**M S Ramaiah Institute of Technology**

(An Autonomous Institute, Affiliated to VTU)

MSR nagar, MSRIT post, Bangalore-54

A Dissertation Report on

Analysis of Indian Suicide Data for Prediction of Number of Suicides Given Features of a Population

Submitted by

**Selva Priya S 1MS12CS100**

**Selva Priyanka S 1MS12CS101**

**Shashank BR 1MS12CS102**

**Sudeep Galgali 1MS12CS114**

*in partial fulfillment for the award of the degree of*

# *Bachelor of Engineering in Computer Science & Engineering*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

**BANGALORE-560054**

[www.msrit.edu](http://www.msrit.edu), **May 2015**

**Acknowledgements**

We would like to thank Dr. K G Srinivasa for giving us this opportunity to perform our study and also for his valuable inputs. We would also like to thank the faculty of MSRIT for their continued support.

# Index

|  |  |  |
| --- | --- | --- |
| Sl no. | Page Heading | Page no. |
| 1 | Abstract | 4 |
| 2 | Introduction | 5 |
| 3 | Implementation | 6 |
| 4 | Analysis and Results | 8 |
| 5 | Conclusion | 10 |
| 6 | Tables | 11 |
| 7 | Figures | 12 |
| 8 | Future Work | 16 |
| 9 | References | 17 |

# Abstract

We present a study aimed at discovering the prime factors that affect the rate of suicides in certain regions of India. The Indian government maintains a database of the reported suicide cases in each region of India; this is made public for the purpose of data analytics. This data is classified under various different sub categories. From our study we were able to identify the prime sub categories that contribute significantly to the suicide rates.

Three datasets were used in our study, the marital status of suicide victims, the Educational level of suicide victims and the Primary Census, all the data was of the year 2011. These datasets was subject to pre-processing to obtain it in the required format. Pearson correlation was performed on them to determine the linearity of the features on the suicide rates and then linear regression was used to develop a model for prediction of suicide rates.

The results we obtained were quite remarkable, 9 features were found to have a significant linear relationship with suicide rates. The prediction model developed using these 9 features gave us a linear fit close to 96% and prediction accuracy of 94%.

**Introduction**

Every year about 800,000 commit suicides all over the world , of these 135,000 (17%) are residents of India,  a nation with 17.5% of world population.  Between 1987 to 2007, the suicide rate increased from 7.9 to 10.3 per 100,000, with higher suicide rates in southern and eastern states of India. In 2012, Tamil Nadu (12.5% of all suicides), Maharashtra (11.9%) and West Bengal (11.0%) had the highest proportion of suicides Among large population states, Tamil Nadu and Kerala had the highest suicide rates per 100,000 people in 2012. The male to female suicide ratio has been about 2:1.

Estimates for number of suicides in India vary. For example, one study projected 187,000 suicides in India in 2010,  while official data by the Government of India claims 134,600 suicides in 2010.

According to WHO data, the age standardized suicide rate in India is 16.4 per 100,000 for women (6th highest in the world) and 25.8 for men (ranking 22th).

In this paper an attempt has been made to identify the causes of suicides. This is done by calculating how correlated features of a state are with the number of suicides. The features include marital, professional and educational status of the state’s population. After this the significant features are used to predict number of suicides.

**Implementation**

First step in implementing is creating datasets.

**Dataset Creation:**

The aim of dataset creation is to prepare a single dataset for analysis which is sourced by multiple datasets. The list of datasets used is

* 12 Marital status of population datasets for each state in 2011[7]
* 12 Educational level of population datasets for each state in 2011[7]
* 12 Primary census abstract datasets for each state in 2011[7]

The 12 states considered are

* Andhra Pradesh
* Bihar
* Gujarat
* Karnataka
* Kerala
* Madhya Pradesh
* Maharashtra
* Punjab
* Rajasthan
* Tamil Nadu
* Uttar Pradesh
* West Bengal

The analysis dataset is created in which each state has 69 features. These features are considered only for the part of the population aged between 15 and 29.

Next step is analysis using different techniques. We have used Pearson correlation and Regression modeling.

**Pearosn Correlation:**

Correlation is a technique to measure how two sets of data are related, Pearson correlation also called as the Pearson Product Moment correlation particularly describes the linear relationship . The correlation coefficient always lies between -1 and 1. A value of -1 indicates that there exists a perfectly negative relationship i.e. If we have two variables X (independent) and Y (dependent), Y is expected to yield a line with a negative slope when plotted with respect to X. Conversely a value of 1 indicates a perfect positive relationship. A value of 0 indicates no correlation implying that Y is independent of the variation in X.

We have calculated the Pearson correlations for all the 70 sub categories and selected only those for modeling which have a value greater than 0.5.

**Regression Modeling:**

Linear Regression is the statistical method of modeling the relationship between a dependent variable and one or more independent variables . In the case of a single independent variable the process is termed Simple Linear Regression and in the case of multiple independent variables it is called multivariate linear regression. The variables are modeled by .

*y = f(x) + c*

Where y is the dependent variable and x = x1, x2, x3…..xn are the independent variables and n is the number of independent variables.

**Analysis and Results**

**Analysis of Pearson Correlation:**

The correlations of the 7 significant features with the number of suicides are tabulated in Table 1. From Table 1 we can identify the main sub categories of the population that influence the suicide rates in certain regions of India as

* Separated Females have a very high correlation with the number of suicides, which prompts us to believe that this is the most important factor in predicting the suicide rates.
* The next in line is the farmer, the main cultivator population.
* Men with a technical diploma or certificate not equal to degrees.
* People who own or work in household industries.
* People with a technical diploma or certificate not equal to degrees.
* Divorced Females
* People who do not work or don’t have jobs

**Analysis of Regression Modeling:**

A linear model is developed for the significant features to predict number of suicides The Classical Ordinary Least Squares method is used to estimate the number of suicides.

In order to evaluate a multivariate regression model a test can be carried out known as the F-test. The F-test is basically used to compare an intercept only model with the specified model. It involves setting up a null hypothesis which states that ‘The coefficients of all independent variables involved in the model are equal to zero’.

The alternate hypothesis is that ‘At least one of the independent variable involved in the model has a non-zero coefficient’. Following this, a number of steps are carried out in the F-test.

* Calculation of test statistic (denoted by F) assuming the null hypothesis is true as given by (2).
* Calculation of 95% confidence interval I for degrees of freedom using an F-table or statistical software.
* Acceptance of the null hypothesis if F ∈ I; reject if F ∉ I.
* Calculation of the p-value using statistical algorithms

***F =*  (2)**

The results are summarized in Table 2.Partial regression plot of the model is shown in Image 2 at the end of the report.

**Conclusion**

The results obtained are extraordinary; a percentage accuracy of 99.83% is exceedingly high. The partial regression plot shows that all the features lie within 95% confidence bounds. From this we can conclude that the rate of suicides depends mainly on the identified factors of the total population. If given the total number of separated females, farmers, men with no degrees or just a technical diploma, people working on household industries, total number of people without degrees or with just a technical diploma, people working in household industries, divorced females and people without jobs of a particular state, we can predict with a great accuracy the number of suicides for that state.

**TABLES**

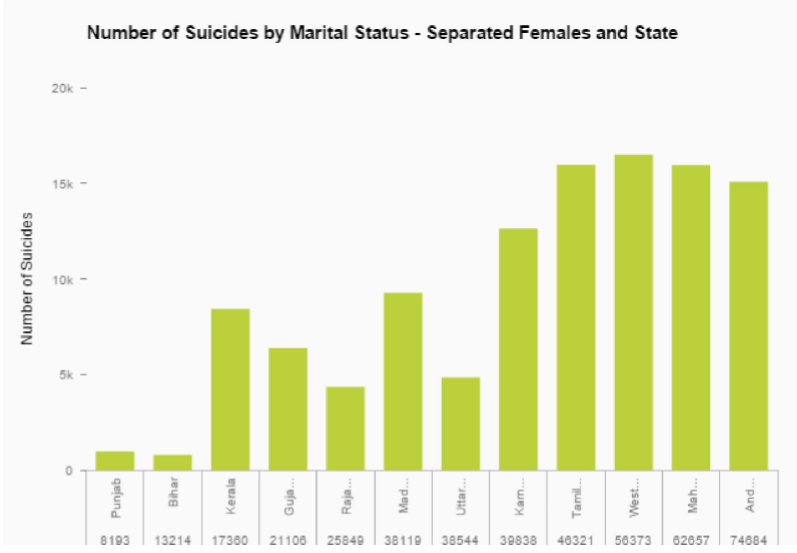
Table 1: Pearson correlations of the 6 significant features

|  |  |
| --- | --- |
| **Feature** | **Correlation co-efficient** |
| Marital Status – Widowed Females | 0.5025 |
| Marital Status - Separated Persons | 0.7209 |
| Marital Status - Separated Females | 0.8618 |
| Marital Status – Divorced Females | 0.5558 |
| Educational level - Technical diploma or certificate not equal to degree - Persons | 0.5842 |
| Educational Level- Technical Diploma or Certificate not equal to Degree - Males | 0.6265 |
| Main Working Population Female | 0.6429 |
| Main Agricultural Laborers Population Female | 0.5914 |
| Main Household Industries Population Female | 0.5452 |

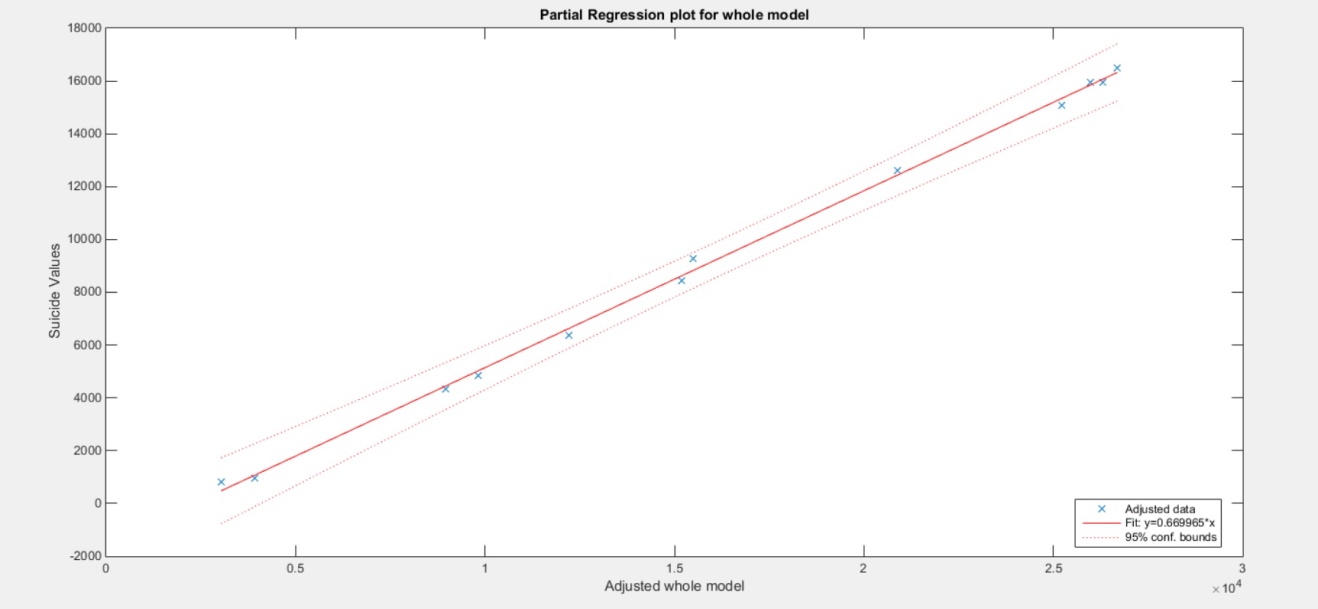
|  |  |
| --- | --- |
| R-squared | 0.961 |
| Percentage Fit | 96% |
| Percentage Accuracy | 94.1363% |
| Ftest Pvalue | 0.00697 |
| RMS Error | 543 |

**Table 2**

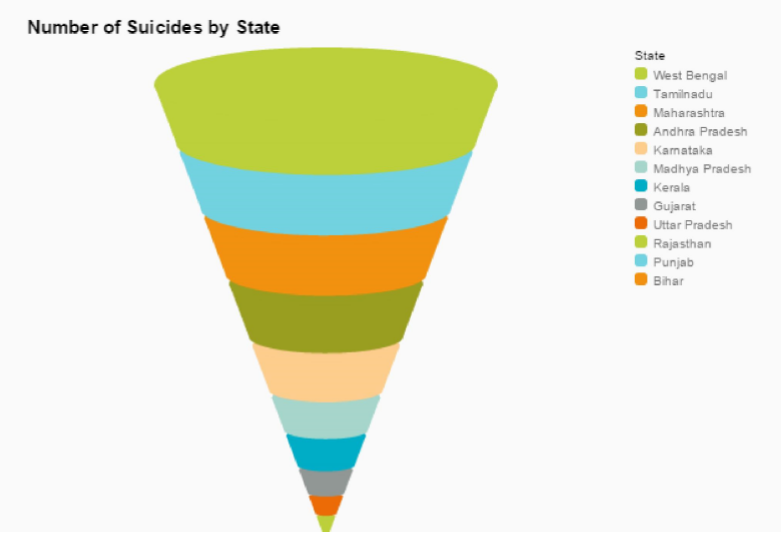
**Screenshots**

****

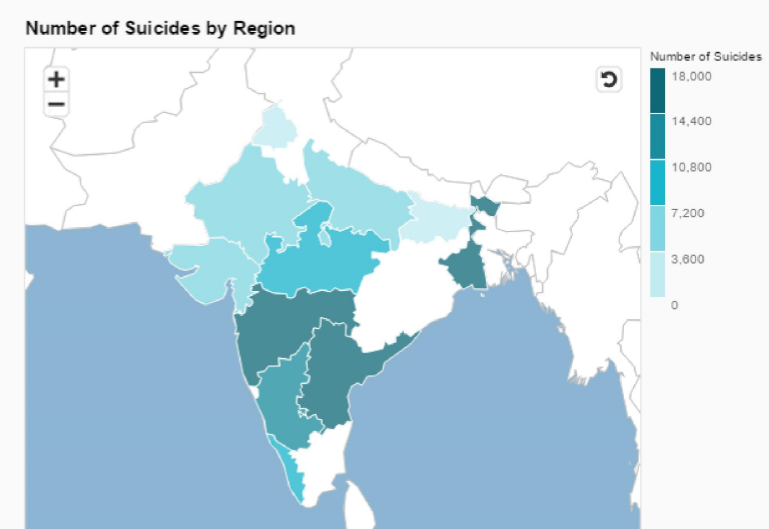
**Image 1**

****

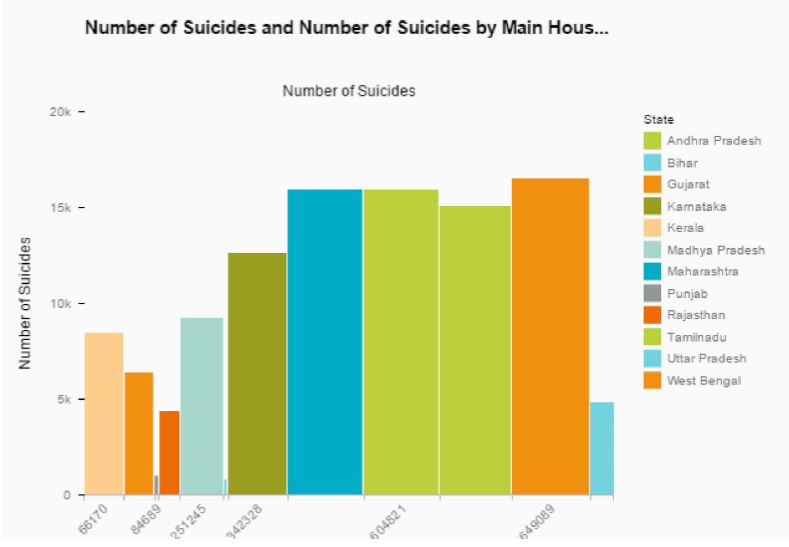
**Image 2**

****

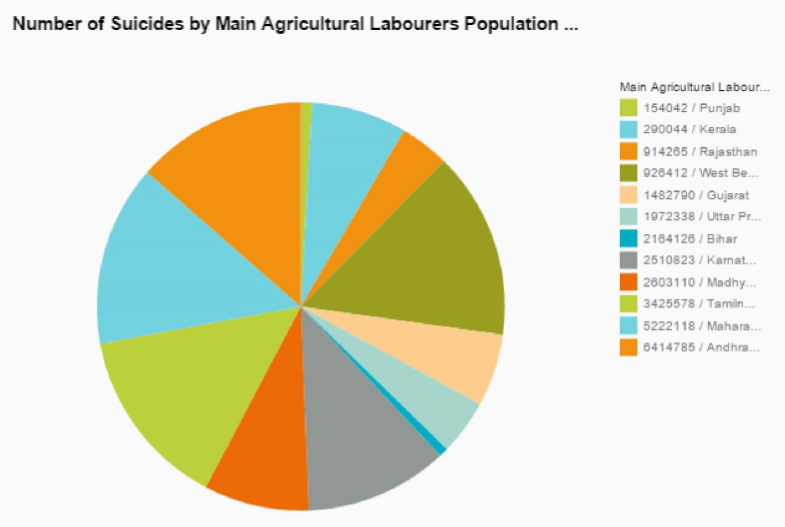
**Image 3**

****

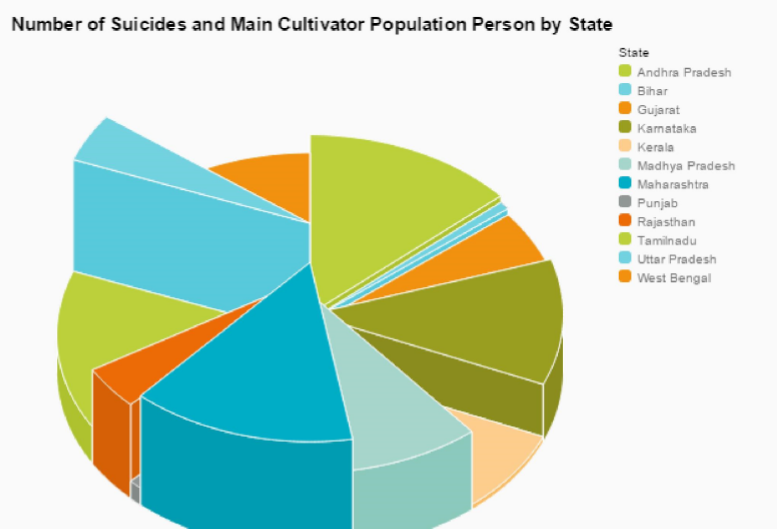
**Image 4**

****

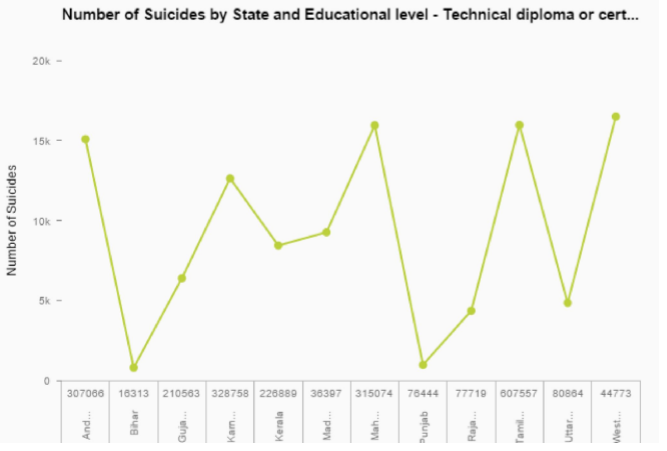
**Image 5**

****

**Image 6**

****

**Image 7**

****

**Image 8**

**Future Work**

We would like to conduct such a study considering the remaining states and find what group of the population has a significant relationship on the number of suicides. We also want to consider features other than the population.

**References**

* Suicide prevention (SUPRE) World Health Organization (2012)
* Suicides in India The Registrar General of India, Government of India (2012)
* Vijaykumar L. (2007), Suicide and its prevention: The urgent need in India, Indian J Psychiatry;49:81-84,
* Polgreen, Lydia (March 30, 2010). "Suicides, Some for Separatist Cause, Jolt India". The New York Times.
* Patel, V.; Ramasundarahettige, C.; Vijayakumar, L.; Thakur, J. S.; Gajalakshmi, V.; Gururaj, G.; Suraweera, W.; Jha, P. (2012). "Suicide mortality in India: A nationally representative survey". The Lancet 379 (9834): 2343. doi:10.1016/S0140-6736(12)60606-0.
* Suicide Rates - Data by country. World Health Organization 2012. Retrieved 30 November 2015.
* Census India 2011
* David A. Freedman, “Statistical Models: Theory and Practice”, Cambridge University Press. P. 26, 2009.Richard Taylor, “Interpretation of the correlation coefficient:A Basic Review” ,JDMS 1:35-39, January/February 1990.
* Hayashi, Fumio (2000). Econometrics. Princeton University Press.ISBN-0-691-01018-8