Understanding secondary factors contributing to child mortality

Team XYZ - Karan Shah, Komal Agarwal, Narayan Acharya, Vikas Dhyani

Section 1: Introduction

Tremendous progress has been made in the past two decades in containing child mortality across the world. Yet, one child or young adolescent died every 5 seconds in 2018. The vast majority of these deaths (85%) occur in the first 5 years of life, of which almost half (47%) happen during the first month. According to UNICEF report [1], 'Under 5' mortality rates are largely due to treatable infectious diseases. The report mentions that *"...52 million children under 5 years of age, will die between 2019 and 2030*". Even though we are seeing progress towards this problem, it might not be enough to reduce the disturbing projections for the upcoming decade and we need to explore more measures.

As part of this project, we wish to understand and study the impact of *secondary factors* on child mortality. We explore environment, lifestyle, and health of a mother as well as her access to amenities like medication and hospitalization, access to clean water, etc. The results of this analysis, if significant, can then be used by local government bodies, NGOs, and people responsible for decision making to take informed actions in order to accelerate the decrease in child mortality rates.

Our project comes under the umbrella of Sustainable Development Goal (SDG) 3 - *"Ensure healthy lives and promote well-being for all at all ages."* We believe that containing child mortality can help the mental and physical well-being of mothers and help children live long and healthy lives.

We first study what research has already been done on this topic (*Section 2*) and then use Big Data tools and methodologies to address the large amounts of data (*Section 3*) that need to be processed and analyzed. To address these challenges, we frameworks like Dask for EDA and PySpark with Hadoop Data File System (HDFS) for processing on Google Cloud Platform's (GCP) DataProc cluster using univariate & multivariate linear regression and similarity search (*Section 4*) and then evaluate our results (*Section 5*).

Section 2: Background & Definitions

Our project is slightly inspired by a research study [2] where they investigate disparities in mortality rate between different regions of India using clustering and p-value significance on geo-spatial data. This research was able to ascertain that female illiteracy was indeed a factor contributing to high infant mortality rate (IMR) in India, whereas, lifestyle didn't influence IMR much. Our ideas are also inspired by a research study [3] which Identifies the contribution of non-biological factors towards IMR. This study demonstrates a retrospective analysis of routinely collected data using verbal and social autopsy tools and proves that infectious diseases and lack of better healthcare facilities were responsible for high IMR.

This project moves around IMR which is the number of deaths of infants (age under one year) in 1,000 live births in a year. Here, live birth means a baby's birth, showing any sign of life after exiting the maternal body.

Section 3: Data

The data used in the project is gathered from the Annual Health Survey conducted in 2011 by the Government of India. It comprises information on women living in each district of 9 high density states. The raw dataset is close to 16.81 GB in size and has more than 3M records and more than 200 features per record. Each record has a rich set of features for women – age, medications, access to clean water, educational qualifications, alcohol consumption, smoking or chewing habits, and so on.

Table 1:An example representation of data

State	Woman ID	Smoke	Household	Toilet	Filtered-water
Uttarakhand	ID1	Occasional	Pucca	Open defecation	Boiled
Bihar	ID2	Usual	Kuccha	Community toilet	electronic filters

Section 4: Methods

4.1 Big Data Tools Used - Dask, Spark & Google Cloud Platform

Dask is an open-source project that helps to take advantage of parallel computation and it also gives abstractions over Pandas Data Frames. Dask data frames were used for feature processing, encoding categorical values into numerical and standardizing final data. Apache Spark is an open-source large-scale data processing tool. Python version of Spark - PySpark was used with processed data in following methods i.e. Linear regression and Similarity Search.

We use GCP's Dataproc product with the following specifications for all our processing and analysis: 1 Master Node (e2-standard-2, 64GB Primary Disk) and 2 Worker Nodes (e2-highmem-4, 32 GB Primary Disk). All instances running the 1.4.27-debian9 image.

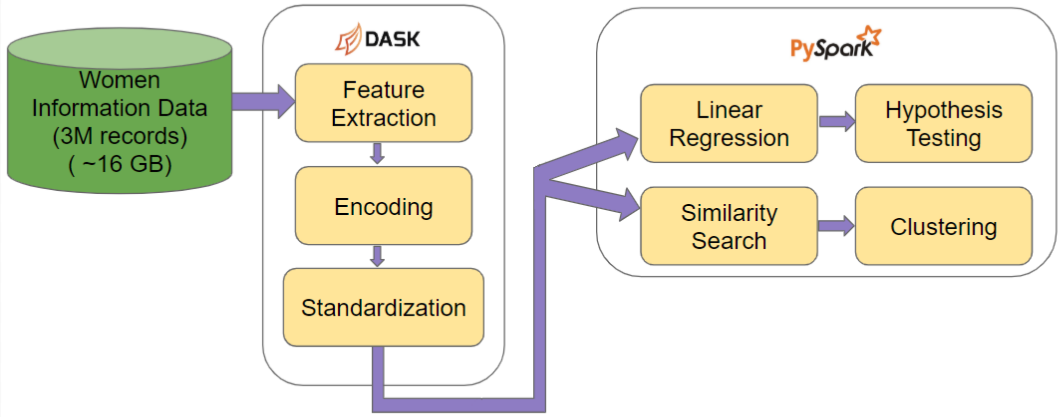

Figure 1: Overview of Implementation

4.2 Exploratory Data Analysis

To get an idea of data, we took a subset of it i.e. CSV of 1 state and analyzed it using Dask data frames on Google Colab before creating a data processing pipeline for the full data. We reduced features from ~200 to ~60 by making assumptions about the impact on mortality. We computed missing data distributions (Figure 2) in the result data and again filtered a few features having a higher ratio of missing data and imputed values based on rudimentary Data Science techniques. The result had ~30 features which we finalized for further procedures.

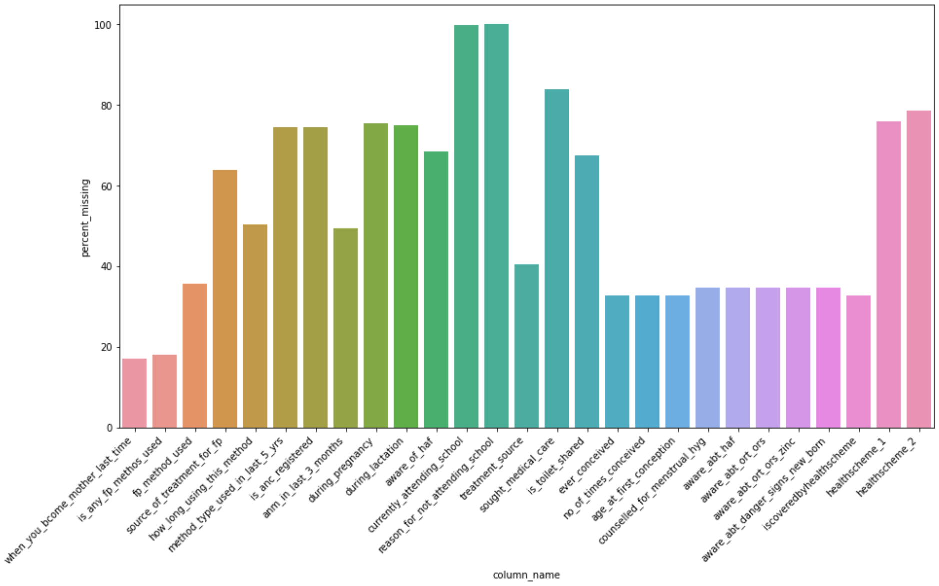

Figure 2:Missing Data Distribution

4.3 Pre-processing

Filtered data of all states is then passed through the data processing pipeline to make it more usable for downstream tasks. First, categorical features are transformed into numerical. A new feature "*children lost*" is created from difference between the *total births* and *total surviving* children available in the dataset and use it as a label. Two attempts were made in encoding categorical features as explained in the table. The Boolean approach resulted in data having only 1 and 0, however, the rating approach data was ranged in [0, n] where n = number of unique values in a feature, which was then scaled and normalized. A sample example with reasoning can be seen in the table below.

Feature	Boolean	Rating
Higher_Education	0: literate, 1: otherwise	0: illiterate, 1: Literate w/o formal education,…, 8: Literate-Post Grad
Smoke	0: never smoked, 1: otherwise	0: never smoked, 1: ex-smoker, …, 4: regular smoker

4.4 Univariate & Multivariate Linear Regression

In order to understand which features contribute most to children lost we use Linear Regression (LR) along with Hypothesis testing to adjudge the significance of our results, both Univariate and Multivariate LR were performed to pick top features that positively correlate with the number of child deaths, for each state. Due to large number of records, we used LR with Stochastic Gradient Descent (SGD) with learning rate 0.1, which allowed us to learn the weights of our model in an incremental fashion by considering batches of records at a time.

As part of univariate LR, we used both Dask and Spark. Since each univariate LR is independent, we parallelize the running of multiple LRs. We use Spark for the actual computation of LR (Figure 3). We achieved significant computation time reduction with this approach (Figure 4). For multivariate LR, we simply use Spark for our computation needs since we only have a single task & nothing to parallelize for. The times reported are for calculating both betas for LR and p-values for Hypothesis Testing.

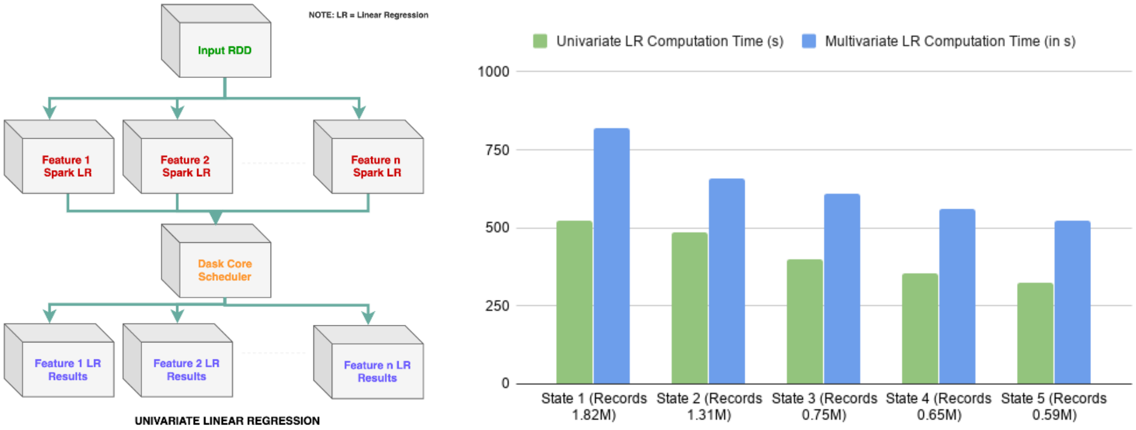

Figure 3,4: Univariate LR, Computation Time Comparison

4.5 Hypothesis Testing

Once we have learned the model weights, beta coefficients of our LR, we perform One-Tailed Hypothesis Testing in order to judge the significance of the beta values of the top 5 features positively correlating with child deaths. We use the t-statistic for our hypothesis testing. We perform hypothesis testing on beta coefficients learned from both univariate and multivariate linear regression along with Bonferroni correction.

4.6 Similarity Search & Clustering

We used similarity search in a different way in that we used it to group districts based on the similarity of the features and used the aggregated features to correlate with IMR. Spark and Pandas enabled us to perform all the necessary transformations required to complete this analysis in the following steps:

1. Calculate the IMR (Total infant deaths/ (Total population/1000))
2. Select the district with minimum mortality rate as the benchmark district. (e.g. in the state of Bihar, district number 16 - Siwan had the lowest IMR)
3. Now, using our representative features in the categories of education, lifestyle, amenities, etc, we calculated cosine similarity of all districts with the benchmark district.
4. This essentially clustered the similar and dissimilar districts; which was further verified using K-prototypes clustering for numerical and categorical data. We then checked features in the benchmark district and other dissimilar districts and compared our observations for each feature used.

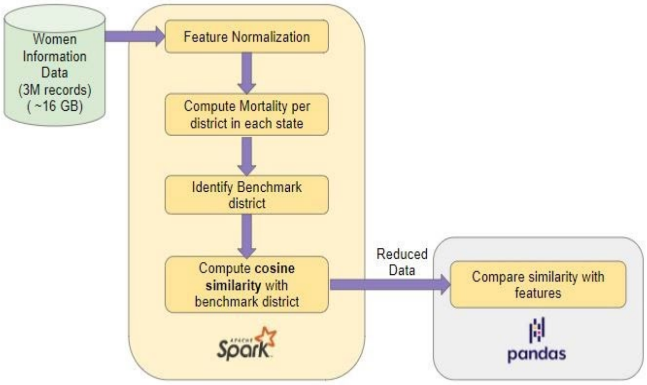
Women Information Data (3M records) (~16 GB)
Feature Normalization
Compute Mortality per district in each state
Identify Benchmark district
Compute cosine similarity with benchmark district
Spark
Reduced Data
Compare similarity with features
pandas

Figure 5: Similarity Analysis Implementation

Section 5: Evaluation/Results

5.1 Linear Regression & Hypothesis Testing Results

From the results of the LR, we derived the top 5 features that most correlated with child deaths and found that Alcohol and Smoking were part of the top 5 features for 4 out of 5 states while Awareness about Acute Respiratory Infection (ARI, Pneumonia) and Chewing Tobacco also appeared but they were not as frequent. The two results, Modern and Traditional Methods of Contraception were frequent too, but we could not ascertain why that could be the case. More domain knowledge is required to comment on the results for these features. Table 2 shows results of the univariate LR for Uttar Pradesh state.

State 1 (Records 1.82M)	beta	p-value
modern_methods_used	0.0438	0.0
alcohol	0.0434	0.0
smoke	0.0426	3.52E-06
traditional_methods_used	0.0398	5.52E-06
aware_of_the_danger_signs	0.0387	8.82E-06

Table 2: LR Results for State 1

5.2 Similarity Search Results

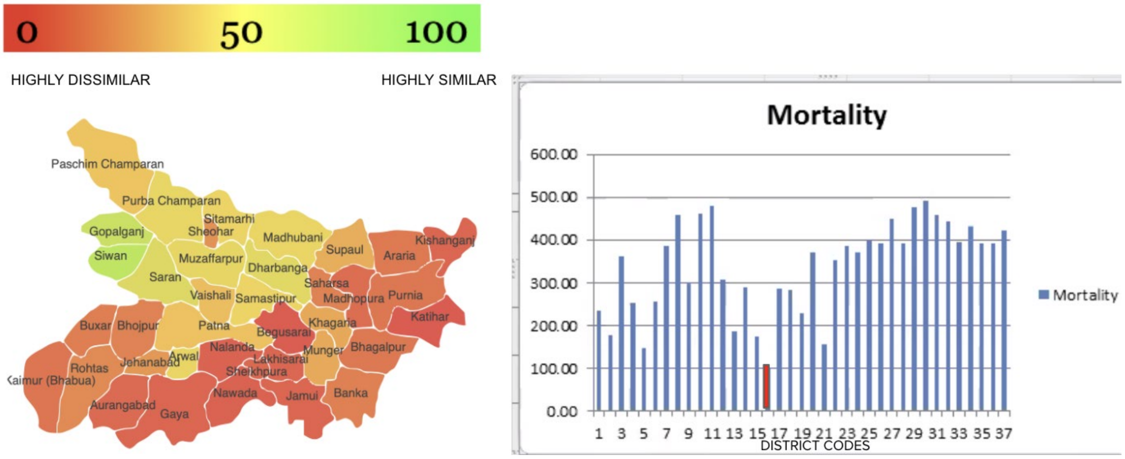

Figure 6,7: Similarity Amongst Districts, Mortality Rate of Districts in Bihar

In the Figure 6 we can see the result of step 4 described above in methods i.e. district coloured as green is our benchmark district, Siwan (District no. 16), and the remaining districts are coloured based on their similarities, with red colour indicating the least similar district. It is clearly evident from fig1 that similar districts have clustered together and the same was visible in K-Prototypes Clustering too.

Further validation of our assumption that non-biological factors indeed contribute to high IMR was obtained when we checked the mortalities of each district. For instance, consider District no. 10 Katihar; that is least similar when compared to Siwan and it also has a very high IMR as evident from Figure 7.

We then checked the correlation of our features with mortality, and our hypothesis was further validated when we saw high correlation or mortality with government schemes, education, lifestyle, and amenities (Figure 8).

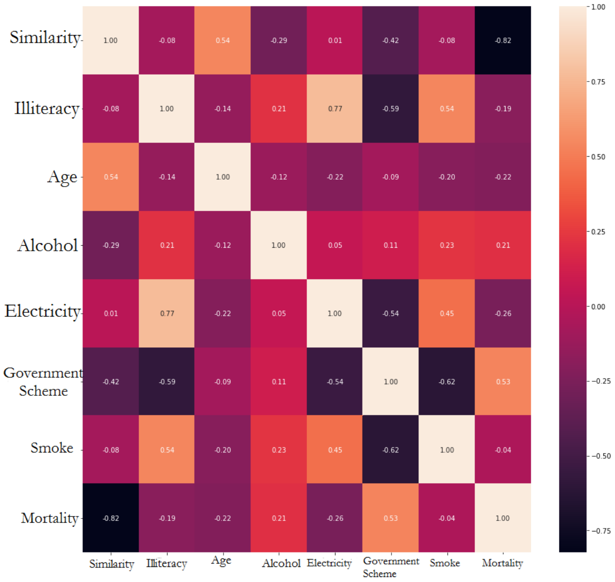

Figure 8: Correlation Matrix of Mortality vs Features

Our analysis using similarity search further cemented the fact the education has a significant role to play in curtailing the infant mortality rate; this was even highlighted in the study presented in our background research, even though their approach was entirely different - they used geospatial data to form clusters. The result obtained by them and ours is strikingly similar; pointing to the possible correctness of our assumptions.

Section 6: Conclusion

Considering SDG 3, our results conclude that non-biological factors also impact IMR to some extent. The analysis reveals actions and awareness measures towards non-biological factors such as education, alcohol consumption, smoking habits, etc. that need to be considered in order to reduce IMR. Our analysis using linear regression on a large dataset across districts in India, helped us identify features which are not intuitive directly but have some impact on infant mortality rate. Similarity search ascertained this by grouping districts having similar features together. Thus, our results restate our hypothesis that secondary factors like education, awareness about government schemes, lifestyle, and basic amenities do contribute to a high IMR. This result can help the government, NGOs, and planning commission to take necessary actions in the right direction that will help bring down the IMR.

References

1. UNICEF Report - https://data.unicef.org/resources/levels-and-trends-in-child-mortality/
2. Research 1 - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026856
3. Research 2- https://bmjopen.bmj.com/content/7/8/e012856.full
4. KPrototype algorithm - https://pypi.org/project/kmodes/
5. Data source - https://data.gov.in/catalog/annual-health-survey-woman-schedule
6. SDG - https://sustainabledevelopment.un.org/?menu=1300
7. WHO Infant Mortality - https://www.who.int/data/gho/data/indicators/indicator-details/GHO/number-of-infant-deaths-(thousands)