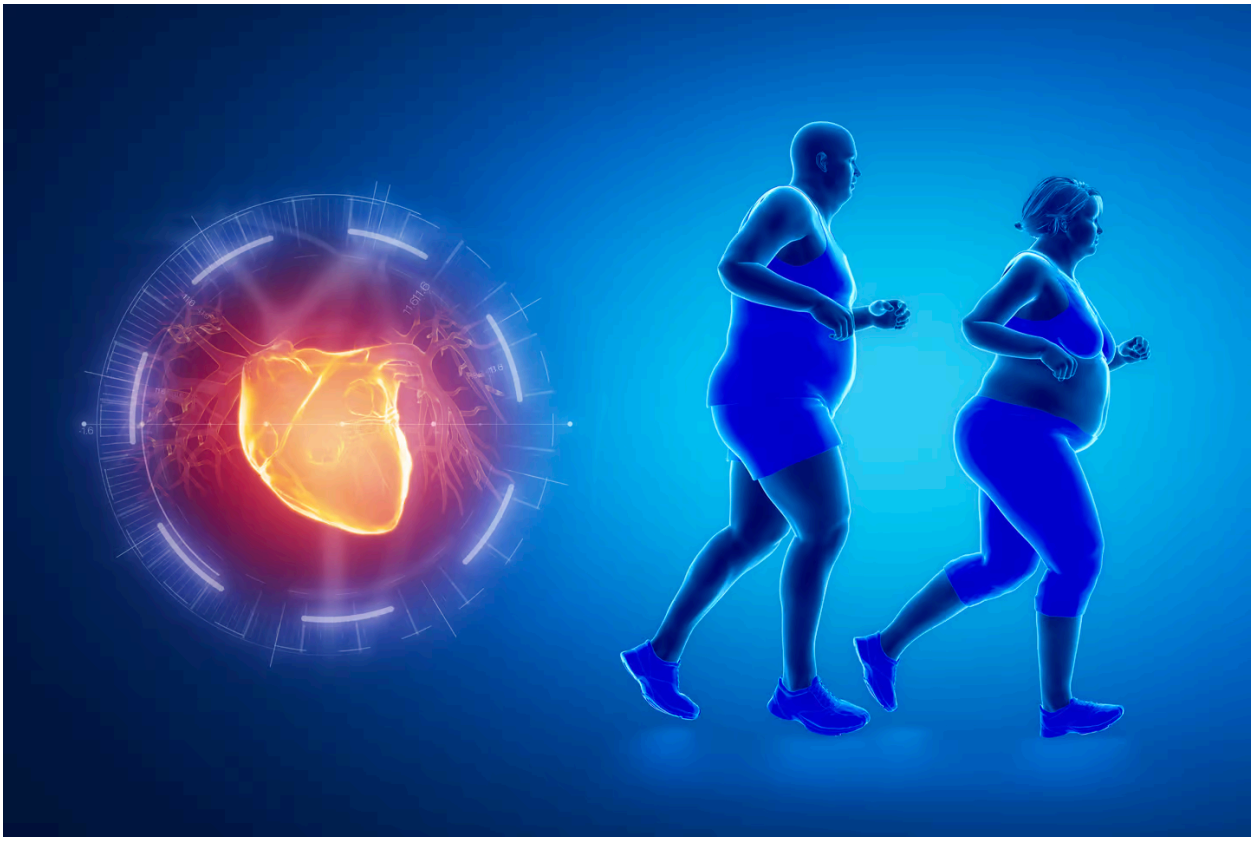
Elie Park
Mentor: Ben Bell
Springboard Project
March 20, 2022



Obesity Prediction

Introduction

Obesity puts people at risk for many other diseases, such as heart disease, and cancer. Across OECD countries, being overweight explains 71% of all treatment costs associated with diabetes, 23% of costs related to cardiovascular diseases, and 9% of costs related to cancers[1]. A 2009 meta-analysis of the co-morbidities of obesity showed that of the 18 weight-related diseases studied, diabetes particularly had the strongest association with obesity. Of individuals with BMIs of 30 or higher, men had a risk 7 times higher, and women had a risk 12 times higher than those in the normal weight range of developing type II diabetes[2].

Canada is now facing an obesity epidemic. According to the Canadian Community Health Survey, 64 per cent of adults equalling 24 million people are now overweight or obese in 2020. 9 billion dollars are spent every year on obesity-related direct healthcare costs, including physicians, hospitalizations, and medication costs. For instance, patients with a high BMI and co-morbidities that are too severe to tolerate surgery or equipments spend a long time in the hospital and experience a reduced quality of life. In Alberta alone, reducing extended stays by

---

[1] OECD. (2019). The Heavy Burden of Obesity. *The Economics of Prevention.* https://www.oecd-ilibrary.org/social-issues-migration-health/the-heavy-burden-of-obesity_67450d67-en

[2] Guh, D., Zhang, W., Bansback, N., Amarsi, Z., Birmingham, C. L. & Anis, A. H. (2009). The Incidence of Co-Morbidities Related to Obesity and Overweight: A Systematic Review and Meta-analysis. *BMC Public Health.* https://pubmed.ncbi.nlm.nih.gov/19320986/

just one night for people with a body weight of over 250 pounds would save the province 14 million dollars annually[3].

Governments around the world with a public or hybrid health care systems should aim to lower health costs related to obesity proactively by promoting healthier lifestyle choices. Other charitable organizations and associations such as Obesity Canada and Dietitians of Canada may benefit from what lifestyle features most affect obesity risk, as they help connect members of the public who are affected by obesity with researchers, health care professionals, and registered dieticians. Weight loss apps such as Lose it or Fitbit also may consider this data for healthy behaviour interventions. Weight loss services in Canada currently have a growing market of $350M.

Following my grandfather and aunt's diagnoses of diabetes, I was curious which lifestyle features are most attributed to obesity risk, since obesity is complex and difficult to treat and managing it is a lifelong process. Logistic Regression, Random Forest Classifier, and XGB Classifier were used for binary classification of overweight vs. not overweight based on lifestyle features. These models did not prove to be useful in predicting obesity with a high enough precision and recall, but the exploratory data analysis showed certain trends that with enough data, could be used to build a useful model in the future.

The dataset was retrieved from the UCI machine learning repository and was originally used in developing an obesity level estimation software[4]. This dataset consists of 17 attributes and 2111 records, surveying Latin American undergraduate students online. These attributes include age, height, weight, gender, family history of obesity, vegetable consumption, food and water intake, alcohol and nicotine usage, physical activity, transportation methods, technology usage, calorie consumption monitoring, and obesity level based on adult WHO classifications. The last 77% rows containing synthetic data created with the Weka tool and SMOTE filter were dropped for this analysis.

Obesity was provided with 7 levels, but Body Mass Index was numerically calculated with the given height and weight through BMI = weight/height^2. The distribution of BMI was right-skewed, with a mean of 24.3 and median of 23.7, both of which fall under but is closer to the upper limit of the normal BMI range. The feature importances were calculated with Random Forest Regressor at 4 and 7 levels of obesity. The final modeling was done with 2 levels: obese and not obese. The final dataset consists of 21 columns (new age group and 3 new variables defining obesity), and 479 rows.

---

[3] Hrvatin, V. (2019). Canada's $9 Billion Obesity Problem. *Chatham Daily News.* https://www.chathamdailynews.ca/diseases-and-conditions/it-costs-canada-9b-to-treat-obesity-while-barely-any-money-is-put-into-preventative-care

[4] De-La-Hoz-Correa, E., Mendoza-Palechor, F. E., De-La-Hoz-Manotas, A., Morales-Ortega, R. C. & Hernandez, S. (2019). Obesity Level Estimation Software Based on Decision Trees. *Journal of Computer Science.* https://thescipub.com/pdf/jcssp.2019.67.77.pdf

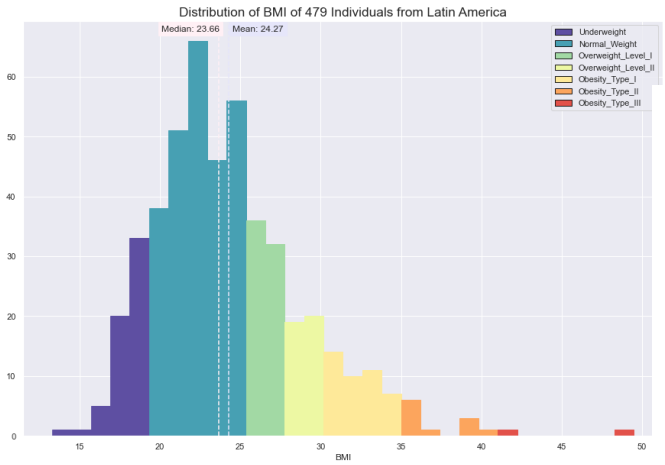


Figure 1. Distribution of BMI

Figure 2. WHO BMI Index

Univariate Data Analysis: Age

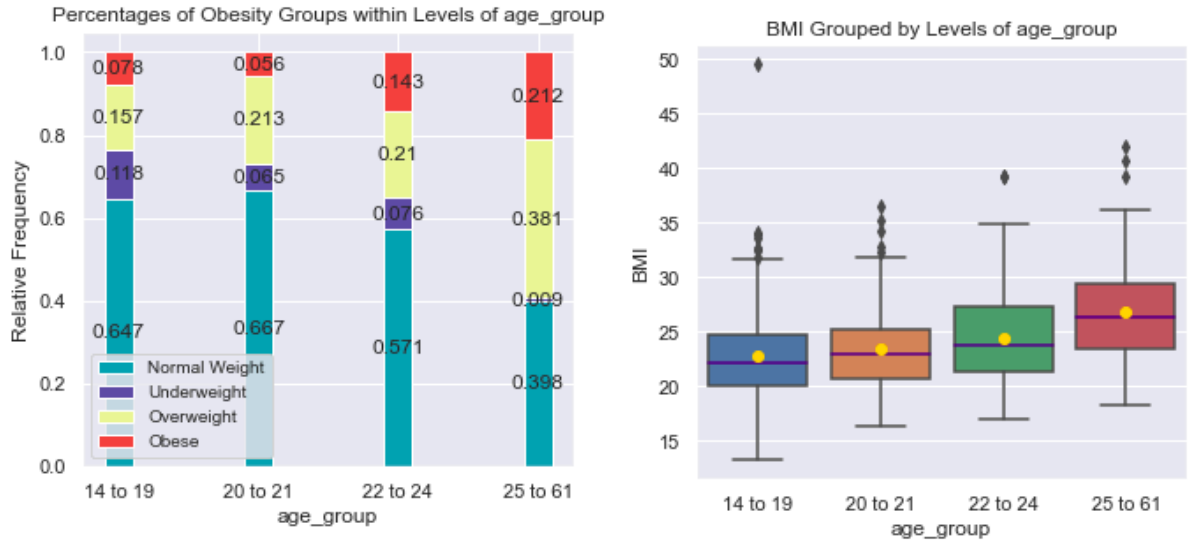


Figure 3. Distribution of BMI within Age Groups

Age was a metric variable which was converted into 4 different age groups based on quartiles. Vegetable consumption, number of main meals, daily water intake, physical activity, and time using technological devices contained floats of ordinal values corresponding to levels of consumption or activity, some of which were interval values before conversion by the original authors of the dataset. Food between meals and alcohol consumption contained ordinal survey answers which were then translated to numerical values. Gender, family history of obesity, high calorie food consumption, calorie consumption monitoring, and nicotine usage were dichotomous variables. Transportation method contained answers pertaining to 5 different

transportation methods. Answers with less than 20 value counts within each column were dropped.

Figure 1 showed that we have many counts of people with normal weight. Because this dataset was based on an internet survey targeting the younger population, it potentially was not an accurate representation of the whole population of Latin America. The distribution of age as shown in Figure 2 may explain some of the skewness of BMI, pushing the median to the left. The age group from 25 to 61 have the highest BMI and proportion of overweight and obese individuals. A chi-squares test confirmed that the means of BMI were higher in older age groups ($p << 0.001$). The lowest age group from ages 14 to 19 had the highest count of individuals. It should be noted that obesity ranges for individuals of ages less than 20 were age and gender specific (based on the CDC growth chart), and which were not accounted for in this study.

Univariate Data Analysis: Family History of Obesity

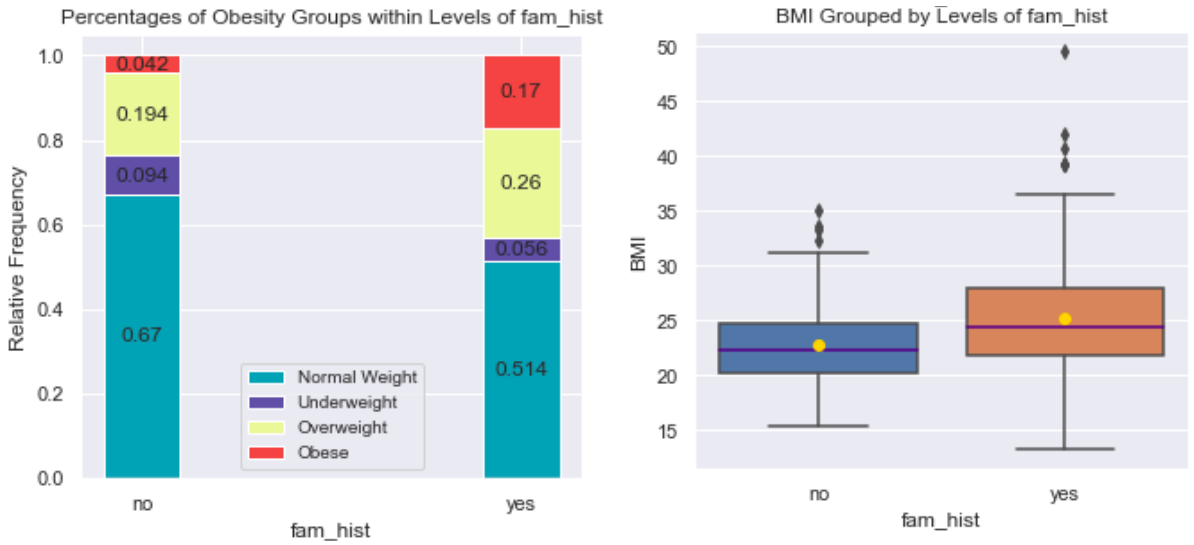


Figure 4. Distribution of BMI within Groups Having Family History of Obesity

Family history of obesity was one of the variables with the highest difference in relative frequencies and group means of obesity. Having a family history of obesity seems to result in an overall higher BMI and especially more counts of extreme obesity. A pooled t-test for group differences showed that there was a difference in means of BMI scores between family history groups ($p << 0.001$). Genetics or behavioural patterns in the family may be important features in determining obesity risk.

Univariate Data Analysis: Physical Activity

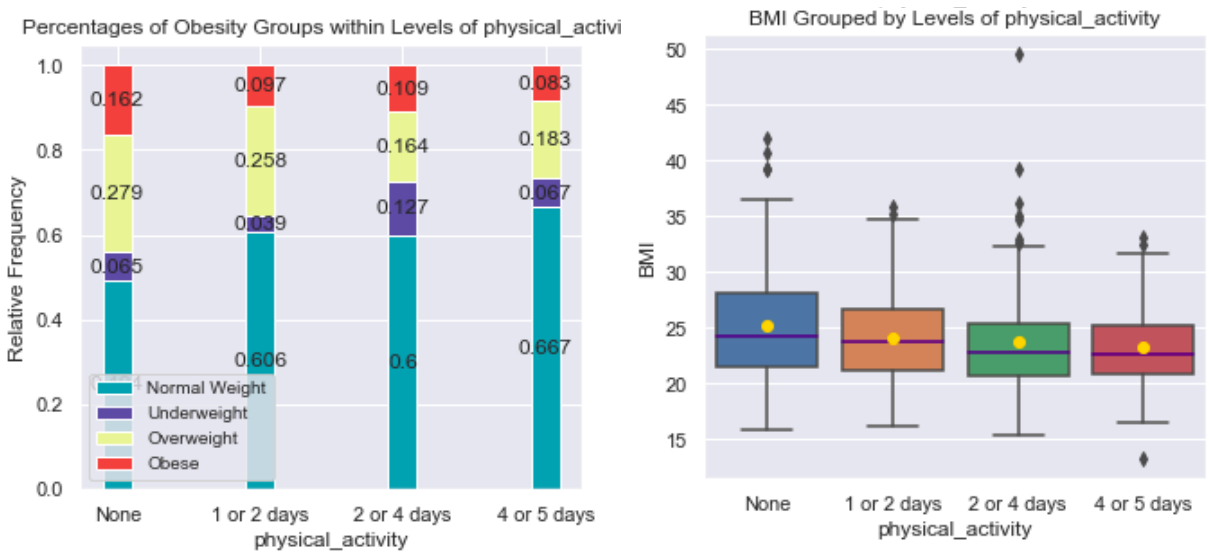


Figure 5. Distribution of BMI within Physical Activity Groups

Obvious trends were also found in physical activity levels. Those who were more physically active per week tended to be less obese. Although the counts of people who never exercised vs. those who exercised 1 or 2 days per week were similar (162 to 158), the relative frequencies of obesity were much lower in those who did exercise for even just 1 or 2 days. A chi-squares test showed that there was a difference of BMI among these levels of physical activity ($p < 0.05$).

Bivariate Data Analysis

The below table shows the results of pooled t-test for dichotomous group differences and chi-squares test for multiple levels, from lowest to highest p-values (most to least significant result). The t-test tested for differences in numerical BMI means between dichotomous levels. The p-values from t-tests were divided by two to mimic a one-tailed t-test since `ttest_ind` from `scipy.stats` is a two-sided test by default. The chi-squares test tested for a difference in levels of obesity (underweight + normal weight vs. overweight + obese) across multiple levels within the categorical variables, with a BMI >= 25.0 being overweight/obese. A chi-squares test is always one-tailed so the p-value was not adjusted.

	Statistical Test	Group with Highest BMI scores	P-value
Age Group	Chi-squares	14 to 19 / 20 to 21 / 22 to 24 / **25 to 61**	6.09E-09
Family History of Obesity	Pooled t-test	**Has a family history of obesity** / no family history of obesity	3.19E-08
Nicotine Usage	Pooled t-test	**Smokes nicotine** / not a nicotine smoker	6.64E-04
Number of Main Meals	Chi-squares	**Between 1 and 2** / 3 / More than 3 Meals per day	6.66E-04
Transportation	Chi-squares	**Automobile** / Public Transportation / Walking	7.44E-04
Time Using Technological Devices	Chi-squares	0 to 2 hours / 3 to 5 hours / **More than 5 Hours** per day	0.0016
Gender	Pooled t-test	**Male** / Female	0.0027
Alcohol Consumption	Chi-squares	**Frequently** / sometimes / no	0.0027
Daily Water Consumption	Chi-squares	Less than 1L / Between 1 and 2L / **More than 2L** of Water per day	0.007
Physical Activity	Chi-squares	**No Physical Activity** / 1 to 2 days / 2 to 4 days / 4 to 5 days per week	0.016
Food Consumption Between Meals	Chi-squares	Result not significant	0.074
Calorie Consumption Monitoring	Pooled t-test	Result not significant	0.08
Vegetable Consumption	Chi-squares	Result not significant	0.26
High Calorie Food Consumption	Pooled t-test	Result not significant	0.42

Table 1. Statistical Testing of Differences in Obesity Levels

At the 95% confidence level, 10 of 14 features were found to be related to obesity. Top features, such as age and family history of obesity were not lifestyle choices an individual could make. One trend which stood out was the number of main meals consumed per day, since it was counterintuitive that eating less meals per day was correlated with being overweight or obese. Time using technological devices was also not a linear relationship as spending 3 to 5 hours was attributed to a lower frequency of obesity than spending 0 to 2 hours. There were no correlations found among these features (Pearson's Coefficient & Cramer's V).

Preliminary Feature Importances

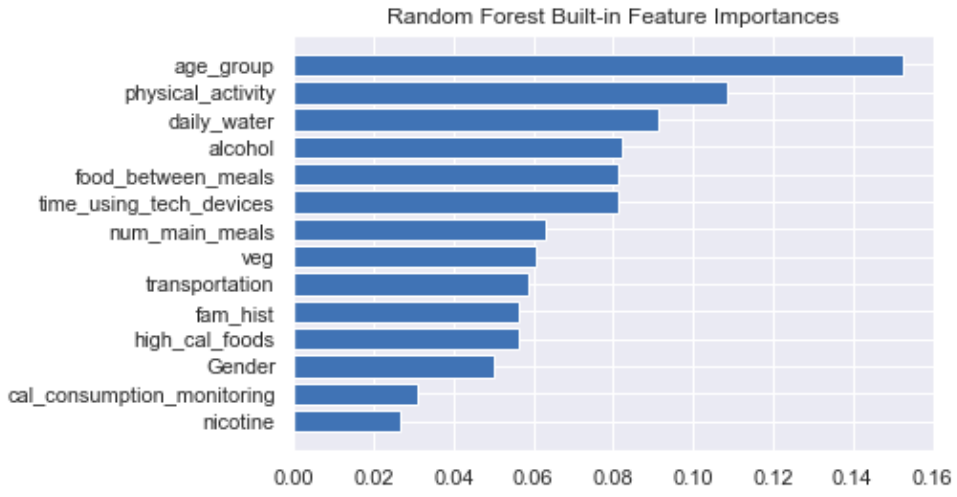


Figure 6. Preliminary Random Forest Build-in Feature Importances

Figure 6 shows built-in, impurity-based feature importances. Potential errors arise from the preference of high cardinality features with numerical values or more unique categories due to over-fitting. The built-in importances are also computed on and are not able to extrapolate values that fall outside of a training set. Here, we see that most features with 4 levels are listed on top and binary features on bottom.

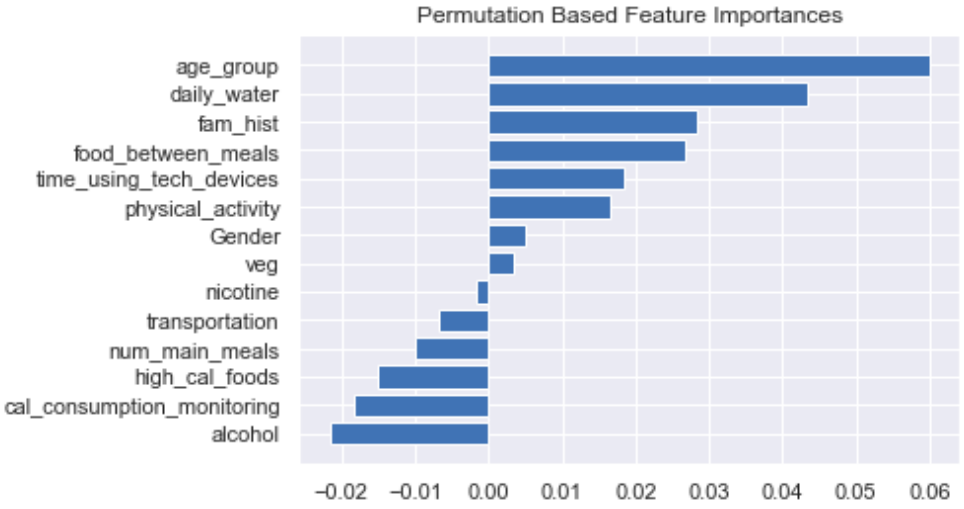


Figure 7. Preliminary Random Forest Permutation (Mean Decrease Accuracy) Based Feature Importances

Permutation importances can be computed on a test set as opposed to the built-in random forest feature importances which is limited to the training set. This method calculates importances by the decrease in the model performance when a single feature is randomly shuffled. We can see more features that have less cardinality being ranked higher, such as

family history, gender, and nicotine. For instance, family history, a dichotomous variable, jumped from rank 10 to rank 3. Negative values are shown for some variables which returned better performance when shuffled. In other words, when the feature is replaced with noise, the model performs better.

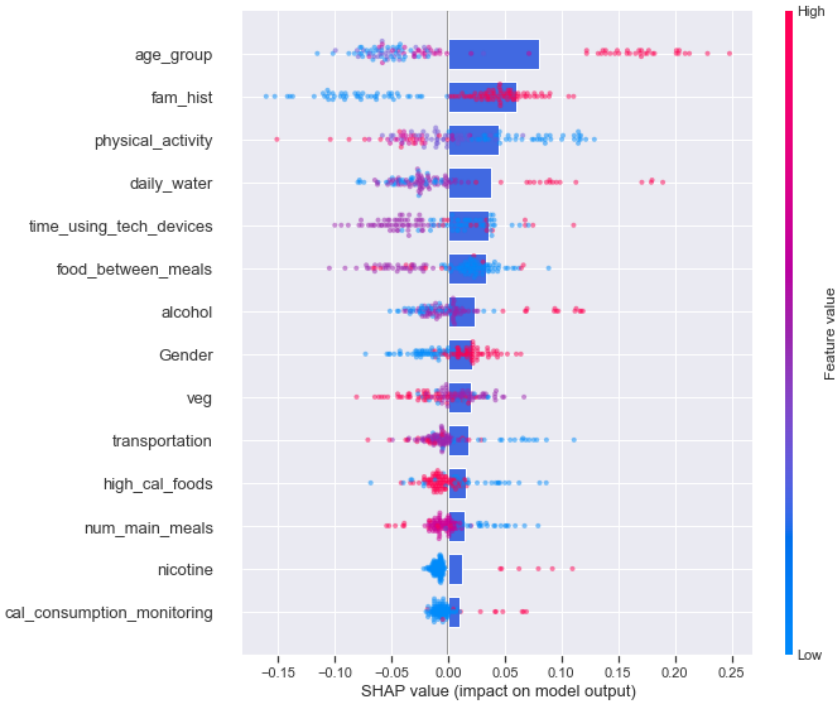


Figure 8. Preliminary SHAP Value Based Feature Importances

SHAP Importances in Figure 8 are based on Game Theory, and hence use Shapley values - the magnitude of marginal feature contributions to the prediction. SHAP is efficient when used in conjunction with tree-based models, and similarly to permutation-based importances, we can calculate importances on the test set. Permutation-based feature importances don't tell us much about how the individual features impact each prediction. SHAP plots show this by colour of the scatter plot, representing the value of each feature from low to high, and by horizontal location, representing the degree of impact of the value on the prediction. For binary variables, a higher value corresponds to 'yes'.

As expected, age for people who are younger doesn't normally affect obesity risk much, but higher age affects obesity risk by a higher magnitude. Having a family history of obesity or lower physical activity leads to a higher chance of obesity. We can also observe non-linear trends through SHAP values. Obesity risk seems to be consistent in people who drink less or moderate amount of water but peaks in individuals who drink a lot of water. Further, those who spend the least time or most time using technological devices seem to positively affect model prediction, and those who spend a moderate amount of time do not.

The SHAP values seems to favour balanced (in terms of the independent feature) features over unbalanced features such as nicotine and calorie consumption monitoring. For instance, there were 32 individuals who responded ‘yes’ to smoking nicotine versus 465 who responded ’no’. Although the pooled t-test proved the BMI means within these groups were different and the distinction between SHAP values according to feature value is clear in Figure 8, there were simply not enough data points for people who responded ‘yes’ for the feature to be deemed important. Hence, the actual importance of these features may differ in reality.

The results in all summary plots showed age group to be the top most feature impacting obesity risk. The remaining top features were shuffled depending on which method was used, with the most consistent features being: age group, family history of obesity, physical activity, and daily water intake. These features were also found to have a statistically significant relationship with BMI at the 95% confidence level.

Modeling

	Best Parameters	Optimal Features	ROC_AUC
Logistic Regression	{‘C’: 0.1}	Age group, family history of obesity, physical activity, daily water intake, vegetable consumption, food consumption between meals, alcohol consumption, nicotine usage	0.7419
Random Forest Classifier	{‘criterion’: ‘entropy’, ‘max_depth’: 3, ‘max_features’: ‘log2’, ’n_estimators’: 50}	Age group, family history of obesity, physical activity, daily water intake, food between meals, vegetable consumption, gender, high calorie food consumption	0.7565
XGB Classifier	{‘colsample_bytree’: 0.1, ‘max_depth’: 1, ’n_estimators’: 50}	Age group, family history, physical activity, gender, daily water intake	0.7512

Table 2. Comparison of Three Classifiers

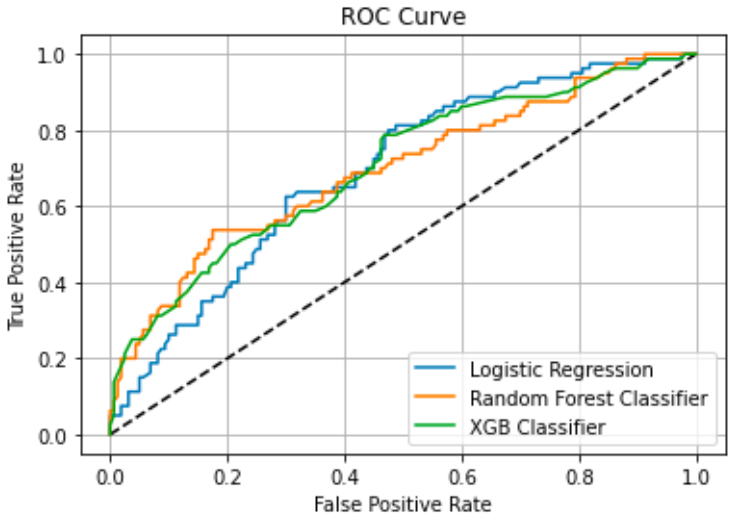


Figure 9. ROC Curves of Three Classifiers

Three models were tested for binary classification between being overweight (overweight & obese classes) and not being overweight (normal & underweight classes), using logistic regression, random forest classifier, and XGB classifier. Despite its name, logistic regression is a linear model for classification, not regression. Random forest on the other hand can capture non-linear and more complex relationships by building trees in parallel. XGB classifier, being a more regularized form of gradient boosting, was also considered to make use of its fast calculation time. Because we only had 169 counts of overweight versus 310 not overweight data, a test size of 50% was used. The random forest classifier was found to be the best model, but the three scores were very close in proximity among the three models (Figure 9). The best hyper-parameters for each classifier listed in Table 2 were tuned using ROC-AUC scoring and 10-fold cross validation on `gridsearch_cv` .

The imbalance in the dataset was accounted for by using the area under the ROC curve rather than the accuracy score as the scoring method, because it divides the accuracy into true and false positive rates. The model could then be chosen based on the respective balance thresholds and importance of the classes. The true positive rate, also referred to as recall, shows us what proportion of the positive class was correctly classified, while the false positive rate shows us the incorrectly classified proportion of the same class. The threshold is then one minus the chosen FPR. Figure 9 shows that the random forest classifier behaves slightly better at a lower false positive rate (a higher threshold), but logistic regression gives a better recall around a threshold of 0.5 and lower.

Feature Importances: Random Forest Classifier

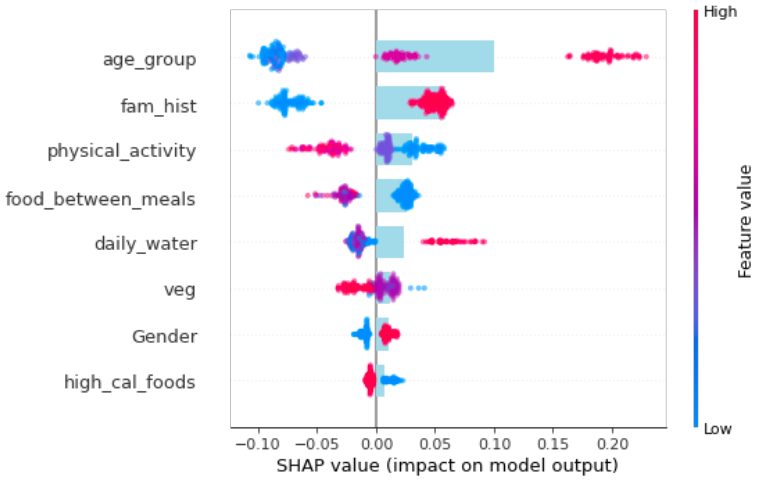


Figure 10. SHAP Value Based Importances for the Tuned Random Forest Model

The resulting Random Classifier model showed clearer SHAP values than the preliminary SHAP graph with all features. Results were consistent with those in Figure 8. Being older, having obese family members, or less physical activity increased obesity risk. It was counter-intuitive that having no food between meals, drinking more water, or having less high calorie foods also increased obesity risk. Perhaps some of these features were an effect of being overweight rather than it causing obesity. Having more vegetables or being female decreased obesity risk.

Thresholding: Random Forest Classifier

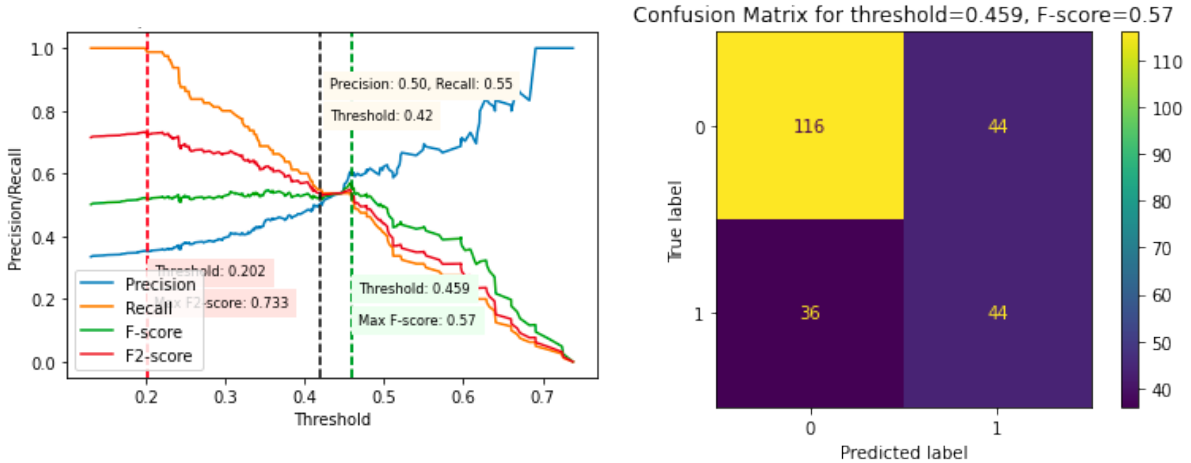


Figure 11. Precision/Recall vs. Threshold & Confusion Matrix for the Selected Threshold (Random Forest Classifier)

For the case of tuning this model so that it can be used by health care practitioners and nutritionists to help obese patients achieve normal weight, we want to minimize false negatives - the false predictions that someone is normal weight when they are actually overweight or obese. We can achieve this by maximizing the true positive rate (recall).

The binary classification labeled 0 as normal weight and 1 as overweight. The Random Forest Classifier mislabeled obese individuals (false negatives) more than those who weren't (false positives). The thresholding chart for the Random Forest Classifier showed that picking the threshold based on the maximum F-score resulted in a recall below 0.5, and picking the threshold based on the maximum F2-score resulted in a precision below 0.5. Picking a threshold of 0.459 allowed us to pick the highest recall of 0.55 without precision going below 0.5, deeming this model not useful.

	Precision	Recall	F1-Score	Support
0	0.76	0.72	0.74	160
1	0.5	0.55	0.52	80
Accuracy			0.67	240
Macro Avg	0.63	0.64	0.63	240
Weighted Avg	0.68	0.67	0.67	240

Table 3. Classification Report for the Selected Threshold (Random Forest Classifier)

Feature Importances: Logistic Regression

Since the ROC curve shows Logistic Regression to perform better at lower thresholds and we are more interested in maximizing recall, we also plotted the feature importances and thresholding charts for Logistic Regression. Figure 12 below shows the permutation importances of Logistic Regression (SHAP does not support regression models- it only uses tree-based models). Top results are very similar to the importances for the tuned Random Forest Model.

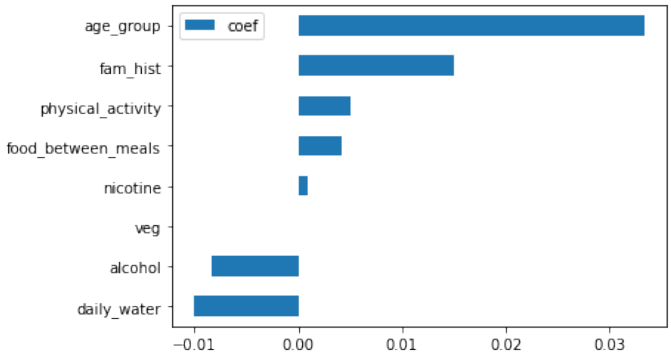


Figure 12. Permutation Based Importances for the Tuned Logistic Regression Model

Thresholding: Logistic Regression

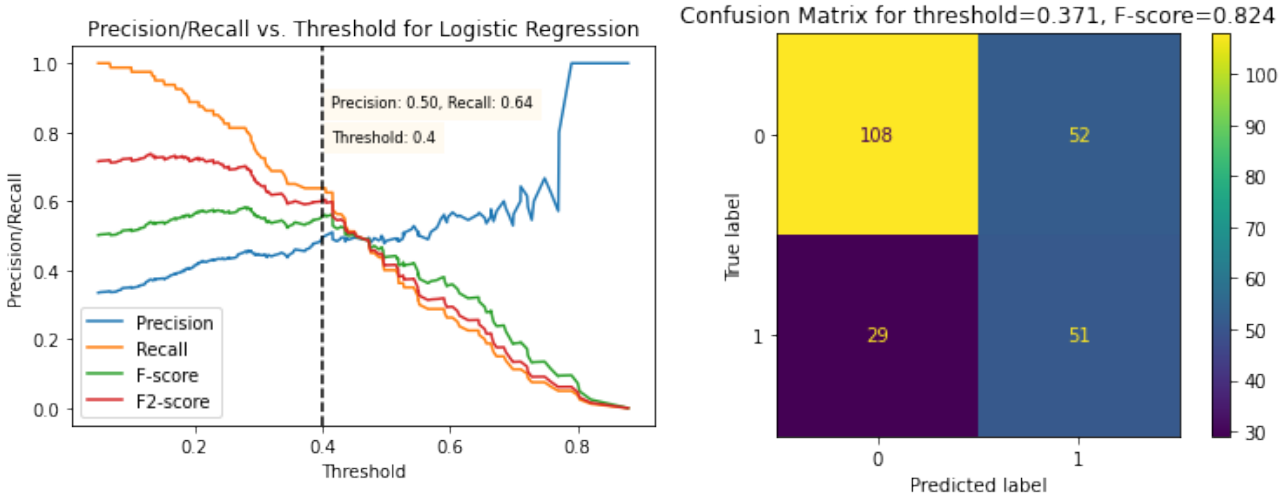


Figure 13. Precision/Recall vs. Threshold & Confusion Matrix for the Selected Threshold (Logistic Regression)

	Precision	Recall	F1-score	Support
0	0.79	0.68	0.73	160
1	0.5	0.64	0.56	80
Accuracy			0.66	240
Macro Avg	0.64	0.66	0.64	240
Weighted Avg	0.69	0.66	0.67	240

Table 4. Classification Report for the Selected Threshold

A threshold of 0.371 allowed us to achieve a recall of 0.64 without going below a precision of 0.5. Although this model performs better than the Random Forest Classifier, these numbers are still too low for this model to be used.

Conclusion

In this project, we attempted to find top lifestyle features which may contribute to obesity risk through survey answers of 479 Latin American individuals, and tune a model to predict whether or not an individual's lifestyle encourages obesity. If the model performed well, it could be used by health care practitioners to recommend healthier lifestyle choices to both obese and non-obese patients whose lifestyle is congruent to studied obese individuals. Obesity counts were

mildly imbalanced (169 overweight vs. 310 underweight). Linear Regression, Random Forest Classifier, and XGB Classifier were tested. The ROC AUC score was used to assess the three models tested, since it does not have the same bias that accuracy scoring has with models that perform well on the minority class at the expense of the majority class. We attempted to account for the small data size by using a larger test size of 0.5. Some feature importances were underestimated due to the imbalance in the individual feature survey counts.

Although the three models were similar in their respective ROC-AUC scores, Logistic Regression performed better under a lower threshold (< 0.5). Random Forest Classifier had the highest score, but when prioritizing recall, did not perform as well as the other two models. However, these differences in the ROC curves were not enough to deem any model useful. Logistic Regression performed best with a recall of 0.64 and precision of 0.5, which are too low for the model to be directly used in a professional setting.

Age group, family history of obesity, physical activity, food between meals, and daily water intake were features that were consistently important in all three models. Out of these features, age group and family history were features that were uncontrollable by the individual. Exercising for even 1 or 2 days per week reduced obesity risk. Having more food between meals surprisingly decreased obesity risk. Number of main meals, proven to have a relationship with obesity through the chi-squares test but not an optimal feature in any of the models, also decreased obesity risk as its numbers rose, although these two features were not correlated, with a Cramer's V coefficient value of only 0.2. More daily water intake increased obesity risk according to the model, but this feature may have been more of a necessary outcome of obesity, as hydration threshold is potentially harder to reach in obese individuals[5].

Future recommendations include balancing important feature survey counts, incorporating clearer habitual indicators which may lead to obesity over time, and tagging the target feature more accurately according to age. Nicotine groups, for instance had a statistically significant mean BMI difference of close to 3 points - the highest out of dichotomous feature groups, but had low feature importance when modeling. The reason may be that only 32 individuals said yes to smoking nicotine versus the 465 who did not. Therefore, the model may have underestimated the importance of this feature. Further, it was unclear whether some of the features such as calorie consumption monitoring and daily water intake were causes or effects of obesity. Direct measures of calorie or nutrition intake may be a more accurate measure of food consumption rather than yes or no answers. Finally, obesity ranges for youth less than 20 are age and gender specific. As a result, 147 individuals of ages 14 to 19 are inaccurately categorized by adult obesity standards. Manual tagging in accordance with the CDC growth chart for childhood obesity is encouraged in future studies.

---

[5] Oakland, M. (2016). Weight Loss and Water Consumption Appear to be Linked. *Time Magazine.* https://time.com/4403276/drink-water-hydration-weight-loss/