Determinants of the Neonatal Mortality Risk in Brazil, 2006-2016: A Machine Learning Models Approach

Luciana Correia Alves (Department of Demography and Population Studies Center ‘Elza Berquó’ (NEPO), University of Campinas, Brazil)

Carlos Eduardo Beluzo (Federal Institute of São Paulo, Brazil)
Natália Martins Arruda (Federal Institute of São Paulo, Brazil)
Rodrigo Campos Bresan (Federal Institute of São Paulo, Brazil)
Tiago Carvalho (Federal Institute of São Paulo, Brazil)

Outline

1. Background

- Infant Mortality Rate
- Neonatal Mortality Rate

2. Methods

- Datasets
- Variables
- Machine Learning Algorithm

3. Results and Discussion

- ROC curve
- Confusion matrix
- Feature Importance
- Key findings

1 Background

Background

- Infant Mortality Rate dropped to 16.2 deaths per 1,000 live births between 1991 and 2010

- Life expectancy at birth increased from about 50 to about 73 years over the same period

- Reduction in infant mortality for infectious and parasitic diseases (better life and sanitary conditions, hygiene, nutrition, access and care health)

- In 2018, Infant Mortality Rate (IMR) was 12.8 deaths per 1,000 live births in Brazil

Background

- Brazil achieved this Millennium Development Goal (persistent inequalities remaining between geographic regions and population groups)
- Regions and populations with lower incomes are at greater risk of infant death
- Neonatal death (between 0 and 27 days) did not reduce satisfactorily (in 2017, Neonatal Mortality Rate (NMR) = 9 deaths per 1,000 live births)
- NM are complexly articulated and influenced by the maternal and newborn biological characteristics, social conditions and the care provided by the health services

Background

- Neonatal mortality is a complex phenomenon, involving interactions of several characteristics and requiring a large volume of data for its full understanding.
- Machine Learning (ML) methods are a powerful tool that helps to understand better the interactions between various factors.
- The application of Machine Learning is innovative to the Brazilian reality and for mortality studies in demography.
- We believe that traditional regression models may not be enough to understand this problem.

Aim of this Study

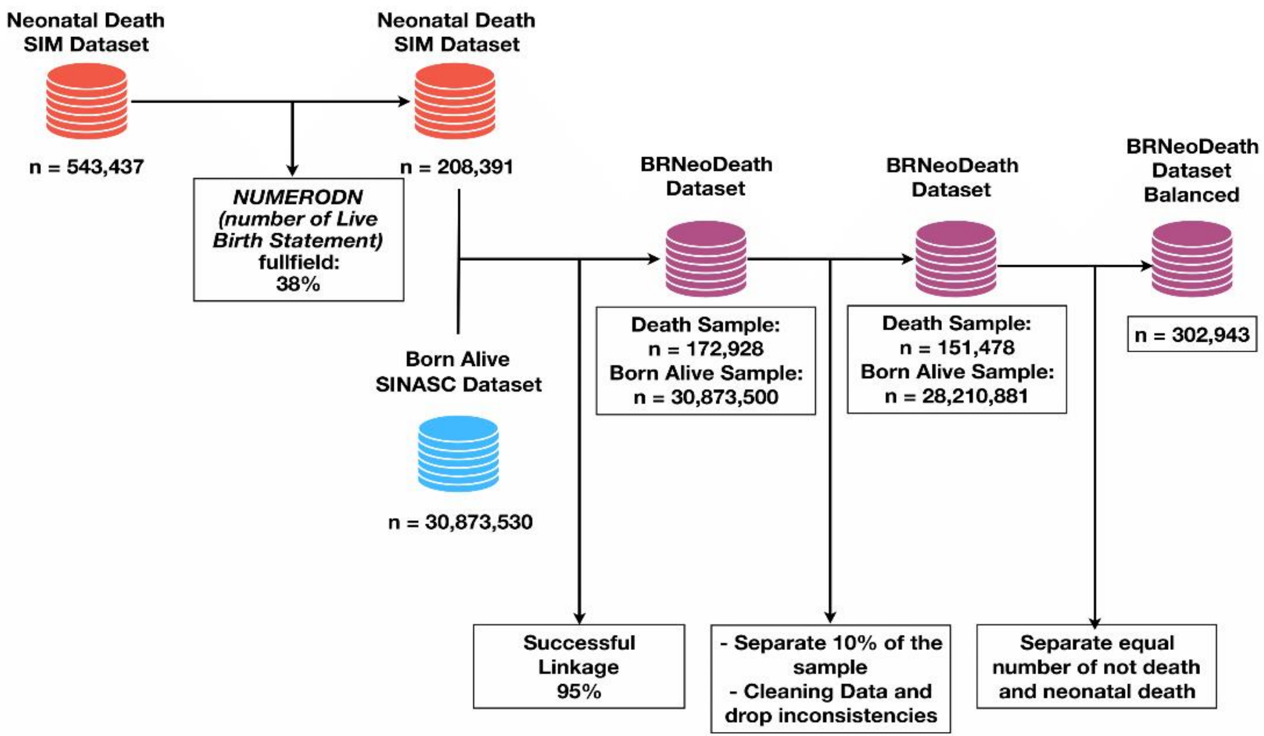
To predict risk of neonatal death and to assess the feature importance based on machine learning approaches between 2006 and 2016 in Brazil

Methods 2

Data

- Data came from by Mortality Information System (SIM) and Live Birth Information System (SINASC) from DATASUS
- Linkage technique was used to relate the datasets through the application of the deterministic method
- We use the standard variable for both systems, Number of Live Birth Statement (NUMERODN)
- It was possible to join the two datasets in 95% of the cases resulting in a large-scale dataset that we call BRNeoDeath with initially 30,873,500 observations (Figure 1)

Figure 1: Flowchart of the process to linkage the data from SIM and SINASC for balanced dataset



Data

- After of the linkage we applied a data cleaning to remove inconsistencies such as duplicate observations
- BRNeoDeath have unbalanced class distribution (percentage of death class samples are outnumbered by the percentage of living class samples), being 99.4% of the living class, and just 0.6% of the dead class
- Sub-sampled dataset (final sample=302,943) consists of all the positive samples (death class), and the same amount of negative samples (alive class)

Variables

Demographic

Socioeconomic maternal

Maternal obstetrics

Related to the newborn

Related to previous care

Robson 10-groups classification*

** Robson Classification:a global standard for assessing, monitoring and comparing caesarean section rates both within healthcare facilities and between them. The system classifies all women into one of 10 categories that are mutually exclusive and, as a set, totally comprehensive. The categories are based on 5 basic obstetric characteristics that are routinely collected in all maternities (parity, number of foetuses, previous caesarean section, onset of labour, gestational age, and fetal presentation).*

Analysis

- Descriptive statistics were calculated for all variables
- Machine Learning models were used
- The algorithms were able to classify the determinants, highlighting the most powerful determinants
- Machine learning models were estimated across one round of experiments using different metrics to measure its effectiveness in neonatal death risk classification task
- Three algorithms were tested and analyzed given their good results achieved on health problems:

Analysis

- **Random Forest (RF):** Construction of decision trees, evaluate the importance of each predictor so that it is possible to identify the relevant variables in the construction of each tree.
- **Extreme Gradient Boosting (XGBoost):** Decision trees that differ from RF since they start with weak learning trees and build more reliable trees based on the residuals from the predictions of each previous tree.
- **Support Vector Machine (SVM):** It creates a border that best separates the data through support vectors.

Analysis

- It was used supervised learning in balanced dataset
- The variables transformation adopted was that of one-hot encoding (that is the process by which all categorical variables were transformed into dummy variables)
- Data were separated by 80% for training and 20% for testing
- Metrics were calculated: accuracy, sensitivity, specificity, Receiver Operating Characteristic Curve (ROC Curve) and Area Under a Curve (AUC)
- The last step: construct and evaluation the performance of three machine learning algorithms (the methods were implemented using Python programming language (3.6))

Results

3.1

Results

- In the sample, the IMR decreased over the years
- NMR has represented a higher proportion of IM cases.
- Between 2006 and 2016, IMR dropped to approximately 13 deaths per 1,000 live births
- NMR decreased from 11 deaths to about 9 per 1,000 live births

Participation of the Neonatal Mortality in the Infant Mortality, Brazil, 2006 – 2016

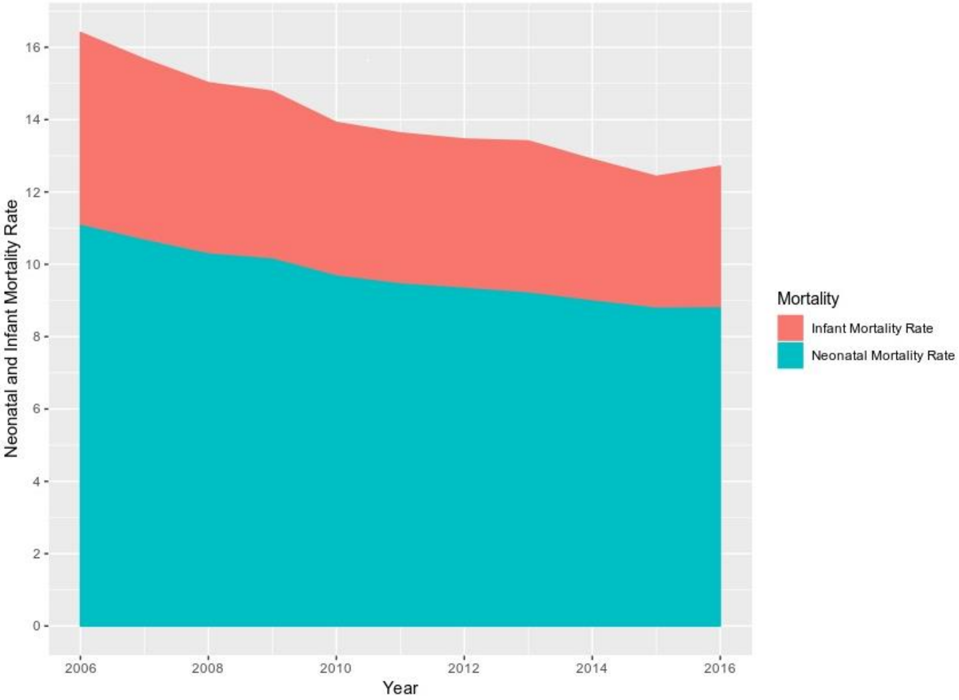
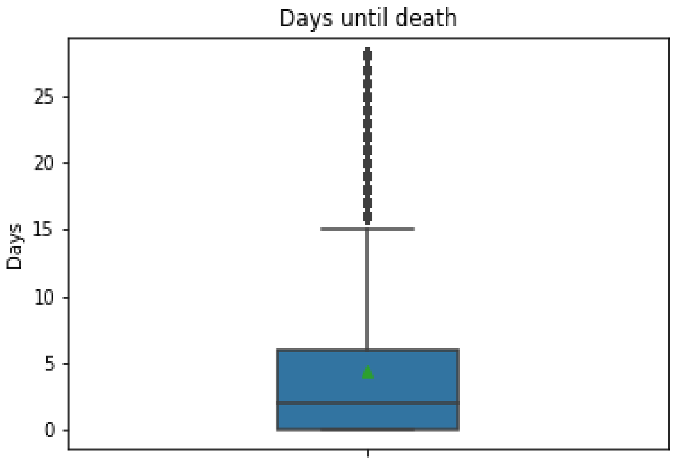
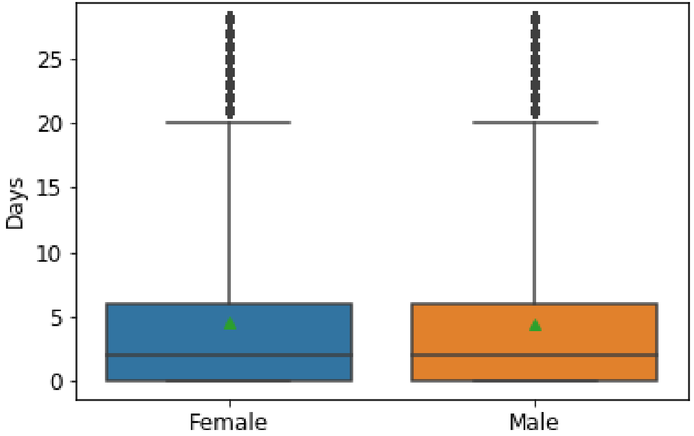


Figure 2 (a) – Days until Death of the Neonatal Mortality, Brazil, 2006 - 2016



Source: SIM, SINASC, 2006-2016.

Figure 2 (b) – Days until Death of the Neonatal Mortality by sex, Brazil, 2006 - 2016

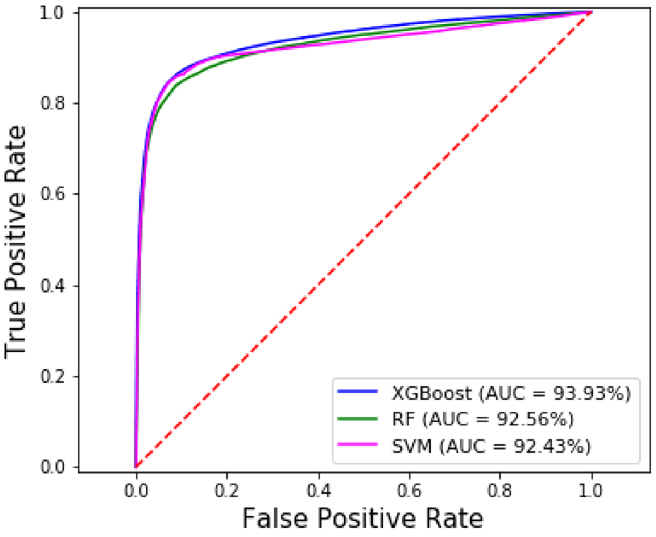


Source: SIM, SINASC, 2006-2016.

Results

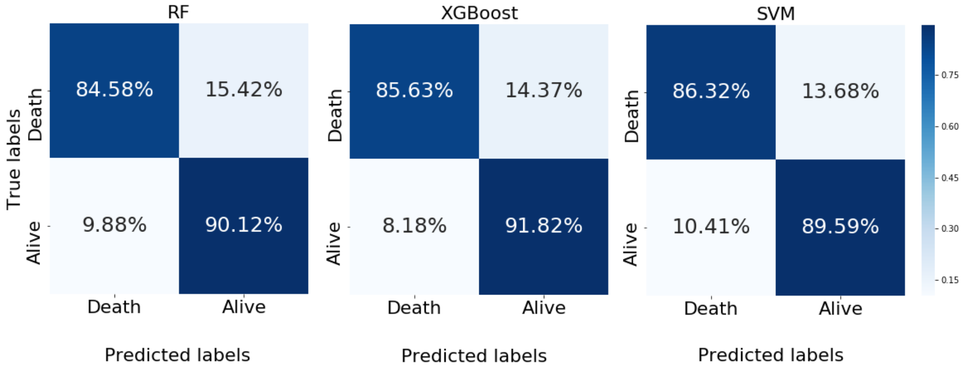
- All evaluated methods present a similar performance on average
- ROC curves almost overlapped
- AUC of 93.93%, 92.56% and 92.43% for XGBoost, Random Forests (RF) and Support Vector Machine (SVM), respectively.
- Accuracy reported on optimal ROC curves for Random Forest, Support Vector Machine and XGBoost are 87%, 88% and 89%, respectively.

Figure 3 - ROC Curve for all the evaluated models



Source: SIM, SINASC, 2006-2016.

Figure 4 - Confusion matrix at the optimal ROC Curve point for evaluated classifiers

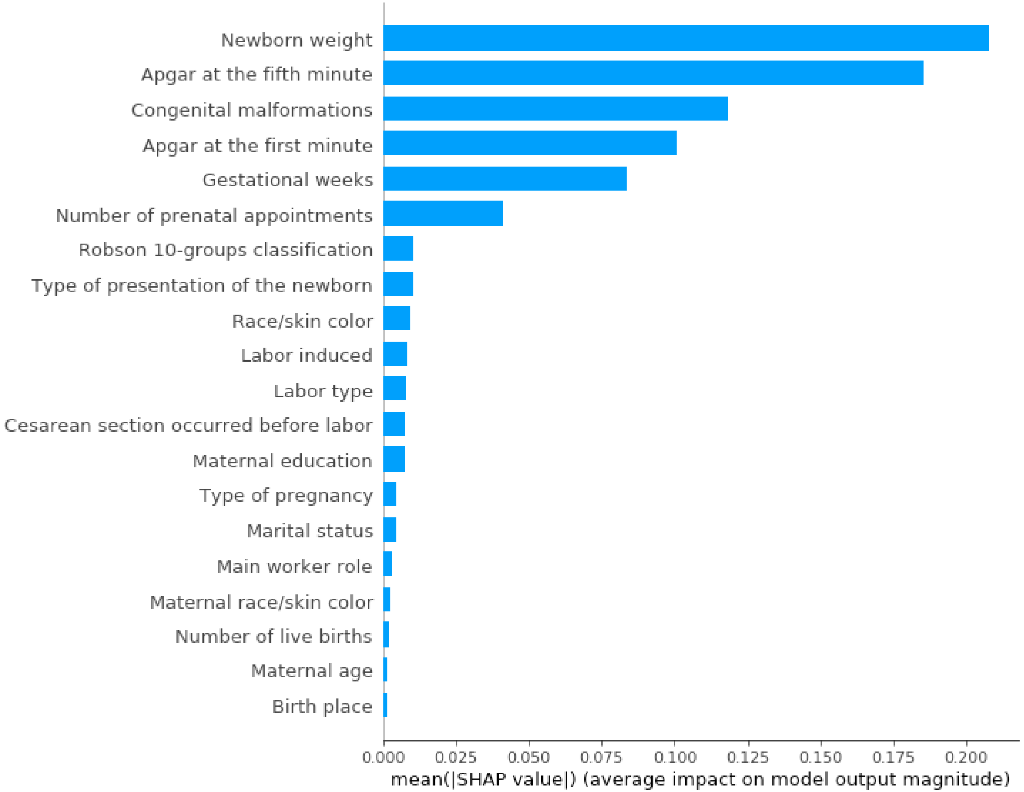


Source: SIM, SINASC, 2006-2016.

Results

- The feature importance is a measurement that can point out the features with major relevance
- XGBoost was the model with the best performance
- Variables presenting a higher degree of importance (highly correlated with label class): newborn weight, Apgar at the fifth minute, congenital malformations, Apgar at first minute, gestational weeks and number of prenatal appointments.
- Maternal characteristics was not important

Figure 5 - Feature Importance for XGBoost model



Source: SIM, SINASC, 2006-2016.

Discussion and Conclusion

3.2

Discussion and Conclusion

- Despite the fact that women show a lower percentage of death in the neonatal mortality rate, we found that the number of survived days was similar for men and women babies
- The prenatal and biological attribute were key determinants
- Low birth weight infants are more vulnerable to pulmonary immaturity problems and metabolic disorders
- Low weight and poor ratings in Apgar 1 and 5 minutes are warnings for possible future complications in the child, creating an alert on the risk of this newborn dying in the first days of life

Discussion and Conclusion

- Some congenital genetic, infectious, or environmental-related anomalies can be prevented through the implementation of public policies and a fair offer of health services
- The present study is in line with other research that focused on an insufficient number of prenatal visits and increased neonatal deaths (70.7% of the mothers that lost their babies made less than seven prenatal appointments)
- The study found no high importance between neonatal mortality and maternal profile

Discussion and Conclusion

- We created new datasets, comprising more than 30 million samples for the problem of neonatal mortality
- We proposed a new method to expose neonatal death risk-based in a combination of machine learning classifiers and demographic features
- This finding reinforces the importance of adequate surveillance of deliveries and qualified care addressed to the newborn as a way to reduce infant morbimortality
- As a decision support tool, this kind of method can be useful to help health experts to take decisions if more intensive care is necessary for newborns in Brazil

Funding

We gratefully acknowledge financial support from the Bill & Melinda Gates Foundation (Process no: OPP1201970) and Ministry of Health of Brazil, through the National Council for Scientific and Technological Development (CNPq) (Process no: 443774/2018-8). It was also supported by NVIDIA, that donated a GPU XP Titan used by the research team.

THANK YOU

lcalves@unicamp.br