Suicide Rate Analysis and Prediction

BY

CHANDRADEEP REDDY KASARALA

ANJI JESHWATH NAIK

Introduction:

To forecast and stop suicide attempts, this research uses machine learning to examine suicide rates and the factors that influence them more accurately. Data on suicide rates and their relationships to several possible contributing factors, such as depression and environmental conditions, will be analyzed first. Any patterns or trends can be better understood with the use of visualizations. Next, in an effort to identify the machine learning model with the lowest error rate, a number of them will be developed and tested in order to determine which one can most effectively forecast suicide rates based on the many influencing factors. In the end, the study investigates how machine learning might be used to predict suicide rates and pinpoint the primary causes behind them.

IMPORTANCE OF CURRENT STUDY :

- As the world's greatest cause of mortality, suicide is a serious global health concern. As the 18th leading cause of death worldwide, suicide claims the lives of an estimated 800,000 individuals annually, according to World Health Organization estimates. As a contributing factor to a decline in average life expectancy, suicide ranked as the tenth most common cause of death in the US in 2018. Every adult suicide death has resulted in over 20 attempted suicides, and suicide affects people of all ages. Suicide is a problem that affects people all around the world; in 2016, 79% of suicides were in low-income nations.

- Many internal and environmental variables are involved in today's epidemic of severe physical and mental diseases. Although depression is more common in people in their 30s and 40s, stress from school and interpersonal interactions can equally impact youngsters and the elderly. Because mental illness is frequently stigmatized in society, many people choose to hide their sickness. The state of the economy, drug and alcohol abuse, and self-harm are all factors that can impact suicidal thoughts and attempts.

- As a means of precisely identifying people who are at risk of suicide within a specific timeframe and facilitating the implementation of preventative measures, one of the first steps in suicide prevention can be looked at as a categorizing task. Still, there hasn't been much progress in suicide prediction over time, according to a large meta-analysis that examined 365 papers and showed that forecasts based on individual risk or protective factors had poor predictive accuracy.

EXPECTATIONS FROM CURRENT STUDY:

- Data analysis is performed to classify data for future prevention. This review identifies reasons of suicide in a given state and year as well as tracks shifts in suicide rates associated with specific causes. It also enables us not only to come up with strategies but also allows for comparison whether there has been any change in suicide rate over time by analyzing this information and putting them into different groups that can be

prevented or controlled against. Such findings may direct policy makers towards different approaches of dealing with the problem. The study involved comprehensive global suicide data; therefore, subsequent parts were arranged systematically to enhance understanding among readers.

- Being able to identify those who are prone to suicide at an early stage is very important if prevention measures are to work. Artificial intelligence has shown itself as a possible option for this purpose. What we want to do with our study is create a model using artificial intelligence which can predict suicide attempts. In order to achieve this, we employ various models like K Nearest Neighbor, Decision Trees and Random Forest Regression then compare their performance under such conditions.

OUTCOMES INTENDED TO ACHIEVE:

- This project attempts to investigate how different algorithms of machine learning can be used for suicide rate prediction. Various models of machine learning will be employed on the basis of relevant factors and data collected from a dataset so as to forecast and estimate numbers of suicide rates. The analysis aims at showing the potentiality and efficiency in making correct predictions about suicide rates by these algorithms with respect to available data features.
- Furthermore, governments and organizations involved in prevention programs or counseling services may find useful knowledge through the results gained by this study. Therefore, it is possible to identify areas that need improvement or more attention based on machine-learning-model performances and outcomes.While implementing these findings into practice, such entities should take effective strategies towards addressing critical issues surrounding suicide prevention as well as providing appropriate counselling support.

METHODS AND TECHNIQUES

Data Acquisition:

The dataset used in this study is Suicide Rates Overview 1985 - 2016 which was obtained from Kaggle. It has a total of 27820 rows and 12 columns. This dataset was made by combining data from four different sources namely United Nations Development Program (HDI), World Bank, World Health Organization (WHO) and Szmali to determine if there are any factors that may correlate with suicide rate worldwide. The file is in CSV format which means it can be easily downloaded and accessed.

Cleaning and Normalization:

Once the dataset is acquired, null or redundant rows will be removed, repeated columns will be dropped as well as carrying out an outlier analysis and treatment. SkLearn's Label Encoder will

then be used to convert country, year, sex, age and generation which are non-numerical labeled columns into numerical labels. Many machine learning estimators require dataset standardization i.e., if individual features do not appear like standard normally distributed data it might cause them to perform poorly hence SkLearn's RobustScaler is applied in normalizing numerical columns such as population gdp_per_capita ($).

Exploratory Data Analysis:

At this point, many techniques for data mining are employed to expose concealed patterns in the dataset and establish relationships between variables; also, several graphs are plotted to reveal tendencies in suicide rates and determine numerous causes that contribute to such deaths. NumPy, seaborn, matplotlib etc., which are Python libraries have been used. Different figures have been presented to help grasp the information more effectively. Heatmaps and scatter plots help indicate correlations.

Some Key Findings in EDA:

- Suicide rate and Suicide count are higher in male compared to female.
- Lithuania is the country with the highest suicide rate for male and Sri Lanka is the country with highest suicide rate for Female.
- Lithuania is the country with the highest suicide rate.

Machine Learning Models:

Now we shall train many machine learning models on our dataset with validation technique for checking overall fit as well. At the end of it all the best model will be shown for suicide prediction. K Nearest Neighbor, Decision Tree Regression, Random Forest Regression and XGBoost Regression were implemented and evaluated based on accuracies and RMSE scores.

IMPLEMENTATION

Dataset:

The dataset has been taken from Kaggle. It has 27,820 rows with 12 columns. Some of the columns are numerical types which include GDP per capita, HDI for year, suicides_no, while others like country, age, sex, generation etc. are categorical. It includes data from over 100 countries from 1985 to 2016.

Columns:

No.	Variable	Type
1	country	object
2	year	int64
3	sex	object
4	age	object
5	suicides_no	int64
6	population	int64
7	suicides/100k pop	float64
8	country-year	object
9	HDI for year	float64
10	gdp_for_year ($)	object
11	gdp_per_capita ($)	int64
12	generation	object

Evaluation Metrics:

Two of the evaluation metrics were employed namely accuracy and RMSE scores. Accuracy is a common measure used in classification problems; it is the number of correct predictions made divided by all predictions made. One of the most popular ways to determine precision for continuous data is through RMSE. This is because large errors are more weighted in RMSE than MAE hence it should be preferred when large errors are not wanted. The best model could be XGBoost Regression as it has highest accuracy and lowest RMSE.

Model Training and Evaluation

Our Machine Learning algorithms are K Nearest Neighbor, Decision Tree, Random Forest, XGBoost. Below are the accuracies and RMSE's of each model.

K Nearest Neighbor Regression : Accuracy - 0.771, RMSE- 0.279

Decision Tree Regression : Accuracy - 0.967, RMSE- 0.105

Random Forest Regression - Accuracy - 0.988, RMSE- 0.063

XG Boost Regression - Accuracy - 0.997, RMSE- 0.029

RESULTS DISCUSSION:

Data analysis helped us understand several underlying trends in suicide attempts over the years 1985 and 2016. – Now about the performance of four machine learning models - Out of all trained models, XGBoost had highest accuracy and lowest RMSE because its execution Speed & model performance is very good among other things. Random forest came with an accuracy rate at 98.3 percent followed by Decision Tree while K-Nearest Neighbors had 79.1% accuracy.

Conclusion:

The objective of this study was to demonstrate how different machine learning algorithms can predict suicide rates using a dataset that contains relevant variables. Although many models have achieved high levels of accuracy, there is still room for improvement. Initial investigations uncovered some interesting findings such as that teenage boys are more prone to suicide than girls. XGBoost and Random Forest Regression consistently outperformed other models in terms of accuracy and precision. Additionally, this analysis will identify areas for government intervention as well as inform other organizations involved with suicide prevention about the most effective strategies they should adopt.

FUTURE WORK :

• To improve this project further, it would be possible to bring together more than one dataset on suicides so that deeper evaluations can be made.

• Several statistical tests (e.g., hypothesis testing) could also be conducted which might yield useful insights.

• Sentiment Analysis could be employed to determine where individuals feel most comfortable discussing their mental health issues via social media.

REFERENCES

1) Boudreaux, E. D., Rundensteiner, E., Liu, F., Wang, B., Larkin, C., Agu, E., Ghosh, S., Semeter, J., Simon, G., & Davis-Martin, R. E. (2021). Applying Machine Learning Approaches to Suicide Prediction Using Healthcare Data: Overview and Future Directions. Frontiers in psychiatry, 12, 707916. https://doi.org/10.3389/fpsyt.2021.707916

2) Gen-Min Lin, Masanori Nagamine, Szu-Nian Yang, Yueh-Ming Tai, Chin Lin, Hiroshi Sato, “Machine Learning Based Suicide Ideation Prediction for Military Personnel”, IEEE Journal of Biomedical and Health Informatics, vol. 24, issue: 7, July 2020

3) Hasmitha Bhutham, “Suicide rates analysis and prediction” , December 2020.

4) Mrs. B. Ida Seraphim , Subroto Das , Apoorv Ranjan, 2021, A Machine Learning Approach to Analyze and Predict Suicide Attempts, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 04 (April 2021).