Obesity Prediction Based on Lifestyle Habits Using Linear Regression

Andrea Octaviani
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta Barat, Indonesia 11480
andrea.octaviani@binus.ac.id

Bryan Ferdinand Teddy F.
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta Barat, Indonesia 11480
bryan.fiersdy@binus.ac.id

Dimas Ramdhan
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta Barat, Indonesia 11480
david01@binus.edu

David David
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta Barat, Indonesia 11480
david01@binus.edu

Abstract

Obesity among adolescents and adults continues to rise every year, making it a significant health concern in Indonesia. Obesity occurs due to an imbalance between calorie intake and the calories needed for bodily activities. Unhealthy lifestyles, such as staying up late, lack of physical activity, and consumption of fast food, are the main factors contributing to the increase in obesity cases. The purpose of this paper is to identify whether an individual is obese or not based on eating patterns, sleep hours, age, gender, activities, and other factors. The scope of this paper will focus solely on Indonesia, targeting adolescents and adults aged 15 and above. This study uses a linear regression model to predict obesity in Indonesians aged 15 and above, with the dataset sourced from Kaggle. The dataset is required for preprocessing and evaluation to determine the model's accuracy by dividing the data into 30% test data and 70% training data. After preprocessing, the linear regression model achieved an accuracy of approximately 86.57%.

keywords - obesity, linear regression, dataset, preprocessing, accuracy

1.INTRODUCTION

Obesity is a health problem characterized by the accumulation of fat exceeding the normal weight threshold. It has become a global health issue with rapidly increasing prevalence. Patients with obesity are at risk of developing various diseases such as heart disease, high blood pressure (hypertension), diabetes, liver cancer, stroke, joint and muscle disorders, respiratory problems, and even death, among others.[1], [2], [3], [4] While this health issue is often prevalent in industrialized nations, in many cases, developing countries surpass industrialized nations in the prevalence of overweight and obesity. According to the World Health Organization (WHO), obesity is the epidemic of the 21st century[2].

The global incidence of obesity nearly tripled from 1975 to 2016, with approximately 340 million children and adolescents affected by obesity. In 2016, around 39%, or approximately 650 million, of the global population aged 18 and over were obese and overweight. Additionally, about 340 million children and adolescents aged 5-19 were obese, with an additional 39 million children under the age of 5 also affected by obesity [3], [5], [6].

In Indonesia, data from Riskesdas in 2016 showed that the obesity rate among adults was 20.7%, an increase from 15.4% in 2013. In 2018, the prevalence of obesity in Indonesia was 13.5% for overweight individuals and 28.7% for those classified as obese among adults aged 18 and over[3], [7].

Various factors contribute to obesity, including genetic, physiological, environmental, socioeconomic, gender, family prosperity, and educational level factors[2], [5], [6], [8], [9], [10]. Obesity can be prevented through several steps such as dietary habits (avoiding eating right before bedtime),

reducing oily food intake, engaging in physical activity, and ensuring adequate sleep or rest.

The purpose of this paper is to identify whether an individual is obese or not based on eating patterns, sleep hours, age, gender, activities, and other factors. The scope of this paper will focus solely on Indonesia, targeting adolescents and adults aged 18 and above.

2.LITERATURE REVIEW

Obesity is a worldwide health problem that presents a substantial threat to persons' health and welfare[11]. Anticipating obesity by analyzing lifestyle behaviors can aid in promptly identifying those who are susceptible, facilitating measures to avert or control obesity.

Various industries such as marketing, banking, and healthcare are utilizing machine learning. Machine learning algorithms are increasingly prevalent due to their ability to learn from data and make predictions later obesity[12]. Linear regression is a widely used machine learning algorithm for determining the linear relationship between dependent and independent variables.

Several studies have used linear regression to examine the connection between lifestyle choices and obesity. Fan et al. (2022), for example, used linear regression to predict obesity in relation to food, sleep, and lifestyle variables like physical exercise[13]. The study included 28,048 children between the ages of 6 and 17. Using multiple linear regression, relationships between PA, SB, sleep, and food were examined. The study found that diet and physical exercise had a greater influence on obesity prediction than sleep habits[13], [14].

Tang et al. (2022) used linear regression in different research to predict obesity in a China multi-ethnic population[15]. DASH and aMED data were collected from 65,699 subjects from seven ethnic groups. Linear regression and x tests were used to compared continuous and categorical variables. This research found that obesity and metabolic health outcomes were positively associated with increased adherence to the DASH diet as opposed to the aMED diet. These findings make the DASH diet have better possibility dietary advice for preventing obesity and related metabolic illnesses in the Chinese population[15].

Additionally, Jindal et al. (2018) used ensemble machine learning techniques to predict obesity[16]. Information on obesity level, BMI, BMR, RMR, BFP, and protein RDA is what this study seeks to gather. Using the R ensemble prediction models, which include the Generalized Linear Model, Partial Least Square, and Random Forest models with Python interface, the average accuracy of the predicted values for obesity is 89.68%[16].

The study by Ferdowsy et al. (2021) emphasizes machine learning algorithms for estimating the risk of obesity so that people are aware of their danger and the causes of their obesity[17]. They compared nine machine learning algorithms: Decision Tree, Gradient Boosting Classifier, k-NN, Random Forest, Logistic Regression, Multilayer Perceptron (MLP), SVM, Naive Bayes, and Adaptive Boosting (ADA Boosting). With a result of 97.09%, Logistic Regression had the highest accuracy[17].

Three distinct machine learning techniques are used by Bag et al. (2023): logistic regression, random forest, and extreme gradient boosting[14]. These machine learning techniques use a tree-based machine learning approach to categorize obesity levels while taking dietary and physical activity habits into account. 498 participants, ranging in age from 14 to 61, provided the data for the analysis. The Logistic Regression model produced the best performance measurements across all variables, according to the findings[14]. Conversely, using the chosen features, Random Forest and Extreme Gradient Boosting produced better outcomes[14].

3. METHODOLOGY

This section explains the methodology, which consists of several main phases including data collection, data pre-processing, and the model to be used for calculating a person's weight using Linear Regression.

3.1 Datasets

The obesity database is a dataset that includes information about overweight-related diseases suffered by individuals. This dataset covers information related to a person's obesity. We use an obesity database consisting of 2112 data points from different individuals aged 15 years and older who suffer from obesity. Several important aspects used include gender, age, height, weight, family history, and smoking habits. The dataset can be accessed through the following link: https://www.kaggle.com/datasets/cahyaalkahfi/obesitas/data

3.2 Preprocessing Data

At this phase, we are using z-score normalization or standardization to normalize the obesity dataset. Z-score normalization will normalize values based on the mean and standard deviation. We define z-score normalization as the equation is shown in eq.1 and eq.2:

$$Xstand = \frac{x - mean(x)}{stddev(x)} \quad (1)$$

$$stddev(x) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x - mean(x))^2} \quad (2)$$

Where Xstand is standardized value, x is original data, mean(x) is the mean of the data, stddev is standard deviation of the data. After normalization, we will split the data into 30% test data and 70% training data.

3.3 Linear Regression

Our proposed method and model is the Linear Regression model, which is used to predict obesity based on weight and height. Linear regression is a statistical approach used to calculate the relationship between two or more variables, with at least one variable acting as the dependent variable and the other variables acting as independent variables. Fig. 1 below represent our Linear Regression Model:

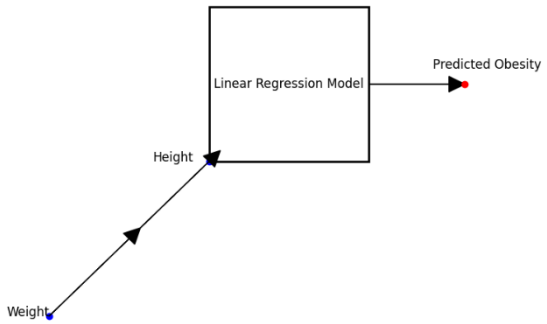

Fig. 1 Proposed Linear Regression Model

Based on Fig. 1, the input features, weight and height, are used as predictors to predict the target variable, obesity.

The Linear Regression model is chosen for its simplicity and interpretability. It is easy to comprehend and apply because it matches a linear relationship between the input data and the target variable.

To find the coefficients that best fit the data, the model is trained using the input of weight and height. Several metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared (R^2), are used to evaluate the model performance after training. These metrics provide a thorough evaluation of the accuracy and generalizability of the model for new data. The experiment will use the scikit-learn library in Google Collab.

3.4 Support Vector Machine (SVM)

SVM in this paper is useful as a comparison to see the best accuracy value between linear regression and Support Vector Machine. Support Vector Machine is a machine learning algorithm used for classification and regression, especially in this paper SVM is used for regression calculations.

3.5 Model Evaluation

We evaluate our model based on regression evaluation metrics, namely MSE, MAE, RMSE, and R^2. The results can be calculated using the following equations:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \widehat{yi})^2 \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \widehat{yi}| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \widehat{yi})^2} \quad (5)$$

$$R^2 = 1 - \frac{\sum i(yi - \widehat{yi})^2}{\sum i\,(yi - \bar{y})^2} \quad (6)$$

where n is the length of the test data array used for evaluation, yi is the true value, and yi represents the predicted value from the linear regression model.

where n is the length of the test data array used for evaluation, y_test is used. The yi values come from y_test, where i denotes the i-th index. The yi values represent the predicted values from the linear regression model, referred to as linear_reg_prediction.

and $\bar{Y}$ represents the sum of the squares of the differences between each actual value y and its mean $\bar{Y}$.

3.6 Model Comparison

Comparing the results of evaluating the accuracy of the model aims to find out whether the linear regression model is better than other models. or not. The results of the accuracy of the linear regression model will be compared with the results of the accuracy of the support vector machine (SVM) model.

4. RESULT AND DISCUSSION

4.1 Model Evaluation

After preprocessing, we proceeded to evaluate the linear regression model on each metric, and here are the results we obtained:

TABLE I. Linear Regression Model Evaluation

	Linear Regression
Mean Absolute Error	0.31738618222094356
Mean Squared Error	0.15481738307829374
Root Mean Squared Error	0.39346840162622176
R-Squared Score	0.8657680507216372

The mean absolute error result obtained is approximately 0.31738618222094356. MAE explains the average of the differences between predicted and actual values.

The mean squared error result obtained is 0.15481738307829374. MSE describes the average of the squared differences between predicted and actual values, providing a higher penalty for larger errors.

The root mean squared error result obtained is 0.39346840162622176, indicating predictions are off by approximately 39.34%. RMSE is the square root of MSE.

RMSE returns the error metric to the same units as the original data, making it easier to interpret.

The linear regression model achieved an R^2 of 0.8657680507216372, or 86.57%. This indicates a high level of prediction accuracy for the linear regression model. R-Squared (R^2) is the proportion of the variance in the dependent variable that is predictable from the independent variables.

The evaluation results obtained for the regression model are influenced by several factors such as good and clean data quality or preprocessing to ensure more accurate data, the features or variables used to calculate the model's accuracy, and the size of the data used for testing and training. Therefore, the results obtained are quite accurate. Since we have already implemented these measures, the obtained results are sufficiently high or accurate, specifically at 86.57%.

4.2 Model Comparison

The obtained result, which is the accuracy or R-Squared, will be compared with another model to determine which one is better. The linear regression model will be compared with the Support Vector Machine (SVM) model.

TABLE II. Linear Regression and SVM Comparison

	Linear Regression	SVM
Accuracy	0.8657680507216372	0.884984025559105

After comparison, it can be seen from table II that the accuracy of SVM is still better compared to Linear Regression in determining the prediction of obesity with a difference of about 0.02.

5. CONCLUSIONS

In this paper, Linear Regression was chosen as the model used to predict obesity due to its advantage in analyzing independent variables, resulting in more accurate outcomes. We also compared the Linear Regression model with another model, such as Support Vector Machine (SVM).

The results show that the prediction accuracy using Linear Regression is 86.57%, while SVM achieves an accuracy of 88.50%. Although the performance of Linear Regression is below that of SVM, the 86.57% accuracy level is still quite high, indicating that this model is effective in predicting obesity in individuals.

Thus, despite SVM having an approximately 1.93% advantage in terms of accuracy, Linear Regression remains a reliable model for predicting obesity, even though the Support Vector Machine (SVM) model is more accurate.

REFERENCES

[1] S. M. Fruh, "Obesity: Risk factors, complications, and strategies for sustainable long-term weight management," *J Am Assoc Nurse Pract*, vol. 29, pp. S3–S14, Oct. 2017, doi: 10.1002/2327-6924.12510.

[2] D. Minos, "Overweight and obesity in low-and middle income countries: A panel-data analysis," 2016.

[3] B. A. Pratama, "Literature Review: Faktor Risiko Obesitas Pada Remaja Di Indonesia," *Indonesian Journal on Medical Science*, vol. 10, no. 2, Jul. 2023, doi: 10.55181/ijms.v10i2.443.

[4] S. P. McGuire *et al.*, "Obesity Worsens Local and Systemic Complications of Necrotizing Pancreatitis and Prolongs Disease Course," *Journal of Gastrointestinal Surgery*, vol. 26, no. 10, pp. 2128–2135, Oct. 2022, doi: 10.1007/s11605-022-05383-0.

[5] N. U. Dewi, I. Tanziha, S. A. Solechah, and Bohari, "Obesity determinants and the policy implications for the prevention and management of obesity in Indonesia," *Current Research in Nutrition and Food Science*, vol. 8, no. 3, pp. 942–955, Dec. 2020, doi: 10.12944/CRNFSJ.8.3.22.

[6] M. J. Shan *et al.*, "Systematic estimation of BMI: A novel insight into predicting overweight/obesity in undergraduates," *Medicine (United States)*, vol. 98, no. 21, May 2019, doi: 10.1097/MD.0000000000015810.

[7] D. S. Harbuwono, L. A. Pramono, E. Yunir, and I. Subekti, "Obesity and central obesity in indonesia: Evidence from a national health survey," *Medical Journal of Indonesia*, vol. 27, no. 2, pp. 53–59, Jun. 2018, doi: 10.13181/mji.v27i2.1512.

[8] T. M. Fordham *et al.*, "Metabolic effects of an essential amino acid supplement in adolescents with PCOS and obesity," *Obesity*, Apr. 2024, doi: 10.1002/oby.23988.

[9] M. Meroni *et al.*, "Hepatic and adipose tissue transcriptome analysis highlights a commonly deregulated autophagic pathway in severe MASLD," *Obesity*, 2024, doi: 10.1002/oby.23996.

[10] 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," *Computers in Biology and Medicine*, vol. 136. Elsevier Ltd, Sep. 01, 2021. doi: 10.1016/j.compbiomed.2021.104754.

[11] A. Okunogbe, R. Nugent, G. Spencer, J. Ralston, and J. Wilding, "Economic impacts of overweight and obesity: Current and future estimates for eight countries," *BMJ Global Health*, vol. 6, no. 10. BMJ Publishing Group, Nov. 04, 2021. doi: 10.1136/bmjgh-2021-006351.

[12] T. M. Dugan, S. Mukhopadhyay, A. Carroll, and S. Downs, "Machine learning techniques for prediction of early childhood obesity," *Appl Clin Inform*, vol. 6, no. 3, pp. 506–520, Aug. 2015, doi: 10.4338/ACI-2015-03-RA-0036.

[13] C. Ding *et al.*, "Association between Physical Activity, Sedentary Behaviors, Sleep, Diet, and Adiposity among Children and Adolescents in China," *Obes Facts*, vol. 15, no. 1, pp. 26–35, Jan. 2022, doi: 10.1159/000519268.

[14] H. G. Gozukara Bag *et al.*, "Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits," *Diagnostics*, vol. 13, no. 18, Sep. 2023, doi: 10.3390/diagnostics13182949.

[15] L. Chen *et al.*, "Association of dietary patterns with obesity and metabolically healthy obesity phenotype in Chinese population: a cross-sectional analysis of China Multi-Ethnic Cohort Study," *British Journal of Nutrition*, vol. 128, no. 11, pp. 2230–2240, Dec. 2022, doi: 10.1017/S0007114521005158.

[16] K. Jindal, N. Baliyan, and P. S. Rana, "Obesity prediction using ensemble machine learning approaches," in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, pp. 355–362. doi: 10.1007/978-981-10-8636-6_37.

[17] F. Ferdowsy, K. S. A. Rahi, M. I. Jabiullah, and M. T. Habib, "A machine learning approach for obesity risk prediction," *Current Research in Behavioral Sciences*, vol. 2, Nov. 2021, doi: 10.1016/j.crbeha.2021.100053.