Predicting Obesity Level using Machine Learning Models: Random Forest, XGBoost and LightGBM

Anggota Kelompok
Nicholas Farandi Harjanto 2602065553
Bryan Orville Audric 2602160750
Edrico Putra Pramana 2602078133
Darren King Wijaya 2602145232

Abstract—*Obesity is one of the biggest health concerns globally, with one eighth of the population affected in the entire world. This study proposes the usage of machine learning to predict obesity levels with the aim of creating a tool that can act as a preventive measure against obesity. The dataset mainly includes measured features of a person, including their lifestyle as well as their obesity level. This paper evaluates three models: Random Forest, XGBoost, and LightGBM. The performance of the models was assessed using accuracy, precision, recall, and F1-score. Results show that both XGBoost and LightGBM achieved the accuracy of 89%, higher than Random Forest models.*

I. INTRODUCTION

In recent years, obesity has grown into a concerning global health issue. According to WHO, in 2022, one out of eight people in the world were living with obesity. Obesity is often overlooked by society, which is one of the main reasons for its growth. Knowing its causing factors will help to facilitate greater awareness of their health conditions. The impact of obesity is profound, affecting individuals and societies in multiple ways. Health-wise, obesity significantly increases the risk of cardiovascular disease, type 2 diabetes, various forms of cancer, respiratory issues, and musculoskeletal disorders. Psychologically, it can lead to higher rates of depression, anxiety, and low self-esteem, compounded by social stigma and discrimination. Consequently, people are more likely to improve their lifestyles and nutritional intake to avoid obesity [1].

Obesity can easily be identified with the help of Body Mass Index (BMI). Using a person's weight and height as part of the calculation, BMI is a measure to estimate the ideal body fat percentage. By using specific BMI cutoffs, a person's obesity level can be categorized as underweight, overweight, or obese. This method helps in identifying individuals at a higher risk of severity due to obesity [2][3]. There are many causing factors that contribute to one's obesity. An excess of energy intake and biological determinants alone are insufficient to account for obesity. Instead, obesity arises from an intricate and complex interaction among multiple factors. Identifying causing factors can be key factors into knowing what can be done to prevent obesity. Adopting a health-promoting lifestyle plays significant roles in mitigating or exacerbating these influences [4].

Over the past few years, the field of machine learning has achieved a range of remarkable growth in advanced learning algorithms and pre-processing techniques. Each of machine learning technique have their own characteristics, whether they are easy to interpret, robust to outlier, able to compute complex dataset, easy to implement or have the ability to prevent overfitting, a phenomena where prediction result between training and testing vary greatly. The aim of this study is to propose a machine learning models that is able to predict the obesity level based on the given factors as preventive measure against obesity. There are some attempts in predicting obesity level such as the Implementation of the K-Nearest Neighbour (KNN) Algorithm for Obesity Level Classification conducted by Ayu Made Surya Indra Dewi and Ida Bagus Gede Dwidasmara, they utilized the KNN algorithm and achieved an accuracy of 78.98% with a value of k = 2 [5]. The accuracy is relatively low due to the use of KNN which is sensitive to the outlier/noise in the dataset. Therefore, in this study, we are attempting to improve the accuracy, using the same dataset [5], with different machine learning models such as Random Forest, XGBoost and LightGBM to reach higher accuracy in order to for this model to be a viable way as a preventive measure against obesity.

II. METHODOLOGY

The research process is shown on Figure 1. The process starts by selecting the correct dataset to be used. The dataset is then cleaned by removing the data with null values and outlier. The value of Age and Weight category is also standardized. The features or columns are then filtered using Pearson correlation. The dataset is divided into training set and testing set with a ratio of 80% for training set and 20% for testing set. The training set will be used for the machine learning models such as Random Forest, XGBoost and LightGBM to train while the testing set is used for evaluating the performance of the model. Several metrics such as accuracy, recall, precision and f1 score will be used to evaluate the performance of the model.

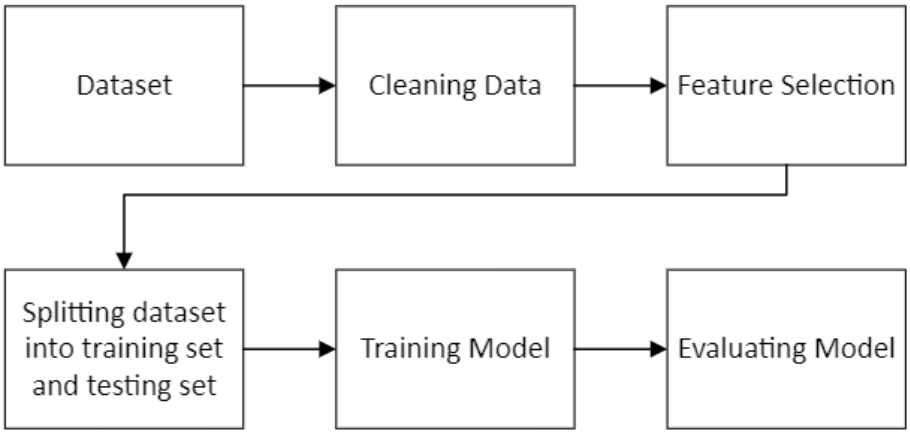

Figure 1. Research Process

A. Dataset

Since the goal is to improve the accuracy of previous research [5], the same dataset is implemented. The dataset contains 20751 data which consists of 16 features that represent obesity factors. The features are shown in Table I. and the label are shown in Table II. Further details of the dataset can be found in [6].

TABLE I. FEATURES OF DATASET

Features	Data type	Description
Gender	Categorical	-
Age	Numerical	-
Height	Numerical	-
Weight	Numerical	-
Family History with_Overweight	Categorical	-
FAVC	Categorical	Frequent consumption of high caloric food
FCVC	Numerical	Frequency of eating vegetables
NCP	Numerical	Number of main meals
CAEC	Categorical	Frequency of food consumption between main meals
SMOKE	Categorical	Smoking status
CH20	Numerical	Daily frequency of water consumption.
SCC	Categorical	Calories consumption monitoring
FAF	Numerical	Frequency of physical activity
TUE	Numerical	Time using technology devices
CALC	Categorical	Frequency of alcohol consumption
MTRANS	Categorical	Transportation used
NObeyesdad	Categorical	Obesity levels (Target)

TABLE II. LABEL OF DATASET

No	Body Weight Category
1	Insufficient weight
2	Normal weight
3	Overweight I
4	Overweight II
5	Obesity type I
6	Obesity type II
7	Obesity type III

B. Pearson Correlation

Pearson's correlation coefficient, a statistical metric, is instrumental in the feature selection process. This coefficient quantifies the linear relationship between features and the target variable, discerning their relevance. By calculating correlation coefficients for each feature in the dataset with respect to the target variable, researchers can gauge their strength of association. Features exhibiting strong correlations, closer to 1 or -1, are considered more pertinent for distinguishing obesity level. Conversely, features with coefficients near 0 indicate minimal linear relationship and are thus deemed less relevant for classification. The formula for calculating the Pearson correlation coefficient between two variables is given by:

$$r_{xy} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$

Where r_{xy} is the Pearson correlation coefficient between variables X and Y, X_i, and Y_i are individual values of variable X and Y , and $\overline{X}$ and $\overline{Y}$ are the means of the values of X and Y, respectively. This formula measures the extent to which variables X and Y vary together, providing a deeper understanding of the relationship between features in the dataset and the target variable.

C. Random Forest

Since the previous research[5] has trouble achieving high accuracy due to the existent of outliers in the dataset, Random Forest is used for this research as it is less sensitive to the outliers. Random Forest is a widely used ensemble learning model comprised of multiple decision trees. Each tree is constructed independently and randomly, with a subset of features selected randomly during the tree-building process. This feature selection technique helps mitigate overfitting and decorrelate the trees within the ensemble. Through bootstrap aggregating or bagging, Random Forest generates multiple bootstrap samples from the training data and builds a decision tree on each sample. During classification tasks, the model employs a majority voting mechanism, where each tree "votes" for the most prevalent class [7].

D. XGBoost

XGBoost is a scalable tree boosting system widely used in machine learning for its effectiveness and efficiency[8]. It works by sequentially building an ensemble of decision trees, where each tree corrects the errors of the previous ones through boosting. By optimizing an objective function that combines a loss function and regularization term, XGBoost minimizes errors while preventing overfitting. Similar to Random Forest, it is also less sensitive to outliers which is a huge issue in[5]. Leveraging gradient boosting, XGBoost fits trees to the gradient of the loss function, learning from previous mistakes. With parallel and distributed computing capabilities, XGBoost can efficiently utilize multiple CPU cores and scale to large datasets. It also provides insights into feature importance and is designed for scalability, handling billions of examples with minimal resources.

E. LightGBM Classifier

LightGBM is a model made as an improvement of XGBoost model. It is a state-of-the-art Gradient Boosting Decision Tree (GBDT) algorithm that revolutionizes traditional implementations by introducing novel techniques like Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS optimizes the training process by sampling data instances based on gradient values, focusing on instances with significant gradients to accurately estimate information gain while reducing computational overhead. EFB enhances efficiency by bundling related features together, improving cache hit rates and spatial locality. By leveraging these techniques, LightGBM accelerates model training, reduces memory consumption, and maintains high accuracy, making it a powerful tool for handling large feature dimensions and datasets in machine learning tasks [9].

F. Accuracy

Accuracy is one of the commonly used evaluation metrics in classification modeling to assess how well a model can classify data correctly overall. In this context, "correct" means the model's predictions match the true labels of the observed data. Accuracy is calculated by dividing the number of correct predictions by the total number of predictions made by the model. The formula for calculating the accuracy of a classification model is as follows:

$$Accuracy = \frac{Number\ of\ Correct\ Prediciton}{Total\ Number\ of\ Prediction}$$

G. Precision

Precision is a crucial evaluation metric in classification tasks, particularly in scenarios where minimizing false positives is important. It measures the accuracy of positive predictions made by the model, i.e., the proportion of correctly predicted positive instances out of all instances predicted as positive. The formula for calculating precision in a classification setting is as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

H. Recall Score

Recall, also known as sensitivity or true positive rate, is an essential evaluation metric in classification tasks, especially when it's crucial to capture all positive instances. It measures the ability of the model to correctly identify positive instances from all actual positive instances in the dataset. The formula for calculating recall in as classification setting is as follows:

$$Recall\ Score = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

I. F1-Score

The F1 score is a commonly used evaluation metric in classification tasks, which combines precision and recall into a single metric. It provides a balance between precision and recall, making it particularly useful when both false positives and false negatives need to be minimized. The formula for calculating the F1 score is:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where Precision and Recall are as previously defined. This formula computes the harmonic mean of precision and recall, emphasizing the balance between the two metrics. The F1 score ranges from 0 to 1, where a higher score indicates better performance. It is a useful metric for evaluating classification models,

especially in scenarios where achieving a balance between precision and recall is important, such as in information retrieval systems or medical diagnosis.

RESULT & DISCUSSION

A. Selected Features

The features selected using Pearson's correlation method are shown in Table III. Every feature has an absolute correlation coefficient value above a threshold value of 0.1. The threshold value was determined to be the best since any value than 0.1, the accuracy of the models started to decrease significantly. It is interesting to note that the feature with lowest correlation coefficient is Time with Technology represented as 'TUE' in the dataset and the feature with highest correlation coefficient is Weight.

TABLE III. SELECTED FEATURES

Features	Correlation Coefficient
Age	0.3670194416789297
Height	0.16803239219481939
Weight	0.922249548022407
Family History with Overweight	0.5086358618078486
FAVC	0.1914821991592605
FCVC	0.20358950303742987
CAEC	-0.35152261670757046
CH20	0.2571571349005148
SCC	-0.17869528879049948
FAF	-0.211909968337177
TUE	-0.11872702963270912
CALC	0.15792181183708395

B. Selected Features

The performance of the models is shown in the tables below:

TABLE IV. EVALUATION RESULT OF RANDOM FOREST

Metric	Insufficient Weight	Normal Weight	Overweight I	Overweight II	Obesity Type I	Obesity Type II	Obesity Type III
Precision	0.94	0.83	0.79	0.78	0.88	0.96	0,99
Recall	0.94	0.86	0.72	0,83	0.86	0.97	1
F1-Score	0.94	0.85	0.76	0,80	0.87	0.97	0,99
Accuracy	0.88						

TABLE V. EVALUATION RESULT OF XGBOOST

Metric	Insufficient Weight	Normal Weight	Overweight I	Overweight II	Obesity I	Obesity II	Obesity III
Precision	0,94	0,85	0,81	0,80	0,90	0,96	0.99
Recall	0,93	0,87	0,78	0,85	0,84	0,97	1
F1-Score	0,94	0,86	0,80	0,83	0,87	0,97	0.99
Accuracy	0.89						

TABLE VI. EVALUATION RESULT OF LIGHTGBM

Metric	Insufficient Weight	Normal Weight	Overweight I	Overweight II	Obesity I	Obesity II	Obesity III
Precision	0,93	0,86	0,81	0,81	0,88	0,96	0.99
Recall	0,94	0,86	0,78	0,85	0,85	0,96	0.99
F1-Score	0,94	0,86	0,79	0,83	0,86	0,96	0,99
Accuracy	0.89						

All three models already achieve higher accuracy results when compared to [5], which was previously 78%. Both models that stood out were XGBoost and LightGBM achieving an accuracy of 89% slightly higher than Random Forest. It is also seen that the variation of precision, recall and f1-score of XGBoost and LightGBM are relatively higher than Random Forest. However, it is not significant enough to suggest that XGBoost and LightGBM can predict more reliably than Random Forest.

It is interesting to note as well that predicting the categories "Insufficient Weight", "Obesity II", and "Obesity III" has the highest Precision, Recall, and F1-Score, indicating that the models are much more consistent at detecting the outside classes instead of the classes in between.

C. Impact in Real Life Setting

In real-life scenarios, obtaining the features of dataset used in the study may not always be straightforward. Parameters like frequency of vegetable consumption (FCVC), frequency of food consumption between meals (CAEC), smoking status (Smoke), daily water consumption (CH2O), and other lifestyle-related data require detailed personal information, which may be an inconvenience for some to gather. However, the information needed is still feasible to be obtained easily. If the information is collected and given to the models, it can be preventive measure to detect early signs of obesity before making a further consultation with the doctor.

CONCLUSION

This study demonstrated the use of machine learning models, which are XGBoost, Random Forest, and LightGBM to predict the obesity level. There was an improvement made as the accuracy of all the models were higher than what was achieved in [5]. It is shown that despite close competition between all three models, XGBoost and LightGBM provide the best result with 89% accuracy higher compared to Random Forest. These models can serve as a great tool for preventive healthcare with the aim of helping non-professionals to be able to check their conditions independently, before deciding it is necessary to make further appointments with the doctor.

FUTURE DIRECTION

Future research could go deeper on finding more accurate model as well as finetuning our method to further increase the performance in determining the classification of obesity level. Other ideas might be to collaborate with biology or healthcare professionals to create tools that can automatically collect parameters such as frequency of vegetable consumption (FCVC), frequency of food consumption between meals (CAEC), smoking status (Smoke), daily water consumption (CH2O), and other lifestyle-related data accurately, which enables machine learning models to function accurately.

REFERENCES

[1] K. D. da S. Ribeiro, L. R. S. Garcia, J. F. dos S. Dametto, D. G. F. Assunção, and B. L. L. Maciel, "COVID-19 and Nutrition: The Need for Initiatives to Promote Healthy Eating and Prevent Obesity in Childhood," *Childhood Obesity*, vol. 16, no. 4, pp. 235–237, Jun. 2020, doi: 10.1089/chi.2020.0121.

[2] D. Mohajan and H. K. Mohajan, "Obesity and Its Related Diseases: A New Escalating Alarming in Global Health," *Journal of Innovations in Medical Research*, vol. 2, no. 3, pp. 12–23, Mar. 2023, doi: 10.56397/JIMR/2023.03.04.

[3] Y. C. Chooi, C. Ding, and F. Magkos, "The epidemiology of obesity," *Metabolism*, vol. 92, pp. 6–10, Mar. 2019, doi: 10.1016/j.metabol.2018.09.005.

[4] F. Lehmann, G. Varnaccia, J. Zeiher, C. Lange, and S. Jordan, “Influencing factors of obesity in school-age children and adolescents - A systematic review of the literature in the context of obesity monitoring.,” *Journal of health monitoring*, vol. 5, no. Suppl 2, pp. 2–23, May 2020, doi: 10.25646/6729.

[5] A. M. S. I. Dewi and I. B. G. Dwidasmara, “Implementation Of The K-Nearest Neighbor (KNN) Algorithm For Classification Of Obesity Levels,” *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, vol. 9, no. 2, p. 277, Nov. 2020, doi: 10.24843/JLK.2020.v09.i02.p15.

[6] F. M. Palechor and A. de la H. Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico,” *Data Brief*, vol. 25, Aug. 2019, doi: 10.1016/j.dib.2019.104344.

[7] G. Biau and G. B. Fr, “Analysis of a Random Forests Model,” 2012.

[8] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[9] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” [Online]. Available: https://github.com/Microsoft/LightGBM.