Peer review

1.ML models are used to predict events or classify data into categories (among other uses). I did not see a train-validation-test split for the data (17232 responses). How many of those responses was the model trained with ? How many responses were used to validate the model ? How many responses (unseen by the model) were used to test the model ? The confusion matrix and AUC curves are used to display test data (not the entire dataset). There seems to be some misconception in how ML models work and are used. Please discuss and describe in the manucript.

2.You state "…..Pairwise Spearman correlation test was used to identify highly correlated questions. One question per theme was identified using Linear Regression by selecting the one question with the highest slope with respect to suicide attempts. Consequently, the survey questions within a theme were reduced to one question that best predicts suicide attempts……" I am unsure as to how exactly this was calculated. I need to see thorough discussion , presentation and explanation of at least one theme in the manuscript. For example, how did you use linear regression to select a question ? Where did the slope in the equation (what equation, how derived, from what data) originate from? Please introduce a new section into the manuscript and describe how exactly (methodology) the highly correlated questions were identified using linear regresssion, and how exactly (methodology) these questions were reduced to one question.

3.How is the KS statistic calculated ? Goodness of fit to what ? KS statistic is calculated as the difference between the training and testing cumulative curve. Where are these curves ? In this regard (see point 1). Please describe in detail as to which curves are being compared ? Please note that it should be possible for anyone reading your manuscript to be able to replicate your work.

4.You have not presented the AUC cross validation for the logistic regression model. Please do so.

5.Class imbalance is inherent in this study based on the lower incidence of suicide ideators as a total percentage of the population. How was this class imbalance dealt with in the three models you used ? In this context, refer to point 1, where you do not seem to have split the data into train-validate-test. Please discuss and describe in the manuscript. See: https://doi.org/10.1007/s10964-023-01892-6

6.In data where class balance exists, the AUPRC is a better indicator of model performance than the AUC. Please present the Area under the precision recall curve. In order to do this, you will need to calculate the precision, recall and F1 for your data for the three models used in the study. Please present these values as well in the manuscript.

7.References need to be numbered in sequential order in the body of the text. Please revise and make sure the references in the references section correlate with the sequential numbering in the text.

8.Why were those three models chosen ? Please provide a rationale.

In summary, the methodology and the use of ML models in this manuscript seems conceptually flawed. This manuscript will require extensive revision and possibly rework before another round of review.

---

1. ML models are used to predict events or classify data into categories (among other uses). I did not see a train-validation-test split for the data (17232 responses). How many of those responses was the

model trained with ? How many responses were used to validate the model ? How many responses (unseen by the model) were used to test the model ? The confusion matrix and AUC curves are used to display test data (not the entire dataset). There seems to be some misconception in how ML models work and are used. Please discuss and describe in the manucript.
**Answer:** We addressed this review feedback in the revised paper in the last paragraph of the 'Methods' section under 'Data Processing' - '3. Model Methodology.' Instead of using a train-test split, we employed stratified 3-fold cross-validation on the full population dataset. According to a reference paper, when the target rate is around 10%, this approach provides a more reliable evaluation of the model's performance by preserving class balance across folds, resulting in a more consistent and accurate assessment compared to a train-test split, which may not always adequately address class imbalance.

2. You state "…..Pairwise Spearman correlation test was used to identify highly correlated questions. One question per theme was identified using Linear Regression by selecting the one question with the highest slope with respect to suicide attempts. Consequently, the survey questions within a theme were reduced to one question that best predicts suicide attempts……" I am unsure as to how exactly this was calculated. I need to see thorough discussion , presentation and explanation of at least one theme in the manuscript. For example, how did you use linear regression to select a question ? Where did the slope in the equation (what equation, how derived, from what data) originate from? Please introduce a new section into the manuscript and describe how exactly (methodology) the highly correlated questions were identified using linear regresssion, and how exactly (methodology) these questions were reduced to one question.
**Answer:** We addressed the review feedback in the revised paper under the 'Methods' section – the 'Data Processing' subsection - the '1. Data Reduction Step' portion. This section includes four new paragraphs under Figure 1, along with two additional figures, Figure 2 and Figure 3. We used Theme 5, Sexual Behavior Theme, as an example to explain the variable reduction process in detail through linear regression.

3. How is the KS statistic calculated ? Goodness of fit to what ? KS statistic is calculated as the difference between the training and testing cumulative curve. Where are these curves ? In this regard (see point 1). Please describe in detail as to which curves are being compared ? Please note that it should be possible for anyone reading your manuscript to be able to replicate your work.
**Answer:** We addressed this review feedback in the revised paper under "Machine Learning Procedure" section – Step 6 – 6.3 portion. The mathematical formula for the K-S statistic has been added below its formal definition. Figure 9 is the K-S curve for our Logistic Regression Model.

4. You have not presented the AUC cross validation for the logistic regression model. Please do so.
**Answer:** In Figure 16, you can see cross validation mean AUC=0.7091

5. Class imbalance is inherent in this study based on the lower incidence of suicide ideators as a total percentage of the population. How was this class imbalance dealt with in the three models you used ? In this context, refer to point 1, where you do not seem to have split the data into train-validate-test. Please discuss and describe in the manuscript. See: https://doi.org/10.1007/s10964-023-01892-6
**Answer:** We addressed the class imbalance issue by employing stratified 3-fold cross-validation on the full population dataset for all 3 model methodologies. Model key performance evaluation metrics has added in AUPRC. The changes are primarily in four areas of our paper: The last paragraph of the "Methods" section under "Data Processing" - "3. Model Methodology" The "Machine Learning Procedure" section, Steps 5 and 6 The "Discussion" section, under "2. Further Study using Machine Learning Models" The middle part of the first paragraph in the "Conclusion" section

6. In data where class balance exists, the AUPRC is a better indicator of model performance than the AUC. Please present the Area under the precision recall curve. In order to do this, you will need to calculate the precision, recall and F1 for your data for the three models used in the study. Please present these values as well in the manuscript.
**Answer:** Yes, we have incorporated AUPRC as one of our model performance key metrics. You can find these updates throughout our revised paper. For a quick review, you may search for the keyword "AUPRC" to locate the changes in the following sections: Abstract Methods: Data Processing - 3. Model Methodology Machine Learning Procedure: Step 6 Results Discussion

7. References need to be numbered in sequential order in the body of the text. Please revise and make sure the references in the references section correlate with the sequential numbering in the text.
**Answer:** We addressed this review feedback at the main body and reference section in the revised paper.

8. Why were those three models chosen? Please provide a rationale.
**Answer:** We have addressed this feedback in the 'Methods' section under 'Data Processing,' specifically in paragraphs 1 through 6 of the 'Model Methodology' sub-section. In these paragraphs, we provide a detailed rationale for the selection of the three models, explaining the criteria and considerations that guided our choices.

In summary, the methodology and the use of ML models in this manuscript seems conceptually flawed. This manuscript will require extensive revision and possibly rework before another round of review.
**Answer:** Thank you for your feedback on our paper. Hopefully our revision has addressed all your concerns and comments.

---

Thank you for addressing my comments. However, for comment 1, you still need to train on 2 stratified folds and test on the other stratified fold (3 times). I still don't see where in the manuscript you have trained the model with the 2 stratified folds and then tested with the 1 stratified fold ( I don't see training and testing data). Please address this comment.

---

Thank you for addressing my comments. However, for comment 1, you still need to train on 2 stratified folds and test on the other stratified fold (3 times). I still don't see where in the manuscript you have trained the model with the 2 stratified folds and then tested with the 1 stratified fold ( I don't see training and testing data). Please address this comment.
**Answer:** We addressed this comment in the revised paper mainly in two places:
1) the last paragraph of the 'Methods' section under 'Data Processing' - '3. Model Methodology.' We added in two figures – figure 5 and 6 - to show the sample sizes for training and test split in the stratified 3-fold cross validation.
2) Machine Learning Procedure section – step 4 & 5. Other statisticians successfully modeled the full data and evaluated with k-fold cross validation. More details are in Empirical evaluation of internal validation methods for prediction in large-scale clinical data with rare-event outcomes: a case study in suicide risk prediction – https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9890785/,
We followed this reference and determined feature importance from a model built on full population. And the model is evaluated with cross validation which uses training-test split in each fold.
BTW - We really appreciate all the feedback and have learned a great deal from it. Please let us know if you still have any concerns.

Accepted