1 Introduction

Suicide causes one death every 40 seconds, which is nearly 800,000 deaths every year worldwide [2]. A lot of research to date has been about factors that increase risk of suicide, like depression. Most of this research has examined a few variables and examined empirical evidence to find causes of suicide. We are going to try a new approach by using machine learning to predict if someone will think about or attempt suicide based on data collected from people about their lives.

In this paper, we will examine how Machine learning regression methods perform in trying to predict thoughts of suicide and suicide attempts. Our dataset, the National Longitudinal Study of Adolescent to Adult Health [1], provides a large set of possible variables to predict suicide thoughts and attempts. The primary contribution of this paper will be finding a better method to predict these events using machine learning.

The organization of this paper is as follows. In Section 2, we discuss possible methods to use and the important parameters for these methods. In Section 3, we lay out the results from each method. In Section 4, we analyze the results and give a detailed discussion of which methods work best. The final section summarizes our work and provides suggestions about how future work can build on our findings.

2 Methods

2.1 Data Used

2.1.1 Data Source

Our data comes from the National Longitudinal Study of Adolescent to Adult Health [1] (Add Health). The Add Health project is made available for public use via the inter-university consortium for political and social research website. The Add Health project is a longitudinal study, meaning they collected information from the same respondents over the course of four unique waves.

The first of these waves started with an in-school questionnaire taking place during the 1994-1995 school year. This was a nationally representative sample of more than 90,000 students of grades 7-12, gathered information related to social demographics, parental education, home life, physical health, social well-being, and self-esteem. All students who participated in the in-school questionnaire were then reviewed for a follow up at-home interview including a parents or guardian. This interview included additional questions related to decision-making processes, future aspirations, employment, sexual history, substance use, and criminal activity. Parents were additionally asked about health, relationships, neighborhood characteristics, community involvement, employment, income, parent-adolescent communication/interaction, and familiarity with their adolecent's peers.

Wave II data was collected from April to August of 1996, and included almost 15,000 in-home interviews from adolescents of wave I. Questions were similar to wave I but included additional questions related to health and nutrition.

Wave III began in 2001 when the respondents were age 18-26 and aimed to collect information such as the respondent's family structure, sexual relationships, childbearing, education, employment history, community involvement, religion/spirituality, mental health, and criminal history.

Wave IV was conducted in 2008 and 2009 when respondents were of age 24 to 32. Questions focused and expanded upon education history, economic/financial history, sleep quality, nutrition, illnesses, physical activity, emotional content, relationship history, maltreatment/abuse history, pregnancy history, criminal history, and military history, as well as dates and circumstances of personally life shaping events. Additional biometrics were recorded such as height, weight, blood pressure, pulse, and more.

2.1.2 Data Quality Overview

The Add Health data is broken into 31 separate pieces that we appropriately combined to form each unique wave. As each wave contains thousands of reponses, the quality of each wave depends upon the proportion of valid responses in the data. With the goal of identifying the quality of each wave, we used the Add Health variable API to identify all invalid responses. In the following figures we show how many missing values exist for each attribute of the data in each wave, along with a reference for the number of instances and attributes in the data.

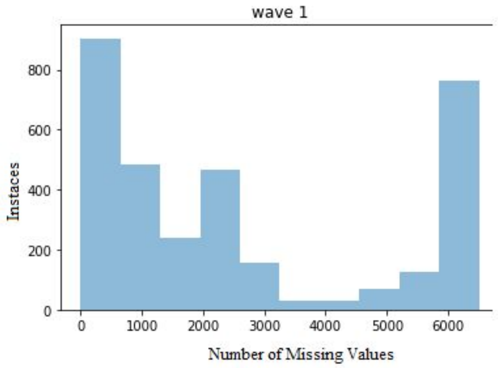


Fig . 1 Wave I: Y instances of features having X missing values

Total Instances: 6,504
Total Attributes: 3,265

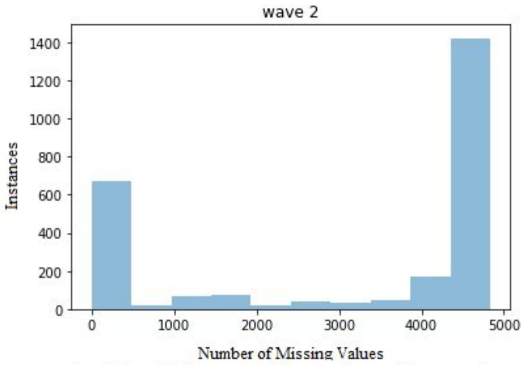


Fig . 2 Wave II: Y instances of features having X missing values

Total Instances: 4,834
Total Attributes: 2,565

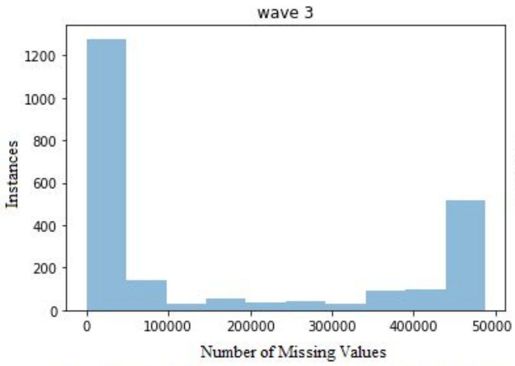


Fig . 3 Wave III: Y instances of features having X missing values

Total Instances: 487,315
Total Attributes: 2,325

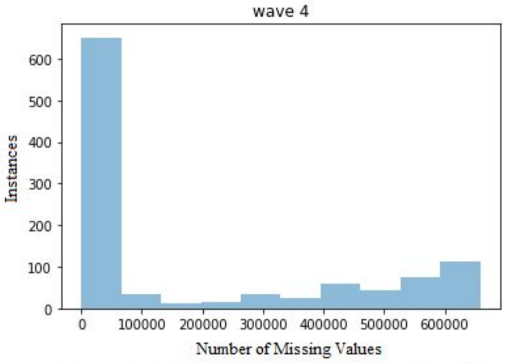


Fig . 4 Wave IV: Y instances of features having X missing values

Total Instances: 658,706
Total Attributes: 1,058

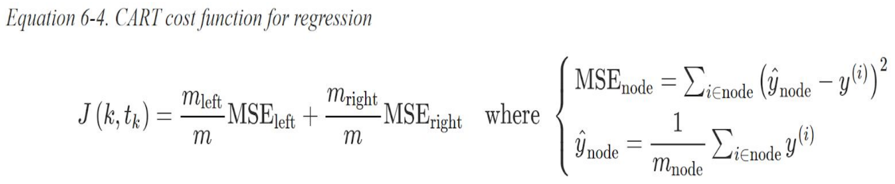
2.2 Regression Algorithm

2.2.1 Regression Trees

The first regression algorithm we used to make predictions is a Decision Regression Tree. For this project, Regression Trees were created using the DecisionTreeRegressor [3] class from the scikit-learn library.

Decision trees are defined as “versatile Machine Learning algorithms that can perform both classification and regression tasks, and even multioutput tasks.” One of the many qualities of Decision Trees is that they require very little data preparation and don’t require feature scaling or centering at all [13]. Decision Trees are called white box models, since they are intuitive and their decisions are easy to interpret.

The algorithm the Scikit-Learn uses is called the CART algorithm (Classification and Regression Tree). The algorithm attempts to split the training set in a way that minimizes the MSE (Mean Squared Error). Below is the cost function that the CART algorithm attempts to minimize. Unfortunately, the CART algorithm is a greedy algorithm: it greedily searches for an optimum split at the top level, then repeats the process at each subsequent level. It does not check whether or not the split will lead to the lowest possible impurity several levels down. A greedy algorithm often produces a solution that’s reasonably good but not guaranteed to be optimal [13].



Equation 6-4. CART cost function for regression

$$J(k, t_k) = \frac{m_{\text{left}}}{m}\text{MSE}_{\text{left}} + \frac{m_{\text{right}}}{m}\text{MSE}_{\text{right}} \quad \text{where} \begin{cases} \text{MSE}_{\text{node}} = \sum_{i \in \text{node}} \left(\hat{y}_{\text{node}} - y^{(i)}\right)^2 \\ \hat{y}_{\text{node}} = \frac{1}{m_{\text{node}}} \sum_{i \in \text{node}} y^{(i)} \end{cases}$$

Figure 5. CART cost function for regression.

By using the default hyperparameters, the tree is prone to overfitting, as seen in the two graphs below. It is best to set at least the max_depth, so the tree can stop splitting at a reasonable point. The main issue with Decision Trees is that they are very sensitive to small variations in the training data, and the smallest change can generate wildly different trees [13].

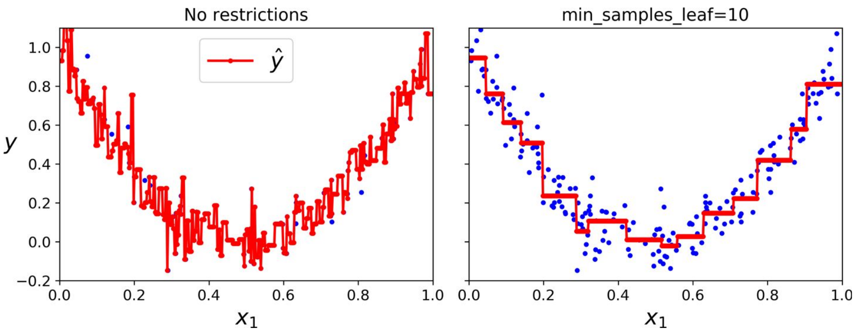


Figure 6. Two examples of Decision Tree results plotted. The left graph has no restrictions, while the right's min_samples_leaf parameter is set to 10.

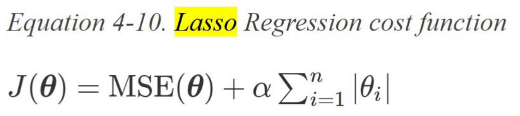
The important parameters for the Decision Regression Tree are:

- Maximum Depth, which sets the maximum depth for the tree. Branches of the tree will not grow beyond the maximum depth. Setting a maximum depth makes the tree more efficient, but may lower the accuracy. This creates a efficiency-accuracy trade off. However, as the maximum depth parameter is increased, the increase in accuracy decreases. This means that little to no additional benefit may come from increasing the maximum depth beyond a certain point. The maximum depth parameter can also be used to help prevent overfitting.
- Criterion, which sets the function used to measure the quality of the split. The options are MSE, which uses mean square error, Friedman MSE, which uses mean square error and a Friedman improvement score, and MAE, which uses mean absolute error. The default, MSE, is used for this project.

2.2.2 Lasso Regression

Another regression algorithm used to make predictions is a Lasso Regression. For this project, Lasso Regressions were created using the LassoCV [4], LassoLarsCV [5], and LassoLarsIC [6] from the scikit-learn library.

Lasso Regression, or *Least Absolute Shrinkage and Selection Operator* Regression, is a *regularized* version of Linear Regression. Regularized means constrained, which is useful since the fewer degrees of freedom a model has, the harder it will be for it to overfit the data [13]. The cost equation for Lasso is below. Lasso Regressions tend to eliminate the weights of the least important features (set them to zero), and attempts to minimize the sum of the Mean Squared Error.



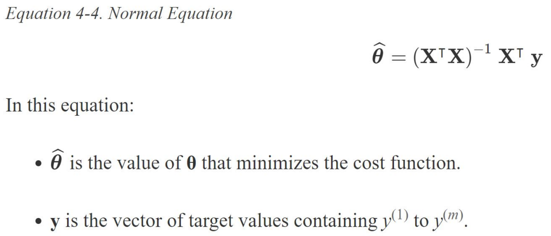
Equation 4-10. Lasso Regression cost function

$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + \alpha \sum_{i=1}^{n} |\theta_i|$$

Figure 7. Lasso Regression cost function. **θ**: the model's parameter vector, containing the bias term θ0 and the feature weights θ1 to θn. **α**: controls how much you want to regularize the model, if the value is 0, then it is a normal Linear Regression; however, if it is very large then all the weights will be close to zero.

A basic Linear Regression makes a prediction by simply computing a weighted sum of the input features, plus a constant called the bias term. We want to minimize **θ** so as to minimize the Mean Squared Error [13]. First, we need to find the value of θ that minimizes the cost function. This can be found using the Normal Equation below. $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ is a square matrix whose diagonal elements are the sums of squares [13]. We then can compute our Lasso Regression cost function above from i=1 to n (n being the number of features), and we can graph the resulting Lasso Regressions.

There are three different Lasso Regression algorithms: LassoLarsIC (least angle regression with *BIC/AIC* criterion), LassoCV (coordinate descent), and LassoLarsCV (least angle regression) [16]. Coordinate descent works to incrementally find the minimum of a function, which in this case it is the sum of squared errors. Least angle regression adds variables one at a time to see what impact they have on the model. BIC stands for Bayes Information Criterion, and it is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred [18]. AIC stands for Akaike Information Criterion, and it is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data [17].



Equation 4-4. Normal Equation

$$\widehat{\theta} = \left(\mathbf{X}^{\intercal}\mathbf{X}\right)^{-1}\ \mathbf{X}^{\intercal}\ \mathbf{y}$$

In this equation:

- $\widehat{\theta}$ is the value of $\boldsymbol{\theta}$ that minimizes the cost function.
- $\mathbf{y}$ is the vector of target values containing $y^{(1)}$ to $y^{(m)}$.

Figure 8. Normal Equation for Lasso Regression.

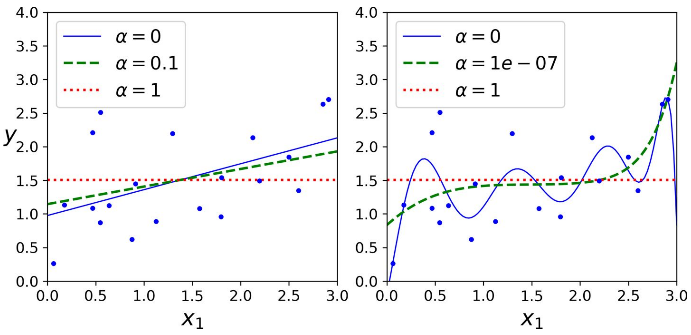


Figure 9: Two example Lasso Regressions plotted. The left Lasso Regression is for a linear function, while the right Lasso Regression is for a polynomial function.

The important parameters for the Lasso Regression are:

- Cross Validation Splitting Strategy (CV), which sets the strategy for cross validation splitting from a choice of K-Fold using an integer provided as an argument, a CV splitter, or a train-test split. This project uses the K-Fold strategy. K-Fold splits the data into K number of folds [8] [9]. There are K number of iterations where fold number K is used to test the other iterations, which are used as training data. The K-Fold cross validation predicts the accuracy of the model.
- Random State, which selects a random feature to update using a random number generator given a seed by the random state parameter. This project uses a random state for LassoCV regressions.
- Criterion, which sets the criterion to the Akaike Information Criterion (AIC) or the Bayes Information Criterion (BIC). AIC focuses on finding the effects of each variable on the class, while BIC focuses on creating a model of the most important predictors of class [10]. This project uses BIC.
- Alpha, which is a parameter used for regularization. Regularization is a process that uses the alpha parameter to limit the number of features selected [11]. As alpha increases, fewer features are selected.

2.2.3 Support Vector Machine (SVM)

The last type of regression algorithm used to make predictions in the project is a Support Vector Machine. For this project, Support Vector Machines were created using the SVR [7] class from the scikit-learn library.

A Support Vector Machine (SVM) is defined as "a powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression, and even outlier detection." To use an SVM for regression, the objective is to fit as many class instances on the *street* (parallel dashed lines that divide classes) while also limiting *margin violations* (fitting the wrong class onto the margin). The SVM Regression can be controlled by a hyperparameter ϵ, which is how wide the street is. An example is pictured below. Since adding more training instances to the model does not affect its predictions, the model is ϵ-insensitive [13].

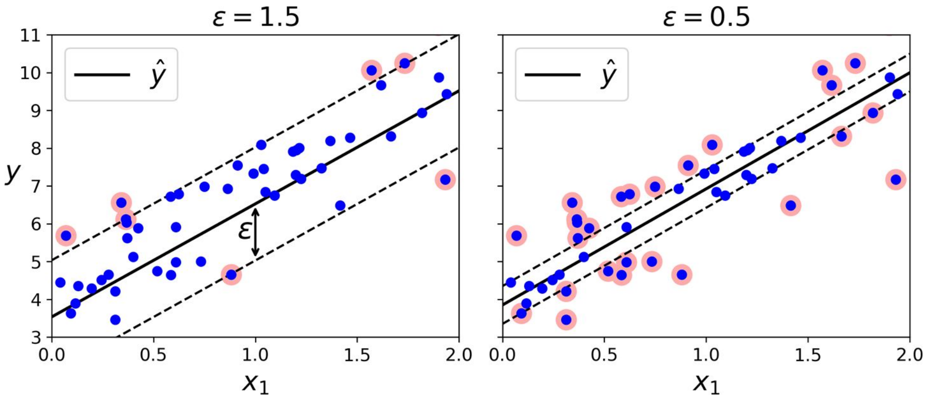


Figure 10. Two example Support Vector Machines (SVM) plotted. The left SVM has an ϵ of 1.5, while the right SVM has an ϵ of 0.5.

There are several different versions of SVM Regression, such as LinearSVR (useful for linear regression tasks), and SVR (uses kernels). A *kernel* is a function used to map a lower dimensional data into a higher dimensional data [15]. The LinearSVR is the regression equivalent of the LinearSVC class, while the SVR is the regression equivalent of the SVC class. The LinearSVR class scales linearly with the size of the training set (just like the LinearSVC class), while the SVR class gets very slow when the training set grows large (just like the SVC class) [13]. Several different kernels are: Linear (Fits margin around linear equation, much faster than other kernels, able to handle a large set of features), Polynomial(Fits margin to a polynomial equation, costly, but has good results), and Gaussian RBF (Slower than polynomial, sometimes produces better results) [13].

At its most basic level, what SVM algorithms do is to find a separating line (or *hyperplane*) between data of two classes (a simple example). A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts [14]. In SVR it can be defined as the line that will help us predict the continuous value or target value [15]. The SVM algorithm finds the best/optimal line by first finding the points closest to the line from both classes (called *support vectors*). Then, we compute the distance between the line and the support vectors (the distance is called the *margin*).

In SVM and SVR, there are two lines other than the Hyperplane which create a margin . The support vectors can be on the *boundary lines* or outside of it. This boundary line separates the two classes [15]. In SVM Classification, we want to maximize the margin, and the hyperplane for which the margin is *maximum* is the optimal hyperplane. If data is nonlinear, we have to convert it to linearly separable data by adding additional dimensions [14]. When it comes to SVR, we are trying to fit the error within a certain threshold. In simple regression, we are trying to minimize the error rate [15]. We are doing the opposite of SVM Classification and are trying to *minimize* the margin and *minimize* ϵ.

The important parameters for the SVM are:

- Kernel, which specifies the kernel type used in the algorithm. Kernels transform linearly inseparable data so that it can be classified, and the kernel parameter specifies the function used in this transformation [12]. For this project, the SVMs used the linear, polynomial, and radial basis function (RBF) kernels. The linear kernel is limited to making decisions along a linear boundary, while polynomial and RBF kernels can make decisions along non-linear boundaries. However, the kernels decrease in efficiency from linear to polynomial to RBF. This means that there is a efficiency-accuracy trade off.
- Maximum Iterations, which sets the maximum number of iterations the SVM will run for. Without a maximum number of iterations, the SVM can take a very long time to run. A drawback to setting a maximum number of iterations is that it can decrease accuracy. For these reasons, the maximum iterations parameter has an efficiency-accuracy trade off. However, as the maximum iterations parameter is increased, the increase in accuracy

decreases. This means that little to no additional benefit may come from increasing the maximum iterations beyond a certain point.

- Epsilon, which sets the distance that the regression can from the actual value without a penalty in the training loss function.

2.3 Feature Selection Algorithms

2.3.1 PCA

PCA stands for Principal Component Analysis, and is an *unsupervised* algorithm used primarily for *dimensionality reduction*. An unsupervised algorithm is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision [19]. Dimensionality reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables [20]. However, PCA can also be used for data visualization, noise filtering, feature extraction and engineering, and more [21]. PCA is affected by scale, so it is a good idea to scale the features before running the PCA on them [22].

There are several different versions of PCA, each used for a different purpose such as Linear PCA (the basic, default one), Randomized PCA (finds an approximation of the first *d* principal components), Incremental PCA (split the training set into mini-batches and feed an IPCA algorithm one mini-batch at a time), and Kernel PCA, which is often good at preserving clusters of instances after projection, or sometimes even unrolling datasets that lie close to a twisted manifold.

For our purposes, we used Linear PCA to reduce the dimensions of our data. This means we were decreasing the number of features that we were dealing with and obtaining a set of features that correlated the best with our target classes. This would give us a much smaller amount of variables to work with, and help us narrow down which variables were key predictors of suicide.

The PCA algorithm identifies the *hyperplane* that lies closest to the data, and then projects the data onto it. PCA then identifies the axis that accounts for the largest amount of variance in the training set, and then continues identifying axes until it has as many axes as the number of dimensions in the dataset [13]. The PCA algorithm can then be run in two different ways. The first way specifies the number of dimensions that the model will have, and then the PCA is fitted to the training data and can be tested with the testing data. The second way is declaring a certain number of variance that one would like the PCA to keep, or how similar one would want the reduced number of components to be to the original data. For example, 95% variance would mean the reduced number of components was 95% similar to the original data. After fitting the PCA to your data, it can be determined how many components were needed for 95% of the data to be retained.

It is also possible to display which attributes made up each of the components, and this will allow for the reduction of the overall number of features, which was our goal. Also, one could plot the explained variance as a function of the number of dimensions to determine the approximate number of dimensions to maintain a high variance.

In summary: in PCA dimensionality reduction, the information along the least important principal axis or axes is removed, leaving only the component(s) of the data with the highest variance [21]. The resulting reduced-dimension dataset is considered "good enough" to encode the most important relationships between the points: despite reducing the dimension of the data, the overall relationship between the data points are mostly preserved. The picture below is an example: the top right hand graph keeps the most variance, the middle less, and the bottom the least of the three.

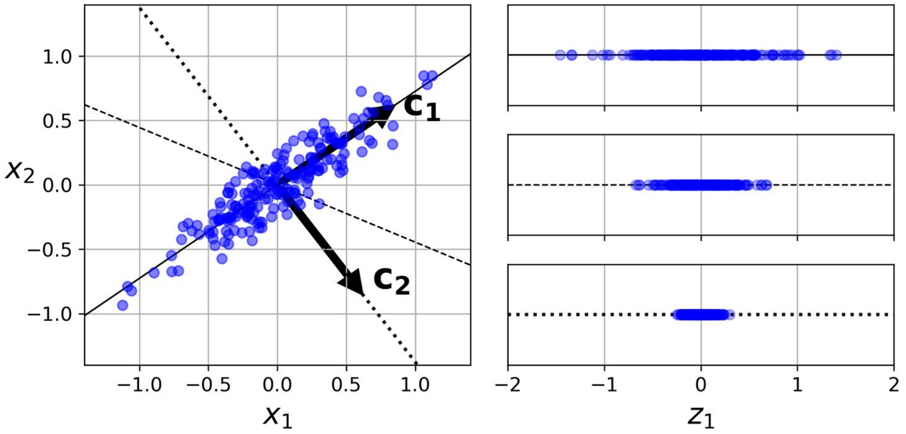


Figure 11. The leftmost graph is a graph of data points, while the rightmost graphs are different variances of the data. The top right graph keeps the most variance, the middle right less, and the bottom right the least.

2.3.2 Tree-based Feature Selection

Tree-based estimators can be used to compute *feature importances*, which could then be used to discard irrelevant or unnecessary features. Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature [24]. The SelectFromModel class in Scikit-Learn library is a meta-transformer that can be used with any estimator that has either a *coef_* (coefficient) or a *feature_importances* (how important the feature is) attribute after training data is fitted to the estimator. The features are considered unimportant and removed, if the corresponding coef_ or feature_importances_ values are below the provided threshold parameter. The threshold can be specified numerically, or through a built-in heuristic such as mean or median [23]. Therefore, the end result can tell us which features are important in the suicide model, and which we can safely discard/ignore.

First, the training data is fitted on a tree. The tree can be a DecisionTreeRegressor, an ExtraTreeClassifier, or any of the Scikit-Learn trees. The splits for the tree are based on the Mean Squared Error, and the DecisionTreeRegressor (which is what we used) runs exactly as explained above in Section 2.2.1. Then, a SelectFromModel is fitted to the resulting DecisionTreeRegressor, in which each feature has been given a feature_importance. As previously mentioned, the features are considered unimportant and removed, if the corresponding coef_ or feature_importances_ values are below the provided threshold parameter. The resulting “important” features can then be extracted after transforming the x training data with the SelectFromModel we created.

2.3.3 Lasso Regularization

Lasso Regression and Lasso Regularization are essentially the same. A Lasso Regression determines which features x, correlate best with y (the class whose value is to be predicted). After performing a Lasso Regression on data, the features and their weights can be extracted from the model. Any feature with a weight greater than zero is considered an “important” feature, and any feature with a weight equal to zero is considered unimportant [13].

First, a Lasso Regression is performed on the data using the same aforementioned cost equation. Then, the features whose weights are greater than 0, also known as the important features, are extracted.

3 Results

Key for graphs:
X-axis: Average Cross-fold Validation Score
Y-axis: Attribute Set-Algorithm-Suicide Attribute

Sets:
All = All attributes
PCA = PCA selected attributes
Lit = Attributes from literature
LaT = Attributes from Lasso and Tree-based method

Algorithms:
Tr = Regression Tree
La = Lasso Regression
SV = Support Vector Machine

Suicide Attribute:
T1 = “During the past 12 months, did you ever seriously think about committing suicide?”

Wave 1: H1SU1 Wave 2: H2SU1
"During the past 12 months, have you ever seriously thought about committing suicide?"
Wave 3: H3TO130 Wave 4: H4SE1

T2 = "During the past 12 months, how many times did you actually attempt suicide?"
Wave 1: H1SU2 Wave 2: H2SU2
"During the past 12 months, how many times have you actually attempted suicide?"
Wave 3: H3TO131 Wave 4: H4SE2

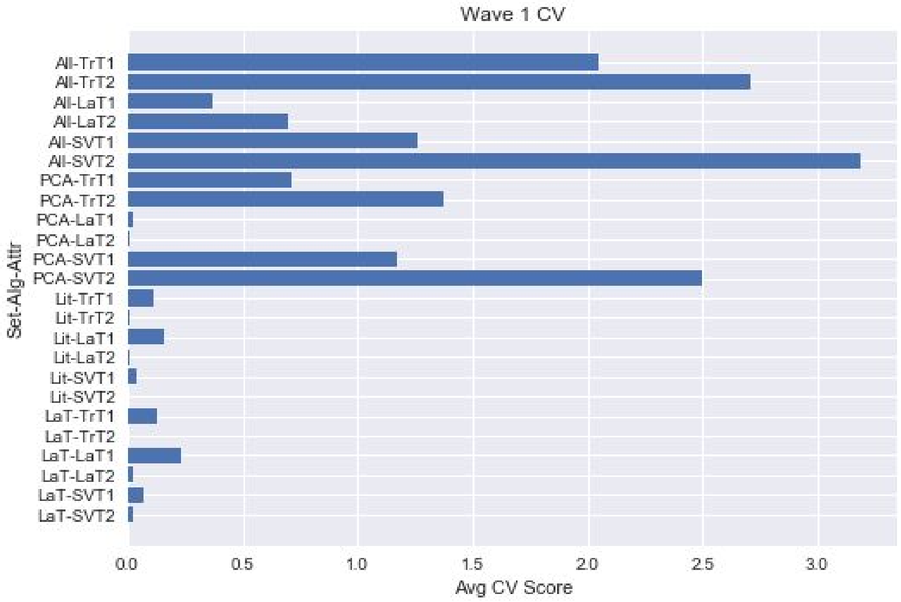


Figure 12. Average cross-fold validation scores for Wave 1.

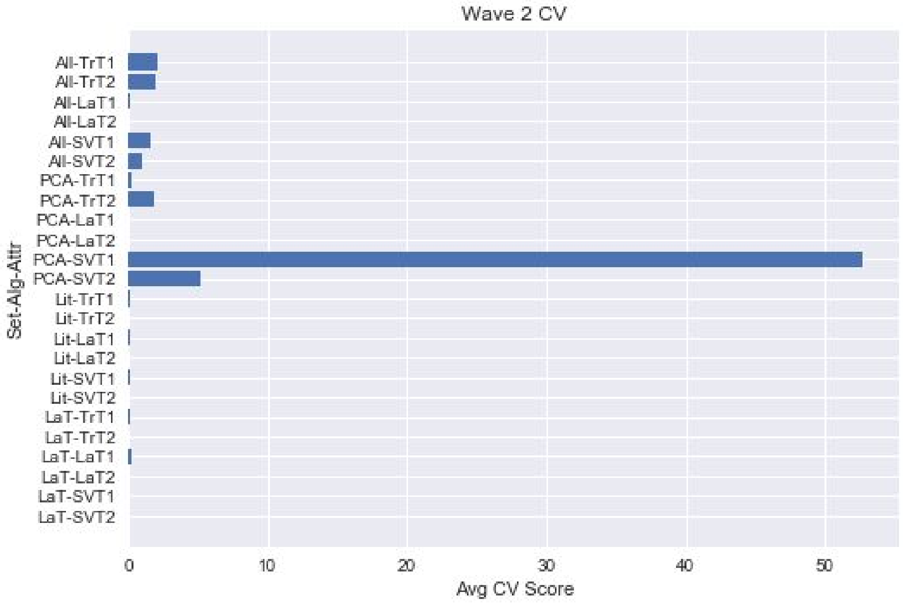


Figure 13. Average cross-fold validation scores for Wave 2.

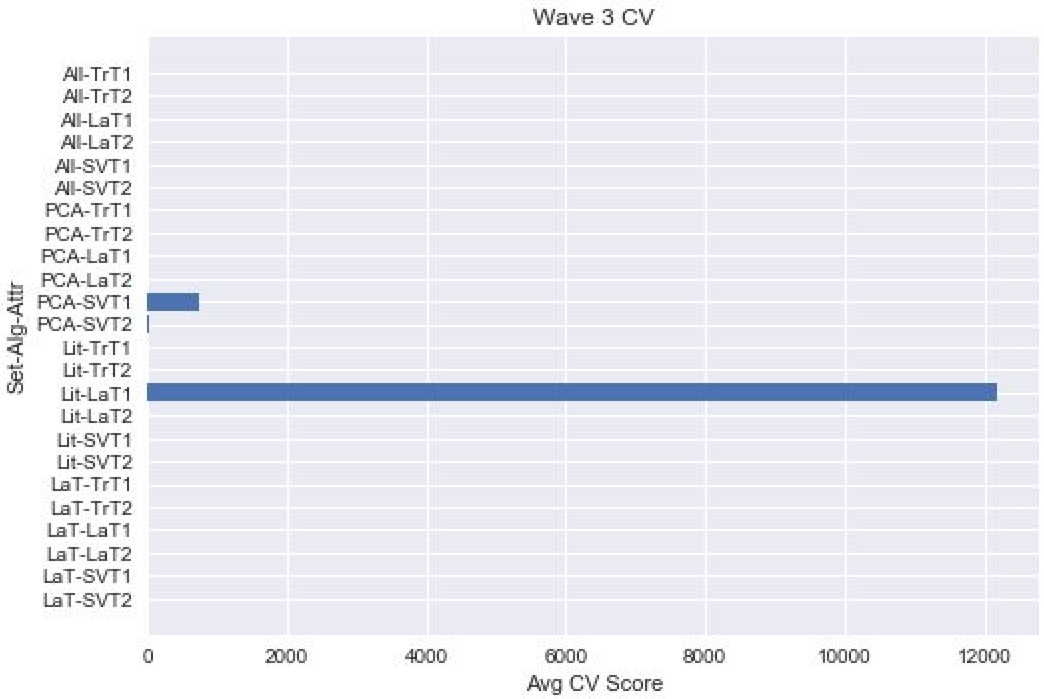


Figure 14. Average cross-fold validation scores for Wave 3. Missing some data for attributes identified in literature and attributes identified by Lasso and tree-based method.

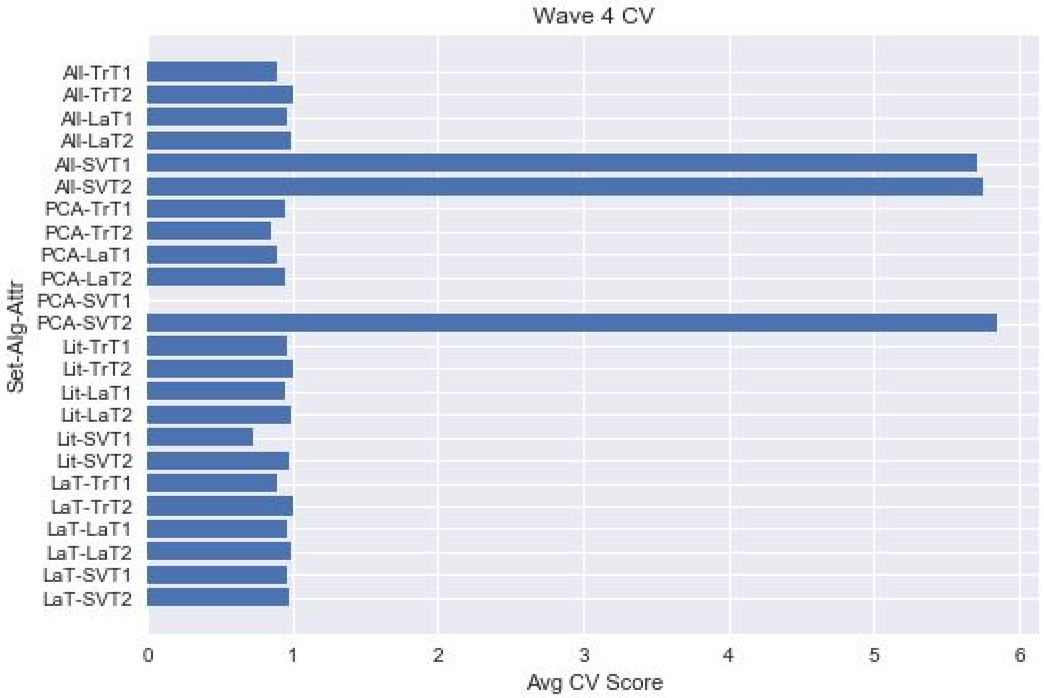


Figure 15. Average cross-fold validation scores for Wave 4. Missing average cross-fold validation score for PCA attribute set using Support Vector Machine algorithm on H4SE1.

Wave	Regression	Thoughts/Attempts	Average Cross-Validation Score
1	Tree	Thoughts	-2.05
1	Tree	Attempts	-2.71
1	Lasso	Thoughts	-0.37
1	Lasso	Attempts	-0.70
1	SVM	Thoughts	-1.26
1	SVM	Attempts	-3.19
2	Tree	Thoughts	-2.12
2	Tree	Attempts	-1.99
2	Lasso	Thoughts	0.13
2	Lasso	Attempts	-0.01
2	SVM	Thoughts	-1.58
2	SVM	Attempts	-1.03
3	Tree	Thoughts	-0.19
3	Tree	Attempts	1.00
3	Lasso	Thoughts	0.58
3	Lasso	Attempts	0.74
3	SVM	Thoughts	-13.36
3	SVM	Attempts	-1.05
4	Tree	Thoughts	0.90
4	Tree	Attempts	1.00
4	Lasso	Thoughts	0.96
4	Lasso	Attempts	0.99
4	SVM	Thoughts	-5.71

4	SVM	Attempts	-5.75

Table 1. Average cross-validation score for all attributes.

Wave	Regression	Thoughts/Attempts	Average Cross-Validation Score
1	Tree	Thoughts	-0.71
1	Tree	Attempts	-1.37
1	Lasso	Thoughts	0.02
1	Lasso	Attempts	-0.01
1	SVM	Thoughts	-1.17
1	SVM	Attempts	-2.50
2	Tree	Thoughts	-0.29
2	Tree	Attempts	-1.92
2	Lasso	Thoughts	-0.01
2	Lasso	Attempts	-0.001
2	SVM	Thoughts	-52.71
2	SVM	Attempts	-5.23
3	Tree	Thoughts	-1.83
3	Tree	Attempts	0.03
3	Lasso	Thoughts	-0.79
3	Lasso	Attempts	0.40
3	SVM	Thoughts	-736.92
3	SVM	Attempts	-27.62
4	Tree	Thoughts	0.95
4	Tree	Attempts	0.86
4	Lasso	Thoughts	0.90
4	Lasso	Attempts	0.95

4	SVM	Thoughts	?
4	SVM	Attempts	-5.85

Table 2. Average cross-validation scores for attributes identified by PCA algorithm.

Wave	Regression	Thoughts/Attempts	Average Cross-Validation Score
1	Tree	Thoughts	0.11
1	Tree	Attempts	-0.01
1	Lasso	Thoughts	0.16
1	Lasso	Attempts	0.01
1	SVM	Thoughts	0.04
1	SVM	Attempts	-0.001
2	Tree	Thoughts	0.12
2	Tree	Attempts	0.002
2	Lasso	Thoughts	0.14
2	Lasso	Attempts	0.003
2	SVM	Thoughts	-0.12
2	SVM	Attempts	-0.0002
3	Tree	Thoughts	0.46
3	Tree	Attempts	0.36
3	Lasso	Thoughts	0.43
3	Lasso	Attempts	0.16
3	SVM	Thoughts	-12186.829187
3	SVM	Attempts	-16221.972882
4	Tree	Thoughts	0.96
4	Tree	Attempts	1.00
4	Lasso	Thoughts	0.95

4	Lasso	Attempts	0.99
4	SVM	Thoughts	0.73
4	SVM	Attempts	0.98

Table 3. Average cross-validation scores for attributes identified in literature.

Wave	Regression	Thoughts/Attempts	Average Cross-Validation Score
1	Tree	Thoughts	0.13
1	Tree	Attempts	0.001
1	Lasso	Thoughts	0.23
1	Lasso	Attempts	0.02
1	SVM	Thoughts	0.07
1	SVM	Attempts	-0.02
2	Tree	Thoughts	0.08
2	Tree	Attempts	-0.01
2	Lasso	Thoughts	0.23
2	Lasso	Attempts	0.02
2	SVM	Thoughts	0.04
2	SVM	Attempts	-0.004
3	Tree	Thoughts	.77
3	Tree	Attempts	?
3	Lasso	Thoughts	.90
3	Lasso	Attempts	?
3	SVM	Thoughts	?
3	SVM	Attempts	?
4	Tree	Thoughts	0.89
4	Tree	Attempts	1.00

4	Lasso	Thoughts	0.97
4	Lasso	Attempts	0.99
4	SVM	Thoughts	0.97
4	SVM	Attempts	0.98

Table 4. Average cross-validation scores identified by Lasso and tree-based algorithms.

4 Discussion

The intent of our work thus far has been to identify feature subsets and regression algorithms that best predict thoughts of suicide and suicide attempts in our data. We used principal component analysis, manual selection through research, and LASSO/Tree-based methods, to form subsets of features for our regression algorithms to model. Additionally, we compared the results using these subsets to the results using the full feature set. The regression algorithms we used to model the data included tree regression, LASSO regression, and support vector regression.

By comparing the cross-validation scores for each feature set and each regression algorith, we deduce which methods are most appropiate for predicting thoughts of sucicide and suicide attempts according to the Add Health data. Below, we discuss the varying accuracies reported for each regression algorithm and feature set.

Tree-based regression reported noteable results for both waves I and IV, with much higher accuracy in wave I. Additionally, tree-based regression in wave I was only half as accurate when constrained to features identified by PCA. In wave IV, tree-based cross-validation scores were significantly lower than for wave I, however there was no difference in scores for the complete feature subset and PCA subset.

Support vector regression generally outperformed other regression algorithms for each wave and for each feature set. As a bonus, support vector regression typically retains over 80% of its accuracy after being limited to PCA identified features, and, in wave IV, modeling PCA features led to increased accuracy compared to the full feature set. With the highest accuracy across all waves, support vector regression is the optimal candidate for regression algorithms in this context.

LASSO regression performed significantly worse than both tree-based regression and support vector regression. Cross-validation scores for LASSO remained low across each feature set, but showed slightly higher accuracy for manually selected features and LASSO/Tree-based feature selection.

In terms of feature subsets, PCA clearly identifies the most important features more appropriately than LASSO/Tree methods and manual selection through research. This is most clear in waves II and III, where the only notably accurate results came from support vector regression on the PCA feature set. In waves I and IV, we observe similar accuracy for considering all features and limiting to PCA features.

A considerable limitation to our work has been the cost of running regression algorithms and parameter tuning both in terms of time and space. Support vector machines in-particular do not scale well to data of this size, and thus can be extremely expensive to build. However, as costly aspects of our work, such as PCA and SVM, yielded the most accurate results, we would suggest that future researchers implement these techniques with significant consideration to memory space available. Additionally, it is crucial to note that SMOTE methods were used to impute synthetic data into the wave IV class outcome, suicide attempts, as instances of this class were insufficient in the original data.

5 Conclusion

Using common regression techniques, we have demonstrated that it is possible to build models that retain nearly their full accuracy when presented with a limited subset of our data's features. Specifically, we have shown that principal component analysis appropriately identifies those features when used to build a regression model. Additionally, we have shown that support vector regression reports the highest general accuracy for our dataset and indeed performs best using PCA identified features.

For future work in predicting thoughts and attempts of suicide, we advise against manual selection of features to build a predictive model on. Attributes identified from our manual researcher yielded poor results, suggesting that factors society considers an indicator of suicidal behavior may not be accurate predictors at all. Rather, researchers should opt to identify important features using algorithmic techniques such as principal component analysis.

6 References

[1] Harris, K. M., &amp; Udry, J. R. (2018, August 6). National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]. Retrieved April 22, 2020, from https://www.icpsr.umich.edu//icpsrweb/ICPSR/studies/21600

[2] Suicide Statistics and Facts. (n.d.). Retrieved April 22, 2020, from https://save.org/about-suicide/suicide-facts/

[3] Sklearn.tree.DecisionTreeRegressor. (n.d.). Retrieved April 22, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[4] Sklearn.linear_model.LassoCV. (n.d.). Retrieved April 22, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html

[5] Sklearn.linear_model.LassoLarsCV. (n.d.). Retrieved April 22, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoLarsCV.html

[6] Sklearn.linear_model.LassoLarsIC. (n.d.). Retrieved April 22, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoLarsIC.html

[7] Sklearn.svm.SVR. (n.d.). Retrieved April 22, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

[8] M, S. (2018, November 13). Why and how to Cross Validate a Model? Retrieved April 22, 2020, from https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

[9] Krishni. (2018, December 21). K-Fold Cross Validation. Retrieved April 22, 2020, from https://medium.com/datadriveninvestor/k-fold-cross-validation-6b8518070833

[10] Phil12phil12, Glen_b -Reinstate Monica, AdamO. (2014, May). When do you use AIC vs. BIC. Retrieved April 22, 2020, from https://stats.stackexchange.com/questions/57133/when-do-you-use-aic-vs-bic

[11] Albon, C. (2017, December 20). Effect Of Alpha On Lasso Regression. Retrieved April 22, 2020, from https://chrisalbon.com/machine_learning/linear_regression/effect_of_alpha_on_lasso_regression/

[12] Chen, L. (2019, January 07). Support Vector Machine - Simply Explained. Retrieved April 22, 2020, from https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496

[13] Géron Aurélien. (2019). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. Beijing ; Boston ; Farnham ; Sebastopol ; Tokyo: OReilly.

[14] Pupale, R. (2019, February 11). Support Vector Machines (SVM) - An Overview. Retrieved from https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

[15] Bhattacharyya, I. (2018, July 12). Support Vector Regression Or SVR. Retrieved from https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

[16] Lasso model selection: Cross-Validation / AIC / BIC. (n.d.). Retrieved from https://scikit-learn.org/stable/auto_examples/linear_model/plot_lasso_model_selection.html#sphx-glr-auto-examples-linear-model-plot-lasso-model-selection-py

[17] Akaike information criterion. (2020, April 13). Retrieved from https://en.wikipedia.org/wiki/Akaike_information_criterion

[18] Bayesian information criterion. (2020, April 16). Retrieved from https://en.wikipedia.org/wiki/Bayesian_information_criterion

[19] Unsupervised learning. (2020, April 8). Retrieved from https://en.wikipedia.org/wiki/Unsupervised_learning

[20] Dimensionality reduction. (2020, April 5). Retrieved from https://en.wikipedia.org/wiki/Dimensionality_reduction

[21] VanderPlas, J. (2016, November). In Depth: Principal Component Analysis. Retrieved from https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

[22] Galarnyk, M. (2019, November 4). PCA using Python (scikit-learn). Retrieved from https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

[23] 1.13. Feature selection. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/feature_selection.html

[24] Ronaghan, S. (2019, November 1). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. Retrieved from https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

Appendix

A.1 Attributes Identified in Literature

Attribute	Description
H1HS5	In the past year, have you attended a drug abuse or alcohol abuse treatment program?
H1SU2	During the past 12 months, how many times did you actually attempt suicide?
H1SU4	Have any of your friends tried to kill themselves during the past 12 months?
H1SU6	Have any of your family tried to kill themselves during the past 12 months?
H1SU3	Did any attempt result in an injury, poisoning, or overdose that had to be treated by a doctor or nurse?
H1FS6	How often was each of the following things true during the past week? You felt depressed.

H1HS12B	Where did you receive this care? Community health clinic
H1TO54D	What kind of gun is available? Other
H1PF35	You feel socially accepted.
H1GH21	Please tell me how often you have had each of the following conditions in the past 12 months. Frequent crying
H1FS15	How often was each of the following things true during the past week? You enjoyed life.
H1FS8	How often was each of the following things true during the past week? You felt hopeful about the future.
H1JO8E	What kind of drugs had you been using? Other illegal drugs
H1JO6E	What kind of drugs had you been using? Other illegal drugs
S62O	How strongly do you agree or disagree with each of the following statements? I feel socially accepted.
S60K	In the last month, how often: Did you feel depressed or blue?
BST90P08	Modal Marital Status
PC41	Does (he/she) drink alcohol at least once a month?
PA54	How did it end?
BST90P23	Unemployment Rate
PA44	How did it end?
PA10	What is your current marital status?

PA49	How did it end?
PA61	How often do you drink alcohol?

Table 5. Wave 1 attributes names and descriptions for attributes identified in literature.

Attribute	Description
H2HS7	In the past year have you attended a drug abuse or alcohol abuse treatment program?
H2SU4	Have any of your friends tried to kill themselves during the past 12 months?
H2SU6	Have any of your family members tried to kill themselves during the past 12 months?
H2SU3	Did any attempt result in an injury, poisoning, or overdose that had to be treated by a doctor or nurse?
H2FS6	How often was each of the following things true during the past seven days?: You felt depressed.
H2PF26	You feel socially accepted.
H2JO8E	What kind of drugs had you been using?: Other illegal drugs
H2JO6E	What kind of drugs had you been using?: Other illegal drugs
BST90P08	Modal Marital Status
BST90P23	Unemployment Rate

Table 6. Wave 2 attribute names and descriptions for attributes identified in literature.

Attributes	Description
H3LM11C	Out of these categories of [Q.11B], which one best describes your first job?
H3LM26C	Out of these categories of [Q.26B], which one best describes this job?

H3EC57	What do you think are the chances that each of the following things will happen to you?: You will be divorced by age 35.
H3MR5M_A	In what month (and year) was your marriage annulled? In what month (and year) were you divorced? In what month (and year) did {YOUR SPOUSE} die?
H3MR5M_B	In what month (and year) was your marriage annulled? In what month (and year) were you divorced? In what month (and year) did {YOUR SPOUSE} die?
H3MR5M_C	In what month (and year) was your marriage annulled? In what month (and year) were you divorced? In what month (and year) did {YOUR SPOUSE} die?
H3MR5Y_A	In what month (and year) was your marriage annulled? In what month (and year) were you divorced? In what month (and year) did {YOUR SPOUSE} die?
H3MR5Y_B	In what month (and year) was your marriage annulled? In what month (and year) were you divorced? In what month (and year) did {YOUR SPOUSE} die?
H3MR5Y_C	In what month (and year) was your marriage annulled? In what month (and year) were you divorced? In what month (and year) did {YOUR SPOUSE} die?
H3MR4_B	How did this marriage end?
H3MR4_A	How did this marriage end?
H3RD124	How did your marriage to <PARTNER> end?
H3EC20	In the past 12 months, was there a time when {YOU WERE/ YOUR HOUSEHOLD WAS} evicted from your house or apartment for not paying the rent or mortgage?
H3ID15	Have you ever been diagnosed with

	depression?
H3SP9	You were depressed, during the past seven days.
H3HS7G	What type of problems were you having at that time? Had a problem that could be related to severe stress, depression, or nervousness
H3ID26F	For which of the following conditions have you taken prescription medication in the past 12 months? Depression or stress
H3GM6	Have you ever gambled to relieve uncomfortable feelings such as guilt, anxiety, helplessness, or depression?
H3HS10	Where do you usually go when you are sick or need health care?
H3PG16	Where did {SHE/YOU} go for most of {HER/YOUR} prenatal care?
H3PC8	[If the respondent is male and Q.4 = 1:] Where does she [if Q.6 > 1, add: "usually"]... [If the respondent is male and Q.5 = 1:] Where will she... [If the respondent is female and Q4 = 1:] Where do you [if Q.6 > 1, add: "usually"]... [If the respondent is female and Q.5 = 1:] Where will you go for prenatal care?
H3CJ108A	How long did you serve in jail or prison?
H3CJ108B	How long did you serve in jail or prison?
H3CJ131A	How long did you serve in jail or prison this most recent time?
H3CJ131B	How long did you serve in jail or prison this most recent time?
H3CJ156A	How long did you serve in jail or prison for these other adult convictions?
H3CJ156B	How long did you serve in jail or prison for these other adult convictions?

H3LM38	Have you ever been in the military reserves?
H3HR25	Have you ever stayed in a homeless shelter?
H3HR24	Have you ever been homeless for a week or longer--that is, you slept in a place where people weren't meant to sleep, or slept in a homeless shelter, or didn't have a regular residence in which to sleep?
H3HS7A	What type of problems were you having at that time? Needed a routine check-up
H3OD38	Did you experience neglect, or physical or sexual abuse while you were in the custody of a biological parent?
H3OD33	[If Q.23=1 and Q.31=0, ask:] Were you ever in an adoptive placement in which you experienced neglect, or physical or sexual abuse? [If Q.23=0 and Q.31=1, ask:] Were you ever in a foster placement in which you experienced neglect, or physical or sexual abuse? [If Q.23=1 and Q.31=1, ask:] Were you ever in an adoptive or foster placement in which you experienced neglect, or physical or sexual abuse?
H3ID33	In the past five years, have you spent a day or more in a facility where you were treated for a mental illness?
H3MA3	How often had your parents or other adult care-givers slapped, hit, or kicked you?
H3MA4	How often had one of your parents or other adult care-givers touched you in a sexual way, forced you to touch him or her in a sexual way, or forced you to have sexual relations?

Table 7. Wave 3 attribute names and descriptions for attributes identified in literature.

Attributes	Description
H4SE3	Did any attempt result in an injury, poisoning, or overdose that had to be treated by a doctor

	or nurse?
H4MH2	How often do you feel isolated from others?
H4SE5	Have any of them died as a result?
H4TR29	How did your {insert number if married to partner more than once} marriage to {initials} end?
H4EC12	In the past 12 months, was there a time when {YOU/YOUR HOUSEHOLD} were evicted from your house or apartment for not paying the rent or mortgage?
H4ID5H	Has a doctor, nurse or other health care provider ever told you that you have or had: depression?
H4MH22	Now, think about the past seven days. How often was each of the following things true during the past seven days: You felt depressed.
H4TO116	During the first few hours of not using {favorite drug}, do you experience one or more withdrawal symptoms such as craving {favorite drug}, feeling depressed, anxious, restless or irritable, having trouble concentrating, feeling tired or weak, having trouble sleeping, or a change in appetite?
H4TO59	Have you ever continued to drink after you realized drinking was causing you any emotional problems (such as feeling irritable, depressed, or uninterested in things or having strange ideas) or causing you any health problems (such as ulcers, numbness in your hands/feet or memory problems)?
H4TO88	During the first few hours of not using marijuana, do you experience withdrawal symptoms such as craving marijuana, feeling depressed, anxious, restless or irritable, having trouble concentrating, feeling tired or weak, having trouble sleeping, or a change in

	appetite?
H4TO117	Have you ever continued to use {favorite drug} after you realized using {favorite drug} was causing you any emotional problems (such as feeling depressed or empty, feeling irritable or aggressive, feeling paranoid or confused, feeling anxious or tense, being jumpy or easily startled) or causing you any health problems (such as heart pounding, headaches or dizziness, or sexual difficulties)?
H4TO89	Have you ever continued to use marijuana after you realized using marijuana was causing you any emotional problems (such as feeling depressed or empty, feeling irritable or aggressive, feeling paranoid or confused, feeling anxious or tense, being jumpy or easily startled) or causing you any health problems (such as persistent cough, sore throat or sinus problems, heart pounding, headaches or dizziness, or sexual difficulties)?
H4HS6	Where do you usually go when you are sick or need health care?
H4MI1	Have you ever been in the military?
H4MH29	What do you think was the main reason for these experiences? Choose only one reason.
H4ID5I	Has a doctor, nurse or other health care provider ever told you that you have or had: post-traumatic stress disorder or PTSD?
H4DS17	In the past 12 months: Someone slapped, hit, choked, or kicked you?
H4SE32	Have you ever been forced, in a non-physical way, to have any type of sexual activity against your will? For example, through verbal pressure, threats of harm, or by being given alcohol or drugs? Do not include any experiences with a parent or adult caregiver.
H4SE4	During the past 12 months, have any of your

	family or friends tried to kill themselves?

Table 8. Wave 4 attribute names and descriptions for attributes identified in literature.

A.2 Attributes Identified by Lasso and Tree-based Algorithms

Attributes	Description
H1GH2	Please tell me how often you have had each of the following conditions in the past 12 months. How often have you had a headache?
H1DS7	How often did you run away from home?
H1FS2	How often was each of the following things true during the past week? You didn't feel like eating, your appetite was poor.
H1PF35	You feel socially accepted.
H1FS4	How often was each of the following things true during the past week? You felt that you were just as good as other people.
H1TO44	How often have you taken such a drug using a needle?
H1FS8	How often was each of the following things true during the past week? You felt hopeful about the future.
H1GH28	How do you think of yourself in terms of weight?
H1TO12	Have you had a drink of beer, wine, or liquor-not just a sip or a taste of someone

	else's drink-more than 2 or 3 times in your life?
H1PF33	You like yourself just the way you are.
H1HS3	In the past year, have you received psychological or emotional counseling?
H1DS11	How often did you use or threaten to use a weapon to get something from someone?
H1NR3	Have you ever given someone sex in exchange for drugs or money?
H1RI30_3	Was a condom used when you had sexual intercourse with {INITIALS}?
H1FV4	During the past 12 months, how often did each of the following things happen? Someone cut or stabbed you.
H1RI1M_2	In what month [and year] did your relationship with {INITIALS} begin? That is, when did you first consider {INITIALS} a special friend?
H1PF26	You have a lot of energy.
H1PR8	How much do you feel that your family pays attention to you?
H1GH21	Please tell me how often you have had each of the following conditions in the past 12 months. Frequent crying

PD4C	When the twins were young children: How often were their siblings confused?
H1TO34	How old were you when you tried any kind of cocaine-including powder, freebase, or crack cocaine-for the first time? If you never tried cocaine, enter "0."
H1GH50	What time do you usually go to bed on week nights? Type in time in this format HH:MM A for AM or HH:MM P for PM. Please remember that midnight is 12:00A and noon is 12:00P!
H1SU6	Have any of your family tried to kill themselves during the past 12 months?
H1TO31	During your life, how many times have you used marijuana?
H1MF2A	FIRST or ONLY MALE FRIEND [If SCHOOL YEAR:] Does {NAME} go to school? [If SUMMER:] Did {NAME} go to school during the 1994-1995 school year?
H1TO53	Is a gun easily available to you in your home?
H1DS3	In the past 12 months, how often did you lie to your parents or guardians about where you had been or whom you were with?
H1TO13	Do you ever drink beer, wine, or liquor when you are not with your parents or other adults in your family?

H1DS15	How often were you loud, rowdy, or unruly in a public place?
H1PL7	Do you use an artificial hand, arm, leg, or foot?
H1DS2	In the past 12 months, how often did you deliberately damage property that didn't belong to you?
H1GH17	Please tell me how often you have had each of the following conditions in the past 12 months. Poor appetite
H1SU4	Have any of your friends tried to kill themselves during the past 12 months?
H1GH20	Please tell me how often you have had each of the following conditions in the past 12 months. Moodiness
H1PR3	How much do you feel that your parents care about you?
H1GH26	Has there been any time over the past year when you thought you should get medical care, but you did not?
H1MF9A	FIRST or ONLY MALE FRIEND Did you talk to {NAME} about a problem during the past seven days?

H1FS19	How often was each of the following things true during the past week? You felt life was not worth living.
H1DS13	How often did you steal something worth less than $50?
H1DA9	How many hours a week do you watch videos?
H1PR6	How much do you feel that you want to leave home?
H1FP5M	On what month [and day] did your most recent period begin?
H1TO51	Is alcohol easily available to you in your home?
H1FS6	How often was each of the following things true during the past week? You felt depressed.
H1WP17G	Which of the things listed on this card have you done with your {MOTHER/ADOPTIVE MOTHER/STEPMOTHER/FOSTER MOTHER/etc.} in the past 4 weeks? Had a serious argument about your behavior
H1PR5	How much do you feel that people in your family understand you?
H1RF5	Does he work for pay?

S45A	Receive alter mean: s45a
S55A	Does that condition involve: A heart problem?
PB4_6	What is (his/her) Hispanic background? Other Hispanic
H1NR10_3	Did you ever hold hands with {INITIALS}?
H1PF10	You never get sad.
H1TO39	During the past 30 days, how many times did you use inhalants?
H1GH6	Please tell me how often you have had each of the following conditions in the past 12 months. Feeling physically weak, for no reason
H1GH18	Please tell me how often you have had each of the following conditions in the past 12 months. Trouble falling asleep or staying asleep
H1RE6	How often do you pray?
H1PF34	You feel like you are doing everything just about right.
H1FS15	How often was each of the following things true during the past week? You enjoyed life.
S60K	Receive alter mean: s60k

H1FS3	How often was each of the following things true during the past week? You felt that you could not shake off the blues, even with help from your family and your friends.
PC42E	How much do you agree or disagree with each of the following statements? Talking about birth control with {NAME} would only encourage (him/her) to have sex.
H1TO43	During your life, have you ever injected (shot up with a needle) any illegal drug, such as heroin, or cocaine?

Table 9. Wave 1 attribute names and descriptions for attributes identified by Lasso and tree-based algorithms based on H1SU1.

Attributes	Description
H1GH21	Please tell me how often you have had each of the following conditions in the past 12 months. Frequent crying
H1MF9A	FIRST or ONLY MALE FRIEND Did you talk to {NAME} about a problem during the past seven days?
H1FS19	How often was each of the following things true during the past week? You felt life was not worth living.
AXRS59B	Ego Net Denominator axrs59b

H1TO42	During the past 30 days, how many times did you use any of these types of illegal drugs?
H1DS3	In the past 12 months, how often did you lie to your parents or guardians about where you had been or whom you were with?
H1FF4A	FIRST or ONLY FEMALE FRIEND [If SCHOOL YEAR:] Does {NAME} go to your school? [If SUMMER:] Did {NAME} go to your school during the 1994- 1995 school year?
H1FS9	How often was each of the following things true during the past week? You thought your life had been a failure.
H1DS15	How often were you loud, rowdy, or unruly in a public place?
AXRS45B	Ego Net Denominator axrs45b
H1FS3	How often was each of the following things true during the past week? You felt that you could not shake off the blues, even with help from your family and your friends.
H1RI26Y2	In what [month and] year did you have sexual intercourse with {INITIALS} most recently?
H1SU4	Have any of your friends tried to kill themselves during the past 12 months?
H1PF33	You like yourself just the way you are.

H1FS6	How often was each of the following things true during the past week? You felt depressed.
H1PR5	How much do you feel that people in your family understand you?
H1TO30	How old were you when you tried marijuana for the first time? If you never tried marijuana, enter "0."
H1RI12_1	In what year did your relationship with {INITIALS} end?

Table 10. Wave 1 attribute names and descriptions for attributes identified by Lasso and tree-based algorithms based on H1SU2.

Attributes	Description
H2GH18	How often have you had skin problems, such as itching or pimples?
H2NU17	Yesterday, did you eat avocadoes?
H2HR8F	How old is {NAME}?
H2GH4	Because of a physical, learning, or emotional condition you have had for at least a year...do you have limitations in doing strenuous activities such as running, swimming, or other sports?
H2FV14	Is a gun easily available to you in your home?

H2FS15	How often was each of the following things true during the past seven days?: You enjoyed life.
H2IR16	Number of interruptions during the interview.
H2GH20	How often have you had chest pains?
H2GH43	During the summer, what time do you usually go to bed on week nights?
H2RI20Y1	In what month [and year] did your romantic relationship with {INITIALS} end?: Year
H2GH12	How often have you felt physically weak, for no reason?
H2UV9	During the summer, how often do you sunbathe, or lie in the sun, to get a tan?
H2GH42	During the school year, what time do you usually go to bed on week nights?
H2TO58	Since {MOLI}, have you tried or used any other type of illegal drug, such as LSD, PCP, ecstasy, mushrooms, speed, ice, heroin, or pills, without a doctor's prescription?
H2RP2	How many people do you know who have AIDS? Include people who are deceased.
H2NU40	Did you eat roast beef, steak, pork, or lamb?

H2SU4	Have any of your friends tried to kill themselves during the past 12 months?
SMP08	Full sibling sample flag
H2ED11	Since school started this year, how often have you had trouble getting along with your teachers?
H2DS11	How often did you steal something worth less than $50?
H2DS5	How often did you run away from home?
H2PA1	How would your mother feel about each of the following things?:How would she feel about your having sex at this time in your life?
H2FS9	How often was each of the following things true during the past seven days?: You thought your life had been a failure.
H2PF24	You like yourself just the way you are.
H2FS8	How often was each of the following things true during the past seven days?: You felt hopeful about the future.
H2NR46C2	What other method of birth control did you or your partner use?
H2IR27	Is the respondent deaf?

H2TO68	Are illegal drugs easily available to you in your home?
BST90P11	Modal Migration Status
H2GH23	How often have you had trouble falling asleep or staying asleep?
H2RP9	Imagine that sometime soon you were to have sexual intercourse with someone just once, but were unable to use any method of birth control for some reason. What is the chance that you would get pregnant/get your partner pregnant?
H2HS5	In the past year have you received psychological or emotional counseling?
H2RI32	What is {INITIALS}'s sex?
H2JO13	Since {MOLI}, have you gotten into a fight when you had been using drugs?
H2DS2	In the past 12 months, how often did you deliberately damage property that didn't belong to you?
H2SU6	Have any of your family members tried to kill themselves during the past 12 months?
H2FS6	How often was each of the following things true during the past seven days?: You felt depressed.

H2FS19	How often was each of the following things true during the past seven days?: You felt life was not worth living.
H2TO42	Is alcohol easily available to you in your home?
H2NU55	Yesterday, did you eat cheese, processed cheese, or cheese spreads?
H2MF4A	Does {NAME} go to school?
H2DA10	How many hours a week do you play video or computer games?
H2GH22	How often have you had a poor appetite?
H2PR5	How much do you feel that people in your family understand you?
H2GH26	How often have you cried frequently?
H2RP10	Suppose that sometime soon you had sexual intercourse for a whole month, as often as you wanted to, without using any protection. What is the chance that you would get the AIDS virus?
H2GH25	How often have you been moody?
H2TO54	Since {MOLI}, have you tried or used inhalants, such as glue or solvents?

H2DS3	In the past 12 months, how often did you lie to your parents or guardians about where you had been or whom you were with?
H2RI14Y1	In what month [and year] did {INITIALS} first do this?: Year
H2PF23	You have a lot to be proud of. H2PF23
H2FS4	How often was each of the following things true during the past seven days?: You felt that you were just as good as other people.
H2PF26	You feel socially accepted.
H2IR15	When you went to the respondent's home, did you feel concerned for your safety?
H2PR6	How much do you feel that you want to leave home?
H2RI50C2	What other method of birth control did you or {INITIALS} use?
H2EE14	You will be killed by age 21.
H2RR2B	Have you had a special romantic relationship in the last 18 months with any other person?
H2TO44	Since {MOLI}, have you tried or used marijuana?

H2TO15	Since {MOLI}, have you had a drink of beer, wine, or liquor (not just a sip or a taste of someone else's drink) more than two or three times?
BST90P20	Modal Educational Attainment of Individuals Aged 25 Years and Over
H2FV3	Someone shot you.
H2DS12	How often did you act loud, rowdy, or unruly in a public place?

Table 11. Wave 2 attribute names and descriptions for attributes identified by Lasso and tree-based algorithms based on H2SU1.

Attributes	Description
BST90P21	Dispersion in Educational Attainment of Individuals Aged 25 Years and Over
H2PF26	You feel socially accepted.
H2PF7	{MOM NAME} usually knows what is going on in your life.
H2PF25	You feel like you are doing everything just about right.
H2FS15	How often was each of the following things true during the past seven days?: You enjoyed life.

H2FS6	How often was each of the following things true during the past seven days?: You felt depressed.
H2SU4	Have any of your friends tried to kill themselves during the past 12 months?
H2MO12	If you got pregnant/if you got someone pregnant you would be forced to grow up too fast.
H2GH41	In the last month, how often did a health or emotional problem cause you to miss a social or recreational activity?
H2GH12	How often have you felt physically weak, for no reason?
H2FS3	How often was each of the following things true during the past seven days?: You felt that you could not shake off the blues, even with help from your family and your friends.
H2GH42	During the school year, what time do you usually go to bed on week nights?
H2RI20Y1	In what month [and year] did your romantic relationship with {INITIALS} end?: Year
H2GH26	How often have you cried frequently?
H2UV9	During the summer, how often do you sunbathe, or lie in the sun, to get a tan?

H2GH43	During the summer, what time do you usually go to bed on week nights?
H2RI7E	In what ways did you know {INITIALS} before your romantic relationship began?: You were friends.
H2HR4A	What is {NAME}'s relationship to you?

Table 12. Wave 2 attribute names and descriptions for attributes identified by Lasso and tree-based algorithms based on H2SU2.

Wave 3:
Tree-based: ?
Lasso: ?

Attributes	Description
H4CJ7B	What were you charged with (the first time)?: other alcohol-related offenses (underage purchase or consumption; open container; public intoxication; disorderly conduct; other liquor law violations)
H4MH20	Now, think about the past seven days. How often was each of the following things true during the past seven days: You felt you were just as good as other people.
H4RD15	On average, how often (do/did) you or {initials} use a contraceptive method of birth control or disease prevention?

C	Did you {TWO/ AND H3MR_C_C} ever marry?
H4LM8	Next I'd like to record a description of your first full-time job. When you see the list of categories, please tell me which best describes what you did at your first full time job.
H4LM13	How many total hours a week do you usually spend at these jobs?
H4SE23	Considering all types of sexual activity, with how many female partners have you had sex in the past 12 months?
H4SE36G	Have you ever been told by a doctor, nurse, or other health professional that you had any of the following sexually transmitted diseases? Select all of the diseases you have had.: hepatitis B (HBV)
VERSION4	Version number of the instrument administered
H4PE30	How much do you agree with each statement about you as you generally are now, not as you wish to be in the future?: I don't worry about things that have already happened
H4PG2Y	What is the expected due date (year)?
H4RD14T	On average, how often (did/do) you have sexual relations with {initials}? By 'sexual relations' we mean vaginal intercourse, oral

	sex, anal sex, or other types of sexual activity. unit - week/month/year
H4PE4	How much do you agree with each statement about you as you generally are now, not as you wish to be in the future?: I have frequent mood swings
PRISON4	[if PRISON4=0, ask version:] How old were you the last time you went to jail, prison, juvenile detention or other correctional facility? [if PRISON4=1 ask, version:] How old were you when you went to jail, prison, juvenile detention or other correctional facility this time?
H4ED3E	What is the fifth most recent degree you have received?
H4WP5	How old were you when your biological mother went to jail or prison (the first time)? [years]
H4MI16C	During your combat deployment, did you see anyone wounded, killed, or dead? Select all that apply.: Yes, civilian
H4ID5E	Has a doctor, nurse or other health care provider ever told you that you have or had: heart disease?
H4WP6	How old were you when your biological mother was released from jail or prison (most recently)? [years]

H4TO26	During the past 30 days, on how many days have you used chewing tobacco (such as Red Man, Garrett, or Beechnut) or snuff (such as Skoal, Skoal Bandits, or Copenhagen)?
H4SE6	Have you ever had vaginal intercourse? (Vaginal intercourse is when a man inserts his penis into a woman's vagina.)
H4MI4C	In which branches of the military have you served? You may give more than one answer.: Marine Corps
H4WP30	(Has/did) your (father figure) ever (spent/spend) time in jail or prison?
H4MH25	Now, think about the past seven days. How often was each of the following things true during the past seven days: You enjoyed life.
H4SE4	During the past 12 months, have any of your family or friends tried to kill themselves?
H4TO20	When you smoked the most, did you smoke cigarettes even if you were so ill that you were in bed most of the day?
H4SE37N	In the past 12 months, have you been told by a doctor, nurse, or other health professional that you had any of the following sexually transmitted diseases? Select all of the diseases you have had.: any other sexually transmitted disease

H4MI1	Have you ever been in the military?
H4IR8A	Specify other location.
H4WP39	How many times has your (father figure) paid your living expenses or given you $50 or more to pay living expenses during the past 12 months?
H4ID23	In the past 12 months, have you had any problem with your voice? By any problem, we mean was there any time when your voice was hoarse, raspy, breathy, weak, or, generally, did not work, perform, or sound as you feel it normally would?
H4TO30	How many times have you tried but been unable to quit smoking or using tobacco for at least one month?
H4LM2	Have you ever worked for 9 weeks or more at a paying job that was at least 10 hours a week? Do not include military service.
H4WP25	How many times has your (mother figure) paid your living expenses or given you $50 or more to pay living expenses during the past 12 months?
H4MH16B	4-7-3-9-1-2-8 (8-2-1-9-3-7-4) Did respondent accurately repeat the set backwards?
H4RD4	(Do/did) either you or {initials} have a residence other than the one you share?

H4MH6	In the last 30 days, how often have you felt that difficulties were piling up so high that you could not overcome them?
H4WP14	Is your (mother figure) still alive?
H4MH2	How often do you feel isolated from others?
H4LM21B	(Does/Did) your employer make the following available to you: retirement benefits (such as 401k, 403b, or a company pension plan)?
H4WP28	Is your (father figure) still alive?
H4CJ17	Have you ever spent time in a jail, prison, juvenile detention center or other correctional facility?
H4DS17	In the past 12 months: Someone slapped, hit, choked, or kicked you?
H4WP29Y	In what year did your (father figure) die?
H4DA19	Do you own a computer?
H4HR11YA	Please tell me all of the states you have lived in. If you lived in the same state during two or more separate periods of time, list that state once for each period of time you lived there. If you have lived outside the US and its territories, select 'not in the US.' In what year did you move to {STATE}?

H4MI14	During your combat deployment, did you ever kill or think you killed someone?
H4SE34	Have you ever been physically forced to have any type of sexual activity against your will? Do not include any experiences with a parent or adult caregiver.
H4ED6	Are you currently attending a college, university, or vocational/technical school where you take courses for academic credit? If you are enrolled but on school break or vacation, count this as attending.
H4RE10	How often do you pray privately, that is, when you're alone in places other than a church, synagogue, temple, mosque, or religious assembly?
H4OD2A	Which of the following languages do you speak or write?: English
H4OD1M	1M. Respondent's date of birth - month
H4CJ15I	What charges were you convicted of or did you plead guilty to the last time?: simple assault (assaults and attempted assaults where no weapon is used and the victim is not seriously injured)
H4RE1	What is your present religion?
H4PG9	Were you and {initials} married to each other at the time of (pregnancy/birth)?

H4ID9A	Have you had gum disease (gingivitis; periodontal disease) or tooth loss because of cavities, Gum disease (gingivitis; periodontal disease) or tooth loss because of cavities in the last four weeks?
H4ID10B	Have you had fever in the last two weeks?
H4DS12	In the past 12 months, how often did you hurt someone badly enough in a physical fight that he or she needed care from a doctor or nurse?
H4CJ15J	What charges were you convicted of or did you plead guilty to the last time?: fraud, forgery, or embezzlement
H4RD16	As far as you know, during the time you and {initials} (have had/had) a sexual relationship, (has/did) {initials} ever (had/have) any other sexual partners?
H4SE36K	Have you ever been told by a doctor, nurse, or other health professional that you had any of the following sexually transmitted diseases? Select all of the diseases you have had.: urethritis
H4HR10	Have you continuously lived in {CURRENT STATE} since {LAST INTERVIEW: 1995/1996/2001/2002}?
H4ID5I	Has a doctor, nurse or other health care provider ever told you that you have or had: post-traumatic stress disorder or PTSD?

H4PE5	How much do you agree with each statement about you as you generally are now, not as you wish to be in the future?: I have a vivid imagination
H4PG13	How many weeks pregnant were you at the time of your first prenatal care visit?
H4CJ15D	What charges were you convicted of or did you plead guilty to the last time?: other drug offenses (unlawful possession, sale, use, or manufacturing of other narcotic drugs)
H4PE39	How much do you agree or disagree with the following statements?: There are many things that interfere with what I want to do.
H4TO65A	Have you every used any of the following drugs?: steroids, anabolic steroids or 'body building' drugs
H4WS2	How many have died?
H4MH23	Now, think about the past seven days. How often was each of the following things true during the past seven days: You felt that you were too tired to do things.
H4CJ13I	What charges were you convicted of or did you plead guilty to (the first time)?: simple assault (assaults and attempted assaults where no weapon is used and the victim is not seriously injured)

H4ED5B	Have you received any other degrees or certificates from a college, university, or vocational/technical school?
H4GH7	How do you think of yourself in terms of weight?
H4CJ22Y	How much time were you sentenced to serve? [years]
H4TO114	When you decided to cut down or quit using {favorite drug}, were you able to do so for at least one month?
H4DA16	When you go outside on a sunny day for more than one hour, how likely are you to use sunscreen or sunblock?
H4ID10E	Have you had blood in stool (feces) or in urine in the last two weeks?
H4RD8	How much do you love {initials}?
H4DA17	During a typical summer week, how many hours do you spend outdoors in the sun during the day?
H4SP5	How often did you have trouble falling asleep?
H4DS1	In the past 12 months, how often did you deliberately damage property that didn't belong to you?

CRP	HIGH SENSITIVITY C-REACTIVE PROTEIN (hsCRP) (MG/L)
H4PE34	How much do you agree with each statement about you as you generally are now, not as you wish to be in the future?: When making a decision, I go with my 'gut feeling' and don't think much about the consequences of each alternative
H4TO65E	Have you every used any of the following drugs?: other types of illegal drugs, such as LSD, PCP, ecstasy, heroin, or mushrooms; or inhalants
H4LB9D	How many weeks or days (before/after) the due date was {baby's first name} born (days)?
H4ID9B	Have you had active infection in the last four weeks?
H4SE22	What is your best estimate, is it
H4ID10F	Have you had frequent urination in the last two weeks?
H4MH27	Now, think about the past seven days. How often was each of the following things true during the past seven days: You felt that people disliked you, during the past seven days.

H4ID7	In the past 12 months, have you suffered any serious injuries? For example, broken bones, cuts or lacerations, burns, torn muscles, tendons or ligaments, or other injuries that interfered with your ability to perform daily tasks.
H4RD19	How often (has/did) {initials} (slapped/slap), hit or (kicked/kick) you?
H4MA6	How old were you the first time this happened?
H4WP12	How old were you when your biological father was released from jail or prison (most recently)? [years]
H4MH19	Now, think about the past seven days. How often was each of the following things true during the past seven days: You could not shake off the blues, even with help from your family and your friends.
H4TO102	Was there ever a time when you used {favorite drug} more than you do now?
H4TO65B	Have you every used any of the following drugs?: marijuana (hash, bhang, ganja)
H4MH26	Now, think about the past seven days. How often was each of the following things true during the past seven days: You felt sad.

H4LM1	Have you ever worked full time at least 35 hours a week at a paying job while you were not primarily a student? Do not include summer work.
H4CJ15A	What charges were you convicted of or did you plead guilty to the last time?: driving under the influence (DUI; DWI)
H4SE36B	Have you ever been told by a doctor, nurse, or other health professional that you had any of the following sexually transmitted diseases? Select all of the diseases you have had.: gonorrhea
H4IR12	Did the respondent's boredom or impatience negatively affect the quality of the interview?
H4CJ9J	That last time, what were you charged with?: fraud, forgery, or embezzlement
H4PG12	During this pregnancy with {initials} did (you/{initials}) ever visit a doctor, nurse-midwife or other health care provider for prenatal care, that is, for one or more pregnancy check-ups?
H4WP13	We would like to know about the woman you feel raised you. This may be your biological mother, or it may be a step-mother, adoptive mother, grandmother, etc. If you have more than one mother figure, choose the one who is most important to you. What is this person's relationship to you?

H4TO115	How many times have you tried but been unable to cut down or quit using {favorite drug} for at least one month?
H4WP23	You are satisfied with the way your (mother figure) and you communicate with each other.
H4EC2	Now think about your personal earnings. In {2006/2007/2008}, how much income did you receive from personal earnings before taxes, that is, wages or salaries, including tips, bonuses, and overtime pay, and income from self-employment? NOTE: Smallest 5 and largest 5 values are displayed.
H4RD25	How often (have/did) you (insisted/insist) on or (made/make) {initials} have sexual relations with you when (he/she) didn't want to?
H4EC15	In the past 12 months, was there a time when {YOU/YOUR HOUSEHOLD WERE/WAS} worried whether food would run out before you would get money to buy more?
H4ED4B	In what year did you receive your next most recent degree/certificate?
H4HS9	In the past 12 months have you received psychological or emotional counseling?
H4HS2A	Why do you not have health insurance? Please tell me yes or no as I read the list of

	options: You are not offered health insurance through work or school.
H4ID10G	Have you had skin rash or abscess in the last two weeks?
H4HR8G	What is {INITIAL'S} relationship to you?
H4PE16	How much do you agree with each statement about you as you generally are now, not as you wish to be in the future?: I rarely get irritated
H4MH22	Now, think about the past seven days. How often was each of the following things true during the past seven days: You felt depressed.
H4HR11YE	In what year did you move to {STATE}?
H4TO45	During the period when you drank the most, how many drinks did you usually have each time?
H4GH10	In the past 7 days, how many diet or low-calorie drinks did you have? Include diet sodas, unsweetened tea or coffee, or other drinks sweetened with artificial sweeteners. (Enter 99 for 99 or more.)
H4TO13	Do you smoke cigarettes even if you are so ill that you are in bed most of the day?

H4EC9	Suppose you and others in your household were to sell all of your major possessions (including your home), turn all of your investments and other assets into cash, and pay off all of your debts. Would you have something left over, break even, or be in debt?
H4GH11T	At what time did you last eat or drink anything other than water, including sugar-containing candy or gum? [am/pm]
H4TO111	Have you often used more {favorite drug} or used {favorite drug} longer than you intended?
H4ID4	Do you use a brace, cane, wheelchair or other device because of a physical condition?
H4SE33	How old were you the first or only time this happened?
H4SP7	Based on what you have noticed or what others have told you, are there times when you snore or you stop breathing during your sleep?
H4KK13I	Has a doctor ever told you that {child name} has any of these conditions? You may select more than one answer: Sickle cell anemia
H4CJ9E	That last time, what were you charged with?: robbery (taking or attempting to take something using a weapon or physical force)

H4SE3	Did any attempt result in an injury, poisoning, or overdose that had to be treated by a doctor or nurse?
H4MH13B	4-9-6-8 (8-6-9-4) Did respondent accurately repeat the set backwards?
H4CJ9F	That last time, what were you charged with?: theft (taking something without using force, such as larceny, burglary, or shoplifting)
H4TO55	Has there ever been a period of time when you wanted to quit or cut down on your drinking?
H4SE36L	Have you ever been told by a doctor, nurse, or other health professional that you had any of the following sexually transmitted diseases? Select all of the diseases you have had.: vaginitis
H4DS14	In the past 12 months: You saw someone shoot or stab another person?
H4IR9A	Specify other location.
H4HR11YH	In what year did you move to {STATE}?
H4SP6	How often did you have trouble staying asleep throughout the night? For example, you woke up several times at night or woke up earlier than you planned to?
H4WP7	Is your biological father still alive?

H4ED5D	Have you received any other degrees or certificates from a college, university, or vocational/technical school?
H4ARM	Which arm for blood pressure measurement
H4DA2	In the past seven days, how many times did you bicycle, skateboard, dance, hike, hunt, or do yard work?

Table 13. Wave 4 attribute names and descriptions for attributes identified by Lasso and tree-based algorithms based on H4SE1.

Attributes	Description
H4TO117	Have you ever continued to use {favorite drug} after you realized using {favorite drug} was causing you any emotional problems (such as feeling depressed or empty, feeling irritable or aggressive, feeling paranoid or confused, feeling anxious or tense, being jumpy or easily startled) or causing you any health problems (such as heart pounding, headaches or dizziness, or sexual difficulties)?
H4CJ7I	What were you charged with (the first time)?: simple assault (assaults and attempted assaults where no weapon is used and the victim is not seriously injured)
H4RD2M	What is the total amount of time that you (have been/were) involved in a romantic or sexual relationship with {initials} - months?

H4ED3C	What is the third most recent degree you have received?
H4MH16A	8-1-2-9-3-6-5 (5-6-3-9-2-1-8) Did respondent accurately repeat the set backwards?
H4LM8	Next I'd like to record a description of your first full-time job. When you see the list of categories, please tell me which best describes what you did at your first full time job.
H4LM16Y	In what year did you last work at this job? [year]
H4SP3H	On the days that you don't have to get up at a certain time, what time do you usually wake up? [Hour]
H4TO120	How old were you when you first experienced these symptoms in the same 12 month period?
H4TO3	Have you ever smoked cigarettes regularly--that is, at least one cigarette every day for 30 days?
H4SE28	In the past 12 months, how many times have you paid someone to have sex with you or has someone paid you to have sex with them?
FAST	FASTED FOR NINE HOURS OR MORE
H4DA17	During a typical summer week, how many hours do you spend outdoors in the sun during the day?

H4SE29	Are you romantically attracted to females?
H4LM21C	(Does/Did) your employer make the following available to you: paid vacation or sick leave?
PRISON4	[if PRISON4=0, ask version:] How old were you the last time you went to jail, prison, juvenile detention or other correctional facility? [if PRISON4=1 ask, version:] How old were you when you went to jail, prison, juvenile detention or other correctional facility this time?
H4LM7	Since you left your first full-time job, have you had another paying job that was at least 10 hours per week? Do not include military service.
H4DS19	In the past 12 months: You pulled a knife or gun on someone?
PRETEST4	Pretest interview
H4RD12	Select the picture, by entering the number under the picture, which best illustrates how close you feel to {initials}.
H4TO106	How often have you had legal problems because of your {favorite drug} use, like being arrested for disturbing the peace or anything else?

H4CJ9I	That last time, what were you charged with?: simple assault (assaults and attempted assaults where no weapon is used and the victim is not seriously injured)
H4ID5I	Has a doctor, nurse or other health care provider ever told you that you have or had: post-traumatic stress disorder or PTSD?
H4SE3	Did any attempt result in an injury, poisoning, or overdose that had to be treated by a doctor or nurse?
H4ID5N	Has a doctor, nurse or other health care provider ever told you that you have or had: Hepatitis C?
H4MH1	Did any of the following happen: interruption during memory task?
H4IR1	How physically attractive is the respondent?
H4DA5	In the past seven days, how many times did you participate in individual sports such as running, wrestling, swimming, cross-country skiing, cycle racing, or martial arts?
H4TO66	Have you ever injected (shot up with a needle) any illegal drug, such as heroin or cocaine?
H4EC18	Between {1995/2002} and {2006/2007/2008}, did you or others in your

	household receive any public assistance, welfare payments, or food stamps?
H4LM4	Thinking back over the period from 2001 to the previous year, how many times have you been fired, let go or laid off from a job?
FASTTIME	HOURS SINCE LAST ATE
H4HS1	Which of the following best describes your current health insurance situation?
FASTTIME	HOURS SINCE LAST ATE
H4TO69	Have you used marijuana more than 5 times?
H4SE4	During the past 12 months, have any of your family or friends tried to kill themselves?
H4RD2Y	What is the total amount of time that you (have been/were) involved in a romantic or sexual relationship with {initials} - years?
H4MH10	Did any of the following happen: interruption during memory task?

Table 14. Wave 4 attribute names and descriptions for attributes identified by Lasso and tree-based algorithms based on H4SE2.