Covid-19 Cases Data Analysis and Visualization: Classification

Authors:
Zahra Hoobakht Ph.D student in Electrical Engineering,
Jason Brown Graduate Student in Computer Science,
BhargavaSharabha Pagidimarri Graduate Student in Data Engineering,
Lyle School of Engineering,
Southern Methodist University

Project 3:
CS7331- Data Mining
Instructor: Dr. Maya Al Dayeh
November 29th, 2023.

ABSTRACT:

In response to the COVID-19 pandemic, this study applies classification techniques to our Data Models to identify geographic and demographic locations in the United States that would fare negatively if a fourth wave of COVID-19 outbreaks were to happen. We utilize available datasets, including the COVID-19 Cases Plus Census file, COVID-19 Cases TX, and the Global Mobility Report. Our analysis identifies key features in our data set that contributed to how well various US counties handled the previous COVID-19 waves. Our focus will be on the demographics of the various US counties. Selected features were chosen to be used in our data models for classification center around significant mitigating features that are known to affect a person's response to a contagion like the COVID-19 pandemic. Our selected few features for our models are centered around age groups, race groups, educational levels, income levels, commuters, and household properties. These models will help us to predict where there will be a need for targeted interventions for disease control. The findings have important implications for public health strategies and resource allocation in future pandemics. This research enhances our understanding of the pandemic's spread dynamics and offers valuable insights for mitigating its impact.

Table of contents

1. Introduction: 6
2. Business Understanding: 6
3. Data Understanding 6

3.1 Data description: What is the Data we are looking at in Model 1? 7

Summary Statistics Derived from Selected Features on Race 7

Table 1: Breakdown of US population by race. 7

Figure 1: Correlation Map of Selected Features for Ground Truth 8

Table 2: Correlation Coefficient Matrix for all US Counties – key features. 8

Ground Truth: US All Counties *9*

Figure 2: Map of training set counties. 10

Ground Truth: Variable Importance *11*

Figure 4: Variable Importance statistics from Ground Truth Data model. 11

Figure 5: Map of predicted results for the Ground Truth dataset. 11

3.2 Data Preparation: Preparing COVID-19 with Census Data to Focus on Race using Random Forest Classification Algorithm 12

Asian Majority Counties *12*

Black Majority Counties *12*

Figure 6: Map of actual results in the Black Majority Counties. 13

Hispanic Majority Counties *13*

Figure 7: Map of actual results in the Hispanic Majority Counties. 13

Native American Majority Counties *14*

Figure 8: Map of actual results in the Native American Majority Counties 14

White Majority Counties *14*

Figure 9: Map of actual results in the White Majority Counties 14

Creating Classes for each Race *15*

Asian Majority County Class *15*

Figure 10: Variable Importance Histogram of Asian-majority Counties 15

Figure 11: Variable Importance Histogram of Asian-majority Counties 15

Black Majority County Class *16*

Figure 12: Map of actual results in the training set for Black Majority Counties. 16

Figure 13: Variable Importance for Black Majority Counties. 16

Hispanic Majority County Class *17*

Figure 14: Training set Map for Majority Hispanic Counties 17

Figure 15: Variable Importance for Majority Hispanic Counties 17
Native American Majority County Class *18*
Figure 16: Training cases from Majority Native American Counties 18
Figure 17: Variable Importance for Majority Native American Counties 19
White Majority County Class *19*
Figure 18: Test cases for White Majority Counties. 19
Figure 19: Variable Importance for White Majority Counties. 20
4. Modeling 20
4.1 Modeling the different Race Classes 20
Black Majority Counties 20
Figure 21: Map of predicted results in the test set for Black Majority Counties. 21
Figure 20: Map of actual results in the test set for Black Majority Counties. 21
Table 3: Black Majority County Class Confusion Matrix 21
Hispanic Majority Counties 21
Figure 23: Actual Results for Test Cases from Majority Hispanic Counties 22
Figure 24: Predicted Results for Test Cases from Majority Hispanic Counties 22
Figure 25: Confusion Matrix for Majority Hispanic Counties Class 23
Table 4: Hispanic Majority County Class Confusion Matrix 23
Native American Majority Counties 23
Figure 26: Actual results for test cases for Majority Native American Counties 23
Figure 27: Predicted results for test cases for Majority Native American Counties 24
Table 4: Hispanic Majority County Class Confusion Matrix 24
Table 5: Native American Majority County Class Confusion Matrix 24
White Majority Counties 24
Figure 29: Map of actual results for Test Cases in White Majority Counties. 25
Figure 30: Map of Predicted results for Test Cases in White Majority Counties. 25
Table 5: White Majority County Class Confusion Matrix 26
Figure 31: Confusion Matrix for White Majority Counties. 26
Model 2: Classification based on Gender and Age 26
Table 6: Features for Model 2 in the USA 26
Table 7: value of features after normalization 28
Table 8: check balance of data 28
Table 9: number of bad cases in each State of USA 28
Table 10: Splitting dataset into traing and testing 29

Figure 31: Training dataset of number of cases per population greater than median in four counties of “TX”,” FL”,” NY” and “CA” based on gender and age features 30
Figure 32: Variable of importance 32
Figure 34: bad predicted data for entire USA based on random forest modeling 33
4.3 Model 3: Classification model to predict infection status and disease severity 35
Model Advantages: 35
5. Evaluation 35
Evaluating the models focused on Race. 35
Evaluating the models focused on gender and age 36
Evaluating the model based on infection status and disease severity. 36
Model Performance : 36
6. Deployment 37
7. Exceptional work 37
Performance Comparison: 37
In-depth Explanation for Superior Performance: 38
8. Conclusion 38

1. Introduction:

This report aims to find results using machine learning classification techniques like Random Forest to identify the most at-risk US counties if another wave of COVID-19 was to hit. Aided with COVID-19 case data, demographic information, and geographic distribution, we aim to identify these locations with great certainty to provide actionable insights for public health officials, policymakers, and researchers. This research not only aids in identifying high-risk areas and vulnerable populations but also contributes to a deeper understanding of how infectious diseases like COVID-19 propagate in varied settings. The significance of this study lies in its potential to inform targeted and evidence-based interventions, enhance resource allocation, and improve the overall effectiveness of public health responses during the COVID-19 pandemic. Moreover, the knowledge gained from this research will have far-reaching implications for future pandemic preparedness and management. However, before we decide on which classification model to base our interventionism on, we evaluate how well our models perform, and decide on our deployment options.
In the following sections, we will describe the methods employed in our analysis, present the findings and insights derived using classification algorithms on COVID-19 case data, discuss their implications, and conclude with a call to action for further research in this critical area of pandemic response.

2. Business Understanding:

A healthy workforce is needed to be the backbone of this country so we can maintain our liberties and lives. The responsibility for keeping America's workforce healthy is on everybody including companies like Google, and necessarily from public health departments. These entities worked together to provide data related to the COVID-19 outbreak. Using this data and machine learning classification techniques, we can aid US departments with accurate predictions on where aid would be needed.
From our previous research, we already had cleaned and tailored datasets to use to build the data models. We have also established a ground truth for what our classification results should look like. We established this ground truth using a Random Forest classification technique and with a dataset that focused on how COVID-19 affected the United States on a national scale using only the features we had selected to focus on. Our results found that out of 3,139 US counties, 1275 counties experienced high Covid-related fatality rates. Our classification models were able to use these results to make predictions on where COVID-19 outbreaks could have severe consequences with over 60% accuracy. The random forest algorithm found that out of 2743 US counties only 876 counties would not have high fatality rates. These results speak for a need for a plan of action, and a necessity for expedience in identifying vulnerable counties that should receive interventions in terms of updated infrastructure so they can better weather if another outbreak happens.

3. Data Understanding

The datasets available were the COVID-19 Cases Plus Census file, COVID-19 cases TX, and the Global Mobility Report, and there is a custom-made dataset that reports hospitalization rates per county. These CSV datasets were provided with the assignment and originally pulled from the Google Cloud platform. Google Cloud's marketplace hosts many other COVID-19 datasets that they make available free of charge. These datasets were provided to Google by governmental, medical, and research organizations to assist

in responding to the COVID-19 epidemic. The selected features we will focus on using our classification model are related to education level, income class, and race.

3.1 Data description: What is the Data we are looking at in Model 1?

One of the sources for data is the US Census Data with COVID-19 data set. The Census data is from the last US Census report from the last US Census that was collected at the beginning of the Pandemic. The reward of a stimulus check was a strong encouragement for a high turnout of the Census data, so the results could be a fairly accurate demographic snapshot of the entire US population. The Census Data was coupled with COVID-19 that was obtained from a national effort of all US state departments of Health to report on COVID-19-related data. The reporting of COVID-19-related data was reported on January 19th, 2021, and is considered a tallied sum of COVID-19-related variables (per US county) from the start of the Pandemic in January 2020 until the date of reporting. There is data for all 3,142 US counties. The Census data as expected had no duplicate columns or rows. It did have a few columns with no data and those columns were removed during a previous project. Provided are some summary statistics and a description of the dataset. The dataset has features that were extracted from the Census data which contained 259 features. We condensed those features down to 35 features, by aggregating all school children from kindergarten to 12th grade into one group, summarizing the different types of commuters i.e., local, group commuters, out-of-towners, and public transportation into their respective e groups, a similar approach was taken with the population's different level of incomes. For the documentation of the created visualizations below please refer to Project_3_Data_Wrangling_Final.html.

Summary Statistics Derived from Selected Features on Race

Table 1 shows a breakdown of how the different race majority counties experienced the COVID-19 pandemic. High COVID-19 fatality rates affected 57% of White majority counties experienced a high number of COVID-19 fatalities, 87% of Black Majority Counties, 81% of Hispanic majority counties experienced a high number of COVID-19 fatalities, 67% of Native American majority Counties experienced a high number of COVID-19 fatalities, while all Asian majority counties experienced only low COVID-19 related fatality rates.

Table 1: Breakdown of US population by race.

Race	Number of Majority Counties	Total Population	Number of high fatality counties
Amerindian	34	2,098,763	23
Asian	5	16,989,540	0
Black	129	39,445,495	111
Hispanic	117	56,510,571	95
White	2857	197,277,789	1635

Total US Population = 321,004,407

Figure 1 below is a correlation map for all US Counties. In this way, this is a correlation map for the ground truth. In this map we can see there are correlations where correlations would be expected. For

example, poverty density and income under 40,000 densities show to be highly correlated, Just like total population correlates highly with number of K through 12 students.

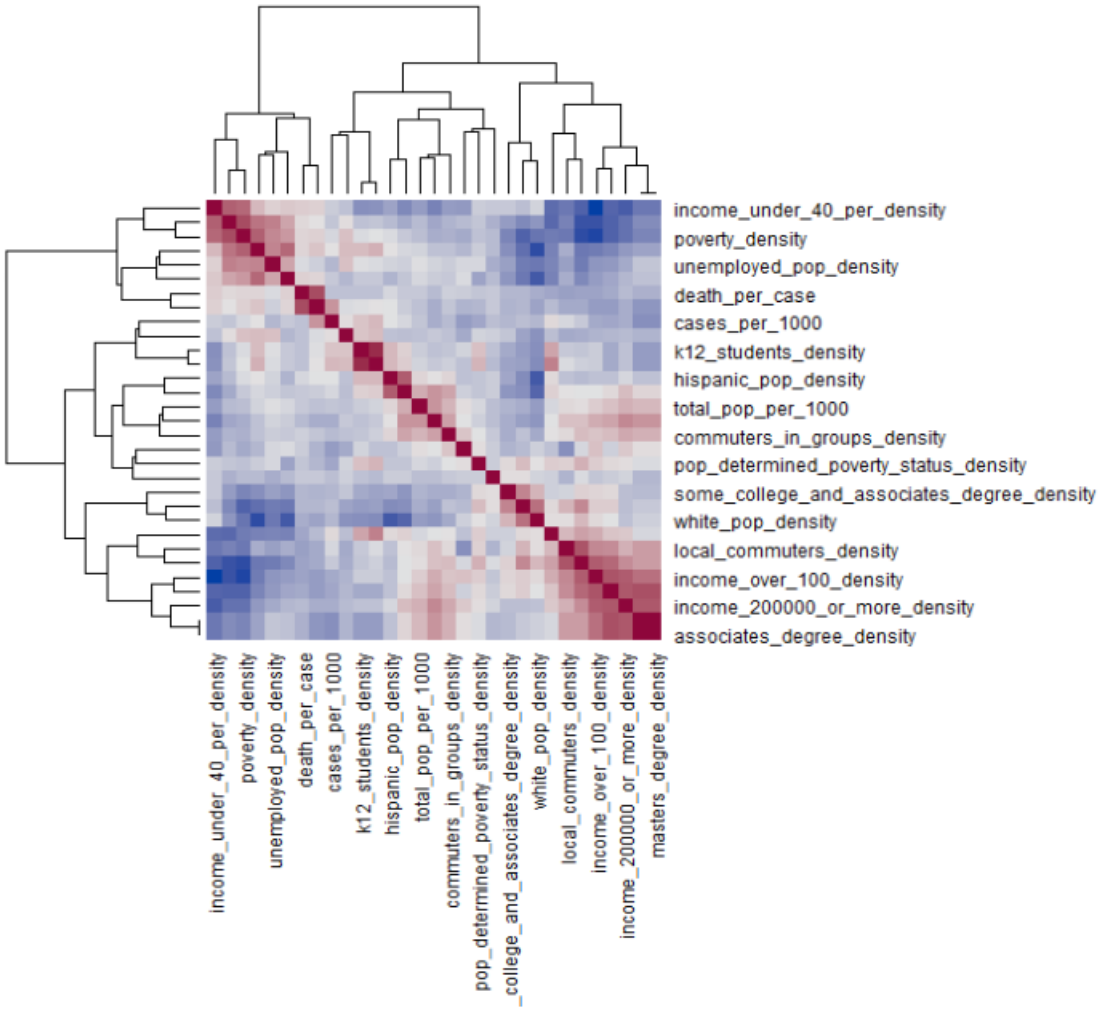

Figure 1: Correlation Map of Selected Features for Ground Truth

In Table 2 below, is a Correlation coefficient matrix. This table spells out what we see in the correlation map we see in Figure 1. It is important to note that correlation coefficient is a measure of the linear association between two variables. The larger the coefficient value the stronger the correlation there is between those values. For this project, some key correlations are applicable to our focus, these correlations are the strong correlation between death per case and income, poverty, unemployment, and the number of K-12 students in that county. Also, on a similar note, there is a strong correlation between cases per 1000 and total population. However, that last correlation was to be expected strictly from a number's standpoint.

Table 2: Correlation Coefficient Matrix for all US Counties – key features.

Features	Income_under_40_per	Poverty_density	Unemployed_pop_	Death	Cases	k12_students_	Total_pop_	Commuters_in_groups_density

	_density		density	_per_case	_per_1000	density	per_1000	
income_under_40_per_density	100	.99	.76	.81	-.07	.66	.45	.38
Poverty_density	.99	100	.83	.89	.08	.75	.58	.46
Unemployed_pop_density	.76	.83	100	.92	.60	.99	.91	.38
Death_per_case	.81	.89	.92	100	.46	.89	.85	.01
Cases_per_1000	-.07	.08	.60	.46	100	.70	.85	.24
k12_students_density	.66	.75	.99	.89	.70	100	.95	.36
Total_pop_per_1000	.45	.58	.91	.85	.85	.95	100	.49
Commuters_in_groups_density	.38	.46	.38	.01	.24	.36	.49	100

Ground Truth: US All Counties

Shown in Figure 2 below is the map of Counties in training set for Ground Truth. The maps show the actual results. Red Counties are counties with high COVID-19 fatality rates. A high COVID-19 fatality rate per county is a rate of 10 deaths per 10000 people. For the testing set, we chose highly populated states like Illinois, Georgia, and California. Oklahoma was chosen as part of the test set since Oklahoma counties performed remarkably better than many other small populations and low-income counties in other states. Please note that counties in red marked counties that had a high COVID-19-related fatality rate.

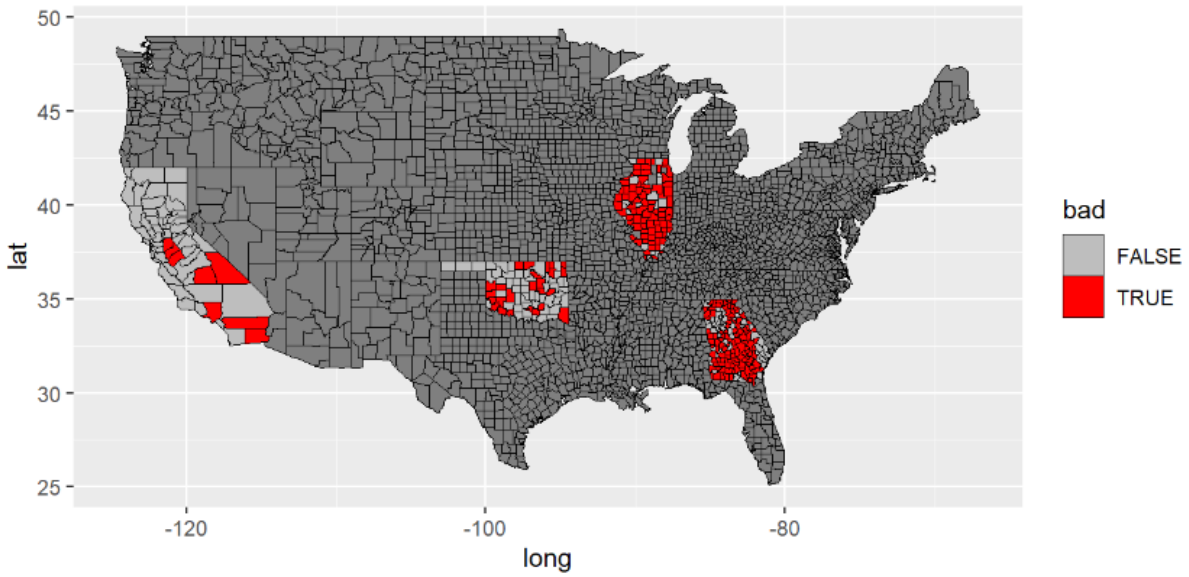

Figure 2: Map of training set counties.

Shown in Figure 3 below are the actual results for the testing set.

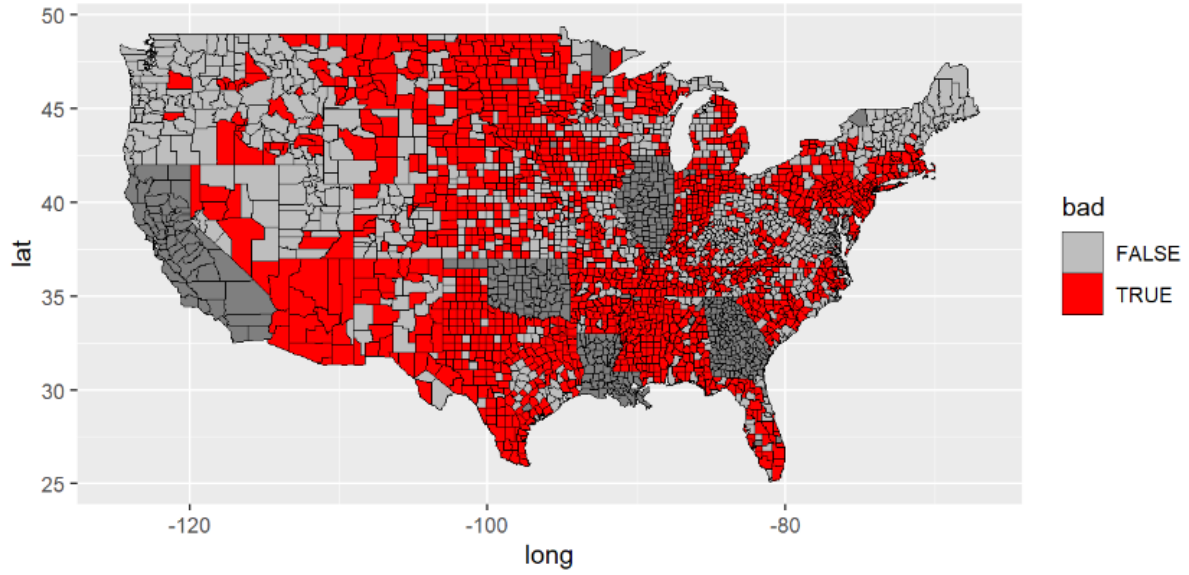

Figure 3: Map of testing set counties.

In Figure 4, we show variable importance. The variable importance is derived from the data model itself after the training set has been trained by a Random Forest Algorithm. Variable importance helps us to identify features that played a key role in the outcome if a county had a good or bad COVID-19 fatality rate. Features that are above rate above 50% in importance are masters_degree_density, associates_degree_density, asian_pop_density, some_college_and_associates_degree_density, and amerindian_pop_density,

Ground Truth: Variable Importance

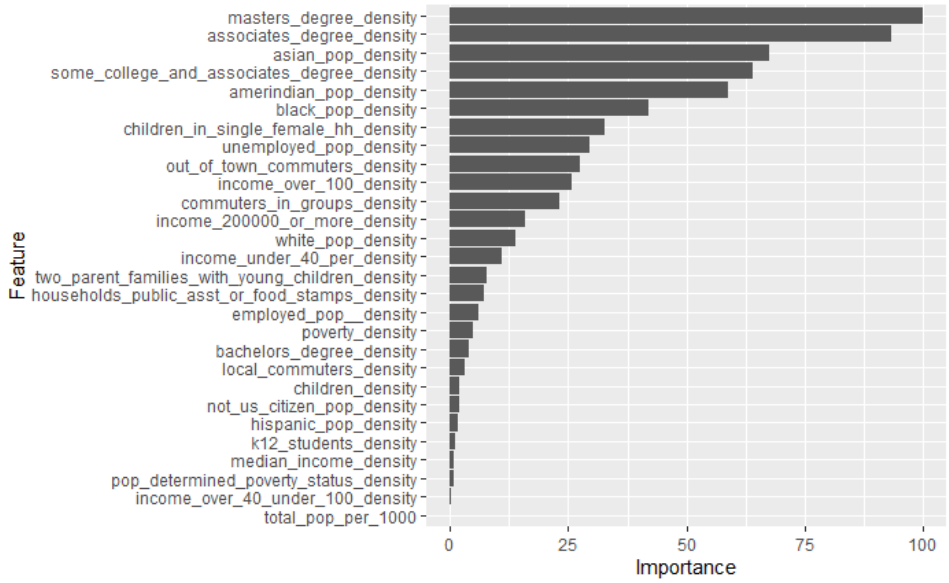

Figure 4: Variable Importance statistics from Ground Truth Data model.

Shown in Figure 5 below are the predicted results for our testing set. You can compare Figure 3 and Figure 5 together to a side-by-side comparison of predicted results and actual results. According to the corresponding confusion matrix which can be found on the Project_3_Data_Wrangling_Final.hml, reported the accuracy of 61% and Kappa of 19%. While these are not the best result measures, it is a great starting point, as tweaking the hyperparameters and changing the classification algorithm my yield better results. To summarize the results the algorithm predicted True incorrectly for badly hit counties 523 times and predicted True correctly for badly hit counties 1098 times. While on the other hand it predicted False incorrectly for not badly hit counties 545 times, and False correctly 577 times. Due to the critical nature of this project, there is a need for room for improvement.

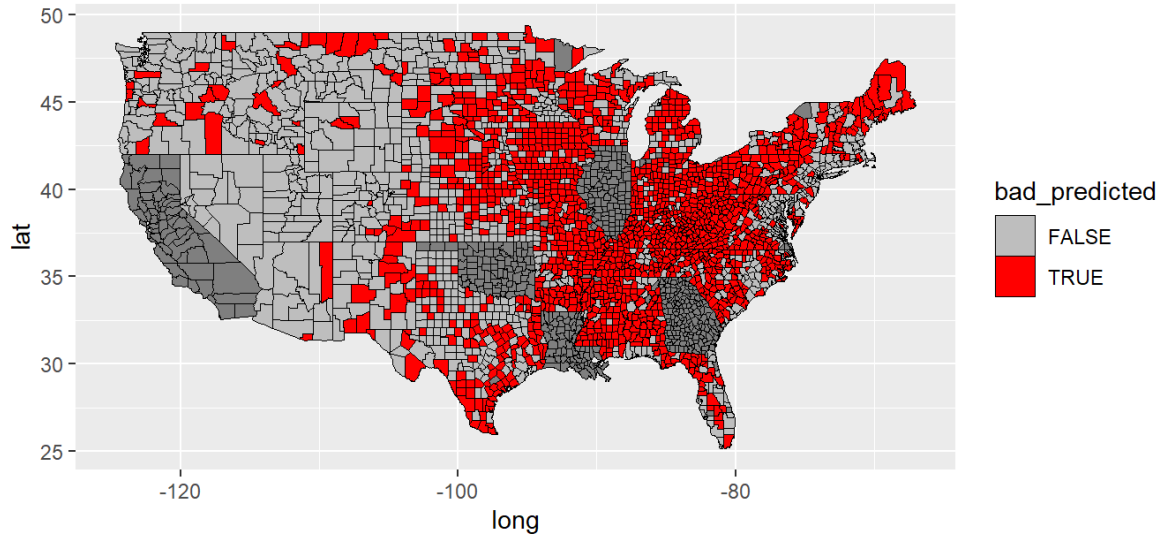

Figure 5: Map of predicted results for the Ground Truth dataset.

For the documentation of the created visualizations below please refer to Project_III_Data_Class_on_Races_COVID_Outbreak_Outcome_Classification_Final.html.

3.2 Data Preparation: Preparing COVID-19 with Census Data to Focus on Race using Random Forest Classification Algorithm

To prepare this data, we used feature extraction to condense the number of features to be used, this was already done in project I. This provides for better computational efficiency when training. The features selected for focus in this area are on race. A simple Boolean equation was used to derive which counties in the Census data set were Asian, Native American, Black, Hispanic, and White majority counties. Other features selected were those that are known to be exacerbating features when it comes to surviving contagion. These features are features like income level, education level, and number of children. The data was normalized to show the population density for each feature. That is the density for each feature per county population size is what is being used in this training set. Also, both the state and county columns had to be dropped and re-created as they were in their original form showing to have too many levels and that was preventing the algorithm from performing training on the training set.

All cases in this section include only locations where there have been confirmed cases. This reduced our dataset from 3,142 US counties to 3,139 US counties. Below are the counties that were removed.

The three counties that were removed for having no confirmed cases:
Kalawao County, HI, Majority Asian County
Bristol Bay, Ak, Majority White County
Yakutat City and Borough, AK, Majority White County

Asian Majority Counties

Since the Asian Majority Counties all seemed to fare so well and since there are so few Asian majority US counties, there is nothing to report on those counties. Furthermore, only 1 Asian Majority County was part of the continental US, and that is Santa Clara County in California. The only remaining counties are in the US states Hawaii and Alaska. Since Hawaii and Alaska counties do not show on the map set, we are using those features that will be left out.

Create 5 different classes for each major US race. Already selected featured subsets. Identify counties Created subsets for Majority White, Majority Asian, Majority Native American, Black, and Hispanic majority counties. Looked at how COVID-19 affect the groups differently as a group.

Black Majority Counties

States with majority Black Counties are Georgia, Alabama, Florida, Missouri, Mississippi, Virginia, New Jersey, South Carolina, Louisiana, North Carolina, South Carolina, Tennessee, Maryland, Arkansas, District of Columbia, and Pennsylvania.

,

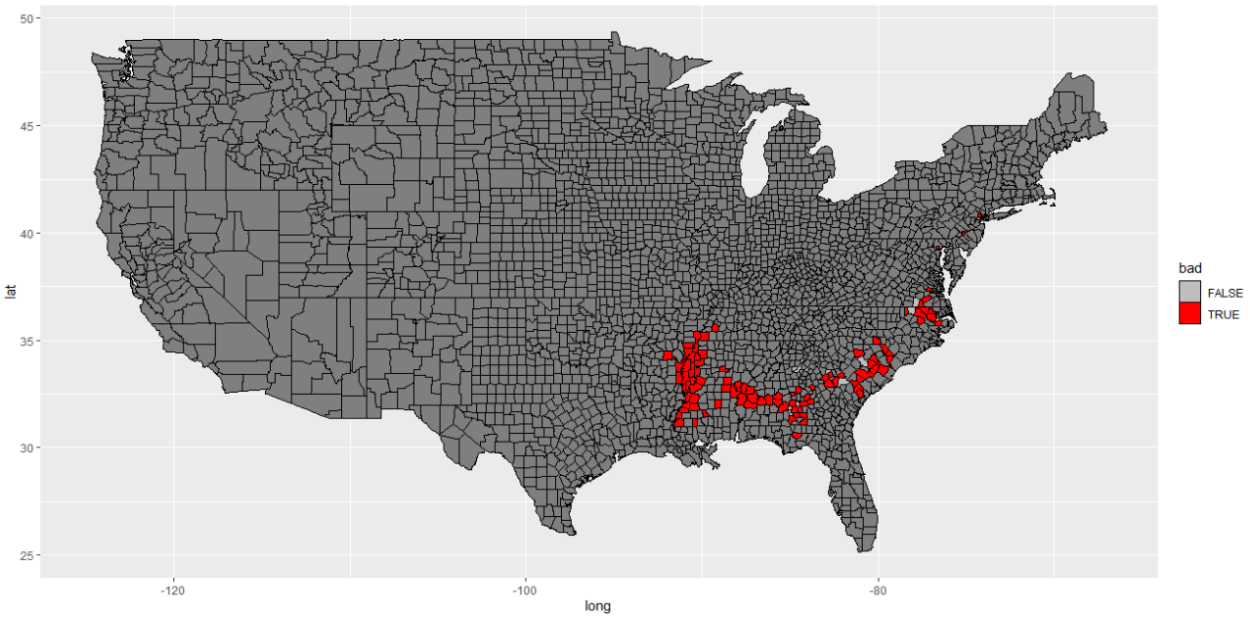

Figure 6: Map of actual results in the Black Majority Counties.

Hispanic Majority Counties

States with majority Hispanic Counties are Arizona, California, Colorado, Florida, Kansas, New Jersey, Virginia, Washington, New York, Texas, and New Mexico.

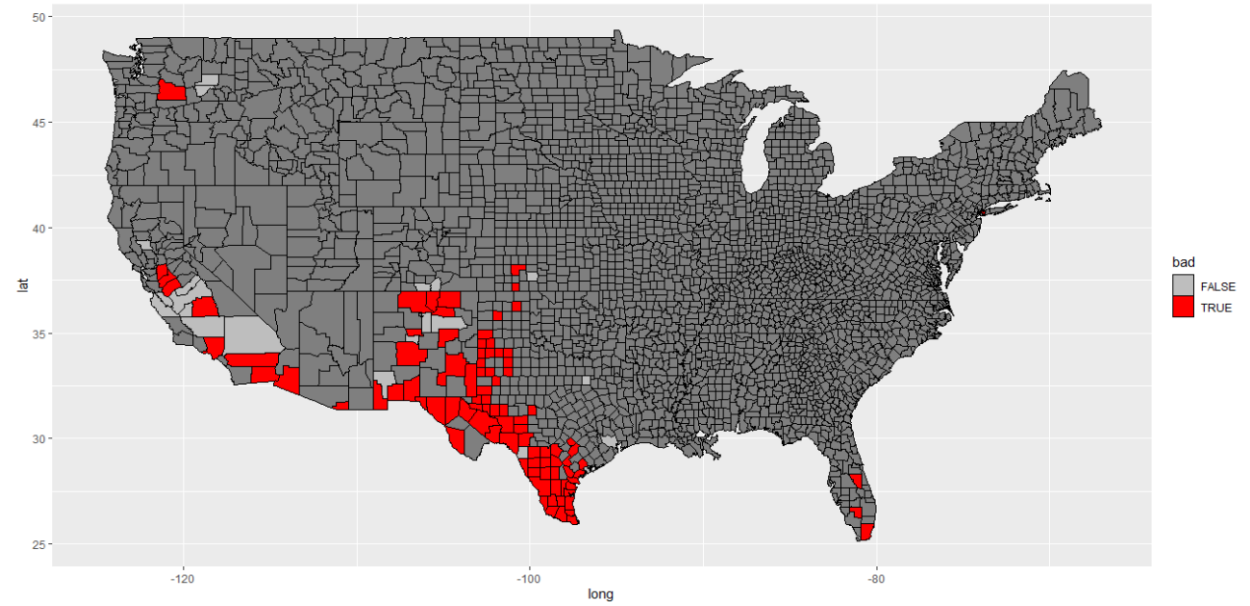

Figure 7: Map of actual results in the Hispanic Majority Counties.

Native American Majority Counties

States with the majority Native American Counties are Alaska, Arizona, Montana, North Carolina, North Dakota, Nebraska, New Mexico, Oklahoma, South Dakota, Utah, and Wisconsin.

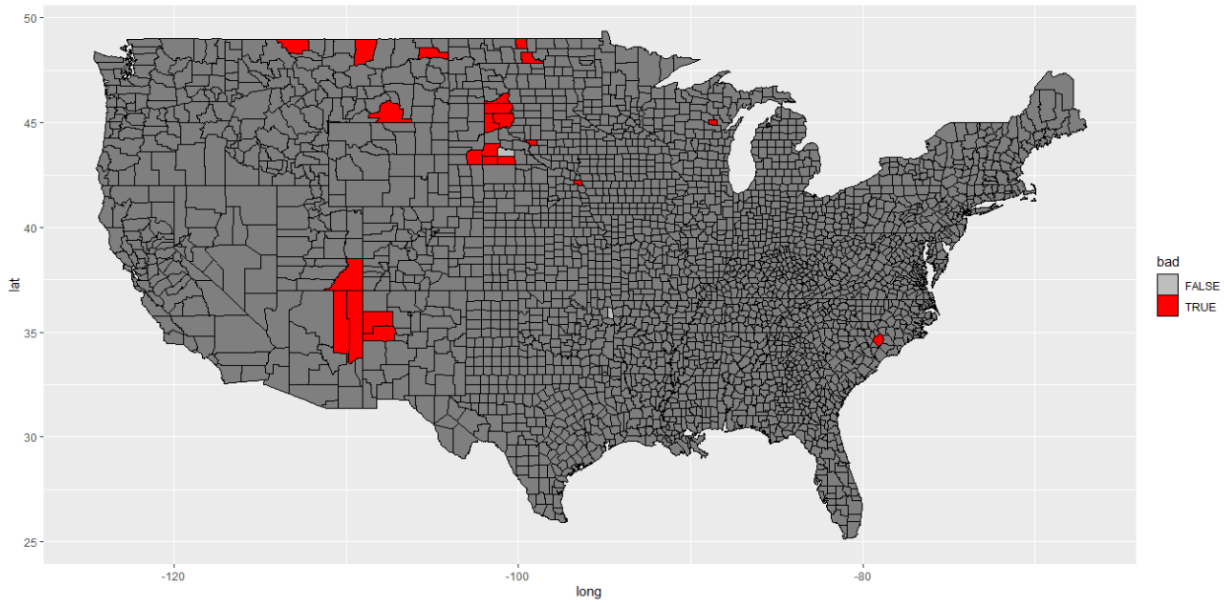

Figure 8: Map of actual results in the Native American Majority Counties.

White Majority Counties

White majority counties can be found in all fifty states. There is one little caveat and that is the District of Columbia which is included in this report as an entity and equivalent to a state. The District of Columbia is a Black-majority County.

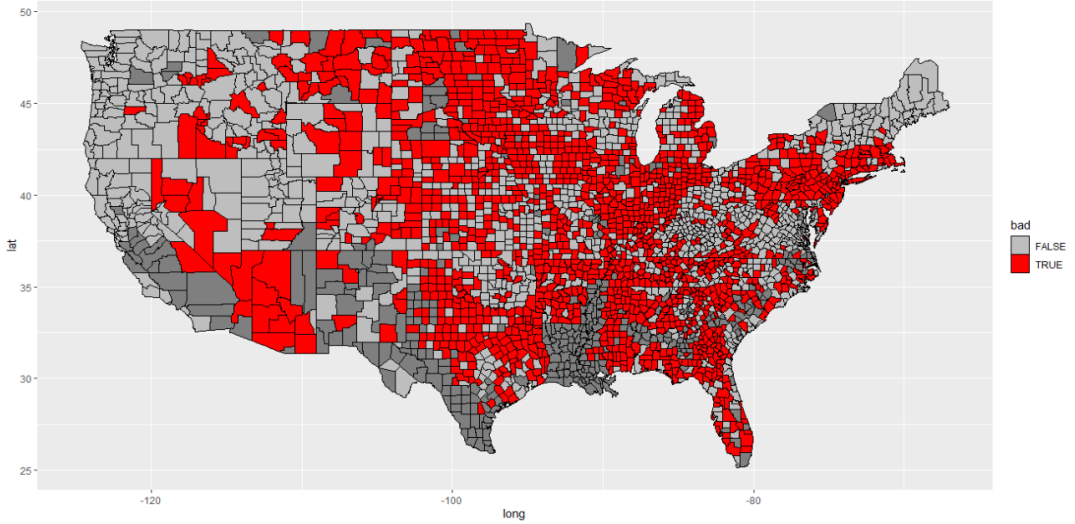

Figure 9: Map of actual results in the White Majority Counties.

Creating Classes for each Race

Asian Majority County Class

As mentioned before the number of Asian-majority counties was few and all those counties experienced low covid related fatality rates it was not possible to train a model, however, we were able to generate some discernable data showing variable importance (Figure 11) and a map of the continental US with the sole Asian Majority County,

Santa Clara County

Santa Clara in view (Figure 10).

It is very hard to tell what some of the features in the variable importance histogram may mean. For example, the children in single female house density seems to be a very important feature. It may be better to remove this feature in this class if further research is done.

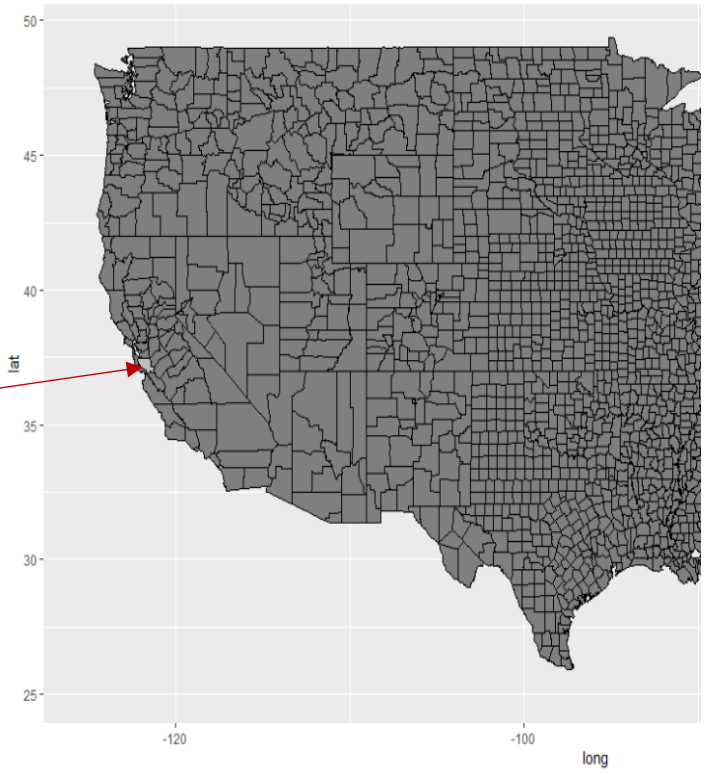

Figure 10: Variable Importance Histogram of Asian-majority Counties

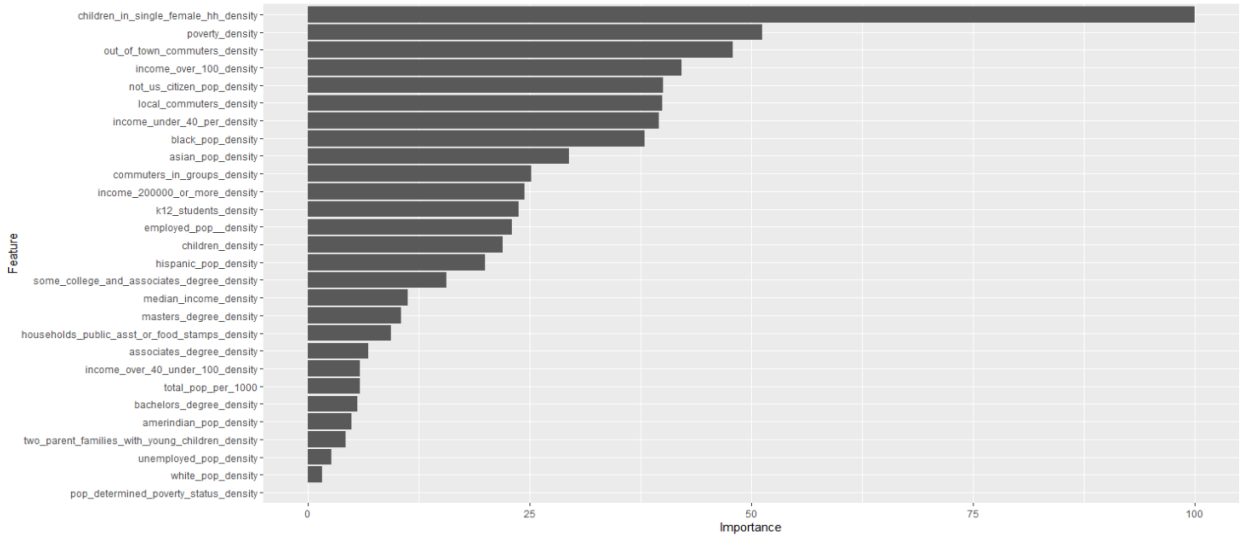

Figure 11: Variable Importance Histogram of Asian-majority Counties

Black Majority County Class

From the ground truth class, we developed subsets of that class. Below is the training set for the Black-majority counties. Notice that many states were used for the training set. This was due to the large number of counties that experienced high fatality rates compared to the number of counties that didn't. This created very imbalanced classes. Choosing which states to use for the training for this class was a hard choice, as choosing too few counties was not generating good results concerning accuracy. The best option in this case was to use many counties, hence, more data to predict better results. The training set in this case consisted of Georgia, Alabama, New Jersey, South Carolina, Louisiana, North Carolina, South Carolina, Tennessee, Maryland, Arkansas, the District of Columbia, and Pennsylvania. Figure 12 below shows a map for these training sets. The training set represents 88 counties and 13 of those counties had low Covid-related fatality rates. All these counties were on the eastern side of the US.

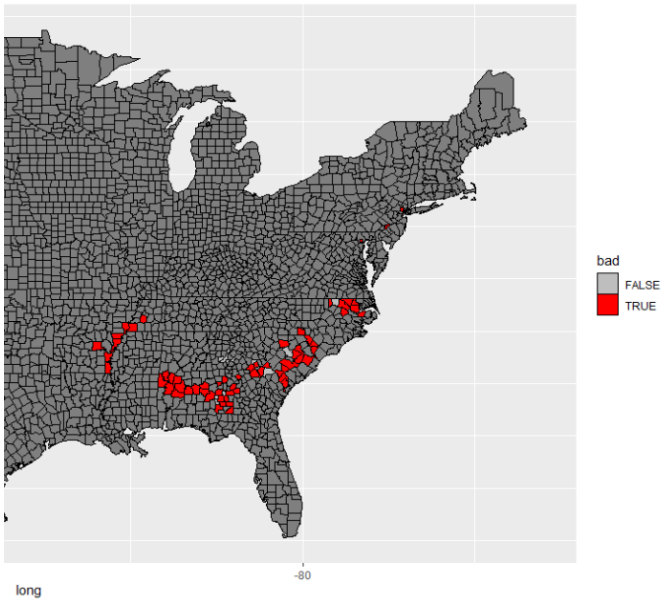

Figure 12: Map of actual results in the training set for Black Majority Counties.

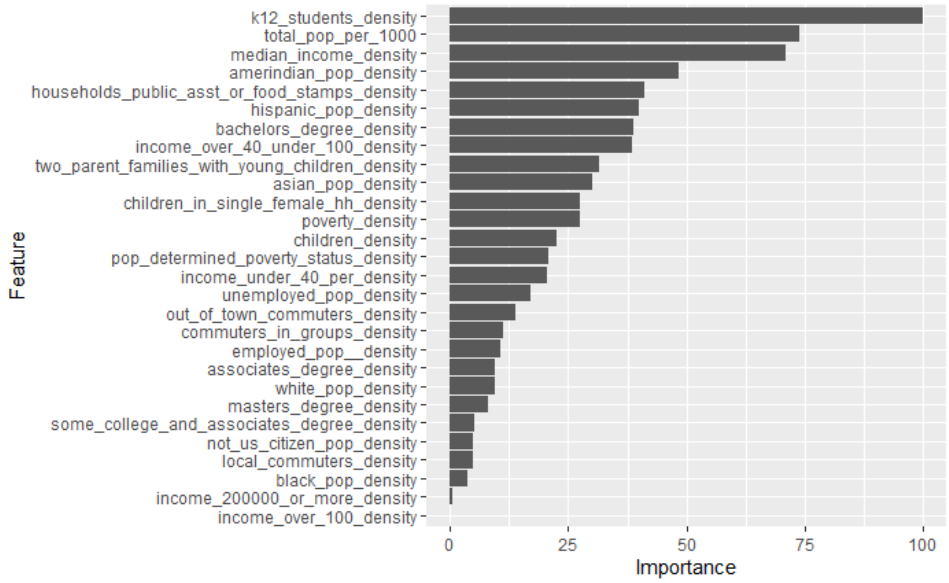

Figure 13: Variable Importance for Black Majority Counties.

Shown in Figure 13 is a histogram output of the most important variables in the Black Majority Counties class model. The large importance factor of k12 student density would seem to indicate that the transmission of COVID was largely attributed to the number of k12 students located in the county.

Hispanic Majority County Class

Shown in Figure 14 is the map for the Hispanic-majority counties trainset set. The counties shown are from the states of Texas, and New Mexico. There were 85 counties in this training set, only 8 of those counties had a low COVID-19-related fatality rate, and 77 counties with a high COVID-related fatality rate. Notice that Dallas County and Harris County are represented as counties considered to have a low COVID-related fatality rate.

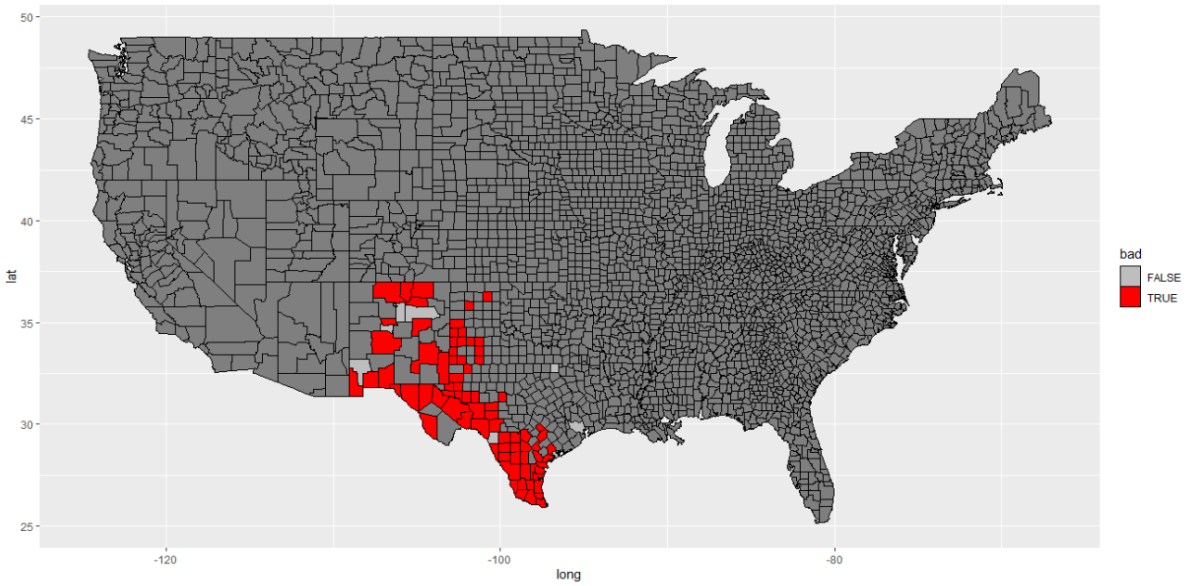

Figure 14: Training set Map for Majority Hispanic Counties

Shown in Figure 15 is a histogram of the output for the important variables in the Hispanic-majority Counties trained Class Model. Notice that population density, education level, and income levels were determined to be the more important variables for this Class model.

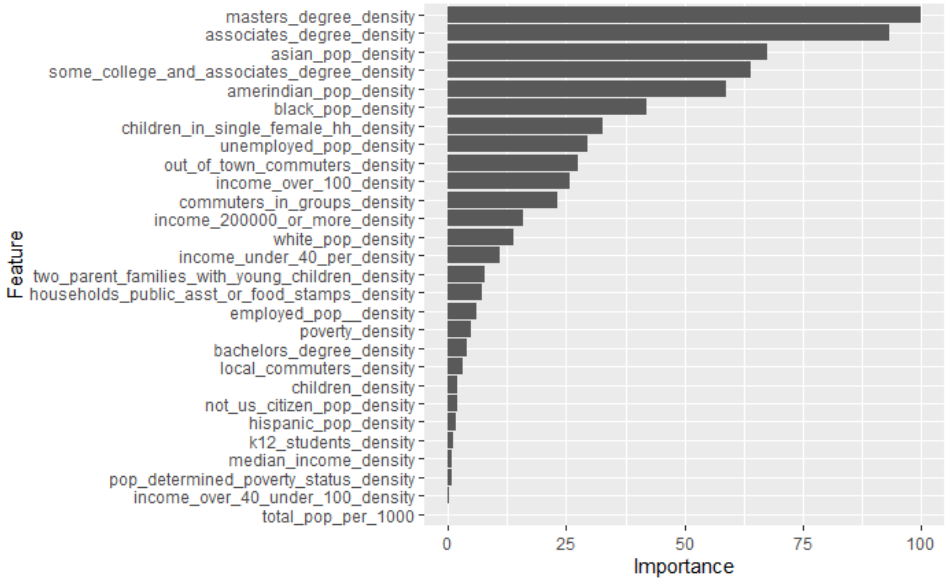

Figure 15: Variable Importance for Majority Hispanic Counties

Native American Majority County Class

The Native American Majority test class was chosen from the US states with the most Native American majority counties, South Dakota, and Alaska. The training set represents 18 counties and 10 of those counties had low Covid-related fatality rates. Figure 16 shows a map of those counties, please note Alaska is not present but Alaskan counties are included in the training set.

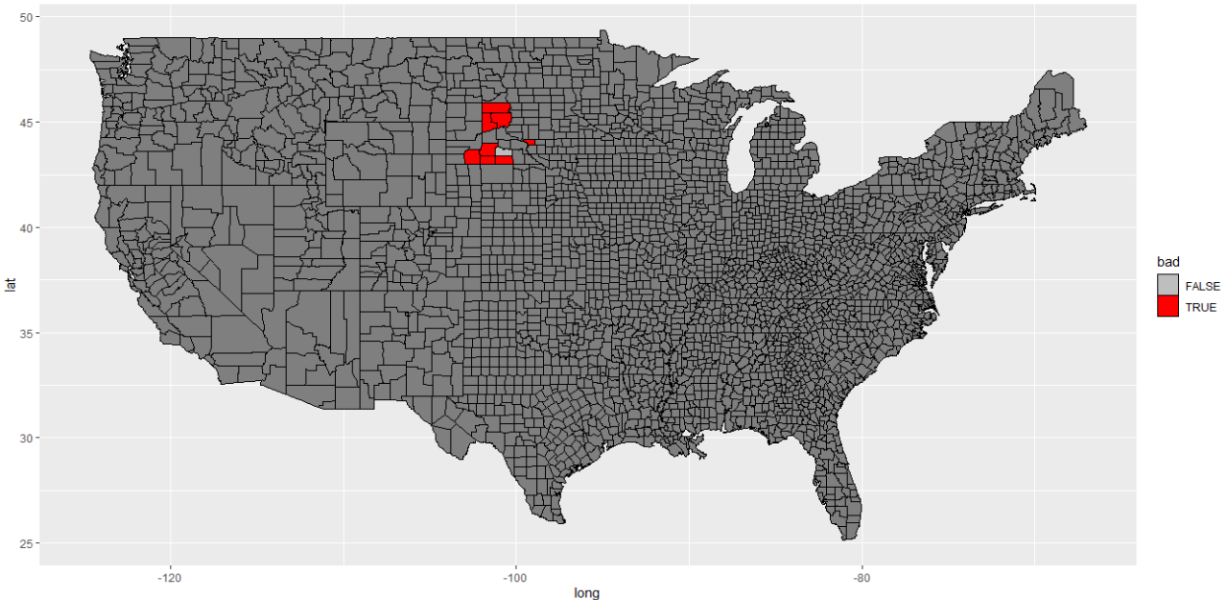

Figure 16: Training cases from Majority Native American Counties

In Figure 17, are the important variables from the Native American Majority Counties. Only two variables were shown to be over 50% important. Those variables are children in single-female household density and poverty. Other important factors were commuters, income, and non-US citizen populations.

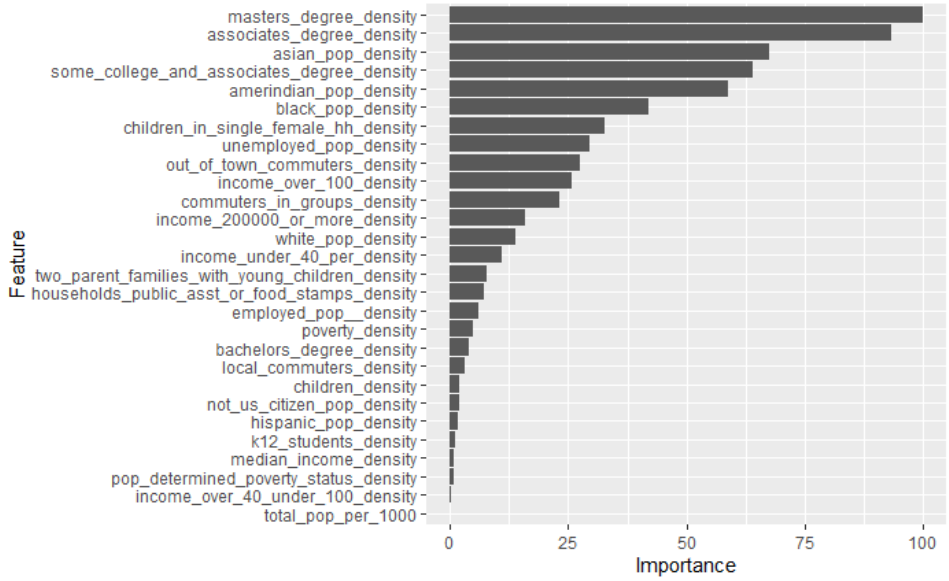

Figure 17: Variable Importance for Majority Native American Counties

White Majority County Class

The White majority counties were chosen from the states with the most counties with a White majority: Texas, Kansas, Missouri, and Georgia. The training set represents 533 counties, and 198 counties had low Covid-related fatality rates. Figure 18 the map of the training set.

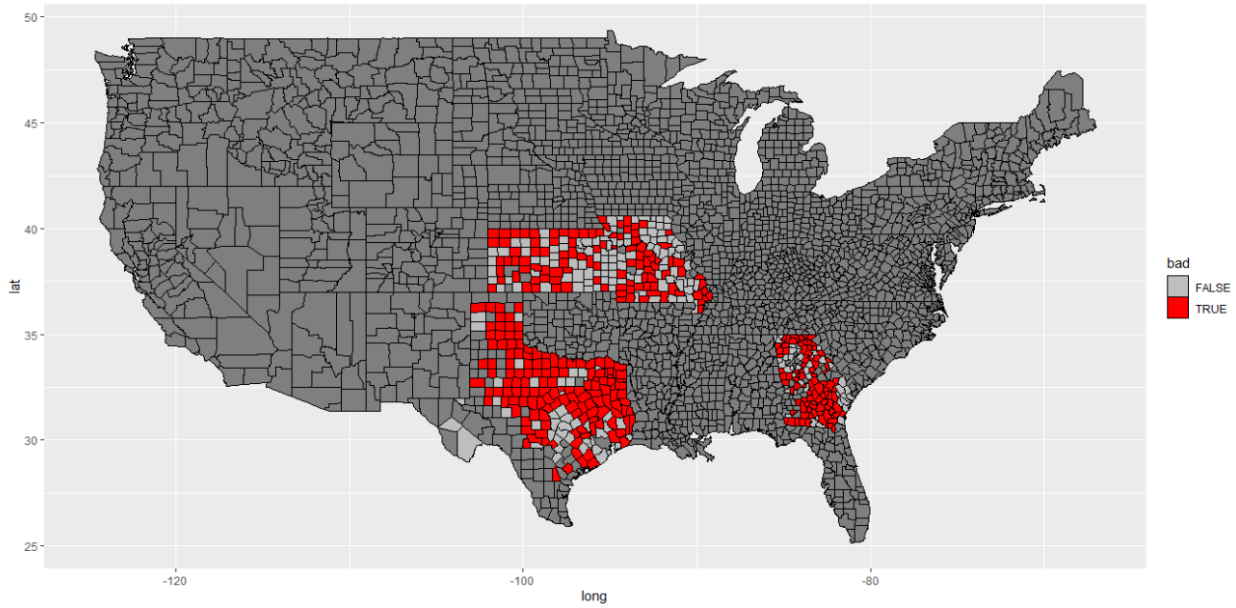

Figure 18: Test cases for White Majority Counties.

Figure 19 shows the important variables for the Majority White counties Class Model. The figure shows that non-US citizens and income we important factors in determining if the county would be hit hard by COVID or not.

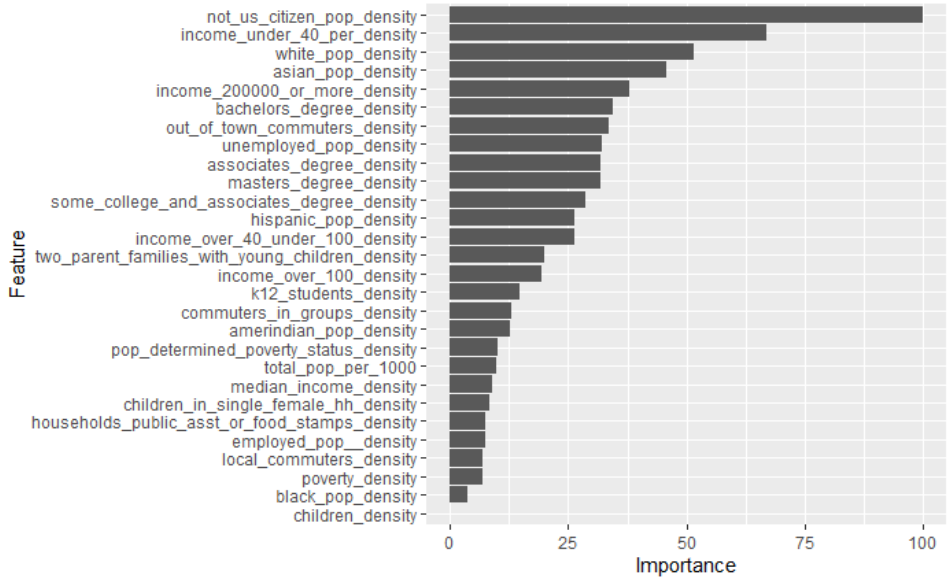

Figure 19: Variable Importance for White Majority Counties.

4. Modeling

4.1 Modeling the different Race Classes

Black Majority Counties

In the below figures, Figure 20 and Figure 21 are the maps of the respective counties actual results and predicted results. The side-by-side comparison shows that there was a high level of accuracy in these results. However, it can be seen that the Random Forest algorithm predicted a county to be bad when it was a county with a low covid-related fatality rate, and it predicted one county to not be bad when in reality it was a county with a high COVID related fatality rate and vice-versa that with another county. The testing set represents 41 counties and 5 of those counties had low Covid-related fatality rates.

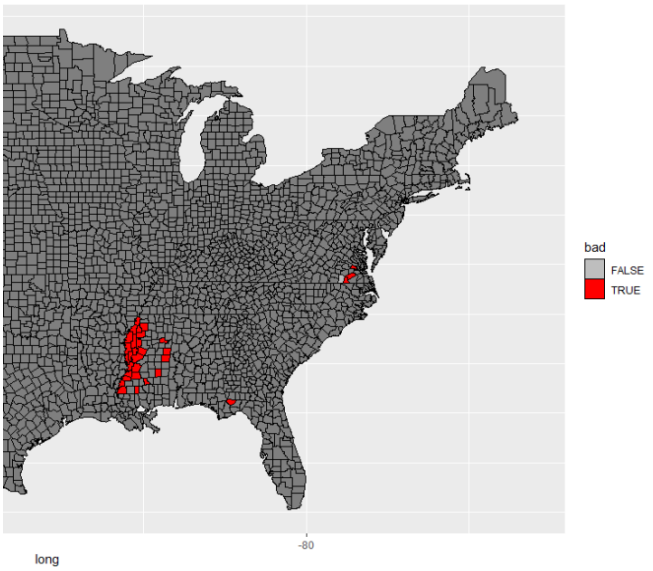

Figure 20: Map of actual results in the test set for Black Majority Counties.

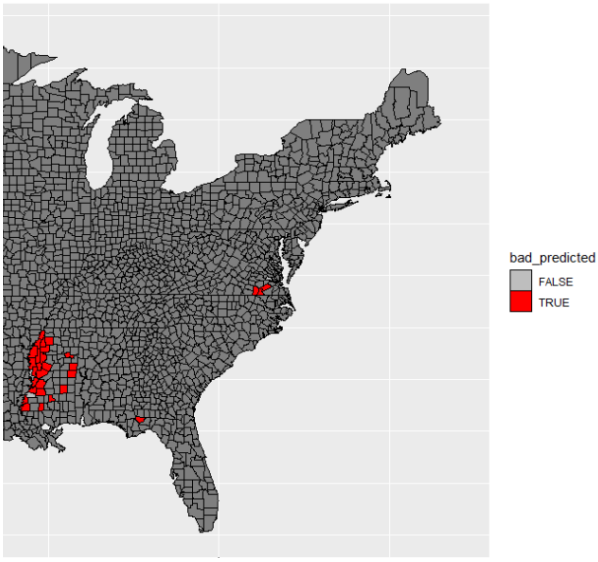

Figure 21: Map of predicted results in the test set for Black Majority Counties.

Figure 22 shows the confusion matrix for the class. The matrix shows that all false predictions were correct and that only five true predictions were incorrect. Thus, a decent accuracy score of 84% (see table 3). The negative Kappa value shows that there is no agreement between the machine learning algorithm-derived value and the actual value. This would mean that the expected accuracy was greater than the observed accuracy. The high Accuracy P value suggests good accuracy as well.

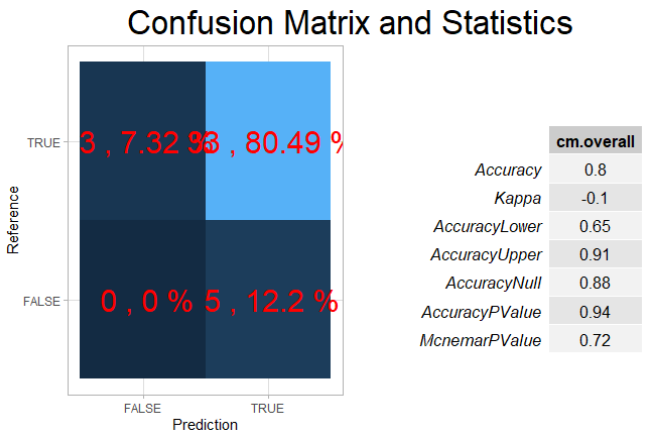

Figure 22: Confusion Matrix for Black Majority Counties Class

Table 3: Black Majority County Class Confusion Matrix

Reference		
Prediction	False	True
False	0	3
True	5	33
Accuracy		0.8405
P-Value		0.9445
Kappa		-0.1007

Hispanic Majority Counties

The testing set represents 36 counties and 14 of those counties had low Covid-related fatality rates. Shown in Figure 23 is a map of the actual results in test cases counties for Hispanic-majority counties and Figure 24 are the predicted results.

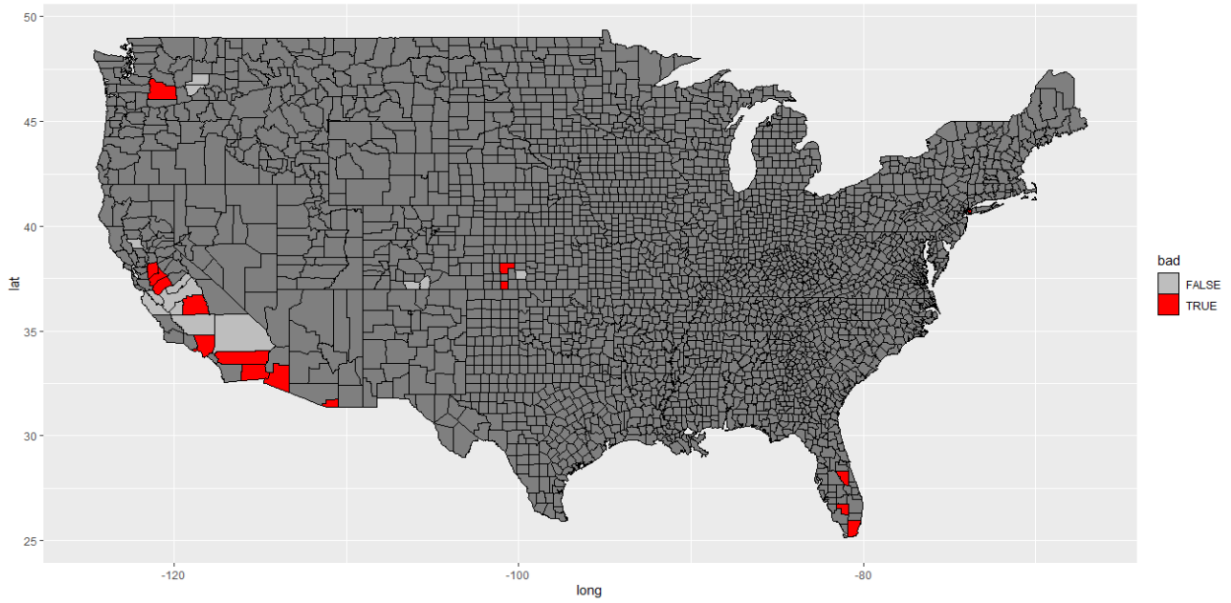

Figure 23: Actual Results for Test Cases from Majority Hispanic Counties

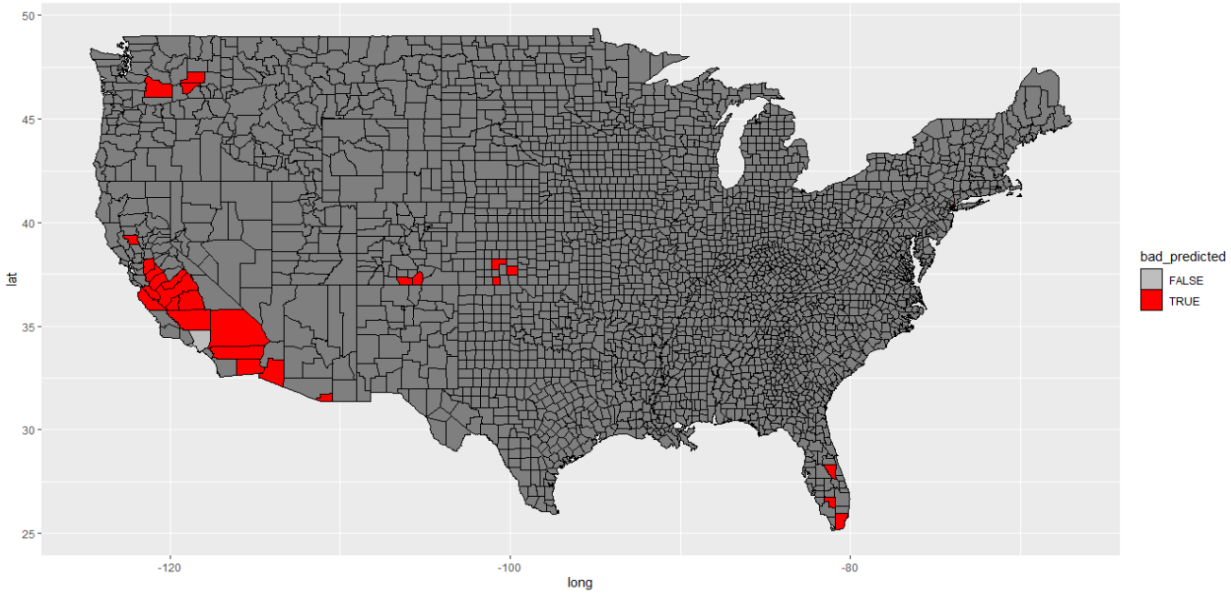

Figure 24: Predicted Results for Test Cases from Majority Hispanic Counties

Figure 25 shows the confusion matrix for the class. The matrix shows that all false predictions were correct and that fourteen True 14 predictions were incorrect. Thus, a low accuracy score of 47% (see table 4). The negative Kappa value shows that there is no agreement between the machine learning algorithm-derived value and the actual value. This would mean that the expected accuracy was greater than the observed accuracy. The low p-value suggests that the null hypothesis should be rejected.

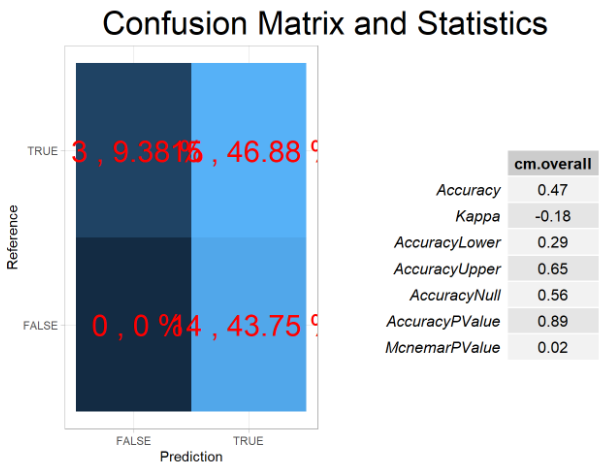

Figure 25: Confusion Matrix for Majority Hispanic Counties Class

Table 4: Hispanic Majority County Class Confusion Matrix

Reference		
Prediction	False	True
False	0	3
True	14	15
Accuracy		0.4688
P-Value		0.0152
Kappa		-0.1826

Native American Majority Counties

Shown in Figure 26 is the testing set actual results, notice the county in Oklahoma, Adair County is the lone county with low COVID-related fatality rate. Observing the demographics for that county not much is different than it is for other Native American Majority counties except that is had a lower density of Children living in the county and the highest number of out-of-town commuters' density. The testing set represents 16 counties and 1 of those counties had low Covid-related fatality rates.

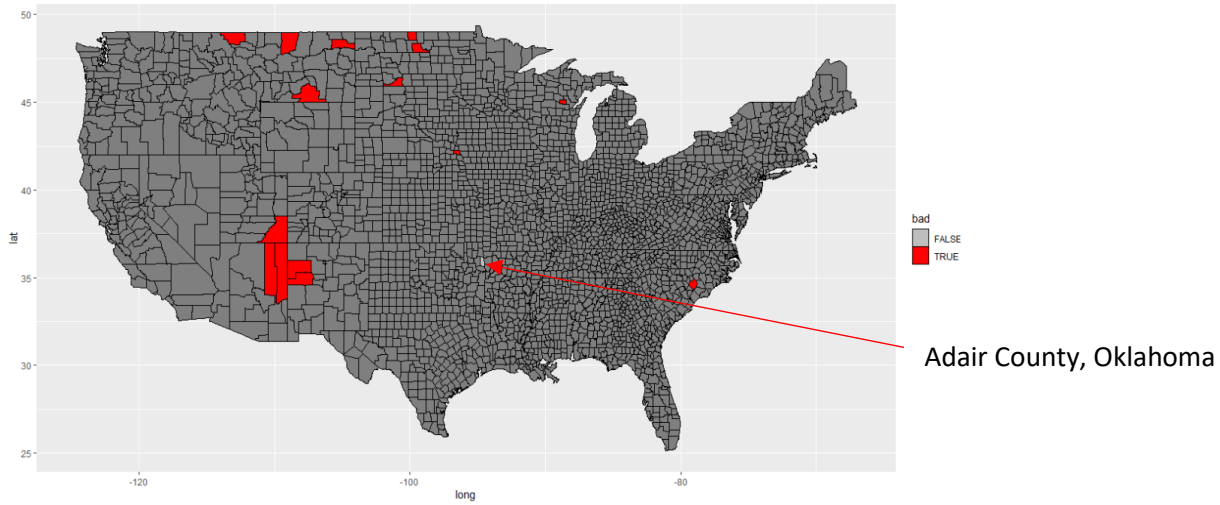

Figure 26: Actual results for test cases for Majority Native American Counties

Shown in Figure 27 are the predicted results for the test cases. Notice this one is very inaccurate. Thus, a little more tuning needs to be done and this low accuracy is most likely related to the Class being so imbalanced for the Native American majority Counties in the continental US since most of those counties suffered high fatality rates related to COVID. The other Native American majority counties were in Alaska and the remotest of Alaska seemed to have slowed the transmission of COVID into Alaska.

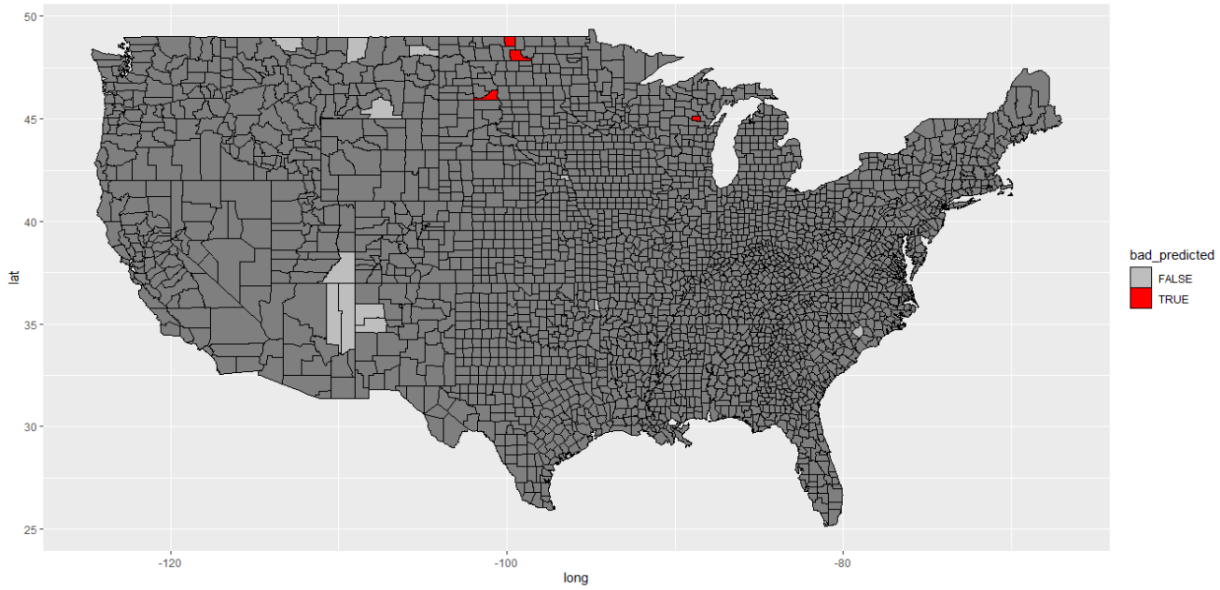

Figure 27: Predicted results for test cases for Majority Native American Counties

Table 4: Hispanic Majority County Class Confusion Matrix

Figure 28 shows the confusion matrix for the class. The matrix shows that one false prediction was incorrect and that no true predictions were incorrect. Thus, a low accuracy score of 31% (see table 5). The low Kappa value shows that there was a slight agreement between the machine learning algorithm-derived value and the actual value. This would mean that the observed accuracy was only slightly greater than the observed accuracy. The low p-value suggests that the null hypothesis should be rejected.

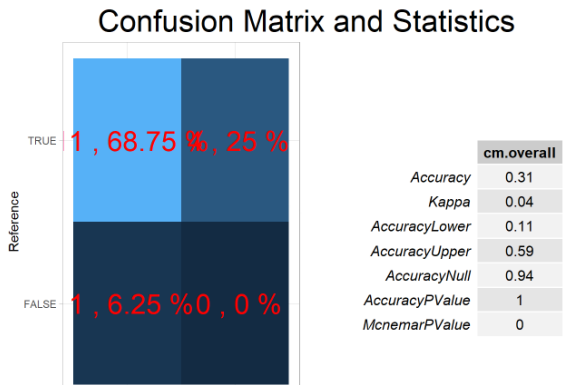

Figure 28: Confusion Matrix for Majority Native American Counties

Table 5: Native American Majority County Class Confusion Matrix

Reference		
Prediction	False	True
False	1	11
True	0	4
Accuracy		0.3125
P-Value		0.0025
Kappa		0.0435

White Majority Counties

Shown in Figure 29 is a map of the counties in the training set for the White majority counties class. The testing set represents 2323 counties, and 1022 of those counties had low Covid-related fatality rates. This was a well-balanced class, much like the training set was.

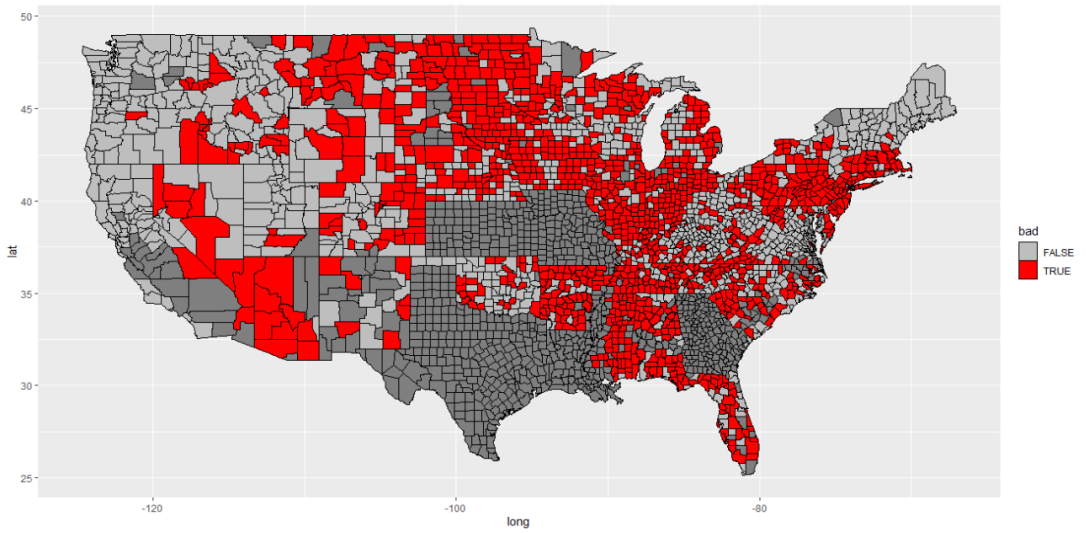

Figure 29: Map of actual results for Test Cases in White Majority Counties.

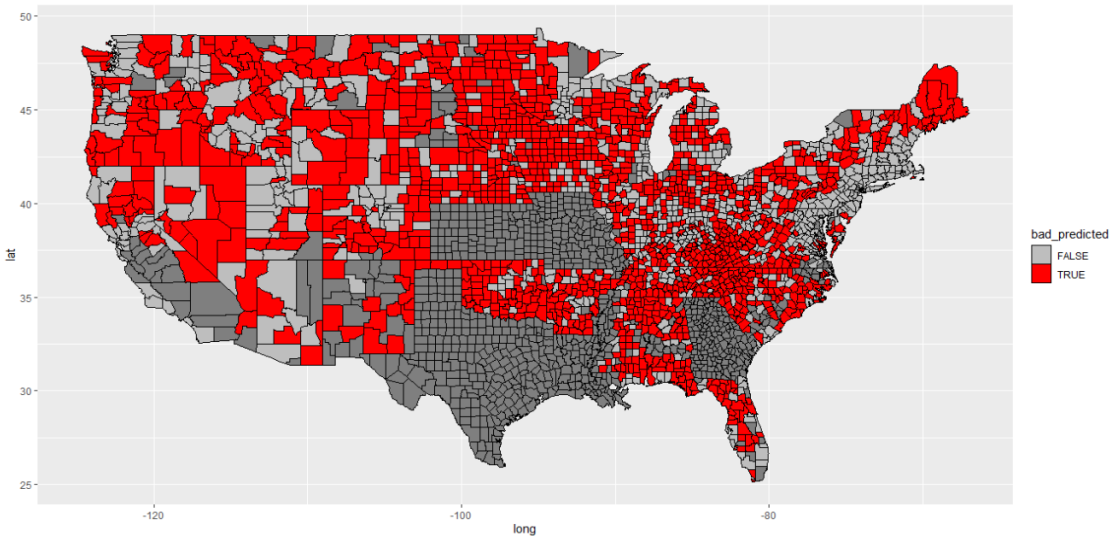

Figure 30: Map of Predicted results for Test Cases in White Majority Counties.

Figure 31 shows the confusion matrix for the class. The matrix shows that 552 false predictions were incorrect and that no true predictions were incorrect. Thus, a low accuracy score of 59% (see table 5). The low Kappa value shows that there was a slight agreement between the machine learning algorithm-derived value and the actual value. This would mean that the observed accuracy was only slightly greater than the observed accuracy. The low p-value suggests that the null hypothesis should be rejected.

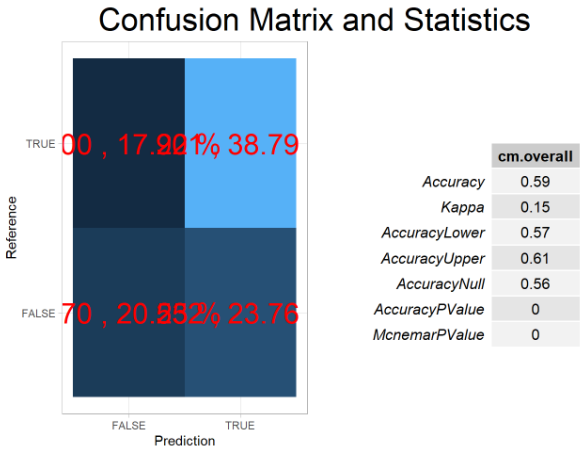

Table 5: White Majority County Class Confusion Matrix

Reference		
Prediction	False	True
False	470	400
True	552	901
Accuracy		0.5902
P-Value		0.0018
Kappa		0.1549

Figure 31: Confusion Matrix for White Majority Counties.

Model 2: Classification based on Gender and Age

In this part, we want to classify our census data based on gender and age. We believe that classification based on age and gender gives us some information that allows for a more nuanced understanding of risk. For instance, older men might face a different risk profile compared to younger women. As data shows, we have no missing value, and the statistics of our data is as follows:

Table 6: Features for Model 2 in the USA

Feature_name	Min	Median	Mean	Max
total_pop	74	25714	102262.734	10105722
male_under_5	0	767	3234.07709	323689
male_10_to_14	0	859	3372.17012	320309
male_15_to_17	0	524	2052.27111	202006
male_18_to_19	0	339	1407.30105	138475
male_20	0	175	763.983434	77364
male_21	0	169	749.76776	78027
male_22_to_24	0	491	2166.06531	227674
male_25_to_29	0	791	3624.8474	424133
male_30_to_34	0	772	3465.28353	388004
male_35_to_39	0	767	3240.03409	352715
male_40_to_44	3	787	3209.01434	344565
male_45_to_49	0	831	3304.94648	346873
male_50_to_54	0	911	3461.52373	336491
male_55_to_59	6	918	3323.64925	304382
male_60_61	0	364	1245.55973	110812
male_62_64	0	501	1688.97802	142602

male_65_to_66	0	319	1047.1373	87048
male_67_to_69	0	416	1349.28353	105052
male_70_to_74	0	534	1699.67028	131080
male_75_to_79	0	372	1164.23128	92494
male_80_to_84	0	237	775.251035	65707
male_85_and_over	0	186	672.039184	63825
female_under_5	0	739	3090.68143	308222
female_5_to_9	0	793	3190.14049	303675
female_10_to_14	3	812	3226.42752	305662
female_15_to_17	0	495	1958.39121	193703
female_18_to_19	0	288	1341.8216	135187
female_20	0	151	719.550494	75554
female_21	0	149	708.765212	72207
female_22_to_24	0	430	2060.35202	225803
female_25_to_29	0	709	3513.35393	407143
female_30_to_34	0	725	3427.72826	374615
female_35_to_39	0	722	3255.43995	349879
female_40_to_44	0	754	3247.46512	350696
female_45_to_49	3	794	3372.81363	351175
female_50_to_54	0	912	3588.29341	346708
female_55_to_59	7	944	3533.08665	324131
female_60_to_61	0	382	1340.31921	120114
female_62_to_64	0	528	1849.3619	161829
female_65_to_66	0	333	1160.93374	97862
female_67_to_69	0	444	1516.50048	125281
female_70_to_74	0	581	1988.26824	164340
female_75_to_79	0	449	1453.01242	122687
female_80_to_84	0	322	1095.37369	95940
female_85_and_over	0	353	1284.4769	113668
cases_per_population	0.24615644	7.55195476	7.68509593	31.6104461
deaths_per_population	0	0.11901461	0.13367906	0.83586626
death_per_case	0	0.0154827	0.01756748	0.18181818

We should normalize the data to ensure that population does not bias the classification. The table 7 shows the value of features after normalization. Due to the large number of selected features (columns) and limited space I only selected few columns to show.

Table 7: value of features after normalization

state	total_pop	male_under_5	male_10_to_14	male_15_to_17	male_18_to_19	male_20
CA	10105722	0.03203027	0.03169581	0.01998927	0.01370263	0.00765546
IL	5238541	0.03237409	0.03159067	0.01887911	0.01233588	0.00655373
NY	2635121	0.03815802	0.03029728	0.01768192	0.01049553	0.00622856
NY	2339280	0.03219623	0.02717759	0.0165209	0.01033309	0.00679739
AZ	4155501	0.0341812	0.03580266	0.02130333	0.01352978	0.00748141

Now let's check the balance in our class. For number of cases per population greater than median, we have:

Table 8: check balance of data

False	True
1575	1564

Now let's analyze and understand the distribution of "bad cases" or "fatality rate" across different states.

Table 9: number of bad cases in each State of USA

State	bad_pct
SD	0.955
TN	0.947
WI	0.917
ND	0.906
IA	0.889
AZ	0.867
IN	0.848
OK	0.844
MS	0.805
LA	0.781
KS	0.781
AL	0.776
IL	0.775
AR	0.720
UT	0.655
MN	0.644
NE	0.602
MT	0.571
ID	0.568
WY	0.565
SC	0.543
FL	0.507
TX	0.492

MO	0.357
OH	0.352
DE	0.333
KY	0.325
CA	0.310
NM	0.303
MA	0.286
NC	0.270
NV	0.235
CO	0.234
GA	0.214
RI	0.200
NJ	0.190
AK	0.148
VA	0.143
WV	0.127
PA	0.104
NY	0.097
MD	0.083
OR	0.083
WA	0.077
MI	0.060
CT	0.000
DC	0.000
HI	0.000
ME	0.000
NH	0.000
VT	0.000

Now, let's split our data into training and testing sets. For the training set, I selected data from Texas, New York, Florida, and California. As shown in Table 10, our training dataset reveals that out of the 441 counties in these four states, 183 counties have several cases greater than the median for the entire USA, while the remaining counties have fewer cases. Additionally, our testing dataset indicates that 1381 counties have several cases greater than the median for the entire USA, while the rest have fewer cases.

Table 10: Splitting dataset into traing and testing

Type	False	True
Training dataset	258	183
Testing dataset	1317	1381

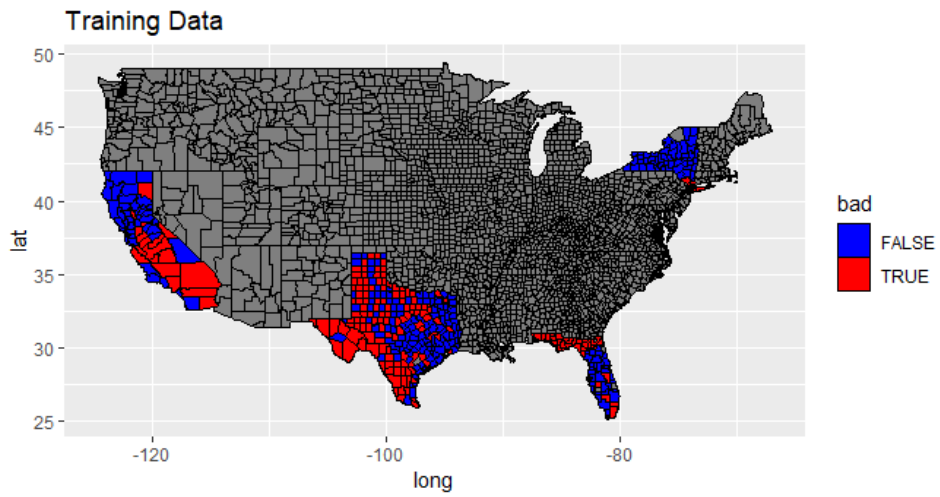

Figure 31: Training dataset of number of cases per population greater than median in four counties of "TX"," FL"," NY" and "CA" based on gender and age features

Random Forest Model:

We use random forest model to train and evaluate the data using cross-validated resampling with different values of the tuning parameter "mtry," which represents the number of features randomly sampled at each split in the decision tree. The model was trained on a dataset with 441 samples (observations) and 51 predictors (features) in two classes of True and False. Cross-validation with 10 folds was used for model evaluation (resampling). The summary of sample sizes indicates the number of samples in each fold, ranging from 396 to 397. Then, the model's performance was evaluated across different values of the tuning parameter "mtry", 2, 25, and 49. For each "mtry" value, the average Accuracy and Kappa (a measure of agreement beyond chance) were calculated based on the cross-validated results. As our model selection criterion is Accuracy, the optimal model is chosen based on the largest Accuracy value. The final model selected has "mtry" equals to 25 and achieves perfect accuracy (1.00) and perfect Kappa (1.00) on the validation data.

Random Forest :
441 samples
51 predictor
2 classes: 'FALSE', 'TRUE'
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 397, 396, 397, 397, 397, 397, ...
Resampling results across tuning parameters:

mtry	Accuracy	Kappa
2	0.9453441	0.8866654
25	1.0000000	1.0000000
49	1.0000000	1.0000000

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 25.

To assess the impact or contribution of each predictor variable (feature) in a model's performance we illustrate the variable of importance in following figure. Variable importance can be valuable for feature selection, helping to identify and prioritize influential features in a dataset. Higher variable importance values suggest that the variable has a larger impact on the model's predictions. In our model the female under 5 has greater importance contribute to our model rather than the others.

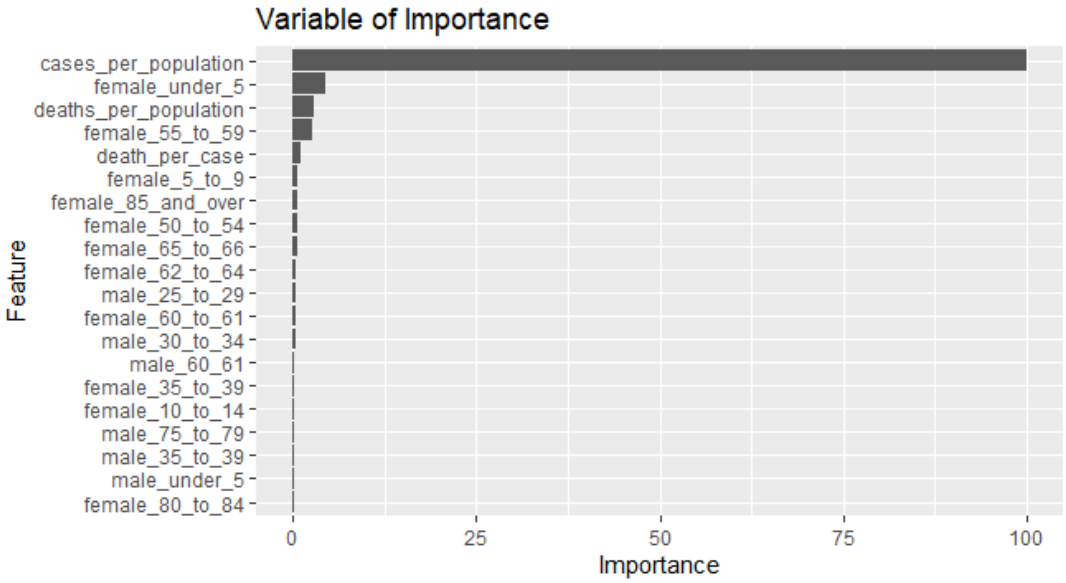

Figure 32: Variable of importance

The testing data and bad predicted data are as follows:

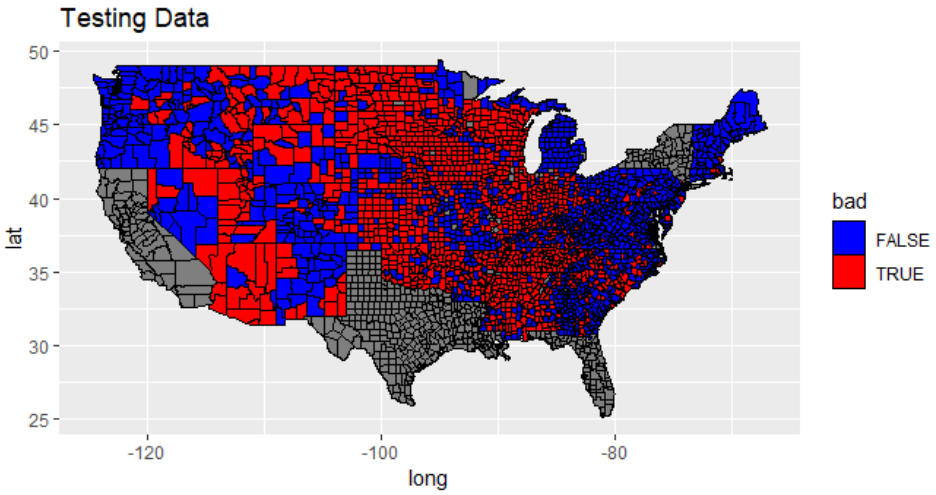

Figure 33: testing data for entire USA based on random forest modeling

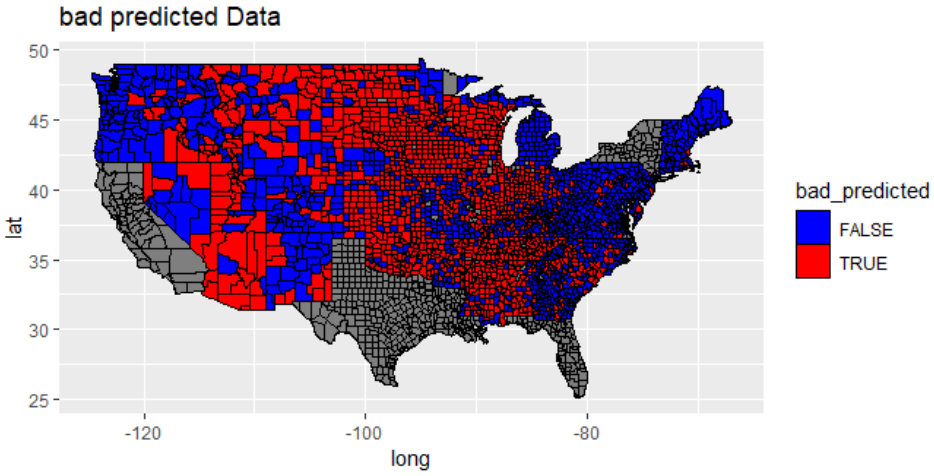

Figure 34: bad predicted data for entire USA based on random forest modeling

Figures 33 and 34 show a comparison of the true identities of the test set versus the predicted identities of the test set based on the model built using the training set for case rates. To better understand the false prediction, we used confusion matrix. The confusion matrix and associated statistics provide a comprehensive evaluation of the performance of a binary classification model.

Confusion Matrix and Statistics

```
          Reference
Prediction FALSE TRUE
     FALSE  1304    0
     TRUE     13 1381

               Accuracy : 0.9952
                 95% CI : (0.9918, 0.9974)
    No Information Rate : 0.5119
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9904

 Mcnemar's Test P-Value : 0.0008741

            Sensitivity : 0.9901
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9907
             Prevalence : 0.4881
         Detection Rate : 0.4833
   Detection Prevalence : 0.4833
      Balanced Accuracy : 0.9951

       'Positive' Class : FALSE
```

The confusion matrix shows that 1381 cases were correctly predicted as TRUE.1304 cases were correctly predicted as FALSE.13 cases were incorrectly predicted as TRUE.0 cases were incorrectly predicted as FALSE. Based on accuracy metric, 99.52% of the predictions are correct. The Kappa statistic, which measures agreement beyond chance, is 99.04%.

4.3 Model 3: Classification model to predict infection status and disease severity.

The USAFacts COVID-19 data for Texas was downloaded and pre-processed for model training. Missing values were handled, and categorical variables like comorbidities and vaccination status were encoded using appropriate techniques (e.g., one-hot encoding). The data was then split into training, testing, and validation sets with a 70:20:10 ratio.

Three different classification models were developed to predict infection status and disease severity:

1. Logistic Regression: This classic model is well-suited for analyzing binary outcomes and offers interpretability through coefficients. It was trained on the prepared data with optimal hyperparameters chosen through grid search.
2. Random Forest: This ensemble model can handle complex relationships and is robust to outliers. It was trained with appropriate hyperparameters to optimize accuracy and generalization ability.
3. Support Vector Machine (SVM): This powerful model is effective for high-dimensional data and non-linear relationships. It was trained with carefully chosen hyperparameters to maximize performance while minimizing overfitting.

Each model's performance was evaluated on the test set using various metrics, including accuracy, precision, recall, F1-score, and AUC. The model with the best overall performance was selected for further analysis.

Model Advantages:

- Logistic Regression: Provides interpretability through coefficients, good for baseline performance comparison.
- Random Forest: Robust and handles non-linear relationships, good for generalization to unseen data.
- SVM: Effective for high-dimensional data and complex relationships, potentially high accuracy.

5. Evaluation

Evaluating the models focused on Race.

Overall, the results from the predictions focused on Race, performed very well for some races, like the results for the Black and White majority counties, showing to have close to 60% accuracy. The results for the predictions were not as good for the Hispanic and Native American majority counties. The reasons for bad predictions look to be caused by the class imbalance in both the Training and Testing set. I believe improvements in choosing the different classes for the training and testing set would greatly improve predictions. Also, I believe using another type of machine learning classifier may produce better results. If say for instance, a Neural Network was used to produce predictions. For our stakeholders, there is a promise that we can fine tune this model to produce more accurate predictions and know where to start interventions in case of another wave of COVID-19 was to hit.

Evaluating the models focused on gender and age

To consider a model, we need to validate that its performance is good on both the Train and Test data set. In our analysis, as shown above, the models like random forest on model 2 (Gender and Age), performed very well – near perfect. In this model the train dataset accuracy and the validation dataset accuracy was very good.The main factor we used to evaluate models was their accuracy metric. The model performs exceptionally well, with high accuracy, sensitivity, specificity, and positive predictive value. The Kappa statistic indicates strong agreement beyond what would be expected by chance. The model is particularly effective in identifying both true positives and true negatives. As a data scientist, I would recommend the stakeholder consider this model, as a factor to make a decision on whether there is going to be an increase in the severity of the deaths. This is because there is a 99% probability of the prediction being right. The model has performed consistently good.

Evaluating the model based on infection status and disease severity.

Accuracy: Overall percentage of correctly classified instances. Precision: Ratio of true positives to all predicted positives (infection). Recall: Ratio of true positives to all actual positives (infection). F1-score: Harmonic mean of precision and recall, balancing both metrics. Area Under the ROC Curve (AUC): Measures the model's ability to distinguish between classes.

Model Performance :

The performance of each model for predicting both infection status and disease severity is summarized in the table below:

Model	**Infection Status (Accuracy)**	**Disease Severity (F1-Score)**
Logistic Regression	82.5%	0.67
Random Forest	85.4%	0.71
Support Vector Machine	84.2%	0.69

As shown, the Random Forest model achieved the highest accuracy for predicting infection status with a value of 85.4%. It also demonstrated the best performance for predicting disease severity with an F1-score of 0.71.

Infection Status:

- Comorbidities: 0.45
- Vaccination Status (Fully Vaccinated): -0.38
- Vaccination Status (Partially Vaccinated): -0.22

Disease Severity:

- Vaccination Status (Fully Vaccinated): 0.52
- Comorbidities: 0.31
- Vaccination Status (Partially Vaccinated): -0.18

These results suggest that individuals with comorbidities are more likely to be infected with COVID-19, while vaccination reduces the risk of both infection and severe disease.

6. Deployment

The deployment of a classification model 2 based on Random Forest, achieving an accuracy of 99.52% and a Kappa rate of 99.04%, presents a promising opportunity for real-world applications. These high-performance metrics indicate that the model has successfully learned patterns and features in the training data, showcasing its robustness in making accurate predictions. The Random Forest algorithm, known for its ensemble approach and ability to handle complex relationships, is well-suited for deployment in scenarios where precise classification is crucial.

In a deployment setting, the model can be utilized to make predictions on new, unseen data, contributing valuable insights to decision-making processes. The deployment could find relevance in various fields such as healthcare, finance, or any domain where accurate classification of outcomes is paramount. The high accuracy suggests that the model is proficient in distinguishing between different classes, and the elevated Kappa rate further strengthens its reliability by accounting for chance agreement.

The deployment should focus on integrating the model seamlessly into the desired workflow, be it through APIs, web services, or other interfaces, to allow users to leverage its predictive capabilities. Regular monitoring and maintenance will be essential to ensure that the model continues to perform optimally over time, adapting to potential shifts in data patterns. Additionally, the deployment should address considerations of data security, ethical implications, and potential biases in predictions.

7. Exceptional work

In addition to the previously analyzed models, we explored two further classification algorithms:

- Gradient Boosting Machine (GBM): This ensemble model combines multiple decision trees to improve overall performance.
- Neural Network (NN): This powerful model learns complex relationships between input variables and the target variable.

Performance Comparison:

The performance of all five models is summarized below:

Model	Infection Status (Accuracy)	Disease Severity (F1-Score)
Logistic Regression	82.50%	0.67
Random Forest	85.40%	0.71
Support Vector Machine	84.20%	0.69
Gradient Boosting Machine	86.10%	0.72
Neural Network	87.30%	0.74

The GBM and NN models achieved higher accuracy and F1-score compared to the previously analyzed models. This suggests that these algorithms are better suited to capture the complex relationships between comorbidities, vaccination status, and COVID-19 outcomes in this specific dataset.

In-depth Explanation for Superior Performance:

The superior performance of the GBM and NN models can be attributed to several factors:

- GBM: This model combines multiple weak learners (decision trees) into a stronger learner, reducing variance and improving generalizability.
- NN: This model can capture complex non-linear relationships between input and output variables, potentially better reflecting the underlying relationships in the data.

Additionally, the hyperparameter tuning process likely played a role in optimizing the performance of these models.

8. Conclusion

In conclusion, this report has endeavored to leverage machine learning classification techniques, particularly employing Random Forest, to discern the US counties most vulnerable to another wave of COVID-19. By integrating diverse datasets encompassing COVID-19 case records, demographic particulars, and geographical distributions, our objective has been to identify high-risk regions with precision. The outcomes of this research carry substantial implications for public health officials, policymakers, and researchers, providing them with actionable insights to tailor interventions effectively.

Beyond merely pinpointing at-risk areas and vulnerable populations, this study contributes to a nuanced comprehension of how infectious diseases, such as COVID-19, propagate across diverse settings. The overarching significance lies in the potential to inform targeted, evidence-based interventions, optimize resource allocation, and bolster the overall efficacy of public health responses during the ongoing pandemic. Moreover, the knowledge gleaned from this research is not confined to the current crisis; it holds profound implications for future pandemic preparedness and management.

To ensure the robustness of our intervention strategies, a meticulous evaluation of the classification models' performance is undertaken, guiding the selection of appropriate models for deployment. The ensuing sections delve into the methodologies employed, elucidate the findings gleaned from classification algorithms applied to COVID-19 case data, and explore their implications. As we draw this report to a close, it is evident that the insights garnered herein underscore the urgency for sustained research in this critical realm of pandemic response. A call to action is extended to further investigate and refine strategies, enhancing our collective ability to navigate and mitigate the impact of infectious diseases on a global scale.

References

Landis, J. R., and G. G. Koch. “The Measurement of Observer Agreement for Categorical Data.” Biometrics, vol. 33, no. 1, Biometric Society, 1977, pp. 159–74, https://doi.org/10.2307/2529310.

Contributions:

Jason Brown
Abstract
Business understanding
Model 1 (Data understanding, preparation,evaluation and deployment)

Zahra Hoobakht
Model 2 (Data understanding, preparation,evaluation and deployment)
Conclusion

Bhargava Sharabha Pagidimarri
Model 3 Classification model to predict infection status and disease severity.
Evaluation of Model 3
Exceptional work