ADDRESSING CHILDHOOD ANEMIA IN NIGERIA THROUGH PREDICTIVE MODELING

Business Understanding.

Problem Statement:

Anemia poses a serious public health challenge in Nigeria, significantly affecting the physical health, cognitive development, and overall quality of life of millions of children. This project aims to create a predictive model that classifies the severity of anemia in children (mild, moderate, or severe) by analyzing key socioeconomic indicators, such as wealth index, parental education, and residence type. By identifying the major predictors of anemia, the model will facilitate data-driven, targeted health interventions.

Objectives

- **Develop a Predictive Model**: To develop a machine learning model that classifies the severity of anemia in children using socioeconomic indicators, providing a non-invasive, accessible tool to aid in early detection and intervention efforts.
- **Analyze Key Indicators**: Use the findings from the anemia severity predictive model to identify socioeconomic factors most strongly associated with anemia (e.g., family income, parental education, access to nutrition, and healthcare availability).
- **Inform Targeted Interventions**: Provide actionable insights for effective resource allocation and health interventions. Leverage predictive insights to guide precise allocation of healthcare resources, prioritizing high-risk communities and designing tailored health interventions to effectively reduce anemia prevalence in children.
- **Enhance Policy Development**: Support the creation of evidence-based public health policies to reduce childhood anemia.

Why This Topic?

Childhood anemia remains a pressing health challenge in low- and middle-income countries, particularly in Nigeria. Its effects extend beyond individual health, impacting education, productivity, and economic growth by affecting children's cognitive and physical development. This project's goal is to contribute meaningful insights toward improving children's health, fostering human development, and breaking cycles of poverty associated with poor health outcomes. Addressing childhood anemia can create lasting benefits for individuals, families, and society, making this topic crucial for both immediate and long-term health progress.

Industry and Domain Context

This project is relevant to public health, maternal and child health, and healthcare policy, particularly in the realm of public health interventions in developing nations. The insights are valuable to government health ministries, healthcare organizations, NGOs, and international bodies like WHO and UNICEF that focus on reducing health inequities and improving outcomes for vulnerable populations.

Target Audience

This project's findings are intended for policymakers, healthcare providers, public health organizations, NGOs, and researchers dedicated to improving child health outcomes. The project's insights could also support governmental and non-governmental organizations in crafting evidence-based policies and interventions to reduce anemia rates.

Real-World Impact of Solution

A predictive model that accurately classifies anemia severity could drive effective resource allocation, enabling healthcare providers and policymakers to implement targeted nutritional, educational, and healthcare interventions. Such a model would help direct resources to the area's most in need, improving the efficacy of public health campaigns and reducing anemia prevalence and severity. By improving childhood health, this solution could support overall societal development, contributing to stronger future generations and economic progress in Nigeria.

Existing Research and Domain Knowledge

The project relies on data from the 2018 Nigeria Demographic and Health Survey (NDHS) and builds on studies linking childhood anemia to socioeconomic factors such as parental education, wealth, and healthcare access. Previous research has demonstrated that maternal education, malaria prevention, and iron supplementation programs can reduce anemia rates, providing a foundation for this analysis. Insights from WHO and UNICEF research further validate the importance of addressing anemia in children through data-driven strategies.

Motivation

This project is driven by an urgent need to reduce childhood anemia rates in Nigeria, where socioeconomic conditions exacerbate health disparities. Personal motivation stems from a commitment to leveraging data to address public health challenges and to drive sustainable health improvements for children in developing countries. This analysis aims to equip health organizations and policymakers with actionable insights to enhance childhood health and break cycles of poverty through improved health outcomes.

Data Understanding

Data Collection and Source.

This analysis uses data originally from the 2018 Nigeria Demographic and Health Survey (NDHS), obtained via Kaggle, which curated the dataset for ease of access. The NDHS dataset is known for its reliability, capturing a wide range of socioeconomic, demographic, and health-related variables crucial for analyzing anemia prevalence and severity in children.

Data Acquisition Plan.

With the dataset already acquired from Kaggle, no additional data collection is necessary. The data is comprehensive and ready for preprocessing and analysis, containing essential variables relevant to the study.

Feature Description and Relevance.

The features selected include

Socioeconomic Factors: Wealth index, parental education, and type of residence (urban/rural) to gauge the impact of economic background and location on anemia.

Demographic Factors*:* Child's age, sex, and regional location to account for variations in anemia risk.

Health-Related Factors: Maternal health indicators and healthcare access variables, which are linked to nutritional status and anemia in children.

These features are foundational for a predictive model to assess anemia severity

Prior Work on This Dataset.

Previous studies using NDHS data have primarily focused on exploring correlations between socioeconomic factors and health outcomes like anemia. This project, however, builds on existing research by developing a machine learning classification model to predict anemia severity (mild, moderate, severe), which could enable more targeted and data-driven public health interventions. This predictive approach aims to extend the impact of previous research by supporting actionable insights for health policy and resource allocation.

Data Preparation

Data Storage: The dataset is stored in a tabular (CSV) format with 33,924 rows and 17 columns.

Data Types of Variables: Categorical variables (13 columns) like "Type of place of residence," "Highest educational level," and "Anemia level" which are in the `object` type.

Numerical variables (4 columns), which include integers for "Births in last five years" and "Age of respondent at 1st birth," and floats for "Hemoglobin level adjusted for altitude (g/dl - 1 decimal)."

Some columns, such as "Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)" and "Anemia level," contain missing values.

Preprocessing Steps

Handle Missing Values- Address the missing values in columns like "Hemoglobin level adjusted for altitude" and "Anemia level."

Encoding Categorical Data -Encode categorical variables for model compatibility.

Feature Selection - Retain relevant features based on correlation and relevance to anemia severity.

Cleaning/Pre-processing Challenges

High Missing Values: Columns like "Hemoglobin level adjusted for altitude and smoking" have substantial missing values, which may affect model accuracy.

Duplicate Features: Columns like "Anemia level" and "Anemia level.1" may contain overlapping information.

Data Imbalance: Severity levels may not be equally represented, affecting model performance.

Minimum Number of Row

Given missing values, the dataset could potentially reduce from 33,924 rows to approximately 10,000-13,000 rows after cleaning

Visualization Plan

Distribution of Hemoglobin Levels: To show anemia severity distribution.

Anemia Levels by Key Demographics: Using bar plots for factors like age group, education, and wealth index.

Correlation Heat map: To identify relationships between hemoglobin levels, anemia severity, and demographic features.

Modelling

For our project, the most appropriate modeling techniques include Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting Machines (GBM) or XGBoost. The target variable is the severity of anemia in children, categorized into groups such as "None," "Mild," "Moderate," and "Severe," making it a multi-class classification problem. We plan to use Logistic Regression as our baseline model due to its simplicity and interpretability, allowing us to establish an initial performance benchmark for comparison with more complex models. Overall, this project will focus on classification rather than regression, as we aim to predict discrete categories of anemia severity based on the selected features.

Evaluation

Performance of our classification model will be evaluated using metrics such as accuracy, precision, recall, F1 score, confusion matrix, and ROC-AUC curve. The Minimum Viable Product (MVP) will involve creating a basic model with a cleaned dataset, implementing the baseline model, and generating simple visualizations to demonstrate performance. Over the next week, we aim to complete data cleaning, model implementation, and a summary report. Stretch goals for enhancing the project include exploring advanced models, feature engineering, implementing cross-validation, addressing class imbalance, and developing advanced visualizations, alongside comprehensive documentation of the process.

Deployment

The final results of our project will be reported through a comprehensive presentation that includes an executive summary outlining the project's objectives, methods, and key findings, along with visualizations of model performance and data insights.

We will also provide detailed documentation of the methodology, data sources, preprocessing steps, modeling techniques, evaluation metrics, and conclusions, and potentially create an interactive dashboard using tools like **Dash** or **Streamlit** for dynamic data exploration.

For deployment, we plan to develop a web application using frameworks like **Flask** or **Django** and host it on free platforms such as **Render** or **Vercel**. This will enable users to input socioeconomic data and receive predictions about anemia severity. These free hosting platforms offer reliable and scalable options for initial deployment, making the application accessible while minimizing costs.

The web app will include a user-friendly interface, displaying predictions with confidence scores, and provide visual insights and educational resources on anemia prevention.

Tools and methodology.

For our project predicting the severity of anemia in children using the 2018 Nigeria Demographic and Health Survey data, we plan to utilize essential Python libraries such as Pandas and NumPy for data gathering and cleaning, along with Matplotlib, Seaborn, and Plotly for data exploration and visualization. Our modeling will involve algorithms like Logistic Regression, Decision Trees, Random Forests, and XGBoost, with initial analyses performed on our local machines using Jupyter Notebook, and potential deployment in cloud environments like Google Colab for scalability. We will initially store the data locally but transition to cloud storage solutions for secure access and deployment in the web application, facilitating user interaction and reporting of final results.