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. Policies targeted at increasing graduation rates of minorities, increasing teachers' pay and qualifications, and increasing access to students' opportunities to learn how academic work could be applied towards their potential career paths early on could be effective strategies to indirectly reduce obesity rates.

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, it seems possible that the local and state governments can predict their regions' future obesity rates using metrics of local population's highest educational degrees. Therefore, we suggest several policy recommendations related to education in the later section to indirectly reduce future obesity rates.

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

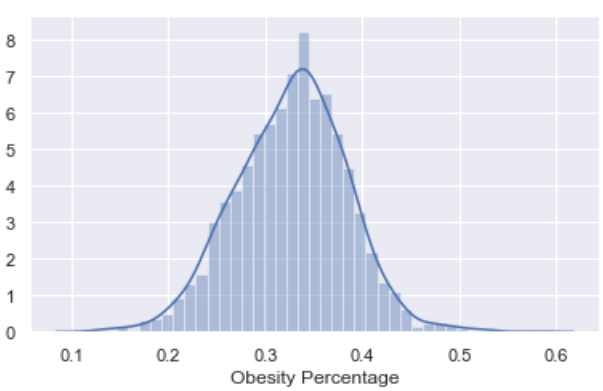
adult male and female in the U.S. have BMIs of 26.6 and 26.5, respectively (Centers for Disease Control and Prevention). BMI levels ranging from 30 to 35 are classified as Class 1 obesity, from 35 to 40 as Class 2 obesity, and at or above 40 as Class 3 obesity, sometimes referred to as 'extreme obesity' or 'severe obesity.'

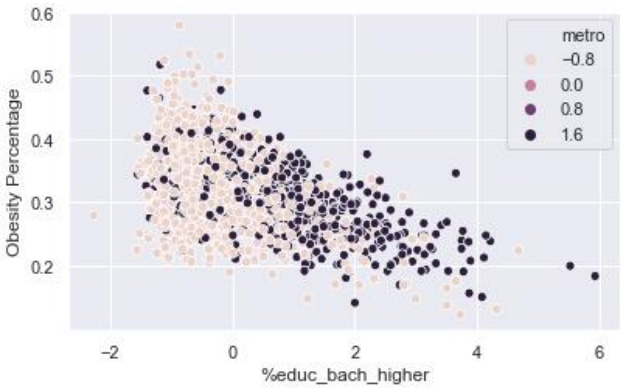
- Label (Coronavirus disease): obtained data on daily cumulative counts of coronavirus disease cases and deaths at county level as of June 1st, 2020. This is a public dataset updated by New York Times on a daily basis pulled from GitHub (https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv).

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').

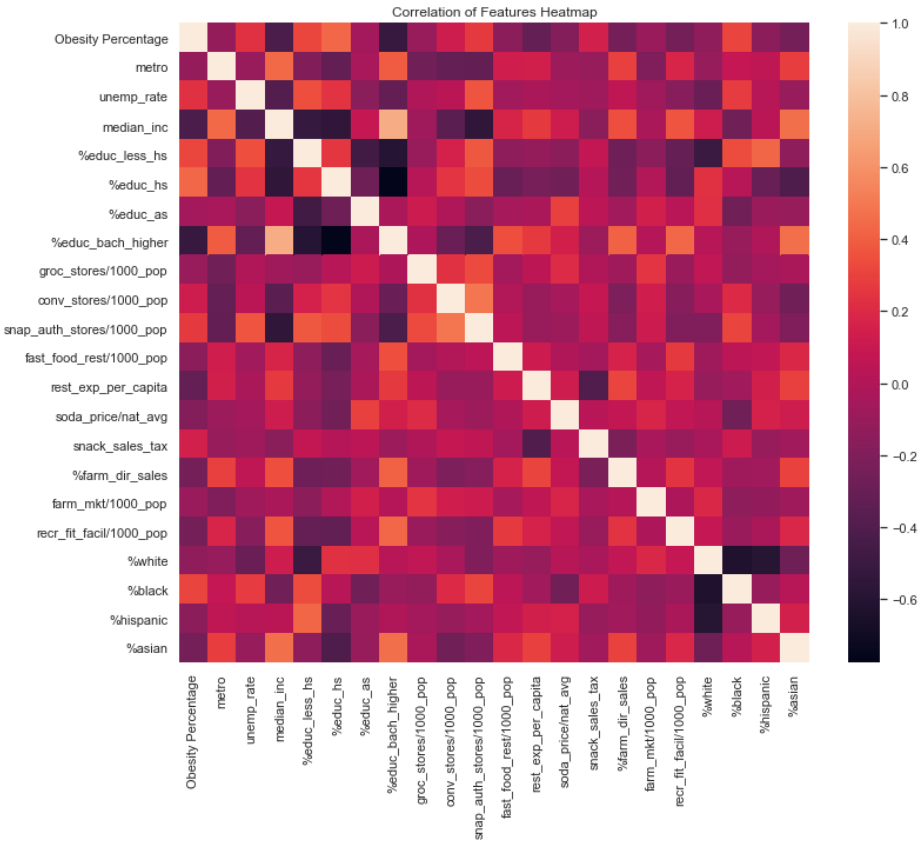
The following are some examples of visualizations we've used in our data analysis.





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.

Model	Parameters	R2_score	MAE	MSE
Random Forest	{'max_depth': 50, 'max_features': 6, 'min_samp...	0.460500	0.031971	0.001690
Gradient Boosting	{'max_depth': 3, 'max_features': 4, 'min_sampl...	0.445780	0.032445	0.001736
Ridge	{'alpha': 10.0}	0.391720	0.034422	0.001906
K-Nearest Neighbor	{'n_neighbors': 25}	0.390366	0.034225	0.001910
Lasso	{'alpha': 0.1}	-0.001746	0.045241	0.003138
Elastic Net	{'alpha': 0.1}	-0.001746	0.045241	0.003138

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 with test set, which is over the best accuracy rate (0.83) calculated in the cross validation. It 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.

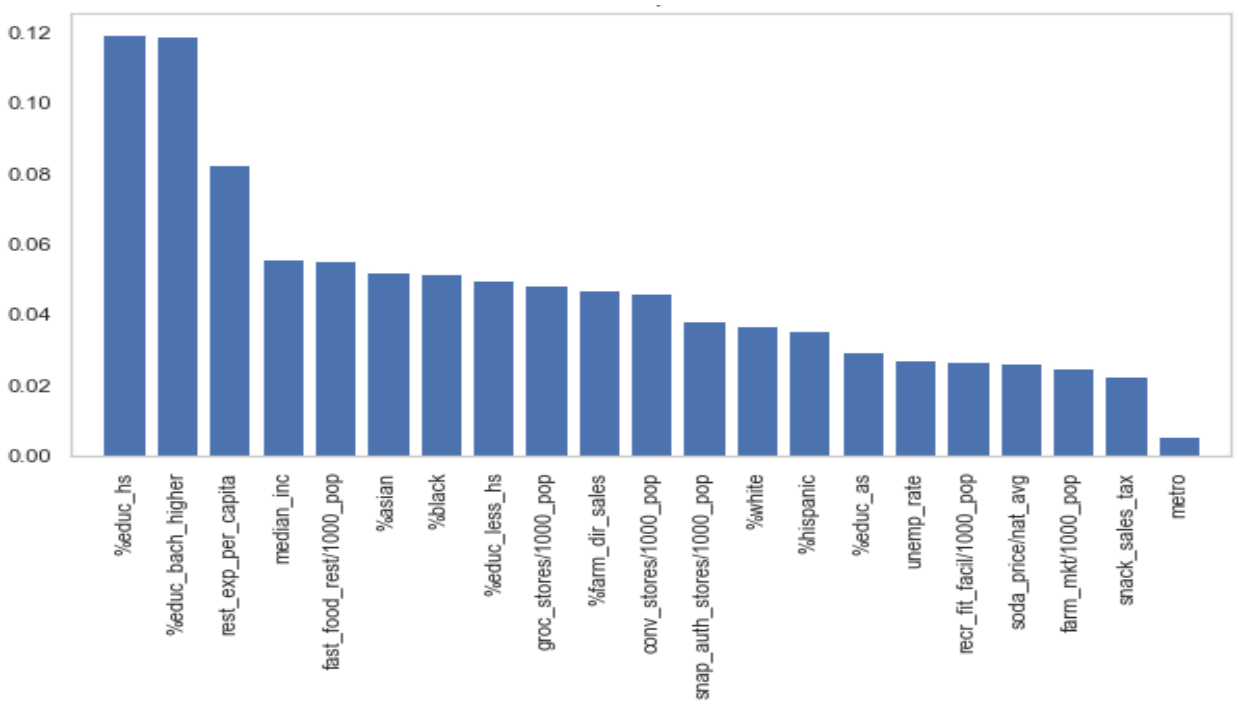
Model	Parameters	Accuracy Score	Precision Score	Recall Score
Gradient Boosting	{'max_depth': 10, 'max_features': 3, 'min_samp...	0.852442	0.712026	0.504171
Random Forest	{'max_depth': 30, 'max_features': 3, 'min_samp...	0.848195	0.556409	0.466223
K-Nearest Neighbor	{'n_neighbors': 8}	0.830149	0.509250	0.457000
Linear SVC	{'C': 1.0, 'gamma': 0.1, 'kernel': 'linear', '...	0.825902	0.543694	0.414288
GaussianNB	{'priors': None}	0.728238	0.520315	0.628456

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.

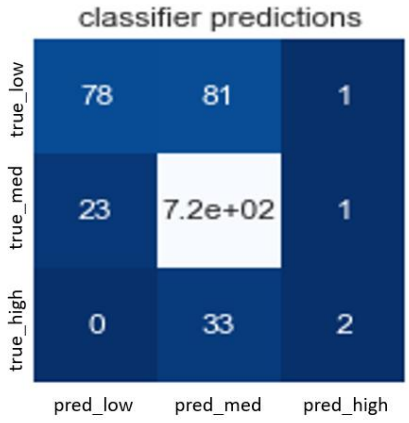
6. Evaluation and Results:

Based on the results of our prediction, our policy recommendation for the local and state governments is to spend more taxes at investing in teachers and students, including minorities, to indirectly reduce obesity rates by improving the highest earned education degree (discussed in more detail in the next section). The governments using our machine learning results would thus need to predict the level of obesity rates to spend taxes wisely. Since our audience would need to consider not only the true positives but also the true negatives, we chose accuracy as our evaluation metric and chose the Gradient Boosting classification as our best predicting machine learning model. As previously mentioned in the earlier section summarizing the results of various prediction models, our accuracy score with testing data turned out to be higher than that of training data. Therefore, we believe our best Gradient Boosting classification model does not suffer from being overfit to the training data.

to further analyze 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.



On the left is confusion matrix of our best Gradient Boosting classification model to assess which types of counties we are under-predicting or over-predicting as part of our evaluation. The result shows that we are predicting the true medium counties better than any of the true low or true high obese counties. However, it turns out that we are also over-predicting the low obese counties as medium obese counties and under-predicting the high obese counties as medium obese counties to a certain extent. From analyzing data, we observed that low obese counties have certain characteristics different from high obese counties in terms of various features in that low obese regions are more likely to be located in metro area, have population with higher education level, have less unemployment rate, have more access to farmers markets and have lower Black population. Therefore, we made a note of this prediction error in our policy recommendation to prevent from having significant bias arise.

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:

Based on our results, it seems possible that local and state governments can predict their regions' future obesity rates from the percentages of local population's highest education degrees. Therefore, it seems possible for governments in high obese regions to reduce future obesity rates through investing in high education. We acknowledge that issues and policies related to education are often complex and that our models contain certain level of prediction errors, but here are some of our policy recommendations based on research for improving the number of high school and college graduates.

Regards to improving the percentage of high school diploma holders, here are some ideas:

- Policies interventions targeted at increasing graduation rates of minorities could be effective, as research shows that high school dropout rates were especially high among minorities. In 2017, there were 2.1 million or 5.4% high school dropouts between ages of 16 and 24, but the dropouts of American Indian/Alaska Native, Hispanic, and Black were 10.1%, 8.2% and 6.5%, respectively (NCES).
- Better train teachers like professionals with higher pay so that qualified professionals won't avoid teaching profession and that excellent teachers will become more incentivized to stay with high-need schools. Research on teachers' pay suggest that teachers make only about 60% of other professionals with similar education and that teachers' pay in the U.S. is much lower than in other OECD countries (OECD). The idea

behind this policy is that better trained and qualified teachers would encourage more students to gain interest in academics and understand how their education would pay off.

- Additional interventions, such as introducing flexibility for students to attend school from 9am to 5pm to better align with parents' working schedules and providing options for breakfast, lunch or dinner at school regardless of parents' income, could be effective.

Additionally, the number of college graduates could be improved through the following policies:

- Increase access to opportunities to learn firsthand how academic work could be applied towards students' potential career paths and provide more opportunities for exposure to workplace experience.
- Since improved high school graduation rates are likely to lead to increased college graduation rates, education policies targeted at reducing high school dropout rates would be effective policies for improving college graduation rates.

The above policies combined with additional funding could create more incentives for students to continue their studies and could in turn also help to reduce obesity rates indirectly.

8. Ethics:

Although we tried our best to include as many features as we could that seemed related to predicting obesity rates, we faced certain limitations in including additional features such as data on health insurance enrollment status or areas more populated with certain age groups. Since our current policy recommendations are based on the best predicting feature from available dataset, our results may suffer from potential bias where there could be more directly related and significant predictors for obesity rates, in which case would serve as better policy interventions.

Besides conducting formal audit of our project, including more features and assessing correlations among features to identify proxy variables would be a way to evaluate potential bias. In order to prevent suffering from ethical or potential bias issues to the extent possible, we performed various sanity checks along the process including applying hyperparameter regularization to models, analyzing areas for under-prediction or over-prediction, and checking for in-sample error vs. out of-sample error to ensure our model is not overfitted.

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

"Dropout rates," National Center for Education Statistics, Institute of Education Sciences, https://nces.ed.gov/fastfacts/display.asp?id=16

"Education at a Glance 2017: OECD Indicators" (2017), Organization for Economic Cooperation and Development (OECD), http://www.oecd-ilibrary.org/docserver/download/9617041e.pdf?expires=1519148041&id=id&accname=guest&checksum=42796EF455E675E79827B115C2A9ADA9