

MACHINE LEARNING

IAPS: INVESTIGATION ANALYSIS TO PREDICT SUICIDE

Mehak Beri | mxb166430@utdallas.edu

Manish Biyani | mxb172930@utdallas.edu

INDEX

SR. NO	TITLE	PAGE NO.
1	Introduction	3
2	Problem definition and algorithm	4
3	Experimental evaluation	6
4	Related work	18
5	Future work	20
6	Conclusion	21
7	Bibliography	22

• Introduction

- The idea is to create a software solution with police officials and the investigating team of a death scenario as audience.
- The application shall ask a series of questions to the officers and help them determine the probability of the cause of death being suicide or not, in suspicious cases, based on a model trained on data fed into the application backend.
- The data used can be found [here](#) and covers the Indian subcontinent only.
- The model is created by considering factors which might push a person to the brink of suicide. When a police officer enters data of current scene of investigation into the application, he/she will get to know the prediction made by the application.
- We intend to create this application as an effort to reduce time and money expenditure on determining the true cause of death in suspicious death scenarios, on the part of the family of the victim as well as the judiciary system.
- We intend to train the model, and make predictions using real time data fed into the application using a combination of methodologies:
 - Data modelled as a Classification problem (using SVM, Neural network, Logistic regression, Decision Trees and Naïve Bayes)
 - Data modelled as a One class classification problem (finding outliers as non-suicide cases)
- The data consists of 237520 rows of data. The file entitled “CSV2Arff.java” has been used to convert csv file into arff for use in WEKA software.

- Problem Definition and Algorithm

2.1 Task Definition

The task is to predict the probability of a death being a suicide based on the data collected about the victim. The following are the parameters of importance being used to predict the same:

- ✓ Indian State
- ✓ Gender
- ✓ Age group
- ✓ Type-code = [Causes of death, Means adopted, Education status, Professional Profile, Social Status]

The raw data was of the following form:

State	Type_code	Type	Gender	Age_group	Total
A & N Islands	Causes	Illness (Aids/STD)	Female	0-14	0
Andhra Pradesh	Education_Status	Primary	Female	0-100+	1076
Chandigarh	Professional_Profile	Service (Private)	Female	0-14	0

The following is the description of the raw data:

- State: The state of the Indian subcontinent to which the victim belonged
- Type_code: The parameter affecting the suicide
- Type: A subtype of Type_code: eg illness is a subtype of “causes” which could cause a case of suicide
- Gender: self explanatory
- Age group: self explanatory
- Total: total number of suicide cases reported for that particular combination of state, type code, type, gender and age-group

The output shall be the prediction accuracy on this data calculated by 10 cross validation using two methodologies:

- Supervised learning [classification problem]
- Unsupervised learning [one class classification]

This problem is very interesting because it will save judiciary a lot of time by having a machine predict the probability of a case being a suicide. In India the judiciary process is very slow. Our software aims to expedite the process and deliver justice earlier than before.

2.2 Approach Description

In order to solve this problem, we took the following approach:

1) Data Modelling

We modelled the data so as to extract the maximum information from the data. We got rid of the “year” attribute in the original data present at Kaggle.

We added a column in the data called “Target” which gives the class as “Yes” – there was a suicide in this category, or “No” there was no suicide in this combination of factors. This value has been filled depending on the value of “total” column. If the total column value is 0, then the target is “No”, else it is “Yes”. The significance of this is that we are keeping a tab on the combination of parameters for which suicide cases have been recorded in the past.

Then, we split all type codes into 5 separate domain specific files containing Causes of death, Means adopted, Education status, Professional Profile, Social Status respectively. For example, the data for “causes of death” domain looks like:

State	Causes	Gender	Age_group	Total	target
A & N Islands	Illness (Aids/STD)	Female	0-14	0	no
Chhattisgarh	Causes Not known	Female	0-14	13	yes

Similarly, the data for “Educational status” domain looks like:

State	Education_Status	Gender	Age_group	Total	target
A & N Islands	Diploma	Female	0-100+	0	no
A & N Islands	No Education	Female	0-100+	4	yes

2) Conversion to arff files:

With the help of a script written in Java called “CSV2Arff.java”, we convert all the above obtained csv files into arff format. The java files utilizes the WEKA software’s APIs for java for the conversion. The conversion has been done for data processing using the WEKA software.

3) Using the data:

Use the data thus obtained in WEKA software by experimenting with different classifiers as well as Bayesian network creation tools.

- Experimental Evaluation

3.1 Methodology

We have thought of making the predictions for the suicide based on the following two methodologies:

- Supervised learning method:

Assumption: The total number of deaths is taken as a parameter and the classification is merely done by considering number of deaths > 0 as “yes” and less than zero as “No”

We do NOT take the “total” number of suicides as a parameter, but instead take another column called target derived from it and put “yes” and “no” as the predicted class and then use techniques like Support vector machine, neural networks, logistic regression, naïve bayes and decision trees to predict the accuracy of calculation. The accuracy will be measured using 10 fold cross validation.

This shall be done for all 5 domains. Then, a supervised learning technique which gives the best accuracies for the domains is chosen.

We assume that given an unseen data, it will be subjected to the above chosen supervised learning technique and the classification of the case as Suicide or not shall be predicted accordingly.

- Unsupervised learning method:

Observation: the data consists of total number of deaths in each row of data for that combination of parameters, but it does not take into account the total number of “non-suicide” cases. Hence we have to use “one class classifier”

According to Wikipedia: In machine learning, one-class classification, also known as unary classification or class-modelling, tries to identify objects of a specific class amongst all objects, by learning from a training set containing only the objects of that class. This is different from and more difficult than the traditional classification problem, which tries to distinguish between two or more classes with the training set containing objects from all the classes. An example is the classification of the operational status of a nuclear plant as 'normal' In this scenario, there are few, if any, examples of catastrophic system states; only the statistics of normal operation are known. A feature of one-class classification is that it uses only sample points from the assigned class, so that a representative sampling is not strictly required for non-target classes.

We build a Bayesian network and learn the following CPTs:

- (1) $P(\text{Gender})$
- (2) $P(\text{Age Group})$
- (3) $P(\text{State})$
- (4) $P(\text{Causes} \mid \text{Gender, Age group, State})$
- (5) $P(\text{Education} \mid \text{Gender, State})$ — Notice that Education does not depend on age
- (6) $P(\text{Means adopted} \mid \text{Gender, Age group, State})$
- (7) $P(\text{Professional Profile} \mid \text{Gender, Age group, State})$

We estimated the conditional probabilities (CPTs) from data, using 1-laplace correction. For example,

(4) can be estimated using $\text{Count}(\text{causes, Gender, Age, State}) / \text{Count}(\text{Gender, Age, State})$

Once a sample is generated from the Bayesian network, the data will contain the following 7 attributes: Gender, Age Group, State, Causes, Education, Means adopted, Professional Profile.

Then run classifiers on the data thus generated to get accurate prediction percentages.

3.2 Results

Part 1

Supervised learning classification problem (using 10 cross validation)

TYPE_CODE	LR (%)	Naïve Bayes (%)	Decision Tree (%)	Mean (%)
Social status	93.64	93.04	94.42	93.7
Causes	84.54	84.21	88.14	85.63
Education status	93.80	92.68	93.05	93.17667
Professional status	79.06	78.83	87.79	81.89333
Means adopted	85.39	84.96	87.46	85.93667

Looking at the results, we choose: Decision Tree

Based on this chosen classifier, following are the detailed results for each domain:

1) Social Status

=== Summary ===

Correctly Classified Instances	4306	94.4298 %
Incorrectly Classified Instances	254	5.5702 %
Kappa statistic	0.8149	
Mean absolute error	0.0911	
Root mean squared error	0.2169	
Relative absolute error	29.8421 %	
Root relative squared error	55.5156 %	
Total Number of Instances	4560	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.970	0.167	0.962	0.970	0.966	0.815	0.968	0.991	yes
	0.833	0.030	0.865	0.833	0.849	0.815	0.968	0.883	no
Weighted Avg.	0.944	0.141	0.944	0.944	0.944	0.815	0.968	0.971	

=== Confusion Matrix ===

a	b	<-- classified as
3592	111	a = yes
143	714	b = no

2) Causes

=== Summary ===

Correctly Classified Instances	96255	88.1456 %
Incorrectly Classified Instances	12945	11.8544 %
Kappa statistic	0.7393	
Mean absolute error	0.1765	
Root mean squared error	0.3049	
Relative absolute error	38.2172 %	
Root relative squared error	63.4484 %	
Total Number of Instances	109200	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.928	0.200	0.891	0.928	0.909	0.741	0.929	0.948	no
	0.800	0.072	0.862	0.800	0.830	0.741	0.929	0.893	yes
Weighted Avg.	0.881	0.154	0.881	0.881	0.880	0.741	0.929	0.928	

=== Confusion Matrix ===

```
      a      b  <-- classified as
64622  5047 |      a = no
 7898 31633 |      b = yes
```

3) Educational Status

=== Summary ===

Correctly Classified Instances	6789	93.051 %
Incorrectly Classified Instances	507	6.949 %
Kappa statistic	0.7172	
Mean absolute error	0.1022	
Root mean squared error	0.2314	
Relative absolute error	39.3147 %	
Root relative squared error	64.1816 %	
Total Number of Instances	7296	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.707	0.029	0.816	0.707	0.758	0.720	0.941	0.811	no
	0.971	0.293	0.948	0.971	0.959	0.720	0.941	0.986	yes
Weighted Avg.	0.931	0.252	0.928	0.931	0.928	0.720	0.941	0.959	

=== Confusion Matrix ===

```
      a      b  <-- classified as
 792  328 |      a = no
 179 5997 |      b = yes
```

4) Professional Profile

=== Summary ===

```

Correctly Classified Instances      43250          87.7941 %
Incorrectly Classified Instances    6013          12.2059 %
Kappa statistic                    0.7545
Mean absolute error                 0.1781
Root mean squared error            0.3081
Relative absolute error             35.7338 %
Root relative squared error        61.7079 %
Total Number of Instances         49263

```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.908	0.155	0.867	0.908	0.887	0.755	0.936	0.932	no
	0.845	0.092	0.891	0.845	0.867	0.755	0.936	0.932	yes
Weighted Avg.	0.878	0.125	0.879	0.878	0.878	0.755	0.936	0.932	

=== Confusion Matrix ===

```

      a      b  <-- classified as
23571 2401 |      a = no
3612 19679 |      b = yes

```

5) Means Adopted

=== Summary ===

```

Correctly Classified Instances      58776          87.4643 %
Incorrectly Classified Instances    8424          12.5357 %
Kappa statistic                    0.744
Mean absolute error                 0.1813
Root mean squared error            0.3088
Relative absolute error             36.8438 %
Root relative squared error        62.2658 %
Total Number of Instances         67200

```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.905	0.164	0.877	0.905	0.890	0.744	0.937	0.943	no
	0.836	0.095	0.872	0.836	0.853	0.744	0.937	0.924	yes
Weighted Avg.	0.875	0.134	0.875	0.875	0.874	0.744	0.937	0.935	

=== Confusion Matrix ===

```

      a      b  <-- classified as
34253 3610 |      a = no
4814 24523 |      b = yes

```

So on average we obtain the following statistics:

Domain	Relative absolute error	Root relative squared error	Correctly classified instances
--------	-------------------------	-----------------------------	--------------------------------


Social status	29.84	55.51	94.42
Causes	38.21	63.44	88.14
Educational status	39.31	64.18	93.05
Professional status	35.73	61.7	87.79
Means adopted	36.84	62.26	87.46
Average	35.986	61.418	90.172

Part 2

One class classification – unsupervised learning


CPT Tables:

(1) P(Gender)

 Probability Distribution Table For Gender ×


target	Female	Male
no	0.531	0.469
yes	0.458	0.542

(2) P(Age Group)

 Probability Distribution Table For Age_group ×

target	0-14	15-29	30-44	45-59	60+	0-100+
no	0.261	0.15	0.156	0.188	0.231	0.015
yes	0.094	0.245	0.236	0.193	0.135	0.097

(3) P(State)

 Probability Distribution Table For State

target	A & N Islands	Andhra Pradesh	Arunachal...	Assam	Bihar	Chandi...	Chhattis...	D & N H...
no	0.04	0.012	0.042	0.022	0.027	0.042	0.018	0.042
yes	0.013	0.051	0.01	0.038	0.031	0.011	0.043	0.009

(4) P(Causes | Gender, Age group, State)

Probability Distribution Table For Causes

i»çState	Gender	Age_group	target	Illness (Aids/STD)	Bankruptcy ...	Cancell...	Physica...	Dowry Di...	Family Pr...	Ideolog...	Other Prolo...	Property ...
A & N Islands	Female	0-14	no	0.041	0.037	0.041	0.041	0.041	0.031	0.041	0.041	0.041
A & N Islands	Female	0-14	yes	0.015	0.015	0.015	0.015	0.015	0.106	0.015	0.015	0.015
A & N Islands	Female	15-29	no	0.05	0.046	0.046	0.05	0.05	0.026	0.05	0.002	0.046
A & N Islands	Female	15-29	yes	0.005	0.005	0.016	0.005	0.005	0.071	0.005	0.136	0.016
A & N Islands	Female	30-44	no	0.046	0.042	0.042	0.046	0.046	0.028	0.046	0.002	0.046
A & N Islands	Female	30-44	yes	0.007	0.007	0.022	0.007	0.007	0.081	0.007	0.184	0.007
A & N Islands	Female	45-59	no	0.043	0.039	0.043	0.043	0.043	0.029	0.043	0.009	0.039
A & N Islands	Female	45-59	yes	0.01	0.01	0.01	0.01	0.01	0.094	0.01	0.219	0.031
A & N Islands	Female	60+	no	0.041	0.038	0.041	0.041	0.041	0.038	0.041	0.008	0.041
A & N Islands	Female	60+	yes	0.014	0.014	0.014	0.014	0.014	0.042	0.014	0.292	0.014
A & N Islands	Male	0-14	no	0.04	0.037	0.04	0.04	0.04	0.037	0.04	0.037	0.04
A & N Islands	Male	0-14	yes	0.019	0.019	0.019	0.019	0.019	0.056	0.019	0.056	0.019
A & N Islands	Male	15-29	no	0.053	0.04	0.044	0.053	0.053	0.032	0.053	0.002	0.053
A & N Islands	Male	15-29	yes	0.005	0.024	0.024	0.005	0.005	0.053	0.005	0.121	0.005
A & N Islands	Male	30-44	no	0.05	0.046	0.042	0.05	0.05	0.018	0.05	0.002	0.046
A & N Islands	Male	30-44	yes	0.006	0.006	0.028	0.006	0.006	0.096	0.006	0.14	0.017

Randomize Ok Cancel

(5) $P(\text{Education} \mid \text{Gender, State})$ — Notice that Education does not depend on age

Probability Distribution Table For Education_status

target	Gender	i»çState	Diploma	No Education	Post Graduate ...	Middle	Graduate	Hr. Secondary/l...	Primary	Matriculate/Sec...
no	Female	A & N Islands	0.278	0.06	0.298	0.026	0.176	0.096	0.033	0.033
no	Female	Andhra Pradesh	0.097	0.137	0.094	0.143	0.116	0.13	0.142	0.142
no	Female	Arunachal Pradesh	0.278	0.06	0.298	0.026	0.176	0.096	0.033	0.033
no	Female	Assam	0.097	0.137	0.094	0.143	0.116	0.13	0.142	0.142
no	Female	Bihar	0.278	0.06	0.298	0.026	0.176	0.096	0.033	0.033
no	Female	Chandigarh	0.097	0.137	0.094	0.143	0.116	0.13	0.142	0.142
no	Female	Chhattisgarh	0.278	0.06	0.298	0.026	0.176	0.096	0.033	0.033
no	Female	D & N Havelli	0.097	0.137	0.094	0.143	0.116	0.13	0.142	0.142
no	Female	Daman & Diu	0.278	0.06	0.298	0.026	0.176	0.096	0.033	0.033
no	Female	Delhi (Ut)	0.097	0.137	0.094	0.143	0.116	0.13	0.142	0.142
no	Female	Goa	0.278	0.06	0.298	0.026	0.176	0.096	0.033	0.033
no	Female	Gujarat	0.097	0.137	0.094	0.143	0.116	0.13	0.142	0.142
no	Female	Haryana	0.278	0.06	0.298	0.026	0.176	0.096	0.033	0.033
no	Female	Himachal Pradesh	0.097	0.137	0.094	0.143	0.116	0.13	0.142	0.142
no	Female	Jammu & Kashmir	0.278	0.06	0.298	0.026	0.176	0.096	0.033	0.033
no	Female	Jharkhand	0.097	0.137	0.094	0.143	0.116	0.13	0.142	0.142

Randomize Ok Cancel

(6) $P(\text{Means adopted} \mid \text{Gender, Age group, State})$

Probability Distribution Table For Means_adopted

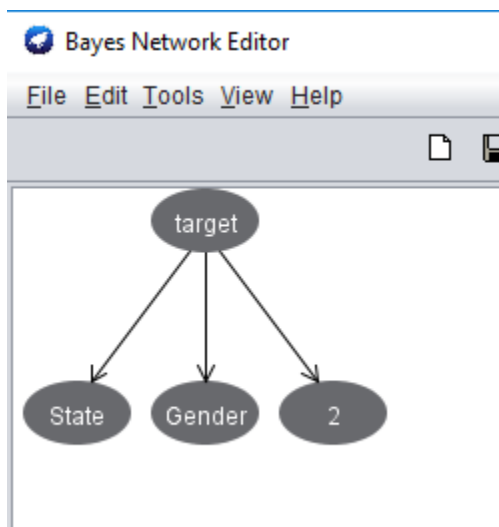
target	Gender	Age_group	i»çState	By Con...	By Han...	By Jum...	By touc...	By Mac...	By Fire...	By Jum...	By Othe...	By Self I...	By Over...	By Con...	By comi...	By Over...	By Jum...	By Dro...	By Fire...	By Othe...
no	Female	0-14	A & N Islands	0.067	0.019	0.067	0.067	0.067	0.057	0.067	0.057	0.067	0.067	0.051	0.067	0.067	0.067	0.067	0.067	0.008
no	Female	0-14	Andhra Pradesh	0.005	0.005	0.124	0.092	0.135	0.005	0.124	0.005	0.103	0.135	0.005	0.016	0.092	0.049	0.005	0.092	0.005
no	Female	0-14	Arunachal Pradesh	0.061	0.024	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.061	0.066	0.066	0.061	0.066	0.008
no	Female	0-14	Assam	0.042	0.004	0.079	0.079	0.094	0.057	0.079	0.011	0.087	0.087	0.004	0.079	0.094	0.094	0.011	0.094	0.004
no	Female	0-14	Bihar	0.032	0.018	0.083	0.09	0.09	0.032	0.09	0.004	0.09	0.09	0.011	0.09	0.083	0.083	0.018	0.09	0.004
no	Female	0-14	Chandigarh	0.065	0.044	0.065	0.065	0.065	0.055	0.065	0.055	0.065	0.065	0.06	0.065	0.065	0.065	0.065	0.065	0.008
no	Female	0-14	Chhattisgarh	0.005	0.005	0.114	0.05	0.114	0.005	0.105	0.005	0.096	0.096	0.005	0.059	0.114	0.105	0.014	0.105	0.005
no	Female	0-14	D & N Havelli	0.063	0.053	0.063	0.063	0.063	0.063	0.063	0.058	0.063	0.063	0.063	0.063	0.063	0.063	0.063	0.063	0.008
no	Female	0-14	Daman & Diu	0.064	0.053	0.064	0.064	0.064	0.059	0.064	0.059	0.064	0.064	0.059	0.064	0.064	0.064	0.064	0.064	0.008
no	Female	0-14	Delhi (Ut)	0.047	0.004	0.069	0.083	0.09	0.025	0.09	0.011	0.076	0.09	0.011	0.076	0.09	0.09	0.061	0.083	0.004
no	Female	0-14	Goa	0.061	0.014	0.072	0.072	0.072	0.055	0.072	0.043	0.072	0.072	0.049	0.072	0.072	0.072	0.043	0.072	0.009
no	Female	0-14	Gujarat	0.004	0.004	0.1	0.108	0.108	0.004	0.091	0.03	0.1	0.108	0.004	0.074	0.091	0.082	0.004	0.082	0.004
no	Female	0-14	Haryana	0.019	0.042	0.095	0.08	0.095	0.042	0.095	0.004	0.087	0.087	0.004	0.049	0.087	0.095	0.019	0.095	0.004
no	Female	0-14	Himachal Pradesh	0.055	0.032	0.072	0.072	0.072	0.061	0.072	0.061	0.072	0.072	0.02	0.072	0.072	0.072	0.043	0.072	0.009
no	Female	0-14	Jammu & Kashmir	0.063	0.052	0.068	0.068	0.068	0.057	0.068	0.035	0.068	0.068	0.046	0.063	0.068	0.068	0.063	0.068	0.008
no	Female	0-14	Jharkhand	0.034	0.034	0.087	0.049	0.094	0.042	0.049	0.004	0.087	0.094	0.049	0.087	0.079	0.087	0.034	0.087	0.004

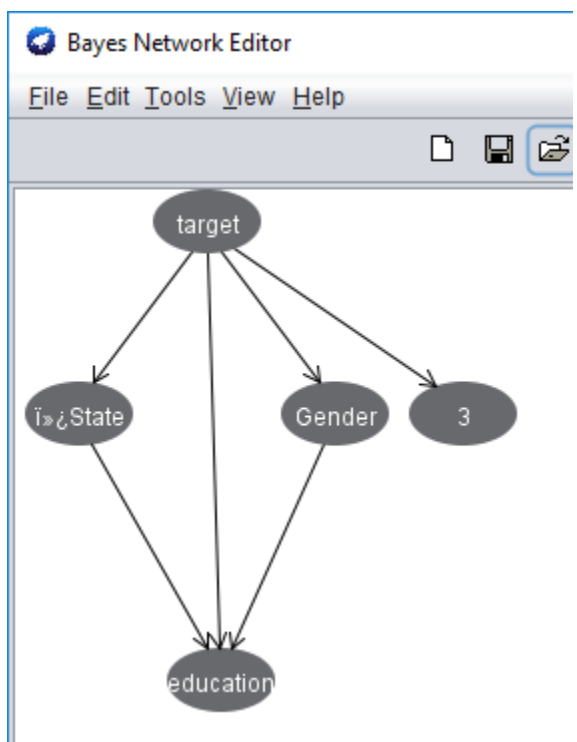
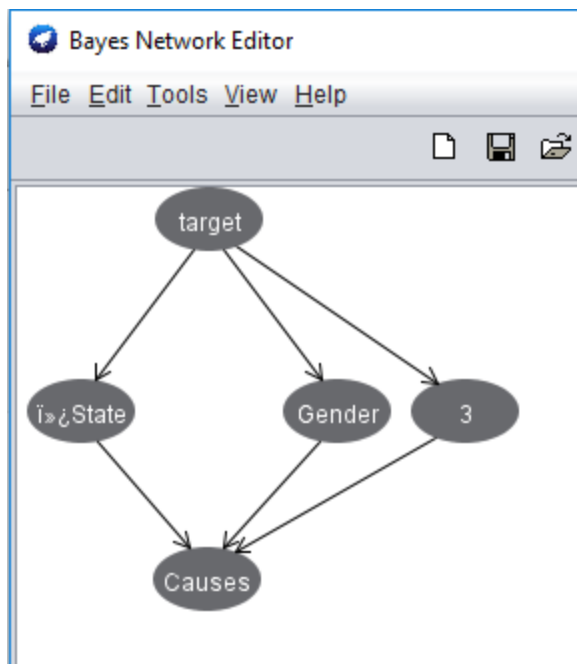
Randomize Ok Cancel

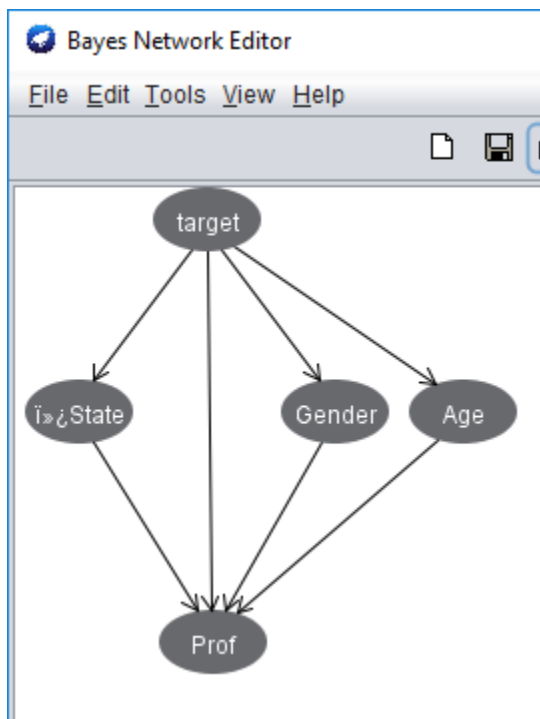
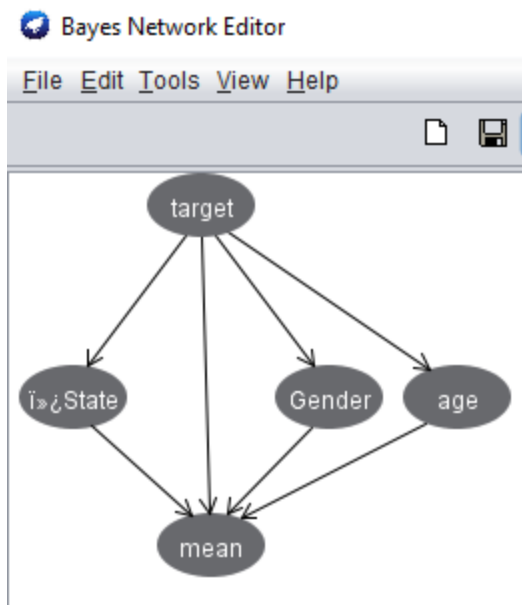
(7) $P(\text{Professional Profile} \mid \text{Gender, Age group, State})$

target	Gender	State	Age_group	Retired P...	Unemplo...	Public Se...	Service (...)	House W...	Self-empl...	Professio...	Student	Others (P...	Farming/...	Service (...)
no	Female	A & N Islands	0-14	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.02	0.098	0.098	0.098
no	Female	A & N Islands	15-29	0.134	0.08	0.134	0.048	0.005	0.123	0.134	0.005	0.123	0.134	0.08
no	Female	A & N Islands	30-44	0.099	0.107	0.107	0.09	0.004	0.107	0.107	0.107	0.107	0.099	0.064
no	Female	A & N Islands	45-59	0.104	0.095	0.104	0.095	0.004	0.104	0.104	0.095	0.104	0.104	0.087
no	Female	A & N Islands	60+	0.101	0.093	0.101	0.093	0.004	0.101	0.101	0.101	0.101	0.101	0.101
no	Female	Andhra Pradesh	0-14	0.159	0.108	0.146	0.121	0.108	0.121	0.07	0.006	0.006	0.006	0.146
no	Female	Andhra Pradesh	15-29	0.697	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
no	Female	Andhra Pradesh	30-44	0.556	0.037	0.037	0.037	0.037	0.037	0.037	0.111	0.037	0.037	0.037
no	Female	Andhra Pradesh	45-59	0.164	0.127	0.018	0.018	0.018	0.055	0.091	0.382	0.018	0.018	0.091
no	Female	Andhra Pradesh	60+	0.047	0.196	0.121	0.047	0.009	0.009	0.103	0.234	0.009	0.009	0.215
no	Female	Arunachal Pradesh	0-14	0.1	0.084	0.1	0.1	0.1	0.1	0.1	0.036	0.092	0.084	0.1
no	Female	Arunachal Pradesh	15-29	0.135	0.081	0.135	0.103	0.005	0.103	0.135	0.005	0.135	0.059	0.103
no	Female	Arunachal Pradesh	30-44	0.113	0.077	0.113	0.068	0.005	0.113	0.113	0.113	0.113	0.095	0.077
no	Female	Arunachal Pradesh	45-59	0.099	0.099	0.099	0.099	0.02	0.099	0.099	0.099	0.091	0.099	0.099
no	Female	Arunachal Pradesh	60+	0.093	0.093	0.093	0.093	0.071	0.093	0.093	0.093	0.093	0.093	0.093

Bayesian Network formation by utilizing the CPT tables formed:







Accuracy achieved using Bayesian Network:

=== Summary ===

Correctly Classified Instances	195214	82.1888 %
Incorrectly Classified Instances	42305	17.8112 %
Kappa statistic	0.6301	
Mean absolute error	0.2809	
Root mean squared error	0.3594	
Relative absolute error	57.3104 %	
Root relative squared error	72.6101 %	
Total Number of Instances	237519	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.897	0.278	0.811	0.897	0.852	0.635	0.905	0.928	no
	0.722	0.103	0.841	0.722	0.777	0.635	0.905	0.866	yes
Weighted Avg.	0.822	0.203	0.824	0.822	0.820	0.635	0.905	0.902	

=== Confusion Matrix ===

a	b	<-- classified as
121550	13931	a = no
28374	73664	b = yes

3.3 Discussion

From the results obtained we see that decision trees perform best on the data, the best being 92%, ignoring the “total” parameter and using it as a single classification as yes or a no.

The accuracies achieved by Bayesian networks are notable at 82% as these are most accurate because the CPTs utilized incorporate in them the relations between different parameters.

- Related Work

1. Predicting Risk of Suicide Attempts Over Time Through Machine Learning

This study overcomes the limitations of accuracy and scale of risk detection for the dangerous behaviors in terms of suicide attempts by applying machine learning to electronic health records within a large medical database.

There were 5,167 adult patients with a claim code for self-injury (i.e., ICD-9, E95x); expert review of records determined that 3,250 patients made a suicide attempt (i.e., cases), and 1,917 patients engaged in self-injury that was nonsuicidal, accidental, or nonverifiable (i.e., controls).

This machine learning algorithm accurately predicted future suicide attempts (AUC = 0.84, precision = 0.79, recall = 0.95, Brier score = 0.14).

Moreover, accuracy had improved from 720 days to 7 days before the suicide attempt, and predictor importance shifted across time.

These findings represent a step toward accurate and scalable risk detection and provide insight into how suicide attempt risk shifts over time.

2. Classification of Suicide Attempts through a Machine Learning Algorithm Based on Multiple Systemic Psychiatric Scales

The purpose of this study was to investigate whether the information from multiple clinical scales has classification power for identifying actual suicide attempts.

Patients with depression and anxiety disorders ($N = 573$) were included, and each participant completed 31 self-report psychiatric scales and questionnaires about their history of suicide attempts.

They trained an artificial neural network classifier with 41 variables (31 psychiatric scales and 10 sociodemographic elements) and ranked the contribution of each variable for the classification of suicide attempts.

To evaluate the clinical applicability of their model, they measured classification performance with top-ranked predictors.

Their model had an overall accuracy of 93.7% in 1-month, 90.8% in 1-year, and 87.4% in lifetime suicide attempts detection.

The area under the receiver operating characteristic curve (AUROC) was the highest for 1-month suicide attempts detection (0.93), followed by lifetime (0.89), and 1-year detection (0.87). Among all variables, the Emotion Regulation Questionnaire had the highest contribution, and the positive and negative characteristics of the scales similarly contributed to classification performance.

Performance on suicide attempts classification was largely maintained when they only used the top five ranked variables for training (AUROC; 1-month, 0.75, 1-year, 0.85, lifetime suicide attempts detection, 0.87).

The findings indicate that information from self-report clinical scales can be useful for the classification of suicide attempts. Based on the reliable performance of the top five predictors alone, this machine learning approach could help clinicians identify high-risk patients in clinical settings.

3. Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms

In this, they have proposed a suicidal ideation detection system, for predicting the suicidal acts using Twitter data that can automatically analyze the sentiments of these tweets.

Then they have investigated a tool of data mining to extract useful information for classification of tweets collected from Twitter based on machine learning classification algorithms.

Experimental results show that their method for detecting the suicidal acts using Twitter data and the machine learning algorithms verify the effectiveness of performance in term of recall, precision and accuracy on sentiment analysis.

Their model accurately predicted future suicide attempts (precision = 75.4%, recall = 72.%, F-measure- 63.8%).

4. Our project

The purpose of our project is to help determine the probability of the cause of death being suicide or not, in suspicious cases, based on a model trained on data fed into the application backend.

The predictions are made using real time data fed into the application using a combination of methodologies:

- Data modelled as a Classification problem (using SVM, Neural network, Logistic regression, Decision Trees and Naïve Bayes)
- Data modelled as a One class classification problem (finding outliers as non-suicide cases)

The data consists of 237520 rows of data.

Our problem doesn't predict the suicide with respect to time but it takes the causes of a person's death and classifies it as a suicide/no suicide.

- Future Work

Currently our model predicts whether the victim has committed suicide or not using the inputs provided. In an extension to the current project, we can predict the timeline in which a person, whose features match with the features /parameters in our database is likely to commit suicide.

Currently we are using Bayesian network and multiple classifiers to predict the occurrence of suicide. In future, we can learn mixtures of Bayesian networks using EM and use it for the same. We can also use the “total” parameter which we have currently excluded from our models as a form of weightage to the combination of each parameter.

We were initially advised to use one class parameters but that required utilization of third party libraries/ code implemented from Github for implementation of one class classifier in weka. In lieu of plagiarism avoidance, we have only utilized functions and libraries freely available with weka. In the future we can work on our own version of a one class classifier to identify outliers on unseen data.

As defined by ‘Learning Mixtures of Bayesian Networks’ in ‘Technical Report MSR-TR-97-30’, Mixtures of Bayesian networks is a heuristic method for learning mixtures of Bayesian Networks (MBNs) from possibly incomplete data. The considered class of models is mixtures in which each mixture component is a Bayesian network encoding a conditional Gaussian distribution over a fixed set of variables where some variables maybe hidden or have missing observations.

Another possible improvement is, utilizing socio-geographical structure of Indian states to better predict suicides based on the national surveys taken every 4 years.

Lastly, we can extend the same model to other countries and continents and introduce respective parameters that would help eliminate time and judiciary resource wastage in the world.

- Conclusion

We achieve notable results on making important assumptions about our data. It has to be kept in mind that the data is based on the total number of suicides which have already occurred and does not keep in account the total number of non-suicidal deaths. Hence we make the following major assumptions to obtain results:

- 1) For classification problem, we assume that the data rows are independent of the “total” parameter- that is how many total deaths have occurred until now. We have added another column which displays a “yes” or “no” based on combination of data in that row. Decision tree gives best results for this configuration. It was important to leave out “total” parameter, else the decision tree would just consist of 3 nodes: one with $\text{total} \leq 0$ gives No and other with $\text{total} > 0$ gives Yes.
- 2) We have created a Bayesian network and utilized CPTs to incorporate relationships between different type-codes and parameters. The corresponding screenshots as well as accuracies have been reported.

- Bibliography

- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5632514/>
- <http://journals.sagepub.com/doi/abs/10.1177/2167702617691560>
- [http://www.biologicalpsychiatryjournal.com/article/S0006-3223\(17\)30741-2/pdf](http://www.biologicalpsychiatryjournal.com/article/S0006-3223(17)30741-2/pdf)
- <http://weka.sourceforge.net/manuals/weka.bn.pdf>
- <http://www.cs.cmu.edu/~arielpro/15381f16/slides/781f16-bn.pdf>
- <https://www.cs.waikato.ac.nz/ml/weka/>
- <https://stackoverflow.com/questions/10341701/convert-csv-to-arff-using-weka>