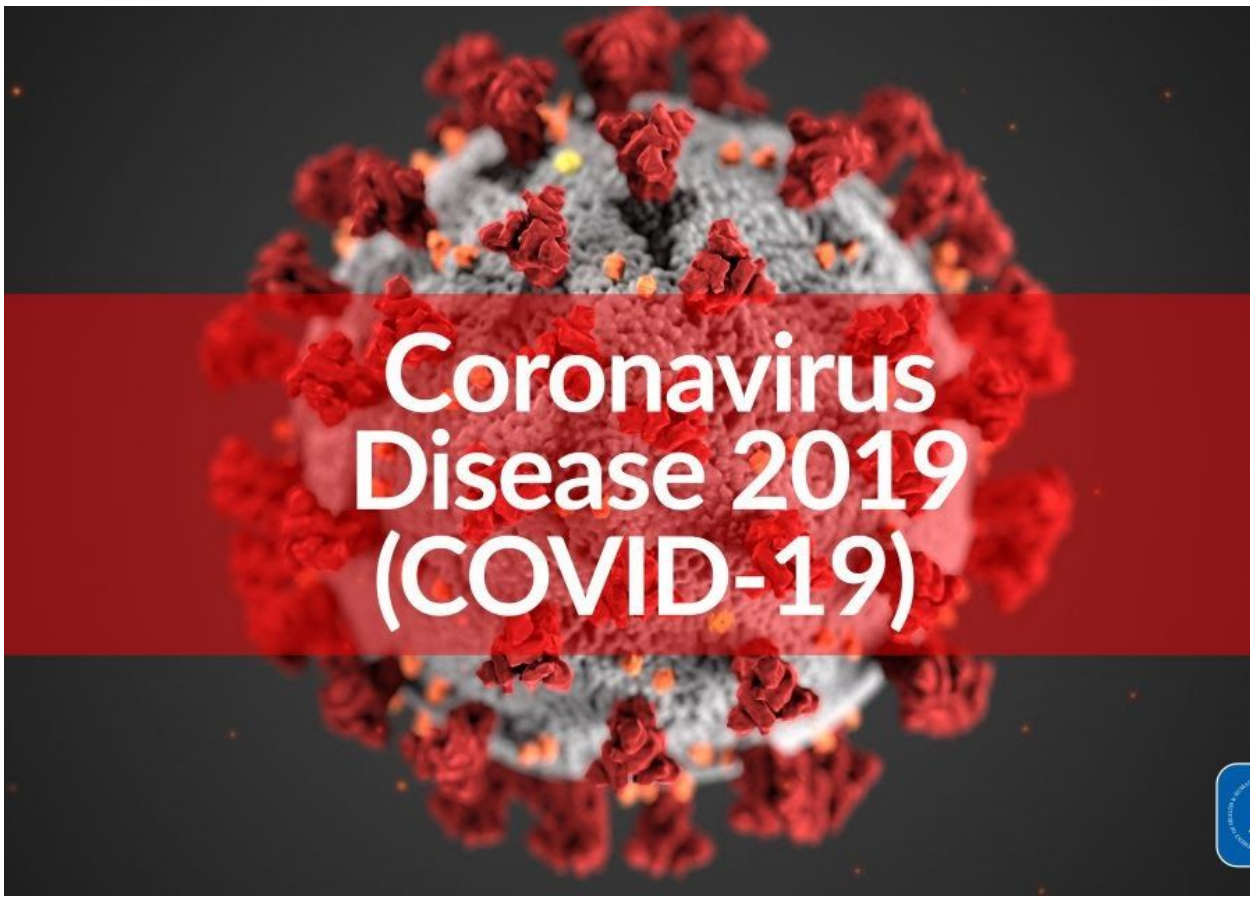
COVID-19 Classification: Deaths and Cases Prediction Model by County

By Starr Corbin



Executive Summary

The goal of the project was to create a model that could predict a county's risk of COVID-19 death and confirmed cases. Variables used in the prediction model included population ethnicity, population density, employment, social distancing scores, and commuting in public transportation attributes. The classifier (risk category) for the prediction model was based on the calculation of a county's peak values for COVID-19 cases and deaths and the most recent date the county experienced these peak values. Classification models such as Conditional Inference Decision Tree, C4.5 Decision Tree, Rules-based Classifier (PART), Artificial Neural Network, and Random Forest were used to create the ideal classification training set for the prediction model. Attributes that had the greatest weights in determining a county's risk for COVID-19 cases and deaths included population ethnicity. Most notably, the attribute that had the most weight in determining a county's risk for COVID-19 deaths and cases was the white population percentage. On average, the classification models calculated that counties with white populations greater than 80% were assigned a low risk for both COVID-19 cases and deaths. As a result, factors such as a county's population makeup and the last time a county experienced a peak in cases and deaths should be primarily taken into account when developing state policy to address the COVID-19 pandemic.

Table of Contents

Executive Summary	**2**
Table of Contents	**3**
Business Understanding	**4**
Data Set Description	**4**
Verifying Data Quality	**6**
Statistical Summary of COVID-19 Predictor Models	**7**
Analysis and Visualization of Selected Data Types and Prediction Models	**11**
Case Category Modeling	**13**
Deaths Category Modeling	**24**
Conclusion	**34**
References	**35**

Business Understanding

Using county data from the United States the analyst's goal of this project is to predict the high, medium, or low risk of U.S. counties as it relates to COVID-19 confirmed cases and deaths. Given a prediction model based on a training set from specific states experiencing the various peaks in cases and deaths, which counties in other U.S. states are similar and therefore share a similar risk for COVID-19 cases and deaths moving forward?

Stakeholders for this information include policy makers and health officials. Having the ability to predict a county's risk for COVID-19 cases and/or deaths can help this group make more informed policy decisions for assessing county risk and developing state policy to combat the spread of COVID-19. .

Data to support this project was sourced from USAFacts, the U.S. 2010 Census, the CDC, Unacast, Johns Hopkins University (JHU) and ArcGIS. The data on COVID-19 trends is as of August 6th, 2020.

Data Set Description

The base starting file that was used to assess COVID-19 trends and data for counties in the United States was pulled from Google's public data set and primarily sourced from Johns Hopkins University (JHU), USAFacts, and the CDC. Data from USAFacts, the CDC, JHU, ArcGIS and Unacast was combined to provide insight into U.S. Census data, COVID-19 cases plus death trends, social distancing trends and population density by county. A description of this data file is given in Table 1. Also, the data type of each feature is provided.

Table 1. **The description of features in the raw data**

Features	Data Type	Description
n_grade_total	Ordinal	Total grade count taken by adding n_grade_encounters and n_grade_distance together and dividing by 2. Variable provides a more holistic view to the social distancing grade by county in Texas.
confirmed_cases_US Facts	Ratio	Count of confirmed cases for COVID-19

Deaths	Ratio	Count of deaths from COVID-19
total_pop	Ratio	Total population for each county per the U.S. 2010 Census
pop_density	Ratio	Population density by county (total number of people per square mile)
black_pop	Ratio	Normalized data. Number of reported black people per county divided by the county total population variable.
white_pop	Ratio	Normalized data. Number of reported white people per county divided by the county total population variable.
asian_pop	Ratio	Normalized data. Number of reported asian people per county divided by the county total population variable.
hispanic_pop	Ratio	Normalized data. Number of reported asian people per county divided by the county total population variable.
amerindian_pop	Ratio	Normalized data. Number of reported American Indian people per county divided by the county total population variable.
employed_pop	Ratio	Normalized data. Number of reported employed people per county divided by the county total population variable.
unemployed_pop	Ratio	Normalized data. Number of reported unemployed people per county divided by the county total population variable.
commuters_by_public_transportation	Ratio	Normalized data. Number of reported public transportation commuters per county divided by the county total population variable.
male_pop	Ratio	Normalized data. Number of reported men per county divided by the county total population variable.
female_pop	Ratio	Normalized data. Number of reported women per county

		divided by the county total population variable.
deaths_category	Nominal	Risk variable labeled as High, Medium and Low. This variable was created after determining the peak value that occurred in a rolling average of newCOVID-19 deaths in a county.
cases_category	Nominal	Risk variable labeled as High, Medium and Low. This variable was created after determining the peak value that occurred in a rolling average of new COVID-19 cases in a county.
county_name	Nominal	Name of the Texas county
state	Nominal	Name of each State in the United States

Source: USAFacts, Unacast, U.S. Census, Johns Hopkins University, ArcGIS

Verifying Data Quality

The quality of raw data is a crucial factor to generate reliable results. If the data isn't properly prepared for calculating or analysis (i.e. missing data, duplicated data and outliers), those data will affect the result of data mining and performances of the models.

The quality of the new data and data cleaning is discussed:

deaths_category:

There is no missing data for the deaths_category variable. Counties that had no new deaths averaging for the past seven days (i.e. 0) were labeled as "Low" risk. Data could be duplicated by county because the frequency of the same number was more than one. For example, multiple counties averaged only two deaths in the past week and therefore shared the value of "2".

cases_category:

There is no missing data for the cases_category variable. Counties that had no new deaths averaging for the past seven days (i.e. 0) were labeled as "Low" risk. Data could be duplicated by county because the frequency of the same number was more than one. For example, multiple counties averaged only five cases in the past week and therefore shared the value of "5".

Statistical Summary of COVID-19 Predictor Models

The original goal of this project was to determine which attribute or attributes in Table 1 had the greatest weight in predicting the risk classification ("High", "Medium" or "Low") for COVID-19 cases and deaths across all counties in the U.S. The data source for COVID-19 death and cases was pulled from Johns Hopkins University since they separated out the number of COVID-19 cases and deaths by date, thus allowing the data analyst to review cases and deaths over periods of time. The"High", "Medium" or "Low" risk classification (deaths_category, cases_category) resulted from calculating the seven day rolling averages for deaths and cases per county, their peak values and when those peak values occurred in a period of time. A county that had a peak in cases or deaths within the past 30 days was labeled as "High" risk. A peak of new cases or deaths in the past 60 days was considered "Medium" risk and anything greater than 60 days was classified as "Low" risk. Table 2 lists a sample of the original data set used to determine these outcomes and their corresponding risk labels (cases_category and death_category).

***Table 2*. Time Since Peak New Cases and Deaths with Associated Risk Label**

County	State	Days Since Last Peak Case	Days Since Last Peak Death	Cases Risk Category (cases_category)	Death Risk Category (deaths_category)
Dallas County	Texas	19	8	High	High
New York County	New York	107	111	Low	Low
Los Angeles County	California	14	98	High	Low
Miami-Dade County	Florida	1	0	High	High

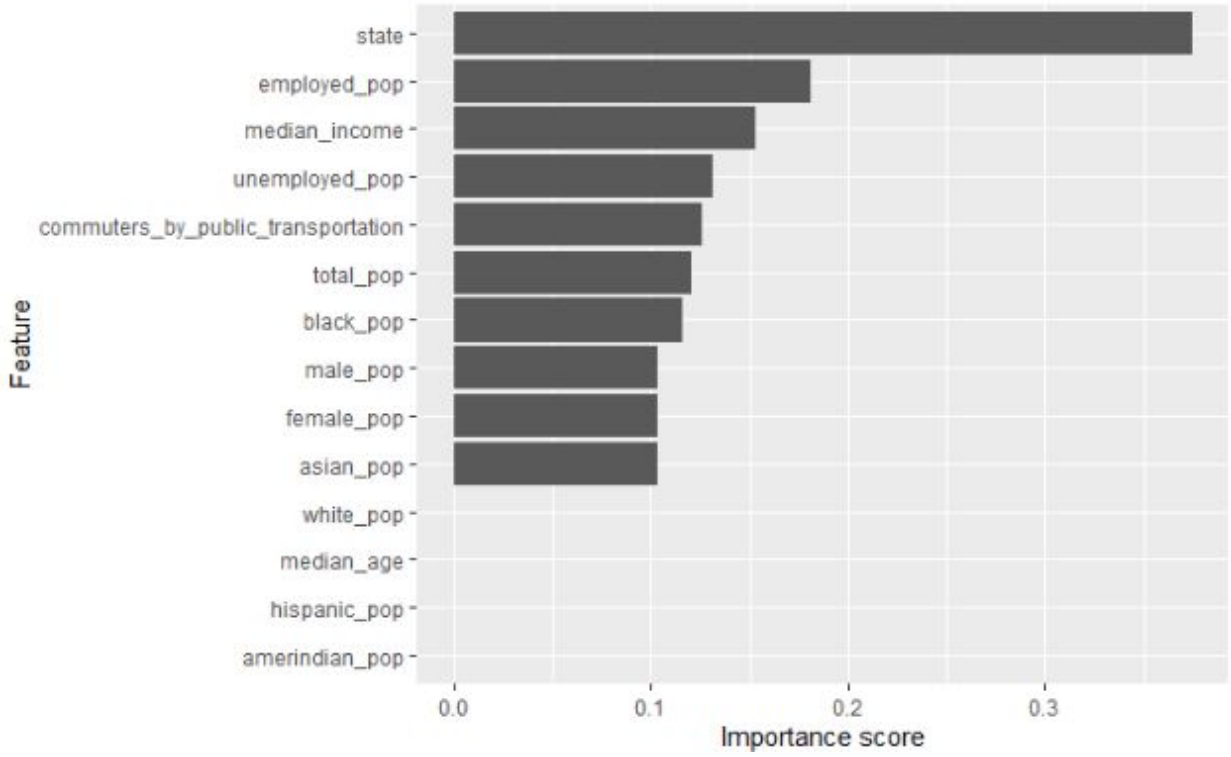
Source: Johns Hopkins University, U.S. Census

To prepare the data set for classification, the risk data (based on time since a county's peak of death and confirmed cases as noted in Table 2) was combined with U.S. Census data, social distancing data from Unicast and population density data from ArcGIST into a single table for analysis. Classification

models were then created from this combined table and included attributes such as state, population density, ethnicity (black, hispanic, asian and american indian), commuting via public transportation, employment and unemployment. The goal of using these attributes was to determine which one correlated the most, in terms of weight, to the risk classification of “High”, “Medium” and “Low”. Classification models to train the data for prediction included Conditional Inference Decision Tree, C 4.5 Decision Tree, PART (Rules-based classifier), Random Forest Fit and Artificial Neural Network (ANN).

In order to best determine the most important attribute to base the classification model from, chi-squared and univariate importance calculations were run on all counties in the United States. The original attribute these calculations predicted had the most weight was “State” for both COVID-19 deaths and cases. As a result, the training set created to predict the county risk prediction models would be based on select States. Figure 1 shows an example of the feature importance scores for Cases Category. It’s important to note that causation factors for “State” were not in the scope of this project. Speculation leads the analyst to assume though that the varying degrees to which the States addressed the spread of COVID-19 (i.e. school and business closures, mask and interstate travel policy) could be a reason why “State” was the leading attribute for a COVID-19 classification model.

***Figure 1*. Bar Graph of Attribute Importance Scores for Cases Category Class**



At the time of this analysis, the top three states with the highest ratings of death were Florida, South Carolina, and Arizona. The top three states with the highest ratings of cases were South Carolina, Florida, and Tennessee. Unfortunately the combination of these states created a "High" risk class imbalance for both death and case categorization. In order to remove the class imbalance in the training set, the classification models were built with State data that, when combined, achieved a balance of "High", "Medium", and "Low" risk.

As a result, the three states used for the case training set were New York, California and South Carolina. Chi-squared and univariate calculations were used to determine attribute importance for this training set. The top attributes with the most importance were white population, median income, black population and hispanic population. Table 3 lists the chi-squared attribute importance scores for each of these attributes.

***Table 3*. Chi-squared Attribute Importance Scores for Confirmed Cases Risk**

States	Feature	Attribute Importance
New York, California, South Carolina	White Population (white_pop)	0.53
	Median Income (median_income)	0.37
	Black Population (black_pop)	0.37
	Hispanic Population (hispanic_pop)	0.34

Source: Johns Hopkins University, U.S. Census

The three states used for the death training set were New York, California and Florida. Chi-squared and univariate calculations were also run to determine attribute importance for this training set. The top attributes with the most importance were white population and hispanic population. Table 4 lists the chi-squared attribute importance scores for each of these attributes.

***Table 4*. Chi-squared Attribute Importance Scores for Deaths Risk**

States	Feature	Attribute Importance
New York, California and Florida	White Population (white_pop)	0.42
	Hispanic Population (hispanic_pop)	0.42

Source: Johns Hopkins University, U.S. Census

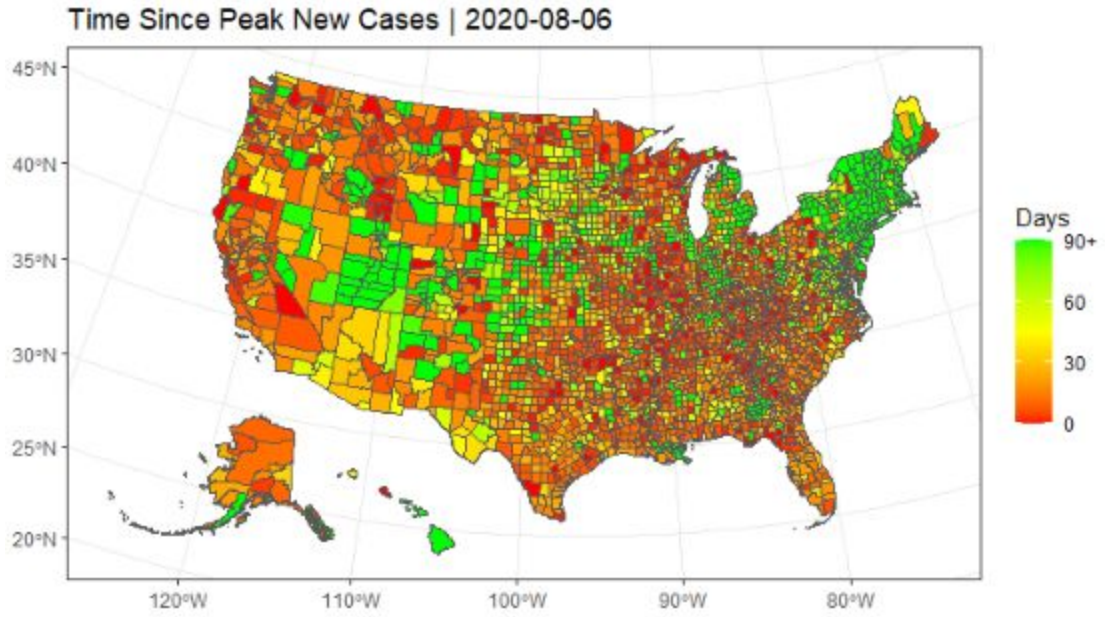
Analysis and Visualization of Selected Data Types and Prediction Models

Time Since Peak New Cases and Deaths

In order to establish “High”, “Medium” or “Low” risk class variables for the prediction models, peak COVID-19 death and confirmed case data over a period of time was analyzed from Johns Hopkins University. This data was used to then create a heat map (red, orange, yellow and green) based on that county’s timeframe of peak results for deaths and confirmed cases. A color range of red to orange notes that a county’s peak for cases and deaths was under 30 days ago. A color of light orange to yellow describes a county’s peak for cases and deaths was between 30 to 60 days ago. The color green notes peak cases and death figures greater than 60 days.

Figures 2 and 3 are heat maps that show the analysis for each county in the United States based on the time since their peak for new cases and deaths.

***Figure 2*. Time Since Peak New Cases**

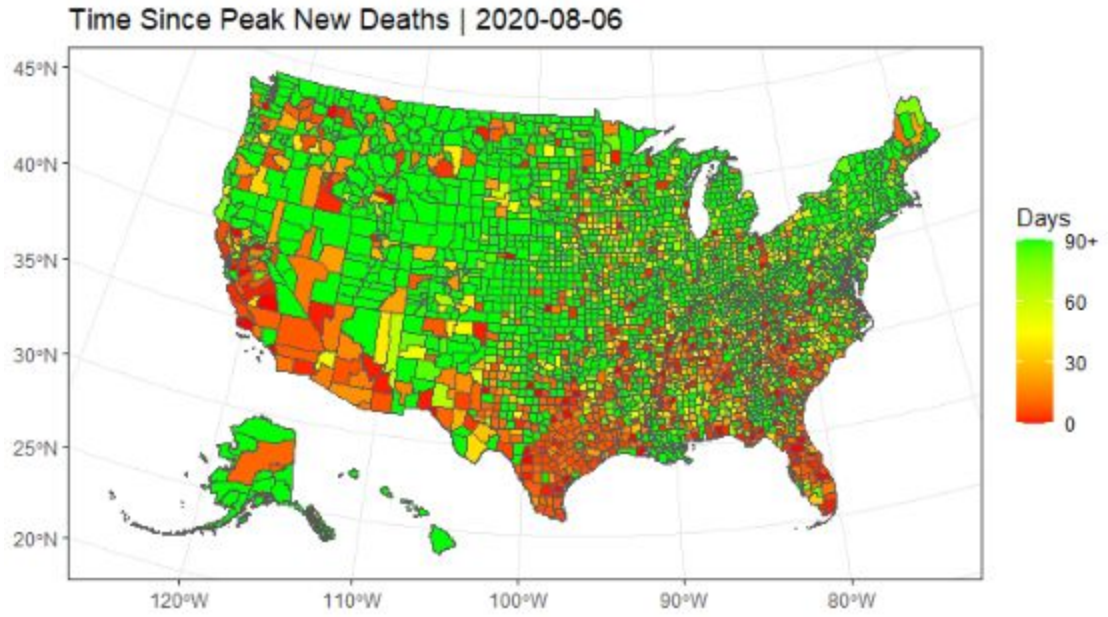


Source: Johns Hopkins University, U.S. Census

The heat map for peak new cases confirms that counties in the South, West, Mid-West and North West are experiencing the most recent peaks in new cases across the United States. Interesting

enough, the East coast areas and New York, which was at the center of the COVID-19 crisis in May have not, on average, had a peak in COVID-19 cases in the past 60 days.

***Figure 3*. Time Since Peak New Deaths**



Source: Johns Hopkins University, U.S. Census

The data for peak number in deaths is more concentrated in key regions of the U.S. than peak number of cases. According to the data, a majority of West and South counties lead in the most recent peaks of COVID-19 deaths. Leaders in these regions include California, Arizona, Texas and Florida.

Cases Category and Death Category Mapping

With the heat map data created, category labels of "High", "Medium" and "Low" were assigned to each county based on their timeline of peak values for cases and deaths. Two new columns: 1). cases category (cases_category) and 2). deaths category (deaths_category) were added to the table to house the category label assignments. Table 5 lists the peak value ranges for each category label.

***Table 5.* Category Label Assignments for Deaths and Confirmed Cases**

Peak Value Range	Category Label
0 to 30 days	High
30 to 60 days	Medium

60+ days	Low

Source: Johns Hopkins University, U.S. Census

Case Category Modeling

A Chi-squared Test for Independence was used to rank the features in Table 1 that had the greatest attribute importance for the case category class. The goal for the data analyst when using a chi-square test is to determine whether there is a significant association between the categories of the feature set listed in Table 1 and Case Category. As stated earlier in the summary (Table 3), "white population" was the independent variable with the greatest weight (0.53) for the chi-squared test. Feature subset selection, where the algorithm tries to find the best set of features using a blackbox approach and a greedy search strategy, also concluded white population, median income, black population, hispanic population and population density had the greatest attribute importance. Univariate importance and Correlation based Feature Selection (CFS) algorithms were also used to test association importance between the features in Table 1 and Case Category. White population had an attribute importance score of 0.36, followed by black population (0.14), hispanic population (0.11) and median income (0.11).

As noted earlier, New York, California, South Carolina data were used to balance the case category class data for the case training set. Their percentage for the "High" risk category is listed in Table 6.

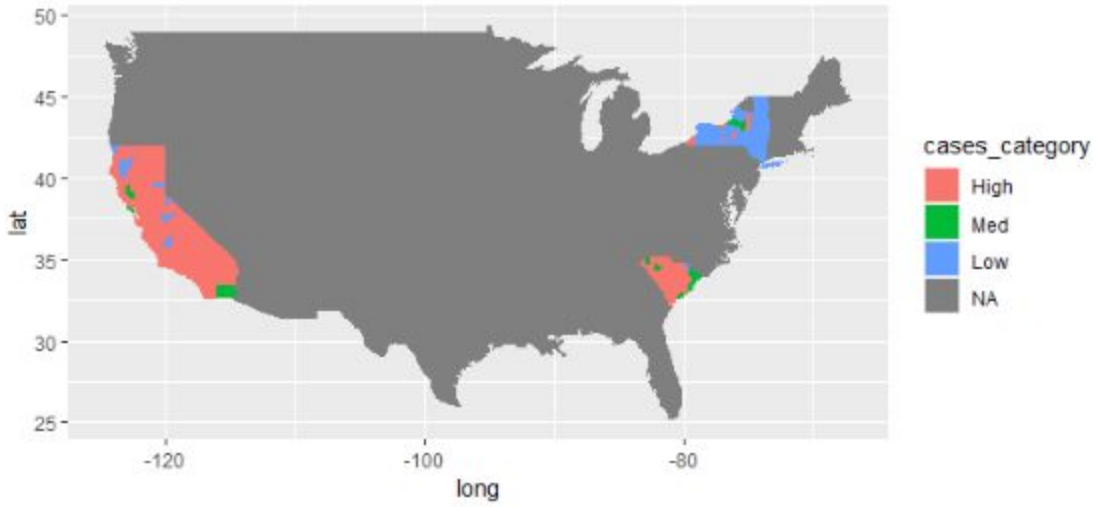
Table 6. **Training States' Percentage of High Risk Labels for Cases**

State	High Percentage
California	0.84
South Carolina	0.82
New York	0.08

Source: Johns Hopkins University

Out of these three states, 92 counties were labeled as "High" risk, 12 were "Medium" risk and 62 were "Low" risk. Figure 4 shows the map breakdown for these states for confirmed cases.

***Figure 4*. Case Category Risk Results for California, New York, South Carolina**



Source: Johns Hopkins University

Ten fold cross validation was then used to split the data set and build different case classification models using training data from these three states. The classification models used for case category classification were RPart Decision Trees, Conditional Inference Tree, C 4.5 Decision Tree, the Rules-based classifier PART, Random Forest and Artificial Neural Network (ANN). The decision to use Decision Trees (Rpart, Conditional Inference and C.45) was because they were fast at classifying the merged data set listed in Table 1, and the results would be easy to interpret (i.e. easy to read visuals for the decision tree that make the classification easy to understand). The rules based classifier PART was also chosen because of the easy to follow logic of "if and then" rules that could easily be interpreted if the number of rules returned were small. In order to provide more depth though to the decision tree classification, Random Forest was used since it's based on generating a large number of decision trees with subsets selected at random. The random decision trees are then used to generate accuracy values to select the optimal model using the largest value.

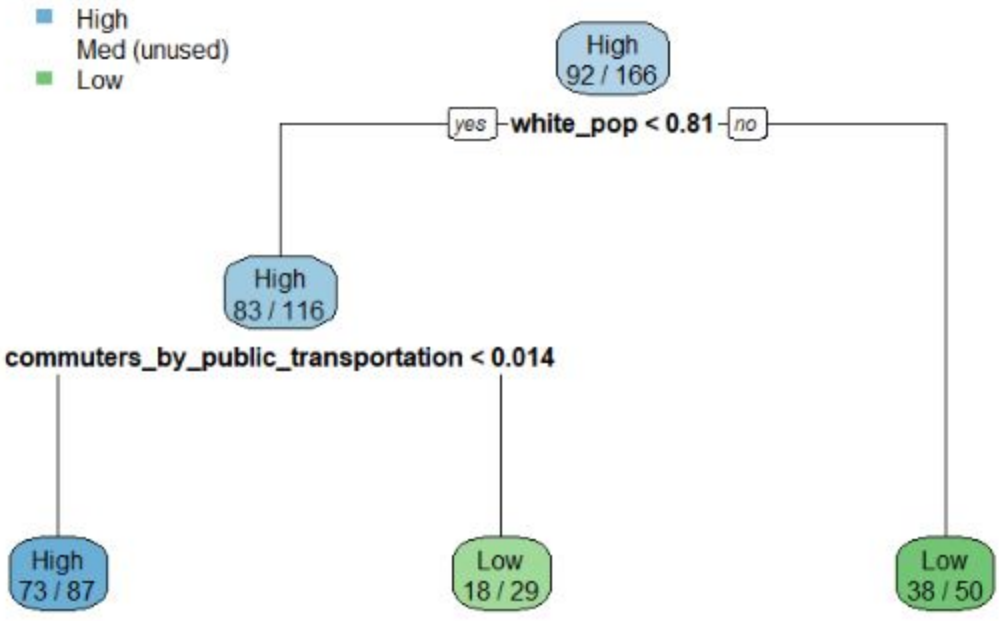
Decision Tree using "rpart"

The rpart decision tree was created from the training data derived from states listed in Table 6. Figure 5 shows the results from the rpart analysis.

***Figure 5*. Rpart Analysis for Highest Risk of Confirmed Cases**

```
CART

166 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 150, 150, 149, 147, 150, 150, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.08108108  0.6877902  0.3990613
  0.10810811  0.6703947  0.3616928
  0.39189189  0.6145124  0.1668783

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.08108108.
```

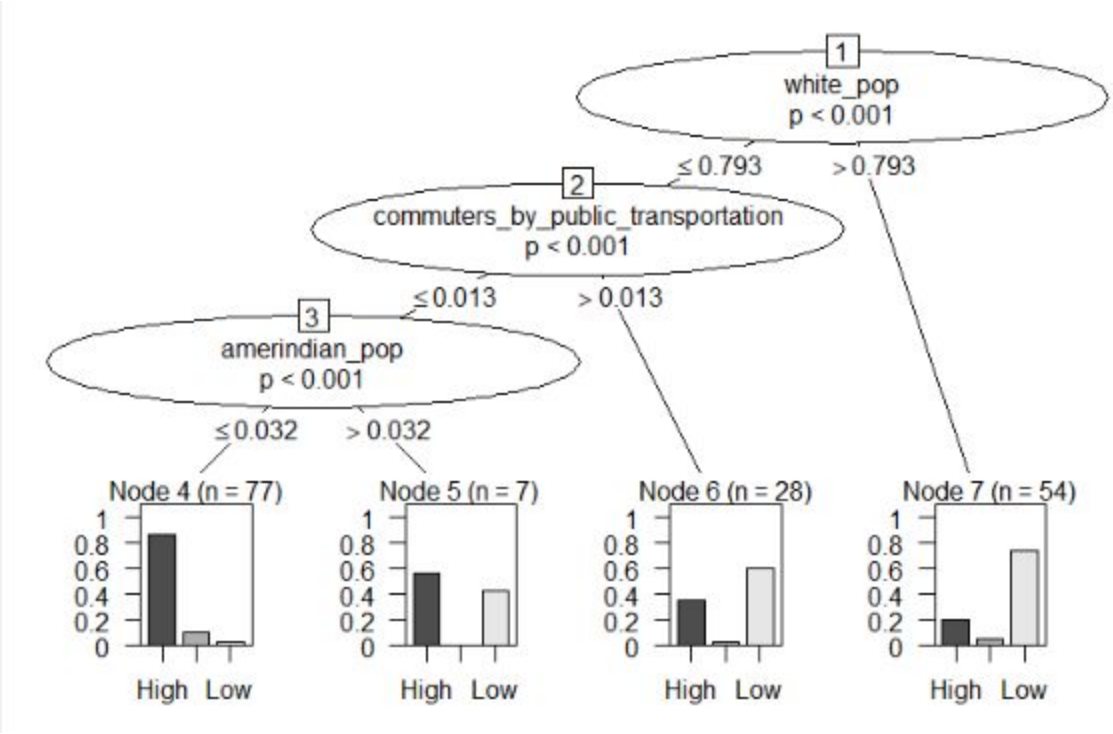
The rpart decision tree model resulted in a top accuracy rate of 0.68 and a kappa value of 0.39. The rpart decision tree also built a classification model that started with whether the white population value for a county was less than 0.81 (81% of the total population). If the white population was not less than 0.81 then the county risk for COVID-19 cases was “Low”. If it was less than 0.81 then the decision tree split to assess the public transportation commuters attribute. If the county value was less than 0.014, then the risk was “High” for cases.



Conditional Inference Decision Tree

The accuracy and kappa results for the Conditional Inference Decision Tree were similar to rpart Decision Tree. Utilizing a ten fold training set yielded a highest accuracy value of 0.73 and a kappa value of 0.50 (Figure 7). Figure 6 shows the results from the conditional inference Decision Tree analysis.

***Figure 6*. Conditional Inference Analysis for Highest Risk of Confirmed Cases**



Similar to the rpart decision tree analysis, the conditional inference analysis concluded that county's with a white population greater than 0.79 (79%) had, on average, greater numbers of "Low" risk categories of COVID-19 cases.

Figure 7. **Conditional Inference Analysis for Confirmed Cases Category**

```
Conditional Inference Tree

166 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 150, 149, 149, 150, 149, 149, ...
Resampling results across tuning parameters:

  mincriterion  Accuracy   Kappa
  0.01          0.6978758  0.4170876
  0.50          0.7228758  0.4590022
  0.99          0.7316176  0.5010148

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mincriterion = 0.99.
```

C45 Decision Tree Analysis

The results of the C45 decision tree algorithm analysis also aligned to rpart and conditional inference decision tree results. The highest accuracy and kappa values were greater than the initial decision tree algorithm results. The optimal model had an accuracy of 0.74 with a kappa value of 0.50 (Figure 8).

***Figure 8*. C45 Tree Analysis for Confirmed Cases Category**

```
C4.5-like Trees

166 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 149, 149, 149, 150, 150, 149, ...
Resampling results across tuning parameters:

  C      M  Accuracy   Kappa
  0.010  1  0.7325980  0.4797733
  0.010  2  0.7443627  0.5086684
  0.010  3  0.7443627  0.5037482
  0.255  1  0.7399510  0.4987212
  0.255  2  0.7333742  0.4878142
  0.255  3  0.7271242  0.4783249
  0.500  1  0.7340686  0.4900033
  0.500  2  0.7333742  0.4878142
  0.500  3  0.7271242  0.4783249

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were C = 0.01 and M = 2.
```

Similar to the other decision tree models, the C45 decision tree started the split with an assessment of a county's white population (Figure 9). If the white population was equal to or greater than 0.80 then the county risk for COVID-19 cases was "Low". If it was less than 0.81 than the decision tree split to assess the public transportation commuters attribute (0.008) and assigned a case category of "High".

***Figure 9*. C45 Tree for Confirmed Cases Category**

```
J48 pruned tree
------------------

white_pop <= 0.808271
|   commuters_by_public_transportation <= 0.008552: High (80.0/12.0)
|   commuters_by_public_transportation > 0.008552
|   |   hispanic_pop <= 0.197377: Low (18.0/2.0)
|   |   hispanic_pop > 0.197377
|   |   |   commuters_by_public_transportation <= 0.108764: High (14.0)
|   |   |   commuters_by_public_transportation > 0.108764: Low (4.0)
white_pop > 0.808271: Low (50.0/12.0)

Number of Leaves  :     5

Size of the tree :      9
```

Rules Based Classifier - PART

The PART rules-based classifier yielded a highest accuracy rating of 0.79 with a kappa value of 0.45 (Figure 11). White population percentage was also a major attribute in the rules based classification

model for cases category. For example, one rule noted that a county with a white population that had value greater than 0.80, a population density less than 128 and a male population greater than 0.49 had a “Low” risk case category (Figure 10).

***Figure 10*. PART Rules-Based Analysis Risk of Confirmed Cases**

```
Rule-Based Classifier

166 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 149, 150, 150, 149, 150, 149, ...
Resampling results across tuning parameters:

  threshold  pruned  Accuracy   Kappa
  0.010      yes     0.7099265  0.4589938
  0.010      no      0.6808824  0.4111677
  0.255      yes     0.6812500  0.4012675
  0.255      no      0.6808824  0.4111677
  0.500      yes     0.6812500  0.4012675
  0.500      no      0.6808824  0.4111677

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were threshold = 0.01 and pruned = yes.
```

***Figure 11*. PART Rules-Based Decision List for Confirmed Cases Risk**

```
PART decision list
------------------

white_pop > 0.808271 AND
pop_density <= 128.2 AND
male_pop > 0.495126: Low (28.0/2.0)

commuters_by_public_transportation <= 0.008008 AND
white_pop <= 0.820627 AND
male_pop <= 0.526696 AND
asian_pop > 0.014354: High (33.0)

median_income <= 46967 AND
asian_pop <= 0.014341 AND
unemployed_pop <= 0.059462 AND
pop_density <= 62.8: High (20.0)

median_income > 46967 AND
amerindian_pop > 0.001583 AND
hispanic_pop > 0.052059: High (27.0/5.0)

commuters_by_public_transportation > 0.003214 AND
asian_pop > 0.006936 AND
pop_density <= 522.2 AND
white_pop > 0.717014 AND
employed_pop > 0.442869: Low (13.0)

commuters_by_public_transportation > 0.008008 AND
black_pop > 0.038994: Low (14.0)

amerindian_pop > 0.003808 AND
male_pop <= 0.520028: Med (6.0/1.0)

amerindian_pop <= 0.016753 AND
black_pop > 0.009927 AND
asian_pop <= 0.013033 AND
asian_pop > 0.003982: High (14.0)

white_pop > 0.627979: Med (8.0/1.0)

: Low (3.0/1.0)

Number of Rules  :      10
```

Artificial Neural Network

Artificial Neural Network (ANN) is an algorithm model that is an assembly of inter-connected nodes and weighted links. The output node sums up each of its input values according to the weights of its links. The ANN results for case category was a 13-3-3 network with 54 weights (Figure 13). Ten-fold

cross validation was also used in this classification model which yielded an highest accuracy rating of 0.68 with a kappa value of 0.33 (Figure 12).

***Figure 12*. ANN Analysis for Confirmed Cases Risk**

```
Neural Network

166 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 150, 150, 150, 150, 149, 150, ...
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
  1     0e+00  0.5607198  0.02213740
  1     1e-04  0.5548375  0.00000000
  1     1e-03  0.5669698  0.03359491
  1     1e-02  0.5846169  0.08085740
  1     1e-01  0.5919698  0.10173607
  3     0e+00  0.5548375  0.00000000
  3     1e-04  0.5610875  0.01652174
  3     1e-03  0.6154992  0.15758022
  3     1e-02  0.6838816  0.33758869
  3     1e-01  0.6247872  0.20435780
  5     0e+00  0.5791022  0.06611434
  5     1e-04  0.5732198  0.05011665
  5     1e-03  0.5860875  0.08115724
  5     1e-02  0.6643963  0.28601122
  5     1e-01  0.6168537  0.21160191
  7     0e+00  0.5732198  0.05011665
  7     1e-04  0.5732198  0.05011665
  7     1e-03  0.5923375  0.09290970
  7     1e-02 0.6599845 0.31340264
 7    1e-01 0.6098491 0.18732051
 9    0e+00 0.5791022 0.06611434
 9    1e-04 0.5849845 0.08119773
 9    1e-03 0.6570240 0.29422163
 9    1e-02 0.5918537 0.14461049
 9    1e-01 0.6055728 0.19946221

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were size = 3 and decay = 0.01.
```

***Figure 13*. ANN Results for Confirmed Cases Risk**

```
a 13-3-3 network with 54 weights
inputs: pop_density male_pop female_pop median_age white_pop black_pop asian_pop hispanic_pop amerindian_pop commuters_by_public_transportation median_income employed_pop
unemployed_pop
output(s): .outcome
options were - softmax modelling  decay=0.01
```

Random Forest Fit

The Random Forest Fit yielded the best accuracy results out of all the case category classification models (Figure 14). The best value had an accuracy of 0.78 and a kappa value of 0.57.

***Figure 14*. Random Forest Analysis for Case Category Risk**

```
Random Forest

166 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 150, 150, 149, 150, 149, 150, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.7831699  0.5728614
   7    0.7828023  0.5739984
  13    0.7651144  0.5407019

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

As a result, Random Forest Fit was the chosen classification model to best predict the counties that had the highest risk of COVID-19 confirmed cases. Figure 15 compares all of the classification models used.

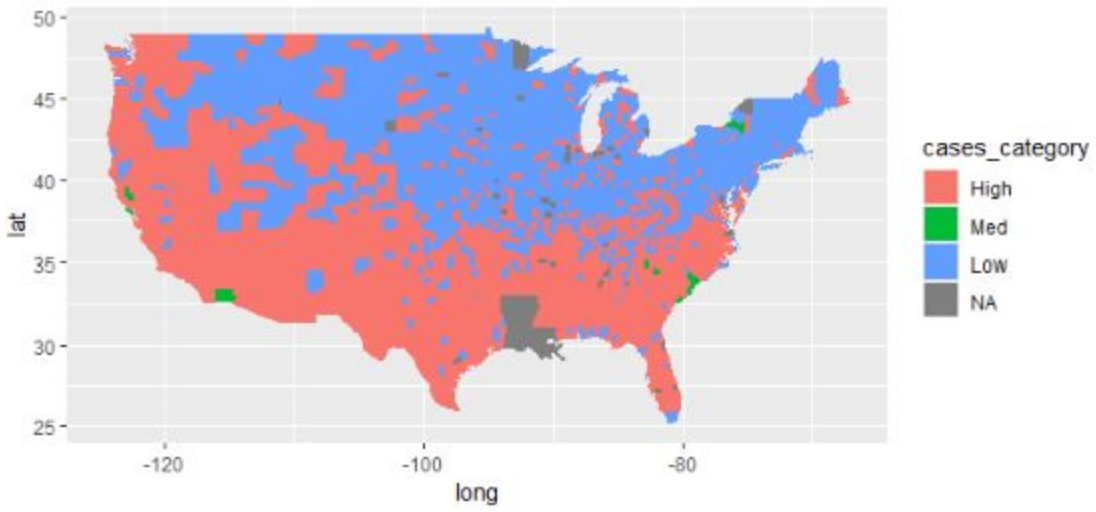
***Figure 15*. Classification Model Comparison for Cases Category**

```
Call:
resamples.default(x = list(decision = fit, ctree = ctreeFit_cases, rules = rulesfit_cases, randomForest
 = randomForestFit_cases, NeuralNet = nnetFit_case, C45fit = C45fit_cases))

Models: decision, ctree, rules, randomForest, NeuralNet, C45fit
Number of resamples: 10
Performance metrics: Accuracy, Kappa
Time estimates for: everything, final model fit

Call:
summary.resamples(object = resamps)

Models: decision, ctree, rules, randomForest, NeuralNet, C45fit
Number of resamples: 10

Accuracy
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
decision     0.7777778 0.8205882 0.8840336 0.8669909 0.9090909 0.9117647    0
ctree        0.5000000 0.6920956 0.7058824 0.7316176 0.8125000 0.8750000    0
rules        0.5625000 0.6470588 0.6966912 0.7099265 0.8088235 0.8750000    0
randomForest 0.6250000 0.7169118 0.7712418 0.7831699 0.8645833 0.9375000    0
NeuralNet    0.5263158 0.5625000 0.6672794 0.6838816 0.7968750 0.9375000    0
C45fit       0.6250000 0.6875000 0.6966912 0.7443627 0.8235294 0.8823529    0

Kappa
                  Min.     1st Qu.    Median        Mean   3rd Qu.      Max. NA's
decision     -0.09090909 -0.04545455 0.0000000 -0.02308103 0.0000000 0.0000000    0
ctree         0.10000000  0.43706897 0.4697824  0.50101481 0.6263027 0.7647059    0
rules         0.20567376  0.36802885 0.4222508  0.45899377 0.6136364 0.7480315    0
randomForest  0.20661157  0.44327250 0.5649351  0.57286137 0.7304790 0.8796992    0
NeuralNet     0.00000000  0.00000000 0.3477131  0.33758869 0.6054299 0.8769231    0
C45fit        0.25581395  0.39849624 0.4299083  0.50866840 0.6671863 0.7687075    0
```

Case Category Prediction Results

Using the Random Fit classification model to predict which counties have the highest risk for COVID-19 cases, Figure 16 was created to map the county results. The results show that most Western and Southern regions are rated as “High” risk with the Midwest and East Coast regions having more counties that are “Low”risk. Therefore inferring that confirmed cases across the United States is not easing but only predicted to grow.

***Figure 16*. Prediction Model Results for Risk of Confirmed Cases in the U.S.**



Deaths Category Modeling

The same approach that was used for case category prediction modeling was used for deaths category prediction modeling. The three states used for the training set were New York, California, and Florida. These states were used to balance the deaths category class data for the deaths training set. Their percentage for risk is listed in Table 7.

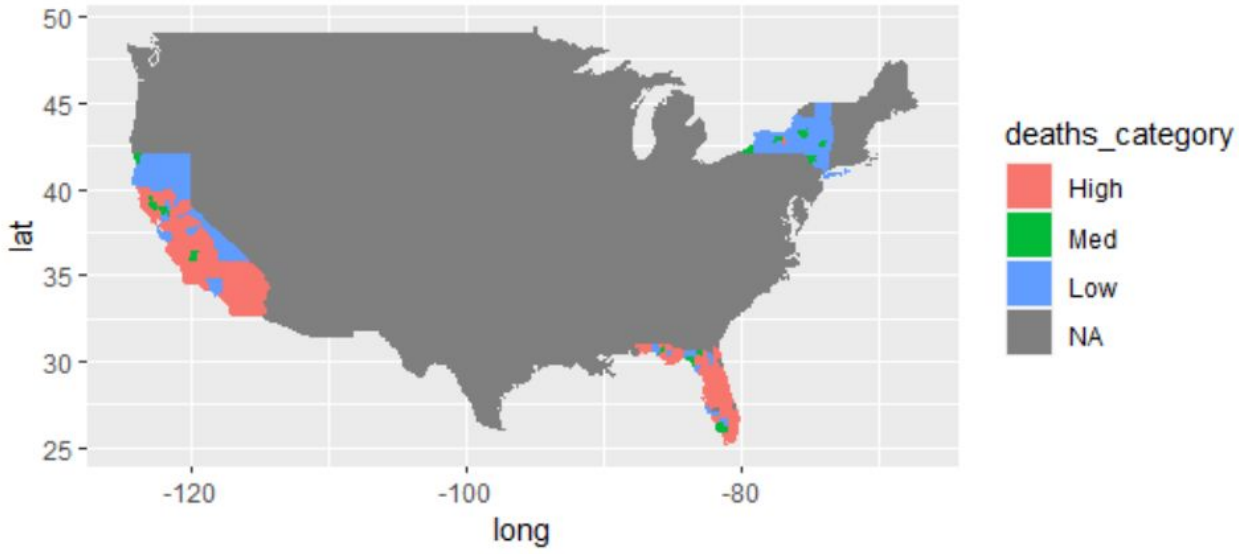
Table 7. **Training States' Percentage of High Risk Labels for Deaths**

State	High Risk Percentage
California	0.55
Florida	0.77
New York	0.01

Source: Johns Hopkins University

Out of these three states, 85 counties were labeled as "High" risk, 14 were "Medium" risk and 88 were "Low" risk (Figure 17).

Figure 17. **Death Category Risk Results for California, New York, Florida**



A Chi-squared test for Independence was used to rank the features in Table 1 that had the greatest attribute importance for the deaths category class by state. The goal for the data analyst when using a chi-square test is to determine whether there is a significant association between the categories of the feature set listed in Table 1 and Deaths Category. Table 4 shows the attribute importance breakdown. "White population" and "Hispanic Population" were the independent variables with the greatest weight (0.42 each) for all of the chi-squared test.

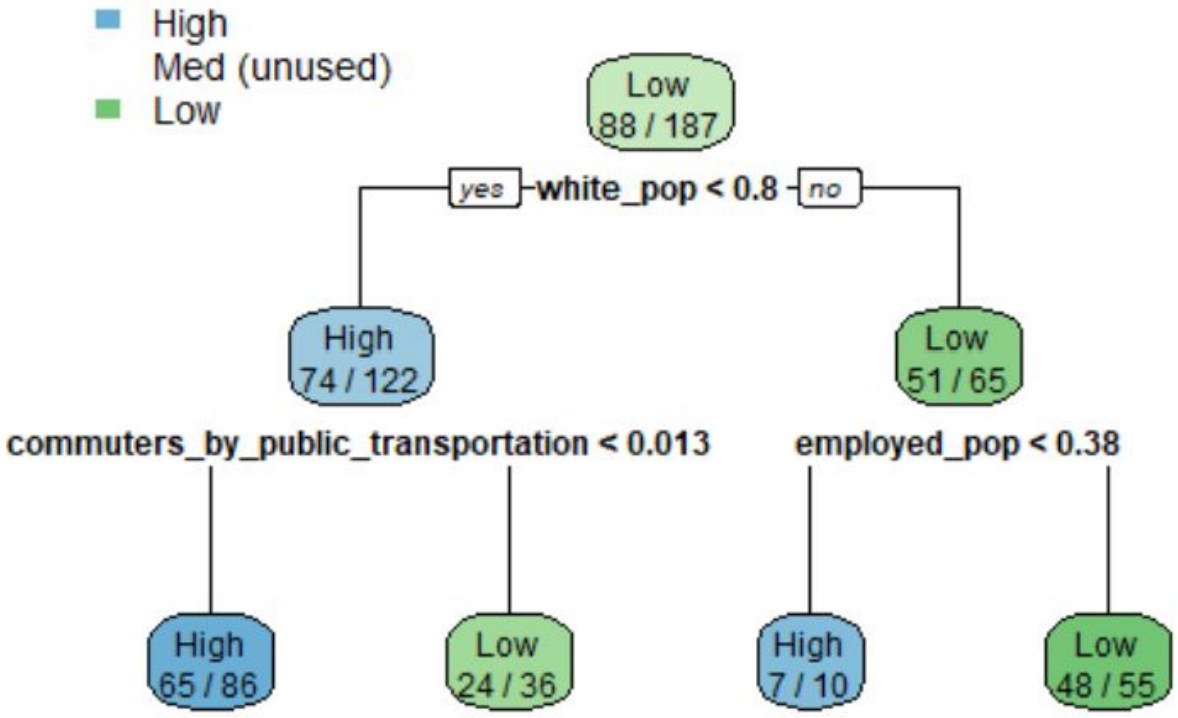
Feature subset selection, where the algorithm tries to find the best set of features using a blackbox approach and a greedy search strategy, concluded white population, hispanic population, population density, male population and female population had the greatest attribute importance. Univariate importance and CFS filter algorithms were also used to test association importance between the features in Table 1 and Case Category. White population had an attribute importance score of 0.27, followed by hispanic population (0.22).

Ten fold cross validation was then used to split the data set and build different case classification models using training data from New York, California and Florida. The classification models used for the deaths category were RPart Decision Trees, Conditional Inference Tree, C 4.5 Decision Tree, the Rules-based classifier PART, Random Forest and Artificial Neural Network (ANN). Like Cases Category, decision trees and rules based classification were used for their simplicity and easier ability to interpret whereas Random Forest Fit was used since it's based on generating a large number of decision trees with subsets selected at random.

Decision Tree using "rpart"

The rpart decision tree was created from the training data derived from states listed in Table 9. Figure 18 shows the results from the rpart analysis. The rpart decision tree model resulted in a top accuracy rate of 0.67 and a kappa value of 0.39. The rpart decision tree also built a classification model that started with whether the white population value for a county was less than 0.80 (80% of the total population).

Figure 18. **Rpart Decision Tree Results for Deaths Category**



C45 Decision Tree

The results of the C45 decision tree algorithm analysis also aligned to the rpart decision tree results. The optimal model had an accuracy of 0.70 with a kappa value of 0.45 (Figure 19).

***Figure 19*. C45 Tree Analysis for Deaths Category**

```
C4.5-like Trees

187 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 167, 169, 168, 168, 169, 168, ...
Resampling results across tuning parameters:

  C      M  Accuracy   Kappa
  0.010  1  0.6818541  0.4148309
  0.010  2  0.7041108  0.4457982
  0.010  3  0.6932921  0.4276060
  0.255  1  0.6525370  0.3772351
  0.255  2  0.6748228  0.4125164
  0.255  3  0.6956123  0.4378434
  0.500  1  0.6630633  0.3964271
  0.500  2  0.6853492  0.4325164
  0.500  3  0.7011386  0.4509952

Accuracy was used to select the optimal model using
 the largest value.
The final values used for the model were C = 0.01 and
 M = 2.
```

Similar to the other decision tree models, the C45 decision tree started it's split with a county's white population (Figure 20). If the white population for a county was greater than 0.89, the death risk category was "Low". Interestingly, if the white population was equal to or less than 0.89 *and* the population density was less than or equal to 7.8 the county risk for COVID-19 cases was "Low".

***Figure 20*. C45 Tree for Deaths Cases Category**

```
J48 pruned tree
------------------

white_pop <= 0.893494
|   pop_density <= 7.8: Low (9.0)
|   pop_density > 7.8
|   |   commuters_by_public_transportation <= 0.013076
|   |   |   white_pop <= 0.789884
|   |   |   |   commuters_by_public_transportation <= 0.000089: Low (4.0/1.0)
|   |   |   |   commuters_by_public_transportation > 0.000089: High (74.0/11.0)
|   |   |   white_pop > 0.789884
|   |   |   |   white_pop <= 0.869671: Low (27.0/9.0)
|   |   |   |   white_pop > 0.869671: High (6.0/1.0)
|   |   commuters_by_public_transportation > 0.013076
|   |   |   amerindian_pop <= 0.0027
|   |   |   |   pop_density <= 207.2: Med (2.0)
|   |   |   |   pop_density > 207.2: Low (28.0/5.0)
|   |   |   amerindian_pop > 0.0027
|   |   |   |   amerindian_pop <= 0.003987: High (5.0)
|   |   |   |   amerindian_pop > 0.003987: Low (3.0)
white_pop > 0.893494: Low (29.0/1.0)

Number of Leaves  :     10

Size of the tree :      19
```

Conditional Inference Decision Tree

The Conditional Inference Decision tree started the tree branch with white population rates greater than or less than .79. The best value derived from this decision tree had an accuracy value of 0.68 and a kappa value of 0.41 (Figure 21).

***Figure 21*. Conditional Inference Decision Tree Analysis for Deaths Category**

```
Conditional Inference Tree

187 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 168, 168, 169, 168, 168, 169, ...
Resampling results across tuning parameters:

  mincriterion  Accuracy   Kappa
  0.01          0.6842105  0.4184414
  0.50          0.6730994  0.3920624
  0.99          0.6517544  0.3522753

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mincriterion = 0.01.
```

PART Rules-based Classification

The PART rules based classifier results were not entirely conclusive against the attribute for white population. For example, the model results in Figure 21 indicate that white populations greater than 0.89 and also less than 0.86 were classified as “Low”. Population density rates less than 7.8 were also classified “Low”. The highest accuracy rating for the PART model was 0.64 with a kappa value of 0.35.

***Figure 21*. PART Rules results for Deaths Category**

```
PART decision list
------------------

white_pop > 0.893494: Low (29.0/1.0)

pop_density <= 7.8: Low (9.0)

commuters_by_public_transportation > 0.013076 AND
amerindian_pop > 0.0027 AND
amerindian_pop <= 0.003987: High (5.0)

commuters_by_public_transportation > 0.013076 AND
pop_density > 207.2 AND
median_age <= 39: Low (15.0/1.0)

commuters_by_public_transportation > 0.013503 AND
pop_density > 218.9 AND
median_age > 39.5: Low (12.0/1.0)

white_pop <= 0.789884 AND
commuters_by_public_transportation > 0: High (81.0/15.0)

white_pop <= 0.869671: Low (30.0/9.0)

: High (6.0/1.0)

Number of Rules  :      8
```

Random Forest Fit Classification

Random Forest Fit provided the best accuracy and kappa results out all the classification models used for deaths category (Figure 23). The number of variables tried at each split was 2. The highest accuracy result was 0.72 with a kappa value of 0.49 (Figure 22). The total number of trees used for this model was 500 and with an error rate of 28.8%

***Figure 22*. Random Forest Fit results for Deaths Category**

```
Random Forest

187 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 168, 168, 169, 168, 168, 168, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.7267974  0.4921772
   7    0.7110079  0.4636858
  13    0.7265050  0.4965926

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

***Figure 23*. Classification Model Comparison for Deaths Category**

```
Models: decision, ctree, rules, randomForest, NeuralNet, C45fit
Number of resamples: 10
Performance metrics: Accuracy, Kappa
Time estimates for: everything, final model fit

Call:
summary.resamples(object = resamps)

Models: decision, ctree, rules, randomForest, NeuralNet, C45fit
Number of resamples: 10

Accuracy
                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
decision     0.5000000 0.6315789 0.6842105 0.6763158 0.7222222 0.7777778    0
ctree        0.5263158 0.6250000 0.7032164 0.6842105 0.7368421 0.7894737    0
rules        0.5263158 0.5500000 0.6055556 0.6431166 0.7083333 0.8421053    0
randomForest 0.4444444 0.6842105 0.6842105 0.7267974 0.8374613 0.9473684    0
NeuralNet    0.4736842 0.5614035 0.6476608 0.6643567 0.7675439 0.8421053    0
C45fit       0.5294118 0.6293860 0.7111111 0.7041108 0.7763158 0.8888889    0

Kappa
                  Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
decision      0.09090909 0.3324937 0.4178218 0.3994593 0.4782476 0.5739645    0
ctree         0.16990291 0.3250000 0.4383721 0.4184414 0.5133249 0.6200000    0
rules         0.11111111 0.2070485 0.2916449 0.3502749 0.4482864 0.7000000    0
randomForest -0.05882353 0.4000000 0.4285391 0.4921772 0.7018425 0.9000000    0
NeuralNet     0.02564103 0.1863636 0.3443808 0.3726918 0.5573529 0.7121212    0
C45fit        0.11111111 0.3037572 0.4594156 0.4457982 0.5900000 0.8000000    0
```

Artificial Neural Network

As noted previously, Artificial Neural Network (ANN) is an algorithm model that is an assembly of inter-connected nodes and weighted links. The output node sums up each of its input values according to the weights of its links. The ANN results for deaths category was a 13-7-3 network with 122 weights. Ten-fold cross validation was also used in this classification model which yielded an highest accuracy rating of 0.66 with a kappa value of 0.37 (Figure 24).

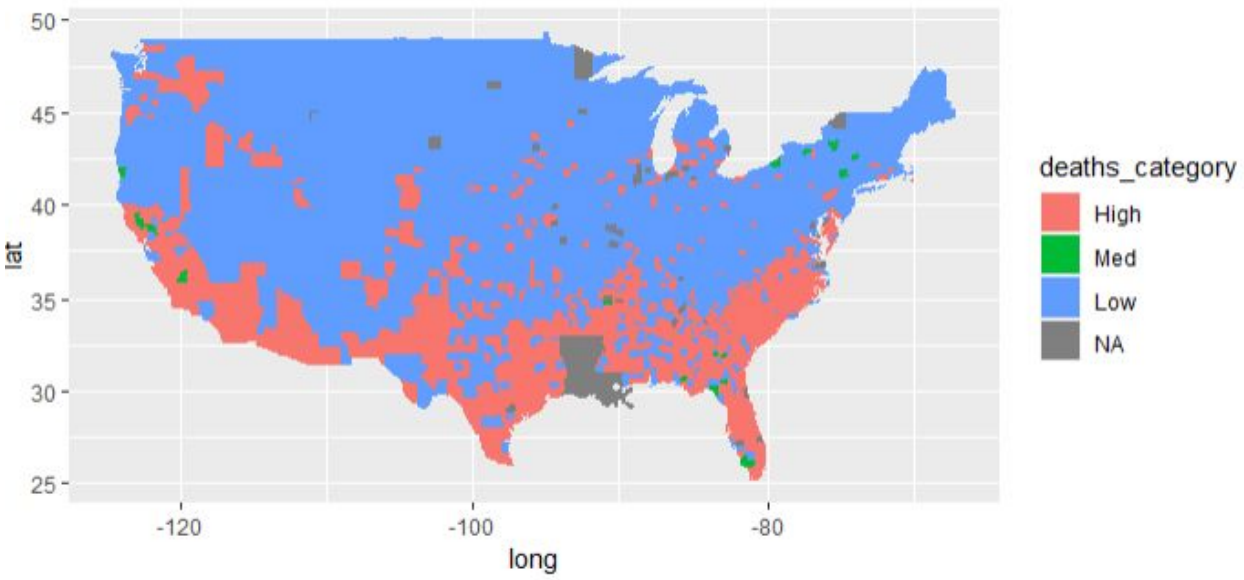
Figure 24. ANN Analysis for Deaths Category Risk

```
Neural Network

187 samples
 16 predictor
  3 classes: 'High', 'Med', 'Low'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 168, 169, 167, 168, 169, 168, ...
Resampling results across tuning parameters:

  size  decay  Accuracy   Kappa
  1     0e+00  0.4713158  0.000000000
  1     1e-04  0.4657602  0.000000000
  1     1e-03  0.4713158  0.009497207
  1     1e-02  0.5625439  0.187429471
  1     1e-01  0.5572515  0.162568327
  3     0e+00  0.4602047  0.003958469
  3     1e-04  0.4812865  0.043117406
  3     1e-03  0.5362281  0.132318063
  3     1e-02  0.6158480  0.277622488
  3     1e-01  0.5953216  0.239220624
  5     0e+00  0.4707310  0.019968411
  5     1e-04  0.4707310  0.010000000
  5     1e-03  0.5075731  0.097954501
  5     1e-02  0.6324561  0.309771029
  5     1e-01  0.6061111  0.260912159
  7     0e+00  0.4859357  0.056673409
  7     1e-04  0.4707310  0.023997207
  7     1e-03  0.5520468  0.154661486
  7     1e-02  0.5661404  0.187573673
  7     1e-01  0.6643567  0.372691793
  9     0e+00  0.4546491  0.010868103
  9     1e-04  0.4762865  0.035710799
  9     1e-03  0.5797953  0.223132105
  9     1e-02  0.5920760  0.243991802
  9     1e-01  0.5803509  0.218443887

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were size = 7 and decay = 0.1.
```

Deaths Category Prediction Results

Using the Random Fit classification model to predict which counties have the highest risk for COVID-19 deaths, Figure 25 was created to map the county results. The results indicate that areas predicted to have the highest risk for COVID-19 deaths are regions in the Western, Southwestern and Southern parts of the U.S.

***Figure 25*. Prediction Model results for risk of Deaths in the U.S.**



Conclusion

The prediction results conclude that percentages of white population for a given county is the most important attribute in predicting the current risk a county has for COVID-19 cases and deaths. On average, white populations greater than .80 have a "Low" risk prediction for COVID-19 cases and deaths. Prediction results for the classification categories do seem to indicate that blanket state or federal policy to reduce cases might not be necessary in areas of the U.S. that are predicted to have a "Low" risk of deaths and cases based on data as recent as August 6th, 2020.

References

[1]Google Public Data Set for Corona Virus.
https://cloud.google.com/blog/products/data-analytics/free-public-datasets-for-covid19

[2]Unacast Social Distancing Scores. https://www.unacast.com/covid19/social-distancing-scoreboard

[3] USAFacts data sets. [https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/]

[4]COVID-19 Projections Using Machine Learning. [https://covid19-projections.com/]

[5]. R Code References for Classification Models by Michael Hahsler.
https://michael.hahsler.net/SMU/EMIS7332/