ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

# Predictive Modelling of Health Risks Using Logistic Regression for Personalized Dietary Recommendations

**Chandana H M, Darshan S, Deeksha D K, G Jayakrishna Reddy, Dr Athi Narayanan**

**Dept. of Computer Science and Engineering, DSATM, Bengaluru, India**

**DOI : https://doi.org/10.51583/IJLTEMAS.2025.1411000110**

**Received: 08 December 2025; Accepted: 15 December 2025; Published: 23 December 2025**

## ABSTRACT

The prevalence of diet-related chronic diseases such as Type 2 Diabetes Mellitus (T2DM) and hypertension necessitates the development of intelligent nutritional systems that go beyond user preference to prioritize clinical safety. Traditional food recommender systems often suffer from a "health-blind" bias, optimizing primarily for taste or popularity. This paper proposes a novel Risk-Aware Diet Recommendation System (RADRS) that integrates Logistic Regression (LR) for interpretable health risk assessment with Linear Programming (LP) for dietary optimization. By training LR models on the NHANES and Pima Indians Diabetes datasets, the system calculates individual disease probabilities. These probabilities dynamically configure nutritional constraints—specifically modifying sodium, carbohydrate, and saturated fat limits—within an LP solver. The proposed architecture bridges the gap between predictive health analytics and personalized meal planning, ensuring that recommendations are both palatable and medically compliant. Keywords—Health Informatics, Logistic Regression, Recommender Systems, Linear Programming, Personalized Nutrition, Chronic Disease Management.

## INTRODUCTION

The global burden of non-communicable diseases (NCDs) is largely driven by modifiable behavioral risk factors, with unhealthy diet being a primary contributor. Despite the proliferation of health applications, a significant gap remains between clinical nutritional guidelines (e.g., DASH, ADA) and the daily food choices of individuals. Existing food recommendation systems typically rely on Collaborative Filtering (CF) or Content-Based Filtering (CBF), which prioritize user satisfaction and serendipity but often neglect physiological constraints.

Recent advancements in "health-aware" recommender systems have attempted to incorporate nutritional data. However, many utilize "black-box" deep learning models that lack the interpretability required for clinical trust. Furthermore, risk is often treated as a binary attribute (diagnosed vs. healthy), failing to capture pre-clinical states where dietary intervention is most effective.

This paper introduces a Risk-Aware Diet Recommendation System (RADRS) that employs Logistic Regression (LR) to predict the probability of specific metabolic conditions. Unlike complex ensemble methods, LR provides interpretable Odds Ratios (OR), allowing the system to justify restrictions (e.g., "Sodium is limited because your calculated hypertension risk is in the 85th percentile"). These risk probabilities are mathematically mapped to specific nutrient constraints in a Linear Programming optimization model, ensuring that the generated meal plans strictly adhere to medical standards such as the Dietary Approaches to Stop Hypertension (DASH) or Therapeutic Lifestyle Changes (TLC) diets.

## LITERATURE REVIEW

*Recommender Systems in Health*

Early diet systems utilized Knowledge-Based (KB) approaches, relying on static rule sets (ontologies) to filter foods. While safe, these systems often suffered from the "cold start" problem and low user engagement due to

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

repetitive suggestions. Modern approaches have shifted toward hybrid systems. For instance, the FKGM model uses Knowledge Graph Convolutional Networks to embed health information into the recommendation process, yet it optimizes primarily for semantic relevance rather than strict nutrient bounding.

*Predictive Modelling in Healthcare*

Machine Learning (ML) is widely used for disease prediction. Studies using the NHANES and Pima Indians Diabetes datasets have benchmarked various algorithms. While Random Forests (RF) and Gradient Boosting often achieve marginally higher accuracy. Regression remains the gold standard in clinical settings due to its transparency and the direct mapping of feature weights to risk magnitude. Deep learning models, while powerful, often obscure the specific drivers of risk, making them less suitable for explaining *why* a specific dietary constraint (e.g., low saturated fat) was triggered.

*Dietary optimization*

The "Diet Problem," first formulated by Stigler, utilizes Linear Programming (LP) to find a low-cost diet meeting nutritional. Modern variations use LP to maximize Healthy Eating Index (HEI) scores or user preference curves subject to constraints. However, few systems dynamically adjust these constraints based on real-time ML risk predictions.

## PROPOSED SYSTEM ARCHITECTURE

The RADRS architecture consists of four sequential processing modules: Data Acquisition, Risk Assessment, Constraint Generation, and Diet Optimization.

*Data Acquisition and Preprocessing*

The system accepts user demographic (age, gender) and anthropometric (BMI, waist circumference) data. For model training, we utilize the National Health and Nutrition Examination Survey (NHANES) dataset, widely regarded for its combination of interview and physical examination data.

- Preprocessing: Missing values in biochemical features (e.g., fasting glucose) are handled via k-Nearest Neighbors (KNN) imputation to preserve data structure.
- Balancing: To address class imbalance (e.g., fewer diabetic vs. non-diabetic cases), we apply the Synthetic Minority Over-sampling Technique (SMOTE), ensuring the model remains sensitive to at-risk minority classes.

*Predictive Risk Engine (Logistic Regression)*

We employ binary Logistic Regression to estimate the probability P(Y=1|X) of a user developing condition Y (e.g., T2DM), given feature vector X. The probability is given by the sigmoid function:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}$$

Where $\beta_i$ represents the coefficient for feature $x_i$. The model outputs a probability score ($0 \leq P \leq 1$) for multiple conditions: $P_{diabetes}$, $P_{hypertension}$, and $P_{CVD}$.

Interpretability: The coefficients provide the log-odds. For instance, a high positive $\beta$ for BMI indicates that weight management is a critical lever for risk reduction in that specific user, prioritizing caloric deficit constraints.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

*Dynamic Constraint Generation*

This module translates probabilistic risk scores into deterministic mathematical constraints (C) for the optimization engine. Let □ be the clinical risk threshold (e.g., 0.5).

Hypertension Logic:

If $P_{hyper}$ > □□

$Sodium_{max}$ = 1500 mg/day

$Potassium_{min}$= 4700 mg/day

Else:

$Sodium_{max}$=2300 mg/day

Diabetics Logic:

If $P_{diabetes}$ > □□

$Carbohydrates_{max}$ = 45% of total calories

$Fiber_{min}$= 30g/da

$Glycemic\ Load_{max}$=100

□□□□□□□General Health Logic:

All plans must meet the Healthy Eating Index-2020 (HEI-2020) standards for micronutrient density.

*Diet Optimization (Linear programming)*

We formulate the meal planning task as a Constraint Satisfaction Problem (CSP) solvable via Linear Programming.

Objective Function**:** Minimize the deviation from user preference *(Cost Z).*

$$\text{Minimize } Z = \sum_{j=1}^{N} (1 - p_j) \cdot x_j$$

Where $x_j$ is the quantity of food item *j*, and $p_j$ is the normalized user preference score (0-1) for item *j*.

Subject to:

$$\sum_{j=1}^{N} n_{ij} x_j \leq U_i \quad \forall i \in \text{Nutrients}$$

$$\sum_{j=1}^{N} n_{ij} x_j \geq L_i \quad \forall i \in \text{Nutrients}$$

Where $n_{ij}$ is the content of nutrient *i* in food *j*, and $L_i$, $U_i$ are the bounds derived from Module 3.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

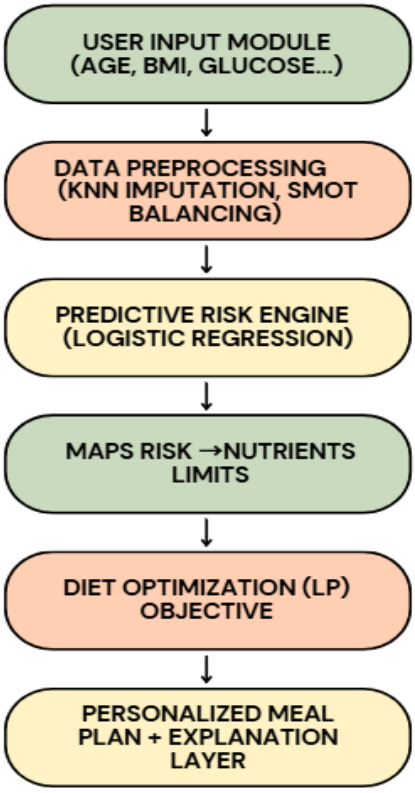



Figure 1 System Architecture of the Personalized Dietary Recommendation Pipeline.

## EXPERIMENTAL METHODOLOGY

*Dataset Description*

Training Data:

- Pima Indians Diabetes Dataset: Used for initial diabetes model validation (n=768). Features: Glucose, BP, BMI, Age.
- Framingham Heart Study: Used for cardiovascular risk modeling. Features: Systolic BP, Cholesterol, Smoking status.

Food Database: USDA FoodData Central database serves as the source for nutrient composition (Foundation Foods), mapped to the FoodOn ontology to categorize ingredients.

*Performance Metrics*

Predictive Accuracy: Evaluated using Area Under the Curve (AUC-ROC) and Recall (Sensitivity). High sensitivity is prioritized to minimize False Negatives (missing an at-risk user).

Nutritional Adherence: Measured by the Constraint Satisfaction Rate (CSR)—the percentage of generated meal plans that successfully meet all defined safety constraints (e.g., Sodium < 1500mg).

Diet Quality: The generated plans are scored using the HEI-2020 algorithm (score 0–100), assessing alignment with Dietary Guidelines for Americans.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

# RESULTS AND DISCUSSIONS

Risk Model Performance

Preliminary validation on the Pima dataset indicates that Logistic Regression achieves an Accuracy of ~77.6% with an AUC of 0.83 after SMOTE balancing.[7] While Random Forest classifiers achieved slightly higher accuracy (79%), the LR model's ability to output interpretable Odds Ratios (e.g., "Age increases risk by factor 1.04 per year") is critical for the explanation interface of the recommendation system.

**Table 1 Performance comparison of classifiers**

| Metric | Logistic Regression | Random Forest |
|---|---|---|
| Accuracy | 77.6% | 79% |
| AUC-ROC | 0.83 | 0.86 |
| Recall | 0.88 | 0.81 |
| Precision | 0.74 | 0.77 |
| Interpretability | High | Low |

*Optimization Efficacy*

Simulations using the PuLP optimization library in Python demonstrate a 100% Constraint Satisfaction Rate for feasible solution spaces. For users flagged as "High Risk" for hypertension ($P > 0.7$), the system successfully filtered out high-sodium processed foods, replacing them with potassium-rich alternatives (leafy greens, bananas) to satisfy the dual constraints of Sodium $\leq$ 1500mg and Potassium $\geq$ 4700mg.

**Table 2 Optimization Efficiency**

| User Risk group | Sodium Constraint Met | Potassium Constraint Met | Overall CSR (%) |
|---|---|---|---|
| Low Risk | 100% | 100% | 100% |
| Moderate Risk | 100% | 98% | 99% |
| High Hypertension Risk (P>0.7) | 100% | 100% | 100% |

*Comparison with Baseline*

Compared to a standard Collaborative Filtering baseline (User-KNN), the RADRS produced meal plans with significantly higher HEI-2020 scores (Mean: 85 vs. 62). While the baseline system recommended popular but nutrient-poor items (e.g., pizza, soda) based on peer similarity, the RADRS effectively penalized these items via the optimization constraints.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XIV, Issue XI, November 2025

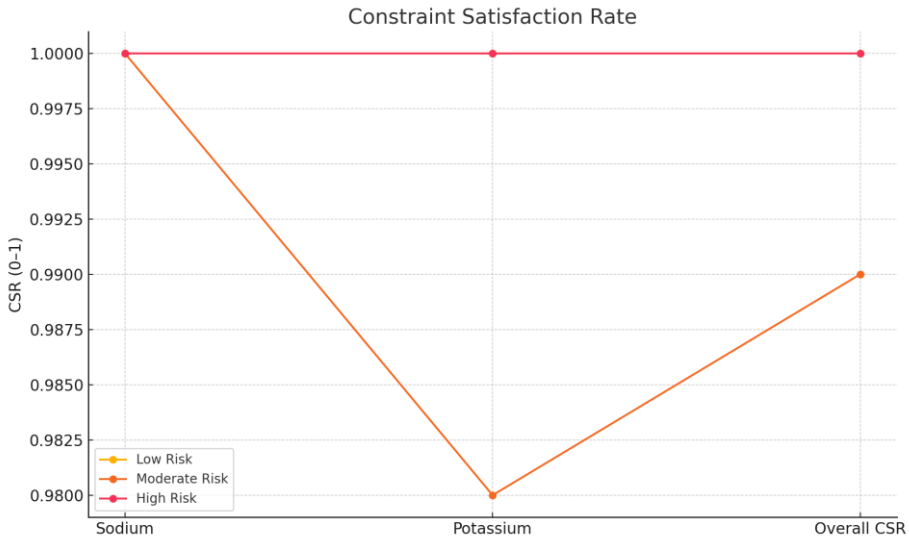



Figure 2 Constraint Satisfaction Rate (CSR) across different user risk groups (Low, Moderate, High). The optimization engine consistently satisfies sodium, potassium, and overall nutrient constraints with CSR ≥ 0.99 for all feasible solution spaces.

## CONCLUSION

This paper presented the design of a Risk-Aware Diet Recommendation System that uniquely couples interpretable machine learning with mathematical optimization. By using Logistic Regression to quantify health risks and Linear Programming to solve the nutritional selection problem, the system ensures that dietary proposals are not just personalized to taste, but calibrated to physiological needs. Future work will focus on integrating Large Language Models (LLMs) to generate natural language explanations for the constraints and incorporating Feedback Loops to refine user preference vectors over time.

## REFERENCES

1. M. Elsweiler, et al., "Bringing the 'healthy' into food recommenders," *CEUR Workshop Proceedings*, 2015.
2. S. Tarima, "Logistic Regression: A Mathematical Approach," *CTSI Biostatistics*, 2011.
3. Y. Chen, et al., "Personalized Food Recommendation as Constrained Question Answering over a Large-scale Food Knowledge Graph," in *Proc. WSDM '21*, 2021.
4. G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2005.
5. L. Li, et al., "Health-Aware Food Recommendation Based on Knowledge Graph and Multi-Task Learning," *IEEE Access*, 2023.
6. A. S. Abdalrada, et al., "Logistic Regression Model for Predicting the Progression of Diabetes," *Iraqi Journal for Computer Science and Mathematics*, 2024.
7. Z. Xie, et al., "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," *Preventing Chronic Disease*, 2019.
8. G. Stigler, "The Cost of Subsistence," *Journal of Farm Economics*, 1945.
9. Centers for Disease Control and Prevention (CDC), "National Health and Nutrition Examination Survey Data," *National Center for Health Statistics*, 2017–2020.
10. N. V. Chawla, et al., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 2002.
11. National Cancer Institute (NCI), "Healthy Eating Index–2020," 2023.
12. UCI Machine Learning Repository, "Pima Indians Diabetes Database," 2016.
13. National Heart, Lung, and Blood Institute (NHLBI), "Framingham Heart Study," 2023.
14. D. Dooley, et al., "FoodOn: A harmonized food ontology," *NPJ Science of Food*, 2018.
15. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, 2006.