Journal of Applied Business and Technology

Optimization of Body Mass Index Classification Using Machine Learning Approach for Early Detection of Obesity Risk

Article History
Received
10 February 2025
Received in revised form
30 August 2025
Accepted
15 September 2025
Published Online
30 September 2025

Dewi Nasien [a*], Steven Owen [a], Fenly Fenly [a], Johanes Johanes [a], Frendly Lombu [a], Leo Leo [a], Zirawani Baharum [b]

[a] Faculty of Computer Science, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Indonesia
[b] Technical Foundation Section, Universiti Kuala Lumpur, Malaysia

*Corresponding author
dewinasien@lecturer.pelitaindonesia.ac.id

Abstract
This study aims to optimize the classification of obesity risk at an early stage using Principal Component Analysis (PCA), which is an important technique in machine learning. PCA is used to reduce the dimensionality of data, maintain important information without losing data, and has the advantage of reducing complexity which usually increases the risk of overfitting. The obesity dataset will be classified using algorithms such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting Linear, and XGBoost. Specifically, each algorithm is chosen because of its respective advantages: KNN for nonlinear data, SVM for high-dimensional data, and Random Forest and XGBoost for complex data patterns. Evaluation is carried out using metrics such as accuracy, precision, recall, and F1-score to assess the performance of the algorithm. The results show that the Random Forest and XGBoost algorithms provide the best performance in terms of accuracy, especially when all dataset features are used without PCA reduction. This study is expected to be a consideration in determining the best algorithm for obesity classification, supporting early detection, and facilitating decision making in health analysis.

Keywords: Obesity; PCA; Classification; Machine Learning
DOI: https://doi.org/10.35145/jabt.v6i3.201
SDGs: Good Health and Well-Being (3); Industry, Innovation, and Infrastructure (9); Quality Education (4); Partnerships for the Goals (17)

1.0 INTRODUCTION

Obesity is defined using body mass index (BMI) cutoffs to reflect the associated health risk. Various cutoffs for BMI or the amount or distribution of body fat serve as an indicator of overweight and obesity in descriptive statistics (Dhurandhar, 2022). Obesity contributes to reduced life expectancy, impaired quality of life, and disabilities, mainly in those individuals who develop cardio- vascular diseases, type 2 diabetes, osteoarthritis, and cancer. However, there is a large variation in the individual risk to developing obesity-associated comorbid diseases that cannot simply be explained by the extent of adiposity (Blüher, 2020). This can be triggered by high-calorie foods that are not balanced with physical activity (Klaten, 2022). According to the World Health Organization, (2024), obesity and overweight are age-specific. Overweight is defined as BMI ≥ 25 and obesity as BMI ≥ 30 for adults. In children, this category uses the standard deviation of weight and height from the median according to age.

Obesity has increased significantly over the past few decades and is expected to continue to increase. This condition can lead to various chronic diseases, including diabetes, hypertension, and heart disease, which are among the leading causes of morbidity and mortality worldwide. Therefore, early detection of obesity is very important to support prevention and early intervention efforts. One approach that can be used is data classification, where patient data is analyzed to assess the risk of obesity.

However, health datasets often contain many interrelated variables, which can reduce model accuracy and increase computational complexity. To overcome this challenge, feature extraction methods such as Principal Component Analysis (PCA) have been introduced (Dewi & Pakereng, 2023). PCA can reduce the dimensionality of data without losing important information, simplify analysis, and reduce the risk of overfitting (Baiq Nurul Azmi et al., 2023). The main advantage of PCA is its ability to transform large datasets into simpler dimensions while retaining important information from the original data (Nurdiansyah et al., 2024).

Previous studies have also shown the effectiveness of PCA in various cases, such as identifying worm infection factors in school children (Nur Muhammad Ali Al Faizi et al., 2023) and reducing the dimensionality of data related to tomato fruit quality (Murdika et al., 2021). With its flexibility and reliability, PCA is an appropriate method to support more efficient and accurate classification of obesity data.

These features are used in combination with PCA to classify obesity data sets with machine learning algorithms such as KNN, SVM, Decision Tree, Random Forest, Gradient Boosting Linear, and XGB. These algorithms are chosen because of the advantages of KNN, which is simple and effective for non-linear data; SVM shows higher accuracy in high dimensions; decision trees can be interpreted intuitively; random forest reduces instability and overfitting; while gradient boosting linear and XGB can handle very complex patterns robustly. The best model is achieved from the model comparison, in this case, supporting better obesity health analysis.

2.0 LITERATURE REVIEW

KNN (KNeighborsClassifier)

The KNN algorithm is a method used to classify data based on the shortest distance to the data object (Cholil et al., 2021). The KNN algorithm is generally used to classify objects based on learning data that has a small difference value and the distance of the nearest neighbor to the object (Maskuri et al., 2022). The KNN algorithm is useful when you are performing a pattern recognition test. It classifies a data point based on it's neighbor's classification and stores all available cases (Sawant & Khadapkar, 2022). The KNN algorithm is able to train on obesity datasets to see the negative impact of the loss of imputation values and solutions for healing.

SVC (Support Vector Classification)

The SVC algorithm is an algorithm known as one of the classification methods that has high results in predicting potential classification of data (Hovi et al., 2022). The SVC algorithm is included in one form of method in the SVM algorithm. SVC has a strong theoretical basis and performs more accurate classification than most other algorithms in many applications (Idris et al., 2023).

Decision Tree Classifier

Decision tree is a machine learning technique that uses a hierarchy of sequential structure classification rules by recursively partitioning the training dataset (Nadiah et al., 2022). According to (Permana et al., 2021) the advantage of Decision Tree is its flexible nature so that it can improve the quality of the resulting decisions, while the disadvantage of this algorithm is that there will be overlap if using data that has a very large number of classes and criteria. The selection of suitable attributes in the Decision Tree algorithm is one that allows objects to be divided based on their class. Attributes that are selected heuristically produce the most "purest" node attributes (Pratiwi et al., 2024).

Random Forest Classifier

Random Forest is an ensemble-based algorithm built on the Decision Tree algorithm and is known to have good performance. The ensemble-based algorithm is a combination of several machine learning techniques that are combined into one predictive model. The algorithm is designed to reduce errors, bias, and improve prediction accuracy (Sari et al., 2023). In the traditional RF formulation, each deci- sion tree is randomly created by sampling roughly two-thirds of the training data with replacement while the other third is kept out of training (training data bagging) (Georganos et al., 2021).

Gradient Boosting Classifier

Gradient Boosting Algorithm (GBA) is a machine learning algorithm that is included in the ensemble learning category, where a number of weak models (weak learners) are arranged gradually to form a stronger model. GBA works by reducing the prediction error of the previous model and focusing on unexpected patterns in the data. Specifically, this algorithm uses a weighting approach to each model, placing more emphasis on data that is difficult to explain by the previous model (Septian, 2023).

XGBClassifier

The XGBoost method is a development algorithm from gradient tree boosting based on an ensemble algorithm, which can effectively overcome large-scale machine learning cases (Herni Yulianti et al., 2022). This algorithm adopts an ensemble approach from decision trees, where a number of decision trees are built sequentially to improve the overall performance of the model (Sajiwo et al., 2024).

3.0 METHODOLOGY

In this research stage, several stages are used. Starting from the collection of obesity datasets, then continued to the stages of Exploration data, Preprocessing data, Label Encoding, Feature Selection, Splitting Data, the classification process using the selected classification algorithm and finally the evaluation process.

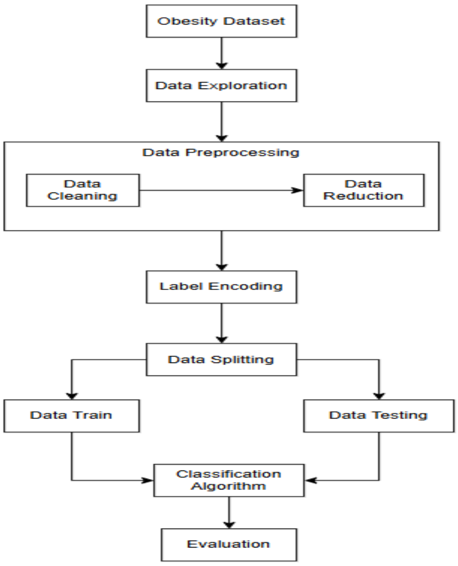


Figure 1. Research Flow

Obesity Dataset

This initial stage involves collecting datasets related to obesity. These datasets contain raw data that will be analyzed later.

Data Exploration

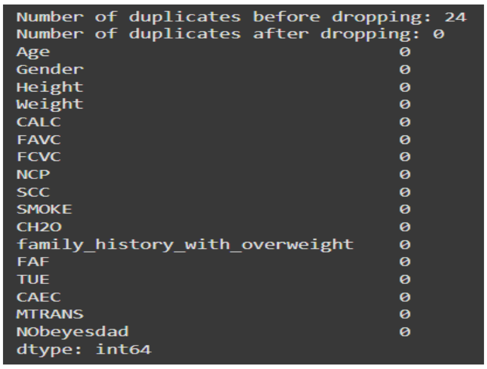
Data Exploration is the first step in data analysis that aims to understand the structure, characteristics, and patterns in the dataset. This process can help in identifying anomalies, patterns, trends, and relationships between variables in the Obesity dataset that can provide in-depth insights into the data. In Data Exploration, it will be done by displaying the form of data visualization in the obesity dataset in order to understand the structure of the dataset.

Data Preprocessing

At this stage, through several stages of Data Cleaning and Data Reduction. This Preprocessing stage cleans unnecessary data such as duplicate words, missing values, and data reduction using PCA.

a) **Data Cleaning**

This process cleans the data from anomalies such as missing values or duplicates that can interfere with the analysis results.

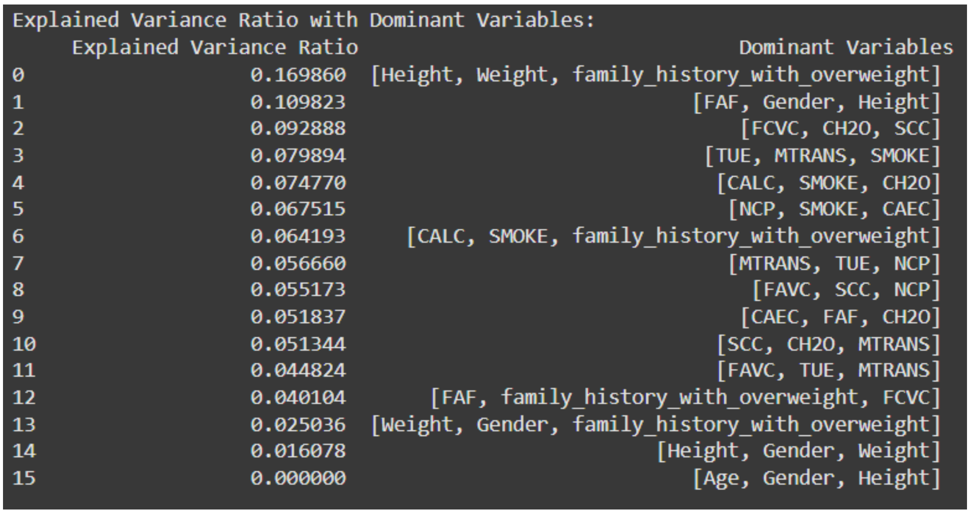


```
Number of duplicates before dropping: 24
Number of duplicates after dropping: 0
Age                               0
Gender                            0
Height                            0
Weight                            0
CALC                              0
FAVC                              0
FCVC                              0
NCP                               0
SCC                               0
SMOKE                             0
CH2O                              0
family_history_with_overweight    0
FAF                               0
TUE                               0
CAEC                              0
MTRANS                            0
NObeyesdad                        0
dtype: int64
```

Figure 2. Data Cleaning

b) Data Reduction

Reducing the number of features or data that are not relevant for further analysis. The goal is to simplify the dataset without reducing important information. In the data reduction process, PCA is used to reduce data in the obesity dataset.



```
Explained Variance Ratio with Dominant Variables:
    Explained Variance Ratio                             Dominant Variables
0                   0.169860  [Height, Weight, family_history_with_overweight]
1                   0.109823                          [FAF, Gender, Height]
2                   0.092888                             [FCVC, CH2O, SCC]
3                   0.079894                           [TUE, MTRANS, SMOKE]
4                   0.074770                            [CALC, SMOKE, CH2O]
5                   0.067515                             [NCP, SMOKE, CAEC]
6                   0.064193    [CALC, SMOKE, family_history_with_overweight]
7                   0.056660                             [MTRANS, TUE, NCP]
8                   0.055173                               [FAVC, SCC, NCP]
9                   0.051837                              [CAEC, FAF, CH2O]
10                  0.051344                            [SCC, CH2O, MTRANS]
11                  0.044824                            [FAVC, TUE, MTRANS]
12                  0.040104       [FAF, family_history_with_overweight, FCVC]
13                  0.025036  [Weight, Gender, family_history_with_overweight]
14                  0.016078                         [Height, Gender, Weight]
15                  0.000000                            [Age, Gender, Height]
```

Figure 3. PCA Results

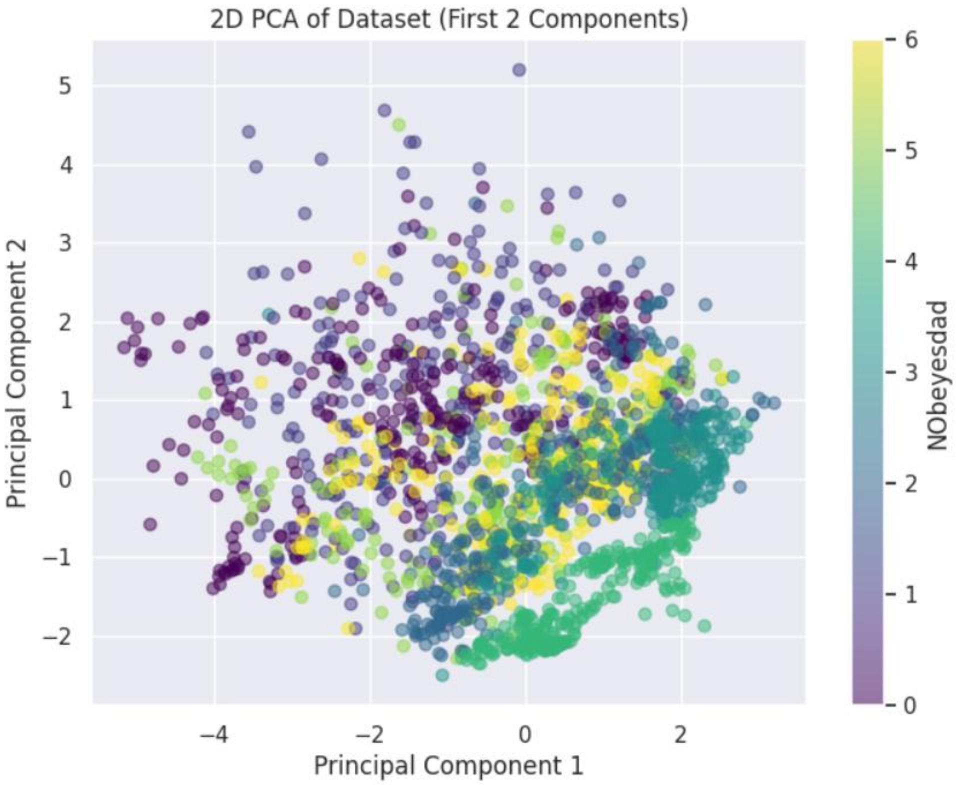


Figure 4. Visualization of PCA Results

Figure 3 showing the ratio of explained variance from principal component analysis (PCA) and the dominant variables contributing to each principal component. Each row in this table shows the ratio of variance explained by a particular principal component and the variables that have the greatest influence on that component, while Figure 4 provides a visualization of the data distribution in the two principal dimensions extracted by PCA, Different colors of the dots indicate different levels or categories of obesity.

c) Label Encoding

Converts categorical variables into numeric form, so that they can be used by machine learning algorithms that require numeric input.

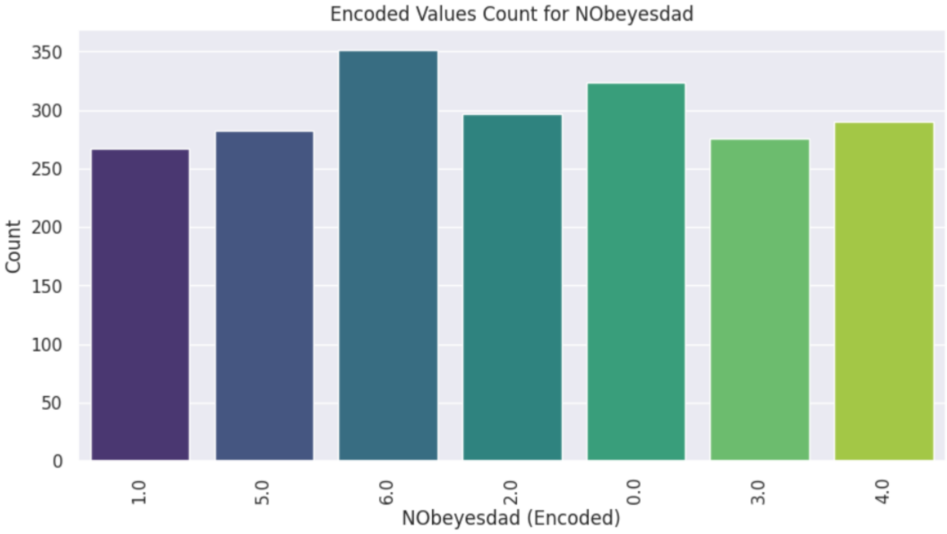


Figure 5. Label Encoding Results of one of the attributes in the obesity dataset

The results of the Label Encoding produce a NObeyesdad variable that is relatively balanced between categories, making it easier to perform a classification model, although some categories are slightly more dominant.

4.0 RESULTS AND DISCUSSION

Splitting Data

In the data splitting process, data is divided by dividing the data by 80:20 to be able to carry out the classification algorithm process.

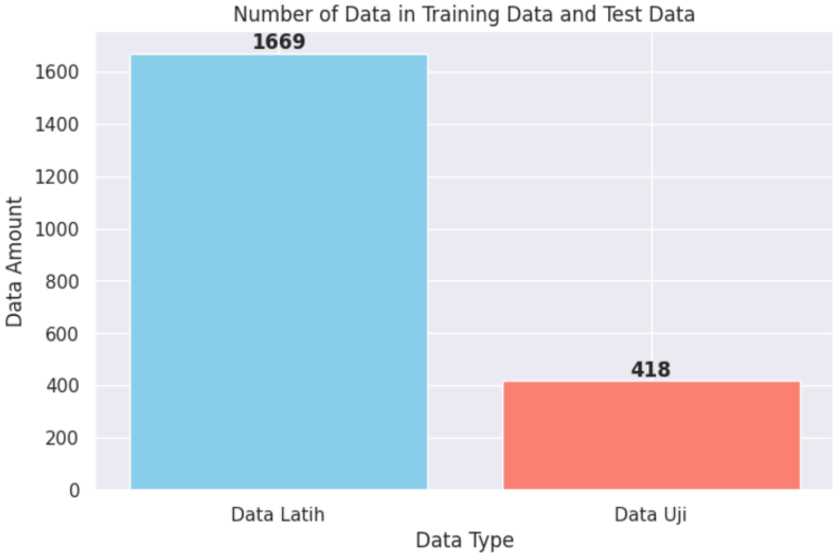


Figure 6. 80:20 Data Splitting Results

Figure 6 illustrates the results of data splitting, which shows the amount of data in the training data is 1669 and the test data is 418. The larger data split for training data aims to provide the model with a lot of information for obesity, while the test data is used to evaluate the model's performance in classifying obesity data.

Classification Algorithm

After splitting the data, the next step is to perform a classification algorithm on the obesity dataset using the KNN (KNeighbors Classifier), SVC (Support Vector Classification), Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, & XGBoost Classifier algorithms using the python programming language with the Google Colab text editor.

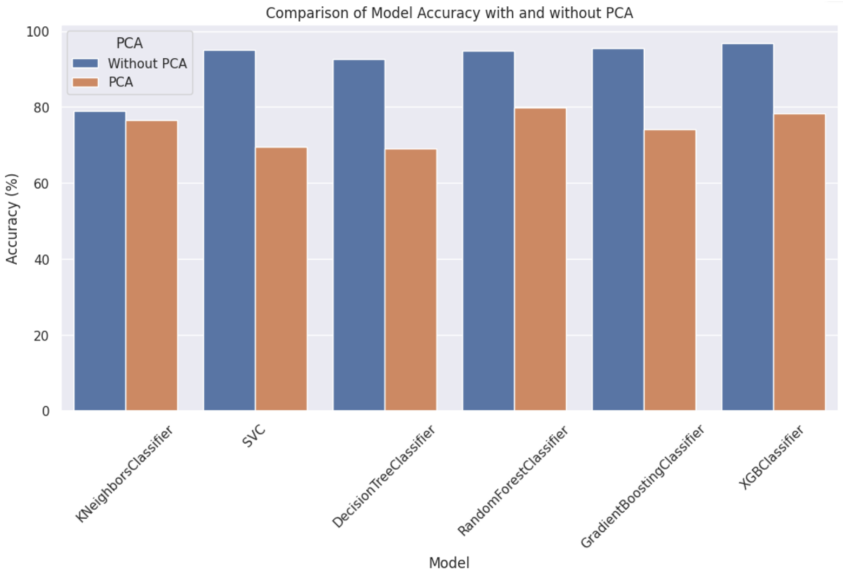


Figure 7. Classification Algorithm Results with PCA and Without PCA

In the figure 7, the accuracy results for Without PCA (Blue Bar) on the Model using all original features from the dataset produce higher accuracy in most models, indicating that information from all features contributes significantly to prediction. Then, the accuracy results With PCA (Orange Bar) on the Model using features reduced to principal components produce accuracy that tends to be lower than without PCA because PCA discards some information to reduce dimensions, although it helps in speeding up computation and reducing overfitting in some cases.

Evaluation

The final step is to evaluate the model performance using testing data. Evaluation methods such as accuracy, precision, recall, or F1-score can be used to assess the effectiveness of the model in classifying obesity data.

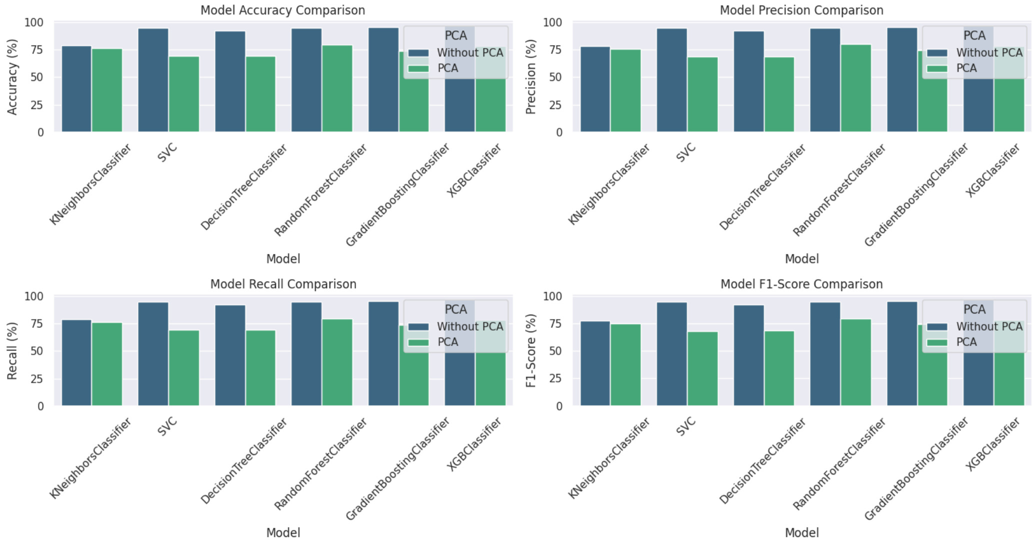


Figure 8. Test Results on the Classification Algorithm

In Figure 8, the test results show a comparison of the performance of several classification models, such as KNN, SVC, Decision Tree, Random Forest, Gradient Boosting, and XGB, in predicting obesity, with and without the application of PCA (Principal Component Analysis). PCA has varying impacts on each evaluation metric, including accuracy, precision, recall, and F1-score. In some models, such as Random Forest and Gradient Boosting, the application of PCA improves performance, especially in precision and F1-score, while other models, such as KNN, show relatively stable results.

5.0 CONCLUSION

Based on the research results in the journal, the Random Forest and XGBoost algorithms show more dominant performance in determining whether someone is obese or not. Both algorithms excel in handling complex datasets, including after the application of Principal Component Analysis (PCA), which functions to reduce data dimensions. Random Forest provides stable and accurate predictions with an ensemble approach from decision trees, while XGBoost offers high efficiency and the ability to handle complex data patterns. Although PCA helps speed up computation and reduces the risk of overfitting, the results of the study show that classification accuracy is higher when using all features without PCA. Therefore, for maximum accuracy in obesity classification, it is recommended not to use PCA and utilize all features of the dataset.

References

Baiq Nurul Azmi, Arief Hermawan, & Donny Avianto. (2023). Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *4*(4), 281–290. https://doi.org/10.35746/jtim.v4i4.298

Blüher, M. (2020). Metabolically healthy obesity. *Endocrine Reviews*, *41*(3), 405–420. https://doi.org/10.1210/endrev/bnaa004

Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *6*(2), 118–127. https://doi.org/10.31294/ijcit.v6i2.10438

Dewi, S., & Pakereng, M. A. I. (2023). Implementasi Principal Component Analysis Pada K-Means Untuk Klasterisasi Tingkat Pendidikan Penduduk Kabupaten Semarang. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *8*(4), 1186–1195. https://doi.org/10.29100/jipi.v8i4.4101

Dhurandhar, N. V. (2022). What is obesity?: Obesity Musings. *International Journal of Obesity*, *46*(6), 1081–1082. https://doi.org/10.1038/s41366-022-01088-1

Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., & Kalogirou, S. (2021). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, *36*(2), 121–136. https://doi.org/10.1080/10106049.2019.1595177

Herni Yulianti, S. E., Oni Soesanto, & Yuana Sukmawaty. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, *4*(1), 21–26. https://doi.org/10.31605/jomta.v4i1.1792

Hovi, H. S. W., Id Hadiana, A., & Rakhmat Umbara, F. (2022). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM). *Informatics and Digital Expert (INDEX)*, *4*(1), 40–45. https://doi.org/10.36423/index.v4i1.895

Idris, I. S. K., Mustofa, Y. A., & Salihi, I. A. (2023). Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM). *Jambura Journal of Electrical and Electronics Engineering*, *5*(1), 32–35. https://doi.org/10.37905/jjeee.v5i1.16830

Klaten, T. P. R.-R. dr. S. T. (2022). *Obesitas*. KEMENKAS. https://yankes.kemkes.go.id/view_artikel/429/obesitas

Maskuri, M. N., Harliana, Sukerti, K., & Herdian Bhakti, R. M. (2022). Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke Stroke Desease Predict Using KNN Algorithm. *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, *4*(1), 130–140.

Murdika, U., Alif, M., & Mulyani, Y. (2021). Identifikasi Kualitas Buah Tomat dengan Metode PCA (Principal Component Analysis) dan Backpropagation. *Electrician*, *15*(3), 175–180. https://doi.org/10.23960/elc.v15n3.2240

Nadiah, N., Soim, S., & Sholihin, S. (2022). Implementation of Decision Tree Algorithm Machine Learning in Detecting Covid-19 Virus Patients Using Public Datasets. *Indonesian Journal of Artificial Intelligence and Data*

Mining, *5*(1), 37–43. https://doi.org/10.24014/ijaidm.v5i1.17054

Nur Muhammad Ali Al Faizi, Mursyidul Ibad, Kuuni Ulfah Naila El Muna, & Budhi Setianto. (2023). Implementasi Principal Component Analysis dalam Analisis Faktor Kecacingan pada Anak Sekolah Dasar di Kabupaten Jember. *SEHATMAS: Jurnal Ilmiah Kesehatan Masyarakat*, *2*(3), 700–710. https://doi.org/10.55123/sehatmas.v2i3.2327

Nurdiansyah, N., Muliadi, M., Herteno, R., Kartini, D., & Budiman, I. (2024). Implementasi Metode Principal Component Analysis (Pca) Dan Modified K-Nearest Neighbor Pada Klasifikasi Citra Daun Tanaman Herbal. *Jurnal Mnemonic*, *7*(1), 1–9. https://doi.org/10.36040/mnemonic.v7i1.6664

Permana, A. P., Ainiyah, K., & Holle, K. F. H. (2021). Analisis Perbandingan Algoritma Decision Tree, kNN, dan Naive Bayes untuk Prediksi Kesuksesan Start-up. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(3), 178–188. https://doi.org/10.14421/jiska.2021.6.3.178-188

Pratiwi, S. A., Fauzi, A., Lestari, S. A. P., & Cahyana, Y. (2024). KLIK: Kajian Ilmiah Informatika dan Komputer Prediksi Persediaan Obat Pada Apotek Menggunakan Algoritma Decision Tree. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(4), 2381–2388. https://doi.org/10.30865/klik.v4i4.1681

Sajiwo, A. F. B., Rahmat, B., & Junaidi, A. (2024). Klasifikasi Indeks Standar Pencemaran Udaran (Ispu) Menggunakan Algoritma Xgboost Dengan Teknik Imbalanced Data (Smote). *Jurnal Informatika Dan Teknik Elektro Terapan*, *12*(3), 2190–2200. https://doi.org/10.23960/jitet.v12i3.4699

Sari, L., Romadloni, A., & Listyaningrum, R. (2023). *Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random*. *14*(01), 155–162. https://doi.org/10.35970/infotekmesin.v14i1.1751

Sawant, N., & Khadapkar, D. R. (2022). Comparison of the performance of GaussianNB Algorithm, the K Neighbors Classifier Algorithm, the Logistic Regression Algorithm, the Linear Discriminant Analysis Algorithm, and the Decision Tree Classifier Algorithm on same dataset. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 1654–1665. https://doi.org/10.22214/ijraset.2022.48311

Septian, F. (2023). Optimasi Klusterisasi pada Lama Tempo Pekerjaan Berbasis Gradient Boost Algorithm. *Indonesian Journal Of Information Technology*, *10*(2), 1–5.

World Health Organisation. (2024). *Obesity-and-Overweight*. World Health Organisation. https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight