Nutritional Efficiency Analysis Using Statistical and Machine Learning Techniques

Overview

This portion of the hackathon project involves the application of statistical and machine learning methods to analyze a nutritional dataset. The primary goals are to create a robust model to predict calories and identify the most nutrient-efficient foods. The analysis involves topics in regression, Gaussian Mixture Models (GMM), and confidence interval. The goal is to to collect data and provide our own recommendation to the students on what foods to look out for when you go for food pantries.

Objectives

1. **Linear and Nonlinear Regression Analysis:**
    - Perform regression analysis on selected nutritional features to predict calorie content.
    - Evaluate the goodness of fit using the R^2 score.
    - Incorporate confidence intervals into the regression analysis to provide a range of predicted calorie values.
    - Normalize the features to enable standardized comparisons across different nutritional components.
2. **Gaussian Mixture Model (GMM) for Classification:**
    - Apply GMM to classify foods based on their sugar and calorie content.
    - Identify and remove foods classified in the high-sugar group to focus on more nutrient-efficient options.
3. **Efficient Food Selection:**
    - Calculate the difference between actual and predicted calories.
    - Filter out foods where the actual calorie count falls within a 65% confidence interval of the predicted value, ensuring the selection of foods that are truly nutrient-efficient.

Dataset

The dataset used for this analysis includes various nutritional components for different foods, such as:

- Total Fat
- Vitamins (A, B12, B6, C, D, E)
- Protein
- Carbohydrate
- Fiber
- Sugars
- Calories (target variable)

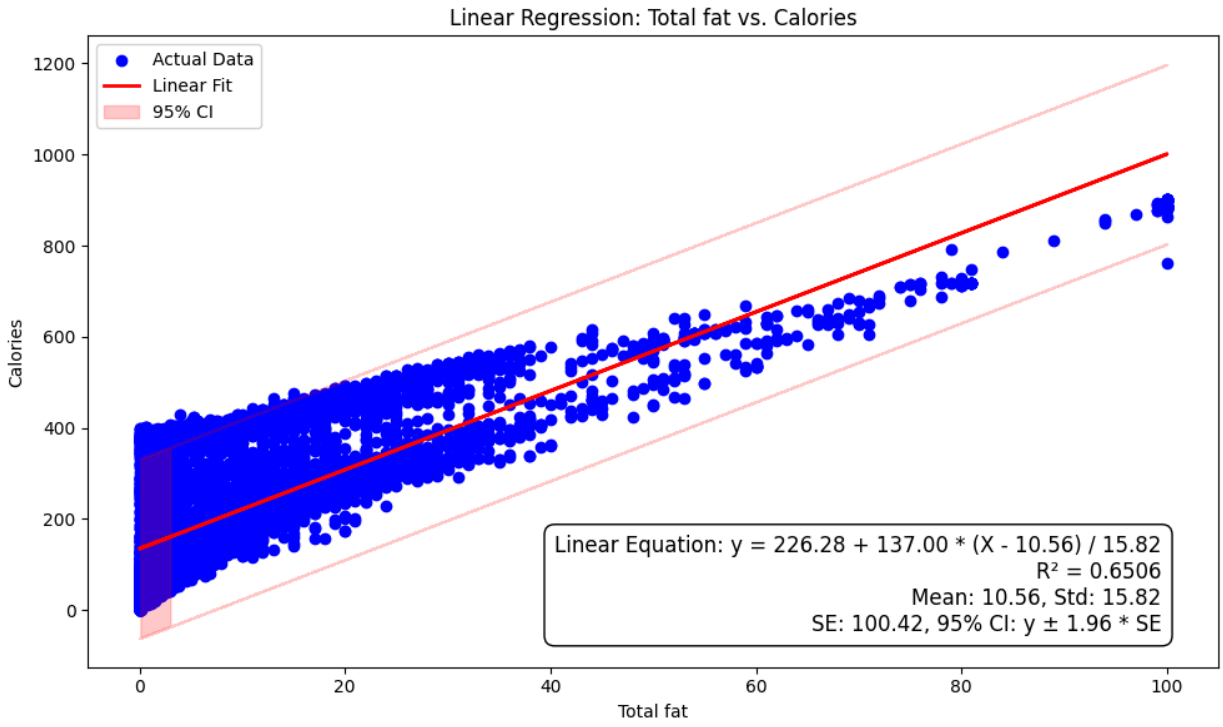
Methodology

1. Data Preprocessing

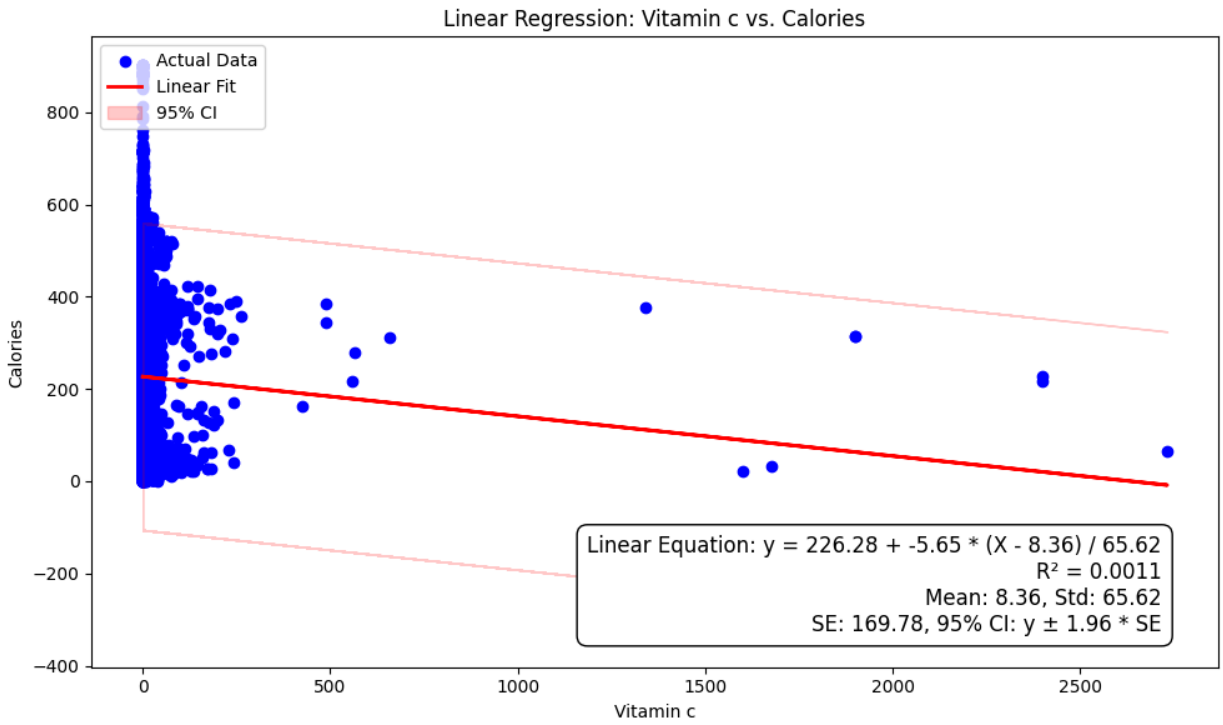
- **Unit Removal**: Measurement units (e.g., "g", "mg", "IU") are removed to ensure numerical consistency across features.
- **Normalization**: Nutritional features are normalized using StandardScaler to standardize the range of the data.

2. Regression Analysis

- **Linear Regression**: Applied to each selected nutritional feature to visualise the correlationt. The regression equation includes normalization terms to allow for real-world application. The point is to see correlation between features and calories.

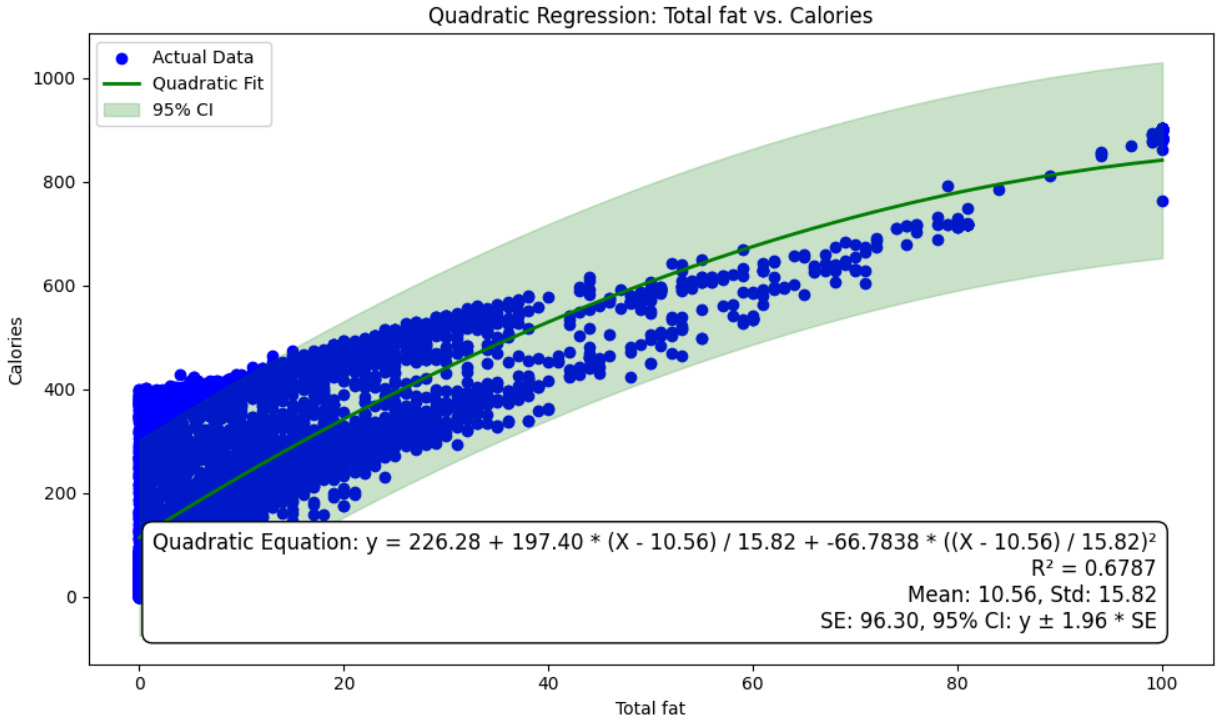


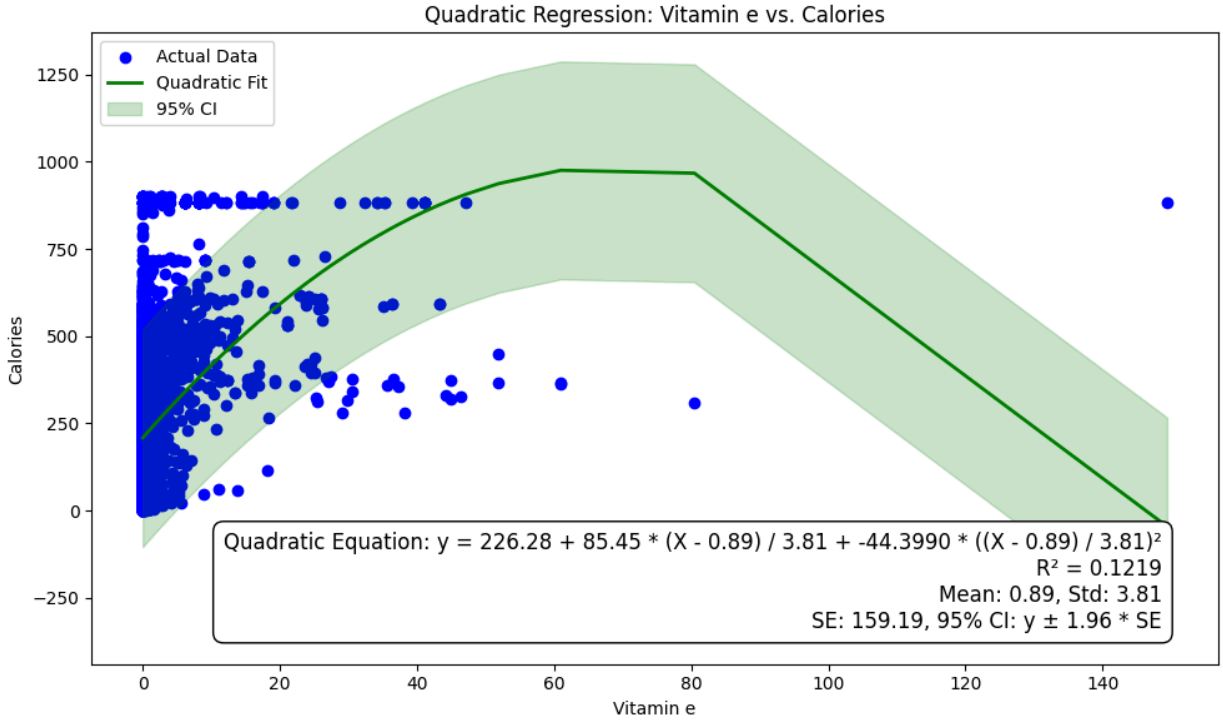
An example of high correlation between features and label.



Example of low correlation

- **Nonlinear Regression (Quadratic)**: Extends the linear model by including a squared term for the selected feature, providing a more flexible fit for the data.





- **Confidence Intervals**: For both linear and quadratic models, confidence intervals are calculated and visualized, offering insight into the reliability of the predictions.

- **R^2 (Coefficient of Determination)**: The R^2 explains the proportion of the variance in the dependent variable (calories) that is predictable from the independent variables (nutritional features). It is an important value to test since it describes the effectiveness of the feature.
- **Adjusted R^2 and Multivariable Regression**: Adjusted R^2 adjusts for the number of predictors in the model and provides a more accurate measure when building multivariable models. It helps to prevent overfitting by penalizing the model for adding predictors that do not significantly improve the model's ability to explain the variance in the data.

$$y = 226.28 + 139.33 * ((\text{Total fat} - 10.56) / 15.82) + 107.45 * ((\text{Carbohydrate} - 22.12) / 27.26) + 42.70 * ((\text{Protein} - 11.35) / 10.53) + -7.67 * ((\text{Fiber} - 2.04) / 4.27) + -1.33 * ((\text{Vitamin c} - 8.36) / 65.62) + -1.58 * ((\text{Vitamin b6} - 0.26) / 0.47) + 0.67 * ((\text{Vitamin b12} - 1.20) / 4.27) + -0.62 * ((\text{Vitamin a} - 676.32) / 3694.53) + 0.43 * ((\text{Vitamin d} - 14.56) / 123.64) + -0.31 * ((\text{Vitamin e} - 0.89) / 3.81)$$
$$\text{Adjusted } R^2 = 0.9915$$
$$SE = 15.62,\ 95\%\ CI = y \pm 1.96 * SE$$

Model fitted by adding features until the adjusted R^2 no longer increases. The model also includes the normalisation of the variables so that you can test it out by hand.

$$y = 226.28 + 139.42 * ((\text{Total\_fat} - 10.56) / 15.82) + 103.25 * ((\text{Carbohydrate} - 22.12) / 27.26) + 41.88 * ((\text{Protein} - 11.35) / 10.53)$$
$$\text{Adjusted } R^2 = 0.9895$$
$$SE = 17.38,\ 95\%\ CI = y \pm 1.96 * SE$$

A simpler yet highly precise model for easier computation.

This part of the analysis focuses on the analysis of the features in order to make a well define model such that it can find food that has good nutrition while having low calories.

3. Gaussian Mixture Model (GMM) Classification

- **GMM**: Used to classify foods into two groups based on their sugar and calorie content.
- **Analysing Clusters:**We take the mean and standard deviation of the data's calories in both clusters. The highest mean would result that cluster being classified as the "high-sugar cluster."
- **Filtering High-Sugar Foods**: Foods identified in the high-sugar cluster are removed from further analysis.

The reasoning for implementing GMM is to counteract the limitation of the regression model. The issue has to do with sugar, where the regression model to filter out bad foods, it still kept sugary foods in. GMM helps with that since it can cluster the data as well classification due to the probabilistic nature of GMM.

4. Efficient Food Selection

- **Calorie Difference Calculation**: The difference between actual and predicted calories is computed for each food.
- **Sorting:** The food items are then sorted by the lowest numerical calorie difference value
- **Filtering by Confidence Interval**: Foods with actual calorie counts within a 65% confidence interval of the predicted value are removed, leaving only the most nutrient-efficient foods.
- **Output**: The filtered list of efficient foods is saved to a text file for further analysis or decision-making.
- **Analysis:** We look through the text file to find common food in the list and determine it to be an efficient food. We will state the result of the analysis by sharing the most nutrient dense food to users of the chatbot when they ask for the location of the nearest pantry

Files

- nutrition.csv: The original dataset containing nutritional information for various foods.
- analysis.py: The Python script that performs the full analysis, from preprocessing to the final selection of efficient foods.
- ml_predict_calories.py: The Python script that constructs the multivariable function through adjusted R^2 method.
- ml_sort.py: The Python script that sorts through our dataset that removes food that doesn't fit the criteria from the regression and GMM analysis.
- filtered_negative_calorie_difference_items.txt: The final output file containing the list of most nutrient-efficient foods, excluding those with high sugar content and low-calorie differences.
- Regression_Plots: Linear and nonlinear regression of all features vs calories.

Results

Based on our statistical analysis, the optimal choices for a well-rounded selection from a food pantry include oat-based products, potatoes, an assortment of nuts, peanut butter, seeds, raisins, beans, noodles, pasta, tomatoes, broccoli, kale, salmon, chicken, ground turkey, and lean beef. These items provide a balanced combination of nutrients to support a healthy diet. Although the algorithm turned 8788 amount of food data to only 167, we still had to look through 167 items in order to determine the common foods. Although this can be automated through another machine learning algorithm, it wasn't implemented due to time constraints. Other analysis on housing in New York and mental health couldn't be done due to the limited data as well as time restriction. However, it allowed us to develop real experience as well as learn more about nutrition.