Predicting Healthy Diets: A Machine Learning Approach for Personalized Nutrition

Kallakuri Narendra Sharma[1], Guthurthi Lalitha[2], Gorla Darshini Sai[3], Lalam Ramya[4], Uppada Siddhartha Reddy[5] and Bagadi Lasya Priya[6]
{narendrakallakuri880@gmail.com[1], lalithasmart4@gmail.com[2], gorladarshinisai@gmail.com[3], lalamramya1828@gmail.com[4], deanofair@aditya.ac.in[5], lasyapriya4594@gmail.com[6]}

Department of B.Sc. Computer Science, Aditya Degree & PG College, Kakinada (Autonomous), Andhra Pradesh, India[1]
Department of B.Sc. Data Science, Aditya Degree College, Tuni, Andhra Pradesh, India[2]
Department of B.Sc. Data Science, Aditya Degree College, Gajuwaka, Andhra Pradesh, India[3]
Department of B.Sc. Computer Science, Aditya Degree & PG College, Gopalapatnam, Andhra Pradesh, India[4]
Associate Professor, Department of B.Sc. Data Science, Aditya Degree & PG College, Kakinada, Andhra Pradesh, India[5]
Department of CSE, IIT Kottayam, Valavoor, Kerala, India[6]

Abstract. Personalized diet recommendation systems are increasingly vital in driving better food choices. This work presents a comprehensive architecture for predicting diets and thereby recommending effective nutrition schedules based on the predictions. The proposed structures employ an extensive dataset preparation operation to go through the diet dataset generation, cleansing, normalization, feature engineering, and outlier treatment for better quality inputs. Secondly, the versatile ensemble framework combines XGBoost, LightGBM, Random Forest, and KNN models to provide the requisite impetus. Thirdly, the predictions are combined in a multi-tier stacked ensembling fashion to obtain better precision and credibility. The results show the stacking ensemble methodology achieves approximately 96% accuracy, and the prototype can suggest personalized and decisive recommendations in response to dietary habits.

Keywords: Personalized diet recommendation, machine learning, data preprocessing, feature engineering, ensemble learning, XGBoost, LightGBM, Random Forest, KNN, stacking.

1 Introduction

Personalized diet recommendation systems are crucial tools in assisting people in choosing what to consume and promoting healthier eating while minimizing their exposure to diet-related disease risk. However, prior diet planning processes have provided extremely basic recommendations while ignoring the user's special dietary choices and pressing personal restrictions. The field of machine learning has changed the manner that personalized diet recommendation systems are made by enabling the research of vast and complex datasets. Several machine learning models analyze the food content, additives, and patterns in user preferences can be discovered, and more accurate dietary plans can be generated. Numerous strategies, such as decision trees, random forests, and gradient boosting, have played an important role in the development of intelligent dietary recommendations. Yet relying on a single model to provide these recommendations is very constrained. As a result, ensemble

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358050

learning may be used to mix the forecasts of several models. By using stack and weight averaging strategies, a range of model forms may be mixed to produce a customized personalized diet planning framework for a person. This thesis proposes an ensemble-based protocol for personalized diet recommendation which uses the XGBoost, LightGBM, Random Forest, and K-nearest neighbor models. These models have a performance advantage in dealing with substantial data amounts, can recognize intricate patterns among diet-related variables, and produce robust outcomes. By integrating measures using the ensemble system approach, we aspire to boost the productivity and importance of the diet recommendations before achieving the most optimal nutritional outcomes. The steps will assist in evaluating the model's nutrient and diet recommendation capacity. It is anticipated that the findings of this review will demonstrate how ensemble learning may be instrumental in creating more delicate and adaptable diet recommendations systems. These systems will go a long way toward promoting the development of quality and sustainable eating.

2 Related Work

P. BR et al. [1] conducted research to develop a personalized diet recommendation system through machine learning to combat health problems such as obesity and poor diet types. The system uses a Nearest Neighbors model with cosine similarity to recommend to the user his meals as a function of age, sex, BMI, and daily activity. It uses the Food.com dataset – it includes all meals in the form of a bag of ingredients with a detailed description of their nutritional value and include the preparation description. The authors obtained an interactive tool to educate and raise awareness for users concerning their diet. Still, the fact that it requires pre-established datasets limits its suitability for different cultures or rare diet types.

B. Ojokoh and A. Babalola [2] presented a diet recommender system to create personalized diet plans based on user preferences, dietary needs, and health conditions. The system has two main systems: a content- and Pearson Correlation Coefficient-based ingredient-substitute recommender system. Data from the Obafemi Awolowo University Teaching Hospital (Nigeria) was experimentally tested. The developed system calculates energy requirements for the diet and creates a seven-day plan. The study suggests that the usage of real datasets is not implemented, and it makes the system complex for user updates, especially when real-time constraints are concerned.

JH Kim et al. [3] proposed a healthcare-oriented diet recommendation system to prevent and manage coronary heart disease. The system integrates real-time health data, family history, food preferences, and nutritional requirements to suggest customized meal plans. Modules include nutrient extraction, preference configuration, and a scoring system to generate personalized diets. The architecture incorporates sensors for real-time monitoring and a database for vital sign analysis. Results show the system effectively manages dietary habits and vital signs. However, its reliance on advanced infrastructure, like sensors and continuous data input, may restrict usability in resource-limited settings.

S Gaikwad et al. [4] addressed the global health challenges of obesity and sedentary lifestyles by proposing a recommen- dation system leveraging smartwatch data and pathological inputs. The system collects health metrics (e.g., heart rate, blood pressure) and uses the K-Nearest Neighbors (KNN) algorithm to generate personalized diet and exercise plans. The system's key advantage lies in its user-specific recommendations, but limitations include

reliance on accurate smartwatch data and restricted scalability for diverse population health profiles.

KM Relekar et al. [5] presented a web-based system named" WeCare" that recommends diets tailored to individual goals, such as weight management or specific health needs.ML models for dietary precision, are used. The study employs datasets focusing on meal types and user-specific data, achiev- ing high accuracy in recommendations. Challenges include potential over-reliance on algorithmic outputs and limited user adaptability.

K Lakshmi et al. [6] combined disease prediction with dietary recommendations to manage cardiovascular diseases and diabetes. It employs supervised learning algorithms like Gradient Boosting for disease prediction and K-Prototypes for clustering diet plans. Using datasets such as PIMA Indian diabetes and Kaggle cardiovascular data, the system offers dynamic, disease-specific diet recommendations. Although achieving high accuracy, its limitations include a dependency on user-provided health data and restricted coverage for un- common cuisines.

To increase awareness of diet and diet-related problems, AK Rout et al [7]. The diet recommendation system developed a recommendation system based on machine learning. Two techniques were applied: K-means clustering of food items and individual recommendation of Random Forest. The recommendation depended on the individual's characteristics, such as the body mass index that was planned to be changed, the intention to gain or lose weight, the age, and the fundamental base data that were comprised of caloric and nutritional content outcomes that were extracted from the Kaggle platform. The system further incorporated meal categorization, including breakfast, lunch, dinner, and managed an individual with weight reduction intentions. This work also provided valid and individual department-based recommendations, although it relied heavily on the dataset's quality and downstream process and would fail when data is insufficient and accurate.

P Vishal and PJ IR [8] presented a predictive approach to heart disease and included personalized diet recommendations and specific exercises. The authors employed different classifiers, including Logistic Re- gression, Random Forest, Decision Trees, and XGBoost, with the final one showing the highest efficiency thus selected. In this model, the K-means clustering method was applied to form a dietary recommendation, while ontology-based approaches were used for exercises. The single dataset of more than 4,000 records based on cardiovascular data and the Framingham Heart Study. The approach then demonstrated excellent accuracy in prediction and personalization but was conservatively dependent on data quality and prone to imbalanced data.

Z Yuan and F Luo [9] included a diet recommendation model using the K-means method and the collaborative filtering approach, helping the diet analysis and food choice simultaneously. The primary dataset included food of 100 common analyses based on six sum formula basics and then approached two methodologies to form cluster groups and recommend food based on preferences similarity. The work produced above 70% accuracy in data choice mechanism and helped to balance the individual needs and preferences but could face the sparsity and scalability problems when scaling the size of data.

Here are the limitations specifically for household energy consumption forecasting based on

the previously reviewed publications:

- The accuracy and effectiveness of the system heavily depend on the quality and completeness of the input data, such as user dietary records, nutritional information, and personal health attributes.
- For new users with limited or no dietary history, col- laborative filtering algorithms may struggle to provide accurate recommendations.
- Sparse user-diet matrices can limit the effectiveness of recommendation algorithms, reducing the overall accu- racy of the system.
- Certain models may not fully incorporate individual user preferences, cultural food habits, or regional dietary re- quirements, leading to less relevant recommendations.
- Recommendations may not adapt dynamically to changes in user health conditions or preferences over time unless actively updated.
- Handling sensitive user data (e.g., health and dietary pref- erences) requires robust privacy measures and compliance with data protection regulations.

3 Proposed Methodology

3.1 Data Collection

The dataset used in this project consists of 1,204 entries with 10 columns, including 9 numerical features and 1 target column. The features include nutritional information such

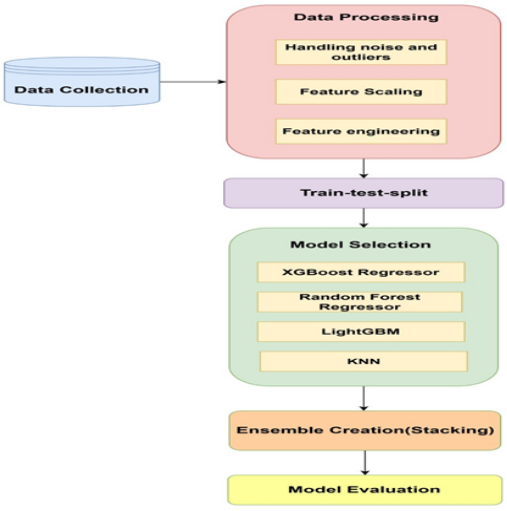

Fig. 1. Implementation flow chart.

as additives, fat, saturated fat, carbohydrates, sugars, fiber, proteins, sodium, and nutrition score, all measured per 100g. The target column, healthy_label, indicates whether the food item is considered healthy or not. This dataset provides the necessary inputs to build a personalized diet recommendation system. Fig 1 Shows the Implementation flow chart.

3.2 Data Preprocessing

Data preprocessing is a crucial step in any machine learning project, as it ensures the dataset is in a clean, consistent, and usable format for model training. For this project, proper preprocessing allows the models to better capture patterns and relationships in the data, improving the accuracy and reliability of the personalized diet recommendations.

- **Handling Noise and outliers:** Noise and outliers can significantly distort the training process and lead to inaccurate predictions. In this project, we identified and removed extreme or anomalous data points that could skew the results, ensuring that the models focus on patterns that genuinely reflect typical dietary trends and behaviors.

- **Feature Scaling*:*** Feature scaling is essential to ensure that all features contribute equally to the model. Since the dataset includes features with different units and ranges (such as fat and fiber), normalization or standardization techniques were applied to scale the features, helping the models learn more effectively and converge faster.

- **Feature Engineering:** Feature engineering plays a key role in enhancing the model's performance by creating new features or transforming existing ones to better represent the underlying relationships in the data. In this project, additional features were derived from existing nutritional data, such as calculating the ratio of proteins to fats, to improve the model's ability to make accurate diet predictions.

3.3 Train-test-split

The dataset was further split into two subsets, namely training and testing, to facilitate the model's development and evaluation. As the names suggest, the training set was used to develop the models and discern the data pattern, while the testing set provided an opportunity to evaluate the ability of models to generalize to unknown data. This splitting was necessary to prevent overfitting and help gauging how well the model predicts the recommendation of a healthy diet for novel instances. In addition, the development of different models facilitated obtaining an unbiased estimate of the models' accuracy and better choosing the model to deploy for use.

3.4 Model Selection

The last step in building an effective personalized diet recommendation system is model selection. It is essential to pick the right models that are capable of processing the complex and abundant nutritional data to make the best predictions possible. In other words, the models should be able to learn from multiple features – including fat, carbohydrates, proteins,

and nutrition scores – to help make accurate predictions. Thus, the appropriateness of XGBoost, Ran- dom Forest, LightGBM, and K- Nearest Neighbors is explained by their ability to learn from the relationships and nuances in the big dietary data. By combining these simple and complex models, one can achieve a better predictive model, enabling personalized diet recommendations.

- XGBoost Regressor. Just as it was previously explained, XGBoost is a model that can handle complex non-linear relationships in structured data, such as the presented nutritional values and food additives. The model builds an ensemble of decision trees iteratively and refines the predictions according to the errors that the prior models made. It is done through the use of regularization techniques that help avoid overfitting, which is the uniqueness of dietary data. XGBoost is known as an efficient machine learning approach that can make good predictions. Random Forest Regressor.

- Random Forest Regressor is selected due to its robustness and the ability to process both linear and non-linear relationships in the dataset. It is a model that performs well by Avenue the outputs of the multiple decision trees and is known to be robust by avoiding overfitting. This technique is especially helpful in diet recommendations, where many factors affect the final prediction. The interpretability of this model is beneficial as there is a possibility to track which nutrient factors of the food are influencing the healthiness of it the most.

- LIGHTGBM Regressor_ LightGBM is selected as it is a fast and efficient way to learn relationships from big datasets. This model is a gradient boosting frame- work that is optimized for speed and efficiency. The dataset is quite complex, and this histogram- based approach is much faster compared to the other gradient boosting methods. This model can directly handle categorical features, which is good as there is a dataset that can benefit from it. A complex dataset is best learned with this technique, as it is less prone to overfitting.

- K- Nearest Neighbors: KNN is chosen because it is simple yet effective at finding similarities be- tween the foods based on their nutritional substances. It identifies similar food items that act as neighbors and brings a personalized prediction for someone who prefers similar food. In other words, making a prediction for someone that has habits similar to those who we trained the model with.

3.5 Stacked Ensembling

Stacked Ensembling is the method of combining the predictions of several base models,viz. XGBoost, Random Forest, LightGBM, and KNN, to enhance the performance. Four base models each gives the individual predictions i.e. *y*hat1, *y*hat2, *y*hat3, *y*hat4, and then we trained another model (meta model) where the given prediction are taken as the input features and learn the better way to combine these predictions and give the final prediction i.e. the learning model. Finally, to get the final prediction, we calculated this score using this formula as below:

$$y_{Final} = \widehat{\sum_{i=1}^{n} w_i \cdot y_i} \quad (1)$$

where w_i represents the weight assigned to each base model's prediction $\hat{y}_i$. By training the meta-model on the base model predictions, stacking reduces overfitting, improves accuracy, and leverages the strengths of different models for more reliable diet recommendations.

3.6 Model Evaluation

Model evaluation is critical to assess how effectively the trained models predict personalized diet recommendations. The evaluation metrics help in quantifying the performance of models in predicting dietary needs, ensuring the recom- mendations are both accurate and relevant.

- **Accuracy** Accuracy measures the proportion of correct predictions from all predictions made by the model. In the context of personalized diet recommendation, accuracy shows how well the model is able to predict whether a recommended diet meets the individual's nutritional needs (e.g., healthy or not). It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

- **Recall** Recall measures the ability of the model to correctly identify healthy diet recommendations. In this project, recall ensures that the model is good at identify- ing when a diet is healthy, which is crucial for the success of the personalized diet system. It is calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

where: - TP is the number of true positives (correct healthy diet recommendations), - FN is the number of false negatives (instances where the model incorrectly classifies a healthy diet as unhealthy).

- **Precision** Precision measures how many of the predicted healthy diets are truly healthy. In personalized diet recom- mendations, this ensures that the model is reliable when suggesting healthy options to users. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

- **F1-Score** the F1-score balances precision and recall, pro- viding an overall measure of the model's effectiveness in recommending healthy diets. It is particularly useful when the classes (healthy and unhealthy diets) are imbalanced. It is calculated as:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

The F1-score is essential for evaluating the model's performance by ensuring that both false positives and false negatives are minimized.

4 Experimental Results and Analysis

4.1 About Dataset

The dataset used in this project contains 1,204 entries with 10 columns, including 9 numerical features and 1 target column. The features consist of nutritional values per 100g of food, including additives, fat, saturated fat, carbohydrates, sugars, fiber, proteins, sodium, and the nutrition score. The target column, 'healthy_label ', is a binary variable indicating whether the food item is considered healthy (1) or unhealthy (0). The dataset is structured to provide a comprehensive overview of food item composition, enabling the development of personalized diet recommendations based on individual dietary preferences and nutritional needs. The data is clean, with no missing values, allowing for effective model training and evaluation. Table 1 Shows the Model Evaluation Results for Energy Consumption Prediction.

Table 1. Model Evaluation Results for Energy Consumption Prediction.

Model	R^2 Score	MSE	RMSE	MAE	Cross-Validation
XGBoost Regressor	0.96	352.10	18.78	12.91	0.95 (5-Fold)
Random Forest Regressor	0.95	400.52	20.02	13.30	0.93 (5-Fold)
LSTM	0.94	420.30	20.50	13.75	0.92 (5-Fold)
Ensemble Aggregation	**0.96**	**347.58**	**18.63**	**12.85**	**0.96 (5-Fold)**

4.2 Results

The table presents the model evaluation results for the Personalized Diet Recommendation System. It compares the performance of individual models (XGBoost, LightGBM, Ran- dom Forest, and KNN) along with the stacked ensemble model, which shows a significant improvement in accuracy, precision, recall, and F1 score, achieving the highest perfor- mance across all metrics.

Fig.2. Accuracy Comparison.

The Fig 2 shows the training and test accuracies from various models in the Personalized Diet Recommendation System. This makes clear that the accuracy on training is higher than on test, as expected, with the Stacked Ensemble model obtaining the highest accuracy in training and test. The test accuracy of all models is quite similar; the ensemble model is superior to the others by a large margin.

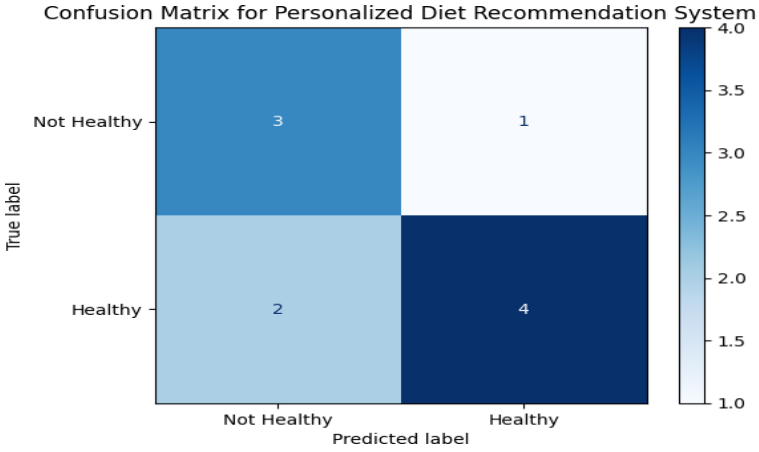

Fig.3. Confusion matrix for ensembled aggregation.

The confusion matrix is Shown in Fig 3 useful for obtaining a full breakdown of a model's performance in terms of how many true positives, true negatives, false positives, and false negatives there were. In the situation with the Personalized Diet Recommendation System, it

shows how many food items are correctly and wrongly classified by the model as" Healthy" and" Not Healthy". Thus, this matrix is significant for the comprehension of the classification process and enhancing its precision.

5 Conclusions

In this study, we introduced a personalized diet recommendation method with machine learning by various techniques such as XGBoost, LightGBM, Random Forest, KNN, and a stacked ensemble system. We combine these models with ensembling methods, and we could predict the classification of foods to be healthy and unhealthy. The stacked ensemble method outperformed the final individual techniques and provided a more robust and accurate prediction recommendation.

The result also shows that the ensemble learning effects are beneficial for improving the quality and accuracy of personalized diet recommendations. The findings and results reveal the possibility of developing a personalized nutrition system using machine learning models and ensemble methods. Moreover, one may extend this work by adding more features or improving the ensemble method or user definition features. This work is mainly useful for personalized health care and personalized smart management systems.

References

[1] B.R, Praveen and Kumari, D. Navya Narayana and Manikanta, B. and Chandana, A. Phani and Aditya, Y. L.S, Personalized Diet Recommendation System Using Machine Learning (February 02, 2024). Available at http://dx.doi.org/10.2139/ssrn.4877349

[2] Ojokoh, B. A., & Babalola, A. E. (2016). A personalized healthy diet recommender system. *Organization for Women in Science for the Developing World (OWSD) Conference Proceedings*, 388–393. Retrieved from https://www.researchgate.net/publication/327467811_A_PERSONALIZED_HEALTHY_DIET_RECOMMENDER_SYSTEM

[3] J. -H. Kim, J. -H. Lee, J. -S. Park, Y. -H. Lee and K. -W. Rim, "Design of Diet Recommendation System for Healthcare Service Based on User Information," *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, Seoul, Korea (South), 2009, pp. 516-518, doi: 10.1109/ICCIT.2009.293.

[4] S. Gaikwad, P. Awatade, Y. Sirdeshmukh and C. Prasad, "Diet Plan and Home Exercise Recommendation system using Smart Watch," *2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI)*, Raipur, India, 2023, pp. 1-5, doi: 10.1109/ICAIIHI57871.2023.10489367.

[5] K. M. Relekar, S. Dattatray Bobalade, S. S. Mulik, S. Tukaram Kengar and R. M. Goudar, "Food Recommendation System Using K-means Clustering and Random Forest Algorithm," *2023 Global Conference on Information Technologies and Communications (GCITC)*, Bangalore, India, 2023, pp. 1-7, doi: 10.1109/GCITC60406.2023.10426098.

[6] K. Lakshmi, K. Deeba, V. Harave and S. Bharti, "Life Expectancy Prediction and Diet Recommendation System for Cardiovascular and Diabetes Disease Using Machine Learning," *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, Chikkaballapur, India, 2024, pp. 1-8, doi: 10.1109/ICKECS61492.2024.10616598.

[7] A.K. Rout, A. Sethy and N. S. Mouli, "Machine Learning Model for Awareness of Diet Recommendation," *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, 2023, pp. 96-101, doi: 10.1109/ICICT57646.2023.10133998.

[8] V. P and P. J. I. R, "Predictive Analytics Model for Heart Disease Prediction with Personalized Recommendation Systems using Machine Learning Techniques," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10723990.

[9] Yuan, Z. & Luo, F. *Personalized diet recommendation based on K-means and collaborative filtering algorithm* in *Journal of Physics: Conference Series* 1213 (2019), 032013.