Review on nhess-2022-198-v2: Using machine learning algorithms to identify predictors of social vulnerability in the event of a hazard: Istanbul case study

After reviewing the materials provided by the authors for this submission, I applaud the authors' efforts in revising their manuscript and offering exciting responses to the previous comments of the reviewers. However, at this stage, I still do not feel that the current version of the manuscript is good enough for publication in NHESS. I still have a number of questions and concerns at the medium or minor level. The authors need to address them. In addition, I strongly suggest the authors, when finishing this round of revision, carefully read their modified manuscript and make serious efforts in polishing their writing and patiently conducting some editing works to their manuscript. Below are listed my medium and minor questions and concerns.

Medium and Minor Issues

1. L23: "CART" is short for "classification and regression tree". Please use either "classification tree (CT)" or "classification and regression tree (CART)" to avoid confusion. The same for the rest of the manuscript.

2. L40: When used as a countable noun, "vulnerability" usually means a weak link, a loophole, a fragile element, etc., of a system that may be exploited by hazardous agents to result in loss to the system. Is that what the authors mean here? If not, please use the uncountable version of the noun "vulnerability".

3. L86: I suggest removing “in fact” because the statement here is still about an intellectual guess or belief.

4. L150-151: Very large earthquakes (over Mw7.0) do seem to be rare for Istanbul. However, earthquakes with a magnitude 4 or above can still cause significant damage to communities (see, e.g., Wang and Sebastian 2022 https://doi.org/10.5194/nhess-22-4103-2022). These earthquakes shouldn’t be rare at all according to the estimated 100-year return period for an earthquake with Mw7.0 and above around Istanbul. In addition, as the manuscript has changed its focus from earthquake to all hazards, the large hazardous events should be more frequent than merely large earthquakes. Even the authors themselves mention on L203-204 that the study area “is in a region that is prone to natural hazards where a large-scale disaster happens every seven to eight years (Baris, 2009)”. Moreover, at the household level, many families do not have to wait for a large-scale disaster to occur before experiencing loss unfortunately. More impacts to households are likely to be caused by the much more frequent smaller-scale hazardous events that may not even be considered or defined as disasters.

5. L202-210: This paragraph is inappropriate for hazards in general. Please revise it.

6. L213-214: Since the focus is on hazards in general instead of earthquake now, I suggest the authors explain a little bit regarding why the survey conducted by an earthquake-related organization can be used for all hazards.

7. L231-234: There are grammatical problems associated with this sentence “It considers … and socio-economic status”. Please modify it.

8. L303-305: The authors claim that for “different tuning parameter alternatives, the choice of the optimal tuning parameter was determined by the largest area under the curve (AUC) value of the receiver operating characteristic (ROC) curve using the automated grid search”. However, in the supplementary file 3 p. 2, the authors clearly state that the parameter K of KNN is “determined with the square root of the number of points in the training data set”. These two statements are inconsistent with each other. Why? Moreover, the parameter **ntree** of RF is also determined arbitrarily by the authors without grid search.

9. L403-404: The sentence “The prevalence … among 41,093 households” needs to be modified for all hazards.

10. L422-424: Sensitivity and recall are the same thing. Positive prediction value and precision are the same thing.

11. L435: Fig. 3 may not be friendly enough to colorblind readers.

12. L468-472: It is still unclear in the manuscript how the relative importance of predictors is measured. What methods or algorithms do the authors use for quantifying predictor importance?

13. L475: In Fig. 4, it is unclear whether it is the size of the circle or the hue that is supposed to correspond to the number next to the circle in the legend. Also, the scale of the circle size does not cover the small value as for debt in Fig. 4A. In addition, what do the colors of the bars in Fig. 4b mean? If they indicate the variable importance in terms of percentages, then these colors provide redundant information that is confusing.

14. L506-507: I find it difficult to see the connections between this sentence "For many decades…derived variables (Di Franco and Santurro, 2020)" and the rest part of this paragraph. I suggest the authors make modifications to the paragraph.

15. L513-521: This paragraph reads awkward. What the main point is here is unclear. The authors need to revise the paragraph. Also, it is dangerous to assume that the trained non-linear structure of the neurons of an ANN represents well the relationships between the input variables. Every training may result in a totally different internal structure of the ANN.

16. L530-531: Why is it important whether the training data has to be balanced? If the identification of high social vulnerability is preferred, why don't the authors use a reversely imbalanced training data to boost the sensitivity, etc., even more?

17. L538-549: What is the main theme of this paragraph? Are the authors trying to discuss the methods for measuring importance of input variables or the important input variables identified in their study? Also, as I have asked previously, what is actually the method or algorithm that the authors use for their quantification of variable importance?

18. L538-585: The writing of these 3 paragraphs needs to be improved.

19. L566-568: It does not make sense that prevalence of social security in low vulnerability households is lower than in high vulnerability households.

20. L569-585: The material in this paragraph involving earthquake needs to be modified to fit the tone for all hazards.

21. L624: What are "hazard risks"? I suggest the authors use "hazards" here instead.

22. L641-642: Why “fault lines”? Is the manuscript supposed to be for all hazards now?

23. The authors’ final ANN model has an accuracy of less than 75%. That is quite low for identifying high vulnerability households. Isn’t this an obvious limitation? If I were the public official to look at vulnerable households, I would definitely need a prediction model with accuracy over 90%, or even 95% or 99%. How could I afford to miss those actually vulnerable households while providing a lot of resources to households that are actually not highly vulnerable?

24. Supplementary file 3, p.2, 4. SVM: Why do the authors use the radial basis function (RBF) kernel? A linear kernel can correspond better to a hyperplane in a vector space with the same number of dimensions as the number of input variables. When an RBF kernel is applied, it is equivalent to transforming the original vector space into a vector space with an infinite number of dimensions. Such a transformation can be demonstrated with the application of a Taylor expansion when we are using the Lagrange multipliers to solve the optimization problem for calibrating the SVM.

25. Supplementary file 3, p.2, 7. ANN: ANN is not necessarily non-linear. When all the activation functions are linear, an ANN is equivalent to a linear model. ANN is also not necessarily based on deep learning. Deep learning involves an ANN with at least 2 hidden layers. If the authors only use 1 hidden layer, that is not deep learning. Also, what activation functions do the authors use for their multilayer perceptron ANN model?