Proceedings of the IISE Annual Conference Expo 2022
K. Ellis, W. Ferrell, J. Knapp, eds.

Prediction of community transmission level of Covid-19 using machine learning algorithms based on CDC Social Vulnerability Index

Saviz Saei, Yibin Wang, Mohammad Marufuzzaman, Nazanin Morshedlou, Haifeng Wang
Department of Industrial and Systems Engineering
Mississippi State University, Mississippi State, MS 39762

Abstract

Response to hazardous events is crucial in every community, whether they are natural or anthropogenic disasters. CDC Social Vulnerability Index (SVI) helps people who need support. Social vulnerability refers to the number of adverse effects on external stress include natural causes or disease outbreaks like the Coronavirus Disease 19 (Covid-19) pandemic on human health. The SVI dataset possesses California state of the US, subdivisions of counties of 15 features into four groups as related themes (i.e., socioeconomic status; household composition and disability; minority status and language; and housing type and transportation). In addition to SVI dataset, the recent Covid-19 data tracker for each county is posted by CDC shows the new cases per 100,000 persons in the last seven days. The transmission values are classified into low, moderate, substantial, and high. The impact of SVI on Covid-19 attracts the attention of researchers to find the relationships between SVI and Covid-19 incidence. This paper aims to incorporate SVI data and the incidence in the United States using ten machine learning algorithms for Covid-19 transmission level classification. The experimental results show the proper prediction based on the community transmission level of Covid-19 by considering the features of SVI. Based on the percentage of various performance metrics accuracy, precision, and recall, random forest achieved the best performance.

Keywords
Social Vulnerability Index (SVI); Coronavirus Disease 19(Covid19); Machine learning algorithms; Prediction

1. Introduction
Coronavirus Disease 19(Covid19) followed by SARS-Cov-2 identified in December 2019 in the Wuhan of China. This virus, which originated from bats [1], spread to one another through contact and droplet, airborne, and fomite transmissions [2]. Since January, the virus has rapidly spread across the globe, and the earliest COVID-19 case came from Wuhan was reported in Snohomish County, Washington on the 19th of January [3], and the earliest COVID-19 death reported in Santa Clara County, California on 6th of February [4].The virus continued and up until June 30th, there was 2,618,817 cases and 126,623 deaths associated with COVID-19 in the US [5]. Researchers have begun to examine the spatial patterns and underlying risk factors of COVID-19. Mollalo et al. (2020) found that place-based factors like median household income, income inequality, percentage of nurse practitioners, and percentage of Black female population explain significant variation in COVID-19 incidence. Others have shown strong evidence of the spatial effects of COVID-19 with county-level socioeconomic factors in neighboring counties influencing incidence in the US [6].This paper aims to find a prediction machine learning method in the disruption scenarios of SVI to transmission level of Covid19. The remainder of this paper is organized as follows. Section 2 addresses the relevant studies that discuss SVI and Covid19 focusing on Machine learning techniques.Next, the details of each technique for modeling are presented. Section 4 summarizes experimental results, and section 5, the research findings of this work and future research are discussed.

2. Literature Review
From 2014 to 2018 (5 years), American Community Survey(ACS) divided the vulnerability into four categories as Themes: "Socioeconomic", "Household Composition and Disability", "Minority Status and Language", and "Housing

Type and Transportation". The overall vulnerability 1 based on this definition is: "The degree to which a community exhibits certain social conditions may affect that community's ability to prevent human suffering and financial loss in the event of a disaster"[7]. Though the primary goal of SVI, Covid19 is one of the disasters as a disease outbreak that influence the community ability. However, SVI failed to sufficiently determine vulnerability for the unprecedented situation, and CDC foundation developed COVID-19 Community Vulnerability Index (CCVI) [8] to rectify vulnerability assessment.

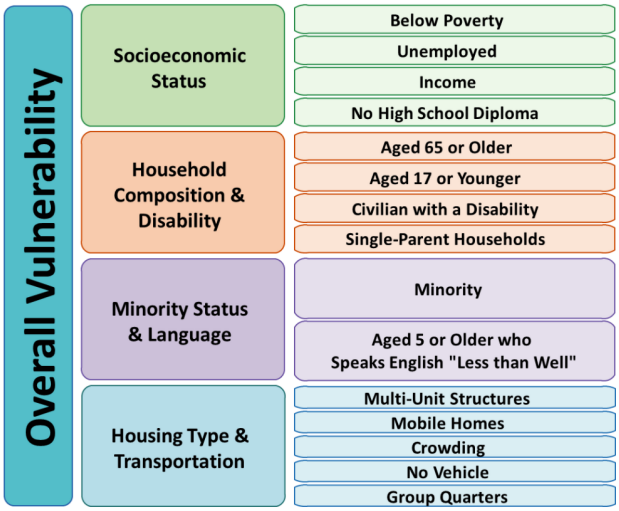

Figure 1: Overall vulnerability

Wylezinski[9] merged county-level COVID-19 testing data with COVID-19 vaccination rates and SDOH information to predict COVID-19 incidence for each Tennessee county with An ensemble of generalized linear and tree-based machine learning models. Tiwari[10] proposed a Random Forest machine learning based on vulnerability model using CDC's socio-demographic and COVID-19-specific themes for identifying and mapping vulnerable counties. This paper use machine learning techniques for leveraging the preparedness of vulnerable counties to reduce the COVID-19 burden in California State [11].

3. Methodology

3.1 Data and Prepossessing

In this paper, we focus on one of the states of the United States, California, with 58 counties and 602,315 observations to find out the effect of SVI features on community transmission level.

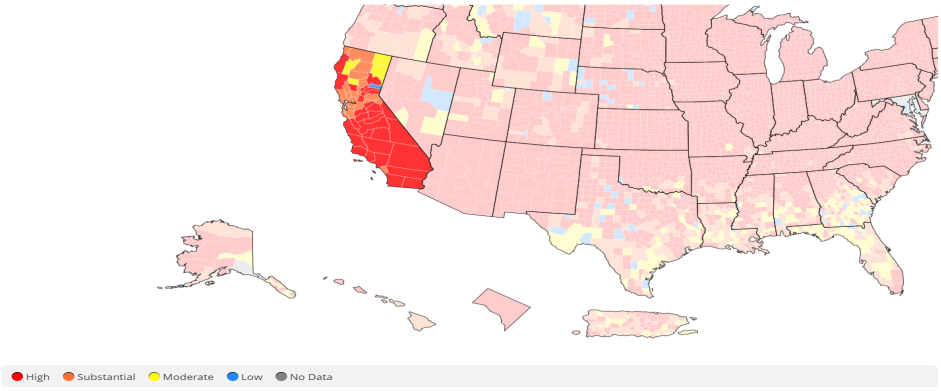

Figure 2: Level of community transmission (high:0, low:1, moderate: 2, substantial: 3)

At first, four summary theme ranking variables, which are RPL-Theme1(Socioeconomic), RPL-Theme2(Household

Composition and Disability), RPL-Theme3(Minority Status and Language), and RPL-Theme4(Housing Type and Transportation) are considered as an independent variables which contains all 15 variables [7]. This dataset has missing values that we screened out any variables with more than %45 missing values imputed with the median. For analyzing Covid 19 dataset, the data of 3 months of August, September, and October 2021 are merged with The SVI per each county, and the factor variable of "community-transmission-level" is converted to numerical based on high:0, low:1, moderate: 2, substantial: 3. This problem is a classification problem because of categorical (or discrete) responses (Figure 2). This data splits to 80% for training and 20% for testing to predict the model with machine learning techniques (Figure 3).

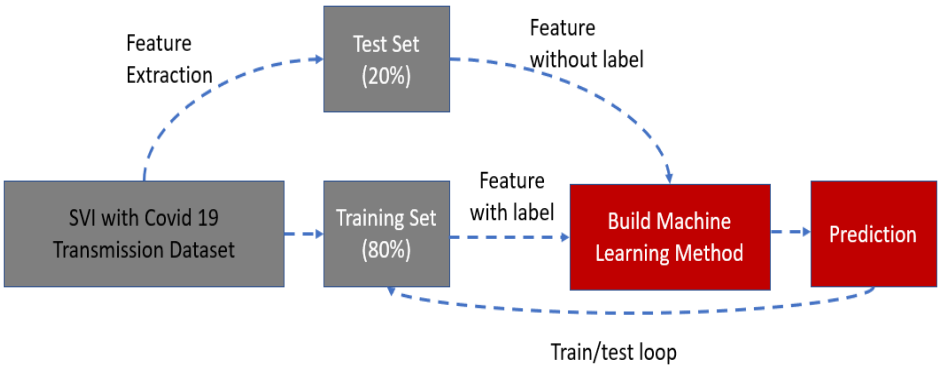

Figure 3: Architecture of the model

3.2 Machine learning Techniques

To find out the impact of SVI data on transmission level of Covid-19, ten machine learning methods are measured and compared which some of them are detailed in this section. The model performance of each methods shows the metrics of Precision, Recall, F1-score calculated using confusion matrix(True positives (TP), False positives (FP), True negatives (TN), False negatives (FN). The definition of these metrics are:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1-score = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

3.2.1 Random Forest

Random Forest is a supervised classification method based on combination of bagging[12], which grow each tree in a random selection of features, without replacement [13].The procedure of random forest is a classifier which consists of a collection of tree classifiers as $h(x,\Theta_k), k = 1,...$ where a random vector of Θ_k is generated from *kth* tree with the same distribution of $\Theta_1,...,\Theta_{k-1}$ and independent from the past random vectors. This results in a large number of trees, and each tree casts a unit vote for the most popular class at the input is x [12] (Figure 4) [14].

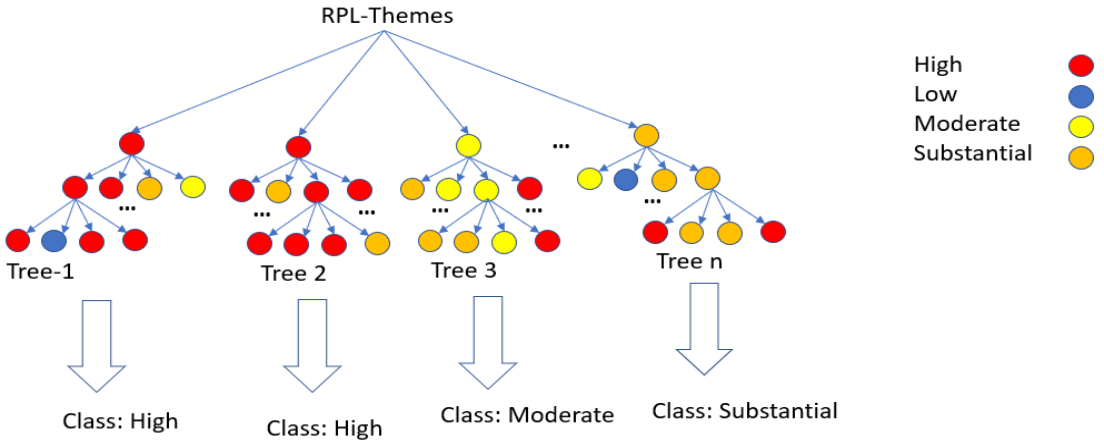

Figure 4: Random Forest model

3.2.2 Support Vector Classification(SVC)

Support Vector Classification(SVC) [15] is a type of supervised machine learning technique for classification applied in real-life scenarios. The problem of training SVCs attracts the attention of researchers because of the change of data set sizes from hundreds to millions. SVC training defines a hyper-plane to bounder the training data into the determined classes. However, the performance of SVC depends on kernel function, which can do complex data transformation [16].There are different kernel functions such as Linear, polynomial, Gaussian, Radial Basis Function (RBF), and Sigmoid. For this problem, linear, poli, RBF, and Sigmoid Kernels are implemented by Scikit-learn library in python.

3.2.3 Logistic Regression

A logistic regression model calculates the class membership probability for one of the two categories in the data set $P(1|x,\alpha) = 1 - \frac{1}{1+e^{-}(\alpha.x)}$ and $P(0|x,\alpha) = 1 - P(1|x,\alpha)$. The complexity of the model is already low, especially if interaction terms and variable transformations are used with little or no interaction. Thus, over-fitting is less of an issue. For this problem, x is defined by RPL-Themes for four categories of Community transmission level.

3.2.4 Adaboost

Adaptive boosting (AdaBoost) is one of the machine learning algorithms formulated by Freund and Schapire [17]. AdaBoost combines with weak classifiers by weighted average to build a learning algorithm with stronger classifiers. e method is used for combination.

The general form of adaboost is:

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)) \tag{5}$$

Where f is the weak learner and theta is the weight associated with the specific weak classifier. For this problem, x is defined by RPL-Themes for four categories of Community transmission level.

3.3 Experimental Results

As table 1 shows, support vector Classification algorithms achieved the best test accuracy of 0.99 with the Radial Basis Function (RBF) kernel, and it has a great potential for predicting the effect of SVI on COVID-19 results. linear kernel performed well in classification with 0.986 test accuracy, and The sigmoid kernel is not applicable according to its low-performance measurements(0.79). In addition, Random Forest has a great potential for predicting the effect

of SVI on COVID-19 considering features of Random Forest n-estimators, random-state, and bootstrap are 2000,1, and False respectively. The results of test accuracy shows 100 which is highlighted in the table; however, AdaBoost is the lowest performance among all 10 methods with an accuracy of 13%. In addition, AdaBoost is sensitive to noise data, and it is highly affected by outliers that try to fit each point perfectly. More analysis on the result of AdaBoost illustrates that different n-trees of 10, 50, 100, 500,1000, and 5000 has 65% accuracy, however with increasing learning rate from 1.9 to 2, the accuracy of the model decrease from 65% to 13% suddenly. This shows that this method is not reliable to predict SVI on the Covid-19 transmission level.

Table 1: Performance measure of the model

Methods	accuracy	Performance	precision	recall	f1-score
Random Forest	**1**	macro avg	**1**	**1**	**1**
		weighted avg	**1**	**1**	**1**
Support Vector Classification	**0.99**	macro avg	**0.98**	**0.97**	**0.97**
		weighted avg	**0.99**	**0.99**	**0.99**
Logistic Regression	0.66	macro avg	0.29	0.26	0.24
		weighted avg	0.59	0.66	0.57
Adaboost	**0.13**	macro avg	0.32	0.35	0.12
		weighted avg	0.69	0.13	0.12
Naive Bayes	0.65	macro avg	0.28	0.26	0.24
		weighted avg	0.58	0.65	0.57
GradBoost	0.66	macro avg	0.41	0.28	0.28
		weighted avg	0.60	0.66	0.58
Knn	0.65	macro avg	0.42	0.34	0.36
		weighted avg	0.62	0.65	0.63
BaggingKnn	0.67	macro avg	0.61	0.31	0.32
		weighted avg	0.63	0.67	0.64
BernoulliRBM	0.67	macro avg	0.61	0.31	0.32
		weighted avg	0.63	0.67	0.64
xgboost	0.67	macro avg	0.55	0.29	0.29
		weighted avg	0.61	0.67	0.60

4. Conclusion

In this paper, we analyzed the dataset of SVI with community-transmission-level of Covid-19 in California state of the United States to find the relationships between community transmission level of Covid-19 and the features of SVI. The most suitable algorithm for each metric is predicted by ten optimizers namely Random Forest, Support Vector Classification, Logistic Regression, Adaboost, Naive Bayes, GradBoost, Knn, BaggingKnn, BernoulliRBM, and xgboost. Among all, the Random Forest optimizer has outperformed the remaining with an accuracy of 100%. This Analysis can be used for all of States of the United States.The important direction for future work is the predict the community transmission level of Covid19 with the updated SVI factors.

References

[1] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, 579(7798):270–273, 2020.

[2] World Health Organization et al. Transmission of sars-cov-2: implications for infection prevention precautions: scientific brief, 09 july 2020. Technical report, World Health Organization, 2020.

[3] Michelle L Holshue, Chas DeBolt, Scott Lindquist, Kathy H Lofy, John Wiesman, Hollianne Bruce, Christopher Spitters, Keith Ericson, Sara Wilkerson, Ahmet Tural, et al. First case of 2019 novel coronavirus in the united states. *New England Journal of Medicine*, 2020.

[4] Eran Bendavid, Bianca Mulaney, Neeraj Sood, Soleil Shah, Rebecca Bromley-Dulfano, Cara Lai, Zoe Weissberg, Rodrigo Saavedra-Walker, Jim Tedrow, Andrew Bogan, et al. Covid-19 antibody seroprevalence in santa clara county, california. *International journal of epidemiology*, 50(2):410–419, 2021.

[5] USA Facts. Detailed methodology and sources: Covid-19 data. *https://usafacts.org/articles/detailed-methodology-covid-19-data*, (50):410–419, 2020.

[6] Christopher F Baum and Miguel Henry. Socioeconomic factors influencing the spatial spread of covid-19 in the united states. *Miguel, Socioeconomic Factors influencing the Spatial Spread of COVID-19 in the United States (May 29, 2020)*, 2020.

[7] Cdc svi documentation 2018. https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html.

[8] Surgo Foundation. The covid-19 community vulnerability index (ccvi). *Surgo Foundation*, 2020.

[9] Lukasz S Wylezinski, Coleman R Harris, Cody N Heiser, Jamieson D Gray, and Charles F Spurlock. Influence of social determinants of health and county vaccination rates on machine learning models to predict covid-19 case growth in tennessee. *BMJ health & care informatics*, 28(1), 2021.

[10] Anuj Tiwari, Arya V Dadhania, Vijay Avin Balaji Ragunathrao, and Edson RA Oliveira. Using machine learning to develop a novel covid-19 vulnerability index (c19vi). *Science of The Total Environment*, 773:145650, 2021.

[11] Center for diseases control and prevention. https://data.cdc.gov/Public-Health-Surveillance/United-States-COVID-19-County-Level-of-Community-T/8396-v7yb/data.

[12] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] Carmen Lai, Marcel JT Reinders, and Lodewyk Wessels. Random subspace method for multivariate feature selection. *Pattern recognition letters*, 27(10):1067–1076, 2006.

[14] Random forest simple explanation. https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d. Accessed: 2018-04-18.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[16] Alex J Smola, Bernhard Schölkopf, and Klaus-Robert Müller. The connection between regularization operators and support vector kernels. *Neural networks*, 11(4):637–649, 1998.

[17] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.