FTL Global ML bootcamp 3 | Capstone project | Group 13

Malnutrition Risk Prediction ML Model

Concept Note | Implementation Plan

Group members:

- Lina Ahmed
- Linda Adil
- Maha Abdalfedil
- Nada Ali

Contents:

Project Adjustments

Throughout the development of this project, we encountered several unforeseen challenges, including the departure of key team members. These circumstances necessitated a strategic adjustment to our original plans and data sources. As a result, the scope of this report has been refined to focus on malnutrition prediction in (Bangladesh, Ethiopia, Ghana, Guatemala, Honduras, Kenya, Mali, Nepal, Nigeria, Senegal, and Uganda 2004-2016.) rather than the initially intended comprehensive coverage for the whole 33 LDC of africa. ***And instead of using a custom dataset, we have opted to utilize a dataset from Kaggle that aligns closely with our project's objectives.*** These modifications ensure that the project remains aligned with our objectives and leverages the available resources effectively.

Concept Note and Implementation Plan for Malnutrition Prediction Using Machine Learning and Time Series Data

Concept Note

1. Project Overview:

This capstone project focuses on *predicting malnutrition* risks among children in 11 low and lower-middle income developing countries using machine learning techniques.

The project aligns with Sustainable Development *Goal 2 (Zero Hunger)* by aiming to reduce child malnutrition, which is a significant contributor to child mortality and morbidity in these regions.

By developing a predictive model, the project seeks to identify at-risk children early, enabling timely interventions and ultimately improving child health outcomes.

2. Objectives:

The primary objective of the project is to develop a predictive model that can accurately assess the risk of malnutrition among children under five based on health and nutrition data. This project aims to:

- *Improve* early detection of malnutrition risks.
- *Equip* healthcare providers and policymakers with a tool to target interventions more effectively.
- *Contribute* to global efforts in reducing child malnutrition and advancing progress towards Sustainable Development Goal 2 (Zero Hunger).

3. Background:

Malnutrition is a persistent issue in many developing countries, exacerbated by poverty, lack of access to nutritious food, and inadequate healthcare. Current efforts to combat malnutrition often rely on reactive measures rather than proactive prevention. ***Existing solutions include nutritional programs and health monitoring systems, but these often lack the predictive capabilities needed for early intervention.***

A machine learning approach offers the ability to analyze complex patterns in socio-economic and health data, leading to more accurate and timely identification of at-risk populations.

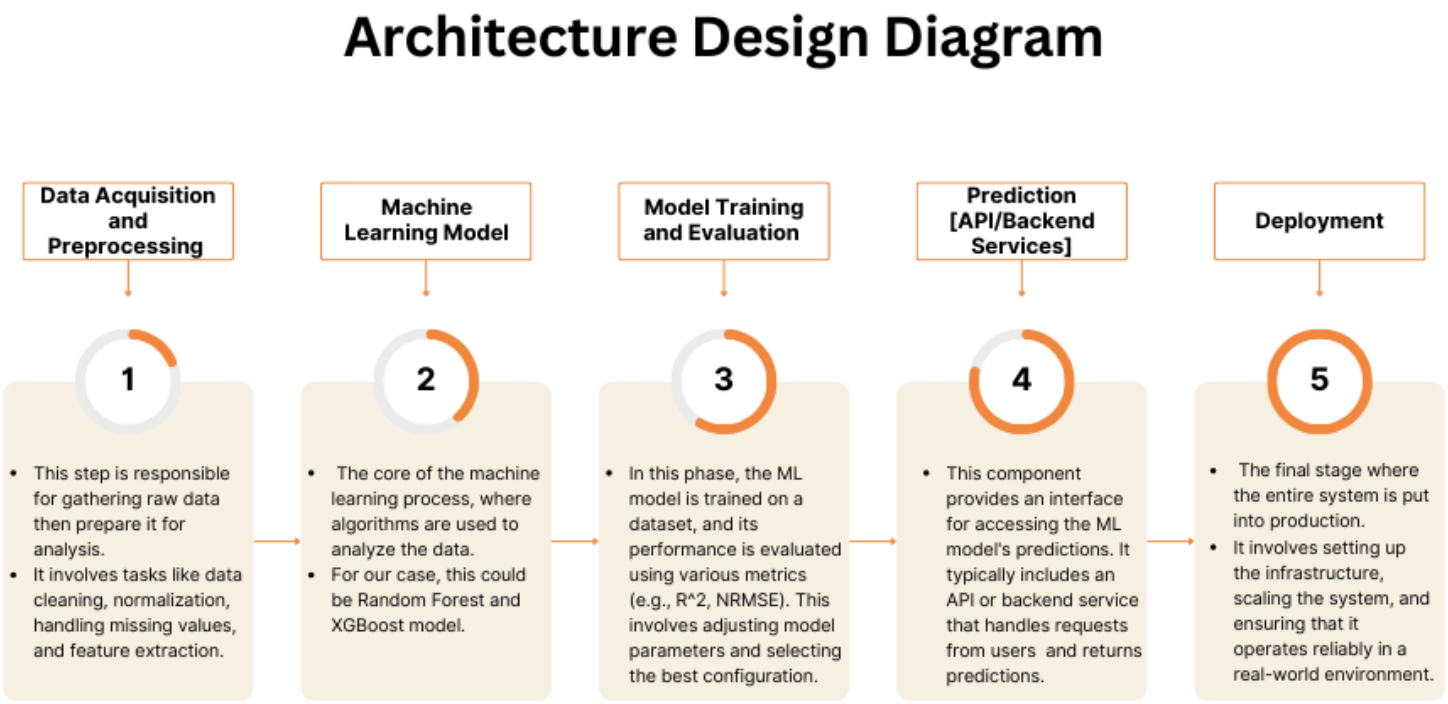
4. Methodology:

The project will utilize machine learning techniques to develop a predictive model for malnutrition. The methodology includes:

- *Data preprocessing* to clean and normalize the dataset.
- *Feature selection* to identify the most relevant predictors of malnutrition.
- *Model development* using algorithms such as Random Forest, XGBoost
- *Model training and validation* to ensure accuracy and generalizability.
- *Hyperparameter tuning* to optimize model performance.
- *Evaluation* using metrics such as accuracy, precision, recall, and F1 score.

5. Architecture Design Diagram:

Here's an overview of the architecture of our project:



Each component role:

- **Data Acquisition and Preprocessing**

  Ensures that the data fed into the ML model is accurate, complete, and in a suitable format for analysis.

- **ML Model**

  Performs the actual prediction or classification based on the processed data. It uses patterns learned from training data to make predictions on new data.

- **Model Training and Evaluation**

  Ensures that the ML model performs well and generalizes effectively to new data. This step involves tuning and validation to achieve optimal performance.

- **Prediction [API/Backend Services]**

  Facilitates the integration of the ML model into applications or services, allowing for real-time or batch predictions.

- **Deployment**

  Makes the system available for end-users (Governments, NGOs) and ensures its functionality in a live setting. It includes monitoring, maintenance, and updates as needed.

6. Data Sources:

The project will leverage data from the "***Multivariate random forest prediction of poverty and malnutrition prevalence***" dataset available on Kaggle ***[Link]***. This dataset, derived from the research paper by Chris Browne, David S. Matteson, Linden McBride, and colleagues, offers a comprehensive repository of social indicators and environmental factors.

The dataset includes crucial metrics such as child stunting, wasting, healthy weight, asset poverty, and underweight BMI. Additionally**,** it provides geographic, environmental, and socioeconomic features critical for understanding the complex factors influencing poverty and malnutrition. Data preprocessing, including cleaning, handling missing values, and feature engineering, will be performed to ensure the dataset is properly prepared for model training and analysis.

7. Literature Review:

Existing literature on predicting malnutrition using machine learning highlights the effectiveness of Random Forest models in capturing complex relationships between health and socioeconomic features. Research has consistently shown that such models can reliably predict indicators like child stunting and wasting using diverse datasets. By focusing on feature relationships and refining the existing model.

Building upon this foundation, our project aims to enhance prediction accuracy. Additionally, developing a user interface will make the model ***accessible and actionable*** for governments and NGOs, facilitating real-time predictions and supporting targeted interventions ***without extending*** the current methodologies or incorporating new variables. This approach builds on established methods to improve usability and practical application in public health initiatives.

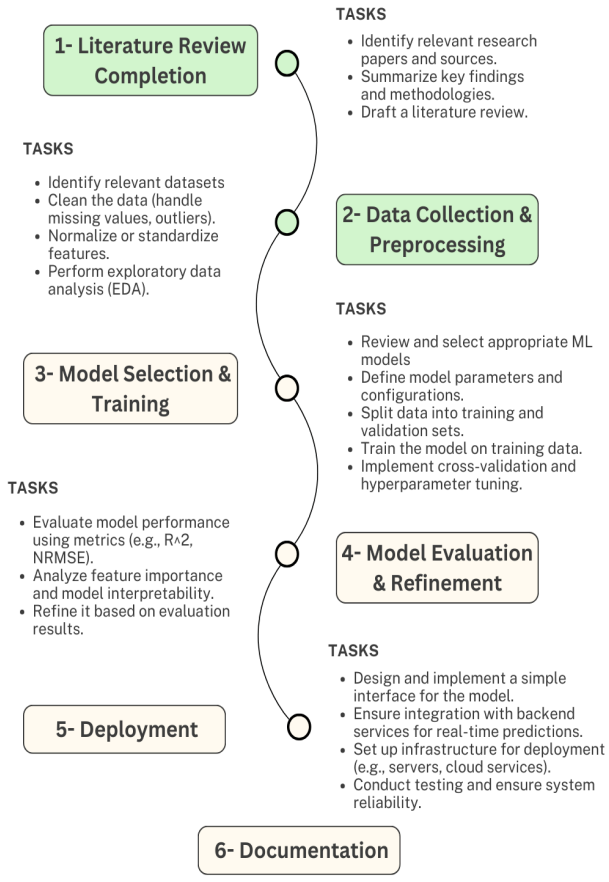
Implementation Plan

1. Technology Stack:

- **Programming Languages**:
  Python.
- **Libraries**:
  Pandas, NumPy, Scikit-learn, XGBoost, Statsmodels.
- **Frameworks**:
  Google collaboration for development, Flask for deployment.
- **Other Tools**:
  Git for version control, Docker for containerization.

2. Timeline:

1. Literature review
   (26.Aug)
2. Data Collection & Preprocessing
   (31.Aug)
3. Model Selection & Training
   (5.sep)
4. Model Evaluation & Refinement
   (10. sep)
5. Deployment ***(15.sep)***
6. Project Documentation ***(20.sep)***



Team members task distribution matrix:

	Linda	Nada	Maha	Lina
1	L	S	S	S
2	S	L	S	S
3	S	S	L	S
4	-	-	-	-
5	-	-	-	-
6	-	-	-	-

L = Lead

S = Support

Empty cells = Not started yet

3. Milestones:

The following milestones highlight the key achievements and critical phases of our project, ensuring structured progress and alignment with our overall goals. These milestones will guide us in tracking our advancements and maintaining focus on the successful completion of the project.

Milestone 1: Finalized literature review document.

Milestone 2: Clean and ready-to-use dataset.

Milestone 3: Trained machine learning model with baseline performance metrics.

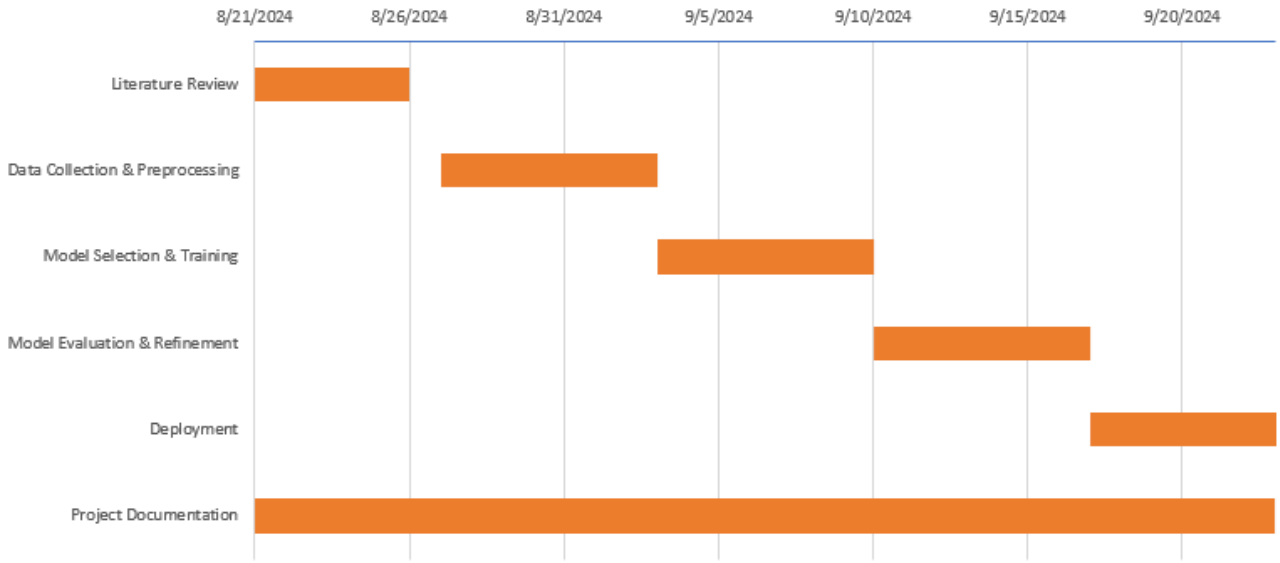
Milestone 4: Achieved target performance metrics (e.g., $R^2 > 0.8$, NRMSE below threshold).

Milestone 5: Prototype or final version of the user interface.

Milestone 6: Model and interface deployed in a live environment.

Milestone 7: Finalized project documentation and user guides.

Gantt Chart:



4. Challenges and Mitigations

1. Data Quality:

Challenge:

Ensuring data accuracy, completeness, and consistency is crucial for the project's success.

Mitigation:

We will implement robust data preprocessing techniques, including imputation methods for missing data, normalization, and outlier detection.

2. Model Performance:

Challenge:

Ensuring the model's effectiveness in handling complex data and achieving high performance across various metrics.

Mitigation:

Employ multiple machine learning models and fine-tune their parameters using hyperparameter tuning. Validate the model's performance using cross-validation.

3. ***Technical Constraints:***

Challenge:

Limitations in computational resources and deployment issues can hinder project progress.

Mitigation:

We will utilize cloud-based platforms like ***Google Colab*** to efficiently manage computational resources and ensure scalability. ***Docker*** will be used for consistent deployment across different environments.

5. Ethical Considerations:

1. ***Data Privacy:***

Anonymization

All personal identifiers will be removed from the dataset to ensure the privacy of individuals.

2. ***Bias:***

Dataset Balancing

Efforts will be made to balance the dataset, especially for underrepresented groups, to reduce bias.

Feature Engineering

Techniques will be employed to minimize the influence of biased features on the model's predictions.

3. ***Impact on Target Community:***

Stakeholder Engagement

Collaboration with local governments, NGOs, and community members will be sought to understand their needs and concerns.

Responsible Use

The model's predictions will be used responsibly to inform policy decisions and interventions that benefit the target community.

Monitoring and Evaluation

The project's impact will be continuously monitored and evaluated to identify any unintended consequences or negative outcomes.

6. References:

[1] Browne, C., Matteson, D. S., McBride, L., et al. (2017). Multivariate random forest prediction of poverty and malnutrition prevalence. *Journal of Applied Econometrics*, 32(4), 818-841.***[link]***

[2] World Health Organization. (2020). *Child Growth Standards and Malnutrition Indicators*. ***[link]***

[3] United Nations. (2020). Sustainable Development Goals: Goal 2 - Zero Hunger. ***[link]***