Identifying Risk Groups of Suicides in India using Visualization and Machine Learning Techniques

Adeel Akhtar
B00775583
Master of Computer Science
Dalhousie University
Halifax, Canada
Adeel.Akhtar@dal.ca

Ashfaq Adib
B00855120
Master of Computer Science
Dalhousie University
Halifax, Canada
ashfaq.adib@dal.ca

Noah Sealy
B00726289
Master of Computer Science
Dalhousie University
Halifax, Canada
noah.sealy@dal.ca

***Abstract*—Suicide has become a more pressing issue, which can no longer continue to be overlooked. Even low suicide rates in a country of such mass as India would mean thousands of deaths each year. Unfortunately, India's suicide rate is far from low. The project developed by Data Benders takes a data driven approach to aid researchers in analyzing this ever-increasing problem. An interactive suicide data dashboard was created to show insights of suicide, using machine learning and visualization methods. Through use of this application, groups of people who are at risk to suicide can be identified based on key demographics.**

1. Introduction

Suicide is one of the leading causes of death worldwide, as per latest WHO Statistics [1], approximately 800K deaths occur every year due to suicide. Young adults between 15 to 49 years of age are among the most affected age groups [2]. The increasing phenomena of suicide is very common across all countries and includes several factors, which ranges from socio-economic reasons, cultural problems, mental health issues etc. In recent years, suicide rate has become higher in Asian countries [3] which is an alarming situation for the governments of these countries. The issue is becoming more challenging to deal with day by day. India has the second largest population worldwide with highest mortality rate among southeast Asian countries [4]. Here, the problem of suicide is more severe, which accounts for the need of thorough systematic analysis of the situation, as well as identifying actions that needs to be taken by the government authorities to overcome this issue.

The main theme of our project is to analyze and visualize the suicide related data of India. This project also aims to identify the risk factors based on demographics to devise timely national level strategies to mitigate this problem that has been prevailing in recent years. To achieve this, we applied machine learning and visualization techniques to better understand the factors associated with suicides in India, as well as developed a tool that can be adopted by experts for goal-specific suicide analysis. Based on the findings from using our developed application, there remains differences among different age groups for causes of suicide. For example, in young children the main cause was identified to be failure in examinations while it was family problems for adults. Moreover, our findings reflect that the suicides in India are plagued by more socio-economic factors than mental health issues as observed in first world countries [5], which emphasizes the need for interventions that can strengthen and aid in such factors to help prevent this prevailing issue.

2. Background

In India, suicide is one of the leading causes of deaths [6]. Globally, it accounts for more than 25 percent of the total suicide deaths [6]. India's share of global suicide deaths has drastically increased from 25.3% in 1990 to 36.6% in 2016 among women, and from 18.7% to 24.3% among men [7], which shows the alarming situation for policy makers. There are multiple factors that are involved in this phenomenon like cultural, social, economic, political issues, and more. Often, these factors are interrelated and complex in nature, so extensive systematic research is needed to identify and understand the risk factors associated. In addition, suicide methods are very different from country to country, and even from state to state within a country. For example, in USA firearms are used in most suicides while in countries with large rural population like India, pesticides are mostly used for suicide [8]. Restricting access to the suicide means is known to be one of the most effective prevention strategies in India [9]. Also, mental health issues are less associated with suicides in developing countries (low-income countries) as compared to developed countries [5].

Most of the Indian literature on suicide focus on farmers. The literature has more emphasis on following risk factors associated with farmers, like slow output growth, financial constraints, crop failure, high unemployment and more [4]. As of 2001, the overall suicide rate for farmers across India was 15.8 per 100,000 people, 50% higher than the general population's suicide rate [10]. As per [10], the leading factors associated with farmer's suicides include indebtedness, increased use of cash crops instead of food crops, introduction of Bt Cotton in 2002 and reduction of bank loans. The reduction of bank loans reflects economic factors affecting the situation [10].

The estimated suicide data provided by Global Burden of Disease (GBD) is higher than the suicide data that is available on National Crime Records Bureau (NCRB), which leads to the potential of under-reporting suicides cases by the government officials of India- NCRB [11]. A study using verbal autopsy investigations of all unnatural deaths in rural areas of India found that the suicide rate is five-fold higher than the national average [8]. This depicts the fact that quality of suicide data in India is quite limited and needs thorough review by the experts. The key findings related to Indian suicides as per [8] are that the age group 20-29-year account for 41-62% of all suicides, 50-66% of suicide victims had low socio-economic status, self-poisoning accounts for 16-49% of all suicides, and hanging accounts for 10-72% of all suicides.

To devise an effective national level strategy for suicide prevention, studies are needed to identify the risk factors, hidden patterns, local issues, and vulnerable/potential group. Then, suicide prevention measures can be implemented to targeted specific groups like awareness programs, education in health care setting and restricted access to means of suicide. A machine learning tool with visualization can assist its users to identify the root causes of these problems as well as to identify the hot spots in India. This will eventually help the government to educate the masses and take necessary steps at national and state level. Ultimately, these steps will contribute to the overall prevention of suicide throughout India.

3. Methodology

Our project implements visualization and machine learning techniques to analyze risk factors associated with suicides in India. To achieve this, we have developed a web application using the Dash framework [12] that offers an interactive interface to analyze the data on suicides in India. This application was developed using the Python programming language [13]. For processing the data, Pandas tool [14] was used along with some basic libraries that comes with Python. For training the generating machine learning models, we used the Scikit-learn [15] tool that offers numerous implementations of machine learning algorithms. Through the help of these tools, we were able to develop a web application where we analyzed our data to fulfill the goals of our project. In the following sections we will discuss the data collected, machine learning methods applied, and visualization techniques employed in our project.

3.1. Data Collection

The suicide related data of India used for this project is available on Kaggle website [16]. The dataset contains data on suicide deaths between year 2001 and 2012, and contains death count in different states stratified by year, age group, and gender. Age group comprises of the following ranges: 0-14, 15-29, 30-44, 45-59, and 60+ years. This data also contains details of these deaths divided into the following categories:

- Causes (e.g. Bankruptcy, Marriage, Mental illness etc.)
- Education (e.g. No Education, Diploma, Graduate etc.)
- Means of Suicide (e.g. Hanging, poison, firearm etc.)
- Profession (e.g. Unemployed, Agriculture, Retired etc.)
- Social Status (e.g. Married, Unmarried, Widowed etc.)

For the categories above, the data contains the number of suicides committed by people of the different age groups (0-14, 15-29, 30-44, 45-59 and 60+ years) and genders (male and female). This is with the exception for Education and Social Status categories where the data only contains suicides from all age groups (0-100+). The data did not contain population data for each state, but it was required find out suicide rate of each state. To fulfill this, we collected supplementary data corresponding to the population of each state.

For the population, we used the publicly available data for the year 2001 and 2011 on the official Government of India website [17]. Since population census is performed once every 10 years, so as an estimate of the population of other years for our data (2002-2010 and 2012) linear interpolation was performed. Normally, mortality rate is defined as per 100,000 of population [18]. Following this norm, for the suicide rates, we calculated suicide rates per 100,000 of population. The equation we use to derive suicide rate is:

$$(total\_state\_suicides * 100000)/state\_population$$

For the machine learning part of our project (clustering), our application uses the entire dataset without population and suicide rates. The focus of this part was to analyze the total number of suicides associated with certain factors. For the interactive map, our application used the suicide data with a combination of population and suicide rate to visualize both the total number of suicides in each state as well as their suicide rates.

3.2. Machine Learning

As mentioned earlier, the data we collected has suicide related information on different states from the years 2001-2012. The data could be separated based on the causes and means adopted in committing suicides, along with the education status, professional profile and social status of the people who took their own lives. This information could also be split based on genders and different age groups. As our goal of the project was to identify risk factors associated with suicides in India, we wanted to extract group of individuals who are at risk and popular causes of or means adopted in suicides.

To achieve our goal, we applied a clustering method which grouped together similar elements in a data while separating the ones that were different. By doing so, we believed that we would be able to find risk factors associated with suicides by analyzing the generated clusters, where data associated with higher number of suicides would be grouped together based on the factors stated above. To apply clustering methods, first, we prepared the data by cleaning and encoding it. There were numerous rows in the data that had 0 suicides associated so we removed them, as they do not incorporate to the risk factors in suicides. We inspected the data for outliers with the help of Pandas profiling. No outliers were found that needed to be dealt with. The data also had no null values that required to be filled. For clustering, we label encoded the Age Group column of the data as it had a natural ordering, i.e. 0-14, 15-29 and so on. The columns- State, Type and Gender were one hot encoded as they are categorical variables. The Total column was normalized so that it would not introduce any imbalance in the distance measurement.

We started with using the K-Means clustering method for our analysis as it was a very easy to understand. To analyze the clusters, we took help of visualization techniques by generating Sankey diagrams (Figure 3), details of which will be discussed in the next section. To observe the quality of the clusters, the model's Silhouette scores [19] were measured, which incorporated the distance between elements of the same cluster and distance between elements of different clusters. This provided a numeric score to the model as an indication of its performance. However, the scores of the models were not sufficient to justify the clustering models, as our goal was to find risk factors from the clusters, however the scores could be high if, for example, the clusters were well separated by the Year or State column. In fact, the scores of the model were between the range 0.2-0.5 using the K-Means algorithm, but the clusters were not distinct enough to find specific risk factors associated. Moreover, K-Means algorithm only supports the Euclidean distance metric for measuring distances. This metric is useful for numeric variables, however, most of our data included binary variables (one hot encoded columns). Considering the limitations of the K-Means method, we used agglomerative clustering algorithm for our project.

The implemented agglomerative clustering method supported various distance metrics and linkage options. So, for our project we wanted to see how different parameters change the models, and which models would help find the clusters that best suit the user's needs. Thus, we provided three options for choosing a distance metric: Euclidean, Manhattan and pre-computed. The pre-computed distance metric combined two metrics to measure a distance matrix i.e. Jaccard coefficient for binary columns, and Euclidean distance for the remaining columns. The four linkage options: Ward, average, single and complete were also included. Finally, the user would choose the number of clusters to generate. The user can perform clustering on the suicide data by selecting a category (Causes, Means Adopted etc.) as stated in Section 3.1. Based on the data and model parameters selected, a model would be generated and displayed to the user using a Sankey diagram (Figure 3). Each cluster of a model could then be analyzed by clicking on it through the visualization tools described in next section. Here, we also allowed the user to see the model's silhouette score by hovering over any cluster that belongs to that model. However, as discussed earlier, the scores are not sufficient to find risk factors.

The agglomerative clustering algorithm seemed appropriate for our project as we could include a pre-computed distance metric. This metric was applicable for that data that had different types of variables. Also, using a Sankey diagram further supports this method as we can see how different clusters are being generated based on changing the parameters of the clustering algorithm. Moreover, no matter which algorithm we choose, it eventually is upon the user of the system to analyze the clusters and find risk factors. Thus, considering the freedom of customization in parameters and choice appropriate metrics, we decided to use agglomerative clustering method for our machine learning model.

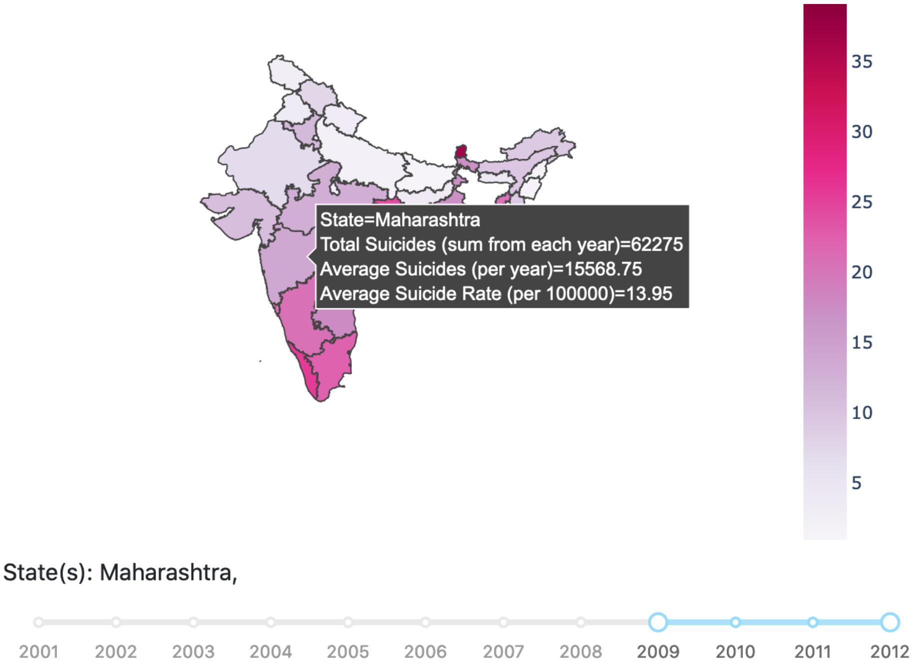
3.3. Visualization

This section discusses the different visualization portions of our project. One of the pages of our application- Interactive Map is heavily focused on visualization techniques. Additionally, the clustering page of our application is only fruitful with the help of the visualization provided. These two visualization parts of our application are discussed below.

3.3.1. Interactive Map Visualization. The interactive map, shown in Figure 1, can be found on the left hand side of the Interactive Map page of the application. The trend figure section (Figure 2) can be found on the right hand side of the same page. These sections of the application are purely devoted to visualizing the data with respect to time and location. This is in order to effectively show areas where suicide may be most common, as well as trends relating to suicide throughout India. In order to efficiently convey this information, both sections of the dashboard have specific features. This section will briefly describe these features, in regards to their functionality, and the development choices behind them.

The interactive map itself provides a data visualization of the suicide data of India with respect to the locations in which those suicides take place. We chose to show visualization on the map as it shows specific state locations of where each suicide occurred, providing a new level of insight for those analyzing the data. Our goal with this project was to provide researchers with a dashboard which allowed them to view suicide data effectively and efficiently; this map allows them to point out problem areas throughout the states of India.

Using the drop-down menu above the map shown in Figure 1, the map's color scale may be selected to show the



State=Maharashtra
Total Suicides (sum from each year)=62275
Average Suicides (per year)=15568.75
Average Suicide Rate (per 100000)=13.95

State(s): Maharashtra,

Figure 1. Visualization of the interactive map's hover feature, showing the suicide data of Maharashtra between the years 2009 and 2012

total number of suicides given each state, or the suicide rate per 100,000 people given each state.

By visualizing the total amount of suicides with respect to state, researchers will be able to know where the most suicides occur, thus where to focus their resources for prevention if the resources are finite. The total number of suicides offers a perspective on the raw magnitude of each state's suicide numbers. This number is not in terms of the other states and their populations though, which may be problematic as a state with a higher population is more likely to have more suicides. This has the potential of overshadowing suicide problems in less populated states.

In order for the number of suicides to be with respect to the population of each state, we provide an option to show the suicide rates per one hundred thousand people. We believe that by offering this additional option, we may be able to tell a more complete story of which states are the most effected by suicide. Ultimately this aims to remove any bias that observers may have towards larger states, as states with smaller populations but larger suicide rates may actually be the area to focus on given finite prevention resources.

In order to have a better idea of the current suicide numbers of these states relative to previous years, we have provided a range slider below the map. This slider allows users to select which years of data they wish to analyze. For example, choosing between 2009 and 2012 will change the color scale of the map to represent either the average total suicides or average suicide rate given each state between the years 2009 and 2012. A simple mouse hover over the data will show the relevant information to the highlighted state. This further tells observers more behind the story of each state's struggle with suicide throughout the years. The slider feature and map hover feature are shown in figure 1.

The year selection slider gives observers the ability to notice key highlights and trends in the data in a way that would prove very difficult from just looking at a spreadsheet of the original data set. This identification of trends is very important in the prevention of suicides. Suicide trends is what the trend figures section of the interactive map page is devoted to visualizing. Upon selecting a state by clicking on it in the interactive map, or by selecting multiple states using the box and lasso select tools, figures are shown plotting suicide totals throughout the years 2002 and 2012 for those states, as shown in Figure 2. In order to further refine the data, the user is able to select which category of data they want to view, the gender, and age groups of the victims.

The trend lines further tell the story of what suicide looked like throughout the states of India between 2001 and 2012. Each trend line corresponds to one class of the selected type. The visibility of its line can be toggled on or off in order to allow for users to specify which data they wish to analyze. Although no claims can be made about the correlations between types, these trends allow users to gain insight on key demographics of those who commit suicide, and even some reasons into why. This insight may allow researchers to identify specific groups at risk throughout the states of India. The results section of the report will walk through some examples of how to identify key information and trends, and potential risk groups using these figures.

Overall, the visualization throughout the interactive map page intends to allow users to view the data in a way that

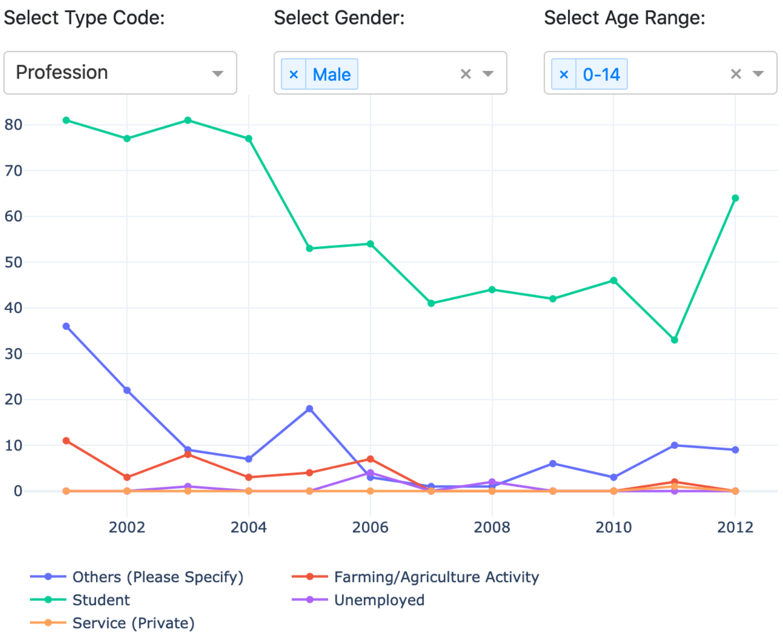


Figure 2. Trend figures section of the Interactive Map page of the application. Showing total suicides per profession, for men of age 0-14 in Maharashtra from 2001 to 2012.

cannot be done from parsing through the raw data. The interactive components allow for a user experience which tells the what, where, when, and perhaps even why of the suicide data in India between 2001 to 2012. This provides a powerful tool for researchers, as it can assist them in investigating and analyzing key risk groups and target zones for the allocation of finite resources that promote suicide prevention.

3.3.2. Clustering Visualization. To visualize the machine learning part of our project, we chose a Sankey diagram 3 as it can display the flow of cluster elements from one model to another. Our application supports trying out different parameters for the clustering model; observing how the clusters are being divided as more clusters are generated, and the change in clusters based on model parameters selected such as, the distance metric or linkage. By observing these changes along with seeing the scores of the models (which could be seen by hovering over a cluster), it would help the user in understanding and selecting a preferred model. We also set the color intensity of a cluster based on the percentage of total suicides a cluster has from the selected data. Thus, by looking at the clusters at a glance, the user can quickly identify which cluster has a higher number of suicides. Ideally, a smaller cluster with higher color intensity could potentially have higher risk factors associated that should be analyzed. By hovering over the clusters, the user can see the silhouette score, distance metric and linkage of the model that it belongs to, providing the user to quickly compare scores and parameters of different models.

Upon clicking on a cluster, a bar chart (Figure 4) is displayed to the user where it summarizes the data inside that cluster. The user can also choose to filter the chart based on gender or can see the totals regardless of gender. The main concept behind having this visualization is to help the user find risk factors by inspecting the data within a cluster. Moreover, the profile of the selected cluster is also displayed to the user where they can see the frequency of the each feature's values as the number of instances within that cluster.

Finally, to see the feature interactions within a clustering model, the user can see multiple scatter plots (Figure 5) that display the interaction between each of the features. By observing this figure, users can identify possible correlations between two features, along with identifying if the model is well separated based on any feature. This figure allows the user to find which feature played a major role in the distance measurement for the clustering model. For example, as shown in Figure 5, it can be easily noticed that the Age_group feature completely separates all the 5 clusters. Thus, it can be concluded that the model in Figure 5 was separated based on age groups of the people who committed suicide.

3.4. Project Feedback

From our project demonstration, we received the following recommendations, which were added to our project.

Silhouette Score on Hover: Initially, we had a section in our clustering page which showed the information of the models generated, i.e. their silhouette scores, parameters etc. Based on feedback, it was desired that the model information be showed on hovering over the clusters. We

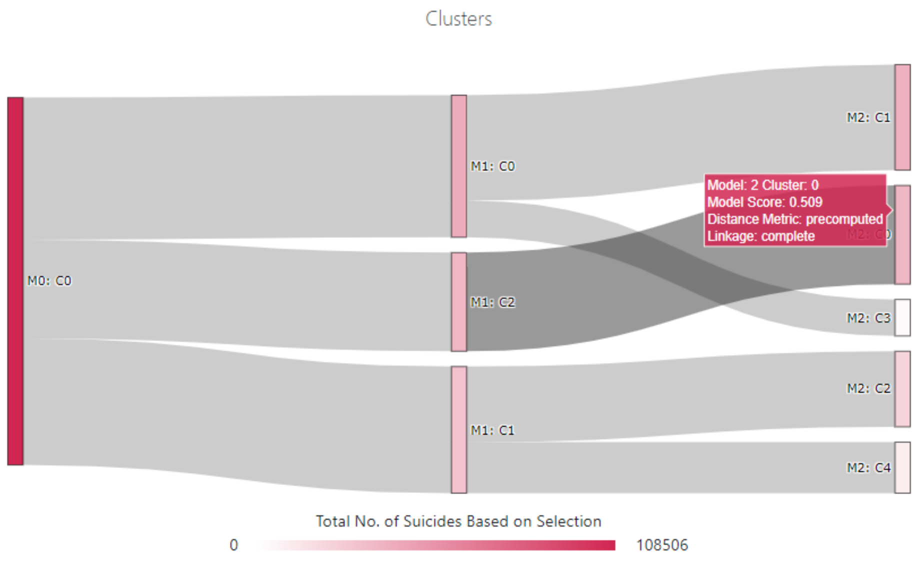


Figure 3. Visualization of the Generated Clustering Models

implemented this feature in our final version of the project as seen in Figure 3.

Suicide Rate: In the interactive map page of the application, the information on the map was only based on the total number of suicides in the states. However, we received a suggestion to consider the portion of suicides in a state with respect to its population to find if suicide rates are different among different states. To incorporate this, we added a drop down menu to switch the data shown in the map between total number of suicides and suicide rates.

Navigation Bar: Our plan for the project included a navigation bar to switch between the interactive map and clustering pages, however, we were unable to implement this before our project demonstration. We added the navigation bar as well as made some other design changes for the final submission.

Remove Cosine Metric: We received feedback that our initial set of available metrics were inappropriate, as we included a Cosine metric. To address the rule of triangle inequality, we excluded the Cosine metric. The metrics that were kept are pre-computed, Euclidean and Manhattan metrics.

Grouped Bar Charts for Gender: Previously, the cluster details bar chart when filtered based on gender was split into two subplots with each facet row showing one gender's data. It was recommended to have male and female bars put beside each other with different color scales for each gender. In our final submission, we have updated the cluster detail bar chart where the data is grouped based on gender with different colors. Along with that, the bars are stacked based on age groups as shown in Figure 4.

In addition to addressing the feedback stated above, we made major changes to the appearance of our application to help different aspects of the application stand out. For instance, all components have been encapsulated with Bootstrap [20] Card elements. The interactive map dashboard page were also changed to have two components side-by-side to avoid scrolling. Some minor design changes were also implemented, i.e. color of graphs, application background etc. Moreover, the visualization of the feature interactions of the models was added after the feedback session, as we considered this would be a nice addition to the project to be able to see how different features are affecting the clustering models.

4. Results

The analysis of the data using our application very much depends on the users and their motivation behind their analysis. However, to justify the usage of our application, we must ourselves be able to make use of it in extracting useful information. This section discusses the results of the analysis we did using our developed application.

4.1. Interactive Map Results

In general, the interactive map page of the application is a tool for making various investigations on the data. As results vary on the type of investigation taken on by the user, this section will briefly cover a few examples of the types of investigations one might make, and the results corresponding to them. These are just a few examples, and many more results and insights can come from this investigation process.

4.1.1. Map Features. The interactive map (Figure 1) provides various results relating to more general data regarding the suicides in India. This is data such as the totals, average total per year, and suicide rates per 100,000 people for each state in India.

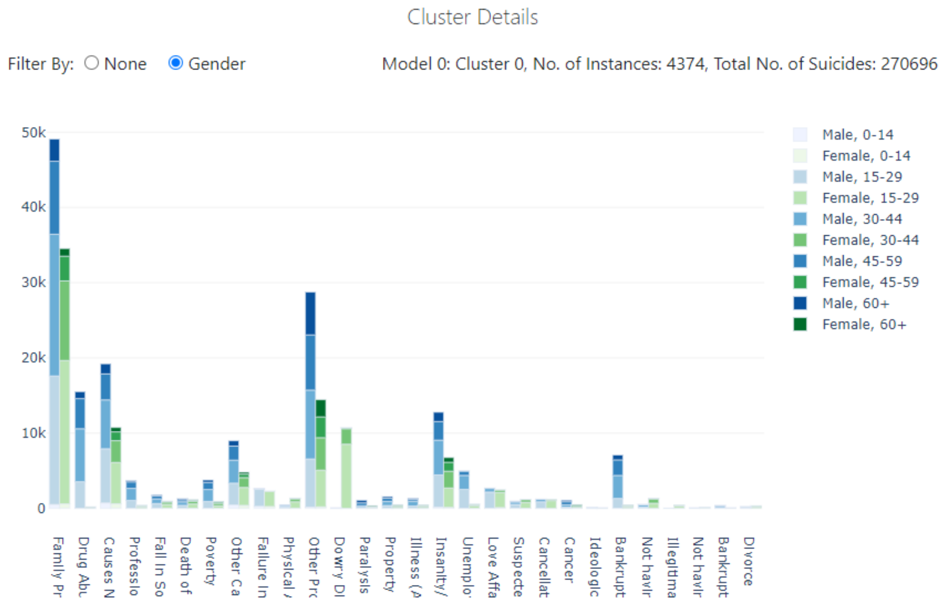


Figure 4. Cluster Details with Gender Filter Applied

There are many insights that can be captured using the hover feature with the range of year slider on the interactive map. If a state displays a high number of suicides over 10 years, but a relatively lower average; that state does not have a uniform distribution of suicides over the years. This indicates that throughout the last ten years, a state may have had a very troubled year. We can delve deeper into this by selecting a smaller range of years in order to pinpoint where that troubled time may have been, and how it may be affecting the current suicide numbers in that state.

Upon analysis of this hover data after selecting 2001 to 2012, most states seem as though they have a uniform distribution of totals throughout the selected years. This is evident because based on our observation, the total suicides are approximately eleven times the average suicide for each state. Despite this, we can still follow an example of how to pinpoint specific numbers using the range of year slider and hover feature of the map. For example, if 2001 and 2012 are first selected on the slider, hovering over Maharashtra will yield the results of 180389 total suicides, with 15032.42 average suicides per year. We can now see how that average changes as we select only more recent years. If we look at the same state, but reduce the range from 2008 to 2012, the average total is now 15329.8. Further, if we reduce the range from just 2010 to 2012, it is shown that the average number of suicides jump to 15991.67. This indicates that the average number of suicides in Maharashtra is increasing in the recent years. Overall, this range slider feature, along with the hover data on the map, allows users to pinpoint certain time periods in order to see suicides throughout the states of India with respect to time.

4.1.2. Trend Figures. The figures shown on the right-hand side of the interactive map page of the dashboard (Figure 2) can be used to gain a lot of insight into the groups of people at risk of suicide throughout India. This section will walk through two examples that show how to use the figures in order to gain these insights. As mentioned, these methods can be generalized to many other examples which users can take on themselves in order to analyze this data.

The first example follows a user who is interested in finding the two most common education status classes which male suicide occurs in those who resided in the Southern states of India. Thus the user chooses the following states: Andhra Pradesh, Goa, Karnataka, Kerala, Lakshadweep, Puducherry, and Tamil Nadu. The trend figure shows that the classes corresponding to “Middle School”, “Primary School”, and “No Education” all range between 7000 and 8000 from the year 2001 to 2012. This data indicates these are the three education status’s that see the most occurrences of male suicide in the selected states. The user can then select a smaller range of states within the subset already selected, perhaps based from suicide rate as shown on the map in order to further refine the data selected. Thus their goal of gaining the sought out insight in a concise visualization is easily achieved; with options available to expand on their insight even more.

Another example of a use case of the trend figures is a comparative investigation between the professions of high suicide totals in different states. Throughout most age groups between 15 and 59 (15-29, 30-44, 45-59), the figures show that the highest suicide totals come from “Farming and Agriculture” related professions for men, and “House Wife” related professions for women. These results are consistent throughout nearly all of the states throughout India. This data taken from the visualization components indicate that these two professions may be key risk group in regards to suicide in India.

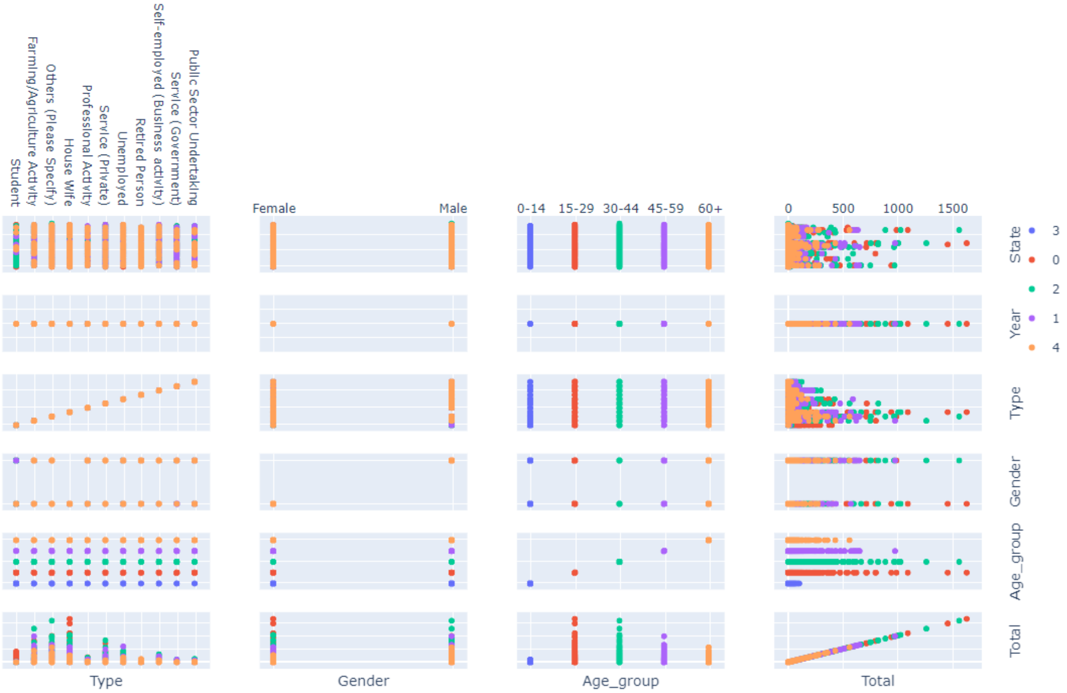


Figure 5. A Portion of Feature Interactions for a Selected Model

Overall, the trend figures shown on the right-hand side of the interactive map page provides users a tool to conduct investigations and analysis to gain insight into the suicide data around India. These investigations can be conducted either by keeping the data very general or honing into very specific demographic groups throughout the states of India.

4.2. Clustering Results

Based on how we implemented the machine learning part of our project, using the clustering models to extract useful information is very much up to the users and their expectations from the model. The application offers various inputs that the user can choose from, and the results of the models rely on it. However, based on our observation, there are certain characteristics of the clusters generated based on the inputs selected. For example, when analyzing a certain state's data for all the years, the clusters seemed to have been formed based on the "Year" column of the data, where each cluster holds mostly data from different years. In contrary, if all the state's data is selected for a specific year, the clusters seem to be separated based on the "Age_group" column of the data. However, these characteristics are expected, i.e. if a state is selected, the clusters need to find separation based on some feature, in this case the feature is the Year. As an example of the usage of our application, we analyzed data of all states for each year from 2001 to 2012 individually to find risk factors associated with suicides in India.

For the analysis, we used pre-computed distance metric as it is more appropriate than only using Euclidean or Manhattan metric due to the nature of our data. We generated 5 clusters as it gave the highest score, especially since all clusters were basically divided based on the 5 age groups. By analyzing the clusters, it was observed that until age 29 the male to female ratio in number of suicides is near to 1, however, for 30+ years of age, the number of suicides in males are higher than in females. To support this, in 2012 between age 30-44 there were twice as many male suicides than females. Analyzing the professional profiles, most people who took their own lives were students for age 0-14, from age 15-44 the most common profession profile was housewife, while for age above 45 it was professions related to farming/agricultural activities. Hanging was predominantly the most common means of committing suicide for all age groups; however, consuming insecticides or poison was also a common mean adopted by people aged between 0-14 years. In terms of gender differences, self-immolation was a noticeable mean of committing suicide among females. Considering the education and social status of the people who took their own lives, the data we collected did not include separation of different age groups. Nevertheless, by observing the age independent data, it was observed that most people who committed suicide were married, and in terms of education, most number of suicides were associated with people having no education or were educated until middle school.

By inspecting the causes of suicide, there were significant differences in different age groups. For people of 0-14 years of age, failure in examination was the most common cause until 2008. Although after that year until 2012 the most common cause was family problems; suicides for failure in examination had similar numbers. From age

15-59, family problems was the main cause for committing suicide based on our findings. For people aged 60 years or older, prolonged illness was the cause for the greatest number of suicides. Moreover, suicides caused due to dowry related problems were significant among females compared to males, while drug addiction was reason for suicides for more males compared to females. It should be noted that for the data associated with causes, there were many entries titled “Other Causes” which had the most number of suicides for all the clusters. We considered the cause with most number of suicides after these entries as the most common cause. This is because ”Other Causes” does not hold much useful information apart from reflecting that there is lacking in properly filing suicide reports in the source of our data.

Considering the findings above, it is clear that suicides are more social factor dependant than other factors such as mental health issues in India. This is very unfortunate considering many of the causes such as failure in examination or family problems that lead to most suicides in India is very much unexpected. This calls for proper investigation as to how these social factors are affecting the lives of people in such a negative way, which leads to such an undesirable outcome of so many lives.

5. Discussion

The project was a fascinating learning experience- combination of machine learning tools with visualization techniques provided a powerful platform to make sense of data. Though the data used for the application was very diverse, the raw data itself was very difficult to analyze and draw useful conclusions from. We were able to identify many useful findings by using our developed tool to assess the suicide related issues in India. This section will discuss our findings in light of our learning process through the development of this project.

5.1. Importance of Proper Representation of Data

The interactive map visualization provided the opportunity to understand the problem of suicides in different states at a glance. The trend figures would provide additional aspects to compliment the map in order to tell a more complete story of the suicide situation in India. Throughout development, various features have been implemented into both the interactive map and trend figures. By switching between total number of suicides and suicide rates in the interactive map, users are able to identify regions which are vulnerable to suicides than other. The regions which are deemed as vulnerable should not be chosen due to the bias of magnitude of population. By displaying suicide rate per 100,000 people, we allow those states to be noticed who may have a low population but a high suicide rate. This change in data representation totally changes the outlook of the problem. The data in the trend figures show that professions relating to farming and agriculture typically have a high amount of suicides. This may indicate that the amount of suicides may be heavily weighted towards more rural areas within a state, with a lower amount of suicides occurring in the cities. Currently, it is not possible to see from our map visualization which states are rural and which are urban, and representing this data further could surface more useful findings. For now, this is simply a hypothesis, and an investigation into it would be possible given more specific data, and a more detailed map. The trends figures themselves provide an elegant visualization of high dimensional data. To analyze the raw data, a user would have to sort the data by state, year, type, and finally by gender and age to even have a close idea of what these trends show. Visual tools such as these are very powerful as they allow for users to gain insight into the data, while saving them the time and confusion of sorting through the raw data.

5.2. Your Model is Only as Good as Your Data

The main potential limitation of our application is the level of detail at which our data is available. A useful aspect of this application is that the user observing suicide data can observe at the level they want to. These levels range from taking a glance at the distribution of suicides throughout the states of India to delving into the common professions of victims in the state of Odisha, for example. These levels of observations allow for multiple perspectives of the data to be shown. Currently, those levels are deepest only at the type of category chosen, age groups and genders; no further detail is provided in the data set. In the most ideal case, a user would be able to analyse many other details, perhaps close to an individual level, in order to get a deeper perspective into the factors that makes someone at risk for suicide in India. These details could include mental health history, family history, relationships, and more. For example, family problems was a leading cause of suicides, however, there may be many interpersonal issues due to family problems that lead to depression, anxiety, or other such mental health issues, which might have caused an individual to take their own life. Overall this application is most limited when it is unable to contribute to offering a deep analysis of the victims of suicide throughout the states in India. Also, there are a few entries in the data throughout each type with names such as ”Other” or ”Unknown”. The overall perspective on the suicide situation in India would be made better, as the current data may be overlooking key insights which are classified under these titles. Unfortunately, this ideal may be far from reality as many victims of suicide leave many mysteries behind, such as the causes of the suicide and identity of the individual.

5.3. Clustering as a Powerful Tool

The clustering models made analyzing groups of data very easy, especially if a specific goal was set in mind. For our case, wanting to see all states’ data for different years and the associated characteristics of suicides was successful; as we observed the clusters were being separated by different age groups. This helped us analyze the effect of age groups in suicides, along with the help of visualization technique to separate male and female suicides proved to be very helpful for information extraction. Analyzing each cluster surfaced findings specific to each age group, which pictures the discrepancies in suicide cases in India for children, adults and older people. Our findings from the data using our developed tool has been discussed in Section 4.2, which can help taking preventive measures to mitigate the problem of suicides in India. It was evident that most suicides were related to socio-economic factors such as, family problems, failure in examination or dowry dispute, which reflects the social constraints of India playing a role in its suicide problem. The number of housewives who committed suicide is also alarming, which requests further investigation whether dependence on spouses are leading to such unfortunate outcomes. Moreover, considering many number of suicides are associated with people with little to no education, this emphasizes the importance of establishing proper education across the country. By using this clustering tool, experts can explore different perspectives of the suicide related issues in India, our take on analyzing different age groups was merely one of the many possibilities.

5.4. Being Aware of Misinterpretations

Although we could extract useful information from the clustering models, the implementation might not always derive desired outcomes. As discussed in Section 4.2, when selecting all states and a single year, the clusters were divided on age groups. We inspected such clusters for our purpose, however the user might be interested in division based on for example total number of suicides. Following the model's decision, it might seem that splitting the data based on age groups is the right way to look at the suicide related data in India. However, this was only due to the inputs selected by the user, given different inputs, the clusters would be different and also the analysis. There are different ways to look into a problem, fixating on the tool's decision may lead to misinterpretations. Our current implementation relies heavily on the inputs selected by the user such as, state or year. To help users understand the model's performance, we have included the model feature interactions graph as show in Figure 5, so the users can understand the reasoning behind the clusters generated. Thus, we encourage the users to make use of these features of our application to avoid any faulty analysis. Additionally, we could provide an option for the user to exclude certain features from being considered in the distance calculation, that way the user could for example, remove the age group bias and focus on other features for the model. Although our clustering tool offers much room for customization in terms of setting input for the models, this addition of selecting specific features for distance calculation would offer new possibilities to the user.

As discussed throughout the report, we set out to provide researchers an application to those researching suicide data. The dashboard page grants users the ability to focus in on areas where suicide is especially a problem. In addition to that, the clustering part of our application could benefit the researchers in finding risk factors based on their selection of data and clustering method parameters. Overall, given a finite amount of resources to spread awareness and prevent suicide, this application provides a method of identifying groups that are at risk of suicide throughout the states of India.

6. Future Work

We were able to develop a tool to analyze suicide data in India, along with using the tool to identify certain risk factors associated with this problem. However, there are rooms for improvement in our project. The clustering model as discussed in Section 5 was generating models that were reliant on specific features, and there was no scope for the user to control this tendency of the models. To address this issue, we wish to implement a feature to let the user determine which attributes of the data should take part in measuring the distance between the generated clusters. Also, the Sankey diagram provided an intuitive visualization of the changes in clusters as different parameters are set for the clustering model, however, our current application does not allow the user to observe which instances are moving from one cluster to another for different models. This would be a useful feature to have if the user could select and investigate such instances. For example, the user might want to know which instances from one cluster of a model are getting separated into building a new cluster as the number of clusters are increased.

Although the total number of suicides and suicide rate per 100,000 for each state gives a quick understanding of the severity of the problem on the interactive map, it would be desirable to learn additional information about each state. This includes information such as whether the states are mainly urban or rural, average income of the people living in a state, literacy rate, and various other factors. This could potentially help find issues that need further attention within the state.

More general future work regarding this dashboard is related to both promotion and expansion of the application. There must be promotion and distribution to the experts for this application to reach its full potential. This process may include market research, additional development iterations, and more testing for the application to reach a benchmark level as intended.

There is also a lot of room for expansion in the application. We intend to create dashboards for other countries around the world in order to provide this tool for not just India. This would include further development to the aspects of this application to fit for the larger scale of data. Suicide is an increasingly common problem everywhere, and a tool like this to aid users in visualizing the data behind that problem is vital for its prevention. However, finding a comprehensive solution to the suicide problem in India is very difficult to achieve given the scope and time of our project. We believe the platform that we have developed would serve as a starting point in adopting a data-driven approach to combat this issue with numerous possibilities for improvements and real-life application.

7. Conclusion

Addressing the issues with suicide is a complex task, with numerous aspects to consider as well as being as cautious as possible. The very reason that there needs to be studies conducted on this issue is unfortunate. The factors related to suicide cannot be dealt with overnight, there must be introduction of interventions that are specific to the targeted population; that supports the economic, social, and geographical characteristics that may be influencing the problem. Our application offers a data-driven approach in dealing with the issues related to suicides in India. In the right hands it has the potential to surface action items that can be adopted by the authorities as steppingstones in battling this prevailing issue. In conclusion, we believe our developed application provided us with a very profound learning experience as well as a platform that has the potential to aid in solving a real-life problem.

References

[1] "Suicide. world health organization." 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide

[2] R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal, S. Y. Ahn *et al.*, "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010," *The lancet*, vol. 380, no. 9859, pp. 2095–2128, 2012.

[3] A. M. Mokhtari, S. Sahraian, S. Hassanipour, A. Baseri, and A. Mirahmadizadeh, "The epidemiology of suicide in the elderly population in southern iran, 2011–2016," *Asian journal of psychiatry*, vol. 44, pp. 90–94, 2019.

[4] R. Rappai, A. V. Cherian, A. Lukose, and L. Vijayakumar, "Suicide research in india: an overview of four decades," *Asian Journal of Psychiatry*, p. 102191, 2020.

[5] D. Knipe, A. J. Williams, S. Hannam-Swain, S. Upton, K. Brown, P. Bandara, S.-S. Chang, and N. Kapur, "Psychiatric morbidity and suicidal behaviour in low-and middle-income countries: A systematic review and meta-analysis," *PLoS medicine*, vol. 16, no. 10, p. e1002905, 2019.

[6] V. Arya, A. Page, D. Gunnell, R. Dandona, H. Mannan, M. Eddleston, and G. Armstrong, "Suicide by hanging is a priority for suicide prevention: Method specific suicide in india (2001–2014)," *Journal of affective disorders*, vol. 257, pp. 1–9, 2019.

[7] R. Dandona, G. A. Kumar, R. Dhaliwal, M. Naghavi, T. Vos, D. Shukla, L. Vijayakumar, G. Gururaj, J. Thakur, A. Ambekar *et al.*, "Gender differentials and state variations in suicide deaths in india: the global burden of disease study 1990–2016," *The Lancet Public Health*, vol. 3, no. 10, pp. e478–e489, 2018.

[8] R. Anil and A. Nadkarni, "Suicide in india: a systematic review," *Shanghai archives of psychiatry*, vol. 26, no. 2, p. 69, 2014.

[9] J. J. Mann, A. Apter, J. Bertolote, A. Beautrais, D. Currier, A. Haas, U. Hegerl, J. Lonnqvist, K. Malone, A. Marusic *et al.*, "Suicide prevention strategies: a systematic review," *Jama*, vol. 294, no. 16, pp. 2064–2074, 2005.

[10] D. Merriott, "Factors associated with the farmer suicide crisis in india," *Journal of epidemiology and global health*, vol. 6, no. 4, pp. 217–227, 2016.

[11] V. Arya, A. Page, G. Armstrong, G. A. Kumar, and R. Dandona, "Estimating patterns in the under-reporting of suicide deaths in india: comparison of administrative data and global burden of disease study estimates, 2005–2015," *J Epidemiol Community Health*, 2020.

[12] "Dash. dash overview." 2020. [Online]. Available: https://plotly.com/dash/

[13] "Python. welcome to pythong.org." 2020. [Online]. Available: https://www.python.org/

[14] "pandas. pandas – python data analysis library." 2020. [Online]. Available: https://pandas.pydata.org/

[15] "scikit-learn. machine learning in python." 2020. [Online]. Available: https://scikit-learn.org/stable/

[16] "Suicides in india. keggle." 2020. [Online]. Available: https://www.kaggle.com/rajanand/suicides-in-india

[17] "Census of india website : Office of the registrar general & census commissioner, india," 2020. [Online]. Available: https://censusindia.gov.in/

[18] "Calculating a rate," 2020. [Online]. Available: https://www.stats.indiana.edu/vitals/CalculatingARate.pdf

[19] "sklearn.metrics.silhouette_score—scikit-learn 0.23.2 documentation," 2020. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

[20] "Bootstrap. build fast, responsive sites with bootstrap." 2020. [Online]. Available: https://getbootstrap.com/