Cek Turnitin

by 1 1

Submission date: 06-Feb-2025 05:13AM (UTC-0500)
Submission ID: 2581126355
File name: 14478-Article_Text-21108-1-2-20250115_-_Revision.pdf (562.22K)
Word count: 5775
Character count: 33504

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume xx, Number xx, month Year
DOI : https://doi.org/10.33395/sinkron.xxx.xxx
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Performance Comparison of K-Nearest Neighbor, Naive Bayes, and Random Forest Algorithms in Obesity Prediction

Submitted :xxxxx | **Accepted** : xxxx | **Published** : xxxxx

Abstract: Obesity is a growing global health issue that has serious impacts on both physical and mental health. According to the World Health Organization (WHO), over 1.9 billion adults worldwide are overweight, with more than 650 million of them categorized as obese. Early detection of obesity is a crucial step to prevent further complications, however, traditional methods such as Body Mass Index (BMI) have limitations in distinguishing between muscle mass and body fat. This study aims to predict an individual's obesity status based on specific attributes using the K-Nearest Neighbor (K-NN), Naive Bayes, and Random Forest algorithms. The dataset used was sourced from the Kaggle platform, containing 2,111 records and 16 attributes, including gender, age, weight, height, frequency of high-calorie foods, physical activity, and water and vegetable consumption patterns. The research process follows the data mining stages, including business understanding, data understanding, data preparation, modeling, evaluation, and documentation. Experiments were conducted using RapidMiner with a 10-fold cross-validation technique to assess the overall model performance. The results show that the Random Forest algorithm outperforms K-NN and Naive Bayes in predicting obesity status. Model evaluation using metrics such as accuracy, precision, recall, and F1-score demonstrates significant results in distinguishing obesity categories. It is hoped that this research can contribute to developing machine learning-based health prediction systems that can support decision-making in the prevention and management of obesity.

Keywords: Obesity, Data Classification, K-Nearest Neighbor, Naïve Bayes, Data Mining.

INTRODUCTION

Obesity has become an alarming global epidemic. The World Health Organization (WHO) reports that more than 1.9 billion adults worldwide are overweight, with more than 650 million of them categorized as obese. This figure continues to increase every year, including in Indonesia, where the prevalence of obesity in adults reached 21.8% in 2021, increasing sharply from 15.4% in 2013 (Khairina et al., 2022). This condition places individuals at high risk for various chronic diseases, such as type 2 diabetes, cardiovascular disease, and cancer.

Detecting obesity early is a crucial step to prevent further complications (Dahlia et al., 2021). However, the primary challenge lies in enhancing the accuracy of obesity risk detection using clinical data and relevant risk factors. Currently, traditional methods such as body mass index (BMI) are often used, but this approach has limitations, such as the inability to differentiate between muscle mass and body fat. The occurrence of obesity is influenced by food intake that exceeds the body's needs, lack of physical activity, and genetic factors (Toro & Lestari, 2023). Efforts to prevent and manage obesity should focus on promoting healthy lifestyles, providing nutritional education, regulating unhealthy foods and beverages, and supporting individuals in adopting a balanced diet and engaging in sufficient physical activity (Mustafa & Simpen, 2019). Therefore, a more accurate and adaptive approach is needed to detect obesity and assess the risk of complications.

Machine learning technology offers innovative solutions to improve the accuracy of obesity detection. Algorithms such as K-Nearest Neighbor (KNN), Naive Bayes, and Random Forest have been widely used for predictive analysis in the health sector. However, there is still a gap in understanding which algorithms are most effective for certain datasets. For example, KNN has high performance on structured datasets but faces efficiency constraints on large datasets. The K-Nearest Neighbor (KNN) algorithm is a method for classifying new objects based on their (K) nearest neighbors (Samosir et al., 2021). The shortest distance between training data will be measured by Euclidean distance and Manhattan distance (Larassati et al., 2022). Naive Bayes is known to be fast

e-ISSN : 2541-2019
p-ISSN : 2541-044X

in the prediction process but less effective when the relationship between features is not independent. In addition, the Naïve Bayes Classifier algorithm has excellent performance in several text classification cases (Khairina, et al., 2022). Random Forest tends to provide more consistent results but requires greater computing resources. Random Forest builds several decision trees, with output in the form of class mode (classification) or average prediction (regression) for each tree.

This study aims to evaluate and compare the performance of the three algorithms above in detecting the risk of obesity. The urgency of research is very high considering the significant increase in the prevalence of obesity every year. Data from the World Obesity Federation (WOF) estimates that by 2030, 1 in 5 adults in the world will be obese if current trends do not change. Without appropriate intervention, health systems in various countries, including Indonesia, will face increasingly heavy burdens.

LITERATURE REVIEW

A literature study was carried out to ensure that the research methodology design was clear, structured, and based on strong references. The literature study will focus on research relevant to the three machine learning algorithms that will be compared, namely K-Nearest Neighbor (KNN), Naive Bayes, and Random Forest. This literature review includes the implementation of these three algorithms in predicting health problems, especially obesity, as well as evaluating their advantages, disadvantages, and performance based on clinical data and risk factors. Thus, this literature study becomes the basis for designing a comprehensive research approach and supports the aim of evaluating the accuracy of each algorithm (Sinaga et al., 2022).

Various previous studies have examined the performance of machine learning algorithms in various domains. Niar (2022) conducted a comparative study of K-NN, Naive Bayes, and Random Forest for text classification, with the results showing that Random Forest had the highest accuracy of 92%, superior in handling text with complex features. Damanik (2021) evaluated a similar algorithm for medical diagnosis, finding that Random Forest had the highest accuracy (93%) compared to Naive Bayes (88%) and K-NN (85%). On the other hand, Sinaga (2022) shows that Naive Bayes is superior in spam detection with an accuracy of 94%, compared to Random Forest (91%) and K-NN (86%), due to its suitability for text data.

Larassati (2022) compared machine learning algorithms for image classification, where Random Forest again showed superiority with 90% accuracy. In the local context, (Damanik et al., 2021) studied the KNN and SVM algorithms for unemployment prediction in Lampung Province, showing the advantages of the KNN approach in classification based on nearest-neighbor similarity. Toro and Lestari (2023) compared classification algorithms for determining promotion locations for new student admissions, where the decision tree achieved the highest accuracy of 100%, followed by K-NN (99.61%) and Naive Bayes (84.78%).

Niar (2022) used the Naive Bayes and Particle Swarm Optimization (PSO) algorithms for anemia prediction, producing an accuracy of 94.02% when using a combination of Naive Bayes and PSO. (Triana, 2023) compared algorithms for classifying heart disease data and concluded that Naive Bayes provided better and optimal results with an average accuracy of 0.91 AUC. Finally, (Kartika et al., 2022) used Random Forest to predict liver disease, achieving an accuracy of 71.33% with an f1 score of 81%. These studies show that machine learning algorithms have different advantages depending on the type of data and application domain.

Obesity

Obesity is a medical condition characterized by excessive body fat accumulation that can harm an individual's health, with a Body Mass Index (BMI) of 30 or more indicating obesity. As a growing global health issue, obesity increases the risk of chronic diseases such as type 2 diabetes, hypertension, coronary heart disease, and certain cancers. Its causes are multifactorial, involving genetic, environmental, behavioral, and psychological factors that interact complexly. Genetic predispositions influence fat storage, appetite regulation, and metabolism, while environmental factors like poor dietary habits, lack of physical activity, and modern sedentary lifestyles also contribute significantly. Additionally, psychological stress can worsen unhealthy eating behaviors, creating a vicious cycle (Hakim et al., 2017). Obesity impacts not only physical health but also mental and social well-being, often leading to stigma, discrimination, depression, and low self-esteem, which further perpetuate unhealthy lifestyles. Effective management requires a holistic approach, including dietary changes, increased physical activity, cognitive behavioral therapy, and, in some cases, medical interventions such as appetite-suppressing drugs or bariatric surgery for severe cases. Preventive measures should start early, emphasizing healthy eating and physical activity through education, community support, and health-promoting policies. Ongoing research explores new insights into obesity's causes, including epigenetics, hormonal regulation, and gut microbiota, paving the way for personalized therapies. Moreover, obesity imposes significant economic burdens due to healthcare costs and reduced productivity, so investing in prevention and treatment programs is a strategic priority. A

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume xx, Number xx, month Year
DOI : https://doi.org/10.33395/sinkron.xxx.xxx
e-ISSN : 2541-2019
p-ISSN : 2541-044X

comprehensive, multi-sectoral approach involving individuals, families, health institutions, policymakers, and society is essential to effectively reduce obesity prevalence and promote a healthier, more productive population.

K-Nearest Neighbor (K-NN)

The K-Nearest Neighbor (K-NN) algorithm is a supervised learning method commonly used for classification and regression tasks due to its simplicity and effectiveness. It classifies new data points based on the majority class of their k-nearest neighbors, which are identified using distance metrics such as Euclidean, Manhattan, Minkowski, or Hamming distances, depending on the characteristics of the dataset (Aldisa et al., 2022). The algorithm is easy to implement, adaptable to various data types, and does not require assumptions about data distribution. However, its performance is significantly influenced by the choice of the k value. A small k can lead to overfitting, making the model sensitive to noise, while a large k may result in underfitting, causing the classification boundaries to become less distinct. Additionally, K-NN requires considerable computational resources when dealing with large datasets since it calculates distances for every new prediction. In this research, the K-NN algorithm is utilized to predict obesity status based on specific health-related attributes such as age, gender, dietary habits, and physical activity levels (Widyaningrum & Yuliana, 2021). The study aims to evaluate K-NN's performance in this context, particularly in comparison to other algorithms like Naive Bayes and Random Forest. Factors such as accuracy, precision, recall, and F1-score are used to measure the model's effectiveness. Considering that the dataset contains diverse attributes with varying scales, preprocessing steps like normalization are applied to ensure balanced distance calculations. This comparison is crucial to determine K-NN's reliability and efficiency in health-related predictive modeling, especially for early detection and prevention of obesity-related complications.

Naive Bayes

The Naïve Bayes algorithm uses a mathematical branching technique by finding the greatest probability of a classification based on the frequency of each classification against the training data, which is often referred to as probabilistic theory. The calculation formula for Naïve Bayes is as follows:

$$P\left(\frac{X}{Y}\right) = \frac{P\left(\frac{Y}{X}\right) x P(X)}{P(Y)} \qquad (1)$$

In this formula, *P(X)P(X)P(X)* is the initial probability of hypothesis X or how likely it is that the hypothesis is true without considering the existing data. Next, *P(Y|X)P(Y|X)P(Y|X)* is the probability of Y based on the conditions of hypothesis X, which represents how likely data Y is to occur if hypothesis X is true. Finally, *P(Y)P(Y)P(Y)* is the probability of Y or the likelihood of data occurring without considering the hypothesis (Ratnawati & Sulistyaningrum, 2020). This approach allows the algorithm to calculate *P(X|Y)P(X|Y)P(X|Y)*, which is how likely it is that hypothesis X is the given data Y. This probability value is then used to determine the most likely class for data Y. Naive Bayes assumes that all features in the data are independent of each other, which simplifies the calculation but still provides effective results in many applications, including obesity risk prediction (Witata & Triloka, 2023).

Random Forest

Random forest is a classification method that is carried out by developing the Decision Tree method based on the selection of random attributes at each node to determine the classification. The classification process is based on the majority vote from the returned decision tree. Random forest can be built using bagging with random attribute selection (Sari et al., 2023). The CART (Classification and Regression Tree) method is used to grow decision trees. The decision tree grows to its maximum size and will not be pruned. So that a collection of trees is produced which is then called a forest as illustrated in Figure 1. Random Forest is a classification method consisting of a structured collection of decision trees where independent random vectors are distributed identically and each decision tree votes a unit for the most popular class on input x (Ramdhani et al., 2022).

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume xx, Number xx, month Year
DOI : https://doi.org/10.33395/sinkron.xxx.xxx
e-ISSN : 2541-2019
p-ISSN : 2541-044X

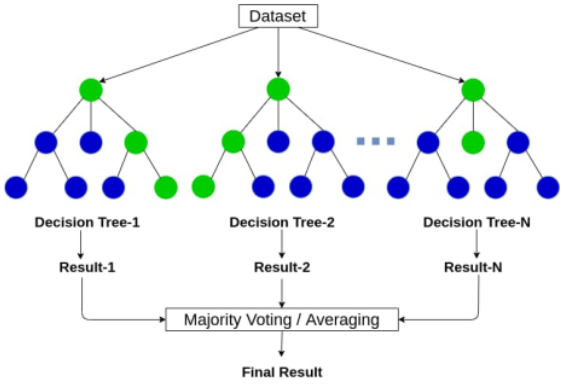

Figure 1. Random Forest Algorithm Model

METHOD

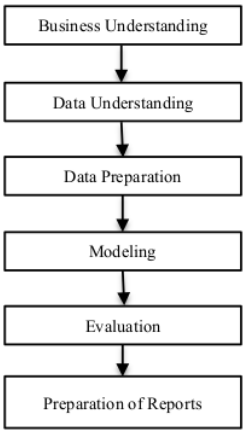

Figure 2. Research Flow Diagram

The stages carried out in this study are shown in Figure 2. The initial stage, business understanding, involves identifying objectives and formulating preliminary strategies. This is followed by data understanding, which includes data collection, exploratory analysis, and evaluation of data quality. Next, the data preparation stage processes raw data into a final dataset ready for use, encompassing variable selection, data transformation, and data cleaning. The modeling stage applies appropriate techniques and calibrations to optimize results, with the flexibility to revisit the data preparation stage if adjustments are needed. Model evaluation is then conducted to ensure the model's effectiveness in achieving research objectives while addressing all key aspects of the problem. Finally, the report preparation stage documents the entire process, research findings, and discussions, presenting them as meaningful literature.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume xx, Number xx, month Year
DOI : https://doi.org/10.33395/sinkron.xxx.xxx
e-ISSN : 2541-2019
p-ISSN : 2541-044X

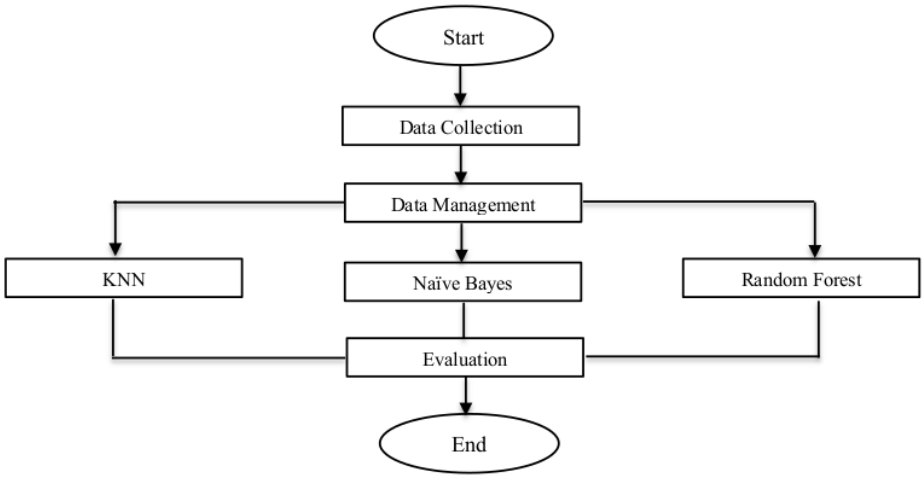

Figure 3. Modeling Flow

The first stage of this research involves data collection, utilizing public data sourced from Kaggle. The dataset used comes from the Kaggle platform (https://www.kaggle.com/code/cahyaalkahfi/klasifikasi-obesitas-dengan-keras-python/input), consisting of 2,111 records. In the data management stage, the main focus is to ensure that the collected data can be processed efficiently, is relevant, and is secure. The dataset includes 16 attributes: Gender, Age, Height, Weight, family_history_with_overweight, FAVC (Frequency of consumption of high-caloric food), FCVC (Frequency of vegetable consumption), NCP (Number of main meals), CAEC (Consumption of food between meals), SMOKE, CH2O (Daily water consumption), SCC (Calorie consumption monitoring), FAF (Physical activity frequency), TUE (Time using technology devices), CALC (Calorie consumption monitoring), and MTRANS (Transportation used). The data processing is conducted using RapidMiner as the primary tool.

The next stage involves the application and testing of methods to achieve the best classification results. The study applies three main algorithms: K-Nearest Neighbors (KNN), Naïve Bayes, and Random Forest. These algorithms are used to train models with the pre-processed dataset, followed by performance evaluation focusing on accuracy metrics as the primary benchmark. The classification results are analyzed to determine which algorithm provides the highest prediction accuracy for obesity classification. The final stage is the evaluation of the results, which involves analyzing the outcomes of the algorithm testing process to assess their effectiveness and identify the most reliable model for obesity prediction.

RESULT

The dataset used uses a dataset from Kaggle https://www.kaggle.com/code/cahyaalkahfi/klasifikasi-obesitas-dengan-keras-r/input with a total of 2111 data using rapid miner tools. The attributes used are 16 attributes, namely Gender, Age, Height, Weight, family_history_with_overweight, FAVC (Frequency of consumption of high-calorie food), FCVC (Frequency of consumption of vegetables), NCP (Number of main meals/Amount of main food), CAEC (Consumption of food between meals), SMOKE, CH2O (Consumption of water daily/ Daily water consumption), SCC (Calories consumption monitoring/ Monitoring calorie consumption), FAF (Physical activity frequency/ Frequency of physical activity), TUE (Time using technology devices), CALC (Calories consumption monitoring/ Monitoring calorie consumption), MTRANS (Transportation used). The methods used in this research are K-Nearest Neighbor, Naive Bayes, and Random Forest. This data set is used to predict whether a patient is likely to be obese.

Selection and application of appropriate modeling techniques is an important step in the data analysis stage. This is done to ensure that the method used can provide accurate prediction results and is relevant to the research objectives. In the context of this research, the focus is on predictions related to obesity diseases. The modeling technique used is data mining, which is the process of extracting significant hidden patterns or useful knowledge from large datasets. In data mining, various algorithms can be used to build K-Nearest Neighbor, Naive Bayes, and Random Forest prediction models.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume xx, Number xx, month Year
DOI : https://doi.org/10.33395/sinkron.xxx.xxx
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The application of data in Rapidminer for Predicting Obesity Using the K-Nearest Neighbor algorithm is shown in Figure 4.

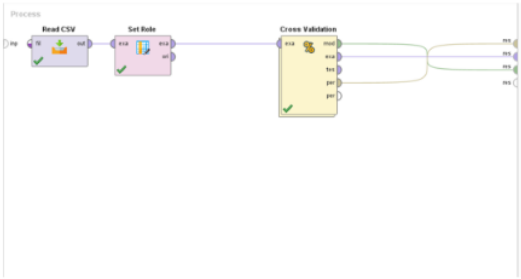

Figure 4. Application of data on Rapidminer for Obesity Disease Prediction

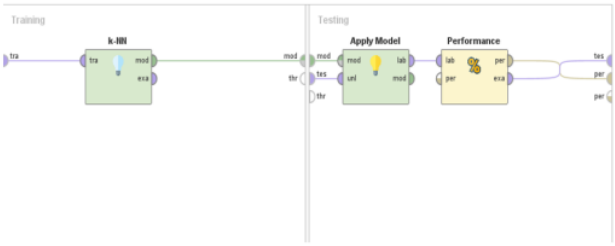

Figure 5. Implementation of Obesity Prediction Using the A-Nearest Neighbor Algorithm on Rapidminer

In Figure 5, the prepared data is implemented in the RapidMiner application to make predictions. This process involves experimentation using the cross-validation method. Cross-validation is a technique commonly used in model evaluation in the fields of data mining and machine learning. This technique directly divides the dataset into two main parts, namely training data and testing data. In this research, the data used is supervised data, which means we have a label or target that we want to predict, namely a person's obesity status based on certain attributes. The algorithm chosen for prediction is K-Nearest Neighbor (K-NN).

In the experimental process using cross-validation, the data is divided into several subsets or folds in this study using Folds 10. Then, the K-NN model is trained using data from these folds alternately, with one fold as testing data and the other fold as training data. This process is carried out repeatedly until all folds become testing data once. The goal of cross-validation is to measure the overall performance of the model by considering variations that may occur in data sharing. Table 1 presents an evaluation of model performance based on accuracy, recall, and precision metrics applied to each class in the analyzed data. This metric is used to assess how well the model can classify data as *True Normal*, *True Overweight*, *True Obesity*, and *True Insufficient*. Accuracy measures the proportion of correct predictions for the entire data, while recall shows the model's ability to detect positive samples in each class.

Table 1.
Model Performance Based on Accuracy, Recall, and Precision for Each Class

Matrix	True Normal	True Overweight	True Obesity	True Insufficient
Akurasi	61,49%			
Class Recall	83,97%	49,66%	62,04%	61,03%

Table 1. shows that the model has an overall accuracy of 61.49%. For recall metrics, the model has the best performance in the class *True Normal* with a recall of 83.97%, which means that most of the data in this category

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume xx, Number xx, month Year
DOI : https://doi.org/10.33395/sinkron.xxx.xxx
e-ISSN : 2541-2019
p-ISSN : 2541-044X

can be detected well by the model. However, the recall is for class *True Overweight* lower, namely 49.66%, which shows that the model has difficulty recognizing data in this category. Class *True Obesity and True Insufficient* have recall values of 62.04% and 61.03% respectively, reflecting moderate model performance in these two classes. Table 2 displays an evaluation of model performance in terms of precision for each prediction class: *Before Normal*, *Pred Overweight*, *Before Obesity*, and Pre *Insufficient*. Precision is a metric used to measure the extent to which a model's positive predictions match the actual class. This metric is important for understanding the accuracy of model predictions in dealing with various categories of data, especially in the context of imbalanced data.

Table 2.
Precision Class Table in Class-Based Prediction

Matrix	Pred Normal	Pred Overweight	Pred Obesity	Pred Insufficient
Class Presisi	27,70%	71,64%	94,66%	82,18%

Table 2 shows that the model has the highest precision in the class *Before Obesity*, amounting to 94.66%, which indicates that positive predictions for this category are almost entirely correct. Precision in class *Pred Insufficient* is also high, namely 82.18%, which reflects the model's good performance in classifying data in this category. Meanwhile, class *Pred Overweight* has a precision of 71.64%, which shows quite good performance. However, precision is for the class *Before Normal* the lowest, namely 27.70%, indicating that many predictions for this category are misclassified into other classes. This can indicate data imbalance or model weakness in recognizing the characteristics of the class.

Research Using Naïve Bayes Algorithm

Table 3 presents a comparison of the performance of two machine learning algorithms, namely Naïve Bayes and Random Forest, in predicting obesity disease. The comparison is based on three main metrics, namely accuracy, recall, and precision, which are evaluated for each class (*True Normal, True Overweight, True Obesity, and True Insufficient*). This assessment aims to determine which algorithm is superior in handling data complexity and providing accurate predictions in each category.

The accuracy results for the two algorithms, Naïve Bayes and Random Forest, as shown in Table 3, were obtained through a series of systematic testing and evaluation processes. The dataset used was sourced from Kaggle, consisting of 2,111 data entries with 16 relevant attributes related to obesity factors. Before applying the algorithms, the data underwent preprocessing steps such as cleaning, normalization, and transformation to ensure optimal model performance. Both Naïve Bayes and Random Forest algorithms were implemented using RapidMiner, with a 10-fold cross-validation technique applied to measure the consistency and reliability of the models. This method divides the dataset into ten equal parts, using nine parts for training and one part for testing in each iteration, ensuring that every data point is used for both training and validation. The accuracy of each model was calculated by comparing the predicted class labels with the actual class labels, and the final accuracy score represents the average performance across all folds. The Random Forest algorithm demonstrated superior accuracy (70.02%) compared to Naïve Bayes (22.08%), likely due to its ability to handle complex data patterns and reduce overfitting through ensemble learning techniques, while Naïve Bayes showed limitations in managing attribute dependencies within the dataset.

Table 3.
Comparison of Naïve Bayes and Random Forest Model Performance in Obesity Disease Prediction Based on Accuracy, Recall, and Precision

Metric	Naïve Bayes	Random Forest
Accuracy	22,08%	70,02%
Class Recall (True Normal)	95,12%	48,43%
Class Recall (True Overweight)	5,86%	36,90%
Class Recall (True Obesity)	4,63%	93,72%
Class Recall (True Insufficient)	41,91%	78,68%
Class Presisi (Pred Normal)	16,62%	64,65%
Class Presisi (Pred Overweight)	32,08%	74,83%
Class Presisi (Pred Obesity)	75,00%	71,34%
Class Presisi (Pred Insufficient)	37,75%	64,26%

Based on Table 3, the Random Forest model has much higher accuracy (70.02%) compared to Naïve Bayes (22.08%), showing better overall performance in predicting obesity. In terms of recall, Random Forest shows

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume xx, Number xx, month Year
DOI : https://doi.org/10.33395/sinkron.xxx.xxx
e-ISSN : 2541-2019
p-ISSN : 2541-044X

consistently higher performance, especially in classes *True Obesity* (93.72%) and *True Insufficient* (78.68%), compared to Naïve Bayes which only reached 4.63% and 41.91% for the two classes. However, Naïve Bayes recorded the highest recall in the class *True Normal* (95.12%) although with overall low accuracy.

For precision metrics, Random Forest also shows superiority over most classes, e.g. *Before Normal* (64.65%) and *Pred Overweight* (74.83%), compared to Naïve Bayes which only reached 16.62% and 32.08% respectively. However, the precision of Naïve Bayes is higher in the class *Before Obesity* (75.00%) compared to Random Forest (71.34%). Overall, Random Forest shows better and more consistent performance than Naïve Bayes in handling obesity disease prediction data, especially in terms of accuracy and recall in classes with data imbalance.

DISCUSSIONS

Table 4 provides a comparative overview of the accuracy levels of three machine learning algorithms, namely K-Nearest Neighbor (KNN), Naïve Bayes, and Random Forest, in predicting obesity based on the data used. The purpose of this analysis is to evaluate the extent to which each algorithm can provide accurate prediction results so that the most effective method can be determined for this case.

Table 4.
Comparison of Accuracy Results of K-Nearest Neighbor, Naive Bayes, and Random Forest Algorithms in Predicting Obesity

Algorithm	Accuracy Level
K-Nearest Neighbor	61,49 %
Naive Bayes	22,08%
Random Forest	70,02%

Table 4 presents a comparison of the accuracy results for three different algorithms: K-Nearest Neighbor (K-NN), Naive Bayes, and Random Forest, based on the experimental findings. The K-Nearest Neighbor (K-NN) algorithm achieves an accuracy rate of 61.49%, classifying new data based on its proximity to existing data points. Naive Bayes, on the other hand, shows a lower accuracy rate of 22.08%. Meanwhile, Random Forest demonstrates the highest accuracy at 70.02%. These experimental results indicate that Random Forest outperforms both K-NN and Naive Bayes in terms of accuracy. The attributes used in this study play a crucial role in determining the accuracy of the models. Based on the analysis, the attributes that contribute the most to the accuracy of the models are as follows:

1. Weight (body weight) and height (height)
   These two attributes have a high correlation with obesity status since obesity is directly defined through the combination of weight and height (e.g., by calculating Body Mass Index/BMI). The model leverages this mathematical relationship to classify the data more accurately.
2. Family_history_with_overweight
   A family history of obesity provides additional information about genetic factors that contribute to obesity. This attribute helps the model identify individuals at higher risk for obesity.
3. FAVC (Frequency of consuming high-calorie foods)
   Consumption of high-calorie foods is one of the main factors causing obesity. This attribute allows the model to predict individual eating patterns that have the potential to increase weight.
4. FAF (Physical activity frequency)
   Physical activity has a negative relationship with obesity. The higher the frequency of physical activity, the less likely a person is to be obese.
5. CH2O (Daily water consumption)
   Water consumption acts as an indicator of a healthy lifestyle. Individuals who consume enough water tend to have a lower risk of obesity.

The implementation data used in the obesity disease prediction experiment was taken from Kaggle, consisting of 2111 data with 16 attributes. The algorithms applied in this research are K-Nearest Neighbor (K-NN), Naive Bayes, and Random Forest, which are processed using RapidMiner software. The experimental process was carried out using the cross-validation method, which divides the dataset into several subsets to train and test the performance of the models built by each algorithm.

The first algorithm evaluated was K-Nearest Neighbor (K-NN). This algorithm classifies data based on the proximity between existing data. Experimental results show that K-NN produces an accuracy of 61.49%, which can be considered quite adequate although not as good as other algorithms such as Random Forest. This algorithm

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume xx, Number xx, month Year
DOI : https://doi.org/10.33395/sinkron.xxx.xxx
e-ISSN : 2541-2019
p-ISSN : 2541-044X

was able to predict the normal class well, producing a recall of 83.97. but the performance is inconsistent in other classes such as Overweight and Obesity.

The second algorithm is Naive Bayes, which assumes that the features in the dataset are independent of each other. However, this assumption often does not match real-world data, so this algorithm showed poor results in this study. The resulting accuracy was only 22.08%, with very poor performance, especially in the Overweight (5.86%) and Obesity (4.63%) class predictions. These results indicate that Naive Bayes is not suitable for the obesity dataset used in this study.

The final algorithm evaluated was Random Forest, a decision tree-based ensemble method. This algorithm gave the best results in the experiment, with the highest accuracy of 70.02%. Random Forest performs excellently in predicting the Obesity class, with a recall reaching 93.72%, and has stable performance in various other classes.

The results of this study demonstrate that Random Forest outperforms K-NN and Naive Bayes in predicting obesity, achieving the highest accuracy of 70.02%. This finding aligns with previous research by Aldisa (2022), which also shows the superiority of Random Forest in handling datasets with complex features. This research supports previous findings that Random Forest is more effective in handling data with complex features. The low performance of Naive Bayes in this study is likely due to its limitations in managing datasets with non-independent features. K-NN demonstrates moderate performance but tends to be less optimal compared to Random Forest. The low performance of Naive Bayes in this study is likely due to its limitations in handling datasets with non-independent features. K-NN shows moderate performance but tends to be less optimal than Random Forest.

CONCLUSION

This research demonstrates the effectiveness of machine learning algorithms—K-Nearest Neighbor (K-NN), Naïve Bayes, and Random Forest—in predicting obesity status based on specific health-related attributes. The study highlights Random Forest as the most robust algorithm, showcasing its superior capability in handling complex datasets and providing stable predictive performance. This finding advances the current state of knowledge by emphasizing the importance of ensemble learning techniques in health prediction models, particularly for conditions with multifactorial causes like obesity. The research contributes to the development of more accurate, data-driven decision-support systems that can assist healthcare professionals in early detection and prevention strategies for obesity-related health risks. Future studies could explore the integration of additional machine learning models, optimize hyperparameter tuning, or incorporate larger, more diverse datasets to enhance predictive accuracy and generalizability across different populations. Additionally, implementing these models in real-world health monitoring systems could provide valuable insights for personalized obesity management and public health interventions.

REFERENCES

Aldisa, R. T., Alfarisi, S., & Abdullah, M. A. (2022). Penerapan Metode Naïve Bayes Dalam Mendiagnosa Penyakit Leptospirosis. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 521–526.

Dahlia, R., Wuryani, N., Hadianti, S., Gata, W., & Selawati, A. (2021). Penerapan Data Mining Terhadap Data Covid-19 Menggunakan Algoritma Klasifikasi. *Jurnal Informatika*, *21*(1), 44–52.

Damanik, A. R., Sumijan, S., & Nurcahyo, G. W. (2021). Prediksi tingkat kepuasan dalam pembelajaran daring menggunakan algoritma Naïve Bayes. *Jurnal Sistim Informasi Dan Teknologi*, 88–94.

Hakim, S. H. F., Cholissodin, I., & Widodo, A. W. (2017). Seleksi Fitur Dengan Particle Swarm Optimization Untuk Pengenalan Pola Wajah Menggunakan Naive Bayes (Studi Kasus Pada Mahasiswa Universitas Brawijaya Fakultas Ilmu Komputer Gedung A). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *1*(10), 1045–1057.

Kartika, Y., Komariah, K., Surip, A., Saputra, R., & Ali, I. (2022). Implementasi Algoritma Naïve Bayes Untuk Prediksi Persediaan Barang Rotan. *KOPERTIP J. Ilm. Manaj. Inform. Dan Komput*, *4*(1), 33–40.

Khairina, N., Sibarani, T. T. S., Muliono, R., Sembiring, Z., & Muhathir, M. (2022). Identification of pneumonia using the K-Nearest neighbors method using HOG Fitur feature extraction. *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, *5*(2), 562–568.

Larassati, D., Zaidiah, A., & Afrizal, S. (2022). Sistem prediksi penyakit jantung koroner menggunakan metode naïve bayes. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *7*(2), 533–546.

Mustafa, M. S., & Simpen, I. W. (2019). Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba. *SISITI: Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *8*(1).

Ramdhani, W., Bona, D., Musyaffa, R. B., & Rozikin, C. (2022). Klasifikasi Penyakit Kangker Payudara Menggunakan Algoritma K-Nearest Neighbor. *Jurnal Ilmiah Wahana Pendidikan*, *8*(12), 445–452.

Ratnawati, L., & Sulistyaningrum, D. R. (2020). Penerapan random forest untuk mengukur tingkat keparahan

e-ISSN : 2541-2019
p-ISSN : 2541-044X

penyakit pada daun apel. *Jurnal Sains Dan Seni ITS*, *8*(2), A71–A77.

Samosir, A., Hasibuan, M. S., Justino, W. E., & Hariyono, T. (2021). Komparasi Algoritma Random Forest, Naïve Bayes dan K-Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung. *Prosiding Seminar Nasional Darmajaya*, *1*, 214–222.

Sari, S. D., Budiman, M. A., Gogo Harahap, R. E., Qonsolanisota, G., Dawami, R., & Alleandra, T. (2023). Peningkatan Pengetahuan dan Deteksi Dini Obesitas pada Remaja di SMA Muhammadiyah 3, Jakarta. *Jurnal SOLMA*, *12*(1), 256–261.

Sinaga, S., Sembiring, R. W., & Sumarno, S. (2022). Penerapan Algoritma Naive Bayes untuk Klasifikasi Prediksi Penerimaan Siswa Baru. *Journal of Machine Learning and Data Analytics*, *1*(1), 55–64.

Toro, R., & Lestari, S. (2023). Perbandingan Algoritma Klasifikasi Untuk Penentuan Lokasi Promosi Penerimaan Mahasiswa Baru Pada IIB Darmajaya Lampung. *Techno. Com*, *22*(1).

Triana, T. (2023). PERANCANGAN APLIKASI BMI CALCULATOR UNTUK MEMPREDIKSI TINGKAT OBESITAS PADA MAHASISWA DENGAN METODE K-NEAREST NEIGHBOR: PERANCANGAN APLIKASI BMI CALCULATOR UNTUK MEMPREDIKSI TINGKAT OBESITAS PADA MAHASISWA DENGAN METODE K-NEAREST NEIGHBOR. *Infokes: Jurnal Ilmiah Rekam Medis Dan Informatika Kesehatan*, *13*(2), 83–89.

Widyaningrum, D. A., & Yuliana, F. (2021). Deteksi Dini Dan Edukasi Tentang Pencegahan Obesitas Di Masa Pandemi Pada Masyarakat Desa Kuwon Kecamatan Karas Kabupaten Magetan. *Jurnal Bhakti Civitas Akademika*, *4*(2), 23–28.

Witata, G. A., & Triloka, J. (2023). Kajian Perbandingan Algoritma KNN Dan SVM Untuk Prediksi Pengangguran Di Provinsi Lampung. *Prosiding Seminar Nasional Darmajaya*, *1*, 218–223.

Cek Turnitin

ORIGINALITY REPORT

17%	15%	15%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Prima Indonesia Student Paper	5%
2	Submitted to University of Muhammadiyah Malang Student Paper	2%
3	Submitted to Arab Open University Student Paper	1%
4	www.coursehero.com Internet Source	1%
5	jomardpublishing.com Internet Source	1%
6	Ashok Kumar, Geeta Sharma, Anil Sharma, Pooja Chopra, Punam Rattan. "Advances in Networks, Intelligence and Computing - International Conference on Networks, Intelligence and Computing (ICONIC-2023)", CRC Press, 2024 Publication	1%
7	www.lindushealth.com Internet Source	1%

8	repository.bsi.ac.id Internet Source	1 %
9	ejournal.kresnamediapublisher.com Internet Source	1 %
10	eitca.org Internet Source	1 %
11	Submitted to RMIT University Student Paper	1 %
12	jurnal.polgan.ac.id Internet Source	1 %
13	www.researchgate.net Internet Source	1 %
14	Submitted to Western Illinois University Student Paper	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		