NUTRITIONAL DEFICIENCY PREDICTION BY REGION USING MACHINE LEARNING

ANNE DASHINI KANNAN

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the methodology used to predict nutrition deficiencies across regions using machine learning techniques. The research workflow, from the initial topic exploration to model evaluation, is described in detail. The dataset and performance evaluation criteria employed are also elaborated.

3.2 Research Framework

The structure of research is going to strive to ensure all the challenges faced when predicting nutrition issues across different areas. It is divided into five key phases: The main steps of the proposed framework include research planning and initial study, data preparation, feature extraction, model development, and model evaluation. Every stage in this process supports the development of an overall understanding of nutrition profiles and main drivers in a region. During formal system analysis phase, the research problem was comprehensively defined in addition to setting achievable objectives. This was then followed by a data description where real data sets were cleaned and preprocessed. During the feature extraction phase, the extracted information in the datasets was analyzed by employing the Machine Learning models to establish relevant patterns. The fourth phase aimed at using clustering methods to cluster regions according to their distinct nutritional aspect. Lastly, in phase five, the authors assessed the validity and accuracy of the developed clustering model through method including ML models.

Figure 3.1 explains the research framework.

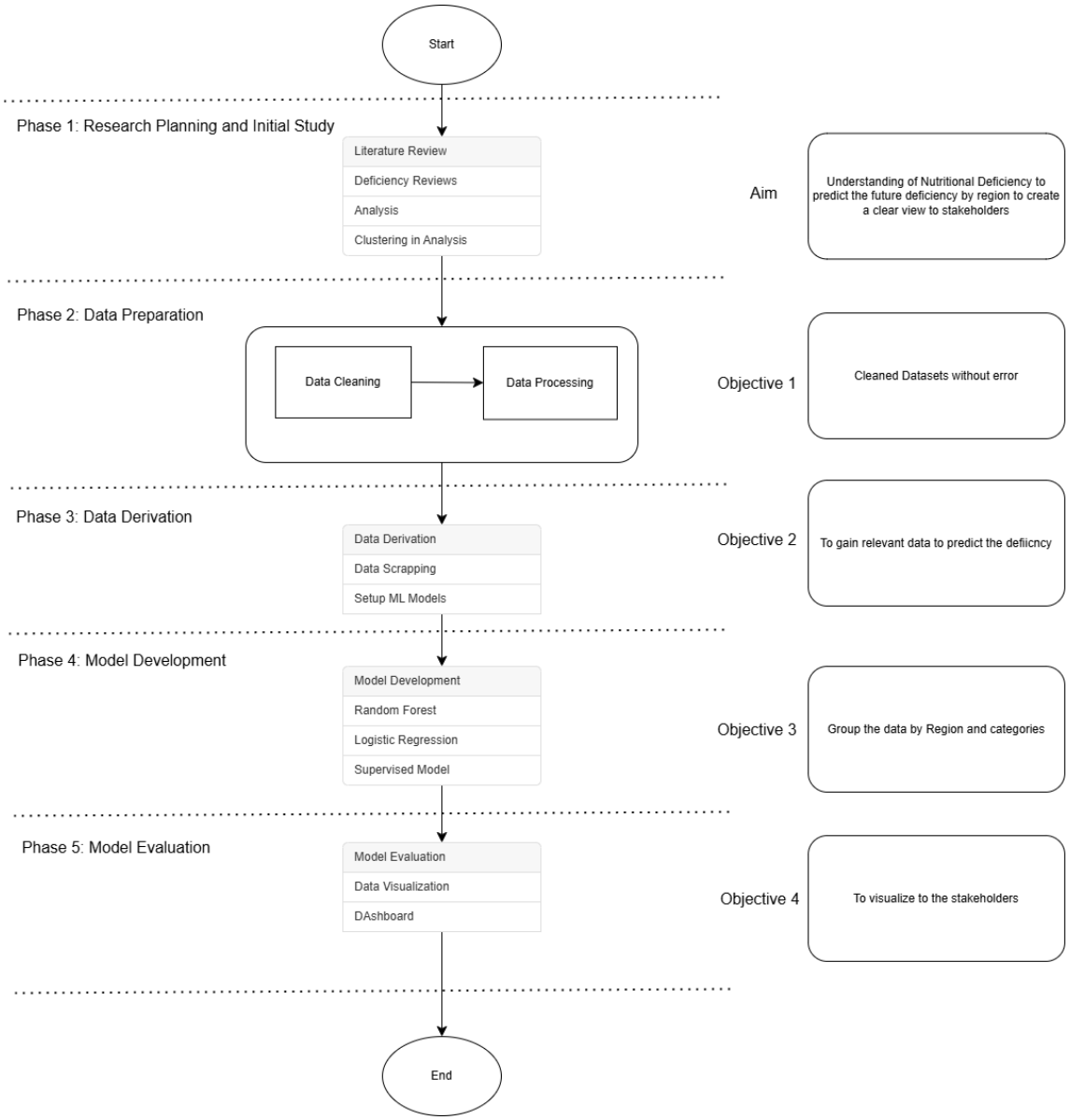

Figure 3.1: Overall Research Framework

3.2.1 Phase 1: Research Planning and Initial Study

The first step of the study was primarily exploratory concerned with identifying the issue of malnutrition and defining research questions. This included a synthesis of the literature on past efforts on malnutrition and lack of nutrients. Literature reviews by Black et al. (2013) and other studies by UNICEF and WHO contributed their invaluable findings regarding the frequency, underlying causes, and consequences of malnutrition. These studies revealed that regional differences regarding diets, standards of living and accessibility to health facilities influenced malnutrition. The literature on methods of assessment also revealed a number of limitations of the conventional approaches like national health surveys and RCT trials where the scientific sample Investigation also exposed a number of weaknesses and shortcomings of the prevalent methods of assessment; for example, national health surveys and RCT trials do not capture the full array of nutrition problems at micro level.

To address these challenges, machine learning was adopted as a cost-effective method for analysing large and large datasets and locating regional characteristics. This phase also defined the purpose of the study, to identify the characteristics associated with poor nutrition and to use machine learning to discover relevant patterns to support public health interventions.

3.2.2 Phase 2: Data Preparation

Before analysis, data has to go through data preparation process to make sure that the collected dataset is accurate and can be used for analysis. The study used different data sources ranging from national health surveys, global nutritional databases, demographic records of regions among others. These datasets offered finer details on aspects including dietary preferences, health status, environmental aspects, and population characteristics.

It must be mentioned that before the analysis the data was cleaned to handle missing values and duplicates as well as inconsistencies. Data that were skipped out were either removed or removed whenever the data was not relevant in assessing the result. The dataset was cleaned up to ensure that the availability of observations was across a different set of variables. Duplicates were deleted to ensure data validity. Secondly, applying text preprocessing techniques, any self-reported data, for example, survey responses, were used. Text was normalized by first converting it to lowercase and secondly all irrelevant terms have been stripped off the text data. Where needed, supplementary relevant data sources included measures of agricultural yields and climate characterizations were added to the dataset to increase the efficacy of the machine learning algorithms.

3.2.3 Phase 3: Feature Extraction

In the process of data analysis, an important step entails selection of relevant, distinguishing features from the dataset. Using regression analyses, the study found that the likelihood of nutritional deficiencies was positively predicted by the dietary diversity scores, perceived prevalence of infectious diseases, mother's education and accessibility of fortified foods. The selection of these variables follows previous studies which confirmed their role in influencing nutritional health.

Textual data mining and, specifically, Large Language Models such as OpenAI's GPT were used to investigate prominent patterns from the collected textual data. For example, the models found words and phrases linked to diet and nutrition issues, health issues, and geographical aspects of life. Feature engineering was also used to convert raw data into derived features including normalized dietary diversity score and Health Risk Indices. These enriched features upset the clustering algorithm to differentiate the regions efficiently cay the nutritional categories they represent.

3.2.4 Phase 4: Model Development

To propose a model for predicting and analysing the states of nutritional deficiencies together, the model consisted of supervised and unsupervised machine learning. Logistic regression was also applied in supervised models to predict region specific probability of the presence of nutritional deficiencies. High accuracy of results and interpretability make these models appropriate and efficient for sets of ordered data. These models were trained under labelled data where scarcity was the elements or variables of concern.

When designing the clusters to partition the regions according to nutritional similarity, two algorithms were used: K-Means and Hierarchal Clustering. The following models partitioned geographical area into different categories according to a common nutritional characteristic, including dietary scarcity, prevalence of deficiency anaemia, nutritional status, and socioeconomic status. K-Means was preferred for its ability to deal with big data while Hierarchical Clustering offered a layering structure of the regions.

There were several stages during the model development. First, hyperparameters of the predictive models were selected by using grid search for their optimal parameters. For instance, features such as the number of trees and the model's maximum depth were optimized for Random Forest, as was the learning rate and the number boosting rounds. The same applies to K-Means where the Number of clusters was decided and to Hierarchical Clustering where clusters were searched.

The numerical results of the feature selection were used to determine crucial predictors associated with nutritional deficiencies and guide policymaking. For instance, dietary diversity scores, the household's access to fortified foods, and maternal education were depicted as significant predictors in the studies. The clustering results were depicted by geographic heat maps and 3D scatter plots for easy interpretation of the regionality.

3.2.5 Phase 5: Model Evaluation

The conclusion of the assessment of the machine learning models incorporated supervised and unsupervised measures for the validity and accuracy of the models in anticipating nutrient depletion and determining geographical zones. To assess the predictive performance of the supervised models, accuracy, precision, recall and F1-score were used. They offered an understanding of how accurately our models were able to assign regions depending on their risk of possible deficiencies.

For example, countries with high prevalence of vitamin A deficiency and low dietary diversity matrix were clustered together, indicating the potential areas suitable for food fortification programs. Data obtained from the clustering were further illustrated using heatmap and geographic maps, which can be easily understood by policy makers and other stakeholders. To strengthen the generalizability of the developed predictive models, the process of cross-validation was used. The dataset was divided into training and testing sets and performance was tested on new data not used for training. For clustering, validation with the regional health reports was conducted to fit the identified clusters with WHO's Global Nutrition Reports for effectiveness of the malnutrition patterns known.

Refinement of the questions took place continuously during the evaluation stage. Main characteristics were included or excluded according to the importance values, and parameters were tuned for better results. This iterative was definitely a strength since it helped in the creation of final models that would be more accurate and most importantly usable for giving out recommendations concerning the nutritional needs of the various regions.

3.3 Dataset

The dataset contained more than 185 000 rows and six main predictors: geographic variables, demographic data, nutritional status, health status, and environmental information. The overall data set of these subsequent products offered a strong baseline for assessing malnutrition in relationships to regions. This approach allowed the study to get an appreciation of the various factors that precipitate malnutrition.

3.4 Performance Measurement

Evaluation of accuracy, reliability and potential usability of the machine learning models created for this study required performance measurement. In the case of the supervised predictive models, the evaluation metrics concerned the classification accuracy, while in the case of the clustering models the quality of the formed clusters was of concern.

To deem, the supervised models of classification such as Random Forest were evaluated using basic classification performance measures. Accuracy defined a ratio of the number of instances which were correctly predicted out of the total number of cases. However, since nutritional deficiency data may have class imbalance problem, using precision, recall and the F1 score offered more valuable information. Among all the positive predictions made by the model, precision Stayed high, indicating that the model did not make many false positives. Recall focused on how many of all actual positives were true positives, showing that the model captured all areas where a risk of nutritional deficiencies existed. Since the F-measure is defined as the harmonic mean between the precision and the recall, their weaknesses were compensated which provided an objective assessment of the system's performance.

3.5 Summary

This chapter provided a detailed analysis of how the nutritional deficiencies at the regional level can be estimated by using machine learning. To promote the reliability and applicability of the obtained results, the framework of the research included data pre-processing, feature extraction, clustering model construction, and assessment. The use of multiple data sources and appropriate analytical methods allowed the researchers to develop practical recommendations concerning deficiency and possible actions in the field of nutrition.