Perinatal and neonatal mortality analysis in Colombia

Problem definition and project overview

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, babies in their first 28 days of life or the neonatal period [1]. Accordingly, perinatal 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].

Our aim is to carry out an 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 the 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.

Impact of a possible solution to the problem.

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.

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, the following set of techniques was applied in conjunction to the aforementioned graphics. Descriptive analysis: Descriptive

analysis of the data was used throughout the project with a view to identifying statistical values to achieve the proposed objectives.

- Correlations between variables that may influentiate the outcome of the pregnancy are explored.
- Clusters analysis: With the help of clustering algorithms, it is intended to find different groups of populations, with a view to predicting the outcome of the pregnancy.
- Decision trees: Based on its branched structure and its conditional control sentences, it will be possible to analyze the different options and consequences, with a view to predictive trends in pregnancy outcomes that depend on temporal, geographic and demographic data.
- Random forest: It is possible to iteratively build decision trees with different parameters within the same data, seeking to reduce possible noisy models, so as to generate more robust predictions.

Exploratory Data Analysis (EDA)

Data Cleaning. In order to concatenate the 3 datasets that were found for each year, some column names were changed and values standardized across datasets. It was observed that the dataset contained death reports with information from people of all ages, so it was reduced to the data that effectively contributed to the objective analysis, that is, data associated with over-one-year-old people was discarded. Afterwards, the missing data rate was computed and it was possible to identify columns with huge amounts of missing or unspecified data. Figure 1 shows a table sorting the columns with the highest missing-value rate, the ones with a score of 70% or more were chosen to be analysed individually, in order to determine which ones can safely be discarded.

	Porcentaje faltantes				
OCUPACION	99.978045	C_ANT12	99.396246	C_PAT1	89.974020
C_MCM1	99.828022	C_ANT22	99.348677	FECHA_NACM	80.965275
C_ANT32	99.758498	C_DIR12	99.235245	C_ANT2	80.723773
MAN_MUER	99.502360	C_PAT2	98.455853	GRU_ED1	73.182334
CODOCUR	99.487724	C_ANT3	91.982875	ULTCURPAD	71.528413
				C_ANT1	69.830583

Figure 1. Missing-values rate on columns

-

A simple exploratory analysis was carried out for each variable, aimed to allow a heuristic determination of the information content. In addition, some columns were imputed based on other variables, namely, the missing values of the field related to the number of alive successful pregnancy deliveries was imputed based on the total of pregnancies and the number of born-dead children, and the other way around to calculate the missing values of born-dead field. Furthermore, some categorical values were relabeled, since they were given as non-self-explanatory integer values, or unified to build new features; e. g. department and municipality were used to create the zip codes and individual day, month and year to create the complete date, and some other transformations.

EDA. The plots shown in figures 2 to 5 were obtained using the totality of the available data for the year 2018, as a means of visualizing the different distributions each variable has, and how they are affected by the different outcomes.
Some easily understood variables were plotted as discriminated distributions (births, fetal deaths and infant deaths), and are presented in figure 2. 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 they do not reach the total of their development. 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, 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.

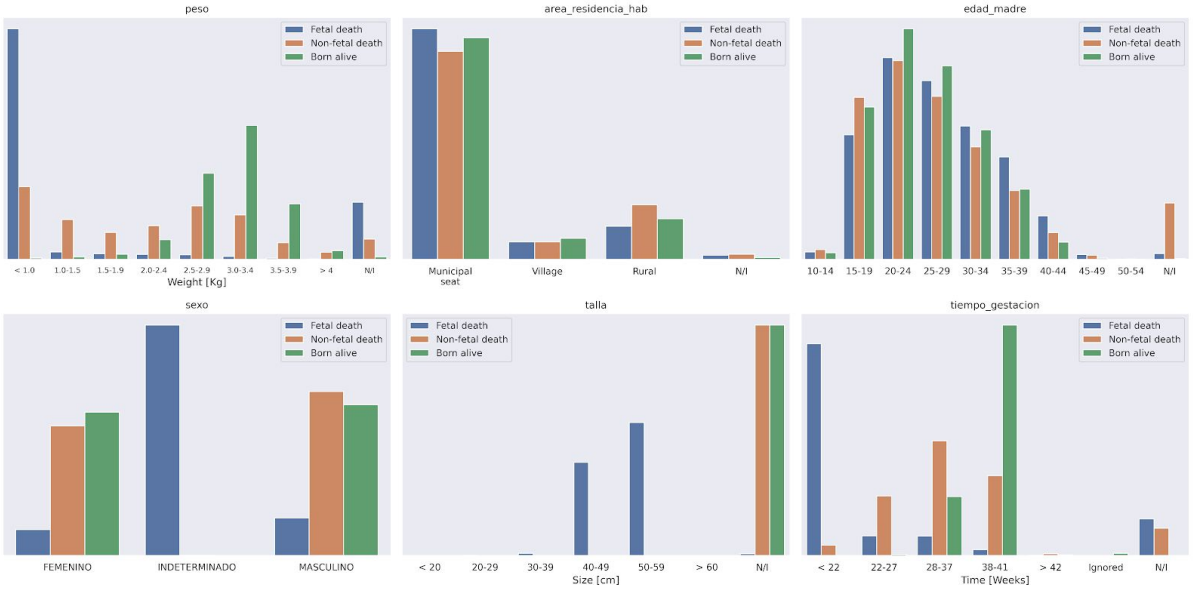

Figure 2. Distribution of some of the variables of interest, discriminated by final pregnancy outcome.

Although distributions for mother's age are similar across categories, it can be seen that pregnancy outcome is also sensitive to this variable. As can be seen from the respective plot, the probability of a fetal death increments with mother's age. 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.
Some general characteristics of mother's age distribution:

- The number of pregnancies in the age range of 10-14 is higher than wanted, as this kind of pregnancies are treated as a public health issue.
- The peak of all distributions is found in the ages of 20-24 years old.
- In the age range 35-49, it would be interesting to see if this observation holds true for the complete data set. Nonetheless, this suggests a greater risk as mothers age, which matches public health information available.

Regarding the sex and pregnancy time plots, most of the fetal deaths occur before 22 weeks, which may affect the sex determination procedure since at this stage the fetus is not fully

developed yet; that way can the sex-undetermined peak in the sex plot (which is due to fetal deaths only) be understood. Additionally, there does not seem to be a relationship between sex and the probability that the infant will be born alive or suffer a fetal or non-fetal death.
Finally, 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.
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.
From the baby length graph, we can see that we only have available data for non-dead newborns.
Some more plots can show further insights on this data and on Colombian population in general. In figure 3 the father and mother's age distribution are shown, and it can clearly be seen some slight differences. Specifically, 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 our country, it would be interesting to dig deeper into this matter and check what is happening to mothers in ages between 10-19 and verify the possible pregnancy outcome in this specific scenario

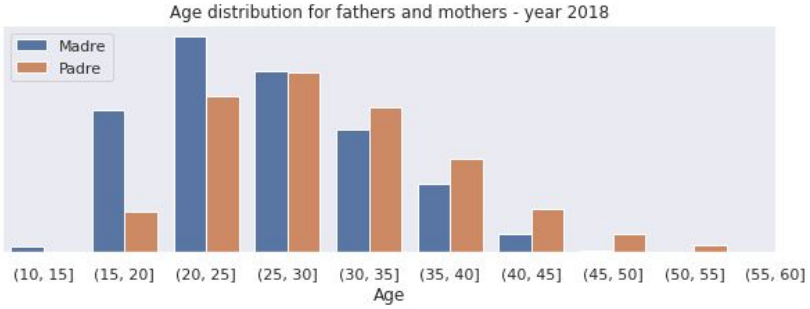

Figure 3. Age distribution of both male and female parents.

The distribution of ages of post-birth deaths is shown in figure 4. Left plot shows the raw numbers categorized by different ages. Here we can see that most of these deaths occur between 1-6 days, and 1-5 months. However, the time-normalized distribution (right plot), 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 where the focus should be set: 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.

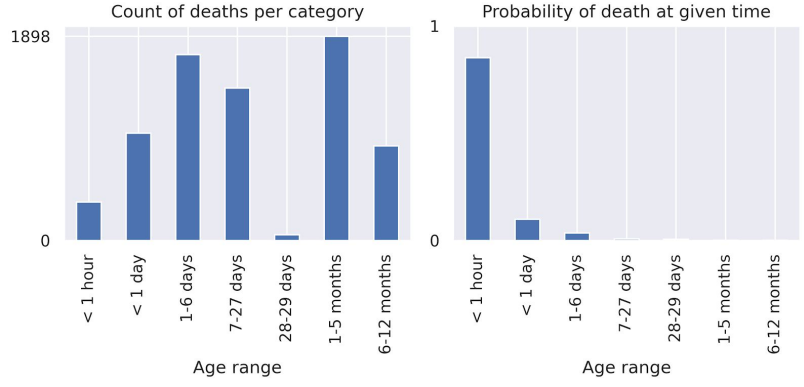

Figure 4. Distribution of death ages for post-birth deaths.
a. Absolute numbers (no normalization).
b. Time normalization, accounting for the different time sizes of the bins. This can be interpreted as a time-dependent death-probability.

Lastly, pregnancy outcome counts were plotted accounting for the different racial and cultural

identification of the person in question (figure 5). 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, this might be due to cultural and/or behavioural differences among these populations; however, we must not discard the possibility of the presence of systematic discrimination against some of these populations, as well as to their territories. This can further be observed in choropleths and other forms of geographical visualization techniques. Most of the conclusions exposed here were obtained from the information observed on plots so a confirmatory analysis will be run in order to make sure that all above are actually true and not only by randomness.

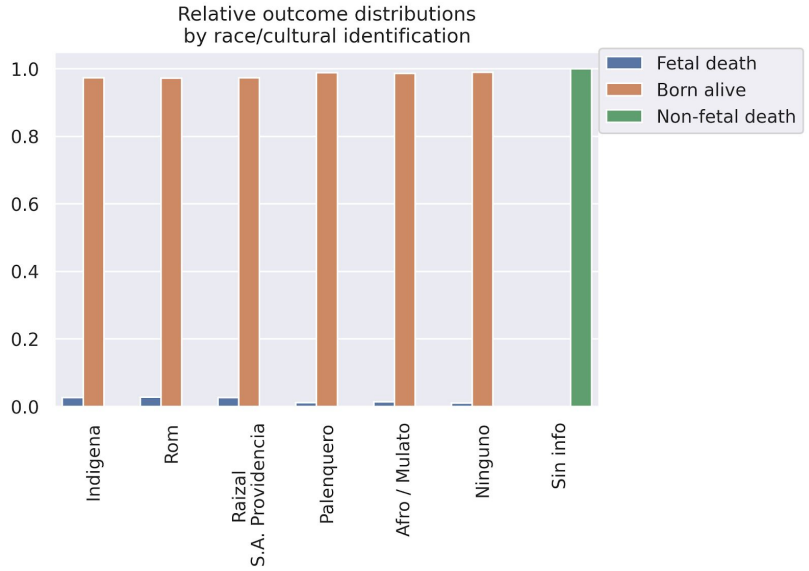

Figure 5. Pregnancy outcome as discriminated by racial/cultural identification. The scale is not the same for every distribution, so as to observe differences in distributions among categories.

Data from 2008 to 2018 was added and plotted, as a result, it was possible to visualize that children born alive are actually a stationary phenomenon. As figure 6 shows, it has very high peaks and low valleys in almost the same periods, which is an interesting seasonal behaviour; further investigation can be done in order to identify the reason why this is happening and maybe correlate this with some other variables.

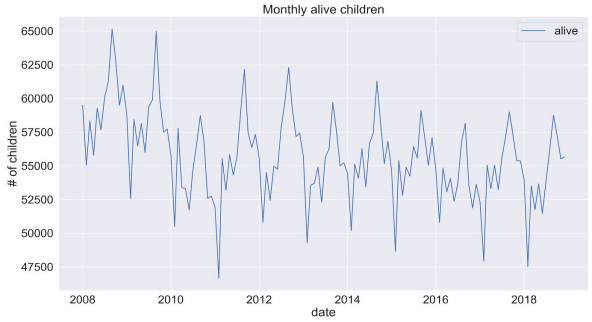

Figure 6. Monthly alive children

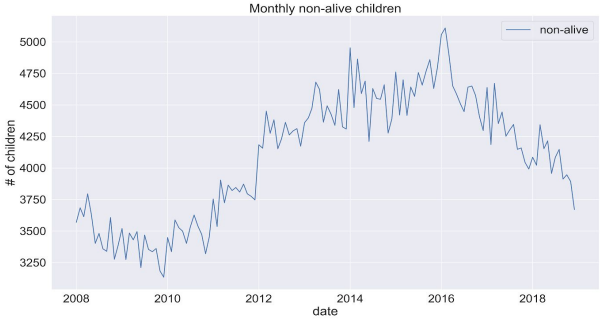

Figure 7. Monthly non-alive children

Separately, the monthly number of non-alive children, both stillborn and new death borns, were plotted as a function of time (figure 7). From this can be seen that in the period from

2010 to 2016 the amount of deaths were increasing but then suddenly it started to decrease. Finally, the non-alive children count was plotted as rates, identified in orange in figure 8, It roughly shows it's stationarity and its maximum value is not greater than 10%.

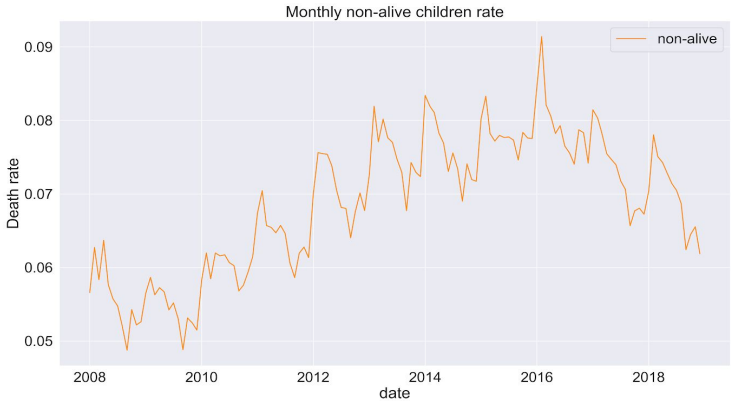

Figure 8. Monthly children death rate.

Death and birth rates by department:

One of the best ways to visually analyze the data, even without using any model, is to observe the mortality rates by departments in a geopolitical fashion. This could allow us to observe and identify strange behaviours in some departments, which can be useful for subsequent analysis.

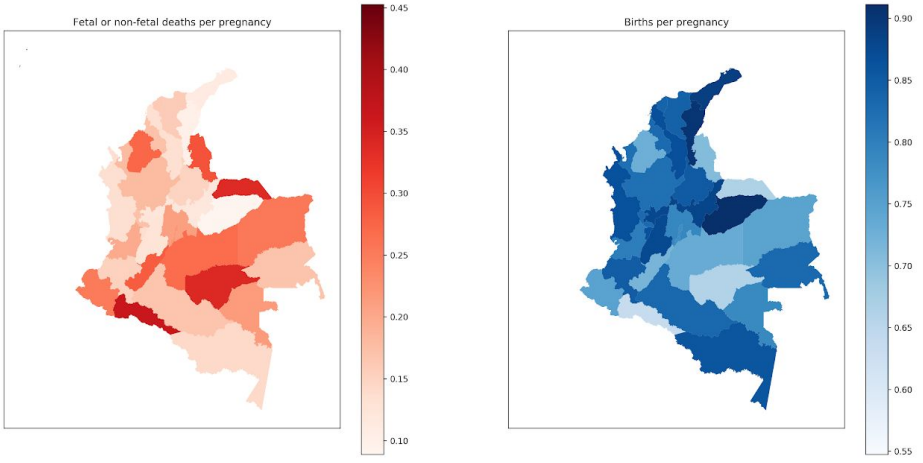

Figure 9.a) On the left, the heat map representing the probability of death (fetal or non fetal) per department. b) On the right, the heat map representing the probability of successful birth per department.

In the first place, in figure 9 it is possible to observe the rates of total deaths (fetal and non fatal) and births per department. At a glance it is possible to identify that Cauca, Putumayo, Arauca and almost all the Orinoquia region are problematic regions when we talk about deaths in the first year of life. In the same way, when we look at the image of births per pregnancy it can be seen that successful pregnancies are more possible in the Andean region, but is important to look carefully what happen in the periferia because many of those regions do not have many inhabitants so the rates could change dramatically for specific problems and we want to keep our attention in the bigger picture.

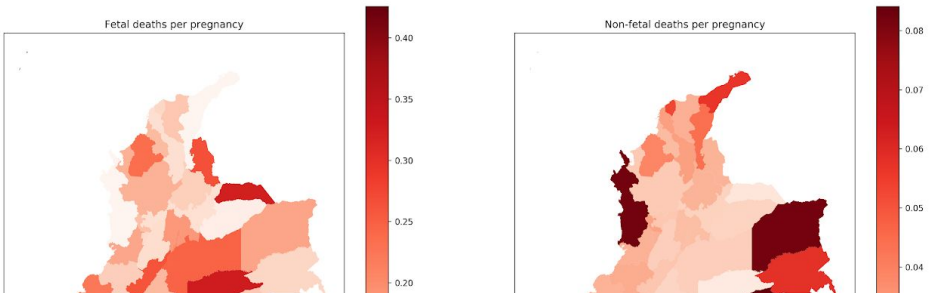

Figure 10.a) On the left, the heat map representing fetal deaths per

pregnancy. b) On the right, the heat map representing the non fetal deaths per pregnancy.

Another important observation is about how deaths occur in each department. In poor, mainly rural departments, the fetal deaths could be undervalued because maybe people do not have the capability to go to hospitals, but non-fetal deaths are different because people must report the baby's death since they are already a citizen. In figure 10 is possible to observe that rural departments in the periferia have much bigger rates when we talk about non fetal deaths but the rate of fetal deaths is almost zero in those departments; on the other hand the central departments of the Andean region have bigger rates for fetal deaths but much lower rates for non fetal deaths. One possible explanation is that in the central region the access to hospitals is better so people could go even for small problems in pregnancy, this on one side increase the possibility of having a healthy baby and on the other side It also increase chance of losing baby while being in the hospital, leading to miscarriage being reported more frequently, this can be contrasted and proved later if we add hospital distribution across the country.

Correlation analysis

The variables were splitted into continuous and categorical, in order to find correlations between them. The correlation methodology was applied to continuous variables, as a correlation matrix presented in figure 12
Here, we find that some of these variables have a correlation between them, for instance, the number of pregnancy controls (NUM_CONSUL) is highly correlated with father's age (EDAD_PADRE) and size of baby (TALLA), also, the variable number of pregnancies (NUMERO_EMBARAZOS) has a high correlation with number of both dead and alive children (N_HIJOSV and N_HIJOSM).

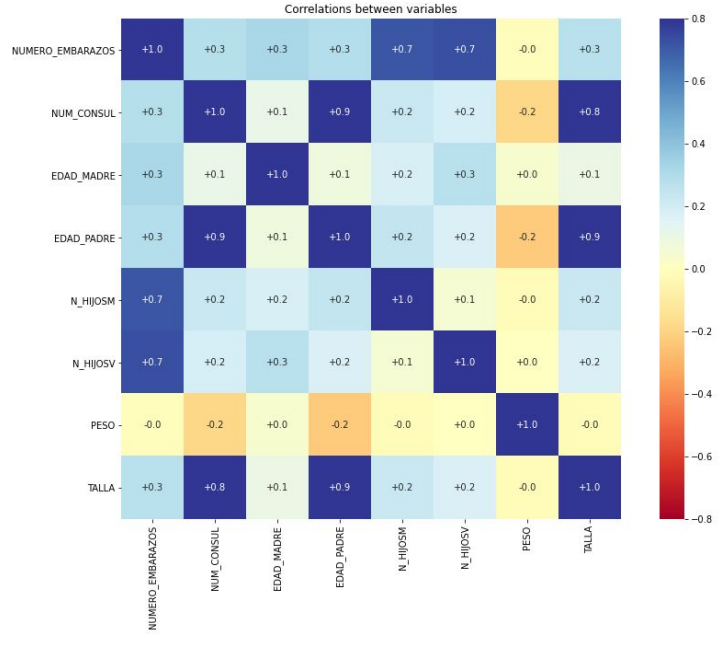

Figure 11. Correlation matrix performed on continuous variables on the dataset.

Subsequently, the exploration of models that could predict the outcome of pregnancy begins with these variables. To do this, a new column called "NACIDO_VIVO_INV" is used where the value 1 represents the subjects who died and 0 the subjects who lived. When training a model (using the natural log) with the variables "NUMERO_EMBARAZOS", "EDAD_MADRE" and "PESO" which correspond to the number of pregnancies of the mother, age of the

mother and weight of the baby respectively, we obtain the results shown in figure 13. From this we can see some interesting coefficients, the R2 is however very low, which indicates that the model does not fit well with our desired result, being this the reason why further work must be conducted over the categorical variables so as to include these in the model.

Front-End application

The final front-end product will feature ideally three pages: an overview page, an analytics page and a recommendations page. The overview page, figure 14, will have some summarized statistics and maybe some context about Colombia's current situation in relation to this topic. Also, some extra information to allow users to effectively understand all the content. The Analytics page contains maps, as a main plot, showing perinatal and neonatal death rates by department as well as by municipalities. Already implemented interactivity allows the user to have a look at the data for all of Colombia while displaying by year and variables of interest; additionally, the user has the option to dig into these same features by municipality, by clicking on a specific department. The recommendations page will also have some visualizations of the historical data, and graphs showing possible correlations between the pregnancy outcome and the most relevant descriptive variables, as well as some results gotten from the models created.

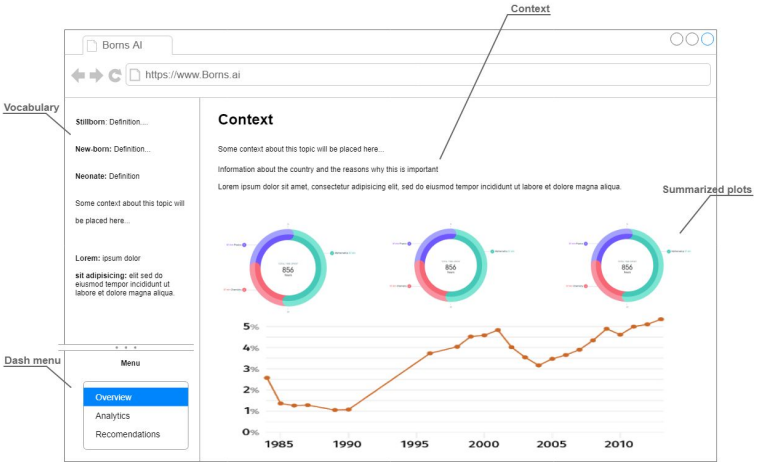

Figure 14. Overview page.

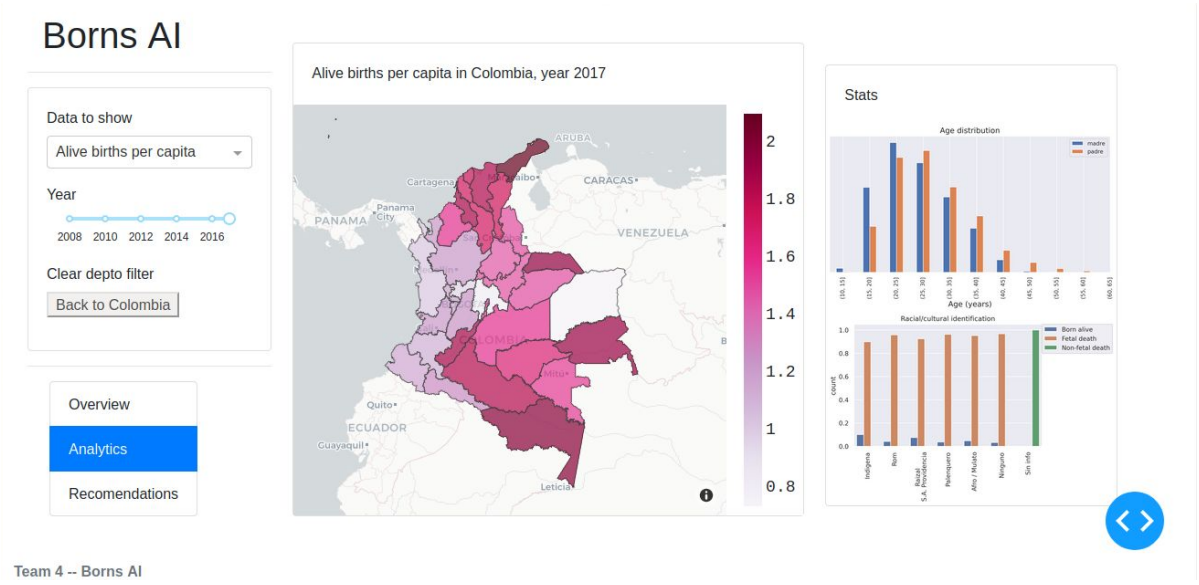

Figure 15. Analytics page, already implemented as of week 8.

Final notes.

This week we decided to divide our problem in two:

- Problem A: compare fetal deaths and births to analyze what makes a pregnancy successful
- Problem B: compare births and infant deaths to analyze infancy survival

For problem B, we have a challenge with the data, which is to identify which births correspond to an infant death. We are currently working on that problem. Related to this, since we are taking into account deaths until one year of age. Information from 2018 was discarded since it is the unique year with different format on the dataset, for these we still don't know if they resulted in death.

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 [3].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, …)

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] COLOMBIA - Estadísticas Vitales. Posted: 24 Jan, 2020. Source; Dane. http://microdatos.dane.gov.co/index.php/catalog/652/get_microdata