Factors affecting the COVID-19 risk in the US counties: an innovative approach by combining unsupervised and supervised learning

Samira Ziyadidegan
Texas A&M University College Station: Texas A&M University
Moein Razavi (moeinrazavi@tamu.edu)
Texas A&M University College Station https://orcid.org/0000-0002-8036-1353
Homa Pesarakli
Texas A and M University College Station: Texas A&M University
Amirhossein Javid
Texas A&M University College Station: Texas A&M University
Madhav Erraguntla
Texas A&M University College Station: Texas A&M University

Research Article

Keywords: multinomial logistic regression, K-means clustering, COVID-19, SARS-CoV-2, meteorological variables

Posted Date: June 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-576376/v1

Version of Record: A version of this preprint was published at Stochastic Environmental Research and Risk Assessment on January 11th, 2022. See the published version at https://doi.org/10.1007/s00477-021-02148-0.

Abstract

The COVID-19 disease spreads swiftly, and nearly three months after the first positive case was confirmed in China, Coronavirus started to spread all over the United States. Some states and counties reported high number of positive cases and deaths, while some reported lower COVID-19 related cases and mortality. In this paper, the factors that could affect the risk of COVID-19 infection and mortality were analyzed in county level. An innovative method by using K-means clustering and several classification models is utilized to determine the most critical factors. Results showed that mean temperature, percent of people below poverty, percent of adults with obesity, air pressure, population density, wind speed, longitude, and percent of uninsured people were the most significant attributes.

1. Introduction

COVID-19 disease is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with common symptoms of fever, dry cough, shortness of breath, and other signs of respiratory-related infections. World Health Organization (WHO) reported that 80% of patients experienced these symptoms mildly. However, older people (>60 years old) and persons with co-morbid diseases are at a higher risk for severe symptoms and death [1,2]. Besides, younger patients with no underlying disease might also experience severe symptoms or even death [3–5].

The first positive case of COVID-19 in the United States was reported in the state of Washington on January 20, 2020. By March 17, 2020, Covid-19 has spread across all US states [6,7]. Figure 1 shows the aggregated COVID-19 positive case and death count maps for all US states until November 6, 2020. Reports showed that on November 6, 2020, the top states for positive COVID-19 cases are California, Texas, Florida, New York, and Illinois, while the top 5 states for death cases are New York, Texas, California, New Jersey, and Florida. [8,9].

Epidemiological models have been used for outbreak estimation and predicting upcoming peak and mortality rate. Accurate outbreak prediction can provide insight into problems caused by COVID-19 and be used to develop new policies [10]. The COVID-19 pandemic has shown a complex nature unlike other recent outbreaks [11]. The COVID-19 can more rapidly spread through human contact comparing to other recent epidemics like Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) [12]. Furthermore, many known and unknown variables affect in the spread of COVID-19. So, it has an uncertain out-break prediction that cannot be estimated accurately by standard epidemiological models. The growing big data of the number of infected subjects, death counts, and possible influential factors requires to be managed and analyzed by innovative solutions [13]. Multiple sources of data provide information about the various region across the country including the growth of infection, geographical, local services and policies, and demographical data. In order to have more precise and long-term prediction, Ardabili et al. [11] showed that the application of various Machine learning algorithm is more effective in estimating covid-19 positive case rate and mortality rate. They also

mentioned that estimating the factors that affect the mortality rate is important in estimating the number of patients and planning new facilities.

Previous studies have been shown that several factors affect COVID-19 spread and mortality rate. They showed that meteorological factors could have an essential role in the spread of a virus [14,15]. Temperature can affect the life cycle and proliferation of viruses; precipitation and rainfall can impact water-borne virus diffusion, while humidity and the wind speed can affect the dissemination of air-borne viruses [14,16–19]. Different studies showed contradictory effects regarding the meteorological attributes on the transmission or lifespan of Coronavirus. Liu et al. [20] studied the correlation between confirmed case counts and weather-related attributes. The results showed that diurnal temperature, ambient temperature, and humidity are significant attributes and have negative correlations with the transmission of COVID-19. Moges Menebo [21] found that the maximum temperature and the normal temperature (average of high and low temperatures) have positive, but precipitation has negative impact on COVID-19 case counts. Ma et al. [22] evaluated the effects of meteorological and air pollutant factors on COVID-19 death rates. They concluded that diurnal temperature range and absolute humidity are positively and negatively associated with COVID-19 mortality rate, respectively. He et al. [23] studied the effects of temperature and relative humidity on daily COVID-19 confirmed positive cases in different Asian cities. Their study revealed that in different cities, the impact of temperature on COVID-19 confirmed cases could differ. In Beijing, Shanghai, and Guangzhou, the correlation between temperature and the number of positive cases was negative, while in Japan, it was positive. The study conducted by Pahuja et al. [24] demonstrated a positive correlation between the wind speed and the COVID-19 case counts. In contrast, Coccia [25] showed that the high wind speed could reduce the number of COVID-19 cases.

In addition to the meteorological factors, there are other factors that impact the transmission and spread of COVID-19 positive case and death case counts. They include air pollutant factors such as the concentration of air pollutants (e.g., PM_{10} and SO_2) [22,25], demographic factors (e.g., elderly people index, population density, etc.) [25], adults smoking history [26,27], location-based factors (e.g., latitude) [25,28–30], and disease-related factors (e.g., cardiovascular or respiratory disease experience) [25,27,31,32]. Besides, the COVID-19 outbreak challenges medical systems worldwide in many aspects like increasing the demands for hospital beds and medical equipment [33]. The shortage of the healthcare resources creates challenges for treating critically ill patients [34].

Reviewing the literature indicated that an analysis that took into account a comprehensive set of factors affecting the rate of positive COVID-19 cases and its death rate was missing. The contradictory findings might be due to confounding factors and a comprehensive study based on a wide range of predictor variables will result in more accurate characterization. Besides, as Figure 1 shows, states with a high number of COVID-19 cases (e.g., Florida, New York, and California) are located in areas that are different in terms of geographical, health, demographical, and meteorological characteristics. However, the majority of the previous studies analyzed the factors affecting COVID-19 transmission in the areas with similar characteristics (e.g., geographical, demographical, etc.). Hence, a comprehensive model that considers more extensive areas (e.g., all US counties) with different geographical, health, demographical,

and meteorological features will be able to characterize the differences in risk across these diverse regions.

In this study, a comprehensive set of factors that affect the risk level of COVID-19 in all US counties have been analyzed. To perform this analysis, we combined unsupervised and supervised learning using clustering analysis and classification models. The results of this study show why some areas have higher risk of COVID-19 than other areas.

This paper is organized as follows. In Section 2, the methodology is described. Section 3 introduces the datasets used in this paper. Then, the process of data preparation and cleaning is explained. The analysis results are presented in Section 4. Finally, the findings are discussed in Section 5.

2. Methodology

Methodology

Figure 2 indicates the flow used in this article to determine the significant factors affecting COVID-19 transmission and the risk level of each county. First, we combined data from different sources and corrected the inaccurate records from the dataset (data preparation). Next, for feature reduction, we calculated the Pearson correlation coefficients between different parameters, and chose a single variable from each set of highly correlated variables. We then performed clustering analysis on COVID-19 data using Elbow method and K-means clustering. Based on clustering results, we labeled the counties into different classes of risk level. In order to profile the clusters and understand the features affecting cluster membership, we applied different classification models on the cluster labels (to overcome the class imbalance, we applied Synthetic Minority Oversampling Technique (SMOTE) [35]). Classification model with the best performance accuracy was selected to determine the significant variables affecting COVID-19 risk levels.

3. Data Preparation

3.1. Data Collection

The data used in this study was taken from various online sources from March 2020 to November 2020, and includes COVID-19 positive cases and deaths, demographic, meteorological, health, and location-based data for each US county (Table 1). To evaluate the risk level of each county, we used the COVID-19 positive cases and death rates (positive cases and death counts divided by the population of the county) instead of their actual numbers. For meteorological data, we took the average of the data over all days of each month to have the average monthly data to match the temporal granularity of other data sets. The maximum and minimum temperatures for each county were derived from the maximum and minimum temperature across all days from March to November. Population density along with the population of elderly people (> 65 years old) and young people (< 65 years old) were also included in the dataset. The parameters used in this study are shown in Table 1.

Table 1
List of all the parameters used in the study

Category	Parameters
COVID-19	positive rate, death rate
location-based	latitude, longitude, county Area, percent of rural areas
meteorological	minimum temperature, maximum temperature, mean temperature, wind speed, air pressure, precipitation, dewpoint
Health	number of ICU beds, number of staffed beds, percent of smokers, percent of adults with obesity, percent of people uninsured, percent of adults with diabetes
Demographic	elderly population, young population, white population, non-white population, eightieth percentile income, twentieth percentile income, percent of people below poverty, total population, population density

3.2. Data Cleaning

After combining the data from various sources, it was found that some values were missing in different variables (Fig. 3). Missing values for the demographic data were replaced with data from the United States Census Bureau website [36]. If no data was available for a specific county, values were imputed using the average of the non-missing values of that parameter across the state for income related data. For weather data, the missing values are imputed with the data available from one neighboring county. For example, the missing precipitation, air pressure and dewpoint values for the “District of Columbia” county were replaced with the corresponding data for “Arlington” county in Virginia state. The total number of counties analyzed in the current study was 3131.

3.3. Removing Collinearities by Correlation

After normalizing all independent parameters, Pearson Correlation was calculated for all variables (correlation matrix shown in Fig. 4), and among the ones with high linear correlation (greater than 0.8) only one was kept. The final parameters for further analyses are shown in Table 2.

Table 2
List of parameters used in this study for analysis

Category	Parameters
COVID-19	positive rate, death rate
location-based	longitude, county area, percent of rural areas
meteorological	mean temperature, Wind Speed, precipitation, air pressure
health	number of ICU beds, percent of smokers, percent of adults with obesity, percent of people uninsured, percent of adults with diabetes
demographic	eightieth percentile income, percent below poverty, population density

4. Results

4.1. Clustering Analysis

To determine the COVID-19 risk level for each county, K-means clustering was performed on COVID-19 positive rates and death rates. Clustering was performed to group the counties based on similarities in their risk profile. To determine the optimal number of clusters that can define the risk level of each county, the Elbow method is used (Figure 5). The Elbow method is a visual method that can determine the optimal number of clusters considering the total within-cluster sum of squares of Euclidean distances (the cost). The optimized k value (k is the number of clusters) is such that adding another cluster (k+1) does not significantly decrease the benefit. In the Elbow method plot, the optimal k value is located at the elbow of the curve [37,38].

Figure 5 shows that k=3 could be the optimal number of clusters. Figure 6 shows the K-means clustering result that clustered the counties into three groups of low positive-low death rate, medium positive-medium death rate, and high positive-high death rate. The county located far apart from the other points (in the top right quadrant in Figure 6) belongs to the New York County, which had a positive rate of 15.5% and a death rate of 1.5%.

Figure 7 and Table 3 show descriptive results of each cluster. Cluster 1 is the low-risk cluster. It contains the highest number of counties with the lowest positive rate and death rate on average. On the other hand, cluster 3 has the lowest number of counties but the highest positive rate and death rate on average, which refers to high-risk counties. Cluster 2 is considered as the medium-risk cluster.

In the next section, classification analysis is applied on the results of clustering to find the significant parameters that affect the risk level of each county. Three COVID-19 risk levels of Low, medium, and high were used as labels for the classification analysis.

TABLE 3 Cluster attributes

Cluster number	Counts	Mean Positive Rate	Mean Death Rate
1	1873	1.231e-2	1.838e-4
2	1024	2.966e-2	6.639e-4
3	234	5.257e-2	18.293e-4

4.2. Model Selection

As can be seen in Table 3, there is size imbalance between different classes. We addressed the class imbalance using SMOTE. Then, to determine the significant factors, different classification models were employed to characterize COVID-19 risk clusters. Based on the accuracy values attained, the best model was used to select the factors with the highest significance in the transmission and mortality rate of COVID-19.

For classification, the data was divided into train and test sets (80% and 20%, respectively). Table 4 shows the classification models used in this study and their respective test accuracies. Among the classification models, Random Forest obtained the best performance on the test data. The linear models including MLR, LDA, and SVM Linear performed similarly. Therefore, to perform feature selection, we select the Random Forest model.

TABLE 1 Train and test errors of the classification models used

Method	Test Accuracy
Multinomial Logistic Regression	72.92%
LDA	70.55%
QDA	64.1%
KNN	80.81%
SVM Linear	73.18%
SVM Radial	85.54%
SVM Polynomial	84.66%
Random Forest	**85.63%**

4.3. Feature Selection

To identify the parameters which affect a county's COVID-19 risk level, the Random Forest model was used. Figure 8 shows the variable importance scores of Mean Decrease Accuracy (MDA, a measure of model accuracy loss by excluding each variable) and Mean Decrease in Gini (MDG, a measure of each variable's contribution to the homogeneity of the nodes and leaves in the resulting Random Forest) [39].

Based on both MDA and MDG criteria, *mean temperature, percent of people below poverty* (people with income lower than the threshold determined by the United States Census Bureau [40]), *air pressure, longitude, percentage of uninsured people* and *population density* are the highest contributing factors to the level of COVID-19 transmission and mortality rate.

In order to verify the consistency of important factors and to better interpret the contribution of each parameter to COVID-19 risk level, we used the Multinomial Logistic Regression (MLR) model. For that purpose, a backward selection approach was used based on the p-values of the coefficients in the MLR model (considering 95% confidence interval, a=0.05). The selected factors using this criterion were *mean temperature, percent of people below poverty, air pressure, longitude, percentage of uninsured people, population density, percent of adults with obesity* (BMI > 30 [41]) and *wind speed*. These selected features by the MLR model are almost the same as the features selected by the Random Forest model. It should be noted that other models with acceptable accuracy values (e.g., LDA, KNN, and SVM) suggested almost the same significant factors as MLR.

4.4. Significant Variables Analysis

In this step, the MLR model is applied only to the significant variables. Table 5 shows the coefficients of significant variables driven from the MLR model. The column *Cluster 2* shows the odds ratio of variables of cluster 2 compared to cluster 1, and the column *Cluster 3* shows the odds ratio of the third columns are for cluster 3 compared to cluster 1.

TABLE 5 Odds ratio of significant variables

Variable	Cluster 2	P-value	Cluster 3	P-value
Intercept	0.545	< 0.001	0.642	< 0.001
Mean Temperature	1.892	< 0.001	1.952	< 0.001
Percent of People Uninsured	1.333	< 0.001	1.660	< 0.001
Percent of People Below Poverty	1.205	0.003	2.410	< 0.001
Air Pressure	0.751	< 0.001	0.468	< 0.001
Percent of Adults with Obesity	1.273	< 0.001	1.349	< 0.001
Population Density	1.705	< 0.001	1.957	< 0.001
Longitude	1.337	< 0.001	1.156	0.014
Wind Speed	1.275	< 0.001	1.551	< 0.001

Table 5 demonstrates that increasing the average temperature, percent of people below poverty, percent of adults with obesity, longitude, wind speed, population density, air pressure, and percent of people uninsured would increase a county's chance to be in a cluster with a higher level of COVID-19 risk. On the other hand, increasing the air pressure would decrease that chance. Among them, mean temperature and population density are the two factors which have the highest impact on the risk level.

Previous studies concluded that there is a positive relationship between temperature and COVID-19 cases [21,22,42]. The results of this study were in line with their conclusion. Higher average temperature belonged to clusters 2 and 3, which had higher COVID-19 positive rate and mortality rate.

Results revealed that the percentage of people below poverty in a county was positively associated with belonging to a cluster with a higher level of COVID-19 risk, as shown in Table 5. Low-income people might have limited access to health products such as masks and sanitizers, which affects virus transmission and mortality [43,44]. They are less likely to work from their homes due to unstable jobs and income or less likely to have reliable and valid information about the COVID-19 [45,46]. Compared to other factors, the percentage of people below poverty has the most significant effect on a county's chance to belong to the high-risk cluster.

Additionally, low-income people are less likely to have health insurance (the higher percentage of people uninsured). So, due to high medical expenses, they prefer not to go to clinics/hospitals or use medications, which might increase the COVID-19 death rate and, as a result, increase the association of a county to a higher risk cluster.

As the center for disease control and prevention (CDC) [47] stated, obesity would increase the risk of COVID-19 death. Besides, obesity would affect the immune system adversely. These are in line with the findings of this study. As the obesity percentage in a county increases, the chance of being in the high-risk cluster would increase as well.

Results demonstrate that air pressure was a significant factor which lowers the chance of a county to belong to a higher risk cluster. Air pressure determines the precipitation, wind, and weather condition. High air pressure is associated with mild wind and calm weather [48]. So, as Coccia [25] found, high pressure decreased the transmission of COVID-19. Research conducted by Takagi et al. [49] also demonstrated that high air pressure would reduce the COVID-19 prevalence. Consistent with these studies, our study shows that the air pressure lowers the probability of a county to belong to higher risk clusters.

Table 5 indicates that counties with dense population have a higher chance of being in higher risk clusters. In dense areas, people cannot keep physical distance from others which is one of the most important factors to prevent the transmission of COVID-19 [50,51].

Studies showed that COVID-19 transmission is dependent upon seasonal dynamics. Longitude is a factor that correlates with seasonal dynamics and affects the COVID-19 transmission [52,53]. Findings of our study show that increasing the longitude would increase the probability of a county's belonging to higher risk clusters.

5. Summary And Discussion

In this paper, we analyzed demographic, meteorological, geographical, and health factors to determine the critical parameters affecting the transmission and mortality level of COVID-19 in the US counties. Aggregated COVID-19 positive cases and death counts for each county were derived, and their rate with respect to population was calculated. Pearson Correlation was used to remove the collinearities. Sixteen features were kept for further analyses. Next, K-means clustering was applied to group the counties based

on COVID-19 positive rates and death rates. According to the Elbow method, three clusters were chosen, representing low-, medium-, and high-risk clusters. The levels obtained from clustering were then considered as the nominal dependent variable for classification.

Using cluster labels, several classification models were applied to the data. Random Forest and Multinomial logistic regression (MLR) models were finally chosen for a more accurate risk-level prediction and better interpretation of the effect of factors, respectively.

Seven factors of mean temperature, percent of people below poverty, percent of adults with obesity, air pressure, population density, wind speed and percent of people uninsured were found to influencing the risk level. Increasing the mean temperature, percent of people below poverty, percent of adults with obesity, wind speed, population density and percent of people uninsured would increase the probability that a county belongs to higher risk clusters. On the other hand, increasing the air pressure would decrease this probability.

Declarations

Author Contributions:

Conceptualization, Ziyadidegan. S., Razavi. M., and Pesarakli. H.; methodology, Ziyadidegan. S., Razavi. M., and Pesarakli. H., Javid. A.; software, Ziyadidegan. S., Razavi. M., and Pesarakli. H.; formal analysis, Ziyadidegan. S., Razavi. M., and Pesarakli. H., Javid. A.; data curation, Ziyadidegan. S., Razavi. M., and Pesarakli. H.; writing—original draft preparation, Ziyadidegan. S.; writing—review and editing, Razavi. M., and Pesarakli. H., Javid. A, Erraguntla. M. “All authors have read and agreed to the published version of the manuscript.”

Data Availability Statement:

Publicly available datasets were analyzed in this study. This data can be found here:

Data combined and used in this paper by authors:

https://github.com/SamiraZiyadg/COVID-19

Covid Data

https://github.com/nytimes/covid-19-data

Health Data

https://opendata.dc.gov/datasets/1044bb19da8d4dbfb6a96eb1b4ebf629_0?selectedAttribute=ADULT_ICU_BEDS

Demographic, Meteorological, and Location-based data

https://www.kaggle.com/johnjdavisiv/us-counties-covid19-weather-sociohealth-data

Acknowledgments:

The first three authors would thank Dr. Darren Homrighausen from Texas A&M University for his valuable feedback on the preliminary draft of this study when he was the instructor of STAT 656: Applied Analytics in 2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Velavan TP, Meyer CG (2020) The COVID-19 epidemic. Trop Med Int Heal 25:278–280. doi:10.1111/tmi.13383
2. World Health Organization *Coronavirus Disease – 2019 (COVID-19) Situation report-41*; 2020
3. Jahromi R, Avazpour A, Jahromi M, Alavi J (2020) Covid-19 With Positive Bronchoalveolar Lavage Fluid But Negative Nasopharyngeal and Oropharyngeal Swabs: a Case Report and Insights. Indian J Case Reports 6:380–382. doi:10.32677/ijcr.2020.v06.i07.010
4. Yousefzadegan S, Rezaei N (2020) Case report: Death due to COVID-19 in three brothers. Am J Trop Med Hyg 102:1203–1204. doi:10.4269/ajtmh.20-0240
5. The Washington Post Hundreds of young Americans have now been killed by the coronavirus, data shows Available online: https://www.washingtonpost.com/health/2020/04/08/young-people-coronavirus-deaths/ (accessed on Mar 8, 2021)
6. Centers for Disease Control and Prevention, C. First Travel-related Case of 2019 Novel Coronavirus Detected in United States
7. Saad B. Omer; Malani P, Rio C del (2020) The COVID-19 Pandemic in the US A Clinical Update. Clin Infect Dis 71:778–785. doi:10.1093/cid/ciaa310
8. John Hupkins University & Medicine Covid-19 Map Available online: https://coronavirus.jhu.edu/us-map (accessed on Nov 13, 2020)
9. 1point3acres Global Covid-19 Trackers and Interactive Charts Available online: https://coronavirus.1point3acres.com/en (accessed on Nov 13, 2020)
10. Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R (2020) COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. SSRN Electron J. doi:10.2139/ssrn.3590821
11. Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM COVID-19 outbreak prediction with machine learning. Algorithms 2020, 13, doi:10.3390/a13100249
12. Mallapaty S (2020) What the cruise-ship outbreaks reveal about COVID-19. Nature 580:18. doi:10.1038/d41586-020-00885-w
13. Sujath R, Chatterjee JM, Hassanien AE (2020) A machine learning forecasting model for COVID-19 pandemic in India. Stoch Environ Res Risk Assess 34:959–972. doi:10.1007/s00477-020-01827-8

14. Wang G, Minnis RB, Belant JL, Wax CL. Dry weather induces outbreaks of human West Nile virus infections. BMC Infect Dis 2010, 10, doi:10.1186/1471-2334-10-38
15. Chien LC, Chen LW (2020) Meteorological impacts on the incidence of COVID-19 in the U.S. Stoch Environ Res Risk Assess 34:1675–1680. doi:10.1007/s00477-020-01835-8
16. Wu X, Lu Y, Zhou S, Chen L, Xu B (2016) Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. Environ Int 86:14–23. doi:10.1016/j.envint.2015.09.007
17. Chen PS, Tsai FT, Lin CK, Yang CY, Chan CC, Young CY, Lee CH (2010) Ambient influenza and avian influenza virus during dust storm days and background days. Environ Health Perspect 118:1211–1216. doi:10.1289/ehp.0901782
18. Pica N, Bouvier NM (2012) Environmental factors affecting the transmission of respiratory viruses. Curr Opin Virol 2:90–95. doi:10.1016/j.coviro.2011.12.003
19. Lowen AC, Mubareka S, Steel J, Palese P (2007) Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathog 3:1470–1476. doi:10.1371/journal.ppat.0030151
20. Liu J, Zhou J, Yao J, Zhang X, Li L, Xu X, He X, Wang B, Fu S, Niu T et al (2020) Impact of meteorological factors on the COVID-19 transmission: A multi-city study in China. Sci Total Environ 726:138513. doi:10.1016/j.scitotenv.2020.138513
21. Moges Menebo M (2020) Temperature and precipitation associate with Covid-19 new daily cases: A correlation study between weather and Covid-19 pandemic in Oslo, Norway. Sci Total Environ 737:139659. doi:10.1016/j.scitotenv.2020.139659
22. Ma Y, Zhao Y, Liu J, He X, Wang B, Fu S, Yan J, Niu J, Zhou J, Luo B (2020) Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Sci Total Environ 724:138226. doi:10.1016/j.scitotenv.2020.138226
23. He Z, Chin Y, Yu S, Huang J, Zhang CJP, Zhu K, Azarakhsh N, Sheng J, He Y, Jayavanth P et al (2021) The Influence of Average Temperature and Relative Humidity on New Cases of COVID-19: Time-Series Analysis. JMIR Public Heal Surveill 7:e20495. doi:10.2196/20495
24. Pahuja S, Madan M, Mittal S, Pandey RM, Nilima; Madan K, Mohan A, Hadda V, Tiwari P, Guleria R (2021) Weather Parameters and COVID-19: A Correlational Analysis. J Occup Environ Med 63:69–73. doi:10.1097/JOM.0000000000002082
25. Coccia M (2020) How do low wind speeds and high levels of air pollution support the spread of COVID-19? Atmos Pollut Res. doi:10.1016/j.apr.2020.10.002
26. Patanavanich R, Glantz SA (2020) Smoking is associated with COVID-19 progression: A meta-analysis. Nicotine Tob Res 22:1653–1656. doi:10.1093/ntr/ntaa082
27. Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J, Li Q, Jiang C, Zhou Y, Liu S et al (2020) Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. J Infect 81:e16–e25. doi:10.1016/j.jinf.2020.04.021
28. Franch-Pardo I, Napoletano BM, Rosete-Verges F, Billa L (2020) Spatial analysis and GIS in the study of COVID-19. A review. Sci Total Environ 739:140033. doi:10.1016/j.scitotenv.2020.140033

29. Jüni P, Rothenbühler M, Bobos P, Thorpe KE, Da Costa BR, Fisman DN, Slutsky AS, Gesink D (2020) Impact of climate and public health interventions on the COVID-19 pandemic: A prospective cohort study. Cmaj 192:E566–E573. doi:10.1503/cmaj.200920
30. Walrand S, Autumn (2021) COVID-19 surge dates in Europe correlated to latitudes, not to temperature-humidity, pointing to vitamin D as contributing factor. Sci Rep 11:1–9. doi:10.1038/s41598-021-81419-w
31. Jordan RE, Adab P, Cheng KK (2020) Covid-19: Risk factors for severe disease and death. BMJ 368:1–2. doi:10.1136/bmj.m1198
32. Bansal M, Cardiovascular disease, COVID-19. Diabetes Metab Syndr Clin Res Rev 2020, 14, 247–250, doi:10.1016/j.dsx.2020.03.013
33. Zoabi Y, Deri-Rozov S, Shomron N (2021) Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digit Med 4:1–5. doi:10.1038/s41746-020-00372-6
34. Shoukat A, Wells CR, Langley JM, Singer BH, Galvani AP, Moghadas SM (2020) Projecting demand for critical care beds during COVID-19 outbreaks in Canada. Cmaj 192:E489–E496. doi:10.1503/cmaj.200457
35. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2020) SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 16:321–357. doi:https://doi.org/10.1613/jair.953
36. United States Census Bureau Available online: https://www.census.gov (accessed on Nov 11, 2020)
37. Kodinariya TM, Makwana PR (2013) Review on determining number of Cluster in K-Means Clustering. Int J Adv Res Comput Sci Manag Stud 1:2321–7782
38. Purnima B, Arvind K, EBK-Means: (2014) A Clustering Technique based on Elbow Method and K-Means in WSN. Int J Comput Appl 105:17–24
39. Han H, Guo X, Yu H Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS* 2016, *0*, 219–224, doi:10.1109/ICSESS.2016.7883053
40. United States Census Bureau How the Census Bureau Measures Poverty Available online: https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html#:~:text=If a family's total income,it is considered in poverty.&text = The official poverty definition uses,Medicaid%2C and food stamps (accessed on Mar 29, 2021)
41. Ogden CL, Fryar CD, Martin CB, Freedman DS, Carroll MD, Gu Q, Hales CM (2020) Trends in obesity prevalence by race and hispanic origin – 1999–2000 to 2017–2018. JAMA - J Am Med Assoc 324:1208–1210. doi:10.1001/jama.2020.14590
42. Islam N, Bukhari Q, Jameel Y, Shabnam S, Erzurumluoglu AM, Siddique MA, Massaro JM, D'Agostino RB (2021) COVID-19 and climatic factors: A global analysis. Environ Res 193:110355. doi:10.1016/j.envres.2020.110355
43. Jahromi R, Mogharab V, Jahromi H, Avazpour A (2020) Synergistic effects of anionic surfactants on coronavirus (SARS-CoV-2) virucidal efficiency of sanitizing fluids to fight COVID-19. Food Chem Toxicol 145:111702. doi:10.1016/j.fct.2020.111702

44. Ramesh N, Siddaiah A, Joseph B (2020) Tackling Corona Virus Disease 2019 (COVID 19) in Workplaces. Indian J Occup Environ Med 24:16–18. doi:10.4103/ijoem.IJOEM
45. Little C, Alsen M, Barlow J, Naymagon L, Tremblay D, Genden E, Trosman S, Iavicoli L, van Gerwen M (2021) The Impact of Socioeconomic Status on the Clinical Outcomes of COVID-19; a Retrospective Cohort Study. J Community Health. doi:10.1007/s10900-020-00944-3
46. Patel JA, Nielsen FBH, Badiani AA, Assi S, Unadkat VA, Patel B, Ravindrane R, Wardle H Poverty, inequality and COVID-19: the forgotten vulnerable. 2020
47. Centers for Disease Control and Prevention, Overweight C & Obesity Available online: https://www.cdc.gov/obesity/data/obesity-and-covid-19.html (accessed on Mar 4, 2021)
48. ThoughtCo Air Pressure and How It Affects the Weather
49. Takagi H, Kuno T, Yokoyama Y, Ueyama H, Matsushiro T, Hari Y, Ando T Higher Temperature, Pressure, and Ultraviolet Are Associated with Less COVID-19 Prevalence – Meta-Regression of Japanese Prefectural Data. *medRxiv* 2020, doi:10.1101/2020.05.09.20096321
50. Centers for Disease Control and Prevention, C. How to Protect Yourself & Others
51. Moein Razavi; Hamed Alikhani; Vahid Janfaza; Benyamin Sadeghi; Ehsan Alikhani An Automatic System to Monitor the Physical Distance and Face Mask Wearing of construction Workers in COVID-19 Pandemic. 2021, 7
52. Skórka P, Grzywacz B, Moroń D, Lenda M (2020) The macroecology of the COVID-19 pandemic in the Anthropocene. PLoS One 15:1–17. doi:10.1371/journal.pone.0236856
53. Keshavarzi A (2020) Coronavirus Infectious Disease (COVID-19) Modeling: Evidence of Geographical Signals. SSRN Electron J. doi:10.2139/ssrn.3568425

Figures

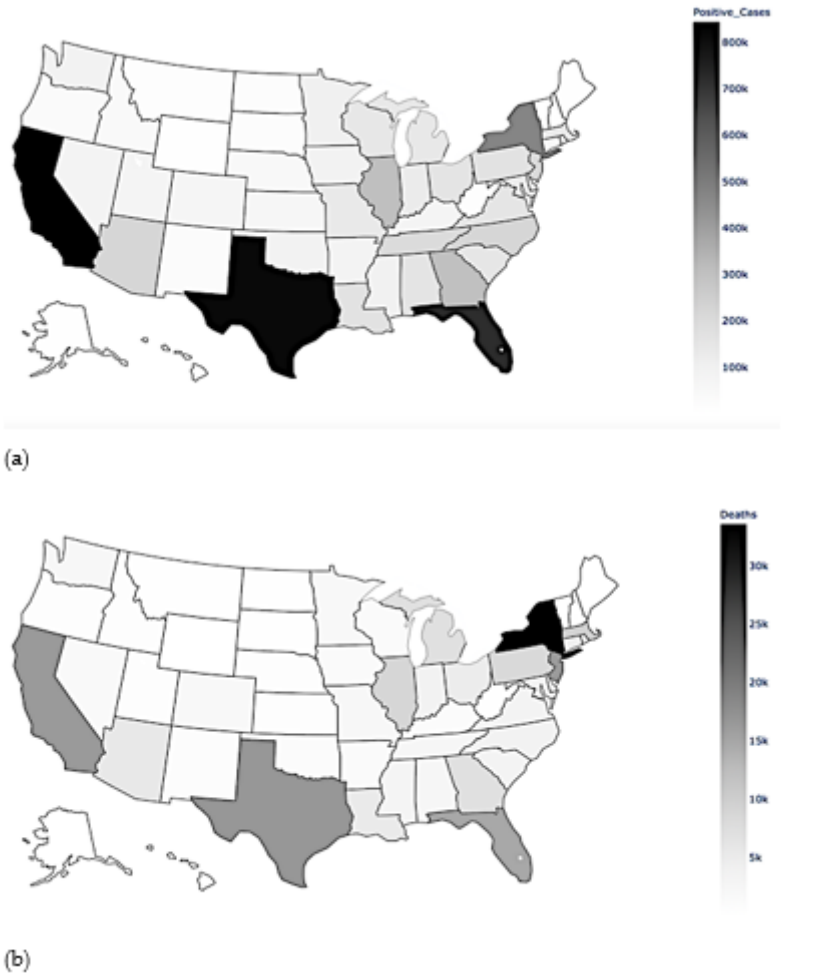

Figure 1

COVID-19 positive case and death counts maps for all states of the US, (a) COVID-19 positive case counts map, (b) COVID-19 Death counts map

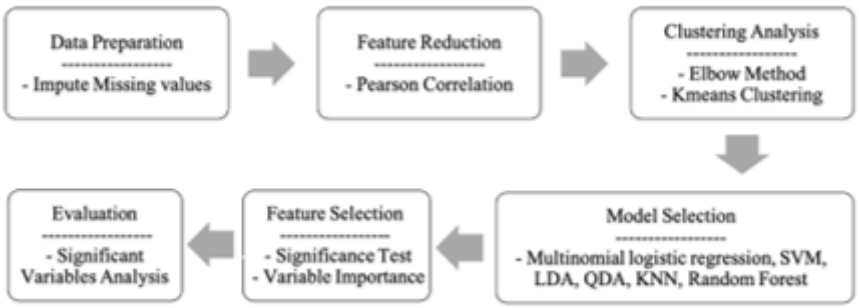

Figure 2

Methodology used in the study

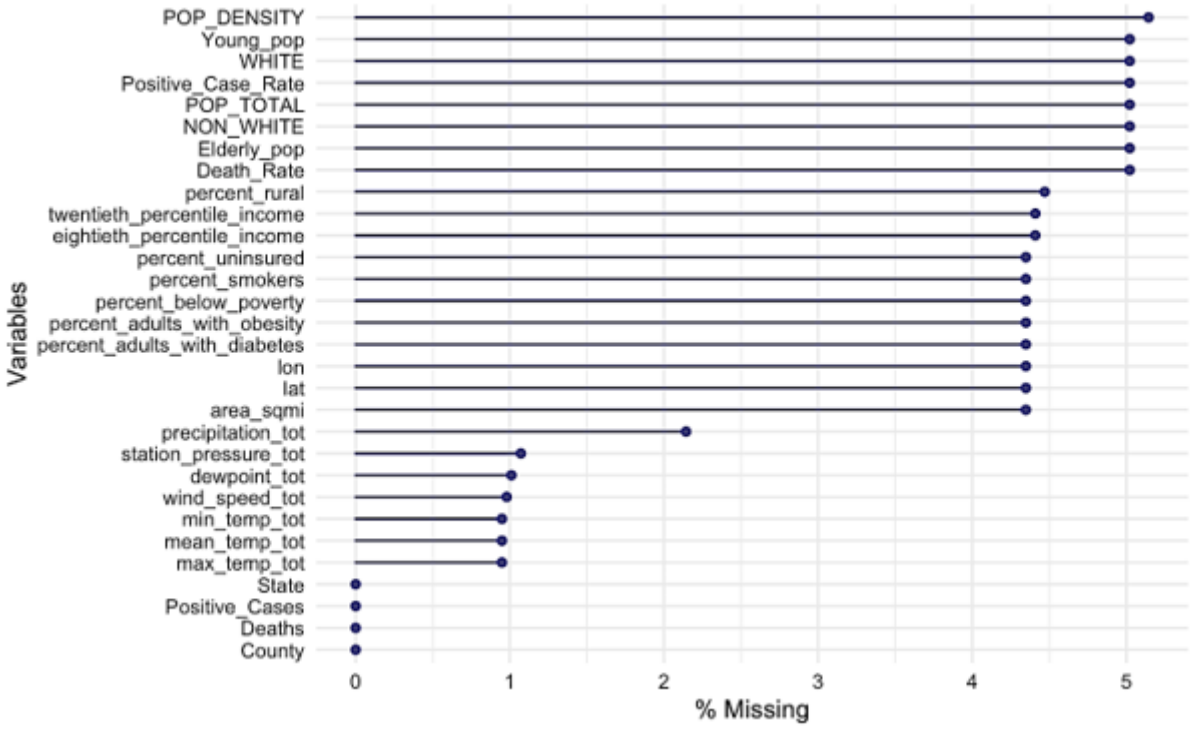

Figure 3

The percentage of missing values plot

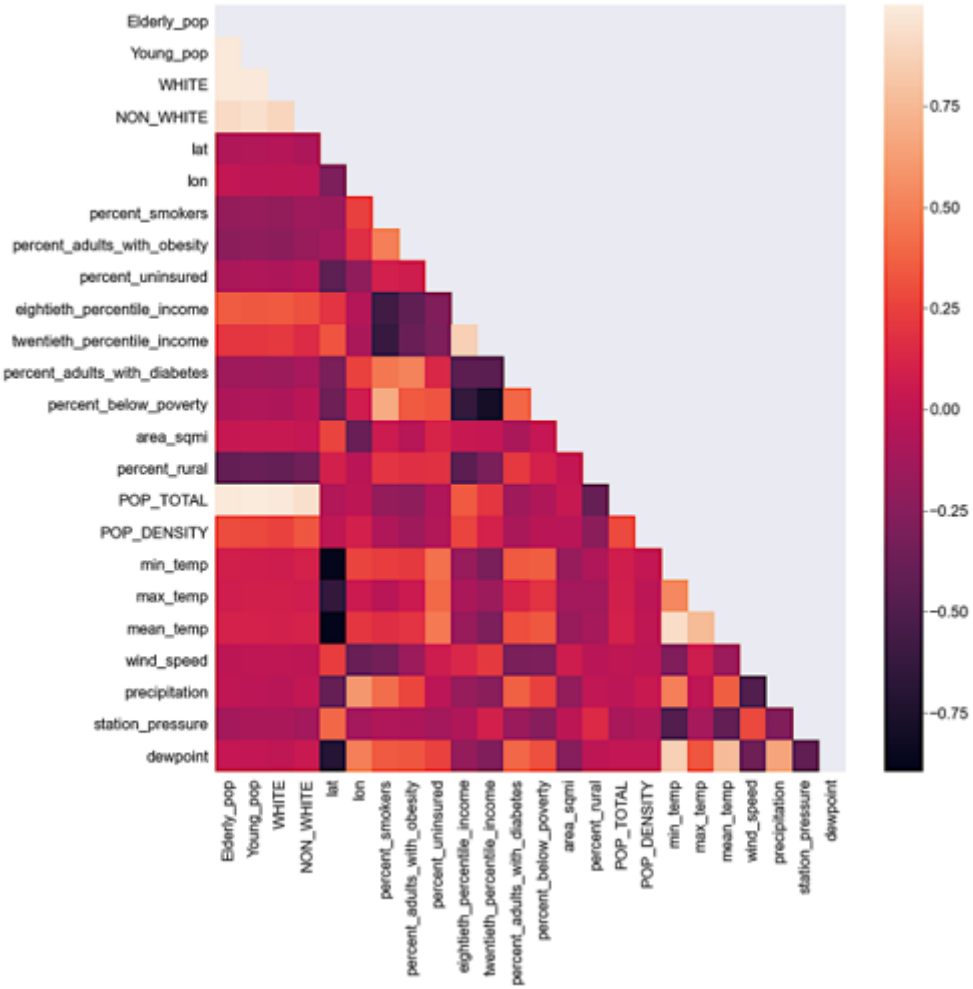

Figure 4

Correlation matrix for all parameters

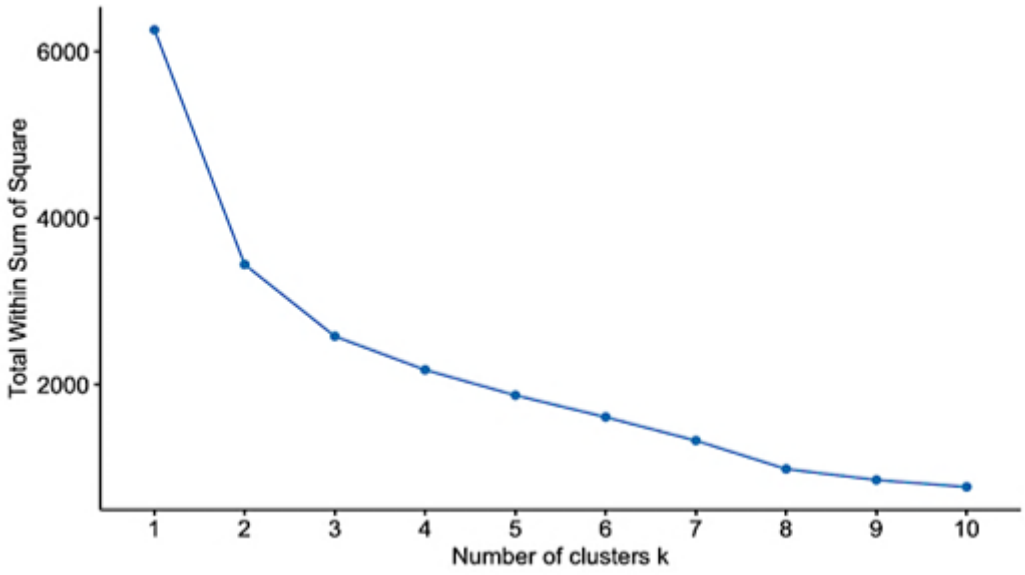

Figure 5

Elbow Method

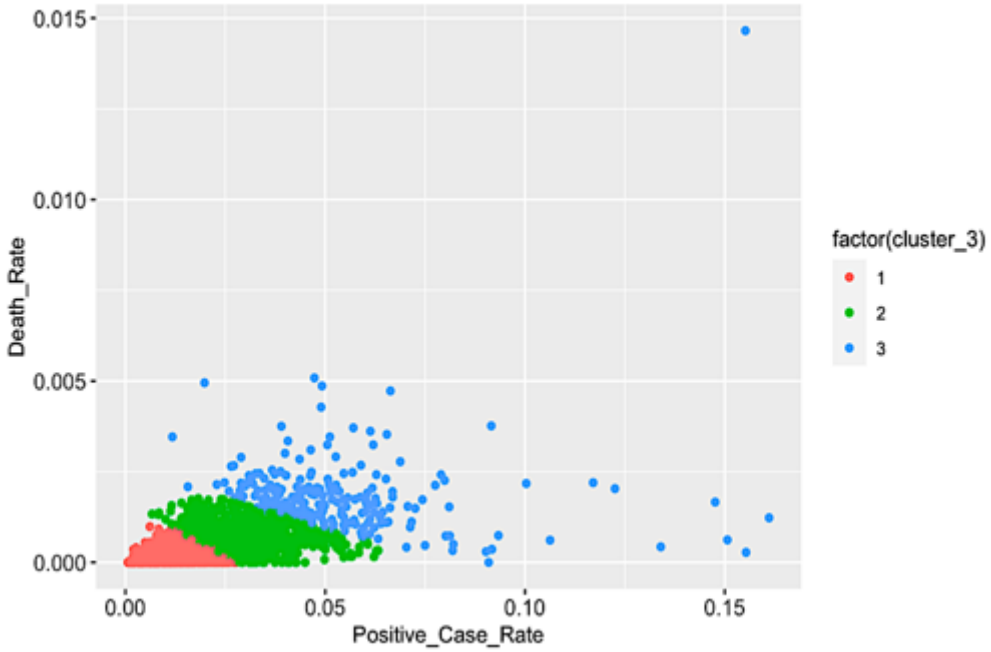

Figure 6

Clustering output plot

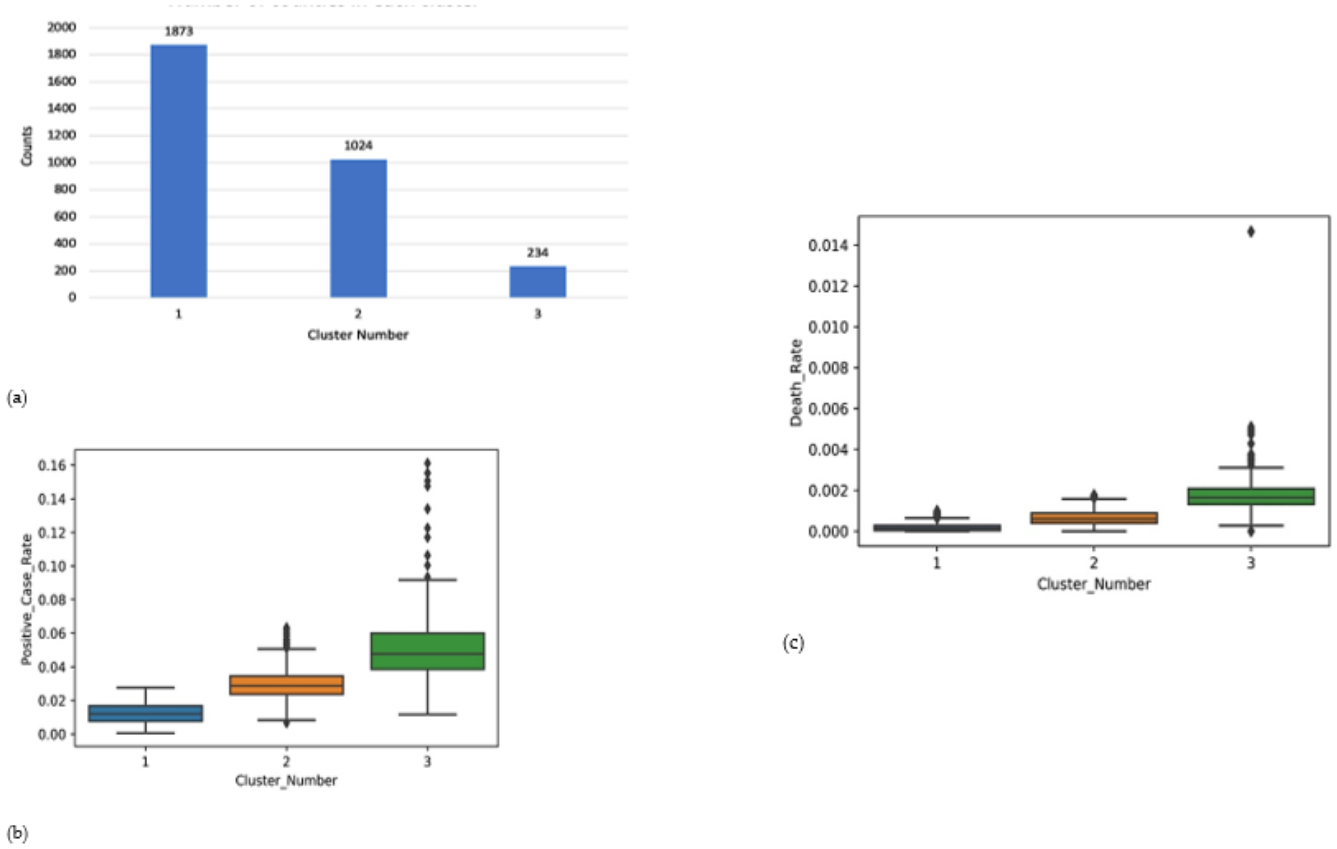

Figure 7

Clustering Results: (a) Number of counties in each cluster; (b) Positive rate; (c) Death Rate

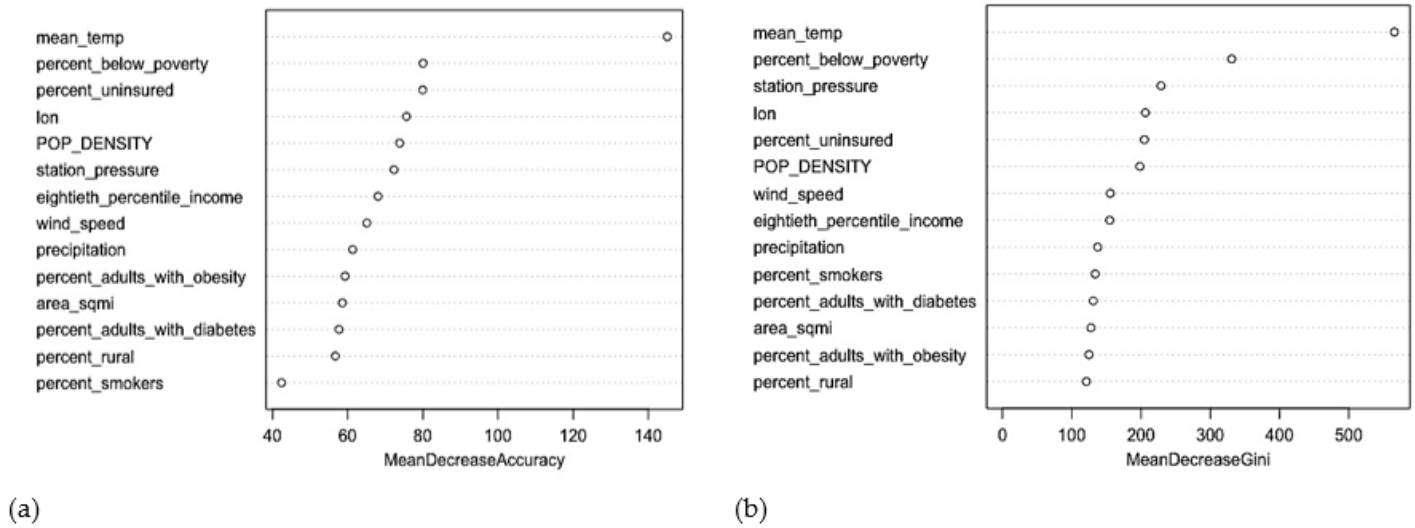

Figure 8

Feature importance plots for Random Forest model by (a) The Mean Decrease Accuracy (b) Mean Decrease in Gini