Predicting Obesity Risk from Behavioral and Demographic Factors

A comparative analysis of tree-based models on behavioral and demographic data.

UNIVERSITÉ LAVAL

Course: MQT7015
Instructor: Prof. Cremona, Severino
Date: December 2025

Presented by Mohammad Ghafourian-Nasiri

Context & Problem Statement

1 Billion+

The number of people globally living with obesity, including 890 million adults and 160 million children (WHO).

$4.32 Trillion

Projected annual global economic impact by 2035, equivalent to nearly 3% of global GDP (World Obesity Federation).

Core Problem: Global prevalence has nearly tripled since 1975, with rising rates in low- and middle-income countries. This drives non-communicable diseases like diabetes and cardiovascular disease.

Objectives & Methodological Framework

Data Preparation

- Real vs. Synthetic Data Separation
- Variable Transformations
- Target Variable Creation
- Train/Test Split
- Class Distribution Verification
- Final Predictor Set

Exploratory Data Analysis

- Target Variable Distribution
- Predictor Distributions
- Relationships with Target

Methodology Implementation

- Deviance-Based Tree
- Gini-Based Tree
- Bagging
- Random Forest
- Variable Importance
- Test Set Performance

Model Comparison

- Best Model Selection
- Comparison Visualization

Conclusions

- Key Findings
- Summary

Dataset Description

Source: Palechor & de la Hoz Manotas (2019)
Population: Individuals aged 14-61 from Colombia, Peru, and Mexico.
Total Observations: 2,111

The Clues (Predictor Variables)

Demographic

Gender, Age, Height

Eating Habits

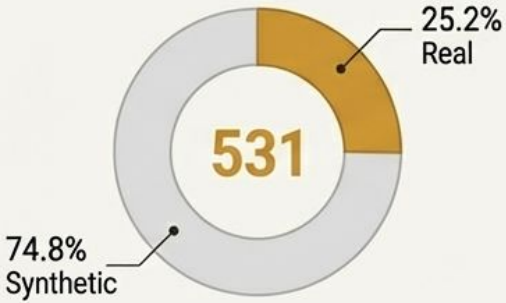
High-Calorie Food (FAVC), Vegetable Consumption (FCVC), Main Meals (NCP), Snacking (CAEC), Water Intake (CH2O), Alcohol (CALC)

Lifestyle

Smoking (SMOKE), Calorie Monitoring (SCC), Physical Activity (FAF), Tech Use (TUE), Transportation (MTRANS), Family History

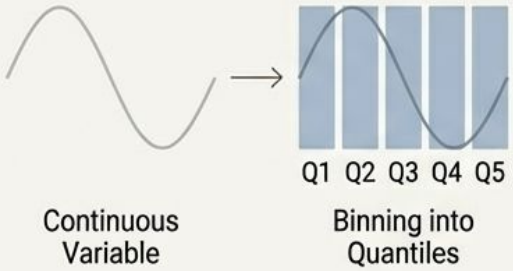
Dataset Preparation

1. Isolating Real Data



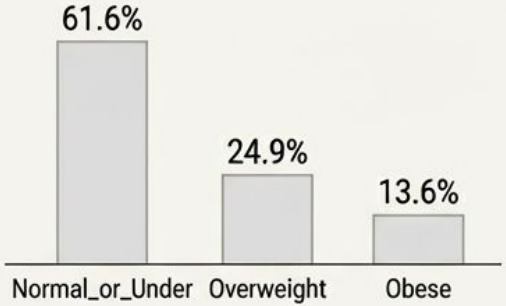
We filtered out 1,580 synthetically generated records to work only with the 531 (25.2%) real survey responses, ensuring our findings reflect genuine human patterns.

2. Refining Predictors



Continuous variables (Age, Height) were binned into 5 quantile groups to capture non-linear effects. Rare categories were merged (e.g., `MTRANS` grouped into 'Active', 'Private Motor', 'Public Transport').

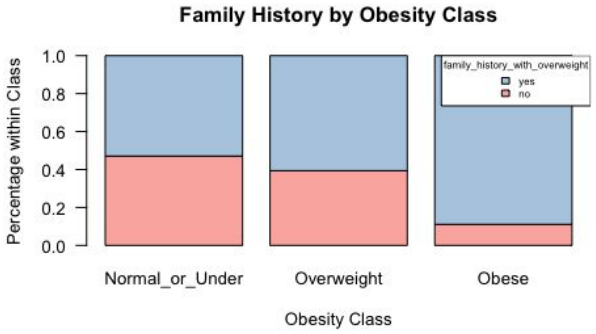
3. Simplifying the Target

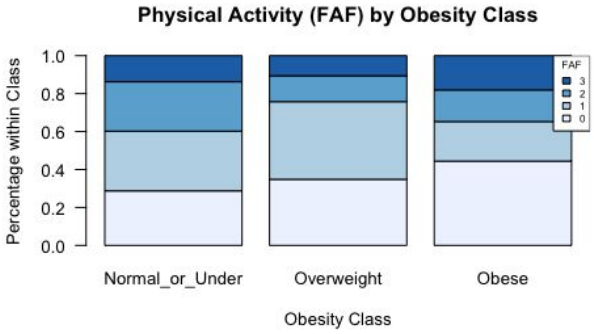


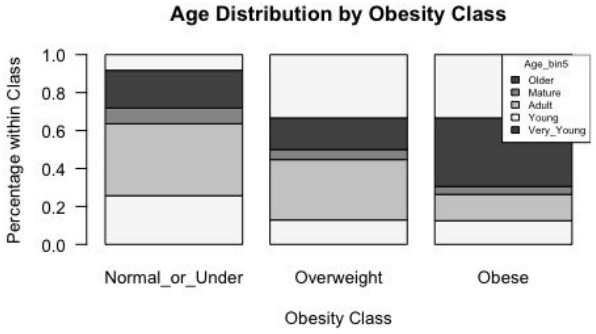
The original 7 obesity levels were consolidated into 3 broader classes for model stability: ``Normal_or_Under`` (61.6%), ``Overweight`` (24.9%), and ``Obese`` (13.6%).

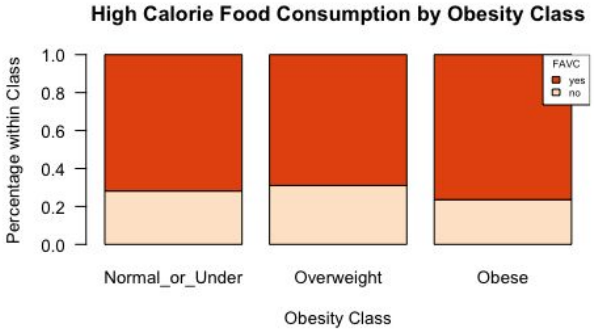
Exploratory Data Analysis (Initial Findings)

high-calorie food consumption is ubiquitous, with nearly 90% of the entire population (regardless of weight class) reporting "Yes." While the Obese group does show a slightly higher saturation of "Yes" responses compared to the Normal group, the lack of a sharp contrast suggests that what people eat (high calorie) is less predictive than how much they eat (NCP) or their genetic background (Family History)





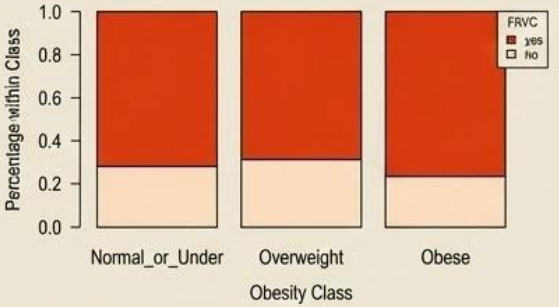




Exploratory Data Analysis (Initial Findings)

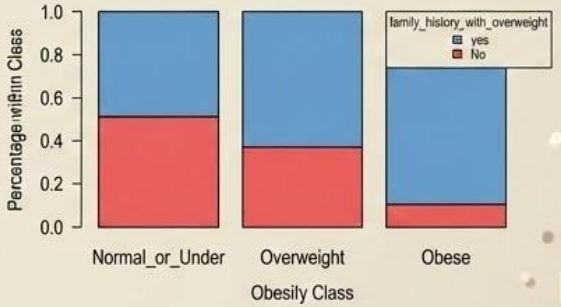
Key Insights on Behavioral and Genetic Factors

High-Calorie Food Consumption (FAVC)



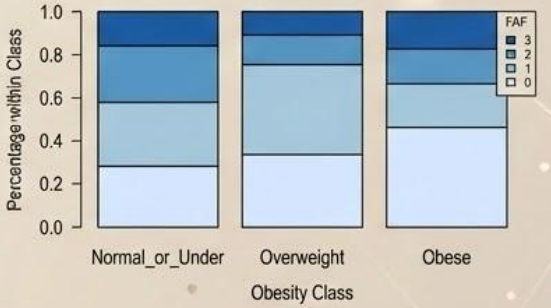
High-calorie food consumption is ubiquitous, with nearly 90% of the population reporting 'Yes'. While the Obese group shows a slightly higher saturation, the lack of a sharp contrast suggests it is less predictive than other factors.

Family History (Genetics)



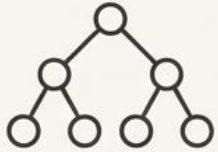
Genetic background (Family History) shows a more distinct pattern, with a significantly higher proportion of the Obese group reporting a family history of overweight, indicating a stronger predictive power.

Physical Activity (FAF)



Physical activity levels are generally lower in the Overweight and Obese groups, suggesting a negative correlation with obesity risk.

Machine Learning Methods Used



Decision Tree (Deviance)

The Baseline.
A single set of rules to classify individuals.

Decision Tree (Gini)

An Alternative Baseline.
Uses a different mathematical logic (Gini impurity) to create splits.

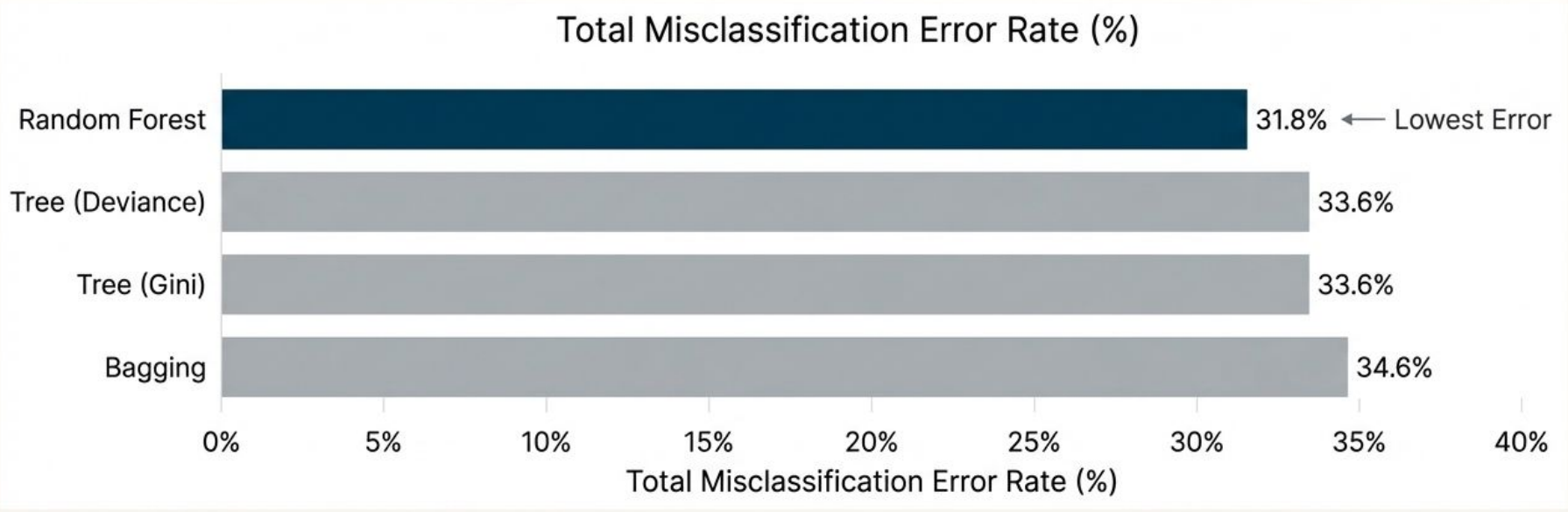
Bagging

The Power of Volume.
Averages predictions from hundreds of trees built on random data samples to improve stability.

Random Forest

The Power of Diversity.
Like Bagging, but forces each tree to consider a different, random subset of predictors, reducing error.

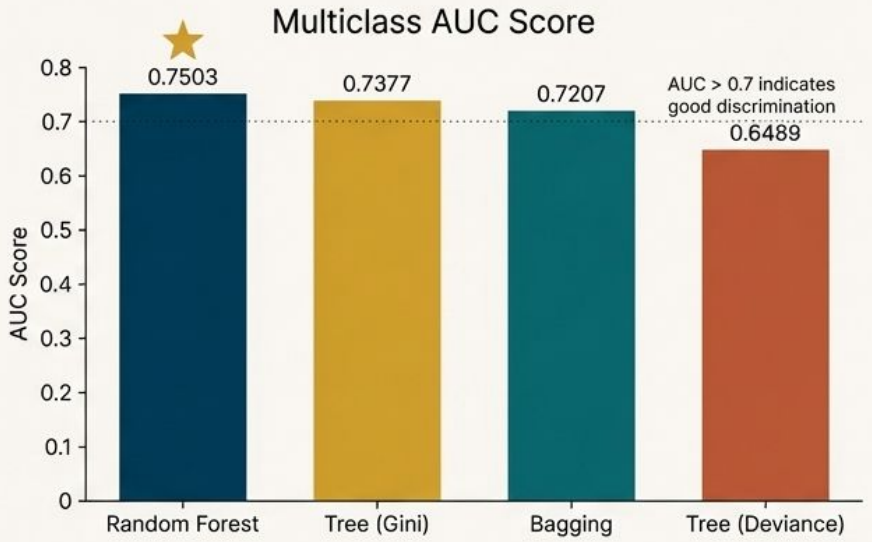
Model Performance Comparison (Error Rate)



Random Forest's strategy of decorrelating trees by using a random subset of features at each split results in the lowest overall prediction error.

AUC Comparison Across Models

The Area Under the Curve (AUC) provides a comprehensive measure of a model's ability to correctly distinguish between the obesity classes. A higher AUC indicates better overall discrimination.



With the highest AUC score, Random Forest is definitively the most robust and reliable model for this classification task.

Class-Specific Error Analysis

The main challenge across all models is classifying overweight and obese cases, as these groups share many characteristics. Demographic factors like age and height are consistently the most important predictors, followed by behavioral factors such as water intake and eating frequency.

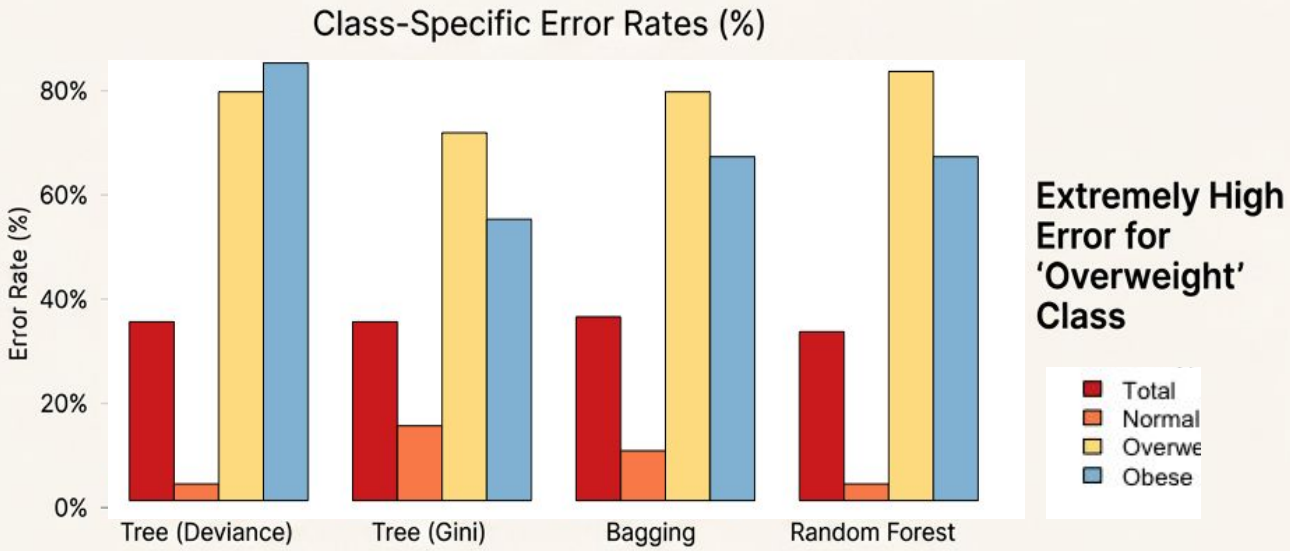
Data Highlights (Overweight Error Rate %)

Tree (Gini)
69.2% (Surprise Best Performer)

Tree (Deviance)
76.9%

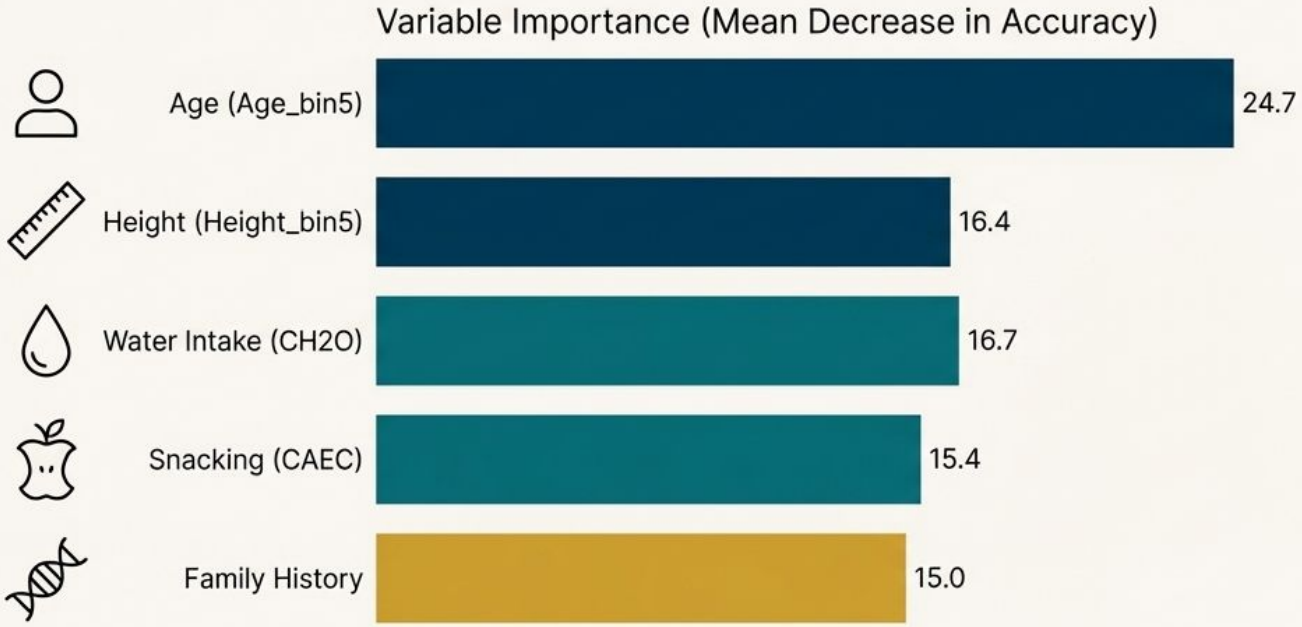
Bagging
76.9%

Random Forest
80.8%



Extremely High Error for 'Overweight' Class

Most Influential Predictors



Observation

Demographic factors (`Age`, `Height`) are the undeniable top drivers, followed by a tight cluster of key behavioral variables.

Conclusion

Best Model: Random Forest

↓ **31.8%** Total Error

↗ AUC **0.75**

Outperformed single deviance & Gini trees and bagging.

Achieves superior performance by reducing overfitting.

Top Predictors & Core Challenge

Strongest Predictors

Water Intake

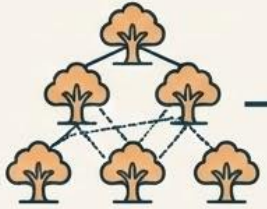
Eating Frequency

Followed by Behavioral Factors

Main Challenge: Classifying Overweight

Difficult due to overlap with normal/obese groups and smaller sample size.

Ensemble Power & Efficiency



→ **Reduced Overfitting**

Efficiency

Random forest balances accuracy and speed, using only 3 random variables per split for faster convergence while maintaining interpretability.

Ghafourian Nasiri, Mohammad

Key Limitations

Data Source

Analysis was limited to 531 real observations, as the original dataset was ~75% synthetic.

Class Imbalance

The 'Normal_or_Under' class dominates (61.6%), making it harder for models to learn minority class patterns.

Feature Scope

The model relies on behavioral data only. Clinical data (e.g., metabolic rates, genetic markers) could significantly improve performance.

Future Work

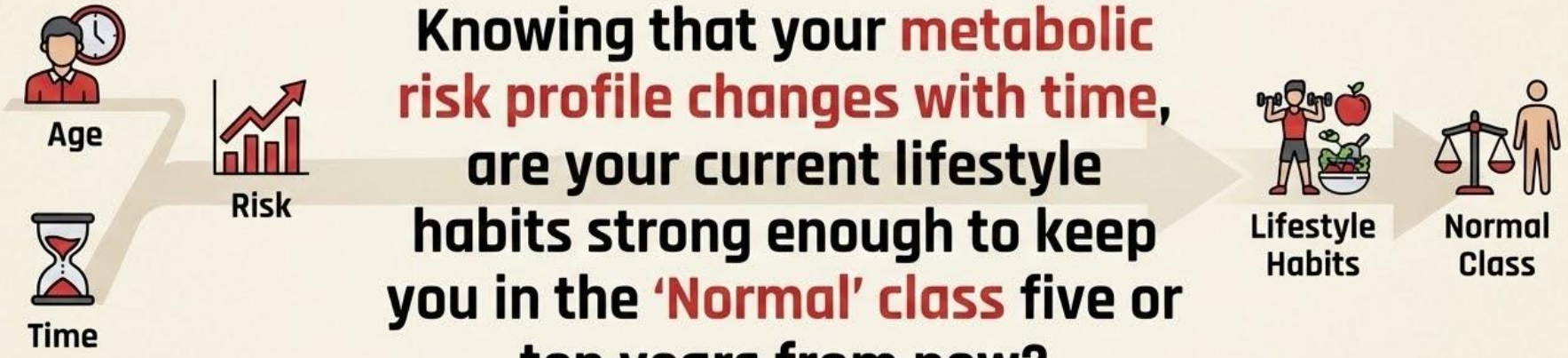
Collect more real, balanced data to address the core limitations.

Explore advanced models like neural networks or gradient boosting.

Conduct regression analysis to predict continuous BMI instead of discrete classes.

Perform separate analyses across demographic groups (e.g., gender, age) to uncover subgroup-specific patterns.

Since Age was the number one predictor in every single model we built, we know risk shifts as we get older.



THANK YOU

Mohammad Ghafourian-Nasiri

MQT7015 - Predicting Obesity Risk from
Behavioral and Demographic Factors