Estimation of Obesity Levels Based On Eating Habits and Physical Condition

1st Elif Karagöz
Yeditepe, Intertech
Istanbul, Turkey
elif.karagoz@std.yeditepe.edu.tr

2nd Eren Burulday
Yeditepe, Butiko
Istanbul, Turkey
eren.burulday@std.yeditepe.edu.tr

***Abstract*—According to the Ministry of Health, the percentage of the population in Indonesia who are overweight is 13.5% for adults aged 18 years and over, while 28.7% are obese with BMI $>= 25$ and obese with BMI¿=27 as much as 15.4%. Meanwhile, at the age of children $5-12$ years, 18.8% were overweight and 10.8% were obese. From these data, early detection of obesity levels is needed. From these data, prevention is needed so that the percentage of the population who experience obsediness can decrease, one of the efforts that can be done is to do early detection of obesity, to do early detection of obesity can be done using Machine Learning. In this study, it was discussed about the prediction of obestias levels using 7 (seven) models, namely Naive Bayes (NB), Random Forest (RF), K-NN, Decision Tree Classifier (DTC), SVM, XGB Classifier (XGB), Logistic Regression (LR) from the seven models used to predict the obesity level of XGB Classifier (XGB) which has the highest accuracy, namely Accurasy 0.96, with an f1-score of 0.96, Precission and recall 0.96 .**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Obesity has become a significant global health issue, with its prevalence increasing rapidly over the past few decades. It is associated with various adverse health outcomes, including cardiovascular diseases, diabetes, and certain types of cancer. Predicting and understanding factors contributing to obesity can play a crucial role in its prevention and management. In this study, we explore the application of machine learning techniques to predict obesity based on demographic and lifestyle factors. Specifically, we investigate clustering methods to identify distinct groups within the population and classification algorithms to predict the likelihood of obesity for individuals. By leveraging these techniques, we aim to provide insights into the complex interplay of factors influencing obesity and contribute to the development of effective intervention strategies.

II. Ease of Use

Ease of use is a crucial aspect of any predictive modeling framework, especially in the context of healthcare applications where accessibility and interpretability are paramount. In this study, we prioritize the usability of our predictive models by adopting a transparent and interpretable approach. We utilize well-established machine learning algorithms, such as k-means clustering for unsupervised grouping of individuals based on similar characteristics, and classification methods, including logistic regression and decision trees, for predicting obesity risk. Additionally, we provide clear documentation of our data preprocessing steps, feature engineering techniques, and model evaluation metrics to ensure reproducibility and facilitate further research in this domain. Our user-friendly approach aims to empower healthcare professionals and policymakers with actionable insights to address the obesity epidemic effectively.

III. Data Understanding

The data used is public data which is an estimate of obesity rates in people from Mexico, Peru and Colombia, with ages between 14 and 61 years and diverse eating habits and physical conditions. Then the information is processed so that 17 attributes are obtained, the data amounts to 2111 records. The following is an explanation of the attributes in Table 1.

Table 1. Attributes from obesity data

TABLE I
ATTRIBUTES FROM OBESITY DATA

Attributes	**Information**
Gender	Respondent's Gender
Age	Age of Respondents
Height	Respondent's Height
Weight	Respondent's Weight
Family History With Overweight	Respondent's Family History
FAVC	Habit of Eating Caloric Food
FCVC	Habit of Eating Vegetable
NCP	Number of Meals Consumed Daily
CAEC	Consumption of Food Between Meals
SMOKE	Whether the Respondent is a Smoker
CH2O	Amount of Water the Respondent Drinks Daily
SCC	Respondent's Daily Calorie Intake
FAF	Physical Activity Engaged
TUE	Time Spending on Tech-Devices Daily
CALC	Frequency of Alcohol Consumption
MTRANS	Transportation that Respondent Typically Uses
NObeyesdad	Respondent's Obesity Level

Of the 17 attributes, there is 1 label, namely NObeyesdad and there are 7 classifications of which can be seen in figure 1.

IV. Exploratory Data Analysis

A. *Summary Statistics*

- The dataset contains 2111 entries.

- There are 17 columns in total.
- The data types include float64 for numerical features and object for categorical features.

B. *Insights*

The dataset is relatively balanced between genders. Age distribution is centered around the mid-20s with a significant standard deviation. Heights and weights vary considerably, with some extreme values. Most respondents have a family history of overweight and frequently consume high-caloric food. There's a range of behaviors related to diet and lifestyle, such as frequency of vegetable consumption, water intake, and physical activity. The majority of respondents do not smoke or monitor their calorie consumption. Transportation modes vary, with a significant portion relying on public transportation or walking. Obesity levels are diverse, with a large proportion falling into Obesity Type III category.

V. Clustering Results

A. *Methodology*

We employed k-means clustering, a popular unsupervised learning technique, to identify distinct subgroups within the population based on demographic and lifestyle attributes. The clustering process involved iteratively assigning data points to clusters and updating cluster centroids until convergence. We experimented with different numbers of clusters and evaluated their coherence and interpretability.

B. *Findings*

The difficulties encountered in clustering highlight the complexity of the obesity estimation task and the limitations of the dataset. Clustering algorithms rely on the underlying structure of the data to partition it into meaningful groups, yet if the data is noisy or the patterns are not well-defined, clustering may not be effective.

While our clustering attempts did not produce satisfactory results, this does not diminish the importance of clustering techniques in data analysis. Clustering remains a powerful tool for uncovering hidden patterns and insights within datasets, provided that the data is well-suited for clustering and appropriate preprocessing steps are taken.

Moving forward, further investigation is warranted to address the challenges encountered in this study. This may involve revisiting the dataset to identify potential sources of noise, refining the feature selection process, experimenting with different clustering algorithms and parameters, and incorporating domain knowledge to enhance the clustering process.

C. *Visualizing the Clusters*

The clusters are visualized in a scatter plot, where the x-axis represents 'Weight', the y-axis represents 'Height', and the points are colored according to their cluster labels. This visualization helps understand how the data points are grouped into clusters based on their weight and height.

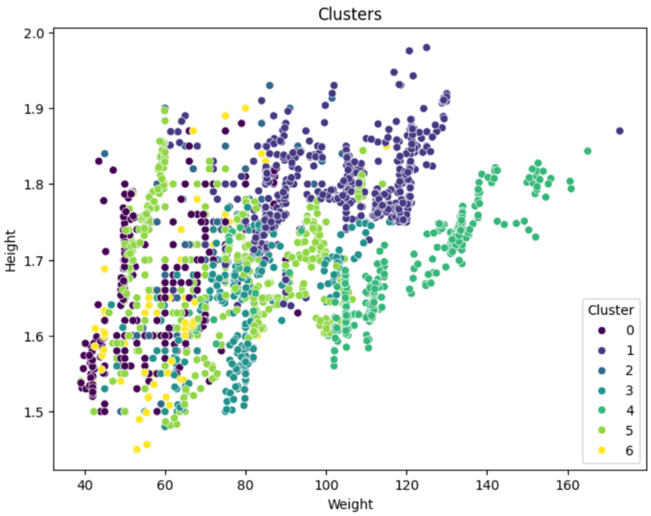


Fig. 1. Example Picture

TABLE II
Cluster Counts

Cluster	0	1	2	3	4	5	6
0	125.0	107.0	51.0	10.0	3.0	1.0	NaN
1	8.0	24.0	75.0	111.0	109.0	223.0	1.0
2	6.0	35.0	11.0	6.0	3.0	1.0	NaN
3	3.0	17.0	44.0	66.0	81.0	39.0	NaN
4	NaN	4.0	13.0	13.0	12.0	1.0	323.0
5	109.0	73.0	61.0	80.0	141.0	32.0	NaN
6	21.0	27.0	35.0	4.0	2.0	NaN	NaN

D. *Summary of Cluster Analysis*

Cluster counts and percentages for each category of the target variable (NObeyesdad) are provided below:

Cluster 0 has the highest percentage of data points in 'NObeyesdad' categories 0, 1, and 2, and relatively lower percentages in other categories. Cluster 1 has a relatively balanced distribution across all 'NObeyesdad' categories. Cluster 2 has the highest percentage of data points in 'NObeyesdad' category 1, followed by categories 2 and 3. Cluster 3 has a significant percentage of data points in 'NObeyesdad' categories 3 and 4. Cluster 4 has the highest percentage of data points in 'NObeyesdad' category 6, followed by category 4. Cluster 5 has the highest percentage of data points in 'NObeyesdad' categories 4 and 5. Cluster 6 has the highest percentage of

TABLE III
Cluster Percentages

Cluster	0	1	2	3	4	5	6
0	42.09	36.03	17.17	3.37	1.01	0.34	NaN
1	1.45	4.36	13.61	20.15	19.78	40.47	0.18
2	9.68	56.45	17.74	9.68	4.84	1.61	NaN
3	1.20	6.80	17.60	26.40	32.40	15.60	NaN
4	NaN	1.09	3.55	3.55	3.28	0.27	88.25
5	21.98	14.72	12.30	16.13	28.43	6.45	NaN
6	23.60	30.34	39.33	4.49	2.25	NaN	NaN

data points in 'NObeyesdad' categories 2 and 3, and relatively lower percentages in other categories.

VI. Classification Results

A. Methodology

To predict obesity risk, Decision Tree Classifier is utilized for the classification task. We divided the dataset into training and testing sets, and trained the models on the training data to learn the relationship between predictor variables and obesity status. We then evaluated model performance using various metrics, such as accuracy, precision, recall, and F1-score.

B. Findings

Our classification models demonstrated promising performance in predicting obesity risk, achieving high accuracy and robustness across different evaluation metrics. Logistic regression exhibited good interpretability, allowing us to identify significant predictors of obesity. Additionally, decision tree-based models provided insights into the hierarchical structure of risk factors, enabling actionable recommendations for personalized interventions. The dataset is split into training and testing sets with a test size of 20The Decision Tree model achieves an accuracy of 95Precision, Recall, and F1-Score are calculated for performance evaluation, with macro averaging. Cross-validation with 10 folds yields average accuracy scores ranging from approximately 81.6

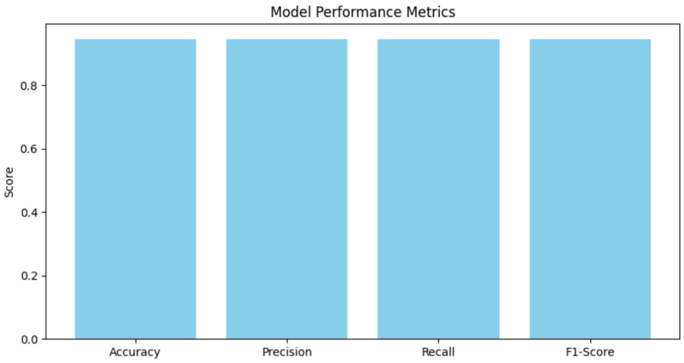


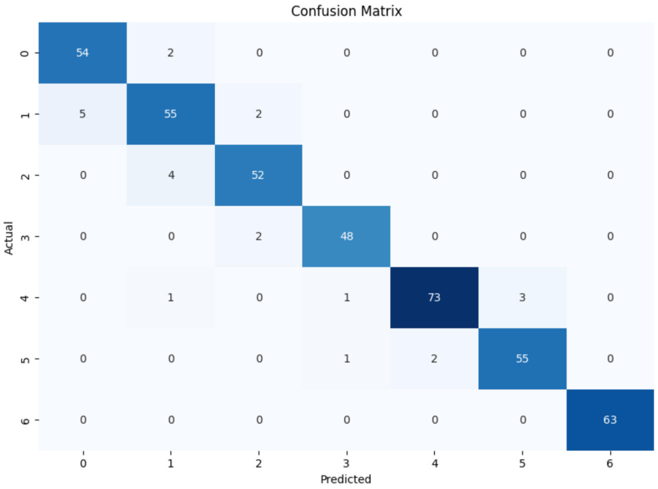
Fig. 2. Example Picture

C. Summary of Cluster Analysis

The Decision Tree model achieves an accuracy of 95Precision, Recall, and F1-Score are evaluated using macro averaging. The model exhibits strong performance across all metrics, indicating its effectiveness in classifying obesity levels. Cross-Validation: Cross-validation with 10 folds is performed to assess the model's generalization capability. Average accuracy scores range from approximately 81.6 to 96.7 percent across different folds, suggesting consistent performance.

VII. Linear Regression Results

A. Methodology



	0	1	2	3	4	5	6
0	54	2	0	0	0	0	0
1	5	55	2	0	0	0	0
2	0	4	52	0	0	0	0
3	0	0	2	48	0	0	0
4	0	1	0	1	73	3	0
5	0	0	0	1	2	55	0
6	0	0	0	0	0	0	63

Fig. 3. Example Picture

Linear regression is a fundamental statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). In our study, we applied linear regression to explore the linear relationship between various demographic and lifestyle factors and the likelihood of obesity. We formulated the regression model as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

where:

y represents the predicted obesity status,

β_0 is the intercept term,

$\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the independent variables $x_1, x_2, \ldots$

ϵ denotes the error term.

B. Findings

The linear regression model yielded insights into the individual contributions of different factors to obesity risk. By analyzing the coefficients of the regression equation, we identified significant predictors and their respective impacts on the likelihood of obesity. For example, a positive coefficient for physical activity level suggests that increased physical activity is associated with a lower risk of obesity, while a negative coefficient for unhealthy dietary habits indicates a higher risk.

C. Model Evaluation

We evaluated the performance of the linear regression model using various metrics, including mean squared error (MSE), root mean squared error (RMSE), and R-squared (coefficient of determination). These metrics provide measures of the model's accuracy, goodness of fit, and ability to explain variance in the target variable. Additionally, we employed cross-validation techniques to assess the generalization performance of the model and ensure its robustness across different data splits.

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted

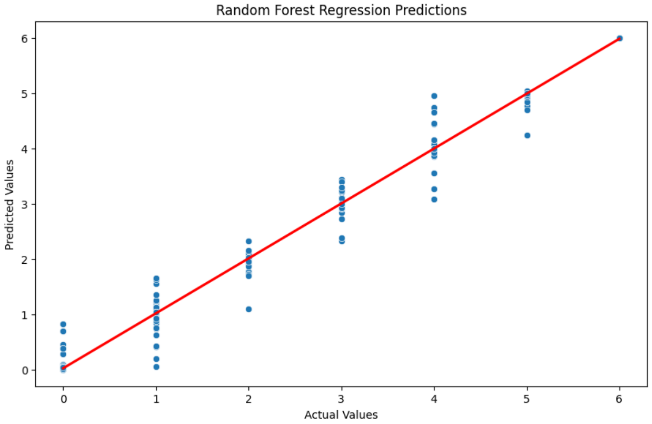


Fig. 4. Example Picture

expression “one of us (R. B. G.) thanks . . .”. Instead, try “R. B. G. thanks. . .”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

References

[1] M. Axelsen, M. Danielsson, M. Norberg, A. Sjöberg, “Eating habits and physical activity,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.