Analysis of Global Suicide Rates and Prediction of Suicidal Demographics

Ananya Veeraraghavan
Computer Science Engineering
PES University
Bangalore, India
ananyav308@gmail.com

Sakshi Shetty
Computer Science Engineering
PES University
Bangalore, India
sakshivshetty@gmail.com

Snigdha S Chenjeri
Computer Science Engineering
PES University
Bangalore, India
snigdhachenjeri00@gmail.com

***Abstract*—Using exploratory data analysis and varied machine learning models, this paper details how our project aims to classify the causes for suicides worldwide, and which aspects of an individual's lifestyle (age, sex, economic situation of the country etc) have the biggest roles in leading to the tendency to commit suicide. With the findings explained in this paper, the analysis, prediction, and the understanding of which groups of people seem to gravitate towards suicidal tendencies will become clearer.**

***Keywords*—EDA, visualisations, suicides, suicide prediction, Logistic Regression, K-fold cross validation, Random Forest, Machine Learning, AI, Artificial Neural Network.**

I. Introduction

As the generations get more advanced—technologically, socially and scientifically—the frequency of cases of depression and other mental health issues has been on a consistent upwards slope. Of course, these aren't the only factors which cause suicides—socio-economic factors, financial circumstances and so on. For a variety of specific reasons, some countries have higher rates than others.

Research shows that approximately 90% of people who have died by suicide were suffering from a mental illness at the time. The most common mental illness reported was depression, impulsivity and substance use, including alcohol and drugs, also warning signs for elevated suicide risk. It is important to remember that suicidal thoughts and behaviors are not the natural consequence of serious life stresses. People who experience a stressful life event may feel intense sadness or loss, anxiety, anger, or hopelessness, and may occasionally have the thought that they would be better off dead. In most people, however, experiences of stressful life events do not trigger recurring thoughts of death, creation of a suicide plan, or intent to die. If any of these are present, it suggests that the person is suffering from depression or another psychiatric disorder and should seek professional treatment.

Suicidal thoughts and behaviors are an international public health problem contributing to 800,000 annual deaths and up to 25 million nonfatal suicide attempts. In the United States, suicide rates have increased steadily for two decades, reaching 47,000 per year and surpassing annual motor vehicle deaths. This trend has prompted government agencies, healthcare systems, and multinational corporations to invest in artificial intelligence-based suicide prediction algorithms. Suicide is one of the leading causes of death in young people. Close to 8,00,000 people die due to suicide every year, which is one person every 40 seconds. According to WHO, effective and evidence-based interventions can be implemented at population, sub-population and individual levels to prevent suicide and suicide attempts.

Broader social factors including socio-economic deprivation are identified as risk factors for suicide. Socio-economic deprivation is often discussed as a link to a more stressful lifestyle, less access to resources including health care, fewer opportunities for educational achievement (in young people), poor housing facilities and lower self-esteem. However, it is rarely discussed in terms of intervention or prevention. In short, while the social origins of depression and hopelessness may be acknowledged, individualistic notions of risk and psychopathology are frequently privileged, in terms of intervention and prevention. As a corollary to this, key risk factors are conceptualized clinically: depression and hopelessness as mental illness rather than reactions to life circumstances, and while the correlation to economic deprivation has been acknowledged there has been little attempt to analyse it. Socio-economic status had a similar magnitude of population attributable risk for suicide as mental disorders. Public health interventions to reduce suicide should incorporate socio-economic disadvantage in addition to mental illness as a potential target for intervention. More distal risk factors associated with suicide, such as measures of socio-economic status, have been shown to have substantially lower relative risk estimates associated with suicide in population-based studies. However, the proportion of the population exposed to low socio-economic status is much greater than the prevalence of psychiatric disorders. Such distal risk factors, with greater population exposure have not traditionally been the focus of suicide prevention targets in public health approaches or in national suicide prevention policies.

This project aims to elucidate on the various factors that statistically seem to affect suicide rates worldwide, and further draw inferences based on visualisations and models which analyse and study the data. The dataset that we have worked on in our project amasses abundant attributes such as sex, generation, age, country, economic status of the country and so on. Upon carrying out various techniques of exploratory data analysis and visualisations, we have inferred which groups seem to be indulging in suicide often, and what factors are primary in causing this.

II. Review of Literature

An analysis of previous works and predecessors is detailed below:

A. Artificial Intelligence Based Suicide Prediction [1]

In this paper, we see how AI, a fast developing tool, is used to predict suicide rates in the US. In "medical suicide

prediction," AI analyzes data from patient medical records. In "social suicide prediction," AI analyzes consumer behavior derived from social media, smartphone apps, and the Internet of Things (IoT). Because medical suicide prediction occurs within the context of healthcare, it is governed by the Health Information Portability and Accountability Act (HIPAA), which protects patient privacy; the Federal Common Rule, which protects the safety of human research subjects; and general principles of medical ethics. Medical suicide prediction tools are developed methodically in compliance with these regulations, and the methods of its developers are published in peer-reviewed academic journals. In contrast, social suicide prediction typically occurs outside the healthcare system where it is almost completely unregulated. Corporations maintain their suicide prediction methods as proprietary trade secrets. The assumptions made in this paper therefore are that the data more or less accurately represents the actual statistics, even though this is not always the case due to concealed/false reasons that are given sometimes for suicides. Despite this lack of transparency, social suicide predictions are deployed globally to affect people's lives every day. Yet little is known about their safety or effectiveness.

The primary approach therefore is using AI. AI may overcome many limitations of traditional suicide screening tools and increase the accuracy of predictions. In this paper, AI-based suicide prediction tools can be divided into two broad categories: The first category involves analysis of patient medical records. It is performed by doctors, public health researchers, government agencies, hospitals, and healthcare systems. The second category involves the analysis of consumer behavior and social interaction derived from retail purchases, smartphone apps, social media, and other activities outside of healthcare.

The limitations were in accurately finding and eliminating confounding variables. The summary is that medical and social suicide prediction tools may be beneficial to individuals and promote public health. However, they also pose a variety of risks to people's safety, privacy, and autonomy. To minimize those risks, new norms and regulations must be developed to control how suicide predictions are made and used. For example, to protect consumer autonomy, suicide prediction methods could be made more transparent, and users could be given unambiguous opportunities to opt-out and delete prediction information; to protect consumer privacy and minimize the risk of exploitation, suicide predictions should not be used for advertising or be shared with third parties; and to protect consumer safety and autonomy, "soft-touch" suicide interventions such as referrals to counseling centers, could be prioritized over "firm-hand" interventions such police-mediated wellness checks. In some cases, healthcare norms and regulations could be imported for use in social suicide prediction. For instance, social suicide prediction research should be approved by independent IRBs, and ongoing suicide prediction programs should be monitored for safety and efficacy by independent data monitoring committees. Though HIPAA does not currently apply to social suicide prediction, to protect consumer privacy, tech companies could voluntarily adopt HIPAA-like standards, or stricter standards could be imposed on them through new privacy legislation.

B. Use of a Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population [2]

This paper primarily deals with suicides in South Korea. However, the techniques and results used in this particular paper served as the primary inspiration for our own work.

The aim in this paper was to develop a model predicting individuals with suicide ideation within a general population using a machine learning algorithm. The approach: Among 35,116 individuals aged over 19 years from the Korea National Health & Nutrition Examination Survey, the authors selected 11,628 individuals via random down-sampling. This included 5,814 suicide ideators and the same number of non-suicide ideators. They randomly assigned the subjects to a training set (n=10,466) and a test set (n=1,162). In the training set, a random forest model was trained with 15 features selected with recursive feature elimination via 10-fold cross validation. Subsequently, the fitted model was used to predict suicide ideators in the test set and among the total of 35,116 subjects.

Basically, a machine learning algorithm was applied here to public health data to develop a model predicting individuals with suicide ideation in the general population. When predicting suicide ideators in the test set, this model showed a good performance (AUC=0.85) with an accuracy of 78.3%. Moreover, the model could predict suicide ideators among the total population of about 35,000 with an accuracy of 82%. The predictive ability of the machine learning model is comparable to that of suicide risk assessment tools used in the clinical setting.

Results: The prediction model achieved a good performance [area under receiver operating characteristic curve (AUC)=0.85] in the test set and predicted suicide ideators among the total samples with an accuracy of 0.821, sensitivity of 0.836, and specificity of 0.807. Conclusion: This study shows the possibility that a machine learning approach can enable screening for suicide risk in the general population. Further work is warranted to increase the accuracy of prediction.

C. Limitations with predecessor work—and our modifications

What's common between the aforementioned papers is that their respective algorithms were focussed on predicting suicide numbers alone, while either not exploring other factors, or only exploring specific factors such as psychology. Our dataset and the work we have done focus on what socio-economic factors affect suicide rates and which generation and sex of individuals are particularly prone to commiting suicide. With our analysis and prediction, we aim to delve further into why certain sexes and generations are more prone to suicides. Our assumption is that the socio-economic factors do indeed play a role in affecting suicides (as mentioned before, an example that proves this is the rising number of suicides among debt ridden farmers in India). Our predictions strengthen this assumption and affirm it.

III. PROPOSED SOLUTIONS

What we intend to do with the dataset is find out what factors seem to affect suicide rates the most. Our dataset primarily focuses on attributes such as age, sex, country,

country's GDP and so on; thus, we are not intending to predict or determine the emotional/mental/psychological reasons behind suicide. While these factors are usually the primary driving forces behind suicides, our dataset and our analysis is more focussed on predicting on how and why socio-economic factors (such as the financial status of the people's place of residence) play a role in influencing suicides, if at all they do. As we explained in our introduction, socio-economic deprivation is often discussed as a link to a more stressful lifestyle, less access to resources including health care, fewer opportunities for educational achievement (in young people), poor housing facilities and lower self-esteem. What this means is that socio-economic factors play a deeper role than what may commonly be assumed. Often, these situations are what can drive individuals to take such extreme measures. One example is the high and alarming frequency of farmer suicides in India due to loan and debt related issues—a clear reflection of how suicides can very easily be caused by financial and economic circumstances.

The data used for analysis compares socio-economic information with suicide rates by year and country from 1986 to 2016. This compiled dataset is pulled from four other datasets linked by time and place and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.

The features used for this analysis include country, year, sex, age group, number of suicides, population, suicide per 100k population, HDI for year, GDP for year, GDP per capita and generation. The Human Development Index is a statistical composite index of life expectancy, education, and per capita income indicators, which are used to rank countries into four tiers of human development and is used to measure a country's overall achievement in its social and economic dimensions. The HDI for each year and country is taken from the United Nations Development Program(2018). Gross domestic product is a monetary measure of the market value of all the final goods and services produced in a specific time period. Gross domestic product (GDP) is the standard measure of the value added created through the production of goods and services in a country during a certain period. As such, it also measures the income earned from that production or the total amount spent on final goods and services (fewer imports). GDP is retrieved from World Bank - World Development Indicators: GDP (current US$) by country: 1985 to 2016.

A. Preprocessing

The dataset consists of 27,820 records of suicide numbers in 101 countries and 12 attributes, with no duplicate records. The variables include both categorical (country, year, sex, age, generation) and numerical (suicides_no, population, suicides/100k_pop, HDI_for_year, GDP_for_year, GDP_per_capita). Age is grouped into 6 categories: 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years and 75+ years. Similarly, the generations are categorized as G.I. Generation, Silent, Boomers, Generation X, Millennials and Generation Z. There are 19,456 empty cells in the dataset, all of which belong to the HDI_for_year attribute. Since 70% of the total values were missing from a single column (HDI_for_year), we decided to drop this feature before analysis.

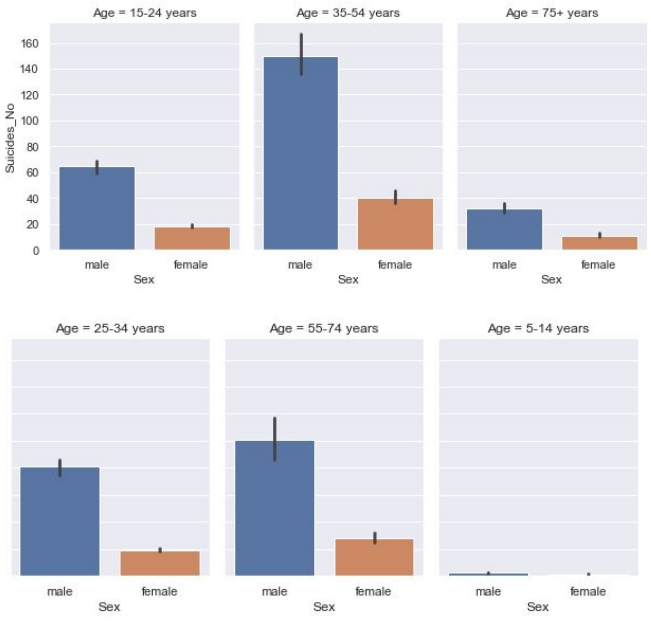
Additionally, to deal with categorical variables before running our models, we applied One-Hot Encoding to transform them into a numerical and workable format. For categorical variables where no such ordinal relationship exists, the integer encoding is not enough. Using regular integer encoding and allowing the model to assume a natural ordering between categories would result in poor performance or unexpected results (predictions halfway between categories), especially since the dataset has large and varying values. This is why one-hot encoding can be applied to the integer representation in our dataset. This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value.

We extensively researched suicides, their causes and the various economic factors contributing to suicide numbers increasing over the years. Our approach to exploratory data analysis of the dataset was directed at figuring out the trends in suicide numbers globally and trying to determine if there are any strong correlations among any attributes, or if some attributes are erroneously predicted. In the case of the latter, we tried to figure out if the errors were caused by missing data, incorrect data, inconsistent data, uneven surveying and the like.

We used Pandas Profiling to get a thorough and complete overview of our data. Following this, we drew conclusions on what visualisations were best suited to give an accurate idea of the relations between the various attributes of our large dataset.

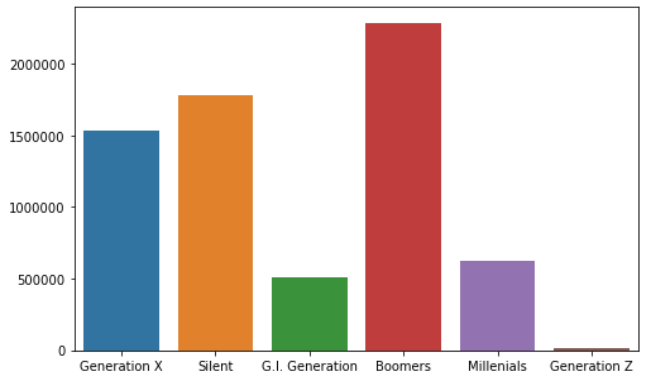
Amongst 11 of our visualisations, we have attached the most relevant graphs and inferences in the subsequent sections.

Sex-wise suicide numbers for every age group



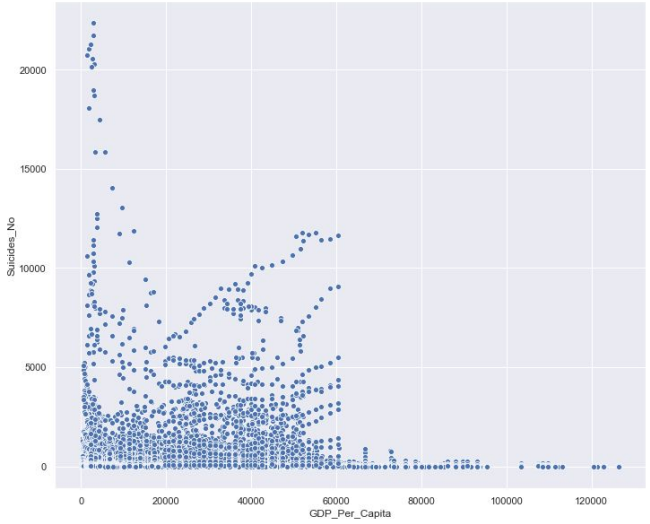
The above graph shows that consistently across all age groups, suicides are higher in the male sex. What's also inferred from this graph is that the age group of 35-54 years has the highest number of suicides. This is a globally well known fact as well. What this means, thus, is that our dataset and predictions are in accordance with actual case studies that have been conducted globally.

Generation having the highest number of suicides



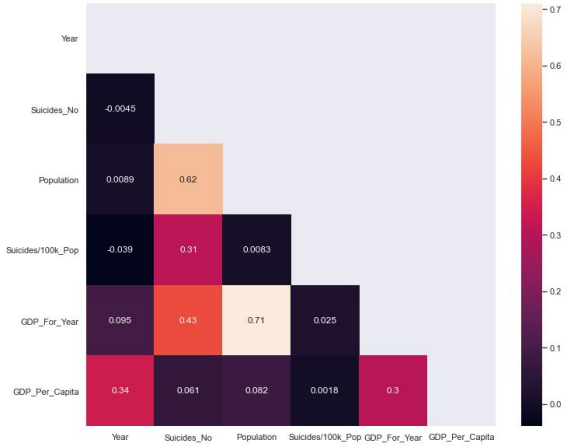
The boomers, silent and X generations are made up of people born until 1976. These are the ones who were most in the age range where most suicides occur. Just check the chart that deals with the age brackets. Boomers fall prey to suicides more often than other age groups (1946-1964).

Does GDP Per Capita influence the suicide rate



Here we study the relation between GDP Per Capita and No of Suicides. According to the above generated visualisation, poorer countries are prone to a larger number of suicides when compared to richer ones. As the income increases, suicide rates appear to decrease. However, from the 20k (GDP_Per_Capita) mark, the number of suicides increases again. The data then seems to show certain flaws beyond the 60k mark.

D. Does GDP Per Capita influence the suicide rate?



As can be seen from the above attached graph, GDP_For_Year and Population have the highest linear correlation, implying that higher population would mean higher gdp. However, the relation between GDP and population is ultimately unrelated to the number of suicides in a country. Population and the number of suicides have the second highest correlation. This is an inherent inference since the number of suicides in larger populations could typically be higher. We also observe that Suicides/100k_Pop vs GDP_Per_Capita has a negligible correlation. This shows that the two aren't linearly dependent on each other.

B. Building Models

We have used three models for predicting two attributes in our dataset: sex and generation of the individual.

The first model we used is **Logistic Regression**. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). Logistic regression performed on generation gave us a satisfactory accuracy of 94.06.

Next is a model using **K-fold Cross Validation with Random Forest Classifier**. Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. We used a 10 fold cross validation model using grid search and a random forest classifier consisting of 1000 trees. Our accuracy for the attribute sex was 60%—which was after many trials of hyperparameter tuning. What we thus concluded was that regular sampling (i.e. splitting the dataset once into a specified ratio) worked better for our dataset.

The third model we used is an **Artificial Neural Network**. Artificial neural networks, usually simply called neural networks, are computing systems vaguely inspired by

the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. When used to predict the sex of the individual, our model gave us an accuracy of 99%.

C. Evaluation

Our conclusion is that an artificial network worked best for the type of dataset we have. The accuracies achieved were highest for logistic regression and artificial neural network and quite low for k-fold cross validation. The clear conclusion to draw from here is that splitting our dataset once into train and test sets worked best for our data, rather than repeated splitting and resampling.

IV. Experimental Results & Insights

We used a range of varied models to predict the two attributes that we were focussed on the most—sex and generation. The best performing models were logistic regression and artificial neural network. K-fold cross validation with a random forest classifier resulted in a low accuracy. We used a 10 fold cross validation model using grid search and a random forest classifier consisting of 1000 trees. Our accuracy for the attribute sex was 60%—which was after many trials of hyperparameter tuning. What we thus concluded was that regular sampling (i.e. splitting the dataset once into a specified ratio) worked better for our dataset.

We looked into why, thus, our model was failing when it came to k-folds. There are situations where proper splitting is particularly hard to achieve, and cross validation becomes infeasible. Consider a problem with a number of confounders. Splitting is easy if these confounders are strictly nested (e.g. a study with a number of patients has several specimens of each patient and analyses a number of cells of each specimen): you split at the highest level of the sampling hierarchy (patient-wise). But you may have independent confounders which are not nested, e.g. day-to-day variation or variance caused by different experimenters running the test. You then need to make sure the split is independent for all confounders on the highest level (the nested confounders will automatically be independent). Taking care of this is very difficult if some confounders are only identified during the study, and designing and performing a validation experiment may be more efficient than dealing with splits that leave almost no data neither for training nor for testing of the surrogate models. This interesting example is a good way of understanding why k-fold and resampling can fail sometimes.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). Logistic regression performed on generation gave us a satisfactory accuracy of 94.06%. The third model we used is an Artificial Neural Network. Artificial neural networks, usually simply called neural networks, are computing systems vaguely inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. When used to predict the sex of the individual, our model gave us an accuracy of 98%.

From our exploratory data analysis as well as prediction models, what we can infer is that the sex and generation of suicidal individuals indicate which portion of the general population across the world is more likely to commit suicide: men, and the boomers. Socio-economic status had a similar magnitude of population attributable risk for suicide as mental disorders. Public health interventions to reduce suicide should incorporate socio-economic disadvantage in addition to mental illness as a potential target for intervention. More distal risk factors associated with suicide, such as measures of socio-economic status, have been shown to have substantially lower relative risk estimates associated with suicide in population-based studies. However, the proportion of the population exposed to low socio-economic status is much greater than the prevalence of psychiatric disorders. Such distal risk factors, with greater population exposure have not traditionally been the focus of suicide prevention targets in public health approaches or in national suicide prevention policies. This project aims to elucidate on the various factors that statistically seem to affect suicide rates worldwide, and further draw inferences based on visualisations and models which analyse and study the data. The dataset that we have worked on in our project amasses abundant attributes such as sex, generation, age, country, economic status of the country and so on. Upon carrying out various techniques of exploratory data analysis and visualisations, we have inferred which groups seem to be indulging in suicide often, and what factors are primary in causing this.

In a report on suicides in men vs women by BBC [3], men were statistically shown to have consistently higher rates of suicides across the world. There are two broad reasons affecting this: psychological and socio-economical. These factors are both unlike and alike each other in interesting ways.

Evaluating this from a psychological point of view, we have several conclusions to draw from our environment alone. Men are always expected to be “masculine” in the traditional sense. One key element is communication. It’s too simplistic to say women are willing to share their problems and men tend to bottle them up. But it is true that, for generations, many societies have encouraged men to be “strong” and not admit they’re struggling. These contribute to being significant psychological reasons behind why men are oppressed and often go to extremes. This does not, in any way, mean that women aren’t oppressed or are freer than men are to express their thoughts and feelings. Depressed men may also try to mask their sadness by turning to other outlets, such as TV, sports and working excessively, or engaging in risky behaviors, such as gambling, smoking, unsafe sex or driving recklessly. Depression is also more likely to show up as anger and irritability in men and teenage boys. When you contrast the societal responses to men being expressive as opposed to women, the former is always shamed as being ‘vulnerable’ and ‘fragile’ in doing so, going against the perceived and ‘accepted’ norms for how men should behave. Such factors lead to bottled up emotions and the eventual inclination towards taking drastic measures to escape the stress and captivity of their own traumas and woes. From a very young age, boys are taught to not cry. “Boys don’t cry” seems like

a harmless phrase simply because it's so commonly and widely perceived. This suppression continues to teenage, adulthood and middle age (note that the boomers age group suffers the most). Although women are hit harder by depression and are more vulnerable to it because of their biology, the illness is missed more frequently in men, Goldstein told Live Science. Health care professionals and even family members may not pick up on depressive symptoms in men, so they can end up with severe depression before it's detected.

Now coming to the other factor: socio-economic. Socio-economic deprivation is often discussed as a link to a more stressful lifestyle, less access to resources including health care, fewer opportunities for educational achievement (in young people), poor housing facilities and lower self-esteem. However, it is rarely discussed in terms of intervention or prevention. In short, while the social origins of depression and hopelessness may be acknowledged, individualistic notions of risk and psychopathology are frequently privileged, in terms of intervention and prevention. As a corollary to this, key risk factors are conceptualized clinically: depression and hopelessness as mental illness rather than reactions to life circumstances, and while the correlation to economic deprivation has been acknowledged there has been little attempt to analyse it. Socio-economic status had a similar magnitude of population attributable risk for suicide as mental disorders. Public health interventions to reduce suicide should incorporate socio-economic disadvantage in addition to mental illness as a potential target for intervention. More distal risk factors associated with suicide, such as measures of socio-economic status, have been shown to have substantially lower relative risk estimates associated with suicide in population-based studies.

V. Conclusion

The reason we chose this dataset is because of our active interests in mental health awareness. In India alone, we hear suicide rates increasing by the year amongst students. Farmer suicides have gripped the nation since years now. There are so many complex psychological, social, economical and environmental factors that go into this. This project made us realise what kinds of people are more affected by such issues, leading to suicides.

Appropriate measures need to be taken by governments across the globe as well as people themselves to ensure that the root causes of suicides can be dealt with and tackled. Support needs to be given at every possible stage to curb suicide rates and aid those in need for help. Although mental health awareness has increased in the past decade, it still has a long way to go.

Overall, this was an educational insight into both the technical and societal aspects of data analytics. This project gave us an insight into how data collected by census can be compared with what machine learning models can predict. Data analytics and machine learning combined can provide outstanding results. This was an extremely insightful project.

In terms of workload, our tasks were evenly distributed amongst us. We each took up one model, and split the EDA portion evenly amongst us.

Data Analytics is a vast ocean to swim in; to have done as much as we have covered in this paper is still just barely scratching the surface.

References

[1] Mason Marks, "Artificial Intelligence Based Suicide Prediction," Harvard University - Harvard Law School; Harvard University - Edmond J. Safra Center for Ethics; Gonzaga University - School of Law; Yale University - Information Society Project, January 29, 2019.

[2] Hyeongrae Lee, Seunghyong Ryu, "Use of a Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population," Hanyang University, Chonnam National University, South Korea, October 2018.

[3] "Why More Men than Women Die by Suicides", a report by BBC.