JOURNAL OF
EMERGING INVESTIGATORS

Article

Identifying factors, such as low sleep quality, that predict suicidal thoughts using machine learning

Maggie Dong[1], **Logan Pearce**[2]
[1] Branham High School, San Jose, California
[2] Department of Psychology, Princeton University, Princeton, New Jersey

SUMMARY

Every year, around 800,000 people die by suicide worldwide, and it is the second leading cause of death for individuals aged 15-24 years. The National Survey on Drug Use and Health (NSDUH) is conducted annually to ask the general United States population aged 12 years and older questions about mental health, substance use, and suicidal thoughts. Some of the factors the survey considers are substance use, including need for treatment and disorders, and mental health topics, including major depressive episodes, suicidal ideation and attempts, and mental illness. In our research, we sought to identify associations between suicidal ideation and relevant variables, such as sleep quality, hopelessness, and anxious behavior, by analyzing survey responses from the 2020 NSDUH. Due to the density of the survey and no obvious direct relationship between any variables with suicidal ideation, through our research we aimed to clearly find and display any association. We hypothesized that frequent anxious thoughts and behavior (such as fidgeting and restlessness), feelings of sadness, and/or low sleep quality would be the factors most predictive of having suicidal thoughts. Using a random forest classifier, we found that sleep problems were highly predictive of suicidal ideation. Professionals and clinicians should keep these findings in mind when developing suicide prevention efforts, such as identifying and supporting people potentially at risk, intervening to improve sleep quality, and teaching coping skills.

INTRODUCTION

Suicide rates increase dramatically in the United States each year, especially for young people aged 15–24 years, for whom suicide has become the second leading cause of death (1, 2). Every year, around 800,000 people die by suicide worldwide, but far more attempt to take their own lives. It was estimated in 2017 that around 20 million people attempt suicide each year, and this rate is expected to continue rising (3).

Every suicide death involves many personal and social factors. The study of which factors affect suicide occurrence is of great importance because understanding the contributors to suicidal ideation could help direct intervention efforts. There is a range of individual and social factors relating to suicidal ideation, such as sex, age, geographic region, and sociopolitical setting (4). For different sexes, people of certain ages tend to attempt suicide more than others, so there seems to be an association between sex and age as well (5). Studies on college students show that depression and mental illnesses are key factors in suicidal thoughts and plans (STPs) (6, 7).

The goal of our research was to address the scientific question of which aspects are most predictive of suicidal ideation through focusing on direct relationships amongst factors, such as analyzing inadequate sleep quality directly with suicidal thoughts, using complex predictive machine learning algorithms. We hypothesized that frequent anxious thoughts and behavior (such as fidgeting and restlessness), feelings of sadness, and/or low sleep quality would be highly predictive of having suicidal thoughts. To answer our research question, we analyzed data from the 2020 National Survey on Drug Use and Health (NSDUH) (8). The Substance Abuse and Mental Health Services Administration (SAMHSA) conducts an annual nationwide survey of the general population aged 12 and older to ask questions about mental health and substance use. The purpose of the data the SAMHSA collects is to lead public health efforts to promote mental health and prevent substance misuse. The analysis we conducted is different from SAMHSA because we directly looked at relationships amongst specific factors, whereas the SAMHSA is looking for overall trends (8). Using machine learning methods, we identified how well various factors predicted suicidal ideation. We found that having trouble sleeping was the most predictive of suicidal ideation compared to all the other factors that we studied. These findings could advance future studies that are directly studying sleep and suicidal ideation and could potentially aid many clinicians to hone in their treatments of mental illness with a focus in sleep.

RESULTS

To determine the impact of different factors on suicidal ideation, we used a random forest classifier. A random forest classifier is a commonly used machine learning algorithm that makes predictions by merging the decisions of many decision-making trees. While a random forest is a machine learning algorithm used for both regression and classification tasks, a random forest classifier specifically implements the random forest algorithm for classification tasks (9). In addition, random forest classifiers successfully generate Relative Importance Values (RIVs), which indicate the relative frequency and dominance values of the species and refer to the statistical significance of factors in the data in terms of their effects on the generated model (10). An RIV can have a value from 0 to 1 with 0 being no impact and 1 being the summation of the impact of all factors. That is, a value of 1 indicates that a factor accounts for 100% of the total of all factors.

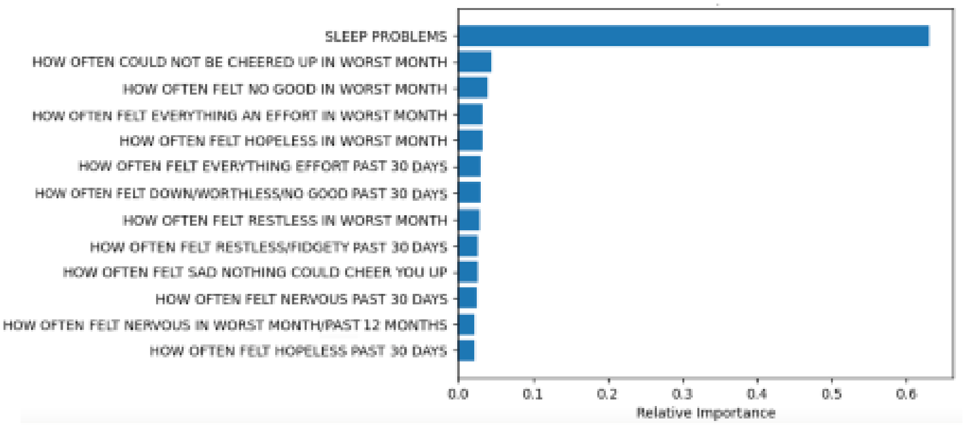

Figure 1. The Relative Importance Value (RIV) of the 13 selected factors in predicting suicidal ideation. We used a random forest classifier to assess the predictive power of 13 individual factors on suicidal ideation. The x-axis represents the 13 selected factors, and the y-axis represents the relative importance of each factor. The bars represent the relative importance of each respective factor with suicidal ideation, ranging from 0 (no impact) to 1 (high impact).

We selected thirteen relevant questions about mental health from the 2020 NSDUH to use as factors that impact suicidal ideation in our random forest classifier. We calculated the RIV of each factor. Sleep Problems were the most important factor in predicting suicidal ideation with an RIV of 0.63, which means that the impact of sleep covers 63% of the total impacts of all the factors (**Figure 1**). The second highest factor was 'How often could you not be cheered up in your worst month?' (rephrased) with an RIV of 0.045.

In the random forest classifier, training data is used to train the model, while the testing data is used to evaluate the correctness of the model calculated from training data. A learning curve is used to show a model's performance as the training set size increases. It is calculated based on accuracy from training data and from testing data to verify how reasonable the model is by testing if there is any overfitting. Overfitting is when a model learns to perform well on the training data but performs poorly on testing data or new unseen data. Another important metric is the cross-validation score, which is the average accuracy score of all the folds (11). Folds refer to subsets of data that are created by splitting the original dataset into *k* smaller subsets of approximately equal size. The accuracy score reflects the overall accuracy of the model (12). We generated the model's learning curve, which reflects the tradeoff between the training data (2/3 of all the data) and the testing data (1/3 of all the data) when the size of the training set changes. Our learning curve showed that both the training score and cross-validation score are higher than 90%, with the cross-validation score being around 94%, proving that our model is relatively accurate; generally, a higher cross-validation score shows better model performance (**Figure 2**) (11, 12). Our findings also show that with a larger training set size, the accuracy score decreases gradually with the training score while the cross-validation score increases until it reaches a training set size of around 15,000.

DISCUSSION

This study employed public-use data from SAMHSA to study factors that relate to people's STPs. Our study showed that sleep problems are highly predictive of STPs. With regards to our findings about the optimal training set size, the training score shows how well a model performs on the training data and is obtained through evaluating the model's performance on the data it was trained upon; the cross-validation score shows how well a model performs on data it hasn't seen and is obtained through assessing a model's performance on training and validation datasets (11, 12).

These results are consistent with previous studies, which also found that sleep problems can relate more to suicidal ideation than psychological problems, such as depression (13–15). Previous research also revealed that the high correlation between sleep problems and STPs was associated with the transmission of serotonin, a hormone that

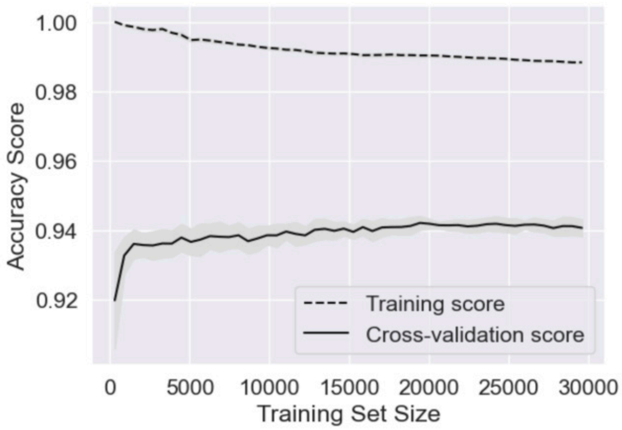

Figure 2. The learning curve (including training score curve and cross-validation score curve) for the suicide prediction model. A learning curve verifies how reasonable a model is by testing if there is any overfitting. Our learning curve (plotted as the accuracy score (y-axis) versus the training set size (x-axis)) shows that both the training score and cross-validation score are higher than 90%. The gray shading around the cross-validation score line represents variability in the model's performance.

https://doi.org/10.59720/23-999

)affects mood, sleep, and other emotional functions (16–18). All this research suggests the importance of sleeping well in suicide prevention.

Our research adds to the existing literature on the relationship between psychosocial factors and suicidal ideation, suggesting that improving sleep quality is critical in preventing suicidal thoughts. Additionally, our research tackles this specific sleep quality/suicidal ideation relationship in parallel with other relationships (including anxiety/suicidal ideation and restlessness/suicidal ideation). Furthermore, our research doesn't solely focus on sleep, but rather a variety of factors such as anxiety, restlessness, and sleep. Fortunately, sleep problems are treatable using cognitive behavioral therapy methods, and people can even use these treatments at home. For example, they can restrict their mobile phone usage, exercise more frequently, and avoid caffeine late in the day (19–21). We also recognize that some individuals may need more substantial treatment; however, making attempts to improve sleep quality is a good first step.

We can also consider an alternate hypothesis—that sleep disturbance was most predictive of suicidal ideation because it may be a symptom of several other psychological factors. That is, each other factor analyzed may have been associated with one particular aspect of cognitive distress, whereas sleep problems may reflect a combination of various aspects of cognitive distress.

Some strengths of our study are that we used a national dataset with a sample size of 30,000+ respondents aged 12 years and older. We also used a random forest classifier, a powerful machine learning tool that is made up of multiple decision trees, which makes its prediction accuracy higher compared to individual classification trees. Additionally, it is able to handle large datasets efficiently with high accuracy and no overfitting (22). However, a random forest classifier does have limitations. The algorithms are fast to train but slow to create predictions once they are trained. Our result suggests that for future research, we can reduce the training set size to 15,000. Additionally, the results from the random forest classifier only indicate which factors are predictive, not which factors *cause* suicidal ideation. Of course, given ethical and moral issues with inducing suicidal ideation, it is difficult for any researcher to establish a causal relationship between factors and suicidal thoughts. Another limitation of our research was that we only looked at a limited number of factors. In the future, researchers can build on our existing work by analyzing more data and more factors pertaining to suicidal ideation.

Additionally, the NSDUH survey asked questions about substance misuse. However, few participants reported that they had any issues with substances. Thus, we were unable to use substance misuse as a factor in our random forest classifier because it was only relevant to a few participants.

Overall, we found that sleep problems are highly predictive of STPs. Future researchers can analyze how and why sleep quality can have such an impact on suicidal ideation, or vice versa.

MATERIALS AND METHODS

Data Collection

We downloaded public-use raw data from the Substance Abuse and Mental Health Services Administration (SAMHSA) to get extensive, nationwide data about mental health (8). SAMHSA conducts a survey called the National Survey on Drug Use and Health Survey (NSDUH) annually to ask the United States general civilian population aged 12 years and older questions about mental illness and substance use. We analyzed survey data from 2020, which received 32,895 responses. The subset of relevant questions that we analyzed is listed below in "Data Analysis".

Data Analysis

We used R-Studio version 2022.7.2.576 and R version 4.2.1 to train the random forest classifier using the factors we selected from the NSDUH survey (**Figure 1**). Each relevant question (below) was a factor in the random forest. Participants could respond to each question on a scale from 1-5: *1 = All the time; 2 = Most of the time; 3 = Some of the time; 4 = A little of the time; 5 = None of the time.* A few participants chose to leave the question blank (around 2-3%). The questions we analyzed were:

- During the past 30 days, how often did you feel nervous?
- During the past 30 days, how often did you feel hopeless?
- During the past 30 days, how often did you feel restless or fidgety?
- During the past 30 days, how often did you feel so sad or depressed that nothing could cheer you up?
- During the past 30 days, how often did you feel that everything was an effort?
- During the past 30 days, how often did you feel down on yourself, no good, or worthless?
- Think of one month in the past 12 months when you were the most depressed, anxious, or emotionally stressed. In that month …
    - … how often did you feel nervous?
    - … how often did you feel hopeless?
    - … how often did you feel restless or fidgety?
    - … how often did you feel so sad or depressed that nothing could cheer you up?
    - … how often did you feel that everything was an effort?
    - … how often did you feel down on yourself, no good, or worthless?
- How often do you have sleep problems?

In the random forest classifier, each question was compared to "any thoughts or plans of suicide" to quantify the predictive power between each factor and STPs. For the suicide question, participants could answer either "NO" or "YES." "YES" was coded as *1 = Has symptom*, and "NO" was coded as *2 = Does not have symptom*. We generated an RIV using the random forest to calculate the importance of each factor in predicting STPs. We first split the selected dataset into two datasets, setting two-thirds of the data as training data and one-third of the data as test data. Finally, we trained the random forest classifier and calculated each factor's RIV. The training score and cross-validation score were used to get the accuracy score.

Received: December 26, 2022
Accepted: May 30, 2023
Published: April 30, 2024

REFERENCES

1. Hedegaard H, et al. "Suicide mortality in the United

https://doi.org/10.59720/23-999

States, 1999–2017.” *NCHS Data Brief*, vol. 330, 2018, pp. 1–8.

2. Li W, et al. “Identifying suicide risk among college students: a systematic review.” *Death Studies*, vol. 40 (7), 2019, pp.450–58, https://doi.org/10.1080/07481187.2019.1578305.
3. World Health Organization & International Association for Suicide Prevention. “Preventing suicide: a resource for media professionals, update 2017.” https://doi.org/10.5617/suicidologi.1847.
4. Turecki G, Brent DA. “Suicide and suicidal behaviour.” *Lancet*, vol 387 (10024), 2016, pp. 1227–39, https://doi.org/10.1016/s0140-6736(15)00234-2 .
5. Rhodes, Anne E et al. “Antecedents and sex/gender differences in youth suicidal behavior.” *World Journal of Psychiatry*, vol. 4 (4), 2014, pp. 120–32, https://doi.org/10.5498/wjp.v4.i4.120.
6. Auerbach, Randy P et al. “Mental disorder comorbidity and suicidal thoughts and behav-iors in the World Health Organization World Mental Health Surveys International College Student initiative.” *International Journal of Methods in Psychiatric Research*, vol. 28 (2), 2019, pp. e1752, https://doi.org/10.1002/mpr.1752.
7. Mortier P, et al. “A risk algorithm for the persistence of suicidal thoughts and behaviors during college.” *Journal of Clinical Psychiatry*, vol. 78 (7), 2017, pp. e828–e836, https://doi.org/10.4088/jcp.17m11485.
8. “National Survey on Drug Use and Health 2020 (NSDUH-2020-DS0001) | SAMHDA.” *Substance Abuse and Mental Health Services Administration*. www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001. Accessed 18 Dec. 2022.
9. Lin, Yi, and Yongho Jeon. “Random Forests and Adaptive Nearest Neighbors.” J*ournal of the American Statistical Association 101*, vol. 474, 2006, pp. 578–90, https://doi.org/10.1198/016214505000001230.
10. Chávez, Daniel, et al. “Spatial correlates of floristic and structural variation in a Neotropi-cal wetland forest.” *Wetlands Ecology and Management*, vol. 28, 2020, pp. 341–56, https://doi.org/10.1007/s11273-020-09718-z.
11. Berrar, Daniel. “Cross-Validation”, *Reference Module in Life Sciences*, 2018, http://dx.doi.org/10.1016/B978-0-12-809633-8.20349-X.
12. Wong, Tzu-Tsung & Yeh, Po-Yang. “Reliable Accuracy Estimates from k -Fold Cross Validation.” *IEEE Transactions on Knowledge and Data Engineering*, 2019, pg. 1, https://doi.org/10.1109/tkde.2019.2912815.
13. Bernert, Rebecca A, and Thomas E Joiner. “Sleep disturbances and suicide risk: A review of the literature.” *Neuropsychiatric Disease and Treatment*, vol. 3 (6), 2007, pp. 735–43, https://doi.org/10.2147/NDT.S1248.
14. Harris, Lauren M et al. “Sleep disturbances as risk factors for suicidal thoughts and be-haviours: a meta-analysis of longitudinal studies.” *Scientific Reports*, vol. 10 (1) 13888, 2020, https://doi.org/10.1038/s41598-020-70866-6.
15. Brüdern, Juliane et al. “Sleep disturbances predict active suicidal ideation the next day: an ecological momentary assessment study.” *BMC Psychiatry*, vol. 22 (1) 65, 2022, https://doi.org/10.1186/s12888-022-03716-6.
16. Agargun MY, Beisoglu L. “Sleep and suicidality: do sleep disturbances predict suicide risk?” *Sleep*, vol. 28, 2005, pp. 1039–40, https://doi.org/10.1093/sleep/28.9.1039.
17. Agargun MY, et al. “Nightmares, suicide attempts, and melancholic features in patients with unipolar major depression.” *Journal of Affective Disorders*, vol. 98 (3), 2007, pp. 267–70, https://doi.org/10.1016/j.jad.2006.08.005.
18. Adrien J. “Neurobiological bases for the relation between sleep and depression.” *Sleep Medicine Reviews*, vol. 6, 2002, pp. 341–51, https://doi.org/10.1016/s1087-0792(01)90200-x.
19. Dolezal, Brett A., et al. “Interrelationship between Sleep and Exercise: A Systematic Review.” *Advances in Preventive Medicine*, vol. 2017, 2017, pp. 1364387, https://doi.org/10.1155%2F2017%2F1364387.
20. O’Callaghan, Frances, et al. “Effects of Caffeine on Sleep Quality and Daytime Functioning.” *Risk Management and Healthcare Policy*, vol. 11, 2018, pp. 263–71, https://doi.org/10.2147%2FRMHP.S156404.
21. He, Jing-wen, et al. “Effect of Restricting Bedtime Mobile Phone Use on Sleep, Arousal, Mood, and Working Memory: A Randomized Pilot Trial.” *PLoS ONE*, vol. 15 (2), 2020, pp. e0228756, https://doi.org/10.1371/journal.pone.0228756.
22. Smith, Paul F., et al. "A comparison of random forest regression and multiple linear regression for prediction in neuroscience." *Journal of Neuroscience Methods*, vol: 220 (1), 2013, pp. 85–91, https://doi.org/10.1016/j.jneumeth.2013.08.024.