Inah Lee

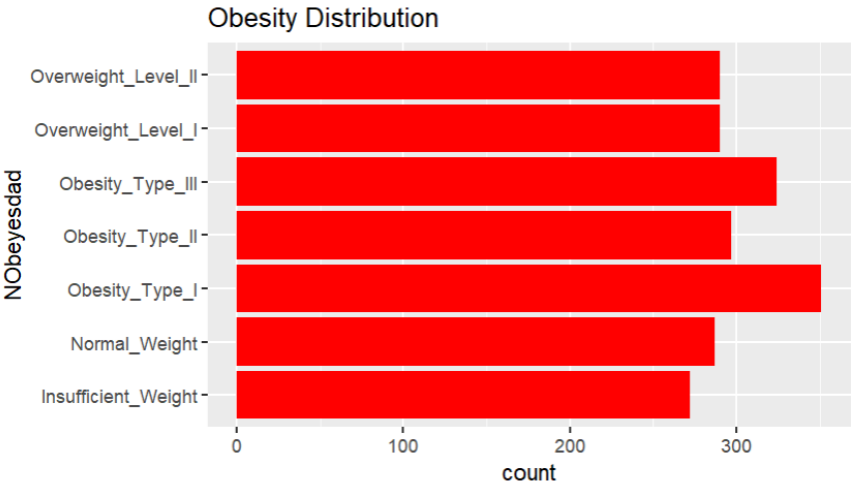
Data Analytics

4/23/2025

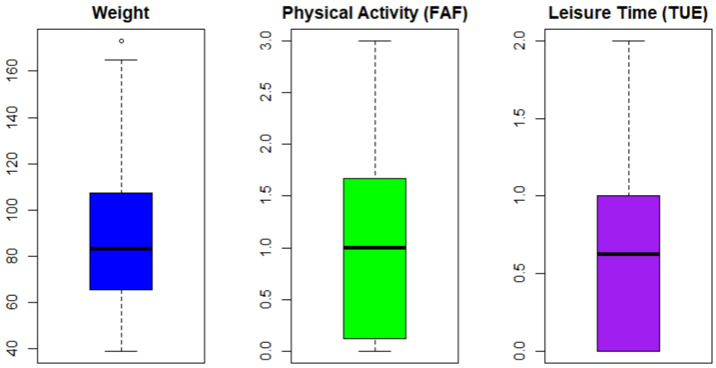
Obesity Classification and Weight Prediction

1. Exploratory Data Analysis

This project involves an exploratory data analysis of the ObesityDataSet_raw_and_data_sinthetic dataset to examine the relationships between statistical characteristics and variables. The dataset contains 2,111 observations and 17 variables, with NObeyesdad serving as the multi-class target variable comprising seven obesity levels.

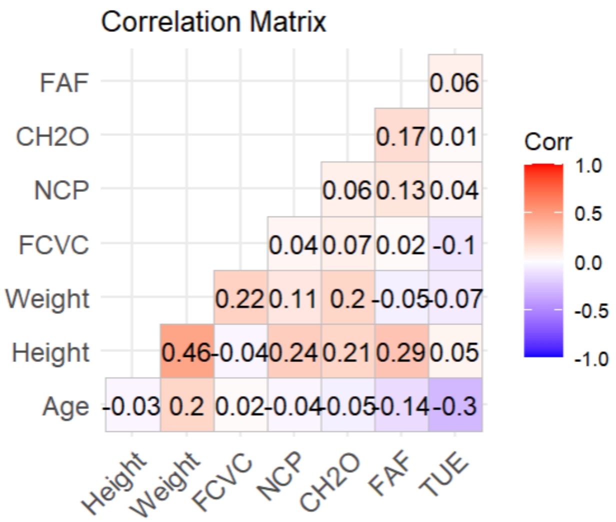


Visualization of the class distribution for NObeyesdad revealed that the majority of samples fall into the Overweight_Level_I, Overweight_Level_II, and Obesity_Type_I–III categories. The Normal_Weight and Insufficient_Weight classes also represented a notable proportion of the dataset. This distribution indicates that there is no significant class imbalance, making the dataset suitable for applying multi-class classification models.



Next, I examined potential outliers in the major variables using box plots. Weight ranged from approximately 40 kg to 165 kg, with some outliers present; however, their removal was deemed unnecessary as they were not considered problematic. FAF (physical activity frequency) was distributed between 0 and 3 hours, with a substantial portion of participants reporting no exercise at all. This variable may serve as an important factor distinguishing obese groups from non-obese groups. TUE (time using technology for leisure) was mostly concentrated within 1 hour, suggesting that many participants have limited leisure time.

Histogram analysis revealed that most numerical variables deviate from a normal distribution, exhibiting clear skewness. For example, Age was concentrated in the early 20s, and variables such as NCP (number of meals per day) and CH2O (daily water intake) displayed sharp peaks at specific values. Due to these distribution patterns, non-parametric classification models such as decision trees and random forests are expected to be more suitable than parametric models that assume normality, such as linear regression.



Finally, correlation analysis of the numerical variables showed that most had correlation coefficients of 0.3 or lower, with the highest being between Weight and Height (0.46). This indicates that there are no major limitations to using multiple features together in the same model.

In summary, the dataset is characterized by weak inter-variable correlations, non-normal distributions, and the presence of certain variables that can serve as clear classification criteria. Therefore, for the modeling stage, a logistic regression model can be applied as a baseline, complemented by a random forest classifier to potentially enhance performance.

---

2. Model Development, Validation and Optimization

For model development, the dataset was split into 60 percent training, 20 percent testing, and 20 percent validation sets. Only numerical variables related to lifestyle and physical characteristics such as Age, Height, FCVC (frequency of vegetable consumption), NCP (number of meals per day), CH2O (daily water intake), FAF (physical activity frequency), and TUE (technology use for leisure) were selected so that both models could be applied to the same feature set. This approach ensured consistency when comparing models and allowed for clear interpretability of results.

For classification models, evaluation metrics included the confusion matrix as well as sensitivity and specificity. For regression models, performance was assessed using prediction accuracy, root mean squared error (RMSE), and the coefficient of determination (R^2).

```
Overall Statistics

               Accuracy : 0.4452
                 95% CI : (0.397, 0.4942)
    No Information Rate : 0.1667
    P-Value [Acc > NIR] : < 2.2e-16

Statistics by Class:

                     Class: Insufficient_Weight Class: Normal_Weight
Sensitivity                             0.46296              0.21053
Balanced Accuracy                       0.70279              0.56256
                     Class: Obesity_Type_I Class: Obesity_Type_II
Sensitivity                        0.47143                0.69492
Balanced Accuracy                  0.64857                0.78790
                     Class: Obesity_Type_III Class: Overweight_Level_I
Sensitivity                           0.9844                   0.15517
Balanced Accuracy                     0.9388                   0.54444
                     Class: Overweight_Level_II
Sensitivity                            0.068966
Balanced Accuracy                      0.513765
```

For the first model, I implemented a multinomial logistic regression using the multinom() function and evaluated its performance with a confusion matrix. The overall accuracy was 44.52%, with substantial variation in sensitivity and specificity across classes. For instance, Obesity_Type_III achieved a very high sensitivity of 0.9844, while Overweight_Level_II recorded only 0.069. The Insufficient_Weight, Normal_Weight, and Overweight_Level_I classes also exhibited low classification accuracy, indicating that the model struggled to classify complex class structures effectively. These limitations may stem from insufficient handling of distributional imbalances or the inability to capture nonlinear relationships between variables.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -209.16290   10.18705 -20.532  < 2e-16 ***
Age            0.79213    0.08635   9.173  < 2e-16 ***
Height       144.23163    5.98030  24.118  < 2e-16 ***
FCVC          10.49734    0.96542  10.873  < 2e-16 ***
NCP            0.44257    0.66874   0.662    0.508
CH2O           5.58272    0.85341   6.542 8.05e-11 ***
FAF           -6.77929    0.65076 -10.417  < 2e-16 ***
TUE            0.32885    0.87691   0.375    0.708
---
RMSE: 21.47276
R-squared: 0.3665857
```

For the second model, I constructed a linear regression to predict Weight as a continuous variable, using the same seven independent variables. The regression results indicated that Height, Age, FCVC (frequency of vegetable consumption), CH2O (daily water intake), and FAF (physical activity frequency) were significant predictors, with Height showing the largest effect (coefficient = 144.2). In contrast, NCP (number of meals per day) and TUE (technology use for leisure) were not statistically significant, suggesting that their influence on weight may be small or nonlinear. The model's coefficient of determination (R^2) was 0.37, meaning that the model explained approximately 37% of the variability in weight. This result highlights the limitation of predicting weight using only lifestyle variables.

When comparing overall performance, the linear regression model provided relatively stable results for numerical prediction tasks and offered strong interpretability in terms of variable influence. The logistic regression model, while valuable for generating probabilistic class interpretations, showed limited classification performance given the complexity of the class distribution. These findings underline both the strengths and constraints of the dataset. Future improvements may include applying nonlinear models, expanding the set of predictor variables, and addressing class imbalances to enhance model performance.

3. Decisions

The classification and regression models developed in this analysis can support health-related decision-making in different ways. The multinomial logistic regression model, which predicts obesity levels, can be useful for identifying risk groups based on individual eating habits and physical activity data. Although the overall accuracy was relatively low at 44.52%, the model demonstrated very high sensitivity for certain classes, such as Obesity_Type_III. This suggests potential as a preliminary screening tool for identifying high-risk individuals who may benefit from preventive programs or additional medical examinations. The practicality of the model is enhanced by the fact that most of its input variables are self-reported lifestyle data, making it adaptable for general use.

The linear regression model, which predicts weight, offers insights into how specific variables influence weight changes. For instance, according to the model coefficients, an increase of one unit in FCVC (frequency of vegetable consumption) is associated with an average weight increase of 10.5 kg, while greater FAF (physical activity frequency) shows a significant negative association with weight. These findings can directly inform practical intervention strategies such as health counseling, diet program planning, and nutrition design. If weight change is the goal, the model can numerically suggest which lifestyle habits should be adjusted.

However, both models have clear limitations that must be considered before direct application in decision-making. In the classification model, very low sensitivity and positive predictive values for some classes raise the risk of incorrect classifications leading to false alarms or misplaced confidence. In the regression model, the R^2 value of 0.37 indicates that a substantial proportion of weight variability remains unexplained, suggesting that many influencing factors are not captured.

In conclusion, while these models are valuable as reference indicators for personal health interventions, they are more appropriate as supportive tools used alongside expert judgment and additional indicators rather than as standalone decision-making systems. Future improvements should focus on increasing data diversity, addressing class imbalances, and applying more sophisticated algorithms to enhance model reliability.

Reference:

Data_Analytics2025Fall_Group3_Evaluating_Classification_Clustering_Models.pdf

Data_Analytics2025Spring_Group2_Generalization_Validation_Optimization.pdf

Data_Analytics2025Fall_Group2_Linear_Models_Random_Forest_0.pdf

Data_Analytics2025Spring_Group4_Model_Validation_Error_Estimation.pdf