International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 4364-4374
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Domain Enhanced Pre-processing for Disease-Aware Recipe Recommendation Systems

Gamini Dhiman[1], Gaurav Gupta[2], Brahmaleen K. Sidhu[3*]

[1]Research Scholar, Department of CSE, Punjabi University, Patiala.
Email: gamini.dhiman99@gmail.com- **ORCID**: 0000-0002-5247-7841

[2]Department of Computer Science & Engineering, Punjabi University, Patiala.
Email: gaurav.shakti@gmail.com- **ORCID:** 0000-0002-5247-7843

[3]Department of Computer Science & Engineering, Punjabi University, Patiala
*** Corresponding Author Email:** brahmaleen.sidhu@gmail.com- **ORCID:** 0000-0002-5247-7842

Article Info:

DOI: 10.22399/ijcesen.2844
Received : 10 March 2025
Accepted : 13 June 2025

Keywords

Data pre-processing
Machine learning
Recipe recommendation
Health informatics
Feature engineering
Disease constraints.

Abstract:

Data pre-processing is an important stage in machine learning, especially for domain-specific applications like personalised nutrition and disease-aware recommendation systems. This study clarifies a hybrid pre-processing framework for disease-specific recipe recommendation that combines general machine learning techniques (missing value handling, feature scaling, encoding, and outlier detection) with domain-specific enhancements (ingredient text normalisation, nutritional profiling, and disease-aware filtering). The suggested strategy promotes dietary compliance for illnesses like diabetes and ulcerative colitis by integrating ingredient appropriateness scores and health-related limitations. The experimental results show that these pre-processing procedures greatly increase recommendation accuracy and personalisation, lowering bias and improving the model's capacity to create health-conscious meal recommendations. This study offers important insights for health informatics, AI-driven personalised nutrition, and machine learning-based food recommendation systems, emphasising the importance of strong data pre-processing pipelines in specialised ML applications.

1. Introduction

The growing interest in AI-driven personalized nutrition has led to the development of various food recommendation systems aimed at helping users make healthier dietary choices. However, most existing models focus on general user preferences and macronutrient profiles, failing to consider specific dietary restrictions for medical conditions like diabetes, ulcerative colitis, and cardiovascular diseases. Traditional recipe recommendation systems, which use collaborative filtering (CF) and content-based filtering (CBF), suggest recipes based on previous user interactions and ingredient similarities [1] While these methods work well for general food recommendations, they lack the ability to filter ingredients at a granular level based on individual health needs. A recipe classified as "healthy" based on macronutrient values may still contain ingredients that are unsuitable for someone with a chronic illness, making it necessary to incorporate disease-aware filtering at the pre-processing stage.One of the major challenges in disease-aware food recommendations is ensuring data quality and consistency. Food datasets often contain missing values, inconsistencies in ingredient names, and outliers in nutritional values, which can distort recommendations and introduce bias in machine learning models [2]. Existing research has attempted to improve nutrition-based food recommendations by integrating high-level filtering mechanisms such as caloric constraints and macronutrient guidelines [3].

To address these challenges, this study introduces a comprehensive pre-processing pipeline that combines standard ML techniques such as handling missing data, outlier removal, feature scaling, and categorical encoding, with domain-specific enhancements like ingredient standardization, disease-aware filtering, and personalized nutrition scoring. Unlike existing models, this approach implements a novel ingredient-level disease scoring

mechanism, which ensures that recipe recommendations align with specific dietary needs for conditions such as diabetes and ulcerative colitis. Additionally, feature engineering techniques such as nutrient-based embedding, recipe complexity scoring, and taste profile modelling are applied to further enhance recommendation accuracy and personalization. By implementing this structured pre-processing strategy, the proposed system significantly improves accuracy, reliability, and personalization in disease-aware recipe recommendations. This research bridges the gap between AI-driven food recommendations and real-world dietary needs, ensuring that users receive meal suggestions that are not only tailored to their preferences but also safe and beneficial for their health.

1.1 Significance of Data Pre-processing in Machine Learning

Data preparation solves issues with data quality such missing values, inconsistencies, noise and redundancy, it directly affects how well machine learning models perform. Pre-processing guarantees:

Data Completeness: Using imputation methods like mean, median, and mode replacement, as well as more sophisticated strategies like KNN and deep learning-based imputation, to handle missing data.

Data consistency: Includes eliminating duplicate data, standardising categorical variables, and fixing inaccurate entries.

Data Relevance: To increase model interpretability, eliminate superfluous features, rank important variables, and use feature engineering.

Feature Representation: Improves the interpretability of the model by producing meaningful features.

Enhanced Performance of the Model: Normalisation, scaling, and encoding are examples of pre-processing methods that support model stability and enhance generalisation.

1.2 Data Pre-processing Constraints

Standard preparation techniques are available, but real-world datasets pose a number of difficulties that need more complex fixes. Among these difficulties are:

Handling Missing Data:A common problem in machine learning applications is missing data, which can be caused by mistakes in data collecting, insufficient records, or malfunctioning sensors. Reduced prediction accuracy and biassed models may result from the lack of key values. Strong imputation techniques are needed to tackle this problem, ranging from straightforward statistical methods (mean, median, mode) to more complex strategies like k-nearest neighbours (KNN) imputation or deep learning-based inference models.

Data Inconsistencies & Cleaning: Errors in manual entry, redundancies created by the system, or the combining of diverse data sources can all result in inconsistent datasets. Duplicate and inaccurate records may cause model learning procedures to be misguided. The dependability of datasets may be greatly increased by putting rule-based cleaning methods, automatic duplication identification, and categorical data standardisation into practice.

Outliers & Noisy Data: Variations in datasets that are unnecessary, or misleading are referred to as noisy data, and they can reduce the accuracy of models. Conversely, outliers are extreme numbers that have the potential to skew data distributions. Techniques including Z-score analysis, interquartile range (IQR) filtering, and clustering-based anomaly identification are necessary for recognising and managing these anomalies.

Increased Dimensionality: There are many characteristics in many datasets, some of which could be unnecessary or redundant. High-dimensional data raises the possibility of over-fitting models and computational complexity. Principal component analysis (PCA) and recursive feature elimination (RFE) are two feature selection methods that effectively reduce dimensionality while maintaining crucial information.

Integration of Data from Various Sources: Integration of data from several sources, such as structured databases, APIs, and sensor feeds, is a common task for machine learning systems. Achieving seamless integration is hampered by variations in data granularity, variable name practices, and schema formats. Effective integration is made possible by techniques like entity resolution, automated data transformation, and schema mapping.

Imbalanced Class in Classification Tasks: Biased model predictions result from class imbalance, which happens when certain classes in a dataset have noticeably more observations than others. This is especially troublesome when it comes to fraud detection and medical diagnosis. Techniques including cost-sensitive learning, class weight modifications in model training, and the Synthetic

Minority Over-sampling Technique (SMOTE) are used to address this issue.

Problems with Categorical Data Encoding: Categorical data must be transformed since many machine learning algorithms demand numerical input. On the other hand, incorrect encoding may result in excessive feature growth or information loss. To successfully handle these problems, entity embedding, label encoding, and one-hot encoding are frequently employed.

2. Literature Review

Data pre-processing plays a crucial role in machine learning as it directly influences the accuracy and effectiveness of models. Without proper pre-processing, raw data often contains inconsistencies, missing values, and noise, leading to unreliable predictions [2]. In fields like personalized nutrition and disease-aware food recommendations, data pre-processing becomes even more critical as it involves domain-specific constraints, such as dietary restrictions and medical guidelines. Despite significant advancements in machine learning, most recommendation systems focus on user preferences rather than health-related considerations, making them ineffective for individuals with specific medical conditions [1].

Traditional recipe recommendation systems predominantly use collaborative filtering (CF) and content-based filtering (CBF). CF-based systems recommend recipes by analyzing user interaction patterns, while CBF models focus on similarities between ingredients and user preferences. However, these techniques fail to incorporate disease-specific filtering, which is crucial for individuals with dietary restrictions [4]. A recipe might be labeled as "healthy" based on macronutrient profiles, but it could still contain ingredients unsuitable for conditions like diabetes, ulcerative colitis, or cardiovascular diseases [3].

AI-driven personalized nutrition has gained attention in recent years, with researchers attempting to integrate nutritional profiling and disease-aware filtering into food recommendation systems. Some studies have explored low-carb meal recommendations for diabetics [5]. However, most of these approaches rely on high-level filtering based on broad dietary labels rather than detailed ingredient-level suitability assessments. For example, while a system may classify a recipe as "diabetes-friendly" based on carbohydrate content, it may not account for the glycemic index of individual ingredients, leading to inaccurate recommendations.

To address these limitations, a structured pre-processing pipeline is needed to enhance the quality and reliability of disease-aware recipe recommendations. Standard data pre-processing techniques like handling missing values, outlier detection, feature scaling, and encoding ensure clean and structured data for machine learning models. However, in the context of personalized nutrition, these techniques must be complemented by health-based transformations, such as ingredient standardization, suitability scoring, and personalized filtering mechanisms [6,7]. Additionally, feature engineering methods such as nutrient-based embeddings, taste profiling, and ingredient interaction modeling can further enhance the accuracy of disease-specific meal recommendations.

Despite growing interest in AI-driven dietary personalization, there are significant gaps in current research. Most food recommendation systems prioritize user preferences over medical constraints, making them ineffective for individuals with strict dietary needs [4]. Furthermore, the lack of robust pre-processing pipelines tailored for disease-aware recommendations limits the practical application of these models in real-world healthcare and dietary planning [2].

Data Pre-Processing Techniques

Data pre-processing is a crucial step in machine learning that ensures the dataset is clean, structured, and relevant before model training. Raw data often contains missing values, inconsistencies, duplicate entries, and outliers, all of which can negatively impact model performance. Given the complexity of the datasets in this study—Recipes, User Profiles, User-Recipe Interactions, and Disease Ingredient Scoring—various pre-processing techniques were applied to improve data integrity and enhance recommendation accuracy. The following sections detail the key pre-processing steps implemented in this research.

Data Collection and Data Cleaning

Data collection is critical in creating an effective disease-specific recipe recommendation system. To provide a comprehensive and diverse collection, we gathered four crucial datasets: user profiles, recipes, user-recipe interactions, and disease-based ingredient scores. These datasets were collected from a variety of sources, including publicly available repositories, online recipe databases, medical guidelines, and user surveys. The data gathered lays a solid foundation for personalised and health-conscious

Table1. *Dataset Description*

Dataset	Description	Key Attributes
Recipes	Contains recipe details, nutritional values, and cooking instructions	Recipe ID, Recipe Name, Ingredients, Diet, Cuisine, Course, Instructions, URL, Calories, Carbs, Saturated Fats, Monosaturated Fats, Polyunsaturated Fats, Total Fats, Carbohydrates, Sugar, Fibre, Protein
User Profile	Stores demographic details, dietary preferences, and allergies	User ID, Age, Height, Weight, BMI, Allergy, Diet, Exercise Level, Preferred Fruits, Non-Preferred Fruits, Preferred Vegetables, Non-Preferred Vegetables
User-Recipe Interactions	Logs user engagement with recipes, ratings, and implicit interactions	User ID, Recipe ID, Recipe Name, Interaction Type, Interaction Sequence, Number of Views, Ingredients Viewed, Cooked Status, Timestamp, Time Spent in Minutes, Explicit Rating, Interaction Weight, Views Weight, Ingredients Weight, Time Weight, Implicit Rating
Disease Ingredient Scoring	Scores ingredient suitability for specific diseases	Ingredient Name, Diabetes Score, Ulcerative Colitis Score

meal recommendations. The collected datasets had missing values, duplicate entries, and inconsistent formats, necessitating data preparation. Missing values in user profiles were imputed using mean/mode imputation, and duplicate recipes were deleted using text similarity metrics. Furthermore, ingredient lists were standardised to ensure consistency among datasets. These pre-processing methods ensure that the collected data is reliable and usable.The datasets used in this study include recipes, user profiles, user-recipe interactions, and disease-specific ingredient scoring. Table 1 provides an overview of each dataset and its key attributes.

Handling missing data

In real-world datasets, missing data is a frequent problem that can occur for a number of reasons, including system malfunctions, human mistake, data corruption, and privacy issues. Missing data can affect the effectiveness of machine learning models, introduce bias, and lower statistical power if not handled appropriately.Missing values were present across multiple datasets, including BMI, weight, and height in user profiles, nutritional values in recipes, and interaction ratings in the user-recipe dataset. Handling these missing values properly was essential to prevent data bias.

- User Profiles: Missing BMI values were filled using the median BMI of users within the same age group to ensure realistic estimations.
- Recipe Dataset: Missing nutritional values such as calories, carbohydrates, and fat content were imputed using the mean values of similar recipes to maintain consistency.
- User-Recipe Interactions: Missing ratings were estimated using KNN-based imputation, leveraging similarities between users with similar preferences.

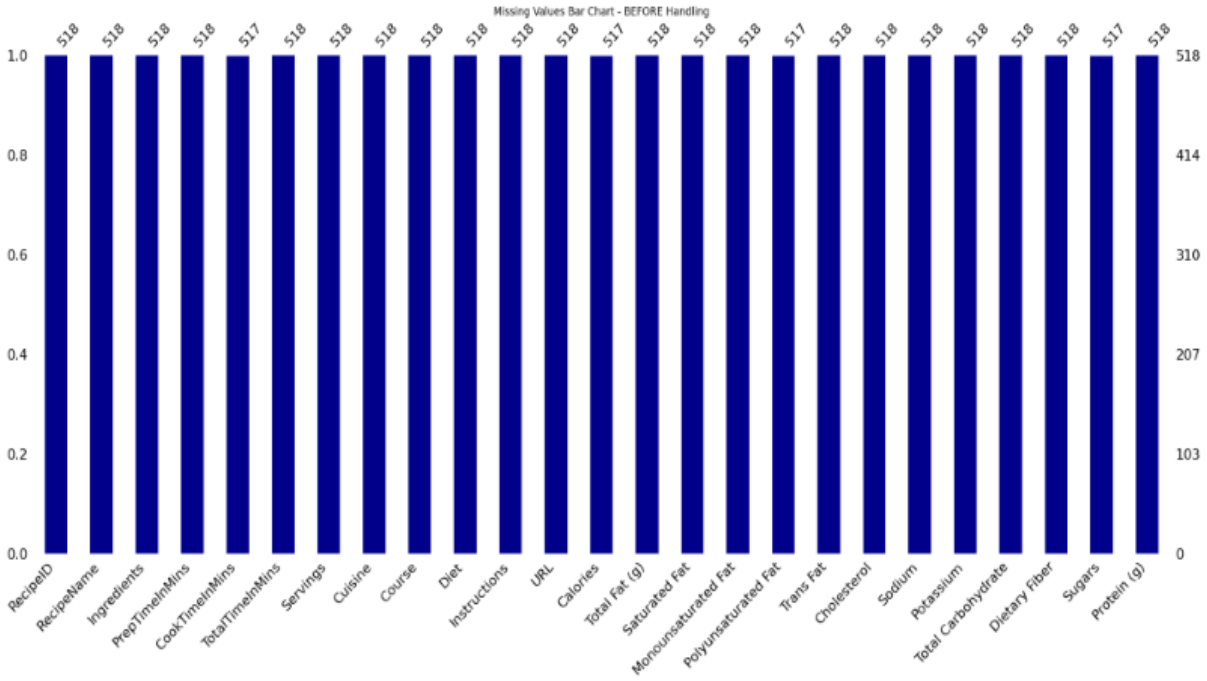

Figure 1. *Missing values before imputation handling of dataset Recipes*

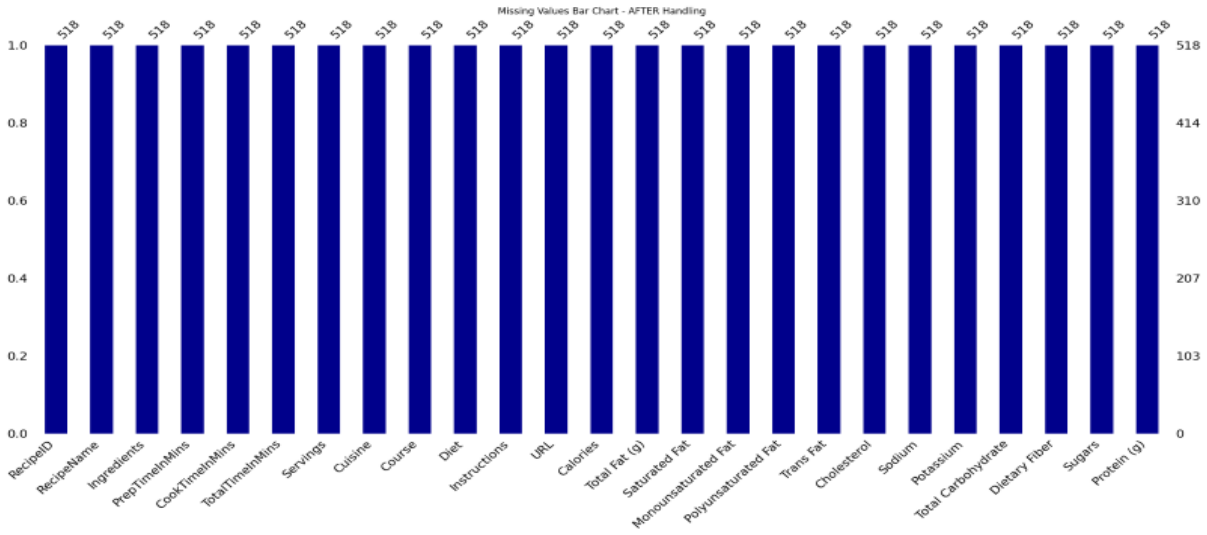

Figure 2. *Missing values after imputation handling of dataset Recipes*

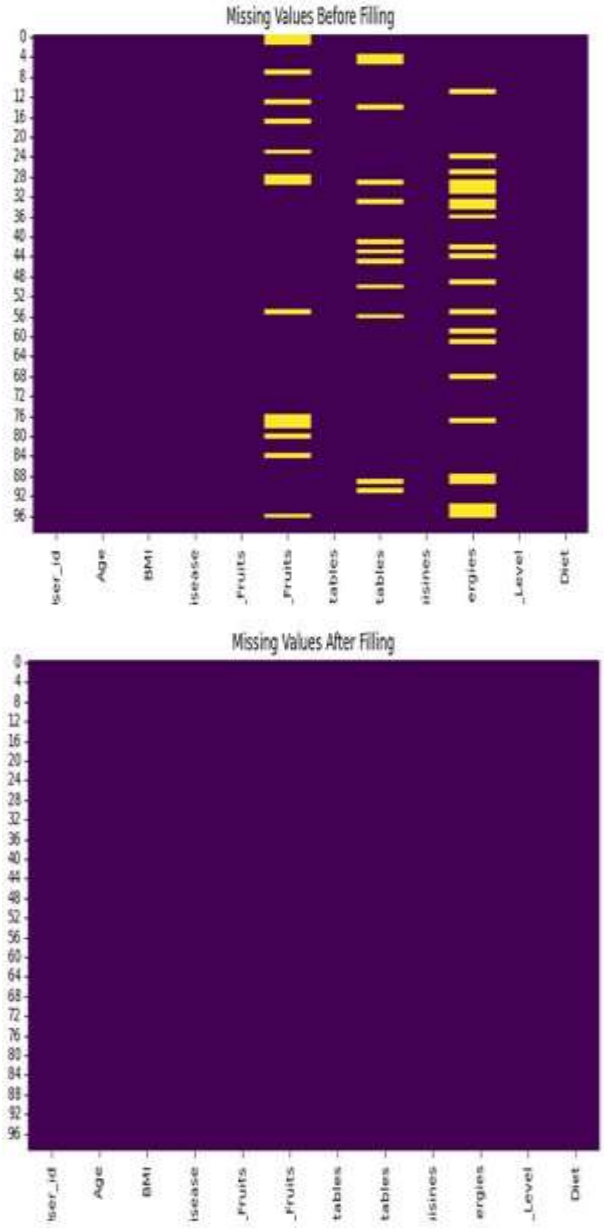

Figure 3. *Missing values before and after mean imputation handling for dataset User-Profile*

Data Cleaning and Standardization

The datasets contained **inconsistent ingredient names, unit variations, and unstructured text data** that required standardization. The following steps were applied:

- **Ingredient Standardization:** Variations in ingredient names, such as "chopped tomatoes" and "tomato," were cleaned to maintain uniformity.
- **Unit Conversion:** All nutritional measurements were standardized to a **consistent unit format (grams, kilocalories, etc.)** to avoid discrepancies.
- **Date Formatting:** Time-related data in the interactions dataset was converted into **a uniform YYYY-MM-DD HH:MM format** for consistency in model processing.

Outliers Detection and Treatment

Extreme values that substantially depart from the bulk of data points are known as outliers. They can have a detrimental impact on statistical analysis, model performance, and data integrity if not managed appropriately. Among the main justifications for managing outliers are skewed data distribution as mean variance and standard deviation makes difficult to draw meaningful insights, model's accuracy is affected as many ML algorithms like neural network and linear regression are sensitive to extreme values which results in a poor generalization, misleading statistical inferences and lastly compromising normalization and feature scaling. Detecting outliers are necessity before determining how to respond to them and the popular methods consist of Visualisation techniques, Statistical methods and Machine learning based methods. Some records had extreme BMI values, abnormally high recipe

calorie counts, and excessive user interaction times, which could introduce noise into the model.

- User Profile Outliers: Unusual BMI values were detected using box plot analysis and the interquartile range (IQR) method to identify outliers beyond the normal range.
- Recipe Dataset Outliers: Recipes with unrealistically high-calorie values were adjusted using Winsorization, a technique that caps extreme values while maintaining distribution integrity.
- Interaction Log Anomalies: Interaction durations that exceeded three standard deviations from the mean were flagged as possible noise and either removed or corrected.

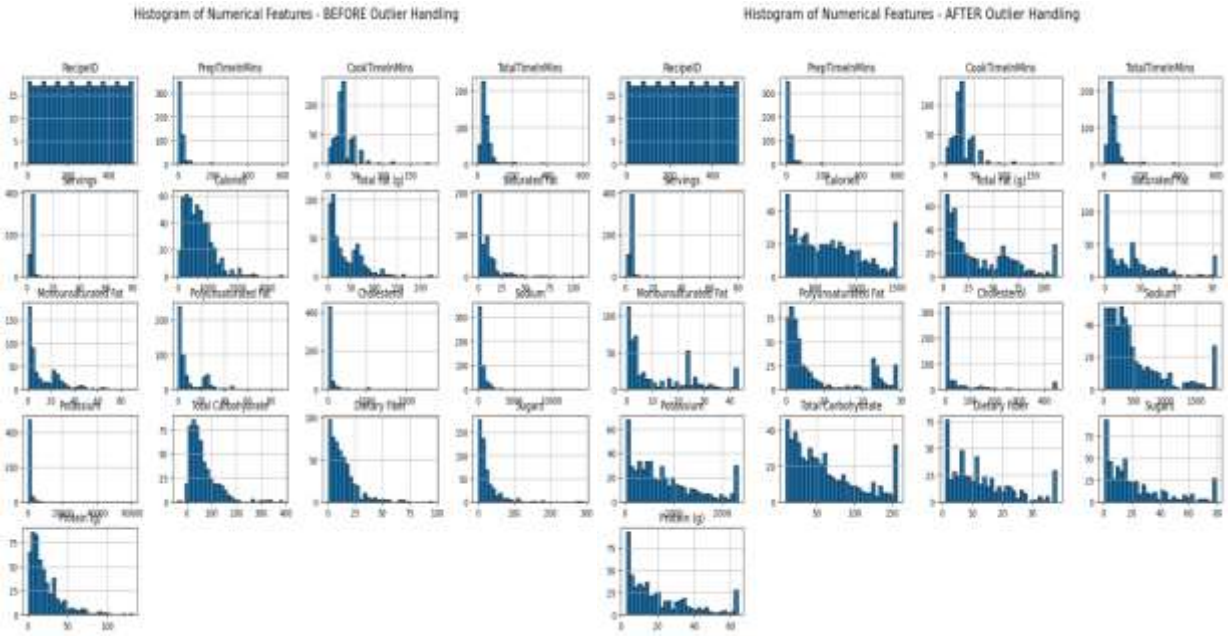

Figure 4. *Histogram for outlier detection and outlier handling using wisorization*

Feature Encoding and transformation

Feature encoding is the process of transforming categorical input into numerical format so that machine learning models can process it efficiently. Many models function with numerical values, therefore categorical variables (e.g., "red", "blue", "green" for colours or "male", "female" for gender) must be converted into numbers. There are multiple encoding techniques like one-hot encoding that converts categorical into numerical values generating new column for each category which is high dimensional data this is suitable for nominal data, label encoding assign unique numerical value to each category which is suitable for ordinal data like red:0, blue:1, green:2. Label encoding can misguide the models that red>blue. In dataset recipe the encoded columns are cuisine, course and diet using the label encoding. Ordinal encoding is similar to label encoding but it preserves the order and considering while considering the ranking. Target encoding replaces every category with the mean of the target variable as 80000 will be targeted as 80 but this prone to data leakage.

The categorical variables in the dataset were encoded accordingly:

- **One-Hot Encoding:** Applied to categorical fields such as diet type (vegetarian, non-vegetarian, eggetarian) to make them machine-readable in both user-profile and recipes dataset.
- **Label Encoding:** Used for ordered categorical features in dataset user-profile and recipes like cuisine type, exercise level which reduces dimensionality while preserving information.
- **TF-IDF Encoding:**Applied to ingredient lists in dataset recipes assigning importance scores to ingredients based on their relevance in different recipes in dataset recipes.Preferred fruits, preferred vegetable, non-preferred fruits and non-preferred vegetable in dataset user profileis encoded with TF-IDF. TF-IDF and word embedding (word2vec) is considered to evaluate the words importance for recommender system.

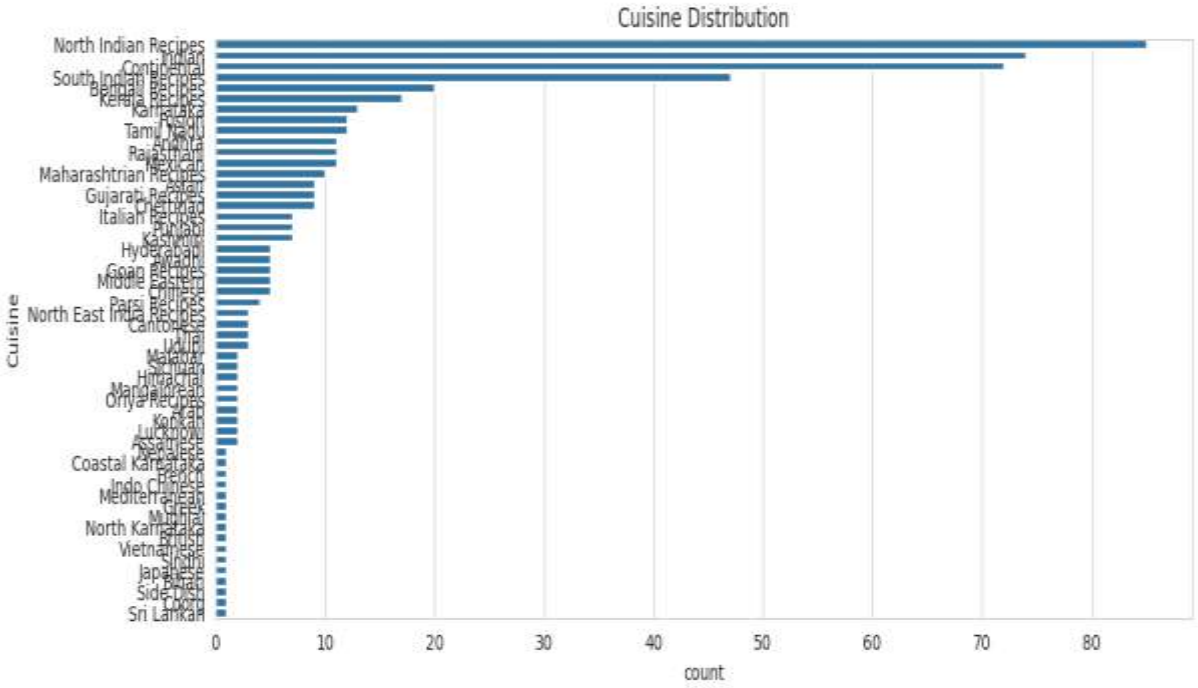

Figure 5*.Label encoding on cuisine column for datasetrecipe*

Feature Extraction

Feature extraction is a critical step in data preparation that converts raw data into structured features, hence boosting the efficiency and effectiveness of machine learning models. This method is critical in several areas, including natural language processing (NLP), computer vision, and structured data analysis. The idea is to reduce dimensionality while keeping relevant information, allowing models to learn patterns more effectively. Text data is converted into numerical representations using feature extraction approaches such as Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BoW), and word embeddings (Word2Vec, GloVe). These numerical features assist in classification, clustering, and recommendation tasks. Similarly, in structured data, statistical measurements such as mean, variance, and principal component analysis (PCA) aid in feature selection and dimension reduction. The disease based nutrition recommendation system uses the dataset recipes, dataset user and interactions where the relevant feature are extracted using word embeddings, TF-IDF.

The figure7 represents the word cloud of the ingredients and recipe is created for the frequency of the ingredients in the recipe which identifies dominating concepts, keywords and trends in the dataset.

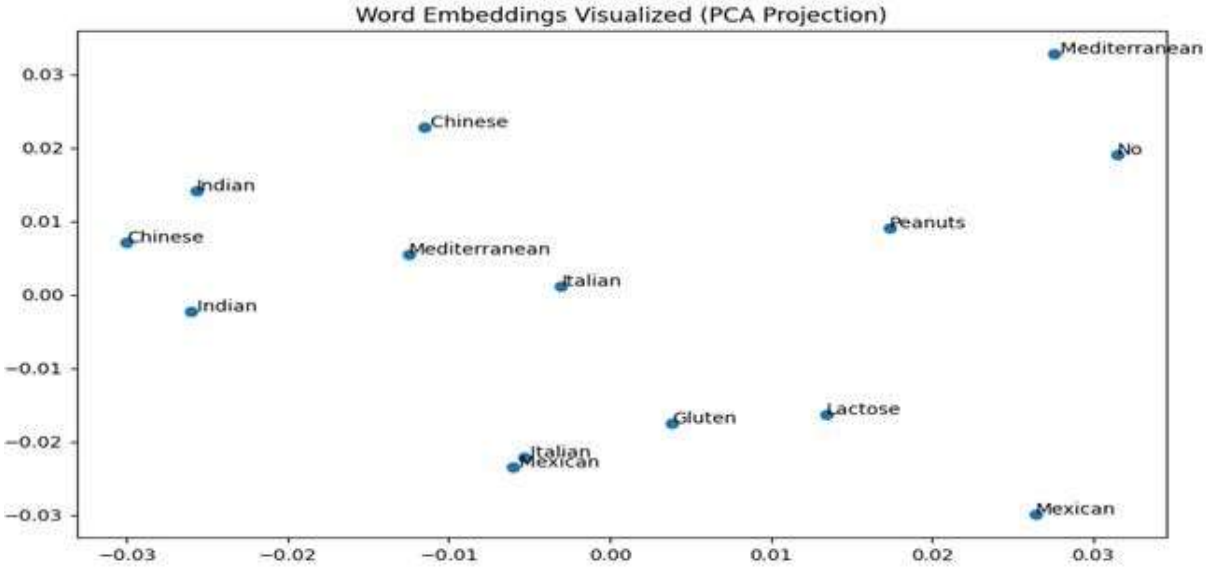

Figure 6. *Word Embedding Visualization*

Figure 7. *World cloud for Recipes Names and Ingredients*

A correlation heat map visually represents the relationship between several numerical elements in a dataset. It enables machine learning practitioners to better comprehend variable correlations, find highly correlated features, and make feature selection and engineering decisions.Correlation assesses the link between two variables. Positive correlation (+1 to 0) indicates increase in one variable with another. Negative correlation (0 to -1) indicates that as one variable increases, the other declines. No correlation (≈ 0) indicates no association between the variables.

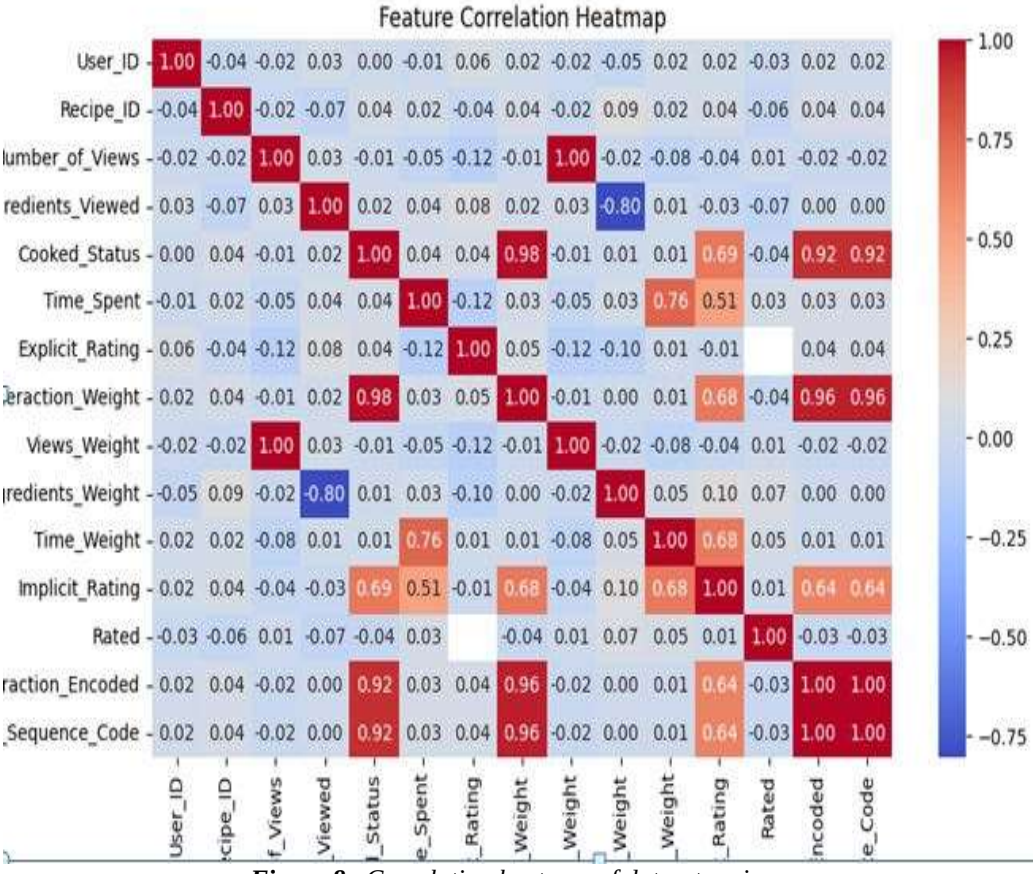

Figure 8 . *Correlation heatmap of dataset recipes*

Feature Scaling and Normalization

Since different datasets contained variables in varying ranges, scaling was essential for **improving model stability and preventing bias**.

- **Min-Max Scaling:** Used for **calories, fat content, carbohydrates, and protein levels** to scale values between 0 and 1, preventing dominant features from skewing predictions.
- **Z-Score Standardization:** Applied to **BMI and exercise levels**, ensuring a normal distribution of user data.
- **Log Transformation:** Used for **time spent on recipe interactions**, as it followed a skewed distribution, making it more interpretable for the model.

Feature Engineering

Feature engineering converts raw data into meaningful features that enhances machine learning model performance.It entails choosing, developing, and altering features to better describe the underlying patterns in the data. To further refine recommendation accuracy, additional features were engineered:

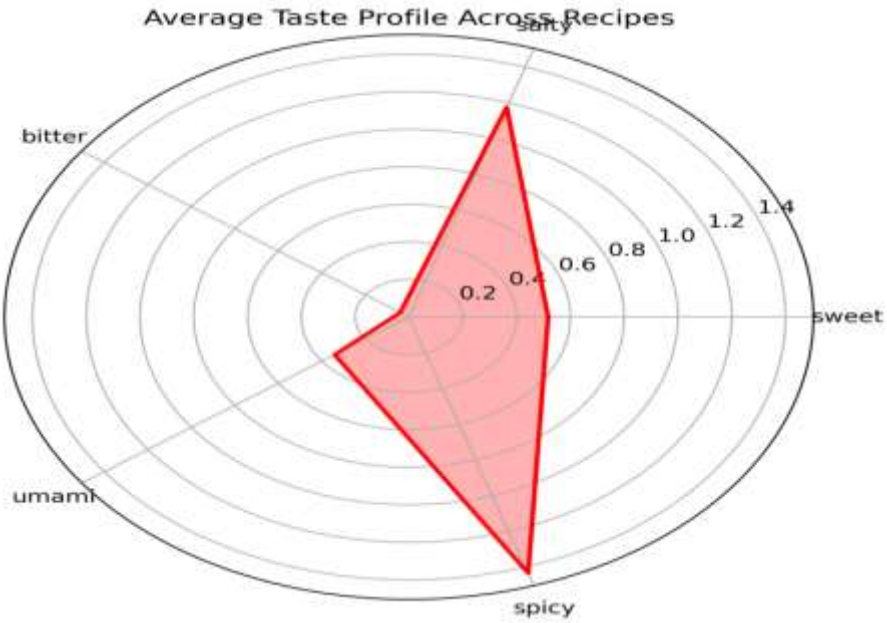

Figure 9. *Radar chart depicting the average intensity of different taste profiles (sweet, bitter, umami, spicy) across all recipes in the dataset recipes*

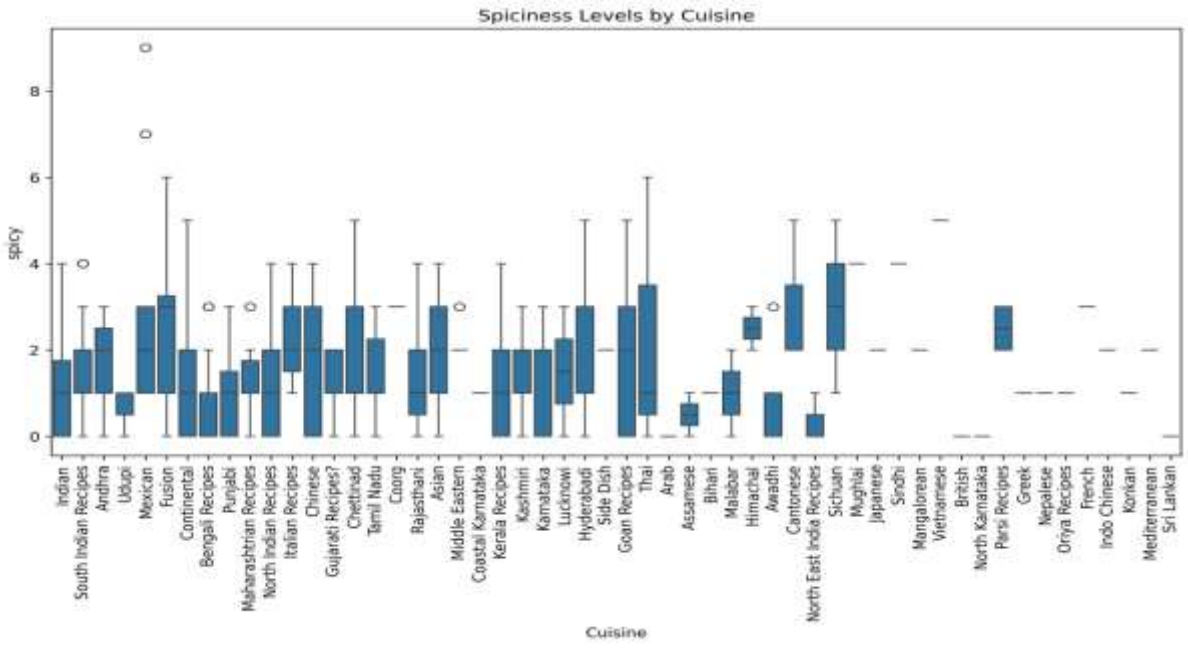

Figure 9. *Box plot fortaste based feature spiciness of cuisines from the dataset recipes*

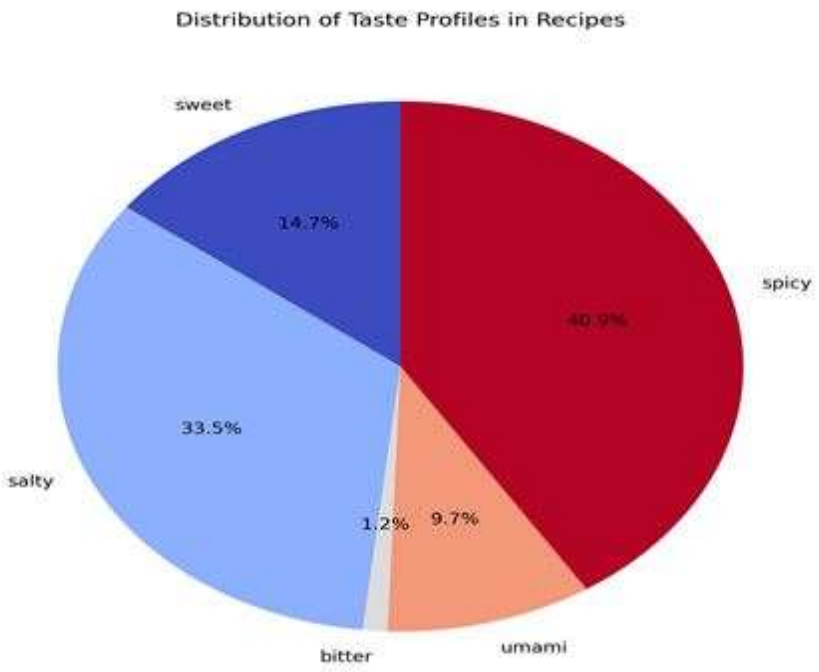

Figure 10. *Distribution of taste profiles in dataset Recipes*

- **BMI-Based Meal Modifications:** Users were grouped into underweight, normal weight, and overweight categories, allowing portion sizes and meal suggestions to be tailored accordingly.
- **Taste Profiles:** Recipes were assigned spiciness, sweetness, bitterness, saltiness, and umami ratings, helping to match meals with user taste preferences.
- **Recipe Complexity Score:** A difficulty score was assigned to each recipe based on ingredient count and cooking steps, helping users choose meals that fit their cooking skills.

Disease-Aware Filtering and Ingredient Suitability Scoring

Unlike generic recipe recommendations, this system required **disease-aware filtering**to ensure that meals matched specific medical conditions.

- **Ingredient Suitability Scores:** Each ingredient was assigned a **score from 0 to 5** based on its impact on **diabetes and ulcerative colitis**, using expert dietary guidelines.
- **Recipe Filtering:** Any recipe containing **high-risk ingredients for a specific disease** was flagged and excluded from recommendations for affected users.
- **Personalized Adjustments:** The recommendation model balanced **user preferenceswithdisease-specific constraints** to optimize meal suggestions.

This pre-processing step allowed the system to generate **medically appropriate meal recommendations**for individuals with dietary restrictions.

Conclusion

Ensuring data quality is essential for building an effective disease-aware recipe recommendation system. In this study, a structured approach was implemented to refine the dataset, addressing inconsistencies and optimizing feature representation. By incorporating advanced pre-processing techniques, the system effectively enhances recommendation accuracy while maintaining personalization and adherence to medical constraints. Additionally, feature engineering methods, such as BMI-based customization and taste profiling, contribute to a more user-centric experience by aligning meal suggestions with both health requirements and individual food preferences. This refined data pipeline not only improves the reliability of predictions but also ensures that dietary recommendations remain practical, relevant, and beneficial for users managing specific medical conditions.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.

- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Elsweiler, D., Trattner, C., & Harvey, M. (2017). Exploiting food choice biases for healthier recipe recommendation. SIGIR '17: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 575-584.*

[2] García, S., Luengo, J., & Herrera, F. (2015). Data Preprocessing in Data Mining. *Springer.*

[3] Ge, M., Elahi, M., & Ricci, F. (2020). Beyond accuracy: Evaluating recommender systems by healthfulness. *Health Informatics Journal*, 26(1), 1-12.

[4] Trattner, C., &Elsweiler, D. (2019). Investigating health aspects in recipe recommendation. *User Modeling and User-Adapted Interaction*, 29(3), 601-633.

[5] Wang, L., Chen, H., & Zhao, J. (2021). AI-based diet personalization for diabetes patients. *Healthcare Technology Letters*, 8(2), 50-59.

[6] Fan, C., Chen, M., Wang, X., Wang, J., &Bufu, H. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research.*

[7] Razzaq, M., Maqbool, F., Ilyas, M., &Jabeen, H. (2023). EvoRecipes: A generative approach for evolving context-aware recipes. *IEEE Access.*