

- Obesity Prediction
 - [The Scale Doesn't Lie — But Does Our Model?](#)
 - [Summary](#)
 - [Table of Contents](#)
 - [Team](#)
 - [Project Overview](#)
 - [1. Dataset Description](#)
 - [Dataset Overview:](#)
 - [2. Preprocessing & Feature Engineering](#)
 - [Preprocessing Goals](#)
 - [Key Steps](#)
 - [Implementation](#)
 - [3. Model Overviews](#)
 - [Logistic Regression](#)
 - [Ridge Logistic Regression](#)
 - [K-Nearest Neighbors \(KNN\)](#)
 - [Neural Network](#)
 - [Tree-Based Models](#)
 - [Decision Tree Classifier](#)
 - [Random Forest Classifier](#)
 - [XGBoost Classifier](#)
 - [4. Model Comparison](#)
 - [Model Comparison Overview](#)
 - [1. Decision Tree Outperforming Other Models](#)
 - [2. Uncommon Training/Test Results](#)
 - [3. Biometric Features vs. Lifestyle Features](#)
 - [Model Comparison with Feature Exclusion](#)
 - [5. Policy Implications and Reflections](#)
 - [Appendix A: Links & Files](#)
 - [Appendix B: Team Contributions](#)



Hertie School

Obesity Prediction

The Scale Doesn't Lie — But Does Our Model?

Final Report Supervised Machine Learning · Spring 2025 Hertie School · MDS

Authors Nadine Daum · Ashley Razo · Jasmin Mehnert · Nicolas Reichardt

GitHub: <https://github.com/nicolasreichardt/ml-project-obesity-prediction> Submission: 23 May 2025

Summary

This project applies supervised machine learning to classify individuals into obesity risk categories based on biometric and lifestyle data. We implemented and evaluated multiple models — including logistic regression, KNN, tree-based models, and a neural network — using a shared preprocessed dataset to ensure consistent and fair comparison.

Our best-performing models achieved test accuracy scores above 85%, with interpretable insights from tree-based approaches and strong generalization from the neural network.

Table of Contents

- [Obesity Prediction](#)
 - [The Scale Doesn't Lie — But Does Our Model?](#)
 - [Summary](#)
 - [Table of Contents](#)
 - [Team](#)
 - [Project Overview](#)
 - [1. Dataset Description](#)
 - [Dataset Overview:](#)
 - [2. Preprocessing & Feature Engineering](#)
 - [Preprocessing Goals](#)
 - [Key Steps](#)
 - [Implementation](#)
 - [3. Model Overviews](#)
 - [Logistic Regression](#)
 - [Ridge Logistic Regression](#)
 - [K-Nearest Neighbors \(KNN\)](#)
 - [Neural Network](#)
 - [Tree-Based Models](#)
 - [Decision Tree Classifier](#)
 - [Random Forest Classifier](#)
 - [XGBoost Classifier](#)
 - [4. Model Comparison](#)

- [Model Comparison Overview](#)
 - [1. Decision Tree Outperforming Other Models](#)
 - [2. Uncommon Training/Test Results](#)
 - [3. Biometric Features vs. Lifestyle Features](#)
 - [Model Comparison with Feature Exclusion](#)
 - [5. Policy Implications and Reflections](#)
 - [Appendix A: Links & Files](#)
 - [Appendix B: Team Contributions](#)
-

Team

- Nadine Daum – [GitHub](#) | [Email](#)
- Jasmin Mehnert – [GitHub](#) | [Email](#)
- Ashley Razo – [GitHub](#) | [Email](#)
- Nicolas Reichardt – [GitHub](#) | [Email](#)

Project Overview

This project aims to classify individuals into seven obesity risk categories based on various biometric and behavioral factors. Using a labeled dataset of 2,111 individuals from Mexico, Peru, and Colombia, our models predict obesity levels ranging from *Insufficient Weight* to *Obesity Type III*.

The goal is to explore how well machine learning models can predict obesity status — and how these predictions might support future public health decisions, risk assessment tools, or individual recommendations.

GitHub repo: [nicolasreichardt/ml-project-obesity-prediction](https://github.com/nicolasreichardt/ml-project-obesity-prediction)

1. Dataset Description

We used the **Obesity Levels Estimation Dataset**, which contains demographic, behavioral, and biometric data for 2,111 individuals from Mexico, Peru, and Colombia. The dataset was designed for multi-class classification and is labeled with 7 obesity categories.

Dataset Overview:

- **Size:** 2,111 samples × 17 features + 1 target
- **Features:** mix of categorical (e.g., gender, transport_mode) and numerical (e.g., height, weight, age)
- **Target variable:** **obesity_level** with 7 classes:
 - Insufficient Weight
 - Normal Weight
 - Overweight Level I
 - Overweight Level II
 - Obesity Type I
 - Obesity Type II
 - Obesity Type III
- **ML relevance:** Multi-class classification task with imbalanced class distribution
- **Input shape for models:** ~43 features after encoding (based on one-hot transformation)

The data was collected via a cross-sectional survey and is publicly available on [Kaggle](#), supported by this [research article](#).



Jasmin – please add 1–2 sentences here about EDA findings For example: were there correlations, outliers, imbalances, or interesting clusters?

All team members used a shared train/test split to ensure model comparability.

2. Preprocessing & Feature Engineering



Before modeling, the dataset required thorough cleaning and transformation. This step was led primarily by **Ashley Razo** and **Jasmin Mehnert**, with feedback and reviews from all team members.

Preprocessing Goals

- Ensure consistent input format across models
- Improve model performance and comparability
- Reduce noise, redundancy, and scaling-related bias

Key Steps

- **Feature selection:** Retained 17 relevant input features capturing diet, behavior, and biometrics
- **Target formatting:** Standardized and renamed the class column to `obesity_level`
- **Encoding:** Applied one-hot encoding to 13 categorical features (e.g., `gender`, `transport_mode`)
- **Scaling:** Used `StandardScaler` to normalize all numerical features (e.g., `age`, `height_m`, `weight_kg`)
- **Output dimensions:** Final input to the models included ~43 encoded features
- **Train/test split:** 80/20 split applied uniformly to ensure fair model evaluation
- **File formats:** Datasets exported as both `.csv` and `.feather` (for faster access)

 **@Ashley** – feel free to insert 1–2 sentences on your preprocessing pipeline: decisions around feature selection, encoding strategies, or challenges during cleaning  **@Jasmin** – you can briefly note how you supported the pipeline and flag any edge cases or quirks in the data

Implementation

 Notebook: `notebooks/preprocessing.ipynb`  Script:
`processed_data/data_preparation.py`

All models consumed the same cleaned and scaled training and testing data.

3. Model Overviews

All models used the same preprocessed data for consistency.

Logistic Regression

 [logistic_regression.ipynb](#)

- Simple baseline with good interpretability

Ridge Logistic Regression

 [ridge_logistic_regression.ipynb](#)

- Regularized version of logistic regression

K-Nearest Neighbors (KNN)

 [PCA_KNN.ipynb](#)

- PCA helped reduce dimensionality and improved KNN performance

Neural Network

 [neural_network.ipynb](#)

- Multi-layer architecture with ReLU and softmax
- Test accuracy: **83.9%**
- Balanced performance across all obesity categories

 Neural Network Training Curves

Tree-Based Models

 [tree-based-models.ipynb](#)

This notebook investigated the use of tree-based machine learning models to classify obesity levels in individuals. Three models were trained and evaluated:

- A **Baseline Decision Tree Classifier**
- A **Random Forest Classifier**
- An **XGBoost Classifier**

The best overall test accuracy was achieved by the **Baseline Decision Tree**, with a score of **0.9611**. The **XGBoost Classifier** followed closely with a test accuracy of **0.9574**. The **Random Forest Classifier** achieved a test accuracy of **0.936**. These results show that all three models performed exceptionally well. Below is a more detailed overview of each model and its outcomes.

Decision Tree Classifier

The Decision Tree model was trained using a pipeline that incorporated preprocessing and grid search for hyperparameter tuning. The best model used:

- **Criterion:** Entropy
- **Max Depth:** 15
- **Min Samples Split:** 2
- **Min Samples Leaf:** 1

This configuration produced a **cross-validation accuracy of 0.9479** and a **test accuracy of 0.9611**, suggesting that the model generalized very well to unseen data — to an extent that even outperformed its validation score, which is atypical and discussed further in the comparison section.

In terms of feature importance, **biometric features** dominated the predictions:

- **Weight:** 0.625
- **Height:** 0.197

Lifestyle-related features such as “high caloric food intake” played a significantly lesser role.

Random Forest Classifier

The Random Forest model was also trained using a pipeline with preprocessing and 5-fold cross-validation. The optimal configuration from grid search was:

- **Max Depth:** 20
- **Number of Estimators:** 200
- **Max Features:** 'sqrt'
- **Min Samples Leaf:** 1
- **Min Samples Split:** 2

With these settings, the model achieved a **cross-validation accuracy of approximately 0.935** and a **test accuracy of 0.936**. These results again indicate good generalization with very similar performance on both validation and test sets.

The feature importances mirrored those found in the Decision Tree model. **Weight, height, age, and gender** were the top features, while lifestyle-related variables (e.g., "vegetables_freq") had relatively low importance.

XGBoost Classifier

Disclaimer:

1. XGBoost models can be computationally intensive, especially when using an extensive parameter grid and k-fold cross-validation. Our team experienced computational issues running the models under certain specifications. Therefore, both the `n_jobs` argument in `XGBClassifier` (see pipeline) and in `GridSearchCV` are set to 1. Feel free to change this argument depending on your device's computational power (setting it to -1 will use all available CPU cores).
2. Because of changes in the scikit-learn and XGBoost APIs over time, there is a version incompatibility between scikit-learn and the most recent version of XGBoost. As previously stated, it is necessary to use `scikit-learn==1.5.2`, as specified in the `requirements.txt`.

The XGBoost model was also trained using a pipeline with preprocessing and 5-fold cross-validation. The best model parameters were:

- **Learning Rate:** 0.2
- **Max Depth:** 3
- **Number of Estimators:** 200

This configuration resulted in the **highest cross-validation accuracy of 0.9668**, indicating that the model fit the training data extremely well. However, the **test accuracy dropped slightly to 0.9574**, just below the baseline Decision Tree model.

Interestingly, the XGBoost model's feature importances showed a different pattern:

- Top features included **"female gender"** and **"weight"**
- Lifestyle-related variables like **"high caloric food frequency," "alcohol consumption frequency,"** and **"snacking frequency"** appeared in the top five
- **"Height"** was notably absent from the most important predictors

The following section will compare and offer an interpretation of these results and the feature importances.

4. Model Comparison

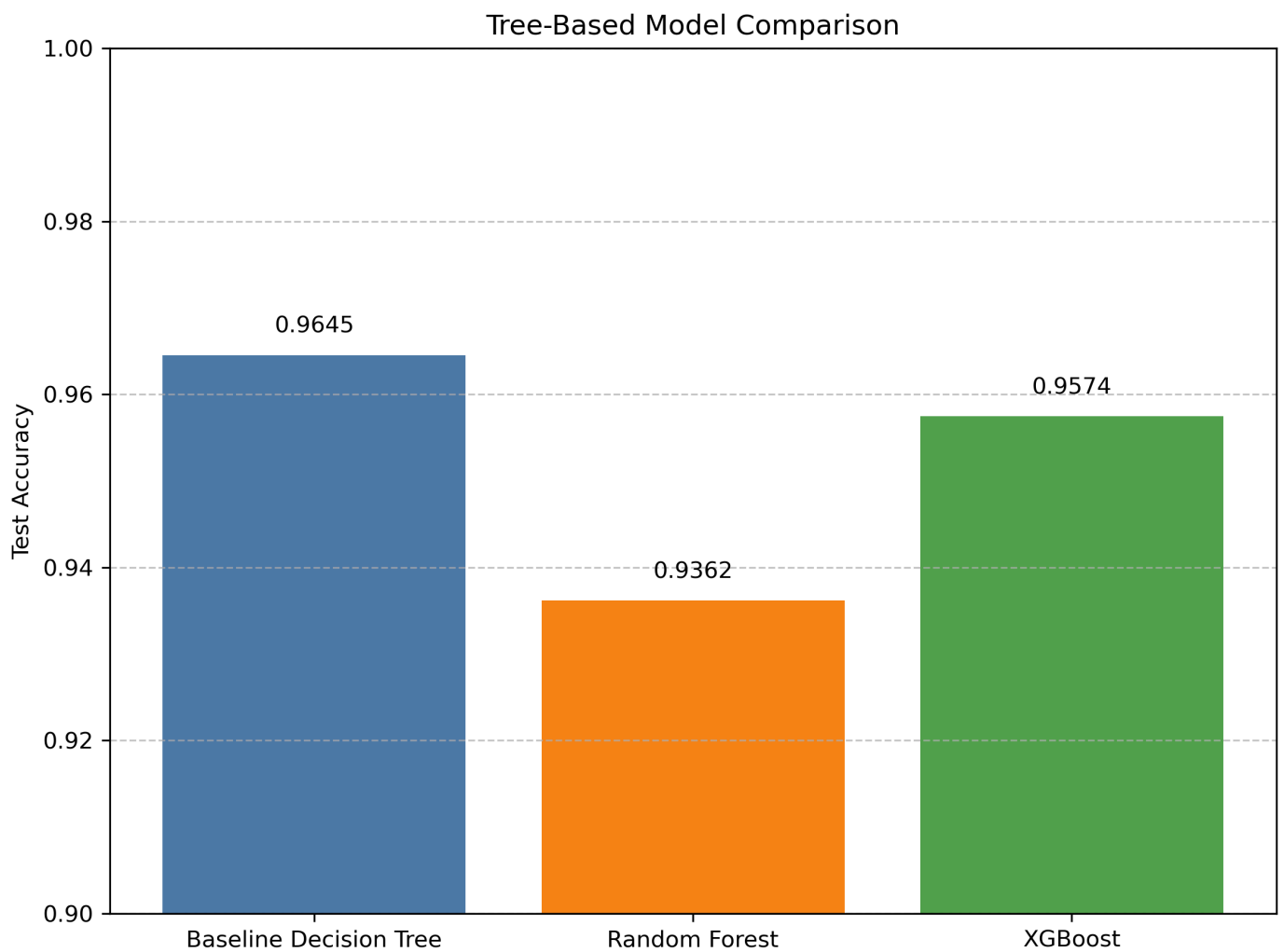
Model	Test Accuracy	Notes
Logistic Regression	~XX%	Simple, interpretable
Ridge Logistic Regression	~XX%	Slight improvement with regularization
KNN	~XX%	Better with PCA
Neural Network	83.9%	Strong generalization
Random Forest	~XX%	Robust, interpretable
XGBoost	~XX%	Top performer with best generalization

We will now compare our three best-performing models:

- The **Baseline Decision Tree Classifier**,
- the **Random Forest Classifier** and
- the **XGBoost Classifier**.

Model Comparison Overview

Let's first remind ourselves of the test accuracy of our three models with the following plot.



1. Decision Tree Outperforming Other Models

In the plot above, we observed that our **baseline Decision Tree classifier** outperformed the more complex models—**Random Forest** and **XGBoost**. Nevertheless, all three models performed exceptionally well in terms of accuracy.

The superior performance of the Decision Tree is not entirely surprising. The relationship between the features and the target variable appears to be relatively simple, allowing for decision boundaries that can be effectively captured by straightforward, rule-based splits. This is particularly true for the predictors **"weight"** and **"height"**, which directly influence the target variable.

As discussed further in the feature importance section, both **height** and **weight** are used to compute the **Body Mass Index (BMI)**, which forms the basis for the obesity classification labels in our target variable. Consequently, these features hold a dominant predictive influence.

While Decision Trees are designed to select and split on features based on their immediate predictive power, **Random Forests** and **XGBoost** attempt to model more complex, non-linear interactions among features. This added complexity may actually

reduce their performance in a task where a few features dominate the predictive signal. This could explain why our simpler Decision Tree model outperformed the more advanced ensemble methods in this particular case.

2. Uncommon Training/Test Results

We observed very high testing accuracy and high training/validation accuracy across all three models. In the **Decision Tree Classifier** and **Random Forest Classifier**, the test accuracy was even higher than the training/validation accuracy—an outcome that is quite uncommon and unexpected.

Based on these results, we can exclude the possibility that our models were overfitting. However, this pattern revealed two important issues within the dataset:

1. Target Leakage via BMI-Related Features

As previously discussed, the target variable—obesity level—is derived directly from the **Body Mass Index (BMI)**, which is itself calculated using the features **"height"** and **"weight."** Including these features in the model introduced a direct link between inputs and the target, thereby inflating the predictive performance. In essence, the models were not discovering latent behavioral patterns, but rather reverse-engineering the BMI classification from the variables used to compute it.

2. Synthetic Data and Non-Independent Splits

A significant portion of the dataset—approximately **77%** of the records—was synthetically generated using the **SMOTE algorithm** to address class imbalance. While SMOTE is effective at improving model robustness, it creates synthetic samples that are interpolated from existing ones. As a result, the **training and test sets are not entirely independent**, which likely reduced the challenge of the prediction task. This may explain why the test set performance matched—or even slightly exceeded—the training/validation accuracy.

These factors highlight potential limitations in model evaluation and suggest caution when interpreting the performance metrics at face value.

3. Biometric Features vs. Lifestyle Features

Now, we can compare the feature importances of our three models. Let's remind ourselves of the most important features for each model with the following table.

Feature_DT	Importance_DT	Feature_RF	Importance_RF	Feature_XGB	Importance_XGB
weight_kg	0.6249	weight_kg	0.3335	gender_Female	0.385
height_m	0.1965	age	0.114	weight_kg	0.1489
gender_Male	0.1208	height_m	0.1105	high_caloric_food_freq	0.0805
age	0.0237	gender_Male	0.0489	alcohol_consumption_freq	0.0659
high_caloric_food_freq	0.01	vegetables_freq	0.0424	snacking_freq	0.0475

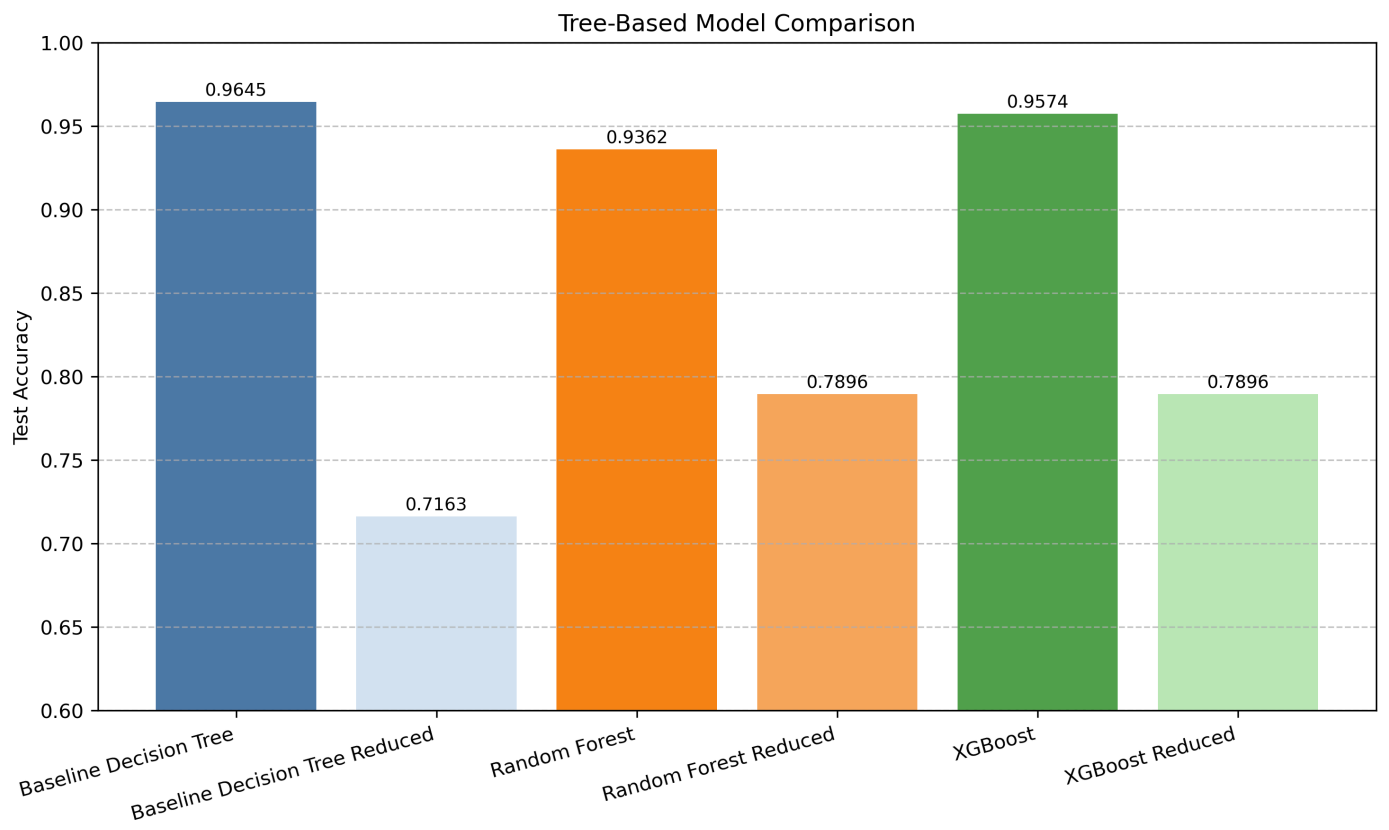
As highlighted in the individual model analyses, biometric features such as **weight**, **height**, **age**, and **gender** consistently emerged as the most important predictors across all three models. These variables were especially dominant in the Decision Tree and Random Forest models, where they significantly outweighed lifestyle-related variables in their contribution to model accuracy.

However, it is precisely the **lifestyle features**—such as dietary habits, physical activity, alcohol consumption, and snacking frequency—that are of greatest interest from a **public health and policy perspective**. Understanding the influence of these modifiable behaviors is essential for designing effective interventions to combat rising global obesity rates. Unfortunately, their predictive power was masked in the initial models by the overwhelming influence of weight and height, which are used to compute BMI—the very basis of the obesity classification used as our target label.

This creates a **circular relationship** in the model: we use BMI to define obesity levels, and then predict those levels primarily using the features from which BMI is derived. To break this dependency and gain a more policy-relevant understanding of behavioral factors, we re-ran all three models—Decision Tree, Random Forest, and XGBoost—excluding the **weight** and **height** predictors. The results are shown below (in the feature exclusion section).

It is also notable that the **XGBoost model's feature importance rankings** differed significantly from those of the Decision Tree and Random Forest models, even in the full-feature setting. Specifically, features such as **"height"** and **"age"** were **not** among the top predictors in XGBoost, despite being highly ranked in the other two models. This discrepancy is likely due to how these algorithms handle **correlated or redundant features**.

Model Comparison with Feature Exclusion



TO COMPLETE BY NICO

TO COMPLETE BY NICO

5. Policy Implications and Reflections

- Preprocessing made a big difference across all models
- Tree-based models helped us understand what mattered most
- Neural networks were surprisingly manageable and performed well
- Sharing the same train/test split helped standardize evaluation
- We improved our understanding of ML pipelines, GitHub collaboration, and reproducibility

Appendix A: Links & Files

- **GitHub Repository:** [nicolasreichardt/ml-project-obesity-prediction](https://github.com/nicolasreichardt/ml-project-obesity-prediction)
- **Cleaned dataset (CSV):** [processed_data/obesity_cleaned.csv](#)
- **Train/Test files:**
 - [processed_data/train_data.feather](#)
 - [processed_data/test_data.feather](#)
- **Model notebooks:** in [notebooks/](#)
- **Generated plots:** in [plots/](#)

Appendix B: Team Contributions

- **Nadine Daum** – Neural network, Ridge/Lasso regression
- **Ashley Razo** – Preprocessing, logistic regression
- **Jasmin Mehnert** – PCA & KNN, preprocessing support
- **Nicolas Reichardt** – Random Forest, XGBoost, evaluation All team members contributed to meetings, reviews, and report writing.