**Food Nutrition Classification Project Report**

**Problem Statement**

In the era of increasing dietary awareness, the ability to classify food items based on nutritional attributes is invaluable. This project focuses on developing a machine learning model that classifies food into multiple categories using tabular data such as calories, proteins, carbohydrates, fats, and sugar. The ultimate goal is to create a robust classification system that can accurately label food types and provide insights into what makes each food category distinct, supporting dietary management systems.

1. Approach

* To address the problem, I adopted a structured data science workflow:
* Data Collection & Understanding : I started with a comprehensive food nutrition dataset containing both numerical (e.g., calories, protein, fat, sugar, sodium, cholesterol, glycemic index) and categorical (e.g., meal type, gluten-free status) features.
* Data Preprocessing & EDA : I performed thorough data cleaning, handled missing values and outliers, visualized feature distributions, encoded categorical variables, and standardized numerical features.
* Feature Engineering & Selection : I identified and selected the most relevant features using feature importance analysis with Random Forest.
* Model Building :I trained and evaluated several classification algorithms to determine the best-performing model.
* Model Persistence : The best models were saved for future use and deployment.

2. Data Analysis

2.1. Data Exploration

The dataset included several hundred food items with a mix of numerical and categorical features.

* Missing values were minimal and were handled by dropping affected rows.
* Outliers were detected using boxplots and capped using the IQR method.
* Duplicate entries were identified and removed to ensure data quality.

2.2. Visualization

* I used KDE plots and boxplots to examine the distribution and outliers in continuous variables like calories, fats, and sugar.
* Count plots were used to visualize the frequency of each food category and meal type.
* Feature importance analysis with Random Forest highlighted the most influential features for classification.

2.3. Feature Engineering

- Boolean features (e.g., gluten-free) were encoded as 0/1, and other categorical features were label-encoded.

- StandardScaler was applied to numerical features for uniformity

.

2.4. Feature Importance

- RandomForestClassifier was used to determine the most relevant features.

- Top features included calories, protein, fat, sugar, and glycemic index.

3. Model Selection and Evaluation

3.1. Models Trained

I trained the following models:

- Logistic Regression

- Decision Tree Classifier

- Random Forest Classifier

- K-Nearest Neighbors

- Support Vector Machine

- XGBoost Classifier

- Gradient Boosting Classifier

3.2. Evaluation Metrics

Each model was evaluated using:

- Accuracy

- Precision

- Recall

- F1 Score

- Confusion Matrix

3.3. Results

- Ensemble methods (Random Forest, XGBoost, Gradient Boosting) generally outperformed simpler models.

- Feature importance analysis confirmed that calories, protein, fat, sugar, and glycemic index are key differentiators among food categories.

- All trained models were saved using `pickle` for future deployment.

4. Insights and Recommendations

- Key Features : Calories, protein, fat, and sugar are the most significant attributes for classifying food categories.

- Model Choice : Ensemble models (Random Forest, XGBoost) are recommended for deployment due to their superior performance and robustness.

- Data Quality : Maintaining clean, well-annotated data with minimal missing values and outliers is crucial for high model accuracy.

- Practical Use : The classification system can be integrated into dietary management apps, food tracking systems, or nutritional recommendation platforms.

By:M.Karthik Chandra

ελπίδα για καλή τύχη

Thanks for the opportunity