Vaccine Hesitancy
Phase 3 Write Up
DS3000
Arshia Mathur

While we aimed to initially evaluate vaccination rates and compare them with hospitalizations and case rates per county, we decided to pivot based on the feedback we received from phase 2 and incorporate the CDC's vaccine hesitancy rates based on county dataset. In order to effectively analyze these two datasets, we first had to merge the data. The FIPS code is a set of numbers that uniquely identifies a geographic area (specific to the county/state) within the U.S.. Both of these datasets incorporated columns listing the FIPS code, so we therefore merged the datasets based on their FIPS code so that each row of data code accurately reflected the rates of hesitancy for their respective county.

We introduced various different algorithms to complexly analyze our data. We began with a decision tree model because it would help us understand which features were the most important, as it picks features based on their best output. It also works well with categorical and numerical data, which makes it a versatile choice for analyzing vaccine data that includes both. We split the model into training and testing sets, with most of the data used towards training the data. To evaluate, we used the testing set. Since our goal was to estimate and predict hesitancy, we used mean squared error to evaluate our accuracy. We believe mse is a good metric for assessing this model and the others because it would capture larger errors better, considering the target variable was already within a smaller range of percentages. We also felt that the mse could be used to compare the three models we have especially after hyper tuning them. We dropped the string columns that aren't categories like county and state, and then fitted the model using the DecisionTreeRegressor model and fitting it. The parameters that we tuned in this model was the maximum tree depth, which helps us understand how deep the tree can grow, leading to a more complex model and better learning from the data. We expect our decision tree to be fairly accurate since it will be analyzing the best features to make its prediction.

We then used a KNN regressor because it is a strong regressor model that could test various neighbors and predict vaccine hesitancy based on the top features determined by the Decision Tree. We split the model into training and testing sets again and used the testing set to evaluate. First, we graphed our regressor against MSE to determine the best k value. However, after using gridsearch, we realized there was a more accurate k value to analyze. Since the mean train and test scores were both high and similar, this led us to believe the model generalizes well to unseen data. The parameters that we tuned in this model were the number of neighbors and the weighting distance. A greater weighting distance assigns greater influence to nearer neighbors, and combined with an increased number of neighbors, the model should be strongly accurate because of these increased opportunities to account for nuances.

We made a random forest model because it was a strong model we studied in class and its ability to aggregate multiple decision trees and work with nonlinear data. Like the previous models, we split the data into testing and training sets and used the testing set to evaluate. The parameters that we tuned in this model were the number of estimators, leading to a more complex model and better learning by averaging predictions and capturing complex patterns. Considering the aggregation of multiple decision trees' predictions, leading to reduced overfitting, we believe that this would be a very accurate model.