Ability Evaluation of Counties against COVID-19 in United States

Wenjun Han
hwenjun@vt.edu

1 Problem Statement

The COVID-19 pandemic is an ongoing global pandemic of coronavirus caused by severe respiratory syndrome. On 11 Match 2020, the World Health Organization declared the outbreak of COVID-19 as a pandemic. Till now, more than 100 million cases have been confirmed. COVID-19 has caused more nearly 3 million deaths around the world. In the United States, the number of total identified cases is nearly 30 million, and total deaths in U.S. is over 500 thousands.

However, since COVID-19 hit the world out of a sudden, it is estimated that the health needs created by the coronavirus pandemic go well beyond the capacity of U.S. hospitals[5]. According to the law, the ICU beds, number of nurses, and other medical equipment are all prepared for certain capacity of patients. The transmission of COVID-19 brings several times of patient number which can lead to situations with insufficient medical resources and lack of appropriate treatments for patients. Without in time treatment and therapy, more people may be infected and transmission of COVID would be even worse. Furthermore, for some of the counties and cities, even they have a large number of patients, the medical resources are still sufficient due to the high GDP or other influencing factors. Due to population structure, economics, ICU beds number, or even more reasons, some areas in the U.S. may be more vulnerable than other areas in terms of COVID.

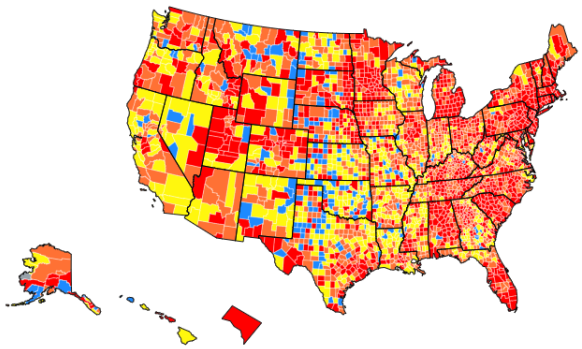

Figure 1. Different transmission level of counties, Blue: Low, Red: High

Considering the difference between counties over the country, the research aims to evaluate the ability of all counties in the U.S. against the COVID-19 outbreak[1] and identify the counties or cities that are more vulnerable. The research will apply clustering techniques to split the counties into groups and evaluate their ability against COVID. It will also apply supervised learning technique to train selected machine learning models for group classification. It is believed that the research would eventually deliver an analytic tool for identifying vulnerable county. The research results would provide solid reference for decision making and greatly help the local governments in resource allocation and COVID-19 related issues[5]. It is expected that this research provide solid reference to the government or other medical organization to maximizing the benefit with limited resource as well as provide evidence for medical decision system.

2 Data Description

For the data collection process, we mainly collect data from CDC, Kaggle and Census Bureau. The vulnerable counties against the COVID-19 outbreak will be identified based on the data sets including total cases, total deaths, underlying medical conditions(including chronic kidney disease, chronic obstructive pulmonary disease (COPD), heart disease, diagnosed diabetes, and obesity), population of the county, population of elder(age 60+), percent of population aged 60+, total ICU beds number, residents aged 60+ per each ICU bed, Gross domestic product(GDP) and poverty condition(log-transformed number of individuals that claimed living in poverty) of the counties over the U.S. The detailed explanation and resource of the data set is described as in the Table 1.

To be notice, for the data we use, some of the data is fixed such as population, underlying disease conditions for each county, ICU beds, GDP, poverty condition. While some of the data we use is changeable by time, such as total cases and total deaths. The total cases and deaths in the project is accumulated cases and projects over time. For this situation, we select a date and its corresponding total cases/deaths data, and then generate models for checking the vulnerability of counties. For the research, February 20(2/20/2021) is chosen as the date and the cumulative cases and deaths for that date are applied for the model. The date is chosen because it contains more counties information and at this stage the vaccination is at the beginning stage which would not influence the results.

Data Title	Description
Total Cases	Data on cumulative coronavirus cases by county and by day[8]
Total Deaths	Data on cumulative coronavirus deaths by county and by day[8]
Underlying medical conditions	Percentage estimates of the prevalence for any of five underlying medical conditions that increase the risk for severe COVID-19-associated illness, including chronic kidney disease, chronic obstructive pulmonary disease (COPD), heart disease, diagnosed diabetes, and obesity[4]
County Population	Population estimation in 2019 by county[3]
County Population of age 60+	Total population of elder people who aged 60+[2]
Percent of population aged 60+	Percentage of population of elder aged 60+[2]
Total ICU beds number	Number of ICU beds in the county[2]
Residents aged 60+ per ICU bed	County residents Aged 60+ Per Each ICU Bed[2]
GDP by dollars	Real gross domestic product by thousands of chained dollars(2019)[7]
Poverty Condition	Individuals (all ages) in the county classified as living in poverty (2019) (log-transformed)[6]

Table 1. Data description and data source

3 Data Pre-possessing

For our research, since our data comes from different sources, there is a need to clean and reformat the data as the procedure in pre-processing. The data are stored as CSV files or Excel xlsx files. All the data is imported and reformatted. They keep the state name, county name and corresponding variable values. For some of the data sources that only used state initials, we changed the state initials to full state names with the aim to match and merge the data sheets. All the data from various sources are combined into one spreadsheet in Python 3 Jupyter notebook.

The columns of the spreadsheet are all the variables that we collected as shown in the Table 1, and they are regarded as preliminary features in the study. Considering the state name and county name, the column number is 16 after data fusion. The total county number in the U.S. is over 3000, the expected number of rows was the number of county number. However, since the data sources have some kinds of variation, when merging the data sheets, some of the counties do not have complete information and they are removed from the spreadsheet. Such that the total number of rows is 2967.

Since the case number and total death number may be varied due to the county size and local population, only count for the number of cases and deaths may not be ideal for the modeling. Thus, to solve the problem, in the pre-possessing process, we considers to use the ratio of total deaths divided by total population, and the ratio of total cases divided by total population in county instead to the pure cases and deaths number. The number of columns remains the same and the pure number of total cases and deaths are eliminated from the spread sheet.

Furthermore, we find that the number of GDP dollars and population is very large, especially for the counties that having large size. The large number of the data may influence the effectiveness of the supervised learning models and data normalization process should be taken. In the research, for all the data in the spreadsheet, z-score normalization method is applied. Sklearn library is used for the normalization process. The data after the normalization process is shown as the Figure 3.

4 Data Exploration

After finishing the normalization process as mentioned in last section, it is necessary to summarize the data and investigate the statistics of each preliminary feature. For each variable, we computed the statistics of features including count, min, max, mean, median, and 25 percentile, 50 percentile(median), and 75 percentile data. The description of the data is shown as the Figure 4.

From the data description, we can see that there are large differences between counties. Take ICU bed number as example, the 75 percentile of ICU bed number is 12, but the maximum bed number is 2126, indicating large variation between counties. We can also learn the variation from the standard deviation(std=84 for the ICU bed number). The situation is similar in other features such as population, population aged 60+, residents aged 60+ per each ICU beds, GDP, cases and deaths. It reveals that the size of counties varies and their population, economics, medical resources are varied. To be noticed, for almost half of the counties(we can see that from the ICU beds summary, in 50 percentile there is still 0), they have no ICU beds and the value of residents aged 60+ per each ICU bed was null. To fill the null value, we use the maximum number of the column plus one. For other features such as underlying disease conditions and poverty condition, since they are percentage of people in the county(underlying disease) or log-transformed index(poverty condition), the variation between counties is more stable than other features. The data description also proves that there is need to normalize the data before further analysis.

state	county	icu beds	total population	population aged 60+	percent of population aged 60+	residents aged 60+ per each icu bed	gdp _dollars	cases	deaths	obesity	heart_disease	COPD	diabetes	CKD	poverty
Alabama	Autauga	6	55036	10523	19.1	1754.0	1501769	5683	69.0	35.8	7.9	8.6	12.9	3.1	10.916415
Alabama	Baldwin	51	203360	53519	26.3	1049.0	6140514	18211	224.0	29.7	7.8	8.6	12.0	3.2	12.279579
Alabama	Barbour	5	26201	6150	23.5	1230.0	762856	1956	40.0	40.7	11.0	12.1	19.7	4.5	9.997843
Alabama	Bibb	0	22580	4773	21.1	8470.0	389547	2309	52.0	38.7	8.6	10.0	14.1	3.3	9.914032
Alabama	Blount	6	57667	13600	23.6	2267.0	869049	5720	100.0	34.0	9.2	10.5	13.5	3.4	10.954973

Figure 2. Data Spreadsheet after Data Fusion

state	county	icu beds	total population	population aged 60+	percent of population aged 60+	residents aged 60+ per each icu bed	gdp _dollars	obesity	heart_disease	COPD	diabetes	CKD	poverty	cases ratio	deaths ratio
Alabama	Autauga	-0.204	-0.138	-0.171	-1.056	-0.934	-0.175	0.173	-0.425	-0.217	-0.061	-0.622	0.463	0.647	-0.266
Alabama	Baldwin	0.332	0.316	0.533	0.246	-1.128	0.016	-1.193	-0.483	-0.217	-0.394	-0.444	1.378	0.174	-0.420
Alabama	Barbour	-0.215	-0.226	-0.243	-0.261	-1.078	-0.206	1.271	1.351	1.278	2.455	1.879	-0.154	-0.339	0.012
Alabama	Bibb	-0.275	-0.237	-0.265	-0.695	0.922	-0.221	0.823	-0.024	0.381	0.383	-0.265	-0.210	0.613	0.800
Alabama	Blount	-0.204	-0.130	-0.121	-0.242	-0.792	-0.201	-0.230	0.319	0.594	0.161	-0.086	0.489	0.507	0.222

Figure 3. Data Spreadsheet after Z-score Normalization

	icu beds	population	people aged 60+	%people aged 60+	residents aged 60+ per each icu bed	GDP	cases	deaths	obesity	Heart Disease	COPD	diabetes	CKD	poverty
count	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00	2,967.00
mean	23.10	100,073.07	20,980.69	24.94	5,132.19	5,743,794.47	8,227.42	132.55	35.03	8.64	9.11	13.07	3.45	10.23
std	84.00	327,317.68	61,097.24	5.53	3,619.17	24,207,952.06	31,440.97	495.70	4.47	1.75	2.34	2.70	0.56	1.49
min	0.00	74.00	29.00	5.80	10.00	22,870.00	1.00	0.00	15.20	3.50	3.50	6.10	1.80	5.02
25%	0.00	10,919.50	2,820.50	21.40	1,176.50	376,257.50	864.50	14.00	32.50	7.50	7.35	11.20	3.10	9.27
50%	0.00	25,574.00	6,262.00	24.60	8,470.00	963,407.00	2,074.00	36.00	35.40	8.60	8.90	12.80	3.40	10.11
75%	12.00	66,275.00	15,814.00	27.90	8,470.00	2,734,368.50	5,399.00	87.00	37.90	9.80	10.60	14.80	3.80	11.07
max	2,126.00	10,105,722.00	1,800,341.00	64.20	8,470.00	726,943,301.00	1,121,349.00	16,854.00	49.90	15.10	19.90	25.60	5.90	16.11

Figure 4. Data Description and Summary

We also want to learn the correlation between the variables. With the correlation matrix, we can easily tell whether linear relationship between the features we have. The Pearson correlation matrix of the data features is shown as Figure 6. From the matrix, we can learn that the ICU bed number, population, population aged 60+, and GDP are closely related. The range of their correlation is above 0.88. However, it seems they have no clear correlation with percent of people aged 60+. Moreover, we can find that poverty is somehow positively related to ICU beds, population, people aged 60+, and GDP, but negatively related to percent of population aged 60+, residents aged 60+ per each ICU beds. We also can find that all the underlying disease are positively correlated with each other, the correlation ranges between 0.54-0.91. We also can find that cases ratio is correlated with deaths ratio though the relationship is not very strong.

Thus, since the ICU beds, population, population aged 60+ and GDP are very closely correlated. We may need to consider if some of the attributes should be deleted since the effects may influence the model building results. In this case, we can remove the first three attributes from the data sets(ICU beds, population, population aged 60+). And keep all other attributes for further analysis.

5 Model Building

With the aim to evaluate the ability of each county against COVID-19, we first apply clustering methods to separate the counties in U.S. into various groups. Then we look into the

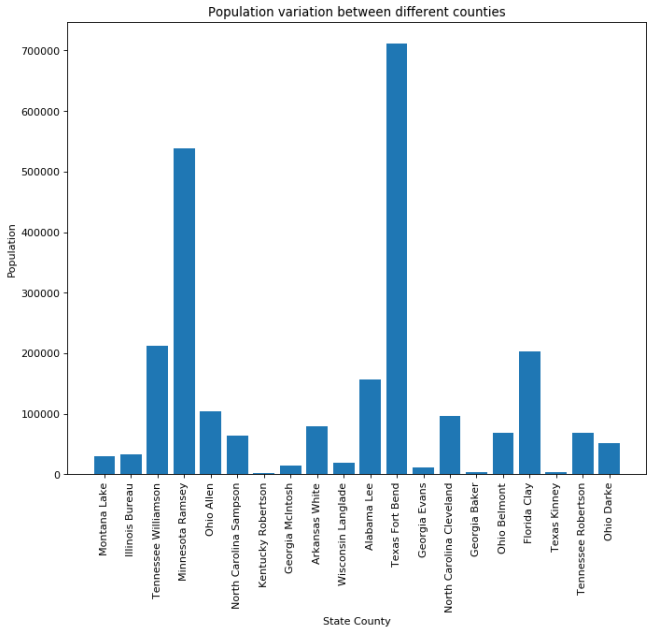

Figure 5. Population Variation between Counties

details of each group, and classify the vulnerability of each group based on the cases, deaths, population, and residents per ICU beds. Then to further apply the model for classification use, we use the clustering results to label the groups, and then separate the data using 5-fold cross-validation methods with 80 percent training sets and 20 percent testing sets. For the labelled data, we apply supervised learning methods to build up machine learning models for county vulnerability classification. The performance of each clustering algorithm and supervised learning model classification will be evaluated by precision, F1 score and recall.

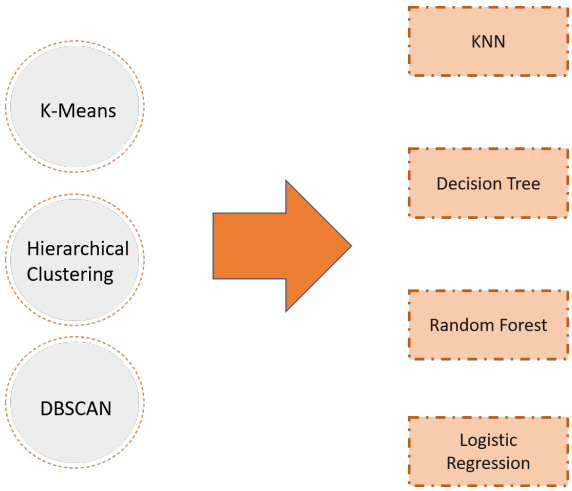

Figure 6. Clustering and Supervised learning road map

For the clustering selection, the study applies K-means clustering algorithm, Agglomerative Hierarchical clustering, and DBSCAN clustering algorithm. And for the supervised learning algorithm, the research applies K-Nearest Neighbour algorithm, Decision Tree algorithm and Random Forest algorithm. The model selection can be summarized as the Figure 7. The research would compare each model combination and corresponding figure to decide which model combination is the best for vulnerable county identification.

5.1 Principle Component Analysis

Before building up the models of clustering and supervised learning, there is a need to select the features and reduce the dimension of the data. Principle Component Analysis(PCA) method is used to transform the data and reduce the data dimension. For the research, first 6 components are selected for representing the original data and they account for 91.8 percent of the variance. The variance ratio of the first 6 components are 0.437, 0.155, 0.135, 0.008, 0.067, 0.039. The percentage of variance vs. number of component plot is shown in the Figure 7.

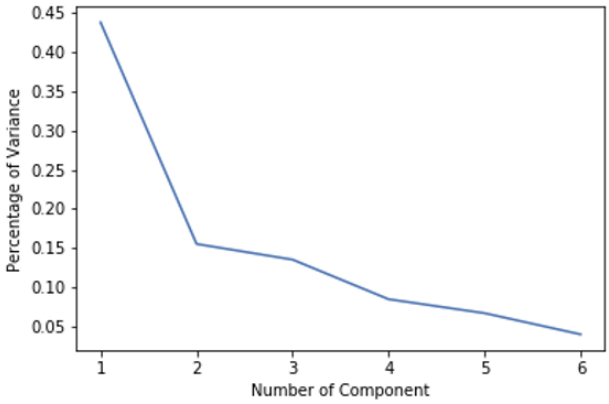

Figure 7. PCA percentage of variance plot

Thus, for the following analysis, PCA data will be used to take the place of the orginal data.

5.2 K-means Clustering

K-means algorithm is one of the most popular clustering methods. Since the algorithm can cluster the data into k groups, we define the k as 3. For the clustering, we plan to have 3 clusters: low risk, moderate risk and high risk group. The random state is 0 for the algorithm.

The algorithm clusters the data into 3 groups and labels them as 0, 1, 2. The cluster plot is shown in the Figure 9. The x and y axes are the first two components of PCA. From the figure we can see that the data is scattered with no pattern and 3 clustered are classified into green, blue and red. The we calculate the data summary from the 3 clusters and part

	icu beds	population	people aged 60+	%people aged 60+	residents aged 60+ per each icu bed	gdp _dollars	obesity	heart_disease	COPD	diabetes	CKD	poverty	cases ratio	deaths ratio
icu beds	1.000	0.924	0.922	-0.209	-0.328	0.887	-0.214	-0.267	-0.216	-0.146	-0.188	0.527	-0.003	-0.043
population	0.924	1.000	0.988	-0.220	-0.284	0.954	-0.245	-0.296	-0.245	-0.173	-0.218	0.541	-0.024	-0.059
people aged 60+	0.922	0.988	1.000	-0.181	-0.304	0.931	-0.267	-0.281	-0.236	-0.174	-0.208	0.568	-0.044	-0.058
%people aged 60+	-0.209	-0.220	-0.181	1.000	0.315	-0.201	-0.026	0.594	0.320	0.259	0.499	-0.453	-0.285	0.039
residents aged 60+ per each icu bed	-0.328	-0.284	-0.304	0.315	1.000	-0.237	0.197	0.336	0.190	0.220	0.303	-0.680	0.017	0.089
gdp _dollars	0.887	0.954	0.931	-0.201	-0.237	1.000	-0.251	-0.283	-0.243	-0.169	-0.211	0.469	-0.032	-0.055
obesity	-0.214	-0.245	-0.267	-0.026	0.197	-0.251	1.000	0.549	0.575	0.690	0.540	-0.308	0.239	0.256
heart_disease	-0.267	-0.296	-0.281	0.594	0.336	-0.283	0.549	1.000	0.884	0.831	0.892	-0.501	-0.050	0.217
COPD	-0.216	-0.245	-0.236	0.320	0.190	-0.243	0.575	0.884	1.000	0.784	0.727	-0.294	-0.058	0.121
diabetes	-0.146	-0.173	-0.174	0.259	0.220	-0.169	0.690	0.831	0.784	1.000	0.916	-0.328	0.045	0.292
CKD	-0.188	-0.218	-0.208	0.499	0.303	-0.211	0.540	0.892	0.727	0.916	1.000	-0.448	-0.015	0.298
poverty	0.527	0.541	0.568	-0.453	-0.680	0.469	-0.308	-0.501	-0.294	-0.328	-0.448	1.000	-0.080	-0.182
cases ratio	-0.003	-0.024	-0.044	-0.285	0.017	-0.032	0.239	-0.050	-0.058	0.045	-0.015	-0.080	1.000	0.467
deaths ratio	-0.043	-0.059	-0.058	0.039	0.089	-0.055	0.256	0.217	0.121	0.292	0.298	-0.182	0.467	1.000

Figure 8. Pearson Correlation Matrix

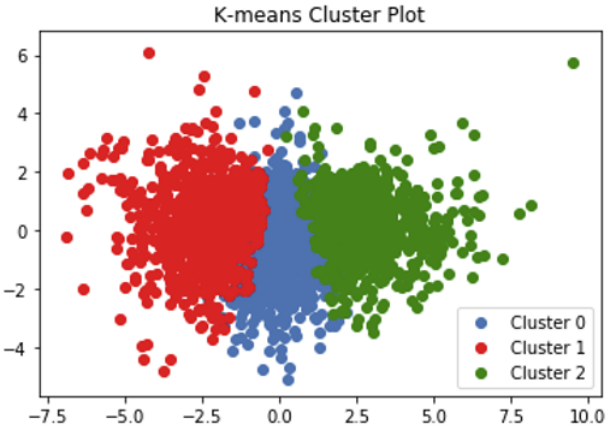

Figure 9. K-means Clustering Result

of the data summary is shown as the Table 2. The data in the table is the mean of the data category in that cluster.

From the Table 2, we can see that cluster 1 has the highest elderly ratio, least average number of ICU bed, highest obesity ratio, heart disease ratio, diabetes disease ratio, and death ratio. However, we can find that the poverty index of cluster 1 is the least one. And it seems like the case ratio is not much related to the clustering. Thus, based on the information, we can assign that cluster 1 should be the high risk group since it has limited medical resources but highest underlying disease ratio and death ratio. Similarly, we assign group 0 as the moderate risk group and group 2 as the low risk group.

5.3 Hierarchical Clustering

Hierarchical clustering algorithm is another popular clustering methods. The algorithm cluster the data into groups with hierarchy. The hierarchical algorithm used in the case is agglomerative hierarchical algorithm. The algorithm cannot pre-define the group number. The planned groups of the data are low risk group and high risk group. The random state is 0 for the algorithm.

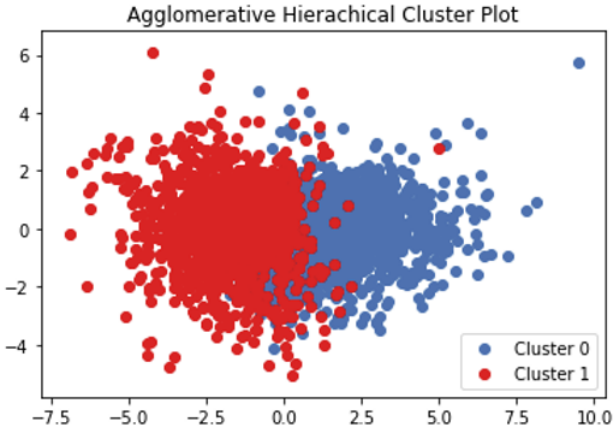

Figure 10. Agglomerative Hierarchical Clustering Result

The algorithm clusters the data into 2 groups and labels them as 0, 1. The cluster plot is shown in the Figure 10. The x and y axes are the first two components of PCA. From the figure we can see that the data is clustered and labeled with blue and red. Similar to the K-means method, we calculate

Cluster	Elderly Ratio	Elderly ICU bed	Obesity Ratio	Heart Disease Ratio	Diabetes Ratio	Poverty Index	Case Ratio	Death Ratio
0	26.1	5996	34.8	8.5	12.4	9.7	0.086	0.0015
1	27	6384	38.4	10.6	16.2	9.58	0.084	0.0019
2	20	2002	31.4	6.6	10.5	11.9	0.081	0.0011

Table 2. Data Comparison between 3 Clusters by K-Means

the data summary from the 2 clusters and part of the data summary is shown as the Table 3.

From the Table 3, we can see that cluster 1 has higher elderly ratio, least average number of ICU bed, higher obesity ratio, heart disease ratio, diabetes disease ratio, and death ratio. The poverty index of cluster 1 is the smaller one. Since the pattern of poverty index is same as K-means method, we may assume that poverty has negative relationship with overall risk of the COVID. Maybe that is because elder people have more property than young people. Based on the data summary, similar to previous analysis, we can assign that cluster 1 should be the high risk group since it has limited medical resources but highest underlying disease ratio and death ratio. We assign group 0 as the low risk group.

5.4 DBSCAN Clustering

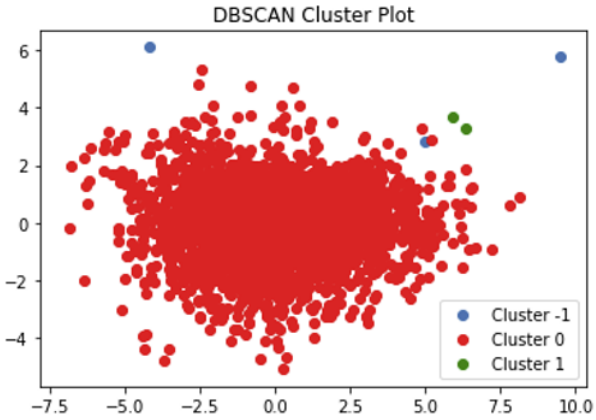

Figure 11. DBSCAN method for 3 clusters

DBSCAN is Density-based spatial clustering of applications with noise. It is a density-based clustering method and suitable for finding clusters with specific patterns or shapes. The algorithm requires users to define the density with the number of points within a specified radius. The algorithm will identify core points, border points and noise points based on the density. However, for the PCA data we use in the study, we can observe that the scatter data does not have a regular pattern or shape, making it is not a good case for DBSCAN method. Since the expected cluster number is 2

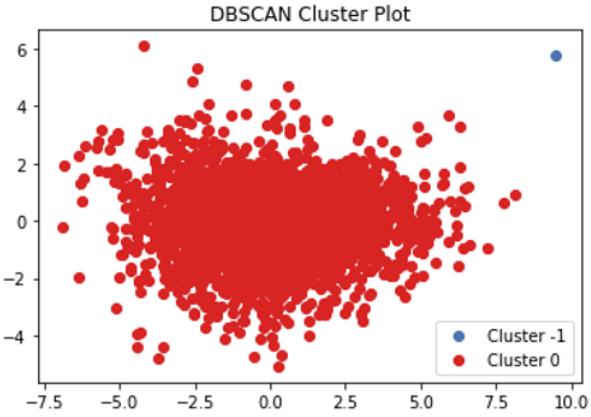

Figure 12. DBSCAN method for 2 clusters

or 3, with the change of number of points within radius(eps). We can obtain the results as shwon in the Figure 11 and 12.

From the figure, we can see that either 3 clusters(eps=2) or 2 clusters(eps=2) are not suitable for modeling since most of the data are clustered into same cluster. There are very few data samples assigned to other clusters. The very imbalanced group assignment making the model not practical for further analysis. Thus, the DBSCAN method will not be applied in the research.

5.5 Supervised Learning

According to the analysis in above sections, the research obtains two available clustering methods and they will be combined with selected supervised learning algorithms. The selected supervised learning methods include K-Nearest Neighbour(KNN), Decision Tree, Random Forest, Logistic Regression. 4 learning methods are applied with 5-fold cross validation. The results of them are presented in Model Evaluation section.

6 Model Evaluation

6.1 KNN

KNN algorithm is an algorithm relies on location and distance, it identify the neighbours of data points and classify them based on its distance to centroid and iterate the process. For the K-means clustering algorithm, since we identify

Cluster	Elderly Ratio	Elderly ICU bed	Obesity Ratio	Heart Disease Ratio	Diabetes Ratio	Poverty Index	Case Ratio	Death Ratio
0	23	3909	33	7.43	11.3	10.8	0.082	0.0012
1	27	6429	37	9.92	15	9.6	0.088	0.0018

Table 3. Data Comparison between 2 Clusters by Agglomerative Hierarchical Clustering

3 clusters, we would have k=3 in KNN modeling. For the Hierarchical clustering algorithm, since we have 2 clusters in total, we would have k=2 for this method. The results of KNN for K-means clustering and Hierarchical clustering methods are shown in Figure 13.

According to Figure 13, both clustering and classification models are performing well in terms of average precision, recall and F1-score, though the details between groups varied. Based on the average results, we can see that the Hierarchical clustering with KNN combination performs slightly better than K-means with KNN.

6.2 Decision Tree

Decision tree method is another popular classification algorithm. It is a tree like model with branch structure representing the outcomes of the classification. One of the advantages of the model is easy for interpretation and explanation. Decision tree method is also applied to this research. The results of decision tree with K-means and Hierarchical clustering algorithm is shown as in the Figure 14.

According to Figure 14, both clustering and classification models are performing well with evaluation metrics. Based on the average results. We can see that the overall results of the Decision tree method is not as good as KNN method.

6.3 Random Forest

Random forest is an ensemble classification with developing multiple decision trees. Random forest method is also applied to this research to compare with other clustering method and supervised learning method combinations. The classification report of the random forest method is shown in the Figure 15.

According to Figure 15, both clustering and classification models are performing well with evaluation metrics. In terms of the overall performing evaluation, it is hard to distinguish their performances from other method combinations.

6.4 Logistic Regression

Logistic regression method is regression method in statistics that can assign probability to each data point and classify them into 0 or 1 group. It is a commonly used model for modeling a binary variable. In this research, logistic regression will be applied to compare its performance with other algorithms. The results of decision tree with K-means and Hierarchical clustering algorithm is shown as in the Figure 16.

According to Figure 16, we find that K-means with logistic regression combination has the highest value of average precision, recall and F1-score. They are 0.95, 0.96 and 0.95, respectively. This result is the most ideal one comparing to previous methods. Thus, K-means combines with logistic regression is the winner and will be used as the clustering and classification model for identifying the COVID risk and vulnerable counties for this research.

7 Results

With the rapid spread of COVID-19, millions of people is U.S. are suffered from the disease and the patient number exceed the capacity of local hospitals. It is a huge challenge to the local government and has been an emergency over the nation even over the world. Fairly allocate medical resources to region with priority is essential but identify those counties with little ability against COVID can be a problem. The research investigate and compare multiple clustering methods and supervised learning methods with the aim to find a useful tool to find counties with priority.

To the end, we find the best combination of the model and it is K-means clustering method combines with logistic regression. It have the highest evaluation metrics with precision, recall and F1-score as 0.95, 0.96 and 0.95 respectively. The classification matrix of the model is also plot as in the Figure 17.The Figure 17 shows that there are very few of the data is mistakenly classified and further prove the high accuracy of the selected model.

8 Real-World Insights

We would like to apply the model to estimate the counties that are vulnerable and provide reference for state government to put those counties into priority. The selected model is supposed to be a powerful tool for government decision-maker to evaluate the risk of severe COVID outcome for each county around the country. Through the model classification, decision-maker can determine the priority level of COVID medical resource allocation or vaccine allocation. We would recommend policy makers to allocate more medical resources to those areas that are in need. Furthermore, for the counties in the same group, decision-makers can have some reference from existing policies from other counties and decide if they should make similar arrangement for the

K-means + KNN

	precision	recall	f1-score	support
0	0.92	0.94	0.93	1354
1	0.93	0.93	0.93	885
2	0.96	0.93	0.95	728
accuracy			0.93	2967
macro avg	0.94	0.93	0.93	2967
weighted avg	0.93	0.93	0.93	2967

Hierarchical + KNN

	precision	recall	f1-score	support
0	0.91	0.99	0.95	1527
1	0.99	0.90	0.94	1440
accuracy			0.94	2967
macro avg	0.95	0.94	0.94	2967
weighted avg	0.95	0.94	0.94	2967

Figure 13. KNN results for 2 clustering methods

K-means + DT

	precision	recall	f1-score	support
0	0.92	0.90	0.91	1354
1	0.92	0.94	0.93	885
2	0.91	0.92	0.92	728
accuracy			0.92	2967
macro avg	0.92	0.92	0.92	2967
weighted avg	0.92	0.92	0.92	2967

Hierarchical + DT

	precision	recall	f1-score	support
0	0.90	0.89	0.89	1527
1	0.88	0.90	0.89	1440
accuracy			0.89	2967
macro avg	0.89	0.89	0.89	2967
weighted avg	0.89	0.89	0.89	2967

Figure 14. Decision Tree results for 2 clustering methods

K-means + RF

	precision	recall	f1-score	support
0	0.92	0.95	0.94	1354
1	0.97	0.94	0.95	885
2	0.94	0.93	0.93	728
accuracy			0.94	2967
macro avg	0.94	0.94	0.94	2967
weighted avg	0.94	0.94	0.94	2967

Hierarchical + RF

	precision	recall	f1-score	support
0	0.90	0.95	0.92	1527
1	0.95	0.89	0.91	1440
accuracy			0.92	2967
macro avg	0.92	0.92	0.92	2967
weighted avg	0.92	0.92	0.92	2967

Figure 15. Random Forest results for 2 clustering methods

K-means + LR

	precision	recall	f1-score	support
0	0.96	0.93	0.95	1354
1	0.94	0.98	0.96	885
2	0.95	0.96	0.95	728
accuracy			0.95	2967
macro avg	0.95	0.96	0.95	2967
weighted avg	0.95	0.95	0.95	2967

Hierarchical + LR

	precision	recall	f1-score	support
0	0.89	0.91	0.90	1527
1	0.90	0.88	0.89	1440
accuracy			0.89	2967
macro avg	0.89	0.89	0.89	2967
weighted avg	0.89	0.89	0.89	2967

Figure 16. Logistic Regression results for 2 clustering methods

county. For this research, it is expected that the analysis results and the models could bring help to government decision making process in terms of resource allocation.

9 Lessons Learned

This advanced analytic course brings great help to me to have a overview of web mining, recommender systems, text mining, social network analysis, time-series, big data platform, data warehousing and etc. It broadens my horizon in machine learning area and help me prepare knowledge for

Moderate risk	1263	54	37
High risk	15	870	0
Low risk	31	0	697
	Moderate risk	High risk	Low risk

Figure 17. Confusion Matrix of K-means + Logistic Regression

my future career. Before the course, I have no experience in web mining, graph modelling and this course gives me a good chance to learn about them and their application in industry. Overall, the course also arise my interest in data analytic area and I wish to learn more about it in future.

Moreover, the course project is another valuable experience for me to perform a complete data analysis for a trending topic related to COVID. The project improves my independent research ability and my coding skills. It pushes me to think what should be a practical research topic and something people really need under this pandemic. The project practice also enhance my ability to search for information and interpret them. I am very thankful about what I have learned during the class and the project in this semester.

References

[1] Nezir Aydin and Gökhan Yurdakul. 2020. Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms. *Applied Soft Computing* 97 (2020), 106792.

[2] Jaime Blasco. [n.d.]. Medical Resource Dataset Source https://www.kaggle.com/jaimeblasco/icu-beds-by-county-in-the-us.

[3] Census Bureau. [n.d.]. County Population Totals: 2010-2019 Dataset Source https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html.

[4] CDC. [n.d.]. Underlying Medical Conditions Dataset Source https://covid.cdc.gov/covid-data-tracker/#underlying-med-conditions.

[5] Ezekiel J Emanuel, Govind Persad, Ross Upshur, Beatriz Thome, Michael Parker, Aaron Glickman, Cathy Zhang, Connor Boyle, Maxwell Smith, and James P Phillips. 2020. Fair allocation of scarce medical resources in the time of Covid-19.

[6] Laurindo Garcia. [n.d.]. Poverty Dataset Source https://www.kaggle.com/laurindogarcia/covid-19-race-gender-poverty-risk-us-county.

[7] BEA Gov. [n.d.]. GDP by County Dataset Source https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas.

[8] MyrnaMFL. [n.d.]. US counties COVID 19 Dataset Source https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset.