How students justify the usability of self-built machine learning models

Katharina Bata, Angela Schmitz, Andreas Eichler

How students justify the usability of self-built machine learning models

Katharina Bata[1], Angela Schmitz[2] and Andreas Eichler[3]

[1] KIT – Karlsruhe Institute of Technology, Karlsruhe, Germany; katharina.bata@kit.edu

[2]TH Köln – University of Applied Sciences, Cologne, Germany

[3]University of Kassel, Kassel, Germany

Using machine learning models is a part of the everyday work of many engineers. In this paper, we investigate how undergraduate engineering students use self-built machine learning models in a concrete application example. A qualitative analysis of excerpts from twelve student groups shows that despite similar models, the tendency to use the models in the application example differs. In addition, justifications for the use of the model are provided on two different levels. One level covers the resilience of the model itself. The other level relates to the applicability of the model statement, which is what we call the information that is obtained by applying a model to the new data set.

Keywords: Machine Learning, Model Building, Model Usability, Data Literacy

Theoretical Background and research question

Machine learning (ML) is at the heart of modern engineering and, through the use of mainly statistical methods and automated data processing, an interesting learning content in the context of statistical and data literacy (Engel & Martignon, 2024). ML can be used to optimize complex systems, to predict failures for maintenance, or to accelerate research by data analysis. Besides the building and the evaluation of ML models, which can be outsourced to other experts if necessary, justifying the usability of a model in an application situation is crucial (Bertolini et al., 2021).

The question of justifying the usability of a model in an application situation is closely related to the user reliance on recommendations from ML models. Due to the increasing use of ML-based methods using the label "AI", some relevant publications about reliance on recommendations from ML models also appear with the label "trust in AI" (Bach et al., 2022; Yang & Wibowo, 2022). "Trust in AI" is investigated in the context of various influencing factors on technological, organizational, contextual, social and personal level (Yang & Wibowo, 2022). Looking specifically at ML-based decisions, there is empirical evidence for different types of influencing factors related to the different levels. For example, on the technological level influencing factors can be the properties of a model, measured by different performance metrics (Yu et al., 2019; Yin et al., 2019). On a personal level, the difference between one's own ideas of a model and the model itself seems to be an influencing factor (Bansal et al., 2019). There is also the assumption that personal preferences and expertise regarding the ML-methods or a respective performance indicator have an influence on the reliance in ML-based decisions of users, i.e. laypersons related to model building (Chiang & Yin, 2022; Yin et al., 2019). Another interesting aspect is that trust in the ML-based decisions can be influenced by an appropriate visual representation of the ML-based decision process (Yang et al., 2020).

Our contribution enriches the field of research on user reliance on a recommendation of a ML model with a perspective that deals with a step prior to the recommendation: The usability of a ML model in a specific situation. The question is: *How do engineering students justify the usability of a self-built ML model?*

The situation in which we investigate students' justifications is characterized by two particular aspects: The investigation is based on the processing of a task in a teaching-learning environment in the context of a design research study (Bata et al., 2022). The task, which is described in the methods section, is an example of an engineering application that is intended to represent a situation from professional life, where students apply a ML model to a new data set. Furthermore, the students in our study use models that they have previously developed themselves, which means that they are explicitly not laypersons but beginners related to model building.

Method

Task

In the investigated task, the students have to apply a model to a new data set. The model was developed by themselves beforehand using the k-Nearest Neighbor classifier to predict steel quality (good quality / bad quality) based on two metric features (number of cracks / width of the biggest crack). In Figure 1 there is a visual representation of the models' prediction for the new dataset (containing samples of old and new steel): In the data set, half of the samples consist of old steel (white), and the other half of new steel (black). The background colors separate the areas classified by the model as good (blue) and as bad (red). The representation shows that all samples of the old steel are classified as of poor quality, and about 80% of the samples of new steel are classified as of good quality. Based on the application context and a visual representation such as in Figure 1, the students are asked to make a recommendation for the new steel in contrast to the old.

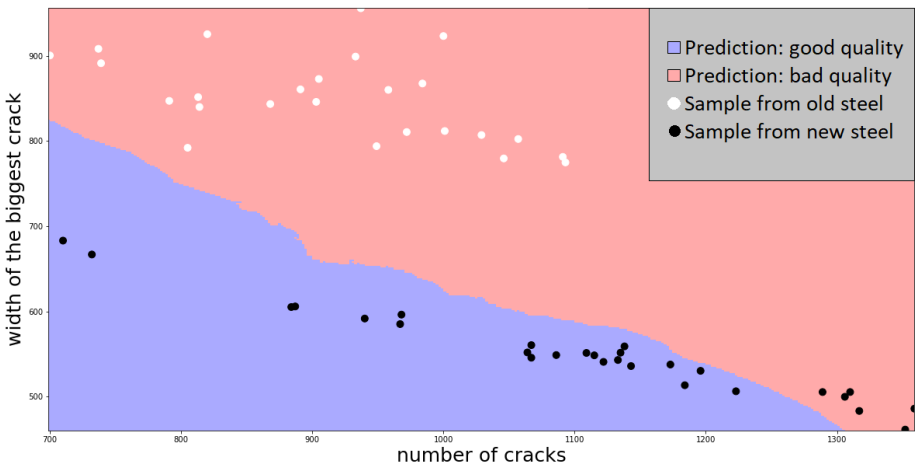

Figure 1: Visualization of a student-built model and the new data set as foundation for the decision

Data from the design experiments

To address the research question, we collected data from three cycles of design experiments in a laboratory setting (Gravemeijer & Cobb, 2006) involving twelve groups with a total of 29

participants. The participants were students from different engineering bachelor programs in the third to seventh semester at a German university, with no previous experience in machine learning. The different bachelor programs had minor discrepancies in the computer science and programming curricula, so a crash course in Python was given to some students before the introductory course started. This was useful in preparing the students for the programming tasks in the introductory course, which were completed using Jupyter Notebooks with Python.

The design experiments were conducted as follows: The twelve student groups were guided individually through the material by the lecturer, which is on top the researcher herself, via an online conference tool in three sessions each (Bata et al., 2022). The work on the material includes both presentation phases by the lecturer and group work on various tasks, one of which is the one described above. The three cycles differed only minimally in their prerequisites, the formulation of the task stayed the same. Each session of each group was documented by video recording.

Qualitative content analysis

The video recordings from the design experiments were transcribed in order to conduct a qualitative content analysis (Kuckartz, 2014). In the course of the qualitative content analysis we developed an inductive category system (Table 1), which was validated by means of consensual coding and by determining the intercoder reliability (Kuckartz, 2014). The intercoder reliabilities, which were calculated on a subset of the material coded by another researcher, of the three main categories range from 0.82 to 1. The value of one is achieved here because the subcategories of the first main category leave little to be interpreted. Taking into account the number of subcategories, the values correspond to a Cohen's kappa between 0.77 and 1 which means a good to very good agreement.

Results

The path from one example to the category system

We will start by showing an excerpt from a transcript of group 2, consisting of the students with pseudonyms Hanna, Wael and Bastian. They are talking about a recommendation looking at the visualization in Figure 1. Wael's statement is an answer to the question "Would you make a recommendation at this time?"

329	Wael:	So I would say, purely intuitively, no, we need much more training data (laughs). But maybe that's also my tic, I don't know.
330	Hanna:	But let's assume that we should now decide on the basis of what we have. Or we don't have no further budget available, then I would say, after 18,000 load changes, the steel, the new steel, is just still good in contrast to the old steel.
331	Bastian:	I think that because the result is so clear it is easier to make a recommendation than if it was a bit scarcer.

In this excerpt the students' decision to use the model for a recommendation is not clear. Two passages are coded regarding their decision, one in [329], coded by "Model is not good enough for recommendation" (C1.3) and the other in [330], coded by "New steel lasts longer / is better" (C1.2). Both codes are part of the first main category "Decision" (C1). This main category distinguishes

whether the students want to make a recommendation at all and if so, in favor or against the new steel. This category serves to frame the further coding with categories C2 and C3 (see Table 1).

Furthermore, we observe that the students provide justifications at two different levels. On the one hand, they give reasons related to the model itself like Wael in [329]. He says that he would like to have much more data to train the model. By mentioning the "Amount of data" (C2.6) he gives an aspect related to the second main category "Model resilience" (C2). Table 1 shows this and the other subcategories of C2. E.g. Students justify with traceability (C2.1), or they state whether their model is good or bad by using some performance metrics (C2.5). On the other hand, the students use justifications that are related to the "Usability of the model statement" (C3). A model statement is the information obtained by applying the existing model to the new data set, e.g. by looking into Figure 1 (all samples of the old steel are classified as of poor quality, and about 80% of the samples of new steel are classified as of good quality). An example for this is, when Bastian points out in [331] that the recommendation using the model is easier because the statement is very clear. This passage is coded as "Clarity of model statements" (C3.2). Table 1 shows the other subcategories of C3. E.g. Students justify with applicability or transferability (C3.3).

Table 1: Category system

Main Category	Subcategory
C1: Decision	C1.1 New steel is not good enough
	C1.2 New steel lasts longer / is better
	C1.3 Model is not good enough for recommendation
C2: Model resilience	C2.1 Clear / Traceable statement
	C2.2 Reality mapping
	C2.3 Parameter selection
	C2.4 Variance and bias
	C2.5 Performance metrics
	C2.6 Amount of data
	C2.7 CRISP-DM
C3: Applicability of the model statements	C3.1 Model quality
	C3.2 Clarity of model statements
	C3.3 Applicability / Transferability
	C3.4 Significance of data / features
	C3.5 Business perspective

Some examples of different subcategories

To illustrate further subcategories, excerpts from the transcripts of two other groups are shown below. First, we look at an excerpt from group 3 with the students Feline and Greta:

182	Lecturer:	[…] Then I have one more question. You wrote that you have a “reliable model”. If somehow, I'll say, something goes wrong in the process at the end. So now you've said we're using the new steel. And it's being used and then something goes wrong. (Greta: Hm (affirmative)). How would you defend that our model was reliable?
183	Feline:	(11) With the uhm criteria. So (..) we were given a data set on the basis of which we were supposed to build a model. And we tried to adapt the model so that the performance indicators were as good as possible and, above all, that the precision was as high as possible. And I would say that you could certainly defend it with that.

In this excerpt, a subcategory of C2 “Model resilience” is coded, namely C2.5 "Performance metrics" in Feline's statement in [183]. Feline explains that the group created the model beforehand with the aim of maximizing the performance metric values, especially the precision. Precision is emphasized here because in the application context, it places work safety above profit, as the students' considerations at another point in the processing show. Another thing, that this excerpt shows, is the lecturer giving a stimulus to motivate the students to give more justifications [182]. Impulses of this kind are discussed later in the paper.

The second excerpt is from group 12 with the students Dieter, Enno and Fabian:

122	Fabian:	[…] Yes, it just depends on the first question, what exactly they want, right? If there are almost no load changes with the steel and it's just permanently installed somewhere. (Enno: Hm (affirmative)) maybe it’s not worth it.
123	Enno:	Or does the company actually want the steel to be replaced as often as possible so that they can sell more parts?

The excerpt shows two subcategories of category 3 “Applicability of the model statements”, namely the two that are particularly broad and cover many different aspects. In Fabian's statement [122] he notes that the given data generated by the simulation of load changes on test components may not be meaningful for the application of the new steel. This formulates the aspect that the data used to build a model or to which a model is to be applied must fulfil certain requirements in order to be meaningful. The statement is therefore coded in category subcategory C3.3 "Applicability / Transferability”. Enno's subsequent statement goes even further than Fabian’s and questions whether the goal of material selection is high quality at all. The associated statement [123] should therefore be categorized as C3.5 "Business perspective".

Both examples highlight the variance in students' justifications, which could already be anticipated by the category system (Table 1): The justifications are at different levels (C2 “Model resilience” & C3 “Applicability of the model statements“) and within the levels concerning diverse topics (e.g.

technical things like performance metrics or contextual things like the business perspective). The next section provides an overview of the results of the coding across all groups.

Summary on the twelve groups

Although all groups start with slightly different models as a result of working with their self-built models, the model statement was mainly similar (most of the old steels as of bad quality and most new steels as of good quality, see Figure 2). The students tend to make a recommendation (C1.3 in only four groups) and when they do, it is more likely to be in favor of the new steel (C1.1 in only four groups), which is indeed the aim of the task. The groups that disagree about using the model or recommending the new steel can be identified by different codes in category 1 (as in the example shown). The disagreeing groups tend to use more and different subcategories in the other two main categories than the agreeing groups. Evaluating all passages coded with the main category 2, we see that not all groups use the resilience of the model (C2) in their justification. Some groups are not even able to do so when asked. A different picture emerges for category 3: The students show a wide range in their ability to justify the usability of the model statement (C3). Each group justifies at least once at the level of main category 3.

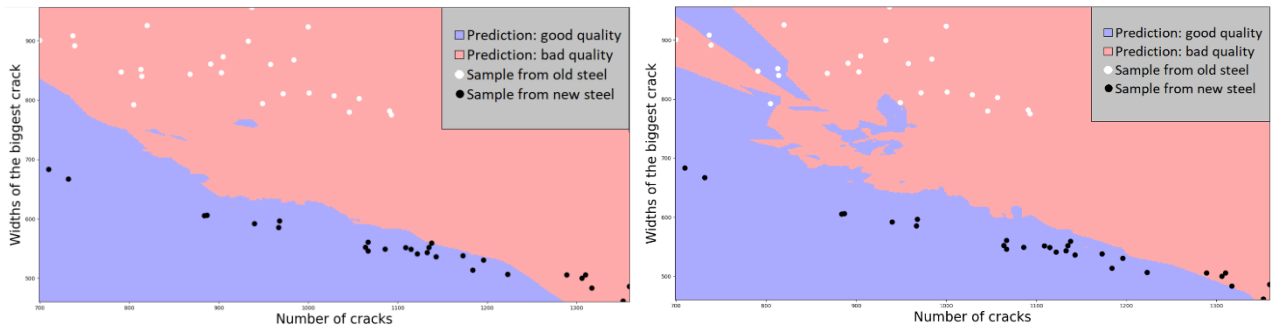

Figure 2: Two different models (group 5 & group 12)

Discussion

The developed category system (Table 1) opens the possibility to identify and analyze the students' justifications for the usability of ML models in an application context. It is interesting to note that although the question is about the application of an existing model to a new data set, the students argue at the C2 level ("Model resilience") in addition to the expected C3 level (**"**Applicability of the model statements"). However, this result is consistent with the existing literature, as the technical characteristics of the models, such as the specification of performance measures, increase the reliance on the model recommendations (Yu et al., 2019; Yin et al., 2019). This is true even when the modelling process has not been carried out by the users themselves in advance, which is the case in the present study. In contrast some groups are not able to argue at C2 level at all even though the lecturer has used a prescribed stimulus (as in the example from group 3) to motivate students to give answers from the category if they have not already done so on their own. Here it would be interesting to understand whether the partial lack of justification is due to the concrete task or to other reasons that lie in the model building processes beforehand (Bata et al., 2024) or in some basic understanding of ML. Despite the partial lack at C2 level, the results show a broad and diverse use of justifications, which is desirable when dealing with recommendations from ML models. This raises the question of

whether it might be useful to involve laypeople in modelling processes or at least to inform them about modelling processes in order to improve their reliance on the model results. This question, or assumption, is in line with Chiang & Yin's (2022) findings that short-term machine learning literacy interventions have an impact on laypeople's interaction with ML models and their outcomes.

Although only twelve groups with 29 students were studied to develop the category system, it is reasonable to assume that the subcategories of categories C2 and C3 are complete and transferable to other applications of supervised learning. This can be assessed from a technical point of view for main category 2, since the usual influences (data, parameters, structured generation) and evaluation parameters (performance metrics, variance and bias, comprehensibility and reality mapping) of supervised learning methods are represented. With regard to the main category 3, usual evaluation parameters of the application of ML models are represented. This is also due to the fact that the categories C3.3 and C3.5 are very broad and cover many different aspects. For example, in subcategory C3.5, the sections vary from “a steel has too long delivery times” to “what is the actual objective of the ordering company”. An investigation of the assumption of completeness in a different setting would be a desirable follow-up study and could expand the picture if necessary.

Although the subcategories of categories C2 and C3 can be assumed to be complete, looking at the differentiation of Yang & Wibowo (2022), there could be other main categories to justify the usability of a ML model. The justifications covered by the category system could, following Yang & Wibowo (2022) be sorted into the "technological" and "contextual" factors. It might be interesting to search intensively for further main categories that analyze the other formulated influences such as personal factors. However, it is quite probable that the learning situation or research format described above is not suitable for such an investigation.

From a design research perspective, there are three further considerations. First, there is the question of how to enable all students to argue not only at C3 level (“Applicability of the model statements”) but also at C2 level (“Model resilience”). Secondly, one could look more closely at whether one can artificially create the disagreement that, at least in the present sample, seems to lead to the use of more and different subcategories. The final consideration, also in relation to the results of Yang et al. (2020), concerns the way in which the model is visualized with the new data set (Figure 1). The question remains as to whether the visualization is suitable for providing students with a basis for justifying the usability of the model and, building on this, whether the type of visualization influences the students' justification.

References

Bach, T. A., Khan A., Hallock H., Beltrão G., & Sousa S. (2022). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction.* https://doi.org/10.1080/10447318.2022.2138826

Bansal G., Nushi, B., Kamar, E., Lasecki, W. S., Weld D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7, 2-11.

Bata, K., Schmitz, A., & Eichler, A. (2024). Processing graphs as an illustration of how engineering students build a machine learning model. In P. Drijvers, H. Palmér, C. Csapodi, K. Gosztonyi, & E. Kónya (Eds.), *Proceedings of the Thirteenth Congress of the European Society for Research in Mathematics Education* (CERME13). Alfréd Rényi Institute of Mathematics and ERME.

Bata, K., Schmitz, A., & Eichler, A. (2022). Insights into the design of an introductory course for data science and machine learning for engineering students. In G. Bolondi, F. Ferretti & J. Hodgen (Eds.), *Proceedings of the 12th Congress of the European Society for Research in Mathematics Education* (CERME12). Free University of Bozen-Bolzano and ERME.

Bertolini M., Mezzogori D., Neroni M., & Zammori F. (2021). Machine Learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications*, 175. https://doi.org/10.1016/j.eswa.2021.114820

Chiang, C.-W., & Yin, Ming. (2022). Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 148-161. https://doi.org/10.1145/3490099.3511121

Engel, J., & Martignon, L. (2024). Data science for informed citizen: Learning at the intersection of data literacy, statistics and social justice. Revista Internacional De Pesquisa Em Educação Matemática, 14(3), 1-13. https://doi.org/10.37001/ripem.v14i3.3816

Gravemeijer, K. P. E., & Cobb, P. (2006). Design research from a learning design perspective. In Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). *Educational Design Research* (pp. 45-85). Taylor and Francis Ltd.

Kuckartz, U. (2014). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung* (2nd ed.). Beltz Verlagsgruppe.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 189-201. https://doi.org/10.1145/3377325.3377480

Yang, R., & Wibowo, S. (2022). User trust in artificial intelligence: A comprehensive conceptual framework. *Electron Markets*, 32, 2053-2077. https://doi.org/10.1007/s12525-022-00592-6

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the CHI Conference on Human Factors in Computing Systems*. https://dl.acm.org/doi/10.1145/3290605.3300509

Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen F. (2019). Do I trust my machine teammate? An investigation from perception to decision. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 460-468. https://doi.org/10.1145/3301275.3302277