Obesity Predictions and Health Recommendations

A Ryan, Sam, and Rick Production (DSI - 523)

Table of Contents

01

Introduction

Background and Problem Statement

02

Data Analysis

Exploratory Data Analysis

03

Models

Regression, Classification, Clustering

04

Conclusions

Recommendations and Follow-Ups

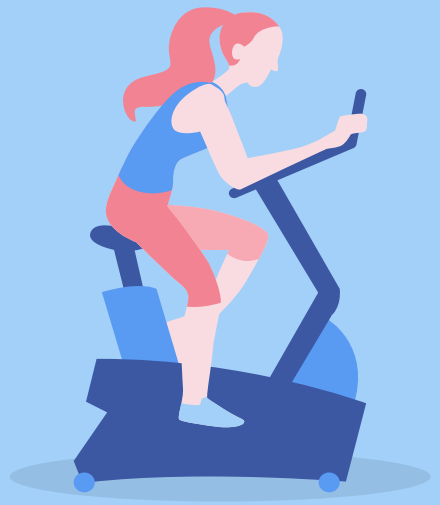
Problem Statement

Obesity is an issue that affects people worldwide. While there are many factors that affect obesity rates, this project will utilize lifestyle survey data to predict if somebody is obese, and give personalized recommendations based on the factors that most contribute to their obesity.

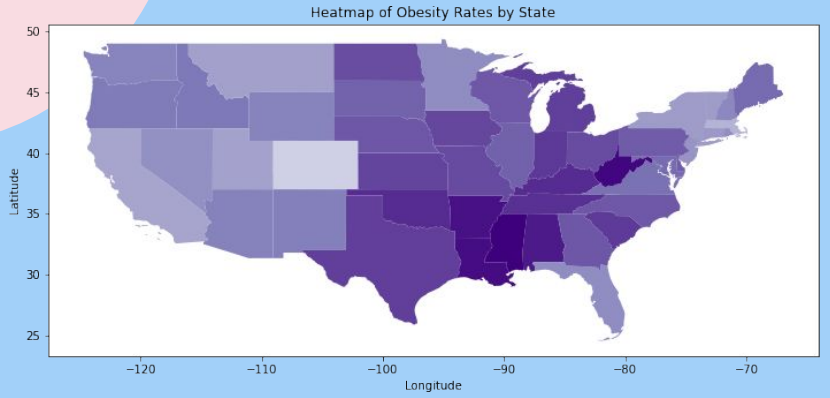
CDC DATA

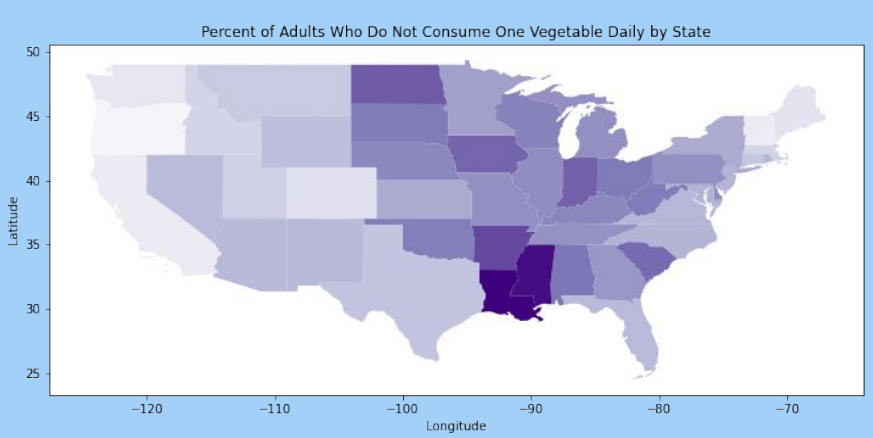
Polling data from each state (2011-2020)

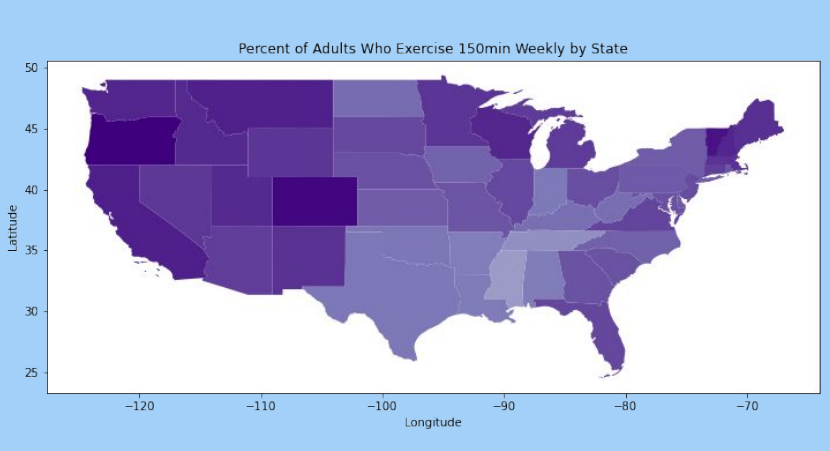
- Obesity rates for different demographics
    - Gender, Race, Income, Education
- Also includes diet and exercise habits

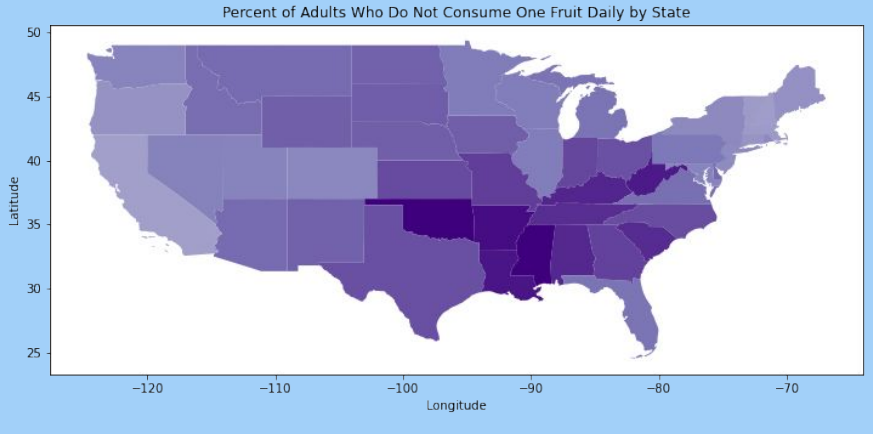


United States Heatmaps



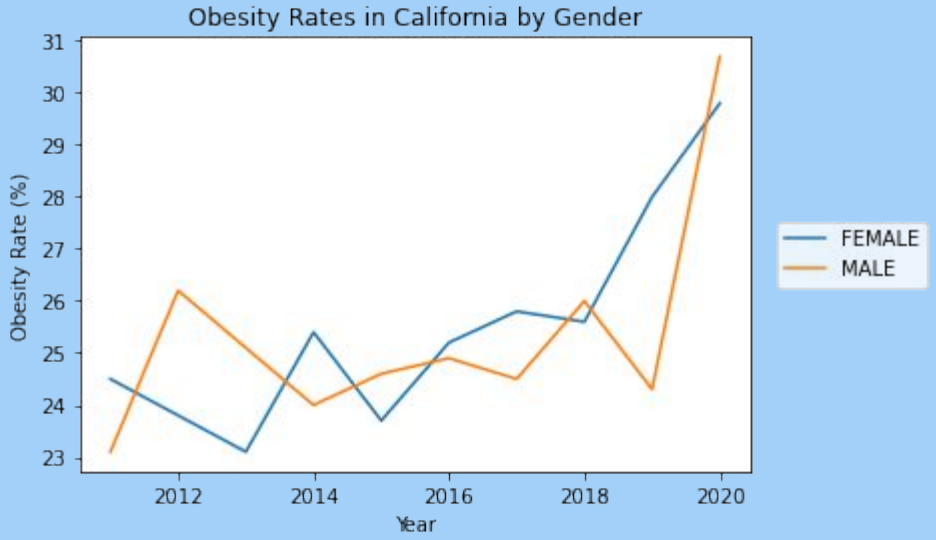


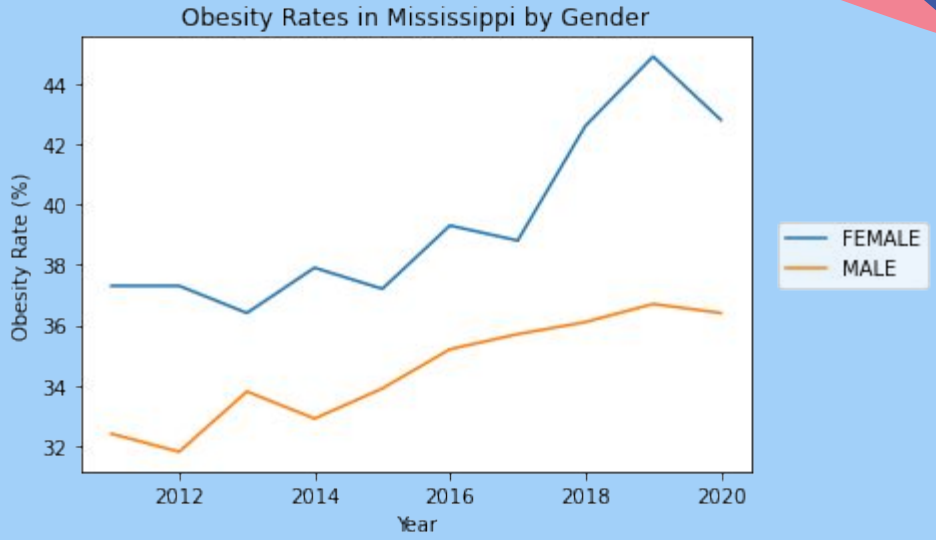




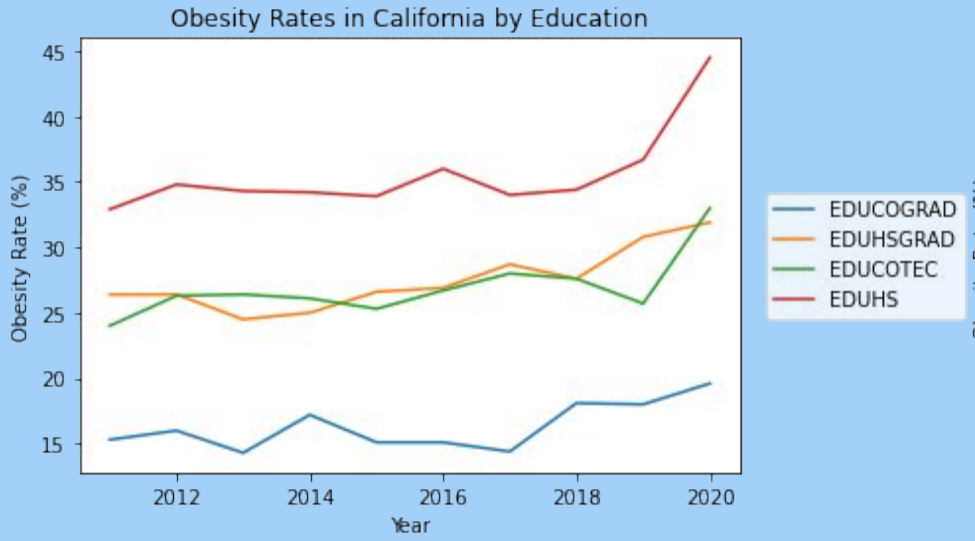
California vs Mississippi

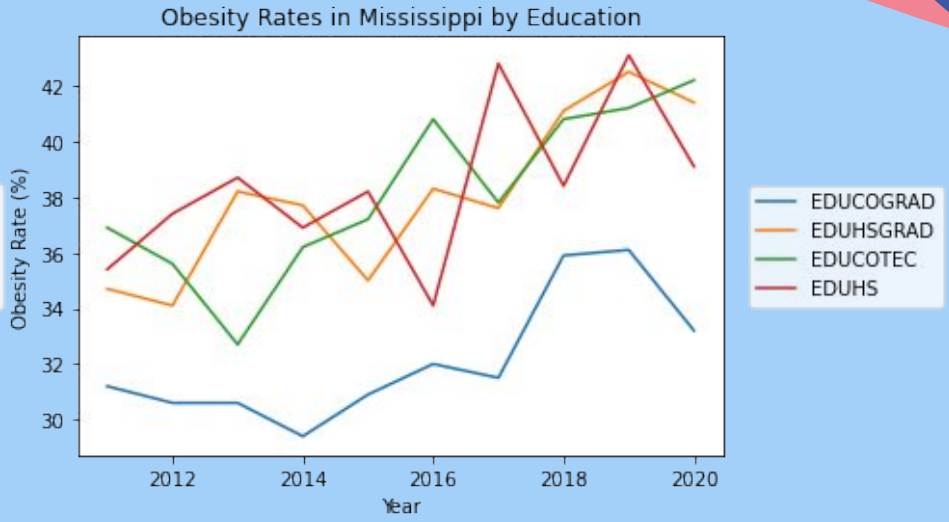
California vs Mississippi



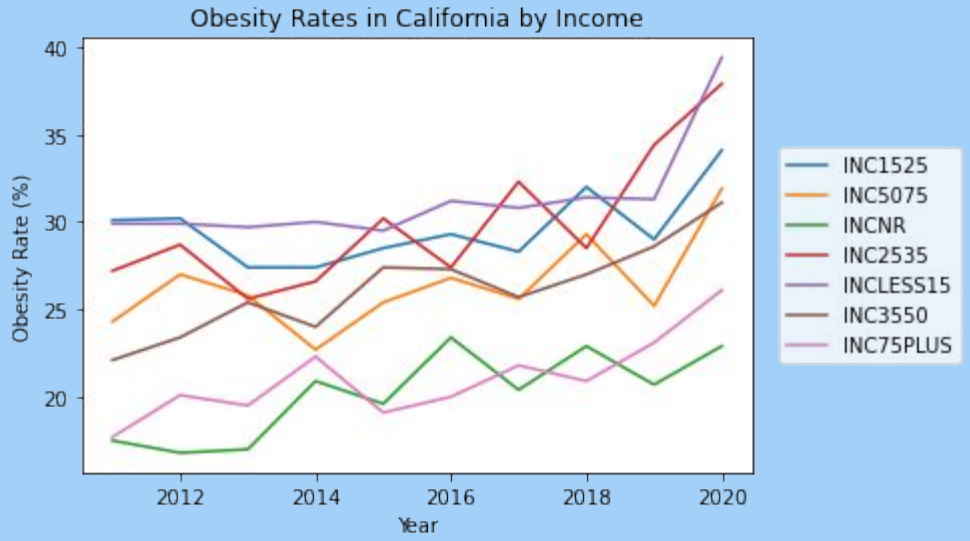


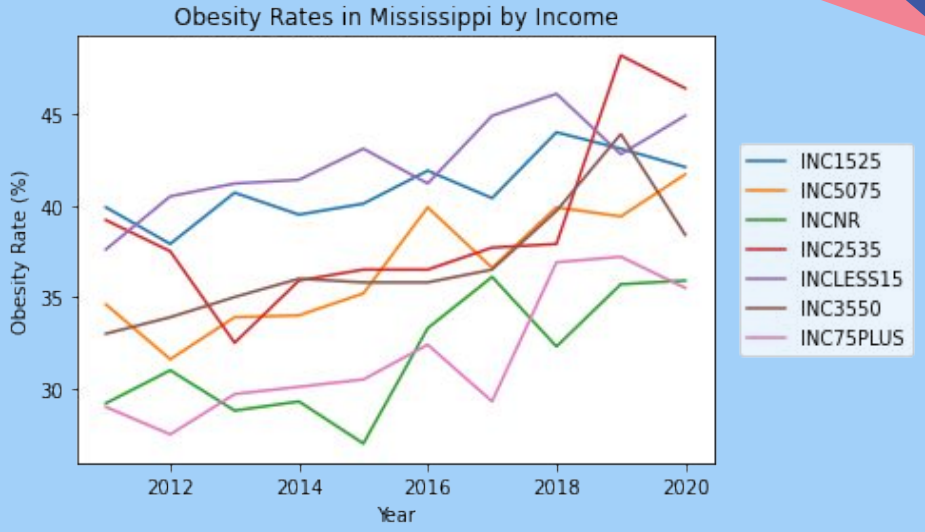
California vs Mississippi



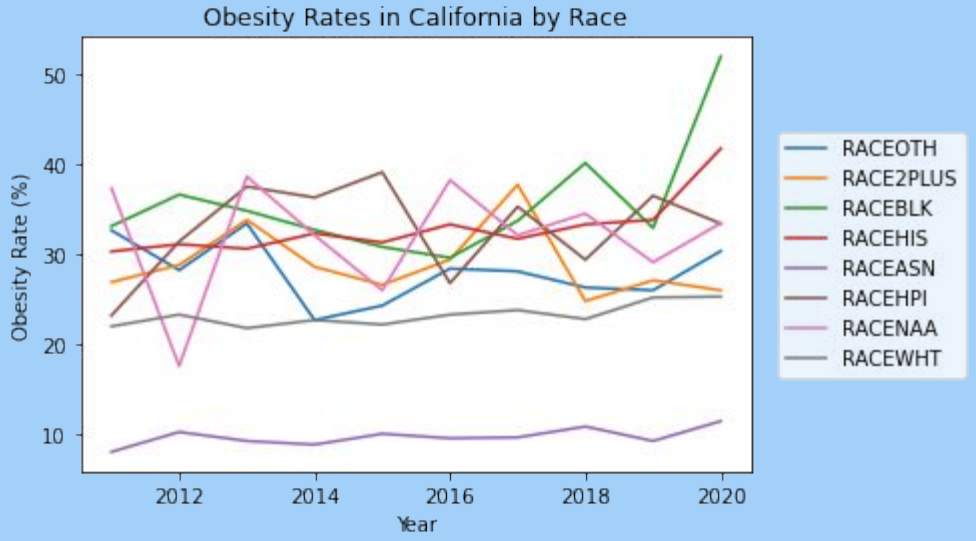


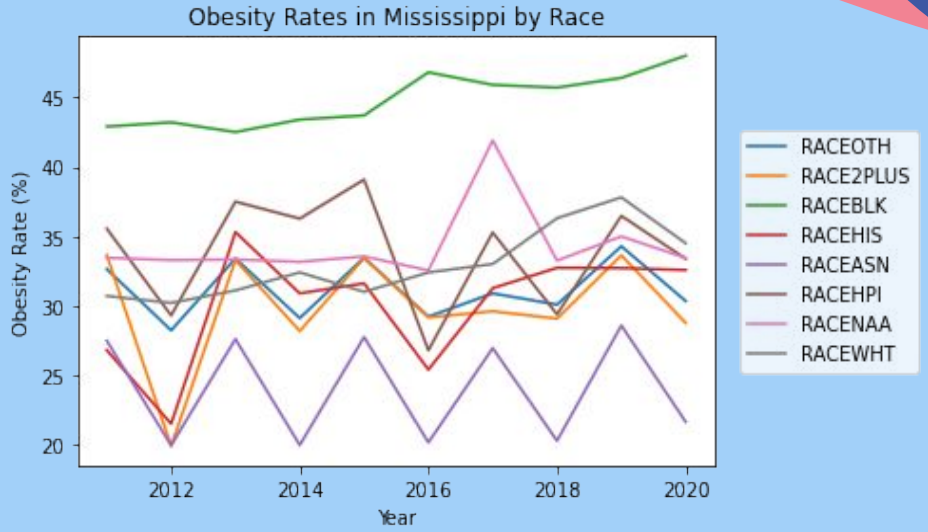
California vs Mississippi





California vs Mississippi



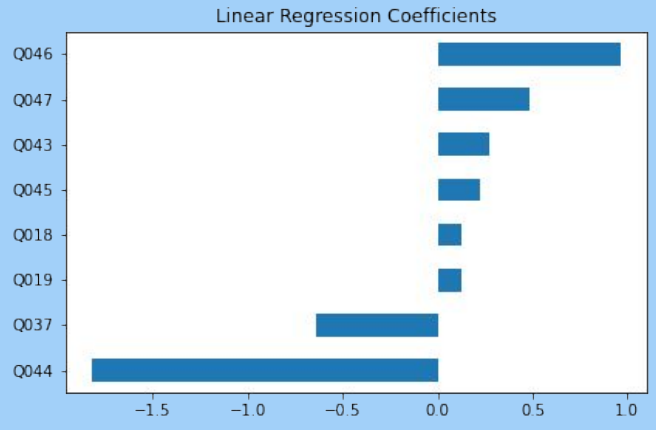


CDC Regression Model

	R2 Score		Mean Squared Error	
Model	Training Score	Testing Score	Testing	Baseline
Linear Regression	.536	.518	7.63	15.86
RandomForest	.936	.576	6.71	15.86
Decision Tree	.995	.176	13.05	15.86
Bagging	.913	.536	7.36	15.86
AdaBoost	.658	.537	7.33	15.86
KNN	.705	.589	6.46	15.86

CDC Regression Model

QuestionID	Question
Q018	Percent of adults who report consuming fruit less than one time daily
Q019	Percent of adults who report consuming vegetables less than one time daily
Q036	Percent of adults aged 18 years and older who have obesity
Q037	Percent of adults aged 18 years and older who have an overweight classification'
Q043	Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)
Q044	Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic physical activity and engage in muscle-strengthening activities on 2 or more days a week
Q045	Percent of adults who achieve at least 300 minutes a week of moderate-intensity aerobic physical activity or 150 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)'
Q046	Percent of adults who engage in muscle-strengthening activities on 2 or more days a week '
Q047	Percent of adults who engage in no leisure-time physical activity



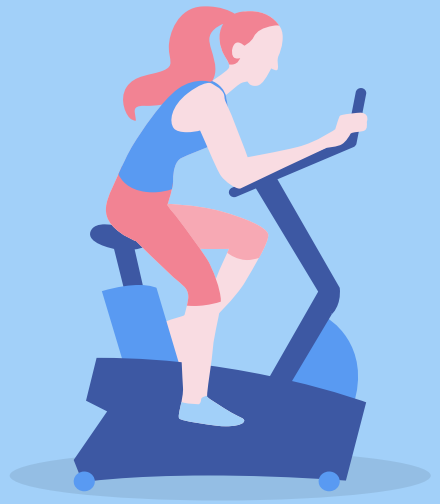
CDC Modeling Limitations

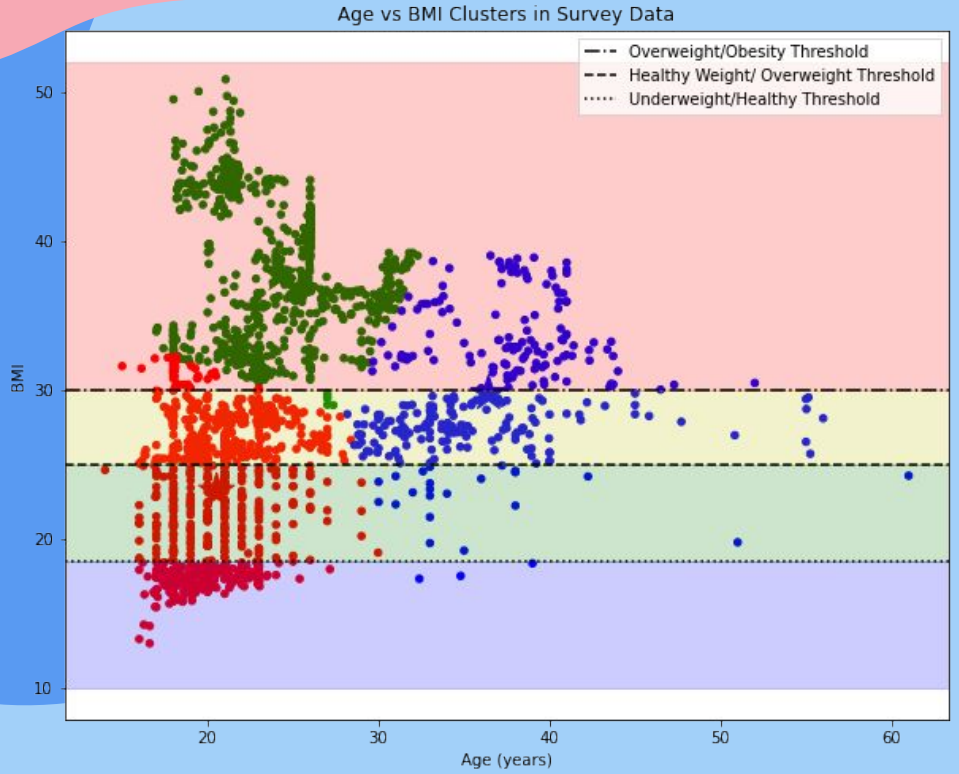
- Data is only on a state by state basis of percentages
- No individual data
- Some of the data has overlap (150 min of physical activity and 300 minutes of physical activity)
- Lots of null values

Survey Questions

Dataset of Survey Results for Obesity in Mexico, Peru and Columbia from 2019 (2111 responses)

¿What is your gender?
¿What is your age?
¿What is your height?
¿What is your weight?
¿Do any of your family members suffer, or have suffered from overweight?
¿Do you eat high caloric food frequently?
¿Do you usually eat vegetables in your meals?
¿How many main meals do you have daily?
¿Do you eat any food between meals?
¿Do you smoke?
¿How much water do you drink daily?
¿Do you monitor the calories you eat daily?
¿How often do you do physical activity?
¿How much time do you use technological devices (cell phone, videogames, tv, computer, etc.)?
¿How often do you drink alcohol?
¿Which form of transportation do you usually use?



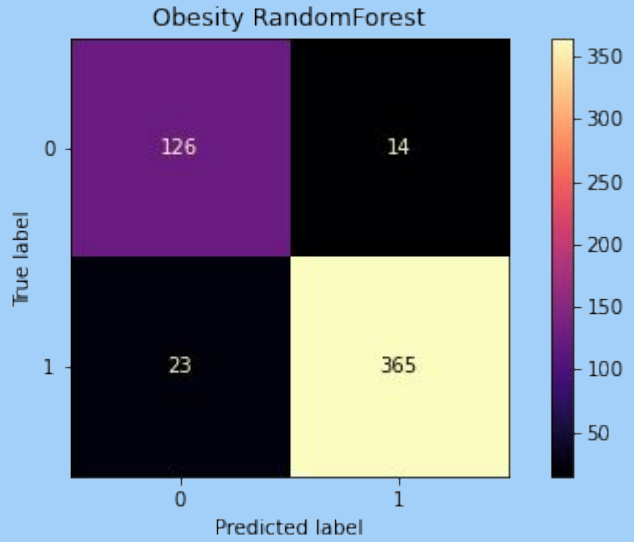


Survey Clustering

- KMeans best clustering method
  - Datapoints too close for DBScan
- As the age of survey participants increased, BMI decreased
- 3 clusters yielded highest silhouette score of 0.494

Classification Model

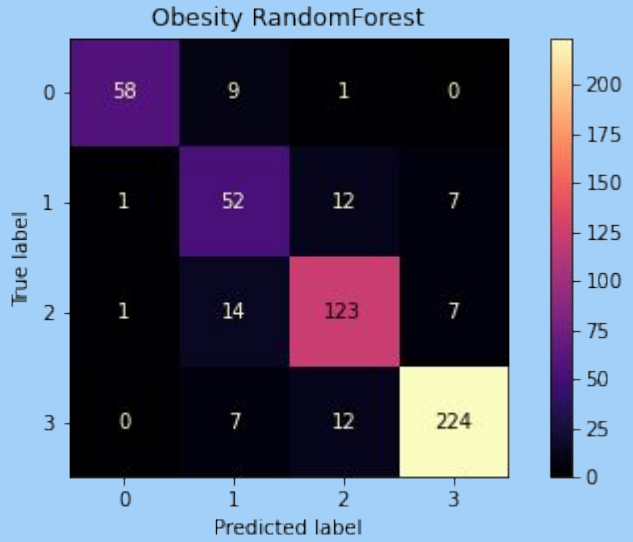
Model	Accuracy	
	Training Score	Testing Score
LogReg	.889	.892
RandomForest	1.0	.930
AdaBoost	.897	.886
GradientBoost	.954	.911



0 = Not Obese, 1 = Overweight or Obese

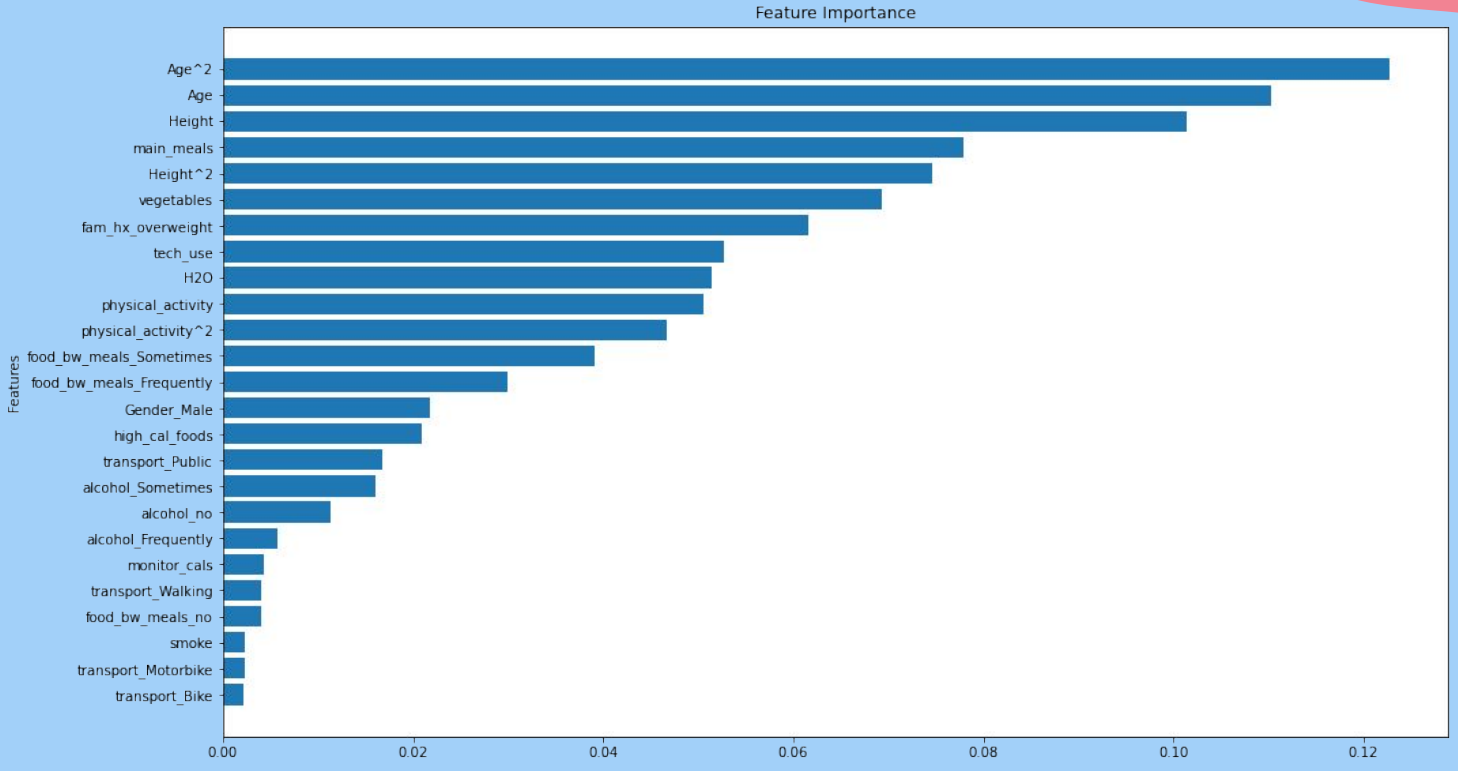
Advanced Classification Model

Model	Accuracy	
	Training Score	Testing Score
LogReg	.663	.655
RandomForest	1.0	.881
AdaBoost	.671	.644
GradientBoost	.944	.837



0 = Underweight, 1 = Normal Weight, 2 = Overweight, 3 = Obese

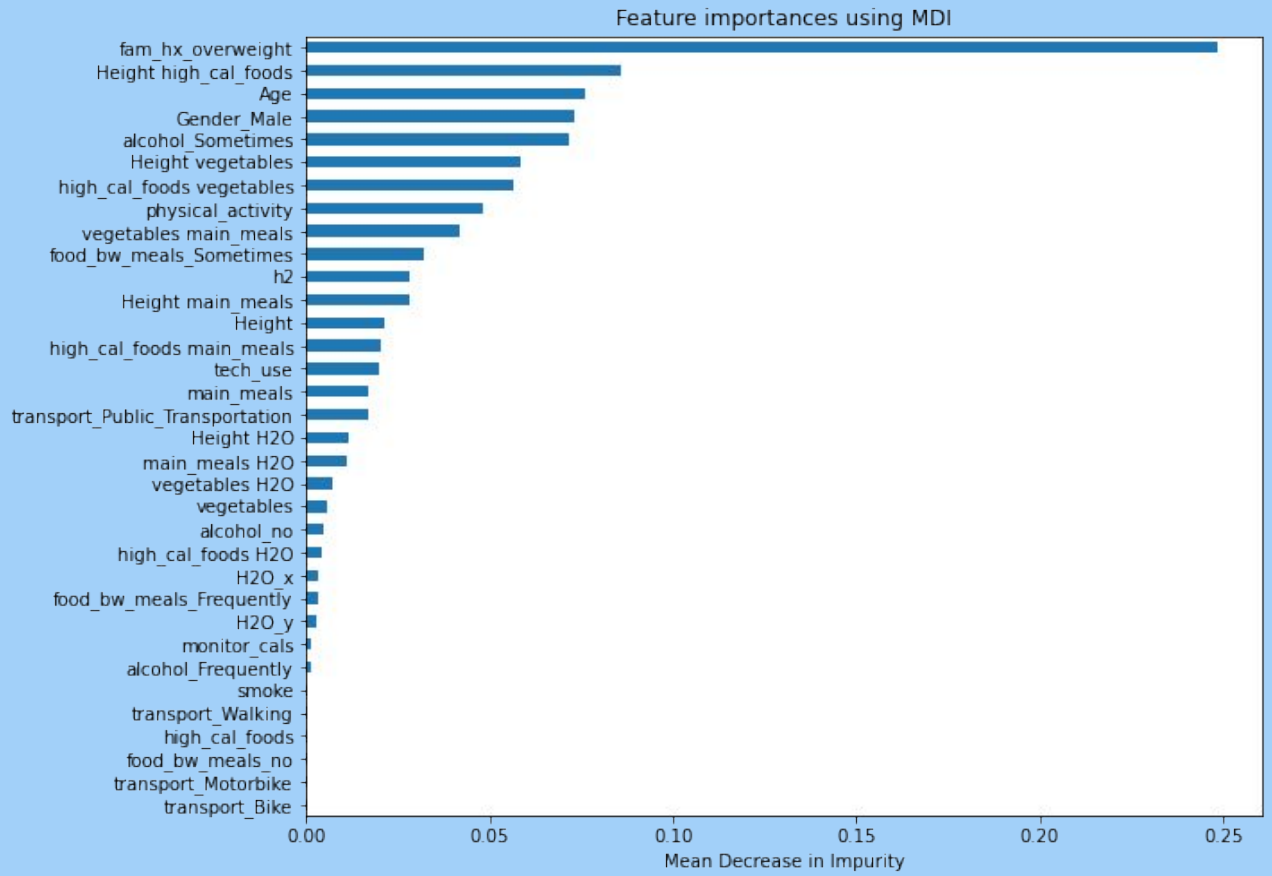
Feature Importances for RF model



Survey Regression Model

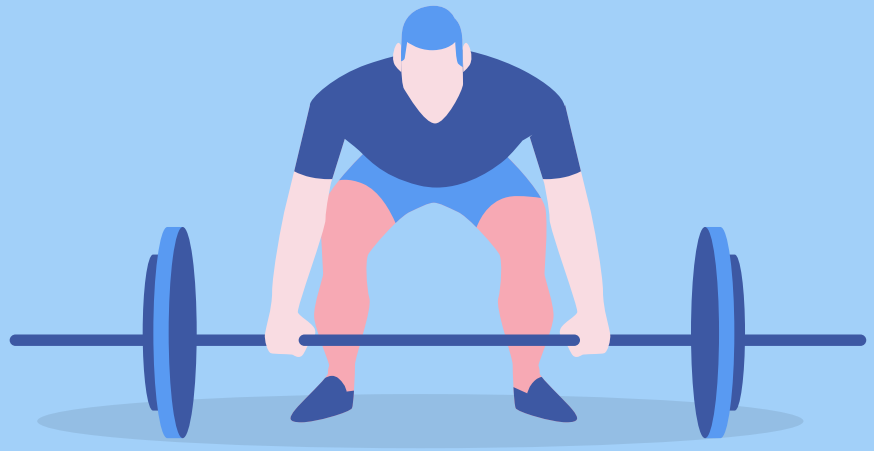
	R2 Score		Mean Squared Error	
Model	Training Score	Testing Score	Testing	Baseline
Linear Regression	.591	..580	297.2	707.7
RandomForest	.987	.862	97.4	707.7
Decision Tree	1.0	.759	170.6	707.7
Bagging	.981	.846	108.9	707.7
AdaBoost	.738	.686	222.46	707.7
KNN	.898	.812	132.9	707.7

Feature Importances



Recommendations

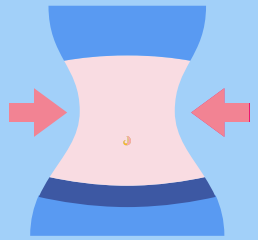
- Diet and exercise is first and foremost
- Reducing alcohol consumption
- Decrease use of technology



Conclusions

Conclusion

Daily Exercise and a healthy diet are common ways to reduce obesity, but simple lifestyle changes such as biking or walking to work (if possible), or monitoring the calories you eat can help manage your weight.



Next Steps

We could improve our results with more data points, our survey had over 2000 responses, but our models and predictions would have been better with more data.
Additionally, our CDC data was missing many values and was all percent based

Questions?