ABSTRACT

Coronavirus pandemic has impacted the normal life in the US since the beginning of year 2020. Recently, experts discussed the possibility of a fourth wave of Covid-19 in the US. The use of classification algorithms can help medical experts and government officials to predict the potential response in the US in case of emerging a new covid-19 outbreak. In this report, classification algorithms are used to identify the bad and good counties in case of the fourth wave of coronavirus hits Texas. First, spread rate is used to classify the data into two possible outcomes (bad/good). The most important features are selected for training and testing. The five most important variable are family households, unemployed population, population, median age, and commuters by public transportations. This is anticipated since Covid-19 can spread fast when large number of people are closed to each other and endanger the life of older people. Different classifiers are performed and compared with each other. The results show that the accuracy and kappa statistic in rule-based classifier are higher than those in decision trees and k-nearest neighbors classifiers. Additionally, the data is classified based on death per case instead of spread rate. The most important feature in this case is median income which is expected since high income people might have better health resources and able to commute by cars unlike low-income people. Similarly, the rule-based classifier has the highest accuracy and kappa statistic.

Content

1 Introduction

In data mining, classification is an expanding area that plays an essential role in mining techniques. Over the years, classification techniques, classification algorithms, and rule-based classification have become trending topics [1].

Classification is a technique that assigns items in a collection to classes for prediction and analysis. The relationship between the predictor variables and the output is found using a classification algorithm summarized by the model in the building process. we can apply this relationship to new data to predict its class [2]. This technique is developed to work for data sets that are complicated and large [3].

The COVID-19 pandemic is an unprecedented crisis that expanded across the world quickly. Predicting the incidence of it is essential to helping government make critical decisions about the disease [4]. There has been a significant increase in data traffic related to COVID-19. Many researchers around the world have produced an enormous collection of literature since the beginning of the pandemic [5]. Classification is a technique for analyzing this information to extract knowledge and provide meaningful insights.

Our focus is on Texas counties to find the pattern in data to categorize these counties into “bad” and “safe” counties concerning two different aspects; “spread rate” and “case per death”. This pattern helps us to predict the spread rate and mortality rate for new data using the same features.

2 Data Preparation

2.1 Data Description

The data file contains four different data sets including US Covid-19 cases plus census data, Texas Covid-19 cases, Google global mobility report, and healthcare resources. US Covid-19 cases plus census data provides total US COVID-19 case and death counts by state and county until 2021-03-14, and census data provides useful information about the population. Texas Covid-19 cases data set provides COVID-19 case and death counts by county in Texas from 2020-01-22 to 2021-03-19. It also provides changes and trends by county. Google global mobility report aims to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. For each category in a region, reports show the changes which compares mobility for the report date to the baseline day. Lastly, the hospital capacity and medical staff in each county in Texas are considered in the data set including total hospital beds, total physician and dentist, and registered nurses which is obtained from [7].

Table 1 shows the description of the selected features that are used in this project. For this project, we focus on Covid-19 cases in Texas counties. The variable population density was created since we used density instead of total population. Moreover, the average number of confirmed cases or deaths increased during 03-13-2021 to 03-19-2021 in each county were created using Texas Covid-19 cases data. The distancing response variable that is used in this report is the overall average changes in mobility from the baseline in each county. “Texas COVID-19 Case Count Data by County” data set provides the daily data about the number of covid-19 cases collected over one year in each county in Texas. The date for the spread rate is found using this data set.

Table 1: The description of selected features used for classification

Feature	Scale	Description
pop_density	Ratio	The population density of the county in people per square mile
family_households	Ratio	Total number of households contain families
median_age	Ratio	Median age
median_income	Ratio	Median income
unemployed_pop	Ratio	Population that are classified as unemployed
commuters_by_public_transportation	Ratio	Number of commuters by public transportation
physicians_and_dentists	Ratio	Total physicians and dentists in each county
registered_nurses	Ratio	Total registered nurses in each county
hospital_beds	Ratio	Total physicians and dentists in each county
cases_per_1000	Ratio	Total confirmed cases per 1000 people in the total population
deaths_per_1000	Ratio	Total deaths per 1000 people in the total population
death_per_case	Ratio	Percentage of number of deaths per total confirmed cases
distancing_response	Ratio	The overall average changes in mobility from the baseline in each county
avg_confirmed	Ratio	The average number of confirmed cases increased during 03-13-2021 to 03-19-2021
avg_deaths	Ratio	The average number of deaths increased during 03-13-2021 to 03-19-2021
Spread rate	Ratio	The slope of the line derived from daily number of cases over one year

2.2 Define Classes

2.2.1 Class based on spread rate

In Modeling 1, the feature related to spread rate is used to classify the counties in Texas into bad and safe counties. After cleaning and normalize the data, the mean for the spread rate is -0.2321 which is used the partition where the counties above the mean are classified as high spread of Covid-19 area (True) while the counties that are lower than the mean is classified as safe area (False). The data is balanced where 124 counties are classified as bad counties (spread rate is very high) while the other 123 counties in Texas are safe counties. The map for counties in Texas with the class of bad area (True and False) is shown in Figure 1.

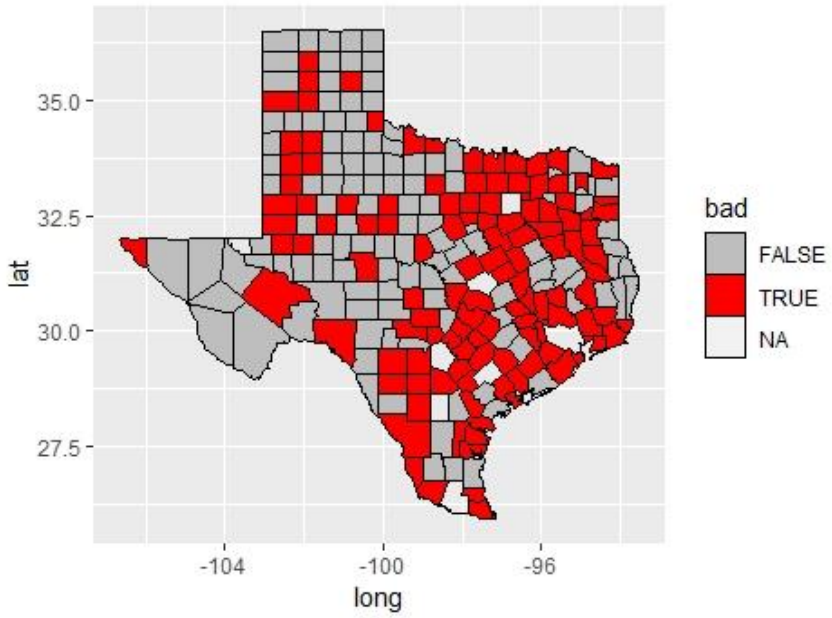

Fig. 1: Texas county map showing the counties classified based on spread rate

2.2.2 Class based on death rate

In Modeling 2, the feature related to death rate (i.e., death per case) is used to classify the counties in Texas into bad and safe counties. After cleaning and normalize the data, the mean for the death rate is 0.0084 which is used the partition where the counties above the mean are classified as bad area (True) while the counties that are lower than the mean is classified as safe area (False). Using the mean of the death rate, the data is balanced where 121 counties are

classified as bad counties while the other 126 counties in Texas are safe counties. It is worth noting that there are 254 counties in Texas; however, there are 7 counties are removed during the cleaning data procedure described in Section 2.4. This class is used in the modeling 1 (Section 4). The map for counties in Texas with the class of bad area (True and False) is shown in Figure 2.

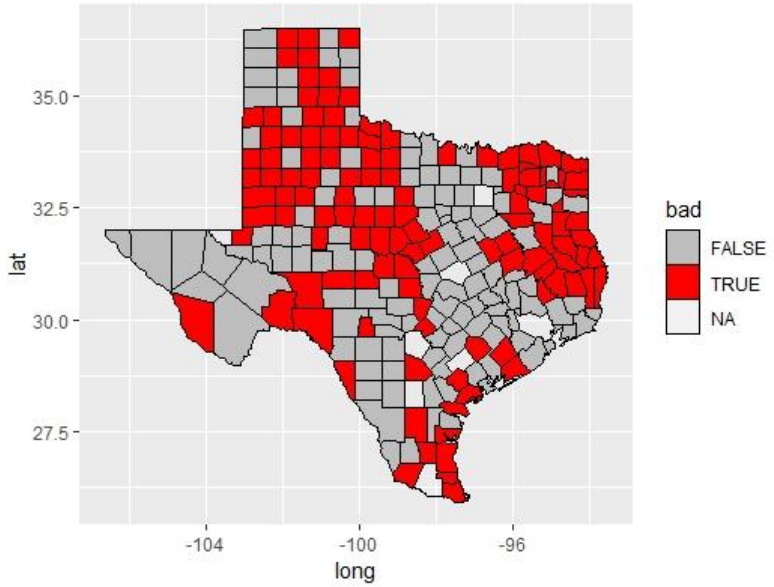

Fig. 2: Texas county map showing the counties classified based on death rate

2.3 Predictive Features

2.3.1 Predictive features for the class based on spread rate

The importance of each features used in the class based on spread rate is shown in Table 2. As shown in the table, the highest importance features are spread rate and county name. This is expected because the spread rate is used to classify the data. For modeling 1 (Section 3), the variables that are related to Covid-19 are removed, i.e., spread rate, death per case, county name, deaths per 1000, case per 1000, and average confirmed cases are removed. The unrelated Covid-19 features are selected as the predictive features.

Table 2: The importance of the features used for predicting the class based on spread rate

Feature	Importance
Spread rate	1
County name	1.0
family_households	0.787
unemployed_pop	0.764
pop_density	0.720
registered_nurses	0.718
hospital_beds	0.649
avg_confirmed	0.609
commuters_by_public_transportation	0.590
avg_deaths	0.517
distancing_response	0.452
median_age	0.402
death_per_case	0.313
cases_per_1000	0.296

Figure 3 show the importance of the predictive features for predict the class based on spread rate in the modeling 1 (Section 3). As it can be seen from this figure, the family households, unemployed population and population density have the highest importance which are expected since the more family members and more population in counties will contribute to the increase of the spread of Covid-19. Median age and commuters by public transportations are the fourth and fifth importance features since the spread of Covid-19 is more likely to affect older people and people who are commuting by public transportation.

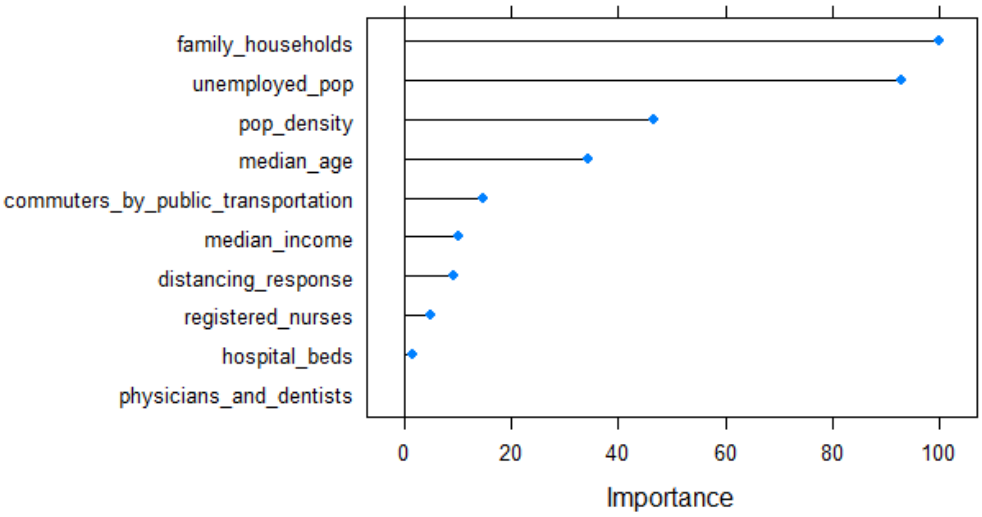

Fig. 3: The importance of the predictive features for the class based on spread rate

2.3.2 Predictive features for the class based on death rate

The importance of each features used in the class based on death rate is shown in Table 3. As shown in the table, the highest importance features are death per case and county name. This is expected because the death per case is used to classify the data. For modeling 2 (Section 4), the unrelated Covid-19 features are selected as the predictive features. Therefore, the features that are removed are: death per case, county name, deaths per 1000, spread rate, case per 1000, and average confirmed cases.

Table 3: The importance of the features used for predicting the class based on death rate

Feature	Importance
death_per_case	1.0
county_name	1.0
deaths_per_1000	0.630
median_income	0.411
spread_rate	0.373
cases_per_1000	0.342

avg_confirmed	0.324
commuters_by_public_transportation	0.321
pop_density	0.314
median_age	0.299
family_households	0.295
unemployed_pop	0.275
registered_nurses	0.249

Figure 4 show the importance of the predictive features for predict the class based on death rate in the modeling 2 (Section 4). As it can be seen from this figure, the median income has the highest importance which is expected since people with more income might have the privileges of more health resources, better work environment (such as working from home) and commuting by cars unlike the people with lower income. The second importance features are family households, population density, and median age which are also expected since counties with more population and older people are more susceptible to have a higher death rate.

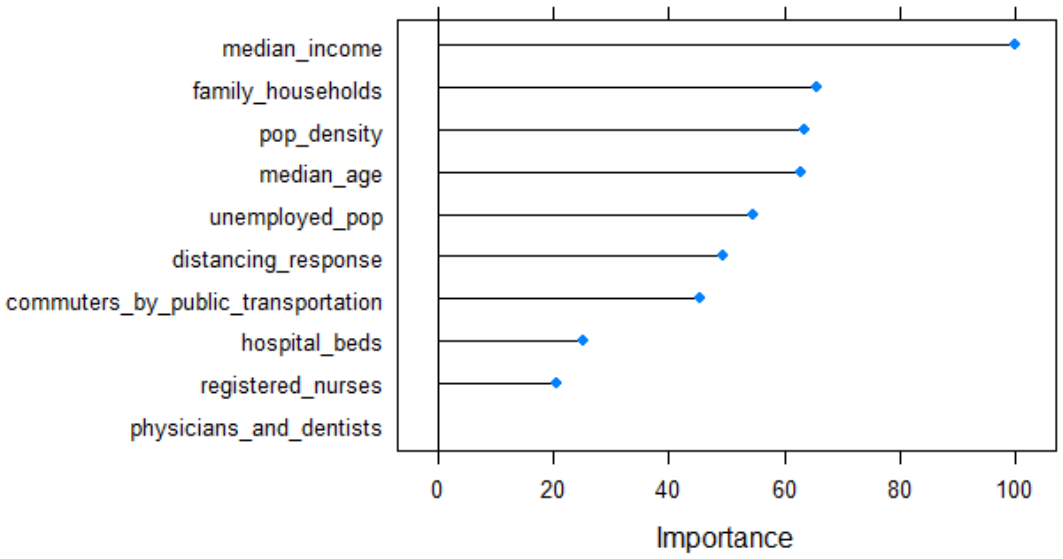

Fig. 4: The importance of the predictive features for the class based on death rate

2.4 Dealing with Missing Values, Outlier Removal and Mean Imputation

The first step is to remove the outliers form the data. By plotting the lof figure (Figure 5), we can set the lof to 2.2 and remove the outliers.

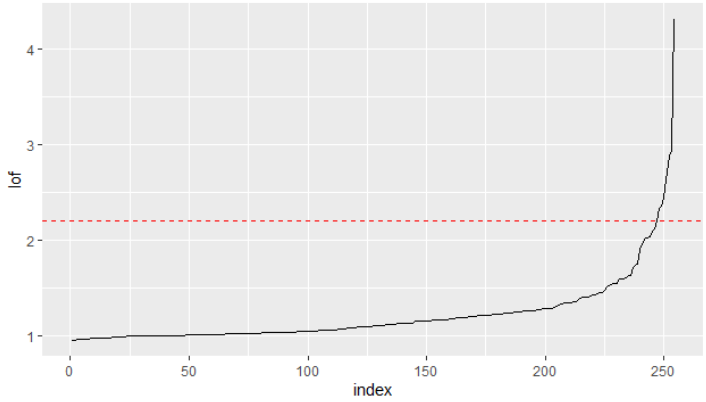

Fig. 5: lof for outlier removal

Form 254 counties, 7 counties are removed. Bell County, Bexar County, Dallas County, Harris County, Hidalgo County, Loving County, McMullen County are the counties that are recognized as outliers and should be removed from the dataset. These seven counties are shown in Figure 6. From these figures, we see that when we plot the outliers for in three dimensions, for this specific combination, some points are not outliers, and they are outlier with respect to other features.

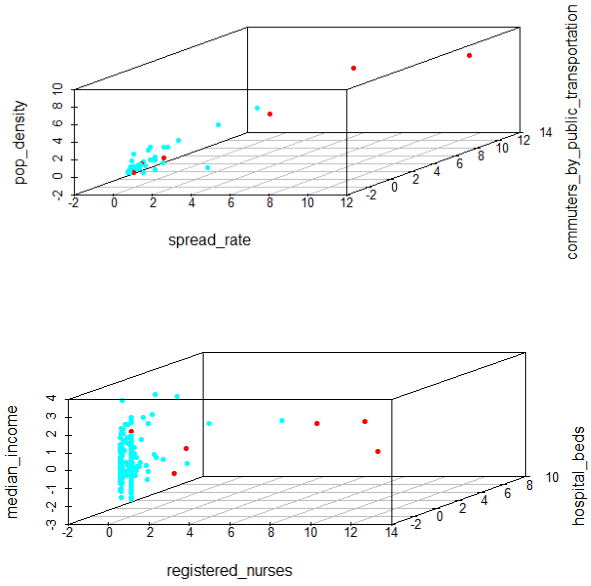

Fig. 6: Outliers with respect to three features

3 Modeling 1

In this section, the Texas counties are classified into bad/ good counties based on the spread rate. As mentioned earlier, there are 124 counties in Texas that are categorized as bad counties while 123 counties are classified as safe area. In this section, three classification techniques are utilized to predict the class of counties based on spread rate. The data was split into 90% training and 10% testing.

3.1 Classification 1: Decision Trees

Using the decision trees to predict the class for the counties in Texas and compare it with the ground truth. As show in the Figure 7 below, the left figure shows the ground truth while the prediction is presented in the left figure. As it can be seen, one county is falsely classified as bad counties, and two counties are falsely classified as safe counties. The accuracy for training set is 0.897, and the Kappa is 0.794. The corresponding confusion matrix is shown in Table 4.

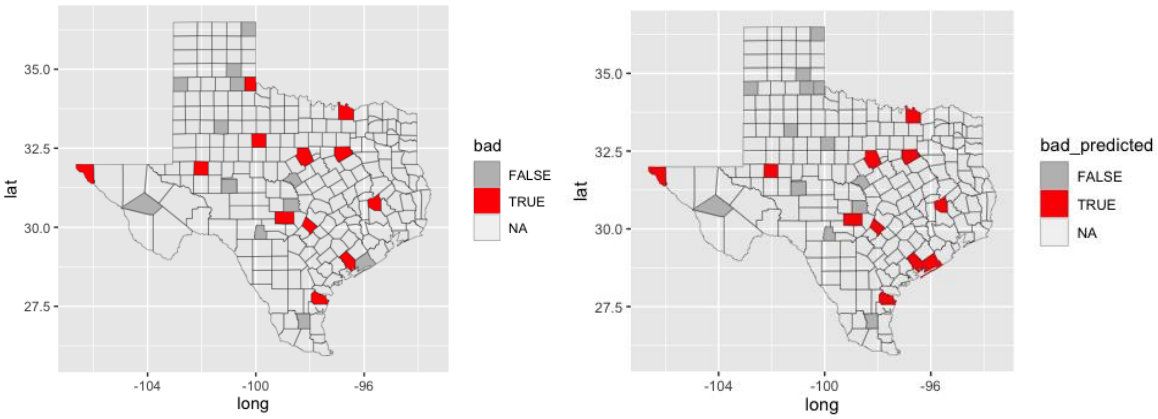

Fig. 7: Ground truth and predict model.

Table 4: Confusion Matrix for training set based on decision tree

Prediction	False	True
False	99	13
True	14	97

Advantage and disadvantage [7]: decision trees classifier is simple and can be visualize; however, it can be unstable due to the variation in the data set.

3.2 Classification 2: K-Nearest Neighbors

The final value used for the model was k = 9. The accuracy for the training set is 0.816, and the Kappa statistics is 0.632. The results of KNN and the confusion matrix for training set is shown in Table 5 and Table 6, respectively.

Table 5: Results of KNN

K	Accuracy	Kappa
5	0.821	0.643
7	0.799	0.600
9	0.826	0.653
11	0.812	0.625
13	0.786	0.572

Table 6: Confusion Matrix for training set based on KNN

Prediction	False	True
False	91	20
True	21	91

Advantage and disadvantage [7]: KNN classifier is simple and not affected by noise in the data set; however, calculating the distance between observations can be computationally high.

3.3 Classification 3: Rule-based classifier (PART)

The final values used for the model were threshold = 0.5 and pruned = no. The number of rules is 6. The results of Rule-based classifier on training set and confusion matrix are shown in Table 7 and Table 8, respectively. The accuracy for the training set is 0.919, and Kappa is 0.839.

Table 7: Results of Rule-based classifier

threshold	pruned	Accuracy	Kappa
0.01	yes	0.915	0.830
0.01	No	0.929	0.858
0.1325	yes	0.920	0.839
0.1325	No	0.929	0.858
0.255	yes	0.920	0.839
0.255	No	0.929	0.858
0.3775	yes	0.920	0.839
0.3775	No	0.929	0.858
0.5	yes	0.920	0.839
0.5	No	0.929	0.858

Table 8: Confusion Matrix for training set based on Rule-based classifier

Prediction	False	True
False	97	3
True	15	108

3.4 Compare Models

Accuracy and Kappa analysis:

The summary statistics are shown in Table 9. Then we performed inference about differences between models, which is presented in Table 10. We created fixed sampling scheme (10-folds), so we compared the different models using exactly the same folds. For each metric, all pair-wise differences are computed and tested to assess if the difference is equal to zero. By default, Bonferroni correction for multiple comparison is used. Differences are shown in the upper triangle and p-values are in the lower triangle. For a better classifier, the p-value should be less than 0.01. We also created a plot of the model evaluation results (Figure 8) and compare the spread and the mean accuracy of each model. In this case, rule-based classifier seems better, but classifiers do not perform statistically differently since all the p-values are greater than 0.05. Therefore, we choose the rule-based classifier model as our best model. The confusion matrix for test data based on rule-based classifier model is presented in Table 11. The accuracy for the test data is 0.833.

Table 9: Summary Statistics of models

	Mean Accuracy/Mean Kappa
Decision Trees	0.897/0.794
KNN	0.826/0.653
Rule-based classifier	0.929/0.858

Table 10: Differences between models

Accuracy:

	Decision Tree	KNN	Rule-based classifier
Decision Tree		0.071	-0.032
KNN	0.256		-0.103
Rule-based classifier	0.681	0.009	

Kappa:

	Decision Tree	KNN	Rule-based classifier
Decision Tree		0.141	-0.064
KNN	0.257		-0.205
Rule-based classifier	0.687	0.009	

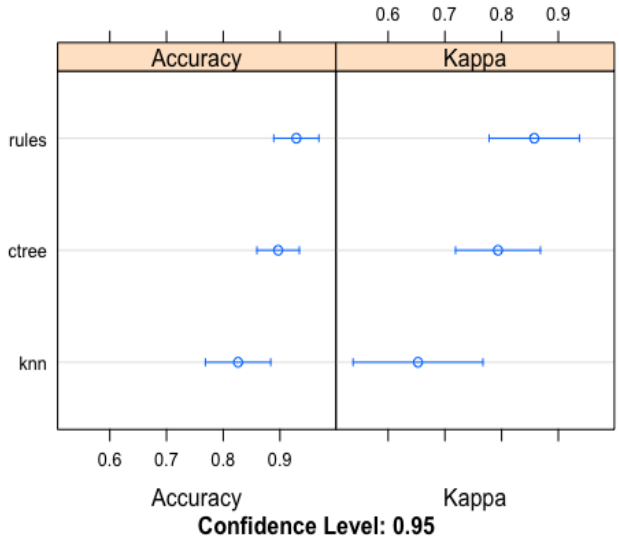

Fig. 8: Model evaluation results

Table 11: Confusion Matrix for test data based on rule-based classifier model

Prediction	False	True
False	9	1
True	3	11

Regression Analysis:

Based on the spread rate, the first classification model has provided good results, and this algorithm can be applied to other new data to predict the level of the spread rate. If we look at the data using the regression model, we see that Multiple R-squared: is 0.946 and Adjusted R-squared 0.9431, which are perfect results. The family households, unemployed population, commuters by public transportation, physicians and dentists, registered nurses, and hospital bed are the significant features. The regression model results confirm the classification ability to get good results based on this set of features.

4 Modeling 2

In this section, the Texas counties are classified into bad/ good counties based on the death rate. As mentioned earlier, there are 121 counties in Texas that are categorized as bad counties while 126 counties are classified as good area. In this section, three classification techniques are utilized to predict the class of counties based on death rate. The data was split into 90% training and 10% testing.

4.1 Classification 1: Decision Trees

Using the decision trees to predict the class for the counties in Texas and compare it with the ground truth. As show in the Figure 9, the left figure shows the ground truth while the prediction is presented in the left figure. As it can be seen, 7 counties are falsely classified as bad counties due to the low accuracy of the model. The accuracy for training set is 0.555, and the Kappa is 0.111. The corresponding confusion matrix is shown in Table 12.

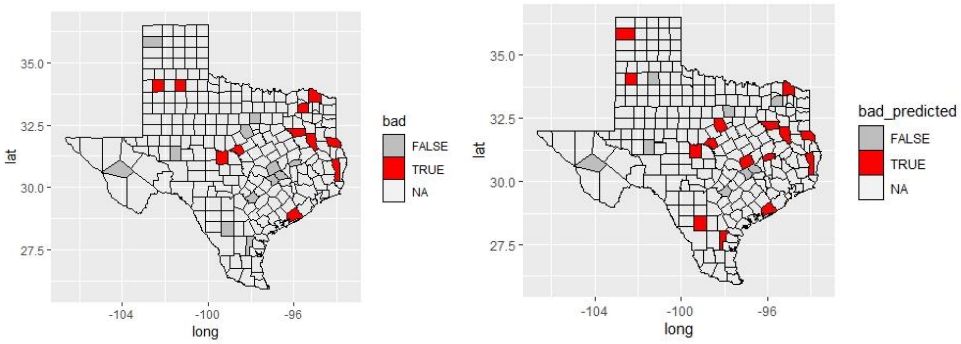

Fig. 9: Ground truth and predict model

Table 12: Confusion Matrix for training set based on decision tree

Prediction	False	True
False	64	50
True	52	57

4.2 Classification 2: K-Nearest Neighbors

The final value used for the model was k = 5. The accuracy for the training set is 0.708, and the Kappa statistics is 0.419. The results of KNN and the confusion matrix for training set is shown in Table 13 and Table 14, respectively.

Table 13: Results of KNN

K	Accuracy	Kappa
5	0.691	0.386
7	0.656	0.316
9	0.664	0.330
11	0.669	0.343
13	0.637	0.278

Table 14: Confusion Matrix for training set based on KNN

Prediction	False	True
False	71	22
True	43	87

4.3 Classification 3: Rule-based classifier (PART)

The final values used for the model were threshold = 0.01 and pruned = yes. The number of rules is 3. The results of Rule-based classifier on training set and confusion matrix are shown in Table 15 and Table 16, respectively. The accuracy for the training set is 0.686, and Kappa is 0.380.

Table 15: Results of Rule-based classifier

threshold	pruned	Accuracy	Kappa
0.01	yes	0.700	0.400
0.01	No	0.668	0.342
0.1325	yes	0.687	0.376
0.1325	No	0.668	0.342
0.255	yes	0.678	0.360
0.255	No	0.668	0.342
0.3775	yes	0.678	0.360
0.3775	No	0.668	0.342
0.5	yes	0.678	0.360
0.5	No	0.668	0.342

Table 16: Confusion Matrix for training set based on Rule-based classifier

Prediction	False	True
False	47	3
True	67	106

4.4 Compare Models

Accuracy and Kappa analysis:

The summary statistics are shown in Table 17. Then we performed inference about differences between models, which is presented in Table 18. In addition, we created a plot of the model evaluation results (Figure 11) and compare the spread and the mean accuracy of each model. In this case, rule-based classifier seems better, and there is some evidence that classifiers perform statistically differently since some p-values are less than 0.05. Therefore, we choose the rule-based classifier model as our best model. The confusion matrix for test data based on rule-based classifier model is presented in Table 19. The accuracy for the test data is 0.708.

Table 17: Summary Statistics of models

	Mean Accuracy/Mean Kappa
Decision Trees	0.555/0.111
KNN	0.691/0.386
Rule-based classifier	0.700/0.400

Table 18: Differences between models

Accuracy:

	Decision Tree	KNN	Rule-based classifier
Decision Tree		-0.136	-0.144
KNN	0.051		-0.008
Rule-based classifier	0.040	1.000	

Kappa:

	Decision Tree	KNN	Rule-based classifier
Decision Tree		-0.275	-0.289
KNN	0.048		-0.014
Rule-based classifier	0.039	1.000	

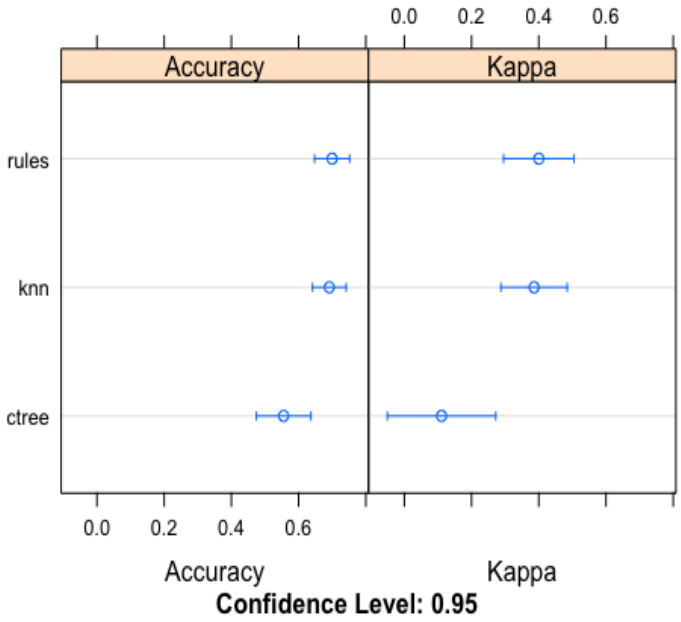

Fig. 10: Model evaluation results

Table 19: Confusion Matrix for test data based on rule-based classifier model

Prediction	**False**	**True**
False	5	0
True	7	12

Regression Analysis:

For the second classification model based on the death rate, using different algorithms, the results for the algorithm performance are not satisfying. To get more insight into the data, we constructed the linear regression model to see the effect of each variable on the output. The results show that only two of the features we use in classification are significant, and we can remove other features from the model. Also, the results show that the Multiple R-squared is 0.2026 and the Adjusted R-squared is 0.165, far from 1, so we can conclude that the features we are using for the classification are not representative of the output.

To see if we can select a better subset of features, we fitted the model using other features and tested the first 50 features of the "COVID-19_cases_plus_census" except the covid related data. Using this subset of features, we see that the Multiple R-squared increases from 0.2026 to 0.3068 and Adjusted R-squared from 0.165 to 0.188, which is still far from 1. Based on the ANOVA analysis and the feature importance, we found the following subset of essential features: median rent, median income, median age, median year structure built, nonfamily households, income per capita, rent_30_to_35_percent, rent_35_to_40_percent, black population, American population are essential features. Therefore, if we fit the classification based on this new feature, the performance will improve, although it is not compelling yet.

5 Evaluation

The classification models presented in this report can help the Texas officials and healthcare providers to predict the safe and high-risk areas based on spread rate and death rate in case of a fourth wave of covid-19 hits in Texas. The proper response will then be implemented based on the class of each county. Moreover, the Covid-19 vaccines can be distributed to the area with high-risk of Covid-19 spread and potential deaths. To utilize the models presented in this report, importance predictive features such as median income, family households, unemployed population, population, median age, and commuters by public transportations should be collected. Subsequently, the rule-base classifier should be used as the classification algorithm since the results show that the highest accuracy is obtained using rule-base classifier.

6 Deployment

Using the classification models in practice, the medical experts and government officials might need collaborate with each other to collect and update the data frequently. Data related to population, family households and other demographic resources might need to be updated yearly. While the data with hospital resources and mobility need to be updated after any responsive actions to combat Covid-19 infections such as additional restrictions and hospital resources. In the counties that are classified as high-risk counties, the possible actions can be delivering government stimulus checks to the people with low income, offering better health insurance especially for older people during the pandemic, and increasing hospital resources such as hospital beds and physicians and nurses on duty in the high-risk area. Moreover, restrictions such as face mask mandates and social distancing might be effective responses to slow down the spread in the high-risk counties.

7 Conclusion

Since early 2020, the US has greatly impacted by the Covid-19 pandemic. Lately, some experts expect and argue over the possibility of a fourth wave of Covid-19 in the US. This project analyzes the potential impacts of the fourth wave of Covid-19 on counties Texas using classification techniques. Two classification models are presented. In the first model, spread rate is used to categorize the counties into good and bad area. Apply feature selection on demographic and hospital resources data, the important features include family households, unemployed population, population, median age, and commuters by public transportations are used as the predictive features. The results show that the demographic features are more important to predict the class of the counties in Texas based on the spread rate than the hospital resources features. In the second model, the counties are classified into good and bad area death rate is used to categorize. The results show that the median income is the most important feature on the death rate. The results in both models show that the accuracy and kappa statistic in rule-based classifier are higher than those in decision trees and k-nearest neighbors classifiers. In case of a fourth wave of covid-19 hits in Texas, medical experts and government officials can utilize the classification models to predict the potential safe and dangerous area in Texas and

determine the proper responses accordingly. The results show that the median income has the highest importance to predict the counties with potential high mortality rate. Therefore, providing stimulus checks to the people in high-risk area might help in reducing the mortality rate. Moreover, the results show that family households, unemployed population and population are the most important features to predict the counties with high spread rate. Enforcing additional restrictions such as face mask mandates and reducing the number of people allowed in gatherings as well as stimulus checks might reduce the spread rate in the counties with potential high spread rate.

References

[1] https://www.quora.com

[2] https://www.jigsawacademy.com/blogs/data-science/classification-in-data-mining/

[3] https://www.lifewire.com/classification-1019653

[4] J. Zhao, M. A. Rodriguez and R. Buyya, "High-Performance Mining of COVID-19 Open Research Datasets for Text Classification and Insights in Cloud Computing Environments," *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)*, 2020, pp. 302-309, doi: 10.1109/UCC48980.2020.00048.

[5] https://www.researchsquare.com/article/rs-21247/v1

[6] https://console.cloud.google.com/marketplace/product/aha-public-data/hospital-capacity?project=smiling-diode-307114&folder=&organizationId=

[7] https://analyticsindiamag.com/7-types-classification-algorithms/