Original Paper

Exploring Socioeconomic Status as a Global Determinant of COVID-19 Prevalence, Using Exploratory Data Analytic and Supervised Machine Learning Techniques: Algorithm Development and Validation Study

Luke Winston[1], MA; Michael McCann[1], PhD; George Onofrei[2], PhD

[1]Department of Computing, Atlantic Technological University, Letterkenny, Ireland

[2]Department of Business, Atlantic Technological University, Letterkenny, Ireland

Corresponding Author:
Luke Winston, MA
Department of Computing
Atlantic Technological University
Port Road
Letterkenny, F92 FC93
Ireland
Phone: 353 862435617
Email: L00162644@student.lyit.ie

Abstract

Background: The COVID-19 pandemic represents the most unprecedented global challenge in recent times. As the global community attempts to manage the pandemic in the long term, it is pivotal to understand what factors drive prevalence rates and to predict the future trajectory of the virus.

Objective: This study had 2 objectives. First, it tested the statistical relationship between socioeconomic status and COVID-19 prevalence. Second, it used machine learning techniques to predict cumulative COVID-19 cases in a multicountry sample of 182 countries. Taken together, these objectives will shed light on socioeconomic status as a global risk factor of the COVID-19 pandemic.

Methods: This research used exploratory data analysis and supervised machine learning methods. Exploratory analysis included variable distribution, variable correlations, and outlier detection. Following this, the following 3 supervised regression techniques were applied: linear regression, random forest, and adaptive boosting (AdaBoost). Results were evaluated using k-fold cross-validation and subsequently compared to analyze algorithmic suitability. The analysis involved 2 models. First, the algorithms were trained to predict 2021 COVID-19 prevalence using only 2020 reported case data. Following this, socioeconomic indicators were added as features and the algorithms were trained again. The Human Development Index (HDI) metrics of life expectancy, mean years of schooling, expected years of schooling, and gross national income were used to approximate socioeconomic status.

Results: All variables correlated positively with the 2021 COVID-19 prevalence, with R^2 values ranging from 0.55 to 0.85. Using socioeconomic indicators, COVID-19 prevalence was predicted with a reasonable degree of accuracy. Using 2020 reported case rates as a lone predictor to predict 2021 prevalence rates, the average predictive accuracy of the algorithms was low (R^2=0.543). When socioeconomic indicators were added alongside 2020 prevalence rates as features, the average predictive performance improved considerably (R^2=0.721) and all error statistics decreased. Thus, adding socioeconomic indicators alongside 2020 reported case data optimized the prediction of COVID-19 prevalence to a considerable degree. Linear regression was the strongest learner with R^2=0.693 on the first model and R^2=0.763 on the second model, followed by random forest (0.481 and 0.722) and AdaBoost (0.454 and 0.679). Following this, the second model was retrained using a selection of additional COVID-19 risk factors (population density, median age, and vaccination uptake) instead of the HDI metrics. However, average accuracy dropped to 0.649, which highlights the value of socioeconomic status as a predictor of COVID-19 cases in the chosen sample.

Conclusions: The results show that socioeconomic status is an important variable to consider in future epidemiological modeling, and highlights the reality of the COVID-19 pandemic as a social phenomenon and a health care phenomenon. This paper also

puts forward new considerations about the application of statistical and machine learning techniques to understand and combat the COVID-19 pandemic.

(JMIR Form Res 2022;6(9):e35114) doi: 10.2196/35114

KEYWORDS

COVID-19; machine learning; data analysis; epidemiology; human development index

Introduction

Background

The COVID-19 pandemic represents the most unprecedented global challenge in recent times. Originally identified in the city of Wuhan, China, the SARS-CoV-2 virus spread across the world, and the situation escalated into an international emergency. Despite widescale containment efforts in 2020, as well as the largest vaccine rollout in history [1], the pandemic continued to challenge the global community in 2021. Research is being conducted to analyze the trajectory of the virus, and to understand why particular populations or countries have been more severely impacted than others [2,3]. This has been supported by increases in data availability, which has enabled researchers to investigate a large range of potential COVID-19 risk factors. These risk factors can be categorized as clinical or nonclinical. Clinical risk factors include obesity [4-6], diabetes [7,8], and smoking [9]. Examples of nonclinical risk factors are cultural differences [10], government containment measures [11], vaccination attitudes [12], and socioeconomic status [13-15].

This paper focuses on the nonclinical risk factor of socioeconomic status as a determinant of COVID-19 prevalence. To provide a reliable empirical metric for socioeconomic status, the Human Development Index (HDI) of the United Nations Development Programme (UNDP) was selected. The HDI calculates the overall socioeconomic status or "well-being" of inhabitants in a country by aggregating its life expectancy, education, and per capita income metrics [16]. It has been applied successfully in previous epidemiological research to map prevalence rates of various diseases [17-20]. Despite its popularity in statistical analysis, the HDI has not yet been widely applied in machine learning COVID-19 modeling. This presents an opportunity to apply statistical and machine learning techniques to examine whether the HDI can be used to accurately predict prevalence rates of COVID-19.

Related Work

Socioeconomic Status in Health Research

Pandemics are as much a social problem as a health care problem [21]. As such, socioeconomic status is an important determinant to consider in pandemic research. The term socioeconomic status is an umbrella term used to describe empirically measurable social or economic factors, such as social class, education, income, and health [22,23]. These factors are applied in a variety of ways to investigate or control their effects on given outcomes, such as health outcomes, and have consistently been found to be statistically significant [24-26]. In terms of health outcomes, higher socioeconomic status has typically been associated with better health. Conversely, lower socioeconomic status is associated with poorer health outcomes [27]. In the literature, lower socioeconomic status has been associated with higher rates of illnesses, such as osteoarthritis, chronic diseases, hypertension, and cervical cancer [28,29].

In relation to COVID-19, socioeconomic status has also been associated with higher prevalence and more severe outcomes. In the United States, the Distressed Communities Index has been used to analyze the impact of socioeconomic status on COVID cases and mortality [30]. Results from this study indicated that lower education and racial differences were associated with poorer COVID-19 outcomes. Another study argued that lower socioeconomic populations are more likely to live in overcrowded accommodation and have less access to outdoor space, making them more vulnerable to COVID-19 infection [31]. Evidently, socioeconomic status is an important determinant of COVID-19 outcomes, which can uncover how the virus affects particular populations.

HDI

The HDI is a composite measure of overall socioeconomic status at the national level, which is annually calculated by the UNDP. The HDI indices include life expectancy, expected years of schooling, mean years of schooling, and gross national income (GNI). Calculating a country's HDI for a given year requires 2 steps. First, values from each of the 4 indices are normalized to an index value between 0 and 1. Maximum and minimum limits for each metric are set by the UNDP. Using the actual value, maximum value, and minimum value, the dimension index can be calculated with the following formula:

$$\text{Dimension index} = (\text{actual value} - \text{minimum value}) / (\text{maximum value} - \text{minimum value})$$

Second, once each individual dimension has been calculated, the equally weighted mean is calculated to provide the overall HDI score of a country [32].

The HDI has been used in health research to analyze both the prevalence rates and mortality rates of specific diseases, which helps to identify disparities in terms of outcome within a country or between countries. It has been applied to understand a range of epidemiological research problems, such as malaria [17], various cancer distributions [19,33,34], hypertension [20], *Blastocystis* parasites [35], and dental health [36]. To provide a specific example, research investigating the relationship between the HDI and thyroid cancer suggested that although higher HDI countries have a higher prevalence of the disease, lower HDI countries have higher mortality rates [34].

The HDI has also been applied to analyze the ongoing COVID-19 pandemic, generating important insights about the disproportionate impact of the pandemic cross-nationally. For example, a study analyzing the HDI and COVID-19 mortality

reported that countries with high HDI scores recorded higher COVID-19 mortality rates [13]. Another study reported significant correlations between the HDI scores of 166 countries and their confirmed cases on March 27, 2020 [14]. Elsewhere, a study focusing on municipal differences in COVID-19 impact in Brazil (using a recalibrated index to analyze municipal differences rather than national differences) found that municipalities with high HDI scores had the highest COVID-19 incidence rate and mortality per 100,000 population as of May 2020 [15]. The index has therefore been recognized as a valuable framework in COVID-19 research.

Multicountry COVID-19 Research

Multicountry COVID-19 research is important for the following 2 reasons: (1) the ability to identify country-specific points of interest, and (2) the ability to uncover common trends or risk factors across countries. In a study of lockdown-associated mental health problems in Egypt, Pakistan, India, Ghana, and the Philippines, it was reported that although lockdowns negatively affected the mental health of respondents in each country, they did so in different ways. For example, respondents from the Philippines coped with lockdowns by increasing self-destructive behaviors, while those from Pakistan sought comfort in religion. Respondents from the 3 remaining countries tended to accept the lockdowns [37]. A similar study in a larger sample of 101 countries analyzed the loneliness and social isolation associated with the pandemic [38]. Other studies have been conducted to analyze cross-national vaccination attitudes [39], the success of containment measures [11,40], and cultural behaviors that impacted cross-national COVID-19 mortality rates [10]. Therefore, multicountry COVID-19 research helps to identify "global risk factors" relating to the pandemic, subsequently aiding evidence-based public health interventions [38]. It also opens up new research questions as to why certain populations behaved or were impacted a certain way during the pandemic.

Modeling Outbreaks Using Machine Learning

When modeling outbreaks, a popular method in epidemiology is the susceptible, infected, recovered (SIR) approach. The SIR approach simplifies the transmission dynamics of infectious diseases by dividing the population into groupings of susceptible, infected, and recovered individuals and analyzes the interaction between these groups over the course of an outbreak. This method has also been deployed to analyze the COVID-19 pandemic [41,42]. However, SIR modeling assumes that complete herd immunity is possible through infection [43], which limits its efficacy in COVID-19 research. It is not yet understood if COVID-19 herd immunity is achievable due to the complex nature of the virus, the questionable long-term efficacy of available vaccines, the emergence of new variants, and the cases of reinfection [44]. Subsequently, the predictive benefits of machine learning may yield better results in relation to this pandemic.

Advancements in machine learning have enabled epidemiological researchers to use a robust data-driven approach facilitated by high-precision algorithms. This has helped to process ever-increasing volumes of data, and to analyze a wider range of factors that impact patient health outcomes [45,46]. For example, naïve Bayes, logistic regression, random forest, and artificial neural network models have been developed to predict hypotension in patients after receiving an anesthetic [47]. Elsewhere, gated recurrent unit neural networks have been designed to identify individuals at risk of in-hospital mortality. This model allows practitioners to map the probability of death longitudinally, and to provide targeted interventions based on the model predictions [48].

Another advantage of machine learning in epidemiology is that it can predict and map disease occurrences and health outcomes in situations where data are limited [49]. Specifically, boosted regression tree models have been used to analyze environmental factors that affect the transmission of diseases, such as dengue fever, Ebola, Crimean-Congo hemorrhagic fever, and Zika virus [50-53]. Another type of machine learning model, the Ensemble Adjustment Kalman Filter, has been used to forecast seasonal outbreaks of influenza [54]. Additionally, several retrospective forecasting studies have been conducted to reconstruct past pandemics, including Ebola, West Nile Virus, and Respiratory Syncytial Virus, by mapping their transmission patterns [55-57].

Regarding COVID-19, epidemiological research using machine learning is emerging in the literature at pace. Generally, studies have involved the design of one or more machine learning models to predict COVID-19 case prevalence [11,58,59], severity [60,61], and mortality/risk of mortality [62,63]. In 1 study, 5 non–time series supervised learning models using random forest and AdaBoost regression were trained to predict the confirmed infection growth (the 14-day growth of the cumulative number of reported COVID-19 cases) of COVID-19 in 114 countries, using nonpharmaceutical containment measures and cultural dimensions as features. Results indicated that confirmed infection growth was predicted to a considerable degree with moderate to high R^2 values (>0.50) [11]. Lastly, a systematic review of machine learning techniques in the prediction of COVID-19 cases found that R^2 values ranged between 0.64 and 1, suggesting that machine learning is a highly valuable method for predicting COVID-19 prevalence, which could support policy makers in shaping future interventions [64].

Description of the Study

This study analyzed the statistical relationship between HDI scores and cumulative COVID-19 cases (total recorded cases up to December 31, 2021) in a sample of 182 countries. It then attempted to predict 2021 COVID-19 cumulative cases in the sample using the previous year's cumulative cases (total recorded cases up to December 31, 2020) and HDI scores. Cumulative cases per million of the population was selected as it provides the number of reported infections proportionate to the population size. Crude rate metrics, such as cases per million, are the most effective for multicountry samples [65]. For example, Afghanistan and Albania reported a similar absolute number of COVID-19 cases in 2020, with values of 51,526 and 58,316, respectively. However, Afghanistan's cases per million was 1324, while Albania's was 20,264. This shows the viral prevalence relative to both populations and indicates that Albania actually had higher case rates in 2020.

To measure socioeconomic status, the HDI indices of life expectancy, expected years of schooling, mean years of schooling, and GNI were used. For the purposes of this study, individual metrics were selected rather than the aggregated HDI value. This approach was used because aggregation can lose important information in the data, which can lead to less accurate predictions [66].

Two predictive models were designed using the open-source integrated development environment Jupyter Notebook, which is compatible with Python programming language. Each model was trained using the following 3 supervised learning regression algorithms: basic linear regression, random forest, and AdaBoost. All algorithms were evaluated using k-fold cross-validation and then compared by calculating their R^2 scores and error statistics. The first model attempted to predict 2021 COVID-19 prevalence using 2020 case numbers to establish a baseline for the performance of the second model. The second model included 2020 case numbers and each country's life expectancy, expected years of schooling, mean years of schooling, and GNI metrics. Due to the uneven progress of the pandemic on a country-by-country basis, this study focused on cross-sectional data rather than time-series data. All data for this study are secondary and publicly available, highlighting the commendable global effort to collect and share data concerning the pandemic.

Methods

Data Preprocessing

COVID-19 case data were downloaded from the COVID-19 OurWorldInData database [65], which in turn retrieves data from the John Hopkins Center for Systems Science and Engineering Data Repository. The OurWorldInData database contains comprehensive COVID-19 metrics for 190 countries, including infection rates, hospitalization numbers, mortality rates, and vaccination uptake figures. Data are uploaded daily, which allows users to track the evolution of the pandemic with up-to-date statistics. This research required each country's "cases per million" figure for December 31, 2020, and the same metric for December 30, 2021. HDI data were extracted from the 2020 Human Development Report Data Center [67]. The report provides each country's overall HDI score and the score for each individual metric.

These data sets were combined so that each observation (country) contained the following metrics: (1) life expectancy, (2) expected years of schooling, (3) mean years of schooling, (4) GNI, (5) COVID-19 cases per million in 2020 (January 1-December 31), and (6) COVID-19 cases per million in 2021 (January 1-December 31).

Countries with missing data were omitted; therefore, the final data set contained data for 182 countries. It was then imported to Jupyter and converted into dataframe format (see Table 1) to begin analysis.

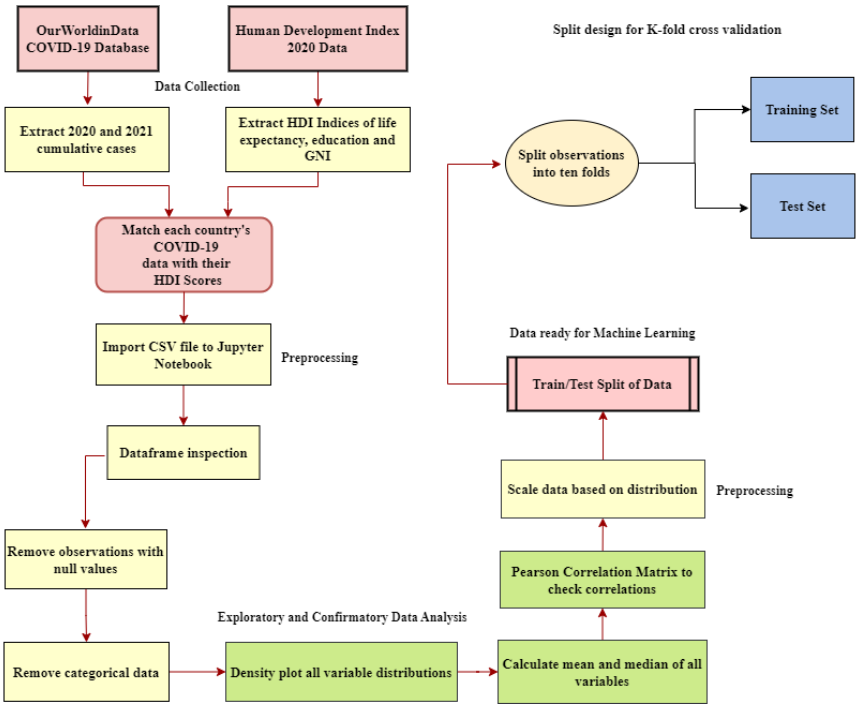
Following this, exploratory data analysis was conducted to explore the distribution of the data and the statistical relationships between the variables. A data scaling method was then selected depending on the distribution of the data. Data scaling is important in machine learning modeling as it prevents measurement differences from negatively affecting the final results [68]. The interquartile range was then calculated to identify outliers in the target variable (2021 COVID-19 cases).

Figure 1 summarizes the workflow for this study, from data preprocessing to model design and exploratory data analysis.

Table 1. Sample of the data set using Human Development Index metrics and COVID-19 cases.

Country	Life expectancy	Expected years of schooling	Mean years of schooling	Gross national income per capita (US$)	Cases 2020 (per million)	Cases 2021 (per million)
Afghanistan	64.8	10.2	3.9	2239	1323.612	3968.427
Albania	78.6	14.7	10.1	13,998	20,264.091	73,173.975
Algeria	76.9	14.6	8.0	11,174	2271.554	4895.753
Andorra	81.9	13.3	10.5	56,000	104,173.947	306,900.742
Angola	61.2	11.8	5.2	6104	534.073	2404.489

Figure 1. A flowchart illustrating the data pipeline, from the collection of COVID-19 and Human Development Index (HDI) data to the cross-validation training and testing process. In addition to designing the predictive models, exploratory data analysis was also conducted to identify trends in the data set. GNI: gross national income.



Machine Learning Algorithm Selection

Supervised machine learning models are trained to make predictions by learning from a data set where the value of the output (dependent variable) is known for each observation. Supervised machine learning produces decisions or "outputs" based on input data during the training process. Implementing different supervised algorithms on a set of data allows for the results to be compared and for the best fitting model to be identified [69,70]. Evaluating a supervised learning model requires robust validation measures [71]. These can be calculated using a variety of accuracy and error metrics, such as the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), or max error. This research compared the performances of linear regression, random forest, and AdaBoost supervised techniques.

Linear Regression

Linear regression is one of the most common machine learning algorithms [72]. Regression in machine learning differs from traditional statistical regression as it partitions the data set into a training set and a test set. Using the input and output data from the training set, algorithms attempt to predict output data in the test set using input data only. This process indicates how accurately a model can make predictions on new data. Linear regression is calculated as follows:

$$y = a_0 + a_1x + \varepsilon$$

where y is the target variable (output), x is the predictor variable (input), a_0 is the intercept, a_1 is the coefficient, and ε is random error.

Random Forest

Random forest is an ensemble of decision tree algorithms that can be used for either classification or regression problems. It is based on the concept of bagging or bootstrap aggregation, which creates an ensemble of learner trees [73]. Each learner tree (K) is trained on separate samples drawn from the original data set (input vector x), and the overall prediction is obtained by calculating the mean of K regression trees as follows:

$$\frac{1}{K}\sum_{k=1}^{K} h_k(x)$$

Random forest is beneficial for reducing model variance compared to individual decision trees. It also helps to prevent model overfitting (when a model fits too closely to training data and poorly to test data) [74].

AdaBoost

AdaBoost or adaptive boosting is a sequential ensemble technique that is based on the principle of developing several weak learners using different training subsets drawn randomly from the original training data set. Using this technique, the training algorithm begins with 1 decision tree, identifies the observations with the highest error, and adds more weight to these. The weights are recalculated after every iteration so that incorrectly classified observations by the previous decision tree receive higher weights [75]. Using Python programming language, the number of trees that the algorithm will deploy can be chosen, with the default set at 50 iterations.

Model Design and Evaluation

Two feature models were created (Feature Model 1 and Feature Model 2). Feature Model 1 was trained to predict 2021 COVID-19 prevalence using 2020 cases only. Feature Model

2 was trained to predict 2021 COVID-19 prevalence using 2020 case data as well as life expectancy, expected years of schooling, mean years of schooling, and GNI per capita. Each feature model was trained using linear regression, random forest, and AdaBoost techniques. Hyperparameters were set for each algorithm, and results were evaluated using a 10-fold (k=10) k-fold cross-validation.

Model Hyperparameters and Validation

Rather than partitioning the data into training and test sets using the train/test split, this research used k-fold cross-validation. K-fold cross-validation has a single parameter called *k* that represents the number of subsets or "folds" that a data set will be split into, which the user selects. As shown in Figure 2, each fold uses a different grouping of data as the test set, and the process is then repeated *k* number of times (for example, 5 times in Figure 2). It is evaluated by the cross-validation score, which is the mean of all scores from each k-fold subset. K-fold cross-validation provides a more generalizable and less biased performance estimate when working with smaller data sets [76,77]. This is because it maximizes the number of observations that can be used for both training and testing. In other words, a model using cross-validation does not depend on a single train/test split.

Using sklearn, the mean cross-validation score defaults to the scoring metric for the specific algorithm being cross-validated. For each algorithm in this study, the default scoring metric was the coefficient of determination (R^2). Therefore, the mean cross-validation score computed was the average R^2 for each algorithm across all k-folds. R^2 represents the goodness of fit of a regression model and explains how much variance in the dependent variable can be explained by one or more independent variables. It is calculated by dividing the residual sum of squares by the total sum of squares and subtracting the derivation from 1, as follows:

> R^2 = 1 – (residual sum of squares / total sum of squares)

R^2 was the primary measure under observation in this study. In machine learning, R^2 is the most informative validation measure with the least interpretive limitations [78].

Table 2 presents the hyperparameters unique to each algorithm. A 10-fold validation was selected for the k-fold cross-validation, which is a generally recommended number of subsets to apply [76,77].

Alongside R^2, 4 error metrics were also calculated to assess performance. First, MAE provides the average of the absolute error between the predicted values and true values. It is calculated as follows:

$$MAE = \frac{\sum_{i=1}^{n} |yi - xi|}{n}$$

where y_i is the prediction value, x_i is the actual value, and n is the number of observations.

Second, MSE measures the average squared difference between the predicted values and true values. It is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where n is the number of data points, Y_i is the actual value, and $\hat{Y}_i$ is the predicted value.

Third, RMSE calculates the square root of the mean of squared errors of a model. It is calculated as follows:

$$RMSE = \frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}$$

where i is the variable i, N is the number of data points, x_i is the actual value, and $x\hat{}_i$ is the predicted value.

Finally, max error computes the maximum residual error, which captures the worst case error between the predicted value and the true value. It is calculated as follows:

$$Max\ error\ (y, \hat{y}) = max(|y_i - \hat{y}i|)$$

where $\hat{y}$ is the predicted value of the i-th sample, and y_i is the corresponding true value.

Figure 2. An example of the 5-fold k cross-validation method where k=5. The overall accuracy score is calculated as the mean value of each fold's accuracy score.

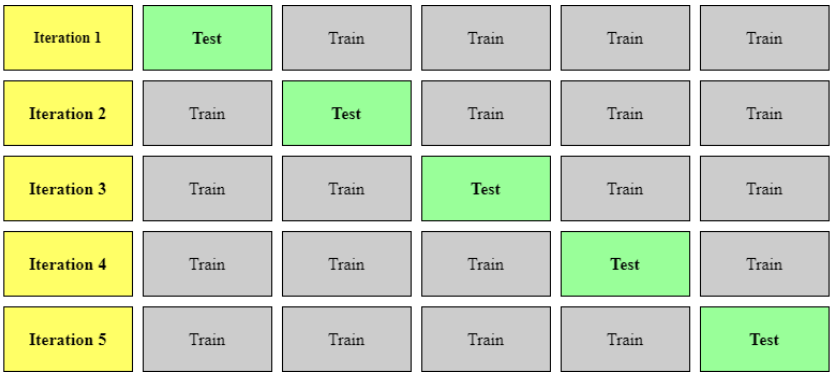


Table 2. Supervised learning model hyperparameters using cross-validation.

Algorithm	Hyperparameters
Basic linear regression	Folds: 10; random state: 1
Random forest	Folds: 10; random state: 1; estimators: 100
AdaBoost	Partitions: 10; estimators: 50; random state: 0

Results

Exploratory Data Analysis

Exploratory data analysis was carried out to identify and visualize trends in the data, and to statistically analyze the variables. In 2020, the mean number of COVID-19 cases per million in the sample was 15,880.41, with a median of 6822.98. In 2021, the mean number of COVID-19 cases per million was 64,479.58, with a median of 50,764.73. Table 3 presents the key descriptive statistics of all variables in the study.

Distplots were created to inspect the distribution of all variables. The resulting plots showed that all variables, with the exception of expected years of schooling, were skewed in the sample. The distribution of 2021 COVID-19 prevalence was positively skewed in the sample (see Figure 3). Calculation of the interquartile range revealed that 4 countries (Andorra, Montenegro, Serbia, and Seychelles) were statistical outliers, which had recorded unusually high rates of COVID-19 (>250,000 per million population). The Seychelles recorded the highest prevalence with 217,096.35 cases per million.

To investigate the statistical relationship between the features and the target variable, a Pearson correlation matrix was implemented (see Figure 4). All chosen features correlated statistically with 2021 COVID-19 prevalence, with R values ranging from 0.55 to 0.85. Moreover, 2020 COVID-19 cases had the strongest correlation with 2021 case data (R=0.85), followed by mean years of schooling (R=0.66), life expectancy (R=0.61), expected years of schooling (R=0.58), and GNI (R=0.55).

Table 3. Statistical measurements (mean and median) of all variables in the study.

Variable	Mean value	Median value
2020 COVID-19 cases per million	15,880.41	6822.98
2021 COVID-19 cases per million	64,479.58	50,764.73
Life expectancy	72.72	74.20
Expected years of schooling	13.31	13.15
Mean years of schooling	8.63	8.95
Gross national income per capita (US$)	20,453.40	13,112.50

Figure 3. A series of density plots illustrating the distribution of each variable under observation (the target variable). The target variable 2021 COVID-19 cases per million is right-skewed in the sample. Expected years of schooling is the only variable with a normal distribution in the sample. CASES_2020: 2020 COVID-19 cases per million; CASES_2021: 2021 COVID-19 cases per million; EXP_SCHOOLING: expected years of schooling; GNI: gross national income per capita; LIFE_EXP: life expectancy; MEAN_SCHOOLING: mean years of schooling.

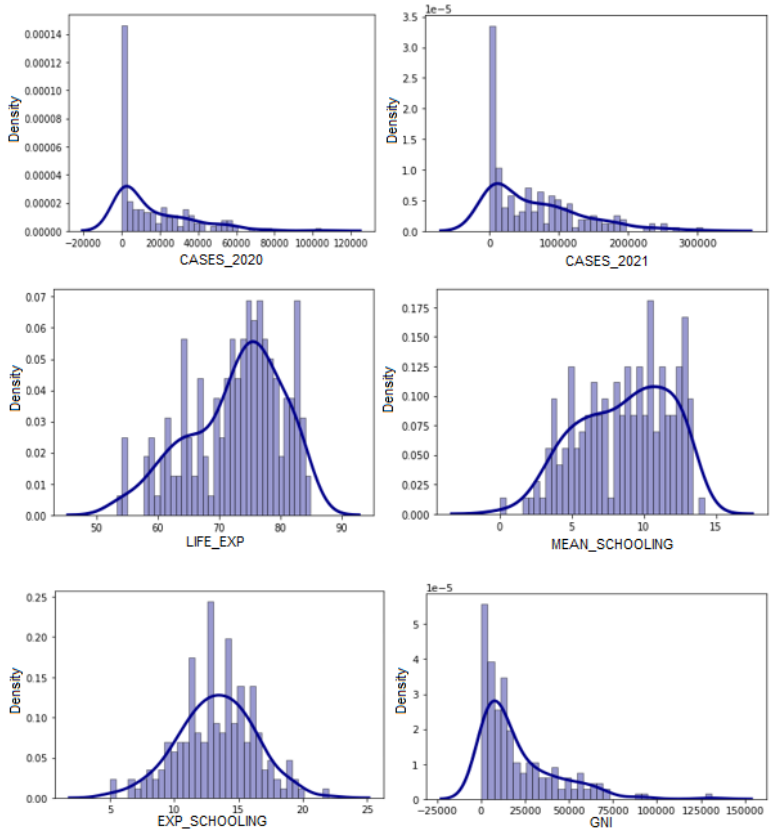
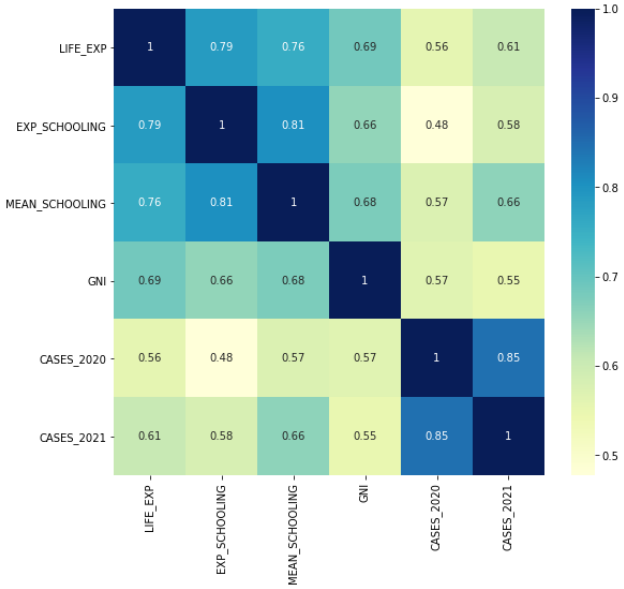


Figure 4. Pearson correlation matrix mapping the correlation between all variables. Results show that all features have a statistical correlation with 2021 COVID-19 cases. CASES_2020: 2020 COVID-19 cases per million; CASES_2021: 2021 COVID-19 cases per million; EXP_SCHOOLING: expected years of schooling; GNI: gross national income per capita; LIFE_EXP: life expectancy; MEAN_SCHOOLING: mean years of schooling.



Supervised Learning Models

Tables 4 and 5 summarize the performances of all regression algorithms in both feature models, while Figure 5 visualizes their performances. Feature Model 1 was trained to predict 2021 COVID-19 cases per million using 2020 cases per million (n=182). Feature Model 2 was trained to predict 2021 COVID-19 cases per million using 2020 cases per million as well as life expectancy, mean years of schooling, expected years of schooling, and GNI (n=182). Both data sets were divided into 10 folds for cross-validation (k=10).

In Feature Model 1, linear regression was the most accurate learner with a mean R^2 of 0.693, followed by random forest (0.481) and then AdaBoost (0.454). The variation in performance was considerable, with a 23.9% difference between the most precise and least precise algorithms. In Feature Model 2, the basic linear regression model was also the strongest learner (R^2=0.762), followed by random forest (0.722) and AdaBoost (0.679). The MAE, MSE, RMSE, and max error statistics of the algorithms were all lower in Feature Model 2 than in Feature Model 1. Feature Model 2 also exhibited closer performances between the algorithms than Feature Model 1, with the strongest learner being 8.4% more precise than the least.

Although it was the best learner on the data in both models, linear regression showed the least improvement with the inclusion of socioeconomic indicators in Feature Model 2 (R^2 improved by 7%). Additionally, its error statistics did not improve as significantly as those of random forest or AdaBoost. For example, the MAE of linear regression decreased by 0.009 (0.079 in Feature Model 1 and 0.070 in Feature Model 2) compared to decreases of 0.026 in random forest and 0.014 in AdaBoost.

Tables 6 and 7 outline the performance accuracy of each individual fold. The widely varying R^2 scores indicate that the cross-validation approach used in this study yielded the most reliable results.

Table 4. Evaluation of Feature Model 1 using linear regression, random forest, and AdaBoost.

Evaluation measure	Linear regression[a]	Random forest[a]	AdaBoost[a]
R^2	0.693	0.481	0.454
MAE[b]	0.079	0.096	0.104
MSE[c]	0.014	0.021	0.020
RMSE[d]	0.117	0.143	0.142
Max error	0.315	0.359	0.355

[a]All results were evaluated using k-fold cross-validation (k=10).

[b]MAE: mean absolute error.

[c]MSE: mean squared error.

[d]RMSE: root mean squared error.

Table 5. Evaluation of Feature Model 2 using linear regression, random forest, and AdaBoost.

Evaluation measure	Linear regression[a]	Random forest[a]	AdaBoost[a]
R^2	0.763	0.722	0.679
MAE[b]	0.070	0.070	0.090
MSE[c]	0.011	0.013	0.015
RMSE[d]	0.107	0.114	0.124
Max error	0.265	0.308	0.300

[a]All results were evaluated using k-fold cross-validation (k=10).

[b]MAE: mean absolute error.

[c]MSE: mean squared error.

[d]RMSE: root mean squared error.

Figure 5. A series of subplots showing the predictive performances of the linear regression, random forest, and AdaBoost algorithms in both Feature Models 1 and 2. Each observation represents a prediction of 2021 COVID-19 cumulative cases per million, with the regression line being the true value. With the addition of Human Development Index metrics, the linear regression algorithm improved from R^2=0.693 to 0.763. The random forest algorithm improved from R^2=0.481 to 0.722. The AdaBoost algorithm improved from R^2=0.454 to 0.679. Data points were calculated using cross_val_predict, which shows the predicted output from each test set within each k fold.

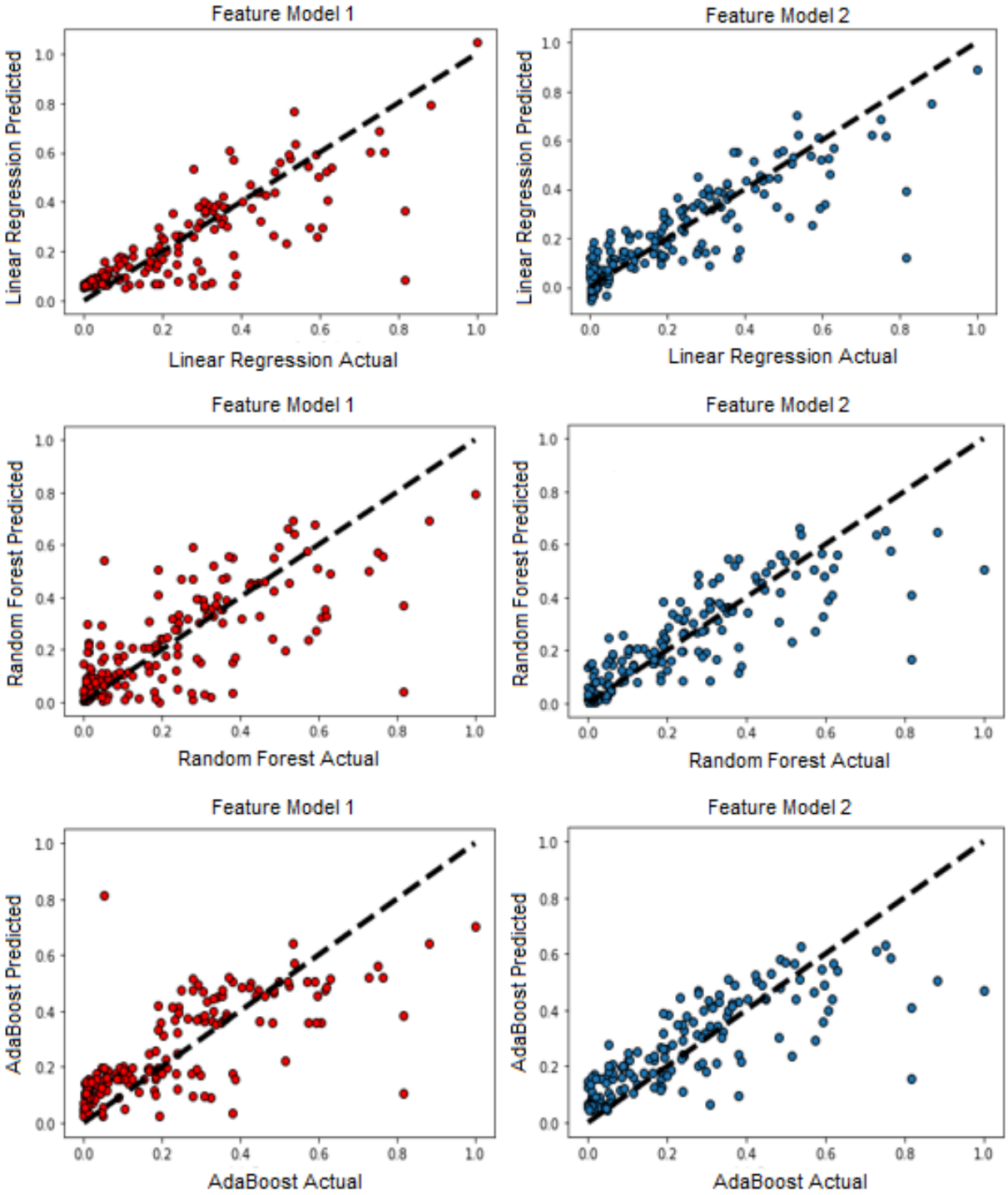


Table 6. Accuracy for each algorithm's individual fold (k=10) in Feature Model 1.

Iteration	Linear regression	Random forest	AdaBoost
Fold 1	0.877	0.799	0.759
Fold 2	0.768	0.687	0.342
Fold 3	0.657	0.464	0.584
Fold 4	0.803	0.530	0.629
Fold 5	0.747	0.153	-0.696
Fold 6	0.733	0.553	0.766
Fold 7	0.804	0.628	0.652
Fold 8	0.035	-0.287	0.083
Fold 9	0.767	0.627	0.696
Fold 10	0.742	0.657	0.722

Table 7. Accuracy for each algorithm's individual fold (k=10) in Feature Model 2.

Iteration	Linear regression	Random forest	AdaBoost
Fold 1	0.774	0.796	0.679
Fold 2	0.595	0.457	0.485
Fold 3	0.946	0.907	0.882
Fold 4	0.602	0.622	0.551
Fold 5	0.833	0.869	0.824
Fold 6	0.780	0.776	0.720
Fold 7	0.627	0.636	0.626
Fold 8	0.850	0.659	0.536
Fold 9	0.780	0.794	0.851
Fold 10	0.844	0.594	0.629

Discussion

Principal Findings

Results from exploratory data analysis yielded a number of interesting insights. First, the positively skewed distribution of 2021 COVID-19 cases resulted in a mean greater than the median in the sample. In the 182 countries sampled, COVID-19 prevalence was asymmetrical and revealed that a minority of countries recorded very high case numbers. Second, the distribution of 2020 COVID-19 cases was positively skewed and similar visually to the 2021 distribution. This shows that the trajectory of the virus in the sample was relatively consistent in 2020 and 2021 in terms of cumulative reported cases. Third, the 4 outlier countries identified shared an interesting pattern; all had higher than average life expectancy, mean years of schooling, and GNI compared with the means in the sample. This indicates that the outliers can be considered above average socioeconomically. Finally, all HDI metrics correlated positively with COVID-19 cases per million, which points to an important statistical relationship between socioeconomic status and COVID-19 prevalence. Education (expected/mean years) shared the highest correlation, followed by life expectancy and then GNI. This correlation is noteworthy and highlights the unique nature of the COVID-19 pandemic. Typically, lower socioeconomic status is associated with poorer health outcomes, but the results from this study suggest that countries with higher socioeconomic status recorded higher rates of COVID-19 in 2021. This could be because more developed countries tend to have older populations, as well as higher prevalence of known COVID-19 clinical risk factors, such as diabetes and cardiovascular disease [79].

The results from machine learning analysis suggest that 2021 COVID-19 prevalence could be predicted with a reasonable degree of accuracy using the previous year's prevalence rates and the socioeconomic indicators of life expectancy, mean years of schooling, expected years of schooling, and GNI per capita. With socioeconomic indicators included, the R^2 of each learning algorithm was higher than that when trained on only 2020 COVID-19 data, and the error statistics were lower. Including the HDI indices as predictors alongside the previous year's COVID-19 cases in each country improved the predictive accuracy of 2021 cases by an average of 18% across the 3 chosen algorithms. Given that predictive algorithms can struggle with smaller data sets [59], the results of this study (n=182) are noteworthy.

The linear regression algorithm was the strongest learner on the data, but also showed the least improvement (7% increase in mean cross-validation) once the HDI metrics were added. Given

that the other algorithms improved considerably when HDI indices were added, this result represents an interesting outlier. The varying performances between the algorithms may be due to the statistically linear relationships between the variables (as discovered in the Pearson correlation matrix in Figure 4). Despite the strong correlation between 2021 COVID-19 cumulative cases per million and case data from the previous year (R=0.84), Feature Model 1 did not make accurate predictions using random forest or AdaBoost. Unlike linear regression models, which excel at fitting to data where linear correlation exists, decision tree algorithms like random forest and AdaBoost may perform more effectively with nonlinear data [80,81]. Lastly, the widely varying performance of each k-fold iteration justified the use of cross-validation to evaluate the models. In Feature Model 2, for example, the highest scoring fold of the linear regression algorithm had a result of 94.6, a highly accurate R^2 result. However, the lowest scoring fold had an R^2 of 59.5. The cross-validation R^2 score of 76.3 was therefore the most reliable score for the data set.

Follow-Up Analysis

Following the primary analysis, 4 follow-up analyses were conducted. First, Feature Model 2 was trained again without 2020 COVID-19 case data as a feature to analyze how well the HDI metrics could predict COVID-19 cases alone. Without the previous year's case data, the accuracy was low (R^2=0.438 for the best performing algorithm, which was again linear regression). This result highlights the significant importance of 2020 case data in predicting the following year's COVID-19 prevalence. Second, Feature Model 2 was trained again using 1 HDI metric at a time to analyze which was the most important for the prediction of COVID-19 cases. The results showed that expected years of schooling and mean years of schooling had the highest scores (R^2=0.755 for each), followed by life expectancy (R^2=0.739) and then GNI (R^2=0.712). This suggests that education was the most predictive socioeconomic indicator (the education HDI metrics were also the most statistically correlative). However, the results also showed that using all HDI indices is more effective than using them separately for COVID-19 case prediction in this data set. The third follow-up experiment removed the 4 previously identified outlier countries (Andorra, Montenegro, Serbia, and Seychelles) from the data set and implemented both feature models again, using the same cross-validation method as the initial analysis. This yielded interesting results (see Tables 8 and 9). Most notably, random forest became the strongest learner in Feature Model 2 (R^2=0.777). Despite being generally less sensitive to outliers [82], random forest benefitted from outlier removal in this data set. Removing outliers also reduced the gap in performance between the algorithms. With outliers included, Feature Model 1 displayed a 23.9% difference between the best and worst performing algorithms, and with outliers removed, this difference reduced to 19.5%. This reduction was more apparent in Feature Model 2, with just a 2.1% difference between the best and worst performing algorithms with outliers removed (compared with an 8.4% difference in the original sample with outliers included). However, the results indicated that removing the outliers did not significantly improve overall predictive accuracy.

The fourth follow-up experiment sought to compare socioeconomic status as a COVID-19 predictor with a selection of other COVID-19 risk factors. Subsequently, each country's median age, population density (individuals per square kilometer), and percentage of vaccinated individuals were sourced and added to the data set. Each of these variables has been shown to predict COVID-19 prevalence in certain samples [83-85]. Most of the required data were also available in the OurWorldInData database, though a small number of entries had to be sourced from Worldometers and IndexMundi [86,87].

When Feature Model 2 was trained again using these new metrics alongside 2020 case data, predictive accuracy dropped to an average of 0.649 across all 3 algorithms. Using these new features, the most accurate algorithm was 10% less accurate than the most accurate learner in the model with socioeconomic features (see Table 10). This is a significant finding, which suggests that socioeconomic status was more effective in predicting 2021 cumulative cases than a country's median age, population density, and vaccination uptake, highlighting its unique importance as a nonclinical predictor of COVID-19 in the sample of countries.

Table 8. Feature Model 1 comparison (outliers included versus excluded).

Algorithm	Mean R^2 in the sample with outliers included (n=182)	Mean R^2 in the sample with outliers excluded (n=178)
Linear regression	0.693	0.689
Random forest	0.481	0.493
AdaBoost	0.454	0.494

Table 9. Feature Model 2 comparison (outliers included versus excluded).

Algorithm	Mean R^2 in the sample with outliers included (n=182)	Mean R^2 in the sample with outliers excluded (n=178)
Linear regression	0.763	0.754
Random forest	0.722	0.777
AdaBoost	0.679	0.733

Table 10. Feature Model 2 performance comparison of socioeconomic metrics versus other risk factors using linear regression.

Measure	Feature Model 2 with HDI[a] indicators	Feature Model 2 with population density, median age, and vaccination uptake
R^2	0.763	0.661
MAE[b]	0.070	0.075
MSE[c]	0.011	0.016
RMSE[d]	0.107	0.128
Max error	0.265	0.312

[a]HDI: Human Development Index.

[b]MAE: mean absolute error.

[c]MSE: mean squared error.

[d]RMSE: root mean squared error.

Significance of the Results

In order to put the machine learning results of this study into perspective, we compared the best performing algorithm (R^2=0.763) with similar machine learning COVID-19 case predictions. Overall, it fits within the accepted range of COVID-19 predictive modeling studies in the systematic review mentioned earlier, which ranged from 0.64 to 1 [64]. Results from this study align with the findings from another study that attempted to predict COVID-19 cumulative cases in 3109 counties in the United States using a multilayer perceptron neural network. In this previous study, the socioeconomic indicator of median household income ranked fifth among 57 clinical and nonclinical predictor variables of COVID-19 prevalence [88]. Studies, such as this, portray the importance of socioeconomic indicators as determinants of COVID-19 prevalence rates, which further supports the use of HDI in this study to more accurately and precisely predict COVID-19 prevalence in 2021.

This research has a number of implications. First, it showcases the utility of combining statistical and machine learning approaches in pandemic research. Although statistical tests can determine correlations between variables, they cannot provide specific predictions of the target variable. Each method thus addresses a shortcoming of the other. Second, this study indicates that socioeconomic status is an important variable to consider in future epidemiological modeling, and reveals the complex social nature of the COVID-19 pandemic. Socioeconomic status was a better predictor of COVID-19 prevalence than median age, population density, and vaccination uptake. Third, the accuracy of these results in a multicountry sample is noteworthy. Owing to the data taken from 182 countries, this research suggests that socioeconomic status can be considered a “global risk factor” rather than a country-specific factor [38]. This will support evidence-based policy and interventions by decision makers. Fourth, the results indicate that although socioeconomic factors aid in the prediction of COVID-19, there could be other important factors that could further optimize prediction. Finally, the importance of historically reported COVID-19 case data cannot be understated in attempting to predict future COVID-19 prevalence. The 2020 COVID-19 case data correlated strongly with 2021 COVID-19 case data and could be considered the most important machine learning feature.

Limitations

As with all research studies, there are inherent limitations in this study. First, when analyzing COVID-19 cross-nationally, it must be noted that some countries have underreported their number of cases more than others for reasons, such as limited testing capacity [89]. Second, there are other socioeconomic factors that the HDI does not account for, including levels of financial inequality, social exclusion, or discrimination within countries [90]. These factors are worthy of inclusion in future research to assess their impact. Third, national COVID-19 prevalence rates give an overall measure of how severely a country is impacted, which is suitable for cross-country research, but they do not capture the full complexity of transmission patterns within each country. It is recommended that further research be conducted at the regional and municipal levels to assist pandemic forecasting. Lastly, it can be challenging to train reliable machine learning models using small data sets [59]. Cross-validation was used to address this limitation, as it maximizes the data set and minimizes the potential bias of a traditional partitioning approach.

Conclusions

A better understanding of population-level predictors is of crucial importance to better understand and respond to public health crises caused by COVID-19 [91]. This study contributes to the growing corpus of COVID-19 predictive modeling research by showing that socioeconomic status is an important nonclinical risk factor. Using HDI and historical case rates, it was observed that 2021 cross-national COVID-19 cumulative cases could be predicted with a reasonable degree of accuracy. Although COVID-19 represents a long-term challenge for the global society, the data-driven approach of machine learning will continue to support decision makers in understanding the pandemic, formulating response strategies, and predicting future outcomes [92].

Conflicts of Interest

None declared.

References

1. The largest global rollout of vaccines in history just got one step closer. Gavi, The Vaccine Alliance. URL: https://www.gavi.org/vaccineswork/largest-global-rollout-vaccines-history-just-got-one-step-closer [accessed 2021-11-05]
2. Iyanda AE, Adeleke R, Lu Y, Osayomi T, Adaralegbe A, Lasode M, et al. A retrospective cross-national examination of COVID-19 outbreak in 175 countries: a multiscale geographically weighted regression analysis (January 11-June 28, 2020). J Infect Public Health 2020 Oct;13(10):1438-1445 [FREE Full text] [doi: 10.1016/j.jiph.2020.07.006] [Medline: 32773211]
3. Balmford B, Annan JD, Hargreaves JC, Altoè M, Bateman IJ. Cross-Country Comparisons of Covid-19: Policy, Politics and the Price of Life. Environ Resour Econ (Dordr) 2020;76(4):525-551 [FREE Full text] [doi: 10.1007/s10640-020-00466-5] [Medline: 32836862]
4. Földi M, Farkas N, Kiss S, Zádori N, Váncsa S, Szakó L, KETLAK Study Group. Obesity is a risk factor for developing critical condition in COVID-19 patients: A systematic review and meta-analysis. Obes Rev 2020 Oct;21(10):e13095 [FREE Full text] [doi: 10.1111/obr.13095] [Medline: 32686331]
5. Mahase E. Covid-19: Why are age and obesity risk factors for serious disease? BMJ 2020 Oct 26;371:m4130. [doi: 10.1136/bmj.m4130] [Medline: 33106243]
6. Masood M, Aggarwal A, Reidpath DD. Effect of national culture on BMI: a multilevel analysis of 53 countries. BMC Public Health 2019 Sep 03;19(1):1212 [FREE Full text] [doi: 10.1186/s12889-019-7536-0] [Medline: 31481044]
7. Zhou Y, Chi J, Lv W, Wang Y. Obesity and diabetes as high-risk factors for severe coronavirus disease 2019 (Covid-19). Diabetes Metab Res Rev 2021 Feb;37(2):e3377 [FREE Full text] [doi: 10.1002/dmrr.3377] [Medline: 32588943]
8. Lima-Martínez MM, Carrera Boada C, Madera-Silva MD, Marín W, Contreras M. COVID-19 and diabetes: A bidirectional relationship. Clin Investig Arterioscler 2021;33(3):151-157 [FREE Full text] [doi: 10.1016/j.arteri.2020.10.001] [Medline: 33303218]
9. Kashyap VK, Dhasmana A, Massey A, Kotnala S, Zafar N, Jaggi M, et al. Smoking and COVID-19: Adding Fuel to the Flame. Int J Mol Sci 2020 Sep 09;21(18):6581 [FREE Full text] [doi: 10.3390/ijms21186581] [Medline: 32916821]
10. Ibanez A, Sisodia GS. The role of culture on 2020 SARS-CoV-2 Country deaths: a pandemic management based on cultural dimensions. GeoJournal 2022;87(2):1175-1191 [FREE Full text] [doi: 10.1007/s10708-020-10306-0] [Medline: 33020679]
11. Yeung AY, Roewer-Despres F, Rosella L, Rudzicz F. Machine Learning-Based Prediction of Growth in Confirmed COVID-19 Infection Cases in 114 Countries Using Metrics of Nonpharmaceutical Interventions and Cultural Dimensions: Model Development and Validation. J Med Internet Res 2021 Apr 23;23(4):e26628 [FREE Full text] [doi: 10.2196/26628] [Medline: 33844636]
12. Cascini F, Pantovic A, Al-Ajlouni Y, Failla G, Ricciardi W. Attitudes, acceptance and hesitancy among the general population worldwide to receive the COVID-19 vaccines and their contributing factors: A systematic review. EClinicalMedicine 2021 Oct;40:101113 [FREE Full text] [doi: 10.1016/j.eclinm.2021.101113] [Medline: 34490416]
13. Troumbis AY. Testing the socioeconomic determinants of COVID-19 pandemic hypothesis with aggregated Human Development Index. J Epidemiol Community Health 2020 Dec 08:jech-2020-215986. [doi: 10.1136/jech-2020-215986] [Medline: 33293289]
14. Azza A, Sarhan A. Using Human Development Indices to Identify Indicators to Monitor the Corona Virus Pandemic. JVAT 2020 Apr 17;1(1):48-57. [doi: 10.14302/issn.2691-8862.jvat-20-3306]
15. de Souza CDF, Machado MF, do Carmo RF. Human development, social vulnerability and COVID-19 in Brazil: a study of the social determinants of health. Infect Dis Poverty 2020 Aug 31;9(1):124 [FREE Full text] [doi: 10.1186/s40249-020-00743-x] [Medline: 32867851]
16. Stanton EA. The Human Development Index: A History. University of Massachusetts-Amherst. 2007. URL: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1101&context=peri_workingpapers [accessed 2021-07-05]
17. Franco-Herrera D, González-Ocampo D, Restrepo-Montoya V, Gómez-Guevara JE, Alvear-Villacorte N, Rodríguez-Morales AJ. Relationship between malaria epidemiology and the human development index in Colombia and Latin America. Infez Med 2018 Sep 01;26(3):255-262 [FREE Full text] [Medline: 30246769]
18. Pervaiz R, Faisal F. Investigating the Nexus between Gynaecologic Cancer and Human Development Index. Afr J Reprod Health 2020 Mar;24(1):53-61. [doi: 10.29063/ajrh2020/v24i1.6] [Medline: 32358937]
19. Khazaei Z, Goodarzi E, Borhaninejad V, Iranmanesh F, Mirshekarpour H, Mirzaei B, et al. The association between incidence and mortality of brain cancer and human development index (HDI): an ecological study. BMC Public Health 2020 Nov 12;20(1):1696 [FREE Full text] [doi: 10.1186/s12889-020-09838-4] [Medline: 33183267]
20. Zeng Z, Chen J, Xiao C, Chen W. A Global View on Prevalence of Hypertension and Human Develop Index. Ann Glob Health 2020 Jun 29;86(1):67 [FREE Full text] [doi: 10.5334/aogh.2591] [Medline: 32676296]
21. Singu S, Acharya A, Challagundla K, Byrareddy SN. Impact of Social Determinants of Health on the Emerging COVID-19 Pandemic in the United States. Front Public Health 2020 Jul 21;8:406 [FREE Full text] [doi: 10.3389/fpubh.2020.00406] [Medline: 32793544]

22. Darin-Mattsson A, Fors S, Kåreholt I. Different indicators of socioeconomic status and their relative importance as determinants of health in old age. Int J Equity Health 2017 Sep 26;16(1):173 [FREE Full text] [doi: 10.1186/s12939-017-0670-3] [Medline: 28950875]
23. Hellmich SN. What is Socioeconomics? An Overview of Theories, Methods, and Themes in the Field. Forum for Social Economics 2015 Jan 15;46(1):3-25. [doi: 10.1080/07360932.2014.999696]
24. Adler NE, Boyce T, Chesney MA, Cohen S, Folkman S, Kahn RL, et al. Socioeconomic status and health. The challenge of the gradient. Am Psychol 1994 Jan;49(1):15-24. [doi: 10.1037//0003-066x.49.1.15] [Medline: 8122813]
25. Wang J, Geng L. Effects of Socioeconomic Status on Physical and Psychological Health: Lifestyle as a Mediator. Int J Environ Res Public Health 2019 Jan 20;16(2):281 [FREE Full text] [doi: 10.3390/ijerph16020281] [Medline: 30669511]
26. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep 2014;129 Suppl 2:19-31 [FREE Full text] [doi: 10.1177/00333549141291S206] [Medline: 24385661]
27. Walters S, Suhrcke M. Socioeconomic inequalities in health and health care access in central and eastern Europe and the CIS: a review of the recent literature. World Health Organization. 2005. URL: https://apps.who.int/iris/handle/10665/350352 [accessed 2022-03-01]
28. Adler NE, Ostrove JM. Socioeconomic status and health: what we know and what we don't. Ann N Y Acad Sci 1999;896:3-15. [doi: 10.1111/j.1749-6632.1999.tb08101.x] [Medline: 10681884]
29. Hakeberg M, Wide Boman U. Self-reported oral and general health in relation to socioeconomic position. BMC Public Health 2017 Jul 26;18(1):63 [FREE Full text] [doi: 10.1186/s12889-017-4609-9] [Medline: 28747180]
30. Hawkins RB, Charles EJ, Mehaffey JH. Socio-economic status and COVID-19-related cases and fatalities. Public Health 2020 Dec;189:129-134 [FREE Full text] [doi: 10.1016/j.puhe.2020.09.016] [Medline: 33227595]
31. Patel JA, Nielsen FBH, Badiani AA, Assi S, Unadkat VA, Patel B, et al. Poverty, inequality and COVID-19: the forgotten vulnerable. Public Health 2020 Jun;183:110-111 [FREE Full text] [doi: 10.1016/j.puhe.2020.05.006] [Medline: 32502699]
32. Human Development Report 2020. United Nations Development Programme. 2020. URL: https://hdr.undp.org/system/files/documents//hdr2020pdf.pdf [accessed 2021-09-08]
33. Hu Q, Zhang Q, Chen W, Bai X, Liang T. Human development index is associated with mortality-to-incidence ratios of gastrointestinal cancers. World J Gastroenterol 2013 Aug 28;19(32):5261-5270 [FREE Full text] [doi: 10.3748/wjg.v19.i32.5261] [Medline: 23983428]
34. Soheylizad M, Khazaei S, Jenabi E, Delpisheh A, Veisani Y. The Relationship Between Human Development Index and Its Components with Thyroid Cancer Incidence and Mortality: Using the Decomposition Approach. Int J Endocrinol Metab 2018 Oct;16(4):e65078 [FREE Full text] [doi: 10.5812/ijem.65078] [Medline: 30464773]
35. Javanmard E, Niyyati M, Ghasemi E, Mirjalali H, Asadzadeh Aghdaei H, Zali MR. Impacts of human development index and climate conditions on prevalence of Blastocystis: A systematic review and meta-analysis. Acta Trop 2018 Sep;185:193-203. [doi: 10.1016/j.actatropica.2018.05.014] [Medline: 29802845]
36. Pereira FA, de Mendonça IA, Werneck RI, Moysés ST, Gabardo MC, Moysés SJ. Human Development Index, Ratio of Dentists and Inhabitants, and the Decayed, Missing or Filled Teeth Index in Large Cities. J Contemp Dent Pract 2018 Nov 01;19(11):1363-1369. [Medline: 30602642]
37. Shaikh A, Peprah E, Mohamed RH, Asghar A, Andharia NV, Lajot NA, et al. COVID-19 and mental health: a multi-country study—the effects of lockdown on the mental health of young adults. Middle East Curr Psychiatry 2021 Aug 09;28(1):1-10. [doi: 10.1186/s43045-021-00116-6]
38. O'Sullivan R, Burns A, Leavey G, Leroi I, Burholt V, Lubben J, et al. Impact of the COVID-19 Pandemic on Loneliness and Social Isolation: A Multi-Country Study. Int J Environ Res Public Health 2021 Sep 23;18(19):9982 [FREE Full text] [doi: 10.3390/ijerph18199982] [Medline: 34639283]
39. Hawlader MDH, Rahman ML, Nazir A, Ara T, Haque MMA, Saha S, et al. COVID-19 vaccine acceptance in South Asia: a multi-country study. Int J Infect Dis 2022 Jan;114:1-10 [FREE Full text] [doi: 10.1016/j.ijid.2021.09.056] [Medline: 34597765]
40. Jang SY, Hussain-Alkhateeb L, Rivera Ramirez T, Al-Aghbari AA, Chackalackal DJ, Cardenas-Sanchez R, et al. Factors shaping the COVID-19 epidemic curve: a multi-country analysis. BMC Infect Dis 2021 Oct 02;21(1):1032 [FREE Full text] [doi: 10.1186/s12879-021-06714-3] [Medline: 34600485]
41. Chen Y, Lu P, Chang C, Liu T. A Time-Dependent SIR Model for COVID-19 With Undetectable Infected Persons. IEEE Trans. Netw. Sci. Eng 2020 Oct 1;7(4):3279-3294. [doi: 10.1109/tnse.2020.3024723]
42. Calafiore G, Novara C, Possieri C. A Modified SIR Model for the COVID-19 Contagion in Italy. 2020 Presented at: 59th IEEE Conference on Decision and Control (CDC); December 14-18, 2020; Jeju, Korea (South). [doi: 10.1109/cdc42340.2020.9304142]
43. Law KB, Peariasamy KM, Ibrahim H, Abdullah NH. Modelling infectious diseases with herd immunity in a randomly mixed population. Research Square. URL: https://www.researchsquare.com/article/rs-289776/v5 [accessed 2022-09-03]
44. Kadkhoda K. Herd Immunity to COVID-19. Am J Clin Pathol 2021 Mar 15;155(4):471-472 [FREE Full text] [doi: 10.1093/ajcp/aqaa272] [Medline: 33399182]
45. Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. Annu Rev Public Health 2020 Apr 02;41:21-36. [doi: 10.1146/annurev-publhealth-040119-094437] [Medline: 31577910]

46. Anderson AB, Grazal CF, Balazs GC, Potter BK, Dickens JF, Forsberg JA. Can Predictive Modeling Tools Identify Patients at High Risk of Prolonged Opioid Use After ACL Reconstruction? Clin Orthop Relat Res 2020 Jul;478(7):-1618 [FREE Full text] [doi: 10.1097/CORR.0000000000001251] [Medline: 32282466]
47. Kang AR, Lee J, Jung W, Lee M, Park SY, Woo J, et al. Development of a prediction model for hypotension after induction of anesthesia using machine learning. PLoS One 2020;15(4):e0231172 [FREE Full text] [doi: 10.1371/journal.pone.0231172] [Medline: 32298292]
48. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Sci Rep 2019 Feb 12;9(1):1879 [FREE Full text] [doi: 10.1038/s41598-019-38491-0] [Medline: 30755689]
49. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. Am J Epidemiol 2019 Dec 31;188(12):2222-2239. [doi: 10.1093/aje/kwz189] [Medline: 31509183]
50. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature 2013 Apr 25;496(7446):504-507 [FREE Full text] [doi: 10.1038/nature12060] [Medline: 23563266]
51. Pigott DM, Golding N, Mylne A, Huang Z, Henry AJ, Weiss DJ, et al. Mapping the zoonotic niche of Ebola virus disease in Africa. Elife 2014 Sep 08;3:e04395 [FREE Full text] [doi: 10.7554/eLife.04395] [Medline: 25201877]
52. Messina JP, Pigott DM, Golding N, Duda KA, Brownstein JS, Weiss DJ, et al. The global distribution of Crimean-Congo hemorrhagic fever. Trans R Soc Trop Med Hyg 2015 Aug;109(8):503-513 [FREE Full text] [doi: 10.1093/trstmh/trv050] [Medline: 26142451]
53. Messina JP, Kraemer MU, Brady OJ, Pigott DM, Shearer FM, Weiss DJ, et al. Mapping global environmental suitability for Zika virus. Elife 2016 Apr 19;5:e15272 [FREE Full text] [doi: 10.7554/eLife.15272] [Medline: 27090089]
54. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proc Natl Acad Sci U S A 2012 Dec 11;109(50):20425-20430 [FREE Full text] [doi: 10.1073/pnas.1208772109] [Medline: 23184969]
55. Wang L, Yang G, Jia L, Li Z, Xie J, Li P, et al. Epidemiological features and trends of Ebola virus disease in West Africa. Int J Infect Dis 2015 Sep;38:52-53 [FREE Full text] [doi: 10.1016/j.ijid.2015.07.017] [Medline: 26216765]
56. DeFelice NB, Little E, Campbell SR, Shaman J. Ensemble forecast of human West Nile virus cases and mosquito infection rates. Nat Commun 2017 Feb 24;8:14592 [FREE Full text] [doi: 10.1038/ncomms14592] [Medline: 28233783]
57. Reis J, Shaman J. Retrospective Parameter Estimation and Forecast of Respiratory Syncytial Virus in the United States. PLoS Comput Biol 2016 Oct;12(10):e1005133 [FREE Full text] [doi: 10.1371/journal.pcbi.1005133] [Medline: 27716828]
58. Painuli D, Mishra D, Bhardwaj S, Aggarwal M. Forecast and prediction of COVID-19 using machine learning. In: Kose U, Gupta D, de Albuquerque VHC, Khanna A, editors. Data Science for COVID-19. Cambridge, MA: Academic Press; 2021:381-397.
59. Ahmad A, Garhwal S, Ray SK, Kumar G, Malebary SJ, Barukab OM. The Number of Confirmed Cases of Covid-19 by using Machine Learning: Methods and Challenges. Arch Comput Methods Eng 2021 Aug 04;28(4):2645-2653 [FREE Full text] [doi: 10.1007/s11831-020-09472-8] [Medline: 32837183]
60. Assaf D, Gutman Y, Neuman Y, Segal G, Amit S, Gefen-Halevi S, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med 2020 Nov;15(8):1435-1443 [FREE Full text] [doi: 10.1007/s11739-020-02475-0] [Medline: 32812204]
61. Bolourani S, Brenner M, Wang P, McGinn T, Hirsch JS, Barnaby D, Northwell COVID-19 Research Consortium. A Machine Learning Prediction Model of Respiratory Failure Within 48 Hours of Patient Admission for COVID-19: Model Development and Validation. J Med Internet Res 2021 Feb 10;23(2):e24246 [FREE Full text] [doi: 10.2196/24246] [Medline: 33476281]
62. Gao Y, Cai G, Fang W, Li H, Wang S, Chen L, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nat Commun 2020 Oct 06;11(1):5033 [FREE Full text] [doi: 10.1038/s41467-020-18684-2] [Medline: 33024092]
63. Banoei MM, Dinparastisaleh R, Zadeh AV, Mirsaeidi M. Machine-learning-based COVID-19 mortality prediction model and identification of patients at low and high risk of dying. Crit Care 2021 Sep 08;25(1):328 [FREE Full text] [doi: 10.1186/s13054-021-03749-5] [Medline: 34496940]
64. Ghafouri-Fard S, Mohammad-Rahimi H, Motie P, Minabi MAS, Taheri M, Nateghinia S. Application of machine learning in the prediction of COVID-19 daily new cases: A scoping review. Heliyon 2021 Oct;7(10):e08143 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e08143] [Medline: 34660935]
65. Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, et al. Coronavirus Pandemic (COVID-19). Our World in Data. URL: https://ourworldindata.org/coronavirus [accessed 2021-09-01]
66. Pollet TV, Stulp G, Henzi SP, Barrett L. Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. Am J Primatol 2015 Jul;77(7):727-740. [doi: 10.1002/ajp.22405] [Medline: 25810242]
67. Data downloads. United Nations Development Programme. URL: https://hdr.undp.org/data-center/documentation-and-downloads [accessed 2022-03-10]
68. Ahsan MM, Mahmud MAP, Saha PK, Gupta KD, Siddique Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies 2021 Jul 24;9(3):52. [doi: 10.3390/technologies9030052]

69. Patel K, Drucker SM, Fogarty J, Kapoor A, Tan DS. Using Multiple Models to Understand Data. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. 2011 Presented at: Twenty-Second International Joint Conference on Artificial Intelligence; July 16-22, 2011; Barcelona, Catalonia, Spain. [doi: 10.5591/978-1-57735-516-8/IJCAI11-289]
70. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019 Dec 21;19(1):281 [FREE Full text] [doi: 10.1186/s12911-019-1004-8] [Medline: 31864346]
71. Xu Y, Goodacre R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. J Anal Test 2018;2(3):249-262 [FREE Full text] [doi: 10.1007/s41664-018-0068-2] [Medline: 30842888]
72. Maulud D, Abdulazeez AM. A Review on Linear Regression Comprehensive in Machine Learning. JASTT 2020 Dec 31;1(4):140-147. [doi: 10.38094/jastt1457]
73. Biau G. Analysis of a Random Forests Model. Journal of Machine Learning Research 2012;13:1063-1095 [FREE Full text]
74. Belle V, Papantonis I. Principles and Practice of Explainable Machine Learning. Front Big Data 2021 Jul 1;4:688969 [FREE Full text] [doi: 10.3389/fdata.2021.688969] [Medline: 34278297]
75. Wyner AJ, Olson M, Bleich J, Mease D. Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers. Journal of Machine Learning Research 2017;18:1-33 [FREE Full text]
76. Marcot BG, Hanea AM. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Comput Stat 2020 Jun 13;36(3):2009-2031. [doi: 10.1007/s00180-020-00999-9]
77. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial intelligence. 1995 Presented at: 14th International Joint Conference on Artificial intelligence; August 20-25, 1995; Montreal, Quebec, Canada p. 1137-1143. [doi: 10.5555/1643031.1643047]
78. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 2021;7:e623 [FREE Full text] [doi: 10.7717/peerj-cs.623] [Medline: 34307865]
79. Bayati M. Why Is COVID-19 More Concentrated in Countries with High Economic Status? Iran J Public Health 2021 Sep;50(9):1926-1929 [FREE Full text] [doi: 10.18502/ijph.v50i9.7081] [Medline: 34722396]
80. Auret L, Aldrich C. Interpretation of nonlinear relationships between process variables by use of random forests. Minerals Engineering 2012 Aug;35:27-42. [doi: 10.1016/j.mineng.2012.05.008]
81. Lo A, Chernoff H, Zheng T, Lo S. Why significant variables aren't automatically good predictors. Proc Natl Acad Sci U S A 2015 Nov 10;112(45):13892-13897 [FREE Full text] [doi: 10.1073/pnas.1518285112] [Medline: 26504198]
82. Breiman L. Random Forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A1010933404324]
83. Wong DWS, Li Y. Spreading of COVID-19: Density matters. PLoS One 2020;15(12):e0242398 [FREE Full text] [doi: 10.1371/journal.pone.0242398] [Medline: 33362283]
84. Davies NG, Klepac P, Liu Y, Prem K, Jit M, CMMID COVID-19 working group, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med 2020 Aug;26(8):1205-1211. [doi: 10.1038/s41591-020-0962-9] [Medline: 32546824]
85. Wilder-Smith A. What is the vaccine effect on reducing transmission in the context of the SARS-CoV-2 delta variant? The Lancet Infectious Diseases 2022 Feb;22(2):152-153. [doi: 10.1016/s1473-3099(21)00690-3]
86. Countries in the world by population. Worldometers. URL: https://www.worldometers.info/world-population/population-by-country/ [accessed 2022-04-01]
87. Factbook-Countries. IndexMundi. URL: https://www.indexmundi.com/factbook/countries [accessed 2022-04-02]
88. Mollalo A, Rivera KM, Vahedi B. Artificial Neural Network Modeling of Novel Coronavirus (COVID-19) Incidence Rates across the Continental United States. Int J Environ Res Public Health 2020 Jun 12;17(12):4204 [FREE Full text] [doi: 10.3390/ijerph17124204] [Medline: 32545581]
89. Lau H, Khosrawipour T, Kocbach P, Ichii H, Bania J, Khosrawipour V. Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. Pulmonology 2021;27(2):110-115 [FREE Full text] [doi: 10.1016/j.pulmoe.2020.05.015] [Medline: 32540223]
90. Bilbao-Ubillos J. The Limits of Human Development Index: The Complementary Role of Economic and Social Cohesion, Development Strategies and Sustainability. Sust. Dev 2011 May 19;21(6):400-412. [doi: 10.1002/sd.525]
91. Erman A, Medeiros M. Exploring the Effect of Collective Cultural Attributes on Covid-19-Related Public Health Outcomes. Front Psychol 2021;12:627669 [FREE Full text] [doi: 10.3389/fpsyg.2021.627669] [Medline: 33833717]
92. Polonsky JA, Baidjoe A, Kamvar ZN, Cori A, Durski K, Edmunds WJ, et al. Outbreak analytics: a developing data science for informing the response to emerging pathogens. Philos Trans R Soc Lond B Biol Sci 2019 Jul 08;374(1776):20180276 [FREE Full text] [doi: 10.1098/rstb.2018.0276] [Medline: 31104603]

Abbreviations

GNI: gross national income

HDI: Human Development Index
MAE: mean absolute error
MSE: mean squared error
RMSE: root mean squared error
SIR: susceptible, infected, recovered
UNDP: United Nations Development Programme

Edited by A Mavragani; submitted 22.11.21; peer-reviewed by P Wang, S Rostam Niakan Kalhori, M Pradhan; comments to author 24.02.22; revised version received 12.04.22; accepted 27.04.22; published 27.09.22

Please cite as:
Winston L, McCann M, Onofrei G
Exploring Socioeconomic Status as a Global Determinant of COVID-19 Prevalence, Using Exploratory Data Analytic and Supervised Machine Learning Techniques: Algorithm Development and Validation Study
JMIR Form Res 2022;6(9):e35114
URL: https://formative.jmir.org/2022/9/e35114
doi: 10.2196/35114
PMID: 36001798