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

**BSAN 775: Introduction to Business Analytics**

Dr. Justin Keeler

**TEAM**

Name1

Name2

-

-

# INTRODUCTION

Suicide is a very worrying problem that has irreversible effects on individuals, families and society. Addressing the causes of suicide, improving access to mental health care, and fostering supportive communities are essential steps to saving lives. According to the World Health Organization (WHO), nearly 800,000 people die by suicide each year, and for every death there are many more suicide attempts. Suicide rates vary widely by region, but are generally higher among certain demographic groups, including young people, men, and certain ethnic and racial minorities.

Analyzing patterns and relationships within the data can provide insight into the key variables that influence suicide rates, thereby focusing interventions and changes in approach.

Culturally sensitive interventions are needed to address these complex issues. Understanding the factors that contribute to suicide is key to prevention. Mental illnesses such as depression and anxiety contribute significantly to this. However, it is often accompanied by other factors such as, Gender ,GDP,population,Inflation rate, Employment Population ratio.

Data analysis plays an important role in suicide prevention. By examining trends and relationships, researchers can identify high-risk groups and the most effective interventions. For example, research shows that better employment ratio can significantly reduce suicide rates.

In summary, suicide is a complex and multifaceted problem that requires a comprehensive prevention approach. Understanding risk factors and patterns is important to developing effective interventions. By working together, we can save lives and build a more caring and understanding 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 primary objective of this analysis is to identify and analyze socioeconomic indicators strongly correlated with suicide rates and develop a predictive model to accurately identify high-risk groups for suicide,

* Our goal is to develop a model that can accurately predict the high-risk groups for suicide based on several factors such as CauseSpecificDeathPercentage, StdDeathRate,GDP, and population, GDP Percapita,Inflation rate, EmploymentPopulationRatio.
* We intend to use various metrics like significance, collinearity, R-squared and Adjusted R- squared.
* Evaluate the correlation among different attributes which helps to reduce suicide-rates.

**LITERATURE REVIEW**

* The literature review on predicting high-risk groups for suicide based on socioeconomic indicators highlights the complex relationship between socioeconomic factors and mental health outcomes, particularly suicide. Studies have identified unemployment, poverty, education level, and social inequality as significant predictors of suicide rates.
* Platt and Hawton's 2019 study reveals a strong link between unemployment and increased suicide risk, suggesting that economic downturns and job loss can worsen mental health issues. Fountoulakis et al.'s 2020 research emphasizes the impact of poverty and low socioeconomic status on suicide, highlighting financial strain and despair.
* Furthermore, education level has been found to be a protective factor, lower suicide rates are linked to higher educational attainment (Milner et al., 2013). However, studies have shown a correlation between greater suicide rates and social inequality, as shown by differences in wealth and a lack of social cohesiveness. This suggests that attempts to prevent suicide must address social determinants of health. (Rehkopf & Buka, 2006).
* Despite these results, there is still a lack of information in the literature about the creation of predictive models that combine several socioeconomic factors to pinpoint high-risk populations for suicide. By examining a large dataset of suicide rates and socioeconomic characteristics, this study seeks to close this gap by creating a prediction model that may guide focused interventions and preventative measures.

# METHODOLOGY

## Data Collection:

We have collected our data from Kaggle website. The data is about various factors influencing Suiscide 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 Correlation between the Training Data and Test Data tells us if the model is fitting or not. In general,
  + If the Pearson Correlation value of the Test Data is less than that of Training Data, it is an overfit model which indicates that the model works well on the training data but not on

the test data.

* + Figure (8) shows the value of Test Data is .989 and for training data is .991. The model is slightly overfitting since we have many predicting variables and less data available on

Kaggle. But the difference is very little, which is acceptable.

**DISCUSSION**

## FUTURE SCOPE OF USE CASES:

### Early Intervention and Prevention:

Authorities and mental health organizations can put targeted interventions and prevention methods into place in high-risk populations or regions by properly estimating suicide rates. This can entail setting aside funds for community outreach initiatives, mental health services, suicide hotlines, and educational campaigns that encourage seeking help and lessen stigma.

### Policy Development and Resource Allocation:

Evidence-based policies pertaining to social welfare, mental health treatment, and suicide prevention can be developed with the use of insights obtained from predictive modeling. This data can be used by governments and legislators to prioritize funding for mental health services, distribute resources efficiently, and create interventions that are suited to particular geographic or demographic groupings.

### Identification of Vulnerable Populations:

Regression analysis can be used to find environmental, social, and demographic characteristics that are linked to a higher risk of suicide. This data is essential for identifying vulnerable groups, including but not limited to teenagers, veterans, and residents of socioeconomically challenged locations. The specific needs of these groups can then be addressed by developing targeted in

These usage cases show the wide-ranging effects that precise suicide rates estimates may have on public policy, healthcare, and economics, among other societal issues. Remember that careful consideration of data security, privacy, and ethics is necessary for the effective implementation of these use cases.

# CONCLUSION

* + To effectively reduce the suicide rates, nations should give priority to sectors like employment, GDP, Inflation, Population Income.
  + There is no significant association found between inflation and suicides.
  + Suicide rate is negatively correlated with population income, which says if the income is more then the suicide rates are less.

# 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 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*