Random Forest in the early detection of suicide
Bosque Aleatorio en la detección temprana del suicidio

Daniel Alejandro Barajas Aranda, Aurora Torres Soto, María Dolores Torres Soto
Universidad Autónoma de Aguascalientes; Av. Universidad # 940, Ciudad Universitaria,
C.P. 20100, Aguascalientes, Ags. México
alengot@hotmail.com, atorres@correo.uaa.mx, mdtorres@correo.uaa.mx

Fecha de recepción: 17 de abril de 2024
Fecha de aceptación: 31 de julio de 2024

Summary. This article presents a review on the use of Random Forest in the early detection of suicide risk. Random Forest models are a machine learning technique that have been used to identify specific risk factors and improve accuracy in predicting suicide risk compared to other machine learning methods or traditional models. In this article, a methodology is proposed for the use of Random Forest in the early detection of suicide risk, which includes selecting relevant variables, training the model, and evaluating its accuracy. Some examples of results from the application of Random Forest in early detection of suicide risk are also presented. The results of the review indicate that the use of Random Forest can be a promising tool for suicide prevention and protection of human life. However, further evaluation of its clinical efficacy and ethical and responsible implementation in clinical practice is required.

Keywords: Suicide, early detection, Random Forest, machine learning, prevention tools.

Resumen. Este artículo presenta una revisión sobre el uso de Bosques Aleatorios en la detección temprana del riesgo de suicidio. Los modelos de Bosque Aleatorio son una técnica de aprendizaje automático que se ha utilizado para identificar factores de riesgo específicos y mejorar la precisión en la predicción del riesgo de suicidio en comparación con otros métodos de aprendizaje automático o modelos tradicionales. En este artículo, se propone una metodología para el uso de Bosques Aleatorios en la detección temprana del riesgo de suicidio, que incluye la selección de variables relevantes, el entrenamiento del modelo y la evaluación de su precisión. También se presentan algunos ejemplos de resultados de la aplicación de Bosques Aleatorios en la detección temprana del riesgo de suicidio. Los resultados de la revisión indican que el uso de Bosques Aleatorios puede ser una herramienta prometedora para la prevención del suicidio y la protección de la vida humana. Sin embargo, se requiere una mayor evaluación de su eficacia clínica y una implementación ética y responsable en la práctica clínica.

Palabras clave: Suicidio, detección temprana, Bosque Aleatorio, aprendizaje automático, herramientas de prevención

1 Introduction

Suicide is a significant public health issue globally, and its early detection is crucial for effective prevention and management. Recently, machine learning techniques, such as the Random Forest algorithm, have emerged as promising tools for identifying risk factors and predicting suicide risk. The Random Forest algorithm relies on the combination of multiple decision trees to improve the accuracy of predictions. This method has proven effective in identifying complex patterns within mental health data and predicting health outcomes.

In this review, we will explore the literature on the use of Random Forest in the early detection of suicide risk, including the most used predictors and the current limitations of this approach. Early detection of suicide risk is vital to prevent such tragedies, and the use of machine learning techniques like Random Forest can help enhance the precision and effectiveness of this critical task.

Moreover, it is crucial to acknowledge that although the use of machine learning techniques can provide valuable aid in the early detection of suicide risk, these techniques should not be used as substitutes for clinical and psychological evaluations conducted by mental health professionals. It is vital that these methods are used as complementary tools within a comprehensive approach to mental health care, which includes prevention strategies, therapeutic interventions, and ongoing monitoring of at-risk patients.

Finally, when using machine learning algorithms in the context of mental health, it is essential to consider ethical and privacy considerations as sensitive data is managed, and patients' rights and data privacy regulations must be respected.

2 Suicide

Death, as medically defined, is the cessation of biological functions, characterized by the irreversible termination of cardiorespiratory and/or neurocerebral functions, leading to a loss of vital signs that the body cannot maintain independently [1]. Suicide, stemming from the Latin words "sui" (of oneself) and "caedere" (to kill) [2], is the act of intentionally causing one's own death.

Various factors associated with suicide can be categorized into several causes according to Barajas [3]. These include physiological alterations that impact the individual at a biological level, such as genetic predisposition or chemical imbalances in the brain. Psychological factors encompass mental illnesses, including mood disorders like depression, bipolar disorder, and anxiety disorders, or personality disorders. Social factors consider the individual's interactions within their societal and cultural environment, such as isolation, bullying, or the lack of a support network. Finally, environmental factors external to individuals include elements like access to lethal means, exposure to suicidal behavior, or stressful life events such as financial hardship or personal loss.

Understanding these interconnected factors is crucial in developing effective strategies for suicide prevention. It is essential to recognize that suicide is often the result of a complex interplay of multiple factors rather than a single cause. Therefore, suicide prevention efforts must adopt a comprehensive approach, considering all these different elements.

Furthermore, it is vital to dispel the stigma associated with suicide, as it can function as a barrier to individuals seeking help. Public education about suicide, its risk factors, and available resources is a vital component of suicide prevention. Comprehensive suicide prevention also involves improving access to mental health care and promoting early intervention strategies that can identify at-risk individuals before a crisis occurs. The use of predictive tools like the Random Forest algorithm can play a critical role in these early intervention strategies.

3 Random forest

Random Forest is a powerful machine learning algorithm characterized by its adaptability and robustness. Conceptualized as an "ensemble" method, it combines the outputs of multiple decision trees to refine overall prediction accuracy, as noted by Han et al. [4]. The uniqueness of Random Forest lies in the construction of its individual decision trees, each built using a distinct, randomly drawn subset of the overall dataset - a method referred to as bootstrapping. This stratagem not only curtails correlation between trees but also mitigates overfitting, enhancing the model's generalization ability and predictive precision.

A distinguishing facet of the Random Forest algorithm is its inherent feature selection capability. By identifying the most significant predictors, it refines the model's complexity and augments interpretability. This is especially beneficial in fields like healthcare, where understanding the vital risk factors can significantly influence preventive strategies and therapeutic interventions.

In the realm of suicide risk prediction, a diverse set of predictors has been identified and employed in the formulation of Random Forest models. As outlined by Chekroud et al. [5] and Ribeiro et al. [6], these predictors span various demographic, psychological, and situational factors, demonstrating the algorithm's versatility.

It is, however, imperative to remember that the performance of a Random Forest model is intrinsically tied to the quality and relevance of the input data. In suicide risk detection, data often encompass complex and sensitive variables such as mental health history, substance abuse patterns, and subjective experiences. These variables necessitate thoughtful handling to uphold ethical standards and ensure patient confidentiality.

Further, while the interpretability of Random Forest models surpasses that of many other machine learning algorithms, the complexity inherent in the methodology can pose challenges. Thus, despite their utility in providing potentially life-saving insights and predictions, the outcomes of Random Forest models should be viewed as one component in a broader spectrum of information. These models should be used in tandem with other clinical tools and expert judgment for making informed decisions in suicide prevention.

4 Materials and Methods

The proposed methodology for employing Random Forest in the early detection of suicide risk consists of a systematic and structured approach, elaborated as follows:

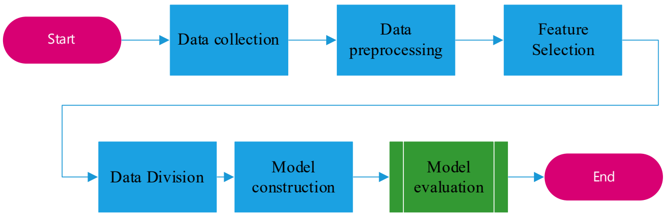

Figure 1. Methodology

Data Collection: The first step entails gathering pertinent data, serving as potential predictors of suicide risk. These include, but are not limited to, demographic attributes, medical and mental health history, substance use patterns, stress factors, and other salient variables that might contribute to suicide risk.

Data Preprocessing: Data preprocessing follows, ensuring that the collected data is complete, consistent, and free from errors. This process involves dealing with missing data, outlier detection, error correction, and data normalization to create a suitable format for further processing and analysis.

Feature Selection: Utilizing feature selection techniques, the most significant variables predicting suicide risk are identified. This process not only streamlines the model by reducing dimensionality but also enhances the model's overall predictive accuracy and interpretability.

Data Partition: Subsequently, the preprocessed dataset is split into two subsets: the training set and the testing set. The training set serves the purpose of building and refining the Random Forest model, while the testing set is reserved for evaluating the model's predictive performance.

Model Construction: The Random Forest model is then constructed using the training dataset. During this phase, model parameters are meticulously tuned to maximize prediction accuracy. This iterative process involves generating multiple decision trees and determining optimal parameters to reduce model bias and variance.

Model Evaluation: Finally, the model's predictive accuracy is evaluated using the testing dataset. Various performance metrics such as precision, sensitivity, specificity, and the area under the ROC curve are used to assess the model's capacity to predict suicide risk accurately.

This comprehensive methodology provides a robust framework for harnessing the power of Random Forest in the early detection of suicide risk. However, it is essential to note that these models should be periodically updated and validated using new data to maintain their predictive accuracy, given the dynamic nature of factors contributing to suicide risk. This iterative process ensures that the models remain relevant and useful in clinical practice over time.

5 Results

The primary classification models utilized in this study were Random Forest models. These models are machine learning algorithms that construct multiple decision trees and amalgamate their results to derive a more accurate prediction. The models were trained on diverse databases, which included relevant information for detecting suicide risk such as mood-related factors, stress, nervousness, and an amalgamation of all these variables.

Different kernels, namely Gini and Entropy, were evaluated, and the models' depths were adjusted to assess their accuracy in the early detection of suicide risk. The precision of the classifiers was assessed using various metrics such as precision, sensitivity, specificity, and area under the ROC curve.

Table 1. Viability

	Mood		Stress		Nervousness		ALL	
	Kernel		Kernel		Kernel		Kernel	
Depth	Gini	Entropy	Gini	Entropy	Gini	Entropy	Gini	Entropy
1	0.73	0.73	**0.73**	**0.73**	0.73	0.73	0.74	0.74
10	0.73	0.73	0.71	0.71	0.73	0.73	**0.76**	0.72
20	0.75	0.75	0.61	0.65	0.73	0.73	0.74	0.7
30	0.71	0.75	0.61	0.63	**0.75**	0.75	0.72	0.74
40	0.71	0.75	0.61	0.67	0.73	0.73	0.68	0.7
50	0.71	0.75	0.65	0.61	0.73	0.73	0.66	0.72
60	0.71	0.67	0.61	0.63	0.75	0.73	0.66	0.7
70	0.73	0.75	0.63	0.61	0.75	0.73	0.66	0.7
80	**0.76**	0.71	0.63	0.61	0.75	0.73	0.72	0.74
90	0.69	0.75	0.63	0.63	0.73	0.73	0.7	0.72
100	0.73	0.75	0.61	0.61	0.73	0.73	0.68	0.7

Table 1, showing the reliability of a Random Forest model in detecting suicide risk, demonstrated that accuracy values across different combinations of depth and impurity function ranged from 0.61 to 0.76. The combination of depth twenty and Gini impurity function yielded the highest accuracy for the "whole" predictor, with a value of 0.74. However, it is critical to remember that the model's accuracy can fluctuate based on the data used and the population studied.

The confusion matrices, presented in Tables 2 to 6, showed the performance of different Random Forest models trained for suicide risk detection using different variables and kernel combinations. It was observed that the models had moderate accuracy in detecting suicide risk, with some instances of false positives and false negatives.

Table 2. Confusion Matrix - mood – depth 80 kernel Gini

Confusion Matrix			
predicted class	suicides	7	6
	not suicidal	7	31
		suicides	not suicidal
		True class	

Table 3. Confusion Matrix – stress – depth 1 kernel gini

Confusion Matrix			
predicted class	suicides	0	0
	not suicidal	14	37
		suicides	not suicidal
		True class	

Table 4. Confusion Matrix – stress – depth 1 kernel entropy

Confusion Matrix			
predicted class	suicides	0	0
	not suicidal	14	37
		suicides	not suicidal
		true class	

Table 5. Confusion Matrix – nervousness – depth 30 kernel gini

Confusion Matrix			
predicted class	suicides	0	0
	not suicidal	14	37
		suicides	not suicidal
		true class	

Table 6. Confusion Matrix – all - depth 10 kernel gini

Confusion Matrix			
predicted class	suicides	2	2
	not suicidal	11	35
		suicides	not suicidal
		true class	

For instance, the "mood" predictor model accurately classified thirty-eight out of forty-five cases, with seven false positives and six false negatives. Conversely, the "stress" and "nervousness" predictor models did not correctly classify any cases as suicidal, indicating that these predictors alone might not be sufficient for early detection of suicide risk.

The "whole" predictor model accurately classified thirty-seven out of forty-eight cases, with two false positives and nine false negatives, suggesting that a combination of different predictors could improve the accuracy of suicide risk detection.

In summary, the results indicate that Random Forest models can offer moderate accuracy in the early detection of suicide risk using different predictors and kernel combinations. However, these results also highlight the need for a more extensive and thorough evaluation to determine the models' clinical effectiveness.

6 Conclusions

The research project aimed to investigate the potential of Random Forest models for early suicide risk detection, using a range of predictors including mood, stress, and nervousness indicators, as well as an amalgamation of these factors. Our findings suggest that these models can offer moderate accuracy in detecting suicide risk, albeit not without room for improvement.

The model that incorporated all variables (mood, stress, and nervousness) yielded the highest accuracy, underpinning the complexity of suicide risk detection and highlighting the importance of a multifactorial approach. This suggests that a single symptom or factor may not suffice for reliable suicide risk prediction, and a holistic view of an individual's mental health status should be considered.

However, it is important to acknowledge that these models, despite their promising results, are not foolproof. The presence of both false positives and false negatives in our findings reinforces this fact. Hence, while AI and machine learning have the potential to augment current suicide risk detection strategies, they should not replace traditional methods, but rather be integrated as supplementary tools.

Furthermore, these models' generalizability to diverse populations is yet to be ascertained. The sample size used for this study, although adequate for an initial investigation, is small. For future studies, incorporating larger and more diverse datasets could enhance the robustness and applicability of the findings.

In conclusion, the use of machine learning, specifically Random Forest models, shows potential in aiding the early detection of suicide risk. However, further research is necessary to refine these models, reduce false positives and negatives, and validate the results in broader and more diverse populations. This study paves the way for future exploration in this promising field.

References

[1] N. Campos, “Diplomado en el Protocolo de Actuación (PROL-SMDIFAGS-SUIC/2016).” 2016.

[2] “Definición de suicidio - Diccionario del español jurídico - RAE.” [Online]. Available: https://dej.rae.es/lema/suicidio. [Accessed: 26-Nov-2019].

[3] D. Barajas, “Identificación de Factores de Riesgo determinantes en el suicidio en Aguascalientes mediante la técnica de Testores Típicos,” Universidad Autonoma de Aguascalites, Aguascalientes Mexico, 2017.

[4] J. Han, M. Kamber, and J. Pei, Data Preprocessing. 2012.

[5] A. M. Chekroud, “Anticipating suicide will be hard, but this is progress,” Am. J. Psychiatry, vol. 175, no. 10, pp. 921–922, Oct. 2018.

[6] J. D. Ribeiro et al., “Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies,” Psychol. Med., vol. 46, no. 2, pp. 225–236, 2016