. 1–9). doi:https://doi.org/10.1080/21681015.2021.1887948.

Diéguez, M., Sepúlveda, S., & Cachero, C. (2012). UMAM-Q: An instrument to assess the intention to use software development methodologies. In *Information Systems and Technologies (CISTI), 7th Iberian Conference on* (pp. 1–6). IEEE.

Doornenbal, B. M., Spisak, B. R., & van der Laken, P. A. (2021). Opening the black box: Uncovering the leader trait paradigm through machine learning. *The Leadership Quarterly*, (p. 101515). doi:https://doi.org/10.1016/j.leaqua.2021.101515.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *CoRR*, *abs/1810.11363*. URL: http://arxiv.org/abs/1810.11363.

Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The Impact of Peer Assessment on Academic Performance: A Meta-analysis of Control Group Studies. *Educational Psychology Review*, (pp. 1–29). doi:https://doi.org/10.1007/s10648-019-09510-3.

Embarak, O. (2021). A New Paradigm Through Machine Learning: A Learning Maximization Approach for Sustainable Education. *Procedia Computer Science*, *191*, 445–450. doi:https://doi.org/10.1016/j.procs.2021.07.055.

Eysenck, H. J. (1994). *The Big Five or giant three: Criteria for a paradigm.* Lawrence Erlbaum Associates, Inc.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research*, *70*, 287–322. doi:10.3102/00346543070003287.

Fang, J.-W., Shao, D., Hwang, G.-J., & Chang, S.-C. (2022). From critique to computational thinking: A peer-assessment-supported problem identification, flow definition, coding, and testing approach for computer programming instruction. *Journal of Educational Computing Research*, . doi:10.1177/07356331211060470.

Feldt, R., Angelis, L., Torkar, R., & Samuelsson, M. (2010). Links between the personalities, views and attitudes of software engineers. *Information and Software Technology*, *52*, 611–624. doi:10.1016/j.infsof.2010.01.001.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of computer and system sciences*, *55*, 119–139. doi:https://doi.org/10.1006/jcss.1997.1504.

Fung, P. L., Zaidan, M. A., Timonen, H., Niemi, J. V., Kousa, A., Kuula, J., Luoma, K., Tarkoma, S., Petäjä, T., Kulmala, M. et al. (2021). Evaluation of white-box versus black-box machine learning models in estimating ambient black carbon concentration. *Journal of Aerosol Science*, *152*, 105694. doi:https://doi.org/10.1016/j.jaerosci.2020.105694.

Gallivan, M. J. (2004). Examining IT professionals' adaptation to technological change: the influence of gender and personal attributes. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *35*, 28–49. doi:https://doi.org/10.1145/1017114.1017119.

Gandomani, T. J., Zulzalil, H., Ghani, A. A., Sultan, A. B. M., & Sharif, K. Y. (2014). How Human Aspects Impress Agile Software Development Transition and Adoption. *International Journal of Software Engineering and its Applications*, *8*, 129–148. doi:10.14257/IJSEIA.2014.8.1.12.

Hardgrave, B. C., Davis, F. D., & Riemenschneider, C. K. (2003). Investigating Determinants of Software Developers' Intentions to Follow Methodologies. *Journal of Management Information Systems*, *20*, 123–151. doi:10.1080/07421222.2003.11045751.

Hinton, G. E. (1990). Connectionist Learning Procedures. In *Machine Learning, Volume III* (pp. 555–610). Elsevier.

Jarillo-Nieto, P. I., Enríquez-Ramírez, C., & Sánchez-Herrera, R. A. (2015). Identificación del factor humano en el seguimiento de procesos de software en un medio ambiente universitario. *Computación y Sistemas*, *19*, 577–588.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, *3*, 114–158.

Kampenes, V. B., Dybå, T., Hannay, J. E., & Sjøberg, D. I. (2007). A systematic review of effect size in software engineering experiments. *Information and Software Technology*, *49*, 1073–1086. doi:10.1016/j.infsof.2007.02.015.

Kappe, R., & Van Der Flier, H. (2012). Predicting academic success in higher education: whats more important than being smart? *European Journal of Psychology of Education*, *27*, 605–619. doi:https://doi.org/10.1007/s10212-011-0099-9.

Könings, K. D., van Zundert, M., & van Merriënboer, J. J. (2019). Scaffolding peer-assessment skills: Risk of interference with learning domain-specific skills? *Learning and Instruction*, *60*, 85–94. doi:10.1016/j.learninstruc.2018.11.007.

Koufaris, M. (2002). Applying the Technology Acceptance Model and Flow Theory to Online Consumer Behavior. *Information systems research*, *13*, 205–223. URL: http://www.jstor.org/stable/23011056.

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American statistical Association*, *47*, 583–621. URL: http://www.jstor.org/stable/2280779.

Lai, P. C. (2017). The literature review of technology adoption models and theories for the novelty technology. *JISTEM-Journal of Information Systems and Technology Management*, *14*, 21–38.

Lazar, I. M., Panisoara, G., & Panisoara, I. O. (2020). Digital technology adoption scale in the blended learning context in higher education: Development, validation and testing of a specific tool. *PloS one*, *15*, e0235957. doi:https://doi.org/10.1371/journal.pone.0235957.

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, *45*, 193–211. doi:10.1080/02602938.2019.1620679.

Li, H., Xiong, Y., Zang, X., L. Kornhaber, M., Lyu, Y., Chung, K. S., & K. Suen, H. (2016). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, *41*, 245–264. doi:10.1080/02602938.2014.999746.

Li, H., Xu, J., Chec, J., & Fan, Y. (2015). A Reliability Meta-Analysis for 44 Items Big Five Inventory: Based on the Reliability Generalization Methodology. *Advances in Psychological Science*, *23*, 755. doi:https://journal.psych.ac.cn/xlkxjz/EN/10.3724/SP.J.1042.2015.00755.

Liu, B. (2020). New Technology Application in Logistics Industry Based on Machine Learning and Embedded Network. *Microprocessors and Microsystems*, (p. 103596). doi:https://doi.org/10.1016/j.micpro.2020.103596.

Liu, N.-F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher education*, *11*, 279–290. doi:10.1080/13562510600680582.

Lounsbury, J. W., Sundstrom, E., Loveland, J. M., & Gibson, L. W. (2003). Intelligence,Big Five personality traits, and work drive as predictors of course grade. *Personality and individual differences*, *35*, 1231–1239. doi:`https://doi.org/10.1016/S0191-8869(02)00330-6`.

Lowther, D. L., Bassoppo-Moyo, T., & Morrison, G. R. (1998). Moving from computer literate to technologically compotent: The next educational reform. *Computers in Human Behavior*, *14*, 93–109. doi:`https://doi.org/10.1016/S0747-5632(97)00034-4`.

Lundberg, S. (2019). SHAP (SHapley Additive exPlanations). `https://github.com/slundberg/shap`.

Lundberg, S. M., & Lee, S.-I. (2017a). Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:1706.06060*, .

Lundberg, S. M., & Lee, S.-I. (2017b). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

Macfadyen, L. P., & Dawson, S. (2012). Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. *J. Educ. Technol. Soc.*, *15*, 149–163.

Mailizar, M., Burg, D., & Maulina, S. (2021). Examining university students behavioural intention to use e-learning during the COVID-19 pandemic: An extended TAM model. *Education and Information Technologies*, (pp. 1–21). doi:`https://doi.org/10.1007/s10639-021-10557-5`.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The annals of mathematical statistics*, (pp. 50–60). doi:`10.1214/aoms/1177730491`.

Martínez, Y., Cachero, C., & Meliá, S. (2013). MDD vs. traditional software development: A practitioners subjective perspective. *Information and Software*

Technology, *55*, 189–200. doi:https://doi.org/10.1016/j.infsof.2012.07.004.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, *60*, 175–215. doi:10.1111/j.1467-6494.1992.tb00970.x.

Menon, H. K. D., & Janardhan, V. (2021). Machine learning approaches in education. *Materials Today: Proceedings*, *43*, 3470–3480. doi:https://doi.org/10.1016/j.matpr.2020.09.566.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, *39*, 102–122. doi:10.1080/02602938.2013.795518.

O'Connor, R. V., & Yilmaz, M. (2015). Exploring the Belief Systems of Software Development Professionals. *Cybernetics and Systems*, *46*, 528–542. doi:10.1080/01969722.2015.1038483.

Panadero, E., Fraile, J., Fernández Ruiz, J., Castilla-Estévez, D., & Ruiz, M. A. (2019). Spanish university assessment practices: examination tradition with diversity by faculty. *Assessment and Evaluation in Higher Education*, *44*, 379–397. doi:10.1080/02602938.2018.1512553.

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review*, *30*. doi:https://doi.org/10.1016/j.edurev.2020.100329.

Panadero, E., Romero, M., & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*, 195–203. doi:https://doi.org/10.1016/j.stueduc.2013.10.005.

Pentreath, N. (2015). *Machine learning with spark*. Packt Publishing Ltd.

Persico, D., Manca, S., & Pozzi, F. (2014). Adapting the Technology Acceptance Model to evaluate the innovative potential of e-learning systems. *Computers in Human Behavior*, *30*, 614–622. doi:https://doi.org/10.1016/j.chb.2013.07.045.

Qu, W., Sun, H., & Ge, Y. (2021). The effects of trait anxiety and the big five personality traits on self-driving car acceptance. *Transportation*, *48*, 26632679. doi:https://doi.org/10.1007/s11116-020-10143-7.

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction*, *60*, 154–165. doi:https://doi.org/10.1016/j.learninstruc.2018.01.004.

Reinholz, D. (2016). The assessment cycle: a model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, *41*, 301–315. doi:10.1080/02602938.2015.1008982.

Rico-Juan, J. R., Cachero, C., & Macià, H. (2022). Influence of individual versus collaborative peer assessment on score accuracy and learning outcomes in higher education: an empirical study. *Assessment & Evaluation in Higher Education*, *47*, 570–587. doi:10.1080/02602938.2021.1955090.

Rico-Juan, J. R., Gallego, A.-J., & Calvo-Zaragoza, J. (2019). Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine learning. *Computers & Education*, *140*, 103609. doi:https://doi.org/10.1016/j.compedu.2019.103609.

Rivers, D. J. (2021). The role of personality traits and online academic self-efficacy in acceptance, actual use and achievement in Moodle. *Education and Information Technologies*, (pp. 1–26). doi:https://doi.org/10.1007/s10639-021-10478-3.

Roth, A. E. (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, *109*, 1049. doi:https://doi.org/10.1037/edu0000190.

Shen, B., Bai, B., & Xue, W. (2020). The effects of peer assessment on learner autonomy: An empirical study in a Chinese college English writing class. *Studies in Educational Evaluation*, *64*, 100821. doi:https://doi.org/10.1016/j.stueduc.2019.100821.

Stapor, K., Ksieniewicz, P., Garcia, S., & Woźniak, M. (2021). How to design the fair experimental classifier evaluation. *Applied Soft Computing*, (p. 107219). doi:https://doi.org/10.1016/j.asoc.2021.107219.

Šumak, B., Heričko, M., & Pušnik, M. (2011). A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types. *Computers in human behavior*, *27*, 2067–2077. doi:https://doi.org/10.1016/j.chb.2011.08.005.

Terzis, V., Moridis, C. N., & Economides, A. A. (2012). How students personality traits affect Computer Based Assessment Acceptance: Integrating BFI with CBAAM. *Computers in Human Behavior*, *28*, 1985–1996. doi:https://doi.org/10.1016/j.chb.2012.05.019.

To, J., & Panadero, E. (2019). Peer assessment effects on the self-assessment process of first-year undergraduates. *Assessment and Evaluation in Higher Education*, *44*, 920–932. doi:10.1080/02602938.2018.1548559.

Toala, G., Diéguez, M., Cachero, C., & Meliá, S. (2018). Evaluating the impact of developers personality on the intention to adopt model-driven web engineering approaches: An observational study. In *International Conference on Web Engineering* (pp. 3–16). Springer.

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In *Optimising new modes of assessment: In search of qualities and standards* (pp. 55–87). Springer.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Wang, Y., Li, H., Feng, Y., Jiang, Y., & Liu, Y. (2012). Assessment of programming language learning based on peer code review model: Implementation and experience report. *Computers & Education*, *59*, 412–422. doi:https://doi.org/10.1016/j.compedu.2012.01.007.

Wanner, T., & Palmer, E. (2018). Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback. *Assessment & Evaluation in Higher Education*, *43*, 1032–1047. doi:10.1080/02602938.2018.1427698.

Webster, B. J., & Yang, M. (2012). Transition, induction and goal achievement: first-year experiences of hong kong undergraduates. *Asia Pacific Education Review*, *13*, 359–368.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, *1*, 80–83.

Wu, J.-Y. (2021). Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement. *Computers & Education*, *163*, 104066. doi:https://doi.org/10.1016/j.compedu.2020.104066.

Yucel, R., Bird, F. L., Young, J., & Blanksby, T. (2014). The road to self-assessment: exemplar marking before peer review develops first-year students capacity to judge the quality of a scientific report. *Assessment & Evaluation in Higher Education*, *39*, 971–986. doi:10.1080/02602938.2014.880400.

Zheng, L., Zhang, X., & Cui, P. (2020). The role of technology-facilitated peer assessment and supporting strategies: a meta-analysis. *Assessment and*

Evaluation in Higher Education, *45*, 372–386. doi:10.1080/02602938.2019.1644603.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: