Identifying Lifestyle Factors Affecting Weight Using Machine Learning Tools

Elizabeth Eid

1. Introduction

Obesity is a disease that has plagued our society for decades, seemingly stemming from fast food, overeating, and laziness. Over the years, researchers have performed numerous studies in an attempt to understand the true causes of obesity, and even more nutritionists and personal trainers have attempted to create diets and exercise routines to help either cure obesity or prevent it. Obesity has been linked to several other diseases, including but not limited to diabetes, heart disease, and cancer.

More recently, researchers have been experimenting with uses of machine learning in relation to obesity, thanks to the wealth of data that is available through smartphones, smartwatches, and other health-tracking devices. Many studies are focused on predicting obesity in childhood or analyzing the effects of obesity. There are many different machine learning methods that have been used in this field, especially decision trees.

This study looks to perform regression on uncover factors that might drive a person's weight using several linear regression models, a linear generalized additive model, and two decision tree ensemble models. The goal is not to identify factors that lead to obesity, but rather to identify factors that might impact weight as a whole. The dataset used in this study consists of broad self-reported lifestyle factors, so the results may provide a starting point for people looking to alter their weight.

2. Data

This study features Fabio Palechor and Alexis Manotas's obesity data, which was published on UC Irvine's Machine Learning Repository (2019). Palechor and Manotas collected this data via an online survey over the course of one month. The survey consisted of 16 questions: five questions about the person's characteristics, six questions about eating habits, and six questions about physical condition (see Table 1). Unlike most datasets involving obesity, the eating habits and physical condition questions were all categorical, even those that could be numerical.

After they collected the survey results, Palechor and Manotas used height and weight to calculate the body mass index (BMI) of each record. BMI is calculated by dividing weight (kilograms) by height (meters) squared, and is used to screen for weight category: underweight, healthy weight, overweight, and obese (CDC 2022). Palechor and Manotas then used these BMI scores to classify each record as the appropriate weight status (see Table 2), as determined by the World Health Organization (WHO) (2010). Palechor and Manotas then graphed the spread of responses by weight category and noticed that the data was skewed heavily with more than half of the responses classified as 'Healthy Weight'. To correct the imbalance, they used the Weka tool and SMOTE filter, resulting in a dataset that is approximately 77% synthetic.

Data Code	Question	Possible Responses
Gender	What is your gender?	Female, Male
Age	What is your age?	Numeric user input
Height	What is your height?	Numeric user input
Weight	What is your weight?	Numeric user input
family_history_ with_overweight	Has a family member suffered or suffers from overweight?	Yes, No
FAVC	Do you eat high caloric food frequently?	Yes, No
FCVC	Do you usually eat vegetables in your meals?	Never, Sometimes, Always
NCP	How many main meals do you have daily?	Between 1 y 2, Three, More than three
CAEC	Do you eat any food between meals?	No, Sometimes, Frequently
SMOKE	Do you smoke?	Yes, No
CH2O	How much water do you drink daily?	Less than a liter, Between 1 and 2 L, More than 2 L
SCC	Do you monitor the calories you eat daily?	Yes, No
FAF	How often do you have physical activity?	I do not have, 1 or 2 days, 2 or 4 days, 4 or 5 days
TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others?	0-2 hours, 3-5 hours, More than 5 hours
CALC	How often do you drink alcohol?	I do not drink, Sometimes, Frequently, Always
MTRANS	Which transportation do you usually use?	Automobile, Motorbike, Bike, Public Transportation, Walking

Table 1

Survey questions used for data collection and the name of the corresponding column in the data.

BMI	Weight Category
< 18.5	Underweight
18.5–24.9	Normal
25–29.9	Overweight
30–34.9	Obesity I
35–39.9	Obesity II
40+	Obesity III

Table 2
Weight categories according to BMI, as determined by WHO.

3. Related Work

Though there has been an increase of machine learning techniques being used to study obesity, there are surprisingly few publications that try to use machine learning to analyze the underlying factors of obesity. There is a wide variety of approaches to this problem, which is unsurprising considering how many different factors could cause obesity, whether they are lifestyle choices, genetic, or otherwise.

One of the very first instances of using machine learning as a tool to gain insight on obesity was a study by Novak and Bigec in 1995. They used an artificial neural network (ANN) to predict obesity in children using data from the mother's pregnancy. The paper delves into ANNs and how they work, but unfortunately there is little emphasis on the results of the research.

Later, two studies utilized data mining to determine patterns related to childhood and adolescent obesity, one by Lazarou, Karaolis, and Matalas (2012), and the other by Pochini, Wu, and Hu (2014). Though the details of Lazarou's study were inaccessible for the purposes of this study, it was determined that among dietary factors, the consumption of junk food is the most associated with the risk of obesity in children. Pochini's study used data from a survey given to high schoolers that consisted of mostly yes or no questions. They then used Pearson Correlation Statistics, logistic regression, and decision trees to classify each high schooler as either overweight, obese, or neither. The Pearson Correlation Statistics demonstrated almost zero correlation between each factor and being either overweight or obese. The logistic regression and one of the decision trees had more success, and Pochini determined that physical activity, breakfast consumption, and smoking were all factors that significantly affected an adolescent's risk of obesity.

In 2017, Weichmann published a preliminary investigation in using a C4.5 decision tree to uncover new insights to child obesity. Weichmann's work with C4.5 decision trees further emphasizes the idea of using machine learning not just to diagnose problems, but to also uncover previously unresearched underlying causes of diseases.

So far, the majority of research in this topic has been conducted using data from children or adolescents. Ramyaa Ramyaa, Omid Hosseini, Giri P. Krishnan, and Sridevi Krishnan (2019),

instead used a dataset of diet and physical activity factors to predict body weight on adult (postmenopausal) women. This study was different for two reasons. First, it had a greater number of factors than the previous studies, and second, it focused on predicting body weight, rather than classifying into 'normal' or 'obese'. With an emphasis on variable importance, Ramyaa used many different algorithms: stepwise linear regression, k-Nearest Neighbors, Gaussian support vector machine (SVM) regression, regression tree, and neural network. They also used some categorical approaches, but the highest accuracy for classifying obesity was only 54%. For the numerical approaches, SVM performed the best with a mean approximate error (MAE) of 6.70, but the R^2 value was just 0.3, which implies there is very low correlation. The neural network and k-nearest neighbor algorithms performed similarly.

4. Methodology

4.1 Preprocessing

Although Palechor and Manotas's data is focused on classifying obesity, this study disregards this completely and insteads performs regression on weight. BMI is cheap, accessible, and has been the standard for nearly 200 years, making it a decent method of screening for weight category; however, it does not take into account age, race, or muscle mass, making it inconsistent for various groups of people (Jitnarin et al., 2014). Weight also does not take these factors into account, but it also does not try to make any broad statements about the person's health. Rather, the goal is to make broad statements of lifestyle habits that may impact a person's weight. Aside from dropping the weight classification column, the data was also limited to ages 20 or older. Since children and adolescents are still growing, it is standard to evaluate measurements such as height and weight differently in children and adolescents than in adults (CDC, 2022).

The data was quite messy and it took some work to convert the data into a usable state. First, the synthetic data was constructed with infinite decimals for the numerical factors. To make the synthetic data reflect human input data more appropriately, these columns were rounded to zero and converted to integers, with the exception being the height and weight columns, which were rounded to two decimal points instead.

Most of the columns, however, are not numeric and had to be converted. The columns with 'yes' or 'no' responses were fixed so 'yes' was 1 and 'no' was 0. The columns with 'no', 'Sometimes', 'Frequently', or 'Always' responses, were converted to 0, 1, 2, and 3, respectively. The gender column was converted so that 'Female' was 1 and 'Male' was 0. The methods of transportation were trickier to rank as the response options were not as ordinal as the others. After much consideration into what might constitute more or less regular physical activity, the responses 'Automobile', 'Motorbike', 'Public Transportation', 'Bike', and 'Walking' were converted to 0, 1, 2, 3, and 4, respectively.

The synthetic data was adjusted for weight categories, not for weight itself, so it was important that the data was still balanced. To do so, each factor was plotted in a histogram to see their distributions (see Figure 1). The majority of the factors were mostly balanced, but there were some factors that were not as balanced as hoped. The most important factors, height and weight, were normally balanced, so the data was deemed usable. Further analysis on the impact of imbalance factors follows in the discussion section. A correlation heatmap was also plotted,

and it shows there is not much linear correlation either between the factors and weight, or even between other factors (see Figure 2).

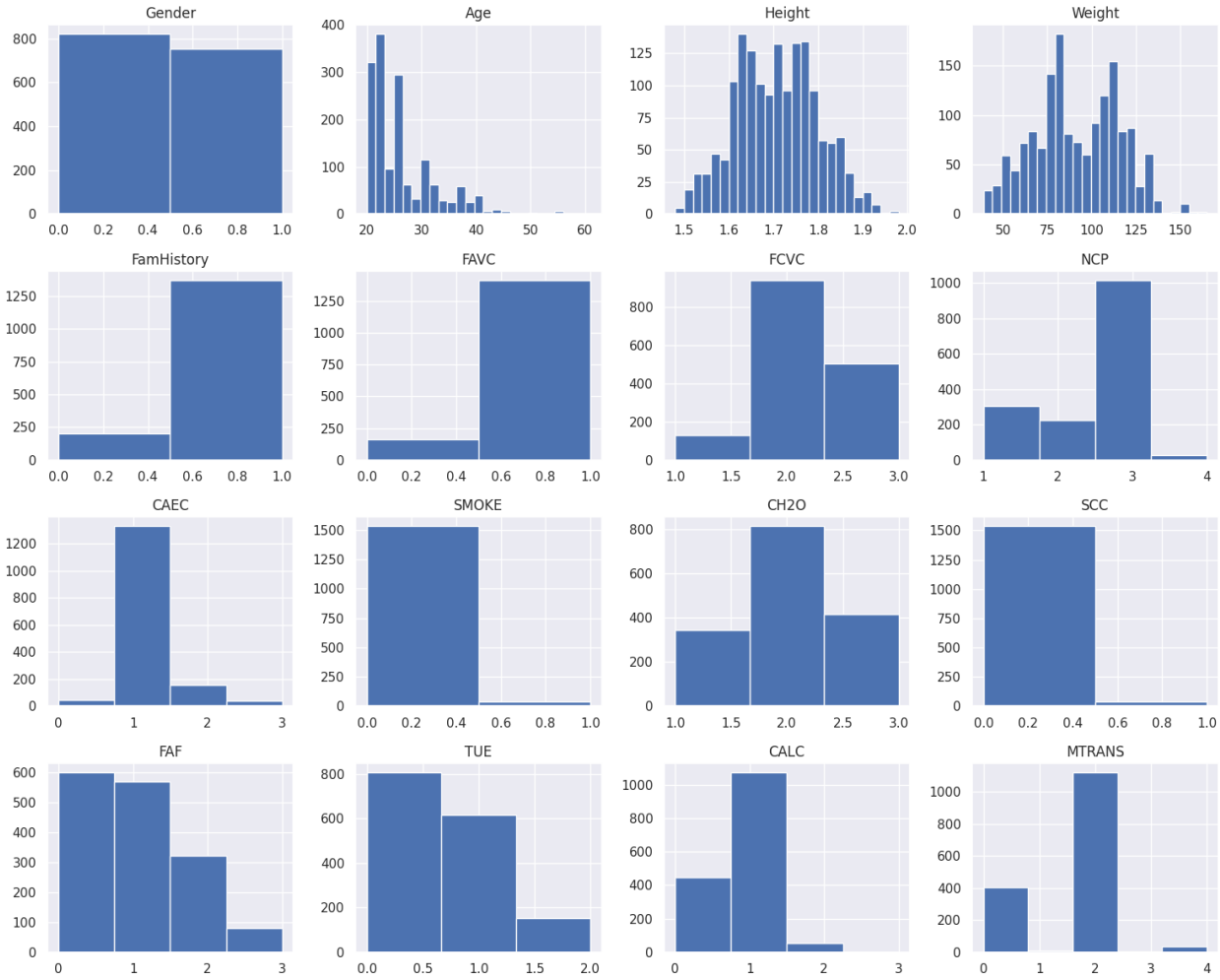

Figure 1

Distribution histograms for each factor (see Table 1 for details). Many of the factors are balanced or mostly balanced, but some of the other factors are less imbalanced, which should be taken into account when conducting analysis on the results.

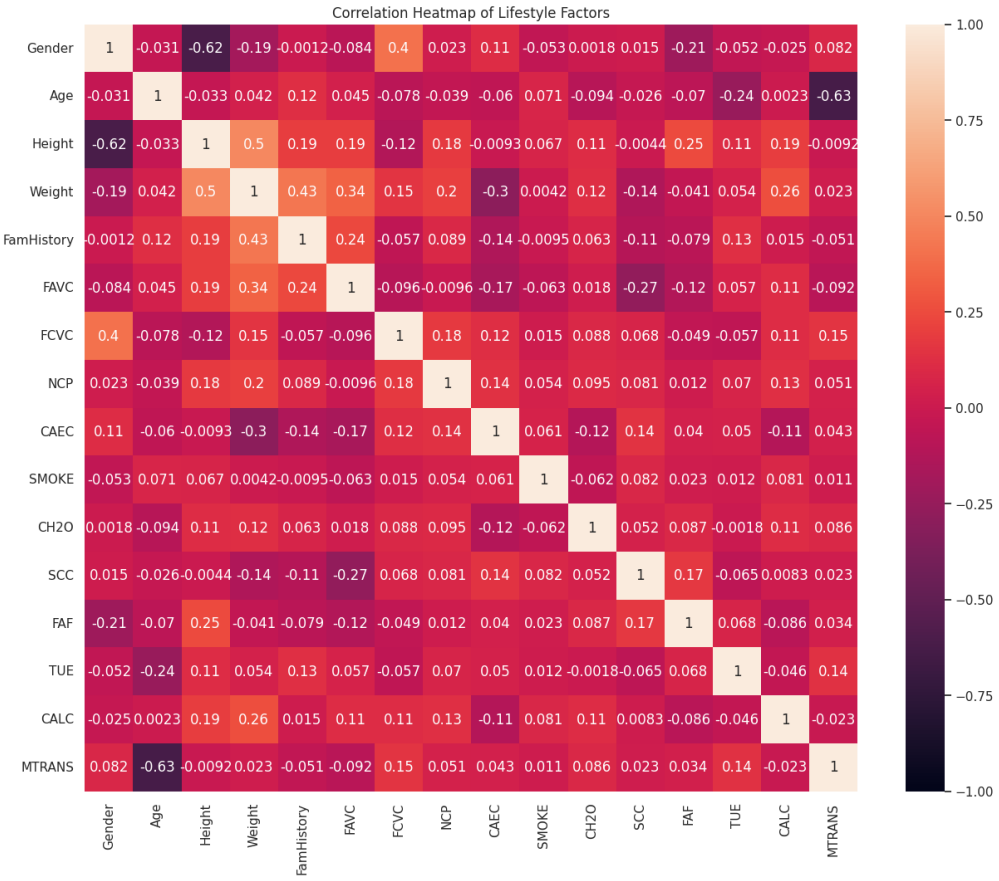

Figure 2

Correlation heatmap built with seaborn. The plot shows very little correlation between each factor.

4.3 General Workflow

In order to assess feature importance as completely as possible, several linear and non-linear methods were used, of which more details will follow. The general procedure for each model was approximately the same. Especially given the ordinality of the data, scaling the data was essential. After much debate, the scaler chosen was scikit-learn's quantile transformer with 9 quantiles, which transforms each feature to follow a uniform distribution.

First, the model parameters were assessed using GridSearchCV from the scikit-learn package (see Figure 3). The grid search algorithm used a 5-fold cross validation and a scoring parameter of maximizing R^2. For some models, it was necessary to perform gridsearch multiple times, once with broad parameter ranges, then again with a narrower range. This was done to limit the size of the grid, and thus limit the execution time. Then, those parameters were used in a 10-Fold cross validation of each model, in which the mean squared error, R^2 value, and coefficients (or feature importances, for the tree methods) were recorded for each fold, then averaged. The only exception to this workflow was LinearGAM, which had a built-in gridsearch method.

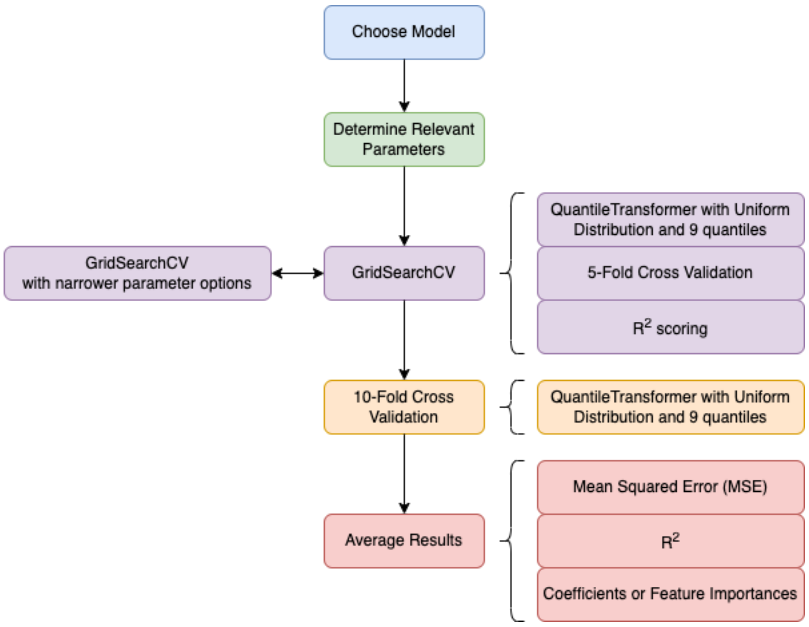

Figure 3

A flowchart depicting the workflow of evaluating each model. Each model had slightly different parameters, but the general setup was the same for every model, except for LinearGAM. This was possible thanks to the scikit-learn standard for building model packages.

4.4 Linear Methods

This study utilizes many different linear models: linear regression, ridge regression, lasso regression, and elastic net regression. Linear regression models attempt to model the relationship between the independent and dependent variables with a straight line or a flat plane. Let $X = \{x_1, x_2, \ldots, x_n\}$ be the independent variables, $B = \{\beta_0, \beta_1, \beta_2, \ldots, \beta_n\}$ be the coefficients, and y be the dependent variable. Then the linear regression model is:

$$y = X \cdot B = \beta_0 + x_1 \cdot \beta_1 + x_2 \cdot \beta_2 + \ldots + x_n \cdot \beta_n$$

The values of B are chosen by minimizing the residual sum of squares (RSS), which measures the difference between the model estimation and the true values. The equation for RSS is:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})^2$$

Ridge, lasso, and elastic net are methods of regularization, which is a measure taken to prevent model overfitting. They use the same framework as linear regression, but each adds a different penalty term to the loss function to prevent coefficients from becoming too large.

Ridge uses Euclidean distance to do so, while Lasso uses what is known as the 'city block' distance, and Elastic Net uses a compromise between both. These penalty terms are multiplied by a parameter, alpha, that determines the tolerance of the penalty term. As alpha increases, more coefficients will converge towards zero, depending on how strong the correlation is.

4.5 Non-Linear Methods

There are three non-linear models that were selected to be used in this study. Linear Generalized Additive Model (LinearGAM), Random Forests, and XGBoost Trees.

LinearGAM uses the linear model framework (described above), but separates the data into a user-defined number of splines, or sections. Each of these splines are fit to a different polynomial function, then added together to create a smooth piecewise polynomial function. The only constraints are that the overall function and its first two derivatives are continuous. Due to the nature of LinearGAM, there are different coefficients for each spline. To account for this, the maximum coefficient for each feature was recorded per fold. Then, the coefficients were averaged like the rest of the models.

Random Forests are XGBoost Trees are both ensemble methods that use decision trees, but there are several key differences. Random forests construct multiple decision trees with randomly selected subsets of features. The trees maximize each split by using the features that are the most effective at splitting the data first, then the second most effective feature, and so on. The prediction is then made by averaging the predictions of each tree. XGBoost trees, unlike random forests, are built sequentially and build on each other with each tree trying to improve on the last. The final prediction is then averaged across each tree's result, with each tree weighted based on its performance.

5. Results

As expected, the decision tree models far outperformed the linear models (see Table 3). The linear models performed so similarly that the variance of both the MSE scores and the R^2 scores were below 0.5. This implies that the data is not linear, but it does not mean that the information we can gain from it is irrelevant. The coefficients of each model were assembled into a table for comparison (see Table 4).

Linear Model	MSE	R^2	Non-Linear Model	MSE	R^2
Linear Regression	255.41675	0.226072	Linear GAM	216.78367	0.427607
Ridge	255.33278	0.214353	Random Forest	75.500748	0.852262
Lasso	256.89311	0.127443	XGBoost Tree	94.780767	0.830279
Elastic Net	255.30190	0.206159			

Table 3

The performance results of each model. Random Forest outperformed all of the models in both MSE and R^2. The linear models (left) all performed almost exactly the same, while the non-linear models (right) had a much greater variance in performance. This makes sense, since the linear models are much more related to each other than the non-linear models are.

	Linear Regression	Ridge	Lasso	Elastic Net	Mean
Gender	1.94	1.69	0.35	1.53	1.38
Age	12.79	12.63	9.76	12.41	11.90
Height	42.36	41.60	38.80	41.21	40.99
FamHistory	17.33	17.42	17.89	17.47	17.53
FAVC	11.40	11.47	11.53	11.50	11.47
FCVC	18.40	18.31	17.98	18.26	18.24
NCP	6.31	6.28	4.92	6.19	5.92
CAEC	-37.79	-36.86	-33.50	-36.39	-36.14
SMOKE	-1.09	-1.09	0.00	-0.90	-0.77
CH2O	-1.17	-1.06	-0.09	-0.93	-0.81
SCC	-7.33	-7.24	-3.26	-6.99	-6.20
FAF	-3.94	-3.84	-3.37	-3.79	-3.73
TUE	1.31	1.34	0.75	1.31	1.18
CALC	6.28	6.47	6.23	6.50	6.37
MTRANS	10.69	10.51	7.72	10.29	9.80

Table 4

The coefficients of the linear models. The correlation is stronger the farther from zero the coefficient is, i.e. a large positive value, like Height, has a strong positive correlation, while a large negative value, like CAEC, has a strong negative correlation. Values closer to zero signify a low correlation, or no correlation at all.

Among the non-linear models, Random Forest outperformed the others, with XGBoost Trees as a close second (see Table 3). LinearGAM's performance was worse than the decision tree-based models, but still outperformed the linear models by a relatively large margin. Since Random Forest and XGBoost trees do not have a coefficient vector, they instead have a feature importance attribute, which assigns each feature a weight based on how high up the tree the feature is. The feature importances also does not account for positive or negative correlation in the data.

To account for the measure differences during comparisons, the absolute values of the coefficients and the feature importances were scaled using the standard scaler. Though taking the absolute value of the coefficients removed the ability to see positive or negative correlation, it made comparisons much easier, as it meant large values imply higher correlation while smaller values imply low or no correlation.

With the exception of XGBoost, the most significant feature was height (see Table 5). This is unsurprising because height is used to generalize weight in BMI. The frequency of food consumption between meals is Random Forest's second most significant—and XGBoost Tree's most significant—feature. It should also be noted that LinearGAM's second most significant

feature was the method of transportation, though it was not deemed very significant for the other three models.

	Linear Models (Mean)	LinearGAM	XGBoost Tree	Random Forest
Height	2.46	2.79	0.18	2.98
CAEC	2.06	1.09	2.80	1.20
FCVC	0.56	0.33	1.17	0.97
FamHistory	0.51	-0.02	1.30	0.12
Age	0.03	-0.19	-0.65	0.07
Gender	-0.85	-0.90	0.16	0.02
CALC	-0.43	-0.51	-0.11	-0.26
NCP	-0.47	-0.43	-0.38	-0.26
FAVC	0.00	-0.28	0.02	-0.47
MTRANS	-0.14	1.49	-0.49	-0.49
FAF	-0.65	-0.69	-0.84	-0.64
CH2O	-0.89	-0.71	-0.81	-0.71
TUE	-0.86	-0.51	-0.80	-0.72
SCC	-0.44	-0.64	-0.76	-0.90
SMOKE	-0.90	-0.82	-0.80	-0.92

Table 5

The absolute and scaled values of the linear models and linearGAM coefficients, and the scaled values of the Random Forest and XGBoost Tree feature importance. The features are ordered by the Random Forest values since it was the best scoring model. This table shows only the mean of the linear model coefficients because they are so similar that it would be redundant if each model was listed in its own column.

6. Discussion

The results of this study are promising. For the most part, the results agree with other medical studies with height, snack consumption, and vegetable consumption being important factors in weight. It was surprising that the family history of being overweight was less significant than expected, though it was in the top five features for most of the models.

The bigger surprise was how insignificant smoking was. Smoking has long been associated with obesity, but this study suggests that it actually has very little effect on weight. It is, however, important to remember that smoking was highly imbalanced in the data, with only about 2.6% of the records being positive for smoking. It is unclear whether the imbalance caused the insignificance, or the data. Calories monitoring has a similar situation, where the models do not consider it significant, but may be influenced by the large imbalance in the data.

It is not possible to draw any true conclusions from this study, but it does invite more experimentation with similar data. It would be good to see data with more variety and more balance, though it was interesting to see data that focuses on lifestyle factors that would not require heavy monitoring, such as nutrition intake. Random Forests were consistently the best models for this type of study. It has definitely been shown that weight does not have a linear relationship with these lifestyle factors, so further study of Random Forests would also be interesting to see.

7. References

Centers for Disease Control and Prevention. (2022, June 3). *About adult BMI*. Centers for Disease Control and Prevention. https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html#InterpretedAdults

Jitnarin, N., Poston, W. S. C., Haddock, C. K., Jahnke, S. A., & Day, R. S. (2014). Accuracy of body mass index-defined obesity status in US firefighters. *Safety and Health at Work*, *5*(3), 161–164. https://doi.org/10.1016/j.shaw.2014.06.003

Lazarou, C., Karaolis, M., Matalas, A.-L., & Panagiotakos, D. B. (2012). Dietary patterns analysis using data mining method. an application to data from the CYKIDS study. *Computer Methods and Programs in Biomedicine*, *108*(2), 706–714. https://doi.org/10.1016/j.cmpb.2011.12.011

Novak, B., & Bigec, M. (1995). Application of artificial neural networks for childhood obesity prediction. *Proceedings 1995 Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, 377-380. https://doi.org/10.1109/annes.1995.499512

Palechor, F. M., & Manotas, A. de la H. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, *25*. https://doi.org/10.1016/j.dib.2019.104344

Pochini, A., Wu, Y., & Hu, G. (2014). Data mining for lifestyle risk factors associated with overweight and obesity among adolescents. *2014 IIAI 3rd International Conference on Advanced Applied Informatics*, 883-888. https://doi.org/10.1109/iiai-aai.2014.175

Ramyaa, R., Hosseini, O., Krishnan, G. P., & Krishnan, S. (2019). Phenotyping Women Based on Dietary Macronutrients, Physical Activity, and Body Weight Using Machine Learning Tools. Nutrients, 11(7), 1681. https://doi.org/10.3390/nu11071681

Wiechmann, P., Lora, K., Branscum, P., & Fu, J. (2017). Identifying discriminative attributes to gain insights regarding child obesity in Hispanic preschoolers using Machine Learning Techniques. *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, 11–15. https://doi.org/10.1109/ictai.2017.00014

World Health Organization. (2010, May 6). *A healthy lifestyle - WHO recommendations*. World Health Organization. https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations