JURNAL MANAJEMEN PELAYANAN KESEHATAN

VOLUME 29 Edisi Khusus 20 - 25

Artikel Penelitian

MODEL PREDIKSI FAKTOR-FAKTOR RISIKO OBESITAS MENGGUNAKAN MACHINE LEARNING

Predictive Modeling of Obesity Risk Factors Using Machine Learning

Husnul Khuluq[1,2] ; Lazuardi Fatahillah H[2] ; Ayu Nissa Ainni[1] ; Tri Cahyani W[1]

[1]Program Studi Apoteker, Universitas Muhammadiyah Gombong, Kebumen, Indoensia

[2] Program Studi Teknik Informatika, Universitas Muhammadiyah Gombong, Kebumen, Indonesia

ABSTRACT

Background: *Obesity is a major global health concern and a key risk factor for various non-communicable diseases, including diabetes, hypertension, and cardiovascular disorders. Despite extensive studies, accurately identifying the key contributing factors remains a challenge.*

Objective: *This study aims to predict the likelihood of obesity using a machine learning algorithm, based on questionnaire-derived clinical and behavioral data. Several supervised machine learning algorithms—logistic regression, naïve Bayes, support vector machine (SVM), and random forest—will be employed to build predictive models. Model performance will be evaluated using accuracy, precision, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).*

Methods: *We used an open-access dataset from Kaggle comprising 2,111 samples with anthropometric, demographic, and lifestyle data. Of these, 972 individuals were categorized as obese and 1,139 as non-obese. The target variable was categorized into binary labels: "Obesity" and "Non-Obesity." Preprocessing included one-hot encoding, label encoding, and train-test splitting. All four ML models were trained and evaluated using accuracy, area under the curve (AUC), precision, sensitivity, and specificity metrics.*

Results: *The model achieved an accuracy of 98.58%, AUC of 99.96%, sensitivity of 98.99%, specificity of 98.21%, and precision of 98.01%. The most influential predictors were weight, frequent consumption of high-caloric food, family history of being overweight, physical activity frequency, and daily water intake.*

Conclusion: *The model demonstrated high performance and identified key lifestyle-related features. These findings support machine learning's potential for obesity screening and public health strategy development.*

Keywords: *Obesity; Machine Learning; Random Forest; Risk Factors; Predictive Modeling*

.

ABSTRAK

Latar belakang: *Obesitas merupakan masalah kesehatan global utama dan menjadi faktor risiko kunci bagi berbagai penyakit tidak menular, termasuk diabetes, hipertensi, dan gangguan kardiovaskular. Meskipun telah banyak dilakukan penelitian, mengidentifikasi faktor penyebab utama obesitas secara akurat masih menjadi tantangan.*

Tujuan: *Studi ini bertujuan untuk memprediksi kemungkinan obesitas menggunakan algoritma pembelajaran mesin (machine learning) berdasarkan data klinis dan perilaku yang diperoleh dari kuesioner. Beberapa algoritma pembelajaran mesin terawasi—logistic regression, naïve Bayes, support vector machine (SVM), dan random forest—digunakan untuk membangun model prediktif. Kinerja model dievaluasi menggunakan metrik akurasi, presisi, sensitivitas, spesifisitas, dan area under the curve (AUC) dari kurva ROC.*

Metode: *Dataset open-access dari Kaggle digunakan, terdiri dari 2.111 sampel yang mencakup data antropometri, demografis, dan gaya hidup. Dari jumlah tersebut, 972 individu dikategorikan sebagai obesitas dan 1.139 sebagai non-obesitas. Variabel target diklasifikasikan ulang menjadi dua label: "Obesitas" dan "Non-Obesitas." Proses pra-pemrosesan mencakup one-hot encoding, label encoding, dan pembagian data menjadi data latih dan uji. Keempat model ML dilatih dan dievaluasi menggunakan metrik akurasi, AUC, presisi, sensitivitas, dan spesifisitas.*

Hasil: *Model mencapai akurasi sebesar 98,58%, AUC sebesar 99,96%, sensitivitas sebesar 98,99%, spesifisitas sebesar 98,21%, dan presisi sebesar 98,01%. Prediktor yang paling berpengaruh adalah berat badan, frekuensi konsumsi makanan tinggi kalori, riwayat keluarga dengan kelebihan berat badan, frekuensi aktivitas fisik, dan asupan air harian.*

Kesimpulan: *Model menunjukkan performa yang sangat tinggi dan berhasil mengidentifikasi fitur-fitur gaya hidup yang berperan penting. Temuan ini mendukung potensi machine learning dalam skrining obesitas dan pengembangan strategi kesehatan masyarakat.*

Kata Kunci: *Obesitas; Machine Learning; Random Forest Faktor Risiko; Pemodelan Prediktif*

*Penulis korespondensi. Email : *husnulkhuluq@unimugo.ac.id*

INTRODUCTION

Obesity is rising globally, affecting both developed and developing countries. In 2022, approximately 2.5 billion adults were overweight, with 890 million classified as obese. The prevalence of obesity is also increasing among children, particularly in Asia. If left unaddressed, the global economic burden of obesity could reach US$18 trillion by 2060[1, 2]. Obesity is influenced by a range of risk factors, including genetic predisposition, unhealthy dietary patterns, low levels of physical activity, and various social and environmental influences. Although these factors have been widely studied, there remains a lack of comprehensive understanding regarding how they interact and contribute to obesity at the individual level[3, 4].

The selected machine learning methods were Logistic Regression, Naïve Bayes, Support Vector Machine, and Random Forest. These methods were chosen because each represents a different modeling philosophy and captures different patterns in the data. Logistic Regression provides an interpretable baseline model and has been widely used in health prediction tasks[5]. Naïve Bayes is efficient for high dimensional categorical datasets and performs well with questionnaire derived variables[6].

With advancements in technology, machine learning (ML) offers significant opportunities to analyze complex data and develop more accurate predictive models for understanding the factors that contribute to obesity[7]. Therefore, this study aims to identify the risk factors associated with obesity by analyzing data using machine learning techniques to produce a more effective predictive model. Based on this background, the research addresses the following questions: (1) What are the significant risk factors contributing to obesity based on questionnaire data collected from individuals? (2) How can machine learning models be applied to accurately analyze and predict obesity-related risk factors? (3) To what extent can the predictive results be used to support obesity prevention efforts among high-risk populations?

Obesity has become a major focus in health research, with numerous studies aiming to identify the risk factors that contribute to this condition. Previous research has commonly employed conventional statistical approaches, such as linear regression or factor analysis, to determine variables associated with obesity. However, these methods have limitations in handling complex datasets involving numerous variables, often resulting in reduced accuracy in predicting obesity risk. With technological advancements, machine learning (ML) has begun to be used to address these limitations. ML algorithms such as random forests, logistic regression, naïve Bayes, and support vector machines (SVM) have been applied to analyze more complex medical and social data, as well as to enhance predictive capabilities related to obesity. Recent studies have demonstrated the potential of machine learning in predicting obesity risk; however, many of these studies are still limited by small sample sizes or rely on specific types of data (e.g., clinical or genetic data only). Some of the latest models have started to incorporate more holistic data, including dietary patterns, exercise habits, environmental factors, and social behaviors. Nevertheless, research that integrates questionnaire-based data on risk factors with machine learning prediction models—particularly in the Indonesian context and for understanding obesity dynamics in broader populations—remains limited.

METHODS

The dataset used was obtained from Kaggle ("Obesity Data Set_raw_and_data_sinthetic.csv") and included 16 input parameters covering anthropometric, behavioral, and demographic aspects. These parameters were: Age, Gender, Height, Weight, CALC (alcohol consumption), FAVC (frequent consumption of high-caloric food), FCVC (frequency of vegetable consumption), NCP (number of main meals), SCC (monitoring of calorie intake), SMOKE, CH2O (daily water intake), family_history_with_overweight, FAF (physical activity frequency), TUE (time using technology devices), CAEC (food consumption between meals), and MTRANS (mode of transportation). The target variable, "NObeyesdad", was recorded into a binary classification: "Obesity" and "Non-Obesity".

The dataset was obtained from Kaggle, but the original source cannot be fully verified, as the dataset does not clearly indicate which survey, institution, or time period the data were derived from. This lack of traceability limits the ability to understand the population context and reduces the generalisability of the findings.

Data Processing The target variable "NObeyesdad" was recorded into a binary class: "Obesity" vs "Non-Obesity." Categorical variables were transformed using one-hot encoding, while numerical values were retained. The data were split into training and test sets (80/20 split).

Model Development: logistic regression, naïve Bayes, support vector machine (SVM), and random forest classifier was trained on the processed data. Model performance was evaluated using confusion matrix metrics: accuracy, sensitivity (recall), specificity, precision, and the area under the ROC curve (AUC).

The model evaluation was carried out using Python and relevant libraries such as scikit-learn. Performance metrics including accuracy, sensitivity (recall), specificity, precision, and area under the

ROC curves (AUC) were computed to assess the effectiveness of each model. Additionally, feature importance analysis revealed the top five most influential parameters in predicting obesity.

Inclusion criteria: Participants were included if they had complete questionnaire data covering anthropometric, demographic, and lifestyle variables required for the analysis. Exclusion criteria: Participants were excluded if any key variables were missing, inconsistent, or contained unrealistic values (e.g., implausible height or weight), or if their records were incomplete for the outcome classification

Ethical Considerations As this study used open-access data with anonymized synthetic information, no ethical clearance was required. Expected Outcomes The study aims to deliver a validated binary classification model for obesity risk and identify actionable predictors. The results are intended to support evidence-based public health strategies and preventative care initiatives.

RESULT

Table 1. Baseline Characteristics of Participants by Obesity Status

Baseline characteristics	All (n = 2111)	Non-Obese (n = 1139)	Obese (n = 972)
Age	22.8 (19.9–26.0)	21.0 (19.0–24.0)	25.1 (21.7–27.9)
Height	1.7 (1.6–1.8)	1.7 (1.6–1.8)	1.7 (1.6–1.8)
Weight	83.0 (65.5–107.4)	68.0 (55.0–80.0)	110.0 (98.9–120.0)
Gender	Male (1068 [50.6%])	Male (577 [50.7%])	Male (491 [50.5%])
SMOKE	no (2067 [97.9%])	no (1117 [98.1%])	no (950 [97.7%])
SCC	no (2015 [95.5%])	no (1046 [91.8%])	no (969 [99.7%])
FAVC	yes (1866 [88.4%])	yes (913 [80.2%])	yes (953 [98.0%])
family_history_with_overweight	yes (1726 [81.8%])	yes (762 [66.9%])	yes (964 [99.2%])
FCVC	2.4 (2.0–3.0)	2.1 (2.0–3.0)	2.7 (2.0–3.0)
NCP	3.0 (2.7–3.0)	3.0 (2.3–3.0)	3.0 (2.9–3.0)
CH2O	2.0 (1.6–2.5)	2.0 (1.5–2.2)	2.1 (1.6–2.6)
FAF	1.0 (0.1–1.7)	1.0 (0.2–2.0)	0.9 (0.1–1.5)
TUE	0.6 (0.0–1.0)	0.8 (0.0–1.0)	0.6 (0.1–0.9)
CALC	Sometimes (1401 [66.4%])	Sometimes (682 [59.9%])	Sometimes (719 [74.0%])
CAEC	Sometimes (1765 [83.6%])	Sometimes (811 [71.2%])	Sometimes (954 [98.1%])
MTRANS	Public_Transportation (1580 [74.8%])	Public_Transportation (821 [72.1%])	Public_Transportation (759 [78.1%])

Source: Secondary Data (2014–2015)

Table 1 summarizes the baseline characteristics of 2,111 participants, comparing demographic, anthropometric, and behavioral variables between obese (n = 972) and non-obese (n = 1,139) individuals.

Table 2. Confusion Matrix for Each Model

Model	TP	FP	FN	TN
Random Forest	197	4	2	220
Logistic Regression	192	6	7	218
Naive Bayes	186	74	13	150
SVM	165	6	34	218

Source: Secondary Data (2014–2015)

Table 3. Evaluation Metrics for Each Model

Model	Accuracy (%)	AUC (%)	Sensitivity (%)	Precision (%)	Specificity (%)
Random Forest	98.58	99.96	98.99	98.01	98.21
Logistic Regression	96.93	99.67	96.48	96.97	97.32
Naive Bayes	79.43	93.06	93.47	71.54	66.96
SVM	90.54	97.58	82.91	96.49	97.32

Based on data from the Indonesian Family Life Survey, Wave 5 (2014–2015)

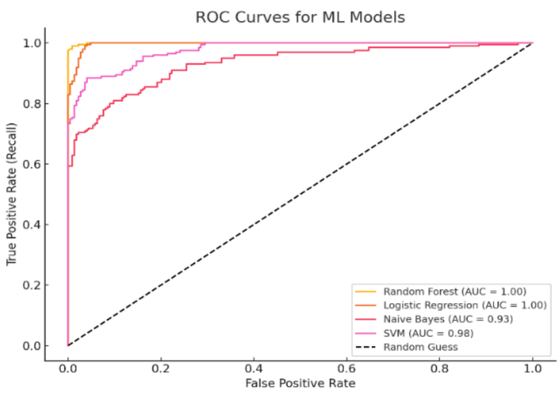

Figure 1. Curve for ML models

Table 4.Top 5 Predictive Features Based on Odds Ratio

Feature	Coefficient	Odds Ratio
remainder__Weight	10.320	30342.593
cat__FAVC_yes	0.530	1.698
cat__family_history_with_overweight_yes	0.368	1.444
remainder__FAF	0.115	1.122
remainder__CH2O	0.065	1.067

Source: Secondary Data (2014–2015)

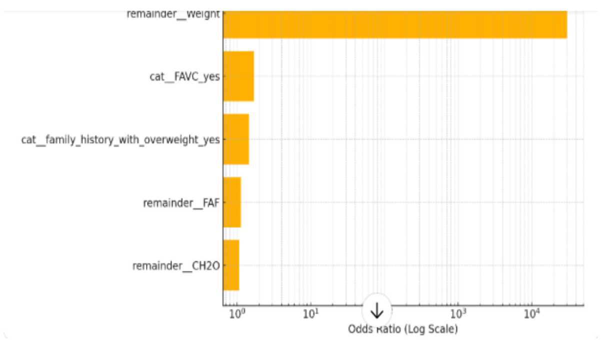

Figure 2. Top 5 Most Influential Features for Obesity Prediction

DISCUSSION

The findings from this study reveal that the Random Forest (RF) algorithm delivered the best performance among the four evaluated machine learning models. With an accuracy of 98.58% and an AUC of 99.96%, RF significantly outperformed Logistic Regression, Naïve Bayes, and SVM (table 3, figure 1). This superior performance can be attributed to several inherent advantages of RF: it is an ensemble method that reduces overfitting, handles both numerical and categorical data well, and can manage complex interactions among features without extensive preprocessing[8, 9]. One of the key strengths of RF lies in its ability to evaluate feature importance[10, 11].

The extremely high accuracy and AUC values reported in the study (exceeding 98%) raise concerns about potential overfitting, especially given the limited validation procedures presented. To enhance the robustness and credibility of the findings, it would be important to incorporate more rigorous evaluation strategies such as k-fold cross-validation, hyperparameter tuning (e.g., grid search or randomized search), and additional performance metrics including log-loss. These techniques are widely recommended in recent machine learning research to prevent model overfitting and ensure generalizability. Without these validation steps, the performance results may be overly optimistic and should be interpreted with caution[12]

In our study the most influential predictors of obesity identified through logistic regression were weight, frequent consumption of high-caloric food, family history of overweight, physical activity frequency, and daily water intake. Among these, weight showed an exceptionally high odds ratio (OR = 30,343), indicating a strong association with obesity. Individuals who frequently consumed high-calorie food had 1.7 times higher odds of being obese, while those with a family history of overweight had 1.44 times higher odds. Physical activity and water intake were also positively associated, though to a lesser extent, with ORs of 1.12 and 1.07, respectively (table 4).

Weight has a direct correlation with obesity, and numerous studies have established that an increase in body weight is one of the strongest indicators of obesity status [13]

The assessment of obesity risk factors in this study is constrained by the exclusive use of odds ratios, which provides only a limited understanding of variable relationships; incorporating additional interpretability techniques such as correlation analysis or model-agnostic feature-importance methods like SHAP would offer a more comprehensive and transparent explanation of how predictors influence model outcomes, consistent with recent recommendations in machine learning interpretability research [14].. Furthermore, variables such as weight, which are intrinsically tied to the definition of obesity, may introduce circular reasoning and artificially inflate model performance, a concern highlighted in recent discussions on data leakage and deterministic predictors in health-related machine learning models [15]. For this reason, the inclusion of such variables should be reconsidered or clearly justified, and the manuscript would benefit from acknowledging their potential impact on validity, interpretability, and generalizability.

Family History (Yes) – A positive family history indicates genetic predisposition and shared lifestyle factors, both of which are well-known contributors to obesity risk[16, 17,18]

Physical activity frequency, higher total physical activity volume and greater frequency/intensity were significantly associated with reduced obesity risk among Australian women over a 21-year period[19] A significant association ($p = 0.027$) between daily physical activity levels and central obesity in adults in Gorontalo, Indonesia those with lower activity levels had a higher risk of obesity[20]

Daily water intake. Adults aged 19–39 who met adequate daily water intake levels had significantly lower odds of abdominal obesity compared to those with lower intake. However, the association was not significant in older age groups[21]. Overweight and obese women who drank more water daily had significantly lower BMI. Specifically, for every 1 mL of additional water consumed, BMI decreased by 0.0021 kg/m^2[22]

Frequently consumed high calorie food. Final-year students with a frequent habit of consuming high-calorie foods had significantly higher odds of being obese. The study found strong associations between high intake of calories, fats, and carbohydrates from these foods and poor nutritional status[23] Adolescents who frequently consumed junk food (high-calorie foods) had a 2.87 times higher risk of obesity compared to those who consumed it rarely. The study also reported positive correlations between junk food intake, BMI, and waist circumference[24]

In addition to the previously noted issues, the manuscript would benefit from a clearer articulation of how the analytical choices—such as the selection of variables, recoding procedures, and model evaluation strategies—shape the interpretation of the findings. Providing stronger justification for these methodological decisions, including why certain preprocessing steps were chosen and how they may influence the results, would help readers better understand the study's internal logic and analytical rigor. Moreover, emphasizing the rationale behind the comparison of different machine learning models and discussing the strengths and limitations of each approach within the study's context would add depth and enhance the overall clarity of the research narrative.

Limitation: while the dataset used in this study is sourced from Kaggle and contains synthetic elements, it is important to clarify that not all components of the dataset are fully artificial. The dataset represents a mixture of real questionnaire-based patterns that have been expanded or partially generated to increase sample size and variability. Nevertheless, the presence of synthetic data means that certain distributions, correlations, or behavioral patterns may not accurately reflect real-world populations. As a result, the generalizability of the findings remains limited, and conclusions regarding obesity risk factors should be interpreted with caution. Future studies should validate these results using fully real, population-based datasets to ensure more robust external validity

CONCLUSION

This study demonstrated the effectiveness of machine learning—particularly the Random Forest algorithm—in predicting obesity risk using questionnaire-based data. Among the four models evaluated, Random Forest achieved the highest performance, with an accuracy of 98.58% and an AUC of 99.96%. The most influential predictors identified were weight, height, age, family history of being overweight, and the number of main meals per day (NCP).

These findings highlight the model's potential practical value for supporting early identification of individuals at higher risk and guiding more targeted public health interventions or clinical decision-making, such as prioritizing lifestyle counseling or screening programs. At the same time, the noted limitations point to clear directions for future research, including validating the model in different population groups, integrating additional data sources (such as clinical measurements or longitudinal records), and evaluating the model's real-world impact when implemented in community or clinical settings to determine its effectiveness and feasibility.

ACKNOWLEDGEMENT

This work was supported by financial assistance from the Institute for Research and Community Service (LPPM), Universitas Muhammadiyah Gombong.

REFERENCE

1. Blüher M. Obesity: global epidemiology and pathogenesis. Nat Rev Endocrinol [Internet]. 2019;15:288–98. Available from: https://consensus.app/papers/obesity-global-epidemiology-and-pathogenesis-bl%C3%BCher/cebb6cda9b2a50b98f003b5df9a3ea9f/
2. Sweis N. The economic burden of obesity in 2024: a cost analysis using the value of a statistical life. Crit Public Health [Internet]. 2024;34:1–13. Available from: https://consensus.app/papers/the-economic-burden-of-obesity-in-2024-a-cost-analysis-using-sweis/e630f8189097548c8cc6c818d9f6653a/
3. Pledger SL, Ahmadizar F. Gene-environment interactions and the effect on obesity risk in low and middle-income countries: a scoping review. Front Endocrinol (Lausanne). 2023;14(August).
4. Kim MS, Shim I, Fahed AC, Do R, Park WY, Natarajan P, et al. Association of genetic risk, lifestyle, and their interaction with obesity and obesity-related morbidities. Cell Metab. 2024 Jul 2;36(7):1494-1503.e3.
5. Bennett M, J. Kleczyk E, Hayes K, Mehta R. Evaluating Similarities and Differences between Machine Learning and Traditional Statistical Modeling in Healthcare Analytics. 2022;1–15.
6. Liu Z, Wen T, Sun W, Zhang Q. Feature-weighting and clustering random forest. International Journal of Computational Intelligence Systems. 2021;14(1):257–65.
7. Marquart I, Marquart EK. RFCC: Random Forest Consensus Clustering for Regression and Classification [Internet]. 2021. Available from: https://ssrn.com/abstract=3807828
8. Alsagri H, Ykhlef M. Quantifying Feature Importance for Detecting Depression using Random Forest. International Journal of Advanced Computer Science and Applications [Internet]. 2020;11. Available from: https://consensus.app/papers/quantifying-feature-importance-for-detecting-depression-alsagri-ykhlef/20a0bd57c68451b691fe58c65b333c38/
9. Miao Y, Xu Y. Random Forest-Based Analysis of Variability in Feature Impacts. 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA) [Internet]. 2024;1130–5. Available from: https://consensus.app/papers/random-forestbased-analysis-of-variability-in-feature-miao-xu/3f1a3a7258d4501683174a33d89f1a1b/
10. Saikia D, Ahmed S, Saikia H, Sarma R. Body mass index and body fat percentage in assessing obesity: An analytical study among the adolescents of Dibrugarh, Assam. Indian J Public Health [Internet]. 2018;62:277–81. Available from: https://consensus.app/papers/body-mass-index-and-body-fat-percentage-in-assessing-sarma-ahmed/3008eff9d94452388b7ca8cc137dcbed/
11. Loos RJF, Yeo GSH. The genetics of obesity: from discovery to biology. Nat Rev Genet. 2022 Feb;23(2):120–33.
12. Mangla A, Dhamija N, Gupta U, Dhall M. Familial Background as a Hidden Cause for Obesity among College Going Girls. J Biosci Med (Irvine) [Internet]. 2019; Available from: https://consensus.app/papers/familial-background-as-a-hidden-cause-for-obesity-among-dhamija-gupta/f9da5a5dbd755e9aa142bc01f414422d/
13. Arsita C, Rachmani E, Isworo S, Kusumangrum L, Anggraini T. Exploring Obesity Risk Factors: Focuses on Family History of Overweight and Smoking Behaviour. Asian Journal of Medicine and Health [Internet]. 2024; Available from: https://consensus.app/papers/exploring-obesity-risk-factors-focuses-on-family-history-arsita-rachmani/f1f6352b034e581cacbc22101022b5ae/
14. Mielke G, Ding D, Keating S, Nunes B, Brady R, Brown W. Physical activity volume, frequency, and intensity: Associations with hypertension and obesity over 21 years in Australian women. J Sport Health Sci [Internet]. 2024;13:631–41. Available from: https://consensus.app/papers/physical-activity-volume-frequency-and-intensity-mielke-brown/000b248221f5514ab8c1d1af184d79a7/
15. Nuryani, Muhdar IN, Ramadhani F, Paramata Y, Adi DI, Bohari B. Association of Physical Activity and Dietary Patterns with Adults Abdominal Obesity in Gorontalo Regency, Indonesia: A

Cross-Sectional Study. Current Research in Nutrition and Food Science Journal [Internet]. 2021;9:280–92. Available from: https://consensus.app/papers/association-of-physical-activity-and-dietary-patterns-muhdar-nuryani/4e5bee9c08cd5139aa54c9a1e362a656/

16. Kim YJ, Oh SN, Kong EK, Seon ES. Association between Water Intake and Abdominal Obesity: The Korea National Health and Nutrition Examination Survey 2019-2021. Korean J Fam Med [Internet]. 2024; Available from: https://consensus.app/papers/association-between-water-intake-and-abdominal-obesity-seon-kong/1074dda12ad8557fa233f0d07d969a03/
17. Rodríguez R, Aparicio A, López-Sobaler A, Ortega R. Importance of water consumption in a group of young women with overweight and obesity. Nutr Hosp [Internet]. 2015;32 Suppl 2:10339. Available from: https://consensus.app/papers/importance-of-water-consumption-in-a-group-of-young-women-ortega-l%C3%B3pez-sobaler/974a65a1228951d4954e19a235ac792c/
18. Putri PA. Association of High Calorie Food and Coffee Consumption Pattern, Sleep Duration and Stress Level with Nutritional Status in Final Year Students. Media Gizi Kesmas [Internet]. 2022; Available from: https://consensus.app/papers/association-of-high-calorie-food-and-coffee-consumption-putri/1231859fa1565623995acbff15b26567/
19. Kotska S, Teguh MM, Santoso AH. Study Analysis of the Role of High-Calorie Food (Junk Food) on the Incident of Adolescent Obesity: A Community-Based Observational Study of Senior High School Adolescents in Tangerang Regency, Indonesia. Community Medicine and Education Journal [Internet]. 2024; Available from: https://consensus.app/papers/study-analysis-of-the-role-of-highcalorie-food-junk-food-on-teguh-kotska/537d279e356d571c9c29106f6ce5cf81/