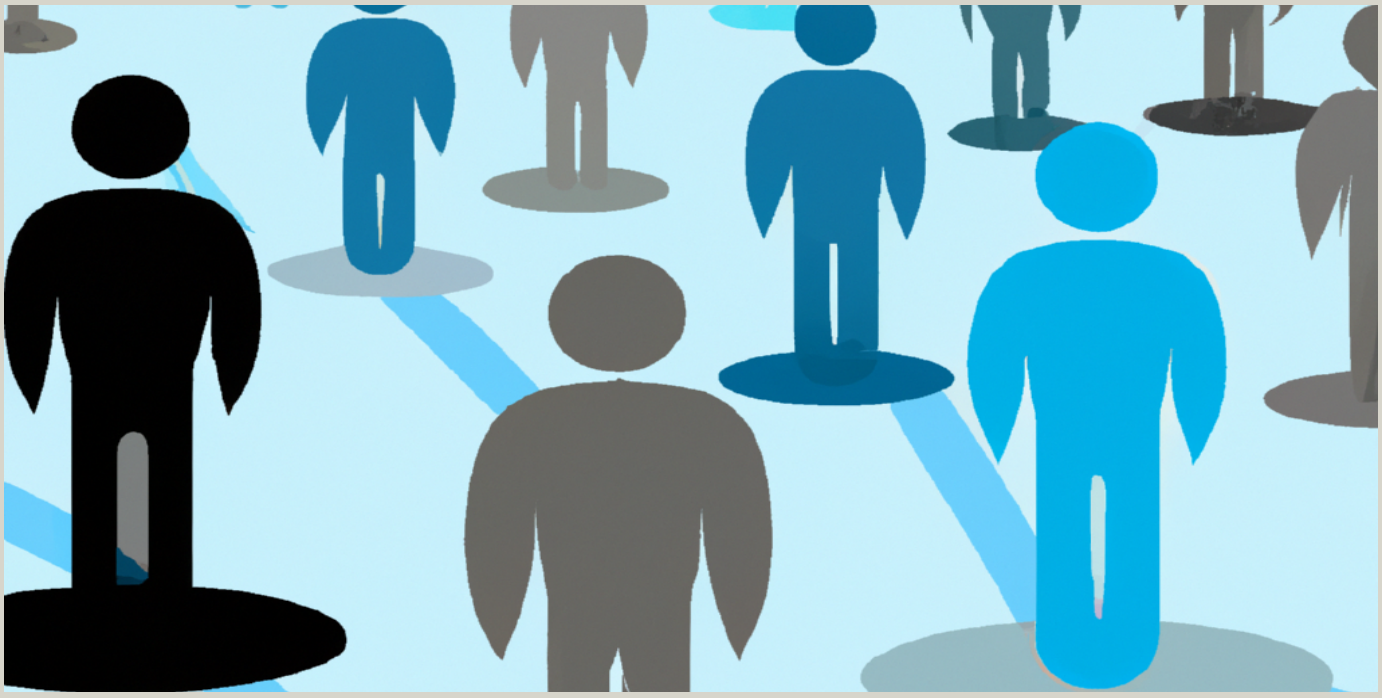


Part 3:COVID-19 Classification Analysis for Policy Recommendations

By Lawrence Lim

Table of Contents

Introduction

❖ Business Understanding

➢ Main Actors in the Field

With the potential for a new COVID wave very high and with the CDC issuing warnings to state health departments and officials, my job as a data scientist for the city of Dallas is to utilize classification to determine

which states and counties will suffer the greatest or least impact from this new wave of COVID. Ultimately, the Texas Department of Health and Safety Services is counting on me to determine which counties could be most susceptible to a COVID outbreak, using a national dataset of various attributes collected regarding a particular county. In our previous research, we were able to quantify the importance of various attributes, reduce dimensionality, and isolate the most important attributes contributing to the number of deaths a particular county faces. Given this, we must determine a methodology for predicting which counties have a high, medium, or low mortality rate by analyzing the attributes. By knowing this, the Department of Health can hopefully develop the infrastructure needed to distribute new vaccines, medication, and hospital resources, in a manner that is proportional to the level of severity that COVID is predicted to have in a particular area.

➢ Relevant Numbers and Studies

The importance of this research stems from the fact that vaccines have a limited period of effectiveness. According to a report in the New England Journal of Medicine, COVID vaccines from Moderna had a gradual decline in antibodies/effectiveness. Thus, the vaccines last approximately one year before requiring a booster shot to keep the immune system resilient to COVID. With this new wave of the virus prepared to hit and with only 23% of individuals having received a COVID booster in Texas aside from the original 2 vaccine shots, it is expected that a large number of the population will require vaccines, leading to a limited supply of the new booster shot, forcing us to make the decision of selecting which counties should receive the vaccines first.

Furthermore, it is also important to note that current national information already supports our hypothesis that communities with a lower Gini index and affluence will have greater deaths. This fact is supported by the Poor People's Pandemic Report, which states that people in penurious counties have died at twice the rate as those in affluent counties. Therefore, it is my job to determine whether that also applies to Texas.

Data Preparation

❖ Defining Classes

➢ Class Definitions and Rationale

Firstly, in order to adequately differentiate different cities and account for the various population sizes, I will utilize the % of deaths in relation to the total population as the classification measure. I will then classify the values as being low, medium, high, or extreme.

Summary Statistics

Min	1st Quartile	Median	Mean	SD	3rd Quartile	Max
0	0.06885	0.11979	0.13509	0.09347343	0.17707	0.84713

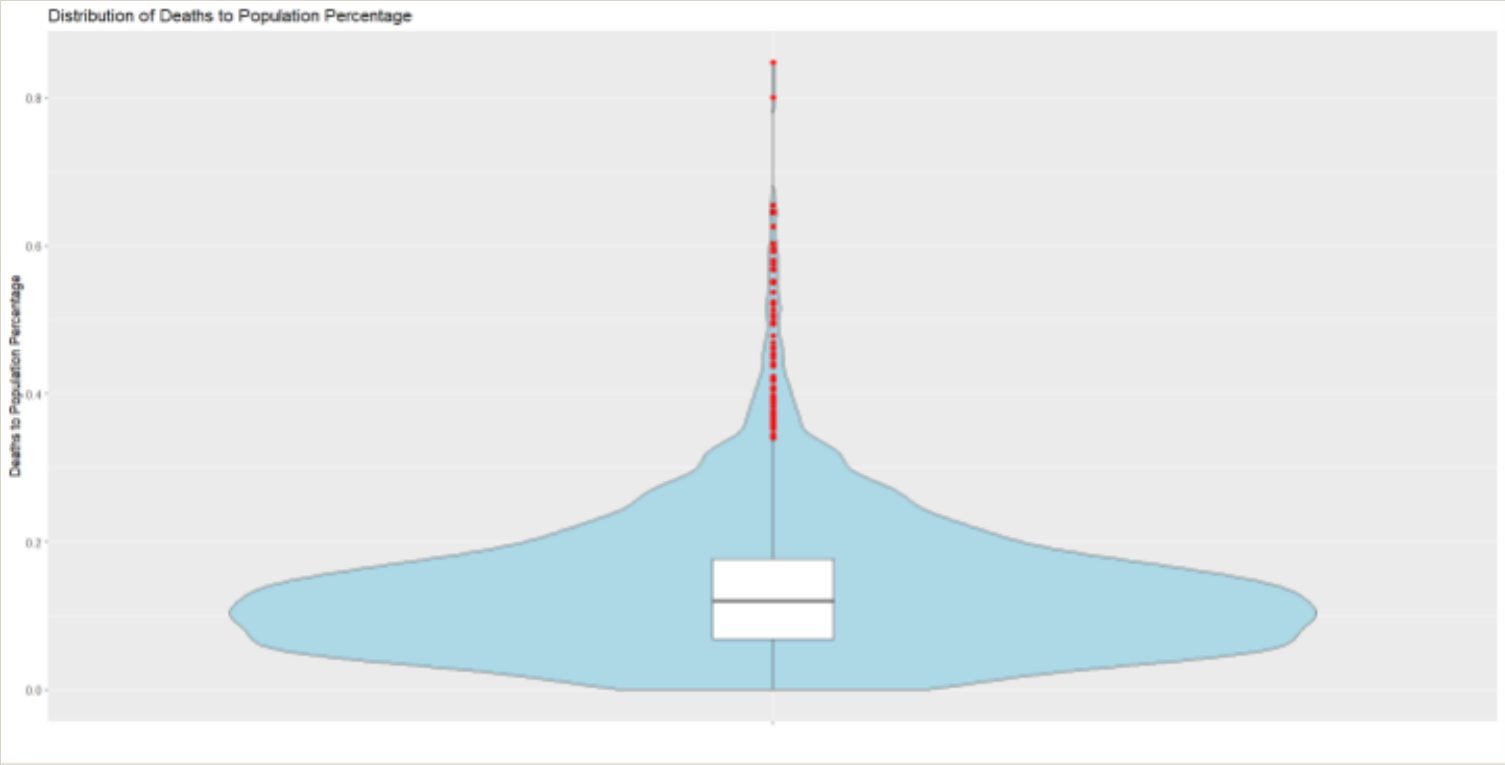


Figure 1: Violin Plot of the Distribution of % of Population that Died

Firstly, it is important to note that the values above are already in % form as I multiplied them by 100. The % of Covid deaths in a county in the US ranges from about 1 in 1700 people as the minimum to close to 1 in 100. When looking at the mean, we see that the mean is greater than the median, indicating that the data is right-skewed. And when looking at the distribution, we see that the vast majority of values fall under the range of about 0.06 to 0.17. The remaining values above and below are thus outliers, that will be handled differently depending on the models used later. Given the quartiles and violin plot, I will use the following as benchmarks:

Low Death Rate	Less than or equal to 0.06
Medium Death Rate	Between 0.07 to 0.16
High Death Rate	0.17 to 0.39
Extreme Death Rate	Greater than or equal to 0.4

❖ Feature Engineering

➢ Dealing with Missing Data

Percent_income_spent_on_rent	1

aggregate_time_travel_work	143
median_year_structure_built	1
Percent_income_spent_on_rent	1

Above are the missing data in the national dataset. The only attribute of concern is aggregate_time_travel_work which is missing 4.54% of its values. Considering this is not an insignificant amount, imputing would be an option. However, in the below feature selection steps, I end up deleting aggregate_time_travel_work anyway. Therefore, once I delete the aggregate_time_travel_work column, I will simply remove the remaining 3 rows with NA values in the other 3 columns.

➢ Dealing With Outliers

As shown above, there are numerous outliers that do not fall in the quartiles. Considering the fact that our goal is to predict the lethality of COVID, it is essential to keep the outlier values, at least for now. Using Cook's statistic, I determined the counties that might have a significant impact on a regression model's coefficients and performance. Out of the over 3000 counties in the national dataset, over 270 of them can be considered outliers. When comparing the outliers to the original data, we see that the outlier counties have an average population of 154,806 compared to 100,970 for the entire data set. Furthermore, when looking at the median income, we see that the outliers earn around $300 less in median income compared to the entire dataset. Therefore, an outlier amount of deaths to population size could be caused by differences in attribute values that are important in understanding the lethality of COVID.

➢ Determining Predictive features

The starting attributes I have chosen to use are derived from those I selected in Project 1 as they eliminate any duplicitous attributes or smaller sub-categories of an overarching, larger attribute.

Column	**Data Type**	**Description**	**Example**	**Potential Use Cases of Features in Clustering**
County Name	nominal	This is the name of the particular county.	Tarrant County	This is a nominal variable and cannot be directly used in clustering algorithms. However, it can be used as a label to identify clusters after the clustering has been performed.
State	nominal	This is the state abbreviation the county resides in.	TX	This is a nominal variable and cannot be directly used in clustering

				algorithms. However, it can be used as a label to identify clusters after the clustering has been performed.
Population_1_year_and_over	ratio	This is the population of a county that is over one year old	1000000	This is a ratio variable that could be used to group counties based on their population size.
Confirmed	ratio	This is the number of confirmed cases (as reported on January 19th, 2021).	1201	This is a ratio variable that could be used to group counties based on their COVID-19 prevalence.
Deaths	ratio	This is the number of deaths (as reported on January 19th, 2021).	321	This is a ratio variable that could be used to group counties based on their COVID-19 severity.
median_year_structure_built	ordinal	This is the median year that a home was built in for a particular county.	1972	This is an ordinal variable that could be used to group counties based on the age of their housing stock.
percent_income_spent_on_rent	ratio	This is the average % amount of income spent on rent for residents in the county.	52%	This is a ratio variable that could be used to group counties based on their housing affordability.
Percent_male	ratio	This is the percentage of males in the county	50%	This is a ratio variable that could be used to group counties based on their gender demographics.
median_income	ratio	This is the median income of an individual in the county.	$60,000	This is a ratio variable that could be used to group counties based on

				their income level.
% graduate degree or more *	ratio	This is the % of individuals that have a graduate degree or higher.	41%	This is a ratio variable that could be used to group counties based on their educational attainment.
aggregate_time_travel _work	ratio	This is the total amount of time (in minutes) spent by individuals traveling to work on January 19th.	1668430	This is a ratio variable that could be used to group counties based on their commuting patterns.
genie_index	Ratio	This is an index that measures inequality in a particular county.	0.417	This is a ratio variable that could be used to group counties based on their level of income inequality.
Deaths_to_Pop_Pe percentage	Ratio	This is a measurement of the proportion of deaths in relation to the population size of a particular county	0.04176%	This is the classifier variable that will be utilized to predict low, medium, high, and extreme numbers of deaths.

One method I will use in feature selection is LASSO. While LASSO and Ridge Regression are both means of handling multicollinearity in features, LASSO is able to reduce the coefficient values of features all the way to zero. Thus, LASSO is able to select for us the features that it considers to be the most important while outright rejecting features that do not contribute any new insights to the dataset. This is important considering that we are starting off with only a subset of the entire dataset of over 40 attributes. Thus, the coefficient for a single attribute may be small, even if it does contribute sufficiently to the variation in the data.In essence, LASSO setting the coefficient to zero better differentiates attributes that don't contribute to the target variable and those that do.

The results in the below table show that the Percent of Male individuals and the aggregate time one takes to travel to work does not contribute to the target variable. Thus, we may eliminate those variables.

Variable	**Coefficient**
genie_index	6.258292e-01
Percent_graduate_or_more	-4.729948e-01
Percent_income_spent_on_rent	-2.793948e-03

median_year_structure_built	-9.196280e-04
Deaths	1.560176e-04
Confirmed	-8.332647e-07
median_income	2.528907e-07
geo_id	-2.387839e-07
Population_1_year_and_over	-9.820098e-08
Percent_Male	0.000000e+00
aggregate_time_travel_work	0.000000e+00

Since relying on one method of feature selection is not good practice, I should verify the results above by utilizing another method. One such method is the VIF score, which is a method of quantifying the level of multicollinearity in the data. Typically a value of 10 would be considered a high level of multicollinearity, which means that an attribute might not be contributing unique variation to the data that another attribute is not already doing.

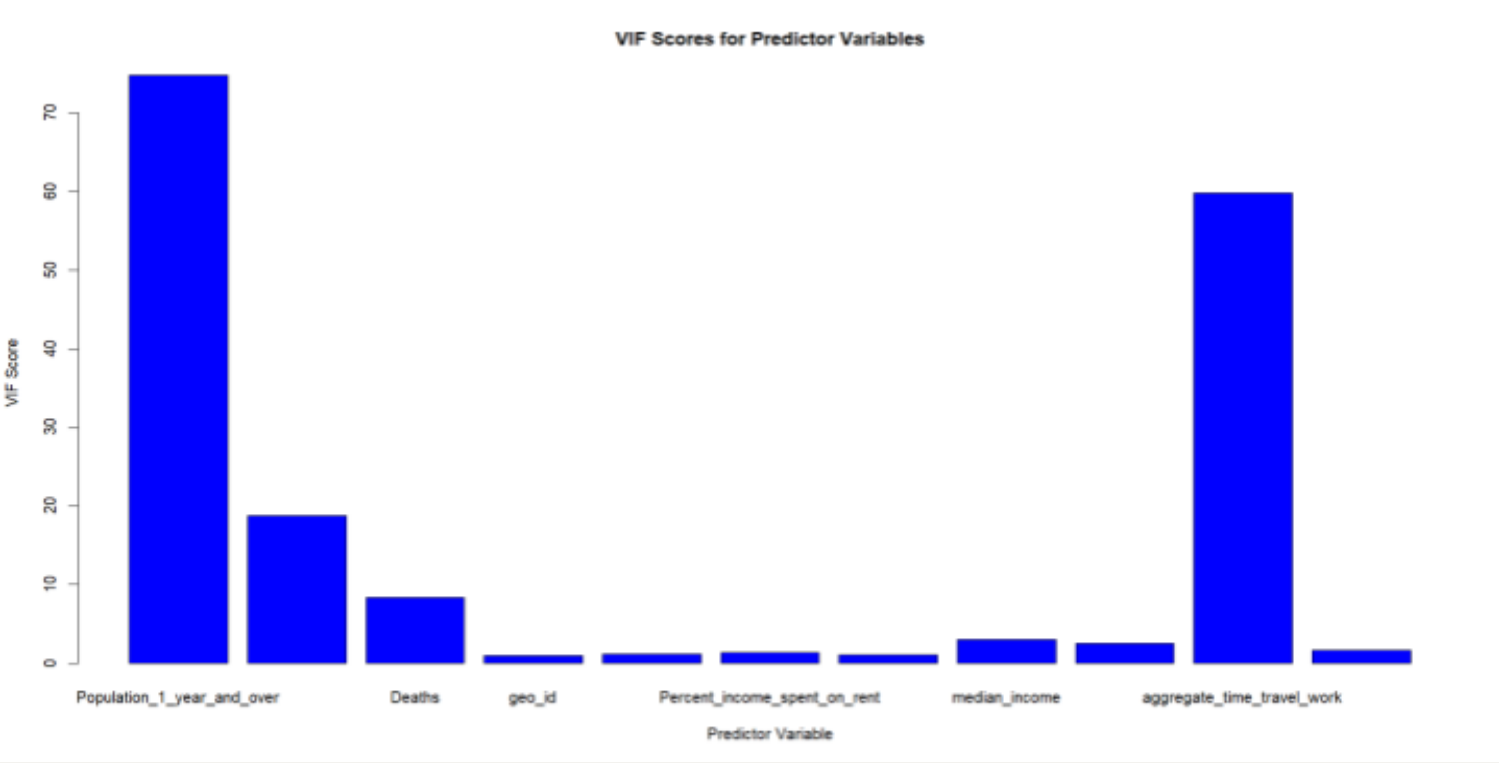


Figure 2: VIF Scores For Variables

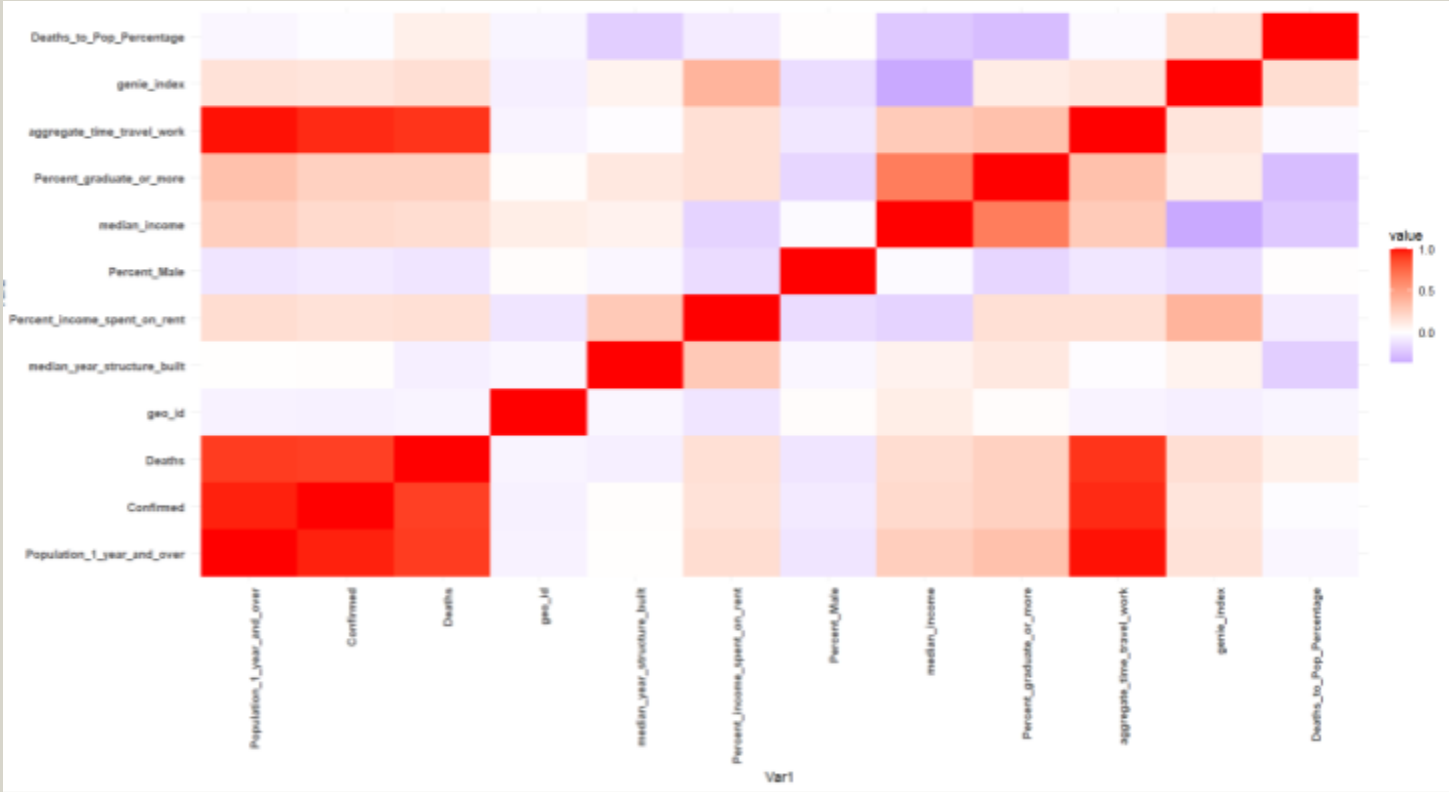


Figure 3: Correlation Matrix for Multicollinearity Analysis

The above results show that potential variables with multicollinearity (score >10) include the aggregate time travel to work, Population size, and the number of confirmed cases. And with the correlation matrix, we can see that the multicollinearity of population, confirmed cases, deaths, and time for traveling to work are all multicollinear. Based on the results of feature selection, I will eliminate the aggregate_time_travel_work, and confirmed cases as a means of reducing multicollinearity. Furthermore, % males will also be eliminated as it does not contribute to the classifier.

➢ Final Attributes Selected

Column	Data Type	Description	Example	Potential Use Cases of Features in Clustering
County Name	nominal	This is the name of the particular county.	Tarrant County	This is a nominal variable and cannot be directly used in clustering algorithms. However, it can be used as a label to identify clusters after the clustering has been

				performed.
State	nominal	This is the state abbreviation the county resides in.	TX	This is a nominal variable and cannot be directly used in clustering algorithms. However, it can be used as a label to identify clusters after the clustering has been performed.
Population_1_year_and_over	ratio	This is the population of a county that is over one year old	1000000	This is a ratio variable that could be used to group counties based on their population size.
Deaths	ratio	This is the number of deaths (as reported on January 19th, 2021).	321	This is a ratio variable that could be used to group counties based on their COVID-19 severity.
median_year_structure_built	ordinal	This is the median year that a home was built in for a particular county.	1972	This is an ordinal variable that could be used to group counties based on the age of their housing stock.
percent_income_spent_on_rent	ratio	This is the average % amount of income spent on rent for residents in the county.	52%	This is a ratio variable that could be used to group counties based on their housing affordability.
median_income	ratio	This is the median income of an individual in the county.	$60,000	This is a ratio variable that could be used to group counties based on their income level.
% graduate degree or more *	ratio	This is the % of individuals that have a graduate degree or higher.	41%	This is a ratio variable that could be used to group counties based on their educational attainment.

genie_index	Ratio	This is an index that measures inequality in a particular county.	0.417	This is a ratio variable that could be used to group counties based on their level of income inequality.
Deaths_to_Pop_Pe percentage	Ratio	This is a measurement of the proportion of deaths in relation to the population size of a particular county	0.04176%	This is the classifier variable that will be utilized to predict low, medium, high, and extreme numbers of deaths.

Modeling

❖ Data Preparation for Modeling

➢ Splitting into training and testing sets

In splitting the data into training and testing sets, I have decided to implement k-fold cross validation as my method. The benefit of k-fold cross validation is that it allows for a more robust estimate of the model's performance. By splitting the data into k subsets and training the model on k-1 subsets while testing on the remaining subset, the model is evaluated on all parts of the data. This can help to reduce the impact of chance variability such as any predefined ordering in the data and provide a more accurate estimate of the model's true performance.

➢ Hyperparameter Tuning

The two options for hyperparameter tuning are grid search and random forest. Grid search involves defining a grid of hyperparameters and exhaustively searching over all possible combinations of values within the grid to determine the optimal hyperparameters. This can be computationally expensive but guarantees that the optimal combination of hyperparameters will be found within the search space. On the other hand, Random search involves randomly sampling hyperparameters from a defined distribution to determine the optimal hyperparameters. This approach can be less computationally expensive than grid search but may require a larger number of samples to find the optimal hyperparameters. Given that we are using three different models, it will be important to generate all possible combinations of hyperparameter values for each model to determine whether the search space is small or large, thus determining how to determine the optimal hyperparameters.

For the Random Forest Model, Support Vector Machine, and Logistic Regression, I get the results below for the # of possible combinations of hyperparameter values:

Algorithm	**# of combinations**
Random Forest	36

SVM	24
Logistic Regression	36

The results above show that the grid space is relatively small for all three models. Thus, grid search remains a good option that will not be too computationally intensive. It also allows for a more thorough exploration of the hyperparameter space.

For Random Forest, we find that the optimal hyperparameters to utilize to train on the overall dataset are:

Mtry: The number of variables randomly sampled as candidates at each split in the decision tree. In this case, the best value for mtry is 2.
num.trees: The number of trees in the random forest. In this case, the best value for num.trees is 200.

For SVM, we find that the optimal hyperparameters to utilize to train on the overall dataset are:

C: the cost parameter, which controls the trade-off between achieving a low training error and a low testing error. In other words, it determines the balance between overfitting and underfitting. In your case, the best value found for C is 5.6.
sigma: the kernel width parameter for the radial basis function (RBF) kernel, which is used by the SVM model. It determines the decision boundary's shape and individual data points' influence. The best value found for sigma is 0.1.

Multinomial Regression does not have a similar optimization process to the algorithms above. It has the regularization parameter that represents the strength of the penalty that is applied to the coefficient values. However, this parameter can be tuned in the cross-validation process during the actual model implementation to maximize performance/reduce the risk of overfitting.

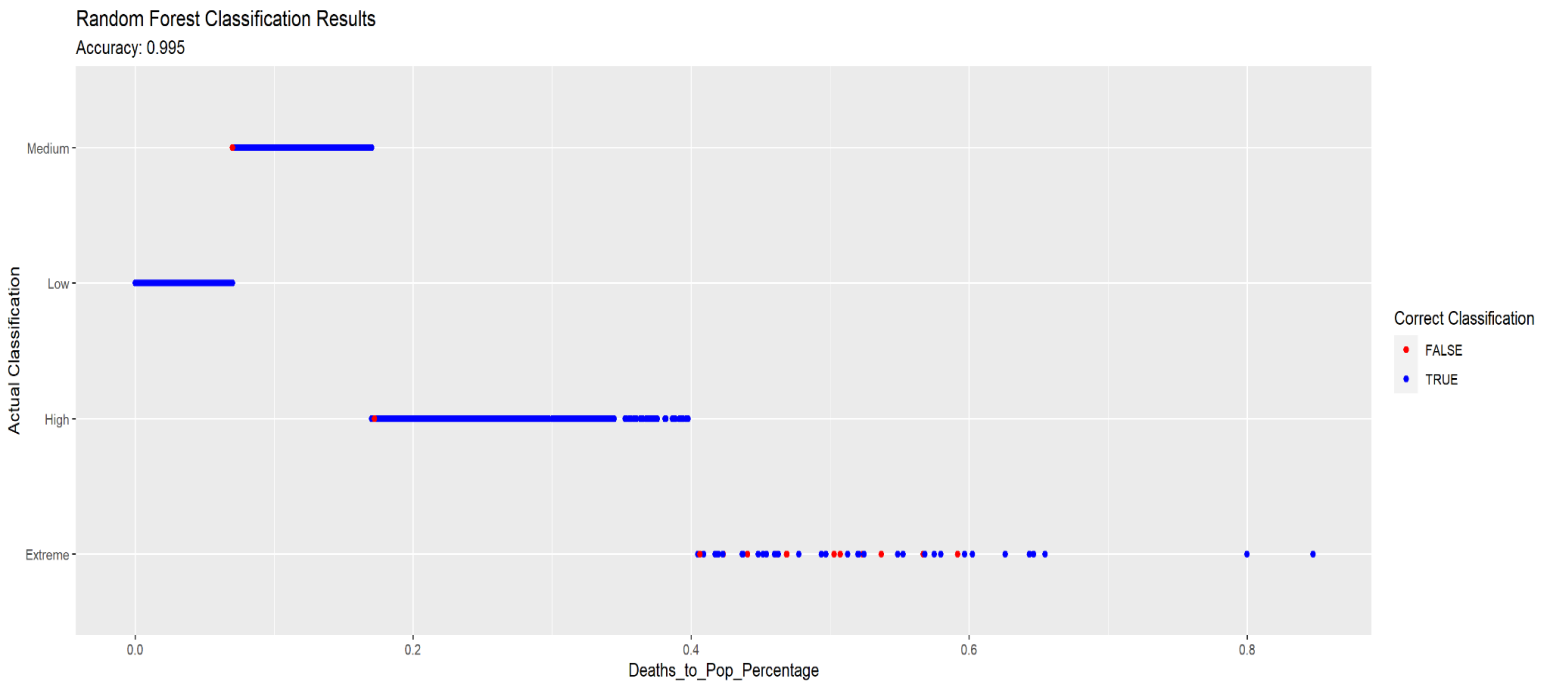
❖ Classification Modeling

➢ Model 1 (Random Forest)

■ Description and Method

In a random forest algorithm, each tree is built from a random subset of the training data, and a random subset of the predictor variables is used to determine the best split at each node of the tree. This randomness helps to reduce overfitting and improve the generalization performance of the model. When predicting, the random forest algorithm aggregates the predictions of all the individual trees to make the final prediction. This aggregation reduces the variance of the model and improves its overall accuracy. Below is the result of random forest classification using the randomForest and caret R packages. The dataset is preprocessed to remove missing values and the target variable, Deaths_to_Pop_Percentage_Classifier, is converted to a factor. The model is tuned using a grid search approach with cross-validation and the optimal hyperparameters are selected. The final model is then trained on the full dataset using the optimal hyperparameters. 5-fold cross-validation is used.

When looking at the cities, we see that Random Forest did a near-perfect job predicting the cities that are Low and medium. The model struggled most with being able to distinguish the boundary between a high and an extreme number of cases. Considering the objective of our algorithm is to determine which cities are going to be impacted the worst, being able to differentiate the extreme cities is essential for applying the model to real life, regardless of whether the overall accuracy is high or not.



Confusion Matrix:

Prediction	Extreme	High	Low	Medium
Extreme	30	0	0	0
High	11	812	0	0
Low	0	0	800	1
Medium	0	0	1	1475

Class	F1 Score	Precision	recall
Low	0.998751560549313	0.998751560549313	0.998751560549313
medium	0.998	0.998	0.998
high	0.985203452527743	0.98641975308642	0.983990147783251
extreme	0.714285714285714	0.697674418604651	0.731707317073171

Accuracy: 99.46%

The ROC curve below is a visualization of the true positive and false positive rate and allows us to see the performance of the model. It is a measure of how well a model is able to differentiate between the classes. Thus, seeing the AUC_ROC for classifying the most affected counties can be valuable. In a model that performs well in identifying the 'extreme' category, like the above, the ROC curve will go straight up the y-axis and then straight across the x-axis and will have an area under the curve close to or exactly 1. If the model did not do well in guessing the extreme values, the curve would appear more jagged/curved with an area closer to 0.5, indicating that the model performance was closer to simply guessing.

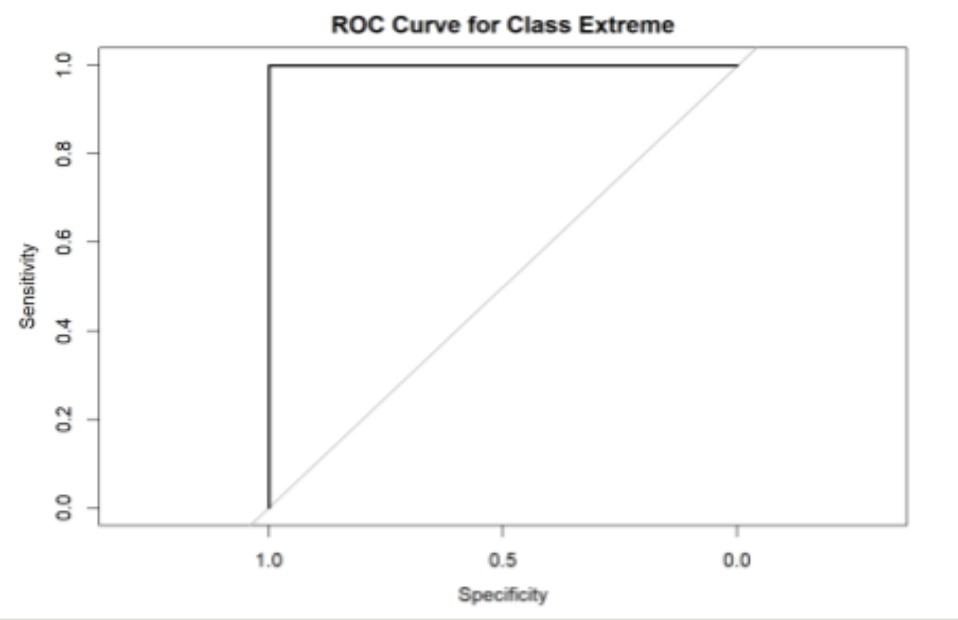


Figure: ROC Curve for Extreme classification in Random Forest

- Potential Advantages

One advantage of this classification over options like Logistic Regression is that it is capable of modeling non-linear regression. Furthermore, compared to Logistic Regression, Random Forest is more resilient to noise and outlier values because it constructs decision trees based on random samples of the data. Furthermore, in comparison with SVM, Random Forest can handle a larger amount of dimensionality due to it being less computationally intensive. Finally, since Random Forest is an ensemble method, it utilizes bootstrap aggregation to combine different models derived from the original data. This can make the prediction more robust and accurate.

➢ Model 2 (Support Vector Machine)

- Description and Method

In the code above, the SVM algorithm is used for classification, where the target variable is "Deaths_to_Pop_Percentage_Classifier" and the predictors are all other variables in the dataset. The data is split into five folds using cross-validation. A hyperparameter grid is defined to confirm optimal values for the cost and sigma parameters using the "svmRadial" method. The "train" function from the "caret" package is used to find the optimal hyperparameters using grid search. The optimal hyperparameters are then used to train the final model on the entire dataset. The "SVM" function is used to fit the model using the radial kernel function. The optimal cost and sigma hyperparameters are set to 0.03125 and 8, respectively.

The results show that though the model did not make a false positive error in classifying a county as extreme, 6% of the high values that it predicted ended up being extreme. The model made the most errors in its medium prediction, where 19% of counties predicted with a medium death ratio actually had a high ratio or low ratio of deaths. This could be indicative that the model was not fully resilient to the fact that the dataset is heavily imbalanced towards counties with medium death rates. Note that the red highlights signify the worst type of error- a false positive in which the prediction is of lower severity than the actual.

Confusion Matrix

Predictions	Extreme	High	Low	Medium
Extreme	11	0	0	0
High	40	641	0	0
Low	0	0	698	32
Medium	0	170	103	1445

Class	F1 Score	Precision	recall
Low	0.916	0.95616	1
medium	0.859249329758713	0.790382244143033	0.941262848751836
high	0.911822338340954	0.871410736579276	0.956164383561644
extreme	0.904538341158059	0.978334461746784	0.841094295692666

Accuracy: 89.01%

- Potential Advantages

Potential advantages of using SVM on this data set include the fact that it is suited for a high number of features and a medium size dataset (a few thousand rows). Furthermore, SVM is technically the most robust method for outliers of the three models as it tries to maximize the margin between the classes rather than using the mean as the method of splitting nodes.

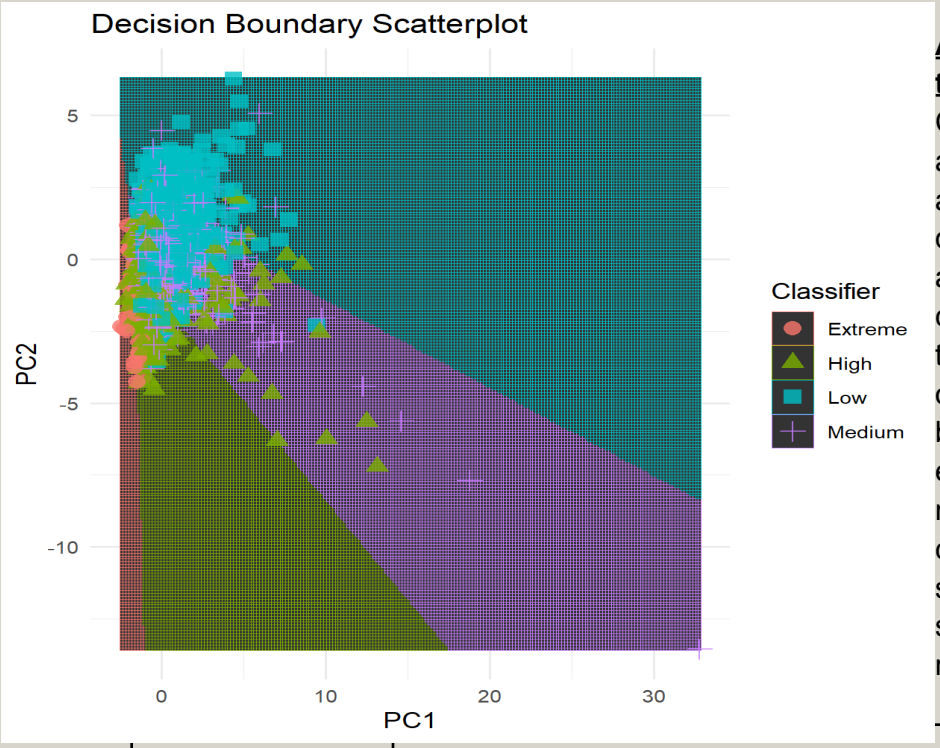
➢ Model 3 (Multinomial Regression)

- Description and Method

Though my original intent was to perform linear regression, I ultimately decided to perform multinomial regression as it can handle classifying predictions based on more than 2 options. Furthermore, unlike logistic regression, multinomial regression can handle curves. Below, we see that the multinomial regression performed well at predicting which counties may not face a high death rate. However, the model has a tendency to make false-negative errors for counties with medium or high death rates. In practice, this could lead us to send resources to cities that are doing alright while the cities that are in need of assistance are not allocated any.

Confusion Matrix

Predictions	Extreme	High	Low	Medium
Extreme	37	30	0	0
High	10	600	0	0
Low	0	0	801	0
Medium	4	211	0	1477



Attempt to Increase Interpretability through Dimensionality Reduction:
Considering that one of the benefits of using a multinomial model is its interpretability, I attempted to utilize PCA to reduce the dimensionality while maintaining an adequate amount of information to differentiate the classifier. The graph to the left was an attempt to reduce the number of features from the original 11 to 2 so that the differences could be easily interpreted. Though we see that the extreme cases are identified successfully, the model is incapable of correctly classifying the counties with a low or high ratio of deaths, as shown by the overlapping blue and green symbols. In summary, dimensionality reduction reduces performance too much.

	The Independent Variables Influence			
The Target Variables Options—>	**Coefficients:**	High	Medium	Low
	Class 1 (Low)	-104.91346	-10.89091	114.75022
	Class 2 (Medium)	-34.87146	85.80392	-57.18802
	Class 3 (High)	134.22129	46.17637	-143.73180
	Class 4 (Extreme)	130.10355	2.84934	-160.21566

The coefficients show the change in the log-odds of the dependent variable for a one-unit increase in each category of the independent variable, relative to a reference category. For instance, in the independent variable high, we see that the target variable being high and extreme is of higher probability while medium and low are less probable. However, we see that an increase in the high attribute by a unit also increases the probability of being in the extreme category, which we saw earlier with the false negative in the confusion matrix.

❖ Overall Performance Comparison

There are distinct performance differences between the three models. We see that Random Forest was the most accurate and also did not face any issues with false negatives, which were present in both the SVM and multinomial models. When comparing the SVM to the multinomial model, we see that the SVM model struggled more in isolating the 'medium' classification when comparing the # of correct medium classifications to the number of incorrect classifications. However, when looking at the false positives, we see that the SVM predicted less counties as being medium, that actually were high, compared to the multinomial model. However, when looking at the high predictions, we see that the situation is reversed. The SVM misclassified 40 extreme death-rate counties as just being high compared to 14 for the multinomial model. When looking at the advantages that each model brings, the results aren't surprising. SVM is the most robust model against outlier values, which though good in theory, might have caused it to not distinguish the extreme COVID values from the high values. The benefits of a multinomial regression is that it is good for interpreting relationships, especially those that are linear, in a simplistic manner that associates each feature with a coefficient representing its relation to the target. However, it doesn't perform as well in handling more sophisticated relationships.

Evaluation

❖ Model Usefulness for Stakeholders

The Random Forest model in particular is useful for the national and state health department in being prepared for future COVID waves. One additional benefit of using the Random Forest rather than the SVM is that the performance is magnitudes better than the SVM. For instance, if the stakeholders decide to narrow each row to being each individual town and city instead of county, we could end up with over a hundred thousand rows which would probably make SVM impossible to run. One potential issue with the model is that it can only be trained on prior COVID variants rather than emerging variants. Thus, it cannot account for changing factors that influence the lethality of the virus. In a study done by Two Augusta researchers during the transition from Delta to Omicron, they found that while Alpha or Delta caused higher death rates in urban areas that had higher population density, Omicron had the opposite effect. With the introduction of the vaccine, researchers found that the least vaccinated areas (more rural) had 1.6 times higher death rate than even the more urban areas with higher vaccination. Since our data was compiled prior to Omicron, it is possible that the model might not work on Omicron or successive variants.

❖ Assessing Model Value in Practice

To assess the model value, I will evaluate its performance with a greater emphasis on the confusion matrix and the avoidance of false positives. Accuracy is still useful to look at a glance, however, it does not allow us to glean a full picture of the model's performance as it is possible to have a high accuracy while making many false negative errors. Furthermore, because our data is imbalanced with an emphasis towards the medium classification, accuracy could be skewed higher as the model simply selects the majority class.

Deployment

In practice, the model would be used to predict the locations with higher COVID death rates to send resources in preparation for when the wave actually hits. One potential implementation that can allow the model to be

adapted to new variants is seeing how COVID waves impact other countries before the United States. COVID variants have spread in other countries like Southeast Asia and even Europe before spreading in the United States. Therefore, if we can collect necessary information on the spread of COVID in other countries' districts and train the Random Forest, we can train the model and prepare more accurately months before the wave hits. In this instance, the model would have to be updated for each successive wave and also updated with the latest information on the spread of the latest variants. Especially during the transition periods between an old variant and new, dominant variant, it will be important that the data fed into the model only pertains to the variant that will become dominant and not on the continued spread of the old variant. For example, if we were to train the model on a set of mostly Delta cases during the transition to Omicron, months later, the model would not accurately predict the areas with new COVID cases.

Works Cited

New study shows omicron variant of COVID-19 hit rural America harder – Jagwire (augusta.edu)

How Long Will Your Coronavirus Vaccination Last? > News > Yale Medicine

A Poor People's Pandemic Report: Mapping the Intersection of Poverty, Race and COVID-19 in the US (unsdsn.org)