UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

MÁSTER EN BIOLOGÍA COMPUTACIONAL

DEPARTAMENTO DE INTELIGENCIA ARTIFICIAL DE LA UNIVERSIDAD POLITÉCNICA DE MADRID

DEPARTAMENTO DE MEDICINA LEGAL, PSIQUIATRÍA Y PATOLOGÍA DE LA UNIVERSIDAD COMPLUTENSE DE MADRID

USE OF SUPERVISED LEARNING TO PREDICT SUICIDAL AND NON-SUICIDAL SELF HARM RISK

TRABAJO FIN DE MÁSTER

Autor: Nicole Alexandra Frontero

Tutor Externo: Alejandro de la Torre Luque

Tutor Académico: Bojan Mihaljević

Julio de 2023

ABSTRACT

Suicide is a very prevalent, tragic public health concern. Artificial intelligence can be a helpful tool for predicting suicide attempts and identifying important factors in suicide risk. In the present work, we use four supervised machine learning techniques - random forest, K-nearest neighbors, classification trees, and generalized linear models - to classify individuals with suicide attempts. We use data from University College London's longitudinal Millennium Cohort Study, which followed a cohort of individuals from age 9 months to age 17 and sampled data over 7 time points. We use variables from wave 6, (age 14) to predict suicidality at wave 7 (age 17). We find that the classification trees we created performed the best, with metrics in the following ranges: accuracy 90.52% – 92.23%; true positive rate 17.00% – 19.43%; true negative rate 95.75% – 97.64%; false positive rate 2.36% – 4.25%; false negative rate 80.57% - 84.64%. Additionally, through modelling, we identified a number of important variables that aligned with the literature, for example, history of self-harm, sex of the participant, and variables implicated in the emotional and behavioral functioning of the participant.

INTRODUCTION

Suicide is a global phenomenon, with more than 700,000 people dying each year due to it in 2019 (World Health Organization, 2021a). For each completed suicide, there are even more suicide attempts, with some estimates putting this ratio at 1:20 (World Health Organization, 2021b). Among children and adolescents, suicide is a leading non-natural cause of death worldwide, with rates of adolescent suicide attempts increasing over the past decades (World Health Organization, 2021a).

There are individual, relationship, community, and societal risk factors for suicide (Center for Disease Control and Prevention, 2023). Individual risk factors include previous suicide attempt(s) (Park et al., 2018), history of depression or other mental illnesses (Bertolote & Fleischmann, 2002), substance use (Nock et al., 2009), impulsive tendencies (Gvion et al., 2015), and aggressive tendencies (Wang et al., 2014), among others. Relationship risk factors include bullying (Center for Disease Control and Prevention, 2014), social isolation (Motillon-Toudic et al., 2022), and high conflict or violent relationships (Brown & Seals, 2019), as well as others. Lastly, societal risk factors include unsafe media portrayals of suicide (Hawton & Williams, 2002), easy access to lethal means (Barber & Miller, 2014), and the stigma associated with mental illness and help-seeking (Center for Disease Control and Prevention, 2023).

Research into suicide risk factors and the increasing incidence of suicide have led to the development of many suicide-prevention programs. Such prevention methods include education of physicians, restricting of access to lethal means, public education, and screening programs (Mann et al., 2005). Examples of effective screening tools used to assess risk of suicidality include the SBQ-R (Osman et al., 2001), the Paykel scale (Fonseca-Pedrero & Pérez de Albéniz, 2020), the Columbia-Suicide Severity Rating Scale (Posner et al., 2011), and the Ask Suicide-Screening Questions toolkit (Horowitz et al., 2012).

Suicide continues to become an increasing problem globally, yet our ability to predict suicides has not improved over the last 50 years (Franklin et al., 2017). Although it is difficult to predict suicide, artificial intelligence and machine learning may be tools well poised to help in the effort of suicide prediction and prevention (Bernert et al., 2020).

Burke et al. (2019) performed a systematic review of the use of machine learning in suicide research and identified three main goals found in research papers pertaining this topic. The first was to improve the accuracy of risk prediction, the second to identify important predictors, and the third to identify subgroups of patients (Lejeune et al., 2022). With regard to risk prediction, AI has the potential to help at an individual level with better identifying individuals in crisis or those who could be at risk of an imminent crisis, while at a population level, AI can help with identifying at risk groups (Fonseka et al., 2019). In their review of AI and suicide prevention, Lejeune et al. (2022) identified that the main algorithms used in the literature were logistic regression, random forest, gradient-boosting algorithms, LASSO, and support vector machines.

One particularly powerful example of the use of AI in predicting suicide was done by Kumar et al. (2022). They created an eXtreme Gradient Boosting (XGBoost) based machine learning model to predict suicides at the county-level (US) with an R^2 of 0.98. Additionally, they were able to extract the most important features to affect prediction results.

Getting data to research suicide can be difficult. However, a data set that is worthy of examination is the Millennium Cohort Study (MCS), otherwise known as the "Child of the New Century" (Connelly & Platt, 2014). The study was conducted by University College London's Centre for Longitudinal Studies (CLS) (Centre for Longitudinal Studies, 2023). The study began with an original sample of 18,818 cohort members from England, Scotland, Wales, and Northern Ireland. These cohort members were studied at 7 different time points (referred to as "sweeps" by the CLS) ranging from age 9 to age 17. At various points throughout these waves of data collection, the data collected could be either by parents/carers/co-residents and/or by the participant (who they referred to as the “cohort member”, or “CM”). To the best of the author’s knowledge, the only research done so far to predict suicidality from the MCS was done by Jankowsky et al., 2023. These researchers used data from wave 6 and wave 7 of the MCS and used 638 variables. They used logistic regressions, elastic net regressions, and GBM.

Given the increase in incidence of suicide among adolescents, the impact that improving predictions of suicide has on research and lives, and the promise that AI and machine learning hold for predicting suicide, there are many compelling reasons to continue to add to the literature. The author seeks to do just that through the present work. In particular, we aim to extend the work done by Jankowsky et al. (2023).

Similarly, to Jankowsky et al. (2023), we aim to classify individuals as positive or negative cases for attempted suicide reported at age 17 (note that participants were asked if they had attempted suicide in the past 12 months). However, we will pursue modeling techniques different from those that Jankowsky et al. employed in an effort to extend their research and see how well other model types perform. Specifically, we will use random forests, classification trees, K-nearest neighbors, and generalized linear models. We aim to identify the classification model types that perform the best. Another goal of ours is to identify the variables most important in predicting suicidality. We will first report on the materials and methods used for the analyses, and then will share the results, and will discuss the results in light of the literature. We will end with limitations, future directions, and concluding remarks.

MATERIALS

The data used in this research comes from the Millennium Cohort Study (Connelly & Platt, 2014).

Prior to the author beginning working on this TFM work and doing an internship with the EPISAM Research Group, members of the research group had analyzed the MCS data and had published papers of their results. Consequently, the author was given R files in which a significant amount of data wrangling had been performed on the original data set and began her data analysis using these materials that she was given.

The data analysis for this project was performed in RStudio using R. A repository with the code used and output files generated for this project can be found at the author's GitHub page ("nfrontero20") in the repository "Masters-Capstone".

METHODS

Preprocessing:

MCS data from waves 1-6 were read into R from a .csv file that the author was given access to and the MCS data from wave 7 was read in from an R object that the author was given access to. Then, only \`MCSID\` (the unique identifier for each participant) and \`GCSUIC00\` (a binary variable coding the participant's response to the question "Have you ever hurt yourself on purpose in an attempt to end your life?" were selected from the wave 7 data as \`outcome_data\`. \`GCSUIC00\` was releveled so that "No" responses were coded as 0's and "Yes" responses were coded as 1's and then was turned into a factor.

Regarding predictor variables, we decided to focus on variables from wave 6 (as opposed to predictors from another wave prior to wave 6, or wave 7). There were a number of reasons behind this decision. First, we were specifically interested in identifying risk factors present at an age that, as the Millennium Cohort Study refers to age 14 as being, is a "key transitional stage between childhood and adulthood" (Centre for Longitudinal Studies - UCL Institute of Education, 2019). Moreover, in wave 6, participants were asked to provide more data than in prior sweeps, which made working with this wave of data appealing (Centre for Longitudinal Studies - UCL Institute of Education, 2019). Additionally, unlike the other sweeps, in wave 6, parents were asked to complete a cognitive assessment that measures psychological distress (the Kessler K-6 scale). Given that parental psychological distress is associated with suicidality in children (Zhu et al., 2023), we wanted to incorporate variables pertaining to the parental Kessler K-6 scale responses in our models.

We performed a \`full_join\` (from the \`dplyr\` R package) to merge the outcome data and the wave 6 predictor data on the \`MCSID\` identifier column. We chose to perform a \`full_join\` (which retained all rows from the predictor and the outcome data) as opposed to a different type of join so that we could maximize the number of observations. We knew that the ratio of negative (0) to positive (1) cases for \`GCSUIC00\` was approximately 14:1 (93.33% negative cases, 6.7% positive cases). Consequently, we knew that the data provided many more negative cases for the models to learn from than positive cases. In order to maximize the ability of the models to learn from the data in both cases, and given the lack of positive cases in the data, we wanted to maximize the number of observations in the data set.

By performing a \`full_join\`, we had 18,936 rows before further filtering (which will be detailed in the following paragraphs). If we were to have performed a \`left_join\` in which all rows from the outcome data were preserved and only matching rows from the predictor data were retained, we would have had 10,917 rows. Therefore, performing a \`full_join\` as opposed to this type of \`left_join\` increased our number of data points by approximately 70%. If we were to have performed a \`left_join\` in which all rows from the predictor data were preserved and only matching rows from the outcome data were retained, we would have had 18,552 rows.

Consequently, performing a `full_join` resulted in approximately 2% more data points compared to this second case.

Since we knew we would be imputing data, retaining as many rows of data as possible that communicated helpful information was deemed valuable, so we performed a `full_join`.

Next, the degree of missing data in the merged dataset was calculated and it was found that 64% of the data was missing. The data set had 18,936 rows. Rows with all NA data were removed, resulting in a loss of 6,669 rows and leaving 12,267 rows remaining. Even after the removal of the rows with all NA's, there was still 44% missingness across the data set.

At this point, we decided to use the percent of missingness found in the columns (variables) as a criterion to decide which variables to include in the analyses. We did this for a few reasons. For one, we knew we would need to inevitably impute data, since classification methods require no missing data in the data sets, but we didn't want to have to impute variables that exhibited such high levels of missingness. Also, we did not want to run classification techniques on 241 variables, which is how many variables came from wave 6.

Consequently, we calculated the percentage of missingness across all of the columns. The percentage of missingness ranged from 7.38% to 100%. We identified the columns with less than 15.00% missingness. However, after this step, there were still 116 variables, which were more than we wanted to use in our analyses. As a result, the author solicited the expertise of the professional tutor to help her identify the variables that would be most important and interesting to study further based on his domain knowledge. This reduced the number of variables to 86. By filtering the data set down to these 86 variables, the data set only had 12% missingness. Instead of listing all 86 variables and their definitions, a smaller subset of all the variables that up being important to some degree and their definitions and data types can be found in Table 1.

Next, the data was imputed using the `mice` function from the `mice `R package. The distributions of the imputed data sets and the original data sets were compared using the `densityplot` function from the `lattice` R package. An .pdf file "densityplots.pdf" where all of the density plots are visualized can be found in the associated GitHub repository. Note that as a result of the imputation, the resulting data set had no missingness.

After imputing, the variables that represented factors were turned into the factor data type in R. Then, another version of the data set was created where additionally, the variables that represented elements of psychological scales were scaled in R.

We knew that we would need a way to perform cross-validation to see how well our models performed and to see how well they generalized. We chose to employ the single hold-out random subsampling technique. Consequently, we created test and train sets for both the original data set (with no scaling) and the data set with scaled variables. We chose to perform a 70% train/30% test split as this falls within the guidelines for a test/train split ratio (Berrar, 2018). It was ensured that a stratified sample with regard to the outcome variable was selected in both versions of the train/test splits. See Table 2 in the Appendix for the breakdown of the incidence of the outcome variables in these data sets.

Modeling:

Note that a complete breakdown of the parameters, data sets (scaled or unscaled), and variables used in each model can be found in the Appendix in Table 3.

Additionally, note that one of the KNN models, two of the classification trees, and 1 of the generalized linear models were created not from all 86 of the variables but rather from a subset of 18 variables that were identified as important through the random forest modelling process (specifically, the mean decrease accuracy variable importance plots, not the mean decrease Gini variable importance plots). We did this because we were aware that having so many variables could introduce noise into the data. We were interested in seeing if the models performed better with less noise.

Finally, note that hyperparameter tuning was done manually as opposed to in a more sophisticated, technical approach. This is discussed in the limitations section.

Random Forest:

Four random forest models were created using the `randomForest` function from the `randomForest` package in R. All modeling was done on the unscaled data. Each model was created by making predictions from `train` and with different parameters of `ntree` and `mtry`. Note that `ntree` refers to the number of trees created and `mtry` refers to the number of variables randomly sampled as candidates at each split. The models were then used to predict the classes of the train data, solely for the purpose of seeing if it was indeed possible to classify the data, and then to predict the classes of the test data. Additionally, variable importance plots, which showed the top 15 most important variables with regard to mean decrease of accuracy and mean decrease of the Gini index, were created for each model. The models were created using the following specifications: Model A: `ntree = 1000`, `mtry = 30`; Model B: `ntree = 1000`, `mtry = 15`; Model C: `ntree = 500`, `mtry = 5`; Model D: `ntree = 500`, `mtry = 25`.

Regarding Model A, we decided to choose a standard number of trees (`ntree = 1000`) and a relatively high value for `mtry`. We knew that an advantage of using large `mtry` values is that it helps decision trees select features that are important to the outcome variable. Because we were interested in creating trees that did a good job at identifying the important variables, we chose to set `mtry = 30` to a relatively high value for Model A. For Model B, we were interested in seeing if we could see a significant difference in the model performance compared to Model A. We know that high `mtry` values have the disadvantage of reducing randomness, which affects model performance. Consequently, we reduced `mtry` for Model B.

The same logic applied for Model C and Model D. Note that we arbitrarily chose to use the lower `mtry` value for the first model we created (Model C) and the higher value for the second model we created (Model D).

KNN:

Three K nearest neighbors models were created using the `knn` function from the `class` package in R. Models A and B were trained on the training data from the `train_scaling` data set and then the model was used to predict the classes of the test data in `test_scaling`. The specifications for Models A and B were as follows: Model A: `k = 3`; Model B: `k = 10`. Model C was created using the 18 variables found across the four random forest models to be the most important variables per the mean decrease accuracy variable importance plots. Model C was run with `k = 5`. Note that k refers to the number of neighbors considered.

Classification Trees:

We created 4 classification trees using the `rpart` function from the `rpart` R package. Model A was constructed from the `train` data set with parameters `minsplit = 2`, `minbucket = 1`, and `cp = 0.001`. Note that `minsplit` refers to the minimum number of observations that must exist in a node in order for a split to be attempted, `minbucket` to the minimum number of observations in any terminal leaf node, and `cp` to “complexity parameter”, signifying that any split that does not decrease the overall lack of fit by a factor of `cp` won’t be attempted. Model B was a tuned version of Model A, with `minsplit = 2` and `cp = 0.001` like in the case of Model A, but with `minbucket = 5` instead of `minbucket = 1`. Model C was created using the aforementioned 18 variables identified from the random forest variable mean decrease accuracy importance plots to be most important. Model C was created with parameters `minsplit = 2`, `minbucket = 3`, and `cp = 0.001`. Model D was a tuned version of Model C, with parameters `minsplit = 3`, `minbucket = 5`, and `cp = 0.001`.

We chose to create Model A with the aforementioned parameters because we were interested in seeing how well a very complex classification tree model would perform and we knew that specifying `minsplit = 2` and `minbucket = 1` would create a complex tree. Additionally, note that for all models, we specified `cp = 0.001` because when we set `cp` to less than 0.001, the trees that were created only had a single node and were not actually trees. We created Model B with the aim of creating a slightly less complex tree than Model A, which we achieved by setting `minbucket = 5`. The difference in complexity between Model A and B can be seen in the apparent difference in the appearance of these trees in Figure 5 and Figure 6, respectively.

With Model C, we wanted to create a tree that was less complex than Model A but more complex than Model B. As Figure 5 shows, very complex trees are nearly impossible to visualize since they are dense and vast. With Model D, we were curious to see how increasing `minsplit` would affect model performance and so we chose `minsplit = 3`, which is the highest `minsplit` used in any of the models.

Generalized Linear Models:

We created two generalized linear models using the scaled data. We used the `glm` function from the `stats` package. The models were created on the `train_scaling` data and then tested on the `test_scaling` data. Model A was created from all 86 of the variables while Model B was created only from the 18 variables identified as important from the random forest mean decrease accuracy variable importance plots.

RESULTS

Note that a breakdown of each model type and metrics for its performance can be found in the Appendix in Table 4. Additionally, a breakdown of variables identified as important from the models can be found in Table 5.

Random Forest:

Overall, the random forest models all performed quite similarly even though we used different parameters for each one. They yielded accuracies that were quite high and very similar to one another (minimum 93.26%, maximum 94.34%). They also yielded high true negative rates (minimum 99.30%, maximum 99.71%), as well as low false positive rates (minimum 0.29%, maximum 0.70%). However, the true positive rates were low (minimum 4.45%, maximum 8.50%) and the false negative rates were very high (minimum 91.50%, maximum

95.55%). The models' high true negative rates and low false positive rates indicate that they were able to identify true negatives well and very infrequently misclassified them as positives. However, the models' low true positive rates and high false negative rates indicate that they did not do a good job at correctly classifying positive cases.

Note that we did not see a significant difference in model performance between Model A and Model B despite decreasing `mtry` from Model A to B, nor did we see a significant difference in model performance between Model C and D despite decreasing `mtry` from Model C to D.

The top 15 important variables from each model were identified via the mean decrease accuracy variable importance plots. These plots can be seen in the Appendix in Figures 1-4. When we look at the top 15 variables with regard to mean decrease in accuracy across the four models, we find 18 variables that arise as important.

Table 5 in the Appendix shows the top 15 most important variables with their ranks for Models A-D. Additionally, we calculated the average ranking for each of these 18 variables by averaging the rankings across all four models. Table 5 features these average rankings. The important variables, in order of average ranking highest to lowest, are the following:

`SDQ_diff.sw6`, `self_harm.sw6`, `feel_hatred.sw6`, `SDQ_peer.sw6`, `SDQ_conduct.sw6`, `SDQ_emot_s.sw6`, `kessler_k6_main.sw6`, `feel_good_others.sw6`, `feel_not_enjoy.sw6`, `SDQ_prosoc.sw6`, `cigarette_freq.sw6`, `cyberbull.sw6`, `feel_no_good.sw6`, `activitiy_status_main.sw6`, `feel_wrong.sw6`, `feel_not_concent.sw6`, `feel_restless.sw6`, and `sex_CM.sw6`.

These variables will be interpreted in the discussion section. For reference on the meaning of these variables, see Appendix Table 1.

KNN:

The KNN models all performed quite similarly despite using a different value of *k* for each one and despite Model C being created from the 18 variables deemed important per the random forest models whereas Models B and C were created using all of the variables. The three models yielded accuracies that were rather high and very similar to one another (minimum 92.45%, maximum 93.23%). They also yielded high true negative rates (minimum 98.60%, maximum 99.89%), as well as low false positive rates (minimum 0.32%, maximum 1.4%).

However, the true positive rates were low (minimum 4.42%, maximum 8.85%) and the false negative rates were very high (minimum 91.16%, maximum 95.58%). The models' high true negative rates and low false positive rates indicate that models were able to identify true negatives well and very infrequently misclassified them as positives. However, the models' low true positive rates and high false negative rates indicate that the models did not do a good job at correctly classifying positive cases.

Classification Trees:

The classification tree models all performed basically the same despite Models A and B being created from all of variables and Models C and D being created from the 18 variables suggested as important from the random forest modeling. Their accuracy was good (minimum 90.52%, maximum 92.23%) although slightly lower than the accuracies seen for the other model types. The true negative rates were all quite high (minimum 95.75%, maximum 97.64%), indicating that the models were able to identify true negatives well, although these values were slightly lower than those seen for the other model types. The classification tree models performed better than any other model type in terms of the true positive rate (minimum 15.38%, maximum 17.81%), although this rate is still not that impressive. Additionally, these models had the lowest false negative rates (minimum 83.00%, maximum 84.62%). However, the false positive rates from the classification trees were the highest across all the model types (minimum 2.36%, maximum 2.71%).

Although we were intentional in tuning the parameters to see if we could improve model performance, the results indicate that our tuning did not induce any improved model performance as all of these models performed about the same. However, it should be noted that Model C (`minsplit = 2`, `minbucket = 3`, `cp = 0.001`) performed the best.

The classification trees can be seen in the Appendix in Figures 5-8. Note that due to the number of variables used, they are nearly impossible to read. However, they were included in the Appendix for reference.

Summaries of each classification tree were generated and stored in .txt files which can be found in the source files for this project. These summary files contain the importance values assigned to the variables that resulted as important in creating the trees. Table 6 in the Appendix shows the importance value assigned to each variable that arose as important for Models A – D. Using these values, we rank ordered the variables in terms of their importance for each model. We also calculated an average ranking for each variable by averaging across the rankings for each model. Table 7 in the Appendix shows includes the variable rankings for each model as well as the average rankings. In the Appendix, we also include breakout tables for each model (Tables 8-11). Each of these tables shows the importance value assigned to each variable that appeared as important as well as the ranking for that variable.

When we look at the average rankings listed in Table 7, we see five variables that show up as the most important: `self_harm.sw6` (rank = 1), `SDQ_diff.sw6` (rank = 2), `SDQ_peer.sw6` (rank = 3), `kessler_k6_main.sw6` (rank = 4th place, tie), `SDQ_emot_s.sw6` (rank = 4th place, tie). The significance of these variables will be interpreted in the discussion.

Generalized Linear Models:

GLM Model A and B performed very similarly overall even though Model A was created from 86 variables and Model B was created from 18 variables. They had high accuracies (93.21% and 93.37% for Model A and B, respectively). Additionally, the specificities, or true negative rates, were high (99.13% and 99.45% for Model A and B, respectively), indicating that the models were able to identify true negatives well.

The models did a poor job at identifying true positives well, with Model A having a sensitivity of 10.84% and Model B having a sensitivity of 8.91%. These rates are generally better than the sensitivities seen for the random forest and KNN models but are worse than those seen for the classification trees. GLM Model A and B had many false negatives (89.07% for Model A and 91.09% for Model B), indicating that often, there were cases classified as 0's that were actually 1's.

It should be noted that the false positive rates were very low (0.87% and 0.55% for Model A and B, respectively) and were comparable to those seen in the random forest and classification tree models. The false positive rate indicates that there were relatively few times when cases were predicted to be 1 and they were actually 0.

.txt files that contain the summary material can be found in the source files for this project. When we look at these files, we can see the variables that are significant. Using this output data, we created Table 12 (see Appendix), which shows the variables that arose as significant at the level of $p < 0.05$ or less.

We see that `self_harm.sw6` ($p < 0.001$) and `sex_CM.sw6` ($p < 0.001$ for Model A, $p < 0.01$ for Model B) were significant in both models.

Then, in Model A, there are significant variables that appear that do not appear as significant in Model B. These variables are `videogame_weekd_hours.sw6` ($p < 0.001$), `alcohol_ever.sw6` ($p < 0.001$), `natural_mother_alive.sw6` ($p < 0.01$), `feel_no_love.sw6` ($p < 0.01$), `antisoc_caution.sw6` ($p < 0.05$), and `feel_bad.sw6` ($p < 0.05$).

In Model B, the variables that appear significant are intercept ($p < 0.001$), `activitiy_status_main.sw6` ($p < 0.01$), `sex_CM.sw6` ($p < 0.01$), and `cigarette_freq.sw6` ($p < 0.05$). Note that `sex_CM.sw6` was coded as 1 = male, 2 = female. Therefore, we interpret the fact that the intercept is significant to indicate that females showed a higher risk of suicide attempt.

The discrepancy between the significant variables that arise for Model A and Model B can, at least in part, be explained by the fact that Model A contained all the variables and Model B only contained the 18 variables deemed most important by the random forest models.

Summary:

Across all model types and all subsets of data, the models performed similarly. The models excelled at accuracy (minimum 90.52%, maximum 94.24%). They also excelled with regard to the true negative rate/specificity (minimum 95.75%, maximum 99.54%), and false positive rate (minimum 0.29%, maximum 4.25%), which indicates that they were able to classify 0's well. However, the models performed poorly with regard to true positive rate/sensitivity (minimum 4.42%, maximum 17.00%) and the false negative rate (minimum 80.57%, maximum 95.58%), indicating that they were unable to classify 1's well.

Across the four model types, the classification trees were best able to identify 1's, as indicated by those models yielding true positive rates/sensitivities higher than those of the other model types and false negative rates lower than those of the other model types. However, the classification trees yielded generally lower values for true negative rate/specificity and generally higher values for the false positive rate.

In comparing the four model types, the classification models also arguably performed the best with regard to identifying 0's. While these models yielded slightly low true negative rates compared to the other model types (in the 95% - 97% range as compared to 99% for the other model types), the false negative rates were noticeably lower in the classification model metrics (in the 80% - 84% range, as compared to 89% - 95% range seen across the other models).

In picking an overall best model, we put forth classification tree Model C as the best. This model has the highest true positive rate out of any of the models and the lowest false negative rate. Additionally, it has a true negative rate only 2% lower than the other models, and it is still quite high (97.35%) and a false positive rate that is relatively low, even though it is one of the higher rates across all model types (2.65%).

With regard to model creation, we theorized that the models with less variables may perform better due to less noise. However, there was no significant difference in model performance between the models constructed from 86 variables and those constructed from 18 variables. This tells us that our attempts to reduce noise by using less variables were unsuccessful, and also, that more variables didn't necessarily translate into better model performance. Additionally, the results indicate that our manual tuning of hyperparameters was unsuccessful in generating meaningful differences in model performance.

With regard to significant variables, 18 variables appeared throughout all four top 15 mean decrease accuracy variable importance plots for the random forest models. These variables were used to create KNN Model C, classification tree Models C and D as well as GLM Model B. Table 5 summarizes the important variables from the random forest models, Tables 6-11 pertain to the important variables identified from the classification tree models, and Table 12 contains information on the important variables found from the generalized linear models.

Table 13 in the Appendix contains a summary of the most important variables found across the various model types. For each model type, we highlight in orange the value associated with the top 5 variable. In the case of the random forest and classification trees, this value is the average ranking associated with the top 5 variable, and for the generalized linear models, this value is the *p* value significance level. For the generalized linear models, we included the variables that were significant at the lowest values of *p.* Note that in the case of random forest, the averaging and rounding used to produce the average ranking resulted in the fourth most important variable having a ranking of 5. Also, six cells are highlighted because there was a tie for the 5th most important variable. Note that in the cases where a variable was a top 5 most important variables for one model type but not for the other model type(s), we listed any pertinent data available for the other model type(s) but did not highlight this information in orange. For example, `alcohol_ever.sw6` was one of the top 5 most important variables per the generalized linear models, but this variable was not one of the top 5 most important variables for random forest nor for classification trees. However, this variable was ranked 8th in variable importance for the classification tree models so we included that information (there was no ranking data to report for random forest since this variable was not ranked in the top 15 across all random forest models).

The following variables are featured in Table 13:

`self_harm.sw6`, `SDQ_diff.sw6`, SDQ_peer.sw6`, `SDQ_emot_s.sw6`, `Kessler_k6_main.sw6`, feel_hatred.sw6`, `SDQ_conduct.sw6`, feel_trust.sw6`, `videogame_weekd_hours.sw6`, `alcohol_ever.sw6`, `sex_CM.sw6`.

These important variables will be interpreted in the discussion.

Note that `self_harm.sw6` was the only variable that appeared as important across all models created: it received the average ranking of #2 across the random forest models; it was on average the #1 ranked important variable for the classification trees; and, it was significant at a value of $p < 0.001$ for both of the generalized linear models.

DISCUSSION

The main goal of this study was to to test the accuracy of varying machine learning models in classifying adolescents with suicidal attempts. We aimed to use different machine learning classification techniques than Jankowsky et al. (2023) used so as to add to the literature. A secondary goal was to identify the classification technique most suitable to classifying the outcome variable among the techniques that we used. Finally, another goal of ours was to identify the variables most implicated and important in the predicted classification of an individual as either a positive or negative case.

To achieve these goals, we created 13 models: 4 random forest models, 3 KNN models, 4 classification trees, and 2 generalized linear models. Most of the models were created using all of 86 of the variables, although four of the models (KNN Model C, and Classification Trees Model C and D, and GLM Model B) were created using the 18 variables gathered from the variable importance plots of the four random forest models to be most important.

Model Performance:

The results indicated that the models generally all performed similarly, even despite some models being created from all 86 variables and some from 18 variables. All models had good accuracy (>90%) and performed well at classifying negative cases, as evidenced by the high true negative rates and the low false positive rates. However, the models struggled in their ability to correctly classify positive cases, as evidenced by the low true positive rates/sensitivities and the high false negative rates.

There can be various goals with models. For example, a priority can be the correct classification of negative cases or instead the correct classification of positive cases. The goal with this study was to prioritize the accurate classification of positive cases. With this goal in mind, we would state that the classification tree models performed the best. These models exhibited higher true positive rates than did the models created by the other model types. In particular, classification tree Model C performed the best with a 19.43% true positive rate.

With regard to prioritizing the model's ability to correctly classify negative cases, there was no single model or type of model that performed the best, although the random forest, KNN, and generalized linear models all performed slightly better than the classification trees with regard to the true negative rate and the false positive rate. However, given that the differential between these models' true negative rates and the classification tree models' true negative rates, as well as the differential between these models' false negative rates and the classification tree models' false negative rates, is only approximately 2%, we would still argue that the classification trees are the best model type on account of how they excelled with correctly classifying positive cases.

Taking all of this into consideration, we put forth classification tree model C as the best model. On the one hand it is unsurprising that the classification trees performed well given that they have qualities that make them stronger modelling techniques as compared to KNN or generalized linear models. For example, classification trees do not require prior preprocessing and are able to handle variable interactions, unlike generalized linear models. Moreover they are robust to outliers and data with different scales. However, we would have expected that the random forest models would perform the best given that random forests are a collection of many trees. A possible explanation for why the classification tree models performed better than the random forest models could potentially be found in how we constructed the random forest trees. It is possible that by choosing to construct our random forest models with relatively high values of `mtry`, the trees were overfit. A future direction could be to keep `mtry` static while we increase the number of trees so as to see how model performance is affected since creating more trees translates to more of a reduction in variance.

Important Variables:

The top 5 variables identified as important for each model type can be seen in Table 13 and their definitions can be found in Table 1. The variables that arose as important make sense in the context of the literature. To start, self-harm (`self_harm.sw6`) was seen as an important variable in predicting suicidality across all models. As Grandclerc et al. (2016), Guertin et al. (2001), and Nock et al. (2006) suggest, self-harm and suicidal behavior are often comorbid with one another. Another variable that appears as significant is past alcohol usage, which is in alignment with the literature: Galaif et al. (2007) highlight in their review paper that substance abuse (including alcohol abuse) is one of the most important risk factors in completed and attempted adolescent suicide.

Videogame usage (`videogame_weekd_hours.sw6`) was also found to be associated with suicidality in this study. In terms of videogames, problematic-gaming has been linked to suicidal ideation (Erevik et al., 2022).

Additionally, having parents who scored high on the Kessler scale (`kessler_k6_main.sw6`) was also associated with positive outcomes in this study. The Kessler scale is a psychological distress scale (Kessler & Mroczek, 1992). Zhu et al. (2023) posit that there is a relationship between parental psychological distress and self-harm and suicide attempts in a child.

Sex of the participant (`sex_CM.sw6` was another significant variable that arose, with females being at higher risk. This is in alignment with the literature: Canetto & Sakinofsky (1998) discuss the “gender paradox” with suicide, which is that while more females than males attempt, males account for more suicides than do females.

High scores on various SDQ scales were also associated with suicide attempts in this study, specifically, `SDQ_peer.sw6` (problems with peers), `SDQ_emot_s.sw6` (emotional symptoms), and `SDQ_conduct.sw6` (conduct problems). `SDQ_diff.sw6`, which represents the summation of points on the aforementioned three scales as well as the hyperactivity/inattention scale (Goodman, 2001), was also associated with suicide attempts. This is supported in the literature, with Burón et al. (2011) finding that suicide attempts and suicidal ideation were related with the overall difficulties score of the SDQ, mainly on emotional symptoms and behavioral symptoms. With regard to our finding that a high value on the SDQ peer scale is associated with suicidality, the literature provides a plausible explanation. Kleiman & Liu (2013) found that social support is a protective factor against suicide attempts, with individuals with more social support over 30% less likely to have a lifetime suicide attempt compared to those with low social support. This finding also can be used to explain why `feel_trust.sw6` was found to be an important variable, as trust is a component of social support.

`feel_hatred.sw6`("I hated myself") was also found in our study to be associated with suicide attempts. There is support for this finding in the literature, with Lieberman et al. (2023) finding in individuals with bulimia nervosa that self-hate was significantly related with suicidal behavior, and with Turnell et al. (2019) finding that high self-hate was associated with suicidal ideation.

There are other variables that arose as important but that weren’t in the top 5 most important variable grouping across all models. We will discuss a few of them here. For example, cigarette frequency (ranked 11th for the random forest models, ranked 7th for the classification tree models, and with significance of $p < 0.05$ for GLM Model B), was found to be associated with suicide attempts by Peprah et al. (2023). Bullying was also found to be associated with suicide attempts in our study. `brother_bull.sw6` (how often the participant got bullied by siblings) and `cyberbull.sw6` (how often the participant was bullied online) were tied with other variables for 5th place in the classification tree average rankings. With regard to bullying, the literature shows that youth who report frequently bullying others and being bullied are at increased risk of suicide-related behavior (Center for Disease Control and Prevention, 2014).

Antisocial behavior was another variable identified as significant through this study. `antisoc_police.sw6` ("Has CM been stopped or questioned by police") and `antisoc_rude12.sw6` ("Past 12 months: CM has been complained for being rude/noisy in public?") were tied for 8th place average ranking for the classification trees. Additionally, for GLM, `antisoc_caution.sw6` ("Has CM ever been given a formal warning or caution from police") was significant at the level of $p < 0.05$ for GLM Model A. The literature shows that antisocial behavior has been found to be strongly related to adolescent suicide (Marttunen et al., 1994).

Limitations and Future Directions:

There are a number of limitations with this work that should be considered. The first topic has to do with the way that missingness and imputing the data were handled. The rationales behind removing rows with all NA's, selecting columns based on low levels of missingness, and imputing the data have already been stated in the Methods section. Therefore, we bring up this topic in the limitations section simply to point out that a more rigorous approach could have been taken to handling missing data and imputing data. This is perhaps the most critical limitation that we have to put forth for this report. For example, the correlations between variables could have been examined prior to imputation and missing data patterns could have been looked for. However, these steps weren't taken due to the vast number of variables we were working with. Also, we could have fit a model to each of the imputed data sets and pooled the results together when modeling.

Another limitation of high level importance in this work is the fact that model tuning was not done through sophisticated hyperparameter tuning approach and that rather, a manual approach was taken. Consequently, a future direction would be to perform hyperparameter tuning on our models.

Additionally, we would like to acknowledge that the foci of this project changed somewhat from the writing of the initial proposal to the final product. Initially, we were interested in predicting suicide attempts as well as non-suicidal self-harm risk, hence the title of this project, "Use of supervised learning to predict suicidal and non-suicidal self-harm risk". However, after submitting the proposal, we discovered that it would be more difficult than originally conceived to predict non-suicidal self-harm risk. There were 8 variables associated with self-harm and in order to perform classification techniques, we would have needed to create an index for self-harm. For example, this could have taken the form of assigning a 1 to an individual who exhibited 3 or more types of self-harm and a 0 to those who exhibited less than 3 types of self-harm. However, we decided that this would be a very imperfect way to predict self-harm risk, and as a result we did not feel comfortable proceeding with the analyses.

Rather than take a considerable amount of liberties in pursuing predictions of self-harm risk, we decided to amplify the work done to predict suicide attempts. Therefore, even though we originally only planned on performing two classification techniques, we expanded the work to include four classification techniques.

Something else that ought to be discussed, although it is not a limitation, is the lack of positive cases in the response variable. Note that the lack of positive cases is inherent to the phenomenon of suicide. Recall that negative cases account for 93.32% of the data while positive cases account for only 6.68% and that in both test sets, there were only 247 positive cases. This vast difference can likely explain all of the models' poor performance with regard to maximizing the true positive rate and minimizing the false negative rate. A future direction could be to impute the data in such a way as to artificially create more positive cases. This would allow for the models to have a better chance at correctly classifying the positive cases, and in turn would allow us to learn more about the model type that may be the most helpful for predicting suicidality, as well as the variables most implicated.

In this same vein, we have to take the variables deemed to be important (per the random forest models, the classification trees, and the generalized linear model) with some caution, given that there were indeed so few positive cases. It is possible that the important variables would change if the analyses were run with more positive cases.

An additional limitation has to do with the fact that more models could have been run in order to diversify the variables used in creating the models. Most of the models were created with all 86 variables while four of them (KNN Model C, classification trees Model C and D, and GLM Model B) were created from the 18 important variables identified in the random forest models. The approach taken is completely reasonable, and a future direction could be to create even more models. For example, perhaps 10 random subsets of 20 variables could be taken from the 86 variables. The benefit in reducing the number of variables used to create models would be to reduce runtime for creating random forest models, to simplify the modeling (instead of having so many variables) and to see if the same variables appear as significant across the subsets that they show up in.

Moreover, a future direction could be to include more variables from wave 6, and to include variables from wave 7 as well, and perhaps also data from earlier waves.

Another future direction could be to perform a Generalized Additive Model (GAM) for the purpose of supervised learning. Also, a potential future direction that could be of interest would be to perform unsupervised learning, for example KNN clustering, hierarchical clustering, or principal component analysis to see what sort of groups are formed, and then to see the proportions of positive and negative cases found in each cluster.

CONCLUSION

Suicide occurs globally and is on the rise. Prevention and risk-identification efforts are essential to reducing the prevalence of suicides. The present work sought to extend the work done by Jankowsky et al.'s (2023) their exploration of predicting suicidality from the MCS data set. The present work differed from Jankowsky et al.'s (2023) work in the decision to focus on a relatively small subset of variables and with regard to the classification methods employed. Three of the four methods that we employed – KNN, classification trees, and generalized linear models – were not used by Jankowsky et al. (2023). This work adds to the literature of using machine learning techniques to predict suicidality and provides many future directions for additional research. Additionally, the present study adds to the body of literature about risk factors for suicidality through the identification of important variables.

BIBLIOGRAPHY

Barber, C. W., & Miller, M. J. (2014). Reducing a suicidal person's access to lethal means of suicide: A research agenda. *American Journal of Preventive Medicine*, *47*(3 Suppl 2), S264-272. https://doi.org/10.1016/j.amepre.2014.05.028

Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnousi, F. (2020). Artificial Intelligence and Suicide Prevention: A Systematic Review of Machine Learning Investigations. *International Journal of Environmental Research and Public Health*, *17*(16), 5929. https://doi.org/10.3390/ijerph17165929

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*.

Bertolote, J. M., & Fleischmann, A. (2002). Suicide and psychiatric diagnosis: A worldwide perspective. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *1*(3), 181–185.

Brown, S., & Seals, J. (2019). Intimate partner problems and suicide: Are we missing the violence? *Journal of Injury and Violence Research*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6420923/

Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders*, *245*, 869–884. https://doi.org/10.1016/j.jad.2018.11.073

Burón, P., Al-Halabí, S., Díaz-Mesa, E., Garrido, M., Galván, G., Rancaño, J. L., Casares, M. J., García-Portilla, P., Sáiz, P., & Bobes, J. (2011). Suicide attempts and suicide ideation in adolescents: SDQ scores in the Spanish sample of "saving and empowering young lives in Europe" (SEYLE) project. European Psychiatry, 26(S2), 1609–1609. https://doi.org/10.1016/S0924-9338(11)73313-2

Canetto, S. S., & Sakinofsky, I. (1998). The gender paradox in suicide. *Suicide and Life-Threatening Behavior*. https://pubmed.ncbi.nlm.nih.gov/9560163/

Center for Disease Control and Prevention. (2014). *The Relationship Between Bullying and Suicide: What We Know and What it Means for Schools*. http://www.cdc.gov/violenceprevention/pdf/bullying-suicide-translation-final-a.pdf

Center for Disease Control and Prevention. (2023, November 2). *Risk and Protective Factors: Suicide*. https://www.cdc.gov/suicide/factors/index.html

Centre for Longitudinal Studies. (2023). *Millennium Cohort Study*. https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/

Centre for Longitudinal Studies - UCL Institute of Education. (2019). *Millenium Cohort Study Seventh Sweep (MCS7) Technical Report*. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://cls.ucl.ac.uk/wp-content/uploads/2020/01/MCS7_Technical_Report.pdf

Connelly, R., & Platt, L. (2014). Cohort profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology*, *43*(6), 1719–1725. https://doi.org/10.1093/ije/dyu001 de la Torre-Luque, A., Essau, C. A., Lara, E., Leal-Leturia, I., & Borges, G. (2022). Childhood emotional dysregulation paths for suicide-related behaviour engagement in adolescence. *European Child & Adolescent Psychiatry*. https://doi.org/10.1007/s00787-022-02111-6

Erevik, E. K., Landrø, H., Mattson, Å. L., Kristensen, J. H., Kaur, P., & Pallesen, S. (2022). Problem gaming and suicidality: A systematic literature review. *Addictive Behaviors Reports*, *15*. https://doi.org/10.1016/j.abrep.2022.100419

Fonseca-Pedrero, E., & Pérez de Albéniz, A. (2020). Assessment of Suicidal Behavior in Adolescents: The Paykel Suicide Scale. *The Psychologist Papers*.

Fonseka, T., Bhat, V., & Kennedy, S. H. (2019). The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors—PubMed. *Australian & New Zealand Journal of Psychiatry*. https://pubmed.ncbi.nlm.nih.gov/31347389/

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, *143*(2), 187–232. https://doi.org/10.1037/bul0000084

Galaif, E. R., Sussman, S., Newcomb, M. D., & Locke, T. F. (2007). Suicidality, depression, and alcohol use among adolescents: A review of empirical findings. *International Journal of Adolescent Medicine and Health*, *19*(1), 27–35.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. Journal of the American Academy of Child and Adolescent Psychiatry, 40(11), 1337–1345. https://doi.org/10.1097/00004583-200111000-00015

Grandclerc, S., De Labrouhe, D., Spodenkiewicz, M., Lachal, J., & Moro, M.-R. (2016). Relations between Nonsuicidal Self-Injury and Suicidal Behavior in Adolescence: A Systematic Review. *PLoS ONE*, *11*(4), e0153760. https://doi.org/10.1371/journal.pone.0153760

Guertin, T., Lloyd-Richardson, E., Spirito, A., & Boergers, J. (2001). Self-mutilative behavior in adolescents who attempt suicide by overdose. *Journal of the American Academy of Child & Adolescent Psychiatry*. https://pubmed.ncbi.nlm.nih.gov/11556630/

Gvion, Y., Levi-Belz, Y., Hadlaczky, G., & Apter, A. (2015). On the role of impulsivity and decision-making in suicidal behavior. *World Journal of Psychiatry*, *5*(3), 255–259. https://doi.org/10.5498/wjp.v5.i3.255

Hawton, K., & Williams, K. (2002). Influences of the media on suicide. *BMJ : British Medical Journal*, *325*(7377), 1374–1375.

Horowitz, L. M., Bridge, J. A., Teach, S. J., Ballard, E., Klima, J., Rosenstein, D. L., Wharff, E. A., Ginnis, K., Cannon, E., Joshi, P., & Pao, M. (2012). Ask Suicide-Screening Questions (ASQ): A brief instrument for the pediatric emergency department. *Archives of Pediatrics & Adolescent Medicine*, *166*(12), 1170–1176. https://doi.org/10.1001/archpediatrics.2012.1276

Jankowsky, K., Steger, D., & Schroeders, U. (2023). Predicting Lifetime Suicide Attempts in a Community Sample of Adolescents Using Machine Learning Algorithms. *Assessment*, 10731911231167490. https://doi.org/10.1177/10731911231167490

Kessler, R., & Mroczek, D. (1992). An update of the development of mental health screening scales for the US National Health Interview Study. *University of Michigan, Survey Research Center of the Institute for Social Research*.

Kleiman, E. M., & Liu, R. T. (2013). Social support as a protective factor in suicide: Findings from two nationally representative samples. Journal of Affective Disorders, 150(2), 540–545. https://doi.org/10.1016/j.jad.2013.01.033

Kumar, V., Sznajder, K. K., & Kumara, S. (2022). Machine learning based suicide prediction and development of suicide vulnerability index for US counties. *Nature Mental Health Research*. https://www.nature.com/articles/s44184-022-00002-x

Lejeune, A., Le Glaz, A., Perron, P.-A., Sebti, J., Baca-Garcia, E., Walter, M., Lemey, C., & Berrouiguet, S. (2022). Artificial intelligence and suicide prevention: A systematic review. *European Psychiatry*, *65*(1), e19. https://doi.org/10.1192/j.eurpsy.2022.8

Lieberman, A., Robison, M., Wonderlich, S. A., Crosby, R. D., Mitchell, J. E., Crow, S. J., Peterson, C. B., Le Grange, D., Bardone-Cone, A. M., Kolden, G., & Joiner, T. E. (2023). Self-hate, dissociation, and suicidal behavior in bulimia nervosa. Journal of Affective Disorders, 335, 44–48. https://doi.org/10.1016/j.jad.2023.05.015

Mann, J. J., Apter, A., Bertolote, J., Beautrais, A., Currier, D., Haas, A., Hegerl, U., Lonnqvist, J., Malone, K., Marusic, A., Mehlum, L., Patton, G., Phillips, M., Rutz, W., Rihmer, Z., Schmidtke, A., Shaffer, D., Silverman, M., Takahashi, Y., … Hendin, H. (2005). Suicide Prevention Strategies: A Systematic Review. *JAMA*, *294*(16), 2064–2074. https://doi.org/10.1001/jama.294.16.2064

Marttunen, M., Aro, H., Henriksson, M., & Lönnqvist, J. (1994). Antisocial behaviour in adolescent suicide. *Acta Psychiatrica Scandinavica*. https://pubmed.ncbi.nlm.nih.gov/8178674/

Motillon-Toudic, C., Walter, M., Séguin, M., Carrier, J.-D., Berrouiguet, S., & Lemey, C. (2022). Social isolation and suicide risk: Literature review and perspectives. *European Psychiatry*, *65*(1), e65. https://doi.org/10.1192/j.eurpsy.2022.2320

Nock, M. K., Hwang, I., Sampson, N., Kessler, R. C., Angermeyer, M., Beautrais, A., Borges, G., Bromet, E., Bruffaerts, R., Girolamo, G. de, Graaf, R. de, Florescu, S., Gureje, O., Haro, J. M., Hu, C., Huang, Y., Karam, E. G., Kawakami, N., Kovess, V., … Williams, D. R. (2009). Cross-National Analysis of the Associations among Mental Disorders and Suicidal Behavior: Findings from the WHO World Mental Health Surveys. *PLOS Medicine*, *6*(8), e1000123. https://doi.org/10.1371/journal.pmed.1000123

Nock, M. K., Joiner, T. E., Gordon, K. H., Lloyd-Richardson, E., & Prinstein, M. J. (2006). Non-suicidal self-injury among adolescents: Diagnostic correlates and relation to suicide attempts. *Psychiatry Research*, *144*(1), 65–72. https://doi.org/10.1016/j.psychres.2006.05.010

Osman, A., Bagge, C. L., Gutierrez, P. M., Konick, L. C., Kopper, B. A., & Barrios, F. X. (2001). The Suicidal Behaviors Questionnaire-Revised (SBQ-R): Validation with clinical and nonclinical samples. *Assessment*, *8*(4), 443–454. https://doi.org/10.1177/107319110100800409

Park, S., Lee, Y., Youn, T., Byung, S. K., Jong, I. P., Kim, H., Lee, H. C., & Hong, J. P. (2018). Association between level of suicide risk, characteristics of suicide attempts, and mental disorders among suicide attempters. *BMC Public Health*.

Peprah, P., Asare, B. Y.-A., Okwei, R., Agyemang-Duah, W., Osafo, J., Kretchy, I. A., & Gyasi, R. M. (2023). A moderated mediation analysis of the association between smoking and suicide attempts among adolescents in 28 countries. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-32610-8

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., Currier, G. W., Melvin, G. A., Greenhill, L., Shen, S., & Mann, J. J. (2011). The Columbia–Suicide Severity Rating Scale: Initial Validity and Internal Consistency Findings From Three Multisite Studies With Adolescents and Adults. *The American Journal of Psychiatry*, *168*(12), 1266–1277. https://doi.org/10.1176/appi.ajp.2011.10111704

Turnell, A. I., Fassnacht, D. B., Batterham, P. J., Calear, A. L., & Kyrios, M. (2019). The Self-Hate Scale: Development and validation of a brief measure and its relationship to suicidal ideation. *Journal of Affective Disorders*, *245*, 779–787. https://doi.org/10.1016/j.jad.2018.11.047

Wang, L., He, C. Z., Yu, Y. M., Qiu, X. H., Yang, X. X., Qiao, Z. X., Sui, H., Zhu, X. Z., & Yang, Y. J. (2014). Associations between impulsivity, aggression, and suicide in Chinese college students. *BMC Public Health*, *14*(1), 551. https://doi.org/10.1186/1471-2458-14-551

World Health Organization. (2021a). *Suicide worldwide in 2019* (p. 7). https://www.who.int/publications-detail-redirect/9789240026643

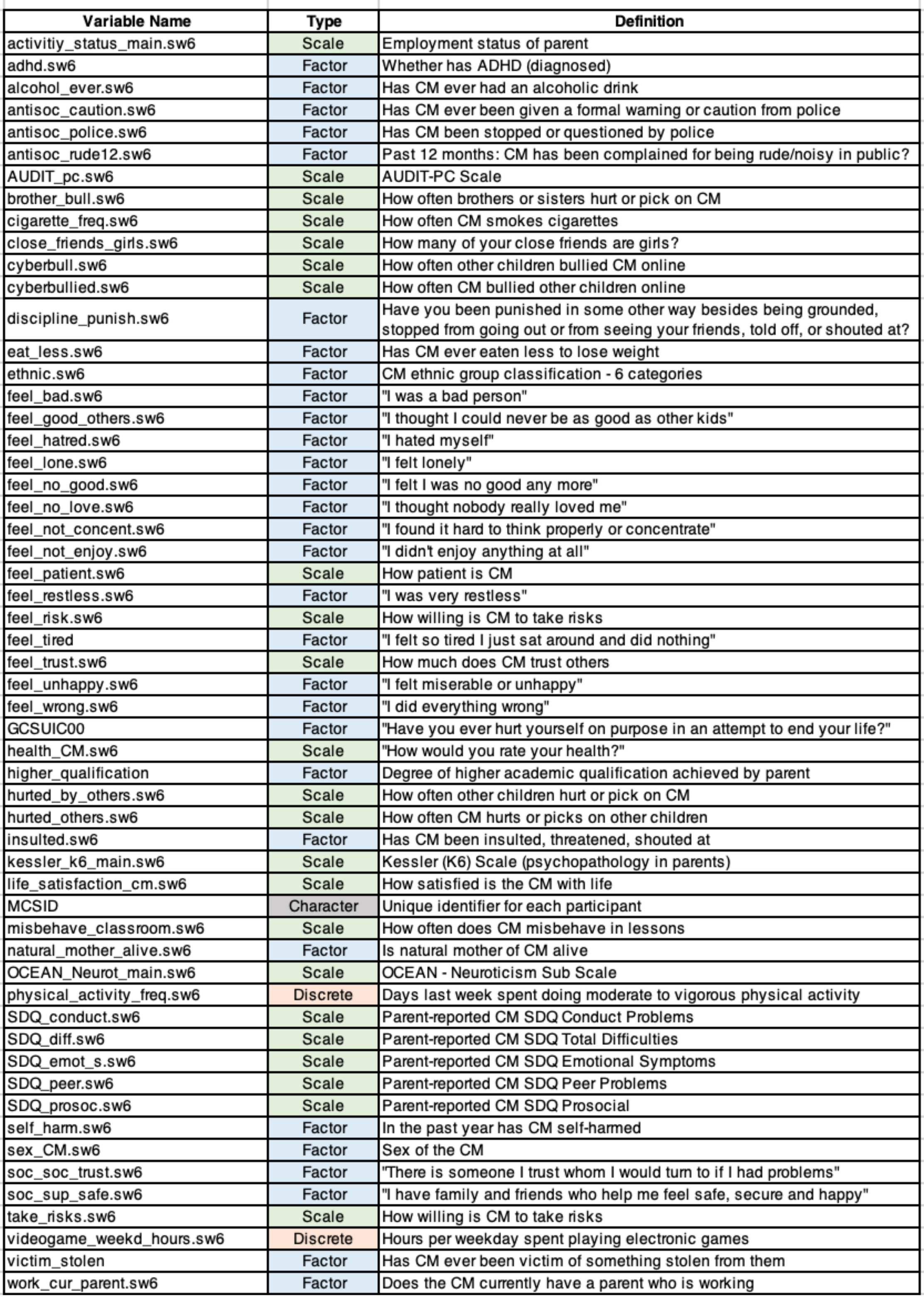
World Health Organization. (2021b). *Mental Health and Substance Use: Suicide data*. https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data

Zhu, X., Griffiths, H., & Murray, A. L. (2023). Co-Developmental Trajectories of Parental Psychological Distress and Child Internalizing and Externalizing Problems in Childhood and Adolescence: Associations with Self-Harm and Suicide Attempts. *Research on Child and Adolescent Psychopathology*, *51*(6), 847–858. https://doi.org/10.1007/s10802-023-01034-3

APPENDIX

Table 1

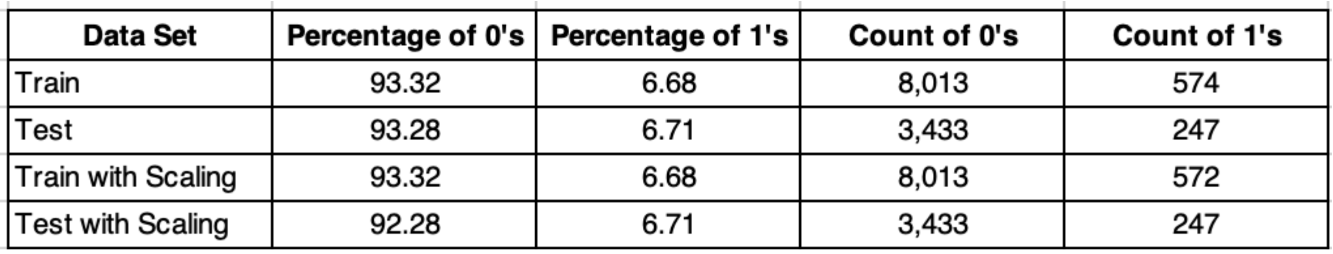
Description of variables (only those that appeared as important in any model)



Variable Name	Type	Definition
activitiy_status_main.sw6	Scale	Employment status of parent
adhd.sw6	Factor	Whether has ADHD (diagnosed)
alcohol_ever.sw6	Factor	Has CM ever had an alcoholic drink
antisoc_caution.sw6	Factor	Has CM ever been given a formal warning or caution from police
antisoc_police.sw6	Factor	Has CM been stopped or questioned by police
antisoc_rude12.sw6	Factor	Past 12 months: CM has been complained for being rude/noisy in public?
AUDIT_pc.sw6	Scale	AUDIT-PC Scale
brother_bull.sw6	Scale	How often brothers or sisters hurt or pick on CM
cigarette_freq.sw6	Scale	How often CM smokes cigarettes
close_friends_girls.sw6	Scale	How many of your close friends are girls?
cyberbull.sw6	Scale	How often other children bullied CM online
cyberbullied.sw6	Scale	How often CM bullied other children online
discipline_punish.sw6	Factor	Have you been punished in some other way besides being grounded, stopped from going out or from seeing your friends, told off, or shouted at?
eat_less.sw6	Factor	Has CM ever eaten less to lose weight
ethnic.sw6	Factor	CM ethnic group classification - 6 categories
feel_bad.sw6	Factor	"I was a bad person"
feel_good_others.sw6	Factor	"I thought I could never be as good as other kids"
feel_hatred.sw6	Factor	"I hated myself"
feel_lone.sw6	Factor	"I felt lonely"
feel_no_good.sw6	Factor	"I felt I was no good any more"
feel_no_love.sw6	Factor	"I thought nobody really loved me"
feel_not_concent.sw6	Factor	"I found it hard to think properly or concentrate"
feel_not_enjoy.sw6	Factor	"I didn't enjoy anything at all"
feel_patient.sw6	Scale	How patient is CM
feel_restless.sw6	Factor	"I was very restless"
feel_risk.sw6	Scale	How willing is CM to take risks
feel_tired	Factor	"I felt so tired I just sat around and did nothing"
feel_trust.sw6	Scale	How much does CM trust others
feel_unhappy.sw6	Factor	"I felt miserable or unhappy"
feel_wrong.sw6	Factor	"I did everything wrong"
GCSUIC00	Factor	"Have you ever hurt yourself on purpose in an attempt to end your life?"
health_CM.sw6	Scale	"How would you rate your health?"
higher_qualification	Factor	Degree of higher academic qualification achieved by parent
hurted_by_others.sw6	Scale	How often other children hurt or pick on CM
hurted_others.sw6	Scale	How often CM hurts or picks on other children
insulted.sw6	Factor	Has CM been insulted, threatened, shouted at
kessler_k6_main.sw6	Scale	Kessler (K6) Scale (psychopathology in parents)
life_satisfaction_cm.sw6	Scale	How satisfied is the CM with life
MCSID	Character	Unique identifier for each participant
misbehave_classroom.sw6	Scale	How often does CM misbehave in lessons
natural_mother_alive.sw6	Factor	Is natural mother of CM alive
OCEAN_Neurot_main.sw6	Scale	OCEAN - Neuroticism Sub Scale
physical_activity_freq.sw6	Discrete	Days last week spent doing moderate to vigorous physical activity
SDQ_conduct.sw6	Scale	Parent-reported CM SDQ Conduct Problems
SDQ_diff.sw6	Scale	Parent-reported CM SDQ Total Difficulties
SDQ_emot_s.sw6	Scale	Parent-reported CM SDQ Emotional Symptoms
SDQ_peer.sw6	Scale	Parent-reported CM SDQ Peer Problems
SDQ_prosoc.sw6	Scale	Parent-reported CM SDQ Prosocial
self_harm.sw6	Factor	In the past year has CM self-harmed
sex_CM.sw6	Factor	Sex of the CM
soc_soc_trust.sw6	Factor	"There is someone I trust whom I would turn to if I had problems"
soc_sup_safe.sw6	Factor	"I have family and friends who help me feel safe, secure and happy"
take_risks.sw6	Scale	How willing is CM to take risks
videogame_weekd_hours.sw6	Discrete	Hours per weekday spent playing electronic games
victim_stolen	Factor	Has CM ever been victim of something stolen from them
work_cur_parent.sw6	Factor	Does the CM currently have a parent who is working

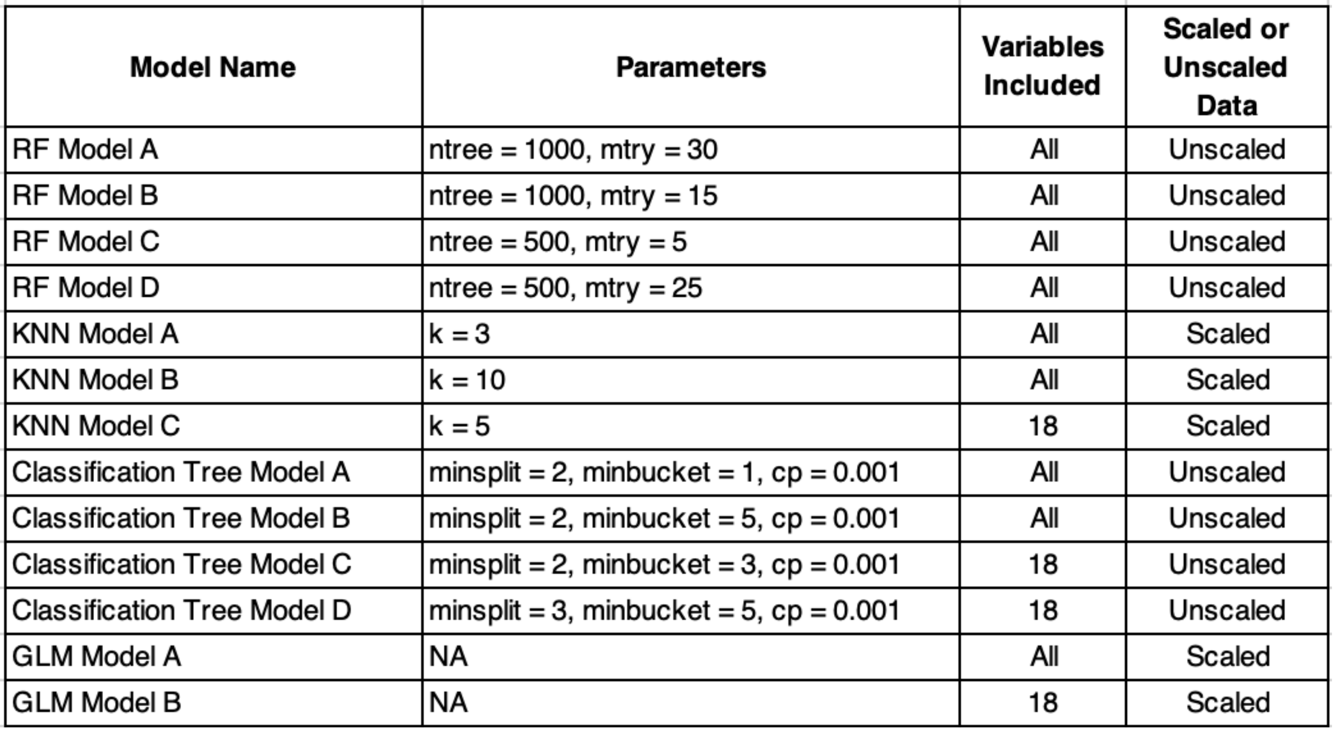
Note. Variables are listed in alphabetical order.

Table 2
Distribution and counts of negative and positive cases in training and test sets



Data Set	Percentage of 0's	Percentage of 1's	Count of 0's	Count of 1's
Train	93.32	6.68	8,013	574
Test	93.28	6.71	3,433	247
Train with Scaling	93.32	6.68	8,013	572
Test with Scaling	92.28	6.71	3,433	247

Table 3
Description of models

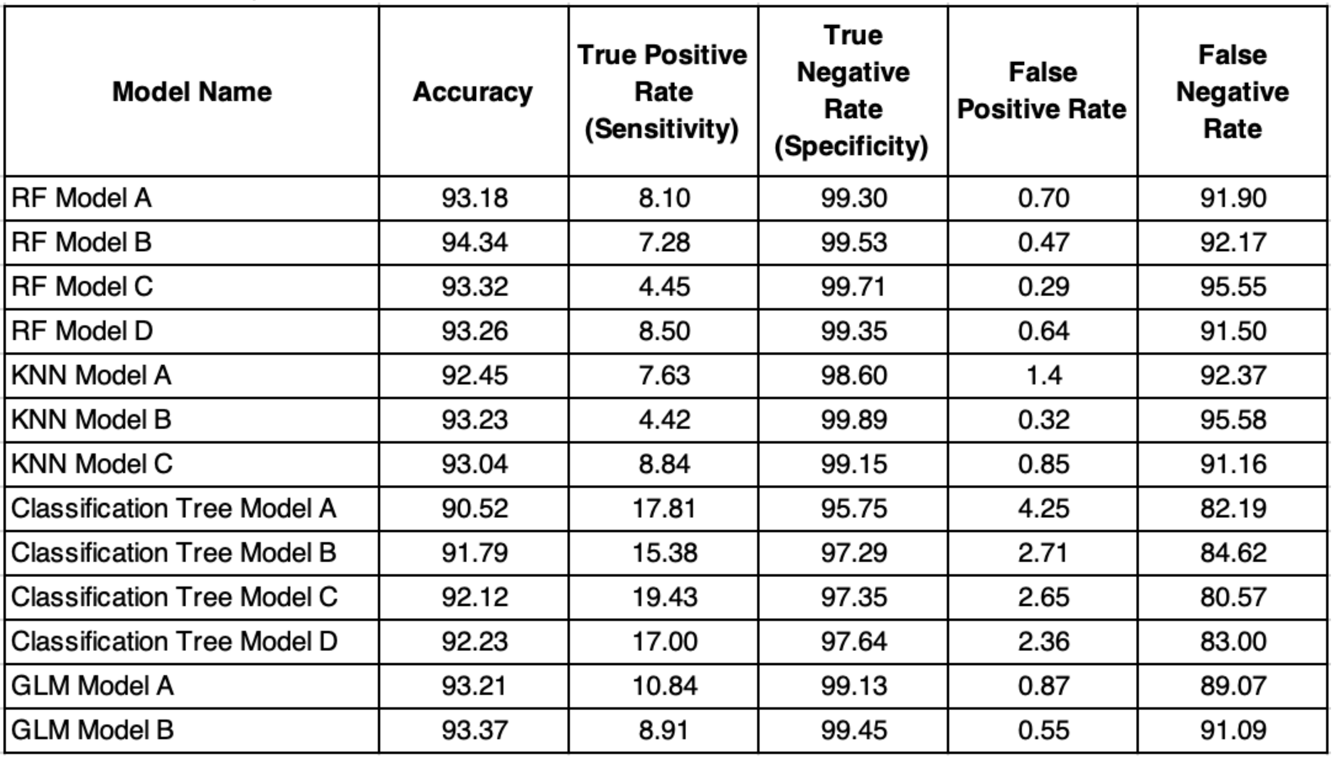


Model Name	Parameters	Variables Included	Scaled or Unscaled Data
RF Model A	ntree = 1000, mtry = 30	All	Unscaled
RF Model B	ntree = 1000, mtry = 15	All	Unscaled
RF Model C	ntree = 500, mtry = 5	All	Unscaled
RF Model D	ntree = 500, mtry = 25	All	Unscaled
KNN Model A	k = 3	All	Scaled
KNN Model B	k = 10	All	Scaled
KNN Model C	k = 5	18	Scaled
Classification Tree Model A	minsplit = 2, minbucket = 1, cp = 0.001	All	Unscaled
Classification Tree Model B	minsplit = 2, minbucket = 5, cp = 0.001	All	Unscaled
Classification Tree Model C	minsplit = 2, minbucket = 3, cp = 0.001	18	Unscaled
Classification Tree Model D	minsplit = 3, minbucket = 5, cp = 0.001	18	Unscaled
GLM Model A	NA	All	Scaled
GLM Model B	NA	18	Scaled

Note. “All” variables refers to the 86 variables that were identified for use in modeling after consulting with the professional advisor. Through his advice and domain knowledge, the data set was reduced from the 126 variables that met the <15.00% missingness criteria to 86 variables.

Table 4

Metrics for model performance

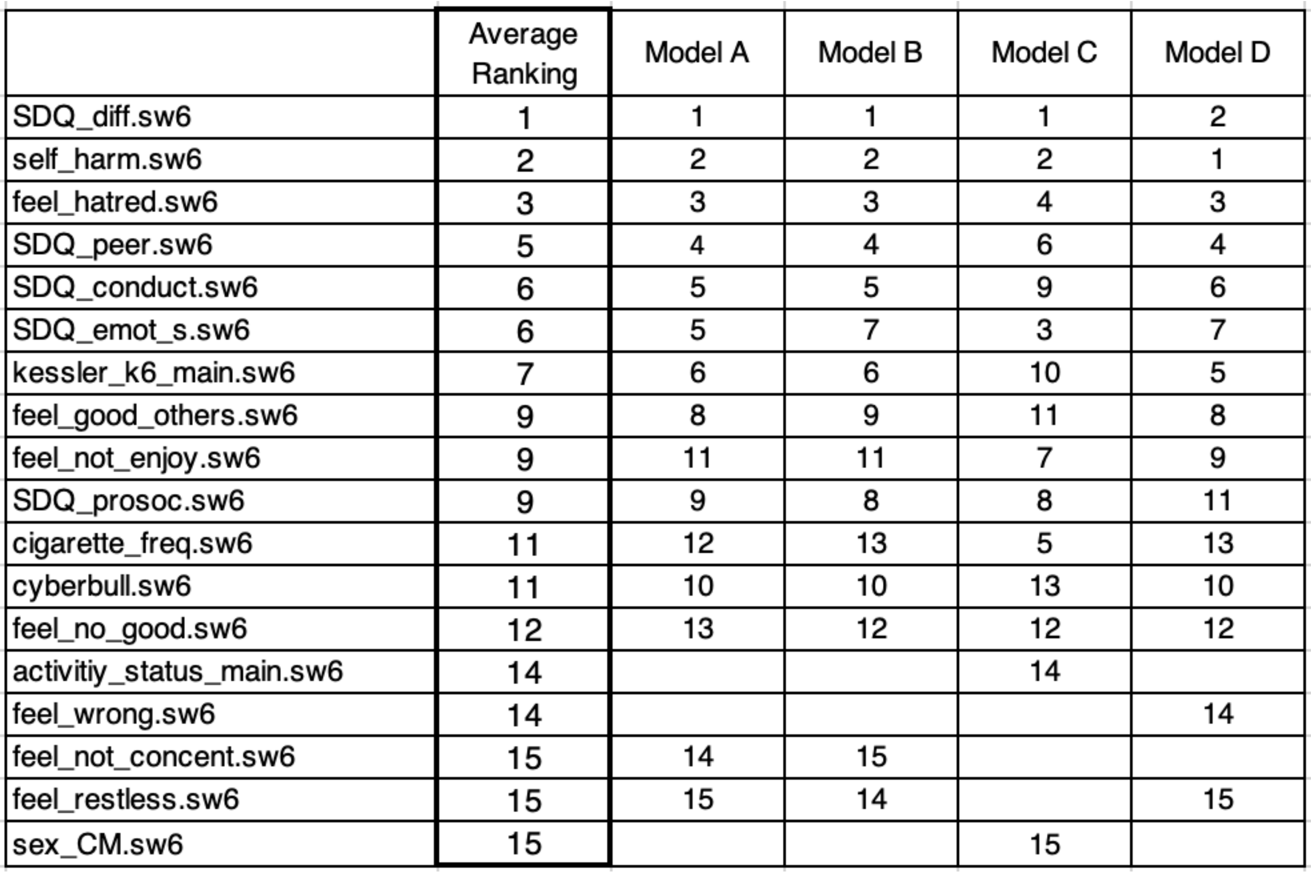


Model Name	Accuracy	True Positive Rate (Sensitivity)	True Negative Rate (Specificity)	False Positive Rate	False Negative Rate
RF Model A	93.18	8.10	99.30	0.70	91.90
RF Model B	94.34	7.28	99.53	0.47	92.17
RF Model C	93.32	4.45	99.71	0.29	95.55
RF Model D	93.26	8.50	99.35	0.64	91.50
KNN Model A	92.45	7.63	98.60	1.4	92.37
KNN Model B	93.23	4.42	99.89	0.32	95.58
KNN Model C	93.04	8.84	99.15	0.85	91.16
Classification Tree Model A	90.52	17.81	95.75	4.25	82.19
Classification Tree Model B	91.79	15.38	97.29	2.71	84.62
Classification Tree Model C	92.12	19.43	97.35	2.65	80.57
Classification Tree Model D	92.23	17.00	97.64	2.36	83.00
GLM Model A	93.21	10.84	99.13	0.87	89.07
GLM Model B	93.37	8.91	99.45	0.55	91.09

Note. The values listed in this table are percentages.

Table 5

Important variables from random forest models, sorted by average ranking from highest to lowest

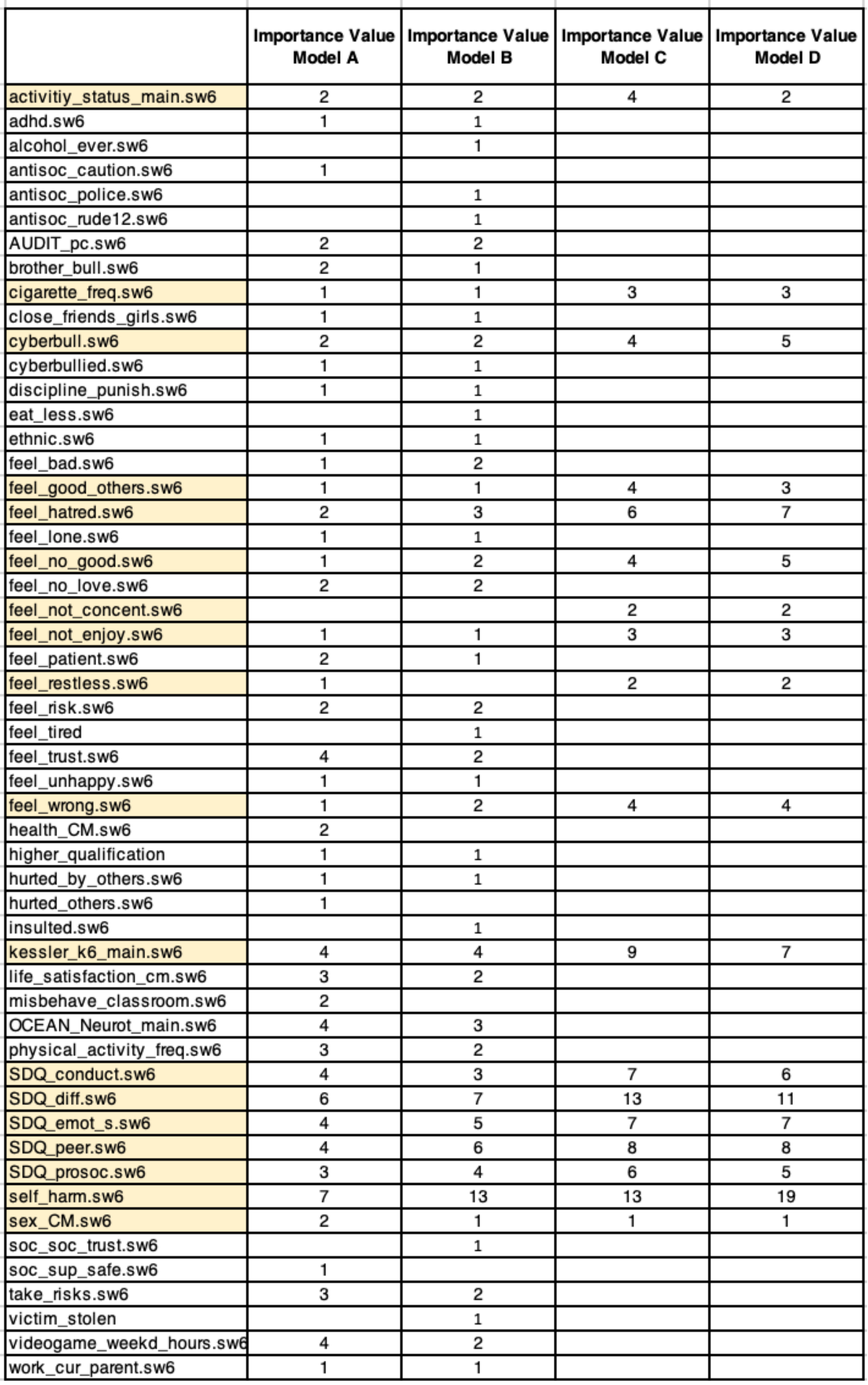


	Average Ranking	Model A	Model B	Model C	Model D
SDQ_diff.sw6	1	1	1	1	2
self_harm.sw6	2	2	2	2	1
feel_hatred.sw6	3	3	3	4	3
SDQ_peer.sw6	5	4	4	6	4
SDQ_conduct.sw6	6	5	5	9	6
SDQ_emot_s.sw6	6	5	7	3	7
kessler_k6_main.sw6	7	6	6	10	5
feel_good_others.sw6	9	8	9	11	8
feel_not_enjoy.sw6	9	11	11	7	9
SDQ_prosoc.sw6	9	9	8	8	11
cigarette_freq.sw6	11	12	13	5	13
cyberbull.sw6	11	10	10	13	10
feel_no_good.sw6	12	13	12	12	12
activitiy_status_main.sw6	14			14	
feel_wrong.sw6	14				14
feel_not_concent.sw6	15	14	15		
feel_restless.sw6	15	15	14		15
sex_CM.sw6	15			15	

Note. This table contains the 15 variables identified from the mean decrease accuracy variable importance plots for each model. The rank of the variable as it appears in these plots is listed for each model. The average ranking was computed by averaging across all models for a given variable. In the case of average rankings that were decimals, we rounded up. The table is sorted by the variables' average ranking from highest to lowest.

Table 6

Variables identified as important in classification tree models with their importance values – all models summary view

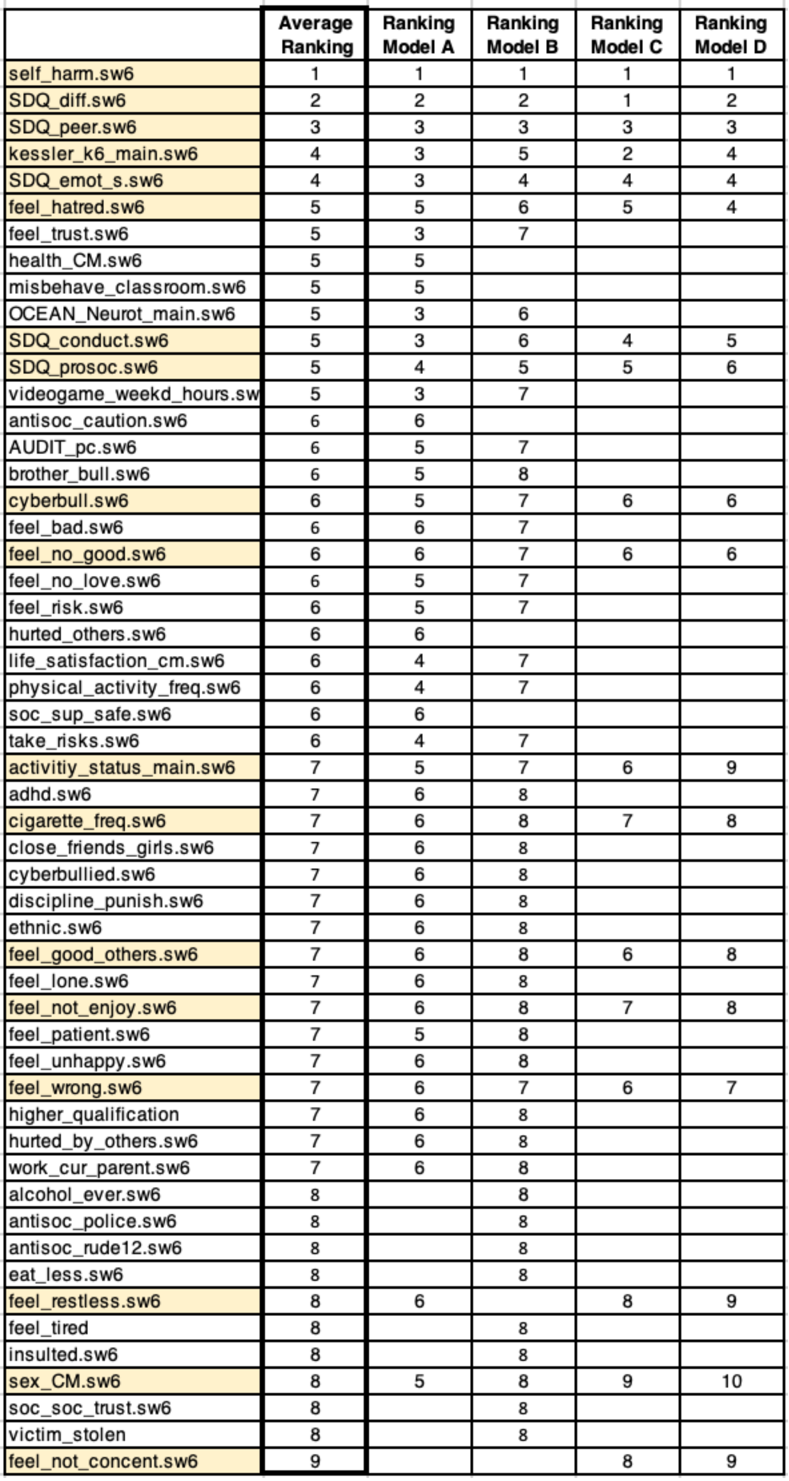


	Importance Value Model A	Importance Value Model B	Importance Value Model C	Importance Value Model D
activitiy_status_main.sw6	2	2	4	2
adhd.sw6	1	1		
alcohol_ever.sw6		1		
antisoc_caution.sw6	1			
antisoc_police.sw6		1		
antisoc_rude12.sw6		1		
AUDIT_pc.sw6	2	2		
brother_bull.sw6	2	1		
cigarette_freq.sw6	1	1	3	3
close_friends_girls.sw6	1	1		
cyberbull.sw6	2	2	4	5
cyberbullied.sw6	1	1		
discipline_punish.sw6	1	1		
eat_less.sw6		1		
ethnic.sw6	1	1		
feel_bad.sw6	1	2		
feel_good_others.sw6	1	1	4	3
feel_hatred.sw6	2	3	6	7
feel_lone.sw6	1	1		
feel_no_good.sw6	1	2	4	5
feel_no_love.sw6	2	2		
feel_not_concent.sw6			2	2
feel_not_enjoy.sw6	1	1	3	3
feel_patient.sw6	2	1		
feel_restless.sw6	1		2	2
feel_risk.sw6	2	2		
feel_tired		1		
feel_trust.sw6	4	2		
feel_unhappy.sw6	1	1		
feel_wrong.sw6	1	2	4	4
health_CM.sw6	2			
higher_qualification	1	1		
hurted_by_others.sw6	1	1		
hurted_others.sw6	1			
insulted.sw6		1		
kessler_k6_main.sw6	4	4	9	7
life_satisfaction_cm.sw6	3	2		
misbehave_classroom.sw6	2			
OCEAN_Neurot_main.sw6	4	3		
physical_activity_freq.sw6	3	2		
SDQ_conduct.sw6	4	3	7	6
SDQ_diff.sw6	6	7	13	11
SDQ_emot_s.sw6	4	5	7	7
SDQ_peer.sw6	4	6	8	8
SDQ_prosoc.sw6	3	4	6	5
self_harm.sw6	7	13	13	19
sex_CM.sw6	2	1	1	1
soc_soc_trust.sw6		1		
soc_sup_safe.sw6	1			
take_risks.sw6	3	2		
victim_stolen		1		
videogame_weekd_hours.sw6	4	2		
work_cur_parent.sw6	1	1		

Note. These data can be found in the .txt files containing the output generated from running the `summary` function on the classification tree objects. These files can be found in the working directory. In this table, higher values indicate more importance. Note that the highest value assigned varies between the models. The 18 variables identified from the random forest models as being important are highlighted in yellow. This table is sorted alphabetically by variable name.

Table 7

Classification tree important variables ranked by importance, with average ranking across all models – all models summary view

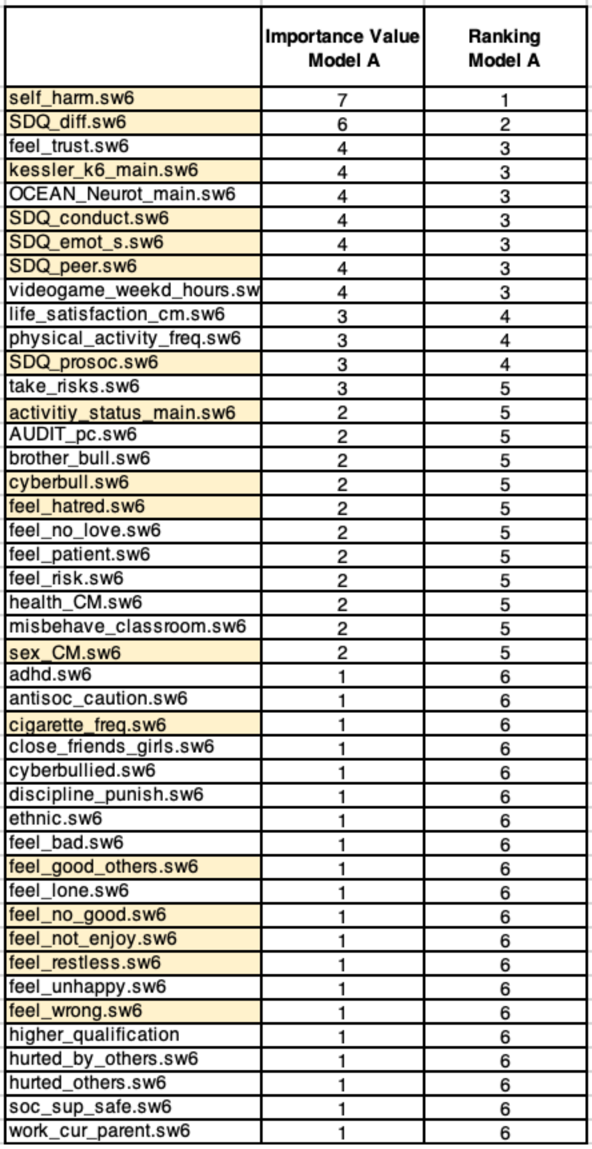


	Average Ranking	Ranking Model A	Ranking Model B	Ranking Model C	Ranking Model D
self_harm.sw6	1	1	1	1	1
SDQ_diff.sw6	2	2	2	1	2
SDQ_peer.sw6	3	3	3	3	3
kessler_k6_main.sw6	4	3	5	2	4
SDQ_emot_s.sw6	4	3	4	4	4
feel_hatred.sw6	5	5	6	5	4
feel_trust.sw6	5	3	7		
health_CM.sw6	5	5			
misbehave_classroom.sw6	5	5			
OCEAN_Neurot_main.sw6	5	3	6		
SDQ_conduct.sw6	5	3	6	4	5
SDQ_prosoc.sw6	5	4	5	5	6
videogame_weekd_hours.sw	5	3	7		
antisoc_caution.sw6	6	6			
AUDIT_pc.sw6	6	5	7		
brother_bull.sw6	6	5	8		
cyberbull.sw6	6	5	7	6	6
feel_bad.sw6	6	6	7		
feel_no_good.sw6	6	6	7	6	6
feel_no_love.sw6	6	5	7		
feel_risk.sw6	6	5	7		
hurted_others.sw6	6	6			
life_satisfaction_cm.sw6	6	4	7		
physical_activity_freq.sw6	6	4	7		
soc_sup_safe.sw6	6	6			
take_risks.sw6	6	4	7		
activitiy_status_main.sw6	7	5	7	6	9
adhd.sw6	7	6	8		
cigarette_freq.sw6	7	6	8	7	8
close_friends_girls.sw6	7	6	8		
cyberbullied.sw6	7	6	8		
discipline_punish.sw6	7	6	8		
ethnic.sw6	7	6	8		
feel_good_others.sw6	7	6	8	6	8
feel_lone.sw6	7	6	8		
feel_not_enjoy.sw6	7	6	8	7	8
feel_patient.sw6	7	5	8		
feel_unhappy.sw6	7	6	8		
feel_wrong.sw6	7	6	7	6	7
higher_qualification	7	6	8		
hurted_by_others.sw6	7	6	8		
work_cur_parent.sw6	7	6	8		
alcohol_ever.sw6	8		8		
antisoc_police.sw6	8		8		
antisoc_rude12.sw6	8		8		
eat_less.sw6	8		8		
feel_restless.sw6	8	6		8	9
feel_tired	8		8		
insulted.sw6	8		8		
sex_CM.sw6	8	5	8	9	10
soc_soc_trust.sw6	8		8		
victim_stolen	8		8		
feel_not_concent.sw6	9			8	9

Note. Important variables are ranked from most important (smallest value) to least important (largest value) for each model. The average ranking was computed for each variable by averaging the rankings of all the models. In the case of average rankings that were decimals, we rounded up. The 18 variables identified from the random forest models as being important are highlighted in yellow.

Table 8

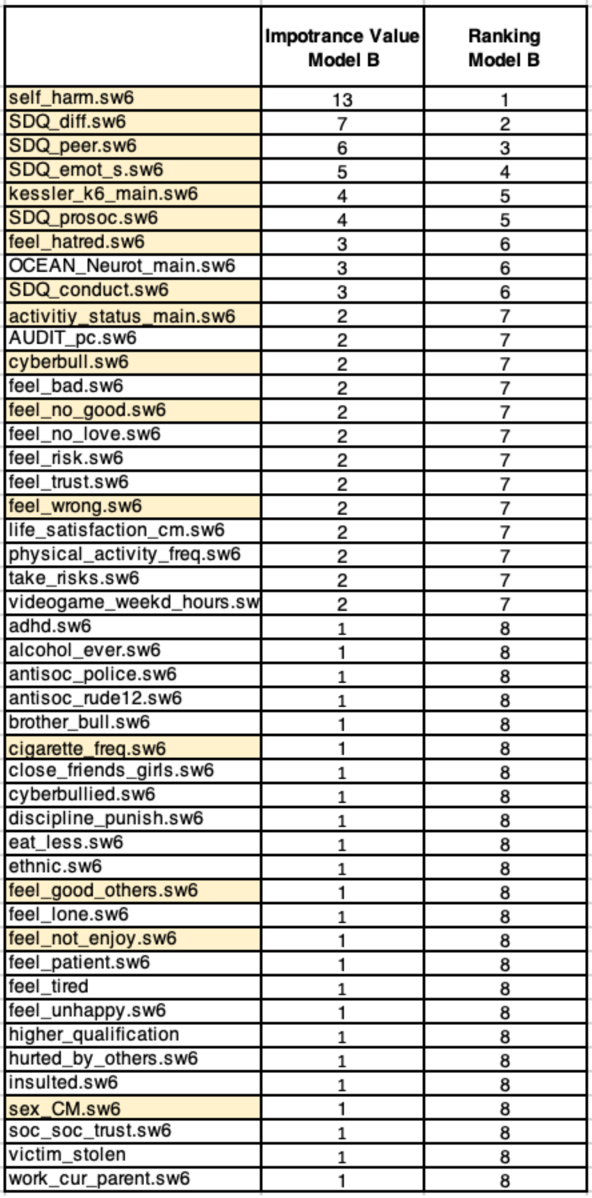
Variables identified as important in classification tree Model A – importance value and ranking



	Importance Value Model A	Ranking Model A
self_harm.sw6	7	1
SDQ_diff.sw6	6	2
feel_trust.sw6	4	3
kessler_k6_main.sw6	4	3
OCEAN_Neurot_main.sw6	4	3
SDQ_conduct.sw6	4	3
SDQ_emot_s.sw6	4	3
SDQ_peer.sw6	4	3
videogame_weekd_hours.sw	4	3
life_satisfaction_cm.sw6	3	4
physical_activity_freq.sw6	3	4
SDQ_prosoc.sw6	3	4
take_risks.sw6	3	5
activitiy_status_main.sw6	2	5
AUDIT_pc.sw6	2	5
brother_bull.sw6	2	5
cyberbull.sw6	2	5
feel_hatred.sw6	2	5
feel_no_love.sw6	2	5
feel_patient.sw6	2	5
feel_risk.sw6	2	5
health_CM.sw6	2	5
misbehave_classroom.sw6	2	5
sex_CM.sw6	2	5
adhd.sw6	1	6
antisoc_caution.sw6	1	6
cigarette_freq.sw6	1	6
close_friends_girls.sw6	1	6
cyberbullied.sw6	1	6
discipline_punish.sw6	1	6
ethnic.sw6	1	6
feel_bad.sw6	1	6
feel_good_others.sw6	1	6
feel_lone.sw6	1	6
feel_no_good.sw6	1	6
feel_not_enjoy.sw6	1	6
feel_restless.sw6	1	6
feel_unhappy.sw6	1	6
feel_wrong.sw6	1	6
higher_qualification	1	6
hurted_by_others.sw6	1	6
hurted_others.sw6	1	6
soc_sup_safe.sw6	1	6
work_cur_parent.sw6	1	6

Note. This table is sorted from most to least important variables. The importance value comes from the output file and the ranking simply ranks the variables in order of their importance. Variables that appear in the list of 18 variables identified from the random forest models as being important are highlighted in yellow.

Table 9
Variables identified as important in classification tree Model B – importance value and ranking

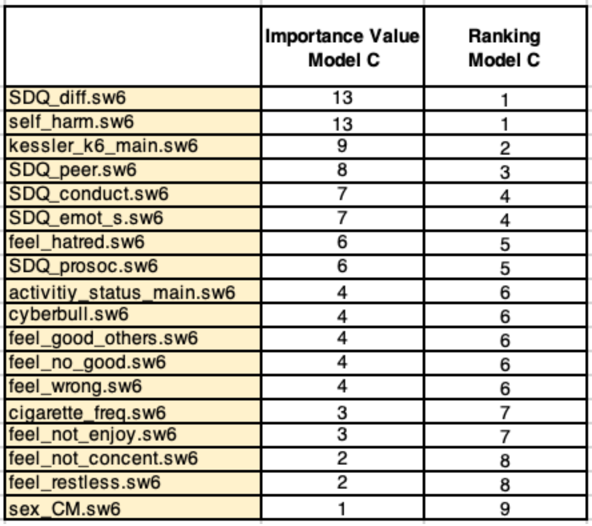


	Impotrance Value Model B	Ranking Model B
self_harm.sw6	13	1
SDQ_diff.sw6	7	2
SDQ_peer.sw6	6	3
SDQ_emot_s.sw6	5	4
kessler_k6_main.sw6	4	5
SDQ_prosoc.sw6	4	5
feel_hatred.sw6	3	6
OCEAN_Neurot_main.sw6	3	6
SDQ_conduct.sw6	3	6
activitiy_status_main.sw6	2	7
AUDIT_pc.sw6	2	7
cyberbull.sw6	2	7
feel_bad.sw6	2	7
feel_no_good.sw6	2	7
feel_no_love.sw6	2	7
feel_risk.sw6	2	7
feel_trust.sw6	2	7
feel_wrong.sw6	2	7
life_satisfaction_cm.sw6	2	7
physical_activity_freq.sw6	2	7
take_risks.sw6	2	7
videogame_weekd_hours.sw	2	7
adhd.sw6	1	8
alcohol_ever.sw6	1	8
antisoc_police.sw6	1	8
antisoc_rude12.sw6	1	8
brother_bull.sw6	1	8
cigarette_freq.sw6	1	8
close_friends_girls.sw6	1	8
cyberbullied.sw6	1	8
discipline_punish.sw6	1	8
eat_less.sw6	1	8
ethnic.sw6	1	8
feel_good_others.sw6	1	8
feel_lone.sw6	1	8
feel_not_enjoy.sw6	1	8
feel_patient.sw6	1	8
feel_tired	1	8
feel_unhappy.sw6	1	8
higher_qualification	1	8
hurted_by_others.sw6	1	8
insulted.sw6	1	8
sex_CM.sw6	1	8
soc_soc_trust.sw6	1	8
victim_stolen	1	8
work_cur_parent.sw6	1	8

Note. This table is sorted from most to least important variables. The importance value comes from the output file and the ranking simply ranks the variables in order of their importance. Variables that appear in the list of 18 variables identified from the random forest models as being important are highlighted in yellow.

Table 10

Variables identified as important in classification tree Model C – importance value and ranking

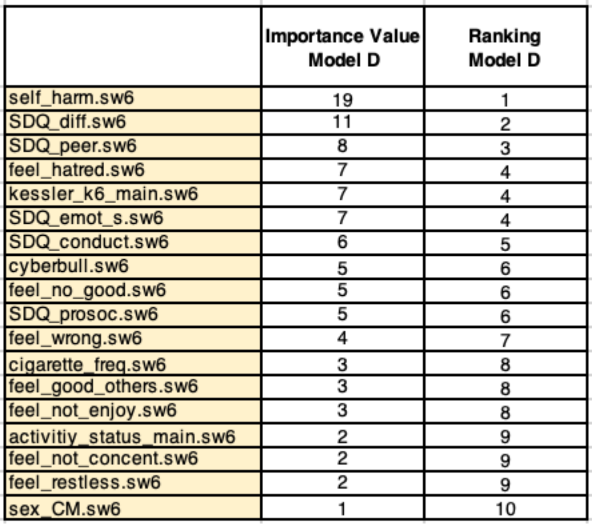


	Importance Value Model C	Ranking Model C
SDQ_diff.sw6	13	1
self_harm.sw6	13	1
kessler_k6_main.sw6	9	2
SDQ_peer.sw6	8	3
SDQ_conduct.sw6	7	4
SDQ_emot_s.sw6	7	4
feel_hatred.sw6	6	5
SDQ_prosoc.sw6	6	5
activitiy_status_main.sw6	4	6
cyberbull.sw6	4	6
feel_good_others.sw6	4	6
feel_no_good.sw6	4	6
feel_wrong.sw6	4	6
cigarette_freq.sw6	3	7
feel_not_enjoy.sw6	3	7
feel_not_concent.sw6	2	8
feel_restless.sw6	2	8
sex_CM.sw6	1	9

Note. This table is sorted from most to least important variables. The importance value comes from the output file and the ranking simply ranks the variables in order of their importance. Variables that appear in the list of 18 variables identified from the random forest models as being important are highlighted in yellow, and consequently, all variables are highlighted in yellow because Model C was created only from these variables.

Table 11

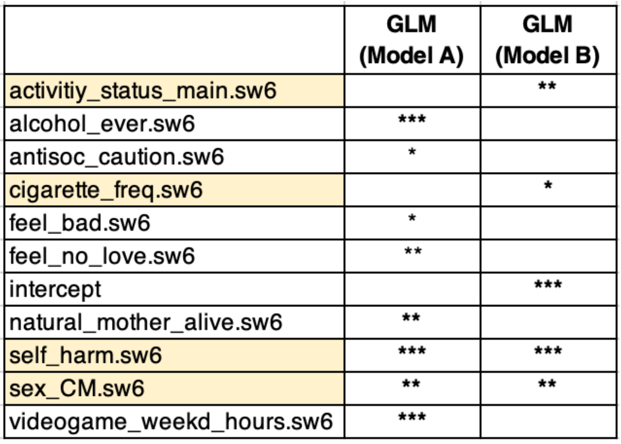
Variables identified as important in classification tree Model D – importance value and ranking



	Importance Value Model D	Ranking Model D
self_harm.sw6	19	1
SDQ_diff.sw6	11	2
SDQ_peer.sw6	8	3
feel_hatred.sw6	7	4
kessler_k6_main.sw6	7	4
SDQ_emot_s.sw6	7	4
SDQ_conduct.sw6	6	5
cyberbull.sw6	5	6
feel_no_good.sw6	5	6
SDQ_prosoc.sw6	5	6
feel_wrong.sw6	4	7
cigarette_freq.sw6	3	8
feel_good_others.sw6	3	8
feel_not_enjoy.sw6	3	8
activitiy_status_main.sw6	2	9
feel_not_concent.sw6	2	9
feel_restless.sw6	2	9
sex_CM.sw6	1	10

Note. This table is sorted by most to least important variables. The importance value comes from the output file and the ranking simply ranks the variables in order of their importance. Variables that appear in the list of 18 variables identified from the random forest models as being important are highlighted in yellow, and consequently, all variables are highlighted in yellow because Model D was created only from these variables.

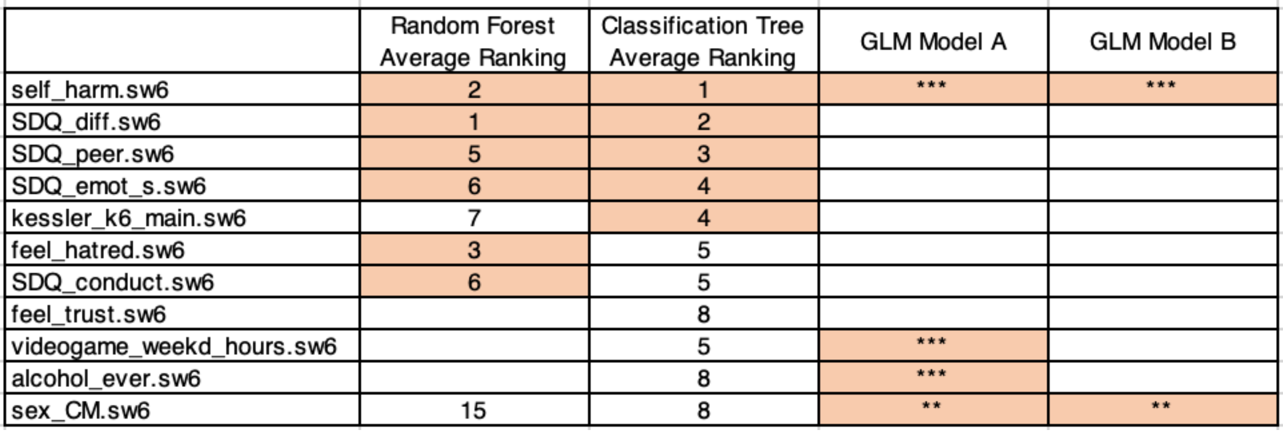
Table 12
Variables identified as important from the generalized linear models



	GLM (Model A)	GLM (Model B)
activitiy_status_main.sw6		**
alcohol_ever.sw6	***	
antisoc_caution.sw6	*	
cigarette_freq.sw6		*
feel_bad.sw6	*	
feel_no_love.sw6	**	
intercept		***
natural_mother_alive.sw6	**	
self_harm.sw6	***	***
sex_CM.sw6	**	**
videogame_weekd_hours.sw6	***	

Note. This table is sorted alphabetically by variable name. $p < 0.001$ is represented by ***, $p < 0.01$ is represented by **, and $p < 0.05$ is represented by *. Variables that appear in the list of 18 variables identified from the random forest models as being important are highlighted in yellow. Recall that Model B was created from these variables.

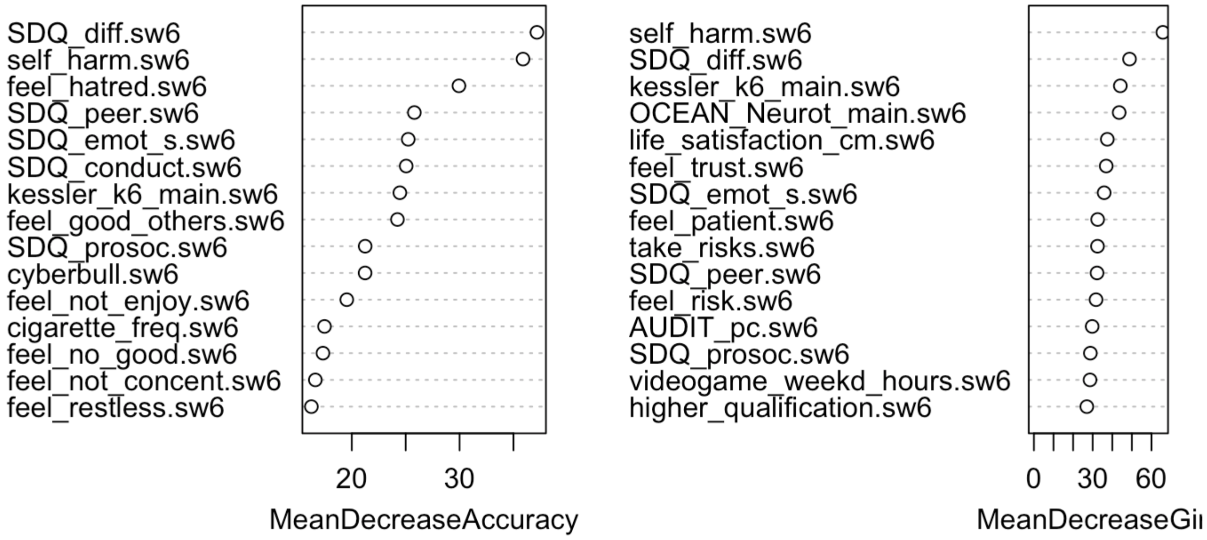
Table 13
Most important variables that appear in the random forest, classification tree, and generalized linear model types



	Random Forest Average Ranking	Classification Tree Average Ranking	GLM Model A	GLM Model B
self_harm.sw6	2	1	***	***
SDQ_diff.sw6	1	2		
SDQ_peer.sw6	5	3		
SDQ_emot_s.sw6	6	4		
kessler_k6_main.sw6	7	4		
feel_hatred.sw6	3	5		
SDQ_conduct.sw6	6	5		
feel_trust.sw6		8		
videogame_weekd_hours.sw6		5	***	
alcohol_ever.sw6		8	***	
sex_CM.sw6	15	8	**	**

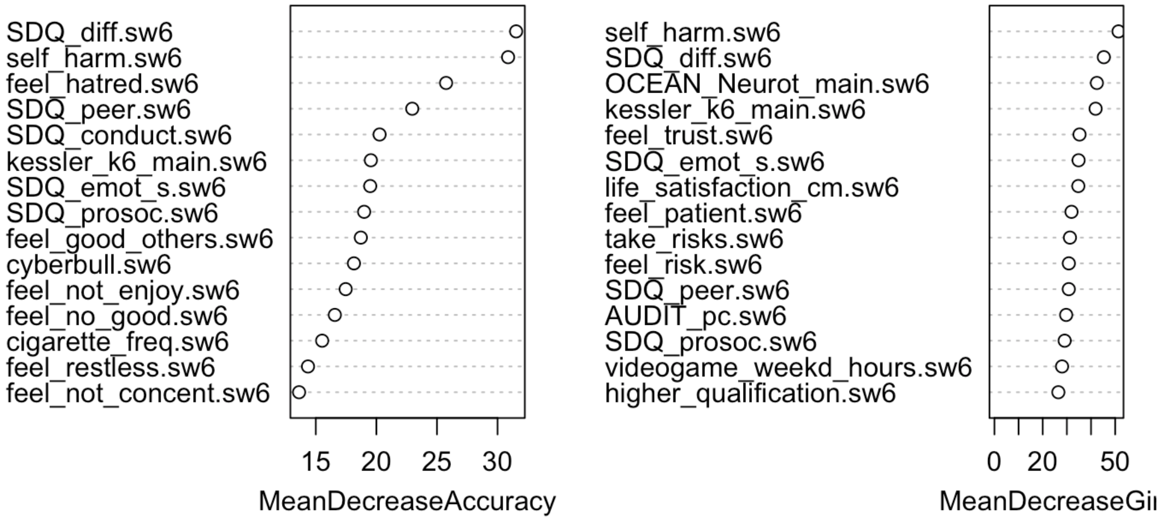
Note. This table shows the top 5 most important variables found for each model type. For each model type, we highlight in orange the value associated with the top 5 variable. In the case of the random forest and classification trees, this value is the average ranking associated with the top 5 variable, and for the generalized linear models, this value is the *p* value significance level. For the generalized linear models, we included the variables that were significant at the lowest values of *p.* Note that in the case of random forest, the averaging and rounding used to produce the average ranking resulted in the fourth most important variable having a ranking of 5. Also, six cells are highlighted because there was a tie for the 5th most important variable. Note that in the cases where a variable was a top 5 most important variables for one model type but not for the other model type(s), we listed any pertinent data available for the other model type(s) but did not highlight this information in orange. For example, `alcohol_ever.sw6` was one of the top 5 most important variables per the generalized linear models, but this variable was not one of the top 5 most important variables for random forest nor for classification trees. However, this variable was ranked 8th in variable importance for the classification tree models so we included that information (there was no ranking data to report for random forest since this variable was not ranked in the top 15 across all random forest models).

Figure 1
Variable importance plot from random forest Model A



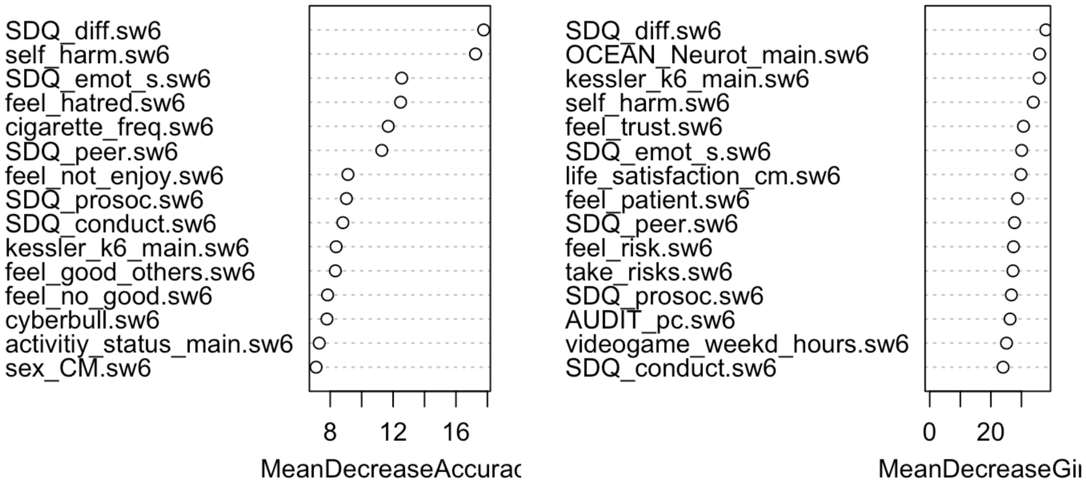
Note. The 18 variables identified across the random forest variable importance plots were identified solely from the mean decrease accuracy plots. No attention was paid to the mean decrease Gini plots in determining the most important variables.

Figure 2
Variable importance plot from random forest Model B



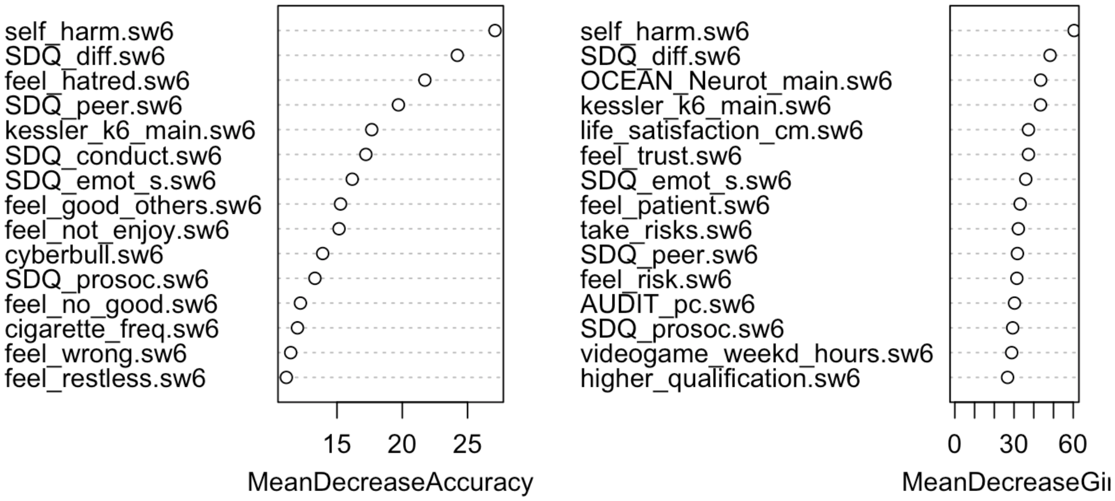
Note. The 18 variables identified across the random forest variable importance plots were identified solely from the mean decrease accuracy plots. No attention was paid to the mean decrease Gini plots in determining the most important variables.

Figure 3
Variable importance plot from random forest Model C



Note. The 18 variables identified across the random forest variable importance plots were identified solely from the mean decrease accuracy plots. No attention was paid to the mean decrease Gini plots in determining the most important variables.

Figure 4
Variable importance plot from random forest Model D



Note. The 18 variables identified across the random forest variable importance plots were identified solely from the mean decrease accuracy plots. No attention was paid to the mean decrease Gini plots in determining the most important variables.

Figure 5

Classification tree Model A

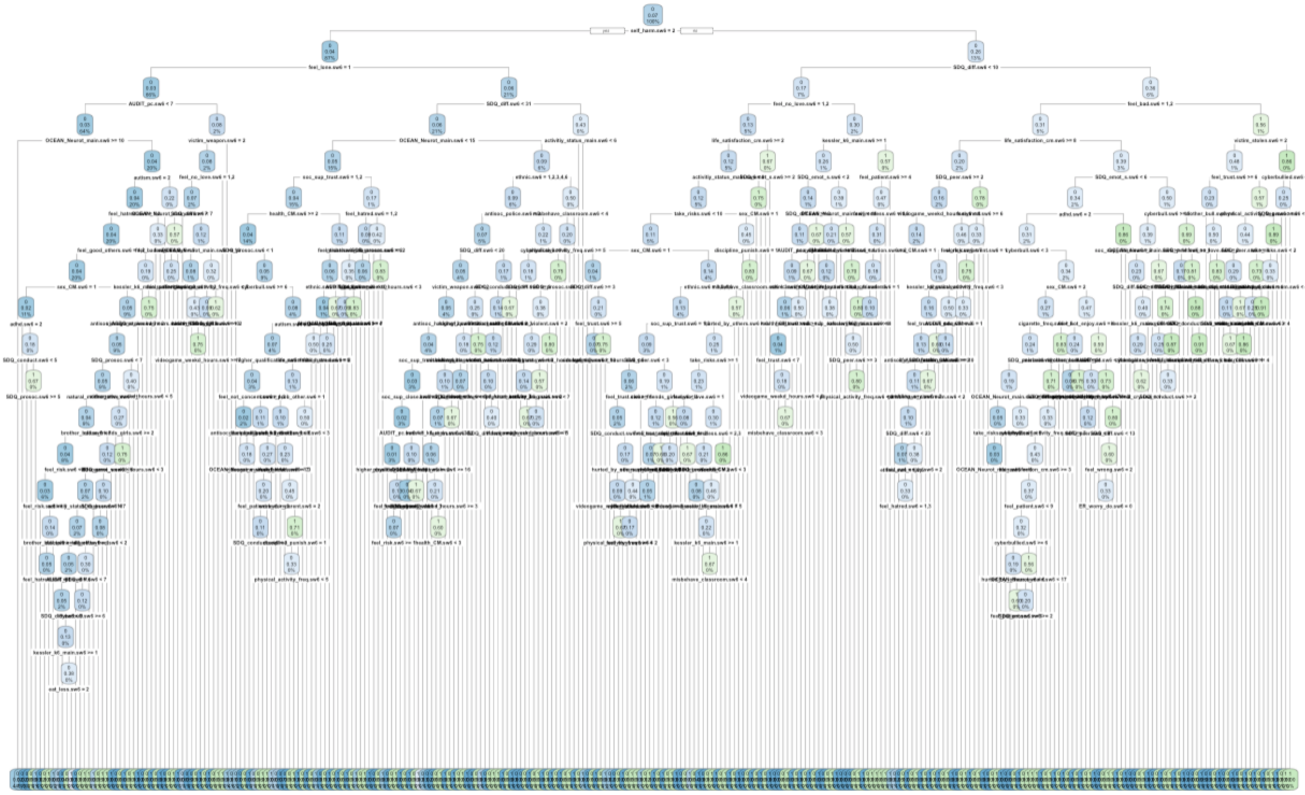


Figure 6

Classification tree Model B

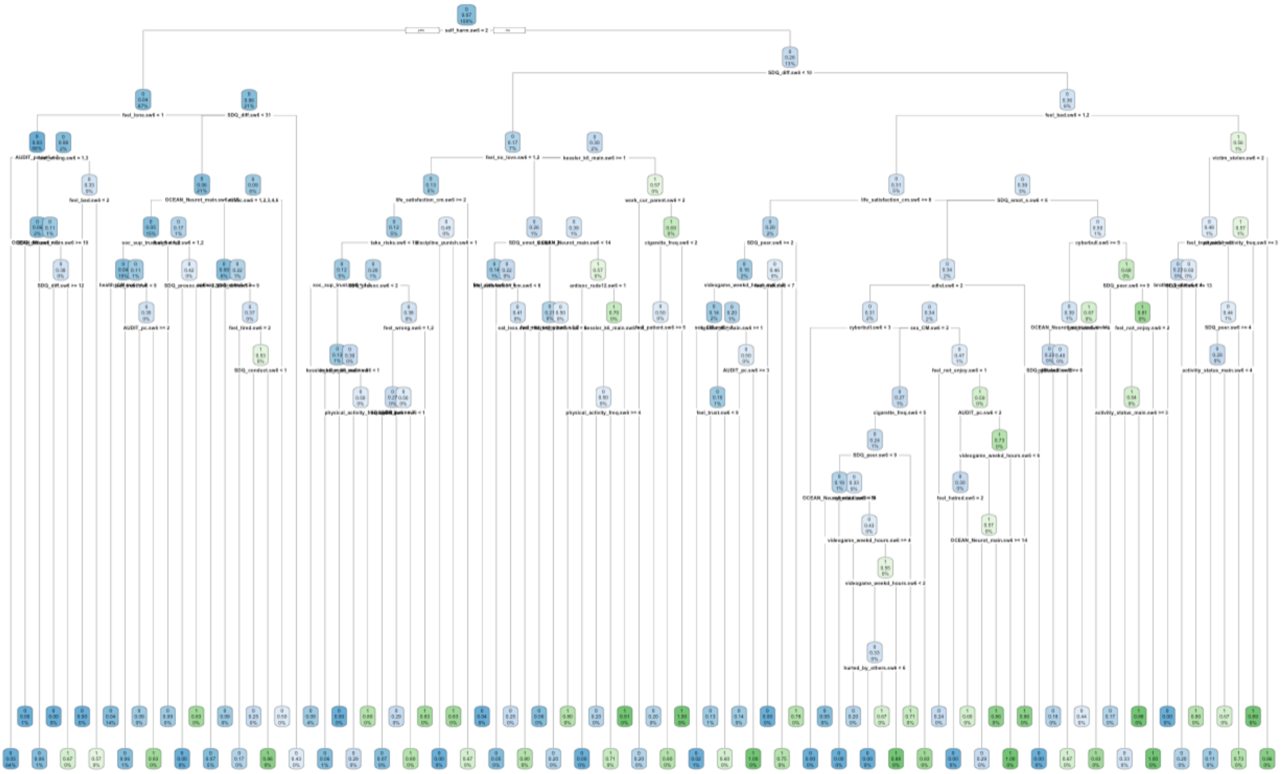


Figure 7
Classification tree Model C

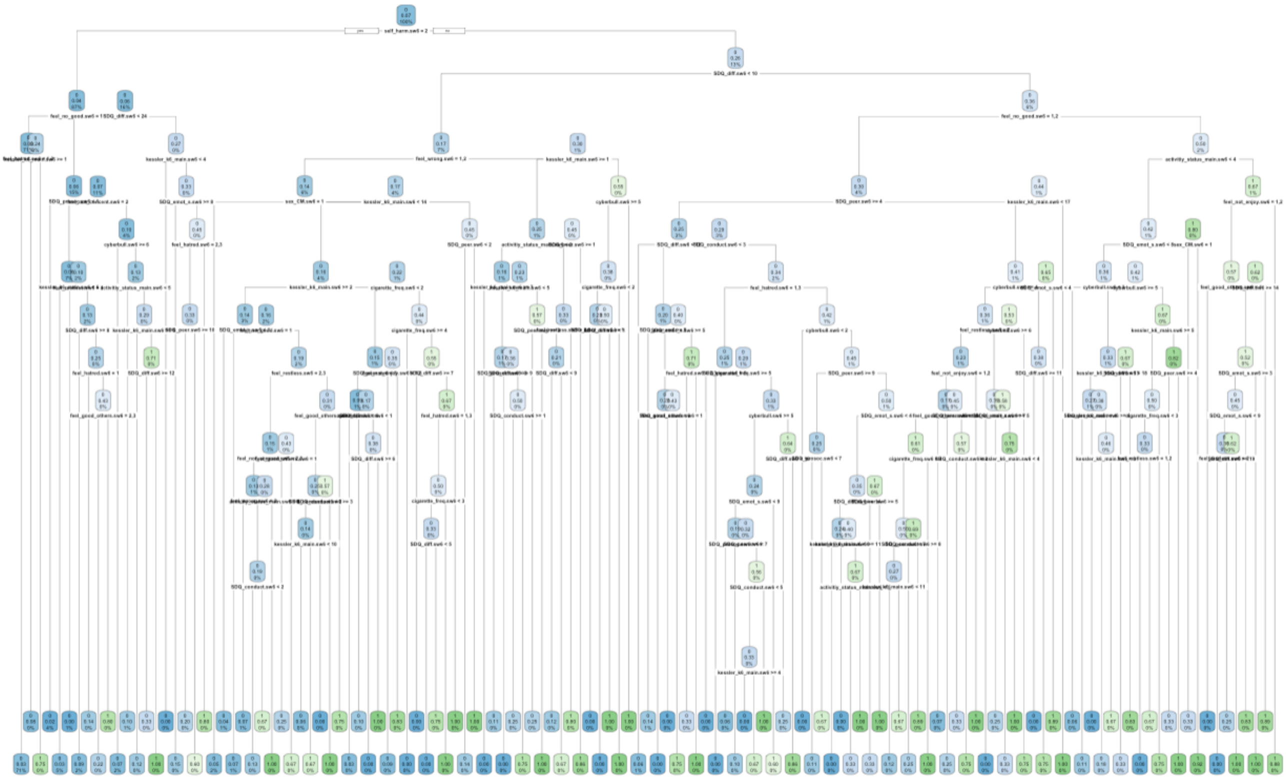
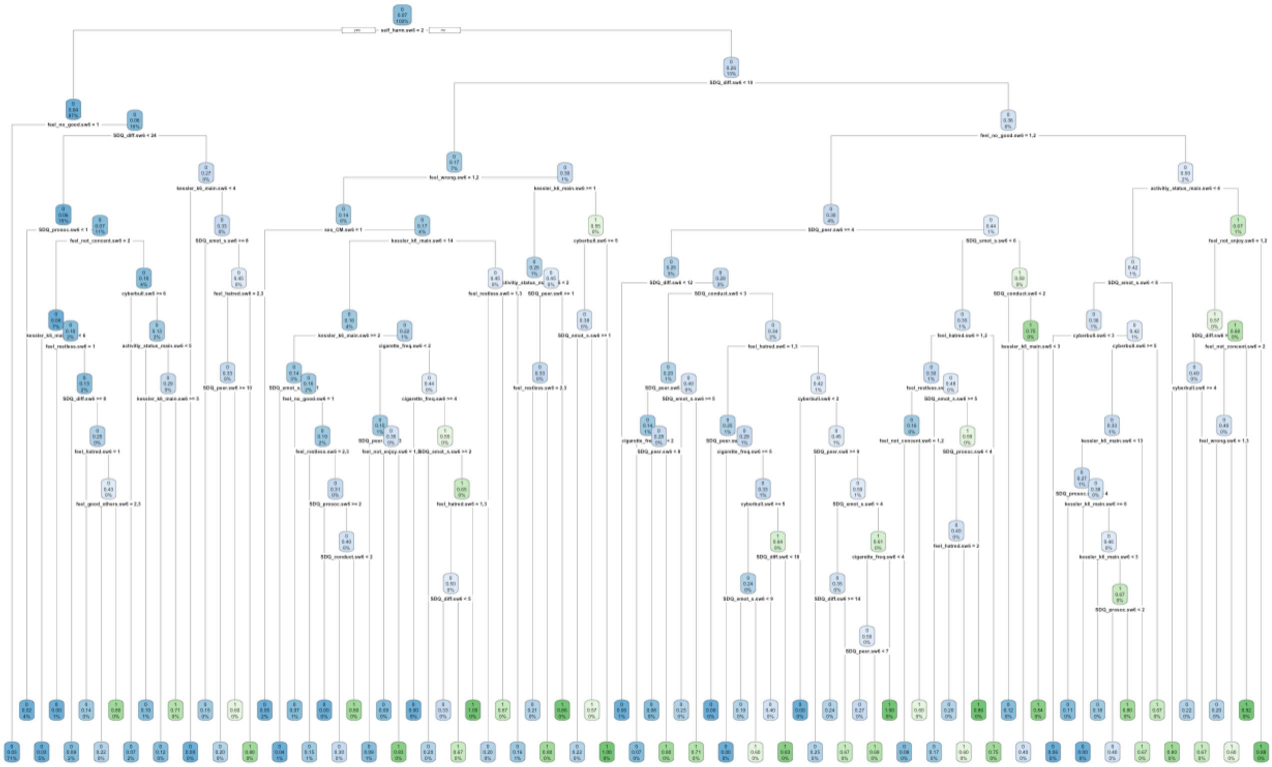


Figure 8
Classification tree Model D


0
0.07
100%
yes
self_harm.sw6 = 2
no
SDQ_diff.sw6 < 10