Assessing obesity levels through an analysis of dietary patterns and physical well-being in developing countries

Vika Li

Georgetown University, McCourt School of Public Policy

May 05, 2024

Assessing obesity levels through an analysis of dietary patterns and physical well-being in developing countries

Introduction

The rising prevalence of obesity in developing countries, particularly in Latin America, represents a significant public health challenge with complex socioeconomic and lifestyle dimensions (Bhurosy & Jeewon, 2014; Popkin, 2001). As countries experience transitions in diet and physical activity due to economic development, obesity rates are expected to continue escalating, compounding existing health issues like diabetes, cardiovascular diseases, and metabolic syndrome (Ellulu et al., 2014; Popkin et al., 2012). This is particularly concerning as studies highlight that the rate of increase in obesity may even surpass those in developed nations (Bhurosy & Jeewon, 2014).

For this data science project, I analyzed a comprehensive dataset from the UCI Machine Learning Repository, which includes 2111 responses from the countries of Mexico, Peru, and Colombia. The dataset encompasses 16 features ranging from demographics such as gender and age to eating habits and physical activities, including questions about the frequency of high-caloric food consumption and exercise. Each participant is distinctly labeled with the NObesity class variable, which categorizes obesity levels from Insufficient Weight to Obesity Types I, II, and III. Importantly, there are no missing values in this dataset, facilitating robust analysis.

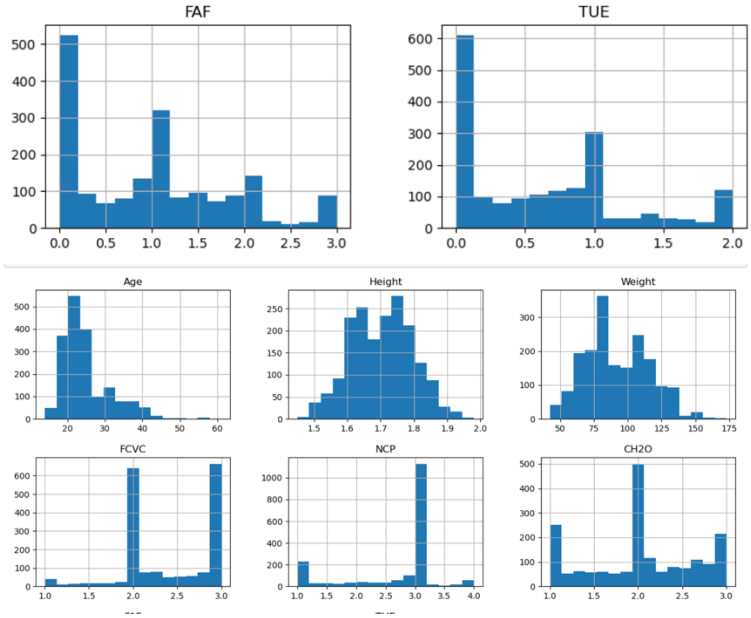
The previous research has identified a variety of factors contributing to obesity, including genetics, dietary choices, and physical inactivity. In developing regions, especially in Latin America, the shift towards higher consumption of calorie-dense and nutrient-poor foods is a significant driver (Monteiro et al., 2004; Uauy et al., 2001). However, the impact of socioeconomic status cannot be underestimated, as lower-income individuals often consume more affordable, energy-dense foods due to economic constraints, which contributes to higher obesity rates (Monteiro et al., 2004; Dinsa et al., 2012).

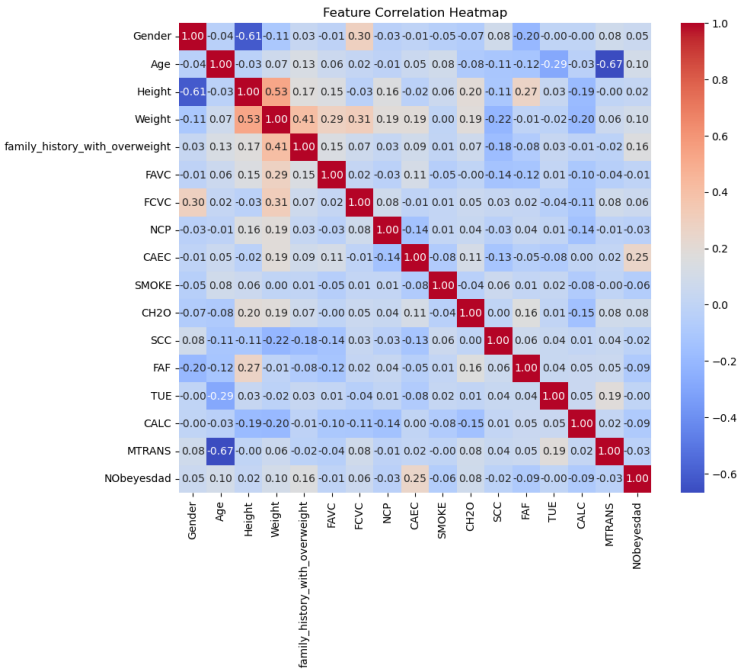
This research aims to explore the specific reasons behind the rise in obesity in Latin American countries and whether these drivers differ significantly from those observed in developed countries. Understanding these nuances is crucial for developing targeted interventions that address the unique challenges faced by populations in developing nations.

Descriptive Statistical Analysis and Preprocessing

The obesity dataset has no missing values and have 17 variables in total with 1068 males and 1043 females. Ages cluster around 22, ranging from 14 to 61. The majority of participants are between 20 and 30 years old. Heights run between 1.45m and 1.98cm. For weight, mode is 80kg and mean in 86kg. Most participants consume around 3 main meals per day, which is typical for most dietary guidelines. I also ran the correlation heat map to check if there any variables were

highly correlated. The map is attached below. Based on the heat map, family history and personal habits such as diet, physical activity, and even modes of transportation appear significantly connected to obesity outcomes.





For the preprocessing, I first encoded categorical and dummy variables into numerical format, including variables gender, family history with overweight, FAVC (high caloric food consumption), CAEC (snacks consumptions), smoke, SCC (calories monitor), CALC(alcohol consumption), MTRANS(transportation preference), and NObeyesdad (obesity level). Please see below for how I encoded these variables. The target variable NObeyesdad involves 7 labels,

running from insufficient weight to normal weight, to overweight, and to obesity. I decided to remove rows with insufficient weight target from the dataset since it was not the focus of the research question. Then, I got 1839 values left. For the remaining unique classifications, I got counts for each as indicated below:

```
d['NObeyesdad'].value_counts()
```

```
Obesity_Type_I         351
Obesity_Type_III       324
Obesity_Type_II        297
Overweight_Level_I     290
Overweight_Level_II    290
Normal_Weight          287
Name: NObeyesdad, dtype: int64
```

Since the target feature classes are relatively balanced, no class balance needed.

Machine Learning Models

In this analysis, I selected five diverse machine learning models to ensure a robust evaluation of the dataset's predictive capabilities. These models included Decision Tree, SVMs, KNN, Random Forest, and Gradient Boosting Machine (GBM). The choice of these models was strategic, each offering unique strengths in handling classification tasks:

- Decision Trees provide a simple and interpretable decision-making process, making it easy to visualize and understand how decisions are made.
- SVMs (Support Vector Machines) are effective in high-dimensional spaces, ideal for datasets with many features, although they require careful tuning of hyperparameters.
- KNN (K-Nearest Neighbors) is a non-parametric method that excels in scenarios where the decision boundary is irregular, and it is very easy to implement.
- Random Forest and Gradient Boosting Machine (GBM) are both ensemble methods that provide higher accuracy and stability by combining multiple models to reduce variance and bias.

I did the preliminary analysis for the five models. See below for the results. The results indicate that ensemble models like Random Forests and GBM outperformed other models, indicating their robustness and effectiveness for this dataset. Models with simpler decision boundaries like SVMs are not performing well. But I think it can be adjusted with hyperparameter testing. I'm very satisfied with the results, as for models like random forests and GBM, their prediction on certain classes can be 100% correct. And I don't think it is a overfitting issue since the accuracy scores are constantly high across these models except for SVMs.

Results for Random Forests:
Accuracy: 0.9538043478260869

	precision	recall	f1-score	support
0	0.86	0.95	0.90	63
1	0.97	0.98	0.98	64
2	1.00	1.00	1.00	66
3	1.00	1.00	1.00	59
4	0.96	0.87	0.91	61
5	0.94	0.91	0.93	55
accuracy			0.95	368
macro avg	0.96	0.95	0.95	368
weighted avg	0.96	0.95	0.95	368

Results for GBM:
Accuracy: 0.970108695652174

	precision	recall	f1-score	support
0	0.97	0.94	0.95	63
1	0.98	1.00	0.99	64
2	1.00	1.00	1.00	66
3	1.00	1.00	1.00	59
4	0.90	0.93	0.92	61
5	0.96	0.95	0.95	55
accuracy			0.97	368
macro avg	0.97	0.97	0.97	368
weighted avg	0.97	0.97	0.97	368

Results for Pruned Decision Tree:
Accuracy: 0.9211956521739131

	precision	recall	f1-score	support
0	0.85	0.95	0.90	63
1	0.91	0.94	0.92	64
2	1.00	0.98	0.99	66
3	1.00	1.00	1.00	59
4	0.90	0.77	0.83	61
5	0.87	0.87	0.87	55
accuracy			0.92	368
macro avg	0.92	0.92	0.92	368
weighted avg	0.92	0.92	0.92	368

Results for K-Nearest Neighbors:
Accuracy: 0.8994565217391305

	precision	recall	f1-score	support
0	0.94	0.70	0.80	63
1	0.86	0.94	0.90	64
2	0.96	0.97	0.96	66
3	0.98	1.00	0.99	59
4	0.84	0.92	0.88	61
5	0.84	0.87	0.86	55
accuracy			0.90	368
macro avg	0.90	0.90	0.90	368
weighted avg	0.90	0.90	0.90	368

Results for SVMs:
Accuracy: 0.5978260869565217

	precision	recall	f1-score	support
0	0.75	0.70	0.72	63
1	0.43	0.34	0.38	64
2	0.88	0.53	0.66	66
3	0.73	1.00	0.84	59
4	0.53	0.48	0.50	61
5	0.38	0.56	0.45	55
accuracy			0.60	368
macro avg	0.61	0.60	0.59	368
weighted avg	0.62	0.60	0.59	368

Then I implemented hyperparameters using grid search and cross validation to improve my model performance. See below for the updated model performance.

Results for K-Nearest Neighbors:
Accuracy: 0.970108695652174
Classification Report:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	63
1	0.98	1.00	0.99	64
2	1.00	1.00	1.00	66
3	1.00	1.00	1.00	59
4	0.92	0.92	0.92	61
5	0.96	0.95	0.95	55
accuracy			0.97	368
macro avg	0.97	0.97	0.97	368
weighted avg	0.97	0.97	0.97	368

Results for SVMs:
Accuracy: 0.970108695652174
Classification Report:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	63
1	0.98	1.00	0.99	64
2	1.00	1.00	1.00	66
3	1.00	1.00	1.00	59
4	0.92	0.92	0.92	61
5	0.96	0.95	0.95	55
accuracy			0.97	368
macro avg	0.97	0.97	0.97	368
weighted avg	0.97	0.97	0.97	368

Results for Random Forests:
Accuracy: 0.970108695652174
Classification Report:

	precision	recall	f1-score	support
0	0.95	0.95	0.95	63
1	0.98	1.00	0.99	64
2	1.00	1.00	1.00	66
3	1.00	1.00	1.00	59
4	0.92	0.92	0.92	61
5	0.96	0.95	0.95	55
accuracy			0.97	368
macro avg	0.97	0.97	0.97	368
weighted avg	0.97	0.97	0.97	368

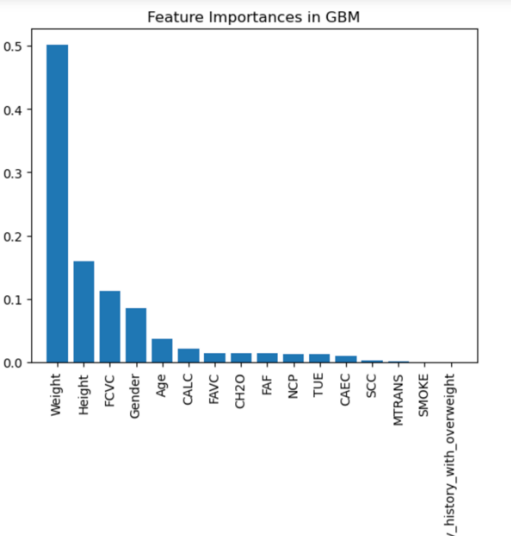
Results for GBM:
Accuracy: 0.970108695652174
Classification Report:

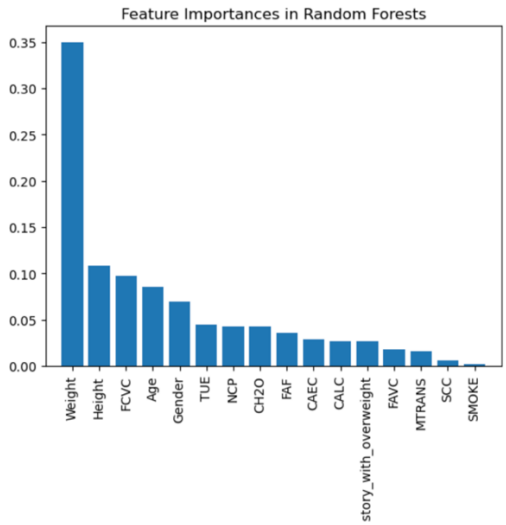
	precision	recall	f1-score	support
0	0.95	0.95	0.95	63
1	0.98	1.00	0.99	64
2	1.00	1.00	1.00	66
3	1.00	1.00	1.00	59
4	0.92	0.92	0.92	61
5	0.96	0.95	0.95	55
accuracy			0.97	368
macro avg	0.97	0.97	0.97	368
weighted avg	0.97	0.97	0.97	368

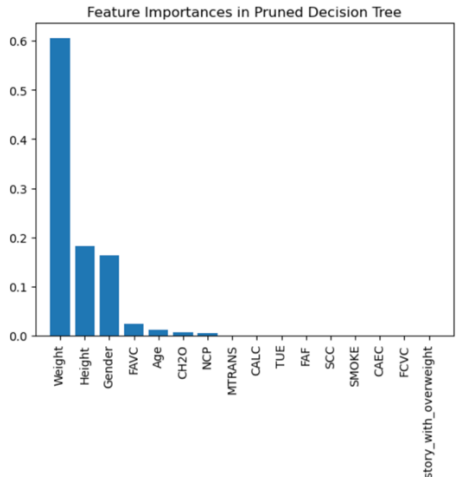
```
Results for Pruned Decision Tree:
Accuracy: 0.970108695652174
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95        63
           1       0.98      1.00      0.99        64
           2       1.00      1.00      1.00        66
           3       1.00      1.00      1.00        59
           4       0.92      0.92      0.92        61
           5       0.96      0.95      0.95        55

    accuracy                           0.97       368
   macro avg       0.97      0.97      0.97       368
weighted avg       0.97      0.97      0.97       368
```

Every model but GBM has improved metrics, showing a high level of accuracy and consistency across all models after hyperparameter tuning. Each model has achieved an accuracy of about 97.01%, and similar precision, recall, and F1-scores for each class in the classification report. The uniformity of results across various models suggests that the dataset may be well-defined and distinct, allowing even diverse algorithms to capture the necessary patterns effectively. The uniform high performance across different models could be indicative of well-defined features that strongly correlate with the outcomes, but it also makes me concern about overfitting. I might need to do another cross validation based on an unseen sample of data.

I also did the feature importance analysis to answer my research question, which features contribute to the obesity issue in Latin American population. And I performed the analysis for decision tree, random forests, and GBM models.







Factors including weight, height, gender, high caloric food consumption, vegetable consumption, and age are the leading contributors to obesity in Latin American countries.

Discussion and Implication

The comprehensive analysis utilizing various models revealed critical factors influencing obesity, such as weight, height, gender, high caloric food consumption, vegetable consumption, and age. These findings underscore the multifaceted nature of obesity, aligning with global research that cites similar contributors in both developed and developing regions (Ellulu et al., 2014; Bhurosy & Jeewon, 2014). The ability of our selected models to pinpoint these factors highlights their utility in epidemiological research and public health strategy formulation.

Particularly, the Random Forest and GBM models provided insights into the relative importance of each feature, guiding potential public health interventions aimed at mitigating obesity in Latin America. However, the uniform high performance across different models, while indicative of the dataset's well-defined nature, also raises concerns about potential overfitting. This necessitates further validation studies, possibly incorporating cross-validation with unseen data to confirm the models' predictive power and ensure they are not overly fitted to the training data.

This project has not only highlighted the critical predictors of obesity but also emphasized the importance of model selection in data science. By choosing models based on their strengths and the specific characteristics of the dataset, we can gain deeper and more accurate insights. Future studies could explore additional models or refine these models further with advanced hyperparameter tuning and feature engineering techniques to enhance predictive accuracy and reliability.

References

Bhurosy, T., & Jeewon, R. (2014). Overweight and Obesity Epidemic in Developing Countries: A Problem with Diet, Physical Activity, or Socioeconomic Status? Scientific World Journal, 2014, 964236. http://dx.doi.org/10.1155/2014/964236

Dinsa, G. D., Goryakin, Y., Fumagalli, E., & Suhrcke, M. (2012). Obesity and socioeconomic status in developing countries: a systematic review. Obesity Reviews, 13(11), 1067-1079.

Ellulu, M. S., Abed, Y., Rahmat, A., Ranneh, Y., & Ali, F. (2014). Epidemiology of obesity in developing countries: challenges and prevention. Global Epidemic Obesity, 2(2). http://dx.doi.org/10.7243/2052-5966-2-2

Monteiro, C. A., Moura, E. C., Conde, W. L., & Popkin, B. M. (2004). Socioeconomic status and obesity in adult populations of developing countries: a review. Bulletin of the World Health Organization, 82(12), 940-946.

Popkin, B. M. (2001). The nutrition transition and obesity in the developing world. The Journal of Nutrition, 131(3), 871S-873S.

Popkin, B. M., Adair, L. S., & Ng, S. W. (2012). Global nutrition transition and the pandemic of obesity in developing countries. Nutrition Reviews, 70(1), 3-21.

Uauy, R., Albala, C., & Kain, J. (2001). Obesity trends in Latin America: transiting from under- to overweight. Journal of Nutrition, 131(3), 893S-899S.

UCI Machine Learning Repository (2019). Estimation of obesity levels based on eating habits and physical condition. https://doi.org/10.24432/C5H31Z