Indian Institute of Technology, Kanpur

CS685A: Data Mining
Project Report

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

---

Supervised By: Prof. Arnab Bhattacharya

26 November 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

To monitor the performance and quality of the health services being provided under the (National Health Mission)NHM, the Ministry of Health Family Welfare, Government of India, is putting in place several mechanisms that would strengthen the monitoring and evaluation systems, through performance statistics, surveys, community monitoring, quality assurance etc. Government releases the Health Management Information System(HMIS) database in public domain, which we are using along with the census data to analyse the total number of births, still births and infant deaths in a year for a particular district.

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 (HMIS)

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, number of still-births and number of infant deaths in each district in a particular year. Data set is available for years 2008 to 2019. Data set is available in the format of Financial years that is from 1st April of a year to the 31st March of the next year.

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

Route to the data set files:
Performance of Key HMIS Indicators(upto District Level)/20XX-20XX/MonthUpToMarch/Select_State/All.xls/Click here to donwload

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.

Data is available for the years 2001 and 2011. From these two files, we have generated projections of census data for the years 2008 to 2018.

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 Transformation

Our training model requires dataset in well structured format. Data we found was in unstructure excel files for both the datasets - Census and HMIS. We need to transform this data into structured csv files. Characterisitics fo these files are as follows:

- There are 70 files of Census data. In which 35 are xls files for the Census-2001 and 35 xlsx files for the Census-2011.
- There are 385 xls files in HMIS data set. There are 11 xls files for each state, each file for a financial year from 2008 to 2019.

4.1 Transformation of HMIS data:

HMIS data is present in unstructured xls files. The attributes are merged into single cell entry in excel file. Figure 1 shows merged cells of the HMIS data. These files present in excel extension but they are written in XML format and then converted into excel by the Ministry of Health. Since these are actually XML files, we can't process them using xldr library.

2	Financial Year: 2018-19																				
3	Provisional Figures for the Period April to March																				
4	These indicators are under compilation (83,84,85,86,197,198)																				
5		1		2		3		4		5		6		7		8		9		10	
6	Indicators	Total number of pregnant women Registered for ANC		Number of Pregnant women registered within first trimester		% 1st Trimester registration to Total ANC Registrations		Number of pregnant women received 4 or more ANC check ups		TT2 given to Pregnant women (numbers)		TT Booster given to Pregnant women (numbers)		% Pregnant Woman received 4 ANC check ups to Total ANC Registrations		% Pregnant women received TT2+ TT Booster to Total ANC Registration		Number of Pregnant women given 180 IFA tablets		% Pregnant women given 180 IFA to Total ANC Registration	
7		2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18	2018-19	2017-18
8	_Jammu & Kashmir	3,95,092	3,99,307	2,66,690	2,58,865	67.5	64.8	2,88,808	2,47,859	1,76,547	1,74,937	55,651	56,517	73.1	62.1	58.8	58	1,40,277	2,14,965	35.5	53.8
9	1 Anantnag	28,918	27,773	24,691	25,353	85.4	91.3	16,029	19,526	20,516	21,546	6,025	6,002	55.4	70.3	91.8	99.2	8,009	14,662	27.7	52.8
10	2 Badgam	12,379	12,592	12,289	12,104	99.3	96.1	9,632	9,707	12,060	11,465	27	1,366	77.8	77.1	97.6	101.9	8,592	10,347	69.4	82.2
11	3 Bandipora	7,140	7,641	5,954	6,279	83.4	82.2	5,155	5,167	5,337	5,436	1,883	2,030	72.2	67.6	101.1	97.7	5,602	4,891	78.5	64
12	4 Baramula	22,241	25,103	18,756	20,940	84.3	83.4	20,931	22,510	15,479	14,386	7,796	9,418	94.1	89.7	104.6	94.8	21,892	23,461	98.4	93.5

Figure 1: HMIS data

To process them we copied the content from all these 385 files into csv (comma separated) files and then processed these csv files. Steps we followed to transform these XML embedded excel files into structure csv is as follows:

1. Download 385 xls files from the HMIS website.
2. Files downloaded from the website have a default name 'All.xls'. So to process different states and years, we renamed these files as per the state names and year. We used the following command of linux terminal to rename all the file belonging to a state at a time:

Command Line

```
$ for file in *; do mv "file""(basename "$file")StateName2016-2017"; done;
```

We executed the above command 35 times, once for each state.

3. To convert the extension of excel files to csv we used the following command of linux terminal:

```
Command Line
$ for i in *.xlsx; do libreoffice –headless –convert-to csv "$i" ; done
```

After performing above steps we got the files in structured form on which we can perform processing using python libraries.

4.2 Transformation of Census data:

Excel files of Census data is semi-structured and to convert them into structured comma separated files (csv), we used **xlrd** library of python to process it. The xlrd is a library for reading data and formatting information from Excel files, whether they are .xls or .xlsx files.

```
Command Line
$ pip3 install xlrd
$ python3 census-transform.py
```

5 Data Pre-processing

In this section, we have discussed the pre-processing of the two data set individually, before integrating them into one. We processed following files:

5.1 Pre-processing of HMIS data:

Characteristics of Data
HMIS data has the following characteristics:

1. There are 385 csv files for every 28 States and 7 Union Territories. Telangana and Ladakh are not considered as they are not in Census data.
2. Data set has 163 attributes.
3. Each files stores the data for two years and hence there are 325 columns in the csv files.
4. One attribute is of category type rest all attributes are discrete variables.

Cleaning

- Data is redundant in each files as each files contains data for two financial years. We removed the data of one financial year as it was present in the other files.
- Many attributes of the data set has missing values, since there was no way to predict these values, we removed such attributes from the dataset.

5.2 Pre-processing of Census data:

Characteristics of Data
Census data for year 2001 has the following characteristics:

1. There are 35 files for every 28 States and 7 Union Territories. Telangana and Ladakh were not created at that time.
2. Data set has 64 attributes.
3. One attribute is of category type rest all attributes are discrete variables.
4. Data is divided on the proximity - Rural and Urban

Census data for year 2011 has the following characteristics:

1. There are 35 files for every 28 States and 7 Union Territories. Telangana and Ladakh were not created at that time.
2. Data set has 94 attributes.
3. Data is divided on the proximity - Rural and Urban

Inconsistency in data
Files contains much inconsistent data:

- 'All ages' attribute includes 'age not stated', i.e. the attribute - 'All ages' has data for the people whose age are not known.
- 'Literate' attribute includes figures for 'literates without educational level' and 'educational levels not classifiable'.
- District IDs are outdated. Since 2011 many districts are newly created which lead to the change of district IDs.
- 'Matric/ secondary but below graduate' includes 'non-technical diploma and certificate not equal to degree'.
- Ever married women includes currently married, widowed, divorced and separated.
- There are missing values in the dataset, represented by NaN.

Cleaning

- We needed statistics for entire district and not in terms of rural and urban regions, so we computed the statistics of entire districts from these two fields.

- Many attributes of the data set has missing values, since there was no way to predict these values, we removed such attributes from the dataset.
- There are cases where some of the attributes' value is NaN. We handle it by setting these values to zero. Imputing with mean, median or mode was not a suitable choice as there are very less concerned data from which these quantities were to be calculated. The concerened data is the data examples for the same district among all the years. Since there are only 11 years, calculating mean, mode or median is not suitable.
- Much of the inconsistencies mentioned above are left in the data. They are handled at outlier analysis stage of the data exploration.

5.3 Generation of Census data for years 2008 to 2019

The Census data was present for the year 2001 and 2011 only but we need the data for each year from 2008 to 2019 to map with the HMIS data. So we have generated the data for the years 2008-2019, 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.

$$(1+r)^{10} = \frac{Census - 2011 \ \ stats}{Census - 2001 \ \ stats}$$

  Growth rate r is constant over here, which means the interpolation is geometric and not linear.

- Data is generated for the remaining years using calculated growth rate(r).

$$\text{Census Stats for the year Y} = (1+r)^{(Y-2001)}$$

Using above method we generated 385 Census files from 70 original files. 11 files for each state, each for one year.

6 Feature Selection

6.1 Feature selection on Census Data

Features selected from this data are: "NAME", 'TOT_P','P_LIT', 'MAINWORK_P', 'MARGWORK_P', 'NON_WORK_P'. These feautures are renamed to relevant names:

- Area Name: It represent the district name.
- Population Persons: It represents the total population of the district in that year.
- Literate Persons: It represents the total number of literate individuals in the district in that year. Literate persons are the people who have completed at least Matric.
- Main workers Persons: It represents the total number of individuals that are involved in any one of the following activities:

  - Culitvators
  - Agricultural labourers
  - Household Industry workers
  - Other work

- Marginal workers Persons: It represents the total number of individiuals that do marginal work in the district.

- Non-workers Persons: It represents the total number of individuals that are unemployed.

6.2 Feature selection on HMIS Data

Out of 163 attributes we filtered out 42 attributes for our project. There are many attributes which represent the percentage of the other attributes. We dropped such attributes as they can be derived from the other attributes otherwise they would increase the problem of multi-collinearity.

There are 8 for which there is no data available so to reduce the strain on missing value treatement step we removed these attributes. There are 6 attributes which contains very low values or the value zero. These discrete values seemed very unrealistic for an entire district, so to reduce the strain on outlier analysis, we removed such attributes from our data set.

7 Data Integration

We have two data sets - HMIS and Census and we need to merge these two data sets to obtain one single data set one which we can learn a regression model to predict the target variables. So in this section we will discuss the techniques we used to integrate the datasets.

Both the datasets has one same attribute. This attribute is 'Area Name' in Census data and it is 'Indicator' in HMIS data. Both represent the name of the district of India. So we performed natural join on this common attribute to obtain the merger of both the data sets.

Problems in Integrating:
To integrate 385 files of Census and 385 files of HMIS we faced following problems:

- District names in both the files are given in different format. In Census data string - "District-" is appended before each of the district's name.

- Some district is represented by different names in different data sets.

- The name of the districts is changed over the years.

- For some of cases the spelling of district names in both the data files do not match.

Solutions Implemented
We handled above discussed problems using various techniques. We removed the additional strings from the 'Area Name' attribute of the Census data.
We used Edit-Distance algorithm to map the cases of different spellings. Target distance for the algorithm was set to 3, on increasing the value greater than that was leading to wrong mapping of district names. There were some cases which Edit-Distance can't handle, so we renamed those explicitly by hard code.

There were many cases in which there were less districts in Census data in comparision to Census data. So in such cases we were forced to drop the data examples from the HMIS data set to map the data completely.
So after doing all the processing we merged 770 csv files to generate one single csv file - 'all_states_merged_8to18.csv'

8 Data Exploration

	Indicator	Total number of pregnant women Registered for ANC	Number of Pregnant women registered within first trimester	Number of pregnant women received 3 ANC check ups	TT2 or Booster given to Pregnant women (numbers)	Number of Pregnant women given 100 IFA tablets	Number having Hb level<11 (tested cases)	Number having severe anaemia (Hb<7) treated at institution	Number of Home deliveries	Number of home deliveries attended by SBA trained (Doctor/Nurse/ANM)	...	Fo Imu
0	1	2253.0	739.0	2007.0	1667.0	7857.0	3496.0	34.0	350.0	123.0	...	
1	2	925.0	247.0	638.0	584.0	1230.0	4459.0	214.0	39.0	29.0	...	
2	1	1511.0	611.0	1491.0	1097.0	2674.0	2715.0	192.0	216.0	89.0	...	
3	2	1043.0	278.0	554.0	485.0	1413.0	3926.0	335.0	62.0	52.0	...	
4	1	1279.0	896.0	1255.0	1115.0	1397.0	1509.0	36.0	190.0	76.0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
6600	12	78766.0	71127.0	4417.0	78.1	105.6	214.0	45.6	0.0	453.0	...	
6601	13	146754.0	108454.0	5773.0	66.4	77.6	527.0	65.0	0.0	1291.0	...	
6602	14	58172.0	53672.0	12079.0	86.3	86.3	386.0	138.4	0.0	2253.0	...	
6603	15	154018.0	134006.0	10871.0	78.3	97.5	416.0	63.0	0.0	7324.0	...	
6604	16	72059.0	63896.0	15796.0	84.3	87.5	113.0	58.5	0.0	12124.0	...	

6605 rows × 47 columns

Figure 2: Dataset of Project

There are xxx data examples in our data set. Each data example has 47 dimensions.

8.1 Univariate Analysis

At this stage, we explore variables one by one. We found that out of **47 varibles/attributes**, only one attribute 'Indicator' is categorical and all other variables are of type discrete. Out of these 46 discrete variables, 5 variables are of integer data type an rest 41 are of float data type.
Since we are solving regression problem, categorical data "Indicator" is of no use in our prediction. So we removed it from the data set.

8.2 Independent and Dependent variable identification

We first, identify **Predictor** (Input) and **Target** (output) variables. We have three regression problems, so for each problem diffrent independent and dependent variables are needed to found.

Case 1: Predicting number of Births in a year
Independent\Target attribute: 'Total Number of reported live births'
Dependent attributes: Rest 45 attributes of the dataset.

Case 2: Predicting number of Still-Births in a year
Independent\Target attribute: 'Total Number of reported Still Births'

Dependent attributes: Rest 45 attributes of the dataset.

Case 3: Predicting number of Infant-Deaths in a year
Independent\Target attribute: 'Total Number of reported Infant Deaths'
Dependent attributes: Rest 45 attributes of the dataset.

8.3 Multivariate Analysis

In this section, we found out the relationship between two variables. Here, we looked at the association and disassociation between variables at a pre-defined significance level.

8.3.1 Collinearity Test

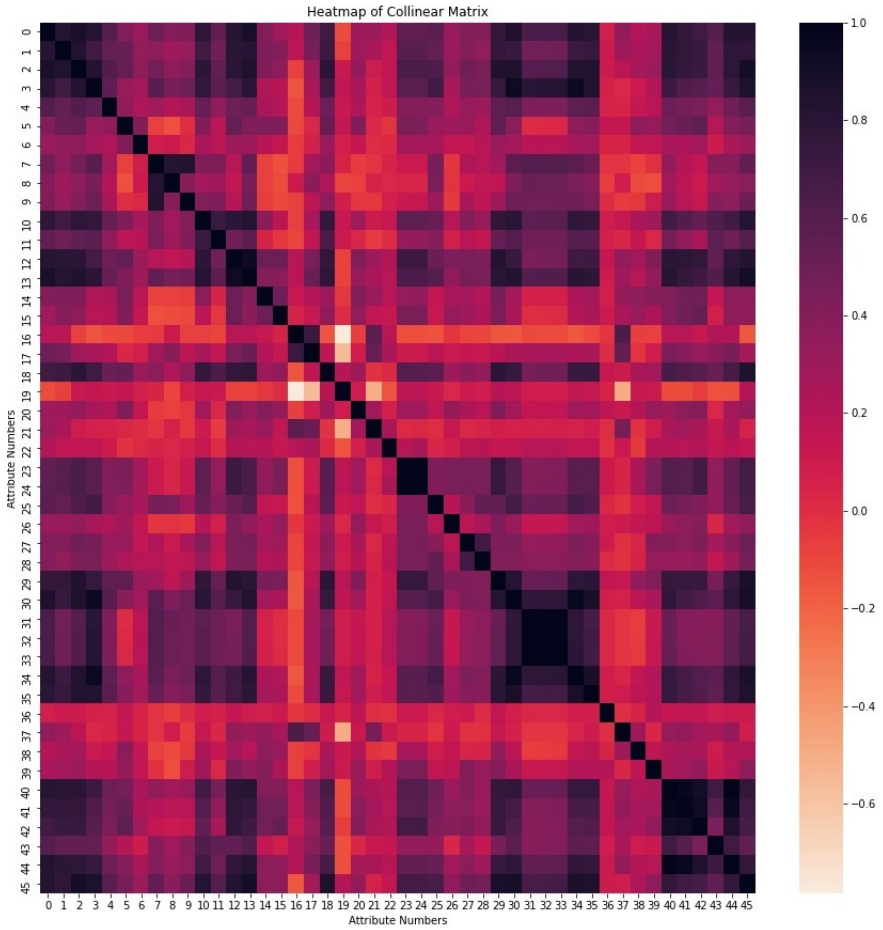

Figure 3: Heatmap of Collinear Matrix of dataset

There are lot of dark squares in above heat map. So it shows that data attributes are highly correlated to each other.
We found that the attribute 'Number of Pregnant women registered within first trimester' is highly correlated to the attributes - 'Deliveries Conducted at Public Institutions', 'Number of pregnant women received 3 ANC check ups', 'TT2 or Booster given to Pregnant

women (numbers)'. This shows that women who registered within first trimester with ANC receives 3 ANC check ups, TT2 supplement are very likely to have delivery at some institution- clinic -public or private. Women those have home deliveries don't register themselves with ANC.

8.4 Missing Value Treatment

Missing Value analysis is most tricky part of this data exploration as there are no NaN values in our dataset at this stage. All NaN values in Census data were taken care at the individual pre-processing of the census files, as discussed in the section 5.2 and the one which cannot be taken care was made equal to zero. In HMIS data there are no NaN values but there are values which are equal to zero.

The tricky part is whether these zero values are missing values or actual value of the data example is zero. So we can't do anything to our dataset at this stage, but our outlier analysis can find our whether these zero values are genuine or not. Outlier Analysis will mark the NaN values written as zero as outlier and will remove them.

8.5 Outlier Analysis

Outlier is an observation that appears far away and diverges from an overall pattern in a sample. In this section, we will discuss two outlier detection techniques that we applied on our data set. These are:

1. IQR Test
2. Z-Score

8.5.1 IQR Test

IQR stands for Interquartile Range. We define IQR as follows:
IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts. IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1. The data points which fall **below Q1 − 1.5 IQR or above Q3 + 1.5 IQR** are outliers.

IQR values of our data set is shown in the Table 1.
Attributes (top 10) having most number of outliers are shown in Table 2.

Table 1: IQR values of attributes of data set

Total number of pregnant women Registered for ANC	0
Number of Pregnant women registered within first trimester	35262
Number of pregnant women received 3 ANC check ups	22257.5
TT2 or Booster given to Pregnant women (numbers)	26596.5
Number of Pregnant women given 100 IFA tablets	28525
Number having Hb level<11 (tested cases)	27815
Number having severe anaemia (Hb<7) treated at institution	19454.5
Number of Home deliveries	790.5
Number of home deliveries attended by SBA trained (Doctor/Nurse/ANM)	4199
Number of home deliveries attended by Non SBA trained (trained TB/Dai)	1135
...	2767.5
Adverse Events Following Imunisation (Others)	...
Number of Major Operations	120
Number of Minor Operations	5182.5
Total Number of Infant Deaths reported	9829.5
Population Persons	337
Literate Persons	1683934.5
Main workers Persons	1101239
Marginal workers Persons	522418
Non-workers Persons	176729.5
Total Number of reported live births	989727

Table 2: Attributes (Top 10) having most number of Outliers

Attributes Name	**No. of Outliers**
Adverse Events Following Imunisation (Others)	610
IUCD insertions done (pvt. facilities)	532
Number of home deliveries attended by SBA trained (Doctor/Nurse/ANM)	414
Number of Vasectomies Conducted (Public + Pvt.)	406
Number of Minor Operations	392
Total Number of MTPs (Public) reported	384
Number of C-section deliveries conducted at private facilities	357
Number having severe anaemia (Hb<7) treated at institution	327
Number of Major Operations	319
Number of home deliveries attended by Non SBA trained (trained TB/Dai)	295
Number of C-section deliveries conducted at public facilities	278
Condom pieces distributed	272
Total Number of Abortions (Spontaneous/ Induced) Reported	267
Number of Home deliveries	256

8.5.2 Z-Score Test

Z score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.
If the z score of a data point is more than 3, it indicates that the data point is quite different from the other data points. Such a data point can be an outlier.

	count	mean	std	min	25%	50%	75%	max
z-score analysis	303830.0	0.0	1.0	-4.903	-0.533	-0.236	0.267	30.46

Figure 4: Z score analysis for outliers

Figure 4 shows our statistics on z-score calculated, where mean of z-scores is 0 and standard deviation of z-scores is 1.0. Minimum value is -10.778 and maximum value is 40.996. To identify the optimum value of threshold for classifying outliers we draw histogram and density plot of obtained z-score. Figure 5 and Figure 6 shows them respectively.

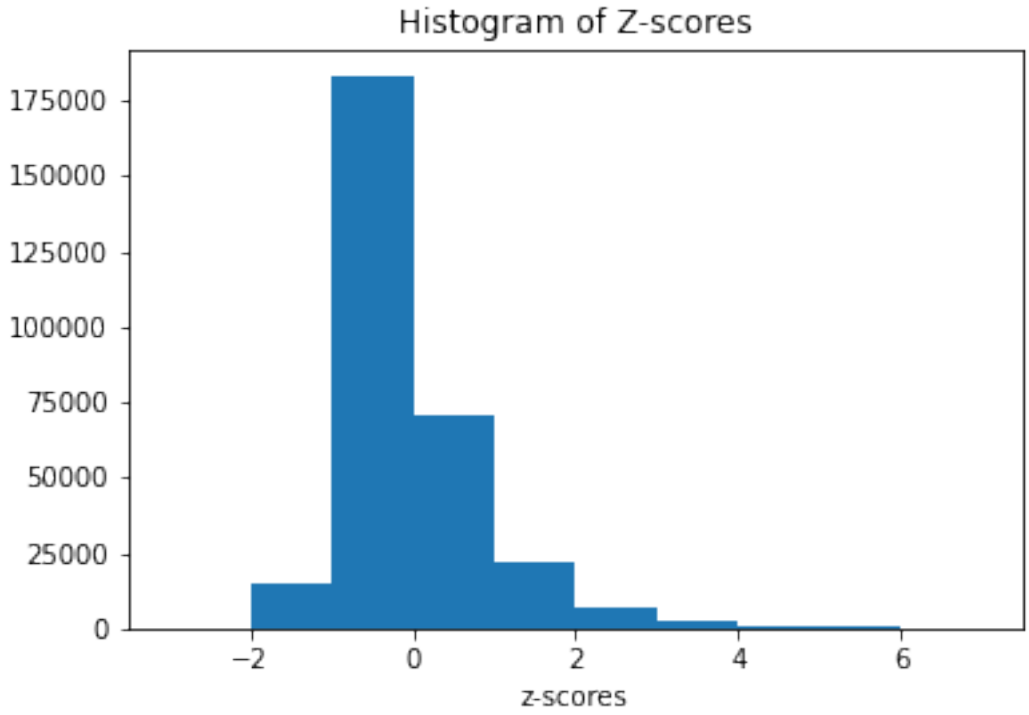

Figure 5: Histogram of z-scores obtained from dataset

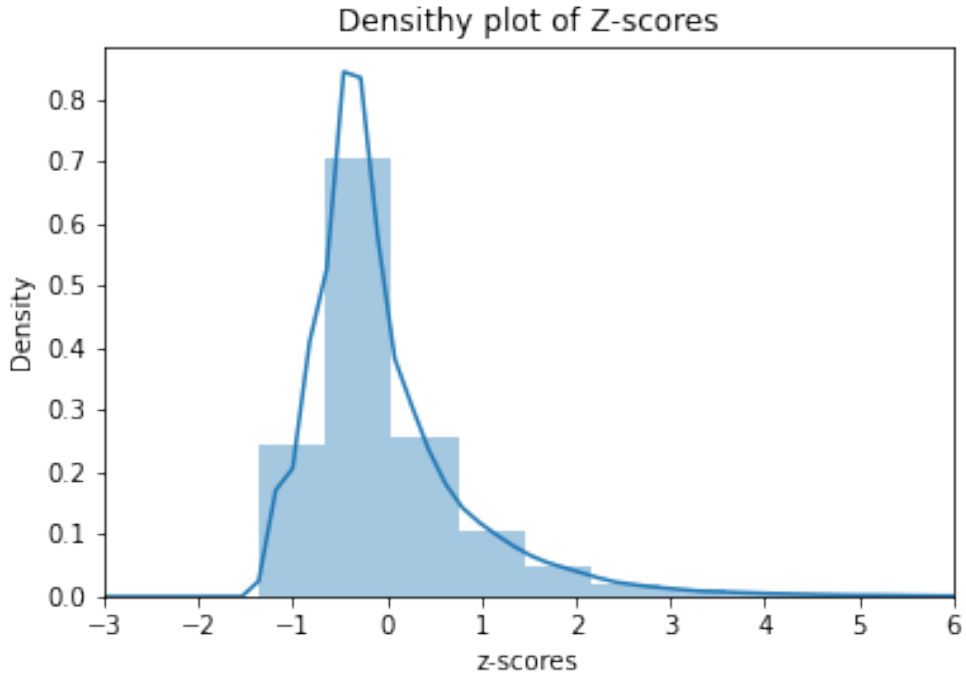

Figure 6: Density plot of z-scores

From the above two plots, we observed that most of the values are near the mean of the z-scores and there is a standard deviation of +/-1. We set the the value of threshold to 4 in our analysis as there are some subsequent amount of data examples whose z-score is around +3.
Table 3 shows Attributes (top 10) having most number of outliers.

Table 3: Attributes (Top 10) having most number of Outliers

Attributes Name	**No. of Outliers**
Number of Vasectomies Conducted (Public + Pvt.)	66
Number of Home deliveries	60
Number of home deliveries attended by Non SBA trained (trained TB/Dai)	59
Number of C-section deliveries conducted at private facilities	58
Number of Major Operations	54
Condom pieces distributed	53
Adverse Events Following Imunisation (Others)	51
Number of Women Discharged under 48 hours of delivery in public facilities	46
IUCD insertions done (pvt. facilities)	46
Number of home deliveries attended by SBA trained (Doctor/Nurse/ANM)	44

Boxplot on the attributes having highest and lowest number of outiler

Z score results shows that every column contains decent amount of outliers, but attribute 'Number of Vasectomies Conducted (Public + Pvt.) ' contains the most outliers and attribute 'Sex Ratio at birth (Female Live Bitrths/ Male Births *1000)' contains least amount of outliers.
To visualize these results we plot a boxplot corresponding these two attributes, this is shown in Figure 7.

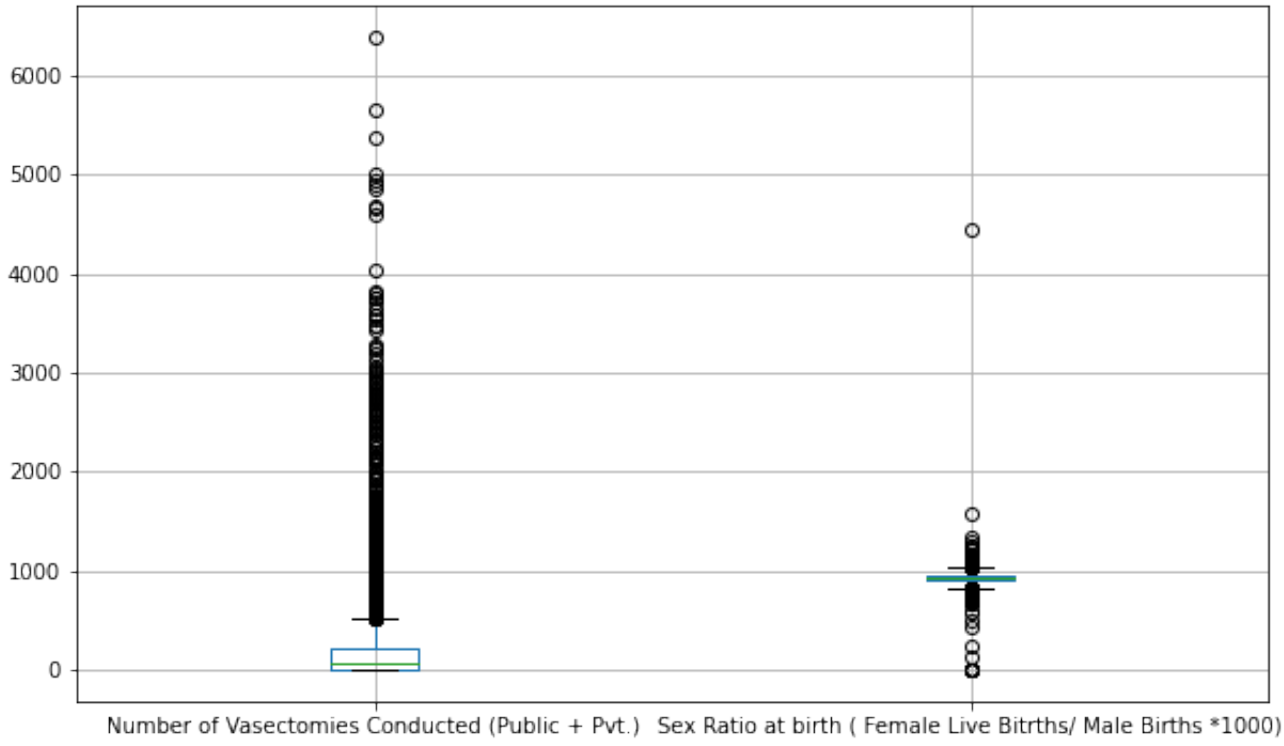

Figure 7: Boxplot on the attributes having highest and lowest number of outiler

8.5.3 Outlier Treatment

We have identified the outliers, we need to treat them. We implemented two types of outlier techniques. From both the tests we draw different conclusions.
Attribute 'Adverse Events Following Imunisation (Others)' is on top of IQR analysis. On checking the values of this attribute we found that most of the values of this attribute is zero and because of this IQR is pointing it out as outlier. So from IQR analysis we found the attributes which has most of the values zero. Since these values are marked as outliers we can conclude these zero values are actually the missing values in the data set marked as value zero.
We confirmed these conclusion by training the regression model after dropping the outliers mentioned by the IQR and our prediction accuracy dropped. So we used z-score results from dropping the outliers.

Number of outliers in Z score test is significantly less than that of in IQR test. So we can say that the outliers pointed by the z-score test are real outlier of the dataset and these are needed to be dropped from the dataset.
We confirmed this conclusion by training the regression model and our accuracy improved.

9 Prediction

We have three problem statements:

1. Prediction of the number of child births in each district of the country.
2. Prediction of the number of still births in each district of the country.
3. Prediction of the number of infant deaths in each district of the country.

Since the all three problems are of regression we applied different models of regression - Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression and Random Forest. We have learned model for each of the problem statement separately with different indepenent and dependent variables. We have discussed them in the following three sections:

10 CASE 1: PREDICTION OF NUMBER OF BIRTHS

Here we predicted the number of child births in each district of India in a year from the data set available. We first learned the Linear Regression model on our data.

We have 47 attributes in the data set out of which we have dropped the categorical attribute - "Indicator" from our dataset. We have divided our variables into independent and dependent set as follows:

Independent\Target attribute: 'Total Number of reported live births'
Dependent attributes: Rest 45 attributes of the dataset.

We used sklearn library for implementing the Linear Regression model.

Linear_Regression.py

```
#importing Linear Regression from sklearn
from sklearn.linear_model import LinearRegression as LR

# Creating instance of Linear Regresssion
lr = LR()

# Fitting the model
lr.fit(train_x, train_y)
```

For Linear Regression model to work good and to give good prediction rates, the data should follow some assumptions/characterstics. These characteristics are as follows:

Assumptions of Linear Regression
Assumptions made while applying the Linear Regression model are:

- **Linear relationship:** Relationship between response and feature variables should be linear.

- **Little or no multi-collinearity:** It is assumed that there is little or no multicollinearity in the data. Multicollinearity occurs when the features (or independent variables) are not independent from each other.
- **Little or no auto-correlation:** Another assumption is that there is little or no autocorrelation in the data. Autocorrelation occurs when the residual errors are not independent from each other.
- **Homoscedasticity:** Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables.
- **Normality:** The errors are generated from a Normal distribution (of unknown mean and variance, which can be estimated from the data).

After learning the model, to check whether the accuracy of the model can be further improved or not, we need to verify whether all the assumptions of linear regression about the data set holds true. If some assumption do not holds true then we need to transform our data set.

10.1 Interpreting Regression Coefficients for Linear Relationships

```
1 # Coefficients of the learned model (that is the weight vector w)
2 np.set_printoptions(suppress=True)
3 coeff = (lr.coef_)
4 print(coeff)
```

```
[-6.67174332  0.036041   -0.02373135  0.13172816  0.00940817 -0.00839384
 -0.03291967 -0.1718669   0.28971957  0.09207437  0.45351398 -0.07844079
  0.03404174  0.49329659  0.0256964   0.01767731  0.22004293 -0.45589662
  0.13425758  0.17781494 11.12483724  0.16010267  0.04308757 -1.20197753
 -1.37039941  1.36645111  0.09024551 -0.08445562  0.00705661 -0.00007458
  0.14291767  0.00347885 -0.03131908  0.21565688 -0.14963718 -0.01346718
  0.03551679 -0.22300692 -0.01655261 -0.00979021  0.97302085  0.01121614
  0.00053409 -0.01220319 -0.01130678 -0.01138641]
```

Figure 8: Coefficients learned by the model

The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.

The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. This property of holding the other variables constant is crucial because it allows you to assess the effect of each variable in isolation from the others. Following figure shows the values of the coefficients learned by the model.

10.2 Coefficient Plot

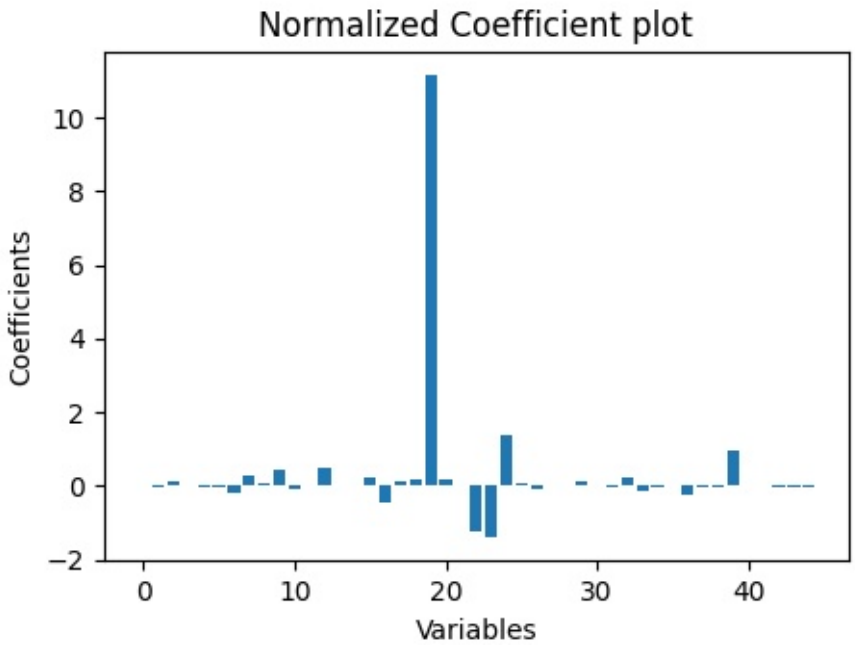

Figure 9: Normalized Coefficient plot

Coefficient plot tells us that our target variable highly depends on 1 attribute and it is positve correlated. Plot shows that coefficient values of 40 attributes is very less, and only 5 attributes majorly contribute to our model's prediction. These attributes are:

	Most Influencial Attributes
0	Sex Ratio at birth (Female Live Bitrths/ Male...
1	Number of Vasectomies Conducted (Public + Pvt.)
2	Number of Tubectomies Conducted (Public + Pvt.)
3	Total Sterilisation Conducted
4	Total Number of Infant Deaths reported

Figure 10: Attributes that influenced the model most

Target attribute is positively correlated with the attributes - Sex Ratio at birth, Total Sterilisation Conducted, Total Number of Infant Deaths reported and is negatively correlated with attributes - Number of Vasectomies Conducted and Number of Tubectomies Conducted. This makes sense too as Vasectomies and Tubectomies makes humans infertile and hence will reduce the number of child births.

10.3 Residual vs Fitted

Residuals are the difference between the observed values and the fitted values. We need to plot the residuals, check their random nature, variance, and distribution for evaluating the model quality. This is the visual analytics needed for goodness-of-fit estimation of a linear model.
If the plot of residual vs fitted do not show any pattern then ideally it suggests linearity in the data set. Since the figure do not show any pattern so it satisfies linearity.

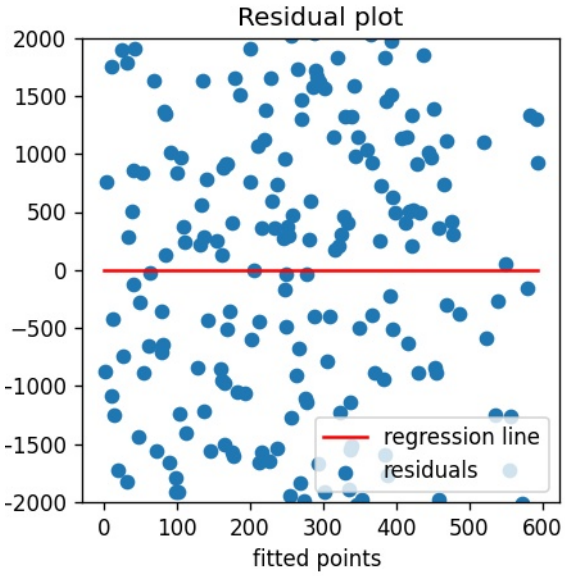

Figure 11: Residual Plot

The points in a residual plot are randomly dispersed around the horizontal axis and they are distributed uniformly randomly around the zero x-axes and do not form specific clusters. Hence a linear regression model is appropriate for the data.
When we plot the fitted response values (as per the model) vs. the residuals, we clearly observe that the variance of the residuals remains constant with response variable magnitude. Therefore, the problem respect homoscedasticity.

10.4 Normality Test

To check the assumption of normality of the data generating process, we plot the histogram and the Q-Q plot of the normalized residuals.

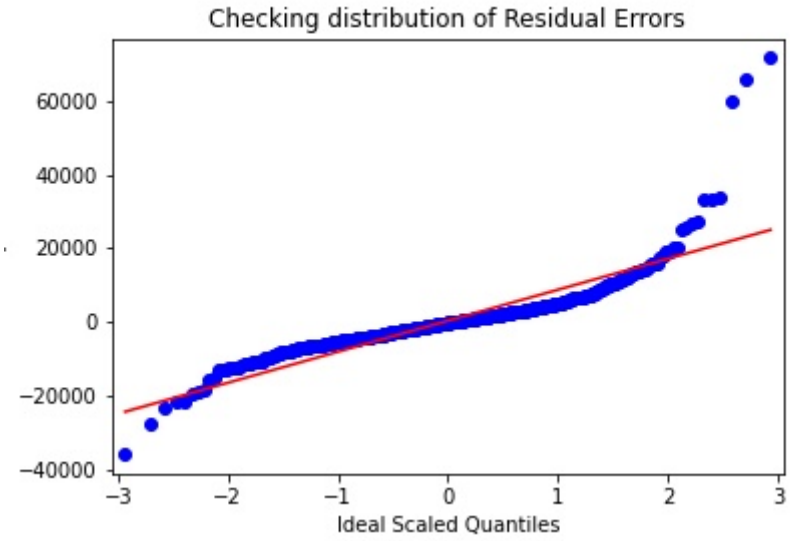

Figure 12: QQ Plot for residuals

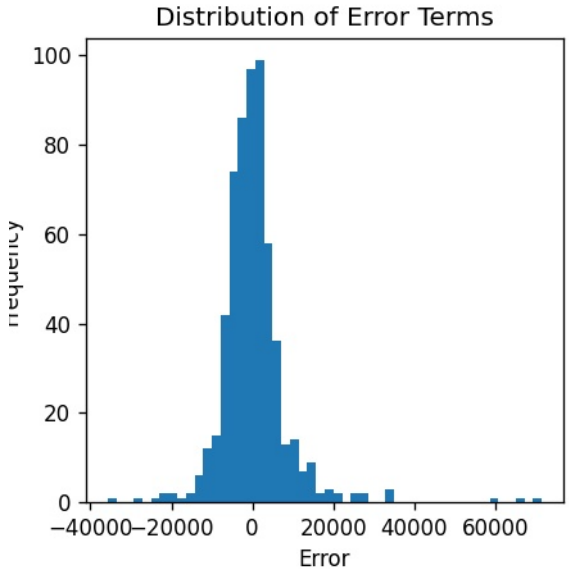

Figure 13: Histograms for residuals

Since the QQ plot is almost linear and fits accross the regression line and histogram plot follows normal distribution, our data set follows normality.

10.5 Evaluation of Model

We evaluated the performance of our model from the following measures:

- **RMSE:** MSE is calculated by the sum of square of prediction error which is real output minus predicted output and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. Root Mean Square Error(RMSE) is the square root of MSE.
- **MAE:** MAE is taking the sum of absolute value of error. MSE gives larger penalisation to big prediction error by square it while MAE treats all errors the same.
- **R Square (Coefficient of Determination):** R Square measures how much of variability in dependent variable can be explained by the model. It is square of Correlation Coefficient(R) and that is why it is called R Square.

Error of the linear regression model for above mentioned measures are shown in the Table 4 below.

Table 4: Results of Linear regression model for Child Births

Measure	**Error**
Training Mean Absolute Error	2828.6132176902
Test Mean Absolute Error	5109.2537316244
Training Root Mean Square Error	5362.25745574993
Testing Root Mean Square Error	8402.35904407914
Training Accuracy(R-square)	95.7276759961367
Testing Accuracy(R-square)	93.3642138572999

So with Linear Regression we are getting 93% prediction accuracy, which is quite good. Since Random Forest gives better results most of the time as compared to linear regression, we learned that model too.

11 Random Forest Model for Predicting Births

Random_Forest.py

```
from sklearn.ensemble import RandomForestRegressor

# create object of Random Forest
regressor = RandomForestRegressor(n_estimators = 25,ran-
dom_state = 0)

# fit the regressor with x and y data
regressor.fit(train_x, train_y)

# predicting for train and test set
train_predict = regressor.predict(train_x)
test_predict = regressor.predict(test_x)
```

Figure 14 shows the plot of actual target values vs the predicted values by the model.

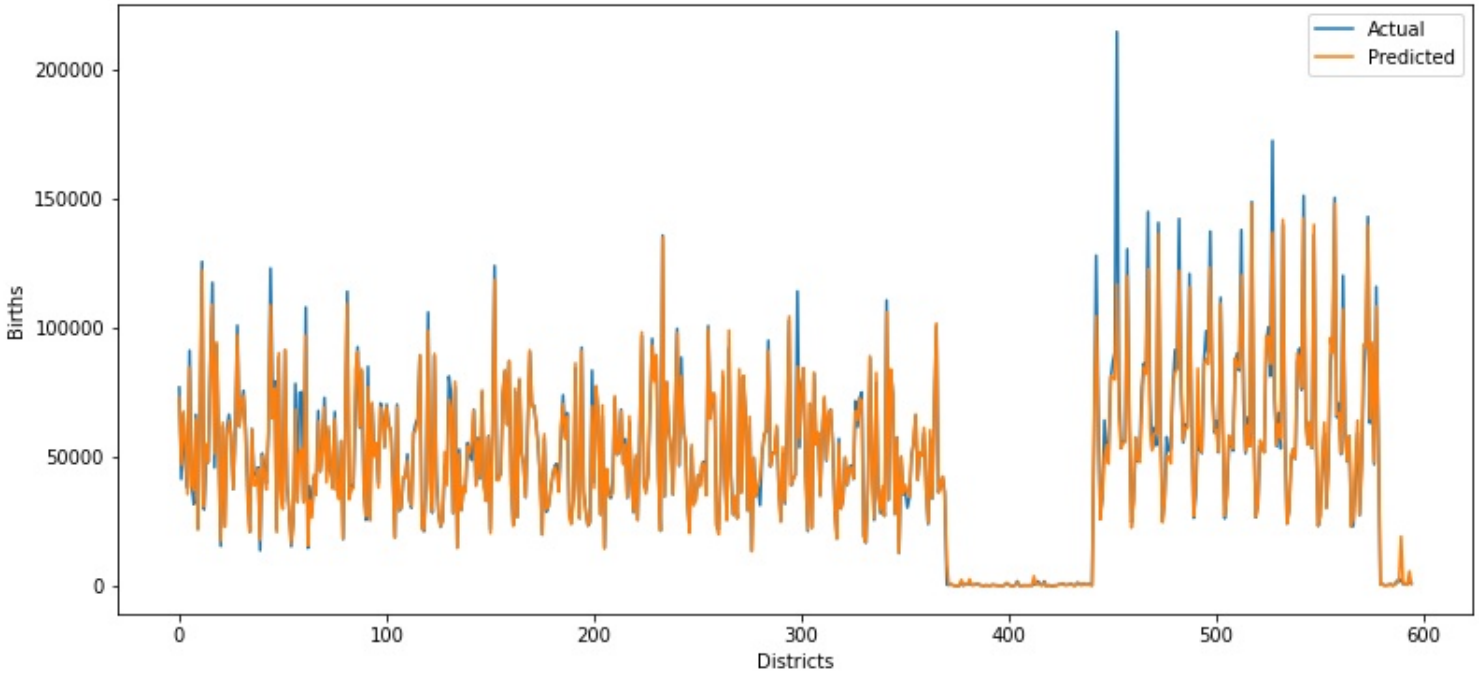

Figure 14: Actual vs Predicted plot for target variable

Table 5 shows the name of the features/attributes on which our target attribute is highly dependent.

Table 5:

Important Features for Random Forest Model
Number of fully immunized children (9-11 months)
Number of pregnant women received 3 ANC check ups
Number of Infants given BCG
Total number of pregnant women Registered for ANC
Number of Infants given OPV 0 (Birth Dose)
Deliveries Conducted at Public Institutions
TT2 or Booster given to Pregnant women (numbers)
Number of Infants given Measles
Number of home deliveries attended by Non SBA trained (trained TB/Dai)
Number of Infants given DPT1

Table 6: Results of Random Forest model for Child Births

Measure	**Error**
Training Mean Absolute Error	574.652419301165
Test Mean Absolute Error	2229.61431932773
Training Root Mean Square Error	1564.56984390149
Testing Root Mean Square Error	5916.33023171844
Training Accuracy(R-square)	99.6362870467858
Testing Accuracy(R-square)	96.7100088929496

Table 6 shows the results of evaluation measures applied on the random forest model learned. Model is having 96% prediction rate, this is 3% more than the linear regression.

12 CASE 2: PREDICTION OF NUMBER OF STILL-BIRTHS

Here we predicted the number of still births in each district of India in a year from the data set available. The Linear Regression gives an accuracy of 92% whereas Random Forest gives an accuracy of 98%.

We have 47 attributes in the data set out of which we have dropped the categorical attribute - "Indicator" from our dataset. We have divided our variables into independent and dependent set as follows:

Independent\Target attribute: 'Total Number of reported Still Births'
Dependent attributes: Rest 45 attributes of the dataset.

12.1 Linear Regression

The Linear Regression model used is same as the one used in CASE 1: Prediction of Number of Births.

Error and Accuracy of the linear regression model for the measures are shown in the Table 7 below.

Table 7: Results obtained for Still Births

Measure	**Error**
Training Mean Absolute Error	1023.99261949344
Test Mean Absolute Error	2390.04152966319
Training Root Mean Square Error	2235.31247540752
Testing Root Mean Square Error	4821.80489352115
Training Accuracy(R-square)	88.124395801068
Testing Accuracy(R-square)	92.4555099954537

So with Linear Regression we are getting 92% prediction accuracy, which is quite good.

The most Influential Attributes as per the Linear Regression model for the predictions of Still Births are mentioned in the Table 8 below.

Target attribute is co-related related to the attributes representing various types of deliveries conducted. The most influential attribute is Number of Home deliveries. This make sense too since there are complications attached with Home deliveries. It is clearly seen from the table that Still Births depend mostly on the no of Deliveries conducted, and also makes sense too. The other attributes : Sex Ratio at birth, Number of Vasectomies/ Tubectomies, Total Sterlisation Conducted are also highly related to the prediction of Still Births.

Table 8:

Most Influencial Attributes
Number of Home deliveries
Number of home deliveries attended by SBA trai...
Number of home deliveries attended by Non SBA ...
Institutional deliveries (Public Insts.+Pvt. I...
Total reported deliveries
Sex Ratio at birth (Female Live Births/ Male...
Number of Vasectomies Conducted (Public + Pvt.)
Number of Tubectomies Conducted (Public + Pvt.)
Total Sterilisation Conducted

Since Random Forest gives better results most of the time as compared to linear regression, we learned that model too.

12.2 Random Forest

Figure 15 shows the plot of actual target values vs the predicted values by the model.

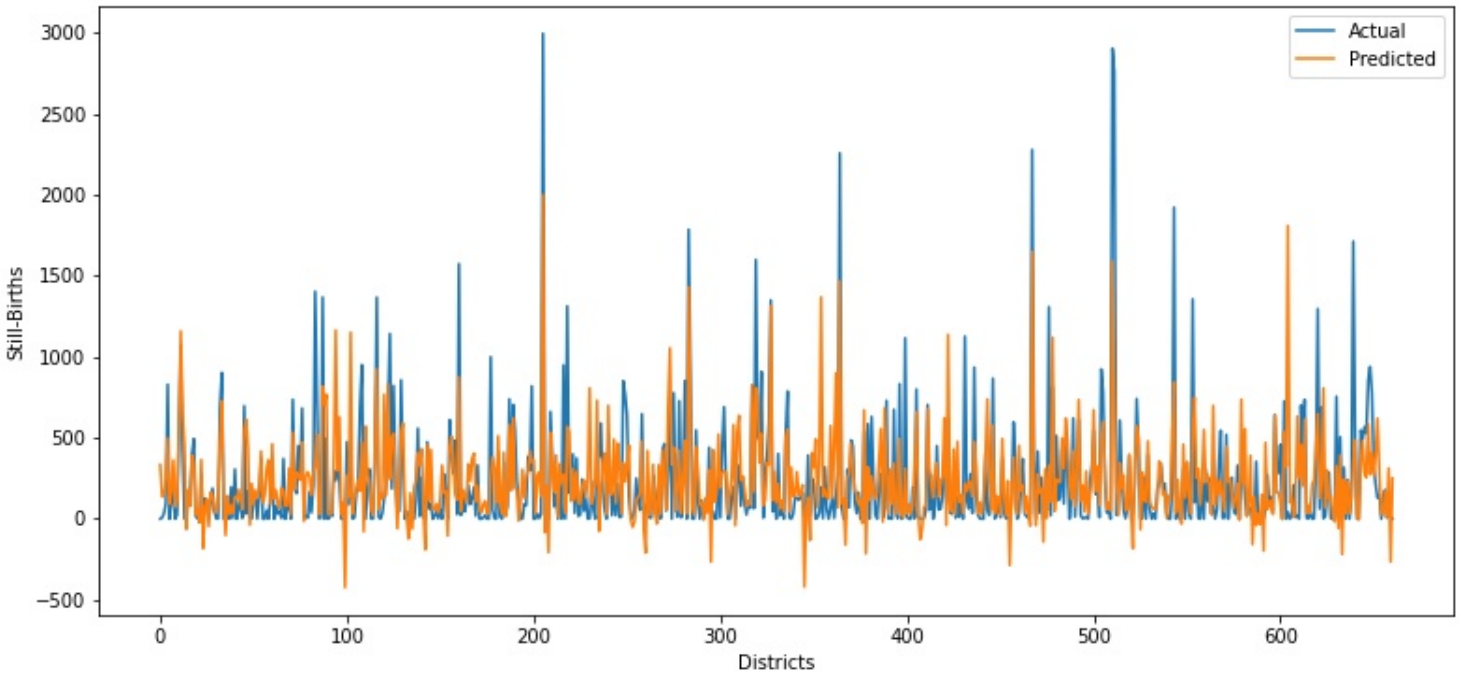

Figure 15: Actual vs Predicted plot for target variable

Table 9 shows the name of the features/attributes on which our target attribute is highly dependent.

Table 9:

Important Features for Random Forest Model
Number of Infants given BCG
Number of Newborns having weight less than 2.5 Kg
Total reported deliveries
TT2 or Booster given to Pregnant women (numbers)
Condom pieces distributed
Institutional deliveries (Public Insts.+Pvt. Insts.)
Total Number of reported live births
Total number of pregnant women Registered for ANC
Number of Pregnant women registered within first trimester
Number of Minor Operations

Table 10 shows the results of evaluation measures applied on the random forest model learned. Model is having 98% prediction rate, this is 6% more than the linear regression.

Table 10: Results of Random Forest model for Still Births

Measure	**Error**
Training Mean Absolute Error	89.88106998654105
Test Mean Absolute Error	541.9993948562783
Training Root Mean Square Error	396.0021534095236
Testing Root Mean Square Error	2139.573840681855
Training Accuracy(R-square)	99.62728718459579
Testing Accuracy(R-square)	98.51452557349563

13 CASE 3: PREDICTION OF NUMBER OF INFANT-DEATHS

Here we predicted the number of infant deaths in each district of India in a year from the data set available.
The Prediction accuracy using Linear Regression is 49% and it increases to 69% for Random Forest. The attributes in the data available on HMIS portal is good for prediction for Births and Still Births but is not very useful in predicting the no of infant deaths.

We have 47 attributes in the data set out of which we have dropped the categorical attribute - "Indicator" from our dataset. We have divided our variables into independent and dependent set as follows:

Independent\Target attribute: 'Total Number of reported Infant Deaths'
Dependent attributes: Rest 45 attributes of the dataset.

13.1 Linear Regression

Figure 16 below shows the Actual vs Predicted plot for target variable using Linear Regression.

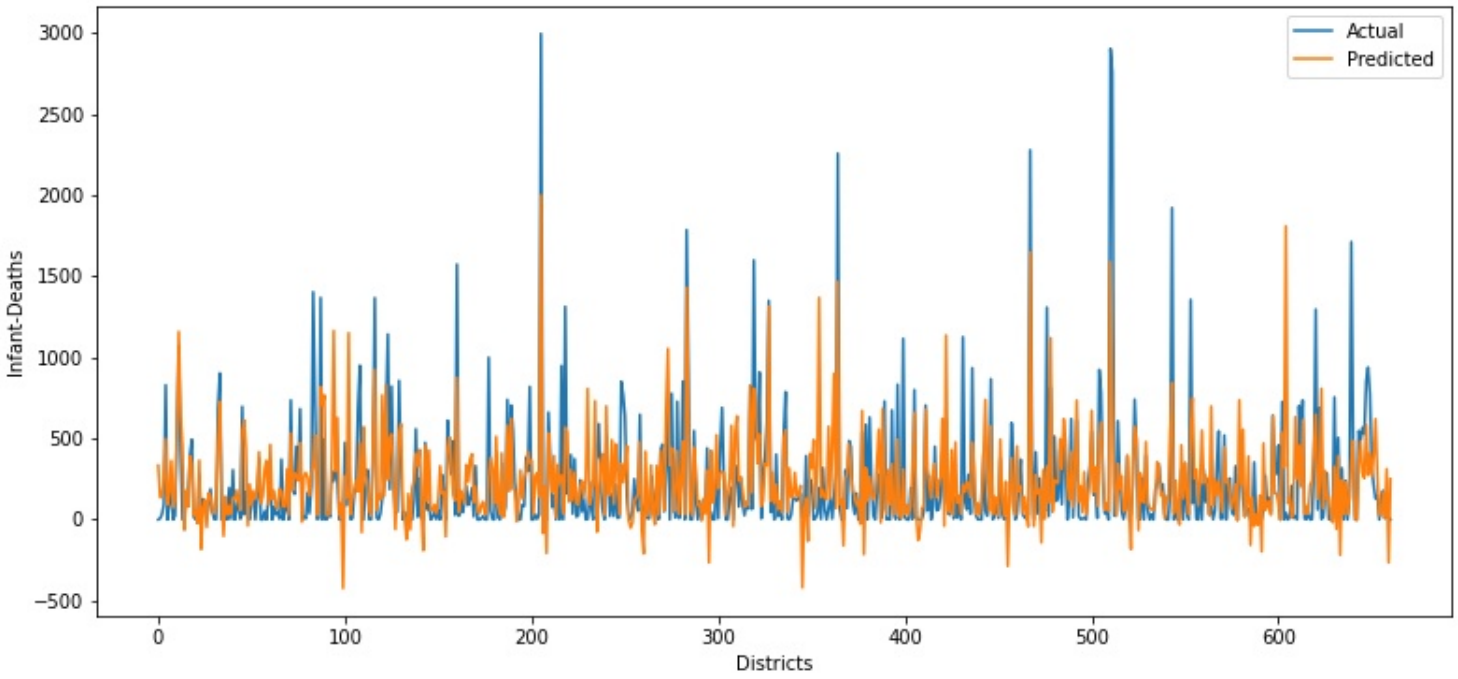

Figure 16: Actual vs Predicted plot for target variable for Linear Regression

Error and Accuracy of Linear Regression for prediction of Infant deaths is shown in the below table 11.

Table 11: Results of Infant Deaths by Linear Regression

Measure	Error
Training Mean Absolute Error	176.98444922111506
Test Mean Absolute Error	171.44992601028855
Training Root Mean Square Error	284.300609512833
Testing Root Mean Square Error	265.06288848278155
Training Accuracy(R-square)	48.09697806858291
Testing Accuracy(R-square)	49.51743097935387

The most important features in the prediction of infant deaths using Linear Regression Model are mentioned in the figure 17.

	Most Influencial Attributes
0	Number of home deliveries attended by SBA trai...
1	Total reported deliveries
2	Number of C-section deliveries conducted at pu...
3	Number of C-section deliveries conducted at pr...
4	Number of Vasectomies Conducted (Public + Pvt.)
5	Number of Tubectomies Conducted (Public + Pvt.)
6	Total Sterilisation Conducted
7	IUCD insertions done (pvt. facilities)
8	Adverse Events Following Imunisation (Others)

Figure 17: Important features of Infant Death Linear Regression

13.2 Analysing Failure of Linear Regression

For Linear Regression model to work and to give good prediction rates, the data should follow some assumptions/characteristics. The main characteristics are Normality, Homoscedasticity, Linear Relationship and Multi-collinearity. The details about these assumptions are discussed earlier in this report. We will now examine these characteristics in this case to reason about the failure of Linear Regression.

13.2.1 Coefficient plot

Coefficient plot tells us that out target variable is highly dependent on 2 variables and it is positive correlated to one variable while negative correlated to another variable.

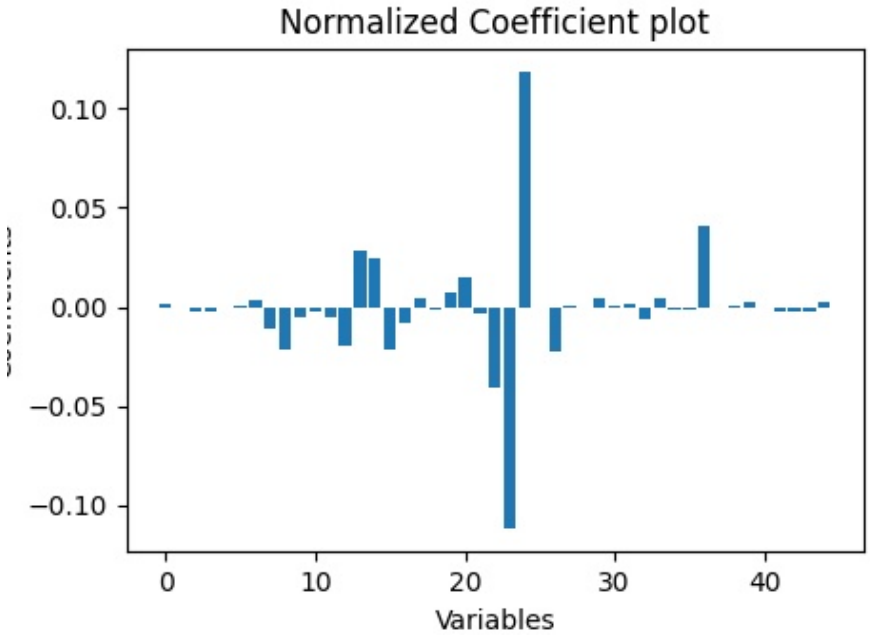

Figure 18: Normalized Coefficient Plot

Plot shows coefficients values of 36 attributes is very less and only 9 attributes majorly contribute to our models prediction. These attributes are:

	Most Influencial Attributes
0	Number of home deliveries attended by SBA trai...
1	Total reported deliveries
2	Number of C-section deliveries conducted at pu...
3	Number of C-section deliveries conducted at pr...
4	Number of Vasectomies Conducted (Public + Pvt.)
5	Number of Tubectomies Conducted (Public + Pvt.)
6	Total Sterilisation Conducted
7	IUCD insertions done (pvt. facilities)
8	Adverse Events Following Imunisation (Others)

Figure 19: Attributes that influence the model most

13.2.2 Residual plot

Residuals are the difference between the observed values and the fitted values. We need to plot the residuals, check their random nature, variance, and distribution for evaluating the model quality. This is the visual analytics needed for goodness-of-fit estimation of a linear model.

If the plot of residual vs fitted do not show any pattern then ideally it suggests linearity in the data set. Since the figure do not show any pattern so it satisfies linearity.

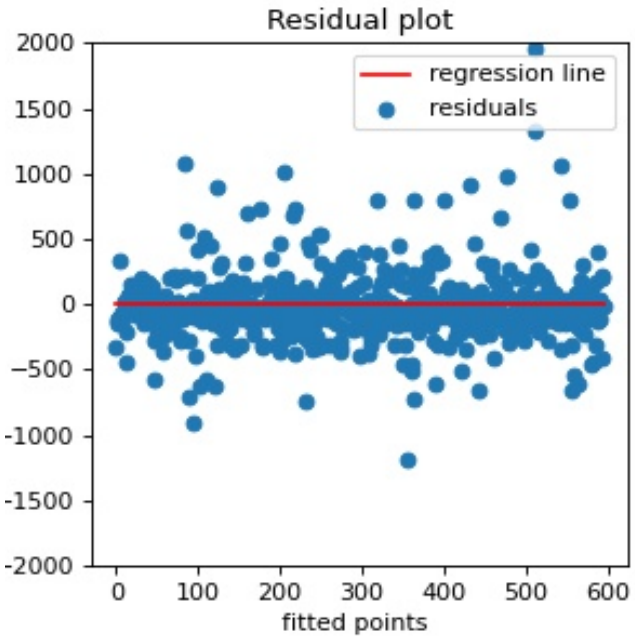

Figure 20: Residual Plot

When we plot the fitted response values (as per the model) vs. the residuals, we clearly observe that the variance of the residuals remains constant with response variable magnitude. Therefore, the problem respect homoscedasticity. But residuals are not distributed uniformly, they are pretty symmetrically distributed, tender to cluster towards the middle of the plot.

13.2.3 Normality Test

To check the assumption of normality of the data generating process, we plot the histogram and the Q-Q plot of the normalized residuals.

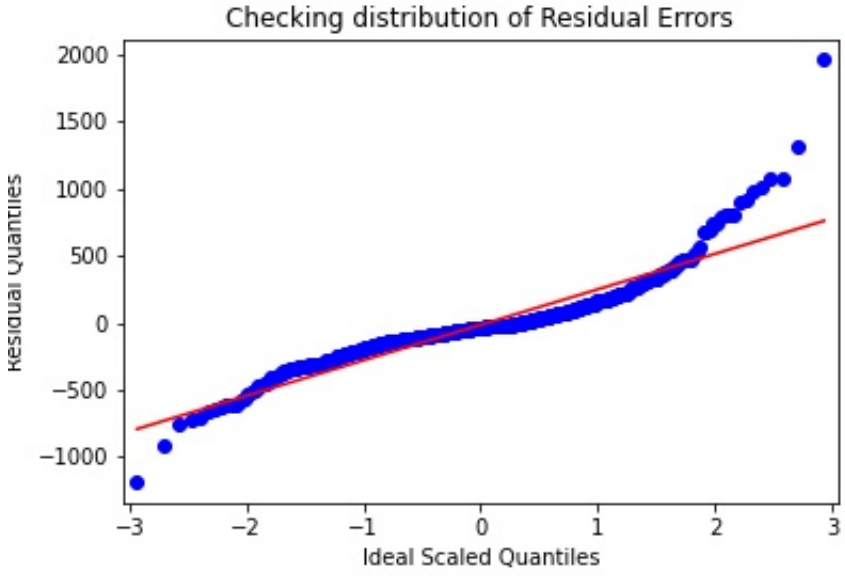

Figure 21: QQ Plot for residuals

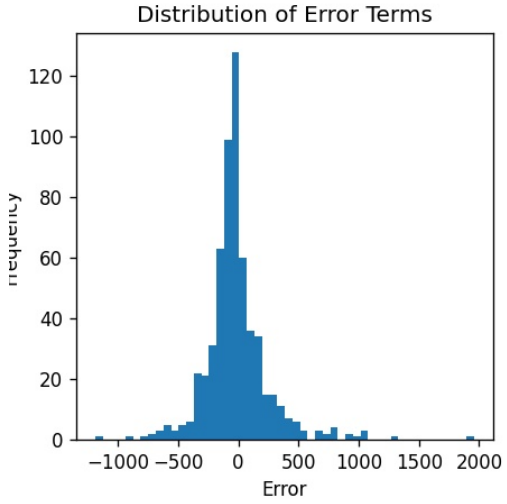

Figure 22: Histograms for residuals

From Histogram and QQ-plot of residuals we can say that data-set in normal for target variable-infant deaths.

Since data satisfies condition of normality, homoscedasticity, Linear relationship. Last thing left is check for Multicolinearity. Variance Inflation Factor test is best suited for it. if VIFk=1, variable k is not correlated to any other independent variable. As a rule of thumb, multicolinearity is a potential problem whn VIFk is greater than 4; and a serious problem when it is greater than 10.

13.2.4 VIF

After running the VIF test, it was seen that 27 out of 46 attributes have VIF greater than 10, it shows that our data suffers from heavy multicollinearity.

	feature	VIF
0	Literate Persons	92467.995672
1	Total Number of reported live births	35924.277775
2	Total reported deliveries	31186.143543
3	Institutional deliveries (Public Insts.+Pvt. I...	21204.626695
4	Marginal workers Persons	8631.852523
5	Total Sterilisation Conducted	8035.113268
6	Number of Tubectomies Conducted (Public + Pvt.)	7671.763229
7	Non-workers Persons	986.105534
8	Number of home deliveries attended by Non SBA ...	828.966345
9	Number of home deliveries attended by SBA trai...	800.939126
10	Number of Infants given DPT2	393.108369
11	Number of Infants given DPT3	236.767165
12	Number of Home deliveries	195.271135
13	Number of Infants given DPT1	153.434271
14	Main workers Persons	107.136348
15	Number of Infants given BCG	57.573416
16	Number of Infants given Measles	51.076832
17	Total number of pregnant women Registered for ANC	34.473289
18	Number of pregnant women received 3 ANC check ups	31.956168
19	TT2 or Booster given to Pregnant women (numbers)	29.394356
20	Number of Vasectomies Conducted (Public + Pvt.)	23.762352
21	Deliveries Conducted at Public Institutions	20.375933
22	Number of Infants given OPV 0 (Birth Dose)	20.028542
23	Number of fully immunized children (9-11 months)	13.840468
24	Number of New Borns Breast Fed within 1 hour	13.000577
25	Number of Pregnant women registered within fir...	12.573843
26	Total Number of reported Still Births	10.430285

Figure 23: VIF1

The model was trained again after dropping the attributes having VIF greater than 10, the accuracy was dropped to 37%. The VIF was then calculated for the remaining attributes.

	feature	VIF
0	Condom pieces distributed	31186.143543
1	Oral Pills distributed	21204.626695
2	Total Number of MTPs (Public) reported	828.966345
3	Total Number of Abortions (Spontaneous/ Induc...	800.939126
4	Sex Ratio at birth (Female Live Bitrths/ Male...	195.271135
5	Number of Pregnant women given 100 IFA tablets	34.473289
6	Number having severe anaemia (Hb<7) treated at...	31.956168
7	Number of Women Discharged under 48 hours of d...	29.394356
8	IUCD Insertions done (public facilities)	20.375933
9	Number having Hb level<11 (tested cases)	12.573843
10	Number of Minor Operations	10.430285

Figure 24: VIF2

Since 11 out of the remaining 19 attributes are having VIF values greater than 10 in the remaining data-sets. We can conclude that attributes of the data-set is highly correlated

with each other and accuracy of regression model will always suffer. Let us apply Random Forest on this problem.

13.3 Random Forest

Figure 25 shows that actual target values vs the predicted values by the model.

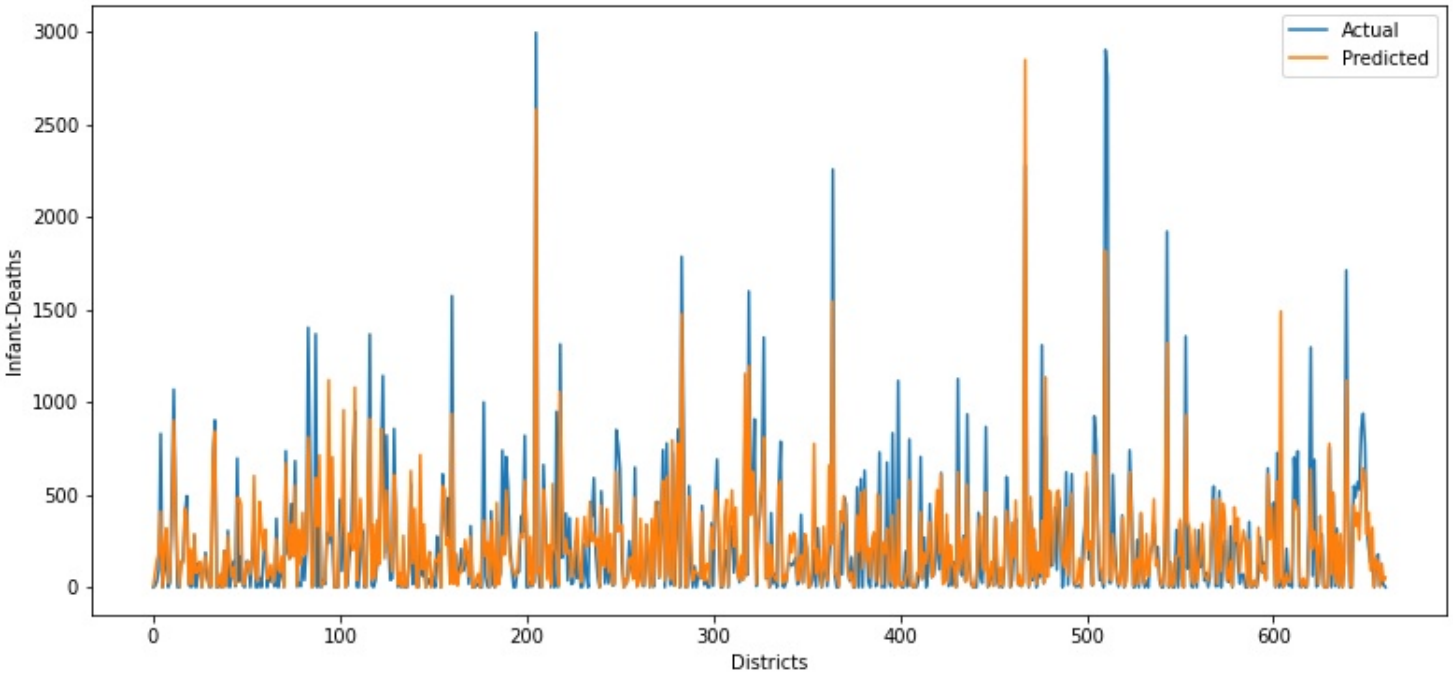

Figure 25: Actual vs Predicted plot for target variable for Random Forest

Table 12 below shows the results of evaluation measures applied on the Random Forest model learned. Model is having 70% accuracy which is greater than linear regression model by 21%.

Table 12: Results obtained for Infant-Deaths by Random Forest

Measure	**Error**
TTraining Mean Absolute Error	44.47679676985195
Test Mean Absolute Error	113.2001815431165
Training Root Mean Square Error	87.23965230976026
Testing Root Mean Square Error	201.14927401387413
Training Accuracy(R-square)	95.12455070018713
Testing Accuracy(R-square)	70.56988968985102

Table 13 below shows the name of attributes/ features on which our target attribute is

highly dependent.

Table 13: Random Forest Important Features

Most Influencial Attributes
Total Number of reported Still Births
Total Number of Abortions (Spontaneous/ Induced) Reported
Total Sterilisation Conducted
Total Number of reported live births
Number of C-section deliveries conducted at public facilities
Oral Pills distributed
Number of Major Operations
Number of home deliveries attended by Non SBA trained (trained TB/Dai)
IUCD Insertions done (public facilities)
Marginal workers Persons

14 Results

The training and testing score using different Machine Learning Techniques:

Model Name	Problem Statement	Training Score	Testing Score
Linear Regression	Child Births	0.957267	0.933956
Random Forest	Child Births	0.993489	0.960923
Linear Regression	Still Births	0.881243	0.924555
Random Forest	Still Births	0.996272	0.985145
Linear Regression	Infant Deaths	0.480969	0.495174
Random Forest	Infant Deaths	0.9512345	0.705698

Table 14: Prediction Accuracy

The Random Forest Regressor Model performs the best in all the three problem statements.

15 Discussion

In this project we have seen that Random forest works very well in the prediction of no of child births and still births against linear regression. The main reasons can be understood from the below points:
Decision Trees are great for obtaining non-linear relationships between input features and the target variable.
Random Forest is an ensemble of decision tress. This is to say that many trees, constructed in a certain "random" way form a Random forest.

The averaging makes a Random Forest better than a single decision tree hence improves its accuracy and reduces over-fitting.
A prediction from the Random Forest Regressor is an average of the predictions produced by the trees in the forest.

16 Future Directions

There are several health schemes run by Government but the data is not available in the public domain for the most of the schemes. If in future we come across any data-sets that might be helpful further in our predictions we plan to improve prediction for the infant deaths and also for the overall project.

17 Important Links and References

1. HMIS Database:`https://www.nrhm-mis.nic.in/hmisreports/frmstandard_reports.aspx`

2. Census Data 2001:`https://censusindia.gov.in/DigitalLibrary/TablesSeries2001.aspx`

3. Census Data 2011:`https://censusindia.gov.in/2011census/population_enumeration.html`

4. GitHub Link for the project: `https://github.com/kuldeeps5/DataMiningFinalProject`

5. Implementation of Lasso and Ridge Regression:
`https://analyticsindiamag.com/hands-on-implementation-of-lasso-and-ridge-regression/`

6. Random Forest Regression:
`https://www.geeksforgeeks.org/random-forest-regression-in-python/`