

Documentation: Nutrition & Disease Prediction Model

1. Objective

The goal of this project is to build a machine learning model that predicts disease categories based on nutritional intake features (Calories, Protein, Carbohydrates, Sugar, Fat, Sodium). This supports research into dietary patterns and chronic disease risk.

2. Workflow Overview

1. Data Preparation

- Selected relevant nutritional features.
- Encoded target variable (Disease) using LabelEncoder.
- Split dataset into training and testing sets (80/20).

2. Feature Engineering

- Target encoding for categorical labels.
- Addressed severe class imbalance (Weight Gain dominates, Weight Loss has only 1 sample).

3. Modeling

- Trained a baseline **RandomForestClassifier**.
- Evaluated using accuracy and classification report.

4. Resampling

- Applied **SMOTE** oversampling to balance minority classes.
- Removed "Weight Loss" class (only 1 sample) to avoid SMOTE errors.

5. Hyperparameter Tuning

- Used **GridSearchCV** to optimize parameters (n_estimators, max_depth, min_samples_split).
- Retrained model with best parameters on resampled data.

3. Results

- **Accuracy improved** from ~0.53 to ~0.56 after balancing and tuning.
- **Weight Gain** predictions remain strong (precision ~0.97, recall ~0.78).
- **Minority classes** (Diabetes, Heart Disease, Acne) show weak precision/recall despite SMOTE.
- Macro average F1 remains low (~0.24), highlighting imbalance challenges.

4. Challenges

- **Extreme imbalance:** Weight Gain dominates (>60% of samples), Weight Loss nearly absent.
- **Minority classes:** Difficult for the model to learn meaningful patterns.
- **Synthetic oversampling limits:** SMOTE cannot generate samples for classes with <2 instances.

5. Next Steps

- Explore **boosting algorithms** (XGBoost, LightGBM) for better handling of imbalance.
- Engineer new features (nutrient ratios, normalized values).
- Use **macro F1-score** and confusion matrix for fairer evaluation.
- Collect or augment more samples for underrepresented diseases.
- Consider merging extremely rare classes into broader categories.