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.

Over ninety percent of people suffering from mental illnesses or diseases lead to suicide. Substance and drug abuse of various kinds such as using light drugs like marijuana and hard drugs like methamphetamine and cocaine. Stress and anxiety in daily lifestyles can lead to a host of problems ranging from minor sorrow to full fledged depression, which can lead to suicidal tendencies building up.

Suicides are a severe problem amongst all age groups. We often hear the present generation's youth spiralling into all sorts of abuse and issues that can lead to depression, anxiety, psychological issues and whatnot compounding massively. There's always a risk factor associated with the way the youth indulges itself in substance abuse and even technology for that matter. While alcohol and drugs have been an age old issue and nothing new or out of the ordinary, technology and electronic media has a huge role to play in what leads to compounding mental health issues and problems in today's generation. The need to look for validation and constantly seek approval from other peers on social media for the sake of getting more likes, shares, reshares, retweets and so on has gripped the generation. People go very far for such social media related validation, and also go to equally far lengths to deal with the induced trauma of not receiving enough attention on social media. On the flip side, cyberbullying is another end of the spectrum when it comes to interaction on social media. Cyberbullying is nothing new; people are driven to commit suicide by threats and pressures online as much as it can happen offline.

While that was an encapsulation of what may affect youth, there's also the problem of what other grounded and socially associated reasons may lead to people taking such drastic and extreme measures. Economical and social reasons are also equally responsible for leading individuals to commit suicide. The financial status of a person matters just as much as the mental well being in an economically underprivileged family. The very real threats of not being able to afford a daily mail, water, clothing, shelter, resources for children and old aged members of the family, and being unable to pay back debts are amongst the leading factors of suicides amongst the poorer sections of the population. This does not mean that only those who are poor tend to commit suicide; we are well aware that mental health sees no status and wealth, and mercilessly attacks whoever falls prey into it. While the rich can afford basic necessities—and often luxuries that surpass basic needs and lead to a socially perceived "comfortable" lifestyle—depression can plague the minds of even the most secure people.

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]

Here, In thisparticular paper, artificial intelligence is used to predict the suicide rates in the United States of America. In this paper there is heavy research done into the basic medical condition of citizens in the US and what sort of access they typically have to resources. Healthcare in the US was in boom when President Barack Obama was in power; however, historically, the capitalistic society of the US has been well known for being very expensive when it comes to medical facilities and dental facilities. They have never been on the cheaper side for anyone, regardless of the economic and social background of the citizens. Often, because of the persistent racism, medical resources can be denied to those who aren't white.

This is why the first thing looked into in this paper is medical resources and records of patients spread across all American states and provinces. This data comprises the records belonging to all publicly owned hospitals in every state of the country, primarily collected by doctors, nurses, receptionists and the like. These records were then cross verified with census numbers and statistics conducted by popular organisations of mental disease numbers, types and deaths caused by individuals committing suicide. The primary approach therefore is using artificial intelligence, since this is the best way to learn from given data and predict the outcome of a neural network.

The issues with this paper however was in accurately spotting and dealing with confounding variables which caused a major issue for the artificial intelligence based prediction model. Confounding variables are those that cause unexpected and practically senseless (i.e. not making sense practically in real life scenarios) relations between two or more seemingly unrelated variables.

Nevertheless, the approach taken with this paper was good and direct, while being inherently simple to understand and easy to gauge. Still, despite the simplicity, the difficulties in handling large amounts of categorical variables and other attributes that were difficult to scale lead to accuracies of the range of eighty to eighty-five percent. Combined with this is the fact that it was after all based on data that was manually collected by individuals and then merged to make one. The report does not deal with global suicide rates, only US ones.

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.

Unlike the previous paper, this paper deals with a model that can study different attributes to predict what may lead to sucide, rather than predicting the final numbers. In other words, this paper focuses on figuring out what the mindset or ideology behind suicidal people will be. This way those who seem likely to commit suicide can be spotted and figured out.

Main techniques used here are based on the vast ocean that is machine learning. Machine learning is often a giant pool of various supervised and unsupervised techniques of prediction. Specifically, this paper deals with multiple models after resampling and downsizing data. A random forest classifier is initially used on data (collected from Korea National Health & Nutrition Examination Survey). After this, a ten fold cross validation model was built to analyse the patterns and provide an accuracy of prediction of suicidal tendencies or ideations.

The result was a decent percentage of accuracy—eighty-two percent for a population of 35,000 and an accuracy of 78% on the test set alone which consisted of a fifth of the aforementioned population size.

Finally, the conclusion of this paper and model is that it's a proof of how truly powerful machine learning is and how it can be used intelligently to solve a number of issues. Suicide, being a grave and serious matter, can be prevented or curbed to an extent by taking surveys of people's mental health alone. Why? Because this paper's model takes input of people's mental health status and conditions based on various factors and predicts what the output may be—suicidal or not suicidal—based on what the machine learning model has accurately learned from previous data.

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 can lead to several issues that lead to unavailability of basic resources, causing extreme stress and depression among the poorer parts of the society. 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 suicide rates by year and country with socio-economic details from the years 1986-2016. The features used for this analysis include country, population, suicide per 100k population, year, GDP for year, sex, age group, HDI for year, number of suicides, GDP per capita and generation. The HDI is just the Human Development Index which gives an idea of the educational levels, income levels and expectancy of life, which are usually linked to categorising countries in the world into different tiers of development (what we hear about third world countries, first world countries and so on); which is all taken directly from the United Nations Development Program (2018). Gross domestic product measures what the value for goods and services produced are before reaching the market and being made available to consumers.

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: 15-24 years, 5-14 years, 35-54 years, 25-34 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.

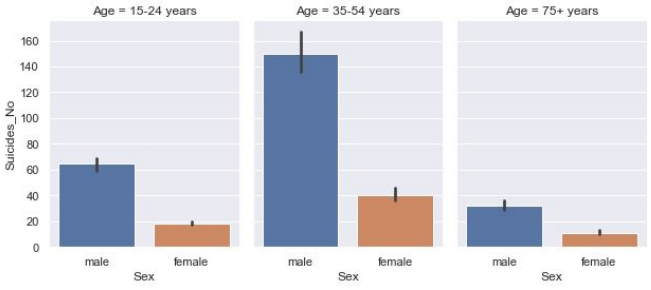
Additionally, to deal with categorical variables before running our models, we applied One-Hot Encoding to transform them into a numerical and workable format. Why one-hot encoding? Say you have a number of variables that are all categorical or maybe even nominal in value. What this method will do is make a sparse matrix of sorts wherein the position indicating a particular value will be marked as 1 and the rest are 0s. For example, if an attribute has three values—red, blue, green—one hot encoding will turn it into [1,0,0] for red, [0,1,0,] for blue anf [0,0,1] for green. This very simple manner allows discrete and unrelated values as well as string type values to be transformed into a computable version.

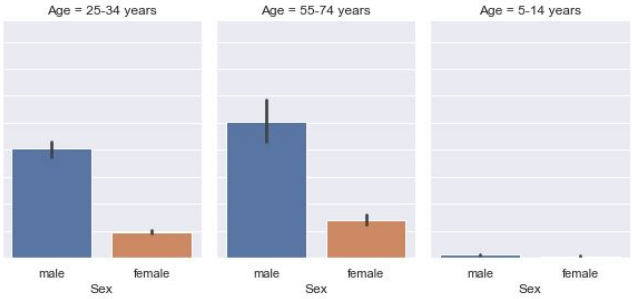
We extensively researched suicides, their causes and the various economic factors contributing to suicide numbers increasing over the years. Our approach to exploratory data analysis of the dataset was directed at figuring out the trends in suicide numbers globally and trying to determine if there are 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.

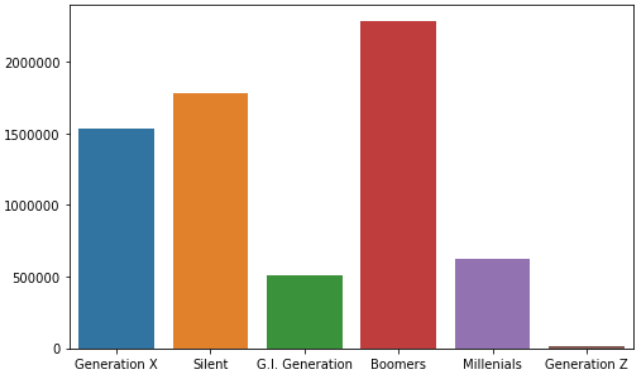
Sex-wise suicide numbers for every age group





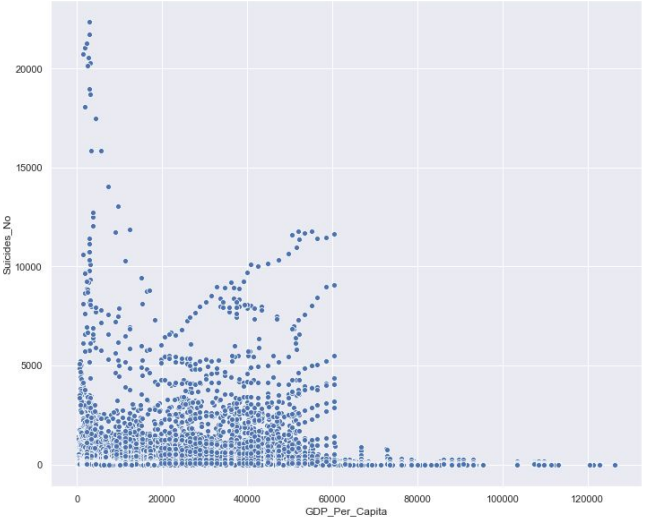
The above graph shows that consistently across all age groups, suicides are higher in the male sex. What's also inferred from this graph is that the age group of 35-54 years has the highest number of suicides. This is a globally well known fact as well. What this means, thus, is that our dataset and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

Generation having the highest number of suicides



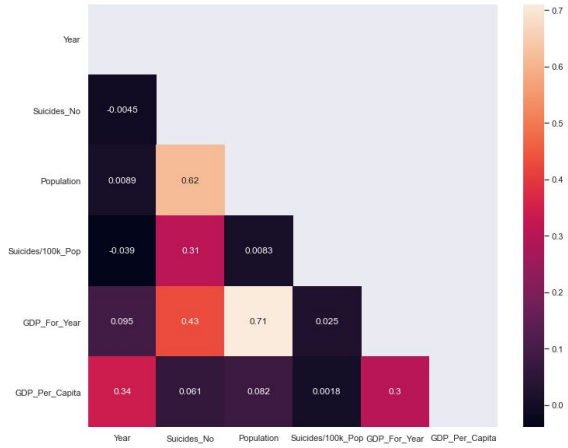
Currently, the generation that is in its 50s or upwards—called boomers or just baby boomers—are the ones that seem to have the highest number of suicides. Boomers fall prey to suicides more often than other age groups (1946-1964).

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.

Correlation Matrix



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 deaths caused by suicides in a country. The second highest correlation is between the number of suicides and total population. 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**. In logistic regression, a binary dependent variable is modeled by using a statistical model in its logistic form (in the simplest type) and can even be more complexly designed to suit larger datasets with more tricky relations between their attributes. In regression analysis on the other hand, the parameters of a logistic model are used for logistic 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 used for evaluating machine learning algorithms and models when the data you have is limited in size and resampling may help. 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. The first step in this was to incorporate random forests or decision forests which are basically exactly like what they sound: forests consisting of many decision trees grouped together. It can be used for regression or classification both. The output prediction is the result of either the mode or the mean (sometimes median) of whatever has been given out by every single tree in the forest. We used grid search with a ten-fold cross validation model along with 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 are like the reflection of the human biological brain; just like how our brains consist of neurons that fire away to give signals and messages to the rest of our body, an artificial neural network makes use of computerised and coded "neurons" that analyse the data fed into them (feed-forward) and give an output of what the result may be. When used to predict the sex of the individual, our model gave us an accuracy of 98%.

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. Let's take an example of a problem consisting of several confounders (confounding variables have been discussed before). Consider a problem with a number of confounders. There won't be issues for nested confounding variables, because you're just splitting at the level of highest sampling in terms of hierarchy. But sudden and unexpected confounding variables may appear in the dataset during transformations and scaling made to make the data suitable. In this case, the splitting needs to be independent on every

level. What this means, in essence, is that there may be confounding relations coming up in each fold and split of the dataset, leading to multiple erroneous conclusions and relations drawn. This is why k-fold cross validation may sometimes fail, because it can cause compounded issues and errors while sampling and resampling the dataset.

In logistic regression, a binary dependent variable is modeled by using a statistical model in its logistic form (in the simplest type) and can even be more complexly designed to suit larger datasets with more tricky relations between their attributes. In regression analysis on the other hand, the parameters of a logistic model are used for logistic 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 are like the reflection of the human biological brain; just like how our brains consist of neurons that fire away to give signals and messages to the rest of our body, an artificial neural network makes use of computerised and coded "neurons" that analyse the data fed into them (feed-forward) and give an output of what the result may be. 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.Since our dataset itself deals with interesting attributes such as the gross domestic product and economic status of a country, it brings us back to the main point we began with: how social and economic factors can lead to suicides in large populations. 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. Added to that is the lack of openness and communication coming forward from men to actually address their own issues. Combine this with the unwillingness to receive help. 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. 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). Sometimes, rather than society being the reason they are shamed, it's also their own internal mindsets—which have been trained by their environments to be this way—which lead to men's increased rates of depression and suicides.

Now coming to the other category of people that are highly affected—the present generation of boomers. As seen from our previous graphs, even amongst boomers, it's once again men who are affected the most. There are many possible social causes for boomers succumbing to suicides. Today's fast paced work life and culture often leads to people neglecting their old parents. While this may sound insignificant, it's actually a massive psychological factor in aggravating suicidal tendencies in old citizens who feel like they are being left behind or neglected. One more reason is illnesses. Euthanasia (commonly termed as mercy killing) is illegal in many countries. Often, the suffering caused by illnesses can lead to them choosing the unfortunate but "easy way out", by taking their own lives.

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.