Project Proposal: Analyzing Global Food Nutrition and Environmental Impact

Group Members:
Weixiao Wang ("Wei"), Trisha Raju, Talin Bedonian, Xiaoyun Yu ("Freda")

Time Zones: Trisha & Talin in EST; Wei and Freda in China time.

Responsibilities:

• Weixiao (Wei): **Data validation** (text & values), Algorithmic and model suggestions (check with course materials), Spark (SparkSQL and MLlib), **algorithm polishing** for efficiency, Before classification: clustering, Dimension reduction (collaborate with Freda), **Neural network** using PyTorch or Keras, Tune hyperparameters (Grid search), Classification: SVM, Classification: Multinomial Naïve Bayes & Random Forest, Regularization (L1 and L2), **XPath** and DOM tree (if any), .ipynb theoretical explanations, Time series data (if any), last minute changing and refining, help with data wrangling

•Trisha: Dataset examination, **Data cleaning, Data wrangling** (collaborate with Talin), Python (pandas, numpy, sklearn, matplotlib, collaborate with Talin), SQL for EDA and refine for efficiency (collaborate with Talin), Power BI, One-hot encoding, Encoding for feeding to NN, **Neural networks (drafting**, collaborate with Wei), **Report refining**

•Talin: Dataset examination, **Data cleaning, Data wrangling** (collaborate with Trisha), Pandas EDA (collaborate with Trisha), SQL (collaborate with Trisha), **Statistical** testings, **math** theories, Logistic Regression etc, Parametric models (collaborate with Freda), Data Visualization, Evaluation metrics examination (collaborate with Freda), guiding group events (collaborate with Freda)

•Xiaoyun (Freda): ER Diagrams (Knowledge Representation), RDBM building (perhaps normal forms), Group coordination (timeline, gather materials), report drafting, **Statistical** theories **in ML**, .ipynb **theoretical explanations**, NLP, Dimension reduction (Subset selection / forward selection / backward selection / PCA / t-SNE before supervised ML), Parametric models (collaborate with Talin), Feed-forward NN, Hold-out validation set/K-fold CV, **Evaluation** metrics (collaborate with Talin), help with data wrangling

Data Source:

Open Food Facts dataset from Kaggle (link) provides global food nutrition and ingredient information, including features on nutrient content, environmental impact, and potential species threats.

Objectives:

To identify which countries produce more nutritious foods and understand ingredient factors influencing nutrition and environmental risks.

Project Plan:

Study Focus:

1. Nutritional Quality by Country: Explore which **countries** produce foods with higher nutritional scores. (Classification. Target y: country)
2. Environmental Threat Analysis: Use an XPath approach to extract a Boolean flag indicating whether a product poses a species threat.
3. Nutritional Trends & EDA: Investigate the correlation between high sugar content, additive levels, and nutrition scores across countries. (Inference)

Modeling:

PCA (Principal Component Analysis)/t-SNE/subset-selection (forward/backward selection/Lasso regularization): Dimensionality reduction for insights into feature contribution and to visualize clusters in nutrition and species threat. Before the supervised ML.

Nutrition Score Analysis:

- Linear Regression & Ridge Regression as regularization: Quantify the influence of nutrients and ingredients on the nutrition score. (Regression. Target y: nutrition score)
- Logistic Regression: Classify products as high or low in nutrition score and assess model accuracy with confusion matrices. (Prediction. Classification. Target y: products)

Species Threat Prediction:

- Decision Trees and Random Forest Regression: Predict whether a product poses a threat to species based on ingredients and nutrient levels.
-

Extended Analysis:

- Correlation Studies: Identify ingredients that correlate with high nutrition scores. (Inference)
- Feature and Error Analysis: Use L2 regularization, cross-validation, and error metrics to ensure accuracy without overfitting.
- Complex Model Evaluation: Explore advanced models to assess performance impact on accuracy.
- Confusion Matrices: Evaluate models' precision in threat prediction and nutrition classification.

Visualizations:

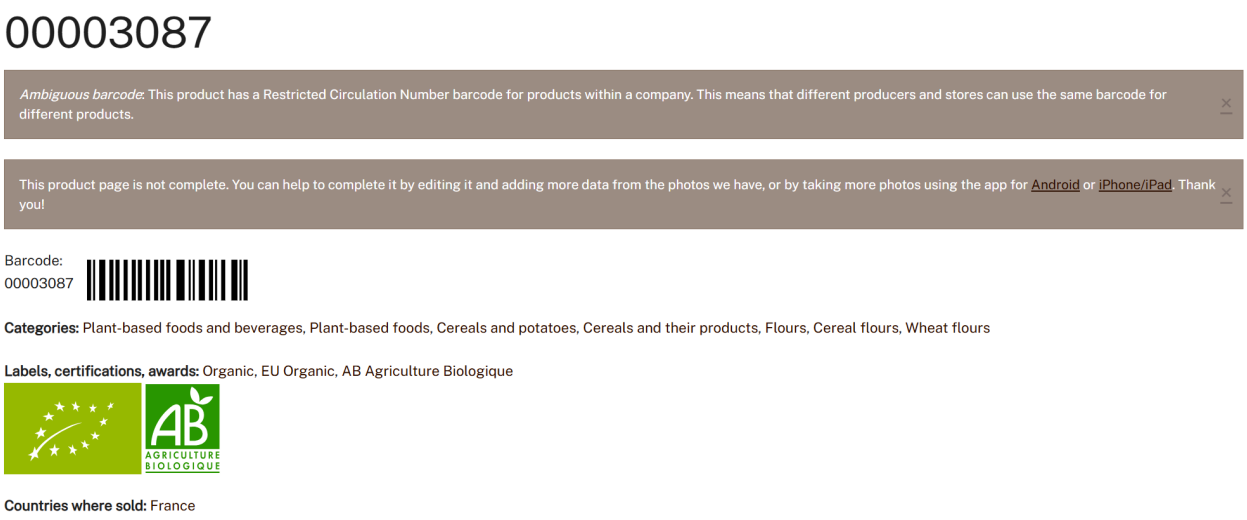
- Nutrition Scores by Country: Heat maps showing countries with products scoring higher nutritionally.
- Species Threat Maps: Visualizations of countries with more products marked as environmentally threatening.

- Sugar Content Analysis: Scatter plots to analyze countries with high sugar content products against their nutrition scores.

Why This Project Is Interesting:

This analysis could inform consumer choices and policies on food production and environmental impact by providing insights into nutritious food trends and ingredient risks. Will form a potential guidance for how people and our environment can exist in harmony. May lead potential fashionable environmental choices, and may even affect country-wise decisions.

By clicking the URLs, you can see very straight-forward colorful and realistic representations for the food barcodes, food facts. This representation can shorten the distance between Data Scientists' perspective and our stakeholders' standpoint, since a graph is better than 1000 words.



Dataset Benefits:

Original dataset has 356028 rows, with blank values, good for wrangling.

163 cols (rich set of predictors), with blank values, good for push-down and dimension reduction.

There are cols with comma-splitted texts, good for expanding and perhaps using JSON.

There are numerical-and-categorical combined values, good for data cleaning and validation.

Have datetime cols, good for exerting SQL and Pandas datetime handling, and stream processing / time series (if we want).

Numerical data have wide ranges, meaning that we can exert standardization.

Challenges:

- Missing data and feature engineering complexity.
- Balancing model complexity with interpretability.
- Neural networks, as most of our teammates do not have experience on using PyTorch.
- Spark MLlib is difficult for all of us.
- Addressing potential gaps in environmental impact data.
- Ethical considerations, as this project is built toward common customers.
- Arranging group events, as we are located in different locations and speaking different languages, and have distinct aspects for food nutrition (Chinese is different from American food preferences!).

TA preferences:

Lareina Liu (can speak Chinese and careful), or Khan Vy (very patient and careful), or Gokul (clearly formed guidance)