Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 2 Iss. 1 July 2024, pp: 26-33
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction

Amanda Iksanul Putri[1*], Nur Alfa Husna[2], Neha Mella Cia[3], Muhammad Abdillah Arba[4], Nasywa Rihadatul Aisyi[5], Chintya Harum Pramesthi[6], Abidaharbya Salsa Irdayusman[7]

[1234]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[5]Department of International Trade and Fianance, Faculty of Business and Management Sciences,
Istanbul Sabahattin Zaim University, Turkey
[6]Department of Clinical Medicine, Faculty International Education College,
Yangtze University, China
[7]Department of Medicine, Faculty of Medicine, Ankara Yildirim Beyazit University, Turkey

E-Mail: [1]12150320068@students.uin-suska.ac.id, [2]12150321301@students.uin-suska.ac.id,
[3]12150324630@students.uin-suska.ac.id, [4]12150314731@students.uin-suska.ac.id,
[5]nrihadatulaisyi@gmail.com, [6]Cehapramesthy28@gmail.com, [7]abidaharbyasalsa@yahoo.com

Received Dec 27th 2023; Revised Mar 14th 2024; Accepted Apr 19th 2024
Corresponding Author: Amanda Iksanul Putri

Abstract

An abnormal or excessive build-up of fat that can negatively impact one's health as a result of an imbalance in energy between calories consumed and burnt is known as obesity. The majority of ailments, such as diabetes, heart disease, cancer, osteoarthritis, chronic renal disease, stroke, hypertension, and other fatal conditions, are linked to obesity. Information technology has therefore been the subject of several studies aimed at diagnosing and treating obesity. Because there is a wealth of information on obesity, data mining techniques such as the K-Nearest Neighbors (K-NN) algorithm, Naïve Bayes Classifier, Support Vector Machine (SVM), and Decision Tree can be used to classify the data. The 2111 records and 17 characteristics of obesity data that were received from Kaggle will be used in this study. The four algorithms are to be compared in this study. In other words, using the dataset used in this study, the Decision Tree algorithm's accuracy outperforms that of the other three algorithms K-NN, Naïve Bayes, and SVM. Using the Decision Tree algorithm, the accuracy was 84.98%; the K-NN algorithm came in second with an accuracy value of 83.55%; the Naïve Bayes algorithm came in third with an accuracy rate of 77.48%; and the SVM algorithm came in last with the lowest accuracy value in this study, at 77.32%.

Keyword: Classification, Decision Tree, K-Nearest Neighbors, Naïve Bayes Classifier, Obesity, Support Vector Machine

1. INTRODUCTION

Obesity is one of the most prevalent health issues in the world, and it's frequently linked to thousands of serious illnesses and even death. Owing to its growing risk to coming generations, this illness has emerged as a worldwide health issue [1]. The definition of obesity, as provided by the World Health Organization (WHO), is "an abnormal or excessive accumulation of fat that can impair health." Additionally, it mentions "it is an imbalance" as one of the main reasons why people become overweight and obese. An imbalance in energy between calories burnt and calories taken is the main cause of obesity and overweight [2][3].

An obese body is one that has an excessive quantity of fat on it. Diet is not the only factor that contributes to obesity; genetic and environmental variables may also play a role. Most people assume that obesity just affects the appearance of the body and has little bearing on health, so they don't give it any thought. The sad fact is that obesity is linked to the majority of illnesses. Epidemic diabetes, heart disease, cancer, osteoarthritis, chronic renal disease, stroke, hypertension, and other fatal illnesses are among the disorders linked to obesity [4]. Humans are divided into three classes based on their Body Mass Index (BMI): "underweight," "overweight," and "obese." To calculate it, divide the weight in kg by the height in m^2 [2].

Because of its exceptional capacity to do predictive analysis, machine learning has attracted a lot of interest [5]. Studies conducted recently have demonstrated the effectiveness of machine learning for analyzing high-dimensional datasets when compared to conventional approaches [6]. While machine learning

DOI: https://doi.org/10.57152/predatecs.v2i1.1110

has demonstrated potential in supporting human analysts of diverse genomes and genetics datasets, the acquired data still has to be examined, deciphered, and utilized appropriately [7].

In a study conducted by Lin et al., 2023 [8] showed that machine learning combined with SHAP can predict obesity risk more accurately in overweight people than previous models. Four key risk variables of female gender, systolic blood pressure, waist circumference, and hip circumference were identified using the CatBoost algorithm, which was found to be the most effective. The findings of this study imply that combining machine learning with SHAP can be a useful strategy to determine disease risk factors and aid in the prevention and management of obesity. Adults who are overweight and at high risk of obesity, as determined by the model predictions, should receive priority attention in terms of preventive measures and treatment options such as medication or lifestyle modification. A different research by Jindal et al (2018) [9] suggested utilizing partial least squares, random forests, and generalized linear models in an ensemble machine learning strategy to predict obesity. This method produced accurate findings, with an average prediction value of 89.68% for obesity. Several machine learning methods were employed in the 2019 study by Hammond et al [10] to predict childhood obesity in five-year-olds. Models for binary regression and classification were trained using data from the first two years. With a quite high degree of accuracy, the system was able to identify kids who would be obese. The models employed in this work included LASSO regression to predict continuous BMI values and logistic regression, random forest, and gradient booster models for the categorization of obesity (Low, Medium and High). To obtain the greatest performance out of the models, 100 bootstraps were conducted. The findings demonstrate the potential use of machine learning algorithms as a tool for early childhood obesity prediction.

According to earlier studies on the prediction of obesity status in 2022 by M.F. Anisat et al., the K-Nearest Neighbors (K-NN) algorithm technique offers a comparatively high accuracy of 95.74% [1]. Additionally, research on methods for classifying obesity levels was conducted in 2022 by Garba Salisu The study's findings show that, in terms of accuracy and precision, the Decision Tree method performs better than the Naïve Bayes algorithm [11]. Then, in a 2019 study, E.K. Kandemir examined three algorithms to identify high school students' estimates of their likelihood of obesity. The outcomes demonstrate that when it comes to predicting obesity, the Naïve Bayes algorithm outperforms both the logistic regression and artificial neural network algorithms in terms of accuracy [12]. To forecast the obesity rate in a country based on food sales, Dunstan et al. (2019) [13] collected data from 79 different countries using three different types of machine learning algorithms. The goal was to find food sales that could provide precise information regarding the synergistic characteristics of the category. Their findings verified that the prevalence of obesity could be predicted with absolute error using five categories, in about 60% of the countries studied, 10% (across the entire prevalence range), and less than 20% for 87% of the countries. They found that baked goods and flour were the most important food groups for predicting obesity. For the models they used Random Forest (RF) had the best performance, followed by XGBoost and then Support Vector Machine (SVM).

Based on research [8][9][10], this research will classify obesity datasets to compare four algorithms, namely K-NN, Naïve Bayes Classifier, SVM and Decision Tree Algorithms. In research [1][11][12], the K-NN algorithm, Naïve Bayes Classifier and Decision Tree Algorithms became superior algorithms in their research. Whereas in research [13] the SVM algorithm became the algorithm with the lowest performance in his research. The novelty of this research is to compare four classification algorithms to find out which algorithm is appropriate for obesity data in this study and whether the SVM algorithm is still the algorithm with the lowest performance.

2. MATERIAL AND METHOD

This research uses an experimental design and follows the specific methodology shown in Figure 1.

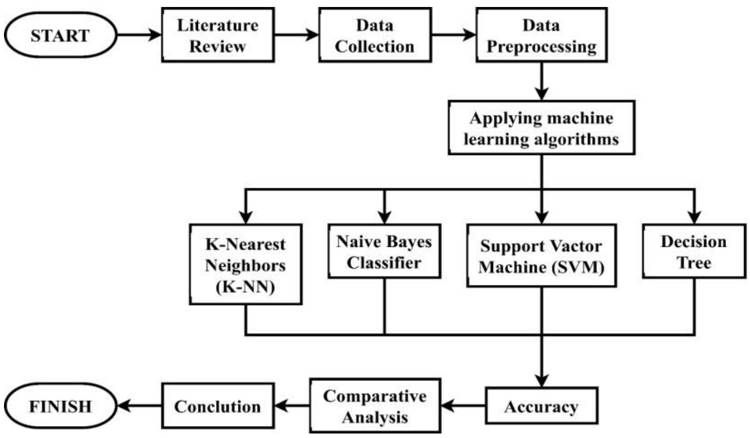

Figure 1. Research Methodology

This study consisted of several stages, including the first stage of reviewing relevant literature and information sources. Sources of information in the form of relevant articles meet the categories including research conducted in the last 5 years, international standard articles, topics related to research (obesity, Machine Learning, K-NN, Naive Bayes, SVM, Decision Tree, etc.). In the second stage, obesity data was collected from individuals aged 14 to 61 years old from Mexico, Peru, and Colombia. The data used in this research comes from the kaggle.com website. Obesity data collected from keggle are 2111 records and 17 attributes.

Furthermore, initial data processing (preprocessing) is carried out, at this stage data processing will be carried out by converting word-shaped data into numerical data. The transformed data is normalized, and the data is then processed using machine learning. In the next stage, the data is classified using the K-Nearest Neighbors (K-NN), Naïve Bayes Classifier, Support Vector Machine (SVM), and Decision Tree algorithms in turn until the accuracy results of each algorithm are compared. The technique for dividing data in this research uses Holdout split by dividing the data into 70:30. The last stage is to compare the accuracy results of the four algorithms that have been processed and analyze this research.

2.1. Obesity

With a complex pathophysiology linked to biological, psychological, socioeconomic, and environmental components as well as variability in the routes and processes causing poor health outcomes, obesity is a multifactorial disease [14]. Obesity or overweight is a condition where there is abnormality or excess fat in individuals who act as one of the factors of diseases that threaten one's health [4], Tables and graphs can be used to determine BMI. The optimal BMI range is 18.5 to 29.9. Adult BMI also indicates dietary status [15]. BMI classification based on the WHO scheme can be seen in table 1 [2].

Table 1. Based on the WHO criteria, BMI is classified as follows: weight in kg/height in meters^2.

Classification	BMI (kg/m^2)	Risk of co-morbidities
B2.5 Underweight	<18.5	Low (although there's a higher chance of further clinical issues)
Normal weight	18.5–24.9	Average
Overweight	25.0–29.9	Mildly increased
Obese	≥30	
Obese I	30.0–34.9	Moderate
Obese II	35.0–39.9	Severe
Obese III	≥40	Very severe

2.2. K-Nearest Neighbors (K-NN)

The K-NN (K-Nearest Neighbor) algorithm is a data categorization technique that determines which group a data point is most likely to belong to in order to determine the likelihood that a data point will join that group[1]. Just one example of a lazy-learning algorithm is K-NN, which only estimates its function locally and doesn't finish all of the computations until the classification stage [16]. Finding groupings of k items in the training data that are most similar to objects in the testing or fresh data is how K-NN is carried out[17]. The steps needed to calculate the K-Nearest Neighbor technique are as follows [16].

1. Determining the parameter K
2. Calculating the distance between training data and testing data

   The most common distance calculation used in calculations on the KNN algorithm is using the Euclidean distance calculation. The formula is as equation 1.

$$euc = \sqrt{(\sum (pi - qi)\ 2\ )\ n\ i{=}1} \tag{1}$$

   Description :

   pi : Training and sample data
   qi : Test information / testing information
   i : Data variable
   n : Data dimension

3. Arrange the created distance.
4. Finding the distance that brings order K closest
5. Assigning matching classes
6. Determine how many classes are in the closest neighbor and designate that class as the data class that will be assessed.

2.3. Naïve Bayes Classifier

The Naïve Bayes technique is based on the Bayes theorem, which calculates the likelihood that each class would correctly anticipate the data[18][19]. This theorem has the benefit of being extremely simple to construct and applicable to big data sets; nevertheless, it has the drawback of assuming that all variables are dependent on one another [20]. Furthermore, estimating the parameters required for the classification process only needs a little quantity of training data when using the Naïve Bayes approach [21]. Equation 2 displays the Bayes theory equation [22].

$$P(B) = P(A)P(A) / P(B) \tag{2}$$

Description :

P(B) : Probability with conditions from A to B
P(A) : Probability conditional on A to B
P(B) : Event probability fom (B)

2.4. Support Vector Machine (SVM)

The Support Vector Machine (SVM) technique is based on the VC dimension of statistical learning and the ideas of structural risk minimization (SRM) [23]. SVM is a supervised learning technique that may identify patterns in data by analyzing it. Regression analysis and classification make use of it [18]. The fundamental concept behind this approach is to choose the optimal separator space from a multiclass dataset. This categorization is done by locating a hyperplane or dividing line that separates one class from another [18]. Equation 3 displays the SVM equation [19].

$$f(x_d) = \sum_{i=1}^{ns} \alpha_i \; y_i x_i x_d + b \tag{3}$$

Description :

ns : Number of support vectors
αi : the weight of each data point
$\vec{y}\, i$: Data class
$\vec{x}\, i$: Support Vector Variable
$\vec{x} d$: Data to be classified
b : Error value

2.5. Decision Tree

A decision tree is a hierarchical tree that is created by partitioning the data into several sets according to input variables as a data mining approach for item classification [24]. Decission tree is further divided into several types including ID3, CART, C4.5 and so on. The ID3 algorithm, created by Quinlan, is extended by the C4.5 algorithm [25]. ID3 is a method used for categorical data, while C4.5 is used for both categorical and numerical data. While C4.5 uses the gain ratio as its attribute selection criterion, the ID3 technique uses information gain as its criterion [26].

The Decision Tree algorithm's phases are as follows [27]:

1. Prepare the training dataset.
2. Locate the decision tree's root.
3. Determine the feature by computing the gain value that will serve as the decision tree's root. The greatest gain value among the available qualities is used to compute gain. The gain value may be computed using the following equation 4.

$$\text{Gain } (S, A) = \text{Entropy}(S) - \sum_{I=1}^{N} \frac{|si|}{|s|} \times \text{Entropy}(Si) \tag{4}$$

4. The procedure for each branch formed, repeat the second step. The other hand, to calculate the entropy value, use the appropriate equation. The equation that can be used the following equation 5.

$$\text{Entropy } (S) = \sum_{i=1}^{N} -\pi \times \log 2\pi \tag{5}$$

5. The decision tree formation process ends when all branches of node N have the same class.

3. RESULTS AND DISCUSSION

3.1. Data Collection

A Kaggle dataset that assesses the obesity rates of people in three countries Mexico, Peru, and Colombia is used in this work. An online survey was used to collect the data, and participants answered questions anonymously. Ten characteristics are as follows: how often high-calorie foods (FAVC) and vegetables (FCVC) are consumed; how many main meals (NCP) and how often food is consumed between meals (CAEC); how much water is consumed daily (CH20); how often physical activity is done (FAF); how often calorie consumption is monitored (SCC); how much time is spent using technology (TUE); and how many vehicles are used (MTRANS). Having determined the four variables (weight, age, height, and gender). The procedure for obtaining the dataset needed to estimate the prevalence of obesity is detailed in Table 2.

Table 2. Obesity rate forecast dataset

Gender	Age	Height	Weight	FAVC	….	NObeyesdad
Male	18	1,73	8,69	Yes	….	No
Male	18,38	1,72	8,47	Yes	….	No
Male	18	1,71	8,44	Yes	….	No
Female	19,43	1,52	8,32	Yes	….	No
Female	19	153	8,31	Yes	….	No
Female	19,63	153	8,25	Yes	….	No
Female	19,94	1,6	8,24	No	….	No

3.2. Implementation Algorithms

The obesity prediction dataset is processed using K-NN, Naïve Bayes Classifier, SVM, and Decision Tree Algorithms to classify patients with obesity disease. "Predicted Obesity" is the class attribute in this dataset. A score of "Yes" denotes obesity, whereas a value of "No" denotes the absence of obesity in the individual.

3.2.1. K-Nearest Neighbors (K-NN)

Table 3 displays the results of the evaluation of the K-NN algorithm's implementation using the confusion matrix.

Table 3. Confusion matrix and K-NN accuracy

Accuracy : 83.55%

	True No	True Yes	Class Precision
Pred. No	277	46	85.76%
Pred. Yes	57	246	81.19%
Class Recall	82.93%	84.25%	

In this case, the K-NN algorithm was tested using data consisting of 246 'Yes' data points and 277 'No' data points. The results show that the algorithm successfully predicted 246 'Yes' data points and 277 'No' data points correctly. There were 46 'Yes' data points that were expected to be 'No,' and 57 'No' data points that were anticipated to be 'Yes,' nevertheless the algorithm was not perfect. The accuracy value obtained from processing obesity data using the K-NN algorithm is 83.55%.

In this research, K-NN is easy to implement and has the ability to handle noisy and multiclass data. However, in this research K-NN has a weakness, namely that it is vulnerable to high dimensions.

3.2.2. Naïve Bayes Classifier

Table 4 displays the evaluation of the Naïve Bayes Classifier algorithm's implementation using the confusion matrix.

Table 4. Confusion matrix and Naïve Bayes Classifier accuracy

Accuracy : 77.48%

	True No	True Yes	Class Precision
Pred. No	212	19	91.77%
Pred. Yes	122	273	69.11%
Class Recall	63.47%	93.49%	

In Table 4 the Naïve Bayes Classifier is tested using data consisting of 212 'Yes' data points and 273 'No' data points. The results show that the algorithm successfully predicted 212 'Yes' and 273 'No' data points correctly. There were 19 'Yes' data points that were expected to be 'No,' and 122 'No' data points that were anticipated to be 'Yes,' indicating that the algorithm was not perfect. The accuracy value obtained from processing obesity data using the Nive Bayes algorithm is 77.48%.

NBC is a simple, fast and efficient classification algorithm. This algorithm is suitable for beginners and situations where data is limited. However, it is not effective if the data used is complex

3.2.3. Support Vector Machine (SVM)

Table 5 displays the results of the evaluation of the SVM algorithm's implementation using the confusion matrix.

Table 5. Confusion matrix and SVM accuracy

Accuracy : 77.32%

	True No	True Yes	Class Precision
Pred. No	254	62	80.38%
Pred. Yes	80	230	74.19%
Class Recall	76.05%	78.77%	

In Table 5 the SVM algorithm is tested using data consisting of 230 'Yes' data points and 254 'No' data points. The results show that the algorithm successfully predicted 230 'Yes' and 254 'No' data points correctly. But the algorithm was not perfect either; 62 'Yes' data points were incorrectly forecasted as 'No,' while 80 'No' data points were incorrectly predicted as 'Yes'. When the SVM technique was used to handle obesity data, the accuracy result was 77.32%.

SVM works well for a variety of regression and classification tasks, particularly when dealing with high-dimensional data with distinct margins. SVM, however, may be challenging to tune and aren't necessarily appropriate for noisy or huge datasets.

3.2.4. Decision Tree

Table 6 displays the results of the evaluation of the Decision Tree algorithm's implementation using the confusion matrix.

Table 6. Confusion matrix and Decision Tree accuracy

Accuracy : 84.98%

	True No	True Yes	Class Precision
Pred. No	324	84	79.41%
Pred. Yes	10	208	95.41%
Class Recall	97.01%	71.32%	

In Table 5 the Decision Tree algorithm is tested using data consisting of 208 'Yes' data points and 324 'No' data points. The results show that the algorithm successfully predicted 208 'Yes' and 324 'No' data points correctly. Ten 'No' data points were forecasted as 'Yes,' and 84 'Yes' data points were predicted as 'No,' indicating that the algorithm was not perfect. 84.98% accuracy was achieved when the Decision Tree method was used to process obesity data.

Decision trees can handle categorical data and are simple to understand. But one needs to be aware of the possibility of bias and overfitting, particularly with big datasets.

3.3. Accuracy Comparison

A comparison of the four algorithms utilized in this study's performance is shown in Table 7.

Table 7. Performance Comparison

No	Algorithm	Accuracy	Recall		Precision	
			True No	True Yes	Pred No	Pred Yes
1	K-NN	83.55%	82.93%	84.25%	85.76%	81.19%
2	Naïve Bayes	77.48%	63.47	93.49%	91.77%	69.11%
3	SVM	77.32%	76.05%	78.77%	80.38%	74.19%
4	Decision tree	84.98%	97.01%	71.23%	79.41%	95.41%

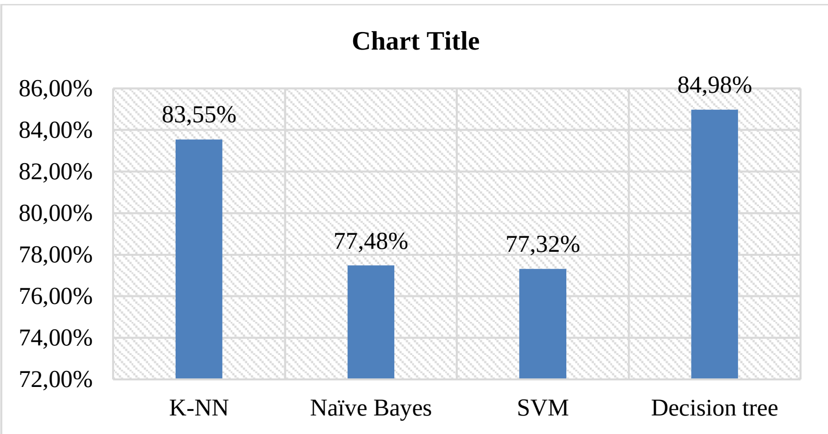

Figure 2. Accuracy Comparison

The study found that the decision tree algorithm had the best performance for predicting obesity in this study. It achieved 84.98% accuracy, making it the most accurate among the four algorithms tested. The second best algorithm in this study, K-NN, has an accuracy of 83.55%, followed by the Naive Bayes algorithm in the third position with an accuracy of 77.48% and followed by the SVM algorithm in the last position whose accuracy is not much different from the Naive Bayes algorithm at 77.32%.

Even though Support Vector Machines (SVM) are very useful for various classification and regression tasks, particularly with high-dimensional data that has clear kelas correlations, SVM may be difficult to fully analyze and possibly not the best choice for large data sets with many noises in them, such as the data set used in this study. Decision tree algorithms are the best in this study because of their strong noise rejection capabilities. This algorithm has the ability to recognize and handle irrelevant input, producing more accurate models.

4. CONCLUSION

According to the study, when applied to an obesity dataset, the Decision Tree approach performed more accurately than the other three algorithms (K-NN, Naïve Bayes, and SVM). The accuracy rate of the Decision Tree algorithm occupies the first position with an accuracy value of 84.98%. With an accuracy rate of 83.55%, the K-NN method took second place, followed by the Naïve Bayes algorithm with an accuracy rate of 77.48%. With an accuracy rate of 77.32%, the SVM method came last in this analysis.

The presence of nois in the data used makes the decision tree algorithm the best algorithm, while the SVM algorithm is the lowest in this study. However, when compared to humans, machine learning algorithms produce more accurate and timely predictions and assessments by quickly analyzing large amounts of data. Machine learning also has the ability to find hidden patterns in the data. Therefore, machine learning research is necessary. To improve the accuracy of algorithms like the one used in this study, the author suggests using optimization algorithms in data management for future research.

REFERENCES

[1] F. Musa, F. Basaky, and O. E.O, “Obesity prediction using machine learning techniques,” J. Appl. Artif. Intell., vol. 3, no. 1, pp. 24–33, 2022, doi: 10.48185/jaai.v3i1.470.

[2] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapi’i, “A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity,” Comput. Biol. Med., vol. 136, no. April, p. 104754, 2021, doi: 10.1016/j.compbiomed.2021.104754.

[3] S. Ibrahim et al., “Overweight and Obesity Prevalence and Predictors in People Living in Karachi,” J. Pharm. Res. Int., vol. 33, no. August 2018, pp. 194–202, 2021, doi: 10.9734/jpri/2021/v33i31b31708.

[4] 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. July, p. 100053, 2021, doi: 10.1016/j.crbeha.2021.100053.

[5] S. Shi et al., “An application based on bioinformatics and machine learning for risk prediction of sepsis at first clinical presentation using transcriptomic data,” Front. Genet., vol. 13, no. September, pp. 1–12, 2022, doi: 10.3389/fgene.2022.979529.

[6] C. Y. Cheng et al., “Evolutionarily informed machine learning enhances the power of predictive gene-to-phenotype relationships,” Nat. Commun., vol. 12, no. 1, pp. 1–16, 2021, doi: 10.1038/s41467-021-25893-w.

[7] M. Rashighi and J. E. Harris, “Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future—Big Data, Machine Learning, and Clinical Medicine,” Physiol. Behav., vol. 176, no. 3, pp. 139–148, 2017, doi: 10.1056/NEJMp1606181.Predicting.

[8] W. Lin, S. Shi, H. Huang, J. Wen, and G. Chen, “Predicting risk of obesity in overweight adults using interpretable machine learning algorithms,” Front. Endocrinol. (Lausanne)., vol. 14, no. November, pp. 1–10, 2023, doi: 10.3389/fendo.2023.1292167.

[9] K. Jindal, N. Baliyan, and P. S. Rana, Obesity prediction using ensemble machine learning approaches, vol. 708, no. January. Springer Singapore, 2018. doi: 10.1007/978-981-10-8636-6_37.

[10] R. Hammond et al., “Correction: Predicting childhood obesity using electronic health records and publicly available data (PLoS ONE (2019) 14:4 (e0215571) DOI: 10.1371/journal.pone.0215571),” PLoS One, vol. 14, no. 10, pp. 1–18, 2019, doi: 10.1371/journal.pone.0223796.

[11] S. Garba, M. Abdullahi, U. A. Umar, and N. T. Wurnor, “Obesity Level Classification Based on Decision Tree and Naïve Bayes Classifiers,” Sule Lamido Univ. J. Sci. Technol., vol. 3, no. 1, pp. 113–121, 2022, doi: https://slujst.com.ng/.

[12] E. Kandemir, Ç. Çavaş, A. E.-U. U. J. Of, and U. 2019, “Comparison of Classifiers for the Risk of Obesity Prediction Among High School Students,” Dergipark.Org.Tr, vol. 2, no. 1, pp. 15–21, 2019, doi: https://dergipark.org.tr/en/pub/uujes/issue/46592/488677.

[13] J. Dunstan, M. Aguirre, M. Bastías, C. Nau, T. A. Glass, and F. Tobar, “Predicting nationwide obesity

from food sales using machine learning," Health Informatics J., vol. 26, no. 1, pp. 652–663, 2020, doi: 10.1177/1460458219845959.

[14] T. M. Powell-Wiley et al., "Obesity and Cardiovascular Disease A Scientific Statement From the American Heart Association," Circulation, vol. 143, no. 21, pp. E984–E1010, 2021, doi: 10.1161/CIR.0000000000000973.

[15] D. Mohajan and H. K. Mohajan, "Body Mass Index (BMI) is a Popular Anthropometric Tool to Measure Obesity Among Adults," J. Innov. Med. Res., vol. 2, no. 4, pp. 25–33, 2023, doi: 10.56397/jimr/2023.04.06.

[16] H. Rajaguru and S. R. Sannasi Chakravarthy, "Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer," Asian Pacific J. Cancer Prev., vol. 20, no. 12, pp. 3777–3781, 2019, doi: 10.31557/APJCP.2019.20.12.3777.

[17] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," Sci. Rep., vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-10358-x.

[18] R. Syahputra, G. J. Yanris, and D. Irmayani, "SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter," Sinkron, vol. 7, no. 2, pp. 671–678, 2022, doi: 10.33395/sinkron.v7i2.11430.

[19] M. F. Fayyad and D. T. Savra, "Sentiment Analysis of Towards Electric Cars using Naive Bayes Classifier and Support Vector Machine Algorithm," Public Res. J. Eng. Data Technol. Comput. Sci., vol. 1, no. July, pp. 1–9, 2023, doi: https://doi.org/10.57152/predatecs.v1i1.814.

[20] H. Yoshikawa, "Can naive Bayes classifier predict infection in a close contact of COVID-19 ? A comparative test for predictability of the predictive model and healthcare workers in Japan," J. Infect. Chemother., vol. 28, no. 6, pp. 774–779, 2022, doi: 10.1016/j.jiac.2022.02.017.

[21] M. Vishwakarma and N. Kesswani, "A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection," Decis. Anal. J., vol. 7, no. April, p. 100233, 2023, doi: 10.1016/j.dajour.2023.100233.

[22] W. Putri, D. Hastari, and K. U. Faizah, "Implementation of Naïve Bayes Classifier f or Classifying Alzheimer ' s Disease Using the K-Means Clustering Data Sharing Technique," Public Res. J. Eng. Data Technol. Comput. Sci., vol. 1, no. July, pp. 47–54, 2023, doi: https://doi.org/10.57152/predatecs.v1i1.803.

[23] Z. C. Dwinnie et al., "Application of the Supervised Learning Algorithm for Classification of Pregnancy Risk Levels," Public Res. J. Eng. Data Technol. Comput. Sci., vol. 1, no. July, pp. 26–33, 2023, doi: https://doi.org/10.57152/predatecs.v1i1.806.

[24] M. R. Anugrah, N. Nazira, N. A. Al-qadr, and N. Ihza, "Implementation of C4 . 5 and Support Vector Machine (SVM) Algorithm for Classification of Coronary Heart Disease," vol. 1, no. July, pp. 20–25, 2023, doi: 10.7910/DVN/76SIQD.

[25] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," J. Appl. Sci. Technol. Trends, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[26] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: A review," Procedia Manuf., vol. 35, pp. 698–703, 2019, doi: 10.1016/j.promfg.2019.06.011.

[27] A. Cherfi, K. Nouira, and A. Ferchichi, "Very Fast C4.5 Decision Tree Algorithm," Appl. Artif. Intell., vol. 32, no. 2, pp. 119–137, 2018, doi: 10.1080/08839514.2018.1447479.