Inside | Jurnal Sistem Informatika Cerdas .

Volume 2 Nomor 1 | bulan July tahun 2024 | halaman 13-18
e-ISSN: 3031-4860

Obesity Risk Prediction Using Random Forest Based on Eating Habit Parameters

Agung Muliawan[1*], Difari Afreyna Fauziah[2], Eko Afrianto[3]
[1,2,3]Institute of Technology and Science Mandala, Jember, Indonesian

Corresponding Author:
Agung Muliawan, Institute Technology and Science Mandala, Jember, 68121, Indonesia
Email: agung.muliawan@itsm.ac.id

Abstract

Obesity is a global health problem associated with multiple chronic diseases, so early detection and risk prediction are important for prevention efforts. As obesity is one of the major health problems that can lead to various chronic diseases, accurate modelling can help in prevention and early intervention efforts. This study aims to develop an obesity risk prediction model using Random Forest technique, which is based on individual eating habit parameters. The dataset used is taken from an open dataset that has variables of eating habits, which includes variables such as frequency of consumption of high-calorie foods, eating patterns, and types of food. The data was processed and analysed with the Random Forest algorithm, an ensemble learning method known to be effective in handling datasets with high dimensionality and non-linear relationships between features. The developed Random Forest model showed good performance with a prediction accuracy of 81.76%. This accuracy indicates that the model can effectively distinguish individuals with high risk of obesity from those with low risk based on their eating habit parameters. The results of this study demonstrate the potential of Random Forest as a useful tool in identifying obesity risk, which can assist in data-driven health prevention and intervention strategies.

Keywords : Random Forest, Obesity Risk Prediction, Eating Habits, Accuracy, Health

1 INTRODUCTION (11 PT)

Obesity has become a significant global health problem, affecting millions of people worldwide. Moreover, obesity has evolved into one of the most pressing public health challenges of the 21st century [1]. The increase in the global prevalence of obesity can be attributed to unhealthy lifestyles, especially poor eating habits [2]. The rising prevalence of obesity has been associated with an increased risk of various chronic diseases such as type 2 diabetes, hypertension, and heart disease [3]. Therefore, a better understanding of the factors that influence obesity risk can play an important role in disease prevention and management[4]. In addition, early detection and prevention of obesity is a top priority in public health policy. Eating habits are one of the main factors that contribute to obesity risk [5].

Eating habits, including the frequency of consumption of high-calorie foods, the types of foods consumed, as well as overall diet, are important components in determining obesity risk [6]. Consumption of high-calorie foods, unhealthy eating patterns and meal frequency can affect energy balance and, in turn, lead to the accumulation of body fat. However, manual analyses of the relationship between eating habits and obesity are often complex and inadequate in providing a clear picture. Data analysis technologies and machine learning methods now offer new approaches to address this challenge. To understand the complex relationship between eating habits and obesity risk, robust data analysis methods are needed to identify significant patterns and provide accurate predictions.

In this context, machine learning techniques such as Random Forest have emerged as an effective tool in analysing and predicting obesity risk [7]. Random Forest is an ensemble learning method that uses multiple decision trees to make more stable and accurate predictions [8]. The advantages of this method lie in its ability to handle high-dimensional data, manage non-linear interactions between variables, and provide easy-to-understand interpretations[9].

Research conducted by M. Patel and R. Kumar describes the ensemble classifer method in random forest in the use of data related to calorie intake and food frequency with the results of providing significant accuracy related to what distinguishes between individuals with high and low obesity risk [10]. In addition, research by Zekun Zhao, Haipeng Lu, etc. on the development of analytical models related to predicting the risk of workers in obesity using SVM with the results of the most influential variables in obesity problems [11]

This study aims to develop an obesity risk prediction model using Random Forest algorithm based on individual eating habit parameters. Using a dataset that includes various parameters related to eating habits, such as frequency of consumption of high-fat foods, daily diet, and calorific intake, the model is expected to provide a more accurate prediction of obesity risk. Through this analysis, we seek to explore how eating habits contribute to obesity risk and to identify key factors that can be used in prevention and intervention strategies.

With the model expected to achieve high accuracy in predicting obesity risk, the results from this study are expected to provide valuable insights for health practitioners, policy makers, and individuals in designing data-driven strategies to address obesity. The use of Random Forest in this context offers significant potential to improve the effectiveness of obesity prediction and prevention through in-depth and comprehensive data analysis.

2 RESEARCH METHOD

The research method carried out in this study has several stages starting from data collection, making Deep learning problems and evaluating and validating to see the accuracy value. The following is a detailed explanation of each stage:

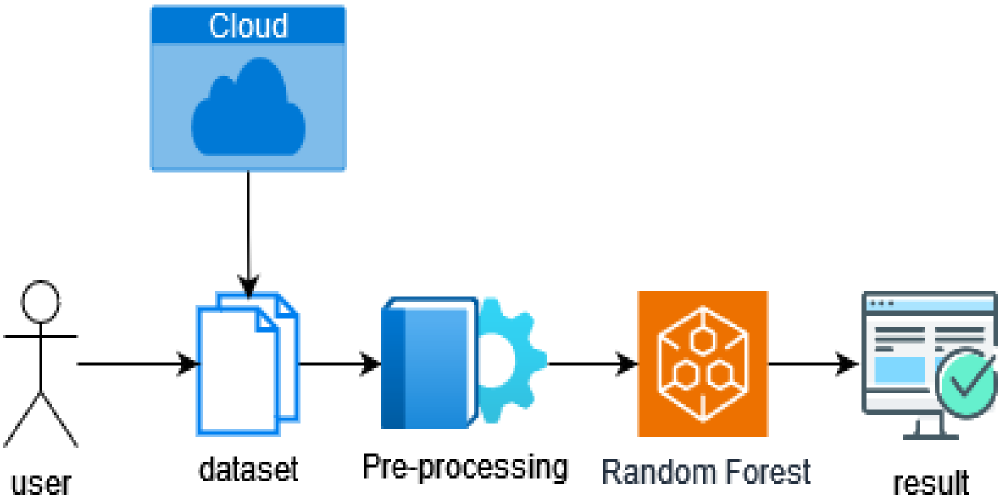


Fig 1. Research Methodology

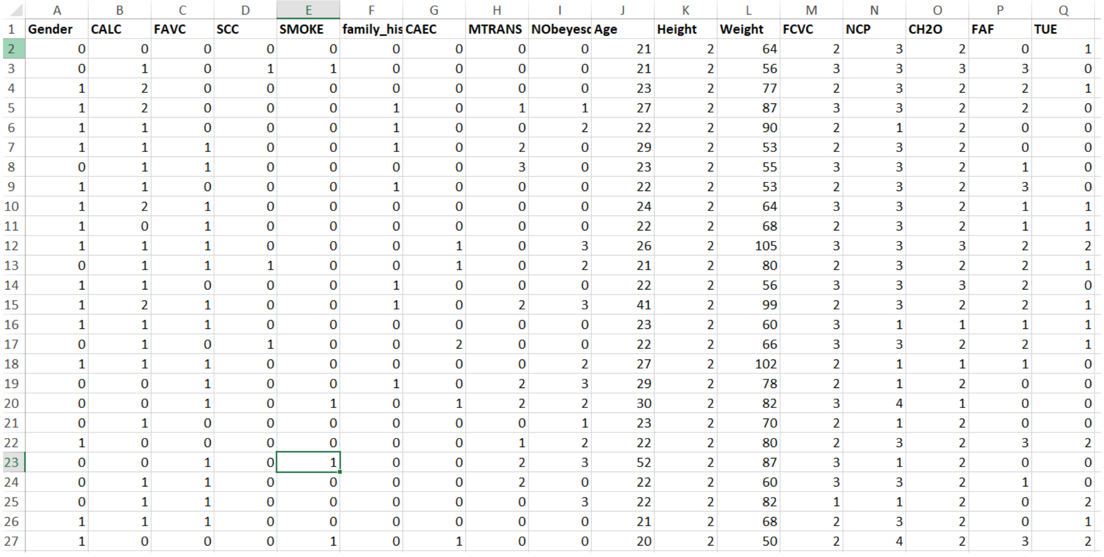
2.1 Pre-Processing

In the data collection process, a dataset is taken through kaggle which contains several variables that will be used as input for the random forest process (Muliawan et al., 2022). The amount of data obtained is 2111 rows with 17 variables covering information on eating habits and other information (Obesity Level Kaggle). The first process is to remove missing values, then normalise the data and transform the data to numeric so that it is easier to process during obesity classification. The following is a table of variables used in this study:

Table 1. Variabel Dataset

No	Variabel	Tipe Data	Describe
1	Gender	Categorical	Gender
2	Age	Continuous	Age
3	Height	Continuous	Height
4	Weight	Continuous	Weight
5	family_history_with_overweight	Binary	Has a family member suffered or suffers from overweight
6	FAVC	Binary	Do you eat high caloric food frequently
7	FCVC	Integer	Do you usually eat vegetables in your meals
8	NCP	Continuous	How many main meals do you have daily
9	CAEC	Categorical	Do you eat any food between meals
10	SMOKE	Binary	Do you smoke
11	CH2O	Continuous	How much water do you drink daily?
12	SCC	Binary	Do you monitor the calories you eat daily
13	FAF	Continuous	How often do you have physical activity
14	TUE	Integer	How much time do you use technological devices such as cell phone, videogames, television, computer and others
15	CALC	Categorical	How often do you drink alcohol
16	MTRANS	Categorical	Which transportation do you usually use
17	NObeyesdad	Categorical	Obesity level

From the dataset that will be somewhat transformed nominally to be processed into classification, the following is an image of the results of the pre-processing process which will be continued in the process to random forest:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Gender	CALC	FAVC	SCC	SMOKE	family_his	CAEC	MTRANS	NObeyesc	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
2	0	0	0	0	0	0	0	0	0	21	2	64	2	3	2	0	1
3	0	1	0	1	1	0	0	0	0	21	2	56	3	3	3	3	0
4	1	2	0	0	0	0	0	0	0	23	2	77	2	3	2	2	1
5	1	2	0	0	0	1	0	1	1	27	2	87	3	3	2	2	0
6	1	1	0	0	0	1	0	0	2	22	2	90	2	1	2	0	0
7	1	1	1	0	0	1	0	2	0	29	2	53	2	3	2	0	0
8	0	1	1	0	0	0	0	3	0	23	2	55	3	3	2	1	0
9	1	1	0	0	0	1	0	0	0	22	2	53	2	3	2	3	0
10	1	2	1	0	0	0	0	0	0	24	2	64	3	3	2	1	1
11	1	0	1	0	0	0	0	0	0	22	2	68	2	3	2	1	1
12	1	1	1	0	0	0	1	0	3	26	2	105	3	3	3	2	2
13	0	1	1	1	0	0	1	0	2	21	2	80	2	3	2	2	1
14	1	1	0	0	0	1	0	0	0	22	2	56	3	3	3	2	0
15	1	2	1	0	0	1	0	2	3	41	2	99	2	3	2	2	1
16	1	1	1	0	0	0	0	0	0	23	2	60	3	1	1	1	1
17	0	1	0	1	0	0	2	0	0	22	2	66	3	3	2	2	1
18	1	1	1	0	0	0	0	0	2	27	2	102	2	1	1	1	0
19	0	0	1	0	0	1	0	2	3	29	2	78	2	1	2	0	0
20	0	0	1	0	1	0	1	2	2	30	2	82	3	4	1	0	0
21	0	1	0	0	0	0	0	0	1	23	2	70	2	1	2	0	0
22	1	0	0	0	0	0	0	1	2	22	2	80	2	3	2	3	2
23	0	0	1	0	1	0	0	2	3	52	2	87	3	1	2	0	0
24	0	1	1	0	0	0	0	2	0	22	2	60	3	3	2	1	0
25	0	1	1	0	0	0	0	0	3	22	2	82	1	1	2	0	2
26	1	1	1	0	0	0	0	0	0	21	2	68	2	3	2	0	1
27	1	0	0	0	1	0	1	0	0	20	2	50	2	4	2	3	2

Figure 2. Pre-processing

2.2 Random Forest

Random forest is a machine learning algorithm used for classification, regression, and other prediction-related tasks. It works by building many decision trees during training and combining the results to

.

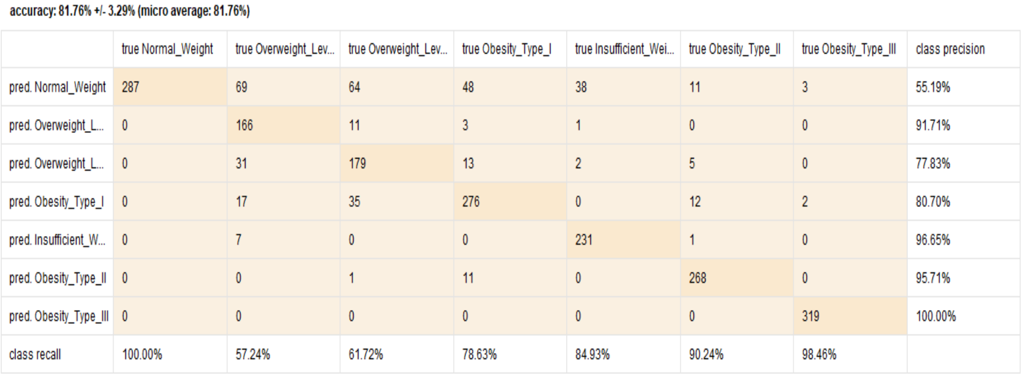
improve accuracy and reduce overfitting. Random forest uses multiple decision trees, it is often more accurate than a single decision tree and can handle missing data and incomplete features.

2.3 Evaluation and Validation

Evaluation and validation of Random Forest models are important steps to ensure optimal performance and good generalisability [12]. By using various metrics and validation techniques, we can ensure that the model is not only accurate in prediction but also reliable and useful in real-world applications.

3 RESULTS AND ANALYSIS

The following are the results of testing the dataset using the rapid miner tool. The accuracy result is 81.76% so this accuracy shows that the model can effectively distinguish individuals with a high risk of obesity from those with a low risk based on the parameters of their eating habits. The following is a picture of the accuracy in this study:



accuracy: 81.76% +/- 3.29% (micro average: 81.76%)

	true Normal_Weight	true Overweight_Lev...	true Overweight_Lev...	true Obesity_Type_I	true Insufficient_Wei...	true Obesity_Type_II	true Obesity_Type_III	class precision
pred. Normal_Weight	287	69	64	48	38	11	3	55.19%
pred. Overweight_L...	0	166	11	3	1	0	0	91.71%
pred. Overweight_L...	0	31	179	13	2	5	0	77.83%
pred. Obesity_Type_I	0	17	35	276	0	12	2	80.70%
pred. Insufficient_W...	0	7	0	0	231	1	0	96.65%
pred. Obesity_Type_II	0	0	1	11	0	268	0	95.71%
pred. Obesity_Type_III	0	0	0	0	0	0	319	100.00%
class recall	100.00%	57.24%	61.72%	78.63%	84.93%	90.24%	98.46%	

Figure 3. Accuracy Random Forest

Furthermore, an analysis was carried out related to the most influential factors in obesity in getting 3 variables that influence physical activity, transportation used and family history affected by obesity. The following is a visual representation of the distribution of influential variables

Variabel : FAF (physical activity)

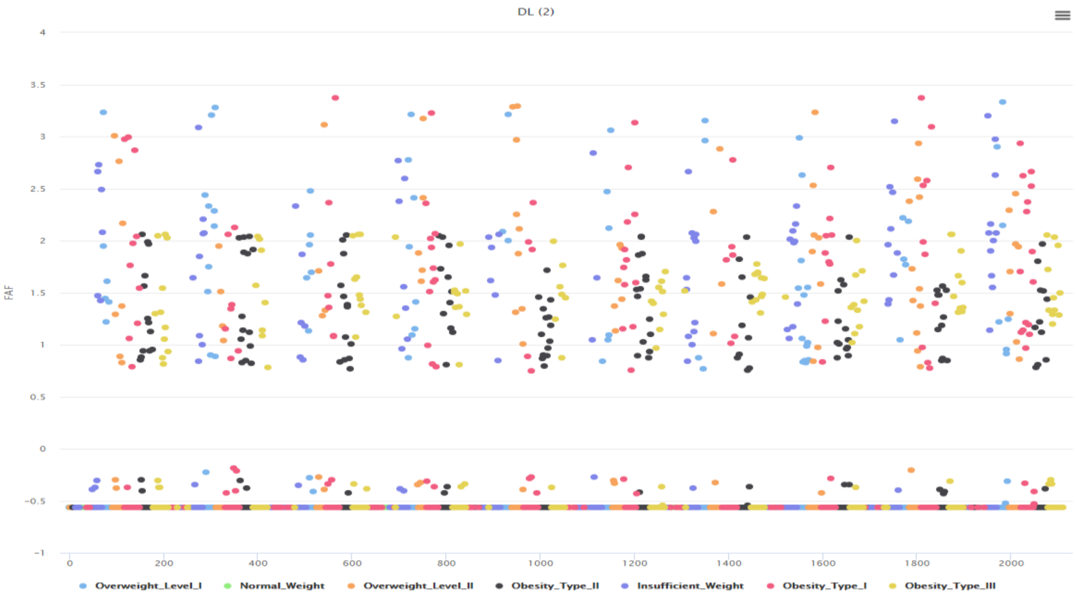


Fig 4. Visualization FAF

Variabel : M trans (Transportation) and family_history_with_overweight

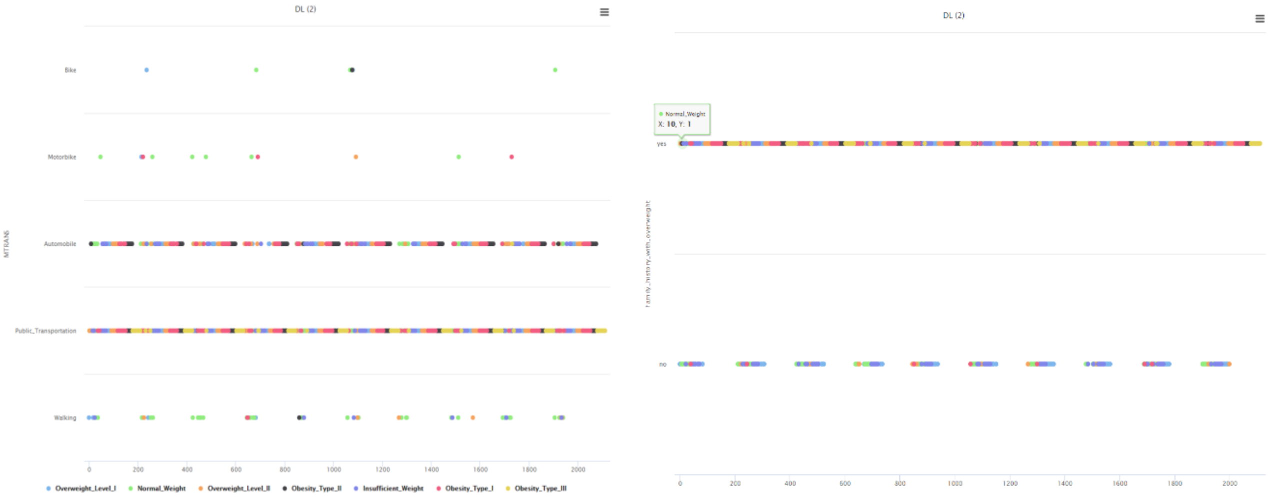


Fig 5. Visualization Mtrans & family_history_with_overweight

4 CONCLUSION

These results show that the Random Forest algorithm is effective in identifying the likelihood of obesity based on an individual's eating habits. The accuracy of 81.76% indicates that the model performs well in predicting obesity risk, with a relatively low error rate. This proves the potential use of the Random Forest algorithm in practical applications for obesity risk evaluation that can assist in more effective obesity prevention and treatment interventions. Furthermore, an analysis was carried out related to the most influential factors in obesity in getting 3 variables that influence physical activity, transportation used and family history affected by obesity. Thus, this study recommends the use of this model as a tool in obesity risk analysis, noting that further development and validation with larger datasets may be required to improve the accuracy and generalisability of the model

REFERENCES (11 PT)

[1] E. D. N. Aini, R. A. Khasanah, A. Ristyawan, and E. Diniati, “Penggunaan Data Mining untuk Prediksi tingkat Obesitas di Meksiko Menggunakan Metode Random Forest,” *Pros. SEMNAS INOTEK Semin. Nas. Inov. Teknol.*, vol. 8, no. 3, pp. 1256–1265, Jul. 2024.

[2] E. Aguirre Rodríguez, E. Rodríguez, L. Nascimento, A. Silva, and F. Marins, “Machine learning Techniques to Predict Overweight or Obesity,” Nov. 2021.

[3] A. Muliawan, A. Rizal, and S. Hadiyoso, “Heart Disease Prediction based on Physiological Parameters Using Ensemble Classifier and Parameter Optimization,” *J. Appl. Eng. Technol. Sci. JAETS*, vol. 5, no. 1, Art. no. 1, Dec. 2023, doi: 10.37385/jaets.v5i1.2169.

[4] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, “Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview,” *Sensors*, vol. 20, no. 9, p. 2734, May 2020, doi: 10.3390/s20092734.

[5] D. N. Fitriani, “Prediksi PREDIKSI TINGKAT OBESITAS MENGGUNAKAN NEURAL NETWORK: PENDEKATAN KLASIFIKASI BINER,” *PARAMETER J. Mat. Stat. Dan Ter.*, vol. 3, no. 01, Art. no. 01, Apr. 2024, doi: 10.30598/parameterv3i01pp85-92.

[6] L. Setiyani, A. N. Indahsari, and R. Roestam, “Analisis Prediksi Level Obesitas Menggunakan Perbandingan Algoritma Machine Learning dan Deep Learning,” *JTERA J. Teknol. Rekayasa*, vol. 8, no. 1, Art. no. 1, Jun. 2023, doi: 10.31544/jtera.v8.i1.2022.139-146.

[7] I. Sabilirrasyad, M. Hermansyah, N. A. Prasetyo, A. Muliawan, and A. Wahid, “Unveiling X/Twitter’s Sentiment Landscape: A Python Crawler That Maps Opinion Using Advanced Search,” *LOREM Comput. Eng. Comput. Inf. Syst.*, vol. 1, no. 1, Art. no. 1, Jan. 2024.

[8] M. Hermansyah, N. A. Prasetyo, A. Muliawan, M. F. Firdausi, and F. Wiranto, “Classification of Student Readiness for Educational Unit Exams: Decision Tree Approach C4.5 Based on Try Out Scores at MTs Nahdlatul Arifin,” *LOREM Comput. Eng. Comput. Inf. Syst.*, vol. 1, no. 1, Art. no. 1, Jan. 2024.

.

[9] R. Dutta, I. Mukherjee, and C. Chakraborty, “Obesity disease risk prediction using machine learning,” *Int. J. Data Sci. Anal.*, pp. 1–10, Jan. 2024, doi: 10.1007/s41060-023-00491-9.

[10] F. Ferdowsy, K. S. A. Rahi, Md. I. Jabiullah, and Md. T. Habib, “A machine learning approach for obesity risk prediction,” *Curr. Res. Behav. Sci.*, vol. 2, p. 100053, Nov. 2021, doi: 10.1016/j.crbeha.2021.100053.

[11] Z. Zhao *et al.*, “Risk factor analysis and risk prediction study of obesity in steelworkers: model development based on an occupational health examination cohort dataset,” *Lipids Health Dis.*, vol. 23, no. 1, p. 10, Jan. 2024, doi: 10.1186/s12944-023-01994-x.

[12] F. Wiranto, M. A. Rohim, M. F. Firdausi, A. Muliawan, and E. Afrianto, “Optimizing Coffee Crop Selection for Plant-Ready Coffee Fields: A Study Using the ‘Predict.in’ Decision Support System at the Coffee and Cocoa Research Center in Jember,” *Prog. Conf.*, vol. 6, no. 1, pp. 42–47, Nov. 2023.