Fetal and Infant Mortality in Colombia

DS4A. Team 4's final project
Angélica Rincón
Melissa Aguilar
Santiago Morales Saldarriaga
Jenny Lancheros Pineda
Pedro Ospina
Andrés C. Marulanda

1. Introduction and context:

As of 2011, The World Health Organization (WHO for its acronym) estimates that about 4 million neonates die every year, and nearly 41% of all under-five child deaths are among newborn infants and specially babies in their first 28 days of life (neonates) [1]. Accordingly, perinatal (period between 20 weeks before, and 4 weeks after birth) and neonatal mortality is now a part of the 2030 agenda for Sustainable Development of the United Nations and is also one of the topics of interest to the Colombian Ministry of Health [2].

This work is aimed to be a comprehensive analysis of different databases related to perinatal and neonatal death, as well as demographic, economic, social and geographic data for Colombia, that can help us understand which features may influence and contribute to said death rate in different locations in the country by looking for possible correlations that may constitute the first step to the search for solutions to this global health issue.

Characterization of the social, economic, demographic and geological variables that can be correlated and may have an impact on the newborn and fetal death rates is of capital importance for the localization of the most affected regions, as well as for the identification of causal relationships between said factors and, upon inclusion of larger databases, identification and prediction of new significant factors, as well as the prediction of future outcomes. Such an analysis may eventually lead to the implementation of more specialized and well-designed social programs, as well as health campaigns, among other humanitarian initiatives.

As an additional motivation for this work, a comparison against other countries and regions is presented in figure 1, where it can be seen that in terms of infant mortality, Colombia does considerably better than some regions such as Sub-Saharan Africa and the world overall, having a decreasing tendency over the years.

A comparison in terms of fetal death rate is shown in figure 2. Here, it can be seen that Colombia shows outstanding behaviour (in a negative way). This may suggest two non-exclusive possibilities: whether the data is being taken differently in other regions, or Colombia is doing outstandingly worse on this front, having fetal death rates more than 7 times those of the more developed regions.

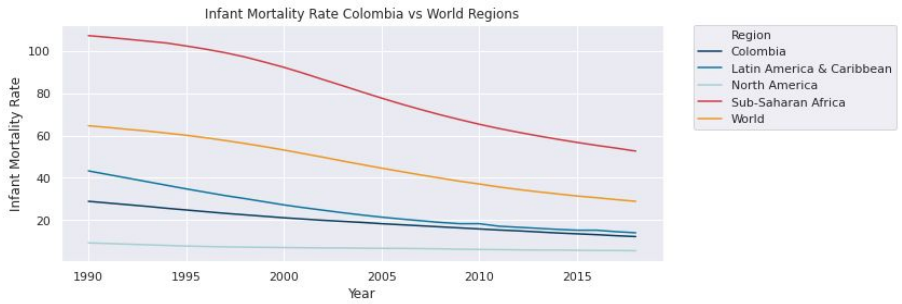

Figure 1. Comparison of Colombia against world regions in terms of infant mortality through the years.

Any of these two possibilities give enough space for a deeper analysis into what is driving such a behaviour; such an analysis is carried out on this work.

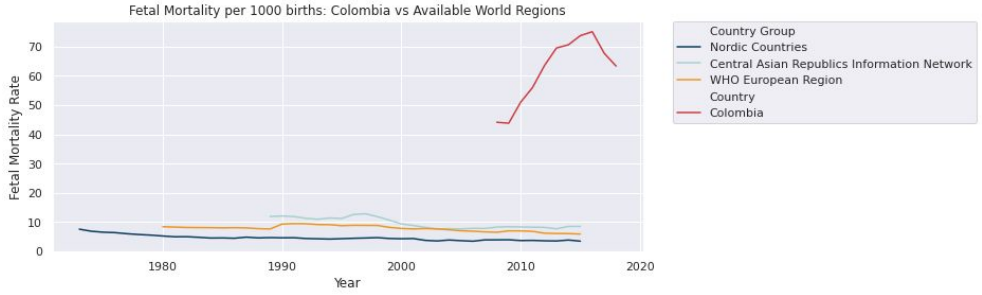

Figure 2. Comparison of Colombia against world regions in terms of fetal mortality through the years.

2. Models and methods:

One of the main objectives of this project is the data visualization, which allows for the identification, comparison, summarization and explanation of the perinatal and neonatal death rates in Colombia. All of this with the objective to find trends and behaviors that can support the programs of prevention and reduction of said rate in this country.

Among the graphics that may potentially serve this purpose are interactive maps, which allow an easy exploration of geographical distributions of the data at a departamental as well as at a municipal level. Interactivity gives an easy way of depleting multiple informative charts while occupying little space in the dashboard. For visualization of differences among the distributions of variables for different groups, distribution charts such as bar plots and histograms are constructed as well. Finally, time-series plots help visualize trends in data, a useful tool when it comes to predicting future outcomes.

For the purposes of finding trends in the data, identifying possible causes for the mentioned death rates, and predicting results for upcoming years, some computational techniques such as correlation analysis, logistic regression and decision trees were applied in conjunction to the aforementioned graphics.

2.1 Exploratory analysis:

After the corresponding data cleaning and preparation, an exploration of the available data was performed so as to extract some useful information, before any model is discussed. For the sake of conciseness, only the data for the year 2018 are shown in figures 3 to 7. Some easily understood variables were plotted as discriminated distributions (births, fetal deaths and infant deaths), and are presented in figure 3. Here, every category was normalized independently from the others so as to allow an easier visualization of the relative differences in the distributions; consequently, the scale is not the same for all the distributions.

These plots show already some important features of the dataset in a more detailed way; particularly, it can be seen that the weight distribution for newborns is a normal-like, and most of the fetal deaths are concentrated below 1.0 kg, which is an expected conclusion since fetal deaths don't make it to the end of the normal nine-month pregnancy period and therefore do not fully develop. Although we have considerably less data for the non-fetal deaths, we can see that the weight distribution for this group differs from that for the newborns in that its mean is displaced to the left and has an additional peak below 1 kg. Further on, it can be noted that post-birth deaths with its seemingly bimodal distribution look already very much like a different population than normal births, with a more normal-like distribution. Weight can be thus proposed to be a good factor indicating newborn viability.

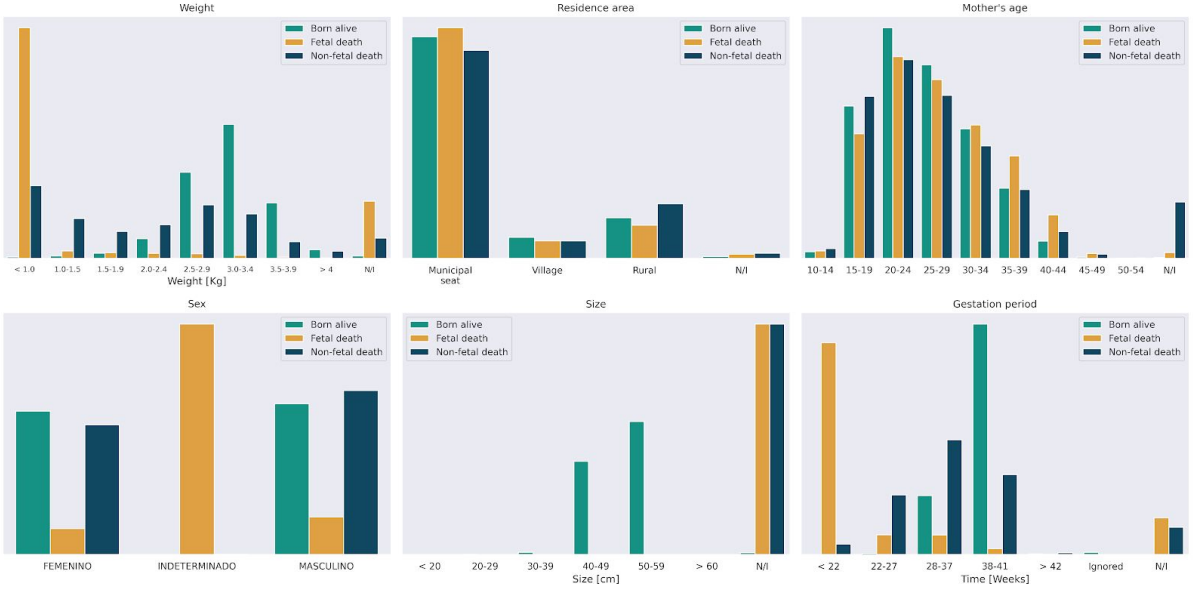

Figure 3. Distribution of some of the variables of interest, discriminated by final pregnancy outcome, for the year 2018.

Although distributions for mother's age are similar across categories, it can be seen that pregnancy outcome is also sensitive to this variable. It is of huge importance to note how the mother's age distribution for fetal deaths has a longer tail (to the right) than the other two categories, suggesting that the probability of a fetal death increases positively with the age of the mother, making this variable an important factor as well. We can also see that the peak is at 20-25 years old, and that there is a worrying number of mothers in the age range 10-14, and even 15-19 years old.

Regarding the sex and pregnancy time plots, it can be seen that most of the fetal deaths occur before 22 weeks, which may affect the sex determination procedure since at this stage the fetus is not yet fully developed. A better picture of this sentence can be seen in figure 4, where the sex-undetermined subjects are mainly concentrated at gestation times below 22 weeks, confirming the above assessment. This variable is thus discarded as a predictive tool for this kind of death due to its lack of information for this population.

Additionally, figure 3 suggests no relationship between sex and the probability that the infant will be born alive or suffer a fetal or non-fetal death. After a more detailed examination, however, significant evidence supporting the non independence of these two variables was found suggesting that male and female populations are actually different. Particularly, a chi-square test between sex and births and infant deaths was run to determine a p-value of 3.00e-36, confirming the significance of an approximate 1.2 male to female ratio for fetal deaths, which cannot be explained by the same ratio for births, which is around 1.0. This is indeed an issue the Unicef periodically monitors [3], which shows this ratio is not coincidental or endemic, but a worldwide systematic behaviour.

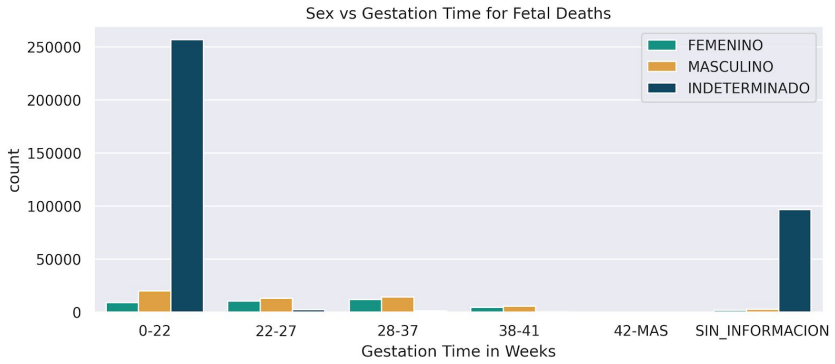

Figure 4. Gestation time distribution as discriminated by sex for fetal deaths.

Following the analysis of figure 3, there does not seem to be substantial differences among distributions in residence area, that is, the fact that some mothers live in a city or in a rural location does not seem to correlate substantially with the outcome of the pregnancy. Same as before though, chi-square hypothesis tests gave a p-value of 4.73e-68 for residence area vs births and infant deaths, and a p-value of 3.50e-264 for residence area vs births and fetal deaths, which shows there is evidence to think this variables are not independent.
From the graph of gestation time, it can be seen that the infant is more likely to suffer a non-fetal death if the infant is born in the 28-37 week period; the effect of this is however apparently weaker as compared to others, such as weight or mother's age, so it needs to be further explored to confirm this claim. At this point this hypothesis can't be confirmed, for this is a model-related issue, but so far it can be said from chi-square hypothesis tests of gestation time vs births and infant deaths, that there is enough evidence to conclude these two variables are not independent, with a p-value of 0. Baby size (figure 3) shows a substantial lack of information for this variable, which allows the disposal of this variable for predictive purposes.

Some more plots can show further insights on this data and on Colombian population in general. In Figure 5 the father and mother's age distribution are shown, and it can clearly be

seen some slight differences. Specifically, the mother's distribution peaks at an early age (20-25) and rapidly drops, while the fathers' peaks at 25-30 and is more spread to the right, showing that male parents tend to be older than females. Although this is a known stereotype of relationships in this country, it would be interesting to dig deeper into this matter and check what is happening to mothers in ages between 10-19 while verifying the possible pregnancy outcome in such a specific scenario.

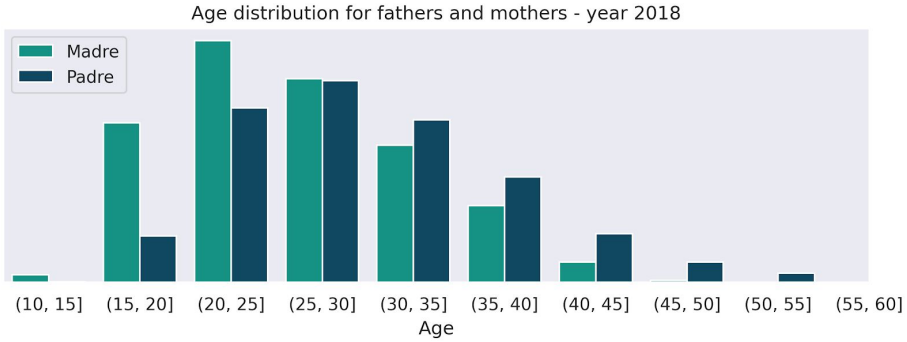

Figure 5. Age distribution of both male and female parents in 2018.

Further dissecting the data, the attention is now turned to the ages of post-birth death subjects. As can clearly be seen from figure 6.a, where the raw counts as categorized by different age spans are shown, most of these deaths occur between 1-6 days, and 1-5 months, which indicates where the attention should be put on, that is, these are presumably the riskier stages of a newborn's early life. However, the time-normalized distribution (6.b), which can be heuristically interpreted as a time-dependent death probability, shows that the first hour of life is critical to determine the survival of newborns. This last point is already an important insight from the data, for it shows how the focus is to be updated: babies between 1-6 days, and 1-5 months, and specially less than 1 hour old newborns and the particular conditions that lead to the so-mentioned behaviour.

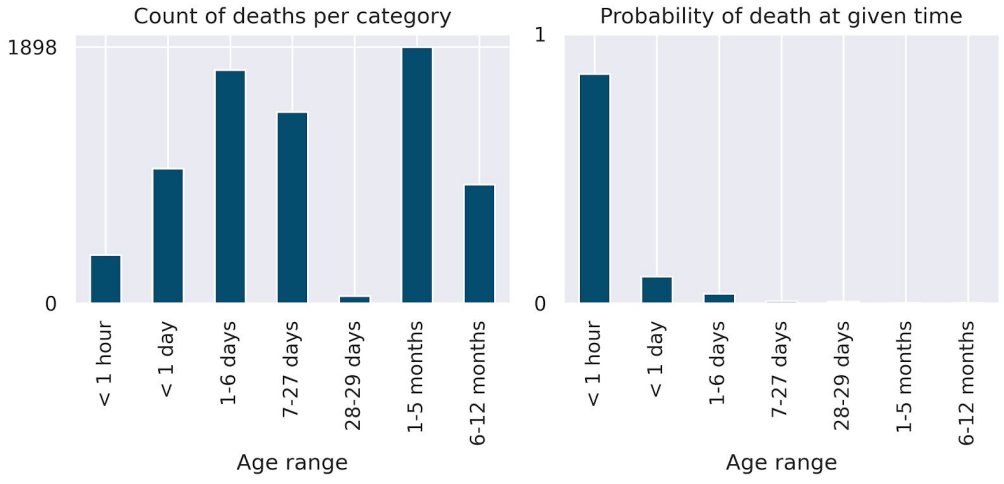

Figure 6. Distribution of death ages for post-birth deaths in 2018. **a.** Absolute numbers (no normalization). **b.** Time normalization, accounting for the different time sizes of the bins. This can be interpreted as a time-dependent death probability.

Pregnancy outcome counts were plotted accounting for the different racial and cultural identification of the person in question (figure 7). Particularly, this plot shows differences

between indigenous, afro and raizal populations, and populations that do not identify themselves with any of these, among others. It can be seen that the probability of post-birth death is higher for the three aforementioned populations, which might be due to cultural and/or behavioural differences among these populations; however, the presence of systematic discrimination against some of these populations as well as to their territories, shall not be discarded as a possibility.

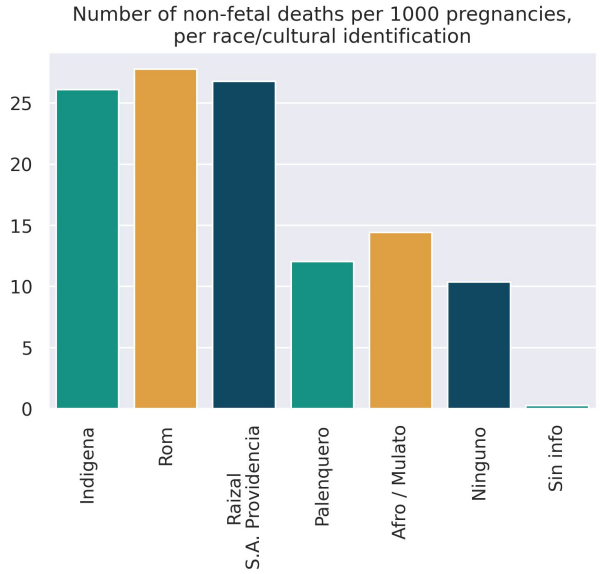

Figure 7. Number of non-fetal deaths per 1000 pregnancies for different cultures/races.

This can further be observed in choropleths and other forms of geographical visualization techniques, which are now introduced. Different behaviours are expected from different departments (geopolitically divided regions in Colombia) given the large differences that occur in this country in terms of territory, resources and most importantly cultural distributions. Such an assessment can be visualized on figure 8, which shows the fetal and non fetal death ratios per department.

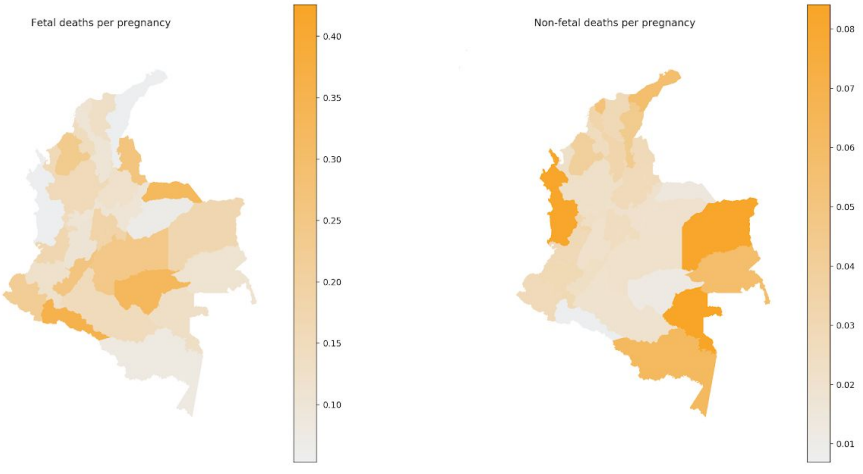

Figure 8. a) Heat map representing fetal deaths per pregnancy (left). **b)** Heat map representing the non fetal deaths per pregnancy (right).

It can clearly be seen from this picture that poor, mainly rural departments such as Chocó (far left) and some in the Orinoquia and Amazonas (far right and bottom), show a dramatic non-fetal death rate as compared to the central departments, the ratio being more than 4 times that of the latter. This is in contrast to the fetal death picture, where most events are

focused on the central regions of the country, which matches with the location of the larger cities and where most of the economical activities are concentrated.

It is proposed that this behaviour is due to the events not being reported on peripheral regions due to the lack of health facilities; the phenomenon could thus be explained accounting for people being obligated to report a non-fetal death since, legally, this corresponds already to a person, while fetal deaths can more easily be unreported. As a complement to this, the ease of access to health facilities can as well account for lower fetal death rates since more thoroughly controlled (in a medical sense) pregnancies are intuitively expected to be more successful.

2.2 Correlation analysis and modelling:

Interactions between variables can be further explored by means of a correlation matrix, whis shows how a given variable changes with a change in another variable, thus making it easier to draw hypotheses about the phenomena behind the data. A correlation matrix was computed (shown in figure 9) for a selected group of 8 variables, and some interesting correlations were found, namely, it was found that the number of medical controls during pregnancy is highly positively correlated with father's age and size of the baby (more than 0.8 and 0.9 correlation respectively).

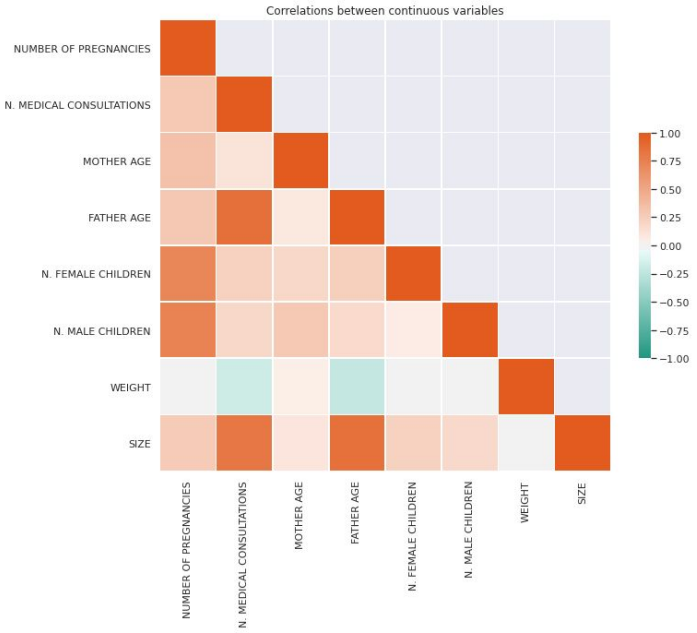

Figure 9. Correlation matrix calculated for continuous variables on the dataset.

Although these are some correlations for which a direct explanation cannot be immediately drawn, it could be proposed that older parents tend to be more prone to attend to the doctor, and on the other hand that the number of controls has a positive effect on babies size. These hypotheses cannot obviously be evaluated from the present dataset but are left for future investigation. One last high correlation is that between father's age and baby size, which naturally comes up from the two correlations above discussed.

Though the number of dead and alive children clearly correlates highly with the total number

of pregnancies, it is surprising that this correlation is not much higher since these two variables are expected to follow a linear equation in that the number of dead and alive children should be equal to the number of pregnancies; this may suggest a slight inconsistency, however we shall not discuss this variable further since it is not explored in-depth on this work.

To some extent, this correlation matrix gives already enough information to start building some simple models. In order to test the predictability of the outcomes, different methodologies were followed among which ordinary linear Regression, logistic regression, decision trees and random forest highlight. The results obtained here are not to be used specifically as a predictive tool, but as a means of exploring the effect of some variables of interest on the outcome of a given pregnancy, and with such information find possible causations and sources for the problems in hand, which is precisely the objective of this study.

Logistic regression:

Many models were tested and the best performance in terms of predictive capabilities for the particular outcome of a pregnancy (born, death before birth) was one with the variables containing the type of social security of the mother, mother's educational level, type of pregnancy (that is, its multiplicity), the area of residence (urban, rural, etc) and the department of residence. These can easily be understood to affect in some way the outcome since these are in many instances determinant (or even a consequence) of the socio-economic conditions of the mother, and may remarkably affect the birth procedure.
As said, the model serves as a means to explore the effect of the variables on the outcome. Having that in mind, the obtained results can now be explored through a set of visualizations of the coefficients, as shown in figure 10.

The coefficients are to be read not in terms of their actual values, but relative to the others. In such terms, a more positive value means an increase in the probability of a fetal death due only to that specific factor, while controlling for the others. Particularly, it can be seen that the most evident trend is that shown by the pregnancy multiplicity coefficients, where a larger multiplicity increases the probability of a fetal death with respect to a multiplicity of 1 (the baseline category). In the same way there are as well differences between coefficients for social security, though a trend can't be drawn for these since there is no unique way to arrange these categories. Aside from very specific results, some interesting behaviour is shown in the map and the educative level plots. Particularly, it can be seen that some departments, where the probability of death is arguably expected to be higher (poor or far from cities) such as La Guajira, Choco and others, turn out to have coefficients far below that of Bogotá, the baseline category. A similar result is shown in the educative level plot, where some populations (Normalista, Profesional) just go off any tendency, arguably showing that the model is heavily influenced by the data imbalance.

The main conclusion that can be drawn from this model is that fetal deaths are systematically not being reported on peripheral and poor regions as much as alive births. In this sense, what the model calls in a way is to conclude that more work should be done in order to

recollect better data on these regions, accounting for the different conditions these people are subjected to.

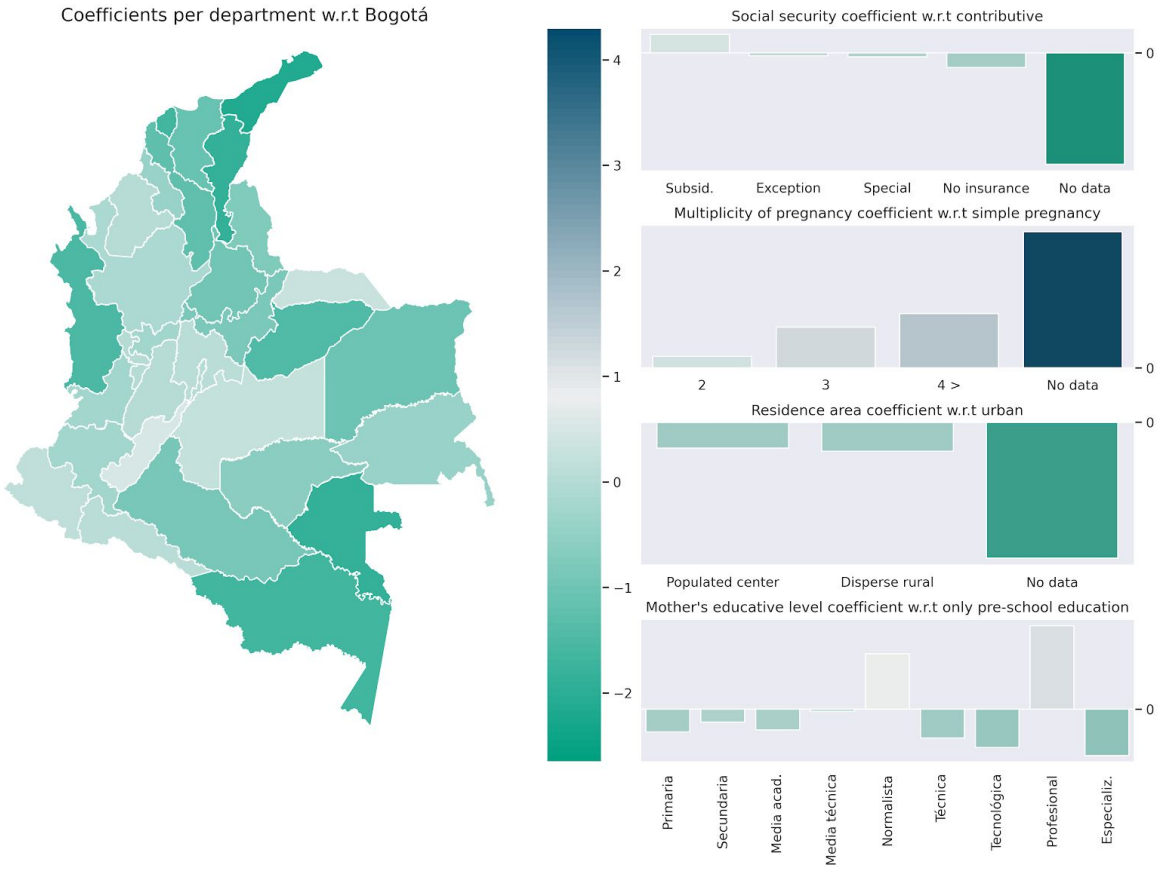

Figure 10. Representation of the coefficients obtained from the trained logistic regression model.

Random Forest:

As a means for obtaining more in-depth information on variable interactions, an additional analysis was performed using the random forest algorithm. Although this particular method does not allow for such a detailed analysis as that done for the logistic regression, it does permit a variable importance analysis, which basically states which variables affect the most the outputs of, in this case, the pregnancies.

With this in mind, a random forest was trained using 100 trees and a maximum depth of 12 levels, while using the same set of variables that were used for the logistic regression (LR). Though the present model performed slightly better than the LR model, the results are -not surprisingly- very similar to those obtained on the latter, as can be seen in figure 11. As was found in the former analysis, the type of pregnancy (that is, the multiplicity) plays a fundamental role for predicting the outcome, having larger coefficients than other variables by factors of 4 in many cases. The number of pregnancies (not shown above) seems to be an important factor as well, which may significantly correlate with the fetus development.

Less importantly -however also slightly prominent- are the residence department and mother's educational level. These results are as well in accordance to the LR model in that there is a difference between pregnancies in different departments, though not so marked as

that given by other variables. The other variables seem to bring small corrections to the model and should not be discussed in further detail.

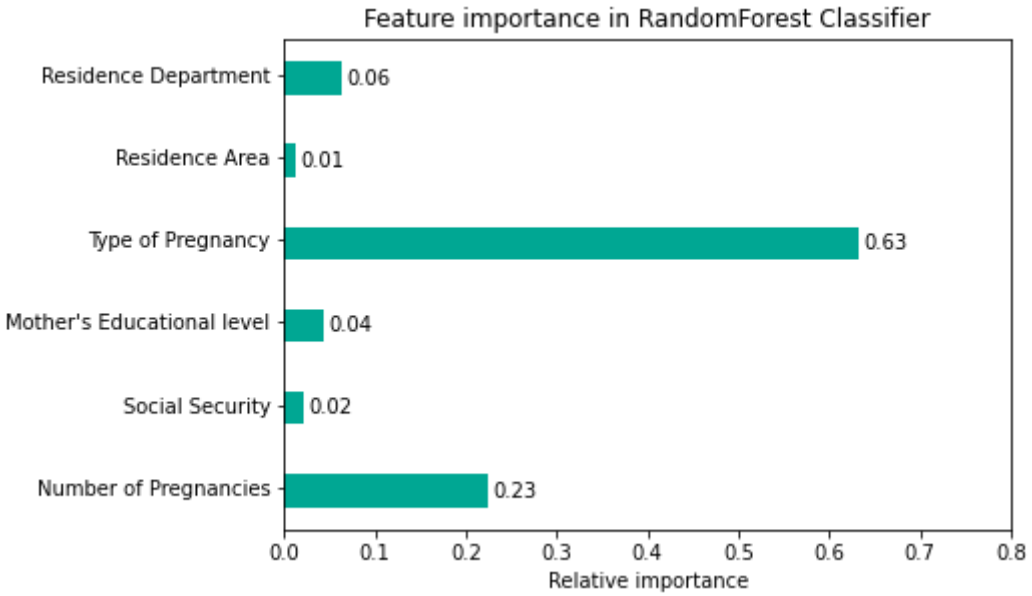

Figure 11. Feature importances obtained using the random forest model.

Looking at the future:

Having completed the analysis of the influential variables, the attention is now turned over the prediction of future outcomes. As a means for obtaining estimates for years whose data is not yet available (such as that of 2019 and 2020), a number of options exist that can help with some level of accuracy. Particularly, the two main options are whether using a new, external covariate for which there is available data for the past and present years so that a model can be trained to predict outcomes based on it, or drawing conclusions from historical trends.

The first one, though more accurate and useful since it makes use of kind of real-time data, was found hard to construct since external covariates that can give useful insights on this topic are usually not available. Variables regarding educational levels in a number of fronts were tested, however gave no useful model as these reduced the number of datapoints to only a few for its lack of granularity.

Given the shortcomings of this approach, the second mentioned approach was used. Namely, the SARIMA (Seasonal Autoregressive Integrated Moving-Average) model was used, which takes into account several of the features of time series such as seasonality and autocorrelation, both of which will shortly be shown to exist in the monthly number of births time-series at hand.

Initially an Augmented Dickey-Fuller test [4] was carried out in order to show the autocorrelation of the time series, that is, the dependence of the present on the previous results. Said test returned a p-value of 0.093, supporting the hypothesis of non-stationarity extracted by looking at figure 12.c, which shows a slight downward trend.

Seasonality was tested by plotting the autocorrelation and partial-autocorrelation functions, shown in figure 12.a and b, respectively.

Figure 12. a) Autocorrelation function (top left). **b)** Partial autocorrelation function (top right) **c)** Monthly number of births in Colombia: Time series

It is clear from the autocorrelation plots that there exists seasonality with a one-year period. After the removal of this frequency and further extraction of other seasonal modes (2 in total), a new time series was obtained which mostly contained white noise. It is after this analysis that the SARIMA model can be constructed, since most of the (6) parameters that must be given to the model are inferred from such an analysis.

An extensive search over the space of possible parameters was performed while testing predictions on the last year of data (2018). The results were shown to be no far from those inferred on the above analysis, which equips the newly trained model with a stronger analytical basis, allowing it to be more easily interpreted and whose predictions are in principle more trusty.

That being said, a model was fitted on the whole time series so as to be able to make predictions for the subsequent years, while accounting for the most recent data. The predicted time-series is shown in figure 13.

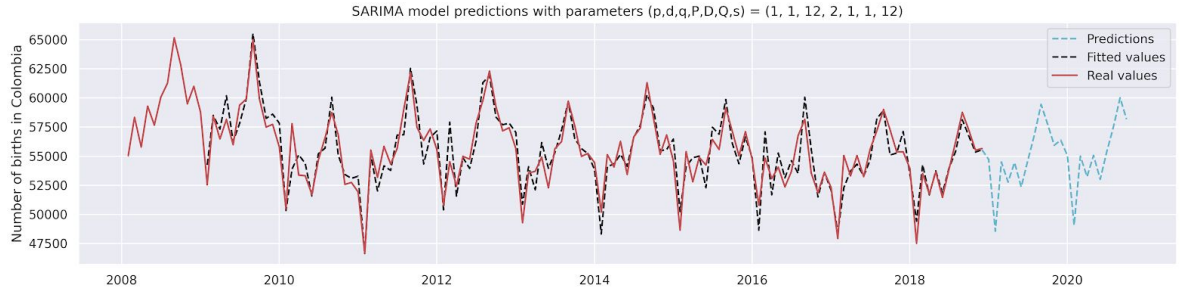

Figure 13. SARIMA model fitted values and predictions for years 2019 and 2020.

It can be seen that the model predicts a slight positive slope for the next two years and the seasonality is reasonably well modelled.

The major drawback of this approach is that it is not possible to include interactions with other variables, that is, the model assumes the phenomenon will occur in vacuum without influence from the world. This is a severe assumption and is deemed to fail at least for the year 2020 with the COVID-19 pandemic, whose social and economical impacts are expected

to reflect negatively on this quantity [5]. A more comprehensive analysis of this topic shall be explored in further detail in the upcoming years, since many of the predictions presented here may completely be shown to not hold. It is noteworthy, however, that these predictions can be used as a baseline against which the actual panorama, the current pandemic left behind, can be compared.

3. Application:

The front-end application (hosted at [7]) kindly introduces the user to the problem through a set of 3 pages, namely “Overview”, “Analytics” and “Models”. The first page (as shown in figure 14) introduces a slides-like presentation of the problem, the data, and important insights that were extracted from it, giving a comprehensive overview of the important topics that shape Colombian populations as well as influentiate the outcomes of pregnancies. Among other features, a brief introduction to the authors is also provided here.

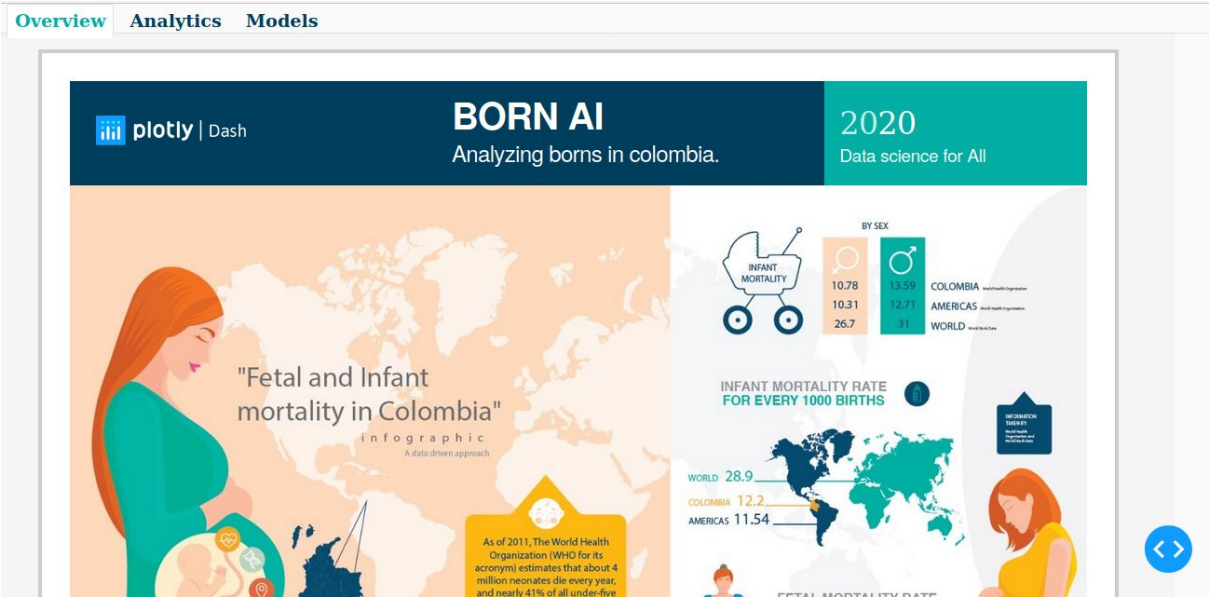

Figure 14. Overview page of the front-end application.

Analytics, the second page (shown in figure 15), features an extended list of topics that can be explored by the user through an exhaustive set of interactive options that, among many others, allow the user to change topics, dates, and geopolitical specificity, that is, she/he is allowed to visualize data for all departments in Colombia, or focus only on a single department’s municipalities just by clicking the department she/he wants to explore. This is as well connected to a bar plot to the side, which changes as the user plays with different parameters.

This tool is the main feature of the application, for it allows users to extract their own conclusions from the data that is being presented. The interface effectively delivers the data in an easy-to-explore way for the user to obtain insights without much guidance.

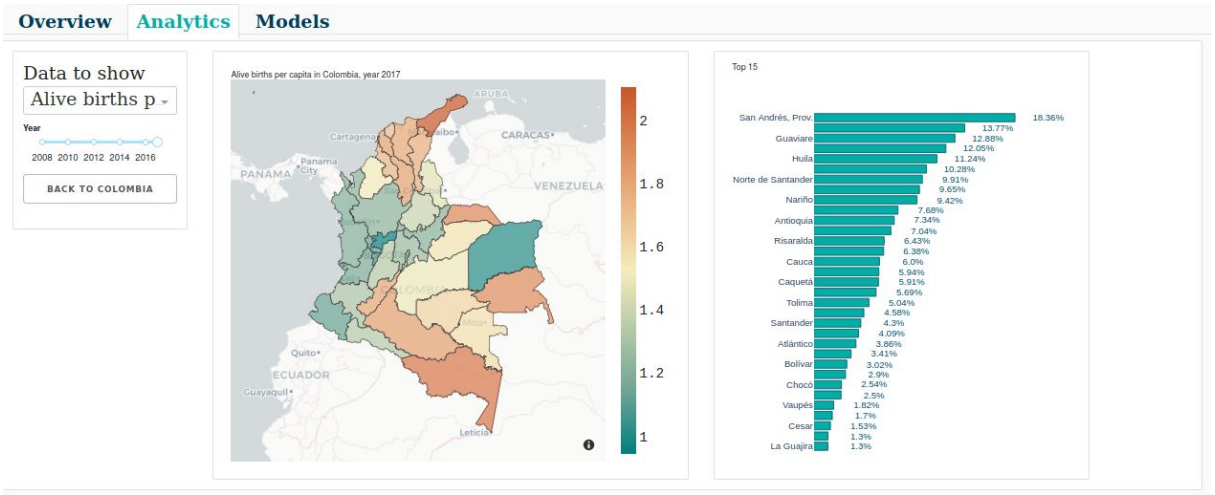

Figure 15. Analytics section.

The final page (figure 16) concludes the presentation of the project with the models that were trained, as well as some additional exploration into the variables that were used, and important insights that were extracted from the models, as discussed in previous sections. Finally, conclusions are drawn that give the user an introduction into what should be done in order to attack the problem in hand.

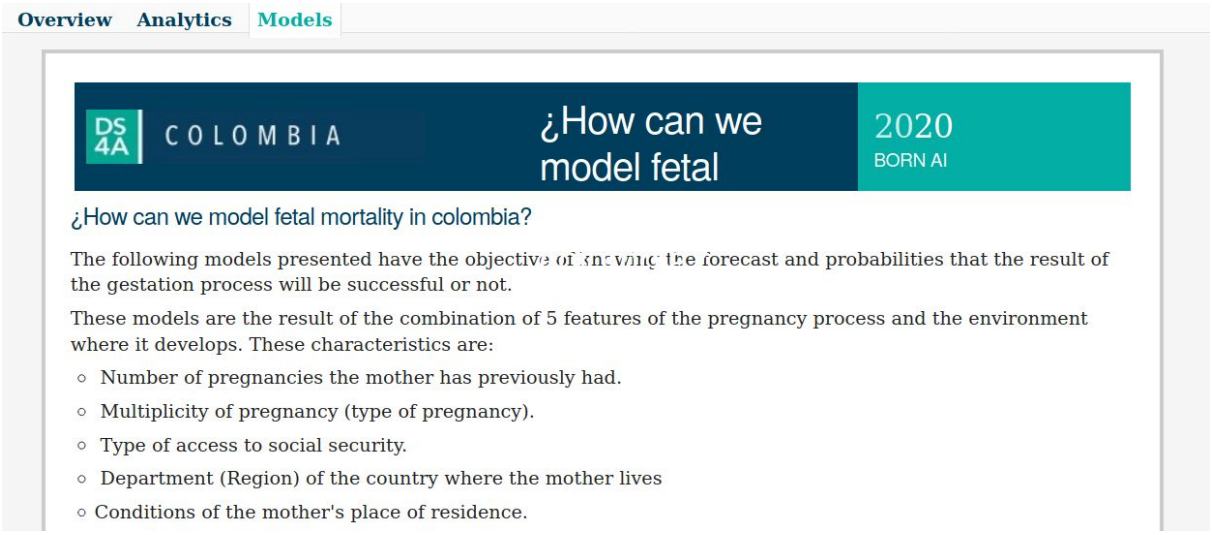

Figure 16. Models section.

4. Concluding remarks

An analysis of different populations as defined by a number of variables pertinent to pregnancies and newborns was performed, and the influence of these variables on pregnancy outcomes was assessed. Some variables -such as pregnancy multiplicity- were not surprisingly found to have an important influence on the probability of fetal death, while others related to geopolitical divisions were found to have a smaller, though as well important effect on the odds of a pregnancy being successful. It is argued that some of the results are heavily affected due to fallencies on the recollection of the data, for some regions show

surprisingly low death rates given their socio-economic conditions and, for some, it is indeed a well known fact that these show the largest infant death rates in the country, as is the case of La Guajira and Chocó, to mention a few.

In terms of future work, it is important to address the challenge of infant mortality. Although such an analysis was begun on the current work, no insightful discussion was presented due to a severe lack of time. It shall be, however, treated as a major priority aim for future works due to the numbers presented in section 1.

5. References.

[1] Newborn death and illness, World Health Organization, 2011. https://www.who.int/pmnch/media/press_materials/fs/fs_newborndealth_illness/en/
[2] Resolution adopted by the General Assembly on 25 September 2015. United Nations, 2015, p. 16 https://undocs.org/A/RES/70/1
[3] “Sex-specific infant mortality rate”, Unicef 2019. https://data.unicef.org/topic/child-survival/neonatal-mortality/
[4] https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test
[5] Half a million fewer children? The coming COVID baby bust. Kearney, Levine, 2020 https://www.brookings.edu/research/half-a-million-fewer-children-the-coming-covid-baby-bust/
[6] COLOMBIA - Estadísticas Vitales. Posted: 24 Jan, 2020. Source; Dane. http://microdatos.dane.gov.co/index.php/catalog/652/get_microdata
[7] Link to the front-end application: http://54.207.229.165:8000/

6. Appendix.

a. Dataset description.

An entire dataset of the Colombian population was sourced from the National Administrative Department of Statistics site (DANE for its acronym in Spanish) with geographical, geospatial and demographic features as well as perinatal and neonatal information [6].The unified information gathered has more than 1’855.962 data points about births, as well as fetal and neonatal deaths, described by about 60 or more columns. This could potentially be complemented by the resource distributions through the country and region characteristics, such as water quality, electricity, health facilities and education access, as well as poverty indicators and some others, sourced from the Humanitarian Data Exchange (HDX).
Since we have several possible databases with lots of features that are not directly related to newborn infants and death rates it is challenging to dig into such an amount of information and find relationships that eventually can lead us to the most insightful ones. Moreover, It will be really time-consuming to clean up, summarize and visualize all of that amount of data in order to show it in an organized and user-friendly way.

This dataset includes in its majority a set of categorical variables, which are counted as ordered natural numbers starting from 1 (1,2,3,...).

From this information we can highlight the following fields:

Common fields on all of the datasets (Fetal death, newborn death, alive newborn):

- **Location:** Department, City/Municipality of death or born. Categorical variables with Colombian standardized encoding called Divipola (for its acronym in Spanish of División Político-Administrativa)
- **Habitual residence**: Country, Department, City. Categorical values. Divipola encoding.
- **Gender**: Categorical variable (Male, Female, Unknown).
- **Born type**: Categorical variable (Natural, Cesarean, Instrumental, Ignored, No data)
- **Pregnancy type**: Categorical variable (Normal, Twin, Triple, Multiple, Ignored)
- **Pregnancy time:** Categorical variable (<22 weeks, <27 weeks, …)
- **Mother's age:** Categorical variable. Encoded in ranges of 4. (10-14,15-19,...)
- **Amount of children alive:** Discrete variable
- **Amount of children not alive:** Discrete variable

Fetal/newborn death

- **Death cause identification method:** Categorical value. (Necropsy, Medical history, Lab. tests, familiar interviews, no data)
- **Death cause:** Categorical variable. Encoded with the International Statistical Classification of Diseases and Related Health Problems, a medical classification list by the World Health Organization (CIE-10)
- **Diagnostic:** Direct, indirect, Precedent medical history, others. Categorical variable. Encoded with CIE-10
- **Probable death:** Categorical variable. (Natural, Violent, In studies)

Alive newborn

- **Weight**: Categorical variable. Encoded on ranges of 0.5 pounds (<1.000, <1.499,...)
- **Height:** Categorical variable. Encoded on ranges of 10 centimeters (<20 , <29, …)