Building a prediction model for neonatal anemia among neonates aged 0-59 months in Nigeria

Ingmar Tulva, Sviatoslav-Oleh Savchak, Subhashini Muhamdiram

Business understanding

Background

Anemia in newborns is a condition where the baby's body has a lower red blood cell count than normal. While it is common for most babies to have neonatal anemia during the first few months after birth, it can be associated with short or long term consequences such as poor growth, decreased activity and limited cardiovascular reserve. Therefore, it is important for the doctor to monitor the babies for signs of anemia closely in order to prevent any adverse consequences.

Infants under six months of age are at high risk of anemia due to their rapid growth and limited iron intake, as breast milk is low in iron. For this reason, they rely mainly on iron from intrauterine life. It is even more devastating when the pregnant mother, who is the source of fetal iron, is anemic, suggesting a critical relationship of mother factors on anemia in neonates. In the current study we have data from 34000 Nigerian children aged 0-59 months and their mothers.

Anemia in neonates can be diagnosed by a blood test which measures hemoglobin (a protein in red blood cells that carries oxygen in the blood). Although anemia is a very big public health concern, newborns, especially in developing countries, are usually overlooked and undiagnosed. Testing or examining all babies for signs of anemia specially from low income countries seems to be challenging as it is financially and technologically intensive. If the anemia is not corrected on time, irreversible long-term complications such as bone diseases, liver and spleen enlargement, growth disorders, decreased motor activity, social inattention, and severe cognitive impairments will follow, with additional health care costs to the provider. According to studies conducted in sub-Saharan Africa, the prevalence of newborn anemia was 35% in Nigeria, 57% in Ghana, 23% in Malawi, and 61% in Benin. Therefore, reducing anemia prevalence among neonates in these countries is still a significant public health challenge.

Business goals

Majority of neonatal anemia is caused by nutritional status of the babies and mother’s health. Identifying factors closely associated with neonatal anemia will allow for relevant and timely interventions, while reducing cost of testing. An efficient model could predict neonates at risk, enabling timely interventions. This way low income countries can cut the cost of expensive blood tests on large cohorts, while identifying as many anemic neonates as possible to avoid adverse consequences. Policy makers can use the findings

of this analysis to plan maternal health education and appropriate health interventions to reduce the problem of anemia among neonates in Nigeria.
The main objective of this project is to build a prediction model to predict the potential risk of neonatal anemia among neonates aged 0-59 months in Nigeria based on their mothers' health and socio-economic factors.

Business success criteria

Achieving around 90% sensitivity and specificity for the prediction model based on the available data

Assessing the situation

Inventory of resources – A dataset, three personal computers, python scripting language with required libraries, and three people with basic knowledge on data science.
Requirements, assumptions, and constraints – All work to be finalised by 11th December. Report, written scripts, poster will be finalised by the given date. All necessary data already obtained.
Risks and contingencies –Time available for the project completion is limited. Data analysis is already initiated and will be updated daily and discussions will be carried out to finish the tasks on time.
Terminology – No nonstandard terminology used.

Costs and benefits – Time and electricity will be spent for the data analysis. Poster printing cost will be covered by the University. As data is already collected, there is no associated cost for data collection. People who analyse the data will gain experience. The predictive model developed during the project (if with a good sensitivity and specificity) can be used to predict the potential risk of an infant for anemia in Nigeria and policy makers can use it to take initiative to prevent it through the timely interventions such as by blood testing and educating mothers. Costs associated with global screening can be reduced by targeting high risk groups. Costs associated with later life health problems of the newborn due to anemia can also be significantly reduced if diagnosed properly. So, the project seems to have more benefits over the cost associated with it.

Defining the data-mining goals

Data-mining goals – Develop model(s) that predict children who likely need laboratory testing for anemia or are at increased risk of having anemia. Designing a poster with a summary of the analysis. There are significantly high missing values for anemia status in the sample, so the developed prediction model will be used to predict their anemia status.

Data-mining success criteria – Predictive model quality will be evaluated primarily by sensitivity and specificity of the model.

Data understanding

Gathering data

Data requirements – The project requires data related to anemia status of neonates of 0-59 months in Nigeria along with their mother's age and other socio-economic factors that might affect neonatal anemia. One file with all relevant data is available.

Data availability – The project will use data freely available from cross-sectional study conducted by Nigeria Demographic and Health Surveys (NDHS) in 2018. It primarily consists of data related to the effect of mothers' age and other socioeconomic factors on anemia levels of children aged 0-59 months in Nigeria. The data is already collected and available in Kaggle. https://www.kaggle.com/datasets/adeolaadesina/factors-affecting-children-anemia-level. The data file is deposited in the Git-hub repository https://github.com/svyat1kk/DSproject.

Selection criteria – Quality of the data will be by assessing factors such as completeness, accuracy, consistency, and timeliness. Data without missing values, inconsistencies, or inaccuracies will be selected for predicting the model.

Describing data – The following information has been collected during the survey. Following column names were found, but there was no proper data dictionary. The data card in Kaggle needs more clarification regarding information in each column. So we have to assume the meaning of some columns.

Age in 5-year groups: Mother's age
Type of place of residence: Place of living of mother and her baby
Highest educational level: Mother's education
Wealth index combined: Mother's income
Births in last five years: Pregnancy status of the mother
Age of respondent at 1st birth: Mother's age when given the first birth
Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal): Mother's Haemoglobin level in blood
Anemia level: Mother's anemia level
Have mosquito bed net for sleeping (from household questionnaire):
Smokes cigarettes: whether mother smokes cigarette
Current marital status: Marital status of the mother
Currently residing with husband/partner
When child put to breast
Had fever in last two weeks: Neonates data?
Hemoglobin level adjusted for altitude (g/dl - 1 decimal): neonatal Heamoglobin levels

Anemia level.1: Neonatal anemia level
Taking iron pills, sprinkles or syrup: External iron supplementation for the neonate

In total there are 33924 rows/cases. There are significant amounts of missing values that need to be cleaned in order to use it for developing a predictive model.
Number of missing values in each column is as follows (columns with no missing values omitted):

Column	**Missing values**
Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)	20788
Anemia level	20788
Currently residing with husband/partner	1698
When child put to breast	12756
Had fever in last two weeks	3323

Additionally, two columns (Taking iron pills… and Had fever…) contain a handful of “Don’t know” values, which may better be treated as missing values as well.

As seen above, the blood parameters for the majority of the neonates and mothers are missing. Few cases in “Currently residing with husband/partner” and “Had fever in last two weeks”, “Taking iron pills, sprinkles or syrup” are also missing. The data in the column in “when child put to breast” is inconsistent and not very clear; given the multiple missing cases, this feature shall likely be dropped from the analysis altogether.

Exploring data – Anemia diagnosis results in 4 different categories (not anemic, mild anemic, moderate anemic, and severe anemia). However, the data in this column will be converted to two values (anemic or non-anemic) and this will be used as the target variable for a binary classification model. Hemoglobin level of the infant can be used as a continuous target variable and can be used in a regression model. Mother’s age group can be converted into a usable numeric feature.

Verifying data quality
Data columns with inconsistencies will be discarded, as will the remaining rows with missing values. After removing all inconsistent and missing data, still we will have around 9515 cases which is adequate for the modelling task. All variables will be

converted to numeric values before modelling. No abnormal distribution or inconsistent data seen after cleaning.

Planning the project

Make a detailed plan of your project with a list of tasks. There should be at least five tasks. Specify how many hours each team member will contribute to each task -

Task	Time in Hours for each student	Team member
Data Preparation Selecting data (removing missing values and inconsistent data), transforming values as needed. Data with missing values for anemia will be prepared separately to use after establishing the model.	4,4,4	Subhashini, Ingmar,Sviatoslav-Oleh
Exploring the data for frequent patterns and associations.	4,4,4	Subhashini, Ingmar,Sviatoslav-Oleh
Data cleaning for modelling. Converting the data values into numeric format. New target variable will be defined in a new column for anemia.	2,2,2	Subhashini, Ingmar,Sviatoslav-Oleh
Selecting appropriate modelling techniques for the data	2,2,2	Subhashini, Ingmar,Sviatoslav-Oleh
Assessing the performance of different models and use it to predict missing data	4,4,4	Subhashini, Ingmar,Sviatoslav-Oleh
Summarising the key findings of the analysis. Report writing and poster preparation	8,8,8	Subhashini, Ingmar,Sviatoslav-Oleh

List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

Regression models (Linear regression, Lasso) for studying the dependence of the child's hemoglobin level on maternal data. With the number of cases vastly exceeding the number of features, overfitting is not likely to be a problem, but lasso regression may provide better insight into the relevance of the factors.

Classifiers (classification trees, random forest, logit regression, others that may come into mind on the way) for studying anemia diagnosis on maternal data. The rationale is that even mild cases of anemia need to be found and treated, therefore it makes most sense to replace the four-level diagnosis with a binary one, an approach which hopefully results in better prediction accuracy than the above regression.

Models resulting from the above approaches to be used for exploring the part of the dataset where the children have not been medically diagnosed, to assess the magnitude of the untested anemia problem.