JoCoSiR
Journal of Computer Science Research ISSN 2986-2337 (Online)

Comparison of Decision Tree and Random Forest Algorithm Performance for Nutrition Classification in Fast Food

Perianus Lombu [a], *Kiki wulandari* [b]

Magister Teknologi Informasi, Universitas Pembangunan Panca Budi.
Email: a perianus.lombu@gmail.com, bkikywullandari@gmail.com.

ARTICLEINFO

Keywords:
Decision Tree
Lifestyle
Classification
Food
Random Forest

IEEE style in citing this article:
Perianus Lombu and Kiki Wulandari, "Comparison of Decision Tree and Random Forest Algorithm Performance for Nutrition Classification in Fast Food," JoCoSiR: Scientific Journal of Information Systems Technology, vol. 3, no. 3, pp. 89-95 2025.

ABSTRACT

Fast food has become an essential part of the busy modern lifestyle, fast food is more popular because it makes eating easy and convenient. Today's young people are very fond of instant food. However, excessive consumption of instant food can trigger various health problems, including obsessive eating patterns. This raises the need to develop more accurate analytical methods for classifying fast food nutritional data, the purpose of classification is to obtain a decision tree model that can be used to anticipate and pay attention to how variables in the data are related to each other. In comparing the performance of the Decision Tree and Random Forest Algorithms in processing fast food nutritional data, it was found that all variables were correlated. The implementation results found that both models have extraordinary capabilities. The performance of the Decision Tree and Random Forest Algorithms on the same dataset, Random Forest outperformed Decision Tree with an accuracy value of 66.67%, while Decision Tree only achieved 55.56%, indicating that Random Forest is able to provide more accurate predictions for the test data class. In addition, the characteristics of the Random Forest group, where several decision trees are combined, provide advantages in handling data complexity and improve model generalization. These results indicate that group learning can improve the performance and reliability of predictions in building classification models, especially in the case of complex datasets.

1. Introduction

Nutritional problems in population development are still a major problem in the global population order [1]. As a result, this problem has become one of the important points in the Millennium Development Goals (MDGs). Health is the most important thing for humans where disease comes through diet [2]. Health is very important because diet can cause disease, especially for individuals who have very busy jobs and do not have time for breakfast and prefer fast food that is easily accessible in many restaurants or cafes. Fast food is known as junk food, defined as junk food or non-nutritious food that is considered not to contain important nutrients for the body [3].

Fast food has become an important part of the busy modern lifestyle, because it makes eating easy and convenient. Teenagers today are very fond of fast food. However, Yaman, et al. Algorithm Journal | Vol. 21 No. 2 | November 2024 | Pages 184-196 185 consuming fast food in excessive amounts can cause various health problems, one of which is obesity. Fast food is also often associated with a lack of nutrition. This is because people often do not have much time to carry out daily routines, which makes them pay less attention to the nutrition in the food they consume. Teenagers should be educated and encouraged to take responsibility for choosing healthy snacks, foods that contain nutrients that are very important for the body to grow and develop [4].

According to data from the Indonesian Central Bureau of Statistics, the average protein consumption per person increased significantly from 2019 to 2023 [4]. This phenomenon can be considered a sign of significant changes in people's consumption patterns. This change can be caused by a number of variables, including increased awareness of nutrition, better economic dynamics and changes in people's preferences regarding lifestyle and food. According to Riskesdas data, most Indonesians aged over ten years consume sweet foods more than once a day (53.1%), fatty, cholesterol-rich, and fried foods more than once a day (40.7%), and almost four out of five Indonesians [5]. This raises the need to develop more accurate analysis methods to classify fast food nutrition data. Classification is a mode for finding patterns that can separate one class of data from another and determine objects that fall into a certain category by looking at how the behavior and characteristics of the defined groups [6]. The purpose of classification is to obtain a decision tree model that can be used to make predictions and see how variables in the data relate to each other [7]. Various classification algorithms are used to improve the efficiency and accuracy of fast food nutrition data management. One of the algorithms used is the Decision Tree and Random Forest algorithms. According to Shafarindu, Decision tree is a fairly popular technique for classification and prediction because it uses a tree structure to determine the sequence of decisions and their consequences [8]. Decision tree is often referred to as a structure similar to a flowchart where each node represents an attribute value, each branch represents a test result, and each leaf represents a class or class distribution [9]. Meanwhile, Random Forest is one of the methods used for classification and regression, Random forest is able to translate as formed

DOI: https://doi.org/10.65126/jocosir.v3i3

from a collection of decision trees as well [10]. Random Forest has several advantages, including the ability to improve accuracy in the presence of missing data and combat outliers, as well as data storage efficiency. In addition, Random Forest has a feature selection process that allows it to retrieve the most important features, which can improve the performance of the classification model. Due to the feature selection process, Random Forest can certainly operate effectively with large data with complex parameters [11]

2. State of the Art

Currently, in the field of applied Machine Learning, particularly for nutritional classification (such as classifying whether a fast food is “High in Salt”, “Low in Fat”, or “Good Source of Protein” based on its nutritional composition), Random Forest (RF) is consistently recognized as an algorithm with superior performance and more robustness compared to a single Decision Tree (DT).

Decision Tree (DT)

Decision Tree is a basic algorithm that converts data into a decision tree structure and decision rules.

Excess:

1. Easy Interpretation (Simplicity):This is a major advantage. The hierarchical tree structure is very intuitive and easy to understand for non-specialists.
2. Minimal Data Pre-processing:Does not require normalization or standardization of nutritional data.
3. Fast:Relatively fast in training and prediction process on small to medium data.

Lack:

1. Prone to Overfitting:If a tree grows too deep and complex, it tends to overfit the training data, so it will perform poorly when tested with new, never-before-seen nutrient data.
2. Data Sensitivity (Instability):Small changes in the training nutrition data can cause large changes in the decision tree structure, making it unstable.
3. Accuracy Tends to be Lower:Prediction accuracy is generally lower than ensemble methods such as Random Forest, especially on complex nutrient datasets.

Random Forest (RF)

Random Forest is an ensemble learning method that works by building a large number (hundreds to thousands) of decision trees. The final classification result is determined by voting (majority) across all trees.

Key Highlights (State of the Art Performance):

1. Superior Accuracy:RF almost always achieves significantly higher accuracy than a single DT. Studies often show significant accuracy improvements (e.g., from 75% on DT to 90-99% on RF, depending on the dataset).
2. Overcoming Overfitting:The ensemble mechanism (combining multiple trees) and random feature selection (selection of random features at each split) effectively reduce model variance and mitigate the risk of overfitting, which is a major problem of DT.
3. Strong Generalization:The RF model is more robust and has better generalization ability to unseen nutritional data (test data).
4. Handling Missing Values and Outliers:RF has built-in mechanisms to handle missing values (blanks) and outliers, which frequently appear in fast food nutrition data.
5. Feature Importance:RF can calculate how important each nutritional feature (e.g., sugar content, total fat, sodium) is in the classification process, providing valuable insights for researchers or nutritionists.

Lack:

1. Less Interpretive (Black Box):Because it involves hundreds of trees, it is difficult to interpret how the final classification decision is made, making it less transparent than a single DT.
2. Longer Training Time:Building hundreds of trees requires more computational resources and training time than a single DT.

3. Method

This study uses a quantitative approach by applying a machine learning-based classification method. The research flow begins with data collection and processing, followed by the application of Random Forest and Decision Tree algorithms to classify fast food menus. Next, the classification results are analyzed using various evaluation metrics to determine the most optimal method in the context of this study. Decision Tree was chosen because of its ability to identify patterns based on certain attributes conceptually and clearly. Meanwhile, Random Forest is used to improve model accuracy by combining the results of several decision trees, thereby reducing the possibility of overfitting on the training dataset. Model evaluation and validation are carried out to measure the effectiveness of the method used. Assessments are carried out based on parameters such as accuracy, precision, recall, and F1-score to ensure the quality of the classification results. The developed model is expected to be implemented in the restaurant industry to support the automation process in managing fast food menus more efficiently.

Data collection

This research began with data collection through literature studies and the public platform kaggle.com. The literature study was conducted to find theoretical foundations and reference sources in books or journals about data mining techniques and was used to examine previous research, methods to be used and review the literature. And data retrieval on the Kaggle.com site. The obtained dataset amounted to 1147 with a total of 14 columns consisting of Company, Item, Calories, Calories from, Total, Saturated, Trans Fat, Cholesterol, Sodium, Carbs, Fiber, Sugars, Protein, Weight.

Algorithmic Modeling

1. Decision Tree
   Each node in the tree represents a tested trait in a tree structure known as a Decision Tree. Each branch serves as a division of the test results, and the leaf node, or leaves, serves as a representation of a particular class group. In a decision hierarchy, the highest level node is the root, or roots, which is usually the trait that has a large impact on a particular class. To solve the problem, Decision Trees usually use a bottom-up search strategy. In the attribute value classification method, the path from the root node (root) to the end node (flower) will be learned, after which the new class will be determined [13]
2. Random Forest
   A set of training data is collected and fed into a tree to start the training process. Within a node, feature selection will be divided and taken randomly. Bagging selects samples repeatedly, using replacement [14]. In the process of designing a Random Forest algorithm model, the Random Forest algorithm model is built using "n" decision trees to find the best "n" value. Furthermore, the Random Forest algorithm produces optimal results [15]. The following is a table of the Random Forest algorithm pseudocode [16]. For classification and regression tasks, the Random Forest Algorithm is a group learning technique based on the idea of randomly generating a large number of decision trees.

4. Results and Discussion

This section discusses the results of testing conducted on the fastfood dataset using the Decision Tree and Random Forest algorithms. The testing process is as follows:

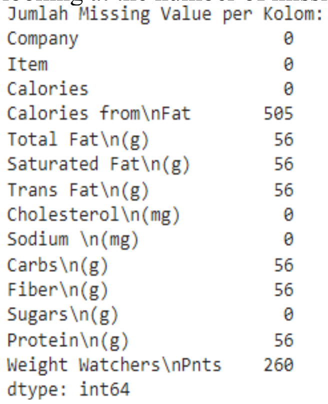
Dataset collection

Data collection from the fast food nutrition dataset on the Kaggle.com site resulted in a dataset totaling 1147 with 14 columns consisting of Company, Item, Calories, Calories from, Total, Saturated, Trans Fat, Cholesterol, Sodium, Carbs, Fiber, Sugars, Protein, Weight. After the data collection stage is carried out, the data import stage is continued. The data import stage is the first stage in the data analysis process and involves loading from various sources into a data analysis environment such as Python.

Data Cleaning

At this stage, namely cleaning the data, the following are the stages in the data cleaning process:

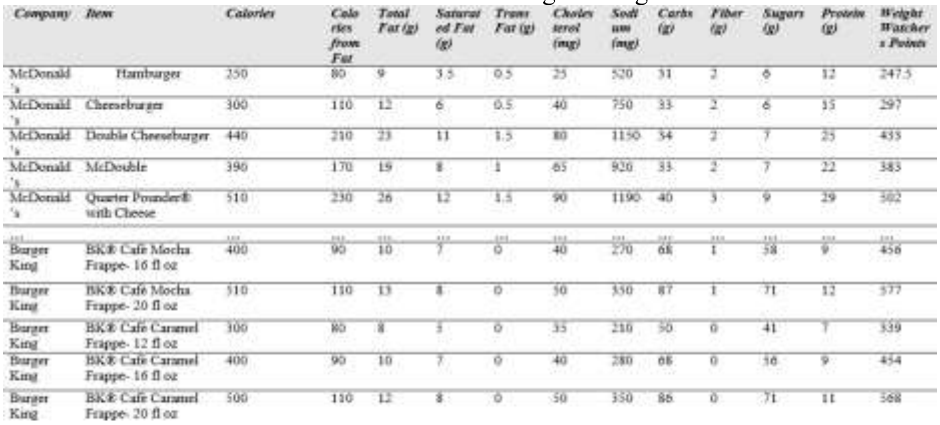
1) Displays the number of missing values
   The data cleaning stage begins by looking at the number of missing values from the dataset used.



```
Jumlah Missing Value per Kolom:
Company                   0
Item                      0
Calories                  0
Calories from\nFat      505
Total Fat\n(g)           56
Saturated Fat\n(g)       56
Trans Fat\n(g)           56
Cholesterol\n(mg)         0
Sodium \n(mg)             0
Carbs\n(g)               56
Fiber\n(g)               56
Sugars\n(g)               0
Protein\n(g)             56
Weight Watchers\nPnts   260
dtype: int64
```

2) Deleting data containing missing values

Table 1: Dataset after removing Missing Values



Company	Item	Calories	Calories from Fat	Total Fat (g)	Saturated Fat (g)	Trans Fat (g)	Cholesterol (mg)	Sodium (mg)	Carbs (g)	Fiber (g)	Sugars (g)	Protein (g)	Weight Watchers Points
McDonald's	Hamburger	250	80	9	3.5	0.5	25	520	31	2	6	12	247.5
McDonald's	Cheeseburger	300	110	12	6	0.5	40	750	33	2	6	15	297
McDonald's	Double Cheeseburger	440	210	23	11	1.5	80	1150	34	2	7	25	433
McDonald's	McDouble	390	170	19	8	1	65	920	33	2	7	22	383
McDonald's	Quarter Pounder® with Cheese	510	230	26	12	1.5	90	1190	40	3	9	29	502
...		...	...	...	...	...	...	...	...	...	...	...	...
Burger King	BK® Café Mocha Frappe- 16 fl oz	400	90	10	7	0	40	270	68	1	58	9	456
Burger King	BK® Café Mocha Frappe- 20 fl oz	510	110	13	8	0	50	350	87	1	71	12	577
Burger King	BK® Café Caramel Frappe- 12 fl oz	300	80	8	5	0	35	210	50	0	41	7	339
Burger King	BK® Café Caramel Frappe- 16 fl oz	400	90	10	7	0	40	280	68	0	56	9	454
Burger King	BK® Café Caramel Frappe- 20 fl oz	500	110	12	8	0	50	350	86	0	71	11	568

3) Displays the number of rows before and after performing Data Cleaning.

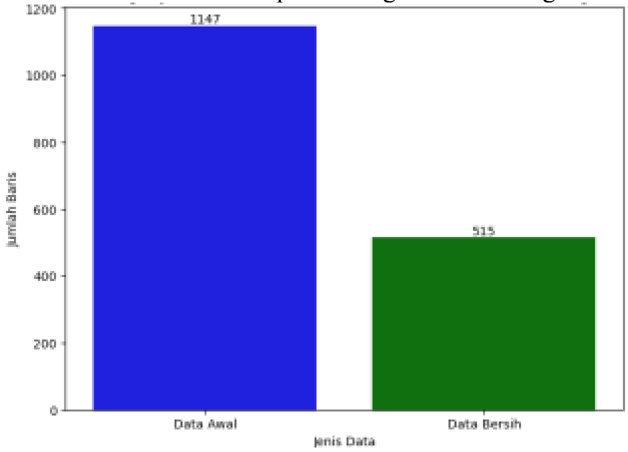


Figure 1: Barplot of Initial Data and Clean Data

Figure 1 above shows that the initial data bar plot displays the distribution or frequency of categorical data before the cleaning or processing process is carried out, which is not yet neat or has several problems such as missing data, duplicates, or entry errors. And the clean data bar plot shows the distribution or frequency of the data after the cleaning or processing process is carried out.

Pre-Processing

This stage is the data processing required if there are problems with the dataset. The following are the preprocessing stages.

1) Convert data type “Object” to “Float”
   a. Reading Dataset
   b. Checking the dataset to find out the data type
   c. Identify columns that have the data type “object”
   d. Convert columns with data type “object” to “float”
   e. Conversion successful

In data processing, the "object" data type can refer to data types that typically store non-numeric values, such as strings, text, or a combination of several data types. For example, a column in a dataframe might contain text such as names, addresses, or categories. However, the "float" data type is a numeric data type used to store real number values. In computer representation, floats are represented as fractional numbers with a decimal point, allowing for the use of both whole and non-whole values. In data analysis, for example, a column with the "float" data type can contain values such as fractional numbers representing measurements, percentages, or other types of numeric values. Converting the data type from "object" to "float" is a common step in data preprocessing before further analysis. Many machine learning algorithms require numeric input, such as classification algorithms. Converting the "object" data type to "float" can make the data type more suitable for use with these algorithms.

DOI: https://doi.org/10.65126/jocosir.v3i3

2) Data correlation using Heatmap

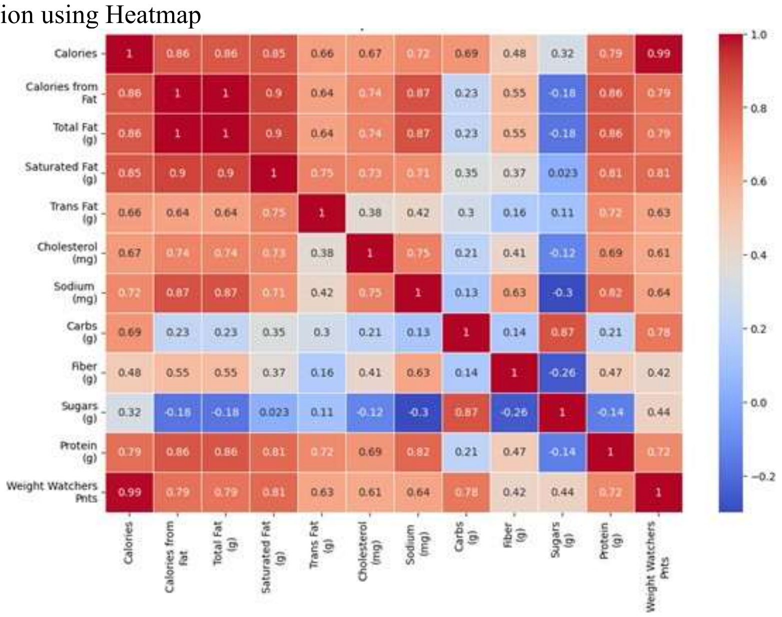


Figure 2: Heatmap of Fast Food Nutrition Correlation

The heatmap visualization above shows a certain range of values. The color scale moves from light to dark to indicate lower values. The colors in between indicate the range of values between them. Each cell in the heatmap contains a numeric value representing the strength of the correlation between factors in the dataset. Values close to 1 indicate a strong positive correlation, while values close to -1 indicate a strong negative correlation. A diagonal line with a value of 1.00 indicates that each factor is perfectly correlated with itself. The image above is compelling because it visually summarizes complex nutritional data and shows how the components relate to each other in fast food. For example, there is a high positive correlation between total fat and calories, or between saturated fat and trans fat.

Processing

1) Balance & Imbalance

The number of samples of each value is displayed in the "Calories" column of the dataset; for example, three samples have a calorie value of 740, two samples have a value of 560, and so on. This is a summary of the number of times each value appears in the dataset. Among the values in the "Calories" column, the sample distribution is relatively even or balanced, meaning that no single value appears significantly more frequently or less frequently than the others, according to the statement "Balanced dataset." Because there is no significant bias toward any particular class, this can be beneficial for training a model.

2) Principal Component Analysis (PCA)

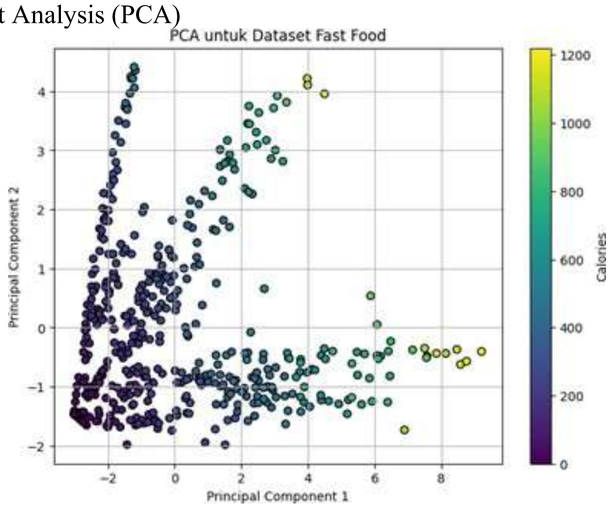


Figure 3. PCA Fastfood

PCA is a statistical method used to reduce the dimensionality of complex datasets while retaining as much information as possible. It is useful in analyzing large datasets with many features or variables. For fast food nutrition datasets, PCA helps identify hidden patterns within the information. For example, consider a dataset with many variables representing nutrients. By reducing dimensionality, PCA helps simplify the interpretation of the information, which can aid in further analysis such as classifying foods based on nutritional value, identifying common eating patterns, or even developing better eat-less recommendations.

3) Training Data and Testing Data
Number of Training Data: 403
Number of Test Data: 112
In the context of testing the performance of two algorithms, this data is used to compare the performance of two different algorithms in classifying the nutrition of fast food.

Algorithmic Modeling

1) Decision Tree

In data mining and machine learning, the Decision Tree algorithm is a popular and easy-to-use learning method. By dividing data into increasingly smaller subsets based on specific features, the algorithm constructs a decision tree. Each node in the tree represents a decision based on the feature being tested, and each subdivision of the tree represents the probability of a test outcome for that feature. A decision tree is a predictive model that uses a tree-like structure to make decisions based on a set of logical rules. Similar to human decision-making processes, the model shows the relationship between input features and target outputs. The model makes decisions based on the value of a specific feature at each node in the decision tree. Each node represents a question or test about a particular feature, and the branches from the node represent the possible outcomes of that test. Finally, the leaves of the tree serve as the model's predictions or labels for the input data.

Table 2: Decision Tree Algorithm Classification Report

Data labels	Precision	Recall	F1-Score	Support
120.0	0.00	0.00	0.00	1
190.0	0.00	0.00	0.00	1
210.0	0.00	0.00	0.00	0
310.0	0.00	0.00	0.00	1
540.0	0.00	0.00	0.00	0
560.0	0.00	0.00	0.00	1
630.0	0.00	0.00	0.00	0
660.0	0.00	0.00	0.00	1
740.0	0.00	0.00	0.00	0
860.0	0.00	0.00	0.00	1
1110.0	0.50	1.00	0.67	1
1130.0	0.00	0.00	0.00	1
1150.0	0.00	0.00	0.00	1
1190.0	0.00	0.00	0.00	0

2) Random Forest

Random Forest is one of the most popular ensemble algorithms in machine learning, particularly for classification and regression tasks. It builds multiple decision trees during training and combines the predictions from each tree to improve the overall model performance. Random Forest is based on decision trees. Each tree in a Random Forest is an independent learning model that makes decisions based on a set of rules derived from the data. These trees are constructed independently using randomly selected subsets of data (bootstrap samples) and random subsets of features.

Table 3: Random Sample using Random Forest Algorithm

Calories	Caloriesfrom Fat	TotalFat (g)	SaturatedFat (g)	TransFat (g)	Cholesterol (mg)	Sodium (mg)	Carbs (g)	Fiber (g)	Sugars (g)	Protein (g)
1110.0	240.0	26.0	16.0	2.0	100.0	350.0	194.0	0.0	168.0	25.0
660.0	360.0	40.0	12.0	1.5	90.0	980.0	49.0	2.0	11.0	28.0
560.0	240.0	27.0	9.0	0.5	260.0	1300.0	56.0	3.0	7.0	24.0
1150.0	710.0	79.0	31.0	3.5	240.0	2150.0	49.0	2.0	10.0	61.0
860.0	220.0	24.0	15.0	1.5	75.0	260.0	144.0	0.0	124.0	18.0

This table shows the values for each variable associated with the fast food data set. Each row shows a fast food sample with values related to calories, fat, cholesterol, sodium, carbohydrates, fiber, sugar, and protein.

Evaluation

One of the simplest and most easily understood classification methods is the Decision Tree Algorithm. It uses various metrics to assess the model's performance in classifying data. The first metric we consider is accuracy. With an accuracy of 55.56%, the Decision Tree model successfully predicted the class for about half of the test data used. However, to gain a better understanding, this accuracy value must be considered in conjunction with other metrics. In contrast, the Random Forest model is an ensemble approach based on Decision Trees. In its performance evaluation, it was found that the Random Forest model had a slightly higher accuracy, reaching 66.67%. This indicates that, overall, the Random Forest model performed better in predicting the class of the test data used. The ensemble nature of this algorithm, where it uses a large amount of test data, can produce better results.

5. Conclusions

Taking into account the research that has been done, it can be concluded that in the performance evaluation of the Decision Tree and Random Forest Algorithms on the same dataset, Random Forest outperforms Decision Tree with an accuracy value of 66.67%, while Decision Tree only reaches 55.56%, indicating that Random Forest is able to provide more accurate predictions for the test data class. Furthermore, the clustering characteristics of Random Forest, where multiple decision trees are combined, provide advantages in handling data complexity and improving model generalization. These results indicate that cluster learning can improve the performance and predictive reliability of classification models, especially in the case of complex datasets. Conversely, performance evaluation doesn't solely rely on accuracy. To gain a better understanding of model performance, additional metrics such as precision, recall, and F1 score should also be considered. For class prediction in the fast food dataset, Random Forest ultimately emerged as the best choice.

6. Acknowledgment

The author would like to thank Mr. and Mrs. Nani Sulistianingsih and Dr. Muhammad Iqbal, M.Kom., as lecturers who consistently provided guidance, advice, and motivation to the author, resulting in the successful completion of this research. The author would also like to thank all parties who have supported and contributed to the preparation of this paper.

7. References

[1] Marini, HS Tambunan, ZA Siregar, AP Windarto, and F. Rizki, “Application of Data Mining for Infant Nutrition Classification Using the Decision Tree C4.5 Algorithm,” ZAHRA Big Data Bulletin, vol. 1, no. 2,
pp. 88–96, 2022.

[2] AP Wahyu, H. Heryono, MB Chaniago, and D. Hamdani, “SmartCanteen: Behavior of Regulating Dietary Patterns by Reading Nutritional Values (Conventional Deep Learning Neural Network),” Jurnal Ilmiah Teknologi Informasi Terapan, vol. 6, no. 2, pp. 115–121, 2020, doi: 10.33197/jitter.vol6.iss2.2020.356.

[3] I. Pamelia, "Fast Food Consumption Behavior in Adolescent and ITS Impact for Health," IKESMA Journal, vol. 14, no. 2, pp. 144–153, 2018.

[4] BP Statistics, “Average daily protein consumption,” 2023.

[5] V. Cyrilla et al., “Research Article on the Relationship between Fast Food Consumption Habits and Body Mass Index in Students of the Faculty of Medicine, Ukrida, Class of 2016, September 2017 Period Relationship Between Habit of Consuming Fast Food ...Fast Food Consumption Against Body Mass Index in St," no. September, 2017.

[6] I. Romli and AT Zy, “Determining Overtime Schedules with Employee Data Classification Using the C4.5 Algorithm,” Journal of Computer Science & Informatics (J-SAKTI, vol. 4, no. 2, pp. 694–702, 2020.

[7] S. Alam, A. Siregar, and A. Juwita, “Application of C4.5 Algorithm for COVID-19 Case Classification,”
… , Technology and …, vol. III, 2022.

[8] AW Wardhana, E. Patimah, AI Shafarindu, YM Siahaan, BV Haekal, and DS Prasvita, "Classification of Sales Data in Supermarkets Using the Decision Tree Method," Senamika, vol. 2, no. 1, pp. 660–667, 2021.

[9] S. Supangat, AR Amna, and T. Rahmawati, “Implementation of Decision Tree C4.5 to Determine Body Weight Status and Energy Needs in Children Aged 7-12 Years,” Teknika, vol. 7, no. 2, pp. 73–78, 2018, doi: 10.34148/teknika.v7i2.90.

[10] M. Mia, AFN Masruriyah, and AR Pratama, "The Utilization of Decision Tree Algorithm In Order to Predict Heart Disease," Global Sisfotek Journal, vol. 12, no. 2, p. 138, 2022, doi: 10.38101/sisfotek.v12i2.551.

[11] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, “Application of Random Forest Algorithm to Determine the Quality of Red Wine,” E-Business: Scientific Journal of Economics and Business, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.

[12] H. Suryono, H. Kuswanto, and N. Iriawan, "Rice phenology classification based on random forest algorithm for data imbalance using Google Earth engine," Procedia Computer Science, vol. 197, no. 2021, pp. 668–676, 2021, doi: 10.1016/j.procs.2021.12.201.

[13] AH Nasrullah, “Implementation of Decision Tree Algorithm for Classifying Best-Selling Products,” Scientific Journal of Computer Science, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.

[14] A. Hidayanti, AM Siregar, SAP Lestari, and YC Cahyana, “Covid-19 Case Analysis Model in Indonesia Using Linear Regression and Random Forest Algorithms,” Petir, vol. 15, no. 1, pp. 91–101, 2021, doi: 10.33322/petir.v15i1.1487.

[15] A. Samosir, MS Hasibuan, WE Justino, and T. Hariyono, “Comparison of Random Forest, Naïve Bayes, and K-Nearest Neighbor Algorithms in Classifying Heart Disease Data,” Proceedings of the Darmajaya National Seminar, vol. 1, no. 0, pp. 214–222, 2021.

[16] H. Tantyoko, DK Sari, and AR Wijaya, “Predicting Potential Earthquakes in Indonesia Using Random Forest and Feature Selection Methods,”IDEALIS : InDonEsiA journaL Information System, vol. 6, no. 2, pp. 83–89, 2023, doi: 10.36080/idealis.v6i2.3036.