Analysis of Suicide Statistics of India

Achyuta Krishna V, Ajith P J, Alvin Ronnie J, Anirudh G

Birla Institue of Technology and Science, Pilani, Hyderabad Campus, India

{f20180165, f20180040, f20180029, f20180217}@hyderabad.bits-pilani.ac.in

***Abstract*—This paper aims to analyse the data about suicides in India to draw patterns and conclusions on how different factors affect the dynamics of suicide cases in India. This study uses datasets provided by the Indian government. These datasets consists of statewise data of various social, economic and educational factors and its variation with gender and time. Various pre-processing techniques have been applied on the datsets to make it suitable for the data analysis.**

Keywords - Data mining, Preprocessing, Suicide, India, Cluster, Outlier, Classification, plot, DBSCAN, Local Outlier Factor, Decision Tree algorithm.

I. INTRODUCTION

In a world where data is abundant, deriving conclusions can be overwhelming. Data Mining is a tool which can be used to find patterns and relationships in data.

Suicide cases in India are alarmingly high and this raises a serious cause of concern (17% of the suicide victims are from India). In this paper we show how Data mining techniques can be used to analyse suicide statistics of India. In addition, we seek to understand how factors such as education level, social status and gender affect the suicide rates.

We begin the analysis through exploratory data analysis. This involves testing for correlation between features, analysis of central tendencies and plotting relevant graphs. This is followed by data cleaning which involves removal of duplicate records. Data pre-processing is then done through aggregation, encoding and attribute subset selection of data.

Then, Data Mining techniques like Clustering and Outlier Analysis are used to draw insights. This is followed by Classification, which is used to create a model to predict useful information.

II. MOTIVATION

Motivation of this paper lies behind the fact that India records a high number of suicide cases every year and it is barely heeded to in the society. Also, the dataset in OGD platform was not easily comprehensible and not many insights could be drawn from it.

III. OBJECTIVES

This paper intends to clean the dataset and make it comprehensible and to derive patterns and insights like

- The effect of social and economic factors on suicide case and to find the dominant cause of the same.
- The variation of suicide rates across different states and union territories of India.
- The influence of an individual's education status on his tendency to commit suicide.

IV. BACKGROUND

A lot of data analysis has been done on suicide datasets of other countries like Korea and the UK whereas not many studies were found to be done on India's suicide scenario.

V. METHODOLOGY

A. Dataset Description

The datasets that have been used in this study is taken from the Open Government Data (OGD) (OGD, n.d.) (OGD, n.d.) (OGD, n.d.)platform and is merged into a single dataset. The merged dataset consists of six attributes namely – States/UT, Year, Cause, Male, Female, Total.

1) States/UT: It is a nominal attribute that consists of 35 unique values representing the various states and union territories of India. It also consists of three more values which denote total of all states, total of all union territories and total throughout India.

2) Year: It is a discrete interval attribute. Years range from 2001 to 2012.

3) Causes: It is a nominal attribute. Causes include reason for death, marital status and educaltional qualification. It also includes total(of all causes) and total illness(of all illnesses).

4) Male: It is a discrete ratio attribute. Each entry denotes the numbers of males who have comitted suicide corresponding to the cause, year and state.

5) Female: It is a discrete ratio attribute. Each entry denotes the numbers of females who have comitted suicide corresponding to the cause, year and state.

6) Total: It is a discrete ratio attribute. Each entry denoted the sum of males and females who have commited suicide corresponding to the cause,year and state.

B. Exploratory Data Analysis

Analysis of the dataset is done from two perspectives. First, the dataset is grouped by state and plots are obtained between states and their corresponding total number of cases.

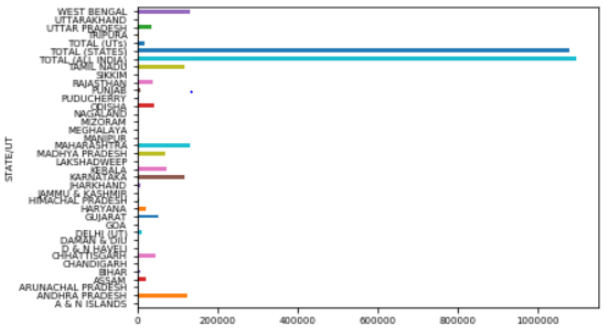

Fig. 1. Bar plot between states and total suicide cases

A visible observation that could be noted is that the dataset has duplicate values as observed from impractically high values for the attribute objects "TOTAL (ALL INDIA)" and "TOTAL (STATES)". Second, the dataset is grouped by year. In this, different line plots are drawn between male, female, total vs year and scatter matrix is obtained among male, female and total. To identify the correlation between variables like male, female and total, the correlation coefficient matrix is obtained for each cause of suicide. Plot is also drawn between the cause and its corresponding total.

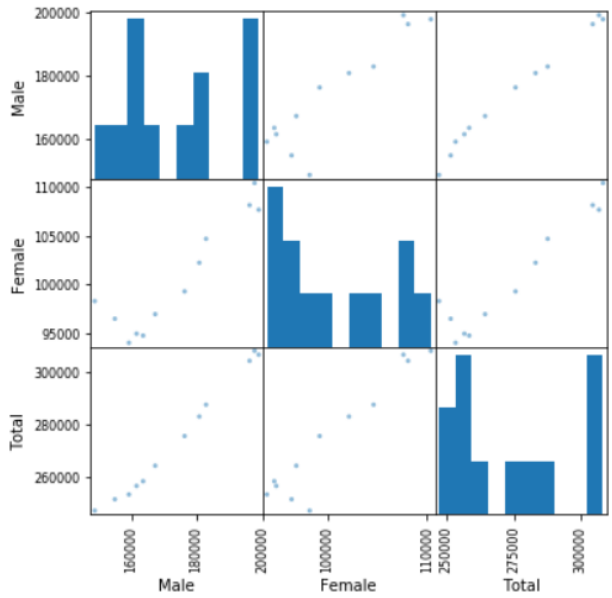

Fig. 2. Scatter matrix among male, female and total

The scatter matrix highlights that Male and Total seem to have a higher positive correlation. Also, some causes are observed to be more correlated to a specific gender. For example: Divorce is observed to have a more effect on female suicides than male suicides.

C. Data Preprocessing

The following preprocessing techniques have been used:

1) Data Cleaning: Redundant atrribute objects namely – "Total"," Total Illness" of the attribute "CAUSE " and "TOTAL (ALL INDIA)", "TOTAL (STATES)" and "TOTAL (UTs)" of the attribute "STATE/UT" have been removed.

2) Aggregation: Aggregation was done separately on two attributes namely- "CAUSE" and "STATE/UT'.The causes were aggregated into three categories which are Physical/Mental Illness, Economic/Social Problems and Other Causes. The attribute objects of "STATE/UT" were aggregated into 6 regions which are North, South, East, West, Central and Northeast *(Administrative Division, n.d.)*.

3) Encoding: Encoding was done on three causes and six regions. The regions were binary encoded which resulted in the binary attributes R1, R2 and R3. The regions were encoded as follows: North (000), West(001), South(010), East(011), Central(100) and Northeast(101). "one hot" encoding is used on the causes which resulted in the binary attributes C1, C2 and C3. The causes were encoded as follows: Physical/Mental Illness(000), Economic/Social Problems(010) and Other Causes(100).

4) Attribute Subset Selection: Attribute Subset Selection was done on both the encoded datasets. The search strategy used is Stepwise Backward Elimination. The objective function that was used to evaluate the candidate attributes is a correlation between them. Correlation is chosen because in exploratory data analysis, one of the observations was that correlation between certain attributes had high positive values. In both the encoded datasets, the pair of male and total had the highest correlation(correlation coefficient>0.98). Between them, total was chosen to be removed.

5) Normalisation: Each state's total male count and female count were used to normalise the corresponding male and female attributes by converting them to percentage value

D. Data Analysis

The following data analysis techniques have been used:

1) Cluster Analysis: Clustering was performed using DBSCAN (Density-based spatial clustering of applications with noise).There are 2 parameters namely: Minpts and ε. First, we begin with an arbitrary point. If there are atleast Minpts number of points in the ε-neighbourhood of this point, then this point becomes a part of a cluster otherwise it is labelled as noise. This process is continued until all the points are processed. The pseudocode for DBSCAN algorithm as implemented here is given below :

```
DBSCAN(D,eps,Minpts)
1.  C=0
2.  For each unvisited point P in dataset D
3.  Mark P as visited
4.  NeighborPts = regionQuery(P,eps)
5.  If sizeof(NeighborPts) < Minpts
6.      Mark P as noise
7.  Else
8.      C = next cluster
9.  expandCluster(P,NeighborPts,C,eps, Minpts)

expandCluster(P,NeighborPts,C,eps,Minpts)
1.  Add P to cluster C
2.  For each point P' in  NeighborPts
3.      If P' is not visited
4.          Mark P' as visited
5.          NeighborPts'=regionQuery(P',eps)
6.          If sizeof(NeighborPts')>=Minpts
```

7. NeighborPts=NeighborPts joined with NeighborPts'
8. If P' is not yet a member of any cluster
9. Add P' to cluster C

regionQuery(P,eps)

1. Return all points with P's eps – neighbourhood (including P)

To find the suitable value for ε, using the "NearestNeighbors",we calculate the distance from each point to its closest neighbour .Two arrays are returned by the method "Kneighbors" .One of the array contains the distance to the closest n_neighbors points and the other contains the index for each of those points. The suitable value for ε is calculated according to the pseudocode given below :

1. For i
2. For j=1 to n
3. d(i,j) ← distance between (xi,xj)
4. Find minimum value of distances to nearest 3
5. End for
6. End for
7. Sort distances in ascending order and plot it
8. Take the "knee" of the curve as the suitable ε value.

2)Outlier Analysis: Outlier analysis was performed on the dataset which resulted after aggregation with respect to regions. "Local Outlier Factor (LOF)", a density based outlier detection technique was used to find the same. This techniques detects outliers as points which have a low relative density compared to its nearest neighbours. It makes use of a concept called Local Reachability Density which is defined as LRD $= \frac{|N_k(o)|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$.

$N_k(o)$ is the set of k-nearest neighbours of o.

$reachdist_k(o' \leftarrow o) = max\{dist_k(o), dist(o,o')\}$.

$dist_k(o)$ is the distance between o and its k[th] nearest neighbour.

$dist(o,o')$ is the Euclidean distance between o and o'.

The LOF of each point is a measure of a point's tendency to be an outlier. Higher the value of LOF, more is the possibility of the given point to be an outlier and vice-versa.

The LOF value is calculated as $LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{LRD_k(o')}{LRD_k(o)}}{|N_k(o)|}$.

3)Classification: Classification was done using Decision Tree algorithm. In this, information gain per feature is calculated using the formula:

$$InforGain(feature_d, D) = Entropy(D) - \sum_{t \in feature} \left(\frac{|feature_d = t|}{|D|} * H(feature_d = t)\right)$$

To grow the decision tree, ID3 algorithm is used. The pseudocode for ID3 is as follows:

ID3(D,Feature_Attributes,Target_Attributes)

1.Create a root node r
2. Set r to the mode target feature value in D
3. If all target feature values are the same:
4. return r
5. Else:
6. pass
7. If Feature_Attributes is empty:
8. return r
9. Else:
10. Att = Attribute from Feature_Attributes with the largest information gain value
11. r = Att
12. For values in Att:
13. Add a new node below r where node_values = (Att == values)
14. Sub_D_values = (Att == values)
15. If Sub_D_values == empty:
16. Add a leaf node l where l equals the mode target value in D
17. Else:
18. Add Sub_Tree with ID3 (Sub_D_values, Feature_Attributes = Feature_Attributes without Att, Target_Attributes)

The feature with largest information gain is then picked and it is assigned as the root node. For each feature value, a branch is grown and unlabelled nodes are added at the end. The dataset is then split along the maximum information gained feature and the feature is removed. The above steps are then done recursively.

VI. RESULTS

Using Exploratory Data Analysis, it was identified that there were duplicate records in the form of "TOTAL (ALL INDIA)", "TOTAL (STATES)" and "TOTAL (UTs)" under the attribute "STATE/UT" and "Total" and "Total Illness" under the attribute "CAUSE". These are removed as a part of data cleaning.

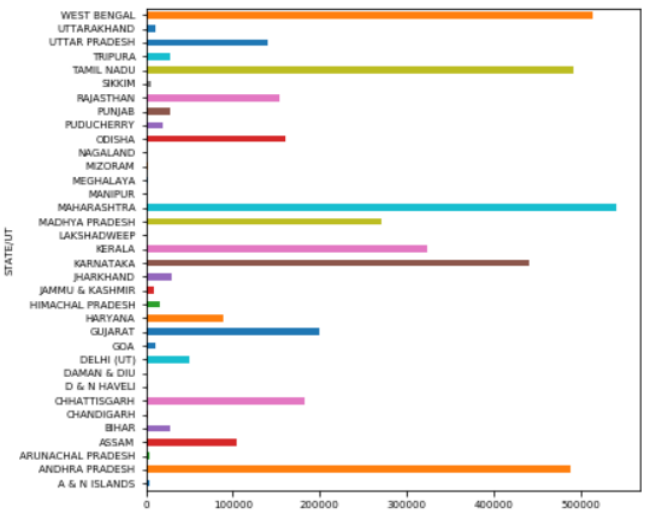

Fig. 3. Bar plot between states and total suicide cases after removing duplicates (*comparing to Fig. 1.)*

Aggregation of the cleaned data resulted in two data sets with reduced sizes – one where 26 causes were aggregated into three categories and the other where 35 were reduced to 6 regions. While aggregation, keeping educational

qualification and marital status lead to duplicates. Hence such records are removed and separate datasets for educational and marital status are created.

	STATE/UT	Year	CAUSE	Male	Female	Total
0	A & N ISLANDS	2001	Physical/Mental Illness	27	19	46.0
0	A & N ISLANDS	2001	Economic/Social Problem	14	9	23.0
0	A & N ISLANDS	2001	Other Causes	38	22	60.0
1	A & N ISLANDS	2002	Physical/Mental Illness	33	16	49.0
1	A & N ISLANDS	2002	Economic/Social Problem	18	18	36.0
1	A & N ISLANDS	2002	Other Causes	40	19	59.0
2	A & N ISLANDS	2003	Physical/Mental Illness	19	13	32.0

Fig. 4.a Snippet of the data set after aggregation w.r.t cause

	CAUSE	Year	Region	Male	Female	Total
0	Bankruptcy or Sudden change in Economic Status	2001	NORTH INDIA	137	38	175.0
0	Bankruptcy or Sudden change in Economic Status	2001	WEST INDIA	383	40	423.0
0	Bankruptcy or Sudden change in Economic Status	2001	SOUTH INDIA	1655	324	1979.0
0	Bankruptcy or Sudden change in Economic Status	2001	EAST INDIA	98	45	143.0
0	Bankruptcy or Sudden change in Economic Status	2001	CENTRAL INDIA	132	40	172.0
0	Bankruptcy or Sudden change in Economic Status	2001	NORTH EAST INDIA	3	4	7.0

Fig. 4.b. Snippet of the data set after aggregation w.r.t region

The encoding of the regions and the causes led to the conversion of nominal attributes to binary attributes.

	CAUSE	Year	R1	R2	R3	Male	Female	Total
0	Bankruptcy or Sudden change in Economic Status	2001	0	0	0	107	31	138
0	Bankruptcy or Sudden change in Economic Status	2001	0	0	1	383	40	423
0	Bankruptcy or Sudden change in Economic Status	2001	0	1	0	1655	324	1979
0	Bankruptcy or Sudden change in Economic Status	2001	0	1	1	98	45	143
0	Bankruptcy or Sudden change in Economic Status	2001	1	0	0	132	40	172
0	Bankruptcy or Sudden change in Economic Status	2001	1	0	1	3	4	7
0	Bankruptcy or Sudden change in Economic Status	2001	1	1	0	30	7	37

Fig. 5.a. Encoding of regions *(binary encoding)*

	State/UT	Year	C1	C2	C3	Male	Female	Total
0	A & N ISLANDS	2001	0	0	1	27	19	46
0	A & N ISLANDS	2001	0	1	0	14	9	23
0	A & N ISLANDS	2001	1	0	0	38	22	60
1	A & N ISLANDS	2002	0	0	1	33	16	49

Fig 5.b. Encoding of causes (*one-hot encoding)*

Additional finding of exploratory data analysis was that few attributes like male and total had a highly positive correlation value (0.994435). Hence the correlation between attributes was used as the objective function for attribute subset selection. This resulted in a dataset where the attribute "Total" was removed.

	State/UT	Year	C1	C2	C3	Male	Female
0	A & N ISLANDS	2001	0	0	1	27	19
0	A & N ISLANDS	2001	0	1	0	14	9
0	A & N ISLANDS	2001	1	0	0	38	22
1	A & N ISLANDS	2002	0	0	1	33	16
1	A & N ISLANDS	2002	0	1	0	18	18

Fig. 6. Snippet of dataset after removing total by doing attribute subset selection

The cause wise encoded dataset was then normalized on each state's corresponding male and female count.

Index	STATE/UT	Year	C1	C2	C3	Male	Female
0	A & N ISLANDS	2001	0	0	1	34.177215	38.000000
0	A & N ISLANDS	2001	0	1	0	17.721519	18.000000
0	A & N ISLANDS	2001	1	0	0	48.101266	44.000000
1	A & N ISLANDS	2002	0	0	1	36.263736	30.188679
1	A & N ISLANDS	2002	0	1	0	19.780220	33.962264

Fig. 7. Snippet of dataset after normalization

After DBSCAN, 6 clusters were identified out of which 3 were prominent and identified with the cluster label 1,2 and 5. Out of 1260 points, cluster 1 has 245 points, cluster 2 has 225 points, cluster 5 has 151 points and there are 43 points identified as noise.

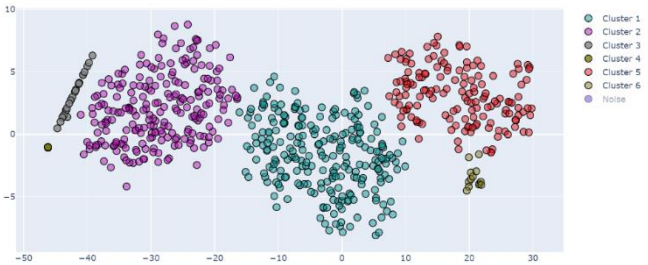

Fig. 8.a. Plot of data after DBSCAN.

On analysing the difference and ratio of number of male and female cases, and drawing plots for the 3 clusters, the following results are obtained :

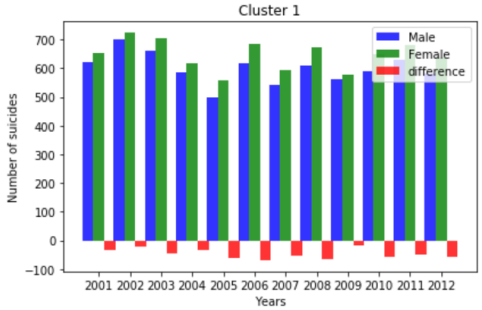

Fig. 8.b. Bar plot of Cluster 1.

Cluster 1 has 47 more female cases than male cases on average. Suicides due to Economic/Social Problems,

impotency, illegitimate pregnancy and physical abuse are greater than the numbers for Physical/Mental illness. As a result of higher number of female suicides (Cluster 1), these causes can be concluded as the primary cause for female suicides.

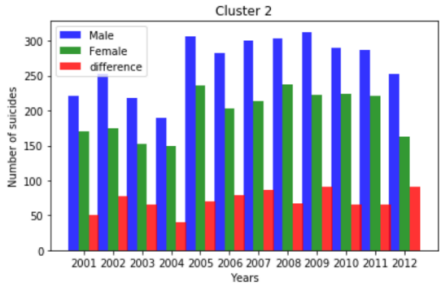

Fig. 8.c. Bar plot of Cluster 2.

Cluster 2 has 71 more male cases than female cases on average. The causes for most of the suicides in this cluster are impotency, failure in Examination, ideological Causes and Hero Worshipping. Hence, these causes can be taken as the primary cause for male suicides.

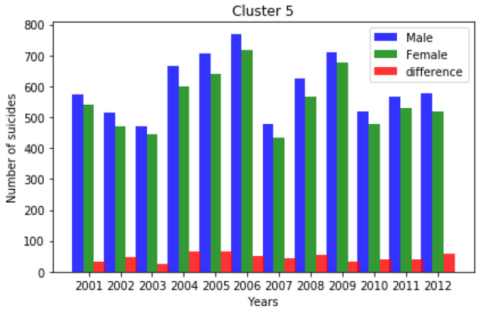

Fig. 8.d. Bar plot of Cluster 5.

Cluster 5 has 46 more male cases than female case on average. Most of suicides are due to Physical/Mental Illness. There are very few cases with other cause for suicide for males. It can be inferred that although Physical/Mental Illness affects both male and female almost equally, its impact is seen to be slightly higher in male.

The algorithm was applied to subset_selection_cause.xlsx which was the dataset obtained after aggregating and encoding with respect to causes. The value of k was chosen to be 25 and the threshold outlier score was chosen to be 1.75. The points obtained as outliers were analyzed and it was found that most of the points with high values of both male and female suicides corresponded to the state of 'MADHYA PRADESH'. Moreover, all the other points obtained had very low values and hence were not very significant.

	Index	State/UT	Year	C1	C2	C3	Male	Female	Outlier
82	27	ARUNACHAL PRADESH	2004	0	1	0	0	4	2.281180
169	56	BIHAR	2009	0	1	0	143	250	2.128948
211	70	CHANDIGARH	2011	0	1	0	5	18	1.943653
271	90	D & N HAVELI	2007	0	1	0	17	25	1.822839
274	91	D & N HAVELI	2008	0	1	0	10	20	1.791194
508	169	JAMMU & KASHMIR	2002	0	1	0	19	38	1.855554
514	171	JAMMU & KASHMIR	2004	0	1	0	6	20	2.011580
688	229	MADHYA PRADESH	2002	0	1	0	1406	1727	1.776889
691	230	MADHYA PRADESH	2003	0	1	0	1462	1745	1.841240
706	235	MADHYA PRADESH	2008	0	1	0	1703	1780	1.841226
709	236	MADHYA PRADESH	2009	0	1	0	1551	1744	1.850450
712	237	MADHYA PRADESH	2010	0	1	0	1569	1955	2.049320
715	238	MADHYA PRADESH	2011	0	1	0	1480	1875	1.986413
787	262	MANIPUR	2011	0	1	0	0	9	2.390989
974	324	PUNJAB	2001	1	0	0	50	114	2.099389
1054	351	SIKKIM	2004	0	1	0	7	23	2.220931
1168	389	UTTAR PRADESH	2006	0	1	0	590	736	1.792551
1240	413	WEST BENGAL	2006	0	1	0	5793	4833	2.094244

Fig. 9.a. Outlier points obtained after applying LOF to subset_selection_cause.xlsx

Hence for further analysis, the points corresponding to 'MADHYA PRADESH' and 'Economic/Social problems'(the cause that occurred with high frequency among the outlier points) were selected from the original dataset before aggregation and outlier analysis was performed on them. The major conclusions that were drawn out of this analysis was that within MADHYA PRADESH 1) The points got divided into 3 visible clusters 2) 'Dowry Dispute' stood out as the only cause that had more number of female suicides than male suicides 3) 'Family Problems' was a notable cause which had unusually high values of both male and female suicides.

Outlier analysis was also performed on subset_selection_regions.xlsx which was the dataset obtained after aggregating and encoding with respect to regions. The value of k was chosen to be 20 and threshold outlier score to be 1.75. The resulting outlier points were mainly from 3 'CAUSES' 1) Dowry Dispute(East and Central India) 2) Family Problems(South India) 3) Causes not known(East India).

	Index	CAUSE	Year	R1	R2	R3	Male	Female	Outlier
255	36	Causes Not known	2001	0	1	1	2083	1741	1.833408
311	44	Causes Not known	2009	0	1	1	2869	1480	2.145365
318	45	Causes Not known	2010	0	1	1	3055	1676	2.359396
319	45	Causes Not known	2010	1	0	0	3015	1569	2.506156
325	46	Causes Not known	2011	0	1	1	3214	1852	1.958993
717	102	Dowry Dispute	2007	0	1	1	12	1044	2.550677
731	104	Dowry Dispute	2009	0	1	1	17	858	1.856192
738	105	Dowry Dispute	2010	0	1	1	19	840	1.801872
739	105	Dowry Dispute	2010	1	0	0	7	839	1.799319
745	106	Dowry Dispute	2011	0	1	1	2	950	2.177820
1052	150	Family Problems	2007	0	1	0	8608	4244	1.860841
1073	153	Family Problems	2010	0	1	0	9298	4773	2.243982
1080	154	Family Problems	2011	0	1	0	9184	4839	2.207673
1087	155	Family Problems	2012	0	1	0	9317	4701	2.239623
2254	322	Other Causes (Please Specify)	2011	0	0	0	2727	948	2.235669

Fig. 9.b. Outlier points obtained after applying LOF to subset_selection_regions.xlsx

On further examining the points corresponding to these states and causes(taken from the original dataset before aggregation), it was inferred that:
1) The states of Madhya Pradesh and West Bengal have very high number of female suicides as a result of 'Dowry Dispute' compared to any other state.

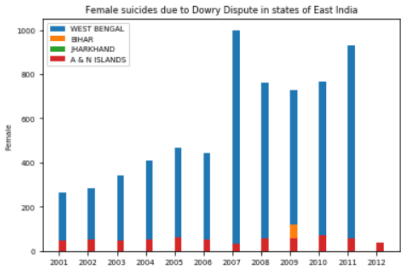

Fig. 9.c. Female suicides due to Dowry Dispute in East India

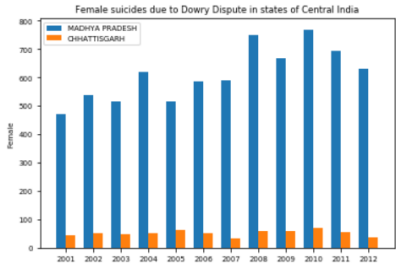

Fig. 9.d. Female suicides due to Dowry Dispute in East India

2) 'Family Problems' leads to a lot of male and female suicides in all the 4 major South Indian states(Andhra Pradesh, Tamil Nadu, Karnataka, Kerala). However Tamil Nadu was found to have unusually high number of suicide cases.

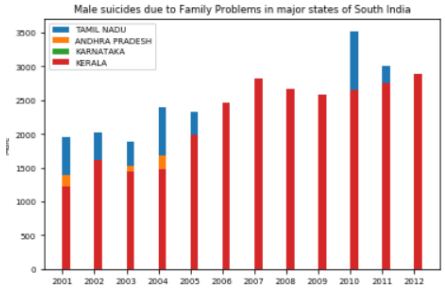

Fig. 9.e. Male suicides due to Family Problems in East India

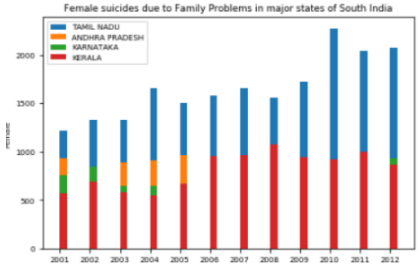

Fig. 9.f. Female suicides due to Family Problems in East India

3) The state of West Bengal also has a large number of suicide cases (both male and female) as a result unknown causes (data points coming under 'Causes Unknown).

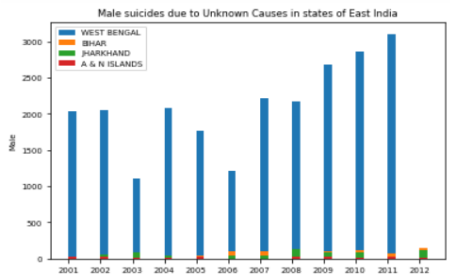

Fig. 9.g. Male suicides due to Unknown Causes in East India

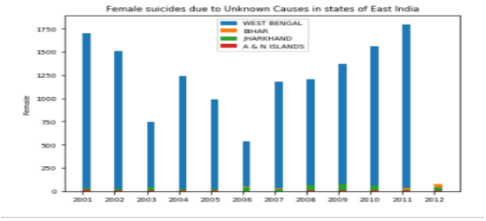

Fig. 9.h. Female suicides due to Unknown Causes in East India

Classification was done on the cleaned dataset which was then aggregated and encoded for ease. The objective was to predict the number of male/female suicide cases given the cause, state, year and total number of cases. The accuracy of prediction was calculated which is as follows:

```
The prediction accuracy is:  94.11764705882352 %
```

VII. BIBLIOGRAPHY

Administrative Division. (n.d.). Retrieved from censusindia: https://censusindia.gov.in/Census_And_You/Administrative_division.aspx

OGD. (n.d.). Retrieved from https://data.gov.in/catalog/stateut-wise-distribution-suicides-causes?filters%5Bfield_catalog_reference%5D=91648&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc

OGD. (n.d.). Retrieved from https://data.gov.in/catalog/stateut-wise-social-status-suicide-victims?filters%5Bfield_catalog_reference%5D=91707&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc

OGD. (n.d.). Retrieved from https://data.gov.in/resources/stateut-wise-educational-status-suicide-victim-during-2001-2012