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:

1.Loaded and examined the training and testing datasets to understand the structure and characteristics of the data.
2.Conducted exploratory data analysis (EDA) to gain insights into the distribution of features and relationships between them.
3.Visualized summary statistics and relationships between features to identify potential patterns and correlations.

Data Preprocessing:

1.Addressed missing values through appropriate imputation techniques such as mean or median imputation. Encoded categorical features using techniques like one-hot encoding to convert them into a numerical format suitable for modeling.
2.Mitigated class imbalance by employing oversampling techniques like Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for minority classes.
3.Engineered new features based on domain knowledge and transformed existing features to ensure meaningful representation.
4.Applied feature scaling to standardize the range of feature values and improve model convergence.
5.Conducted feature selection using methods like Recursive Feature Elimination (RFE) and Random Forest feature importance to identify the most relevant features for classification.

Model Training and Evaluation:

1.Trained multiple classification models on the preprocessed data, including Logistic Regression, Random Forest, and k-Nearest Neighbors (KNN).
2.Evaluated the performance of each model using metrics such as accuracy, precision, recall, and F1-score on both training and validation sets.
3.Utilized techniques like cross-validation and hyperparameter tuning to optimize model performance and prevent overfitting.
4.Selected the best-performing model based on validation set performance for further analysis and deployment.

Results:

1.Logistic Regression Model:
Achieved an accuracy of 81% on the validation set.
Demonstrated varying levels of precision, recall, and F1-score across different classes.

2.Random Forest Classifier:
Outperformed other models with an accuracy of 98% on the validation set.
Exhibited high precision, recall, and F1-score for all classes, indicating robust performance.

3.k-Nearest Neighbors (KNN) Classifier:
Attained an accuracy of 86% on the validation set.

Showcased competitive performance with respect to precision, recall, and F1-score across classes.

Conclusion:

The Random Forest classifier emerged as the top-performing model, achieving the highest accuracy on the validation set.
The interactive interface facilitates seamless interaction and engagement, allowing users to obtain personalized health category predictions for food items.

Recommendations:

1.The Random Forest classifier is recommended for its superior performance in accurately predicting health categories.
2.Further optimization and tuning of hyperparameters could potentially enhance the performance of the models.
3.Continued monitoring and evaluation of model performance are essential to ensure reliable and accurate predictions in real-world scenarios.