Predicting Obesity Risk

...

By the numbers..

64%

of the Canadian adult population were overweight or obese in 2020.

By the numbers..

Being overweight explains

71%

of all treatment costs associated with diabetes in OECD countries.

By the numbers..

$9 Billion

is spent annually on obesity related health costs in Canada.

Stakeholders

Public Sector:

Private Sector:

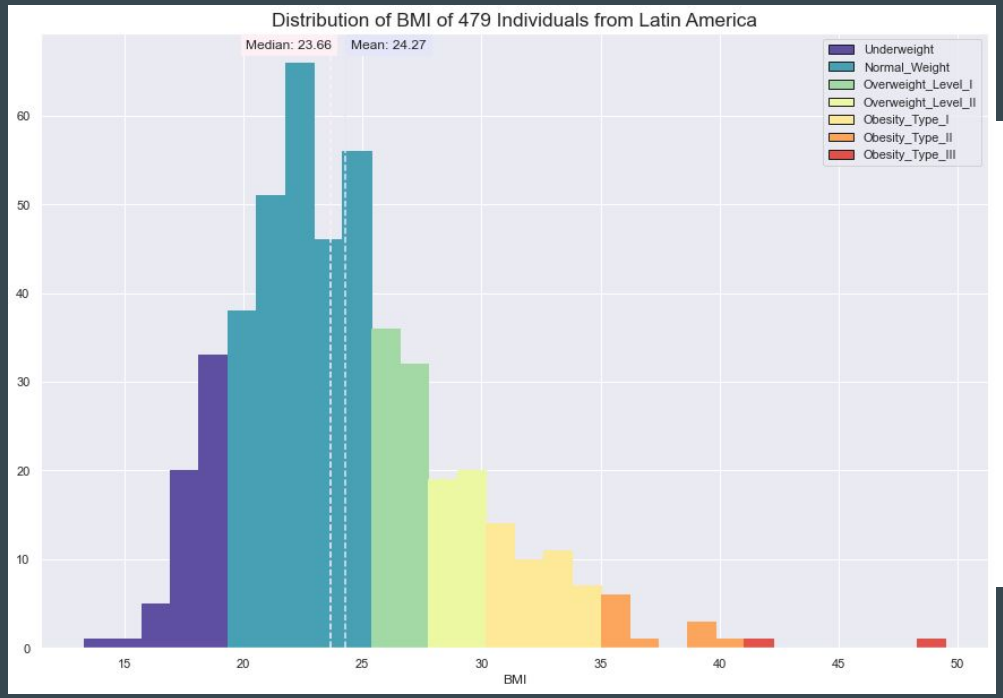
Individuals struggling with overweightness/obesity who want to decrease their risk for obesity

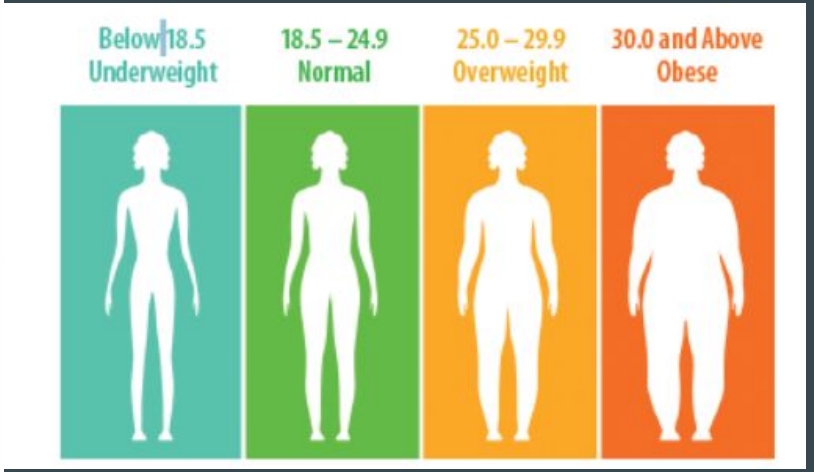
Data

Number of Instances:	2111
Number of Attributes:	17

- Age
- Gender
- Family History of Obesity
- Height & weight
- Vegetable consumption
- Food & water intake
- Alcohol & nicotine usage
- Physical activity
- Transportation methods
- Technology usage
- Calorie consumption monitoring

Data: Obesity Level



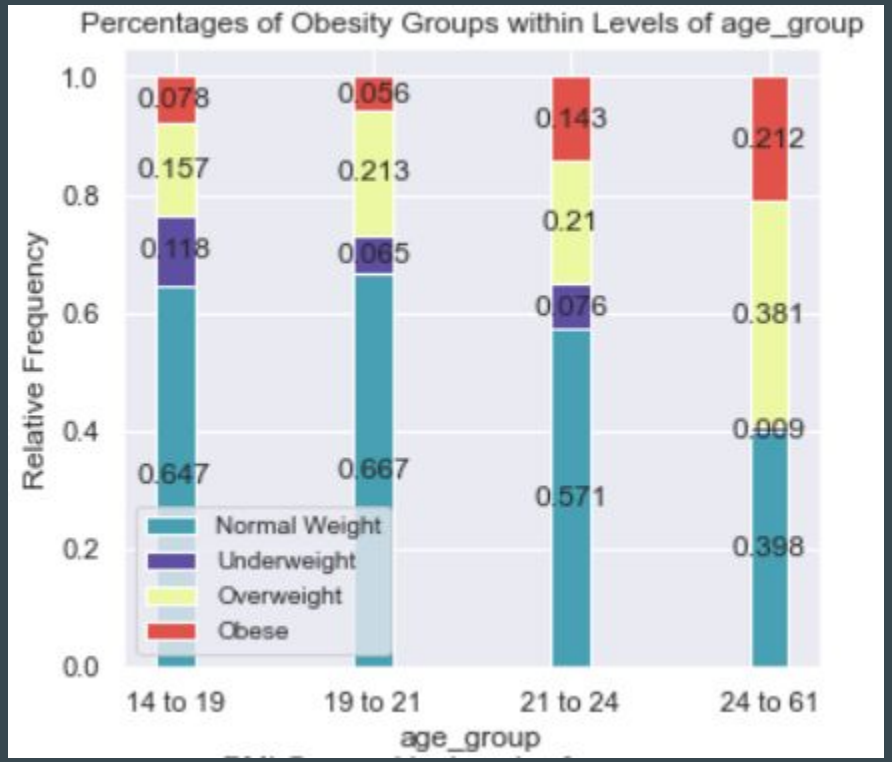


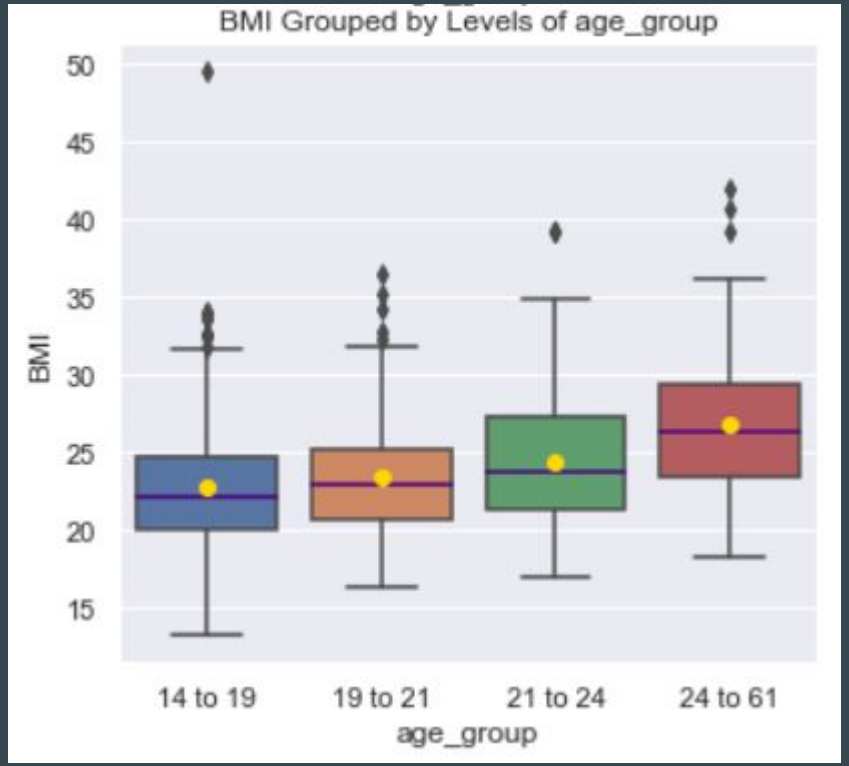
T-Test & Chi-Squares

- Gender (M/F)
- Family history of obesity (Y/N)
- Nicotine use (Y/N)

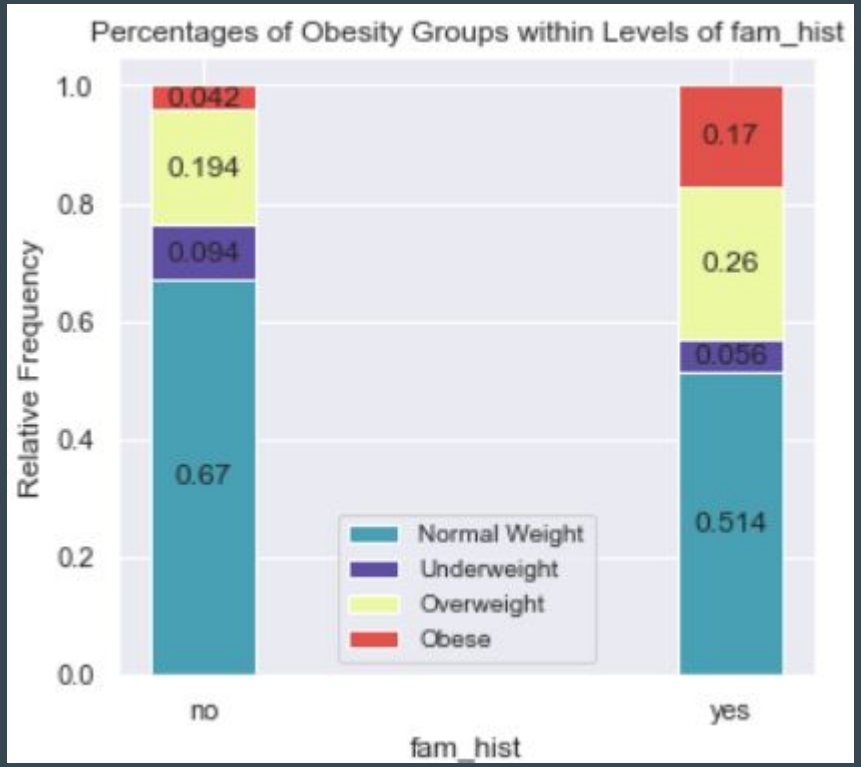
- Age
- Number of main meals
- Daily water consumption
- Physical activity
- Time using technological devices
- Alcohol
- Transportation
- Food between meals (p=0.07)

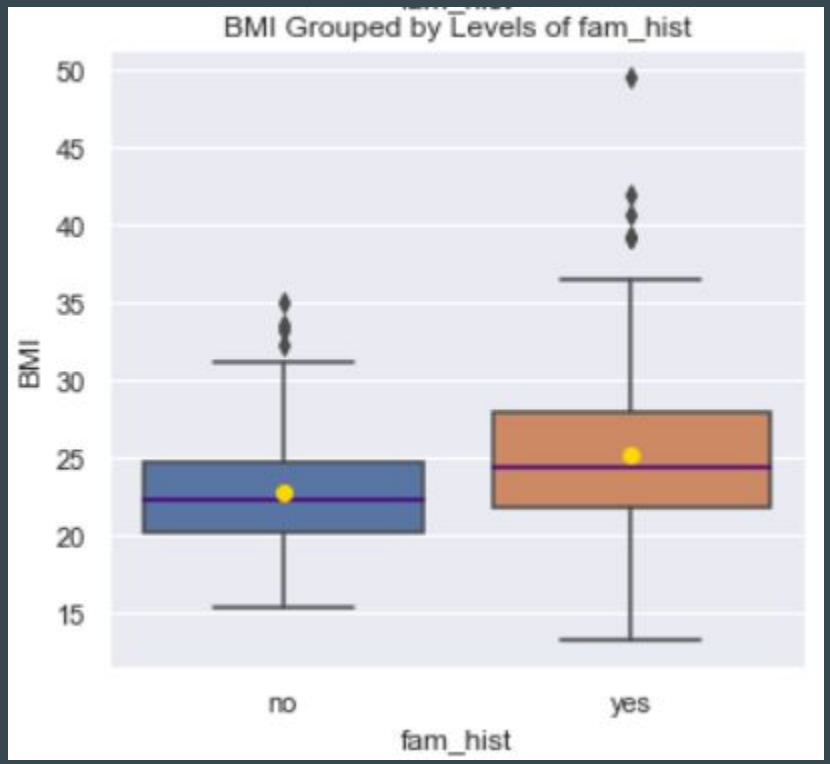
Bivariate Analysis: Age



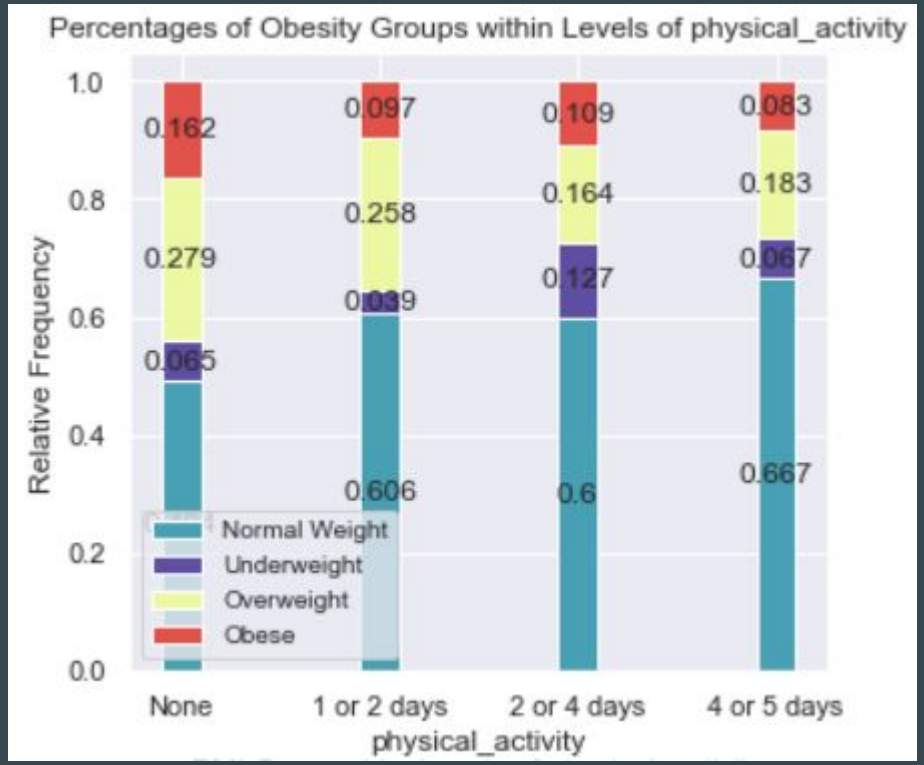


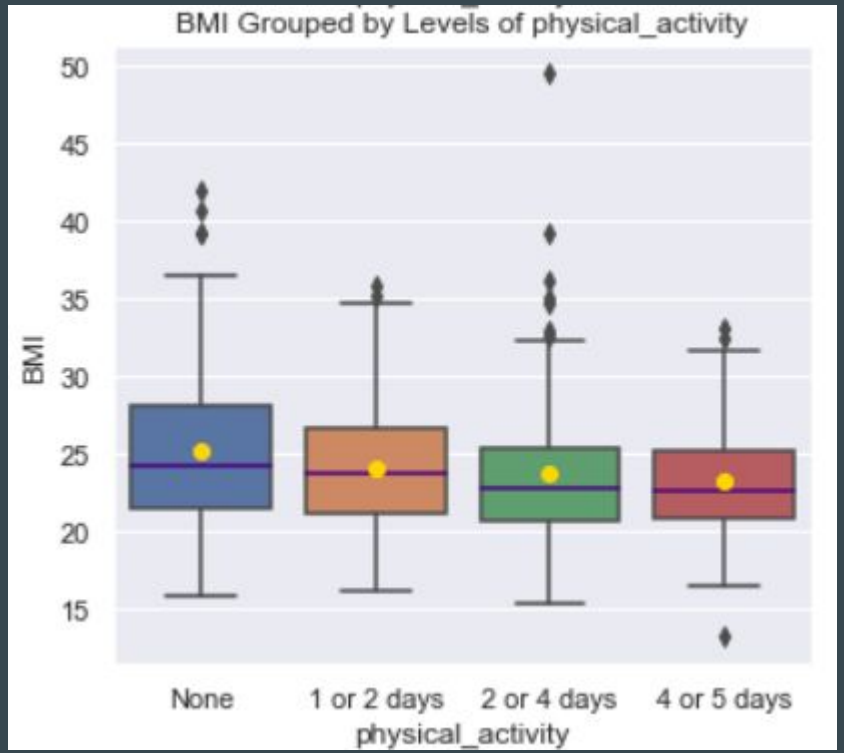
Bivariate Analysis: Family History of Obesity



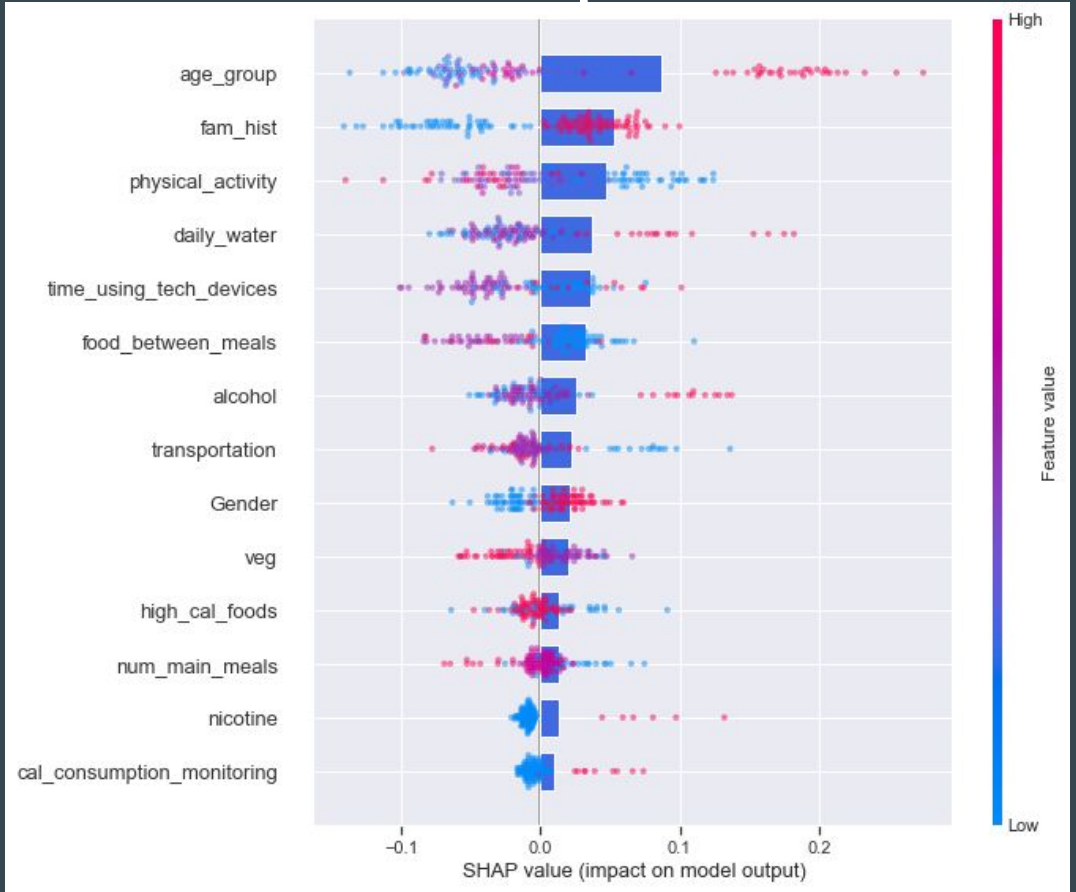


Bivariate Analysis: Physical Activity





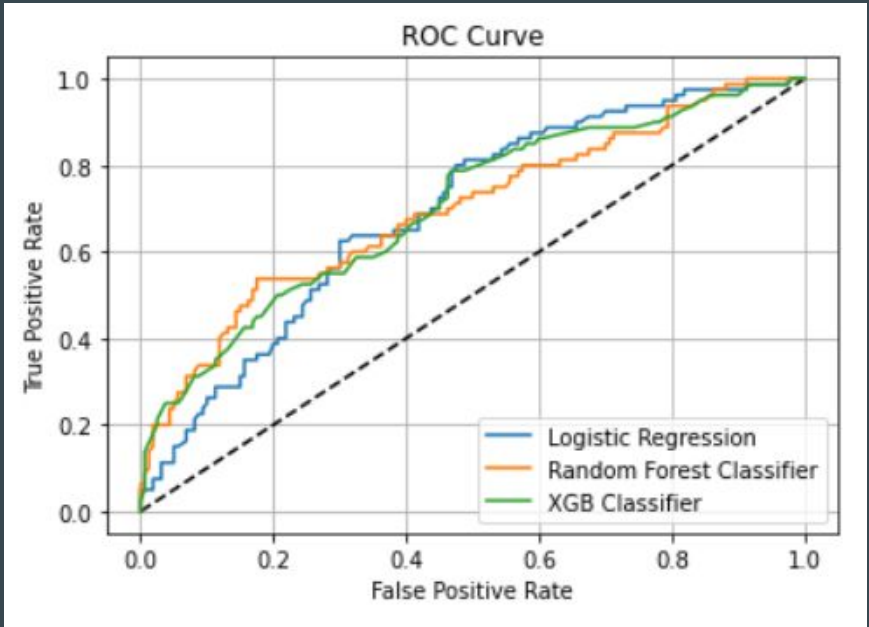
Feature Importances



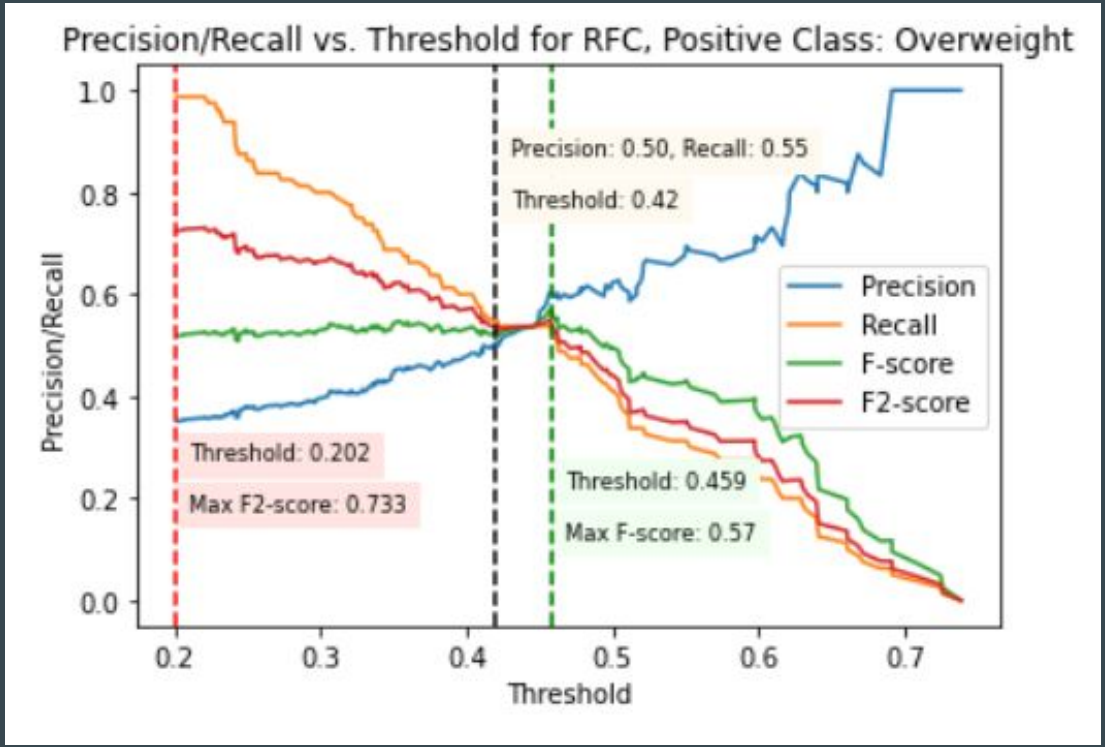
Modeling

	Best_params	Optimal_features	ROC_AUC
Logistic Regression	{'C': 0.1}	[age_group, fam_hist, physical_activity, daily_water, veg, food_between_meals, alcohol, nicotine]	0.741944
Random Forest Classifier	{'criterion': 'entropy', 'max_depth': 3, 'max_features': 'log2', 'n_estimators': 50}	[age_group, fam_hist, physical_activity, daily_water, food_between_meals, veg, Gender, high_cal_foods]	0.756481
XGB Classifier	{'colsample_bytree': 1, 'max_depth': 1, 'n_estimators': 50}	[age_group, fam_hist, physical_activity, Gender, daily_water]	0.750139

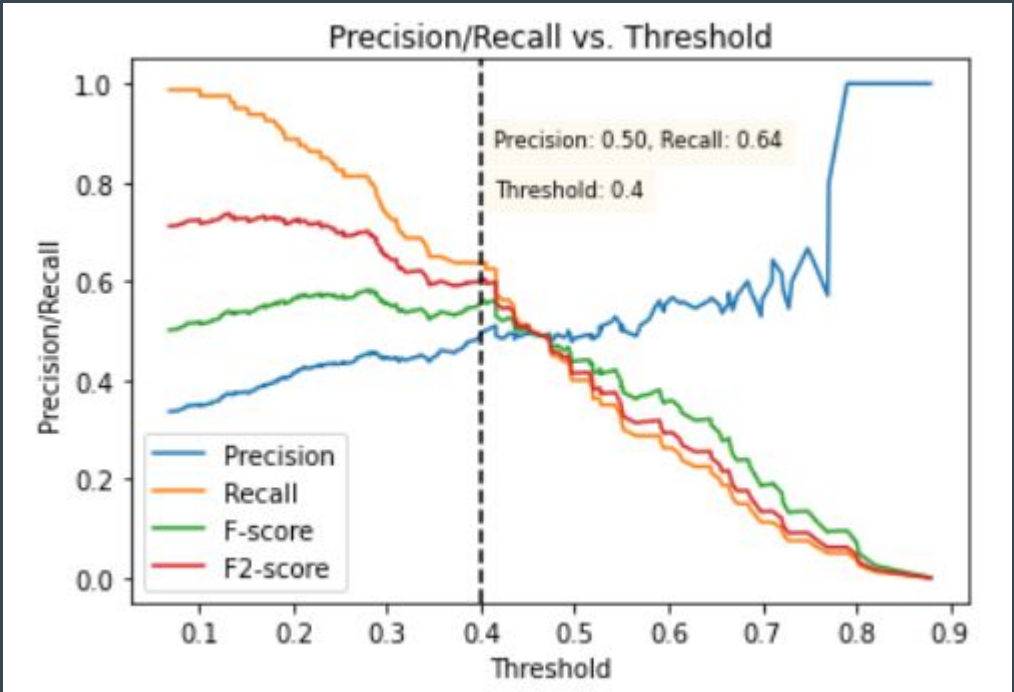
Modeling


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
Random Forest Classifier
XGB Classifier

Thresholding: Random Forest Classification Model



Thresholding: Logistic Regression Model



Conclusions

- Age group, family history, physical activity, daily water intake, food between meals, gender, and vegetable consumption were variables that were consistent among the three models tested.
- The Random Forest Classification and Logistic Regression Models do not suffice to make accurate predictions of obesity.
- We should consider testing other models, or finding new datasets that contain more data or have less of an imbalance between classes.