

The goal of this project was to answer the following research question: Can an individual's obesity level be accurately predicted based on their lifestyle variables? Researching the answer to this question is worthwhile because it would provide individuals and healthcare officials with more information about how different factors like water consumption, smoking habits, and technology use could contribute to obesity, and how clearly personal weight levels are affected by lifestyle habits.

The dataset was found on Kaggle and downloaded, but the data was originally collected by Fatma Hilal Yagin, approved by Inonu University, for the estimation of obesity levels in individuals from Mexico, Peru, and Colombia, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records, and the obesity records are labeled with the class variable NObesity (Obesity Level), which allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Additionally, 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, and 23% of the data was collected directly from users through a web platform. The downloaded data was clean, with no missing values or unclear variables.

To construct a decision tree using Random Forest, there weren't many preprocessing steps necessary aside from encoding categorical variables into numerical representations and splitting the training dataset into a training and validation set. For all of the models, the data was split into a 60/20/20 ratio, with the remaining 20 being the final testing dataset. To assess the model's performance, the cross-validation grid score and sklearn's own metric `accuracy_score()` were used. For the neural network, the preprocessing pipeline scaled numeric features with `MinMaxScaler` and label-encoded the target classes. Each model was evaluated using 5-fold cross-validation to make sure that performance was consistent even with different splits of the data. The neural network's performance was compared using accuracy and F1-score, and we also looked at the confusion matrix. For the penalized regression model, numeric features were scaled and categorical variables were one-hot encoded during preprocessing. The SVM used the same preprocessing setup, applying standardization and one-hot encoding before training and tuning the model.

<u>Model Performances</u>	
Model	Cross-validated Performance Accuracy
Random Forest	0.87
Penalized Regression	0.60
SVM Regression	0.96

Neural Network	0.91
----------------	------

The best model after cross-validation was the SVM, the second best model was the neural network, then the random forest, and finally the penalized linear regression was last. This gives us insight into the structure of our data. We can assume that obesity classification is non-linear and that our data had more complicated (but not too complicated) patterns. Especially with our one-hot encoding, it makes sense that the linear models would struggle compared to the ones that excel with non-linear relationships.

The performance for the best model on the final test set was an accuracy of 0.978, which is very high. Our best model was the support vector machine, and we believe this is because the type of data we were working with is where support vector machines excel. Our data had non-linear relationships and was a multi-class classification problem, which all works very well with SVM's. Additionally, our dataset was small/medium sized and moderately dimensional with 31 variables. Most of them are also binary, which SVM's are good with while other models (like random forest) are not. As expected, the most influential feature overall to determine obesity level was weight, which makes sense because weight is a direct indicator and measure of obesity. In the random forest model, the next most influential feature was vegetable consumption, then age and gender.

Despite the high accuracy of our best model, there are some limitations of our project present. Since the dataset was compiled using data outside of the US, the results might not be very accurate if applied to people within the country. Problems with extrapolation, making predictions outside the scope of the model, will likely be apparent, considering the lifestyle of people in the US is very different from Mexico, Peru, and Colombia. Additionally, a large majority of this data is synthetic, so while the results display possibilities, they may not be truly reflective of patterns in reality compared to a model that used data entirely collected from real people. Another limitation within our data is that we chose to include the weight column within the dataset. We knew that this would have the largest impact, but it had over double the impact of the second feature (vegetable consumption). This is a very significant impact from one variable, so if this was removed, the model accuracy would likely fall significantly.