# Analyzing Global Suicide Rates Using Random Forest Regression within the CRISP-DM Framework

Jayasurya Murali

October 13,2024

**Abstract**

Suicide is a critical public health issue affecting individuals, families, and societies worldwide. Understanding the underlying factors contributing to suicide rates is essential for developing effective prevention strategies. This study employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to analyze a comprehensive dataset on global suicide rates from 1985 to 2016. By leveraging Random Forest Regression and data visualization techniques, the research aims to uncover patterns, trends, and key influencing factors of suicide rates across different countries and demographic groups. The findings provide valuable insights that could inform policymakers and health organizations in implementing targeted interventions to reduce suicide rates globally.

## 1 Introduction

### 1.1 Background

Suicide is recognized as a major global health concern by the World Health Organization (WHO). It ranks among the top 20 leading causes of death worldwide for all ages, accounting for over 800,000 deaths annually [1]. The complexity of factors leading to suicide includes mental health disorders, socioeconomic status, cultural influences, and more. Despite significant efforts to address mental health issues, suicide rates continue to rise in several countries, highlighting the need for data-driven approaches to understand and mitigate this phenomenon.

### 1.2 Purpose of the Study

The aim of this study is to systematically analyze global suicide data spanning over three decades, employing Random Forest Regression within the CRISP-DM framework to develop predictive models. The study seeks to:

- Identify global trends and patterns in suicide rates. - Examine the influence of demographic factors such as age, gender, and generation. - Investigate the relationship between economic indicators, such as GDP per capita, and suicide rates. - Develop a predictive model to forecast suicide rates based on significant variables. - Provide actionable insights and recommendations for policymakers and health organizations.

### 1.3 Significance of the Study

Understanding the factors contributing to suicide is critical for developing effective prevention strategies. The findings of this study can help shape public health policies, inform resource allocation, and guide intervention programs aimed at reducing suicide rates. Moreover, the application of the CRISP-DM methodology offers a structured approach to data mining within public health research.

## 2 Methodology

### 2.1 CRISP-DM Framework

The Cross-Industry Standard Process for Data Mining (CRISP-DM) provides a structured approach, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [3].

### 2.2 Business Understanding

The primary objective of this study is to assist health organizations and policymakers in identifying key factors contributing to global suicide rates and in creating data-driven strategies for prevention. By applying predictive modeling techniques, the research provides valuable insights into demographic and socio-economic influences on suicide rates. This analysis is particularly relevant for stakeholders tasked with allocating mental health resources, developing outreach programs, and formulating policies that address at-risk populations. The business goal is to build a predictive model that accurately forecasts suicide rates across different countries and demographics, enabling more focused and effective intervention measures.

## 2.3 Data Understanding

The dataset used is publicly available on Kaggle, titled "Suicide Rates Overview 1985 to 2016" [2], and contains over 27,000 records from 101 countries. It combines demographic information with socio-economic indicators.

### 2.3.1 Exploratory Data Analysis

An initial exploration of the data revealed notable missing values in the 'HDI for year' variable and significant variability in suicide rates across countries and years. Correlation analysis suggested relationships between suicide rates and factors such as GDP per capita.

## 2.4 Data Preparation

The data underwent several preprocessing steps, including handling missing values by omitting records with missing fields. Categorical variables such as 'age' and 'generation' were transformed using one-hot encoding to make them suitable for modeling. The dataset was divided into training and testing sets with an 80-20 split.

# 3 Modeling

## 3.1 Random Forest Regression

A Random Forest Regressor was selected due to its ability to capture non-linear relationships. The model was trained using 100 estimators, and predictions were made on the test data. The Random Forest model was trained using the following process:

# 4 Evaluation

The Random Forest Regression model was evaluated based on the Mean Squared Error (MSE) and R-squared score. The model achieved a Mean Squared Error of 188.0641 and an R-squared score of 0.3904, indicating that the model explains approximately 39 percent of the variance in the suicide rates. While this represents a moderate level of predictive power, it also highlights the complexity of the factors influencing suicide rates.

# 5 Deployment

The final phase of the CRISP-DM process is deployment, which involves making the results actionable for stakeholders. In this study, the deployment phase would focus on integrating the model into health policy frameworks, where policymakers and public health organizations can utilize the predictions for early intervention programs. Additionally, a user-friendly dashboard could be developed to allow for real-time monitoring and forecasting of suicide rates across different regions and demographic groups. The model's predictive capabilities can be leveraged to prioritize resource allocation and tailor intervention programs to specific at-risk groups. Collaboration with governments, NGOs, and mental health organizations would ensure that the insights gained are translated into meaningful action.

# 6 Results and Discussion

The Random Forest model demonstrated that GDP per capita, population size, and age groups were among the most influential factors affecting suicide rates. The analysis revealed that wealthier countries generally exhibited lower suicide rates, but this pattern varied significantly across regions. Age and gender also played crucial roles, with certain demographic groups being more vulnerable.

The line plot depicting suicide rates over time revealed that older generations, such as the Silent Generation and Baby Boomers, had higher suicide rates in earlier years, while younger generations, including Millennials and Generation Z, have shown increasing trends in recent years.

## 7 Conclusion

This study successfully applied the CRISP-DM methodology to analyze global suicide rates using Random Forest Regression. The model highlighted the importance of economic indicators, demographic factors, and generational trends in influencing suicide rates. The insights gained from this study can aid policymakers in developing targeted interventions, focusing on at-risk groups, and mitigating the factors contributing to rising suicide rates. Future research could enhance the analysis by incorporating additional variables such as unemployment rates and mental health data, as well as exploring advanced modeling techniques.

## References

[1] World Health Organization. (2018). *Suicide data*. Retrieved from `https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/`

[2] Russell, Y. (2017). *Suicide Rates Overview 1985 to 2016*. Kaggle Dataset. Retrieved from `https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016`

[3] Shearer, C. (2000). The CRISP-DM model. *Journal of Data Warehousing*, 5(4), 13-22.