

**Emerald International College**

**Big Data Engineering Project**

**Project title: Predicting the nutritional density of different types of foods based on the nutritional content of food.**

**Group 8 Members:**

1. **Hagos W/senbet**
2. **Melkamu Adigo**
3. **Wassye Yitages**
4. **Wondwossen Alemu**
5. **Zelalem Aysheshim**

**August 22024**

**Addis Ababa, Ethiopia**

Table of Contents

[1. Project Overview 3](#_Toc175555431)

[2. Project Scope 5](#_Toc175555432)

[3. Project Objectives 5](#_Toc175555433)

[9. Expected outcomes and Deliverables of the project 6](#_Toc175555434)

[10. Data Sources 6](#_Toc175555435)

[11. Data Ingestion 6](#_Toc175555436)

[12. Data Storage 6](#_Toc175555437)

[13. Data Pre preprocessing and processing method 6](#_Toc175555438)

[14. Data Analysis and Visualization 6](#_Toc175555439)

[15. Challenges and Solutions 7](#_Toc175555440)

[16. List of our team member 8](#_Toc175555441)

# Project Overview

This project is targeted on a comprehensive Nutritional Food Database that provides detailed nutritional information for a wide range of food items commonly consumed around the world. This dataset aims to support dietary planning, nutritional analysis, and educational purposes by providing extensive data on the macro and micronutrient content of foods. The dataset contains 551 different types of foods with its calorie value, fat, saturated fat, monounsaturated fat, polyunsaturated fat, carbohydrate, protein, sugar, dietary fiber, cholesterol, sodium, water, vitamins, calcium, copper, iron, magnesium, manganese, phosphorus, potassium, selenium and zinc content. The dataset also has a record of nutritional density of each food based on the food content. This dataset is invaluable for researchers in nutritional science, dietitians planning meals, healthcare providers advising on dietary options, and individuals tracking their food intake based on the nutritional density of each food.

Nutritional density refers to the amount of beneficial nutrients in a food relative to its calorie content. Nutrient-dense foods provide essential vitamins, minerals, amino acids (the building blocks of protein), and healthy fats. It is also known that nutrient-dense diet is an anti-inflammatory diet. Chronic inflammation is linked to various health issues—heart disease, high blood pressure, diabetes, and obesity. So, it is recommended that choosing nutrient-dense foods to help the body to fight inflammation is the most crucial step. Therefore, it’s not just about how much you eat; it’s about what you eat. The overall quality of your diet matters. Nutrient intake divided by daily calorie intake—this ratio plays a role in the overall health. So, it’s not just about counting calories; it’s about counting nutrients too.

So, this project aims to predict the most nutritious food based on the given nutrient used as a raw material to manufacture a certain food type. It is also helpful for recommending certain food type for specific populations based on their health conditions by using matching learning algorithms. Machine learning algorithms can analyze trends and patterns in different food consumption, linking them to nutritional impacts. For example, clustering techniques can identify common characteristics of high-calorie foods or those high in certain nutrients like vitamins or minerals. By integrating this dataset with broader dietary data, machine learning models can recommend dietary adjustments to individuals. For instance, a recommendation system could suggest lower-calorie or lower-sugar foods alternatives to users looking to reduce their sugar intake but who still want to enjoy sweets.

With sufficient data linking certain food consumption to health outcomes, predictive models could forecast health impacts based on food consumption patterns. This could be particularly useful for medical research or public health studies examining the effects of certain nutrients and additives on long-term health. Machine learning can help in formulating new food recipes by predicting the nutritional content based on ingredients. This could aid manufacturers in creating healthier food options that maintain flavor while reducing undesirable nutrients like sugars or synthetic additives.

Machine learning models can use classification or regression models and helps businesses to predict consumer preferences for certain types of food based on nutritional information. This insight can drive targeted marketing and product development strategies. It can also be trained to predict the quality and consistency of food based on variations in manufacturing parameters and ingredient quality helps to maintain high standards. By applying natural language processing (NLP) techniques to product reviews and feedback, companies can extract consumer sentiments related to specific nutritional aspects of certain foods. This could guide improvements in product formulations and marketing strategies. Analyzing food sales and nutritional preferences with machine learning can optimize supply chains by predicting demand fluctuations based on health trends and nutritional awareness.

In general, developing interactive machine learning applications that teach users about nutrition helps to visualize the impact of different foods on dietary intake and encourage healthier eating habits. Machine learning models can integrate certain food consumption data into broader dietary tracking tools used in fitness and health apps, providing users with insights into how occasional treats certain foods fit into their overall dietary goals. This project demonstrates how a dataset can have broad applications across health, industry, consumer insights, and education, leveraging the power of machine learning to extract and predict valuable insights.

# Project Scope

This project will cover on provide insights into which factors contribute most significantly to nutrient density and how they can be optimized by collection of data on various foods, including their nutrient content, caloric value not including soil health, crop management practices, and spectral data. It employs statistical and machine learning techniques to build the predictive model. During the implementation there will be direct collaboration with researchers, supply chain companies, farmers and growers. By considering the link between nutrient density and human health it can move beyond simple comparisons and focus on optimizing nutrient density within each food category for specific health conditions.

# Project Objectives

1. To gain a comprehensive understanding of the macro and micronutrient distribution across various food items in the dataset.
2. To provide insights that can assist in dietary planning to promote healthy eating habits based on the nutritional content of foods.
3. To develop predictive models to forecast nutritional values based on selected factors such as food type or preparation methods.
4. To represent the findings visually through effective data visualizations, making the insights user-friendly and actionable for diverse audiences.
5. To understand the most significant determinant of nutritional density among the food contents.
6. To develop predictive model to recommend a certain food type for specific populations based on the health conditions.
7. To assess and understand how nutrient density varies across different food items.
8. To identify which factors, contribute most significantly to nutrient density and how they can be optimized.

# Expected outcomes and Deliverables of the project

* Foods can be ranked based on their nutrient density and help consumers to identify nutrient-rich options.
* The model can also balance nutrients to encourage health-promoting components while limiting fewer desirable ones.
* There could be publicly available models with transparent and validated nutrient profile that inform dietary recommendations and guidelines.
* Food companies can use these models to reformulate products for better nutritional profiles.
* Government agencies can use nutrient profiles to regulate marketing (e.g., restricting unhealthy food advertising to children).

# Data Sources

The dataset is accessed from Kaggle, and it is structured data. It can be accessed through Kaggle web site.

# Data Ingestion

During the project different statistical tool were used to analyses and visualize the data such as Power BI, PySpark and Jupiter notebook

# Data Storage

* Our personal PC

# Data Pre preprocessing and processing method

* 1. Processing procedure: - EDA and feature selection using PySpark in Jupiter notebook
  2. Summery statistics and predicative analysis using machine learning regression model using PySpark in Jupiter notebook

# Data Analysis and Visualization

* 1. Analysis Tools: Jupiter Notebooks, Power BI).
  2. The result described using graph and summery statistics for selected features we are using to make predicative analysis

# Challenges and Solutions

1. **Data Quality and Cleaning**
   1. **Challenge**: The dataset may contain missing values, duplicates, or inconsistencies across entries.
   2. **Solution**: Implement robust data cleaning procedures in PySpark, such as handling missing values with imputation strategies and removing duplicates.
2. **Feature Selection:**
   1. **Challenge**: Identifying the most relevant features for building predictive models can be complex.
   2. **Solution**: We Conduct exploratory data analysis (EDA) and use techniques like feature importance from tree-based models or correlation analysis to determine significant features.
3. **Model Selection and Evaluation:**
   1. **Challenge**: Choosing the best predictive model and avoiding over fitting.
   2. **Solution**: We Experiment with a variety of models, employ cross-validation techniques, and analyze metrics like RMSE or accuracy based on the task type (regression or classification).
4. **Integration with Power BI:**
   1. **Challenge**: Combining PySpark outputs with Power BI for visualization can be challenging if data formats do not match.
   2. **Solution**: We Export processed data from PySpark to CSV format that is compatible with Power BI, ensuring data types and schemas align.
5. **User Engagement and Interpretation:**
   1. **Challenge**: Ensuring users can interpret the visualizations and predictive outputs accurately.
   2. **Solution**: Provide user guides and interpretive notes alongside visualizations, enhancing clarity and understanding.

By addressing these objectives and challenges, the analysis can yield valuable insights into nutritional data, aiding users in making informed dietary choices and promoting health.

# List of our team member

* All our group members are working together both virtual and in person for both project objective setting and selection of our project tittle. In addition we write our project document together in person.

1. Melkamu Adigo: EDA and Power BI visualization
2. Hagos W/senbet: EDA and Power BI visualization
3. Wassye Yitages :EDA and Power BI visualization
4. Wondwossen Alemu: EDA and Predicative analysis using PySpark and creating our group project repository in GitHub.
5. Zelalem Aysheshim: EDA and Predicative analysis using PySpark