# Data Mining (Classification)

Team 10: Jeevan Rai & Abhilash Narayanan

**Executive Summary:**

This report is on analyzing various aspects of social structures that were collected during the COVID-19 pandemic. These social structures are referred as variables that went through various classification models to identify high-risk Texas counties. The models were used to identify feature(s) that have the highest importance in terms of predicting counties with high death rates due to COVID-19. After training the dataset through various classification models, it was found that K-Nearest Neighbor and Random Forest classifier had the highest performance at predicting counties with high death rates based on metrics like accuracy and Kappa value. Amongst various features, it was found that the feature “commute” had the highest importance at predicting high death rates at high-risk counties. The results from the models could be very useful to the agency like Texas Department of State Health Services (DSHS) as they could use such results to implement various preventive measures (as described in evaluation section) to control high mortality rates during such pandemic

**Table of Contents**

[Data Mining (Classification) 1](#_Toc183716702)

[1 Business Understanding 3](#_Toc183716703)

[2 Data Preparation 3](#_Toc183716704)

[3 Modeling 5](#_Toc183716705)

[2.1 Random Forest Classifier 7](#_Toc183716706)

[2.2 K-Nearest Neighbor Classifier 9](#_Toc183716707)

[2.3 Gradient Boosted Decision Trees (XGBoost) 11](#_Toc183716708)

[2.4 Support Vector Machines (SVM) 13](#_Toc183716709)

[2.5 Artificial Neural Network (ANN) 15](#_Toc183716710)

[3 Evaluation 20](#_Toc183716711)

[4 Deployment 21](#_Toc183716712)

[5 List of References 22](#_Toc183716713)

[6 Appendix 22](#_Toc183716714)

# Business Understanding

A widespread disease COVID-19 (also known as coronavirus disease 2019) started at the end of 2019, and it quickly spread throughout the entire world impacting every aspect of human society. It was first identified in December 2019 in Wuhan district in China. Since then, there have been several kinds of studies conducted on the impact of this pandemic. The data on those studies are available for the general public. Among those datasets, we will be looking at four different datasets. The primary focus of the analysis is based on understanding the impact of this pandemic on various aspects of human society all around the world from 2019. These policies were primarily implemented to enforce “social distancing” amongst people. Social distancing involves measures taken to reduce close contact between individuals to slow the spread of infectious diseases such as COVID-19. By implementing measures like social distancing, mask-wearing, and hygiene practices, the goal is to spread out the number of cases over a longer period, resulting in a flatter curve.

Various classification models were used to predict Texas counties with high deaths per 10,000 (i.e. high mortality rate). The models were then compared against each other to find the best performing classification model for the dataset. Also, the feature with the highest importance in making the prediction was identified using the models’ results. Based on the results, several recommendations were made available to DSHS. All of the recommendations are very practical and could be very helpful to prevent high mortality rate at high-risk counties during similar pandemic. The report works on answering following questions:

* From all the selected features for making the prediction, which one has the highest significance?
* After determining the feature with highest significant, what can DSHS do with this finding?
* Which classification model was found to be the most useful model and why?
* What are the counties that were predicted as the high-risk counties by the best performing model? How well did the best performing model performed compared to the ground truth?
* After determining the high-risk counties, what can DSHS do with this finding?

# Data Preparation

The report uses two datasets “COVID-19\_Global Mobility” and “COVID-19\_cases\_plus\_Census”. The primary objective of the report is to use the datasets to classify the US counties as low or high risk in terms of the COVID-19 confirmed cases and death rates. Since the dataset COVID-19\_cases\_plus\_Census only contains records for the date 01/19/2021, the observations recorded on the same date were extracted from the dataset COVID-19\_Global Mobility. The report focuses only on Texas counties. So, the observations for Texas counties were extracted from both datasets. Both of these datasets are then merged. The final merged dataset has observations on 204 Texas counties.

The merged dataset was checked for missing values. There were feature columns with various numbers of missing values (see below). Since most of the features have too many missing values, they were replaced with the corresponding average values.

|  |  |
| --- | --- |
| **Features** | **Number of missing values** |
| retail\_and\_recreation\_percent\_change\_from\_baseline | 91 |
| grocery\_and\_pharmacy\_percent\_change\_from\_baseline | 104 |
| parks\_percent\_change\_from\_baseline | 155 |
| transit\_stations\_percent\_change\_from\_baseline | 114 |
| workplaces\_percent\_change\_from\_baseline | 9 |
| residential\_percent\_change\_from\_baseline | 93 |

Features with missing values

Next, all the irrelevant or less significant features were dropped from the dataset. See below for basic statistics on the final dataset. These predictive features were used to perform all the classifications.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Total Population | 600 | 12,635 | 24,898 | 133,668 | 66,671 | 4,525,519 |
| Hispanic Population | 0.03454 | 0.17705 | 0.25266 | 0.34352 | 0.48529 | 0.99185 |
| Black Population | 0 | 0.01504 | 0.05254 | 0.07161 | 0.10209 | 0.33743 |
| Male (50 and above) | 0.1067 | 0.1564 | 0.1857 | 0.1854 | 0.2116 | 0.3057 |
| Female (50 and above) | 0.1192 | 0.1698 | 0.1949 | 0.2004 | 0.2305 | 0.3554 |
| Income ($50K - $100K) | 0.04964 | 0.09331 | 0.10466 | 0.10275 | 0.11199 | 0.14785 |
| Rent > 50% Income | 0 | 0.01207 | 0.01587 | 0.01745 | 0.0218 | 0.06957 |
| Commute (work outside home) | 0.466 | 0.7021 | 0.7697 | 0.7608 | 0.8285 | 0.9794 |
| Worked at Home | 0 | 0.009951 | 0.014673 | 0.015404 | 0.018924 | 0.045182 |
| Transit Stations % Change from Baseline | -61 | -11.06 | -11.06 | -11.06 | -11.06 | 76 |
| **Metric** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Workplaces % Change from Baseline | -49 | -27 | -21 | -22.69 | -17 | -8 |
| Cases per 10,000 | 231.1 | 590 | 754.7 | 785.7 | 942.4 | 1829 |
| Deaths per 10,000 | 3.309 | 12.016 | 16.478 | 17.705 | 22.083 | 54.608 |
| Death per Case | 0.003789 | 0.015099 | 0.02219 | 0.023541 | 0.029217 | 0.09322 |

Basic statistics on predictive features

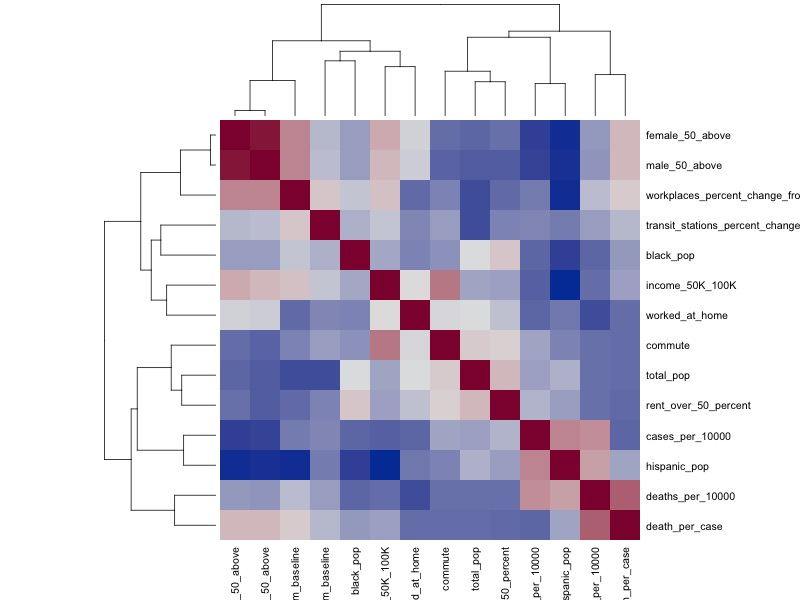
In the table above, the values for the features cases\_per\_10000, deaths\_per\_10000, and death\_per\_case are the number of people per 10,000 out of the total population. The rest of the features (except for “Transit Stations % Change from Baseline” and “Workplaces % Change from Baseline”) are normalized values as their counts were divided by the total population. These two features were not normalized like other features because they represent percent change in the visits to public spaces like transit stations and workplaces from a baseline. These feature processing operations were performed to normalize each feature.

All the Texas counties will be evaluated based on deaths per 10,000. Furthermore, the threshold for the counties to be considered as high is 16. That means the counties are classified as high if their values for deaths\_per\_10000 is greater than 16. This threshold value was strategically chosen as it created a very well-balanced training and testing datasets for the counties that were chosen as training and testing datasets. Based on the counties selected for the training dataset, the class labels for TRUE (high) and FALSE were 11 and 12 for the training dataset respectively. Similarly, the class labels were 88 as FALSE and 93 as TRUE for the testing dataset. There were reasons why the classes were defined this way.

Counties for training dataset: (DFW metropolitan: collin, dallas, denton, ellis, johnson, kaufman, parker, rockwall, tarrant, wise), (from project 2: bell, cameron, el paso, hidalgo, nueces), terry, martin, lubbock, wichita, cherokee, hale, maverick, parmer.

# Modeling

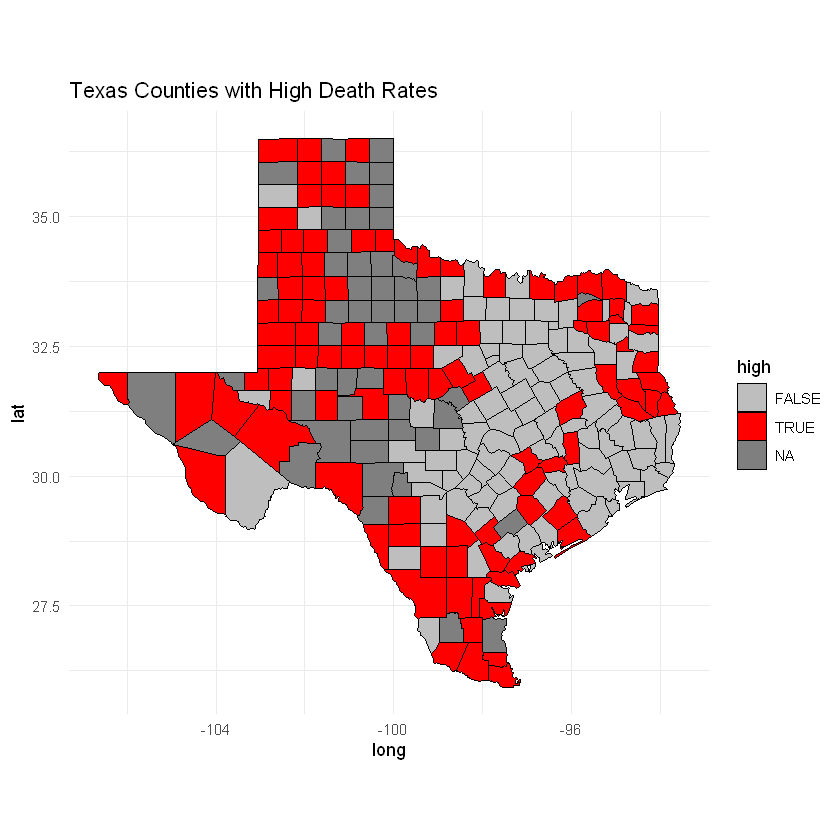
Before beginning to perform classification, the chosen features of the final dataset were evaluated in a correlation plot to get a glimpse of their correlations with deaths\_per\_10000, cases\_per\_10000, and death\_per\_case.



Correlation heat map for dataset features

Looking at the above plot, we can see that the feature “worked\_at\_home” seems to have the highest positive correlation to the target variable “deaths\_per\_10000”. At the moment, this is based on the dataset with all the counties included. This assessment will be validated throughout the experimentation of various classification models

throughout the report.

Texas Counties with high deaths per 10,000

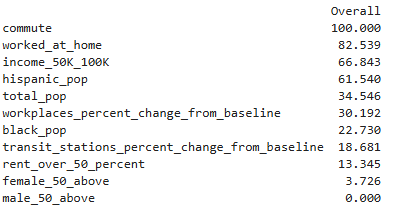
The above heat map shows all the Texas counties with high deaths per 10,000 (labelled as TRUE). There are 104 counties (out of 204) that have high deaths per 10,000. That means about half of the Texas counties fall under this category.

### 2.1 Random Forest Classifier

For the training dataset, variables like county\_name, cases\_per\_10000, deaths\_per\_10000, and death\_per\_case were removed because they can’t be used for training a classifier. The classifier training was performed using 10-folds cross validation to introduce randomness in the selection of training samples. The classifier contained 23 individual data points or observations where each data point represents whether the death rate is high or not. There were 11 features selected for predicting the target variable (deaths\_per\_10000 as being high or not). Thus, there were two classes for the target (TRUE or FALSE).

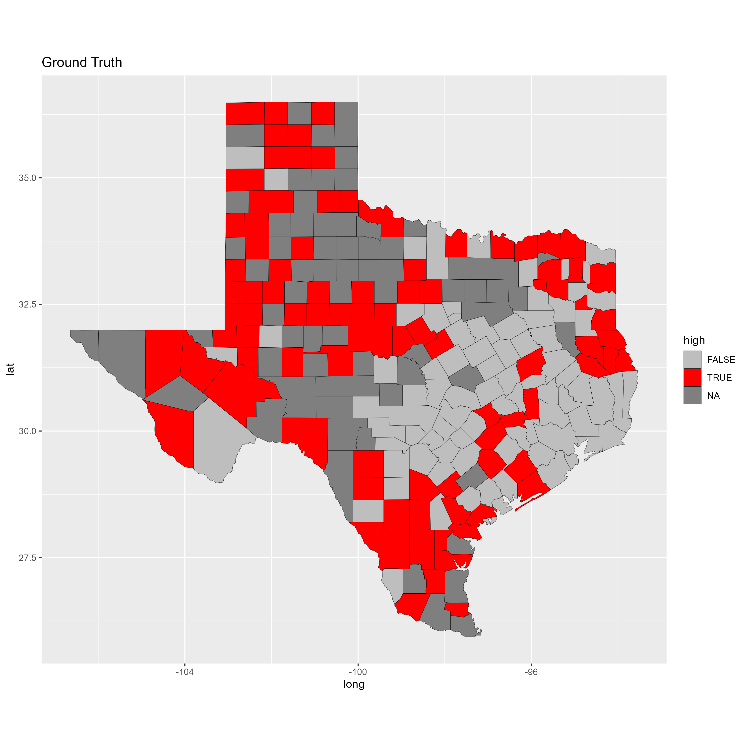
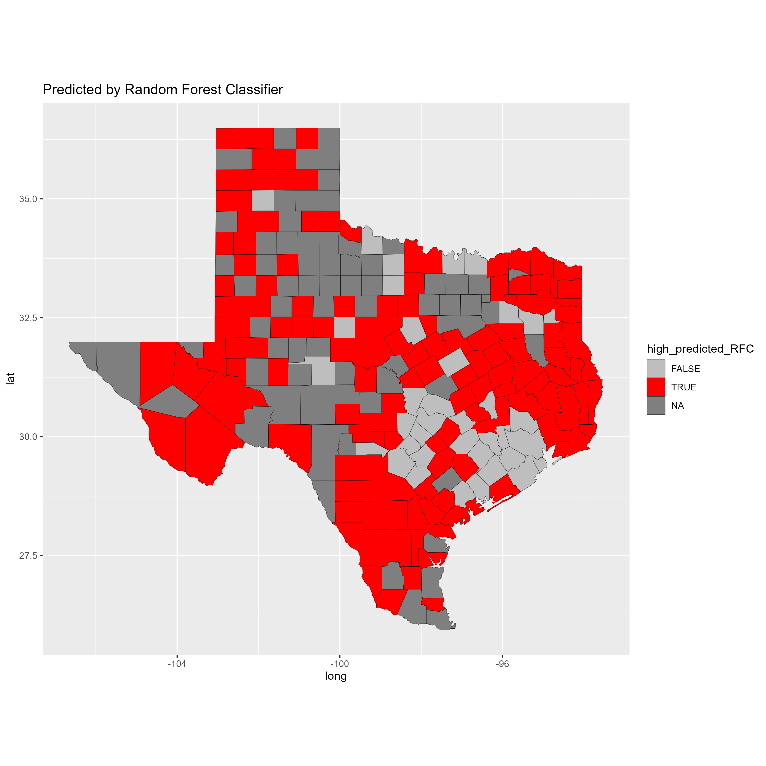
|  |  |  |
| --- | --- | --- |
| **mtry** | **Accuracy** | **Kappa** |
| 2 | 0.85 | 0.70 |
| 6 | 0.75 | 0.50 |
| 11 | 0.67 | 0.34 |

Based on the above table, the final value used for the model was mtry = 2 which means that the algorithm selected 2 features at each split to find the best split. Both accuracy and Kappa metrics indicate that the best model setting was when 2 features were selected at each split. This suggests that considering fewer features at each split (mtry = 2) leads to a better-performing random forest model for the training dataset.



Variable importance for Random Forest Classifier

Now looking at the importance of the variables for predicting the target using this classifier, we can see that the variable “commute” plays the most significant role in predicting the target while the variable “male\_50\_above” is not important at all to predict the target.

Prediction by Random Forest Classifier

Shown above is the comparison between ground truth and prediction made by the model using heat map. From the visual observation, it can be noted that this model has a relatively high accuracy for the prediction as it was able to correctly predict most of the counties.

Confusion matrix + Performance metrics for Random Forest classifier

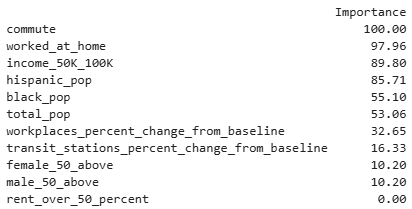
Looking at the confusion matrix, we can see the model has a true negative value of 28 which indicates the number of cases corrected predicted as FALSE (i.e. not high rate). So, it correctly identified 31.82% of the actual FALSE cases. It has a true positive value of 84 which indicates the number of cases correctly predicted as TRUE (high rate). So, it correctly identified 90.32% of the actual TRUE cases. Thus, it has an advantage of high performance in predicting high death rate cases. Overall, the accuracy of the modal is 61.88% (i.e. correctly predicted only 61.88% of the cases correctly). Additionally, it only has 9 cases that it predicts as high but are actually not high. This is a relatively low false negative value which is also an advantage of using such classifier in this kind of pandemic related analysis. There is 95% confidence that the model true accuracy is between 54.35% and 68.98%. With the kappa value of 0.2249, there is a fair agreement between the predicted and actual classifications indicating that there is considerable room for improvement in the performance.

### 2.2 K-Nearest Neighbor Classifier

For the training dataset, variables like county\_name, cases\_per\_10000, deaths\_per\_10000, and death\_per\_case were removed because they can’t be used for training a classifier. The classifier training was performed using 10-folds cross validation to introduce randomness in the selection of training samples. The classifier contained 23 individual data points or observations where each data point represents whether the death rate is high or not. There were 11 features selected for predicting the target variable (deaths\_per\_10000 as being high or not). Thus, there were two classes for the target (TRUE or FALSE). The training dataset was scaled before training the model. Five different k values (1, 3, 5, 7 and 9) were selected for the classifier.

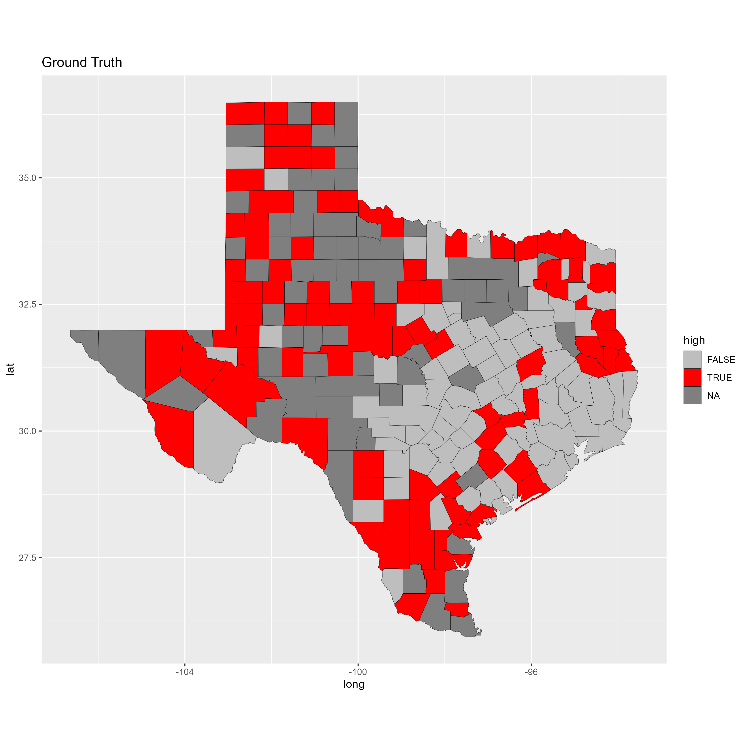
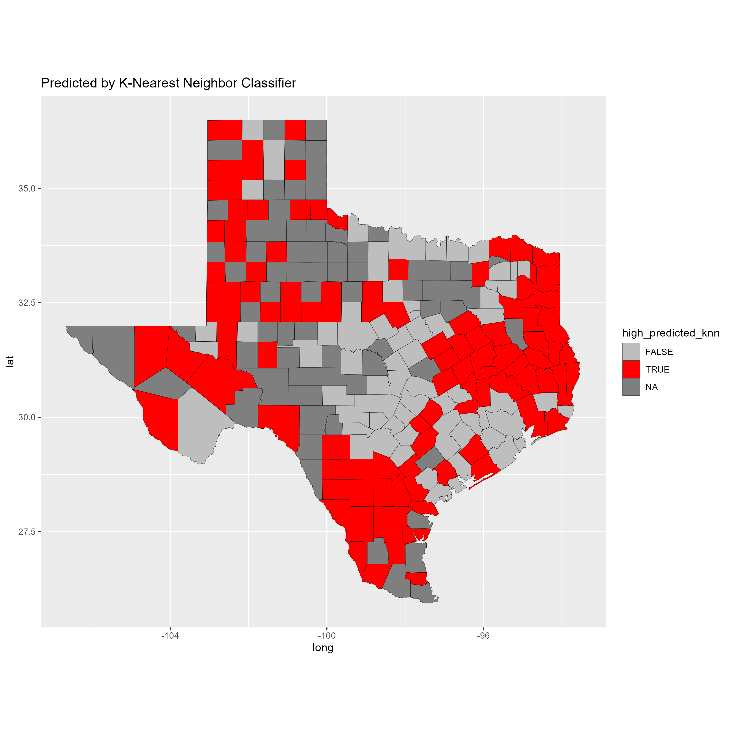
|  |  |  |
| --- | --- | --- |
| **k** | **Accuracy** | **Kappa** |
| 1 | 0.883 | 0.74 |
| 3 | 0.866 | 0.70 |
| 5 | 0.833 | 0.64 |
| 7 | 0.700 | 0.34 |
| 9 | 0.666 | 0.29 |

Based on the above table, the model performed the best with accuracy of 0.883 and kappa value of 0.74 when the number of neighbors was 1.



Variable importance for K-Nearest Neighbor Classifier

Now looking at the importance of the variables for predicting the target using this classifier, we can see that the variable “commute” yet again plays the most significant role in predicting the target while the variable “rent\_over\_50\_percent” is not important at all to predict the target.

Prediction by K-Nearest Neighbor Classifier

Shown above is the comparison between ground truth and prediction made by the model using heat map. From the visual observation, it can be noted that this model has a higher prediction performance than the Random Forest classier. Let’s evaluate that observation through some meaningful metrics.

Confusion matrix + Performance metrics for Random Forest classifier

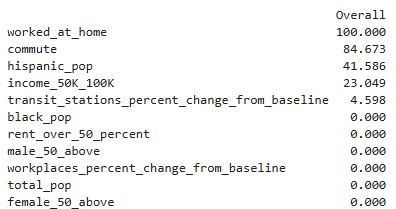
Looking at the confusion matrix, we can see the model has a true negative value of 48 which indicates the number of cases corrected predicted as FALSE (i.e. not high rate). So, it correctly identified 54.55% of the actual FALSE cases. It has a true positive value of 68 which indicates the number of cases correctly predicted as TRUE (high rate). So, it correctly identified 73.12 % of the actual TRUE cases. Thus, the advantage of using this classifier is that it is more reliable in predicting high death rate cases. Compared to other values, it has comparatively lower false negative value (i.e. 25) which means it is not so bad at inaccurately predicting cases not being high while they actually are. This is another advantage of using this classifier in a pandemic related analysis like this. Overall, the accuracy of the modal is 64.09% (i.e. correctly predicted only 64.09% of the cases correctly). There is 95% confidence that the model true accuracy is between 56.64% and 71.07%. With the kappa value of 0.2779, there is a fair agreement between the predicted and actual classifications indicating that there is considerable room for improvement in the performance.

### 2.3 Gradient Boosted Decision Trees (XGBoost)

For the training dataset, variables like county\_name, cases\_per\_10000, deaths\_per\_10000, and death\_per\_case were removed because they can’t be used for training a classifier. The classifier training was performed using 10-folds cross validation to introduce randomness in the selection of training samples. The classifier contained 23 individual data points or observations where each data point represents whether the death rate is high or not. There were 11 features selected for predicting the target variable (deaths\_per\_10000 as being high or not). Thus, there were two classes for the target (TRUE or FALSE). The model has been setup with the maximum depth of the trees as 5 to allow the model to capture more intricate patterns. The value of “colsample\_bytree” has been set to 0.6 indicating that 60% of the features are randomly selected for each tree. The learning rate has been set to 0.1 for the model. As the dataset is not highly complex, the gamma value has been set to 0 indicating that pruning of the tree is not implemented. The parameter to control the minimum sum of weights in a child node has been set to 1. Finally, the 50% of the training data is randomly selected to grow each tree which is implemented through subsample value. All of these parameters collectively help in fine-tuning the model to achieve a balance between bias and variance.

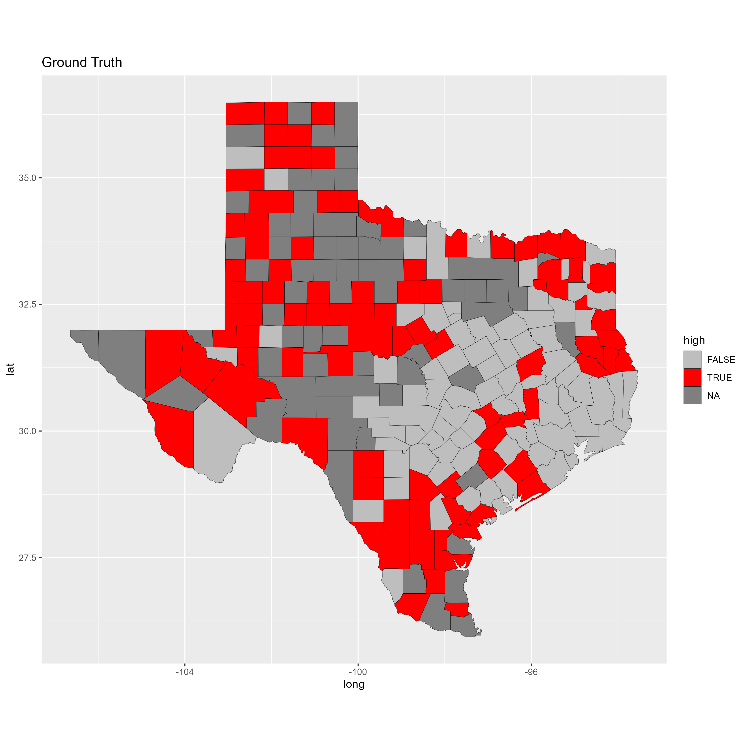
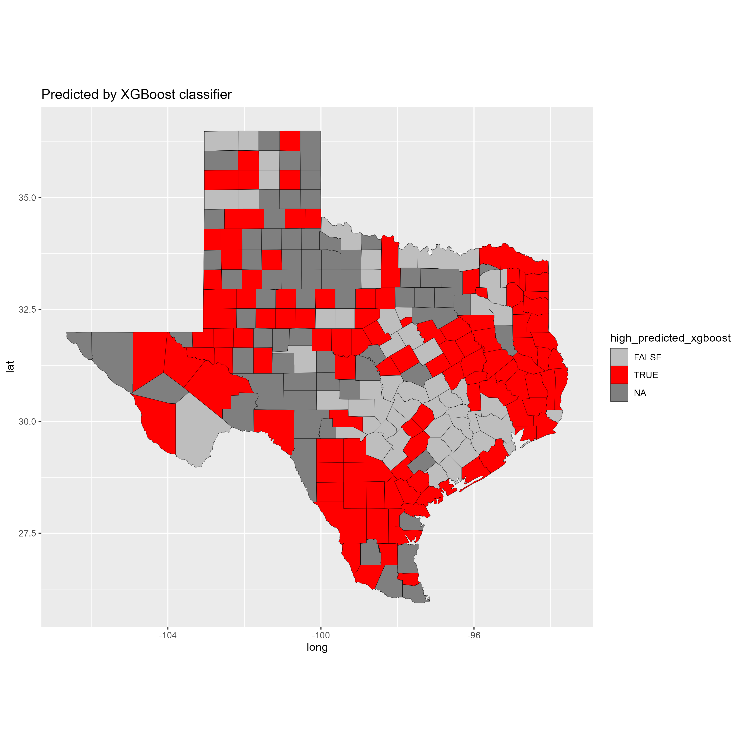
|  |  |  |
| --- | --- | --- |
|  | **Accuracy** | **Kappa** |
| 1 | 0.883 | 0.74 |

Based on the above mentioned setting, the accuracy of model is found to be 0.883 with kappa value of 0.74. Both of these values are far better than the previous two model. Thus, the hyperparameter tuning was not that necessary to achieve better performance.



Variable importance for XGBoost

Now looking at the importance of the variables for predicting the target using this classifier, we can see that the variable “worked\_at\_home” again plays the most significant role in predicting the target. However, the variable “commute” is still a significant variable for predicting the target like the previous two models. Interestingly, there are 6 variables that are not important at all to make prediction of the target. However, the variables that were found to be not significant for the prediction by the previous two models are still part of these 6 variables.

Prediction by XGBoost classifier

Shown above is the comparison between ground truth and prediction made by the model using heat map. From the visual observation, it can be noted that this model seems to have the highest prediction performance than the previous two classifiers. Let’s evaluate that observation through some meaningful metrics.

Confusion matrix + Performance metrics for XGBoost classifier

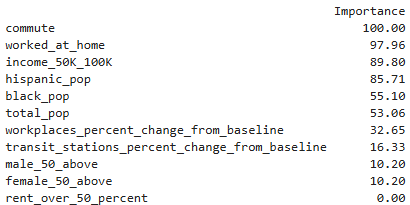
Looking at the confusion matrix, we can see the model has a true negative value of 38 which indicates the number of cases corrected predicted as FALSE (i.e. not high rate). So, it correctly identified 43.18% of the actual FALSE cases. It has a true positive value of 71 which indicates the number of cases correctly predicted as TRUE (high rate). So, it correctly identified 76.37 % of the actual TRUE cases. Thus, this classifier has an advantage of high performance in predicting high death rate cases accurately. Compared to other values, it has comparatively lower false negative value (i.e. 22) which means it is not so bad at inaccurately predicting cases not being high while they actually are. This is another advantage of using this classifier in a pandemic related analysis like this. Overall, the accuracy of the modal is 60.22% (i.e. correctly predicted only 60.22% of the cases correctly). There is 95% confidence that the model true accuracy is between 52.69% and 67.41%. With the kappa value of 0.1969, there is a slight agreement between the predicted and actual classifications indicating that there is considerable room for improvement in the performance. Thus, even though visually this classifier appeared to be better performing classifier, its performance was less than the previous two classifiers.

### 2.4 Support Vector Machines (SVM)

For the training dataset, variables like county\_name, cases\_per\_10000, deaths\_per\_10000, and death\_per\_case were removed because they can’t be used for training a classifier. The classifier training was performed using 10-folds cross validation to introduce randomness in the selection of training samples. The classifier contained 23 individual data points or observations where each data point represents whether the death rate is high or not. There were 11 features selected for predicting the target variable (deaths\_per\_10000 as being high or not). Thus, there were two classes for the target (TRUE or FALSE).

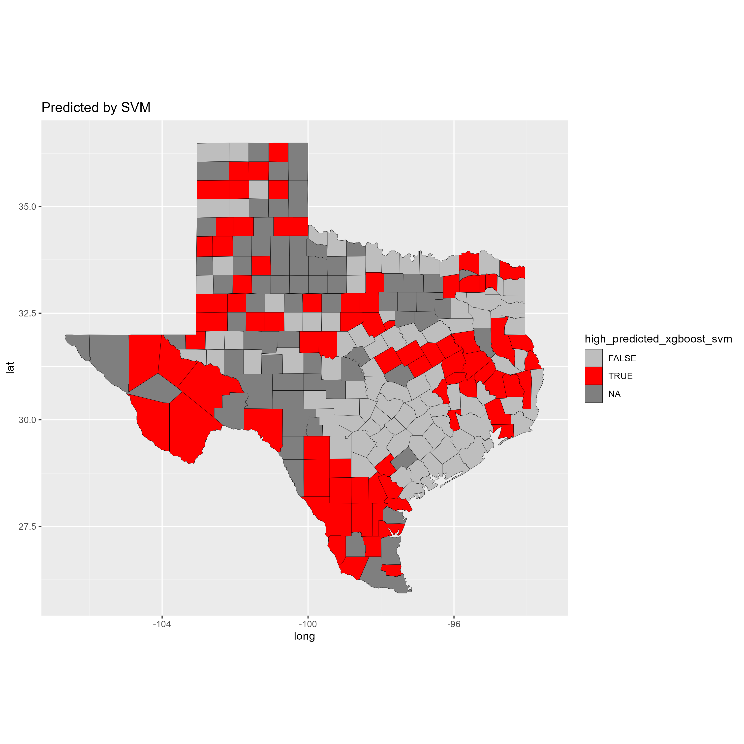
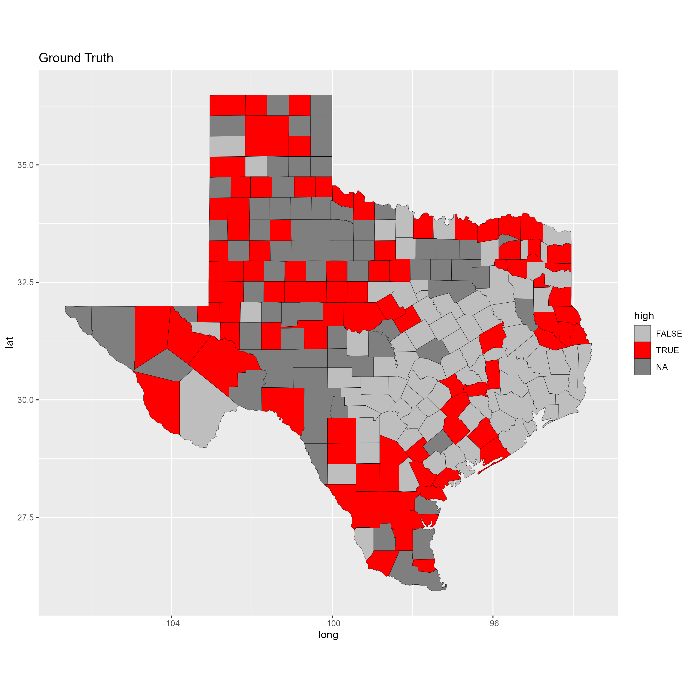
|  |  |  |
| --- | --- | --- |
|  | **Accuracy** | **Kappa** |
| 1 | 0.75 | 0.45 |

Based on the above mentioned setting, the accuracy of model is found to be 0.783 with kappa value of 0.58. Both of these values are worse than the previous three classifiers.



Variable importance for SVM

Now looking at the importance of the variables for predicting the target using this classifier, we can see that the variable “commute” yet again plays the most significant role in predicting the target. Interestingly, the variable related to finance “rent\_over\_50\_percent” was found to have no importance to predict the target. This was different than what was observed in the previous three classifiers.



Prediction by SVM

Shown above is the comparison between ground truth and prediction made by the model using heat map. From the visual observation, it can be noted that this model seems to have lower prediction performance than the previous three classifiers. Let’s evaluate that observation through some meaningful metrics.

Confusion matrix + Performance metrics for SVM

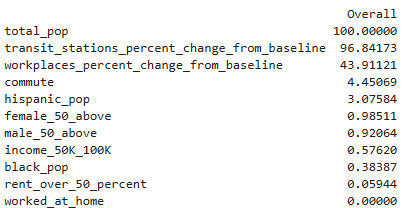
Looking at the confusion matrix, we can see the model has a true negative value of 57 which indicates the number of cases corrected predicted as FALSE (i.e. not high rate). So, it correctly identified 64.77% of the actual FALSE cases. It has a true positive value of 51 which indicates the number of cases correctly predicted as TRUE (high rate). So, it correctly identified 54.84 % of the actual TRUE cases. This means that the classifier is less reliable in predicting high death rate cases which is different than the previous three classifiers. Overall, the accuracy of the modal is 59.67% (i.e. correctly predicted only 56.67% of the cases correctly). There is 95% confidence that the model true accuracy is between 52.14% and 66.88%. With the kappa value of 0.1955, there is a slight agreement between the predicted and actual classifications indicating that there is considerable room for improvement in the performance. Thus, these metrics do support the observation made on the heat map. Although, this classifier does not have the same advantages as the previous three classifiers, it does have a huge advantage while handling non-linear decision boundaries which could be how the dataset is spread. Although, it might not be apparent but SVMs are very memory efficient as they use a subset of training points (support vectors) in the decision function.

### 2.5 Artificial Neural Network (ANN)

For the training dataset, variables like county\_name, cases\_per\_10000, deaths\_per\_10000, and death\_per\_case were removed because they can’t be used for training a classifier. The classifier training was performed using 10-folds cross validation to introduce randomness in the selection of training samples. The classifier contained 23 individual data points or observations where each data point represents whether the death rate is high or not. There were 11 features selected for predicting the target variable (deaths\_per\_10000 as being high or not). Thus, there were two classes for the target (TRUE or FALSE).

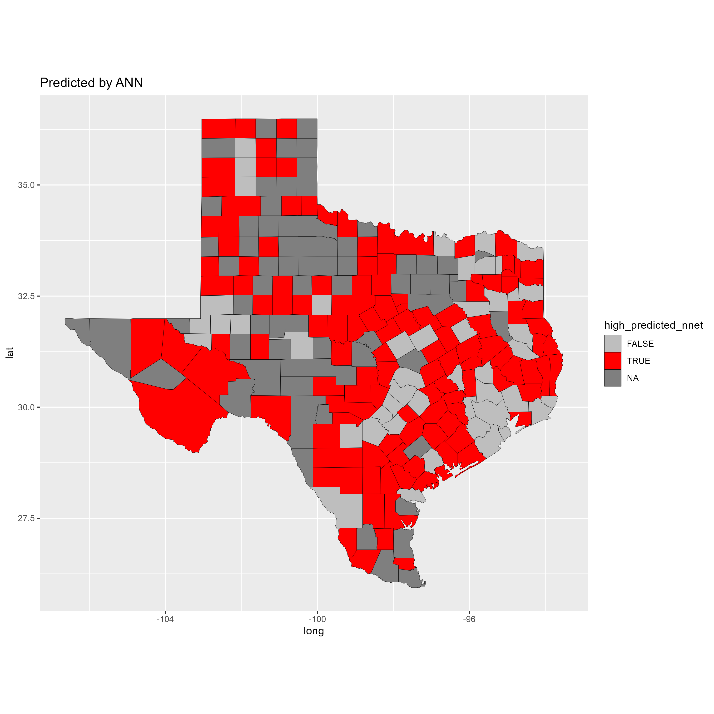
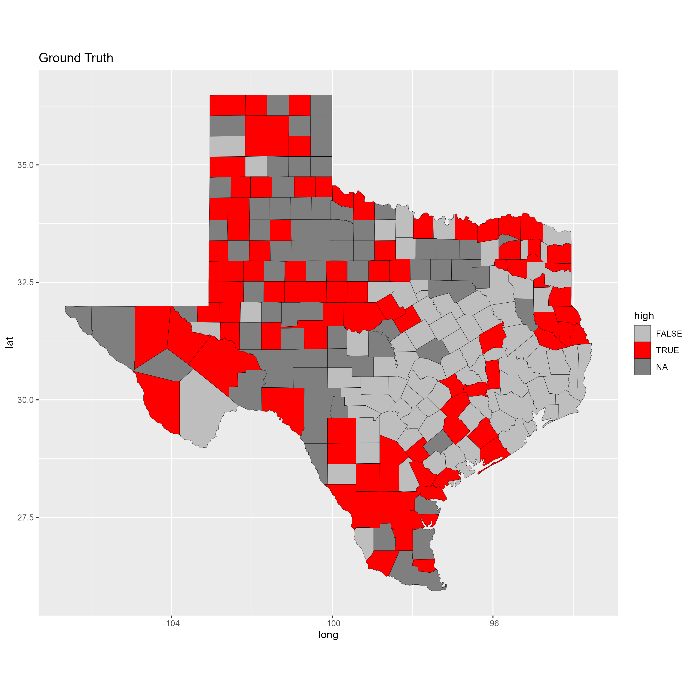
|  |  |  |  |
| --- | --- | --- | --- |
| **size** | **decay** | **Accuracy** | **Kappa** |
| 1 | 0.00E+00 | 0.483333 | 0 |
| 1 | 1.00E-04 | 0.516667 | 0 |
| 1 | 1.00E-03 | 0.55 | 0.1 |
| 1 | 1.00E-02 | 0.716667 | 0.44 |
| 1 | 1.00E-01 | 0.716667 | 0.4 |
| 3 | 0.00E+00 | 0.516667 | 0 |
| 3 | 1.00E-04 | 0.541667 | 0.05 |
| 3 | 1.00E-03 | 0.691667 | 0.35 |
| 3 | 1.00E-02 | 0.675 | 0.35 |
| 3 | 1.00E-01 | 0.75 | 0.5 |
| 5 | 0.00E+00 | 0.516667 | 0 |
| 5 | 1.00E-04 | 0.508333 | 0.05 |
| 5 | 1.00E-03 | 0.7 | 0.4 |
| 5 | 1.00E-02 | 0.725 | 0.45 |
| **5** | **1.00E-01** | **0.825** | **0.65** |
| 7 | 0.00E+00 | 0.483333 | 0 |
| 7 | 1.00E-04 | 0.675 | 0.35 |
| 7 | 1.00E-03 | 0.725 | 0.45 |
| 7 | 1.00E-02 | 0.591667 | 0.19 |
| 7 | 1.00E-01 | 0.75 | 0.5 |
| 9 | 0.00E+00 | 0.516667 | 0 |
| 9 | 1.00E-04 | 0.725 | 0.45 |
| 9 | 1.00E-03 | 0.775 | 0.55 |
| 9 | 1.00E-02 | 0.675 | 0.35 |
| 9 | 1.00E-01 | 0.725 | 0.45 |

Based on the above mentioned setting, the highest accuracy of model is found to be 0.825 with kappa value of 0.65 with 5 neurons in the hidden layer of the neural network. At this configuration, the decay is 0.001 which is a regularization parameter to help prevent overfitting.



Variable importance for ANN

Now looking at the importance of the variables for predicting the target using this classifier, we can see that the variable “transit\_stations\_percent\_change\_from\_baseline ” plays the most significant role in predicting the target. This observation is different than what was observed in the previous four classifies. The variable “worked\_at\_home” was found to have no importance to predict the target. This was also different than what was observed in the previous three classifiers.



Prediction by SVM

Shown above is the comparison between ground truth and prediction made by the model using heat map. From the visual observation, it can be noted that this model seems to have low accuracy as it predicts more counties as “high” than the actual counties. Let’s evaluate that observation through some meaningful metrics.



Confusion matrix + Performance metrics for ANN

Looking at the confusion matrix, we can see the model has a true negative value of 32 which indicates the number of cases corrected predicted as FALSE (i.e. not high rate). So, it correctly identified only 36.36% of the actual FALSE cases. It has a true positive value of 78 which indicates the number of cases correctly predicted as TRUE (high rate). So, it correctly identified 83.87 % of the actual TRUE cases. Thus, this classifier has an advantage of high performance in predicting high death rate cases accurately. The false negative value of 15 does support the observation earlier that the model predicts more counties as high than they actually are. This could be another advantage of using such classifier in cases like COVID-19 pandemic. Overall, the accuracy of the modal is 60.77% (i.e. correctly predicted only 60.77% of the cases correctly). There is 95% confidence that the model true accuracy is between 53.26% and 67.93%. With the kappa value of 0.5138, there is a moderate agreement between the predicted and actual classifications indicating that there is still some room for improvement in the performance even though it has the highest kappa value amongst all the classifiers.



Comparison on highest/lowest importance

Based on the above overall observation on the highest and least important features for each of the classifiers, it can be said that the feature “commute” could be the one that has a high performance in predicting the target variable “high” with high confidence. Similarly, it could be said that the feature “rent\_over\_50\_percent” could be the feature with the lowest importance in making the prediction since it is shown as such by KNN, XGBoost, and SVM classifiers. However, let’s compare the classifiers and further understand this analysis.



Accuracy comparison between classifiers

* KNN classifier shows a high mean and median accuracy, indicating that it performs well on average and often achieves perfect accuracy (as indicated by the Q3 and Max values).
* SVM has the lowest minimum accuracy among the classifiers but maintains a decent mean and median accuracy. The first quartile is also relatively low, indicating that in some cases, SVM's performance may vary more than other classifiers.
* Random Forest performs very well with a high mean and median accuracy. It performs better than KNN. The first quartile is also relatively high, indicating consistent performance.
* XGBoost shows a wide range of accuracy scores, with a minimum of 0.00 but also a high maximum and median accuracy. This suggests that XGBoost can be very powerful but may also fail completely in some cases.
* ANN has a high mean and median accuracy, indicating strong performance overall. The first quartile is slightly lower than KNN and Random Forest but still indicates good consistency.

Based on the accuracy alone, it can be said that Random Forest and KNN classifiers outperform all of the classifiers. They are also equally, if not better, consistent as theirs accuracies swing between 0.5 to 1.



Kappa comparison between classifiers

* KNN shows high median and mean Kappa values, indicating it generally has good agreement between observed and predicted classifications. The first quartile is also high, suggesting consistent performance.
* SVM has a wide range of Kappa values, with the minimum being -0.5, indicating poor performance in some cases. The median is lower than KNN and Random Forest, showing that SVM has greater variability in performance.
* Random Forest demonstrates high median and mean Kappa values, similar to KNN, indicating strong performance and good agreement between observed and predicted classifications.
* XGBoost shows a wide range of Kappa values, with a minimum of -1.0, suggesting very poor performance in some cases. However, it can also achieve perfect agreement (Kappa = 1.0). The median and mean are lower than KNN and Random Forest, indicating more variability.
* ANN has high median and mean Kappa values, showing good performance and agreement between observed and predicted classifications. The first quartile is lower than KNN and Random Forest but still indicates relatively consistent performance.

Overall, KNN and Random Forest classifiers yet again appear to be the most reliable classifiers in terms of Kappa. So, it can be concluded that these are the best performing classifiers for the chosen features and target variable. Based on these observations, it can be said that the features “commute” is indeed the feature that has the highest significance in terms of correctly predicting the counties that have high deaths per 10000 (i.e. above 16).

# Evaluation

The results obtained from the classifications performed on the dataset can be very important for a state agency like DSHS. Such agency could utilize the findings from this analysis to allocate resources and develop targeted interventions to reduce mortality rates in high rick counties in Texas. In the event of future outbreak of virus such as COVID-19, the agency could use classifiers like KNN and Random Forest to accurately predict the counties that could have high death rates. Then, they could implement necessary state policies and provide infrastructures to control the deaths in such critical counties.

* Since commute was found to be the most important predictor, DSHS could develop educational programs focused on safe commuting practices.
* The agency could identify the communities that greatly depend on public transportations in critical counties so that the necessary amenities and healthcare services are prioritized accordingly.
* Community-based health programs could be initiated targeting the high-risk communities or counties thereby providing preventive healthcare services
* DSHS also can work with city and county governments to develop strategies of providing affordable housings near workplaces to minimize long commutes
* DSHS can work with the Texas Transportation Authorities to initiate community-based programs or campaigns to encourage safe commuting practices.

All of these proactive activities could be very helpful in prevent high deaths during pandemic like COVID-19 in future especially in high-risk counties. The classification models are valuable as they provide data-driven approach to identifying high-risk counties and guiding targeted interventions. So, the observations made on these classifications models are very important for the agency such as DSHS.

Assessing the model's value involves monitoring reductions in mortality rates, evaluating the effectiveness and efficiency of implemented programs, gathering stakeholder feedback, and analyzing the economic impact. For example, the mortality rates in high-risk counties could be compared to their previous values against the ones after implementing the recommended interventions. Use statistical methods to determine if the observed reduction in mortality rates is statistically significant, thereby affirming that the change is not due to random variation. Another way to assess the model’s value is to conduct a cost-benefit analysis to assess the economic impact of the interventions. Compare the costs of implementing the recommendations with the financial savings from reduced healthcare expenditures and improved productivity. Then, calculate the return on investment for the interventions to determine their financial viability and long-term benefits. Overall, there are several methods that could be used to assess the model’s effectiveness and their contributions in reducing death rates during similar pandemic.

# Deployment

All the results obtained from the experimenting with various classification models, and the recommendations made on previous section are very practical as they could be used as intended during a pandemic like COVID-19. DSHS could consider all of the recommendations made on the evaluation section, and work with necessary public or private entities to put those recommendations into actions. For example, DSHS could start coordinating with schools in developing plans of teaching students the importance of safe commuting practices and train them on such practices. It takes a lot of time to properly distribute health care related services and infrastructures throughout the state. Thus, DSHS could use the results from this report to convince the responsible authorities of the counties to build strategies of investing such facilities to high-risk counties.

The observations made on these models should remain unchanged as long as new data on the pandemic are not discovered or the current data gets corrected. There could be more data related to the pandemic that are still being collected or assessed. Once such data is collected, the current models and their corresponding results would need to be updated.

Overall, all of these actions are very practical and could play an important role in controlling mortality rates during similar pandemic. Such models could be deployed to various agencies such as DSHS as they could definitely get many benefits from these models.

# List of References

Google. (n.d.). Global mobility report. Retrieved 11/20/2024, from <https://www.google.com/covid19/mobility/index.html>

USAFacts. (n.d.). COVID-19 case and death counts by state and county. In USAFacts Public Data - COVID-19 US Cases.

Retrieved from

[https://console.cloud.google.com/bigquery?p=bigquery-public](USAFacts.%20(n.d.).%20COVID-19%20case%20and%20death%20counts%20by%20state%20and%20county.%20In%20USAFacts%20Public%20Data%20-%20COVID-19%20US%20Cases.)

[data&d=covid19\_usafacts&page=dataset&project=crucial-cycling-338005&ws=!1m4!1m3!3m2!1sbigquery-public](USAFacts.%20(n.d.).%20COVID-19%20case%20and%20death%20counts%20by%20state%20and%20county.%20In%20USAFacts%20Public%20Data%20-%20COVID-19%20US%20Cases.)

[data!2scovid19\_usafacts](USAFacts.%20(n.d.).%20COVID-19%20case%20and%20death%20counts%20by%20state%20and%20county.%20In%20USAFacts%20Public%20Data%20-%20COVID-19%20US%20Cases.)

# Appendix

Jeevan Rai:

* Random Forest classifier, K-Nearest Neighbor classifier
* Data setup, feature selection, feature preprocessing
* Schematics for performing models classification, comparison and overall project
* Write project report
* Develop objectives of the project
* Exceptional work: Artificial Neural Network (See ANN sections for more details on the model)

Abhilash Narayanan:

* Gradient Boosted Decision Trees
* Write project report
* Develop objectives of the project
* Exceptional work: Support Vector Machines (See XGBoost sections for more details on the model)