**1. Project Title and Team**

**Project Title:** Policy recommendation for preventing obesity, a potential high-risk predictor for coronavirus disease.

**Team members:** Tetsuo Fujino (tfujino), Jinyoung Hur (jinahur), Takayuki Kitamura (kitamura), Sarah Woo (sarahwoo)

**2. Executive Summary:**

The main goal of our project was to provide appropriate policy recommendations that would not only help cope with the current pandemic of Coronavirus disease, but also prevent other obesity-related issues in the long-term. We obtained U.S. county-level data within the past 10 years from various government and open data sources, and applied 11 machine learning models (6 regression and 5 classification models) to predict obesity and coronavirus deaths rates for a county in the U.S. using various demographics, health and lifestyle related features.

The results of our prediction show that classification models performed better than regression models at predicting obesity rates in general, and that Gradient Boosting model produced the highest accuracy score of 85.2% at predicting obesity rates with relatively high precision and recall scores of 71.2% and 50.4%, respectively. What is interesting about our prediction results is that the level of education turned out to be the most important predictor for obesity rates among some commonly observed demographic, health and lifestyle features. Some of the features we initially thought would be good predictors of obesity, such as the number of recreational and fitness facilities per 1,000 of population and tax rates on soda and snacks, turned out to be insignificant features for predicting obesity rates. Therefore, we were able to draw a conclusion that improving population's highest education level with targeted education policies would be the most effective strategy for reducing obesity rates within a county. When we focus on improving the percentage of high school diploma holders, we have to take care of specific races such as American Indian/Alaska Native, Hispanic, and Black([source](https://nces.ed.gov/fastfacts/display.asp?id=16)). Thus, we recommend that local and state governments should create or expand grants for minorities who go to high school, which gives them incentives to keep studying. We also suggest that local and state governments give grants to charter schools that teach minorities. These policies reduce the obesity rate in the county indirectly.

When we further applied our prediction model to predict coronavirus death rates, we found surprising, unexpected results that coronavirus death rate prediction was improved when obesity rate was excluded from the set of features. The model for predicting coronavirus death rate is far from perfect yet, with accuracy score of 54% and precision and recall around 45%. However, we found that the single most important predictor for coronavirus death rate is the percentage of black population in a county. Therefore, whereas we had initially expected obesity rates to be a strong predictor for Coronavirus death rates and had targeted to come up with a meaningful policy recommendation to reduce Coronavirus death rates through reducing obesity rates, the results of our project suggest that policy interventions for reducing obesity rates and combating Coronavirus disease should be viewed separately. For reducing Coronavirus death rates, supports provided to areas with high percentages of Black population seems to be mostly in need.

**3. Background and Overview of Solution:**

We read about several news articles and medical research suggesting a possible connection between BMI levels and Coronavirus disease cases, especially for people under the age of 60 and those who need to be ventilated (Fallik). Knowing that obesity is one of the major factors for numerous diseases, specifically for increased risks of cancer, coronary artery disease, type II diabetes, and stroke, we wanted to come up with a meaningful policy recommendation to reduce Coronavirus death rates through policy interventions targeted at reducing obesity rates. According to a report from the Centers for Disease Control and Prevention (CDC), the age-adjusted prevalence of obesity rates among U.S. adults was 42.4% in between 2017 and 2018, and it has been a contributing factor of a lot of deaths and has costed a huge amount of money in the U.S. Hence, we used machine learning for predicting obesity rates in a county using 21 features of demographic, health and lifestyle related attributes (e.g., percentage of population with the highest education level of Bachelor’s degree or higher, median income figures, racial compositions in a county, number of recreational and fitness facilities per 1,000 of population, and tax rates on soda and snacks, etc.) and also applied our models to predict Coronavirus death rates under two different scenarios, one with using the same set of 21 features used for predicting obesity rates and another with using the same set of 21 features plus obesity rate added as a feature (making it a set of 22 features in total).

As our prediction results indicate that the highest level of education is the best predictor for obesity rates, we suggest local and state governments to check changes of percentages of people with high education to predict the future obesity rates. Our findings suggest local governments to invest more in education by grants for minorities and charter schools where they go.

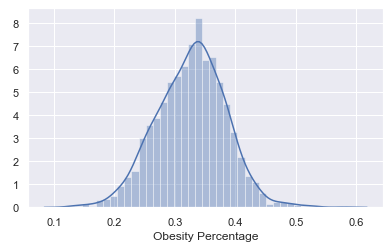
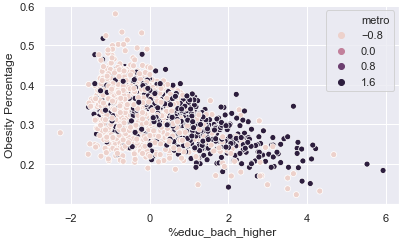
**4. Data:**

We have obtained our data from multiple sources, including the percentage of population who are obese at each county level from The Center for Disease Control, data on coronavirus infection and death rates per county from GitHub updated by The New York Times, and the relevant demographics, health and lifestyle features such as education, income, access to workout facilities and grocery stores from the U.S. Department of Agriculture. The following is a more detailed breakout of each dataset we used.

* Features:
  + Education and income: obtained data on percentage of population with highest educational achievements broken down into categories for a 5-year average of 2014-2018 from the Department of Agriculture’s Economic Research Service. Also obtained a separate dataset on 2018 median household income (<https://www.ers.usda.gov/data-products/county-level-data-sets/>).
  + Lifestyle and locational features: obtained other relevant health, lifestyle and locational features, such as number of recreational facilities in a county, proximity to grocery stores, availability of direct sales of produce from farmers, and access to farmers’ markets from years after 2010 from the Department of Agriculture website (<https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/>).
* Label (obesity rates): obtained 2016 data on obesity percentage (percentage of adult population aged 20 or older who report BMI greater than or equal to 30) at county level provided by CDC (<https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html>). BMI is defined as a person’s weight in kilograms divided by the square of height in meters. An average adult male and female in the U.S. have BMIs of 26.6 and 26.5, respectively (Centers for Disease Control and Prevention). BMI levels ranging from 30 to 35 are classified as Class 1 obesity, from 35 to 40 as Class 2 obesity, and at or above 40 as Class 3 obesity, sometimes referred to as ‘extreme obesity’ or ‘severe obesity.’
* Label (Coronavirus disease): obtained data on daily cumulative counts of coronavirus disease cases and deaths at county level as of June 1st, 2020. This is a public dataset updated by New York Times on a daily basis pulled from GitHub (<https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv>).

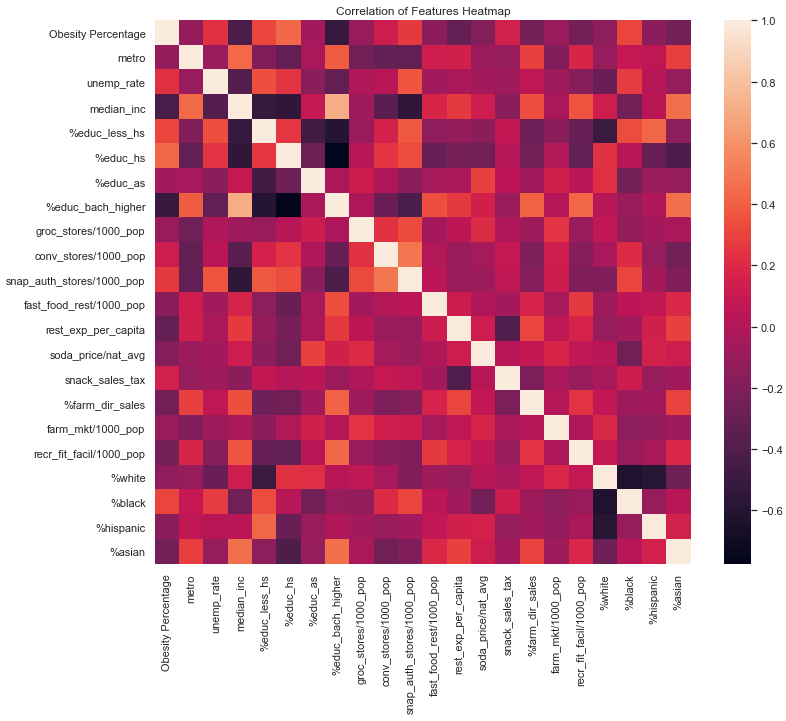
Since all of our datasets were initially obtained at the county level from various sources, we combined our data into a single data frame without losing any observations using county FIPS code and pulled 21 features that we thought were relevant to predicting obesity and coronavirus rates. There were at least 3,000 observations in each dataset and these various datasets with large number of observations allowed our team to produce accurate predictions.

After merging all relevant data, we split the dataset into training and testing data, and applied general data cleaning steps to normalize the features and impute missing or negative values. We also ran some sanity checks on the data to make sure that the ranges of our features made sense and plotted multiple visualizations to better understand the features. Each row in the final dataset contained 1 label (either obesity rate or Coronavirus death rate for a county) and 21 features: whether or not the county is located in a metropolitan area ('metro'), unemployment rate for a county ('unemp\_rate'), median income for a county ('median\_inc'), number of grocery stores per 1,000 of population ('groc\_stores/1000\_pop'), number of convenience stores per 1,000 of population ('conv\_stores/1000\_pop'), number of SNAP authorized stores per 1,000 of population ('snap\_auth\_stores/1000\_pop'), number of fast food restaurants per 1,000 of population ('fast\_food\_rest/1000\_pop'), restaurant expenses per capita ('rest\_exp\_per\_capita'), price of soda compared to national average ('soda\_price/nat\_avg'), dollar amount of taxes on snack sales ('snack\_sales\_tax'), percentage of direct sales from farmers ('%farm\_dir\_sales'), number of fast farmers’ markets per 1,000 of population ('farm\_mkt/1000\_pop'), number of recreational and fitness facilities per 1,000 of population ('recr\_fit\_facil/1000\_pop'), percentage of population whose highest education level is high school or less ('%educ\_less\_hs'), percentage of population whose highest education level is high school ('%educ\_hs'), percentage of population whose highest education level is Associate’s degree ('%educ\_as'), percentage of population whose highest education level is Bachelor’s degree ('%educ\_bach\_higher'), percentage of White population in a county ('%white'), percentage of Black population in a county ('%black'), percentage of Hispanic population in a county ('%hispanic'), and percentage of Asian population in a county ('%asian').

The following are some examples of visualizations we’ve used in our data analysis.

First, on the left above is distribution of our target label, which is the percentage of obese population in each county. From this, we were able to infer that most counties had obesity rates around 0.3 and 0.4, and this also helped us categorize our target label into three different buckets, where we’ve labeled the counties into low, medium, and high. The second graph on the right represents the relationship between our target label and one of the education features, which is the percentage of population with a bachelor’s degree or higher. From this, we inferred that there seems to exist a negative correlation between our target label and high education level.

The visualization below is a correlation matrix of features and label, where darker colors represent higher correlation. From this, we were able to identify education and median income as some of our highly correlated features to the label.



**5. Machine Learning Methods & Details of Solution:**

After processing and exploring data, we ran three different categories of models at a high level. We first ran a simple OLS, a non-regularized linear regression model between features and obesity rates to serve as a basis for comparison, so that we could compare how well our machine learning models with hyperparameters are performing compared to non-regularized model. Then we experimented with both regression and classification models. We used 6 different regression models listed below to predict obesity rates. We also used the categorized obesity rates to run 5 different classification models listed below. By running both regression and classification models, we wanted to compare their results and select a model that best serves our purpose of predicting the labels and producing meaningful policy recommendations.

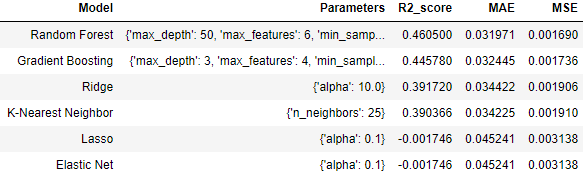
Regression Models:

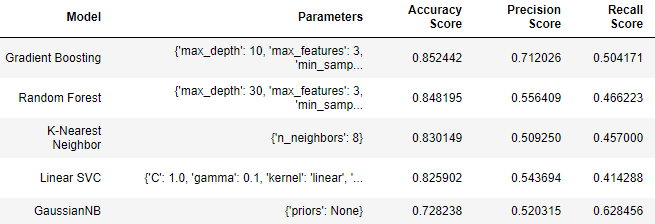
* Ridge
* Lasso
* Elastic Net
* K-Nearest Neighbor
* Random Forest
* Gradient Boosting

Classification Models:

* Gaussian Naïve Bayes
* Linear Support-Vector
* K-Nearest Neighbor
* Random Forest
* Gradient Boosting

We used various parameters and 5-fold cross-validation on regression models, where the result of our models are summarized in a table below. Our best performing regression model was Random Forest in terms of R-squared score of 0.46, and the top predicting features for obesity rates using this best regression model were percentage of population with Bachelor’s degree or higher and percentage of population with high school degree.



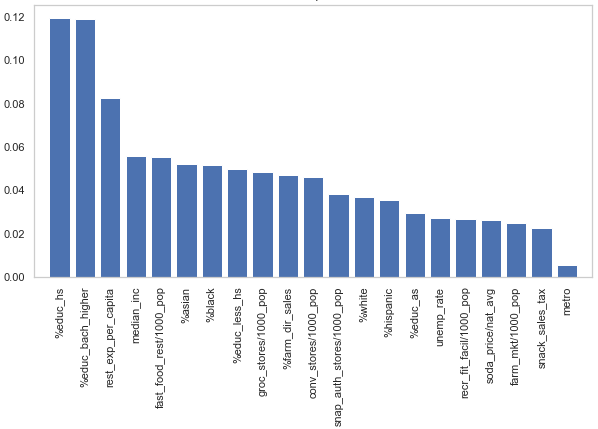
We also explored with different parameters and 5-fold cross-validation on classification models. In this result, Gradient Boosting model with the highest accuracy score of 0.85 also had the highest precision of 0.71 with relatively high recall of 0.50. The results of classification models are summarized in the table below.

The grid search method we learned in class allowed us to easily compare mean train scores and mean test scores, where we selected hyper-parameters that do not make the model’s training error to be significantly lower than its testing error. This was an important step in the process to generate valid predictions and to ensure our model is not overfitted, especially as our model for predicting obesity rates have complex structures with numerous features. We also examined variance-bias tradeoff of these models, as we wanted to reduce variance at the tradeoff of a small increase in bias to ensure that our model is appropriately predicting obesity rates based on given features. [**left this red highlighted section from proposal – just wanted to confirm if the grid we used is the same thing as ‘grid search method’?**]

This comprehensive method gave us a better insight into prediction using various features. Our goal of this project was to provide meaningful results that local governments could incorporate in making decisions and implementing local policies at county and state levels.

**6. Evaluation and Results:**

Based on the results of our prediction, our policy recommendation requires local and state governments to spend more taxes to invest minorities and charter schools. This means that they have to predict if they have high obesity rates or not so that they can spend taxes wisely. Thus, we have to take care not only True Positive but also True Negative, so we use the accuracy as a measure for our models. Since our goal is to provide such policy interventions and this could be best achieved by based on our results, we chose the Gradient Boosting classification as our best predicting machine learning model and analyzed its feature importance.

It is interesting that the percentage of population with a bachelor’s degree or higher and median income, which were the two highest correlated features with obesity turned out to be the second and the fourth best predictors of obesity using the best classification model. Also, we were surprised that some features we thought to be good predictors of obesity, such as the number of recreational and fitness facilities per 1,000 population and the tax rates on soda and snacks were not that significant as predictors. Analyzing this in high level, it seems that the level of education is the most important predictor for obesity rates among some commonly observed demographic, health and lifestyle features. The feature importance of the best model is summarized below.

In overall, we’ve obtained the same top predictors between regression and classification models. When we further analyzed our results by categorizing the predicted obesity rates from simple OLS and the best regression model to compare them with true categories, we observed that the match rate was only slightly improved by the best classification model than the simple OLS. When we sought feedback on this part from TAs, we learned that there could be different reasons why more complex machine learning models might not yield much of an improvement over simple non-regularized regression model. Since our prediction of obesity rates is at the county-level, this comparison result seems appropriate.

We further applied our prediction model to predict Coronavirus death rates, using cumulative death counts as of June 1st. We ran regression and classification models with two scenarios, one with obesity rate included as a feature with other 21 features, and another scenario where obesity rate was excluded from the feature set. We found surprising, unexpected results that Coronavirus death rate prediction is improved when obesity rate is excluded from the features. The model for predicting Coronavirus death rate is far from perfect yet, with accuracy score of 54% and precision and recall around 45%. However, when we analyzed its feature importance, we found an interesting result that the single most important predictor for coronavirus death rate is the percentage of Black population in a county.

These results of the project are far from what we had initially expected. We had initially expected obesity rate to be a strong predictor for Coronavirus death rates, and therefore we were interested in exploring the best predictors for obesity rate in order to come up with meaningful policy recommendations that would reduce obesity rates in highly obese counties. However, the results of our project suggest targeting obesity rates would not make a significant difference on Coronavirus death rates and that policy interventions for reducing obesity rates and combating coronavirus disease should be viewed separately.

In order to reduce obesity rate, improving population's highest education level with targeted education policies seems to be the most effective strategy. For fighting coronavirus disease, supports provided to areas with high percentages of Black population seems to be mostly in need. Our results provide interesting, yet uncomfortably candid view of the world. We conducted research on policies related to higher education and race-related issues to come up with effective policy recommendations.

**7. Policy Recommendations:**

We suggest that local and state governments should predict their future obesity rates from percentages of people with high degrees. Also, they should invest in an education system to reduce obesity rates if they have high obesity rates. We understand that the education issue in the U.S. is very complex and things are different in different areas, but we are going to make a more specific policy recommendation.

When we focus on improving the percentage of high school diploma holders, we have to take care of specific races such as American Indian/Alaska Native, Hispanic, and Black([source](https://nces.ed.gov/fastfacts/display.asp?id=16)). In 2017, there were 2.1 million status dropouts between the ages of 16 and 24 and the overall status dropout rate was 5.4 percent. In particular, American Indian/Alaska Native (10.1 percent), Hispanic (8.2 percent), and Black (6.5 percent) had the highest status dropout rates.

Therefore, supporting minorities is important to raise the overall percentages of people with high degrees in each county. Thus, we recommend that local and state governments should create or expand grants for minorities who go to high school, although there may be grants for minority education for now. It leads to giving them an economic incentive to keep studying. We also suggest that local and state governments give grants to charter schools that teach minorities. There are many charter schools that give minorities opportunities to learn([source](https://www.cbsnews.com/news/minority-and-poor-students-gain-from-charter-schools-study-shows/)). Hence, supporting charter schools results in improving minorities’ abilities. These policies reduce obesity rates in county levels indirectly.

**8. Ethics:**

[**from proposal – need to be updated**] The multivariable regression may include proxy variables or variables that are highly correlated with each other. We will need to perform an in-depth analysis to identify any potential bias that may affect or interfere with the outcome of the project. Moreover, potential limitations in interpreting the results may include inability to obtain or include all possibly related features in our model, such as information on health insurance enrollment status or areas more populated with certain age groups. However, we do not expect these limitations to produce significant bias in policy recommendations because we are using various features, including demographics data, along with regularization methods in the model to provide non-overstated results.

**9. Limitations, Caveats, Suggestions for Future Work:**

We used obesity percentage at county level as our predictand. However, obtaining BMI data at county level that accounts for more detailed levels of obesity could help provide better policy recommendation to prevent obesity-related issues. As of now, the BMI data is only available for certain counties, but we recommend collecting the BMI data at county level and using that dataset to improve the current study.

Further, the current project is limited to 3,139 observations as our scope of the study is focused on the county level. The future study could explore data on individual level, especially when having larger amounts of data could provide useful results for machine learning algorithms.

Only few features of our dataset were missing information: 0.7% of number of SNAP authorized stores per 1,000 of population, 1.0% of price of soda compared to national average, and 2.0% of percentage of direct sales from farmers. Considering how there are only small numerical datasets missing, we imputed missing values by calculating the median of the non-missing values in a column. However, 10 out of 21 features have minimum values of 0. Among which, features named ‘snack\_sales\_tax’, ‘farm\_mkt/1000\_pop’, and ‘recr\_fit\_facil/1000\_pop’ are notable, as they have 2,146, 893, and 1,011 number of observations of zero values, respectively. There is a possibility that some of these zero values are, in fact, missing data.

Our conclusion to come up with targeted education policies to improve population’s education level to lower obesity rates within a county is based on our analysis that 1) the level of education was considered as the most important predictor for obesity, and 2) some of the health and lifestyle-related features, including number of recreational and fitness facilities per 1,000 of population and tax rates on snacks, turned out to be insignificant features for predicting obesity rates. Nevertheless, this conclusion could have been attributed by the massive number of zero values for certain health and lifestyle-related features, and more in-depth understanding of zero values and validation for our data might be necessary for the future work.

Overall, there are some limitations with our current data. The availability of good data for future work could help provide better insights and appropriate recommendations for current pandemic of Coronavirus disease and obesity-related issues.

**10. References for Citations:**

“Prevalence of Obesity and Severe Obesity Among Adults: United States, 2017–2018.” Centers for Disease Control and Prevention, <https://www.cdc.gov/nchs/products/databriefs/db360.htm>

Fallik, Dawn. “COVID-19 is hitting some patients with obesity particularly hard.” Science News, 22 Apr. 2020, <https://www.sciencenews.org/article/coronavirus-covid19-obesity-risk-factor>

“Defining Adult Overweight and Obesity.” Centers for Disease Control and Prevention, <https://www.cdc.gov/obesity/adult/defining.html>

“Healthy weight, overweight, and obesity among U.S. adults.” National Health and Nutrition Examination Survey, Centers for Disease Control and Prevention,

<https://www.cdc.gov/nchs/data/nhanes/databriefs/adultweight.pdf>