**PREDICTING HIGH-RISK GROUPS FOR SUICIDES BASED ON SOCIOECONOMIC INDICATORS**

**BSAN 775: Introduction to Business Analytics**

Dr. Justin Keeler

**TEAM**

Murali Krishna, Katragadda

Sesha Tarun Kuchibhotla

Balu, Vojja

Dhanush naga chowdary, Garikipati

Vyshnavi, Danaboyina

# INTRODUCTION

Suicide continues to top the list of serious challenges to the public health sector across the world, costing a nearly 800,000 cut-off of lives annually, as the records available at the World Health Organization (WHO) indicate. Suicides affect not only the one committing the act but also the families and communities. In this regard, therefore, it is considered one of the vital areas of well-being in society. This essay, therefore, aims at discussing in detail the multivariate factors of suicide while reflecting on the need for accessible mental health care and supportive community frameworks with reference to how data analysis is important for crafting efficient prevention strategies.

Understanding suicide is needed more comprehensively, bearing in mind both the socio-cultural and socio-economic contexts in which people exist. The attitude of society toward mental health has enormous influences not only on committing suicides but also on preventing it from happening. Most cultures continue to attach some stigma to mental illnesses, which make the affected people shy away from seeking the necessary help. It would be very critical in coming up with cultural interventions that are sensitive and designed according to what a certain community deems right and important to be able to overcome such challenges. For example, community-based programs aimed at increasing awareness and understanding about mental health can reduce stigma and encourage individuals to seek help.

The difference in demographic variability of suicide rates really points toward some high-risk specific groups for whom particular interventions are needed. Recent statistical data has demonstrated the presence of higher rates of suicide among young people, males, and some ethnic and racial minorities. That variability could be in a certain way associated with social isolation, discrimination, economic instability, and a lack of proper access to mental healthcare. Effective suicide prevention should address these underlying issues by promoting inclusivity, stabilizing the economy, and universality in all provisions of healthcare services. Economic factors are particularly pronounced in regard to suicide at the crossroads with mental health.

What is suggested is a strong association between economic downturns and an increase in suicide rates. Unemployment and income status, many factors, even inflation, could work to worsen despairing and hopeless feelings, especially among people who are at most risk. Enhancing employment and the population ratio, therefore, help in economic recovery and are preventive actions from suicide by giving people a sense of purpose and financial stability. While this is an important risk factor, mental illness typically interfaces with other psychosocial environmental factors in contributing to the problem of suicidal behavior. Conditions such as depression and anxiety can be exacerbated by the personal predicament of relationship problems, chronic health problems, or substance abuse. The approach to mental illness in the suicide prevention strategies becomes very crucial and has to include professional health care, community support, and further education on better literacy of mental health.

Data analysis brings to light the significance of the role in understanding, and, above all, prevent suicide. Research into patterns and trends of suicide data offers great potential for intervention and identification of risk factors. The advanced analytics technique, like Predictive Analytics, helps in figuring out high-risk groups with reference to the variables like gender, GDP, and socio-economic indicators of population density. Such information is very invaluable in designing targeted interventions that may only focus on the specific risk factors within a population.

In short, suicide prevention needs a collective effort through financial support in the delivery of easily available mental health care, community partnerships, and the use of strategies derived from data. Is a complex, many-sided problem that requires a comprehensive and compassionate solution. The growth in the number of suicides is an indicator: it's high time for society to start building a wide network, the branches of which will come into joint action aimed at bringing suicide mortality down, from the government and health provider to community leaders and laypeople. Through such an approach, only can we hope to save lives and foster a more compassionate and supportive society.

# PROBLEM STATEMENT

A typical Regression Model uses historical data to predict the future and to analyze the trends that influence the target variable. This problem statement aims to predict Suicide rate of the person considering various factors. A lot of studies in the past considered only a few features such as gender, population and factors such as health, day to day habits such as alcohol consumption were not taken into consideration. This gives us the motivation to consider multiple factors.

Statistical tools such as SPSS and Excel can be used to identify the relation between the independent variables.

# OBJECTIVES:

* The present study would be identifying and analyzing those particular socioeconomic indicators having a strong relation with the rate of occurrence of suicide. It would also be the development of a predictive model to appropriately find the high-risk group of suicide. The project will aim to develop a model that predicts high-risk groups based on the following factors: CauseSpecificDeathPercentage, StdDeathRate, GDP, and population; GDP per capita, Inflation rate, EmploymentPopulationRatio, and other related data.
* We are using a variety of metrics, such as the significance, collinearity, R-squared, and Adjusted R-. Examine the relationship between variables that helps in reducing the rate of suicide.

**LITERATURE REVIEW**

The literature review indicates a very complex relationship of the socio-economic factor with the mental health outcomes and suicide mainly. The studies reveal that unemployment, poverty, education level, social inequality point out to the leading predictors of suicide rates.

• A 2019 study by Platt and Hawton attests to strong associations between unemployment and increased suicide risks, providing evidence that points to economic downturns and job losses as factors that exacerbate mental health issues. The study conducted by Fountoulakis et al. in 2020 mentioned financial stress and hopelessness as conditions for suicide among the poor and low-income population.

• Further, it was concluded that the level of education is a protective factor. Therefore, below-average education levels were associated with a low suicide risk (Milner et al., 2013). On the other hand, a study has shown a correlation between greater suicide rates and social inequality as demonstrated in the differences of wealth and a lack of social cohesiveness. It would, therefore, mean that suicide preventive measure must include social determinants of health (Rehkopf & Buka, 2006).

"However, there has been no literature developed regarding predictive models that combine several socio-economic factors in order to identify populations at a high risk of suicide. The paper uses a large dataset on suicide rates and characteristics of socioeconomics, trying, therefore, to fill this gap through the development of a predictive model that could guide focused interventions and measures in the regard."

# METHODOLOGY

## Data Collection:

We have collected our data from Kaggle website. The data is about various factors influencing Suicide rates. It has 10 Independent variables and 1 Dependent variable.

## Data Preprocessing:

Initially there were missing values in the dataset which was referred to. Missing values have been handled before statistical analysis using manual imputation based on the distribution of columns.

Figure (1) shows the Descriptive statistics of the data having the frequencies of the variables.

There are 4698 Valid records and 0 Missing values.

## Exploratory Data Analysis:

It is an approach used to analyze data to find the relationship between the variables in order to understand the main characteristics before modeling.

As part of our EDA, we have generated Histogram, Graphs and Scatter plots.

## Frequency curve:

The Frequency curve depicts the residuals.

Figure (2) shows the frequency curve and it lies between -5 & +5 which shows the residuals have no bias and do not systematically over or under-fit.

## The Normal P-P Plot of Regression:

The Normal P-P Plot of Regression tells us about the relation between the observed and expected cumulative probabilities.

In Figure (3) Expected Cumulative Probability and the Observed Cumulative Probability are closely aligned which tells that the residuals are normally distributed.

## Scatter Plot:

The scatter plot depicts the constant variance and there are no outliners involved in the data.

# RESULTS

We have used Multiple Linear Regression on our data to make the predictions. As, there are several independent variables that make contribution in predicting the Dependent/response variable.

## Multiple Linear Regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory /Independent variables to predict the outcome of a

response/Dependent variable. Multiple regression is an extension of linear regression that uses just one explanatory/Independent variable.

## Multiple Linear Regression on our Data:

For our analysis we have used multiple regression model, as out dependent variable is continuous.

## Interpretation Of Regression Analysis Results:

### Model Summary:

In the Figure (5), we can see the R value is .990 which indicates a strong correlation between the predicted and actual values.

1. The R square value is .981 which indicates the proportion variance in the dependent variable by the independent variables. It has approximately 98.1% variability.
2. The Adjusted R Square value is .981 which is not lagging behind R Square value, which tells the independent variables are contributing properly in predicting the

dependent variable.

***Anova:***

1. From Figure (6) we can see the F value is high which tells the overall regression model explains the significance in predicting the target variable
2. A p value<.005 suggests that model is statistically significant, in our case it is <.001b which show good significance.
3. The df (Degree of Freedom) shows the number of predictors.
4. The above Anova table confirms the statistically significance of the regression model.

**Components of Coefficient table (Figure (4)):**

### Significant Predictors:

We can see in Figure (7) most of the variables has p<0.005 which are significantly associated with the target variable.

### Collinearity:

* + Most of the tolerance values are closer to 1 and VIF values are below 5 except for few which shows minimal collinearity.
  + The entire dataset was divided into 80% Training data and 20% Test data, regression was applied on the training data and predicted value is calculated using the compute variable option in the Transform panel. Which gave the expected value.

### Computing the variable:

Predicted=-28.097+ 0.015 \* Year+7.782E-5 \* SuicideCount+ -0.408 \*

CauseSpecificDeathPercentage+1.125 \* StdDeathRate+ -3.793E-9 \* Population+

1.306E-5\*GDPPerCapita+2.278E-5 \* GNIPerCapita+0.000 \* InflationRate+ -0.058 \*

EmploymentPopulationRatio+0.195 \* Gender

### Correlation between the Training and Test data:

The relationship between training and test data is an important gauge to judge model performance in any machine learning and statistical modeling. This relation, however, makes them through Pearson's correlation coefficient and it quantifies how much a model can generalize from training data to unseen test data. Understanding this relationship helps in diagnosing model behavior, for example, overfitting or underfitting, and guides necessary corrections for bettering the robustness of the model.

Overfitting occurs when a model is overly complex and learns both the actual underlying patterns in the training data and the noise peculiar to it. As a result, while it performs at the training dataset, it works very well, and the representation for the new unseen data will typically be the test dataset. On the other end, underfitting occurs when the model is too simple to learn from the true underlying pattern of the data; it performs dismally on the training data and too poorly at the test data.

In the above scenario, the correlation coefficient of the Pearson Correlation for training data is 0.991, while the same for test data is 0.989. These are very high correlation values, representing a strong linear relationship between the predicted values and the actual values, meaning the model learned quite well the patterns from the training data. The slightly lower value of correlation in the test data, however, insinuates that the model is not performing equally well over the test data and somewhat hints at overfitting.

This points out that small differences between both the correlation values (0.991 on training and 0.989 on test data) are due to overfitting only. This small overfit might have been due to the number of predictors that the model used in relation to the available data number. A model with a lot of predictors has a good chance of probably overfitting a lot of details and noise from the training dataset, hence overfitting the training data but with less ability to generalize.

It should be noted that the available dataset in this Kaggle has a limitation; data scarcity may worsen the overfitting challenges, most importantly if the model is complex with many predictors. In case of a small quantity of data, each datum is given more influence to shape the model; hence, predictions from the model are more likely to reflect the specific tendencies of training data rather than general trends that could apply to broader, unseen data.

It is possible to ameliorate such overfitting by different means. One common way is to reduce the complexity of the model. This may involve a smaller number of predictor variables, obtained either through a careful choice of only the most important or by making use of techniques for dimensionality reduction, such as principal component analysis (PCA). Second, where possible, one may also take an approach to increasing the number of data points or adopting cross-validation techniques, which may give a more robust estimate of model performance over different subsets of the data.

Regularization techniques, such as lasso or ridge regression, also help in constraining the overfitting problem. They modify the learning algorithm in such a way that the model gets penalized for being of excessive complexity, hence bringing in a trade-off between bias and variance, so as to encourage a model generalizing better to unseen data.

In conclusion, although the overfitting in this described setting is very minor, this is something to be viewed cautiously. Due care should be applied to ensure there is a balance between the complexity of the model and adequacy of data so that it may continue to be predictive and stand robustly across different datasets. Some suggestions that help increase the model's performance and reliability involve some form of dimensionality reduction through either feature selection or feature transformation, increasing the size of your dataset, cross-validation, or regularization.

**DISCUSSION**

## FUTURE SCOPE OF USE CASES:

**Early interventions and Prevention:**

Appropriate estimation of suicide rates will enable the authorities concerned and the associated mental health organizations to enforce targeted interventions and methods of prevention in high-risk populations or regions. This may include the allocation of funds to community outreach programs, mental health services, suicide hotlines, and education programs to raise awareness about seeking help and busting stigma.

**Policy development and resource allocation:**

This means that derived insight from predictive modeling can inform an evidence-based policy outlining the threshold level of social welfare and generally mental health treatment and suicide prevention. This information might go a long way to enable the Government and its Legislators to streamline funding priorities in mental health services, facilitate proper allocation of resources, and to tailor-make the interventions in a targeted approach towards the said geographical areas or demographical groupings.

**Vulnerable groups include children and the elderly**

The regression analysis helps to find a set of environmental, social, and demographic characteristics that are related to the increased suicide risks. This is essential in the effort to determine who is vulnerable among teenagers, who are war veterans, or the residents of some socio-economic challenged locations. Lastly, the nature of the targeted interventions allows addressing the specific needs of these groups. These use cases have implications in terms of where an accurate estimate of suicide rates might play out for public policy, health, economic impact, or another scope of society—the reminder of such importance that issues of data security, privacy, and ethics need to be undertaken successfully for the use case implementation.

# CONCLUSION

Key economic aspects like employment, Gross Domestic Product (GDP), inflation, and income of the population are put on the front line for any given nation that takes their suicide rate seriously and wants to intervene to cut it down. Of all the above factors, employment and GDP of the country take the front line in affecting this societal behavior. A strong sector in employment does not avail financial stability but also ensures that people belonging have purpose and sense to belong, which contributes to mental soundness. Equally, a thriving economy, most of the time through effectively grown GDP, tends to enhance the quality of life for its citizens, hence reducing despair and hopelessness related to suicides.

Interestingly, although it could be the immediate cause due to the purchasing power and economic stability of a country, found no strong relationship that showed a significant increase in suicide rates due to inflation. It seems that the rising price of inflation doesn't push up suicides, either because people are tending to adapt to it, or government interventions are cushioning its effect.

On the other hand, the income from the population is critical in this dynamism. To be precise, it can be said that a perfect negative correlation of -1 is found between the income of the population and the suicidal ratio, as per the data. In other words, suicides are expected to be at a lower rate with an increase in income. This relationship puts weight on the importance of financial security and how it impacts mental health. High income is likely to assure good access to health care services, including mental health support, and also gives a cushion against the stress emanating from financial uncertainties.

To sum up, even though inflation seems to affect suicide rates almost insignificantly, other issues like employment, GDP, and population income are of great importance. Governments should therefore lay more emphasis on those economic policies that boost employment opportunities and raise income levels, rather than only being concerned with the inflation rates, as income levels have a far more direct association with suicides.

# REFERENCES:

1. Blakely, T. A., Collings, S. C., & Atkinson, J. (2003). Unemployment and suicide. Evidence for a causal association Journal of Epidemiology & Community Health, 57(8), 594-600.
2. Gunnell, D., & Chang, S. S. (2016). Suicide prevention: A global imperative. Lancet Psychiatry, 3(10), 899-900.
3. Kawachi, I., & Berkman, L. F. (2001). Social ties and mental health. Journal of Urban Health, 78(3), 458-467.
4. Lorant, V., Deliège, D., Eaton, W., Robert, A., Philippot, P., & Ansseau, M. (2005). Socioeconomic inequalities in depression: A meta-analysis. American Journal of Epidemiology, 161(2), 98-112.
5. Nock, M. K., Borges, G., Bromet, E. J., Cha, C. B., Kessler, R. C., & Lee, S. (2008). Suicide and suicidal behavior. Epidemiologic Reviews, 30(1), 133-154.
6. Platt, S., & Hawton, K. (2000). Suicidal behaviour and the labour market. In K. Hawton & K. van Heeringen (Eds.), The International Handbook of Suicide and Attempted Suicide (pp. 309-384). John Wiley & Sons.

# APPENDIX

## Figure(1):

**A screenshot of a graph

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Figure(2): Frequency Curve**

**A graph of a normal distribution

Description automatically generated**

Figure(3): **The Normal P-P Plot of Regression:**

**A graph with a line drawn on it

Description automatically generated**

**Figure(4):**

**Scatter Plot:**

**A graph showing a number of blue dots

Description automatically generated**

**Figure(5): Model Summary**

**A screenshot of a report

Description automatically generated**

**Figure(6): Anova Table**

**8A screenshot of a calculator

Description automatically generated**

## Figure(7): Coefficients Table:

**A screenshot of a table

Description automatically generated**

**Figure(8):**

The Correlations between the Test data and Training data

A screenshot of a computer

Description automatically generated

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* THE END \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*