
Universidad de
SanAndrés

Big Data

Profesoras: María Noelia y Victoria Oubiña

# Who's next? Predicting the suicide rates in Peru

**Samuel Arispe**
**Gonzalo Ochoa**
**Facundo Valle Quintana**

# 1 Introduction

Suicide is a phenomenon that has existed since the birth of humanity. From the Bible narrating the suicide of Judas Iscariot following his betrayal of Jesus (Matthew 27:3-10) to the mass suicide in Guyana in 1978, efforts have been made to explain the reasons why people choose to take their own lives.

Approximately 700 000 people die due to suicide every year (WHO 2023), and it is the second leading cause of death in young people aged 10–24 globally (Patton et al. (2009)). These numbers keep prompting the academia to keep researching to better understand and counter the phenomena. Furthermore, several countries show increasing tendency in suicides by population. More than 77% of suicides occurred in low- and middle-income countries. Risk factors include mental disorders, impulsive crises in situations of extreme stress, and experiences such as conflicts, disasters, abuse, and discrimination.[1]

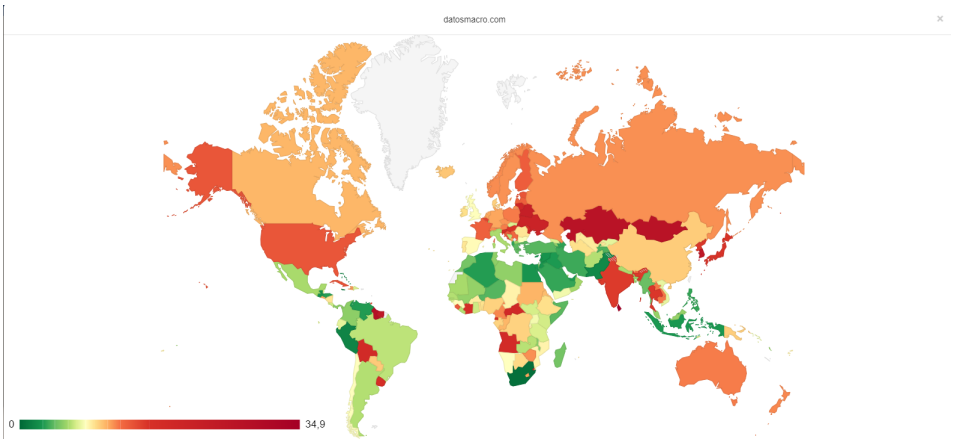


Figure 1: Suicides by population of each country

Becker and Woessmann (2018) show that Protestants, compared to Catholics, have a higher propensity for suicide. Sociological theory suggests that Protestants have "lower utility" for life and a lower cost for committing suicide. Protestant doctrine emphasizes religious individualism, while Catholics have a more integrated religious community. As a result, Protestants will have lower utility for continuing to live and a lower cost for committing suicide compared to Catholics.

Even the community can be affected by the suicide of known individuals or celebrities. Ha and Yang (2021) find this relationship between celebrity suicide and suicides in the Korean population. The groups most affected are women and young people. The public reacts more intensely to incidents of celebrity suicide of the same gender and even imitates the suicide methods used by them.

To prevent suicide, it is important to distinguish between risk factors and warning signs (changes in behavior or functioning indicating distress and the possibility of immediate suicide risk). Risk

[1] In Argentina, more specifically, suicide is the second leading cause of death for Argentine boys and girls aged 10 to 19.

factors can be long-term distal (e.g., genetic risk for suicide) or precipitating events (e.g., recent bullying or humiliation).

An analogous situation is heart disease and death from cardiac arrest. Risk factors for heart disease include genetic loading/family history, high cholesterol, obesity, unmanaged stress and depression, and lack of physical exercise. However, the presence of these risk factors does not necessarily indicate an immediate risk of a heart attack, while warning signs, such as chest pain, shortness of breath, sweating, and nausea, provide signals that a heart attack may be imminent.

The WHO launched the "LIVE LIFE - VIVIR LA VIDA" campaign to prevent suicide globally. In the Americas region, suicides are preventable with timely and low-cost interventions. Factors such as lack of awareness, underreporting, and misclassification hinder suicide prevention. Mitigating risk factors and strengthening protective factors can effectively reduce suicide rates. The PAHO strives to work on suicide prevention in the Americas, and the goal of reducing premature mortality is part of the 2020-2025 Strategic Plan. Despite challenges, continuous country participation is sought to provide data and improve decision-making in suicide prevention.

We are going to take attention for Peru suicides. According to surveillance data from the Ministry of Health of Peru, between 2016 and 2021, 71.5% of suicide attempt cases correspond to individuals aged 15 to 34. The campaign "Say Yes to Life, together we can move forward" aims to raise awareness about the importance of early detection of warning signs and promote self-care behaviors to prevent suicide. The data indicate that suicide is more frequent in women (69%) and in young people aged 15 to 24. The median age of the first suicide attempt was 22 years, and there is an increase in suicide deaths in the adolescent population.

In this work, we will attempt to predict suicides using variables related to crime as a predictors, besides the common features like age, education, size of household, among others. Initially, we believe that certain regional characteristics related to crime can be a valuable tool to prevent these events and conduct campaigns more efficiently. We will use a machine learning method calling Random Forest to train and test a sample and see the principal variables to explain Peru suicides focus relation with crime dataset.

## 2 Litterature review

Suicide has long captured human interest, resonating across generations and becoming a focal point of scholarly inquiry. Durkheim's seminal work in 1897, compiled in Émile Durkheim (2007), solidified its relevance in sociological academia. This discourse has since expanded beyond academic realms, permeating society with global suicide prevention campaigns and diverse interdisciplinary investigations.

Economists have delved into this sphere, presenting evolving economic theories that integrate multiple facets to comprehend this inherently complex phenomenon. Initially, Hamermesh

and Soss (1974) laid the foundation within the utility maximization theory, proposing that individuals contemplate suicide when their discounted lifetime utility falls below a threshold. Within this framework, higher income diminishes suicide rates, while advancing age heightens the likelihood of suicide.

Building upon this paradigm, Koo and Cox (2008) incorporated human capital into the utility model. They emphasized that unemployment impacts not only current income but also diminishes future income potential due to a lack of continuous job training. Their analysis pinpointed middle-aged individuals as more susceptible due to severe human capital depreciation.

Empirical studies often employ aggregated data, analyzing socio-economic factors vis-à-vis suicide rates. Education's influence remains elusive; while higher education might weaken social bonds, potentially elevating suicide rates, it may also foster heightened frustration and stress among individuals. Stress-inducing factors such as economic growth and unemployment emerge across studies. Rodríguez (2005), Viren (2005), Platt (1984), Platt and Hawton (2000), and Watanabe et al. (2006) suggest that economic growth negatively impacts suicide rates, while multiple studies confirm the positive correlation between unemployment and suicide risk. Additionally, income inequality consistently aligns with higher suicide rates across various studies (Chen et al., 2008).

Divorce, another contributing factor to societal stress and disintegration, correlates with elevated suicide rates, particularly affecting male rates more sensitively than female rates (Neumayer (2003), Rodríguez (2005), Watanabe et al. (2006), Chen et al. (2008), Koo and Cox (2008)).

Population density or growth is often construed as a proxy for urbanization, potentially undermining social integration and escalating suicide rates. One-person households and singlehood also exhibit associations with higher suicide odds (Burr et al. (1994), Qin et al. (2003)).

Considering a new sphere of data, Yarborough et al. (2022) explored if the integration of opioid-related data could boost predictive performance for ML models predicting suicide risk prediction models among individuals with mental health diagnoses; using techniques like LASSO regression and cross-validation, the study explored over 600 new variables but found that integrating opioid-related data did not significantly enhance the predictive performance of the models among individuals with mental health diagnoses.

In an extensive study conducted by Lin et al. (2020) on military personnel, the inclusion of comprehensive medical data led to exceptional accuracy in predicting suicidal ideation. This comprehensive dataset encompassed medical examinations, blood tests, and chest x-rays, focusing specifically on elevated psychological stress and heightened suicide risk prevalent among military personnel. After applying a range of machine learning algorithms, including logistic regression, decision tree, random forest, gradient boosting regression tree, support vector machine, and multi-layer sensor the models collectively surpassed 98% accuracy in predicting the presence of suicidal ideation, underlining the pivotal role of medical data in discerning suicidal

tendencies within this population.

Meanwhile, Fortaner-Uyà et al. (2023) endeavored to differentiate between Major Depressive Disorder (MDD) suicide attempters (SA) and non-attempters (nSA) using machine learning algorithms applied to complex grey matter and white matter data from structural MRI scans of 91 depressed MDD patients. Despite the high quality of their data, the models exhibited only moderate performance in classifying SA and nSA. Notably, the Support Vector Machine (SVM) model achieved the highest accuracy among the tested models, surpassing random accuracy rates but falling short of achieving robust predictive performance.

In the realm of model selection, Tate et al. (2020) investigated the efficacy of various machine learning algorithms—Random Forest, XGBoost, Logistic Regression, Neural Network, and Support Vector Machines—utilizing data from the Child and Adolescent Twin Study in Sweden (CATSS). The study revealed that none of the models significantly outperformed the others, with random forest and support vector machine models displaying slightly higher Area Under the Curve (AUC) values. However, no single model emerged as the unequivocal superior choice.

Furthermore, in their study, Turk (2023) employed various Machine Learning (ML) methods to predict suicidal risk among adolescents and young adults. The supervised learning classification algorithms, including ExtraTrees, GradientBoosting, CatBoost, XGBoost, Logistic Regression, and ensemble methods, were assessed using a 5-fold cross-validation and a holdout method. While achieving accuracies ranging between 74% and 77%, the models demonstrated a 66% hit rate in identifying individuals with suicidal thoughts within the past year. Through refinements, particularly the utilization of ensemble models, their predictive models achieved an 82% success rate in identifying suicidal thoughts. However, the study acknowledged limitations such as its cross-sectional nature, reliance on self-report scales, and the absence of additional medical data that could potentially enhance accuracy rates.

Interestingly, the resemblance of machine learning models to logistic regression suggests the potential sufficiency of simpler models, particularly in datasets characterized by predominantly linear relationships between predictors and outcomes. This observation is especially pertinent for smaller datasets, where the added complexity of advanced machine learning techniques might not yield substantial benefits in interpretability and computational resources compared to logistic regression.

# 3 Datasets

We want to use the deaths data of National Death System (SINADEF, acronym in spanish) that reports deaths information since 2017 until now, daily with type and clinical causes of deaths. This data contains also sex, marital status, education level, age, deaths place of the deaths. We want to link theses datasets with National Register of Crimes and Offense Reports (RNDDF, acronym in spanish) to obtain crime rates and incorporate other variables we consider relevant to measure or predict suicides.

We want to recover variables related to mental health of Demographic and Family Health Survey (ENDES, acronym in spanish) that reports mental health issues and several variables related to health in general to obteain good covariates. These variables will be relevants to predict our chosen outcome of suicides. Finally, we will use National Household Survey (ENAHO, acronym in spanish) to recover variables about household characteristics and recover the type of climate in the different locality/districts.

We have to deal with the fact that there is not way to link all datasets individually way because we do not have way to recover exact number of national identification document or similar ID. In this case, we will use the variability of deaths by localities/districts and assign 0 if the localities/districts have not had suicides and 1 if the localities/districts have had at least one suicide death in 2019. We use this year because was the immediately before year to COVID-19 pandemic. In this sense we want to predict suicides at localities/district level. This decision is coherent because not all localities/districts suffered suicides.

First, if we make an exploratory analysis we see in the figure 2 that suicides was in 3rd or 4th place in ranking of type of violent deaths rates of each year having in average near to 2 suicides per 100,000 inhabitants.

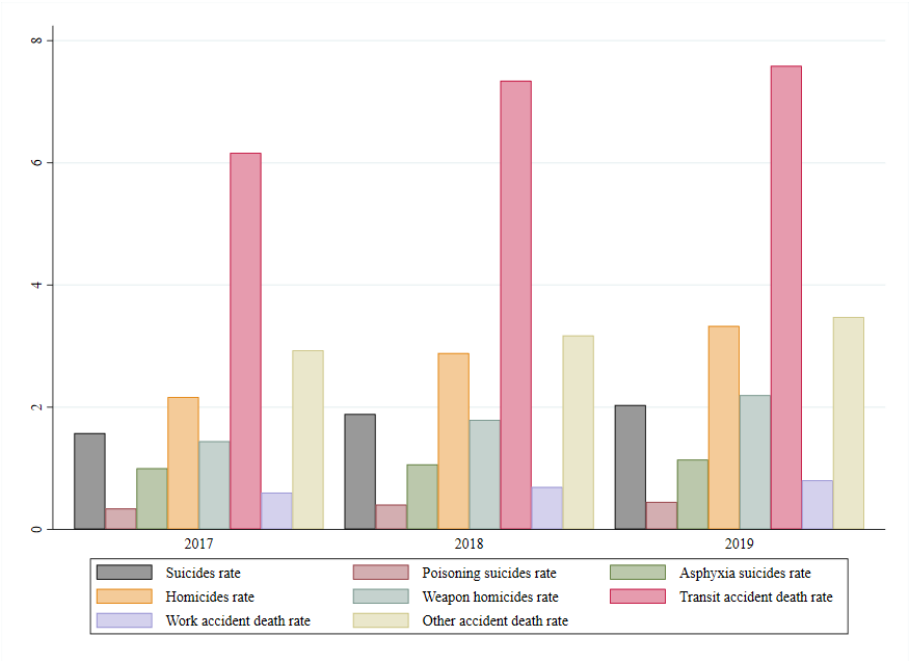


Figure 2: Type of violent deaths rates (per 100,000 population each year) - Peru

On the other hand, in figure 3 we see that 32.9% of the locality/districts suffered at least one suicide in 2019. If we see all type of deaths we have 17.7% of the locality/districts suffered at least one suicide in 2019.

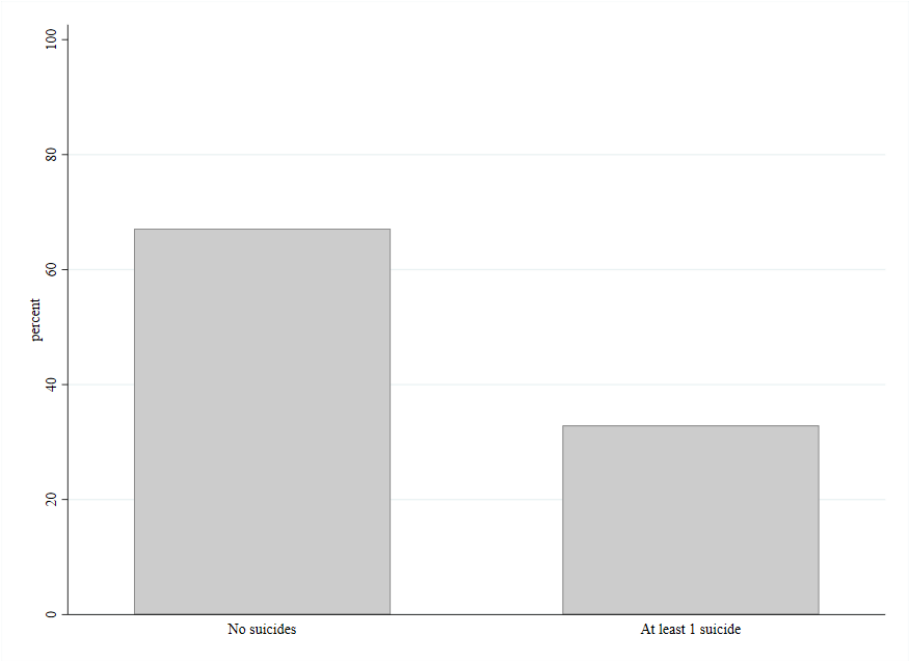


Figure 3: Localities with at least one suicide of total of violent deaths in 2019

# 4 Empirical strategy

We discard to use linear methods because, *a priori*, we think crime and others variables are related with suicides in no linear way. Furthermore, recent studies use several methods different to Random Forest. Because of this, we will use Random Forest and will analyze several metrics of prediction like area under the ROC curve, accuracy, precision and mean square error (MSE). Other type of methods based on trees have issues:

- Trees: in general are not robust against little changes in parameters.
- Bagging: trees are correlated.
- Boosting: is better than before method based on trees but does not adjust the weights of explanatory variables.

For this reason we think Random Forest could be better than other methods based on trees. We want to take advantage of the capture matter explanatory variables that offer Random Forest method. The hyperparameters would be chosen by k-fold Cross Validation, with a $k$ to be defined but in principle it would be between 5 and 10 as recommended by the literature. Additionally, we can use the individual features of suicides to capture heterogeneous prediction related to gender, age, marital status or education level.

# 5 Final remarks

Harnessing the potential of cutting-edge Big Data models, our project proposes an innovative approach to address suicides and their profound repercussions on affected families.

Our primary objective is to explore predictive insights into suicides by integrating variables related to crime with traditional demographic features such as age, education, household size, among others. Our analysis, centered on Peru, draws from an extensive array of datasets including the National Death System (SINADEF), National Register of Crimes and Offense Reports (RNDDF), Demographic and Family Health Survey (ENDES), and National Household Survey (ENAHO). This comprehensive dataset amalgamation enables a holistic evaluation and predictive analysis of suicide incidents.

The inherent challenge of linking datasets due to the absence of specific identification necessitates an inventive approach. Our solution involves predicting suicides at the localities/district level by employing binary indicators based on the presence or absence of suicide occurrences within these areas. This approach acknowledges the disparate occurrences of suicides across localities/districts, enabling more precise and targeted prevention strategies.

In our empirical strategy, we opt for the Random Forest algorithm as our primary machine learning method. This choice is motivated by the anticipated non-linear relationship between crime-related variables and suicides. Random Forest's strength relies in capturing complex non-linear relationships, along with its robustness in comparison to other tree-based methods, which positions it as our preferred technique.

By harnessing these datasets and leveraging the Random Forest algorithm, we aim to study the intricate intrelationship erplay between crime-related variables and suicidal tendencies, our main goal being identifying pivotal predictors that deepen our comprehension of suicide occurrences and thereby facilitating the formulation of more precise and effective prevention strategies; providing nuanced insights essential for effective tailored intervention health policies.

In conclusion, this project endeavors to exploit the potential of comprehensive datasets and advanced machine learning methodologies to broaden our understanding of suicide occurrences in Peru. This approach will lay the groundwork for the implementation of more effective suicide prevention initiatives and informed decision-making in shaping public health policy.

It's important to note a limitation: while our study cannot draw conclusions at the individual level due to constraints, our findings can offer valuable insights at the locality level, contributing significantly to the knowledge base.

# References

Becker, S. and L. Woessmann (2018): "Social Cohesion, Religious Beliefs, and the Effect of Protestantism on Suicide," *The Review of Economics and Statistics*, 100, 377–391.

Fortaner-Uyà, L., C. Monopoli, F. Calesella, F. Colombo, B. Bravi, E. Maggioni, E. Tassi, S. Poletti, I. Bollettini, F. Benedetti, and et al. (2023): "Predicting Suicide Attempts among Major Depressive Disorder Patients with Structural Neuroimaging: A Machine Learning Approach," *European Psychiatry*, 66, S1111–S1112.

Ha, J. and H.-S. Yang (2021): "The Werther effect of celebrity suicides: Evidence from South Korea," *PLoS ONE*, 16, e0249896.

Hamermesh, D. S. and N. M. Soss (1974): "An Economic Theory of Suicide," 82, 83–98, publisher: The University of Chicago Press.

Koo, J. and W. M. Cox (2008): "An Economic Interpretation of Suicide Cycles in Japan," 26, 162–174, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1465-7287.2007.00042.x.

Lin, G. M., M. Nagamine, S. N. Yang, Y. M. Tai, C. Lin, and H. Sato (2020): "Machine Learning Based Suicide Ideation Prediction for Military Personnel," *IEEE Journal of Biomedical and Health Informatics*, 24, 1907–1916.

Patton, G. C., C. Coffey, S. M. Sawyer, R. M. Viner, D. M. Haller, K. Bose, T. Vos, J. Ferguson, and C. D. Mathers (2009): "Global Patterns of Mortality in Young People: A Systematic Analysis of Population Health Data," *The Lancet*, 374, 881–892.

Tate, A. E., R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola (2020): "Predicting mental health problems in adolescence using machine learning techniques," *PloS one*, 15, e0230389–e0230389.

Turk, B. (2023): "Predicting suicidal thoughts in a non-clinical sample using machine learning methods," *Düşünen adam (Bakırköy Ruh ve Sinir Hastalıkları Hastanesi)*, 36, 179–188.

Yarborough, B. J. H., S. P. Stumbo, A. G. Rosales, B. K. Ahmedani, J. M. Boggs, Y. G. Daida, S. Negriff, R. C. Rossom, G. Simon, and N. A. Perrin (2022): "Opioid-related variables did not improve suicide risk prediction models in samples with mental health diagnoses," *Journal of Affective Disorders Reports*, 8, 100346.

Émile Durkheim (2007): *Le suicide*, PUF, f. alcan, paris ed.