**TITLE**

ANEMIA LEVEL PREDICTION IN CHILDREN

**SUBTITLE**

Utilizing Predictive Analytics for Anemia severity in children

**COLLABORATORS**

1. Ruth Kitasi
2. Agatha Nyambati
3. Joseline Apiyo
4. Cecilia Ngunjiri
5. John Mbego
6. Leonard Koyio

**PROJECT SUBMISSION DATE**

22nd November 2024

**EXECUTIVE SUMMARY**

**PURPOSE**

Childhood anemia is a major public health issue, especially in Nigeria, where it often goes undiagnosed and untreated due to limited healthcare access, high diagnostic costs, and low awareness. This condition adversely affects cognitive and physical development, education, and quality of life. To address this, a project aims to develop a classification model to predict the severity of anemia in Nigerian children under 59 months. The model will analyze key socioeconomic factors to identify those that most contribute to the risk and severity of anemia, providing insights for targeted interventions that could improve health outcomes for vulnerable children.

**METHODOLOGIES**

CRISP-DM Methodology

1. **INTRODUCTION**

**OBJECTIVES**

The main objectives of our project are;

* To develop a classification model for Anemia Severity in children. This model will aim at distinguishing between mild, moderate, severe and not anemic cases.
* To identify key socioeconomic determinants influencing Anemia severity. Conduct an in depth analysis to pinpoint specific socioeconomic factors that have the most significant impact on anemia severity among young Nigerian children.
* To evaluate the model’s predictive accuracy for effective Public Health Use. Assess the model’s predictive performance using appropriate metrics ensuring high accuracy and reliability to provide a robust tool healthcare professionals can rely on.
* Classify and Monitor Anemia Severity Level. Using the model to classify anemia cases into severity levels(mild, moderate, severe, not anemic)

**BACKGROUND**

Anemia is a significant global health issue, particularly affecting children under five, with risks to growth, cognitive development, and immunity. The World Health Organization estimates that 42% of children worldwide are anemic, mainly due to iron deficiency, poor diet, and infections. In Africa, the prevalence is even higher, ranging from 60-80%, driven by malnutrition, infectious diseases, and poor healthcare. Nigeria faces one of the highest burdens, with 68% of children under five affected. Socioeconomic factors like poverty, limited access to nutritious food, and inadequate healthcare exacerbate the issue, leading to long-term health and development challenges. Addressing anemia in Nigeria is crucial for improving children's health and national development.

**PROJECT STATEMENT**

Childhood anemia is a critical public health issue, leading to adverse impacts on cognitive and physical development. This project aims to build a predictive classification model to determine anemia status in children based on a comprehensive dataset encompassing demographic, health, and socio-economic factors. By identifying high-risk groups and predicting anemia likelihood, this model will support targeted interventions, optimize resource allocation, and guide public health policies. The ultimate objective is to reduce anemia rates in children, leading to improved health, educational outcomes, and quality of life.

1. **DATA COLLECTION AND DESCRIPTION**

* Data Source

The dataset we used for our project was sourced from ([Kaggle](https://www.kaggle.com/datasets/adeolaadesina/factors-affecting-children-anemia-level))

* Data Variables i.e. columns in the dataset

|  |  |
| --- | --- |
| Column Name | Description |
| Age in 5-year groups | Age of the mother, categorized into 5-year intervals. |
| Type of place of residence | Residential classification: Urban or Rural. |
| Highest educational level | Mother's highest educational attainment, e.g., "No Education", "Secondary". |
| Wealth index combined | Economic status of the mother, ranked as "Poorest", "Poorer", etc. |
| Births in last five years | Count of live births by the mother in the past five years. |
| Age of respondent at 1st birth | Mother's age at her first live birth |
| Hemoglobin level adjusted for altitude and smoking | Hemoglobin level in g/dL, adjusted for altitude and smoking status. |
| Anemia level | Categorical anemia status of the child (e.g., "Not anemic", "Mild", "Moderate", "Severe"). |
| Have mosquito bed net for sleeping | Presence of a mosquito net in the household (True/False). |
| Smokes cigarettes | Mother's smoking status (True/False). |
| Current marital status | Mother's marital status (e.g., "Married", "Single"). |
| Currently residing with husband/partner | Co-habitation status with spouse or partner (True/False). |
| When child put to breast | Timing of initial breastfeeding (hours after birth). |
| Had fever in last two weeks | Fever incidence in the child during the past two weeks (True/False). |
| Hemoglobin level adjusted for altitude (g/dl) | Hemoglobin level in g/dL, adjusted solely for altitude. |
| Anemia level.1 | Alternative or duplicate anemia categorization. |
| Taking iron pills, sprinkles or syrup | Iron supplementation status of the child (True/False). |

* Data Summary

The anemia dataset has 33,924 rows and 17 columns

The dataset is grouped into two categories of data i.e.

**Categorical Data** which includes;

Type of place of residence, Highest educational level, Wealth index combined, Current marital status, Currently residing with husband/partner, Had fever in last two weeks, Anemia level, Anemia level.1,Have mosquito bed net for sleeping, Smokes cigarettes, When child put to breast, Taking iron pills, sprinkles or syrup

**Numerical Data** which includes;

Age in 5-year groups (ordinal), Births in last five years, Age of respondent at 1st birth, Hemoglobin level adjusted for altitude and smoking, Hemoglobin level adjusted for altitude (g/dl)

* Missing Data and Handling Missing Values

During our analysis, we identified missing values in several columns. We decided to drop the following columns: **Anemia level.1**, **Hemoglobin level adjusted for altitude (g/dl)**, **Smokes cigarettes**, and **when child put to breast**. These variables had minimal correlation with our target variable, **Anemia Level**, and dropping them helped streamline our dataset without losing valuable predictive information.

We chose to drop missing values in the **Anemia Level** column. This decision was made to ensure that our data remained consistent and unbiased. Filling missing values in **Anemia Level** with an "Unknown" category could introduce potential biases and distort ours. Instead, by dropping these rows, we preserved the quality of the data and ensured more reliable model performance.

For the missing values in our **categorical data**, we filled the null values with ‘Unknown’, as this allowed us to maintain the integrity of the categorical variable without compromising the analysis. For **numerical data**, we dropped rows with missing values to avoid introducing any uncertainty or errors that could arise from imputation methods.

* Renaming Columns

We renamed several columns for clarity and ease of interpretation. The changes include:

Have mosquito bed net for sleeping (from household questionnaire) to Mosquito net, Highest educational level to Education level, Wealth index combined to Wealth, Currently residing with husband/partner to Living with spouse, Type of place of residence to Area\_Type, Taking iron pills, sprinkles or syrup to Taking\_meds, Age in 5-year groups to Age\_group, Had fever in last two weeks to Had fever, Current marital status to Marital status.

* Encoding of Categorical Columns

Encoding categorical columns is the process of transforming categorical (non-numeric) data into a numerical format so that it can be used effectively in machine learning models and other analytical processes that require numerical inputs. Since our categorical columns contains a meaning order, we proceeded and applied the ordinal encoding.

1. **EXPLORATORY DATA ANALYSIS**

Explanatory Data Analysis involves examining and visualizing the dataset to highlight key characteristics and patterns using statistical graphics.

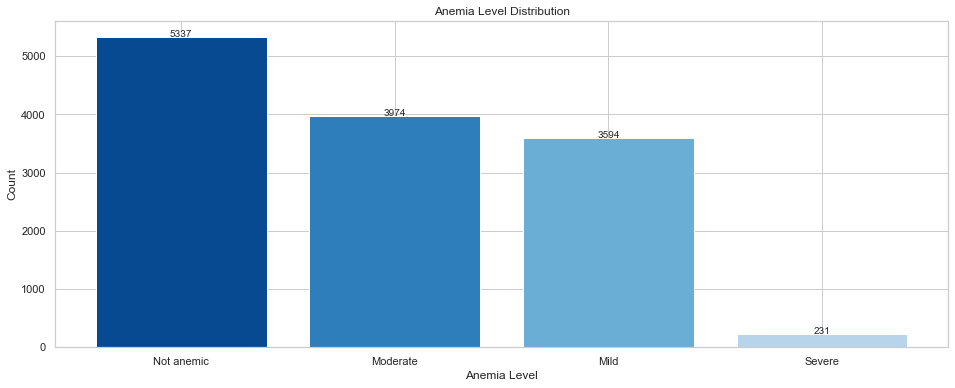
We handled our EDA in 3 Analyses i.e. Univariate analysis, Bivariate analysis and Multivariate analysis

* Univariate Analysis

Univariate analysis examines each variable independently to understand its distribution, central tendency, and variability.

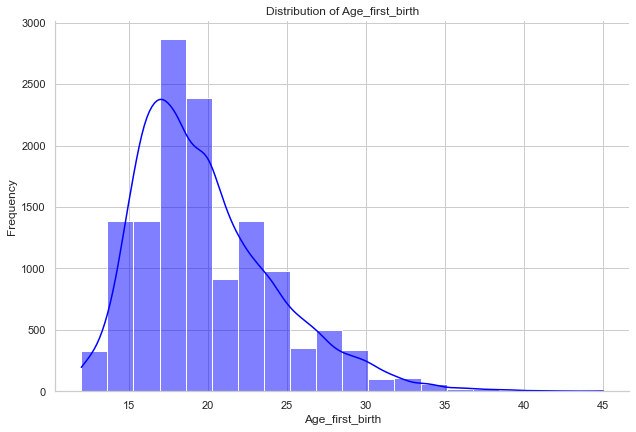
We analyzed;

**Distribution of Anemia Level**



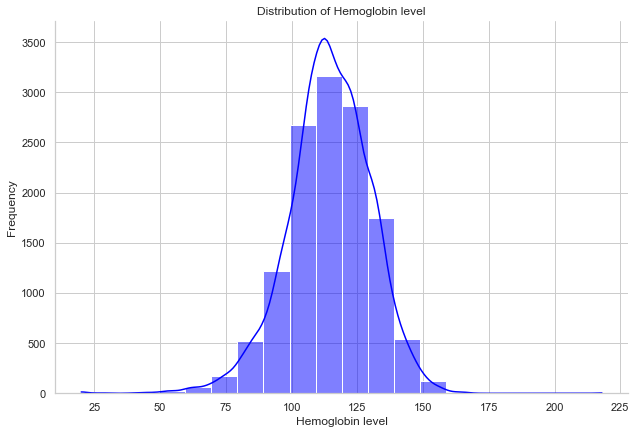
The distribution of anemia levels shows a significant imbalance, with the majority of individuals classified as "Not Anemic." The "Severe" category has the least number of individuals. The class imbalance can pose challenges during model training. If left unaddressed, the model may be biased towards the majority class and struggle to accurately classify instances from the minority classes.

**Distribution of the Age\_first\_birth**



The distribution is skewed to the right, indicating that the majority of respondents had their first child at a younger age, particularly between 15 and 25 years. This distribution suggests that early childbearing is more prevalent in this dataset, while fewer respondents experience first-time childbirth later in life.

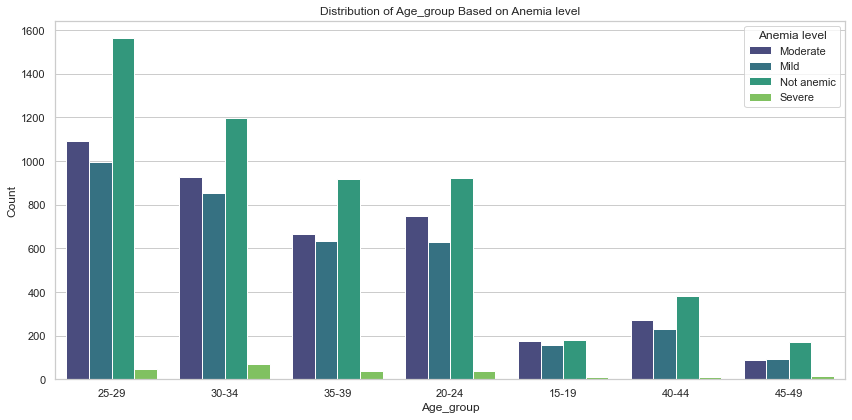
**Distribution of the Hemoglobin level.**



The distribution is approximately bell-shaped and symmetrical, with most hemoglobin levels concentrated between 95 and 125 g/dL. The distribution has a narrow spread, with relatively few respondents having hemoglobin levels below 75 or above 150. This shape implies that hemoglobin levels are normally distributed among the respondents after adjusting for altitude.

* Bivariate Analysis

**Relationship of Anemia level and parent's age\_group**



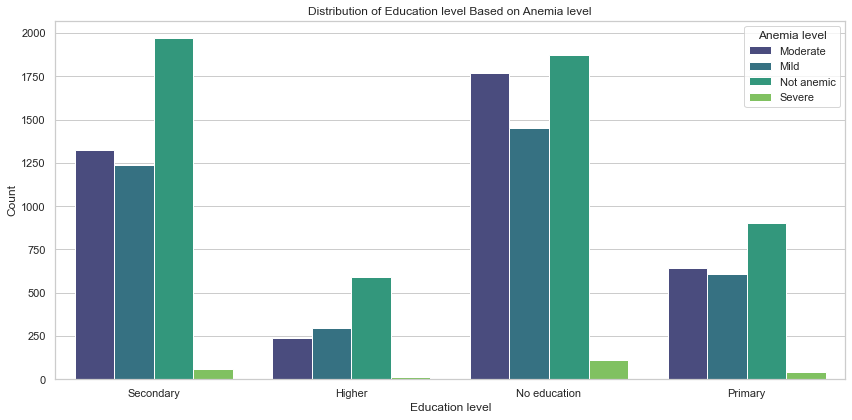
Children of mothers in the 20-39 age range show higher counts of moderate anemia compared to children of mothers in other age groups, with the peak in the 25-29 and 30-34 age groups. This trend may indicate that maternal age influences children's anemia levels, potentially due to factors associated with the health, nutritional status, or socioeconomic conditions of mothers in these age ranges

**Relationship of Anemia level and Area\_Type**



In rural areas majority of children fall under the "Moderate" anemia level, with the count significantly higher than other categories while in urban areas, the counts are more balanced across the anemia levels compared to rural areas, but "Moderate" anemia is still common. This distribution suggests that moderate anemia is more prevalent in rural areas than in urban areas. This could indicate potential socio-economic or healthcare access disparities between rural and urban settings that influence anemia severity.

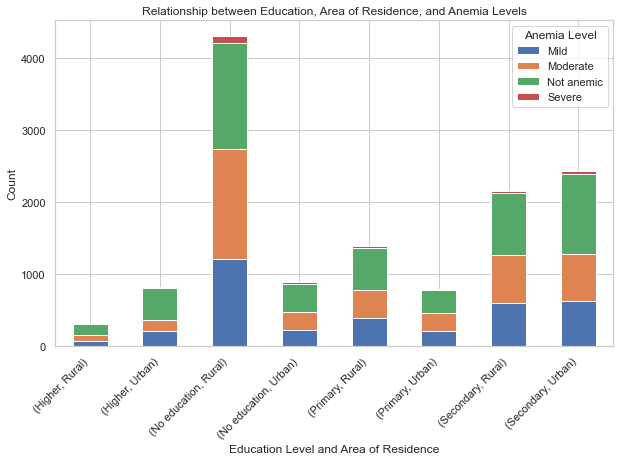
**Relationship of Anemia level and parent's Level of education**



The analysis shows a clear relationship between education level and anemia severity. Children whose parents have no education have the highest prevalence of moderate anemia, suggesting that lack of education is associated with poorer anemia outcomes. In contrast, children whose parents have higher education have the fewest cases of moderate and severe anemia, indicating that education may play a protective role against anemia. This could be due to greater health awareness, better nutrition, and access to healthcare among educated individuals.

* Multivariate Analysis

Relationship of Anemia level, Education level and Area\_type



This chart suggests that **education level** and **area of residence** are both influential factors in anemia status. Higher education levels, particularly in urban areas, appear to be associated with lower rates of anemia. Conversely, individuals with no education, especially in rural areas, show a higher prevalence of anemia, indicating possible socio-economic and environmental factors influencing health outcomes.

1. **DATA PREPROCESSING**

* Data Splitting

Data splitting involves dividing the dataset into separate subsets for training and evaluation. The primary goal is to assess how well a model generalizes to unseen data.

We split our dataset into **X** and **y**. **X** represents all the **independent variables** (features) that will be used as inputs for our model, while **y** is the **target variable** (the outcome we aim to predict), which in this case is the anemia level.

For the independent variables (X), we performed feature selection by identifying and selecting the relevant features from our main dataset and in this case, we used the top 10 features with the highest correlation to Anemia level which are ; Hemoglobin level ,Wealth ,Area\_Type ,Age\_first\_birth, Age\_group ,Mosquito net ,Education level ,Marital status ,Living with spouse, Had fever, Taking\_meds, Births\_last\_5y

* Feature Selection

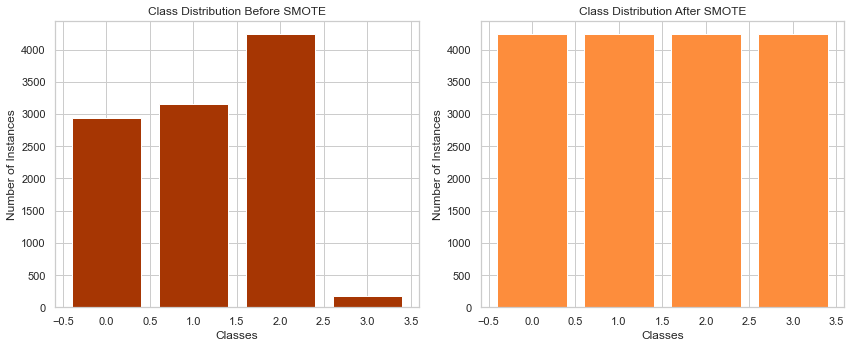
The process of identifying and selecting a subset of relevant features from the main dataset. The primary goal of feature selection is to improve the performance of machine learning models by reducing overfitting, improving accuracy, and decreasing training time.

We did this by selecting the top 10 features with the highest correlation to Anemia Level.

* Handling Class Imbalance

Class imbalance is a common issue in classification problems, occurring when the target variable's classes are not represented equally. In our Anemia level, the severe anemia class is underrepresented compared to other classes, which suggests a class imbalance which we handled using SMOTE (Synthetic Minority Over-sampling Technique). It works by balancing class distribution by increasing the number of samples in the minority class.

A visualization to show before and after using SMOTE



* Standard Scaling

Is a common feature scaling technique in data preprocessing used to transform data so that it has a mean of 0 and a standard deviation of 1. It ensures that all features are on a similar scale, making each feature contribute equally to the model. This reduces bias and speeds up training and improves performance.

1. **MODELING**

For our modeling approach, we will start with three algorithms: Logistic Regression as our baseline model, followed by Decision Trees and Random Forest. Based on the performance of these models, we will then select the most suitable one for hyper parameter tuning.

For our **Logistic Regression model**, it has the following results;

Accuracy: 0.8836, Precision: 0.8912, Recall: 0.8836, F1 Score: 0.8854

Accuracy: Our model correctly predicts whether a child has anemia or not 88% of the time

Precision: When our model predicts a positive case, it is correct 89.12% of the time

Recall: The model captures 88.36% of all true positive cases (the actual people with anemia)

F1 Score: The model has a good balance between precision and recall at 88.54%

For our **Decision Tree model**, it has the following results;

Accuracy: 0.9338, Precision: 0.9353, Recall: 0.9338, F1 Score: 0.9342

Accuracy: Our model correctly predicts whether a child has anemia or not 93.38% of the time

Precision: When our model predicts a positive case, it is correct 93.53% of the time

Recall: The model captures 93.38% of all the true positive cases

F1 Score: The model has a good balance between precision and recall at 93.42%

For our **Random Forest model**, it has the following results;

Accuracy: 0.9547, Precision: 0.9558, Recall: 0.9547, F1 Score: 0.9549

Accuracy: Our model correctly predicts whether a child has anemia or not 95.47 % of the time

Precision: When our model predicts a positive case, it is correct 95.58 % of the time

Recall: The model captures 95.47 % of all the true positive cases

F1 Score: The model has a good balance between precision and recall at 95.49%

For our **K Nearest Neighbors model**, it has the following results;

Accuracy: 0.6533, Precision: 0.6970, Recall: 0.6533, F1 Score: 0.6654

Accuracy: Our model correctly predicts whether a child has anemia or not 65.33 % of the time

Precision: When our model predicts a positive case, it is correct 69.70 % of the time

Recall: The model captures 65.33 % of all the true positive cases

F1 Score: The model has a good balance between precision and recall at 66.54%

The **Random Forest** model outperformed all other models across all metrics, making it the top choice for this classification task. We will proceed with further model tuning to explore if we can improve the accuracy even more.

* Model Tuning (Hyper-parameter Tuning)

We then went ahead to tune our best performing model, the Random Forest, using GridSearchCV. The reason why we used this is because it systematically performs exhaustive search over all possible combinations of hyperpraramters within a specified grid ensuring the best combination is found within the parameter space.

The results of our Tuned Random Forest model

Accuracy: 0.9562, Precision: 0.9574, Recall: 0.9562, F1 Score: 0.9564

Accuracy: Our model correctly predicts whether a child has anemia or not 95.62 % of the time

Precision: When our model predicts a positive case, it is correct 95.74% of the time

Recall: The model captures 95.62 % of all the true positive cases

F1 Score: The model has a good balance between precision and recall at 95.64%

1. **CONCLUSION**

This project developed a classification model to predict anemia severity in children under 56 months, focusing on the impact of socioeconomic factors. The model, validated for accuracy, enables early identification of at-risk children, assisting healthcare providers in efficiently prioritizing resources and tailoring interventions.

Key findings show that socioeconomic factors like wealth and parental education significantly influence anemia severity. These insights can guide targeted public health strategies and policies, addressing underlying causes to reduce anemia sustainably among vulnerable groups. Overall, the model provides a valuable, data-driven tool for improving anemia interventions and informing long-term health policy.

1. **RECOMMENDATIONS**

* Routine Screening: Integrate anemia screening in regular health check-ups, especially in high-prevalence and remote areas, for early detection and intervention.
* Address Inequalities: Provide targeted nutrition support, like iron supplements and fortified foods, to low-income families to improve child health and reduce anemia risk.
* Expand Rural Healthcare: Use mobile units and partnerships with NGOs to improve anemia diagnosis and care access in underserved regions.
* Data-Driven Policies: Use insights from predictive analytics to guide policies that address socioeconomic causes of anemia for more targeted interventions.
* Education and Awareness: Increase community awareness on anemia prevention, emphasizing nutrition and regular check-ups for sustainable health improvements.