Vázquez-Ingelmo, A., García-Holgado, A., García-Peñalvo, F. J., Pérez-Sánchez, P., Antúnez-Muiños, P., Sánchez-Puente, A., Vicente-Palacios, V., Dorado-Díaz, P. I., & Sánchez, P. L. (2023). Are Textual Recommendations Enough? Guiding Physicians Toward the Design of Machine Learning Pipelines Through a Visual Platform. In H. L. Larsen, M. J. Martin-Bautista, M. D. Ruiz, A. T, G. Bordogna, & G. De Tré (Eds.), Flexible Query Answering Systems. 15th International Conference, FQAS 2023. Mallorca, Spain, September 5–7, 2023. Proceedings (pp. 247–255). Springer Nature. https://doi.org/10.1007/978-3-031-42935-4_20

Are Textual Recommendations Enough? Guiding Physicians Toward the Design of Machine Learning Pipelines Through a Visual Platform

Andrea Vázquez-Ingelmo[1](✉), Alicia García-Holgado[1], Francisco José García-Peñalvo[1], Pablo Pérez-Sánchez[2], Pablo Antúnez-Muiños[2], Antonio Sánchez-Puente[2], Víctor Vicente-Palacios[3], Pedro Ignacio Dorado-Díaz[4], and Pedro Luis Sánchez[5]

[1] GRIAL Research Group, Computer Science Department, Universidad de Salamanca, Salamanca, Spain
{andreavazquez,aliciagh,fgarcia}@usal.es
[2] CIBERCV and Biomedical Research Institute of Salamanca (IBSAL), University Hospital of Salamanca, Salamanca, Spain
{pperezsanc,pantunezm,asanchezpu}@saludcastillayleon.es
[3] Philips Clinical Science, Western Europe, Valencia, Spain
victor.vicente.palacios@philips.com
[4] Biomedical Research Institute of Salamanca (IBSAL) and University of Salamanca, Statistics Department, Salamanca, Spain
acho@usal.es
[5] University Hospital of Salamanca, CIBERCV, Biomedical Research Institute of Salamanca (IBSAL) and University of Salamanca, Cardiology Department, Salamanca, Spain
plsanchez@saludcastillayleon.es
https://ror.org/02f40zc51

Abstract. The prevalence of artificial intelligence (AI) in our daily lives is often exaggerated by the media, leading to a positive public perception while overlooking potential problems. In the field of medicine, it is crucial to educate future healthcare professionals on the advantages and disadvantages of AI and to emphasize the importance of creating fair, ethical, and reproducible models. The KoopaML platform was developed to provide an educational and user-friendly interface for inexperienced users to create AI pipelines. This study analyzes the quantitative and interaction data gathered from a usability test involving physicians from the University Hospital of Salamanca, with the aim of identifying new interaction paradigms to improve the platform's usability. The results shown that the platform is difficult to learn for inexperienced users due to its contents related to AI. Following these results, a set of improvements are proposed for the next version of KoopaML, focusing on reducing the interactions needed to create the pipelines.

Keywords: Information system · Medical data management · Artificial Intelligence · Health platform · HCI · Usability · SUS

1 Introduction

Artificial intelligence is present in our daily lives, however media coverage is not always realistic and exacerbates its capabilities [1, 2]. This media attention makes the public's perception of AI positive and overlooks the problems it can cause [3].

It is certain that AI is becoming increasingly present in our daily lives, and medicine is no exception [4]. For this reason, it is important to provide future medical students and healthcare professionals with adequate education in this regard [5]. And despite the fact that future doctors are not afraid of being replaced by AI [6], it is important to let them know its pros and cons [7]. It is also important for clinicians to be aware of the importance of using or creating models that are fair [8], ethical [9] and reproducible [10].

In order to train inexperienced users in all the above-mentioned points, the KoopaML platform was created [11–13]. The main objectives of this platform are (1) to provide a visual and intuitive interface and (2) to offer an educational AI experience. To develop the platform to the needs of inexperienced users, their feedback is necessary.

This work presents the quantitative and interaction analysis results of KoopaML of a usability study involving physicians from the University Hospital of Salamanca. The analysis of the results aims at identifying new interaction paradigms to solve the issues arisen during the usability test.

The rest of this work is organized as follows. Section 2 provides an overview of the KoopaML platform. Section 3 describes the methodology followed for the usability test and analysis. Section 4 presents the test results, while Sect. 5 discusses the results and proposes new methods to interact with the ML pipelines to address the encountered issues.

2 Background

KoopaML [11–13] is conceived as a platform with two main goals; (1) to ease and automate the generation and execution of ML pipelines, and (2) to offer a learning experience to non-expert users on the basics of ML while leveraging its benefits.

These goals are tackled through a graphical interface inspired in building blocks, in which users can add, connect, and execute ML tasks transparently, without programming expertise. Figure 1 shows a pipeline example in which a Logistic Regression algorithm is trained with an input dataset (in dark blue).

Although the platform has allowed the automatization of ML pipelines visually, it is still complex to address their design with no expertise. For these reasons, in previous works, we have included a new feature: a textual recommendation engine that yields information about the potential steps to take given the current state of the workspace (type of ML nodes included, current connections, etc.).

This recommendation engine was included in the workspace in the form of a modal box (Fig. 2, bottom), and the textual recommendations can be easily modified by privileged users (experts) to include new heuristics or explanations.

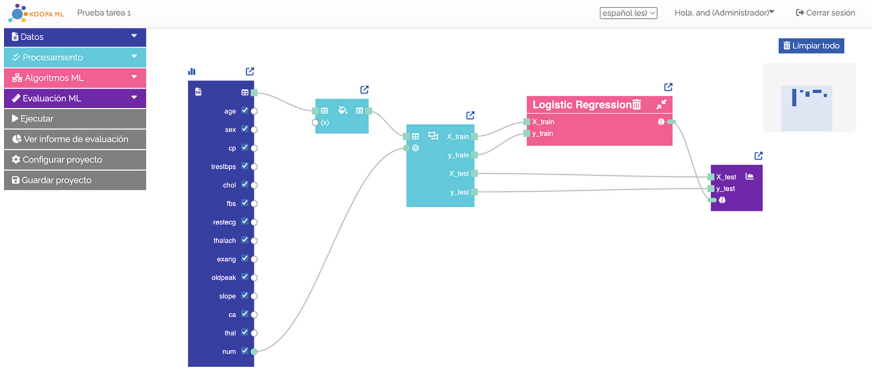

Fig. 1. The KoopaML platform (contents in Spanish).

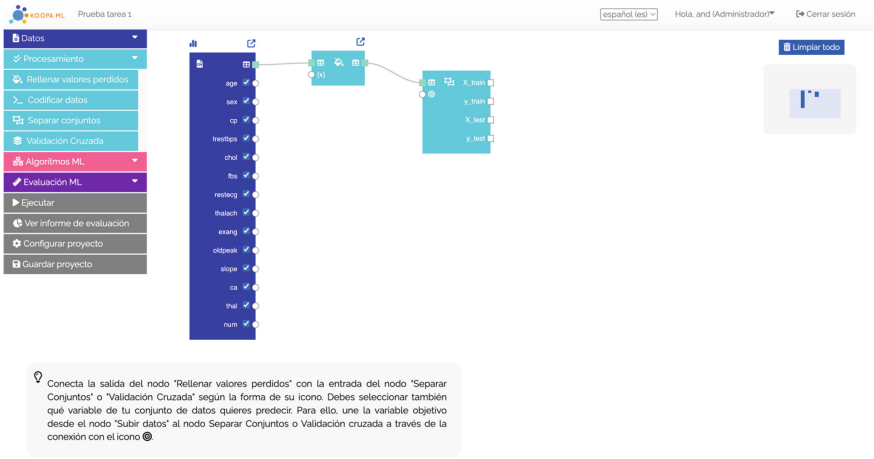

Fig. 2. Recommendation engine (contents in Spanish). In this case, the textual recommendation is explaining how to use the "Test/Train Split" node, and what is the goal of splitting the input dataset.

3 Methodology

A user testing was conducted to test the usability and find issues related to the user experience in KoopaML. The study was conducted with **8 physicians** (with low or no expertise in ML) by using the **think-aloud method** [14, 15], with the goal of analyzing the interactions performed by the users while using KoopaML.

Every participant was introduced to the tool and to the basic concepts of ML through the following video (contents in Spanish): https://www.youtube.com/watch?v=JeQrz2I20TY. The think-aloud method was complemented with a quantitative analysis of the perceived usability.

3.1 Interaction Analysis

The user interactions were captured through Hotjar (https://hotjar.com), a digital data analysis tool that allows the visualization of heatmaps and even recordings of the interactions carried out by the users during the testing. The analysis of interaction allows to better understand the decisions taken by the users while carrying out simple tasks in the platform.

3.2 Perceived Usability Evaluation

For the quantitative analysis, the System Usability Scale was selected as the instrument to assess the platform's perceived usability (SUS). The SUS questionnaire offers a practical, reliable, and valid [16, 17] method for rating a system's usability, and it can be used with different categories of systems [18].

The items of the questionnaires are positive and negative alternated and rated on a 1 to 5 Likert scale [19].

The instrument was implemented using a customized version of LimeSurvey (https://www.limesurvey.org), an Open-Source on-line survey web application.

The interpretation of the results is guided by the System Usability Scale benchmarks [20, 21] which allow SUS score comparisons and provide useful insights about the perceived usability of the system.

4 Results

The qualitative results provided beneficial insights into the current interface of KoopaML. The interaction heatmaps obtained from Hotjar enabled us to understand the parts of the interface that were more prone to interactions and the differences between users that had the textual recommendations enabled and those who did not.

Figure 3 shows one of these heatmaps. In this case, the participant did not have textual recommendations during the study. It is possible to observe that the participant spent more time navigating through the side menu than interacting with and constructing the pipeline. The evaluation shown the same pattern for almost every user, as they were unsure about which node needed to be included into the workspace in order to complete the ML pipeline.

Figure 4, on the other hand, also shows a high number of interactions on the side menu. But in this case, with the textual recommendations enabled, participants could interact more with the pipeline as the system guided the process through the suggestions. However, although a difference between the participants with the recommendation engine can be identified, most participants did not finish the task successfully, leaving the pipeline incomplete despite the system support.

Regarding the quantitative results, six participants that took part in the think-aloud evaluation answered the survey. Although a small sample, it allows to complement the analysis of the results obtained in the qualitative assessment and to get deeper insights.

The guidelines from [19] were followed to compute the SUS score. In this case, the score contributions from each item were added. Given that each item's score must range

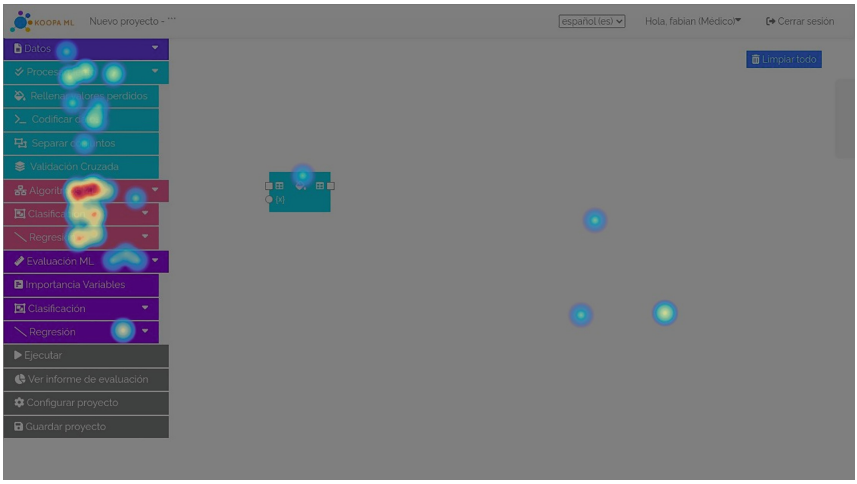

Fig. 3. Heatmap of the interactions made by a participant with the recommendation engine disabled. Obtained through Hotjar.

Fig. 4. Heatmap of the interactions made by a participant with the recommendation engine enabled. Obtained through Hotjar.

from 0 to 4, the positive items of the questionnaire were subtracted 1 point, while the negative items' scores were subtracted from 5, to normalize the sample. The sum of the scores is finally multiplied by 2.5 to obtain the overall value of the SUS between 0 to 100.

The SUS score was calculated following the scoring instructions [19] for every participant's responses. The SUS questionnaire also enables the computation of a learnability score (from items 4 and 10) and a usability score (from questions 1, 2, 3, 5, 6, 7, 8, and 9). Both scores were also calculated and transformed to fit a scale from 0 to 100. The following results were obtained for the KoopaML platform (Fig. 5):

- The average perceived usability of the KoopaML platform is **62.5**, which is considered a borderline acceptable SUS score (interpretation based on the studies done in [20, 21]).
- On the other hand, the perceived usability is significantly higher (**70.31**) than the learnability (**31.25**), which indicates that the platform is complex to learn.

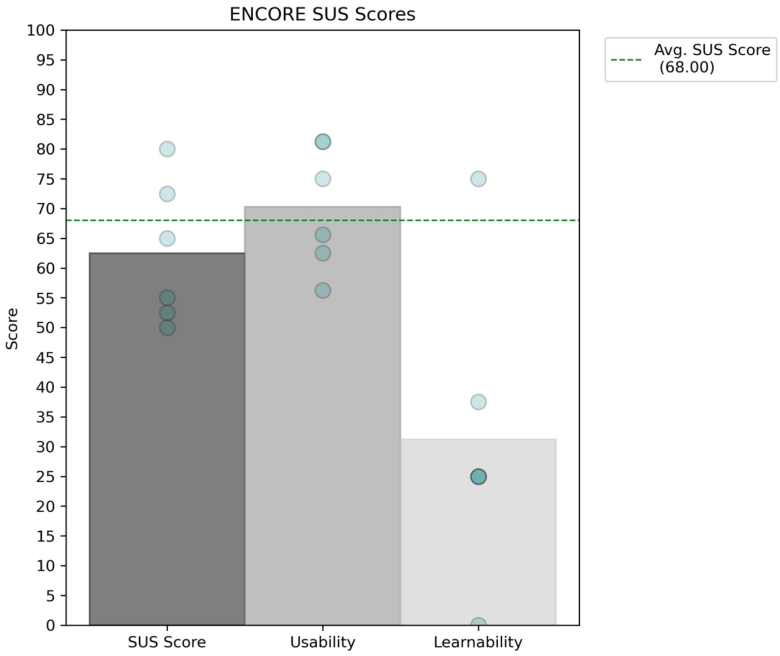

Fig. 5. SUS questionnaire results.

5 Improvements Proposal and Conclusions

Based on the previous qualitative validation and following the SUS results, it is possible to affirm that, although textual recommendations offer support to some extent, the platform is still complex, especially for non-expert users.

This issue is confirmed by the learnability score obtained from the SUS questionnaire. In fact, the usability score (70.31) is considered a "good" score following the SUS interpretation guidelines, however, the learnability (31.25) of the system is poor and not acceptable, which has impacted the overall SUS score (62.5).

The consequences of the learnability score can be observed in the interaction analysis. As shown in Figs. 3 and 4, although the users that had the textual recommendations enabled (Fig. 4) performed more interactions in the workspace and more actions with the nodes, they were not able to create a functioning ML pipeline.

One of the theories of this performance is the location and format of the recommendations. The modal box containing the next steps to perform is at the bottom of the screen, which may provoke it to be overlooked. On the other hand, the textual recommendations can be lengthy even broken down into individual steps due to the complexity of the topic, so they can be considered hard-to-follow.

Regarding the side menu containing the toolbox for creating the ML pipelines, most of the users' interactions were concentrated in this area, meaning that users spent more time searching for the proper nodes than designing the pipeline. In addition, participants were confused about finding the right node to add to the pipeline.

Finally, even if participants included the right nodes to carry out the training of the ML algorithm, they could not properly connect the nodes to create the pipeline, resulting in errors.

After this evaluation, there is a list of improvements to be included in the new version of KoopaML (Fig. 6). The proposal is to provide the recommendations visually by constraining the nodes that can be connected to a specific ML task. This way, the connections can be made almost automatically, saving time for users from connecting each socket in each node.

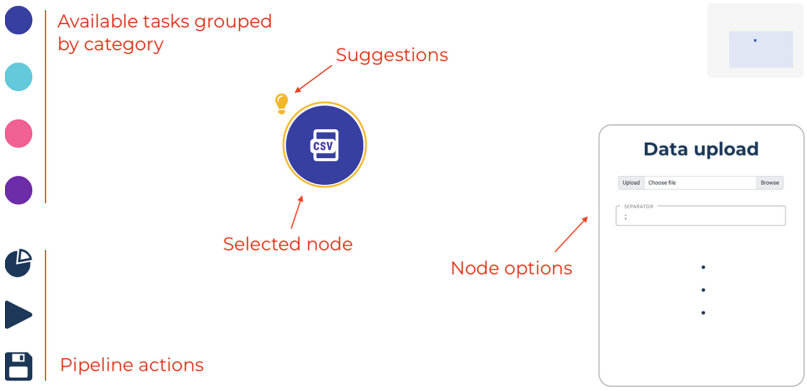

Fig. 6. Prototype sketch of the new interface.

Moreover, by using this graphical approach, the time spent on the side menu would be reduced due to the downsizing of the available nodes to connect. In this sense, nodes can be directly included through the selected node instead of going back and forth to the menu (Fig. 7).

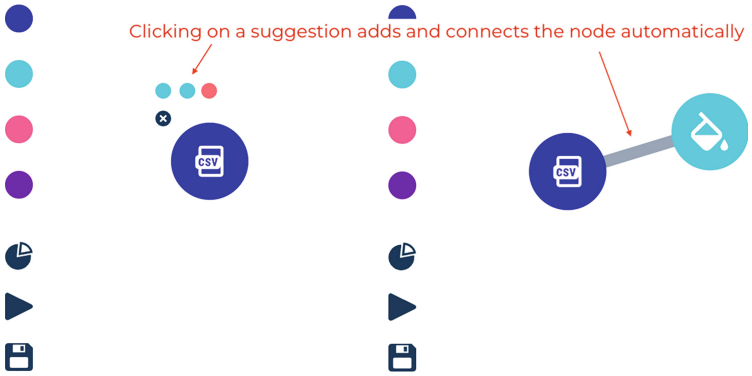

Fig. 7. Automatic connection of nodes based on suggestions.

These improvements are in the prototype phase. Future research lines will involve the implementation of the improvements in KoopaML and further testing to compare and measure the performance of the new version of the platform.

Acknowledgments. This research was partially funded by the Ministry of Science and Innovation through the AVisSA project grant number (PID2020-118345RB-I00). This work was also supported by competitive community grants (GRS 2033/A/19, GRS 2030/A/19, GRS 2031/A/19, GRS 2032/A/19) from the SACYL, Junta Castilla y León; by competitive national grants (PI14/00695, PIE14/00066, PI17/00145, DTS19/00098, PI19/00658, PI19/00656, PI21/00369) from the Institute of Health Carlos III, Spanish Ministry of Science and Innovation and co-funded by ERDF/ESF, "Investing in your future" and; by the CIBERCV (CB16/11/00374) from the Institute of Health Carlos III, Spanish Ministry of Science and Innovation.

References

1. Nemitz, P.: Constitutional democracy and technology in the age of artificial intelligence. Philosophical Trans. Roy. Soc. A Math. Phys. Eng. Sci. **376**, 20180089 (2018)
2. Brennen, J.: An industry-led debate: How UK media cover artificial intelligence (2018)
3. Fast, E., Horvitz, E.: Long-term trends in the public perception of artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence (Year)
4. Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., Biancone, P.: The role of artificial intelligence in healthcare: a structured literature review. BMC Med. Inform. Decis. Mak. **21**, 1–23 (2021)
5. Kolachalama, V.B.: Machine learning and pre-medical education. Artif. Intell. Med. **129**, 102313 (2022)
6. Pinto dos Santos, D., et al.: Medical students' attitude towards artificial intelligence: a multicentre survey. Europ. Radiol. **29**, 1640–1646 (2019)
7. Carbone, M.R.: When not to use machine learning: A perspective on potential and limitations. MRS Bulletin 1–7 (2022)
8. Pfohl, S., Xu, Y., Foryciarz, A., Ignatiadis, N., Genkins, J., Shah, N.: Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1039–1052 (Year)

9. Prabhakaran, V., Mitchell, M., Gebru, T., Gabriel, I.: A Human Rights-Based Approach to Responsible AI. arXiv preprint arXiv:2210.02667 (2022)
10. Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in ML-based science. arXiv preprint arXiv:2207.07048 (2022)
11. Vázquez-Ingelmo, A., et al.: Bringing machine learning closer to non-experts: proposal of a user-friendly machine learning tool in the healthcare domain. Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21), pp. 324–329. Association for Computing Machinery, Barcelona, Spain (2021)
12. García-Peñalvo, F.J., et al.: KoopaML: a graphical platform for building machine learning pipelines adapted to health professionals. International Journal of Interactive Multimedia and Artificial Intelligence (In Press)
13. García-Holgado, A., et al.: User-centered design approach for a machine learning platform for medical purpose. In: HCI-COLLAB 2021, pp. 237–249. Springer, Cham (2021). Doi: .https://doi.org/10.1007/978-3-030-92325-9_18
14. Jääskeläinen, R.: Think-aloud protocol. Handbook of translation studies **1**, 371–374 (2010)
15. Van Someren, M., Barnard, Y.F., Sandberg, J.: The think aloud method: a practical approach to modelling cognitive. London: AcademicPress 11, pp. 29–41 (1994)
16. Brooke, J.: SUS: a retrospective. J. Usability Stud. **8**, 29–40 (2013)
17. Tullis, T.S., Stetson, J.N.: A comparison of questionnaires for assessing website usability. In: Usability Professional Association Conference, pp. 1–12 (Year)
18. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Intl. J. Human-Comput. Inter. **24**, 574–594 (2008)
19. Brooke, J.: SUS-A quick and dirty usability scale. Usability Evaluation Ind. **189**, 4–7 (1996)
20. Bangor, A., Kortum, P., Miller, J.: Determining what individual SUS scores mean: adding an adjective rating scale. J. Usability Stud. **4**, 114–123 (2009)
21. Sauro, J.: A practical guide to the system usability scale: Background, benchmarks & best practices. Createspace Independent Pub, Scotts Valley, CA, US (2011)