# Stat230a Final Project: Predicting Obesity from Behavioral Data

Brandon Xu - brandonaxu@berkeley.edu
Derry Xu - derryxu@berkeley.edu

May 13, 2025

## 1 Introduction

Obesity is one of the most pressing global health issues of the current day, contributing to a variety of chronic conditions such as type 2 diabetes, cardiovascular disease, and certain cancers. As lifestyle and dietary habits evolve, so does the prevalence of obesity continues to rise. This has prompted increasing interest in the predictive modeling of obesity based on behavioral and physiological factors.

The goal of this project is to explore how individual characteristics such as eating habits, physical activity levels, and family health history can be used to predict obesity. By leveraging statistical and machine learning methods—particularly regression and classification techniques—we aim to build models that can accurately estimate an individual's obesity status.

## 2 Data

The data set chosen came from UCI Machine Learning Repository (2022). It contains synthetic data generated through a combination of real survey data and modeling techniques in order to protect individual privacy while maintaining statistical reliability. There are 2111 rows, with each row of the data set representing a patient and relevant features such as demographic attributes (such as age, gender, weight, and height), lifestyle indicators (frequency of high-calorie food consumption, and alcohol intake), and other behavioral indicators (such as smoking habits and family history of being overweight). The distribution of classes are shown in Graph 1. The dataset classifies obesity severity with the following categories: `Obesity_Type_I`, `Obesity_Type_II`, `Obesity_Type_III`, `Overweight_Level_I`, `Overweight_Level_II`, `Normal_Weight`, and `Insufficient_Weight`. To help reduce data sparsity, the positive class was defined as individuals classified under any obesity category (Type I, II, or III), and the negative class included all other weight categories.

For the data pre-processing phase, the dataset was split into training (80%) and test (20%) subsets to avoid data leakage. We also wanted to not rely on BMI for prediction, since BMI is a function of weight and height while also being the criteria for determining obesity levels. Therefore, due to its direct correlation with BMI, including both weight and height would lead to a trivial classification. Following Sun et al. (2024), we proceeded by assuming that weight is known and height is not.

During feature engineering, numerical features underwent logarithmic and exponential transformations to capture non-linear relationships, with new variable names being preceded with `log1p_` and `exp_`, respectively. Additionally, polynomial features up to degree two were generated to capture potential interactions among features. Numerical features were standardized and categorical variables were encoded using one-hot encoding to ensure that all features were on comparable scales. This left us with 64 potential features. We then apply LASSO using three different regularization strengths $(0.1, 1, 10)$ to generate three different sets of features. This is primarily done to reduce the computational power needed to fit our models, rather than actually selecting the best features. Since we apply GridSearchCV later across L1, L2, and Elasticnet penalties, the GridSearch is able to automatically feature select using L1/ElasticNet based on cross-validated accuracy.

## 3 Final Model

A logistic regression model were used to classify individuals into binary categories. Initial models were evaluated using different penalties at varying levels of strength (C values of 0.1, 1, and 10). The best-performing binary classification model was obtained using elastic-net regularization with $C = 1$ and an $l_1$ ratio of 0.75, trained with the `saga` solver. This configuration achieved a cross-validation accuracy of approximately 96%, with similarly strong performance on the held-out test set (accuracy: 96.2%, precision: 95.98%, recall: 95.98%, F1 score: 95.98%). To further evaluate the model's discriminative ability, we plotted the Receiver Operating Characteristic (ROC) curve, shown in Figure 3. The ROC curve illustrates the trade-off between the true positive rate and false positive rate across various decision thresholds. Our model achieved an area under the curve (AUC) of 0.99, indicating near-perfect separability between obese and non-obese individuals. This high AUC confirms the model's robustness and its ability to correctly distinguish positive cases, even across a range of classification thresholds.

To interpret the model, we examined the learned coefficients. Table 1 presents the most influential features, ranked by the absolute magnitude of their coefficients, which reflect each variable's relative importance. Other coefficients not listed in the table were zeroed out using ElasticNet and didn't contribute towards the model. Positive coefficients indicate an increase in the log-odds of being classified as obese, while negative coefficients indicate a decrease. Among the top contributing variables were weight-related interaction terms, such as `FCVC Weight` and `NCP Weight`, along with dietary and behavioral variables like `CH2O TUE`, `log1p_TUE`, and `MTRANS_Walking`. Interestingly, such the negative effect of some coefficients, such as being male or with increased screen time, challenge common assumptions in society.

## 4 Discussion

One key strength of the project lies in its use of elastic-net regularization, which combines the advantages of LASSO (feature selection) and ridge regression (handling multicollinearity). This allowed the model to achieve high predictive performance (96% accuracy).

Despite these strengths, the project also has important limitations. A significant one is the cross-sectional nature of the dataset: all features and the obesity label are measured simultaneously. As a result, the model learns associations present at the same point in time, rather than predicting future obesity based on earlier behaviors. For practical applications such as early intervention or preventive screening, a dataset tracking individuals over time would be more appropriate. Without temporal sequencing, it is difficult to establish causal relationships or assess whether observed predictors truly precede the onset of obesity.

Another limitation is the reliance on self-reported behaviors and habits, which may be subject to recall bias or social desirability bias. Variables such as dietary intake, physical activity, and screen time are often imprecisely measured, potentially reducing model accuracy or introducing noise. Additionally, while the model generalizes well within the dataset, the data measures individuals from Mexico, Peru and Colombia. Its external validity (i.e., how well it performs on different populations) has not been assessed, limiting the scope of its applicability without further validation.

Overall, while the model offers promising insights into the behavioral and demographic correlates of obesity, future work should incorporate longitudinal data and external validation to improve its utility for real-world prediction and intervention efforts.

## 5 Conclusion

This project set out to explore the extent to which obesity can be predicted from lifestyle and demographic variables using logistic regression models. By adding polynomial interactions, non-linear transformations, and one-hot encoded categorical variables—and leveraging elastic-net regularization, the final model achieved strong predictive performance with a test accuracy of over 96%.

Beyond its technical contributions, this analysis shows how simple, accessible information—like someone's eating habits, daily routines, and basic demographics—can be used to build fairly accurate models for obesity prediction. While the model isn't perfect or prescriptive, it gives us a sense of which behaviors and patterns are most strongly associated with obesity in the data. This can be useful not only for researchers and developers, but also for public health workers or even individuals interested in tracking risk factors. This project offers a starting point for thinking about how data-driven tools might help flag signs of obesity before more serious health outcomes develop.

## 6 Additional Work

In addition to the binary classification model, we tried to fit a multinomial logistic regression model that predicts the specific obesity category. The target classes left the labels in the `NObeyesdad` variable as is, meaning `Insufficient_Weight`, `Normal_Weight`, `Overweight_Level_I`, `Overweight_Level_II`, `Obesity_Type_I`, `Obesity_Type_II`, and `Obesity_Type_III` were the classes. The distribution of the classes are shown in Graph 2. This formulation poses a more nuanced classification task, requiring the model to distinguish between gradations of weight status rather than a simple obese/non-obese distinction.

Similar to the binary model, the same feature engineering pipeline of adding polynomial features, logarithmic and exponential transformations, and one-hot encoding was applied to the inputs. LASSO-based feature selection was used to reduce the number of predictors, but, like mentioned earlier, a grid search over regularization strengths and penalty types was conducted to optimize performance and more or less nullified the feature selection that LASSO conducted. The final model used an L1-penalized logistic regression and $C = 10$, which yielded the best cross-validation accuracy of 85.6%, with strong performance in identifying high-risk categories such as `Obesity_Type_II`, `Obesity_Type_III`, and `Insufficient_Weight`, all of which had F1 scores above 0.90. The model struggled somewhat with intermediate categories like `Normal_Weight` and `Overweight_Level_I`, likely due to subtler feature distinctions and potential overlaps in behavioral patterns between adjacent classes. Nonetheless, the model's ability to robustly detect the extremes of the weight spectrum is valuable in practical settings, where identifying individuals at greatest health risk is often the priority.

The coefficient analysis of the multinomial logistic regression model in Table 2 offers additional insight into the predictors that influence specific obesity categories. The most influential variable was `log1p_Weight`, with a large negative coefficient ($\beta = -15.24$), followed by the raw `Weight` variable itself ($\beta = -5.45$). These large negative values likely reflect the model's contrast between high and low BMI categories (e.g., underweight versus obese), depending on the class being referenced in the log-odds computation. Interestingly, `Gender_Male` had a strong positive coefficient ($\beta = 7.39$).

A number of behavioral interactions also appeared among the top predictors, such as `Age NCP` (age and number of main meals), `FCVC CH2O` (vegetable intake and water consumption), and `FCVC TUE` (vegetable intake and screen time). These interactions point to the complex interdependencies between diet and daily routines in determining weight class. Lifestyle indicators such as `CAEC_Frequently` (frequency of eating out) and `CALC_Frequently` (alcohol consumption) were also strong contributors, suggesting that eating out and drinking habits play a meaningful role in predicting obesity category transitions. As in the binary model, time spent using electronic devices (`log1p_TUE`) showed a positive association with certain classes, though its influence was somewhat more moderate.

Ultimately, the multinomial model's coefficients reveal that accurate classification of weight status relies not just on raw weight, but on other variables such as food intake, physical activity, alcohol use, and demographic variables. While some associations aligned with expectations, others—such as the shifting role of gender or the complex impact of alcohol consumption—highlight the importance of context when interpreting coefficients. These findings reinforce the value of rich feature engineering and interpretable models in public health research and support the potential for such models to guide targeted prevention strategies across a spectrum of obesity risk.

# References

Sun, J. Y., Badiani, R., & Matthews, D. W. (2024). Bmi obfuscation improves generalizability in obesity prediction using behavioral data. *Frontiers in Big Data*, *7*. Retrieved from https://www.frontiersin.org/articles/10.3389/fdata.2024.1469981/full DOI: 10.3389/fdata.2024.1469981

UCI Machine Learning Repository. (2022). *Estimation of obesity levels based on eating habits and physical condition.* https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition.

# 7 Appendix

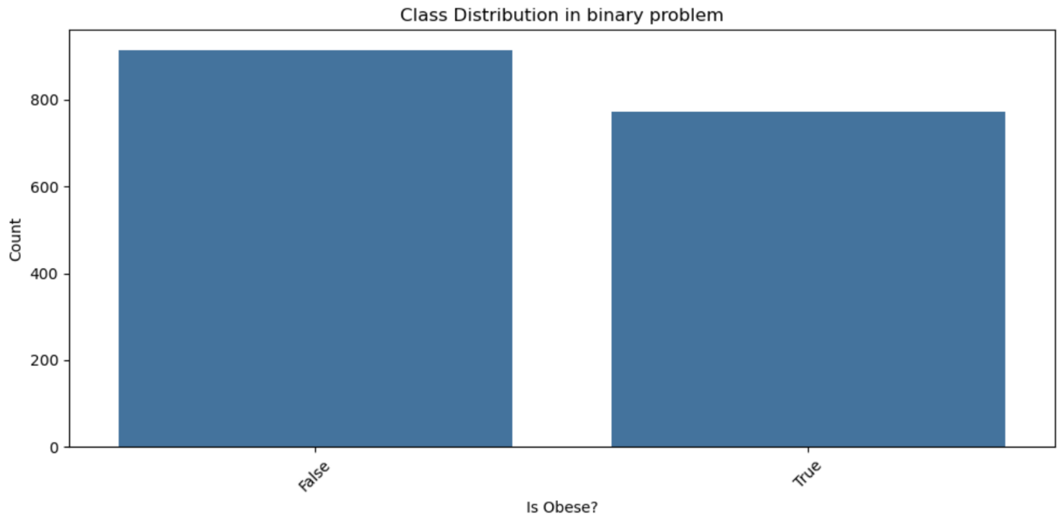

Figure 1: Distribution of Binomial Obesity Categories.

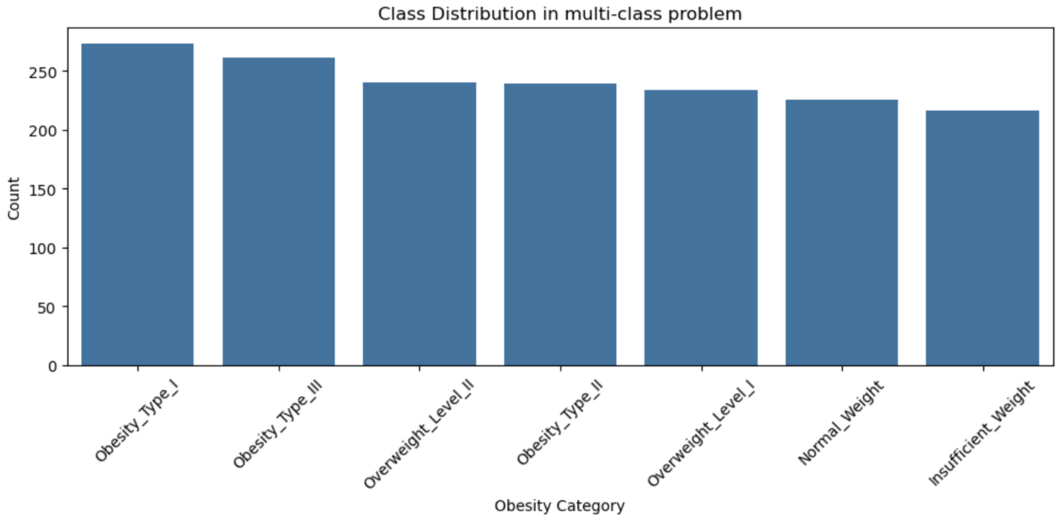

Figure 2: Distribution of Multinomial Obesity Categories.

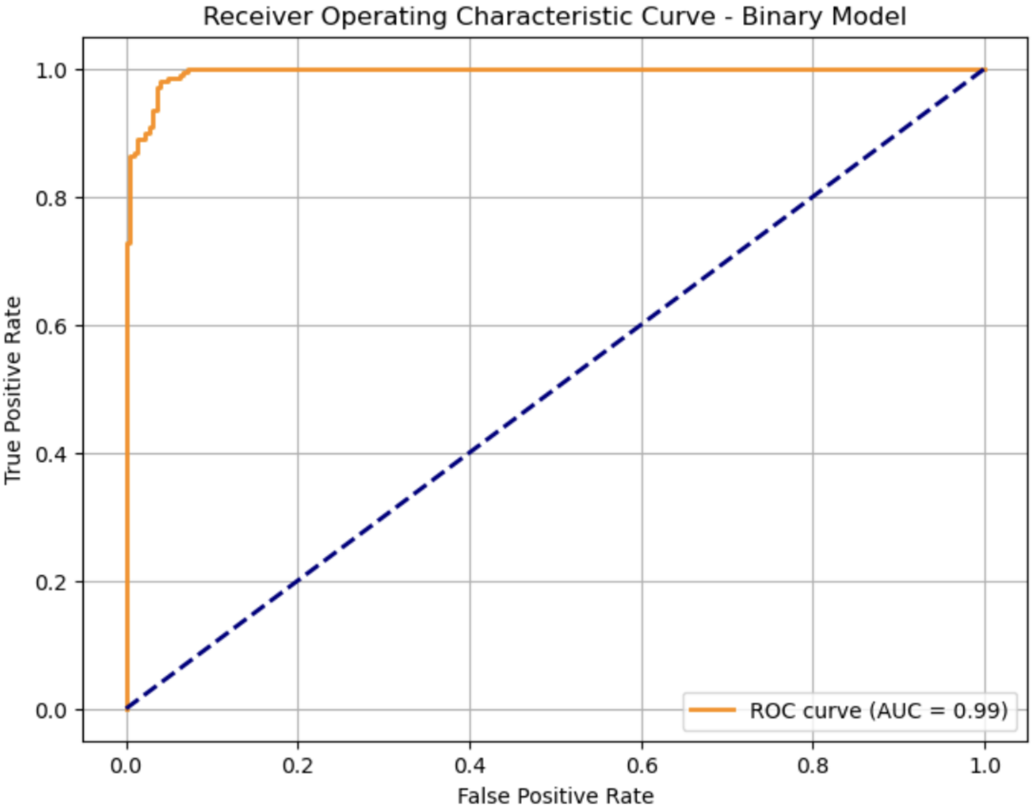

Figure 3: ROC curve for Binomial Logistic Regression

| **Variable** | **Coefficient** | **Absolute Coefficient** |
|---|---:|---:|
| Gender_Male | -4.392389 | 4.392389 |
| FCVC Weight | 2.913306 | 2.913306 |
| NCP Weight | 2.847838 | 2.847838 |
| log1p_Weight | 2.639394 | 2.639394 |
| log1p_TUE | -1.934976 | 1.934976 |
| Weight | 1.920716 | 1.920716 |
| TUE Weight | 1.887823 | 1.887823 |
| Age FCVC | -1.764305 | 1.764305 |
| log1p_FAF | -1.726216 | 1.726216 |
| MTRANS_Walking | -1.527345 | 1.527345 |
| CH2O TUE | 1.067157 | 1.067157 |
| Age CH2O | 1.044735 | 1.044735 |
| log1p_CH2O | -0.962328 | 0.962328 |
| FCVC NCP | -0.912646 | 0.912646 |
| CALC_no | 0.860265 | 0.860265 |
| family_history_with_overweight_yes | 0.781417 | 0.781417 |
| CAEC_Frequently | -0.773137 | 0.773137 |
| MTRANS_Public_Transportation | -0.757257 | 0.757257 |
| NCP TUE | -0.738755 | 0.738755 |
| TUE | -0.737358 | 0.737358 |
| Age NCP | -0.726740 | 0.726740 |
| exp_NCP | -0.609791 | 0.609791 |
| Age FAF | 0.555723 | 0.555723 |
| FAF Weight | 0.455353 | 0.455353 |
| NCP FAF | 0.435209 | 0.435209 |
| Ageˆ2 | 0.408678 | 0.408678 |
| CAEC_Sometimes | 0.383828 | 0.383828 |
| MTRANS_Motorbike | 0.319830 | 0.319830 |
| exp_FAF | -0.302509 | 0.302509 |
| FCVC FAF | 0.228396 | 0.228396 |
| NCP CH2O | -0.216993 | 0.216993 |
| SMOKE_yes | 0.213114 | 0.213114 |
| FAF TUE | 0.206545 | 0.206545 |
| CH2Oˆ2 | -0.193942 | 0.193942 |
| TUEˆ2 | -0.157075 | 0.157075 |
| FAVC_yes | 0.140879 | 0.140879 |
| Age | 0.091797 | 0.091797 |
| FCVC CH2O | -0.076903 | 0.076903 |
| log1p_Age | 0.058754 | 0.058754 |

Table 1: Top binomial logistic regression coefficients

| **Variable** | **Coefficient** | **Absolute Coefficient** |
|---|---:|---:|
| log1p_Weight | -15.240805 | 15.240805 |
| Gender_Male | 7.391642 | 7.391642 |
| Weight | -5.449063 | 5.449063 |
| Age NCP | 4.646445 | 4.646445 |
| FCVC CH2O | 4.379754 | 4.379754 |
| CAEC_Frequently | 3.794199 | 3.794199 |
| CALC_Frequently | -3.764253 | 3.764253 |
| CH2O FAF | -2.617706 | 2.617706 |
| log1p_NCP | -2.459601 | 2.459601 |
| CALC_Sometimes | -2.237846 | 2.237846 |
| log1p_TUE | 2.090039 | 2.090039 |
| FCVC TUE | -2.034605 | 2.034605 |
| NCP Weight | -1.991072 | 1.991072 |
| exp_FAF | -1.813217 | 1.813217 |
| MTRANS_Motorbike | -1.767815 | 1.767815 |
| log1p_FCVC | -1.723881 | 1.723881 |
| FAF^2 | 1.683448 | 1.683448 |
| Age FAF | 1.645654 | 1.645654 |
| CALC_no | -1.531236 | 1.531236 |
| MTRANS_Public_Transportation | -1.372818 | 1.372818 |
| FAVC_yes | 1.287263 | 1.287263 |
| TUE Weight | -1.281764 | 1.281764 |
| log1p_Age | -1.269120 | 1.269120 |
| NCP^2 | 1.255061 | 1.255061 |
| FCVC Weight | -1.253277 | 1.253277 |
| NCP FAF | 1.247274 | 1.247274 |
| Age | -1.228970 | 1.228970 |
| SMOKE_yes | -0.946492 | 0.946492 |
| SCC_yes | -0.902086 | 0.902086 |
| MTRANS_Walking | 0.894073 | 0.894073 |
| Age^2 | -0.813690 | 0.813690 |
| CH2O^2 | -0.795799 | 0.795799 |
| Age Weight | -0.691724 | 0.691724 |
| CH2O TUE | 0.686699 | 0.686699 |
| NCP TUE | 0.641797 | 0.641797 |
| MTRANS_Bike | -0.578800 | 0.578800 |
| exp_NCP | -0.531415 | 0.531415 |
| NCP CH2O | -0.431793 | 0.431793 |
| FCVC NCP | 0.378850 | 0.378850 |
| FCVC^2 | 0.349267 | 0.349267 |
| CAEC_Sometimes | -0.337133 | 0.337133 |
| Age CH2O | -0.326485 | 0.326485 |
| log1p_CH2O | 0.322194 | 0.322194 |
| FCVC FAF | 0.253490 | 0.253490 |
| family_history_with_overweight_yes | 0.250890 | 0.250890 |
| FAF Weight | 0.222676 | 0.222676 |
| log1p_FAF | -0.177887 | 0.177887 |
| TUE^2 | 0.066497 | 0.066497 |

Table 2: Top multinomial logistic regression coefficients