Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 2 Iss. 1 February 2025, pp: 53-62
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Obesity Prediction Using Machine Learning Algorithms

Hanifatus Syahidah[1*], Novila Irsandi[2],
Adila Nur Ajizah[3] , Amelia[4]

[1,2]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[3,4]Department of Business, Faculty of Economics and Administrative Sciences,
Dicle University, Turkey

E-Mail: [1]12250324302@students.uin-suska.ac.id, [2]12250323414@students.uin-suska.ac.id,
[3]nurajizahadila@gmail.com, [4]ameliabubbletea67@gmail.com

Received Dec 29th 2024; Revised Feb 25th 2025; Accepted Feb 27th 2025; Available Online Feb 28th 2025, Published Feb 28th 2025
Corresponding Author: Hanifatus Syahidah

Abstract

This study aims to develop a prediction model for obesity levels by utilizing five machine learning algorithms, namely K-Nearest Neighbors (K-NN), Naïve Bayes Classifier (NBC), Decision Tree, Random Forest, and Support Vector Machine (SVM). The data used in this study were obtained from Kaggle, consisting of 2111 data with 17 attributes covering lifestyle and demographic factors. The research process involved data collection, pre-processing, data division using the Holdout Split method (70% training data and 30% testing data), and the application of machine learning algorithms. Performance evaluation used accuracy, precision, recall, and F1 score metrics. The results showed that the Random Forest algorithm had the best performance with an accuracy of 92.29%, followed by Decision Tree at 90.54%, K-NN at 83.44%, and NBC and SVM which reached 59.15% and 59.08%, respectively. Confusion matrix analysis revealed that NBC and SVM had difficulty distinguishing certain obesity classes. Based on these findings, it can be concluded that Random Forest is the most effective algorithm in predicting obesity levels. The results of this study are expected to contribute to developing a more accurate obesity prediction system that can be implemented in the real world.

Keyword: Decision Tree, K-NN, NBC, Random Forest, SVM, Obesity Prediction

1. INTRODUCTION

Data mining is the process of exploring and analyzing large datasets to uncover patterns or extract valuable information [1][2]. One of the key techniques in data mining is classification, which aims to categorize data into specific groups based on identified patterns [3][4]. This technique is widely applied in various domains, including disease prediction, market analysis, and risk management. In this study, classification is used to analyze obesity levels based on multiple lifestyle and demographic factors. Obesity, which has become a rapidly growing public health issue [5], serves as the primary focus of this research.

Obesity has become a global epidemic that affects all age groups and backgrounds and can occur in both adults and children [6]. In Latin America, especially in countries such as Mexico, Peru, and Colombia, the prevalence of obesity continues to increase significantly [7][8]. This alarming condition is strongly linked to various chronic diseases, including cardiovascular diseases, type 2 diabetes, hypertension, and other metabolic disorders. Identifying the factors that contribute to obesity is crucial for developing effective prevention strategies [9]. Hence, identifying the factors that lead to obesity is crucial for designing effective prevention strategies.

Obesity and Sarcopenic Obesity (SO) are health conditions that are closely related to body composition, lifestyle, and physical activity [10]. Obesity, which is generally associated with increased body adiposity, can increase the risk of cardiovascular disease and diabetes, and contribute to increased mortality related to these diseases [11]. In this study, the obesity dataset obtained from Kaggle was used to develop a predictive model of obesity rates based on demographic and lifestyle data, such as dietary habits and physical activity levels, with a focus on the cultural diversity of individuals from Mexico, Peru, and Colombia. Previous studies have also shown that OS, which is a combination of sarcopenia and obesity criteria in the elderly, is highly correlated with decreased physical function and the risk of other complications, especially in women. Variables such as waist circumference, absolute grip strength, and body composition (body fat) are important predictors in

DOI: https://doi.org/10.57152/ijatis.v2i1.1869

understanding the risk of OS [12]. Obesity in Indonesia has become an increasing health problem in recent years. Based on data from the Ministry of Health, the prevalence of obesity in the population aged ≥15 years rose from 19.1% in 2007 to 28.9% in 2013, with a significant spike in the adolescent and adult age groups. The main causes of obesity in Indonesia include unhealthy diets, lack of physical activity, and the influence of modern lifestyles that make access to high-calorie foods easier.

To achieve this goal, five machine learning algorithms are used, namely K-Nearest Neighbors (K-NN), Naïve Bayes Classifier (NBC), Decision Tree, Random Forest, and Support Vector Machine (SVM). The selection of these algorithms is based on their respective advantages in handling data with different characteristics. For example, previous studies have shown that algorithms such as SVM and K-Means clustering are effective in classifying obesity based on individual physical and lifestyle data[13]. The performance comparison of five machine learning algorithms, namely K-NN, NBC, Decision Tree, Random Forest, and SVM, is expected to reveal the best method for predicting obesity levels accurately, taking into account various factors, such as physical fitness, demographics, and other relevant factors, which can provide deeper insights into the risk of obesity in diverse populations, as has been shown in previous studies that used similar algorithms to predict obesity levels and identify key risk factors [8].

A study conducted by Mahmut Dirik, 2023 [14]shows various machine learning algorithms were tested to predict obesity levels, with results showing significant accuracy. The Random Forest model recorded the highest accuracy of 95.78%, followed by Logistic Regression with an accuracy of 95.22%. Naive Bayes produced an accuracy of 67.41%, while the fuzzy Neural Network classifier achieved an accuracy of 78.16%. Support Vector Machine recorded an accuracy of 84.23%, and the Decision Table classifier achieved an accuracy of 84.89%. Random Forest showed an accuracy of 87.3%, while the Rough Set classifier achieved an accuracy of 87.83%. Multi-layer Perceptron produced an accuracy of 94.36%, and FURIA achieved an accuracy of 95.07%. These findings suggest that while various algorithms show adequate accuracy, models with higher accuracy such as Random Forest and logistic regression have great potential as tools in the early detection and treatment of obesity.

Another study on predicting the risk of obesity in 2023 by Rajbhoj[15], compared four supervised Machine Learning (ML) classifiers, namely SVM, Decision Tree, Random Forest, and Logistic Regression. The results showed that the Random Forest model achieved 100% accuracy in predicting obesity, while Logistic Regression achieved 97.09% accuracy in another study. The performance metrics used to evaluate the models include accuracy, precision, recall, and F1 score, which provide a comprehensive picture of the effectiveness of each model. Additionally, a confusion matrix was employed to analyze the classification outcomes and identify potential areas for enhancement. In conclusion, the Random Forest model showed the most superior performance in obesity prediction, while the Logistic Regression model also produced good results, although slightly lower.

Based on research by [6][14][15], this study will evaluate the performance of each algorithm by comparing five algorithms, namely, K-NN, NBC, Decision Tree, Random Forest, and SVM. In this study [14], the Random Forest algorithm and Logistic Regression are the superior algorithms in his research. In the research [15], the Random Forest algorithm also showed superiority in predicting obesity while Logistic Regression showed lower results. The results of this study aim to determine which algorithm provides the best results on the obesity dataset and whether Random Forest is still the algorithm with the best performance.

This study has the novelty of comparing five machine learning algorithms to predict obesity, showing that Random Forest has the highest accuracy (92.29%), while NBC and SVM are less optimal. Different from previous studies, this research uses a more diverse dataset and confusion matrix analysis to identify classification challenges. The results are expected to contribute to the development of a more accurate obesity prediction system.

2. MATERIAL AND METHOD

This study requires several stages. Based on Figure 1, the first stage is to review the literature from relevant sources or search for review literature, such as Scopus or internationally indexed research articles in the last five years, and related topics such as obesity, machine learning, and algorithms (K-NN, NBC, SVM, Random Forest, and Decision Tree).

The selection of these five machine learning algorithms K-NN, NBC, Decision Tree, Random Forest, and SVM is based on their diverse characteristics in handling classification problems, particularly in medical and health-related datasets. Previous research compared these algorithms in obesity risk prediction and found that different algorithms performed optimally depending on the dataset structure and feature interactions [3]. K-NN is selected due to its simplicity and effectiveness in pattern recognition. NBC, despite its strong independence assumption, remains a fast and interpretable probabilistic classifier. Decision Tree provides an easy-to-interpret model with good performance in structured datasets. Random Forest enhances Decision Tree's robustness by reducing overfitting through ensemble learning. Meanwhile, SVM is known for its effectiveness in high-dimensional spaces, making it a competitive option despite its computational cost. The inclusion of these five algorithms ensures a comprehensive performance evaluation and allows identification of the most suitable model for obesity prediction.

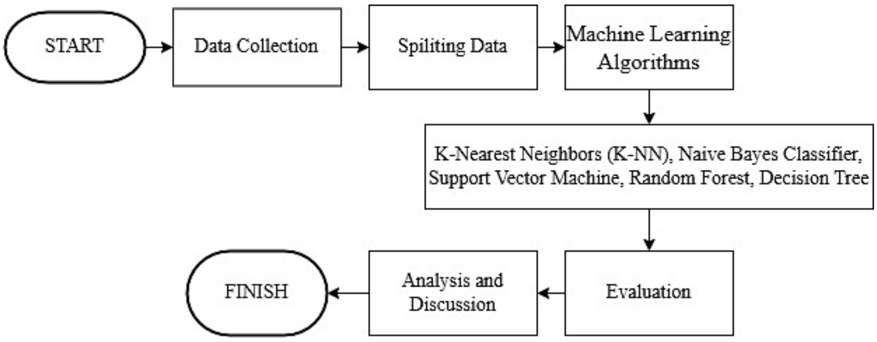
START
Data Collection
Spiliting Data
Machine Learning Algorithms
K-Nearest Neighbors (K-NN), Naive Bayes Classifier, Support Vector Machine, Random Forest, Decision Tree
Evaluation
Analysis and Discussion
FINISH

Figure 1. Research Methodology

The dataset is subsequently classified using five algorithms: K-NN, NBC, SVM, Random Forest, and Decision Tree. The data is split using the Holdout Split technique, which divides the dataset into two subsets: one for training the model and the other for testing its performance, typically in a ratio such as 70:30. The final step involves comparing the accuracy of the five algorithms and analyzing the results.

2.1. Data Collection

Obesity and overweight have been linked to increased risk factors for both morbidity and mortality. In the United States, overweight and obesity contributed to 335,000 deaths and 11.6 million disability-adjusted life years in 2021, making it one of the leading and most rapidly growing risk factors. Overconsumption and obesity not only lead to environmental changes but also impose substantial economic burdens. In 2016, the direct healthcare costs associated with obesity in the United States were estimated to range from $261 billion to $481 billion, and diabetes, a common complication of obesity, has seen an increase of over 140% in the past 30 years[5]. Additionally, the obesity data used in this study, sourced from Kaggle, consists of 2,111 records and 17 attributes related to obesity. The preprocessing phase involved modifying and cleaning the data to make it suitable for machine learning analysis, which included tasks such as removing missing data, changing data formats, and normalizing values. Previous research compared four classification algorithms K-NN, NBC Classifier, SVM, and Decision Tree in predicting obesity risk based on datasets taken from Kaggle. The results showed that the Decision Tree algorithm had the highest accuracy, which was 84.98%, followed by K-NN with an accuracy of 83.55%, NBC with 77.48%, and SVM with 77.32%. The four algorithms were used to classify the "Predicted Obesity" class attribute into "Yes" (obese) and "No" (not obese). The dataset consists of 2111 records with 17 attributes, including data such as frequency of high-calorie food consumption, physical activity, and technology use.

This study compares the performance of five classification algorithms, namely K-NN, NBC Classifier, SVM, Random Forest, and Decision Tree, in predicting obesity risk using datasets from Kaggle. The results showed that the Random Forest algorithm had the highest accuracy of 92.29%, followed by Decision Tree with an accuracy of 90.54%, K-NN with 83.44%, and NBC with 59.15%, and SVM with 59.08%. Based on this evaluation, the Random Forest algorithm proved to be the most effective in handling the dataset and provided the most accurate predictions among the five algorithms.

2.2. Data Preprocessing

The data pre-processing stage was performed on a dataset consisting of 2111 data records and 17 attributes. This dataset has no missing values, so it does not require a data imputation step. Of the 17 available attributes, only 13 attributes were used for further analysis. The attributes that were not used included smoke, CH2O, Age, and Gender. The selection of attributes was based on relevance to the analysis objectives so that attributes that were considered less significant were excluded from the modeling process. This data was also used in research [8] with the title “Obesity Level Prediction Analysis Using Machine Learning and Deep Learning Algorithm Comparison” in 2023.

In addition to attribute selection, the pre-processing process also includes converting categorical data types into numeric representations. This step aims to ensure that the data can be processed by algorithms that require numeric input. Numerical representation is done using techniques such as one-hot encoding or label encoding, depending on the needs and characteristics of the attributes. With this transformation, the data becomes more consistent and ready to be used in the analysis process or predictive model development. This pre-processing stage is very important to ensure optimal data quality and improve the accuracy of the analysis results.

2.3. Splitting Data

In the data splitting stage, the dataset is divided into two parts, namely training data and testing data. This division is done to separate the data that will be used in the model learning process and the data that will be used to evaluate model performance. In this division, a 7:3 ratio is used, where 70% of the data is allocated for training data and the remaining 30% for testing data. Training data is used to build and train the model while testing data is used to measure how well the model can predict data that has never been seen before. With this division, it is expected that the model can have good generalization and provide accurate prediction results when applied to new data.

2.4. Machine Learning Algorithm

Machine Learning is a branch of artificial intelligence that aims to give computers the ability to learn to perform certain tasks even though they have not been explicitly programmed to do so. This method relies on the design of models that learn from data and make decisions or predictions when new data is presented. Artificial Neural Network (ANN), a layered structure, is an evolution of ML. Since the features are extracted automatically, Deep Learning (DL) algorithms require less human involvement. However, DL requires very large data sets to work well, which makes it different from other ML methods. The first computer learning program was written by Arthur Samuel in 1952, and the first neural network was proposed by Frank Rosenblatt in 1957, although ML and DL are new ideas. The development of ML and DL has been significant since the 1990s, mainly due to the increase in computing power and the availability of large amounts of data. Virtual Computers can solve a wide range of problems. We will only examine algorithms that have been used to predict pollutants in this section. We can distinguish those that use regression analysis from those that use artificial neural networks. In addition, we will distinguish between the use of classical regression algorithms and machine learning algorithms in the first category [16].

2.5. K-Nearest Neighbors (K-NN)

K-NN is a method in machine learning that predicts the category or value of data based on other data that is most similar to it [17][18]. When there is new data to be predicted, this algorithm looks at several data that are closest or similar to the data, then decides the category or value of the new data based on what is most often found in the closest data. The more neighbors involved (k value), the more accurate the prediction results, although this method depends on the quality and amount of existing data [19]. To use this method it is necessary to use an equation 1.

$$\mathrm{D(X,Y)} \;= \surd(\textstyle\sum\_(\mathrm{i}=1)\hat{}\,\mathrm{n}\text{▒}(\mathrm{xi}-\mathrm{yi})2) \qquad (1)$$

The Euclidean distance between two data points, denoted as $D(X, Y)$ is calculated based on the square root of the sum of the squared differences between the feature values of each data point. Each squared difference, denoted as (xi−yi)2, represents the difference in the value of the ith feature of data points X and Y. This calculation involves all the features present in each data point, with the total number of features denoted as n. The index i represents each feature used in the distance calculation process.

2.6. Naïve Bayes Classifier (NBC)

NBC is a classification method that uses probability to predict the class of data based on its features[20][21]. This algorithm works by calculating the probability of data belonging to a certain class using Bayes' Theorem, which relates the probability of conditions [22]. Although called “naive” because it assumes that the features are independent, this method often gives very good results in many applications. Naive Bayes is known for its simplicity and speed in processing big data, to use an equation 2.

$$\mathrm{C(X)} = \arg\max\nolimits_{\mathrm{c€C}}\ \mathrm{P\ (c)} \prod\nolimits_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{P\ (ai\ |c)}^{\mathrm{Wi}} \qquad (2)$$

The arg max symbol generally functions to select class c which maximizes the value of all probability expressions. As for P (c) = priori probability or initial probability for class c. And ai is the i-th feature of the data to be classified by P (ai | c)Wi which means conditional probability.

2.7. Support Vector Machine (SVM)

SVM is an algorithm in machine learning that is used for classification and regression tasks [24][25]. SVM works by finding a hyperplane that divides the data into two categories with the widest margin between the two classes [26]. For classification, this hyperplane tries to maximize the distance between data from different classes. The data points closest to the hyperplane are called support vectors, because they determine the position and direction of the hyperplane. SVM is very effective for handling high-dimensional data and is often applied in various fields, such as pattern recognition and text classification [27]. SVM formula can be view equation 3.

$$\underset{\overline{w},b,\overline{\varepsilon}}{\underline{\min}}\ \overline{w}\ .\overline{w}+ C\ i\ .\textstyle\sum_{i=1}^{L} \varepsilon \tag{3}$$

The symbol w is a weight vector that determines the direction and orientation of the hyperplane that separates two classes in the feature space, for b which means the bias that shifts the hyperplane from the origin, then for the symbol C is a regularization parameter that controls the trade-off between maximizing margin and minimizing classification error. And, finally, E is a slack variable for each data.

2.8. Random Forest

Random Forest is a machine learning algorithm employed for both classification and regression tasks. It operates by generating numerous decision trees randomly and then aggregating the predictions from each tree to produce the final result. Each tree is built using a random subset of the training data, with random feature selection at every split in the tree [29]. This approach, called bagging (bootstrap aggregating), reduces the likelihood of overfitting and improves the accuracy of the model [29]. The equation 4 and 5 of this algorithm minimizes the mean square error.

$$\mathrm{Gini(D)} = \textstyle\sum_{k=1}^{k} \mathrm{pk(1 - pk)} \tag{4}$$

$$\mathrm{MSE(D)} = \frac{1}{|D|} \textstyle\sum_{1 \in D} (\mathrm{yi} - \bar{\mathrm{y}}\mathrm{d})^2 \tag{5}$$

For symbol D is a dataset or subset of data being analyzed at a decision tree node, then symbol K is intended for the total number of classes in dataset D, after that Pk is the probability or proportion of data in class k. Next |D| is the total number of data in dataset D. Finally yi for the actual value of data ke=i and yd for the predicted value generated by the model for data ke=i.

2.9. Decision Tree

Decision Tree is a machine learning algorithm that transforms the decision-making process into a tree form. Each node in the tree represents a division of data based on a particular feature, while leaf nodes represent predicted outcomes, such as classes in classification or values in regression [31]. This algorithm divides data into smaller groups based on conditions on features and chooses the most effective division to separate the data. One of the main advantages of decision trees is their ability to produce models that are easy to understand and interpret, and are flexible in handling both numeric and categorical data [32], and view equation 6.

$$\mathrm{G} = 1 - \textstyle\sum_{i=1}^{c} \mathrm{P}_i^2 \tag{6}$$

For the description, the symbol G is the calculated value or metric, Pi is the probability or proportion of data in class I, and C is intended for class or category.

3. RESULTS AND DISCUSSION

The results of this study show the performance of five machine learning algorithms in predicting obesity levels based on lifestyle and demographic data. The dataset used includes 2111 records with 17 attributes, such as eating habits and physical activity. After going through data pre-processing processes, such as attribute selection and transformation of categorical data into numeric, the dataset is divided into two parts using the Holdout Split method with a ratio of 70:30 for training and testing data. The five algorithms are tested based on accuracy, precision, recall, and F1 score to evaluate their effectiveness in classifying obesity levels.

3.1. Implementation of Algorithms

3.1.1. K- Nearest Neighbors (K-NN)

Based on Table 1, the confusion matrix of K-NN provides information about how well the algorithm clusters data based on its closest distance. This algorithm uses a distance metric (e.g., Euclidean distance) to determine the proximity of data to other data. The high recall values in classes 0 and 4 indicate that K-NN can detect these classes with high accuracy. Conversely, the error rates in classes 1 and 5 indicate that the model has difficulty distinguishing between these classes. This may be because the data from these classes are nearby or there are similar patterns between them.

Table1. Confusion Matrix K-NN Algorithm

	True 0	True 1	True 2	True 3	True 4	True 5	True 6	Class Precision
Perd.0	81	17	0	0	0	1	0	94.19%
Perd.1	3	53	1	0	0	7	4	77.94%
Perd.2	0	3	85	2	0	5	7	83.33%
Perd.3	0	0	3	84	1	0	2	95.45%
Perd.4	0	0	1	2	97	0	0	98.98%
Perd.5	2	17	5	0	0	68	5	70.10%
Perd.6	0	3	7	0	0	7	61	78.21%
Class Recall	94.19%	56.99%	83.33%	95.45%	98.98%	77.27%	77.22%	

3.1.2. Naïve Bayes Classifier (NBC) Classifier

Table 2. Confusion Matrix NBC Algorithm

	True 0	True 1	True 2	True 3	True 4	True 5	True 6	Class Precision
Perd.0	70	1	10	0	0	1	0	70.71%
Perd.1	26	32	13	1	1	11	2	60.38%
Perd.2	0	3	55	41	0	2	5	34.16%
Perd.3	0	1	2	84	0	0	2	53.50%
Perd.4	0	0	1	0	96	0	0	98.97%
Perd.5	3	10	40	8	0	24	2	57.14%
Perd.6	0	6	40	23	0	4	14	56.00%
Class Recall	85.37%	37.21%	51.89%	94.38%	98.97%	27.59%	16.09%	

Based on Table 2, the confusion matrix table in the NBC algorithm illustrates the prediction pattern of the probability-based model. Using the principle of feature independence, this model predicts the class by considering the conditional probability of each feature. The good performance in classes 0 and 4 indicates that the model can classify data well in those classes. However, the lower recall in classes 1 and 5 indicates that the model tends to miss some data that should be included in those classes. This may be due to the assumption of feature independence in the NBC algorithm which may not be fully applicable in this dataset.

3.1.3. Support Vector Machine (SVM)

Table 3. Confusion Matrix SVM Algorithm

	True 0	True 1	True 2	True 3	True 4	True 5	True 6	Class Precision
Perd.0	194	78	0	0	0	0	0	68.55%
Perd.1	88	121	0	0	0	61	17	46.92%
Perd.2	0	0	165	64	3	5	114	54.92%
Perd.3	0	1	76	208	12	0	0	69.12%
Perd.4	0	0	0	1	323	0	0	99.69%
Perd.5	0	39	2	0	0	128	121	48.68%
Perd.6	0	6	92	0	0	84	108	50.23%
Class Recall	68.10%	49.59%	45.20%	69.33%	99.69%	31.68%	44.26%	

Based on Table 3, the confusion matrix table of the SVM algorithm illustrates the model's ability to separate data into two or more categories using the optimal hyperplane. SVM tries to maximize the margin between different classes. In this case, the low precision and recall in some classes, such as classes 1, 5, and 6, indicate that the hyperplane may not be able to separate the data effectively. One reason could be the overlapping between these classes. In contrast, the performance in class 4 is quite high, indicating that the hyperplane can separate the data of this class from other classes better.

3.1.4. Random Forest

Based on Table4, it can be seen that the Random Forest algorithm shows the number of correct and incorrect predictions from a collection of decision trees that are combined to improve prediction accuracy. Random Forest works by combining predictions from several decision trees to provide a more stable final result. The high precision and recall values in classes 0 and 4 indicate that the model is very effective in recognizing data patterns in that class. This model also shows advantages in handling varied data and preventing overfitting, thanks to the use of bagging techniques. However, the lower performance in classes 5 and 6 suggests the need for improvement, perhaps through data balancing or adjusting model parameters.

Table 4. Confusion Matrix Random Forest Algorithm

	True 0	True 1	True 2	True 3	True 4	True 5	True 6	Class Precision
Perd.0	253	17	0	0	0	1	0	94.19%
Perd.1	3	53	1	0	0	7	4	77.94%
Perd.2	0	3	85	2	0	5	7	83.33%
Perd.3	0	0	3	84	1	0	2	95.45%
Perd.4	0	0	1	2	97	0	0	98.98%
Perd.5	2	17	5	0	0	68	5	70.10%
Perd.6	0	3	7	0	0	7	61	91.30%
Class Recall	93.01%	87.77%	93.39%	96.94%	100%	84.50%	91.90%	

3.1.5. Decision Tree

Table 5. Confusion Matrix Decision Tree Algorithm

	True 0	True 1	True 2	True 3	True 4	True 5	True 6	Class Precision
Perd.0	83	11	0	0	0	0	0	88.30%
Perd.1	3	74	1	0	0	12	1	80.43%
Perd.2	0	0	89	4	0	0	0	91.75%
Perd.3	0	0	5	84	0	0	1	93.33%
Perd.4	0	0	2	0	98	0	0	97.03%
Perd.5	0	0	0	0	0	73	4	81.11%
Perd.6	0	0	5	0	0	3	73	85.88%
Class Recall	96.51%	85.06%	84.76%	93.33%	100%	81.11%	92.41%	

Based on Table 5, the confusion matrix shows that the model tends to have high accuracy for certain classes, such as class 0 (non-obese) and class 4 (advanced obesity), where most of the predictions are correct. However, for more complex classes, such as classes 5 and 6, there are more significant prediction errors, indicating that the model may have difficulty distinguishing the characteristics of these classes clearly. This could be due to overlapping features or an imbalanced data distribution.

3.2. Accuracy Comparison

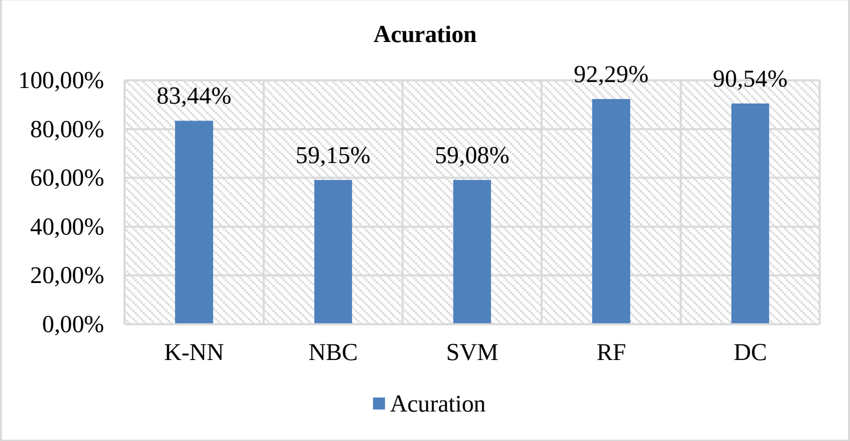

Figure 2. Accuracy of Comparison Algorithm

Based on the research results, the performance of the five classification algorithms K-NN, NBC, SVM, Random Forest, and Decision Tree shows significant variations in accuracy.

1. Random Forest achieved the highest accuracy at 92.29%, making it the best-performing algorithm in this study. The superior performance of Random Forest is likely due to its ensemble approach, which combines multiple decision trees, reducing overfitting and improving the model's generalization ability.
2. Decision Tree ranked second with an accuracy of 90.54%, slightly lower than Random Forest. This difference can be explained by the fact that Decision Tree constructs only a single tree, making it more

prone to overfitting compared to Random Forest, which leverages multiple decision trees for better stability.

3. K-NN achieved an accuracy of 83.44%, which is still relatively high compared to other methods like NBC and SVM. The strong performance of K-NN suggests that the dataset may have a well-defined structure, allowing the distance-based classification approach to work effectively.
4. NBC and SVM showed lower accuracy, at 59.15% and 59.08%, respectively. The nearly identical accuracy values indicate that both algorithms may not be well-suited for the dataset used in this study. NBC's lower performance could be due to violations of the independence assumption among features, while SVM may not have performed optimally due to improper kernel selection or the nature of the dataset not favoring hyperplane-based classification.

From this analysis, it can be concluded that Random Forest and Decision Tree are the best-performing algorithms in this study, with Random Forest being the most effective. K-NN also delivered a reasonably good performance, though not as strong as the tree-based models. Meanwhile, NBC and SVM exhibited lower accuracy, suggesting that probabilistic models like NBC and hyperplane-based models like SVM are less suitable for the given dataset.

The selection of the best algorithm highly depends on the characteristics of the data. If the data structure aligns well with tree-based learning, Random Forest and Decision Tree are the most effective choices. However, for high-dimensional and complex datasets, algorithms such as SVM or Naïve Bayes could perform better with optimal parameter tuning.

3.3. Discussion

This study analyzes the performance of five machine learning algorithms, namely K-NN, NBC, Decision Tree, Random Forest, and SVM, in classifying obesity levels based on demographic and lifestyle data. The dataset used has 2111 entries with 17 attributes and is processed through pre-processing stages, such as selecting relevant attributes and data normalization. From the results of the study, Random Forest proved to be the most reliable algorithm with an accuracy of 92.29%, followed by Decision Tree with an accuracy of 90.54%, and K-NN in third place with 83.44%. In contrast, NBC and SVM performed lower with accuracies of 59.15% and 59.08%, respectively. The advantage of Random Forest lies in its approach that combines the results of several decision trees to produce more stable and accurate predictions while minimizing the risk of overfitting. Although the Decision Tree also showed good performance, this algorithm still faces obstacles in classifying classes with unbalanced data distribution.

To evaluate the effectiveness of the developed model, the results of this study were compared with research conducted [9][10]. In Dirik's study (2023), various machine learning algorithms were tested for obesity prediction, with Random Forest achieving an accuracy of 95.78% and Logistic Regression obtaining an accuracy of 95.22%. Meanwhile, in Rajbhoj's study (2023), Random Forest achieved 100% accuracy, higher than Logistic Regression with 97.09%. These comparisons indicate that the Random Forest model in this study remains competitive, although slightly lower than previous studies. Several factors, including dataset characteristics, the number of attributes, and preprocessing techniques, may contribute to these differences.

On the other hand, NBC and SVM have difficulty in distinguishing certain obesity categories. This is likely due to the assumption of independence between features in NBC, as well as the limitations of the SVM hyperplane in separating data with overlapping patterns. These results are reflected in the confusion matrix, where prediction errors occur more in classes with similar data patterns. Overall, the Random Forest algorithm shows a better ability to handle complex and heterogeneous datasets. However, this study also shows that other algorithms still have the potential to be improved through methods such as data balancing, dimensionality reduction, or model parameter adjustment. The results of this study provide valuable insights into the development of a more accurate obesity prediction system. Further research can be focused on the application of techniques such as oversampling or attribute selection optimization to improve prediction performance, especially in distinguishing obesity categories that are difficult to classify.

4. CONCLUSION

Based on the results of the study, it can be concluded that among the five machine learning algorithms tested, the Random Forest algorithm showed the best performance in predicting obesity levels. With an accuracy of 92.29%, Random Forest proved to be more effective than other algorithms, such as Decision Tree (90.54%), K-NN (83.44%), NBC (59.15%), and SVM (59.08%). The main advantage of Random Forest lies in its ability to manage complex and heterogeneous data through the process of combining predictions from several decision trees. This approach helps reduce overfitting and increase prediction stability. The confusion matrix results reveal that the model can well classify certain obesity classes, especially in the early and advanced obesity classes.

However, algorithms such as SVM and NBC face difficulty distinguishing classes with similar data patterns. This indicates the need for better data distribution management or the introduction of data balancing methods. In conclusion, the Random Forest algorithm is the best choice for predicting obesity levels in the dataset. These findings provide insights that can be utilized by researchers and developers of machine learning-based obesity prediction systems. Future developments can focus on optimizing the model through techniques such as oversampling, dimensionality reduction, or selecting more relevant attributes to improve model performance.

REFERENCES

[1] X. Shu and Y. Ye, “Knowledge Discovery: Methods from data mining and machine learning,” Soc. Sci. Res., vol. 110, no. October 2022, p. 102817, 2023, doi: 10.1016/j.ssresearch.2022.102817.

[2] B. ALTINDIS and F. BAYRAM, “Data mining implementations for determining root causes and precautions of occupational accidents in underground hard coal mining,” Saf. Health Work, no. xxxx, 2024, doi: 10.1016/j.shaw.2024.09.003.

[3] J. C. Macuácua, J. A. S. Centeno, and C. Amisse, “Data mining approach for dry bean seeds classification,” Smart Agric. Technol., vol. 5, no. April 2023, doi: 10.1016/j.atech.2023.100240.

[4] C. Saiprakash, S. R. Kumar Joga, A. Mohapatra, and B. Nayak, "Improved fault detection and classification in PV arrays using Stockwell transform and data mining techniques," Results Eng., vol. 23, no. September, p. 102808, 2024, doi: 10.1016/j.rineng.2024.102808.

[5] U. S. Obesity and F. Collaborators, "Articles National-level and state-level prevalence of overweight and obesity among children, adolescents, and adults in the USA, 1990 – 2021, and forecasts up to 2050," pp. 1–21, 2024, doi: 10.1016/S0140-6736(24)01548-4.

[6] F. Ferdowsy, K. S. A. Rahi, M. I. Jabiullah, and M. T. Habib, “A machine learning approach for obesity risk prediction,” Curr. Res. Behav. Sci., vol. 2, no. May, p. 100053, 2021, doi: 10.1016/j.crbeha.2021.100053.

[7] W. Stroebe, “Is the energy balance explanation of the obesity epidemic wrong?,” Appetite, vol. 188, no. May, p. 106614, 2023, doi: 10.1016/j.appet.2023.106614.

[8] A. I. Putri et al., “Implementation of K-Nearest Neighbors, Naïve Bayes Classifier (NBC) Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction,” Public Res. J. Eng. Data Technol. Comput. Sci., vol. 2, no. 1, pp. 26–33, 2024, doi: 10.57152/predatecs.v2i1.1110.

[9] G. Melo et al., "Structural responses to the obesity epidemic in Latin America: what are the next steps for food and physical activity policies?" Lancet Reg. Heal. - Am., vol. 21, p. 100486, 2023, doi: 10.1016/j.lana.2023.100486.

[10] B. Yu et al., “Sarcopenic obesity is associated with cardiometabolic multimorbidity in Chinese middle-aged and older adults: a cross-sectional and longitudinal study,” J. Nutr. Heal. Aging, vol. 28, no. 10, p. 100353, 2024, doi: 10.1016/j.jnha.2024.100353.

[11] C. Liu et al., “The role of obesity in sarcopenia and the optimal body composition to prevent against sarcopenia and obesity,” Front. Endocrinol. (Lausanne)., vol. 14, no. March, pp. 1–11, 2023, doi: 10.3389/fendo.2023.1077255.

[12] J. H. Bae, J. W. Seo, X. Li, S. Y. Ahn, Y. Sung, and D. Y. Kim, "Neural network model for predicting possible sarcopenic obesity using Korean national fitness award data (2010–2023)," Sci. Rep., vol. 14, no. 1, pp. 1–15, 2024, doi: 10.1038/s41598-024-64742-w.

[13] J. Bae, “Sequential Deep Learning Model for Obesity Prediction Based on Physical Fitness Factors : An Analysis of Data from the 2010 – 2023 Korean National Physical Fitness Data,” pp. 1–20, 2024.

[14] M. Dirik, “Application of machine learning techniques for obesity prediction: a comparative study,” J. Complex. Heal. Sci., vol. 6, no. 2, pp. 16–34, 2023, doi: 10.21595/chs.2023.23193.

[15] Dr. S. M. Rajbhoj, Shweta Shivale, Prof. Vinod. P. Mulik, Sakshi Shirke, and Prof. Amol. P. Yadav, "Obesity Guard: Machine Learning for Early Detection and Prevention," Int. Res. J. Adv. Eng. Hub, vol. 2, no. 07, pp. 2041–2051, 2024, doi: 10.47392/irjaeh.2024.0279.

[16] M. Méndez, M. G. Merayo, and M. Núñez, Machine learning algorithms to forecast air quality: a survey, vol. 56, no. 9. Springer Netherlands, 2023. doi: 10.1007/s10462-023-10424-4.

[17] Y. Chen, P. Tan, M. Li, H. Yin, and R. Tang, “K-means clustering method based on nearest-neighbor density matrix for customer electricity behavior analysis,” Int. J. Electr. Power Energy Syst., vol. 161, no. January 2024, doi: 10.1016/j.ijepes.2024.110165.

[18] A. F. Lubis et al., "Classification of Diabetes Mellitus Sufferers Eating Patterns Using K-Nearest Neighbors, Naïve Bayes Classifier (NBC) and Decision Tree," Public Res. J. Eng. Data Technol. Comput. Sci., vol. 2, no. 1, pp. 44–51, 2024, doi: 10.57152/predatecs.v2i1.1103.

[19] S. S. Shijer, A. H. Jassim, L. A. Al-Haddad, and T. T. Abbas, “Evaluating electrical power yield of photovoltaic solar cells with k-Nearest neighbors: A machine learning statistical analysis approach,” e-Prime - Adv. Electr. Eng. Electron. Energy, vol. 9, no. July, p. 100674, 2024, doi:

10.1016/j.prime.2024.100674.

[20] O. Peretz, M. Koren, and O. Koren, “Naive Bayes classifier – An ensemble procedure for recall and precision enrichment,” Eng. Appl. Artif. Intell., vol. 136, no. PB, p. 108972, 2024, doi: 10.1016/j.engappai.2024.108972.

[21] M. Muta’alimah, C. K. Zarry, A. Kurniawan, H. Hasysya, M. F. Firas, and N. Nadhirah, “Classifications of Offline Shopping Trends and Patterns with Machine Learning Algorithms,” Public Res. J. Eng. Data Technol. Comput. Sci., vol. 2, no. 1, pp. 18–25, 2024, doi: 10.57152/predatecs.v2i1.1099.

[22] A. V. D. Sano, A. A. Stefanus, E. D. Madyatmadja, H. Nindito, A. Purnomo, and C. P. M. Sianipar, “Proposing a visualized comparative review analysis model on tourism domain using Naïve Bayes Classifier (NBC) classifier,” Procedia Comput. Sci., vol. 227, pp. 482–489, 2023, doi: 10.1016/j.procs.2023.10.549.

[23] W. Guo, G. Wang, C. Wang, and Y. Wang, “Distribution network topology identification based on gradient boosting decision tree and attribute weighted naive Bayes,” Energy Reports, vol. 9, pp. 727–736, 2023, doi: 10.1016/j.egyr.2023.04.256.

[24] I. T. Akinola, Y. Sun, I. G. Adebayo, and Z. Wang, "Daily peak demand forecasting using Pelican Algorithm optimized Support Vector Machine (POA-SVM)," Energy Reports, vol. 12, no. June, pp. 4438–4448, 2024, doi: 10.1016/j.egyr.2024.10.017.

[25] W. J. Sari et al., “Performance Comparison of Random Forest, Support Vector Machine and Neural Network in Health Classification of Stroke Patients,” Public Res. J. Eng. Data Technol. Comput. Sci., vol. 2, no. 1, pp. 34–43, 2024, doi: 10.57152/predatecs.v2i1.1119.

[26] Y. tao Zhu, C. Shi Gu, and M. A. Diaconeasa, "A missing data processing method for dam deformation monitoring data using spatiotemporal clustering and support vector machine model," Water Sci. Eng., vol. 17, no. 4, pp. 417–424, 2024, doi: 10.1016/j.wse.2024.08.003.

[27] W. Zhou, H. Liu, R. Zhou, J. Li, and S. Ahmadi, "An optimal method for diagnosing heart disease using a combination of grasshopper evalutionary algorithm and support vector machines," Heliyon, vol. 10, no. 9, p. e30363, 2024, doi: 10.1016/j.heliyon.2024.e30363.

[28] J. S. Pimentel, R. Ospina, and A. Ara, “A novel fusion Support Vector Machine integrating weak and sphere models for classification challenges with massive data,” Decis. Anal. J., vol. 11, no. December 2023, p. 100457, 2024, doi: 10.1016/j.dajour.2024.100457.

[29] J. Du et al., “Maize crop residue cover mapping using Sentinel-2 MSI data and random forest algorithms,” Int. Soil Water Conserv. Res., no. xxxx, 2024, doi: 10.1016/j.iswcr.2024.09.004.

[30] T. Ait tchakoucht, B. Elkari, Y. Chaibi, and T. Kousksou, "Random Forest with feature selection and K-fold cross-validation for predicting the electrical and thermal efficiencies of air-based photovoltaic-thermal systems," Energy Reports, vol. 12, no. March, pp. 988–999, 2024, doi: 10.1016/j.egyr.2024.07.002.

[31] K. Matsumura, K. Hamazaki, H. Kasamatsu, A. Tsuchida, and H. Inadera, “Decision tree learning for predicting chronic postpartum depression in the Japan Environment and Children’s Study,” J. Affect. Disord., vol. 369, no. February 2024, pp. 643–652, 2025, doi: 10.1016/j.jad.2024.10.034.

[32] M. Bagriacik and F. E. B. Otero, “Multiple fairness criteria in decision tree learning,” Appl. Soft Comput., vol. 167, no. PA, p. 112313, 2024, doi: 10.1016/j.asoc.2024.112313.

[33] M. Itzkin, M. L. Palmsten, M. L. Buckley, J. L. Birchler, and L. M. Torres-Garcia, “Developing a decision tree model to forecast runup and assess uncertainty in empirical formulations,” Coast. Eng., vol. 195, no. October 2024, p. 104641, 2025, doi: 10.1016/j.coastaleng.2024.104641.