# Missing Title

Lévi Docters van Leeuwen

2023-06-16

## Introduction

Quantitative research within the social sciences regularly involve survey designs, where numerous questions are asked to respondents so data can be generated and analysed. While the goal generally is to have the data be as complete as possible, data collected will usually contain a fair amount of missing responses. Typically, these come in two flavours, with unit non-response referring to the absence of an entire record of a respondent, and item non-response indicating one or multiple items of a respondent are missing. Both present the researcher with their own unique problems, given the researchers had not intended for them to happen. With unit non-response, the sample might not reflect the true population, leading to a loss of statistical power among other harmful effects (Särndal & Lundström, 2005). A possible solution is weighting of the respondents that the sample did capture, as to construct a more accurate representation of the population. The other kind, item non-response, is a bit more tangible since only certain values are missing. This essentially forces the researcher(s) to make a decision on how to treat these values, as computationally, almost no analyses can be performed. Among the social sciences, the item non-response problem deserves attention and proper treatment, but is often overlooked or neglected in reality, with crude methods like listwise deletion and mean imputation still being widely popular (Bell, Kromrey & Ferron, 2009; Savage et al., 2021). Both methods can introduce bias into the results, which could lead to inaccurate conclusions.

Fortunately, more sophisticated methods like multiple imputation by chained equations (MICE) and full information maximum likelihood (FIML) are receiving more recognition. In short, FIML estimates a likelihood function for each individual based on the variables that are present, thus using all data available. MICE works differently by using a series of regression models and adding a random component to the final prediction to emulate a level of uncertainty. Research has shown that MICE and FIML consistently outperform the aforementioned approaches (Wulff & Jeppesen, 2017), and will produce unbiased imputations under the right circumstances (Wulff & Jeppesen, 2017). Nonetheless, researchers might not feel comfortable implementing these methods, as they will have to determine whether certain assumptions are met. Especially in the social sciences, of which its students have been shown to be the least statistically literate among peers (Pan & Tang, 2004; Berndt et al., 2021), a certain hesitancy might be structurally present to consider these modern instruments. This hesitancy is not entirely misplaced, since if specific conditions are not met, both MICE and FIML might produce biased results without the researchers being aware (Pepinsky, 2018). To grasp what these conditions are, and to gain a better understanding of the underlying causes of missing data, missing data mechanisms have to be acknowledged first.

In modern literature, three distinct missing data mechanisms are identified: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR), also called Not Data Dependent, Seen Data Dependent and Unseen Data Dependent, respectively (Van Buuren & Groothuis-Oudshoorn, 2011). In summary, when data is MCAR, the missingness is unrelated to the observed as well as the unobserved data. An example of this would be if data is not recorded because of an accidental technical difficulty during an online survey. Data is MAR when the probability of missingness is related to the observed data. Imagine a survey is held among a population to monitor anxiety, and sex of the participant is recorded as well. Assume for the sake of the example that males have a harder time answering the questions regarding anxiety, and tend to skip them. Data would then (rather counterintuitively) be called Missing At Random, as the probability of missing is dependent of sex of the respondent. Finally, MNAR implies the reason data is missing because it is related to unobserved data. In the aforementioned example about the anxiety survey, data would be MNAR if sex of the respondent had not been observed, meaning the probability of missingness on questions concerning anxiety would now be dependent on unobserved data.

Accordingly, the sophisticated methods mentioned earlier make use of this observed data available to compute imputations for the missing values, and can therefore only produce unbiased results if data are either MCAR or MAR. Naturally, these methods will introduce bias in their results if the data are MNAR, as the unobserved variables (the underlying cause of the missingness) cannot be accounted for.

Consequently, in order to tackle the missing data problem in a refined manner, a researcher has to be conscious early on in the scientific process. Before data collection starts, the researcher should determine what variables might have missing values, and what other variables might account for their missingness. Data on those variables should then also be collected. However, identifying which variables can account for

missingness presents its own set of challenges. While the general topic of non-response is widely explored, validated information about specific constructs (like what may cause respondents to refrain from answering questions about depression) is missing, leaving the researcher to resort to speculation.

To remove the need for this scientifically ambiguous process, this thesis project will aim to enrich a common knowledge base about these specific variables by predicting missingness with the assistance of algorithmic modelling. This knowledge base could lead to researchers (in particular in the social sciences) experiencing less of a barrier to implement methods like MICE and FIML. Accordingly, more accurate data and interpretations could be produced.

As the necessity and relevancy of a trustworthy scientific process is at the core of this project, principles of Open Science were applied. These principles aim to attain transparency, collaboration and accessibility by ensuring data is publicly available, research is able to be reproduced, ethical considerations and various other practices (Vicente-Saez & Martinez-Fuentes, 2018). Additionally, in consideration of possible harmful biases that may occur when the researchers degrees of freedom are not accounted for, this project utilises the novel concept of multiverse analysis (Steegen, Tuerlinckx, Gelman & Vanpaemel, 2016). The researcher degrees of freedom refer to the inherent nonuniformity of scientific experiments, as researchers can choose from a variety of different methods and approaches to conduct the data collection and analysis processes. Data dredging is one of those instances, for example, where the researcher degrees of freedom are not taken into account; Exhaustive analysis of the data and deliberately only reporting a particular set of results may create a skewed representation of the data (Smith & Ebrahim, 2002). Aiming to counteract these biases, multiverse analysis makes use of multiple methods and assesses whether results congrue among the methods used, thus increasing reliability and transparency. Regarding this project, multiverse analysis was implemented by using various (appropiate) machine learning algorithms, along with different parametrizations. Results of all of the algorithmic models were reported regardless of quality.

Since conducting a new survey for this project might introduce all kinds confirmation bias (whether it be consciously or unconsciously), existing data was analysed. In line with Open Science principles, the data is publicly available (as long as the intention with the data is non-commercial; see appendix A). The data selected for this project is from a survey concerning mental health, parental supervision and alcohol and drug use among Scottish adolescents. Further details about the data are disclosed in the 'Data' section. Hence, the research question of this thesis project is: "What variables are able to consistently predict item non-response among Scottish adolescents through multiverse modelling?"

A short description of the data is given first. This is followed by the methodology, consisting of which steps were taken to process the data as to follow Open Science principles regarding reproducibility. All of the selected algorithmic models are also elaborated on in this section. Subsequently, the results are listed, proceeded by the discussion. Limitations and suggestions for future research conclude the project.

## Data

The survey was conducted in 2018 by Ipsos MORI Scotland, and published in 2020. It consisted of 89 questions regarding topics as alcohol and drug use, parental control and leisure activities. To illustrate with a few examples, "How is your health in general?", "Have you ever had a proper alcoholic drink - a whole drink, not just a sip?", "Have you ever been offered powders or pills that are sold as legal highs?" and "How many close friends would you say you have?" were among the questions included in the survey. A "Strengths and Difficulties" questionnaire was included at the end of the survey, and measured a variety of mental health constructs with statements like "I worry a lot" and "I have many fears, I am easily scared". The target population was adolescents ages 12 through 18, living in Scotland. The final data set provided by Ipsos consists of 635 columns, with a sample of 23.365 respondents.

Skip patterns (also named 'routing') were also present in the data. Skip patterns intend to increase efficiency of a survey by only asking certain questions to a subset of the sample based on answers on earlier questions. As the missingness on these items are (largely) dependent on other items, it could be viewed as an intended version of MAR. The missing responses (if the question was indeed not asked to a respondent as a result of a skip pattern) on these items were coded as '-1' in the data. Missing values as a result of the respondent

not answering the question were coded as '-9'. Naturally, only the latter are of interest in the scope of this project, as the researchers did not intend to have these values unrecorded.

## Methodology

In this section, an overview of the steps taken to process the data is provided, along with information about the algorithmic models and why they were deemed appropiate for the analysis. All of the processing of the data and algorithmic modelling was done in R 4.2.1. To ensure comprehensibility, trivial steps concerning the processing were not described, but are explained in the source code.

### Data preparation

The data was prepared by removing irrelevant data, selecting which variables could be taken as dependent variables based on a cluster analysis, and finally, imputations with MICE.

### Data reduction

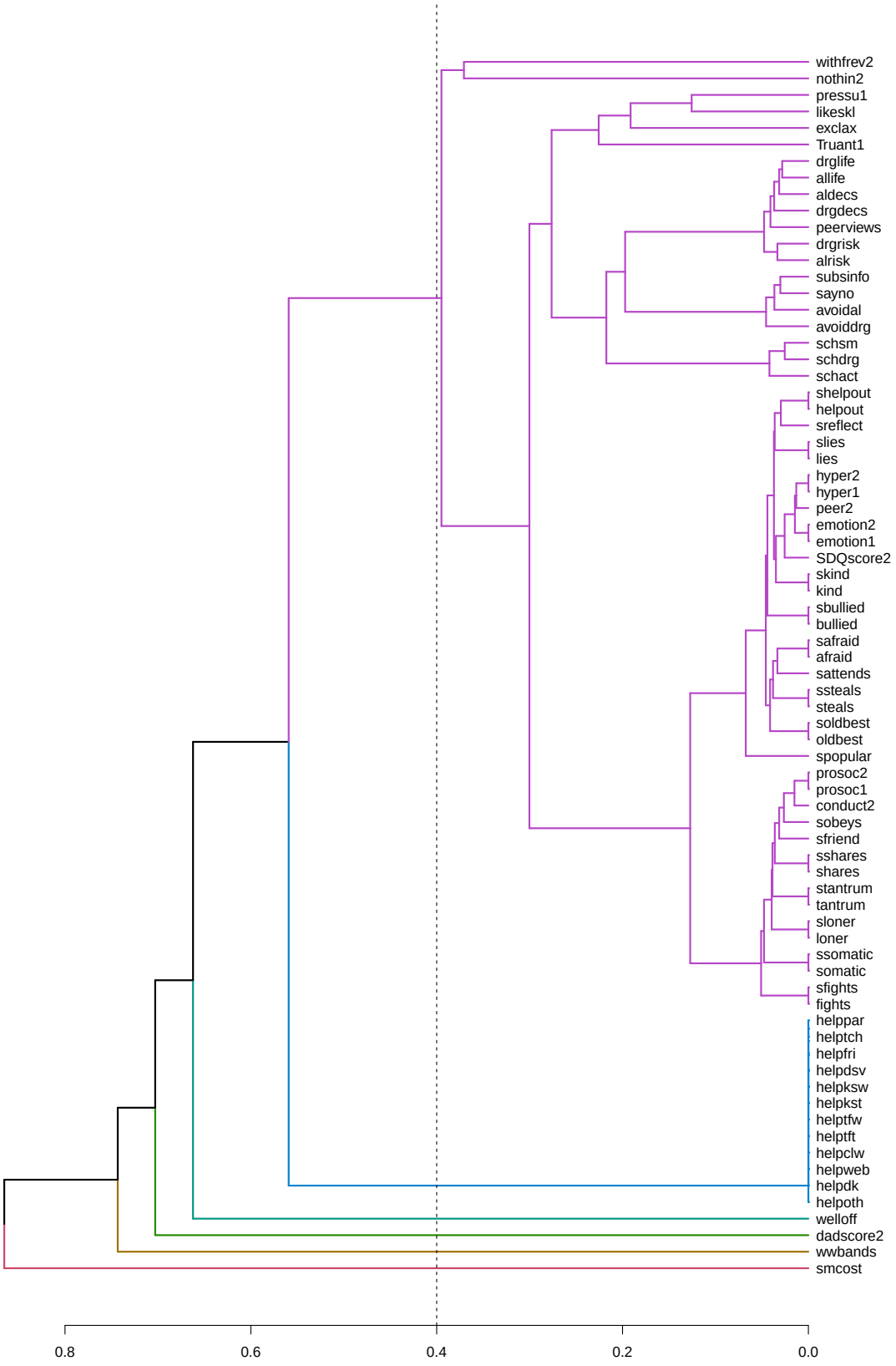
First off, data containing skip pattern questions were deemed irrelevant, and were removed accordingly. While this may appear like unnecessary data reduction, the provided codebook and example survey were examined thoroughly to support this decision. The survey was designed to include every kind of drug, for instance, which were only asked to adolescents who said that they had ever taken drugs. This results in variables that are very highly correlated, and essentially carry the same information. Another example is question 12: "How many cigarettes did you smoke on each day in the last 7 days, ending yesterday?", which end up as 7 columns that all highly correlate with each other, and are therefore redundant for the analyses. Secondly, a correlation analysis was performed to discern whether variables held similar information. Correlations of each pair of variables were calculated with pairwise comparison, and as to not remove essential data, only the first variable of the pair was removed if correlations rose above a coefficient of 0.7. Lastly, a couple of variables were deleted as a result of qualitative assessment. Administrative variables like the respondent ID, variables that convey information about the same construct but were coded differently, and trivial variables about cigarette brands all fall into this category.

### Dependent variable selection

Of the remaining 192 variables, 74 variables were considered a possible contender as a dependent variable, as they contained more than 10% missing value. Next, a cluster analysis was performed to determine whether variables followed a similar missingness pattern. In other words, if the missing values in two or more variables occur in a corresponding order, they should not be analysed separately, since they virtually carry the same information regarding missingness. Including them as a predictive power on one another may result in multi-collinearity, which should be avoided, since this could lead to a loss of statistical power and reduced interpretability.

To realise this, dummies were constructed based on the missingness of these variables. Next, a correlation matrix was calculated, essentially measuring the similarity of missingness of each pair of variables. The hierarchical cluster analysis was carried out by converting the similarity matrix to a distance matrix, and the result is shown by figure 1:

**Figure 1. Cluster dendrogram**




A cutoff score of 0.4 was applied, leading to a total of six clusters. Two of those clusters were selected to

be excluded from the analysis, as `wwbands` is not based on response, and the other `smcost`, held irrelevant information in the scope of the project. One variable of the four remaining clusters was selected arbitrarily for analysis, and are listed below in table 1, along with the amount of missingness.

Table 1: Class Distributions for the Four Target Variables

| Name | Class | Count | Percentage |
|---|---|---|---|
| welloff_DUMMY | 0 | 2631 | 11.26 |
| | 1 | 20734 | 88.74 |
| helptch_DUMMY | 0 | 2782 | 11.91 |
| | 1 | 20583 | 88.09 |
| alrisk_DUMMY | 0 | 2672 | 11.44 |
| | 1 | 20693 | 88.56 |
| dadscore2_DUMMY | 0 | 2786 | 11.92 |
| | 1 | 20579 | 88.08 |

**Imputation with MICE**

Perhaps unsurprisingly, MICE was used (with package `mice`) to impute the missing values in the data. Having extensively described the potential harm of applying MICE when the data might not be MCAR/MAR, this may appear tricky.

[Note: I want to explain this better, but forgot to write down the answer to the question on monday. I will contact you about this.]

Moreover, variables in the same cluster were not used in the imputation process of one another.

**Algorithmic models and parametrizations**

The goal of this project can essentially be translated to a binary classification problem, as the algorithms are trained to predict missingness in a variable by taking a respective dummy (where 0 is coded as 'missing' and 1 as 'present') as the target. As table 1 visualises clearly, all of the four target variables roughly have the same missing rate at 11%. The two classes are decidedly very imbalanced, which might pose a problem; One could predict every case as present, and would be right in 89% of those cases. Balanced accuracy was therefore used to as the main evaluation metric, as it takes both sensitivity and specificity into account. Moreover, algorithmic models whose performance is gravely affected by this, like Logistic Regression, were excluded. Algorithms who have been shown to struggle with large amounts of data (Support Vector Machines or Naive Bayes, for instance) were also rejected. Utimately, the Random Forest, Neural Networks, XGBoost Regression and Adaptive Boosting algorithms were used. In the spirit of multiverse analysis, different training and test sets were formed (with a margin of 65% for the training set) for each algorithm. A baseline model of each algorithm was implemented for all four target variables, resulting in sixteen baseline models. Various tuning methods were utilized, such as threshold tuning and a grid search for parameters. These were all validated by k-fold cross-validation, of which the stratified variant was applied to prevent the folds from having too few data points of the 'missing' class. Tuning was only performed on the training set to avoid overfitting. Details for each implementation are listed below.

**Random Forest**

The Random Forest is an ensemble learning algorithm that combines multiple decision trees to create a powerful predictive model. Each decision tree in the random forest is trained on a random subset of the training data, and the final prediction is obtained by aggregating the predictions of all individual trees. Regarding classification, this aggregation is done by taking the majority vote. Moreover, since each tree

is trained on a different subset of the data, the algorithm is not prone to overfitting. The Random Forest has been shown to be well-suited for classification tasks as well as having the ability to handle large data sets with high dimensionality (Speiser, Miller, Tooze & Ip, 2019), and for these reasons, the algorithm was selected for the analyses.

The `ranger` and `randomForest` packages were used for modelling. The `caret` package was used to tune parameters with a grid search. The mtry, splitrule and min.node.size parameters were tuned. A standard ntree of 500 was wielded.

### Neural Networks

Neural Networks are a class of deep learning models inspired by the structure and functioning of the human brain, hence the name. They consist of interconnected nodes, called neurons, which organized in layers. Each neuron applies a mathematical transformation to its inputs and passes the result to the next layer until the final output is generated. While mostly designed for sequential data, it has also been demonstrated to work well with non-sequential data (Lipton, Berkowitz & Elkan, 2015).

Implementing the Neural Networks algorithm was done with the `nnet` package. The `caret` package was used to tune the `size` and `decay` parameters along a grid search.

### XGBoost Regression

The XGBoost Regression algorithm (short for Extreme Gradient Boosting) constructs an ensemble of weak prediction models in a sequential manner. New models are trained to correct the errors made by the previous ones, gradually improving the final prediction accuracy. It has been shown to excel in efficiency and performance (Ramraj, Uzir, Sunil & Banerjee, 2016), and was therefore selected for the analyses.

The `xgboost` package was used to perform the analyses. The `nrounds`, `max_depth`, `eta`, `gamma`, `colsample_bytree`, `min_child_weight` and `subsample` parameters were tuned with a grid search (in `caret`). The `xgbTree` module was chosen in `caret`.

### Adaptive Boosting

Adaptive Boosting, also called AdaBoost, combines multiple weak classifiers to construct a strong classifier. Assigning weights to each training sample in subsequent iterations, the misclassified samples are emphasised to improve classification accuracy. Research has demonstrated its effectiveness (Margineantu & Dietterich, 1997; Feng et al., 2020), and as it is designed to handle class imbalance, the algorithm was selected for the analyses.

The Adaptive Boosting algorithm was implemented with package `adabag`. Threshold tuning was performed, meaning the best decision threshold was calculated on the training sets. Next, this optimal threshold was applied on the test set.

## Results

Results so far are listed in table 2.

## Discussion

### Limitations

### Conclusion

Table 2: Results

| Target | Model | Version | Sensitivity | Specificity | F1_Score | Balanced_Accuracy |
|---|---|---|---|---|---|---|
| welloff | Random Forest | Base | 0.0000000 | 1.0000000 | NA | 0.5000000 |
| helptch | Random Forest | Base | 0.0010277 | 1.0000000 | 0.0020534 | 0.5005139 |
| alrisk | Random Forest | Base | 0.0010695 | 0.9997238 | 0.0021322 | 0.5003967 |
| dadscore2 | Random Forest | Base | 0.3230769 | 0.9855596 | 0.4519369 | 0.6543182 |
| welloff | Random Forest | Parameter Tuned | 0.0000000 | 1.0000000 | NA | 0.5000000 |
| helptch | Random Forest | Parameter Tuned | 0.0020555 | 1.0000000 | 0.0041026 | 0.5010277 |
| alrisk | Random Forest | Parameter Tuned | 0.0096257 | 0.9995857 | 0.0190074 | 0.5046057 |
| dadscore2 | Random Forest | Parameter Tuned | 0.4010256 | 0.9805610 | 0.5192563 | 0.6907933 |
| welloff | Random Forest | Parameter Tuned + Class Weights | 0.0021739 | 0.9994487 | 0.0043197 | 0.5008113 |
| helptch | Random Forest | Parameter Tuned + Class Weights | 0.0164440 | 0.9980566 | 0.0319043 | 0.5072503 |
| alrisk | Random Forest | Parameter Tuned + Class Weights | 0.0866310 | 0.9946147 | 0.1535545 | 0.5406229 |
| dadscore2 | Random Forest | Parameter Tuned + Class Weights | 0.4502564 | 0.9773674 | 0.5567533 | 0.7138119 |
| welloff | Neural Networks | Parameter Tuned | 0.1836957 | 0.9382580 | 0.2199089 | 0.5609768 |
| helptch | Neural Networks | Parameter Tuned | 0.2219938 | 0.9407274 | 0.2673267 | 0.5813606 |
| alrisk | Neural Networks | Parameter Tuned | 0.3037433 | 0.9487711 | 0.3572327 | 0.6262572 |
| dadscore2 | Neural Networks | Parameter Tuned | 0.4194872 | 0.9450153 | 0.4595506 | 0.6822512 |
| welloff | Neural Networks | Parameter Tuned + Class Weights | 0.5956522 | 0.7145810 | 0.3096920 | 0.6551166 |
| helptch | Neural Networks | Parameter Tuned + Class Weights | 0.5477903 | 0.7451416 | 0.3189707 | 0.6464660 |
| alrisk | Neural Networks | Parameter Tuned + Class Weights | 0.6064171 | 0.7301850 | 0.3281250 | 0.6683011 |
| dadscore2 | Neural Networks | Parameter Tuned + Class Weights | 0.6953846 | 0.8561511 | 0.5042767 | 0.7757678 |
| welloff | XGBoost Regression | Base | 0.1543478 | 0.9816703 | 0.2376569 | 0.5680091 |
| helptch | XGBoost Regression | Base | 0.2651593 | 0.9755691 | 0.3667377 | 0.6203642 |
| alrisk | XGBoost Regression | Base | 0.3058824 | 0.9827396 | 0.4249629 | 0.6443110 |
| dadscore2 | XGBoost Regression | Base | 0.4430769 | 0.9719522 | 0.5369795 | 0.7075146 |
| welloff | Adaptive Boosting | Base | 0.1608696 | 0.9826351 | 0.2479062 | 0.5717523 |
| helptch | Adaptive Boosting | Base | 0.1613566 | 0.9802887 | 0.2468553 | 0.5708227 |
| alrisk | Adaptive Boosting | Base | 0.2641711 | 0.9849489 | 0.3826491 | 0.6245600 |
| dadscore2 | Adaptive Boosting | Base | 0.4307692 | 0.9734796 | 0.5296343 | 0.7021244 |
| welloff | Adaptive Boosting | Threshold Tuned | 0.6097826 | 0.7318082 | 0.3274001 | 0.6707954 |
| helptch | Adaptive Boosting | Threshold Tuned | 0.6515930 | 0.7387562 | 0.3634279 | 0.6951746 |
| alrisk | Adaptive Boosting | Threshold Tuned | 0.7304813 | 0.7214858 | 0.3757909 | 0.7259835 |
| dadscore2 | Adaptive Boosting | Threshold Tuned | 0.7210256 | 0.7997778 | 0.4506410 | 0.7604017 |

## Things that I want to implement

**Points for the introduction:**

- Mention that data collection can always be done after research has concluded, but this is a bit inefficient
- Reasoning behind choosing data about adolescents and this topic especially

**Points for the data section:**

- Add in some visualisations about data distributions / demographics

**Points for the methodology:**

- Something to implement: Hyperparameter tuning (not with a grid)
- Something to implement: More clusters with a lower distance threshold than 0.4

**Points for the results section:**

- Something to implement: interpretation of feature importance

**Points for the discussion:**

- Discussion of results and if determining whether they are good enough
- Mentioning that this project was an exploratory analysis, so results must be carefully interpreted
- Class imbalance will occur a lot in these analyses, as survey data usually does not have items where missingness is - evenly distributed
- Missing data problem extends beyond item non-response

**Points for the conclusion:**

- Distinction between don't know / refusal in the survey
- Item non-response can happen because of multiple reasons
- Maybe recommend to choose smaller data sets for future research, as algorithms like SVM could then also be used
- Really difficult to assume data is completely MAR, and although multiple imputation is great, it cannot be applied agnostically when encountering missing data

## References


Bell, B. A., Kromrey, J. D., & Ferron, J. M. (2009). Missing data and complex samples: The impact of listwise deletion vs. subpopulation analysis on statistical bias and hypothesis test results when data are MCAR and MAR. In Proceedings of the Joint Statistical Meetings, Survey Research Methods Section (Vol. 26, pp. 759-4770).

Berndt, M., Schmidt, F. M., Sailer, M., Fischer, F., Fischer, M. R., & Zottmann, J. M. (2021). Investigating statistical literacy and scientific reasoning & argumentation in medical-, social sciences-, and economics students. Learning and Individual Differences, 86, 101963.

Feng, D. C., Liu, Z. T., Wang, X. D., Chen, Y., Chang, J. Q., Wei, D. F., & Jiang, Z. M. (2020). Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. Construction and Building Materials, 230, 117000.

Ipsos MORI Scotland. (2020). Scottish Schools Adolescent Lifestyle and Substance Use Survey, 2018. [data collection]. UK Data Service. SN: 8615, http://doi.org/10.5255/UKDA-SN-8615-1

Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.

Margineantu, D. D., & Dietterich, T. G. (1997, July). Pruning adaptive boosting. In ICML (Vol. 97, pp. 211-218). Pan, W., & Tang, M. (2004). Examining the effectiveness of innovative instructional methods on reducing statistics anxiety for graduate students in the social sciences. Journal of Instructional Psychology, 31(2).

Pepinsky, T. B. (2018). A note on listwise deletion versus multiple imputation. Political Analysis, 26(4), 480-488. Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications, 9(40), 651-662.

Särndal, C. E., & Lundström, S. (2005). Estimation in surveys with nonresponse. John Wiley & Sons.

Savage, C., Hübner, N., Biewen, M., Nagengast, B., & Polikoff, M. S. (2021). Social studies textbook effects: Evidence from Texas. Aera Open, 7, 2332858421992345.

Smith, G. D., & Ebrahim, S. (2002). Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers. Bmj, 325(7378), 1437-1438.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 134, 93-101.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5), 702-712.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45, 1-67.

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. Journal of business research, 88, 428-436.

Wulff, J. N., & Jeppesen, L. E. (2017). Multiple imputation by chained equations in praxis: guidelines and review. Electronic Journal of Business Research Methods, 15(1), 41-56.