COVID CLASSIFICATION

Table of Contents

Abstract :

In early December 2019, an outbreak of coronavirus disease 2019 (COVID-19), caused by a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), occurred in Wuhan City, Hubei Province, China. On January 30, 2020, the World Health Organization declared the outbreak a Public Health Emergency of International Concern. As of February 14, 2020, 49,053 laboratory-confirmed and 1,381 deaths have been reported globally. The perceived risk of acquiring disease has led many governments to institute a variety of control measures.

We are creating a report for Government officials to indicate/Predict the situation of Covid . ie. either rise or fall in cases based on the different variables of factors that we have considered in our analysis.

We have created a model which gives us better output and can efficiently predict the rise in covid cases with almost a good Accuracy rate. We have made use of several classification models and chosen the best models that give us better results

Problem Statement & Stakeholder :

The main target of this analysis is to create a model that could efficiently predict the rise or fall in covid 19 cases in the target counties and state so that our stakeholder that is the Government official could take necessary action and formulate policies to control the Growth of Covid 19 in their respective Region. We have made use of several Classification models to analyze the output from different models and compared them all together to select the best model that could predict with higher accuracy and classify the counties where there are chances of more than 38 people per 10000 getting infected with covid 19 in the upcoming month.

Background Information :-

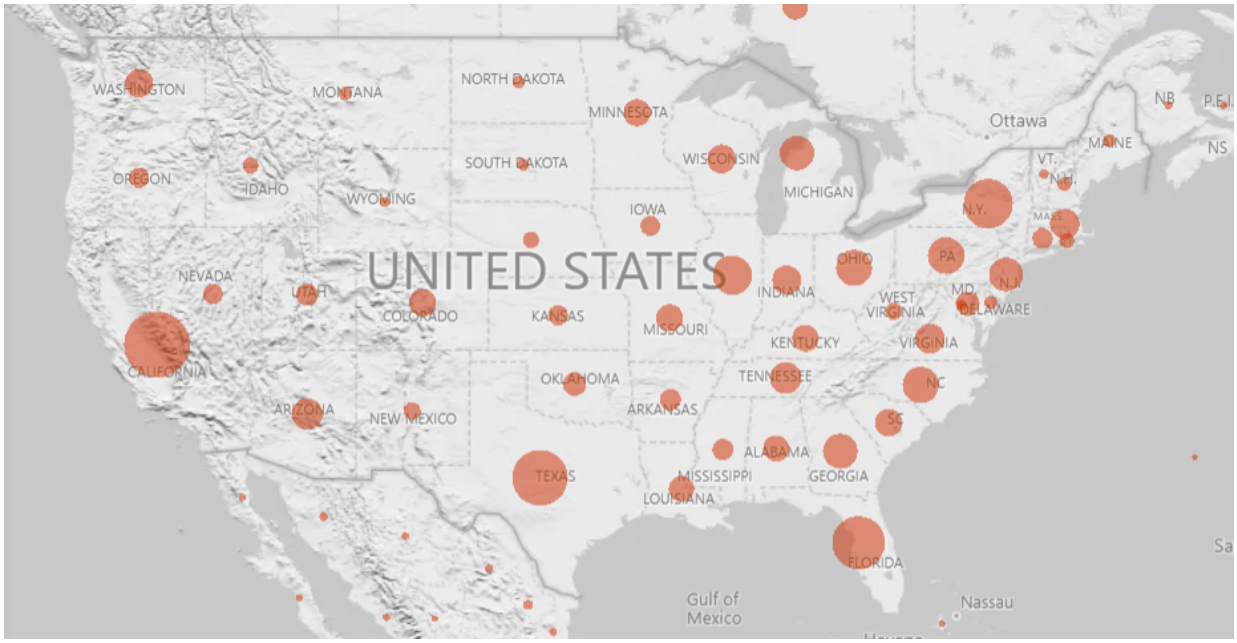

Figure 1.1 : Impact of Number of Cases on the US States

There have been five human coronavirus epidemics in the last twenty years: HCoV-NL63 and HCV-HKU1, which cause mild respiratory symptoms and circulated through the population, and the more serious SARS (2003), Middle East respiratory syndrome (MERS) (2012), and current SARS-CoV-2 in 2019 [CDC, 2019]. SARS-CoV is thought to have origins in bats and civet cats, though no final conclusion is agreed upon, and MERS originated in camels [WHO, 2019]. While it is currently unknown the precise route by which SARS-CoV-2 is transmitted from animals to humans, it is argued to have an origin within the wet markets of Wuhan, China[CDC, 2021], markets where live animals are butchered and bought. Early speculation suggested two routes by which the SARS-CoV-2 originated: either natural selection in a human following zoonotic acquisition or natural selection in an intermediate animal host prior to zoonotic transmission to a human [CDC, 2019].

Many of the first cases of SARS-CoV-2 visited a live animal market in Wuhan, which houses many live animals, suggesting that an animal at the market may have transmitted the virus to these first individuals [CDC, 2021]; however, it is now believed this was not the location of the first zoonotic transmission and rather acted as a super spreader event. As the most notable addition to the coronavirus family, the SARS-CoV-2 pandemic has brought many questions and doubts over the origin of the virus, the threat it poses to human beings, and the risks of fast-approved vaccination on humans. The SARS-CoV-2 virus causes the outbreak of COVID-19 (Coronavirus disease); an infectious disease caused by the SARS-CoV-2 virus

While most coronaviruses circulate amongst animals such as pigs, bats, and cats, recent and notable coronaviruses have shown abilities to infect humans, like in the outbreak of SARS-Covid in February 2003[CDC, 2005] and MERS-covid in September 2012[CDC, 2019]. Furthermore, similar to its siblings, SARS-CoV-2 affects the respiratory system, bringing about symptoms such as fever, fatigue, dry cough, and the novel loss of smell and taste [CDC, 2019].

As of Today, i.e December 1St 2022, There are 61,49,98,361 globally confirmed cases and a total death count of 65,36,643. In The United States, the Death count has reached 10,56,416 which is a large population of US residents.

Data Preparation :-

1.1 Data Description :

The data we use is the COVID-19 plus census data until November 20th, 2022. From the raw dataset, we have computed new variables by calculating each variable per 10,000 population. We have also created Week Over Week change in cases and death for the last month, to account for the trend Table1 has all the features considered. There are few a variable that we have combined togetherfollowswing:

- **income_less_50000_per_10000** – this variable has all the variables that contain income details for less than 50000 per 10,000 population
- **income_50000_100000_per_10000** – this variable has all the variables that contain income details for between 50000 and 10000 per 10,000 population
- **income_100000_or_more_per_10000** - this variable has all the variables that contain income details for more than 10000 per 10,000 population
- **commute_more_10_mins_per_10000** - this variable has all the variables that contain commute details for more than 10 mins per 10,000 population
- **dwellings_less_10_units_attached_per_10000** - this variable has all the variables that contain dwelling units that are less than 10 units attached per 10,000 population
- **dwellings_more_10_units_per_10000** - this variable has all the variables that contain dwelling units that are more than 10 units attached per 10,000 population
- **male_under_20_per_10000** - this variable has all the variables that contain male population under age 20 per 10,000 population
- **male_20_to_59_per_10000** - this variable has all the variables that contain male population between age 20 and 59 per 10,000 population
- **male_over_59_per_10000** - this variable has all the variables that contain male population over age 59 per 10,000 population
- **female_under_20_per_10000** - this variable has all the variables that contain female population under age 20 per 10,000 population
- **female_20_to_59_per_10000** - this variable has all the variables that contain female population between age 20 and 59 per 10,000 population
- **female_over_59_per_10000** - this variable has all the variables that contain female population over age 59 per 10,000 population
- **WOW1_Cases** – this variable is the difference between the covid cases of week 2 and week 1 of October month
- **WOW1_Deaths** – this variable is the difference between the covid deaths of week 2 and week 1 of October month
- **WOW2_Cases** – this variable is the difference between the covid cases of week 3 and week 2 of October month
- **WOW2_Deaths** – this variable is the difference between the covid deaths of week 3 and week 2 of October month

- **WOW3_Cases** – this variable is the difference between the covid cases of week 4 and week 3 of October month
- **WOW3_Deaths** – this variable is the difference between the covid cases of week 4 and week 3 of October month

We have utilized the covid cases and death variables for the past month I.e., Oct 1st to 31st on week over week difference. Keeping in mind to not add future data and made sure there is no data leak.

The following table shows the main characteristics of attributes and their values before normalization.

Variable	min	1st quartile	median	Mean	3rd quartile	Max
total Population	74	10955	25691.5	96920.1	67183.25	5238541
cases per 10000	286.8687	2365.5	2791.583	2792.868	3191.206	47297.3
deaths per 10000	0	26.90573	38.13768	38.60813	49.26858	150
death per case	0	0.009728	0.013308	0.01457	0.017474	0.090278
family households per 10000	1992.189	2442.403	2605.201	2580.676	2742.543	3192.386
rent under 15 percent per 10000	22.55427	222.1252	278.1136	286.1997	345.3583	530.1554
median age	21.6	37.9	41.2	41.14989	44.2	66.4
income less 50000 per 10000	792.5649	1695.529	2016.288	1986.649	2298.403	3200.345
black Population per 10000	0	60.38945	212.413	660.8104	987.0939	2380.16
asian Population per 10000	0	27.03992	57.53309	88.37515	122.0795	265.2639
hispanic Population per 10000	0	204.6069	396.0041	671.0436	925.3349	2014.742
other race Population per 10000	0	0	3.977656	8.110388	12.81876	32.14351
commuters by public transportation per 10000	0	4.078303	13.89617	22.47279	32.48834	75.4069
median income	19264	41118	48048.5	49277.18	55753	77721.63
income per capita	9334	21805	25268.5	25846.51	29109.5	40085.63
million dollar housing units	0	2	19	45.8866	67	165.125
families with young children per 10000	348.4174	587.6032	667.0508	671.0562	747.0813	986.452
income 50000 100000 per 10000	613.7913	1042.19	1188.446	1181.533	1328.539	1756.178
income 100000 or more per 10000	105.9274	484.0368	641.4545	689.3075	835.6618	1364.303
commute less 10 mins per 10000	130.4803	570.082	810.4533	970.2051	1253.688	2274.838
commute more 10 mins per 10000	605.2405	2841.132	3364.022	3346.861	3889.433	6286.019
dwellings less 10 units attached per 10000	0	269.4514	402.384	466.4919	607.5785	1114.795
dwellings more 10 units per 10000	0	65.80859	134.7287	190.1562	271.2225	581.0954
male under 20 per 10000	870.2765	1177.707	1285.651	1284.043	1382.955	1690.897
male 20 to 59 per 10000	1970.556	2389.662	2518.493	2542.509	2668.318	3088.267

Table 1.1 : Summary of Variables used in Analysis

As you can observe, there are few variables, which has outliers, based on the difference in the mean and the 3rd quartile. We have utilized the “Quartile based flooring and capping” for treating the outliers.

If we observe variables like black_population_per_10000, asian_population_per_10000, hispanic_population_per_10000, other_race_population_per_10000, commuters_by_public_per_10000, million_dollar_hosing_units, dwellings_less_10mins_per_10000, and dwellings_more_10units_per_10000 have minimum values and the first quartile to be 0. Hence, treating the outliers for these kinds of variables and replacing them with median, mean or mode, wouldn’t capture the reality. We have identified all the rows having less than the 10th percentile and greater than 90th percentile as outliers. And replacing these values with the 10th percentile and the 90th percentile.

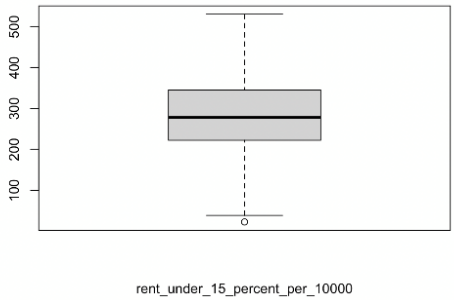

Figure 1.1.1 : Outlier Plot for Rent under 15 Percent per 10000

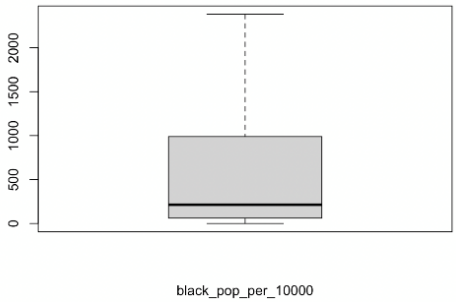

Figure 1.1.2 : Outlier Plot for Black Pop per 10000

Figure 1.1.3 : Outlier Plot for Income less than 50000 per 10000

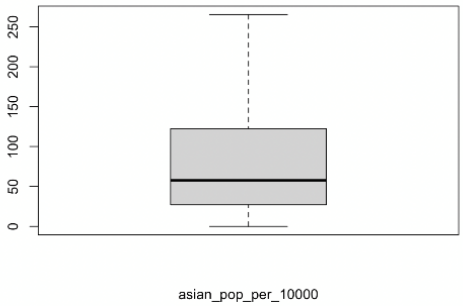

Figure 1.1.4 : Outlier Plot for Asian population per 10000

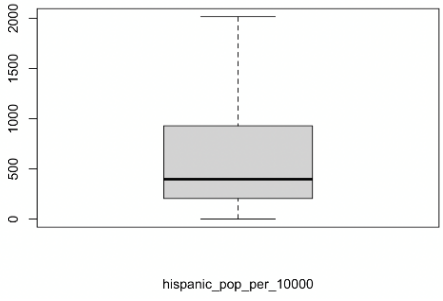

Figure 1.1.5 : Outlier Plot for Hispanic pop per 10000

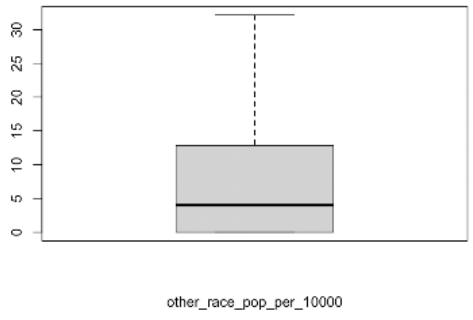

Figure 1.1.6 : Outlier Plot for Other race Pop per 10000

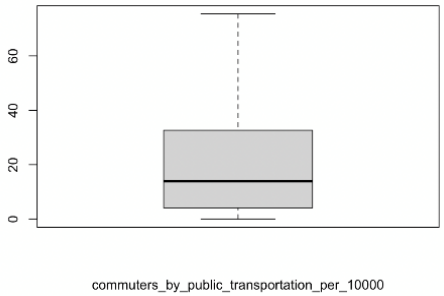

Figure 1.1.7 : Outlier Plot for Commuters per public Transportation per 10000 Transportation per 10000

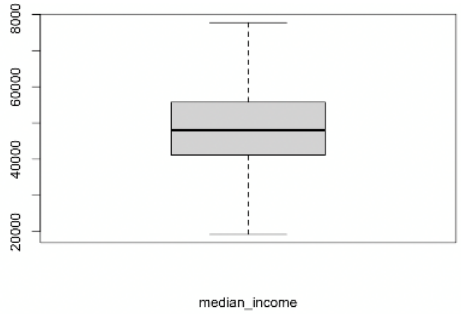

Figure 1.1.8 : Outlier Plot for Median income

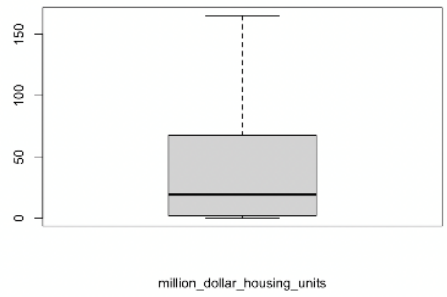

Figure 1.1.9 : Outlier Plot for million Dollar Housing Unit

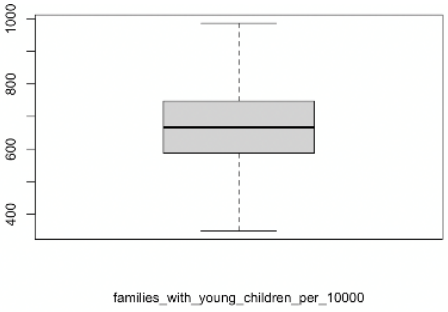

Figure 1.1.10 : Outlier Plot for families with young Children

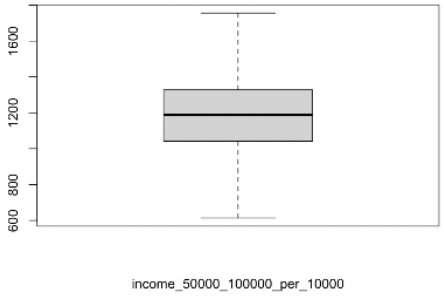

Figure 1.1.11 : Outlier Plot for Income 50000 to 100000 per 10000

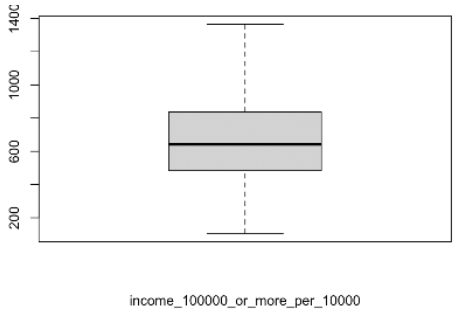

Figure 1.1.12 : Outlier Plot for Income 100000 or More per 10000

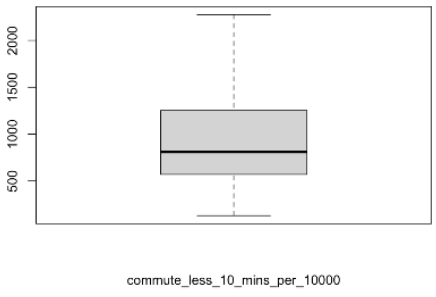

Figure 1.1.13 : Outlier Plot for Commute less than 10 min per 10000

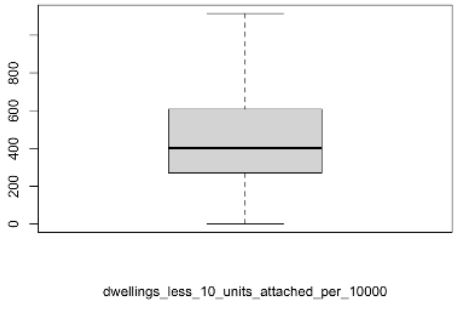

Figure 1.1.14 : Outlier Plot dwelling Less 10 unit per 10000

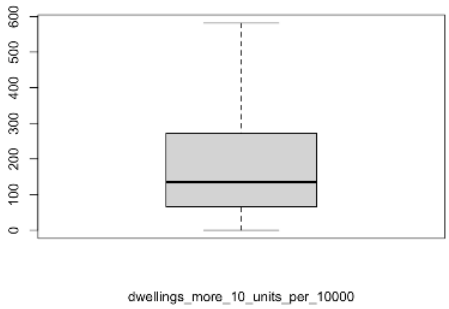

Figure 1.1.15 : Outlier Plot for Dwelling more 10 Units per 10000

Figure 1.1.16 : Outlier Plot for Male under 20 per 10000

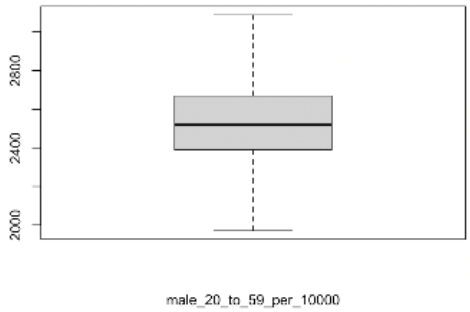

Figure 1.1.17 : Outlier Plot for Male 20 to 59 per 10000

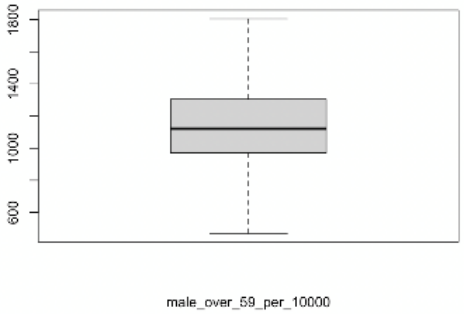

Figure 1.1.18 : Outlier Plot for Male over 59 per 10000

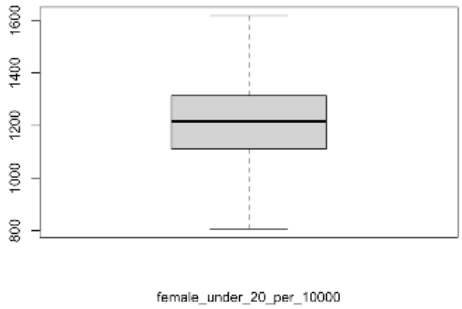

Figure 1.1.19 : Outlier Plot for Female under 20 per 10000

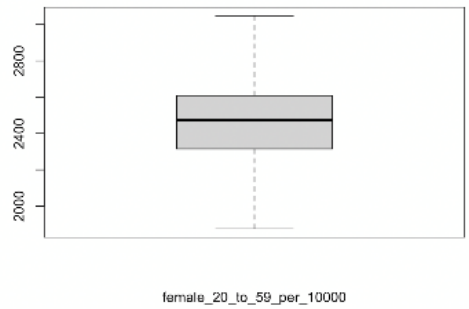

Figure 1.1.20 : Outlier Plot for Female 20 to 59 per 10000

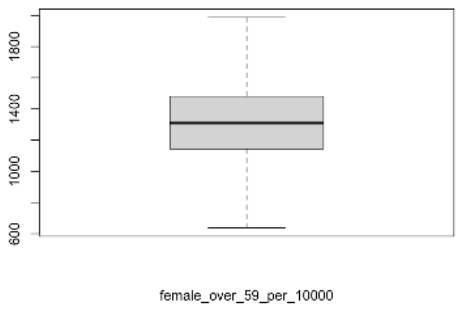

Figure 1.1.21 : Outlier Plot for Female over 59 per 10000

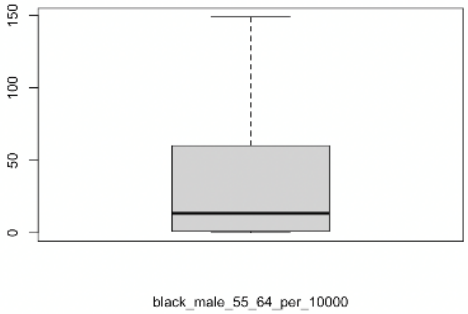

Figure 1.1.22 : Outlier Plot for Black male 55 to 64 per 10000

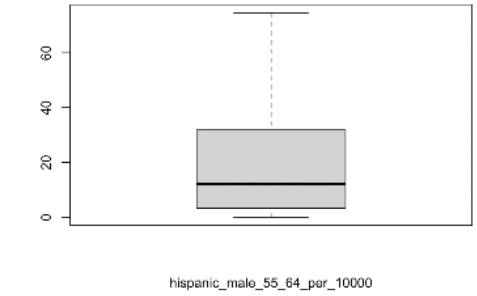

Figure 1.1.23 : Outlier Plot for Hispanic male 55 to 64 per 10000

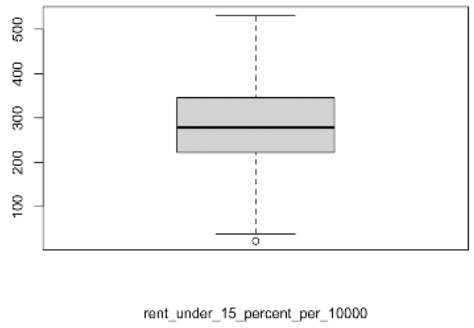

Figure 1.1.24 : Outlier Plot for Rent under 15 percentper 10000

We normalized the data and identified the top 20 features using feature selector with importance greater than 25%.

Features	Feature Importance
million_dollar_housing_units	0.421
asian_pop_per_10000	0.406
commuters_by_public_transportation_per_10000	0.391
income_100000_or_more_per_10000	0.378
commute_more_10_mins_per_10000	0.363
WOW1_Cases	0.362
commute_less_10_mins_per_10000	0.362
median_income	0.362
total_pop	0.35
income_per_capita	0.349
female_20_to_59_per_10000	0.337
hispanic_male_55_64_per_10000	0.323
dwellings_less_10_units_attached_per_10000	0.322
dwellings_more_10_units_per_10000	0.321
other_race_pop_per_10000	0.314
hispanic_pop_per_10000	0.311
WOW2_Cases	0.309
WOW3_Cases	0.306
asian_male_55_64_per_10000	0.294
male_20_to_59_per_10000	0.251

Table 1.2 : Feature Importance

The correlation matrix for all the 20 variables are as follows:

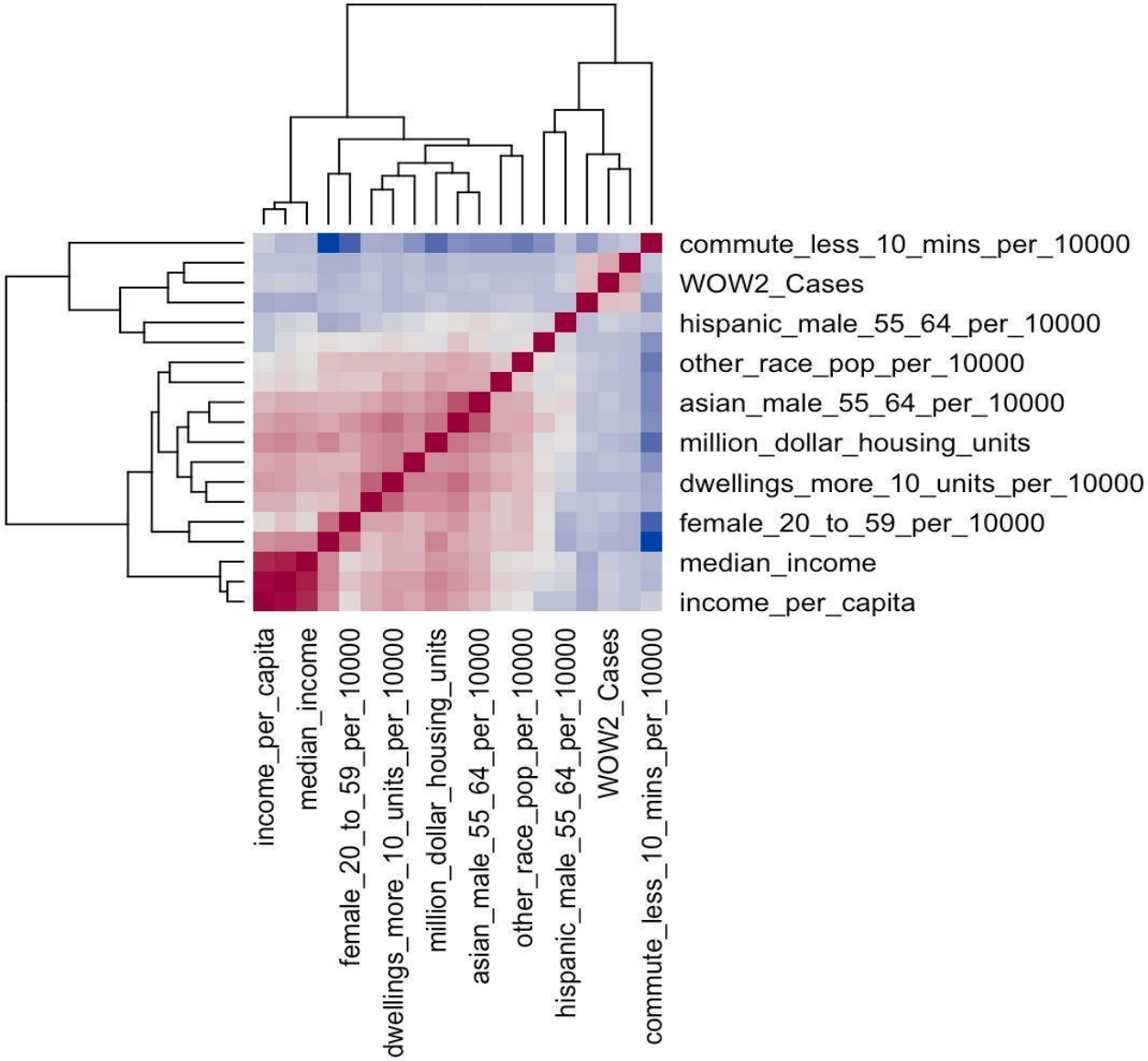

Figure 1.2: Correlation between the variables in the table

To make balanced class variable, we took the median of three variables namely, Deaths per 10000, Cases per 10000, and cases in the last month and made the split based on those. As a result, we appointed the counties with death per 10000 by more and less than 38, cases per 10000 by more and less than 2791, and cases in the last month per by more and less than 0.0024, as counties with good and bad situation in terms of covid spread. The results are as follows:

Class variable	counties with Less than 37 (good)	counties with More than 37 (bad)
Death per 10000	1551	1567

Table 1.2: Counties with a good and bad situation in terms of covid spread

The following table also shows the states with the highest percentage of counties with more than 38 deaths per 10000.

STATE	%	STATE	%	STATE	%	STATE	%	STATE	%
TN	0.884211	WV	0.672727	MT	0.517857	ID	0.372093	WA	0.051282
MS	0.865854	MO	0.672566	IN	0.5	VA	0.354839	UT	0.034483
AL	0.835821	DE	0.666667	NV	0.470588	IL	0.333333	CA	0.018182
AR	0.813333	PA	0.666667	KS	0.457143	WI	0.225352	FL	0.014925
AZ	0.8	LA	0.65625	NJ	0.428571	MA	0.142857	NE	0.010753
TX	0.766798	SC	0.652174	ND	0.415094	MD	0.125	AK	0
GA	0.754717	NM	0.636364	WY	0.409091	MN	0.103448	CT	0
KY	0.733333	SD	0.621212	CO	0.40625	NH	0.1	DC	0
OH	0.681818	IA	0.565657	RI	0.4	OR	0.083333	HI	0
OK	0.68	MI	0.542169	NC	0.38	NY	0.080645	ME	0

Table 1.3: State with the highest percentage of counties

The table above shows some states such as Tennessee and Mississippi have the highest percentages of bad covid situations in terms of covid deaths, while some counties such as Hawaii and Connecticut did a good job controlling covid.

The following table is for cases per 10000.

Class variable	counties with Less than 2791 (good)	counties with More than 2791 (bad)
cases per 10000	1557	1561

Table 1.4: Count of Counties with cases Per 10000

The following table also shows the states with the highest percentage of counties with more than 2791 cases per 10000.

STATE	%	STATE	%	STATE	%	STATE	%	STATE	%
DE	1	AL	0.776119	WY	0.545455	MT	0.375	CA	0.145455
RI	1	NJ	0.761905	MA	0.5	PA	0.363636	CT	0.125
WV	0.927273	MS	0.756098	SD	0.469697	MO	0.353982	VT	0.071429
KY	0.925	NC	0.75	MN	0.45977	ID	0.348837	OR	0.055556
TN	0.905263	KS	0.714286	SC	0.456522	VA	0.330645	GA	0.037736
WI	0.887324	OK	0.693333	OH	0.454545	NY	0.306452	NE	0.032258
AZ	0.866667	ND	0.679245	CO	0.453125	NV	0.294118	DC	0
FL	0.865672	AR	0.666667	TX	0.411067	IN	0.293478	HI	0
IL	0.852941	AK	0.62963	MI	0.385542	IA	0.282828	MD	0

LA	0.828125	**NM**	0.575758	**UT**	0.37931	**WA**	0.230769	**ME**	0

Table 1.5: States with the highest percentage of counties

The table above shows all counties in some states such as Delaware and Rhode Island have all more than 2791 cases per 10000, while some counties such as Maryland and Hawaii did not have even one county with more than 2791 cases per 10000.

The following table is for cases per population in the last month (October 20th, 2022, to November 20th, 2022.

Class variable	counties with Less than 0.0024 (good)	counties with More than 0.0024 (bad)
cases last month	1542	1576

Table 1.6: Cases per population

The following table also shows the states with the highest percentage of counties with more than 0.0024 cases per population in the last month.

STATE	%	STATE	%	STATE	%	STATE	%	STATE	%
CT	1	**KY**	0.958333	**MI**	0.831325	**ID**	0.418605	**DC**	0
DE	1	**MO**	0.946903	**VA**	0.822581	**UT**	0.37931	**GA**	0
HI	1	**OH**	0.943182	**SD**	0.772727	**OK**	0.333333	**IN**	0
ME	1	**MA**	0.928571	**MT**	0.714286	**AL**	0.298507	**LA**	0
NJ	1	**NM**	0.909091	**WY**	0.681818	**MS**	0.292683	**MD**	0
NY	1	**ND**	0.886792	**FL**	0.656716	**CA**	0.290909	**MN**	0
RI	1	**IL**	0.872549	**CO**	0.5	**AK**	0.185185	**NE**	0
WI	0.985915	**AZ**	0.866667	**AR**	0.48	**NV**	0.176471	**NH**	0
PA	0.969697	**KS**	0.847619	**IA**	0.454545	**TX**	0.114625	**OR**	0
NC	0.96	**WV**	0.836364	**VT**	0.428571	**WA**	0.102564	**SC**	0

Table 1.7: States with the Highest Percentage of Counties

The table above shows that states such as Rhode Island, New York, etc. are facing the latest wave of covid harder than other states and some states such as South Carolina, Oregon, etc. are in relatively better condition.

Data Modelling :-

Part 1:

In this part, we first divided data into two datasets for the sake of training and testing. The training data set consists of 6 states New York, Texas, California, Washington, Illinois, Pennsylvania, Florida and North Carolina. The rationale for choosing these states are the most populated states in the U.S. and are very diverse. We believe these 6 states, as samples, can be a good representation of the U.S. for the modeling purpose.

Classification based on the total number of deaths per 10000

First, we do the modeling based on the class variable of death per 10000. The number of counties with death per 10000 more and less than 36 for these 6 states are as follows:

Class variable	counties with Less than 38	counties with More than 38
Death per 10000	236	201

The same table for the rest of the 46 states is as follows:

Class variable	counties with Less than 38	counties with More than 38
Death per 10000	1315	1366

Map 1 visualizes the 6 states selected in the previous steps with their counties. The counties in red represent counties with more than 38 deaths in 10000 people.

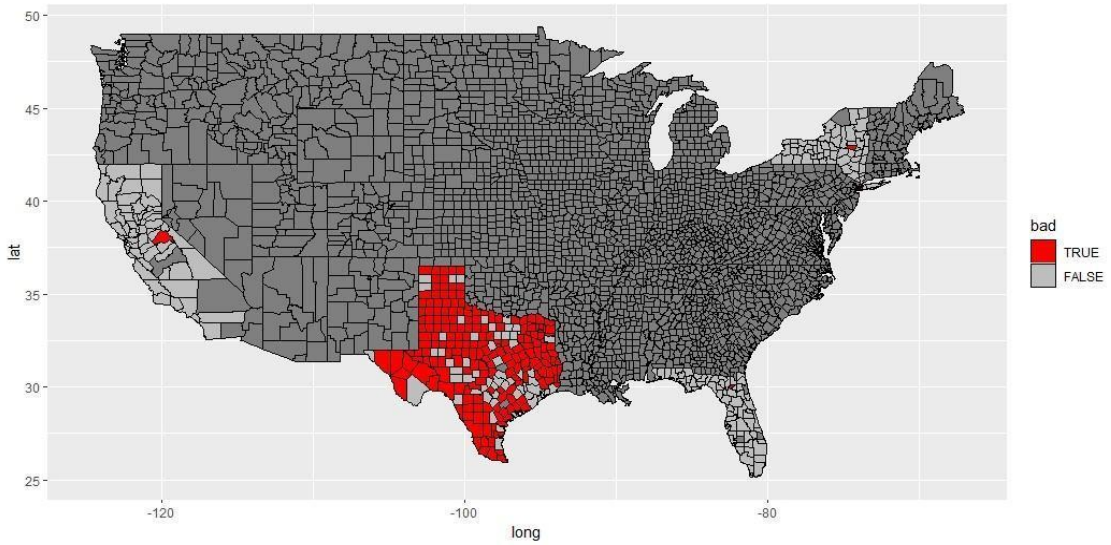

Figure 2: Counties with more than 38 Deaths per 10000 people on selected states

The figure above shows that the state of Texas has the highest percentage of counties with more than 38 deaths per 10000 population. It is while in other states combined, there are only 2 other counties that show such a rate.

The result of analysis shows that in the 6 states above, the importance of selected variables on death rate is as follows:

Features	Feature Importance
million_dollar_housing_units	0.421
asian_pop_per_10000	0.406
commuters_by_public_transportation_per_10000	0.391
income_100000_or_more_per_10000	0.378
commute_more_10_mins_per_10000	0.363
WOW1_Cases	0.362
commute_less_10_mins_per_10000	0.362
median_income	0.362
total_pop	0.35
income_per_capita	0.349
female_20_to_59_per_10000	0.337
hispanic_male_55_64_per_10000	0.323
dwellings_less_10_units_attached_per_10000	0.322
dwellings_more_10_units_per_10000	0.321
other_race_pop_per_10000	0.314
hispanic_pop_per_10000	0.311
WOW2_Cases	0.309
WOW3_Cases	0.306
asian_male_55_64_per_10000	0.294
male_20_to_59_per_10000	0.251

Table 2: Importance of Features

Application of Model for Cases in the Last Month

Random Forest

First, we need to take a look at the actual results A.K.A Ground Truth, so we can compare it to the results of our prediction on the test set to evaluate the model performance.

The figure below shows the ground truth results:

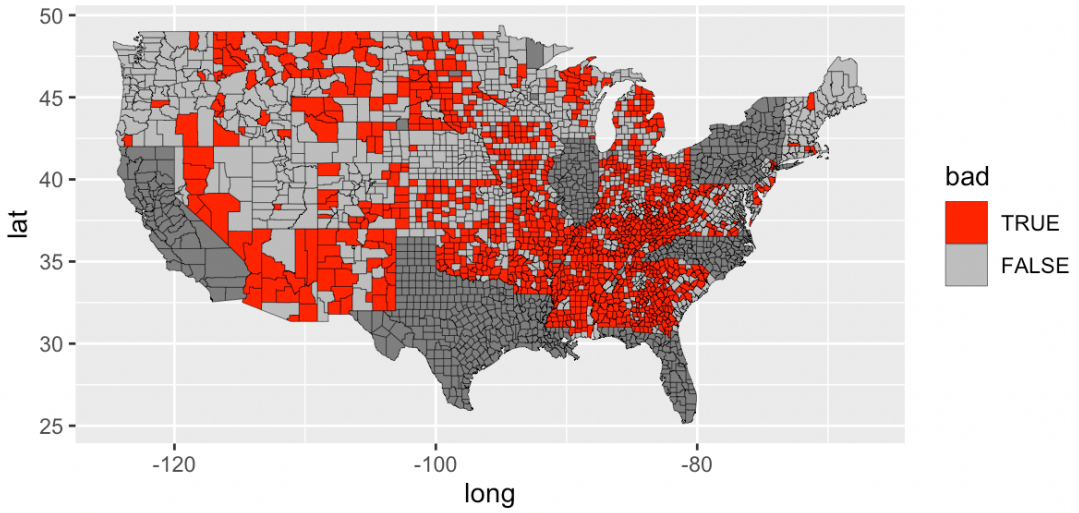

Figure 3.1: Ground Truth result for all States except selected states.

Now, after using model for prediction, the figure looks like the following:

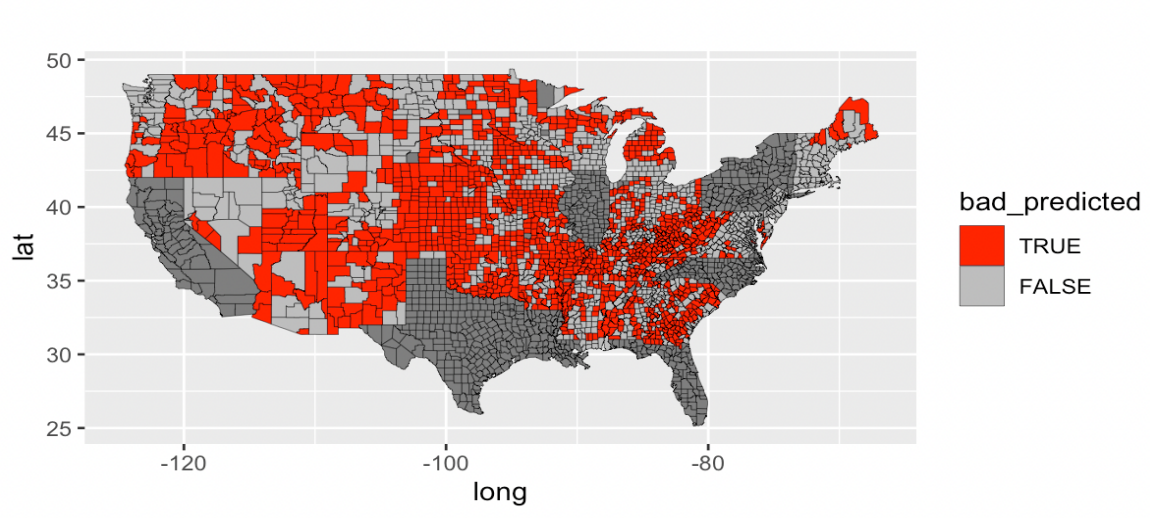

Figure 3.2: Prediction Model results

As it is observed, the model does a good prediction in southern regions, while in some northern regions it seems that model did not do a good prediction. The confusion matrix for the for the results of prediction is as follows:

	FALSE	**TRUE**
FALSE	713	363
TRUE	457	897

Table 3.1: Confusion Matrix for Random Forest

We have used the parameters like “mtry”, “maxdepth” to hyper-tune the trees. With these parameters the depth of the tree and random sampling of variables was set to 3. This helped to prune the tree and obtain the following metrics

Training Set Accuracy	0.7451
Training Set Kappa	0.4867
Testing Set Accuracy	0.6626
Specificity	0.7119
Sensitivity	0.6094
Kappa	0.3222

Table 3.2: Fitting results

The results show that the accuracy of the model is good for training dataset but less for test dataset, indicating overfitting. Looking at, the sensitivity and predictivity of the model the values are good. Considering more data points or addition of new variables or using a different model may improve these metrics.

Classification and regression tree (CART)

In this section, we apply another method of classification called CART or classification and regression tree. We do the same steps as the previous part. CART is known work best with the small dataset.

The figure below shows the results of the application of CART model for prediction on the test set.

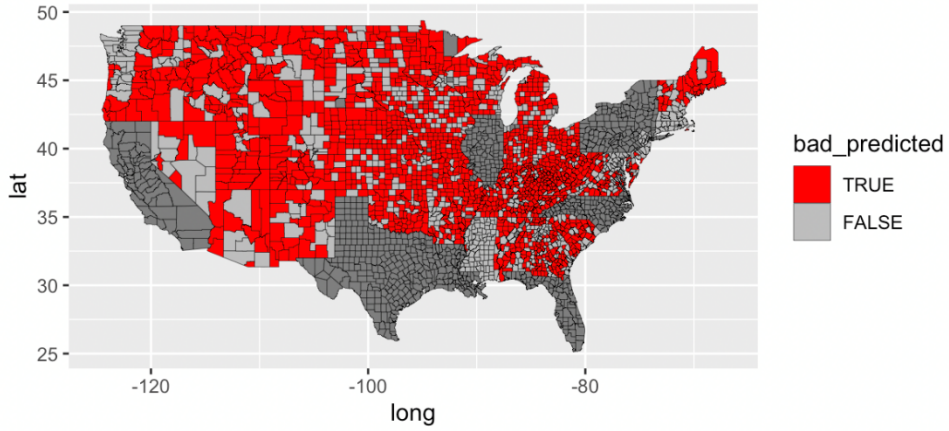

Figure 4.1: Prediction Model results

As it can be observed, the results are relatively close to those of random forest method. However, as for the classification matrix and statistics, we see slight changes. Although it seems like on the northern region, this model is marking every other county to be having covid severity.

	FALSE	**TRUE**
FALSE	890	574
TRUE	280	686

Table 4.1: Confusion Matrix for CART

Training Set Accuracy	0.6486
Training Set Kappa	0.4359
Testing Set Accuracy	0.5926
Specificity	0.5444
Sensitivity	0.7607
Kappa	0.3024

Table 4.2: Fitting results

We see that the overall accuracy of the model has slightly decreased. The accuracy difference between the training dataset and the test dataset is better. The Kappa values seems to be doing good. The specificity is however low.

k-Nearest Neighbors

In this step, we use another mode of classification called k-Nearest Neighbors for the same phenomenon. The results are as follows:

The figure below shows the results of the application of the k-Nearest Neighbors model for prediction on the test set.

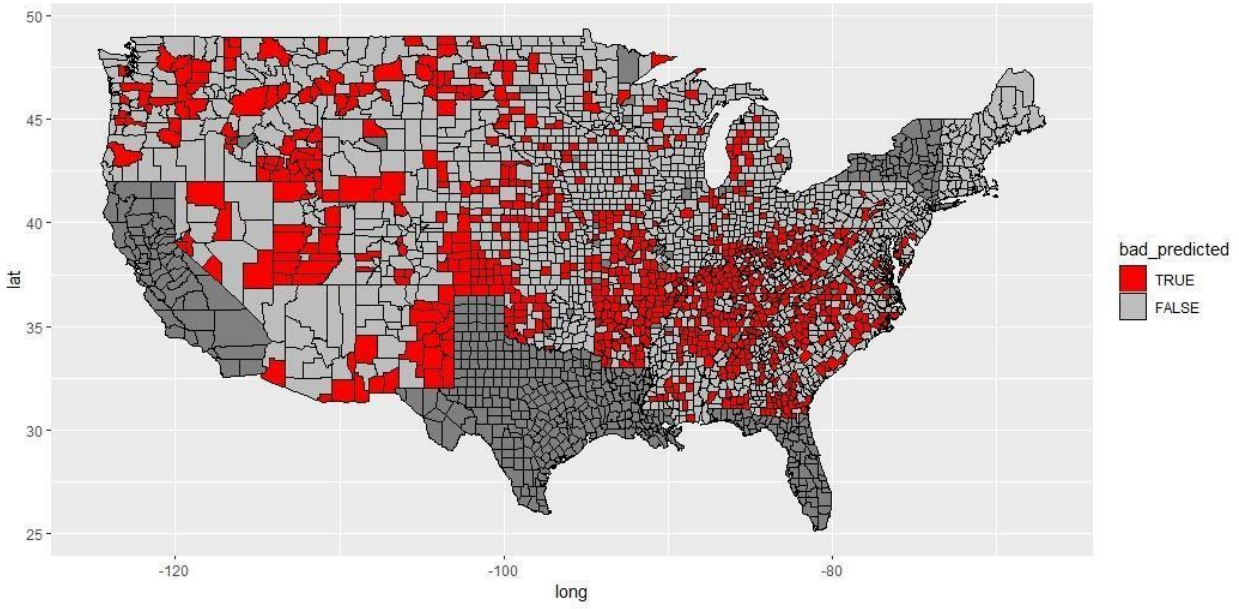

Figure 5.1: Prediction Model results using KNN

It is observed that this method, shows much better results than previous methods. In eastern areas, the number of counties in bad situation has significantly increased in-line with the ground truth.

To better understand the improvements made by this model, we refer to the confusion matrix and statistics of this model.

		FALSE	**TRUE**
FALSE		973	789
TRUE		340	577

Table 5.1: Confusion Matrix for K-Nearest neighbour

Overall accuracy	0.5786
Specificity	0.4224
Sensitivity	0.7411
Kappa	0.1624

Table 5.2: Fitting results

Naïve Bayes

For a smaldatasetset, Naïve Bayes is considered to be the best model. The figure below shows the results of the application of the naïve Bayes classifier model for prediction on the test set.

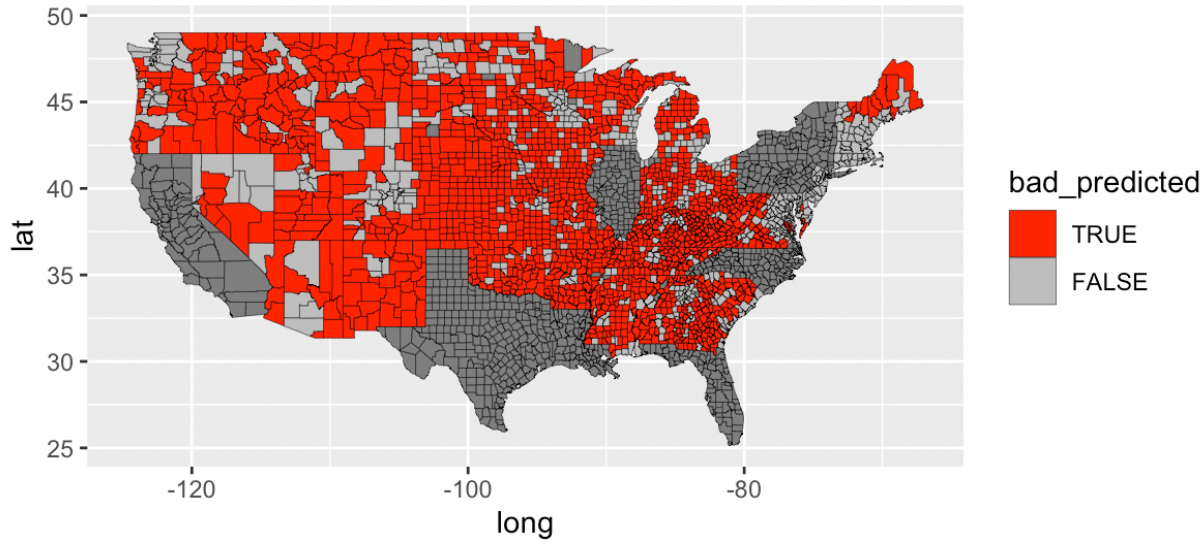

Figure 6.1: Prediction Model results using Naïve bayes

		FALSE	TRUE
FALSE		21	11
TRUE		620	646

Table 6.1: Confusion Matrix for Naïve bayes

Training Set Accuracy	0.7042
Training Set Kappa	0.4234
Testing Set Accuracy	0.5139
Specificity	0.98326
Sensitivity	0.03276
Kappa	0.3682

Table 6.2: Fitting results

This is by far,One of the worst model. Although we tried using different hyperparameters. We obtained the similar pattern

Part-2:

In this section, we apply our first model which is based on random forest classification, that we created using a training set consisting of 23 states Texas, New York, California, Florida, Illinois, North Carolina, Georgia, Ohio, Michigan, New Jersey, Tennessee, Arizona, Virginia, Massachusetts, South Carolina, Washington, Wisconsin, Indiana, Minnesota, Colorado, Missouri, Kentucky, and Alabama into our test set consisting of other 27 states.

This is because, we see that in the previous section, the model's performance is not that good. Although it has been 2 years, since COIVD hit, due to the mutations and vaccines, there is a change in the trend. Hence, we suspect that it is because of less dataset. Thus, in this section, we are considering training datasets to have close to 1000data pointss.

Classification based on the total number of deaths per 10000

First, we do the modeling based on the class variable of death per 10000. The number of counties with death per 10000 more and less than 38 for these 23 states is as follows:

Class variable	counties with Less than 38	counties with More than 38
Death per 10000	919	923

The same table for the rest of 27 states is as follows:

Class variable	counties with Less than 38	counties with More than 38
Death per 10000	641	657

Map 1 visualizes the 23 states selected in the previous steps with their counties. The counties in red represent counties with more than 38 deaths in 10000 people.

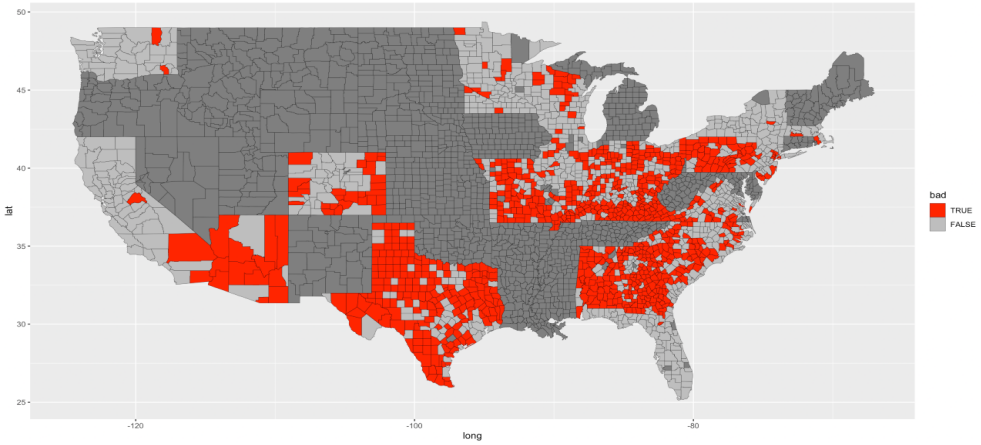

Figure 11 : Counties with more than 38 Deaths per 10000 people on selected states

The figure above shows that the state of Texas has the highest percentage of counties with more than 38 deaths per 10000 population. It is while in other states combined, there are only 2 other counties that show such rate.

The result of analysis shows that in the 23 states above, the importance of selected variables on death rate is as follows:

Random Forest

First, we need to take a look at the actual results A.K.A Ground Truth, so we can compare it to the results of our prediction on the test set to evaluate the model performance.

The figure below shows the ground truth results:

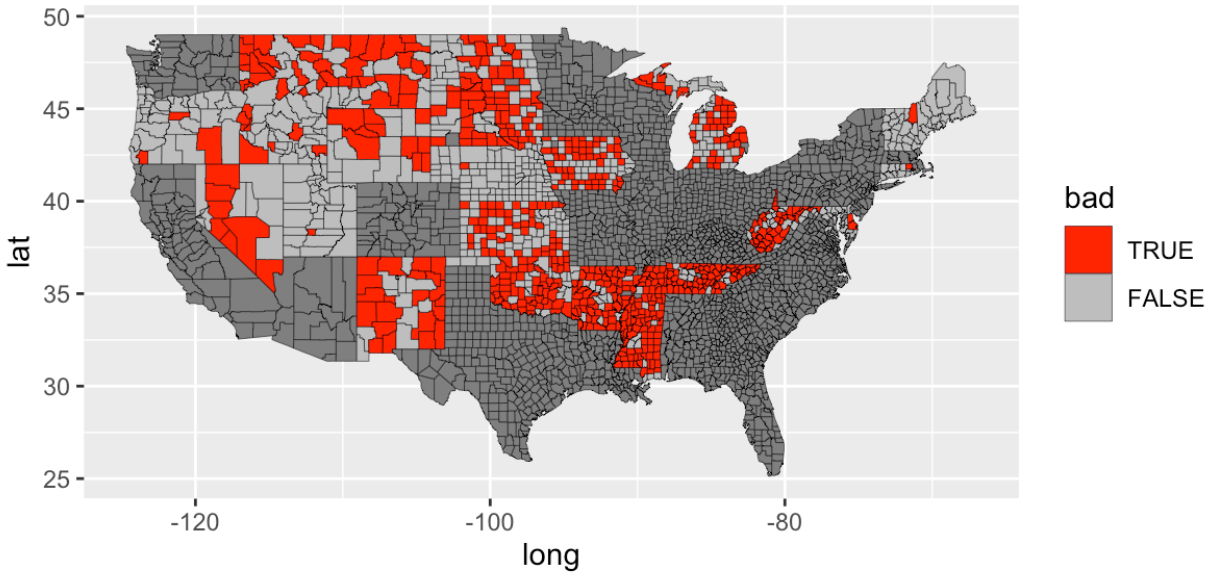

Figure 12.1: Ground Truth result for all States except selected states.

Now, after using model for prediction, the figure looks like the following:

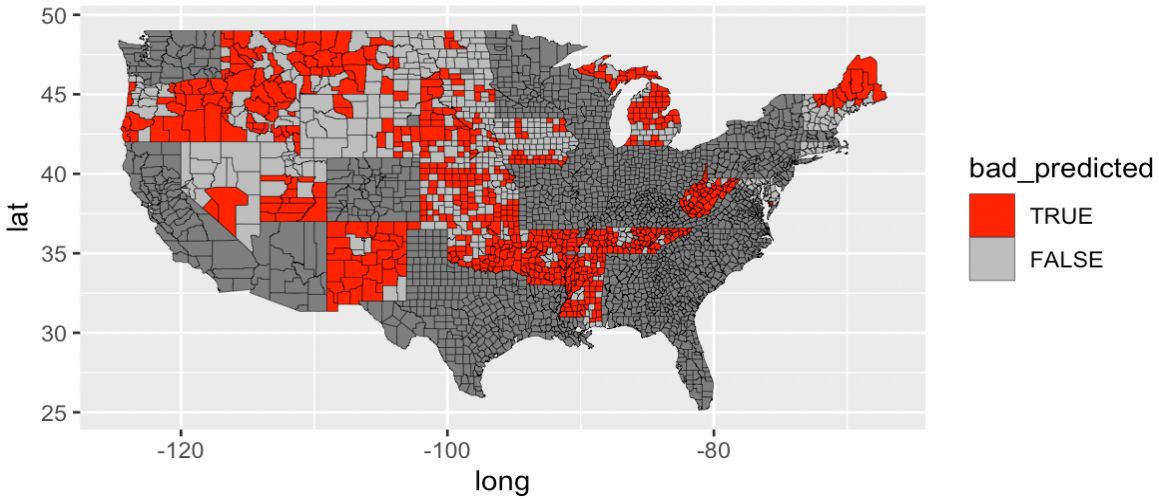

Figure 12.2: Prediction Model results

As it is observed, the model does a good prediction in southern regions, while in some northern regions it seems better than the previous model, but needs improvement. The confusion matrix for the for the results of prediction is as follows:

	FALSE	**TRUE**
FALSE	415	153
TRUE	226	504

Table 12.1: Confusion Matrix for Random Forest

Training Set Accuracy	0.7394
Training Set Kappa	0.4787
Testing Set Accuracy	0.708
Specificity	0.7671
Sensitivity	0.6474
Kappa	0.4151

Table 12.2: Fitting results

The results for this dataset are better than previous models. The training and test datasets show no signs of overfitting or underfitting. The specificity and sensitivity are also good comparatively.

(TP+TN)/(TP+TN+FP+FN) 1298

Classification and regression tree (CART)

In this section, we apply another method of classification called CART or classification and regression tree. We do the same steps as the previous part. CART is known work best with the small dataset.

The figure below shows the results of the application of CART model for prediction on the test set.

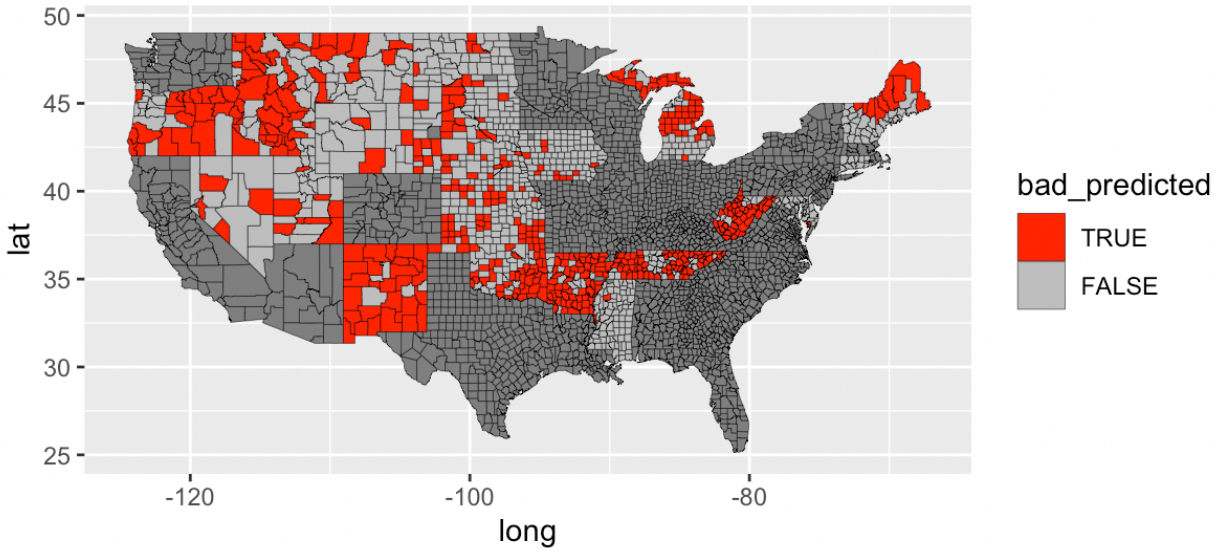

Figure 13.1: Prediction Model results

As it can be observed, the results are relatively close to those of random forest method. However, as for the classification matrix and statistics, we see slight changes.

	FALSE	**TRUE**
FALSE	469	303
TRUE	172	354

Table 13.1: Confusion Matrix for CART

Training Set Accuracy	0.7324
Training Set Kappa	0.4648
Testing Set Accuracy	0.6486
Specificity	0.5388
Sensitivity	0.7317
Kappa	0.2698

Table 13.2: Fitting results

We can observe that the difference between the training dataset and the test dataset has come down to 10%. And we see that the Specificity value has increased a little. The kappa value for both the datasets works well.

Naïve Bayes

Again, for a small dataset, Naïve Bayes is considered to be the best model. The figure below shows the results of the application of the naïve Bayes classifier model for prediction on the test set.

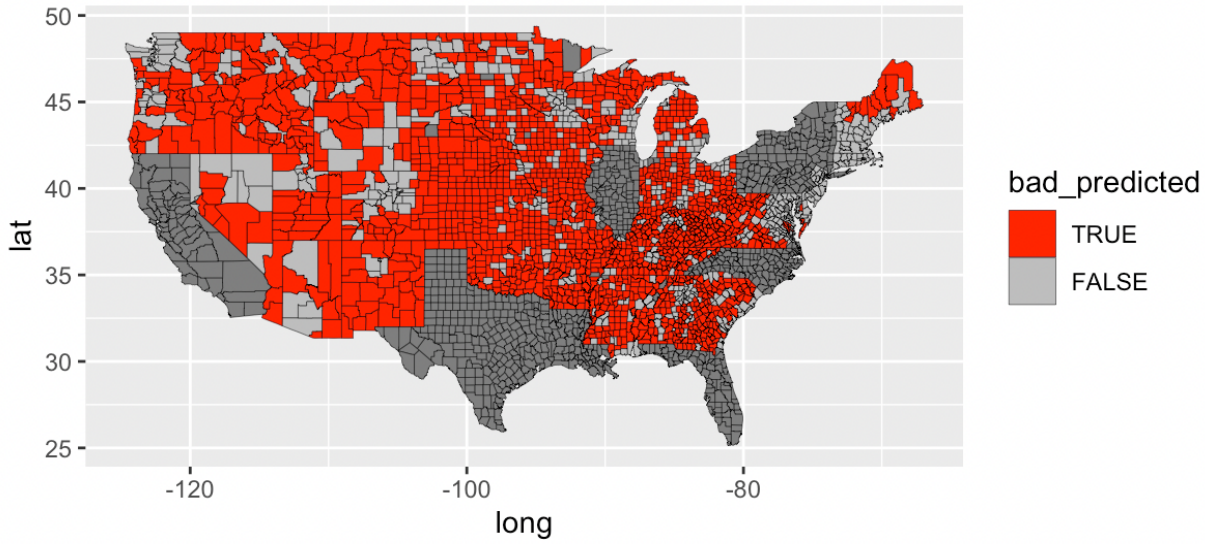

Figure 14.1: Prediction Model results using Naïve bayes

		FALSE	TRUE
FALSE		21	11
TRUE		620	646

Table 14.1: Confusion Matrix for Naïve bayes

Training Set Accuracy	0.7042
Training Set Kappa	0.4234
Testing Set Accuracy	0.5139
Specificity	0.98326
Sensitivity	0.03276
Kappa	0.3682

Table 14.2: Fitting results

This is by far, the worst model. Although we tried using different hyperparameters. We obtained the similar pattern

k-Nearest Neighbors

In this step, we use another mode of classification called k-Nearest Neighbors for the same phenomenon. The results are as follows:

The figure below shows the results of the application of k-Nearest Neighbors model for prediction on the test set.

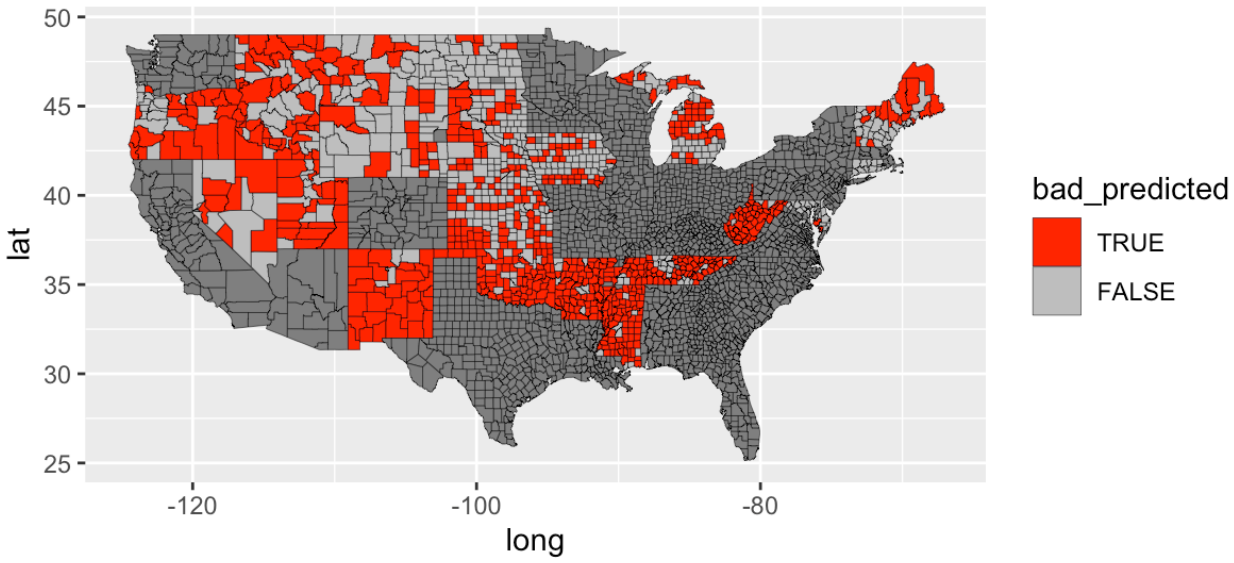

Figure 15.1: Prediction Model results using KNN

It is observed that this method, shows much better results than previous methods. This model is performing well, the northern regions predictions are better than the other models

To better understand the improvements made by this model, we refer to the confusion matrix and statistics of this model.

		FALSE	**TRUE**
FALSE		389	169
TRUE		252	488

Table 15.1: Confusion Matrix for K-Nearest neighbour

Training Set Accuracy	0.7535
Training Set Kappa	0.5145
Testing Set Accuracy	0.6757
Specificity	0.7336
Sensitivity	0.6069
Kappa	0.3363

Table 15.2: Fitting results

Comparison Of Model :-

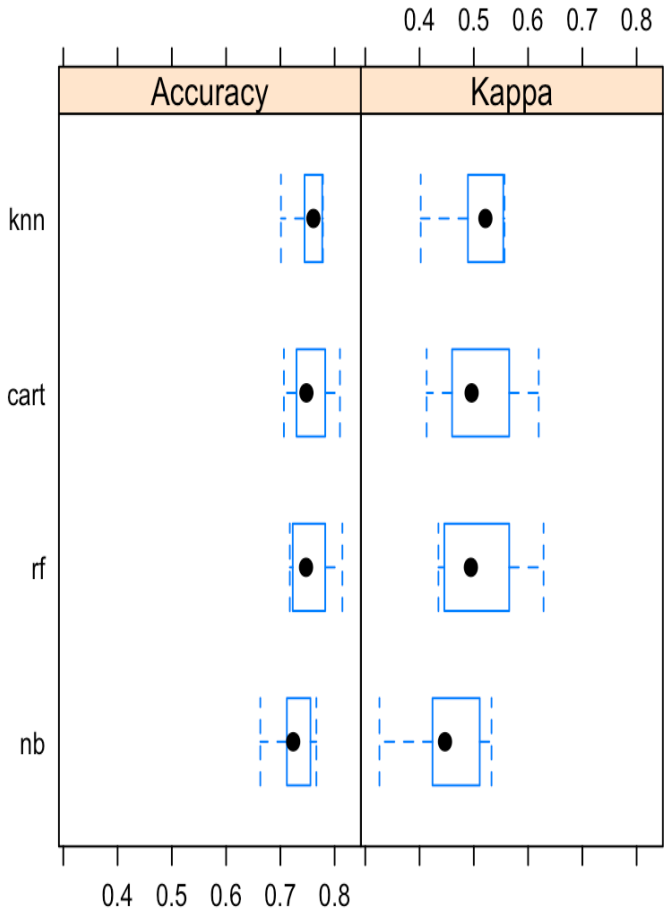

Figure 19.2: Comparison between Base Models

From all the models built, the knn model seems to perform consistently on all the folds.

```
Call:
summary.diff.resamples(object = difs)

p-value adjustment: bonferroni
Upper diagonal: estimates of the difference
Lower diagonal: p-value for H0: difference = 0

Accuracy
     nb     rf        cart      knn
nb          -0.027703 -0.029326 -0.031974
rf   0.2656            -0.001623 -0.004271
cart 0.2414 1.0000              -0.002648
knn  0.7475 1.0000    1.0000

Kappa
     nb     rf        cart      knn
nb          -0.055794 -0.059050 -0.064214
rf   0.2602            -0.003256 -0.008420
cart 0.2358 1.0000              -0.005164
knn  0.7436 1.0000    1.0000
```

The above summary indicates that there is no difference between cart, random forest and knn. Since the random forest has the highest accuracy and kappa value, we will consider that as our model.

Model Evaluation & Deployment :-

From our analysis done so far, the best model obtained is random forest which provides a 70% accuracy. That the prediction of this model is 70% accurate. Where the false positives and true negatives are also around 70%. The government can consider this model to predict the counties that may have death cases more than 38 per 10000 population. That is, the government can consider this model, as a reference to make decision. However, there is 30% possibility that this model may give false positives or negatives.

As stated earlier, although we have 2 years of COVID data, there has been changes to the scenario, this is because most of the people are vaccinated and also the virus is constantly mutating. Hence this model is better bet, for the government to consider in this scenario

As a data scientist, I would recommend the stakeholder to consider this model, as 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 70% probability of the prediction to be right. The model has performed consistently good

This model can be deployed on cloud, or run every week by providing the data required as showcased in the Table 1.2. The model will predict which counties will have more than 38 deaths per 10000 population and which does not . The data considered, has week over week trend data, hence when the model should be run every other week for better results.

We would suggest our stake holder, to consider the predictions, and take measures to prevent deaths on the counties that are predicted to have higher deaths.

This leads us to the question What about the false negatives? The model can also be run on daily basis, by providing the week over week trend calculated from the current day. The government case, ask the people in these counties to take self measures.

For future, we would like to collect more features, may be consider including health related data for each county, hospital patient's data to better the results

References:-

- About Social Determinants of Health (SDOH). (2021, March 10). https://www.cdc.gov/socialdeterminants/about.html
- Angelucci, M., Angrisani, M., Bennett, D., Kapteyn, A., & Schaner, S. (2020). Remote Work and the Heterogeneous Impact of COVID-19 on Employment and Health(No. w27749; p. w27749).
- National Bureau of Economic Research. https://doi.org/10.3386/w27749
- Bhadra, A., Mukherjee, A., & Sarkar, K. (2021). Impact of population density on Covid-19 infected and mortality rate in India. Modeling Earth Systems and Environment, 7(1), 623–629. https://doi.org/10.1007/s40808-020-00984-7
- Classification Measures. (n.d.). Datanovia. Retrieved October 28, 2022, from https://www.datanovia.com/en/lessons/clustering-distance-measures/Ferrazzano, G. F., Ingenito, A., & Cantile, T. (2020). COVID-19 Disease in Children: What Dentists Should Know and Do to Prevent Viral Spread. The Italian Point of View. International Journal of Environmental Research and Public Health, 17(10), 3642. https://doi.org/10.3390/ijerph17103642
- Singh, R., & Adhikari, R. (2020). Age-structured impact of social distancing on the COVID-19 epidemic in India. https://doi.org/10.48550/ARXIV.2003.12055