**JURNAL TEKNISI** **E-ISSN: 2775-0965**
**(Jurnal Teknologi Komputer dan Sistem Informasi)**
February 2026, Vol 6, No 1, Hlm. 1 – 8
Available online at http://jurnal.goretanpena.com/index.php/teknisi

# CLASSIFICATION OF OBESITY USING THE NAÏVE BAYES METHOD AND K-NEAREST NEIGHBOR

**Ilham Asy'ari[1] , Muhammad Amin[2] , Andi Saputra[3] , Irianto[4] , Nuriadi Manurung[5]**
[1.3]Computer Science, Universitas Persada Bunda
[2.4]Information Systems , Universitas Royal
[5]Software Engineering, Universitas Persada Bunda
Email: [1]ilham.asyari@upbi.ac.id, [2]mhdamin7@gmail.com , [3]andi.saputra@upbi.ac.id,
[4]irianto2121212@gmail.com ,[5]nuriadi0211@gmail.com

**Abstract:** Obesity is a major health problem that significantly impacts quality of life and can trigger various chronic diseases. Early detection of obesity levels is crucial for public health management, but traditional methods such as BMI often have limitations. Solution: This study proposes a data mining-based approach using feature engineering techniques to improve the accuracy of obesity classification. The purpose of this study is to classify obesity levels and compare the performance of Naïve Bayes and K-Nearest Neighbor (KNN) methods. This research method includes preprocessing stages, feature extraction using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), feature selection using CFS and Chi-Square, classification with Naïve Bayes and KNN, and model evaluation using accuracy and confusion matrix on 2,111 data sets from Kaggle. The results of this study show that on the original data without LDA, KNN achieves a higher accuracy (88.41%) than Naïve Bayes (63.82%). However, after using LDA, the accuracy of Naïve Bayes increased sharply to 93.61%, surpassing KNN's 92.19%. The study concluded that KNN was more effective on raw data, while Naïve Bayes was more optimal when combined with LDA-based dimensionality reduction.

**Keyword:** classification_ naïve_bayes; data mining; k-nearest neighbor; obesity; LDA; PCA.

**Abstract:** Obesitas merupakan salah satu masalah kesehatan utama yang berdampak signifikan pada kualitas hidup dan dapat memicu berbagai penyakit kronis. Deteksi dini tingkat obesitas sangat krusial untuk manajemen kesehatan masyarakat, namun metode tradisional seperti BMI seringkali memiliki keterbatasan. Solusi: Penelitian ini mengusulkan pendekatan berbasis penambangan data (*data mining*) menggunakan teknik rekayasa fitur untuk meningkatkan akurasi klasifikasi tingkat obesitas. Tujuan penelitian ini adalah untuk mengklasifikasikan tingkat obesitas dan membandingkan kinerja metode Naïve Bayes dan K-Nearest Neighbor (KNN). Metode penelitian ini mencakup tahap prapemrosesan, ekstraksi fitur menggunakan *Principal Component Analysis* (PCA) dan *Linear Discriminant Analysis* (LDA), seleksi fitur menggunakan CFS dan *Chi-Square*, klasifikasi dengan Naïve Bayes dan KNN, serta evaluasi model menggunakan *accuracy* dan *confusion matrix* pada 2.111 data dari Kaggle. Hasil penelitian ini menunjukkan bahwa pada data asli tanpa LDA, KNN mencapai akurasi lebih tinggi (88,41%) dibandingkan Naïve Bayes (63,82%). Namun, setelah penggunaan LDA, akurasi Naïve Bayes meningkat tajam menjadi 93,61%, melampaui KNN yang mencapai 92,19%. Kesimpulan dari penelitian ini adalah KNN lebih efektif pada data mentah, sedangkan Naïve Bayes menjadi lebih optimal ketika dikombinasikan dengan reduksi dimensi berbasis LDA.

**Keywords:** klasifikasi naïve bayes; k-nearest neighbor; obesitas; PCA; penambangan data; LDA

**JURNAL TEKNISI** **E-ISSN: 2775-0965**
**(Jurnal Teknologi Komputer dan Sistem Informasi)**
February 2026, Vol 6, No 1, Hlm. 1 – 8
Available online at http://jurnal.goretanpena.com/index.php/teknisi

## INTRODUCTION

Obesity is one of the most serious and rapidly growing health problems worldwide. According to a report by the World Health Organization (WHO, 2022), [1] , the prevalence of obesity has almost tripled since 1975. In Indonesia, the Ministry of Health (2021) reported that more than 21% of adults are classified as obese. [2] . Obesity not only reduces quality of life, but also increases the risk of chronic diseases such as diabetes, hypertension, stroke, and heart disease [3] .

Factors causing obesity include genetic predisposition, unhealthy diet, and a sedentary lifestyle. Early detection of obesity levels is crucial for public health management, diet planning, and chronic disease prevention [4] . Traditional methods such as Body Mass Index (BMI) are often used, but these methods are limited and unable to capture comprehensive lifestyle patterns. Therefore, approaches based on data mining and machine learning are needed. [5] .

Some popular classification algorithms include Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM). Naïve Bayes is simple and efficient, but weak if there is correlation between features [4] . KNN is flexible for complex data, but sensitive to the amount of data and the choice of parameter k. Previous studies have shown mixed results: finding Naïve Bayes to be quite good for obesity prediction [6] , [7] finding KNN to be superior on health data, while other studies have combined feature engineering (PCA, LDA, Chi-Square, CFS) to improve accuracy.

Based on this, this study aims to compare the performance of Naïve Bayes and KNN in obesity classification with additional stages of feature extraction (PCA, LDA) and feature selection (CFS, Chi-Square) [8] .

## METHOD

### Dataset

The research data uses the Obesity Levels Dataset obtained from Kaggle. This dataset contains 2,111 rows of data with 17 independent attributes and 1 target attribute (NObeyesdad), which are divided into 7 categories of obesity levels: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, Obesity Type III. [9]

**Table 1. Dataset Variables**

| Variables | Description |
|---|---|
| Gender | Type sex respondents |
| Age | Age respondents (year) |
| Height | Height (meter) |
| Weight | Heavy body (kg) |
| Family History with Overweight | History family obesity |
| FAVC | Consumption food tall calories |
| FCVC | Frequency consumption vegetable |
| NCP | Amount food main per day |
| CAEC | Consuming snacks outside of hoursEat |
| FAF | Frequency activity physique |
| NObeyesdad | Class obesity (7 category) |

### Research Stages

Collection data

Dataset downloaded from Kaggle in CVS format.

Preprocessing data:

1). Handling missing values for ensure No There is data empty. 2). Label encoding on variables categorical (Gender, Family History, etc.). 3). Normalization so that variables numeric is at on scale uniform.

### Extraction feature

Feature extraction is the process of transforming original data into a new, more meaningful representation.concise However still maintain information important. Objective mainly is reduce dimensions data, minimize redundancy, as well as increase ability algorithm classification in differentiate between class. On study This used two method extraction feature,

that isPCA (Principal Component Analysis) And LDA (Linear Discriminant Analysis).

**PCA (Principal Component Analysis)**

PCA is used to reduce the dimensionality of data by finding new linear combinations of the original features that can explain the largest variance in the data [10] . Thus, PCA helps eliminate redundancy and retain the main information of the dataset. The basic equation of PCA is written as follows:

$$Z= XW( 1)Z = XW \quad (1)$$

Information:
X = data original (data matrix),
W = matrix eigenvector from the matrix data covariance,
Z = new representation projection results to component main.

Intuitively, PCA looks for directions (new axes) where the data has the greatest spread (variance). The new features resulting from PCA are called principal components, which serve as a summary of the data's information with a smaller dimension.

**LDA (Linear Discriminant Analysis):**

In contrast to PCA, which only focuses on data variance without considering class labels, LDA explicitly considers the target class [11] . The goal of LDA is to project data into a low-dimensional space by maximizing the separation between classes.

$$w = \arg\max \frac{w^T S_b w}{w^T S_w w} \quad (2)$$

Information:
Sb = between-class scatter
Sw = within-class scatter
w = projection weight vector.

The interpretation of this equation is:

**Denominator (Sw):** measures how dense the data is in one class → the smaller the value the better.

**Numerator (Sb):** measures how far the center is between classes → the larger the value getting better.

**Feature Selection**

[12] Feature selection is the process of selecting the best subset of features from all available attributes so that only features relevant to the target are retained. The goal is to reduce model complexity, speed up the training process, reduce overfitting, and improve classification accuracy.

CFS (Correlation-based Feature Selection): selects attributes that have high correlation with the target class but low between attributes.

Chi-Square: measures the significance of the relationship between categorical attributes and class labels.

**Classification**

Classification is one of the main techniques in data mining and machine learning that aims to map new data into certain categories or classes based on patterns learned from training data. In this study, two classification algorithms were used, namely Naïve Bayes and K-Nearest Neighbor (KNN) [13]

Naïve Bayes: calculates the posterior probabilities of classes with C the target class and X the input features.

$$P\ (H\backslash X)= \frac{P(X\backslash H).P(H)}{P(X)} \quad (3)$$

K-Nearest Neighbor (KNN): calculates the Euclidean distance between data:

$$\text{dis} = \sqrt{\sum n}(\quad - ) 2 \quad (4)$$
$$i \ =1 \ \ 1 \ \ 1$$

Then, the next step is selecting the majority class from the k nearest neighbors.

**Model evaluation**

Accuracy : percentage of correct predictions.
Confusion Matrix: view prediction details for each obesity class.
10-Fold Cross Validation: divides the dataset into 10 parts for stable evaluation.

**JURNAL TEKNISI** **E-ISSN: 2775-0965**
**(Jurnal Teknologi Komputer dan Sistem Informasi)**
February 2026, Vol 6, No 1, Hlm. 1 – 8
Available online at http://jurnal.goretanpena.com/index.php/teknisi

## RESULTS AND DISCUSSION

**Table 2. Classification Accuracy Results**

| Method | Tampa LDA | With LDA |
|---|---|---|
| Naive Bayes | 63.82% | 93.61% |
| K-Nearest Neighbor | 88.41% | 92.19% |

### Research Results

Naive Bayes Analysis

Naïve Bayes initially only produced an accuracy of 63.82 %, but after using LDA, this increased sharply to 93.61%. This is because LDA successfully reduced the correlation between features, thus better meeting the Naïve Bayes independence assumption.

K-Nearest Neighbor Analysis

KNN outperforms without LDA with an accuracy of 88.41 % , but this drops to 92.19% after using LDA. This is because LDA's data projection changes the distance representation that KNN relies on.

Comparison

The results show that the performance of the algorithms differs depending on the use of LDA feature extraction:

**Without LDA** → KNN achieved the highest accuracy of 92.19 % , far superior to Naïve Bayes' 63.82%. This is because KNN does not require the assumption of independence between features, making it capable of handling obesity data.

**Which has a correlation between variables. With LDA** → Naïve Bayes increased drastically to 93.61%, surpassing KNN which actually decreased to 92.19%. LDA helps Naïve Bayes by reducing the correlation between features so that the independence assumption is more fulfilled. Meanwhile, in KNN.

### Discussion

This section discusses the experimental results in detail, explaining why they emerged and how the findings address the issues raised in the Introduction. The evaluation was conducted using 10-fold cross-validation and focused on performance changes resulting from the application of feature extraction (LDA) and feature selection.

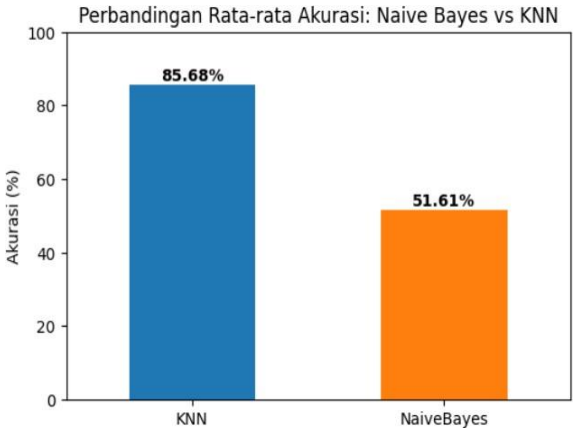


**Image 1.** Chart comparison Naive Bayes And K-Nearest Neighbor

**Table 3. Comparison of the Naive Bayes Algorithm and the K-Nearest Neighbor Algorithm**

| Algorithm | Comparison flat accuracy |
|---|---|
| Naive Bayes | 51.61% |
| K-Nearest Neighbor | 85.68% |

Summary of main results

The main experimental results that form the basis of the discussion are as follows:

Naïve Bayes: 66.82 % (without LDA) → 93.61% (with LDA).

KNN: 88 .41 % (without LDA) → 92.19% (with LDA).

This change shows that feature engineering techniques (especially LDA) have a significant impact on model performance, and this impact differs for each algorithm.

The effect of LDA on model performance

Linear Discriminant Analysis (LDA) searches for a linear projection that maximizes separation between classes. By reducing the dimensionality to features that maximize the between-class/within-class scatter ratio, LDA:

Reducing redundancy and correlation between features that confound models that assume independence. Clarifying the decision boundaries between classes so that simple probability-based models can perform better.

This explains why Naïve Bayes experiences such a large increase in accuracy after LDA: the LDA transformation produces feature representations that better match the assumptions of Naïve Bayes (or at least reduce the violations of these assumptions). In contrast, KNN, which relies on distance metrics in the feature space, shows a smaller improvement — because LDA changes the scale and orientation of the space, so the structure of the relative distances between points can change (sometimes favorably, sometimes not significantly).

a. Discussion of the performance of the naïve Bayes model
   - Before LDA ( 66.82 %): Naïve Bayes underperforms the raw data because many features (e.g., height, weight, diet) are interrelated; violation of the independence assumption leads to biased conditional probability estimates.
   - After LDA ( 93.61 %): LDA projects the data into a lower-dimensional space with clearer separation between classes → the relative feature distribution becomes more discriminatory by class → Naïve Bayes can calculate a more representative likelihood so that accuracy increases drastically.

     Practical implications: when wanting to use a fast, lightweight, and easy-to-implement model (e.g., on a device-based system with limited resources), the combination of LDA + Naïve Bayes is a good choice because it produces high performance with low prediction computational complexity.

b. Discussion of the performance of the k-nearest neighbor model
   - Without LDA ( 88.41 %): KNN already performs better than Naïve Bayes on the original representation because KNN does not require feature independence and is able to capture local patterns in the feature space. However , 88.41 % indicates that there is noise, correlation, and/or class imbalance that limits performance.
   - With LDA ( 92.19 %): LDA slightly improves KNN accuracy (~3.8% improvement). This improvement indicates that the transformation helps clarify class separation, making neighbor distances more informative. However, the improvement is not as significant as with Naive Bayes because KNN already utilizes non-linear/correlated relationships between features.

   Practical note: KNN excels when sufficient training data is available and inference time is not an issue (KNN is computationally expensive in prediction because it calculates distances to the entire training data). If the implementation requires real-time prediction on limited devices, it is worth considering acceleration methods (e.g., kd trees, hashing) or using a computationally less expensive model.

c. Confusion matrix analysis and performance per class. Confusion matrix analysis (see Figure KNN confusion matrix) reveals the following misclassification patterns:
   - The Normal and Obesity Type I classes tend to be well classified (high precision and recall). This indicates that the features that differentiate these classes are quite robust.
   - Errors occur more frequently in the Obesity Type II/III class (severe class), possibly due to: a relatively smaller number of samples per class (class imbalance), overlapping features between the overweight and obesity categories, or a lack of specific medical features that differentiate advanced obesity levels.
   - LDA helps correct some errors by increasing the distance between class clusters, so that recall/precision in classes that were originally ambiguous becomes better — especially seen in Naïve Bayes.

**Table 1. Confuison Matrix Analysis**

| Obesity Class | Success Rate (Recall/Precision) | Observation Notes |
|---|---|---|
| Obesity Class | Success Rate (Recall/Precision) | Features for very low body weight are easily distinguishable. |
| Insufficient Weight | Very High | Possesses robust and strong feature boundaries. |
| Normal Weight | Very High | Frequently misclassified with one another due to thin BMI thresholds. |
| Overweight I & II | Moderate | Successfully classified with high accuracy. |
| Obesity Type I | High | Frequent misclassificatio n occurs due to class imbalance and overlapping features. |

d. Interpreting results against research problems. The issues raised in the Introduction relate to: (1) the effectiveness of classification algorithms for detecting obesity levels, and (2) the role of feature engineering in improving performance. Based on the results:
   - Answer to (1): both KNN and Naïve Bayes can be used; KNN is more reliable on the original representation (its accuracy is higher without transformation), while Naïve Bayes becomes very competitive after applying LDA. So, there is no single answer—the choice of algorithm depends on the preprocessing pipeline.
   - Answer to (2): feature engineering (especially LDA) has a big impact. LDA changes the order of preference of algorithms: without LDA KNN is better; with LDA Naïve Bayes is better. This confirms that model performance is strongly influenced by how features are represented.

e. Limitations of results and validity
   - Class distribution & sample size: Classes with small samples tend to produce less stable metrics. Further analysis requires examining the distribution per class.
   - Hyperparameter tuning: the k value in KNN, the smoothing parameter in Naive Bayes, and the number of components in LDA/PCA can affect the results. Further research should incorporate grid search/optuna.
   - Generalisability: the dataset comes from open sources (Kaggle) — before deployment on local populations, the model needs to be validated on local real data due to differences in demographics and lifestyles .
   - Additional evaluation methods: besides accuracy and confusion matrix, other metrics (precision, recall, F1-score, AUC) need to be reported for a more complete analysis especially in cases of class imbalance.

f. Practical implications and recommendations
   - For fast detection systems (resource-constrained) that require fast inference: use LDA + Naïve Bayes (this combination provides high accuracy and fast prediction).
   - For offline analysis or when large training data is available: KNN on real data can provide good performance but consider computational optimization.
   - To improve the accuracy of the minority class: apply balancing techniques, add relevant features (waist circumference, body fat percentage), or use ensemble models
   - Perform hyperparameter tuning and validation on an external dataset to ensure the model does not overfit.

## CONCLUSION

This research successfully classifies obesity levels using the Naïve Bayes and K-

Nearest Neighbor (KNN) methods with the help of feature engineering techniques in the form of PCA, LDA, CFS, and Chi-Square. On raw data (without LDA), KNN provided better accuracy (88.41 % ) than Naïve Bayes (66.82%). This indicates that KNN is capable of capturing complex data patterns despite the influence of correlation between features. After using LDA, Naïve Bayes improved drastically to 93.61 % , while KNN also increased to 92.19%. Thus, Naïve Bayes became the best algorithm after dimensionality reduction using LDA. These results demonstrate that the choice of classification algorithm is highly dependent on preprocessing and feature representation. KNN excels on raw data, while Naïve Bayes performs better after feature transformation with LDA.

## SUGGESTION

a. Model Development – Further research is recommended to add other algorithms such as Support Vector Machine (SVM), Random Forest, Gradient Boosting, or ensemble learning methods to compare performance more broadly.
b. Advanced Feature Engineering – In addition to PCA and LDA, other methods such as Autoencoder or t-SNE can be tested to see if they can further improve accuracy.
c. Handling Data Imbalance – Use data balancing techniques such as SMOTE or oversampling to improve accuracy in minority classes (e.g., Obesity Type II/III).
d. Real Data Validation – The developed model should be tested on local datasets (e.g. data from hospitals or health centers in Indonesia) to ensure generalizability to real conditions.
e. System Implementation – The results of this research can be implemented into a simple web-based or mobile application to detect obesity levels, thus providing direct benefits to the community.

## BIBLIOGRAPHY

Anisa, D. N., & Jumanto. (2022). Classification of diabetes disease using the Naive Bayes algorithm. *Jurnal Dinamika Informatika*, *14*(1), 33–42.

Aprilita, W. Z., Akbar, R., & Prayogi, R. C. (2023). Comparison of K-Nearest Neighbor (KNN) and Naive Bayes Algorithms in the Classification of Parkinson's Disease. *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*, 188–193.

Argina, A. M. (2020). Application of the K-Nearest Neighbor classification method on a dataset of diabetes patients. *Indonesian Journal of Data Science*, *1*(2), 29–33.

Atmaja, D. M. U. (2019). Application of the K-Nearest Neighbor algorithm. *Jurnal Ilmiah*, *1*, 199–208.

Ikhromr, F. N., Sugiyarto, I., Faddillah, U., & Sudarsono, B. (2023). Implementation of data mining to predict diabetes using Naive Bayes and K-Nearest Neighbor algorithms. *INTECOMS: Journal of Information Technology and Computer Science*, *6*(1).

Ling, J., Kencana, I. P. E. N., & Oka, T. B. (2014). Sentiment analysis using the Naive Bayes classifier method with Chi-Square feature selection. *E-Jurnal Matematika*, *3*(3), 92.

Medea, M. J., Rantung, V. P., & Kembuan, O. (2024). Latent Dirichlet Allocation method in topic modeling of online news headlines about law and crime. *JOINTER: Journal of Informatics Engineering*, *5*(2), 1–7.

Pieters, L. S. (2025). IoT-based pH monitoring of skincare products: A solution for consumer safety. *Edumatic: Journal of Education Informatics*, *9*(1), 236–245.

Ramdan, H., Gunawan, A., & Gunawan, G. (2024). Analysis of cardiovascular effects in Covid-19 cases on obesity using the K-Medoid method. *Indonesian Journal of Computer Science*, *3*(1), 16–24.

Rinanda, P. D., Delvika, B., Nurhidayarnis, S., Abror, N., & Hidayat, A. (2022). Comparison of classification between Naive Bayes and K-Nearest Neighbor on the risk of diabetes in pregnant women. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(2), 68–75.

Sholekhah, F., Putri, A. D., Rahmaddeni, R., & Efrizoni, L. (2024). Comparison of Naive Bayes and K-Nearest Neighbors algorithms for metabolic syndrome classification. *MALCOM: Indonesian Journal of Machine*

*Learning and Computer Science*, *4*(2), 507–514.

Sibi, S. Y., & Widiarti, A. R. (2022). Classification of obesity levels using the KNN algorithm. *Seminar Nasional Corisindo*, *7*(2).

Widiastuti, N. I., Rainarli, E., & Dewi, K. E. (2017). Summarization and support vector machine in document classification. *Jurnal Infotel*, *9*(4), 416.