Group 3: Bilwa Khaparde, Kai Yang, Linh Tran, Rifa Safeer Shah, Sangita Poudel

Project Proposal

Dataset 1: Food Delivery

- **Link:** https://www.kaggle.com/datasets/willianoliveiragibin/food-delivery-time
- **Description of dataset**
    - Each delivery and delivery person has a unique ID for tracking performance. Personal attributes such as age and ratings impact efficiency. Geographical data, including restaurant and delivery coordinates, enable precise distance calculations. Order details (meals, snacks, drinks, buffets) and vehicle type (scooters, motorcycles, bicycles) influence speed and preparation time. Environmental factors like temperature, humidity, and weather_description capture weather-related delays, while traffic_level (low, medium, high) accounts for congestion. The target column represents delivery time in minutes.
- **What can be done**
    - **Predicting delivery time**
        - Algorithms to use: random forest regressor, Gradient boosting regressor (XGboost, LightGBM)
        - Target variable: delivery time (column name Target)
        - Numeric features: distance, temperature, humidity, precipitation delivery_person_age, delivery_person_ratings
        - Categorical features: traffic_leve,type_of_vehicle, type_of_order, weather_description
    - **Classifying delivery status**
        - Algorithms to use: logistic regression, decision tree classifier, random forest classifier, gradient boosting, SVM, neural network classifier
        - Target variable: delivery status (categorical)
        - Numerical features: distance, temperature, humidity, precipitation delivery_person_age, delivery_person_ratings
        - Categorical features: traffic_level, weather_description, type_of_vehicle, type_of_order

Dataset 2: Food Processing Level in Retailed Grocery Items

- **Link:** GroceryDB/README.md at main · Barabasi-Lab/GroceryDB
- **Description of dataset**
    - This dataset contains information about the grocery food items in various grocery stores. The dataset categorizes items into harmonized food categories and provides nutritional details like protein, fat, carbohydrates, sugars, fiber, calcium, iron, sodium, vitamins, and cholesterol. Additionally, the dataset contains pricing details, package weight, and processed food properties (f_FPro, f_FPro_P, f_FPro_class).
- **What can be done**
    - **Predicting level of food processing based on nutrition information and product information**
        - Algorithms to use: random forest regressor, KNN, SVM, polynomial regression
        - Target variable: Level of processing
        - Numeric features: price percal, package_weight, is_Nuts_Converted_100g, Protein, Total Fat, Carbohydrate, Sugars, Fiber, Calcium, Iron, Sodium, Vitamin C, Cholesterol, Fatty acids, Total Vitamin A
        - Category features: food category, has10_nuts
        - **Analysis and Business Insights:** Analyze feature importances (especially with tree-based models) to understand which nutritional or product factors drive commercial food processing levels.
    - **Market segmentation and pricing strategies**
        - Algorithms to use: K-Means Clustering,
        - Target variable: Group of similar products
        - Numeric features: Level of processing, price percal, package_weight, is_Nuts_Converted_100g, Protein, Total Fat, Carbohydrate, Sugars, Fiber, Calcium, Iron, Sodium, Vitamin C, Cholesterol, Fatty acids, Total Vitamin A
        - Category features: food category, has10_nuts
        - **Analysis and Business Insights:**
            - Pricing Strategies: Analyze average pricing (e.g., price percal) within each cluster to identify segments that may support premium pricing or need competitive positioning.

- Product Repositioning: Understand the nutritional and processing profiles of each segment to optimize product portfolios.
- Targeted Marketing: Use segmentation to tailor marketing messages to consumer groups aligned with specific nutritional or processing levels.