Final Project Report: Building a Nutritional Health Classifier

Introduction:

We were tasked with building a model that classifies the health category of food items based on their nutritional content. The categories include "Nourishing," "Balanced," and "Indulgent." Our goal was to leverage machine learning techniques to predict the health category of food items accurately.

Data Loading and Exploration:

- Loaded training and testing datasets.
- Explored summary statistics and visualized relationships between features.

Data Preprocessing:

- Handled missing values using appropriate imputation techniques.
- Encoded categorical features.
- Addressed class imbalance using oversampling (SMOTE).
- Performed feature engineering to create new features and handled division by zero or small denominator values.
- Scaled features to a smaller range.
- Conducted feature selection using Recursive Feature Elimination (RFE) and Random Forest.

Model Training and Evaluation:

- Trained a Logistic Regression model on the preprocessed data and evaluated its performance.
- Split the dataset into training, validation, and test sets.
- Trained a Random Forest classifier and evaluated its performance on the validation set.
- Trained a k-Nearest Neighbors (KNN) classifier and evaluated its performance on the validation set.

Results:

Logistic Regression Model:

- **Accuracy: 81%**
- **Precision:**
  - **Class 0: 77%**
  - **Class 1: 88%**
  - **Class 2: 67%**
- **Recall:**
  - **Class 0: 80%**
  - **Class 1: 86%**
  - **Class 2: 65%**
- **F1-score:**
  - **Class 0: 79%**
  - **Class 1: 87%**
  - **Class 2: 66%**

Random Forest Classifier:

- **Accuracy: 98%**
- **Precision:**
  - **Class 0: 98%**
  - **Class 1: 98%**

    - **Class 2: 97%**
- **Recall:**
    - **Class 0: 96%**
    - **Class 1: 98%**
    - **Class 2: 100%**
- **F1-score:**
    - **Class 0: 97%**
    - **Class 1: 98%**
    - **Class 2: 99%**

k-Nearest Neighbors (KNN) Classifier:

- **Accuracy: 86%**
- **Precision:**
    - **Class 0: 76%**
    - **Class 1: 94%**
    - **Class 2: 91%**
- **Recall:**
    - **Class 0: 92%**
    - **Class 1: 88%**
    - **Class 2: 60%**
- **F1-score:**
    - **Class 0: 83%**
    - **Class 1: 91%**
    - **Class 2: 72%**

User Interaction:

- Created interactive widgets to allow users to input nutritional information.
- Integrated the trained model to predict the health category based on user input.

Conclusion:

- The Random Forest classifier achieved the highest accuracy of 98% on the validation set, followed by the KNN classifier with an accuracy of 86%.
- The interactive interface allows users to input nutritional information and obtain predictions for health categories.

Recommendations:

- The Random Forest classifier is recommended for accurate predictions, while the KNN classifier provides a balance between accuracy and computational efficiency.
- Further optimization and tuning of hyperparameters may improve model performance.