SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Information System Exploration and Research

https://shmpublisher.com/index.php/joiser

p-ISSN 2964-1160 | e-ISSN 2963-6361

Machine Learning Techniques for Classifying Indonesian Foods and Drinks by Nutritional Profiles

Bagus Al Qohar[1*], Yulizchia Malika Pinkan Tanga[2], Aditya Yoga Darmawan[3]
[1,2,3]*Department of Computer Science, Universitas Negeri Semarang, Indonesia*

DOI: https://doi.org/10.52465/joiser.v3i1.528
Received 07 January 2025; Accepted 29 January 2025; Available online 29 January 2025

Article Info

Keywords:
Dietary analysis;
Food categorization;
Indonesian food;
Machine learning;
Nutritional profiling

Abstract

Local ingredients and Indonesia's diverse culinary traditions play an important role in shaping people's health and eating habits. Understanding the nutritional profile of Indonesian food is crucial to promoting healthier food choices. This study aims to classify Indonesian food and beverages based on their nutritional content, with a focus on calories, protein, fat, and carbohydrates. To achieve this, a dataset of 1,346 food items was preprocessed using normalization techniques to improve model performance. Each food item was categorized as High Protein, High Fat, or High Carbohydrate based on its dominant macronutrient content. Five machine learning models which are K-Nearest Neighbors, Decision Trees, Support Vector Machines, Random Forest, and Multilayer Perceptron-were used and compared. Among these models, the Support Vector Machine achieved the highest classification accuracy of 99.1%. These findings demonstrate the potential of machine learning in nutrition research, providing a basis for developing data-driven dietary recommendations tailored to individual nutritional needs. This research bridges traditional dietary research with modern computational approaches, offering insights for public health initiatives and personalized nutrition planning.

1. Introduction

Indonesia's geographical diversity and rich cultural heritage are reflected in its cuisine, which varies greatly across regions. Traditional Indonesian foods and beverages not only offer distinctive flavors but also have unique nutritional compositions, shaped by centuries-old cooking methods and the use of local ingredients. For example, the diverse types of soto in Indonesia illustrate the complexity of analyzing and classifying its nutritional aspects [1]. As global awareness of nutrition and health increases, accurately assessing and classifying traditional foods is becoming increasingly important for dietary research and public health initiatives. Understanding food and nutrient intake is critical to understanding how dietary habits affect public health outcomes. However, traditional dietary

* ***Corresponding Author:***

Bagus Al-Qohar,
Department of Computer Science,
Universitas Negeri Semarang,
Semarang, Indonesia.
Email: bagusximipa6@students.unnes.ac.id

assessment methods often face limitations, especially in culturally diverse countries such as Indonesia [2]. The advent of machine learning offers a promising solution by enabling more precise and efficient analysis of complex food data, thus supporting evolving food trends influenced by both modern and traditional cuisines.

Various machine learning models have demonstrated their ability to classify foods based on nutritional characteristics. K-Nearest Neighbors (KNN), for example, has been effectively used to categorize foods with similar nutritional content, providing an easy and efficient classification method [3]. Despite its simplicity, KNN's reliance on labeled data and sensitivity to data distribution patterns may limit its performance on more complex data sets that require higher accuracy. Decision Tree has also been used as a decision support system in food categorization and other fields such as medicine, thanks to its interpretability and hierarchical structure [4]. In the context of Indonesian cuisine, the Decision Tree facilitates the identification of key nutritional elements that differentiate various food categories, providing valuable insights. However, its susceptibility to overfitting necessitates careful tuning and the use of ensemble techniques to improve performance.

Support Vector Machines (SVM) are another popular choice for high-dimensional classification tasks. SVMs excel at distinguishing sophisticated dietary profiles by constructing a hyperplane that maximizes the margin between classes [5]. However, its performance can be significantly affected by the choice of kernel and class balance, especially when dealing with unbalanced datasets such as Indonesian traditional and modern foods. Random Forest classifiers address this challenge by effectively handling noisy and imbalanced data. By utilizing features such as color and texture, the Random Forest model achieves high accuracy in identifying traditional Indonesian foods [6]. Its ensemble nature helps to reduce overfitting and improve generalization, making it suitable for large and complex food datasets. In addition, the feature importance ranking of Random Forest provides deeper insight into the most influential nutritional factors for classification.

Advanced neural network methods, particularly Multilayer Perceptron (MLP), are well suited to capture complex patterns in nutritional data from dietary supplements, foods, and beverages. MLP can identify complex interactions between nutrients, achieving high accuracy in classifying traditional foods with diverse nutritional profiles [7]. However, the application of MLPs in nutrition research is often constrained by the need for large computational resources and large data sets. A comprehensive framework for analyzing Indonesian food and beverages can be created by combining various machine learning techniques, such as KNN, Decision Trees, SVM, Random Forest, and MLP. Hybrid models and ensemble methods leverage the strengths of each classifier, overcoming their respective limitations and achieving a balanced trade-off between interpretability and accuracy [8]. This integration advances food informatics and supports data-driven public health policies in Indonesia.

The incorporation of machine learning into nutrition studies is in line with recent advances in intelligent food control systems. Multisource data mining techniques can be adapted to classify and manage vast nutrition databases covering a wide variety of foods, including Indonesian specialties [9]. This approach not only improves food classification but also facilitates real-time monitoring and prediction of nutritional outcomes, paving the way for future research. Unlike previous studies that focused on individual models, this research contributes to the growing body of knowledge by providing a systematic approach to classifying Indonesian foods and beverages using multiple machine-learning techniques. By comparing several algorithms on a comprehensive food dataset, this research offers new insights into model performance and practical applications, such as better diet assessment and customized nutrition recommendations for Indonesians [10].

2. Literature Review

The main objective of this research is to classify Indonesian food and beverages based on their nutritional content, specifically focusing on calories, protein, fat, and carbohydrates. By using various machine learning techniques, including K-Nearest Neighbors, Decision Trees, Support Vector Machines, Random Forest, and Multilayer Perceptron (MLP), this research aims to improve the accuracy of diet classification. This classification not only provides insights into the nutritional profiles of local ingredients but also supports the development of data-driven diet recommendations tailored to the Indonesian population.

Previous research has been conducted by Asmara et al. (2020) [11] with the main objective being the creation of a recommendation system for Herbalife Nutrition products made possible through the utilization of Mamdani Fuzzy Logic. This research is capable of effectively managing uncertainty in

decision-making by utilizing fuzzy logic, which ultimately results in increased customer satisfaction and an optimized recommendation process. Another research was conducted by Sulistiani et al. (2020) [12]. The main objective of this research is to determine the optimal nutrition pattern that most supports toddler growth. PSO is used to tackle difficult optimization challenges, making this achievable. As a result of these findings, the practical application of PSO in the process of refining diet planning. Research conducted by Rong (2024) [13] presents a study that optimizes student diet recipes by utilizing the Entropy Weight TOPSIS method along with the Simulated Annealing Algorithm. The objective of this study is to investigate and classify various dietary options. The methods used successfully optimized optimization across various population groups, as evidenced by the fact that the combination of these methods was effective in determining which recipes were most suitable. In addition, Siagian et al. (2024) [14] developed a food menu recommendation system specifically tailored to meet the dietary needs of the Indonesian population. This system can integrate cultural food practices and existing local nutritional standards.

In this paper, we investigate the classification of Indonesian food and beverages based on their nutritional profiles using machine learning techniques. The selection of these models is justified by their ability to accurately classify various types of food by capturing the interactions between essential macronutrients: calories, protein, fat, and carbohydrates. Machine learning models including K-Nearest Neighbors, Decision Trees, Support Vector Machines, Random Forest, and Multilayer Perceptron can process large datasets with many variables, offering a more scalable and efficient method for nutritional analysis compared to conventional manual categorization techniques. A standard scaler for data normalization helps improve classification accuracy, which is crucial for handling challenging and varied datasets. This research provides a methodical and data-driven approach to classifying Indonesian food, offering valuable information that can serve as a guide for appropriate dietary recommendations and guidelines for the Indonesian population. This knowledge will help improve public health policies and create more tailored nutrition plans.

3. Method

To enable effective classification, this study handled and investigated the Indonesian food and beverage dataset including nutritional values such as calories, protein, fat, and carbohydrates. Food type classification using five distinct machine learning models: K-Nearest Neighbors (KNN), Decision Trees, Support Vector Machine (SVM), Random Forest, and Multilayer Perceptron (MLP). Every model was chosen for its particular benefits in problems of classification. SVM excels in high-dimensional spaces, for instance, while Random Forest resists overfitting in big datasets. Its performance was assessed using the evaluation matrix including precision, accuracy, recall, and F1 score following hyperparameter modification using randomized search to acquire the best configurations for every model. Figure 1 gives a general picture of the methodical approach applied in this study, so summarizing the whole technique.

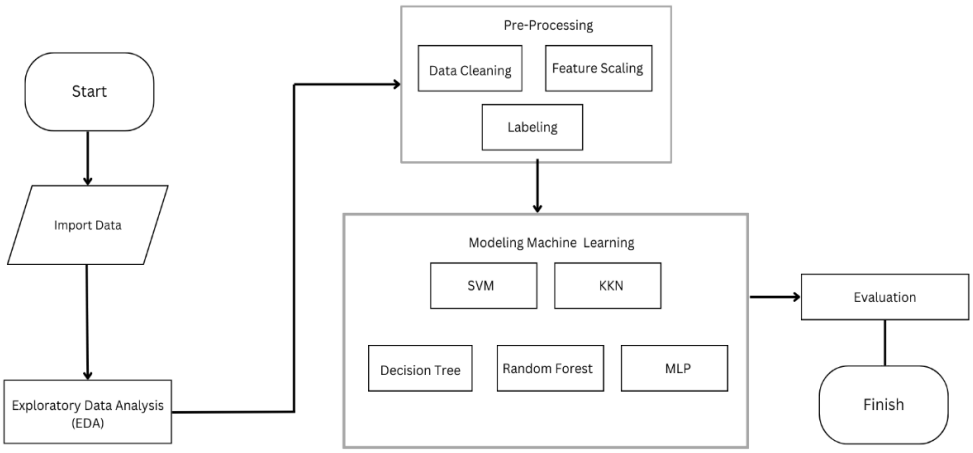

Figure 1. Flowchart of the research method

3.1. Import Data

This study uses nutritional information from Indonesian foods and drink nutrition dataset [15]. With information on calories, proteins, fats, and carbohydrates, the 1,346 entries in the dataset reflect different types of foods. Every entry in the dataset also features a link to an image of the item and a distinct identity. Preprocessing techniques are used to solve data quality concerns including handling missing values and normalizing the numerical nutritional data, so preparing this dataset for machine learning classification. Table 1 shows a summary of the Indonesian food and drink nutrition dataset.

Table 1. Table Summary of Indonesian Food and Drink Nutrition Dataset

id	calories	proteins	fat	carbohydrate	name	image
1	280.0	9.2	28.4	0.0	Abon	https://imgcdn.medkomtek.com/PbrY9X3ignQ8sVuj...
2	513.0	23.7	37.9	21.3	Abon haruwan	https://img-global.cpcdn.com/recipes/cbf330fbd...
3	0.0	0.0	0.2	0.0	Agar-agar	https://res.cloudinary.com/dk0z4ums3/image/upl...
4	45.0	1.1	0.4	10.8	Akar tonjong segar	https://images.tokopedia.net/img/cache/200-squ...
5	37.0	4.4	0.5	3.8	Aletoge segar	https://nilaigizi.com/assets/images/produk/pro...
…	…	…	…	…	…	…
1346	52.0	3.3	2.5	4.0	Yoghurt	https://d1vbn70lmn1nqe.cloudfront.net/prod/wp-...

3.2. Exploratory Data Analysis

Exploratory data analysis (EDA) is an important first step in research, as it helps provide closer knowledge of the structure and features of a dataset. Particularly in clinical settings, Carnevale et al [10] showed how EDA supports feature selection in challenging data sets. Their work on optimizing pediatric laparoscopy shows how EDA can reveal important trends, anomalies, and relationships in the data, thus guiding improved feature selection for better model results. Similarly, Wang et al. [16] used EDA techniques for geoexploration data, thus further highlighting the importance of EDA in uncovering the underlying data structure and therefore enabling discovery in large multidimensional datasets. These cases show how the role of EDA in feature identification precisely shapes data-driven outcomes in many scientific domains.

EDA is a crucial step in the research process, particularly in the context of dietary classification. EDA involves analyzing datasets to summarize their main characteristics, often using visual methods. In this study, EDA played a significant role in feature selection by helping to identify important trends, anomalies, and relationships within the nutritional data of Indonesian foods. By employing EDA techniques, researchers were able to uncover the underlying structure of the dataset, which informed the selection of relevant features for the machine learning models. This process not only enhanced the accuracy of the models but also ensured that the selected features were representative of the nutritional characteristics of the foods being classified.

3.3. Data Preprocessing

The preprocessing stage of this study was crucial to ensure that the data was clean, consistent, and ready for classification. This stage involved multiple steps, including data cleaning, feature scaling, and labeling, which together aimed to improve the accuracy and reliability of the machine learning models used to classify Indonesian foods. Each preprocessing step was selected and executed based on the unique structure and characteristics of the dataset, and it helped to standardize and prepare the data for effective analysis. Table 2 shows the description for each preprocessing step.

Table 2. Table Data Preprocessing for Indonesian Food and Drink Nutrition Dataset

Step	Description	Purpose	Method/Tool
Data Cleaning [17]	Delete rows with missing values or unused data	To ensure that each entry is complete and reduce bias, preventing potential errors and inconsistencies during model training.	Pandas: dropna()
Feature Scaling [18]	Normalize variables such as calories, protein, fat, and carbohydrates on a comparable scale	To prevent features with larger data from dominating the classification	Scikit-learn: Standardscaler()
Labeling [19]	Set target variables based on the dominant macronutrients in each food: High-Protein, High-Fat, or High- Carbohydrate.	To categorize each food based on nutrients, facilitating macronutrient analysis and creating multiclass target variables for supervised learning.	Custom function: labeling_foods()

3.4. Modeling Machine Learning

In this study, five distinct machine-learning models were selected for their unique strengths in handling classification tasks related to Indonesian foods and beverages. Each model was implemented with careful consideration of its hyperparameters to optimize performance. Overall, the methodology employed a systematic approach to model selection and hyperparameter optimization, ensuring that each machine learning model was tailored to effectively classify the diverse Indonesian food dataset based on nutritional profiles.

3. 4.1. KNN

K-Nearest Neighbors (KNN) classifier's training process starts with careful hyperparameter selection and preparation meant to maximize its predictive performance. RandomizedSearchCV helps one to accomplish this using a quick search over a specified range of hyperparameter values. The KNN model's fundamental hyperparameters are distance metric (metric), weighting scheme (weights), and number of neighbors (n_neighbors). Defining these boundaries inside a search grid helps the model to be evaluated with several combinations, so improving its capacity to generalize to fresh data. 5-fold cross-validation guarantees that every combination is tested in several subsets of the training data, so offering a strong estimate of the performance of the model. Following the random search, the model chooses the hyperparameters producing the best cross-validation accuracy. These ideal values then are used to train the last model on the whole training set, so ensuring that it makes use of the best configuration discovered during the search. This careful approach not only optimizes the KNN classifier to fit the particular features of the dataset but also reduces overfitting risk. Automating hyperparameter tuning with RandomizedSearchCV helps the model training page to emphasize the need to use cross-validation and randomized searches to produce a dependable, strong predictive model. The KNN classifier was chosen for its simplicity and effectiveness in classification tasks [20]. The training process involved hyperparameter tuning using RandomizedSearchCV, which allowed for a quick search over various hyperparameter values, including the distance metric, weighting scheme, and the number of neighbors. This method ensures that the model is well-suited to the dataset's characteristics, reducing the risk of overfitting while enhancing generalization to new data. The model's performance was validated through 5-fold cross-validation, ensuring robust accuracy estimates.

3.4.2. Decision Tree

The decision tree model's training is methodically hyperparameter-tuned to maximize predictive accuracy. RandomizedSearchCV is applied in this process to effectively investigate a variety of hyperparameter settings. Combining 5-fold cross-valuation with a randomized search across these parameter values helps the model find the best settings that maximize high accuracy while preserving generalizing capability to new data. Utilizing cross-validation, every set of hyperparameters is tested for consistency over several data splits, so enhancing the model's dependability. From the hyperparameter identification, re-training the Decision Tree classifier on the entire training set using the optimal configuration follows. This retraining phase ensures that the model catches all accessible training data patterns by using the tuned parameters to avoid problems including underfitting and

overfitting. Utilizing RandomizedSearchCV, automating the search for the ideal model configuration enables this process to highlight a successful approach to generate and improve decision trees, thus generating a more accurate and stronger predictive model [21]. Although not detailed extensively in the contexts, Decision Trees were also part of the classification process, providing a straightforward method for interpreting the classification results based on the nutritional features of the foods.

3.4.3. Support Vector Machine (SVM)

The support Vector Machine (SVM) classifier's training process consists of hyperparameter fine-tuning to reach the best performance. This method investigates a predefined set of hyperparameters using RandomizedSearchCV, including the regularization parameter C, several forms of the kernel (linear, rbf, poly), the kernel coefficient gamma, and the polyn degree (specific to the poly kernel). Starting the SVM model using random states helps to produce consistent results. RandomizedSearchCV tests 10 random combinations of these hyperparameters over the search using 5-fold cross-validation to identify the configuration optimizing model accuracy. This cross-valuation approach helps to avoid overfitting by enabling the guarantee of consistent model performance over several data splits. Once discovered, the best hyperparameters print alongside the maximum cross-validation accuracy. The last SVM model is then trained on these optimal conditions using the complete training dataset. This last training phase uses the perfect configuration to increase the SVM's capacity for effective generalizing to new data. This methodical approach of hyperparameter tuning with RandomizedSearchCV shows how efficiently data scientists could maximize SVM models, thus improving prediction accuracy and resilience. SVM was selected for its ability to excel in high-dimensional spaces. The model was configured with a linear kernel, and hyperparameters such as gamma, degree, and C values were optimized [22]. The SVM model achieved high accuracy demonstrating its effectiveness in accurately identifying true instances within the dataset.

3.4.4. Random Forest

RandomizedSearchCV is used in the model training process for a Random Forest classifier to maximize hyperparameters and improve the model's predictive capacity. A strong ensemble learning technique, Random Forest gains much from careful hyperparameter tuning to balance bias and variance. RandomizedSearchCV guarantees that a large spectrum of model configurations is evaluated effectively by setting these parameters and running 10 iterations of randomized testing using 5-fold cross-valuation. This cross-valuation method aids in the identification of the most optimal hyperparameter combination maximizing accuracy and encouraging generalizability. Once the random search is over, the hyperparameters with the highest performance are printed together with their matching cross-valuation accuracy. The last model is then re-trained on the whole training set using these ideal hyperparameters, arming it to generate strong predictions. Driven by the power of RandomizedSearchCV, this all-encompassing training method underlines how methodical hyperparameter tuning can significantly improve model performance, so ensuring that the Random Forest classifier generates consistent and accurate results in many data settings. This ensemble learning method was included for its robustness against overfitting, especially in large datasets [23]. The model's hyperparameters, such as the number of estimators and maximum depth, were fine-tuned to enhance its predictive performance.

3.4.5. Multilayer Perceptron (MLP)

Optimizing several hyperparameters helps a Multi-Layer Perceptron (MLP) classifier to enhance its performance. Randomized Search CV is applied in this process to do an extensive search across a grid of possible hyperparameters. These comprise the hidden layer sizes, the activation function applied inside the hidden layers, and the weight update optimization technique (solver). Furthermore, taken under consideration are the learning rate schedule (learning_rate) which regulates how the learning rate changes during training, and the regularization term (alpha), which penalizes significant weights and helps prevent overfitting. Defining the maximum number of training iterations, the max_iter value guarantees sufficient chances for convergence of the model. Using RandomizedSearchCV which executes 10 randomized iterations under 5-fold cross-validation the optimal hyperparameter combination is chosen depending on accuracy. This method improves the predictive power and dependability of the model by helping to find configurations that generalize well in many data splits [24]. Re-training the final MLP model on the whole training set with the ideal parameters helps to attain the best possible predictive performance. The MLP model was implemented to capture complex,

nonlinear relationships within the data. Hyperparameters such as the learning rate schedule and regularization term were optimized using RandomizedSearchCV, which executed multiple iterations under cross-validation. This approach ensured that the final model was trained on the entire dataset with the best-found parameters.

3.5. Evaluation

The evaluation model uses the confusion metric, with the following composition, F1 score, precision, recall, and precision. The focus of this performance analysis is on the accuracy produced by the proposed model. When analyzing the results using a confusion matrix, there is a mathematical formula for all metrics used.

Accuracy measures the proportion of correctly predicted instances out of the total number of predictions. It is useful for evaluating the overall performance of a model but may not be reliable for imbalanced datasets [25]. The formula of accuracy shown in equation (1).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (1)$$

Precision is the ratio of true positive predictions to the total predicted positives. It reflects the relevance of positive predictions and is crucial in scenarios where false positives have high costs, such as spam detection [25]. The formula of precision shown in equation (2).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall, also known as sensitivity, measures the proportion of actual positives that are correctly identified by the model. It is important in scenarios where missing a positive instance (false negative) carries significant consequences, such as disease detection [25]. The formula of recall shown in equation (3).

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

The F1 score is the harmonic mean of precision and recall, providing a balanced measure when both metrics are crucial. It is particularly effective for imbalanced datasets [25]. The formula of the F1 score is shown in equation (4).

$$F1\ Score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \quad (4)$$

Description:
TN (True Negative): True data that is categorized as negative
TP (True Positive): True data that is categorized as positive.
FP (False Positive): False data that is categorized as positive
FN (False Negative): False data that is categorized as negative

4. Results and Discussion

4.1. Results of Exploratory Data Analysis (EDA)

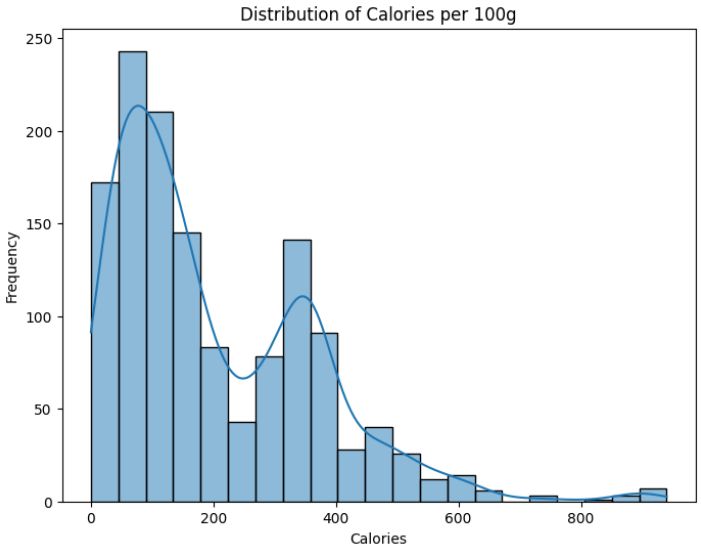

Figure 2. Distribution of Calories

Figure 2 illustrates the distribution of calories per 100 grams across a dataset, providing insights into the caloric density of various food items. The x-axis represents the calorie values, while the y-axis indicates the frequency of items within a particular calorie range. A smooth density curve is superimposed over the histogram to highlight the overall shape and distribution of the data. The distribution appears to be right-skewed, with a majority of items concentrated in the lower calorie ranges, particularly between 0 and 200 calories. This suggests that most food items in the dataset are low in caloric density. The presence of secondary peaks around 400 calories per 100 grams indicates a potential subcategory of food items with higher calorie content. Beyond 600 calories, the frequency sharply declines, indicating that high-calorie items are relatively rare in the dataset. This distribution could be useful for researchers studying dietary patterns or analyzing food choices, as it provides an understanding of the prevalence of high- and low-calorie foods, helping to inform nutritional guidelines and public health recommendations.

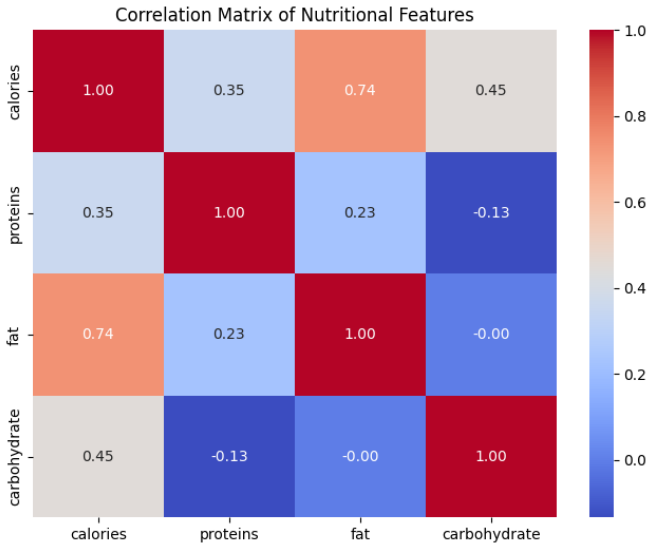

Figure 3. Correlation Matrix

Figure 3 visualizes the relationships between different nutritional features: calories, proteins, fat, and carbohydrates. The color scale represents the strength and direction of the correlation, ranging

from -1 (strong negative correlation) to 1 (strong positive correlation). Red tones indicate positive correlations, while blue tones indicate negative or weak correlations. For instance, calories have a strong positive correlation with fat (0.74) and a moderate correlation with carbohydrates (0.45), suggesting that foods higher in calories are often rich in fats and, to a lesser extent, carbohydrates. Proteins, however, show weaker correlations with other features. There is a moderate positive correlation between proteins and calories (0.35) and a weak positive relationship with fat (0.23), while the correlation with carbohydrates is slightly negative (-0.13). This indicates that protein content does not strongly align with other macronutrient levels in the dataset. Such findings can aid researchers and dietitians in understanding how macronutrients contribute to the caloric density of foods and their interdependence, providing valuable insights for dietary planning and nutritional analysis.

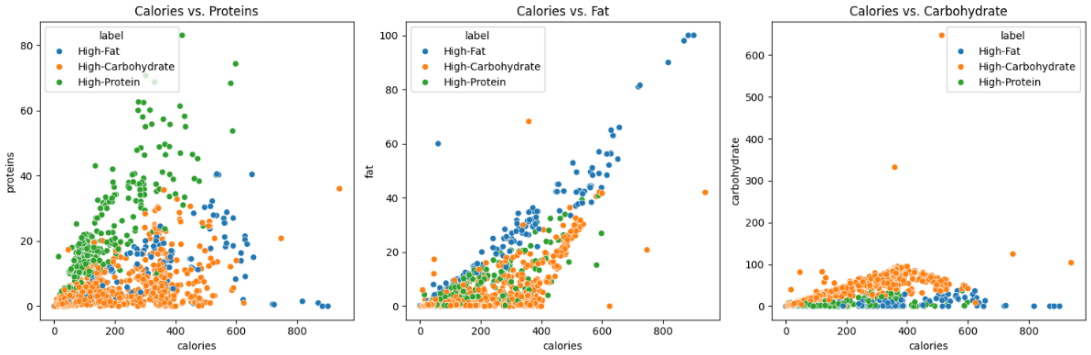

Figure 4. Scatter Plot Relantionship

This set of scatter plots displays the relationships between calories and three nutritional components: proteins, fat, and carbohydrates, with points color-coded by category (High-Fat, High-Carbohydrate, and High-Protein). The first plot (Calories vs. Proteins) indicates that higher-protein foods generally span a range of calorie values, with many high-protein items (green points) clustering in lower to moderate calorie ranges. This suggests that high-protein foods may not always be calorie-dense, contrasting with other macronutrient categories. In the second plot (Calories vs. Fat), there is a strong positive relationship, as expected, between calories and fat, with high-fat items (blue points) dominating at higher calorie levels. The third plot (Calories vs. Carbohydrates) shows that high-carbohydrate items (orange points) are widely distributed, but their calorie contribution does not follow as clear a pattern as fat. These visualizations provide insights into how the balance of macronutrients contributes to the caloric content of food, helping in dietary classification and nutritional profiling for different diet categories.

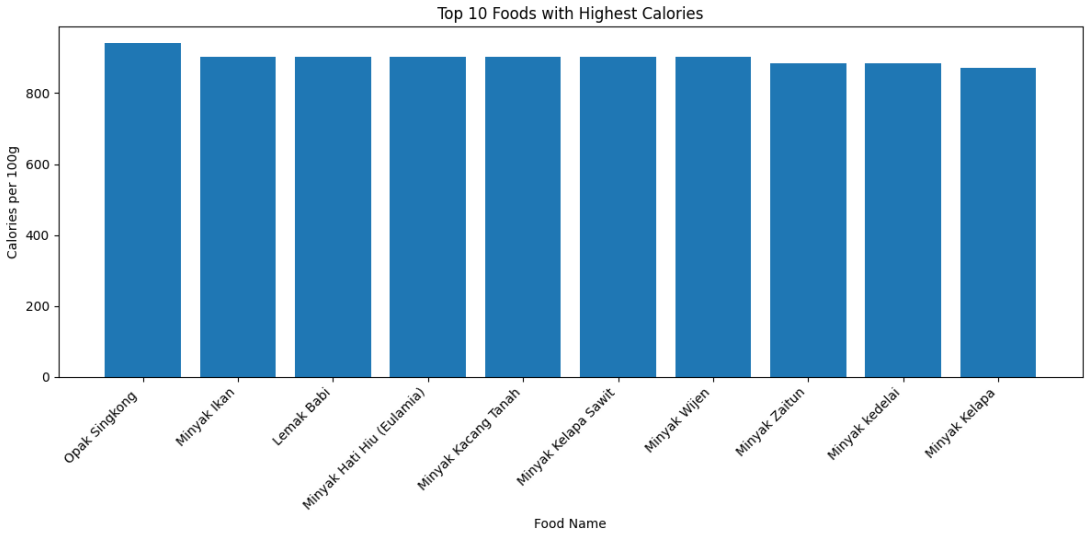

Figure 5. Top 10 Foods with Highest Calories

This bar chart illustrates the top 10 foods with the highest calorie content per 100 grams. The food items listed are predominantly oil-based or fat-rich products, which explains their significantly high caloric density. Foods like "Opak Singkong," "Minyak Ikan," and various types of oils (e.g., palm oil, sesame oil, and olive oil) feature prominently in this list. These items surpass 800 calories per 100 grams, underscoring their concentrated energy value, a characteristic typical of fats and oils. The presence of "Lemak Babi" (pork fat) and "Minyak Hati Hiu (Eulamia)" further reinforces this trend, as animal fats are also known for their high energy content. From a research perspective, such data is critical in dietary studies, especially when analyzing caloric intake patterns or assessing energy-dense food consumption's role in public health. For example, while these foods are energy-rich, excessive consumption could lead to health issues like obesity or cardiovascular diseases if not balanced with nutrient-dense, lower-calorie foods. This chart could also inform food policy or guide recommendations for calorie-conscious individuals, emphasizing the importance of moderation when including these high-calorie foods in daily diets.

Figure 6. Word Cloud Food-Related Words

This image is a word cloud that visually represents the frequency of various food-related words. Larger words, such as "segar" (fresh), "goreng" (fried), "masakan" (cooking/dish), "Daun" (leaf), and "Ikan" (fish), indicate higher usage or relevance within the dataset. The prevalence of these terms suggests a focus on fresh ingredients, traditional cooking methods, and common staples like fish, leaves, and fried dishes in the context being analyzed. This visual tool effectively highlights key themes and patterns in the associated text or data. From a research perspective, word clouds like this provide an intuitive summary of dominant terms, making it easier to grasp the central ideas in qualitative data. For instance, the emphasis on "segar" implies a strong cultural or dietary preference for fresh ingredients, while "goreng" indicates that frying is a prevalent cooking method. This type of analysis can be valuable for culinary studies, consumer behavior research, or even public health planning, as it points to dietary habits and priorities in a given community or dataset.

4.2. Results of Modeling Machine Learning

a. KNN

Table 3. Table Result of KNN Model Training

	Precision	Recall	F1-score	Support
High-Carbohydrate	0.98	0.97	0.98	186
High-Fat	0.92	0.92	0.92	24
High-Protein	0.92	0.93	0.93	60
Accuracy			0.96	270
Macro avg	0.94	0.94	0.94	270
Weighted avg	0.96	0.96	0.96	270

On a dataset comprising three dietary categories High-Carbohydrate, High-Fat, and High-Protein, the performance measures of a KNN model show an accuracy of 96%. The model thus accurately categorized 96% of every example in the dataset. With a precision of 0.98, the model for the High-Carbohydrates class indicates that 98% of the cases projected were accurate. With a similar high recall for this class, 97% of the actual High-Carbohydrates were accurately found. At 0.98, the F1-score which strikes a mix between accuracy and recall also shows great performance for this class. With a precision of 0.92 for the High-Fat class, 92% of the forecasts for that class were accurate. With a recall of 0.92, the model identified 92% of the real High-Fat cases.

With a 0.92 F1-score for High-Fat, this class's balance of recall and precision is likewise good. The model attained a precision of 0.92 for the High-Protein class, so meaning 92% of cases predicted as High-Protein were accurate. At 0.93, the recall is somewhat higher; hence, 93% of the actual High-Protein cases were precisely identified. With an F1-score of 0.93, High-Protein suggests dependable performance for this class. With an average performance across the three categories, the model's precision, recall, and F1-score are each 0.94 looking at the macro-average across all classes. With precision, recall, and F1 score all at 0.96 the weighted average which considers class support is somewhat higher. With somewhat better performance in the more common High-Carbohydrates class, this consistency reflects the strong classification ability of the model over many dietary classes.

b. Decision Tree

Table 4. Table Result of the Decision Tree Model Training

	Precision	Recall	F1-score	Support
High-Carbohydrate	0.97	0.96	0.96	186
High-Fat	0.88	0.88	0.88	24
High-Protein	0.89	0.90	0.89	60
Accuracy			0.94	270
Macro avg	0.91	0.91	0.91	270
Weighted avg	0.94	0.94	0.94	270

On a dietary dataset spanning three categories High-Carbohydrate, High-Fat, and High-Protein the performance measures of a Decision Tree model show an accuracy of 94%. The model thus correctly categorized 94% of all the data set instances. With a precision of 0.97, the model for the High-Carbohydrates class indicates that 97% of the instances projected as such were accurate. With a 0.96 recall for this class, 96% of real high-carbohydrate cases were correctly found. With an F1 score of 0.96, which strikes a mix between recall and accuracy, this class performs rather well. With a precision of 0.88, the model in the High-Fat class found that 88% of the forecasts for High-Fat were accurate. With a recall of 0.88, the model precisely identified 88% of real high-fat cases. With an F1 score of 0.88, High-Fat reflects a balanced class performance.

With a precision of 0.89, the High-Protein class's 89% of the instances projected were accurate. At 0.90, the recall is somewhat higher, meaning that 90% of real High-Protein cases were correctly found. With an F1-score of 0.89, High-Protein shows consistent identification of this class. With the model's precision, recall, and F1 score each 0.91, the average performance in the three categories and the macro-average in all classes show the same. At 0.94, the weighted average which considers the varying number of cases in every class showcases precision, recall, and the F1 score, quite near to the general accuracy. Particularly in the more frequent High-Carbohydrate class, this consistency among measures reflects the balanced and dependable performance of the model over the dietary categories.

c. Support Vector Machine

Table 5. Table Result of SVM Model Training

	Precision	Recall	F1-score	Support
High-Carbohydrate	0.99	0.98	0.99	186
High-Fat	0.96	1.00	0.98	24
High-Protein	0.97	0.98	0.98	60
Accuracy			0.99	270

Macro avg	0.97	0.99	0.98	270
Weighted avg	0.99	0.99	0.99	270

On a dietary classification dataset comprising the High-Carbohydrate, High-Fat, and High-Protein categories, the performance measures of a Support Vector Machine model show an overall accuracy of 99%. The model thus appropriately categorized 99% of all the data points. The model attained a precision of 0.99 for the High-Carbohydrates class, meaning that 99% of the cases projected as High-Carbohydrates were accurate. With a 0.98 recall for this class, 98% of the real high-carbohydrate cases were correctly found. Reflecting great performance in this class, the F1 score balances precision and recall and is 0.99. With a 0.96 precision, the model indicates that 96% of High-Fat class predictions were accurate. With a 1.00 recall, the model accurately recognized 100% of the real High-Fat events. With an F1 score of 0.98, the High-Fat class shows strong performance.

With a precision of 0.97, the model in the High-Protein class found that 97% of the cases forecasted were accurate. At 0.98, the recall is rather higher, meaning that 98% of real High-Protein cases were accurately found. With an F1-score of 0.98, High-Protein shows a great degree of accuracy in class identification. With the macro average across all classes, the model's precision, recall, and F1-score are 0.97, 0.99, and 0.98, respectively, so indicating the average performance over the three categories. With consideration for the varying number of events in every class, the weighted average reveals precision, recall, and F1-score all at 0.99, quite closely matching general accuracy. This consistent performance over all classes and measures reflects the dependability and efficacy of the model in classifying dietary categories with a high degree of accuracy.

d. Random Forest

Table 6. Table Result of Random Forest Model Training

	Precision	**Recall**	**F1-score**	**Support**
High-Carbohydrate	0.99	0.97	0.98	186
High-Fat	0.96	0.92	0.94	24
High-Protein	0.89	0.98	0.94	60
Accuracy			0.97	270
Macro avg	0.95	0.96	0.95	270
Weighted avg	0.97	0.97	0.97	270

On a dietary dataset comprising High-Carbohydrate, High-Fat, and High-Protein categories, the performance measures of a Random Forest model show an overall accuracy of 97%. The model thus correctly categorized 97% of every example in the dataset. With a precision of 0.99, the model for the High-Carbohydrates class found that 99% of the cases forecasted were accurate. With a 0.97 recall for this class, 97% of real high-carbohydrate cases were precisely found. Reflecting great performance for this class, the F1-score balances accuracy and recall and is 0.98. With a precision of 0.96, the model in the High-Fat class makes 96% of all the predictions accurate. At 0.92, the recall is somewhat lower, meaning that 92% of the real High-Fat cases were correctly found. With an F1-score of 0.94, High-Fat shows good but somewhat less performance than High-Carbohydrates class.

With a precision of 0.89 for the High-Protein class, the model correctly 89% of the cases projected as High-Protein. At 0.98 the recall is higher, indicating that 98% of real High-Protein cases were correctly identified. Though accuracy is rather lower, the F1-score for High-Protein is 0.94, indicating dependable performance in defining this class. With the macro average across all classes, the model's accuracy, recall, and F1-score are 0.95, 0.96, and 0.95, respectively, so reflecting the average performance over the three categories. Closely matching the general accuracy, the weighted average which considers the varying number of instances in every class shows the precision, recall, and F1 score all at 0.97. This great consistency in the measures reflects the dependable and balanced performance of the model at all dietary levels.

e. Multilayer Perceptron (MLP)

Table 7. Table Result of MLP Model Training

	Precision	Recall	F1-score	Support
High-Carbohydrate	0.99	0.98	0.99	186
High-Fat	0.96	0.96	0.96	24
High-Protein	0.95	0.98	0.97	60
Accuracy			0.98	270
Macro avg	0.97	0.98	0.97	270
Weighted avg	0.98	0.98	0.98	270

On a three-category dietary dataset, High-Carbohydrate, High-Fat, and High-Protein the performance measures of an MLP model show an overall accuracy of 98%. This shows that in the whole dataset, the model accurately categorized 98% of all events. With a precision of 0.99, the model found that 99% of the High-Carbohydrate class instances projected were accurate. With a recall for this class of 0.98, 98% of real High-Carbohydrates cases were precisely detected. Reflecting extraordinary performance in this category, the F1-score which balances precision and recall is 0.99. With a precision of 0.96, the model in the High-Fat class found that 96% of the High-Fat instances projected were accurate. With a recall of 0.96, which denotes that 96% of real High-Fat cases were precisely found. With a 0.96 F1-score for High-Fat, this class exhibits constant and dependable performance.

With a precision of 0.95, the model forecasts 95% of the High-Protein class events to be accurate. At 0.98, the recall is rather higher, indicating that 98% of real High-Protein cases were correctly found. With a 0.97 F1-score for High-Protein, this category's performance is likewise rather strong. Concerning the three categories, the macro-average in all classes reveals that the precision, recall, and F1 score of the model are 0.97, 0.98, and 0.97, respectively. At 0.98, the weighted average which takes into account the different number of events in every class showcases the precision, recall, and F1 score, so closely matching the general accuracy. This great consistency among several criteria emphasizes the balanced and efficient performance of the model in classifying all kinds of diets.

4.3. Discussion

The discussion that will be discussed is a comparison between research models that have been trained using data. Then, the best model is used as a proposed model. The comparison of research models can be seen in Table 7.

Table 8. Table Model Evaluation Comparison

Proposed Model	Best Hyperparameter	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
KNN	weights: uniform, n_neighbors: 5, metric: manhattan	96.19	94	94	94
Decision Tree	min_samples_split: 10, min_samples_leaf: 1, max_features: None, max_depth: 50, criterion: gini	94.89	91	91	91
SVM	kernel: linear, gamma: 1, degree: 2, C: 100	99.16	97	99	98
Random Forest	n_estimators: 100, min_samples_split: 5, min_samples_leaf: 1, max_features: log2, max_depth: 20	96.74	95	96	95
MLP	solver: adam, max_iter: 1000, learning_rate: adaptive, hidden_layer_sizes: (50,), alpha: 0.0001, activation: logistic	98.04	97	98	97

The Support Vector Machine (SVM) model achieved the highest accuracy at 99.16%, with a linear kernel, gamma set to 1, degree at 2, and a C value of 100. This model also had the highest recall (99%) and F1 score (98%), reflecting its superior ability to accurately identify true instances. The multi-layer perceptron (MLP), configured with an adaptive learning rate and logistic activation function, also performed well, reaching an accuracy of 98.04%, with precision and F1 score at 97% and recall at 98%. Other models, including Random Forest and XGBoost, achieved accuracies above 96%, showing reliable performance with optimized parameters such as n_estimators, max_depth, and sampling parameters. This comparison highlights SVM and MLP as the most effective models for achieving high classification accuracy in this study, followed closely by Random Forest and XGBoost.

The results of this study have significant implications for dietary recommendations, particularly in the context of Indonesian cuisine. The high accuracy rates achieved by the machine learning models, especially the Support Vector Machine indicate that these models can effectively classify foods based on their macronutrient content, such as high carbohydrates, high fats, and high proteins. This classification not only aids in understanding the nutritional composition of various foods but also supports the creation of data-driven dietary guidelines. The findings suggest that individuals can make more informed dietary choices based on the classification results, which can lead to improved health outcomes. Furthermore, the study highlights the potential for machine learning to bridge the gap between traditional dietary research and modern technological innovations. Overall, the integration of EDA and machine learning in this research underscores the importance of data-driven approaches in developing effective dietary recommendations that are both practical and scientifically grounded.

5. Conclusion

This study opens up new avenues for dietary classification and machine learning research. Expanding the model to include more diets and foods is promising. Researchers can improve the model's robustness and applicability across populations and diets by using more datasets. Future studies could also use mobile apps to collect real-time data for continuous dietary monitoring and personalized nutrition recommendations. Future research should examine how preprocessing methods affect model performance. This study normalized nutritional features using standard scaling, but feature engineering or dimensionality reduction may improve classification accuracy. Advanced machine learning methods like deep learning could also be tested for dietary classification. Future research could examine how dietary classifications affect health. Data-driven dietary recommendations can be tested by longitudinal studies that track health metrics and diet. This could lead to public health-focused nutrition interventions. Finally, nutritionists and healthcare professionals must collaborate to turn these findings into dietary guidelines. Engaging nutrition stakeholders can ensure that models and recommendations are scientifically sound, culturally relevant, and accessible to the target population. These future directions show that machine learning and nutrition science can continue to innovate, improving diets and health.

References

[1] B. Yudhistira and A. Fatmawati, "Diversity of Indonesian soto," *J. Ethn. Foods*, vol. 7, no. 1, p. 27, Dec. 2020, doi: 10.1186/s42779-020-00067-z.

[2] A. Wibisono, H. A. Wisesa, Z. P. Rahmadhani, P. K. Fahira, P. Mursanto, and W. Jatmiko, "Traditional food knowledge of Indonesia: a new high-quality food dataset and automatic recognition system," *J. Big Data*, vol. 7, no. 1, p. 69, Dec. 2020, doi: 10.1186/s40537-020-00342-5.

[3] P. K. Fahira, Z. P. Rahmadhani, P. Mursanto, A. Wibisono, and H. A. Wisesa, "Classical Machine Learning Classification for Javanese Traditional Food Image," in *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, Nov. 2020, pp. 1–5. doi: 10.1109/ICICoS51170.2020.9299039.

[4] D. Chrimes, "Using Decision Trees as an Expert System for Clinical Decision Support for COVID-19," *Interact. J. Med. Res.*, vol. 12, p. e42540, Jan. 2023, doi: 10.2196/42540.

[5] M. Senthilmurugan, N. Yamsani, C. M. B. M J, L. S, S. Padmakala, and A. Akilandeswari, "Evaluation of Support Vector Machine and Kernel Neural Network Classification for Fast Food Nutrition Data," in *2023 Second International Conference on Augmented Intelligence and*

Sustainable Systems (ICAISS), IEEE, Aug. 2023, pp. 150–154. doi: 10.1109/ICAISS58487.2023.10250603.

[6] Y. A. Sari *et al.*, "Indonesian Traditional Food Image Identification using Random Forest Classifier based on Color and Texture Features," in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, IEEE, Sep. 2019, pp. 206–211. doi: 10.1109/SIET48054.2019.8986058.

[7] I. Kumar, J. Rawat, N. Mohd, and S. Husain, "Opportunities of Artificial Intelligence and Machine Learning in the Food Industry," *J. Food Qual.*, vol. 2021, pp. 1–10, Jul. 2021, doi: 10.1155/2021/4535567.

[8] J. Kong *et al.*, "Deep-Stacking Network Approach by Multisource Data Mining for Hazardous Risk Identification in IoT-Based Intelligent Food Management Systems," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/1194565.

[9] V. M. Oddo *et al.*, "Evidence-Based Nutrition Interventions Improved Adolescents' Knowledge and Behaviors in Indonesia," *Nutrients*, vol. 14, no. 9, p. 1717, Apr. 2022, doi: 10.3390/nu14091717.

[10] L. Carnevale *et al.*, "Towards a precision medicine Solution for optimal pediatric Laparoscopy: An exploratory data analysis for features Selections," *Biomed. Signal Process. Control*, vol. 88, p. 105321, Feb. 2024, doi: 10.1016/j.bspc.2023.105321.

[11] R. A. Asmara, M. Z. Abdullah, and F. M. Wahyuningtyas, "Herbalife Nutrition Product Recommender System using Mamdani Fuzzy Logic," in *2020 4th International Conference on Vocational Education and Training (ICOVET)*, IEEE, Sep. 2020, pp. 1–5. doi: 10.1109/ICOVET50258.2020.9230303.

[12] Q. D. Ayu Sulistiani, B. Irawan, and C. Setianingsih, "Dietary Habits for Toddler Growth using Particles Swarm Optimization Algorithms," in *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, IEEE, Oct. 2020, pp. 1–6. doi: 10.1109/ICORIS50180.2020.9320841.

[13] M. Rong, "Nutritional Analysis and Optimization of College Students' Dietary Recipes Based on Entropy Weight TOPSIS Method and Simulated Annealing Algorithm," in *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, IEEE, Aug. 2024, pp. 286–290. doi: 10.1109/ICSECE61636.2024.10729414.

[14] A. H. A. M. Siagian *et al.*, "Developing Food Menu Recommendation System Based on Indonesian Nutritional Needs," in *2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, IEEE, Oct. 2024, pp. 105–110. doi: 10.1109/IC3INA64086.2024.10732573.

[15] A. F. Hanif, "Indonesian Food and Drink Nutrition Dataset." Accessed: Nov. 25, 2024. [Online]. Available: https://www.kaggle.com/datasets/anasfikrihanif/indonesian-food-and-drink-nutrition-dataset/

[16] W. Wang, Z. Liu, J. Tang, and C. Yuan, "An enhanced strategy for geo-exploratory data analysis to facilitate the discovery of new mineral deposits," *J. Geochemical Explor.*, vol. 258, p. 107411, Mar. 2024, doi: 10.1016/j.gexplo.2024.107411.

[17] F. Zou, "Research on data cleaning in big data environment," in *2022 International Conference on Cloud Computing, Big Data and Internet of Things (3CBIT)*, IEEE, Oct. 2022, pp. 145–148. doi: 10.1109/3CBIT57391.2022.00037.

[18] L. Shen *et al.*, "Enhanced multi-scale feature adaptive fusion sparse convolutional network for large-scale scenes semantic segmentation," *Comput. Graph.*, p. 104105, Dec. 2024, doi: 10.1016/j.cag.2024.104105.

[19] B. B. C, A. Deshmukh, and A. V Narasimhadhan, "Modulation and signal class labelling with active learning and classification using machine learning," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, IEEE, Jul. 2022, pp. 1–5. doi: 10.1109/CONECCT55679.2022.9865826.

[20] S. Zhang and J. Li, "KNN Classification With One-Step Computation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, 2023, doi: 10.1109/TKDE.2021.3119140.

[21] T. A. Munshi, L. N. Jahan, M. F. Howladar, and M. Hashan, "Prediction of gross calorific value from coal analysis using decision tree-based bagging and boosting techniques," *Heliyon*, vol. 10, no. 1, p. e23395, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23395.

[22] R. Rofik, R. Aulia, K. Musaadah, S. S. F. Ardyani, and A. A. Hakim, "Optimization of Credit Scoring

Model Using Stacking Ensemble Learning and Oversampling Techniques," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, Dec. 2023, doi: 10.52465/joiser.v2i1.203.

[23] K. Zhang, J. Yang, J. Sha, and H. Liu, "Dynamic slow feature analysis and random forest for subway indoor air quality modeling," *Build. Environ.*, vol. 213, p. 108876, Apr. 2022, doi: 10.1016/j.buildenv.2022.108876.

[24] F. K. Oduro-Gyimah, K. O. Boateng, P. B. Adu, and K. Quist-Aphetsi, "Prediction of Telecommunication Network Outage Time Using Multilayer Perceptron Modelling Approach," in *2021 International Conference on Computing, Computational Modelling and Applications (ICCMA)*, IEEE, Jul. 2021, pp. 104–108. doi: 10.1109/ICCMA53594.2021.00025.

[25] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.