Indian Institute of Technology, Kanpur

CS685A: Data Mining
Mid-Term Project Report 2020

Title: Analysis of number of child births and infant deaths in India

---

Supervised By: Prof. Arnab Bhattacharya

16 October 2020

Submitted By: Group 5

Lavlesh Mishra	19111048(lavleshm@iitk.ac.in)
Kuldeep Kumar Solanki	19111045(kuldeeps@iitk.ac.in)
Jaydeep Meda	19111039 (jaydeepm@iitk.ac.in)
Aditya Jain	20111004 (adityaj20@iitk.ac.in)
Rohit Singh	20111418 (kun20@iitk.ac.in)

Contents

1 Abstract

This report narrates few main aspects of the project like, our approach for dataset preparation starting from methods used for data extraction to steps involved in data pre-processing. It also describes the work done till now, with expected results followed by ways to evaluate them.

2 Broad Aims of the project

To create a model that predicts the number of child births and infant deaths in a year for all district of India.

3 Datasets

3.1 Performance of Key Health Management Indicators for each district in India

This data is released by Ministry of Health under National Health Mission flagship program which seeks to provide effective healthcare to the rural population throughout the country.

Data set includes the key indicators which affects health of mother and child during pregnancy and at the time of delivery. Data also reports the number of children born and number on infant deaths in each district in a particular year. Data set is available for years 2008 to 2019.

Link: `https://www.nrhm-mis.nic.in/hmisreports/frmstandard_reports.aspx`

3.2 Census Data 2001 and 2011

The Census of India is conducted every 10 years Ministry of Home Affairs. The Census Data is available in public domain. The data is available at district level for total population, age, disability, education, migration, religion, and various other features.

Link to Census 2001: `https://censusindia.gov.in/DigitalLibrary/TablesSeries2001.aspx`

Link to Census 2011: `https://censusindia.gov.in/2011census/population_enumeration.html`

4 Data Preparation

4.1 Pre-processing of HMIS data:

The data of the Health Management Information System (HMIS) database was given in the form of excel files and in state wise manner for each financial year from 2008-09 to 2018-19. Initially, the dataset contained a lot of unnecessary information that we did not use.

We have filtered the dataset and selected the relevant attributes for the prediction of the number of births in a particular district in a particular year.

4.2 Pre-processing of Census data:

The data of census was given in form of excel files. We have manually pre-processed the excel file to extract the required data. The data was present for the year 2001 and 2011. So we have generated the data for the years 2002-2019, using the Compound Interest formula.

- The Growth Rate(r) is calculated using the 2001 data as Principal(p) and 2011 data as the final amount(p+i). Here i is the Interest.
- Data is generated for the remaining years using calculated growth rate(r).

Correction of district and state names over the years for both the datasets. As new districts are introduced, or name of a district is changed over the years 2008 to 2019.

5 Work Done Till Now.

5.1 Pre-processing:

As a test case we have applied the above pre-processing and different Machine Learning techniques for the Gujrat state.

The data from the year 2008 to 2017 is used to train the models and the testing was done for the year 2018. We have predicted the no. of births for the year 2018 district wise in Gujrat state. There are 26 districts in Gujrat state. Each district is given a district id from the numbers 0-25.

The Features used in training the models from HIMS datasets are:

- Number having Hb level<11 (tested cases)
- Number having severe anaemia (Hb<7) treated at institution
- Number of home deliveries attended by Non SBA trained (trained TB/Dai)
- Number of C-section deliveries conducted at public facilities
- Number of C-section deliveries conducted at private facilities
- Total Number of reported Still Births
- Total Number of Abortions (Spontaneous/ Induced) Reported
- Total Number of MTPs (Public) reported
- Number of Vasectomies Conducted (Public + Pvt.)
- Number of Tubectomies Conducted (Public + Pvt.)
- Total Sterilisation Conducted
- IUCD Insertions done (public facilities)
- IUCD insertions done (pvt. facilities)
- Oral Pills distributed
- Condom pieces distributed
- Adverse Events Following Immunisation (Others)
- Number of Major Operations
- Number of Minor Operations
- Total Number of Infant Deaths reported

The Features used in training the models from census datasets are:

- Population Persons
- Literate Persons
- Main workers Persons
- Marginal workers Persons
- Non-workers Persons

5.2 Results Obtained

The training and testing score using different Machine Learning Techniques:

Model Name	Training Score	Testing Score
Linear Regression	0.9670404857345802	0.8150463694405432
Ridge Regression	0.9670404857345805	0.8150463696084705
Lasso Regression	0.966971710557302	0.8188446199470976
Random Forest	0.9915817817676562	0.9441529668423073

Table 1: Prediction Accuracy

The Random Forest Regressor Model performs the best with training accuracy of 99.15% and testing accuracy of 94.41%.

5.3 Results Evaluation

Figure 1. is the graph for prediction of births in Gujrat state district wise for the year 2018. The 'Districts' axis in the graph is 'district id' and 'Births' axis is the 'no of births in the corresponding district'.

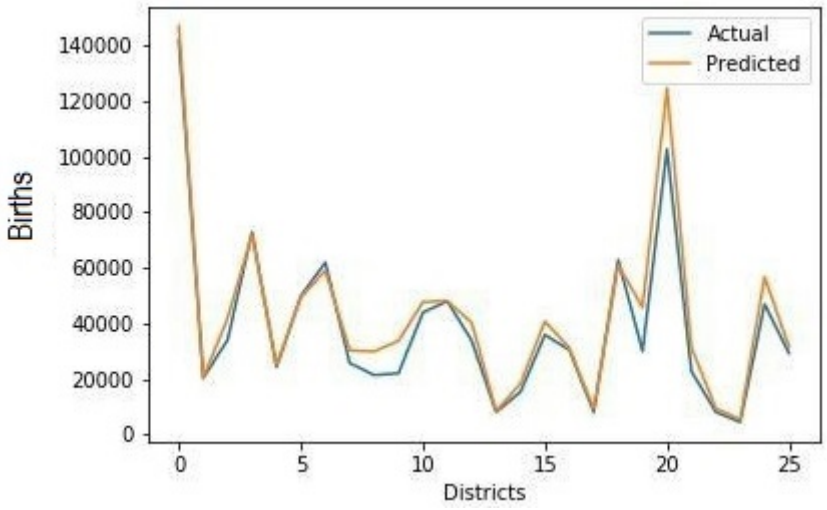

Figure 1: Random Forest Prediction Graph

The most significant features as per the Random Forest Regressor:

- Non-workers Persons
- Population Persons
- Literate Persons
- Number of Tubectomies Conducted (Public + Pvt.)
- Total Sterilisation Conducted
- Main workers Persons
- Total Number of reported Still Births
- Oral Pills distributed
- Number having severe anaemia (Hb<7) treated at institution
- IUCD Insertions done (public facilities)

6 Results to be obtained

The results obtained are:

- Total no. of Child Birth for the Year 2020.
- Total no. of Infant Deaths for the Year 2020.

7 Evaluation of Results

We will divide our datasets in two parts: Training set & Test set. The data from the year 2008 to 2017 will be used as training set and the data for the year 2018 will be used as test set. Accuracy of the model developed will be evaluated based on this test set.

8 Important Links and References

- HMIS Database:`https://www.nrhm-mis.nic.in/hmisreports/frmstandard_reports.aspx`
- Census Data 2001:`https://censusindia.gov.in/DigitalLibrary/TablesSeries2001.aspx`
- Census Data 2011:`https://censusindia.gov.in/2011census/population_enumeration.html`
- GitHub Link for the project: `https://github.com/kuldeeps5/DataMiningFinalProject`
- Implementation of Lasso and Ridge Regression: `https://analyticsindiamag.com/hands-on-implementation-of-lasso-and-ridge-regression/`
- Random Forest Regression: `https://www.geeksforgeeks.org/random-forest-regression-in-python/`