Research Article

Classification and Prediction of Obesity Levels among Subjects in Colombia, Peru, and Mexico Using Unsupervised and Supervised Learning

Suhaila Bahrom [1,*], Anuar Ab Rani[2], and Aisyah Amalina Mohd Noor[3]

[1] Universiti Malaysia Pahang Al-Sultan Abdullah; suhaila_b@iium.edu.my; ORCID ID: 0009-0002-2172-9282
[2] Universiti Malaysia Pahang Al-Sultan Abdullah; csm22010@student.umpsa.edu.my
[3] Universiti Malaysia Pahang Al-Sultan Abdullah; aisyah.amalina@gmail.com
* Correspondence: suhaila_b@iium.edu.my; 0148405251.

Abstract: *This research investigates the multifaceted relationship between various factors and obesity rates in Mexico, Peru, and Colombia using a publicly available dataset. Through Python, the study employs classification and clustering analyses, focusing on logistic regression to predict obesity levels and generate actionable recommendations. Combining exploratory data analysis (EDA) and advanced machine learning techniques, the research aims to unveil nuanced insights into obesity determinants. Unsupervised learning methods segmentize individuals, providing deeper insights into obesity profiles. Supervised learning algorithms like logistic regression, random forest, and adaboost classifier predict obesity levels based on labelled datasets, with random forest exhibiting superior performance. The study enhances understanding of obesity classification through machine learning and integrates data inspection, formatting, and exploration using Excel, Python, and graphical user interfaces (GUIs) such as SweetViz and PandaGui. Overall, it offers a comprehensive approach to understanding and addressing obesity using sophisticated analytical tools and methodologies.*

Keywords: supervised learning; unsupervised learning; machine learning.

Copyright: © 2024 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. INTRODUCTION

Obesity is a complex chronic disease by possessing too much body fat. With the development of contemporary technology, people are making every effort to reduce the negative effects of leading an unhealthy lifestyle, which is a primary cause of obesity. There were several implications of obesity in human society (Corvalán et al., 2017). While obesity was once considered a sign of affluence and social standing in certain societies, many other civilizations tried to comprehend the long-term risks that obesity brought (Omkar & Nimala, 2022). Medical studies seem to be moving toward a more contemporary approach to treating obesity over time. This was also led on by an increase in the population that is obese in society (Du et al., 2022). Although researchers have been able to discover the factors that contribute to obesity, there is lack of investigation on the systematic techniques to identify the level of obesity at the early stage. Modern investigative models can be built with the advances in machine learning to create precise and accurate programming routines at very low computing cost (Babajide et al., 2020; Cuevas et al., 2009). This would ideally involve striking a balance between the quality of the conventional technique elements.

This project employs a comprehensive methodology that includes exploratory data analysis, unsupervised learning for segmentation and classification, and supervised learning algorithm for predicting the key factors influencing obesity levels. The project's scope involves a meticulous examination of data, including detailed inspection and formatting carried out using Excel, followed by Python-based processes for data cleaning and transformation (McKinney, 2013). User-friendly data exploration tools such as SweetViz and PandasGui contribute to a more accessible understanding of the dataset. The utilization of advanced classification and clustering analyses in Python further enhances the depth of insights extracted from the data. The detailed findings derived from the project's analyses are instrumental in constructing practical recommendations. In the realm of health and wellness, where the prevalence of obesity is a mounting concern, our project is strategically designed to address this challenge through a proactive and data-driven approach. With a specific focus on multinomial logistic regression, we aim to predict obesity levels, providing a nuanced understanding of the contributing factors within the dataset.

2. METHOD & MATERIAL

The dataset used in this analysis is from "Estimation of Obesity Levels based on Eating Habits and Physical Condition" dataset in UCI Machine Learning Repository. This dataset include data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 variables and 1866 records. Nine categorical variables used in this analysis are obesity level, usage transportation, frequency of alcohol consumption, daily calories monitoring, smoking, feed between meals, high caloric frequency, family overweight and gender. For numerical variables, includes entertainment spending hours, physical activities frequency, daily water consumption, daily meals, vegetable consumption, weight, height, and age. The variables used in this analysis are listed in detail in Table 1.

Table 1. Types of variables

Variable	Type of Variables	Level of Measurement
Gender	Categorical	Nominal
Age	Numerical	
Height	Numerical	
Weight	Numerical	
Family history of overweight	Categorical	Nominal
Frequently consumed high-calorie food	Categorical	Nominal
Frequency of consumption of vegetables	Numerical	Ordinal
Number of main meals	Numerical	Ordinal
Consumption of food between meals	Categorical	Ordinal
SMOKE	Categorical	Nominal
Consumption of water daily	Numerical	Ordinal
Monitor calorie intake	Categorical	Nominal
Frequency of physical activity	Numerical	Ordinal
Entertainment spending hours	Numerical	Ordinal
Consumption of alcohol	Categorical	Nominal
Type of transportation used	Categorical	Nominal
Level of obesity	Categorical	Ordinal

2.1 Data Preprocessing

Data preprocessing is an important step in the machine learning and data analysis to improve the data's quality and reliability by addressing issues such missing values, outliers, and inconsistent formats (Müller & Guido, n.d.). Figure 1 shows the ETL Pipeline for Pre-Processing and Exploratory Data Analysis.

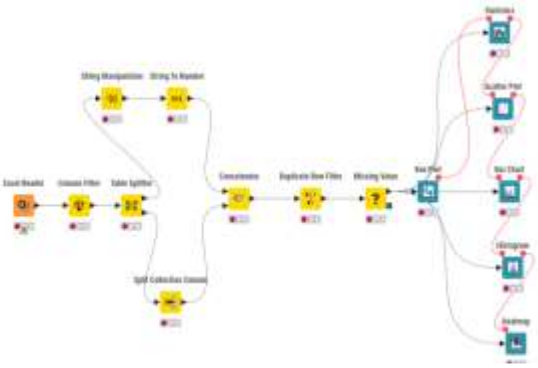


Figure 1. ETL Pipeline for Pre-Processing and Exploratory Data Analysis

2.2 Data Cleaning

Through a systematic data cleaning procedure, the presence of 29 instances of data duplication within the raw dataset was identified. The drop_duplicates() function was used to eliminate duplicated rows. This process is important for ensuring consistency, efficiency, accuracy, and the reliability of the analysis result. The next important step in data preprocessing was to handle the missing values. In this analysis, there is no missing data. The comprehensive Exploratory Data Analysis (EDA) undertaken in the initial phases of the project revealed a dataset entirely devoid of missing values, highlighting a rare and advantageous quality.

2.3 Data Transformation

In this analysis, the outliers exist from numerical and categorical are being assess systematically. There are outliers in four numerical variables: age, weight, height, and daily meals. These outliers have been purposefully left untreated to preserve these anomalies for a closer look down the road and to assess how robust the models are. In this study, data transformation is applied to enhance the quality and usability of the data. Categorical variables are encoded to transform them into a numerical format appropriate for analysis. We use data splitting in this analysis to separate the numerical dataset from the categorical dataset so that we can ensure that the model is tested on a different, previously untested subset, validated on another subset, and trained on.

3. FINDINGS

3.1 Exploratory Data Analysis

Our final dataset consists of 17 columns and 1844 rows. Eight of these are variables are numerical values. Nine categorical variables represent the category variables. For exploratory data analysis in this investigation, we employ fundamental statistical EDA tools including SweetViz and Panda GUI. When searching through the data for patterns, trends, or clusters that can direct future investigations or theoretical frameworks, EDA will be very helpful. Part of the output from EDA using SweetViz is seen in Fig.2 and Fig.3. The best method for locating natural categories in a dataset is clustering. Thus, in order to categorise groups with comparable attributes, we employ clustering analysis. We use PandaGUI to identify the number of cluster suitable to be used in this analysis.

Figure 2. EDA from SweetViz

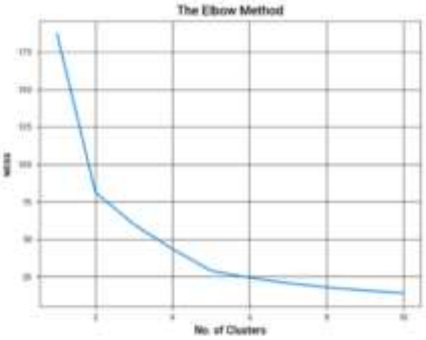
Figure 3. Relationship between variables

3.2 Classification and Clustering Analysis

The two main categories of pattern recognition methods are clustering and classification. By using these methods, measurements that are tangentially connected to the property of interest can be used to classify samples based on a certain attribute (such as the type of fuel responsible for an underground spill). When the measurements and the characteristic of interest are known for a group of samples, an empirical relationship or classification rule can be created. After that, samples outside of the initial training set can be predicted to possess the property using the classification rule (Lavine, 1992).

3.2.1 Unsupervised Learning

Unsupervised learning aims to identify structures, relationships, or patterns in the data without the need for pre-established target labels (Naeem et al., 2023). In this analysis we are using cluster analysis to segregate groups with similar traits and assign them into clusters in unsupervised learning. The purpose of cluster analysis, a multivariate data mining approach, is to group items according to a set of user-selected criteria or characteristics. To segmentize and classify individuals according to their obesity levels using unsupervised learning, K-means clustering proves to be a valuable tool. The first step involves selecting relevant features related to obesity to serve as inputs for the algorithm. Based on the Elbow method shows in Fig. 4, the number of clusters, denoted by 'k,' which could represent the different obesity levels is 5.



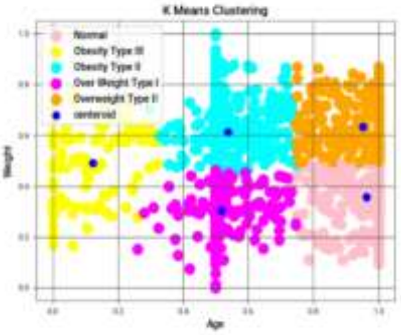


Figure 4. The Elbow Method for Weight vs Age

Figure 5. K Means Clustering for Weight vs Age

The algorithm is then applied to the dataset, iteratively assigning individuals to clusters, and updating cluster centroids until convergence (Chatterjee et al., 2020). The resulting clusters can be interpreted as distinct segments or groups of individuals with similar characteristics. By analysing the features and patterns within each cluster, the obesity level can be classified according to their age and weight. Figure 6 representing a tree or known as dendrogram. In hierarchical clustering, it illustrates the arrangement of the clusters produced by the corresponding analyses. The dendrogram visually depicts the hierarchical relationships and natural groupings within the dataset, highlighting patterns and structure based on the individuals' age and weight (Santisteban Quiroz, 2022). Another variable that is suitable for clustering is between weight and height. The Elbow method in Fig.4 that the individuals can be clustering based on 5 cluster according to their weight and height. This cluster can be visualized in Figure 5.

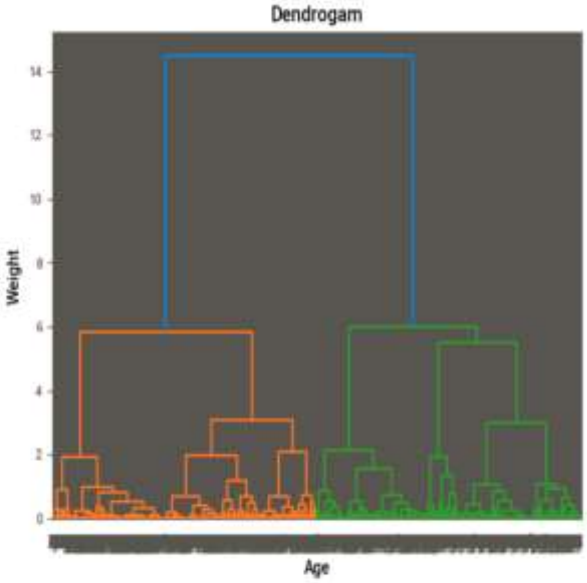
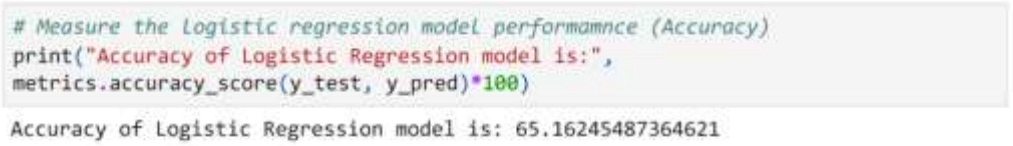


Figure 6. Dendrogram for Weight vs Age

3.2.2 Supervised Learning

3.2.2.1 Logistic Regression

Logistic regression has the capacity to account for several factors and allows the use of continuous or categorical predictors (Stoltzfus, 2011). This project uses logistic regression to predict obesity level among people from Colombia, Peru and Mexico. The accuracy value of 65.16245 for the logistic regression model predicting obesity levels among people in Colombia implies that the model is able to correctly classify the obesity levels for around 65.16% of the individuals in the dataset as shown in Fig.7. In the context of predicting obesity, this means that the model is moderately effective in distinguishing between individuals with different obesity levels based on the features considered in the logistic regression.



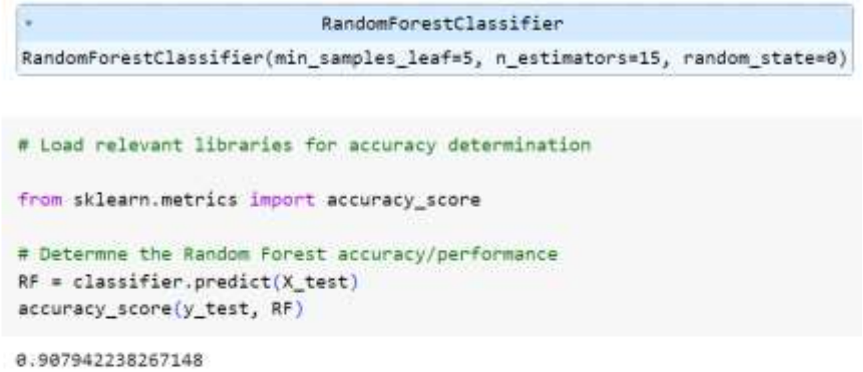
```
# Measure the Logistic regression model performamnce (Accuracy)
print("Accuracy of Logistic Regression model is:",
metrics.accuracy_score(y_test, y_pred)*100)
```

```
Accuracy of Logistic Regression model is: 65.16245487364621
```

Figure 7. Accuracy level from logistic regression

3.2.2.2 Random Forest Classifier

The random forest classifier is one type of machine learning method that belongs to the ensemble learning category. It is a collection of decision trees that are aggregated by majority rule. To increase overall performance, ensemble learning combines the predictions of several different separate models (Chen & Guestrin, 2016). During training, the random forest method creates several decision trees and combines their predictions to improve the classification's robustness and accuracy. Figure 8 shows the accuracy level from random forest classifier which is 90.79%.



```
RandomForestClassifier
RandomForestClassifier(min_samples_leaf=5, n_estimators=15, random_state=0)
```

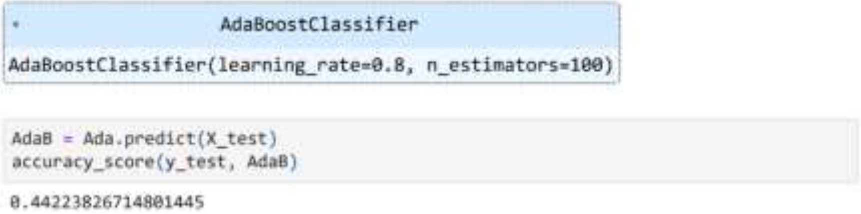
```
# Load relevant libraries for accuracy determination

from sklearn.metrics import accuracy_score

# Determne the Random Forest accuracy/performance
RF = classifier.predict(X_test)
accuracy_score(y_test, RF)
```

```
0.907942238267148
```

Figure 8. Accuracy level from Random Forest Classifier

3.2.2.3 Adaboost Classifier

A machine learning ensemble approach that is specifically meant for classification tasks is adaptive boosting, or adaBoost. AdaBoost, is a boosting technique works by iteratively adding one classifier at a time. Each classifier is trained on a selectively sampled subset of the training instances, where the sampling distribution starts uniform but changes with each iteration by attributing higher selection probabilities to the instances that were misclassified in the previous iteration. This way, the algorithm gathers classifiers that have their training each time more focused on the harder instances of the training set. Figure 9 depicts the results of accuracy from adaboost classifier which is 44.22%.



```
AdaBoostClassifier
AdaBoostClassifier(learning_rate=0.8, n_estimators=100)
```

```
AdaB = Ada.predict(X_test)
accuracy_score(y_test, AdaB)
```

```
0.44223826714801445
```

Figure 9. Accuracy level from AdaBoost Classifier

3.2.3 Confusion Matrix

Confusion matrices represent counts from predicted and actual values. A confusion matrix represents a classification model's performance on a set of test data where true values are known. The illustration of confusion matrix is shown in Figure 10. The diagonal values in a confusion matrix show the number of accurate predictions made by our model for each class. It indicates how many instances were correctly classified. Seeing a positive value in the visualization suggests our model has high precision. To identify the main factors that contribute to the obesity level, we visual using bar chart as shown in Figure 11. The highest factor or variable that contribute to the obesity level is the weight.

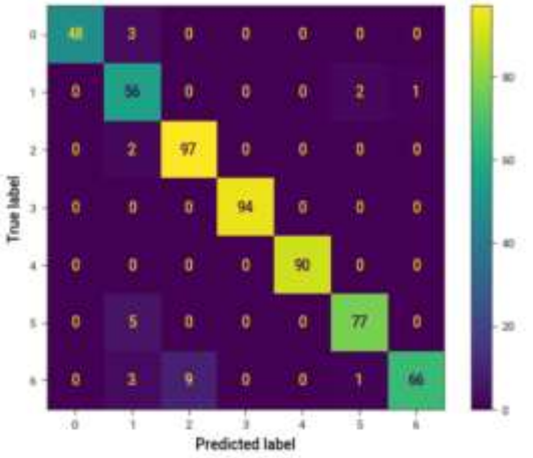


Figure 10. Confusion Matrix

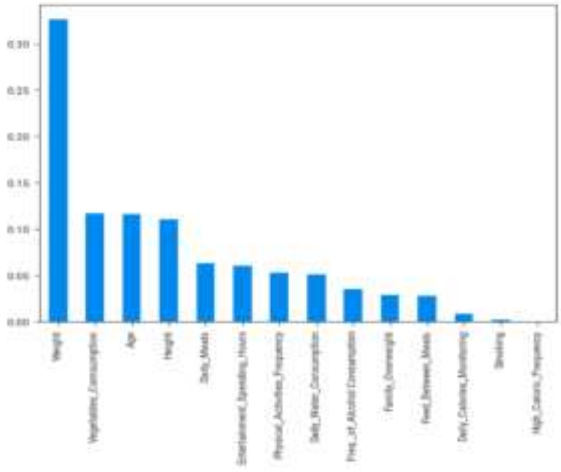


Figure 11. Factor contribute to obesity

4. DISCUSSION

The machine learning techniques can be applied to obesity level prediction. Within the current study, the random forest model emerged as a standout performer in forecasting obesity levels with 90.79% accuracy. These results highlight the precision of the random forest model for classifying and predicting obesity level. Notably, the identification of weight as a variable with a high correlation to obesity levels aligns with established medical knowledge. This finding shows the importance of weight as a main factor in predicting obesity level. In the context of obesity classification, random forest model suggests its suitability for predictive modelling with the high level of accuracy. These findings not only contribute to the refinement of predictive models but also emphasize the critical role of certain variables, particularly weight, in informing interventions and healthcare strategies aimed at addressing obesity across diverse populations (Banna, 2019).

5. CONCLUSION

Based on the scope of this study, the model that shows the level of performance better than that of alternative methodologies is the random forest model. The approach used in this study gives positive impact not only to public health industry, but also to the policies makers and community as a whole. The growth of knowledge for the improvement of global health, the development of interventions, and the empowerment of individuals are the contributions to society. In addition to that, efforts can be made to present more publicly available datasets on obesity, which would be homogenous in terms of the features and descriptions. This would herald a more recent line of inquiry into developing a model capable of accurately predicting and identifying the overall world trend. After careful inspection of the data, relevant patterns, relationships, and information can be discover. Although the data used in this analysis does not represent Malaysian population, the use of data mining techniques is still very beneficial to society in terms of raising awareness and promoting healthy lifestyles all over the world. In response to the research findings on obesity factors and segmentation, suggests tailoring educational initiatives for different population segments, ensuring cultural relevance and accessibility.

Concurrently, the implementation of health screenings, guided by classification models, aims to identify individuals at various obesity risk levels. The subsequent integration of early intervention programs, including personalized health plans and support services, forms a cohesive strategy to proactively address obesity, emphasizing both prevention and timely intervention. This multifaceted approach seeks to enhance the effectiveness of public health initiatives by catering to the unique needs of diverse populations.

References

Babajide, O., Hissam, T., Anna, P., Anatoliy, G., Astrup, A., Alfredo Martinez, J., Oppert, J.-M., & Sørensen, T. I. A. (2020). *A machine learning approach to short-term body weight prediction in a dietary intervention program* (pp. 441–455). https://doi.org/10.1007/978-3-030-50423-6_33

Banna, J. (2019). Obesity prevention in children in latin america through interventions using technology. *American Journal of Lifestyle Medicine, 13*(2), 138–141. https://doi.org/10.1177/1559827618823320

Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Identification of risk factors associated with obesity and overweight—a machine learning overview. *Sensors, 20*(9), 2734. https://doi.org/10.3390/s20092734

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

Corvalán, C., Garmendia, M. L., Jones-Smith, J., Lutter, C. K., Miranda, J. J., Pedraza, L. S., Popkin, B. M., Ramirez-Zea, M., Salvo, D., & Stein, A. D. (2017). Nutrition status of children in Latin America. *Obesity Reviews, 18*(S2), 7–18. https://doi.org/10.1111/obr.12571

Cuevas, A., Alvarez, V., & Olivos, C. (2009). The emerging obesity problem in Latin America. *Expert Review of Cardiovascular Therapy, 7*(3), 281–288. https://doi.org/10.1586/14779072.7.3.281

Du, Y., Rafferty, A. R., McAuliffe, F. M., Wei, L., & Mooney, C. (2022). An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. *Scientific Reports, 12*(1), 1170. https://doi.org/10.1038/s41598-022-05112-2

McKinney, W. (2013). Python for Data Analysis (J. Steele, Ed.; First). O'reilly.

Müller, A. C., & Guido, S. (n.d.). *Introduction to machine learning with Python a guide for data scientists introduction to machine learning with python.*

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: comprehensive review. *International Journal of Computing and Digital Systems, 13*(1). https://doi.org/10.12785/ijcds/130172

Omkar, M., & Nimala, K. (2022). Machine learning based diabetes prediction using with AWS cloud. *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES),* 1–7. https://doi.org/10.1109/ICSES55317.2022.9914160

Santisteban Quiroz, J. P. (2022). Estimation of obesity levels based on dietary habits and condition physical using computational intelligence. *Informatics in Medicine Unlocked, 29,* 100901. https://doi.org/10.1016/j.imu.2022.100901

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic Emergency Medicine, 18*(10), 1099–1104. https://doi.org/10.1111/J.1553-2712.2011.01185.X