Infant Mortality Data Analysis; SDG 3: Team Matrix

Anuarg Arora[1]
Stony Brook University
SBU-ID: 111425080

Keshav Gupta[1]
Stony Brook University
SBU-ID: 111464733

Renu Rani[1]
Stony Brook University
SBU-ID: 111482474

Shayan Ray[1]
Stony Brook University
SBU-ID: 111424665

Abstract

In this project, we have analyzed the infant mortality data from Center for Disease Control and Prevention (1,) and designed a framework to predict the risk of infant death and provide reference to similar pregnancy cases in the past to help the doctors in taking informed decision. Big data technologies such as large scale machine learning, dimensionality reduction, clustering, and similarity search have been employed to make this happen.

1 Introduction

Infant mortality is defined as the death of an infant before his or her first birthday. The infant mortality rate is defined as the number of infant deaths for every 1000 live births. The total infant mortality in the year 2016 was 3.15 million. US lags behind other wealthy nations on infant mortality. In 2010, the U.S. infant mortality rate was 6.1 infant deaths per 1,000 live births, and the United States ranked 26^{th} in infant mortality among Organization for Economic Co-operation and Development countries. (5,). Infant mortality rate is an indication of population health, poverty, and socioeconomic status of a country and quality of health services in a country. To ensure development of a country and healthy population, it is important to reduce infant mortality.

In this project, we analyzed the infant mortality data and designed a framework to improve the overall health of a country by predicting the risk of infant death and find similar cases for reference.

Section 2 describes the Sustainable Development Goal and Background of the project. The data-set used in the project is described in section 3. Section 4 goes over the methods used in the project. In Section 5 we share the results of our analysis. Section 6 and 7 discusses the inference and the main contribution of our work. Finally, in Section 8, we enumerate the team member contributions.

2 Sustainable Development Goal

The sustainable development goal we aimed to tackle in our project is Sustainable Development Goal 3: *Ensure healthy lives and promote well-being for all at all ages* (3,). More specifically, we focused on SDG Goal 3.2 which aim to reduce:

1. neonatal mortality to as low as 12 per 1,000 live births and
2. under-5 mortality to as low as 25 per 1,000(by 40%) live births.

In order to achieve this goal, it would be quite useful to have a framework which can perhaps answer the following questions:

1. Can we predict the risk of infant death?
2. If at risk, can we find similar pregnancy reference cases in the past to aid the medical practitioners in taking well-informed decisions?

The overall framework proposed use Spark data pipelines,Dimensionality Reduction, Clustering, Similarity Search, and Large-Scale Machine Learning.

What others have done:

Big data was used in Indiana to find the areas where there is a need of more focus so that overall infant mortality can be reduced.(7,)

Also, Department of Bio-medical Informatics, University of Pittsburgh is also analyzing infant mortality data to predict infant mortality in Allegheny county, Pennsylvania using Big Data (8,).

3 Data

The data has been taken from Center for Disease Control and Prevention (2,). Specifically, we are using 2014 ”Period Linked Birth-Infant Death Data Files”. The 2014 period linked birth/infant death data set includes 3 data files. The first file includes all US infant deaths which occurred in the 2014 data year linked to their corresponding birth certificates, whether the birth occurred in 2013 or 2014 - referred to as the numerator file. The second file contains information from the death certificate for all US infant death records which could not be linked to their corresponding birth certificates, referred to as the unlinked death file. The third file is the 2014 NCHS natality file for the US, which is used to provide denominators for rate computations. We merged all these 3 files and used the merged file for our analysis. Total number of observations in our data set are 4,021,418. It has approximately 300 features and is over 5 GB in size.

During actual prediction and clustering, the number of features finally used were around 148 after excluding irrelevant features such as method of disposition after death or some flags used to denote whether data is recorded or not.

Refer the following links to download the raw data (more than 5GB after zip extraction) and User Guide for detailed data description :

US-Data

US-Territories-Data

User Guide and Data Dictionary

4 Methods

Spark was used as the data pipeline for data loading, data analysis and feature engineering. Following this, the useful features were converted to numeric values, scaled, normalized, transformed with PCA and then used for prediction and clustering. Refer to Figure 1 for an overall summary of methods employed and Figure 2 for detailed spark machine learning pipelines leveraged for this task. Details follow in the sub-sections.

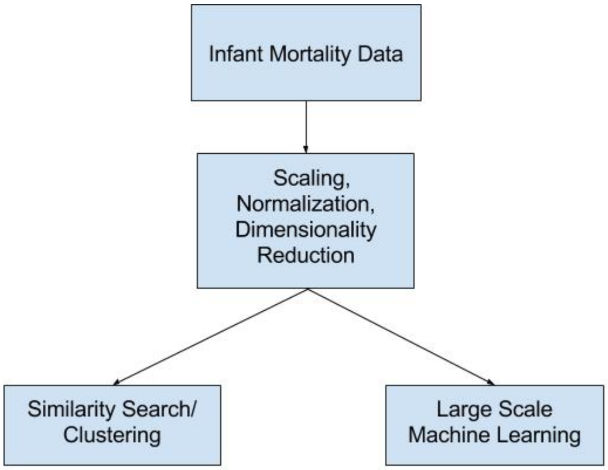

Infant Mortality Data
Scaling, Normalization, Dimensionality Reduction
Similarity Search/ Clustering
Large Scale Machine Learning

Figure 1: Methods Used

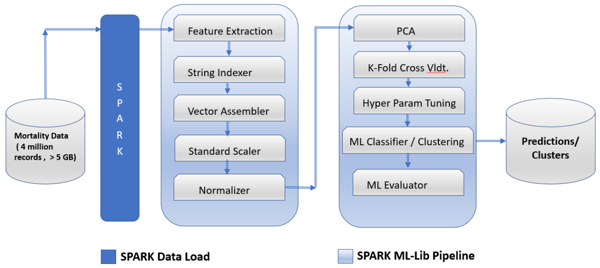


Figure 2: Implementation Framework

4.1 Spark

Apache Spark is an open source large-scale data processing tool. Spark SQL and dataframes were used for data processing. Specifically, pyspark.ml library was used for feature engineering, dimensionality reduction, clustering and for various machine learning models.

4.2 Feature Engineering

After data load and feature selection, those with string representations were converted to numeric form using String indexers. A corresponding row vector was formed from these numerical features and then scaled and normalized. This standardized the data and reduced the frequency of out-liers. Following this, PCA (Principal Component Analysis) was used to perform dimensionality reduction for lower the high-dimensionality(number of features) of our data. The reduction the number of features (from 148 to 50 in our case) would help to avoid over-fitting. That would lead to better prediction accuracy and a more generalized prediction model.

4.3 Exploratory Data Analysis

Some insights gained from data distribution were as follows:

1. Correlation among some key features: (Refer to Figure 2). It became evident from data that the infant death risk was not a clear correlation of a few factors but a complex mix of many parameters.

2. Fetal cause of death distribution: (Refer to Figure 3) The leading cause of death based on the ICD10 (International Cause of Death) was extreme pre-maturity(code=P072).

3. Distribution of some key features in the dataset: (Refer to Figure 4).

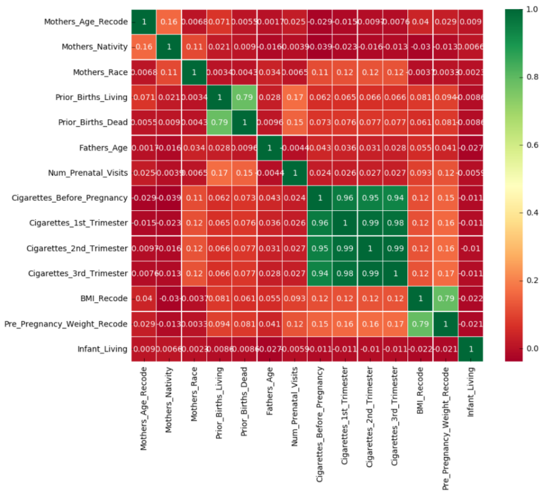


Figure 3: Correlation among few key features

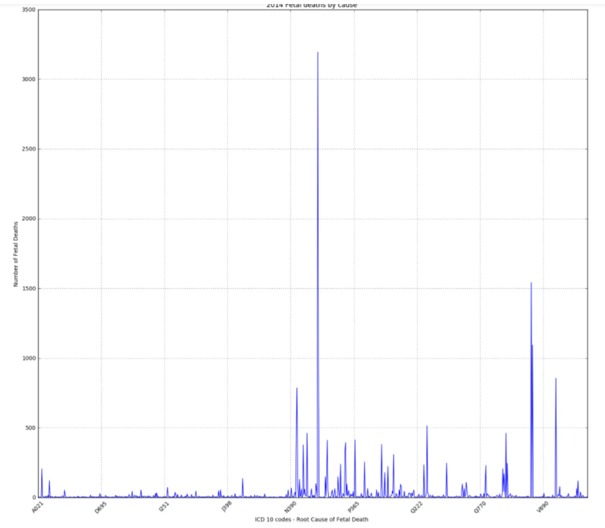


Figure 4: Root Cause of Fetal Death

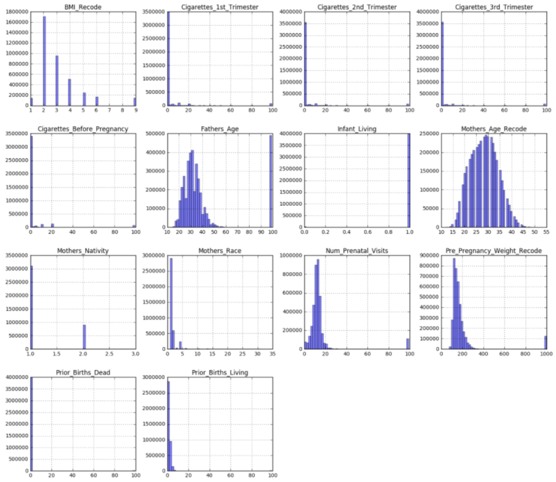


Figure 5: Sample Feature Distributions

4.4 Machine Learning

Supervised machine learning algorithms(in particular, classification algorithms) were used for predicting the risk of death for an infant. The label or the ground truth was the information whether the infant had demised or not, as indicated by ICD10 code(International Cause of Death). Multiple features like mother's attributes, pregnancy attributes, medical information, father's race, and education were factored as the features for prediction and clustering. On a similar note, features like the infant's burial information were removed as these were not going to influence the risk of death.

The original 5 GB data-set was split into 70 percent training and 30 percent testing data-sets. K-Fold Cross Validation was carried out only the training set of the data (70 percent) and followed by hyper-parameter tuning for each of the classification methodologies to produce a finely tuned prediction model. The 30-percent test data-set was held-out for final prediction and accuracy evaluation resulting in a more generalized model and useful accuracy metrics. Machine learning Classification Algorithms used were:

1. Logistic Regression
2. Gradient Boosted Trees
3. Random Forest
4. Decision Tree Classifier

4.5 Similarity Search

We used similarity search to find the similar cases which can aid the medical practitioner to take well informed decision. Cosine similarity is calculated on dimensionality reduced data i.e. after applying the PCA on the observations. Also, before finding the similarity, we are taking a random sample from all the observations and then we apply the similarity search on the randomly chosen sample.

4.6 Clustering Algorithm

K-Means (4,) clusters the observations in multiple clusters (k=20 in this case) such that the observations belong to the cluster with the nearest mean. For any new observation that arrives, it is clustered to the most similar existing cluster. It enables us to find similar reference cases in the past.

5 Results

All the results in HTML format are submitted along with this report archive. Refer to it for detailed dataset analysis and respective results. The following Machine Learning classifiers were employed:

1. Logistic Regression
2. Gradient Boosted Trees
3. Random Forest
4. Decision Tree Classifier

The results for the Machine Learning algorithms are presented in Table 1. AUROC(Area under the Receiver Operating Characteristic) and AUPRC (Area under the Precision Recall Curve) were used to evaluate the performance of these models. Refer to for more details on AUROC and AUPRC.

Logistic Regression	
AUROC	0.5
AUPRC	0.50289
Gradient Boosted Trees	
AUROC	0.99999
AUPRC	1.0
Random Forest	
AUROC	0.99999
AUPRC	0.99989
Decision Tree Classifier	
AUROC	1.0
AUPRC	1.0

Table 1: Machine Learning Results for 5GB data-set

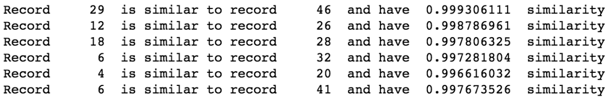
K-Means Clustering	
Clusters	20
Sum of Squared Errors	0.60108

Table 2: Clustering Results for 5GB data-set

In general, the closer these values are to 1, the better the accuracy of the model. From the results it was observed that a naive model like Linear Regression started off with values of 0.5 (AUROC and AUPR) whereas more powerful models like Decision Tree Classifier and ensemble models like Random Forests performed way better with very high accuracy (1).(Refer to Table 1)

For Clustering, the sum of squared errors indicate the accuracy. The closer these values are to 0 the higher the accuracy. (Refer to Table 2)

For similarity search, we are calculating the cosine similarity between different cases.(Refer to Figure 6 for a small sample) This along with clustering can nail down good reference cases in the past and hence help the medical practitioner take informed decisions.



```
Record     29  is similar to record     46  and have  0.999306111  similarity
Record     12  is similar to record     26  and have  0.998786961  similarity
Record     18  is similar to record     28  and have  0.997806325  similarity
Record      6  is similar to record     32  and have  0.997281804  similarity
Record      4  is similar to record     20  and have  0.996616032  similarity
Record      6  is similar to record     41  and have  0.997673526  similarity
```

Figure 6: Similarity Search sample output

6 Discussion

For a pregnant mother, it is possible to predict the risk of infant death with this framework and US

CDC data. If a risk is detected, the medical practitioners can leverage this framework to find the similar pregnancy cases in the past which are a close match to this one and perhaps refer the pregnant mother to a Perinatologist with proven track record(specialist in mother and fetal health) instead of a regular gynecologist. This would definitely reduce the chances of infant mortality, thereby fulfilling SDG 3.2 to quite an extent.

7 Conclusion

This framework serves as a tool to answer the following question:

1. Can we predict the risk of infant death? Yes, with the help of large-scale machine learning models, the risk or no-risk classification has been successfully accomplished.

2. If at risk, can we find similar pregnancy reference cases in the past to aid the medical practitioners in taking well-informed decisions? Using clustering for big-data and similarity search for sampled data, this has been successfully accomplished.

   This framework empowers the medical practitioners to assess the risk of infant death ahead of time and look-up past matching pregnancies as treatment reference. Perhaps this can lead to the pregnant mother's case being transferred to a Perinatologist instead of a regular Gynecologist

8 Team Member Contributions

Data Load: Anurag, Shayan
Data Cleaning: Keshav, Shayan
Feature Extraction: Renu, Keshav
Dimensionality Reduction: Keshav, Anurag
Clustering: Renu, Anurag
Machine Learning: Shayan, Renu
Similarity Search: Anurag, Keshav
ML Evaluation: Shayan, Renu

References

[1] CDC website

[2] CDC Data

[3] UN Sustainable Development website

[4] K-Means Clustering Wikipedia

[5] National Vital Statistics Reports (NVSS) Report on Infant Mortality

[6] Similarity Search

[7] Reducing Infant Mortality in Indiana

[8] Infant Mortality Prediction in Allegheny County