Received 06-04-22; Accepted 12-05-22; Published 19-05-22

Association of Social, Demographic, Health, Nutritional and Environmental Factors With the Incidence and Death Rates of COVID-19; a Global Cross-Sectional Analytical Study.

Supun Sudaraka ssm123ssm@gmail.com
National Hospital of Sri Lanka
Colombo 08
Sri Lanka

Ishanya I. Abeyagunawardena ishanya1993@gmail.com
National Hospital of Sri Lanka
Colombo 08
Sri Lanka

Raahya Lafir raahya.l@icloud.com
National Hospital of Sri Lanka
Colombo 08
Sri Lanka

Samath Dharmaratne Samath.Dharmaratne@mbbs.md
Institute for Health Metrics and Evaluation,
Department of Health Metrics Sciences
School of Medicine, University of Washington, United States of America
Sri Lanka

Corresponding Author: Abeyagunawardena I.A

Abstract

Background: The magnitude of the impact of COVID-19 is dependent on social, demographic, health, nutrition and even environmental factors. These factors act individually and synergistically to impact the incidence, mortality and morbidity of COVID-19. We aimed to evaluate the variables contributing individually to COVID-19 incidence and mortality utilizing techniques to minimize the effects of interaction between these factors.

Method: Data regarding 88 variables for 195 countries over three years were extracted from The Health Nutrition and Population Statistics database and aggregated into a consolidated median. Outliers were eliminated and variables having a completeness of more than 70% were selected. The analysis was done separately for the incidence and mortality of COVID-19. Principal component Analysis (PCA) and Elastic net regression were used to identify the most important single variables. The significant variables of the PCA which explained the most variance were identified. Subsequently, variables with the highest importance (using normalized ranked regression coefficients) in the Elastic Net model were selected and the

Citation: Manathunga SS, et al. Association of Social, Demographic, Health, Nutritional and Environmental Factors With the Incidence and Death Rates of COVID-19; a Global Cross-Sectional Analytical Study. Advances in Artificial Intelligence and Machine Learning. 2022;2(2):23.

intersecting set of variables common to both models was considered as predictors affecting incidence and mortality of COVID-19.

Result: The study revealed communities with a high prevalence of anaemia has a negative correlation with COVID-19 incidence which was furthermore, interestingly seen in multiple age groups. Diphtheria, Tetanus and Pertussis (DTP) Immunization in children was also found to have a negative linear correlation.

Conclusion: A negative individual association was seen between anaemia (in multiple age groups) and DTP immunization in children with the incidence and mortality of COVID 19.

Keywords: Coronavirus, COVID-19, Anemia, DTP vaccination

1. INTRODUCTION

Since its detection in November 2019 the novel coronavirus disease (COVID-19) has spread rampantly causing considerable adversity leading to the WHO declaring the pandemic a Public Health Emergency of International Concern on the 30th January 2020 [1]. The pattern and the magnitude of the spread of the disease depend on many factors, some of which are yet to be uncovered. However, it has been ascertained that socio-demographic factors like population age structure and prevalence of non-communicable diseases (NCDs) have an effect on the incidence, morbidity and mortality of COVID-19. A systematic literature review done by Zlatko et al. on NCDs and COVID-19 reported that certain NCDs predispose to a higher probability of infection by the virus and increase the likelihood of developing severe disease leading to higher mortality [2]. In addition, some studies have shown that malnutrition is an important risk factor as much as NCDs for the COVID-19 mortality [3,4].

There have also been studies reporting associations of COVID-19 incidence and mortality with climatic variables such as temperature and humidity. A study by Malki et al. has identified temperature as the most important weather variable using machine learning algorithms to predict death rates due to the infection [5].

Immunization history, particularly past vaccination of diphtheria and tetanus vaccines has been observed to be associated with lower odds of COVID-19 hospitalization, possibly due to cross-reactive immunity [6].

However, in analyzing these variables with COVID-19 incidence and its mortality a conundrum exists, namely that these variables are not independent of each other and a complex synergistic interplay is present between a multitude of factors. Hence, interpretation of the associations of each variable with the COVID-19 incidence and mortality should be done with caution. To overcome this hurdle, we aimed to identify the most important individual variables explaining the incidence and mortality of COVID-19 infection by utilizing techniques to minimize the effects of interactions between these multiple factors.

2. METHODOLOGY

2.1 Data Acquisition

The Health Nutrition and Population Statistics database of The World Bank Group which is freely available was accessed and data was extracted for 88 variables pertaining to social, demographic, health, nutritional and environmental factors for 195 countries across the globe (which will be collectively called as World Bank data for reference) [7]. The total number of COVID-19 cases and deaths per million population for each country up to 01-11-2021 were obtained from 'Our World in Data', a publicly available database providing comprehensive information on global issues including COVID-19 [8].

In order to maximize yield, data for each variable was extracted across three consecutive years per country and aggregated to a consolidated median after filling the missing values to the nearest neighbour. Data were pre-processed by outlier elimination using the Hampel filter method followed by scaling and centring to achieve a mean of zero and a standard deviation of one [9]. Variables that had a completeness of less than 70% were excluded from the analysis.

2.2 Exploratory Data Analysis and Modelling

Spearman correlation coefficients between the variables were calculated, correcting for multiple comparisons along with statistical significance. A correlogram was generated to identify associations between the variables and to detect multi-collinearity. Histograms and density plots for each variable were generated and the correlation of each variable with COVID-19 cases and deaths per million were visualized with scatterplot matrices, in order of the significance of correlation coefficients. Different dimension reduction and feature subset selection methods, including decision-tree based algorithms, were used to identify the most important single variables explaining the COVID-19 cases and deaths per million. They were Elastic Net regression; a regularization technique for General Linear Models, Principal Component Analysis (PCA), Random Forests and XGBoost[10-12]. PCA maps the data into a new coordinate system in a way that the first coordinate explains the highest variance of data, the second coordinate explains the second-highest variance and so on. The algorithm requires a scaled correlation matrix as input and uses singular value decomposition for calculating the Principal Components [13]. Elastic Net Regression falls under the category of regularized linear regression family. It is considered superior to least absolute shrinkage and selection operator (LASSO) regression due to its ability to perform better in high-dimension, low-observations cases [14]. Random Forests is an ensemble of bootstrap-aggregated decision trees. The ability of this model to rank the predictors to yield a variable importance map is an advantageous property for this analysis of identifying the most important predictor variables [11].

XGboost is an algorithm that implements gradient boosted decision trees optimized for speed and accuracy, which supports parallelization of tree construction and automatic missing data handling. The model incorporates gradient boosting, stochastic gradient boosting and regularized gradient boosting with cache-optimization for very large datasets [12].

The variables extracted from the World Bank database are summarized in TABLE 01.

Table 1: Variables extracted from World Bank database.

	Variable
1	ARI treatment (% of children under 5 taken to a health provider)
2	Children with fever receiving antimalarial drugs (% of children under age 5 with fever)
3	Diarrhea treatment (% of children under 5 receiving oral rehydration and continued feeding)
4	Diarrhea treatment (% of children under 5 who received ORS packet)
5	Incidence of malaria (per 1,000 population at risk)
6	Incidence of tuberculosis (per 100,000 people)
7	Intermittent preventive treatment (IPT) of malaria in pregnancy (% of pregnant women)
8	Malaria cases reported
9	Tuberculosis case detection rate (%, all forms)
10	Tuberculosis death rate (per 100,000 people)
11	Tuberculosis treatment success rate (% of new cases)
12	Use of insecticide-treated bed nets (% of under-5 population)
13	Female headed households (% of households with a female head)
14	GNI per capita, Atlas method (current US$)
15	Labor force, female (% of total labor force)
16	Labor force, total
17	Literacy rate, adult female (% of females ages 15 and above)
18	Literacy rate, adult male (% of males ages 15 and above)
19	Literacy rate, adult total (% of people ages 15 and above)
20	Literacy rate, youth male (% of males ages 15-24)
21	Literacy rate, youth total (% of people ages 15-24)
22	Poverty headcount ratio at national poverty line (% of population)
23	Primary completion rate, female (% of relevant age group)
24	Primary completion rate, male (% of relevant age group)
25	Primary completion rate, total (% of relevant age group)
26	Public spending on education, total (% of GDP)
27	Ratio of young literate females to males (% ages 15-24)
28	School enrollment, primary (% gross)
29	School enrollment, primary (% net)
30	School enrollment, primary, female (% gross)
31	School enrollment, primary, female (% net)
32	School enrollment, primary, male (% gross)
33	School enrollment, primary, male (% net)
34	School enrollment, secondary (% gross)
35	School enrollment, secondary (% net)
36	School enrollment, secondary, female (% gross)
37	School enrollment, secondary, female (% net)
38	School enrollment, secondary, male (% gross)
39	School enrollment, secondary, male (% net)
40	School enrollment, tertiary (% gross)
41	School enrollment, tertiary, female (% gross)

	Variable
42	Unemployment, female (% of female labor force)
43	Unemployment, male (% of male labor force)
44	Unemployment, total (% of total labor force)
45	Immunization, BCG (% of one-year-old children)
46	Immunization, DPT (% of children ages 12-23 months)
47	Immunization, HepB3 (% of one-year-old children)
48	Immunization, Hib3 (% of children ages 12-23 months)
49	Immunization, measles (% of children ages 12-23 months)
50	Immunization, measles second dose (% of children by the nationally recommended age)
51	Immunization, Pol3 (% of one-year-old children)
52	Diabetes prevalence (% of population ages 20 to 79)
53	Prevalence of current tobacco use (% of adults)
54	Prevalence of current tobacco use, females (% of female adults)
55	Prevalence of current tobacco use, males (% of male adults)
56	Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)
57	Total alcohol consumption per capita, female (liters of pure alcohol, projected estimates, female 15+ years of age)
58	Total alcohol consumption per capita, male (liters of pure alcohol, projected estimates, male 15+ years of age)
59	Consumption of iodized salt (% of households)
60	Exclusive breastfeeding (% of children under 6 months)
61	Infant and young child feeding practices, all 3 IYCF (% children ages 6-23 months)
62	Number of people who are undernourished
63	Prevalence of anemia among children (% of children ages 6-59 months)
64	Prevalence of anemia among non-pregnant women (% of women ages 15-49)
65	Prevalence of anemia among pregnant women (%)
66	Prevalence of anemia among women of reproductive age (% of women ages 15-49)
67	Prevalence of overweight (% of children under 5)
68	Prevalence of overweight, female (% of children under 5)
69	Prevalence of overweight, male (% of children under 5)
70	Prevalence of severe wasting, weight for height (% of children under 5)
71	Prevalence of severe wasting, weight for height, female (% of children under 5)
72	Prevalence of severe wasting, weight for height, male (% of children under 5)
73	Prevalence of stunting, height for age (% of children under 5)
74	Prevalence of stunting, height for age, female (% of children under 5)
75	Prevalence of stunting, height for age, male (% of children under 5)

	Variable
76	Prevalence of undernourishment (% of population)
77	Prevalence of underweight, weight for age (% of children under 5)
78	Prevalence of underweight, weight for age, female (% of children under 5)
79	Prevalence of underweight, weight for age, male (% of children under 5)
80	Prevalence of wasting, weight for height (% of children under 5)
81	Prevalence of wasting, weight for height, female (% of children under 5)
82	Prevalence of wasting, weight for height, male (% of children under 5)
83	Vitamin A supplementation coverage rate (% of children ages 6-59 months)
84	Hospital beds (per 1,000 people)
85	Nurses and midwives (per 1,000 people)
86	Physicians (per 1,000 people)
87	Specialist surgical workforce (per 100,000 population)
88	Median annual temperature

3. RESULTS

Two heatmaps were generated, before and after cleaning and pre-processing to visualize the completeness of data as depicted in FIGURE 1 and FIGURE 2. The white spaces of the map represent missing data for each variable.

The variables which had a completeness of more than 70% after data cleaning and pre-processing are summarized in TABLE 02.

The Spearman correlation coefficients for the variables selected after preprocessing were calculated. The correlation matrix for the correlation coefficients is visualized in the following FIGURE 03, ordered by the hierarchical clustering.

As seen in the above correlogram it is clear that multicollinearity exists between these factors.

3.1 Analysis of Factors Affecting Total Cases per Million Using Conventional Dimension-Reduction Techniques

The first ten variables which demonstrated a significant correlation with the total cases per million, ordered by the absolute value of the correlation coefficient are summarized below in TABLE 03.

The principal component analysis revealed that most of the variance of the dataset can be explained by the first few principal components, as seen in the below scree plot (FIGURE 04).

Table 2: Variables selected after cleaning and pre-processing.

	Variable
1	Total cases per million
2	Total deaths per million
3	Incidence of tuberculosis (per 100,000 people)
4	Tuberculosis case detection rate (%, all forms)
5	Tuberculosis death rate (per 100,000 people)
6	Tuberculosis treatment success rate (% of new cases)
7	GNI per capita, Atlas method (current US$)
8	Labor force, female (% of total labor force)
9	Labor force, total
10	Primary completion rate, female (% of relevant age group)
11	Primary completion rate, male (% of relevant age group)
12	Primary completion rate, total (% of relevant age group)
13	Public spending on education, total (% of GDP)
14	School enrollment, primary (% gross)
15	School enrollment, primary, female (% gross)
16	School enrollment, primary, male (% gross)
17	School enrollment, secondary (% gross)
18	School enrollment, secondary, male (% gross)
19	Unemployment, female (% of female labor force)
20	Unemployment, male (% of male labor force)
21	Unemployment, total (% of total labor force)
22	Immunization, BCG (% of one-year-old children)
23	Immunization, DPT (% of children ages 12-23 months)
24	Immunization, HepB3 (% of one-year-old children)
25	Immunization, Hib3 (% of children ages 12-23 months)
26	Immunization, measles (% of children ages 12-23 months)
27	Immunization, measles second dose (% of children by the nationally recommended age)
28	Immunization, Pol3 (% of one-year-old children)
29	Diabetes prevalence (% of population ages 20 to 79)
30	Prevalence of current tobacco use (% of adults)
31	Prevalence of current tobacco use, females (% of female adults)
32	Prevalence of current tobacco use, males (% of male adults)
33	Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)
34	Total alcohol consumption per capita, female (liters of pure alcohol, projected estimates, female 15+ years of age)
35	Total alcohol consumption per capita, male (liters of pure alcohol, projected estimates, male 15+ years of age)
36	Prevalence of anemia among children (% of children ages 6-59 months)
37	Prevalence of anemia among non-pregnant women (% of women ages 15-49)
38	Prevalence of anemia among pregnant women (%)
39	Prevalence of anemia among women of reproductive age (% of women ages 15-49)
40	Prevalence of undernourishment (% of population)
41	Nurses and midwives (per 1,000 people)
42	Median annual temeperature

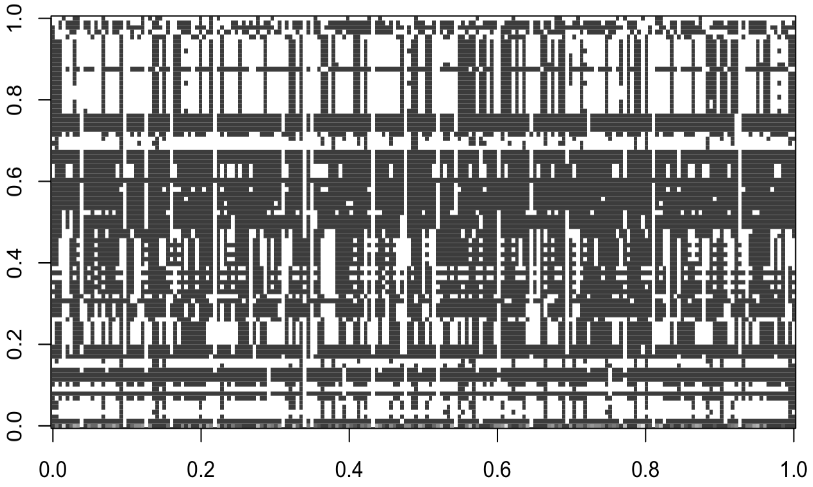

Figure 1: The heatmap generated before cleaning and pre-processing the data.

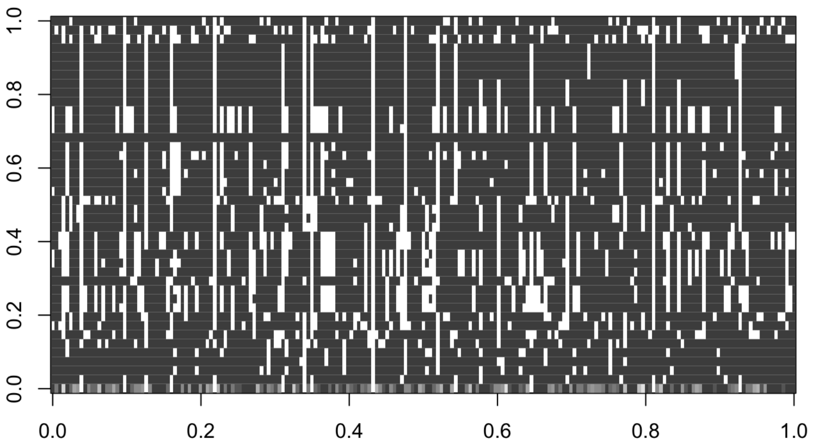

Figure 2: The heatmap generated after cleaning and pre-processing the data.

The dimensions of the PCA model which had the strongest correlations with cases contain components that have the highest explanatory power from the original predictor space. The variables identified by exploring the strength of contribution for the significant dimensions are listed below in TABLE 04.

Table 3: Variables demonstrating the highest correlation with total cases per million.

Variable Number	**Variable name**	**Correlation co-effiecient**	**p value**
1	Total cases per million	1.000	<0.001
20	Unemployment, male (% of male labor force)	0.734	<0.001
17	School enrollment, secondary (% gross)	0.723	<0.001
18	School enrollment, secondary, male (% gross)	0.705	<0.001
21	Unemployment, total (% of total labor force)	0.653	<0.001
19	Unemployment, female (% of female labor force)	0.591	<0.001
36	Prevalence of anemia among children (% of children ages 6-59 months)	-0.536	<0.001
42	Median annual temperature	-0.525	<0.001
10	Primary completion rate, female (% of relevant age group)	0.470	<0.001
7	GNI per capita, Atlas method (current US$)	0.467	<0.001
12	Primary completion rate, total (% of relevant age group)	0.455	<0.001

Table 4: Variables with the highest strength of contribution to the significant dimensions (for the analysis of total cases per million).

Variable of the PCA model - For the total cases per million	**Contribution %**
Prevalence of anemia among children (% of children ages 6-59 months)	7.53
Prevalence of anemia among pregnant women (%)	7.22
Tuberculosis case detection rate (%, all forms)	7.09
Prevalence of anemia among women of reproductive age (% of women ages 15-49)	6.91
Prevalence of anemia among non-pregnant women (% of women ages 15-49)	6.84
School enrollment, secondary (% gross) [SE.SEC.ENRR]	5.25
School enrollment, secondary, male (% gross)	5.21
Incidence of tuberculosis (per 100,000 people)	4.65
Tuberculosis death rate (per 100,000 people)	4.51
GNI per capita, Atlas method (current US$)	4.50

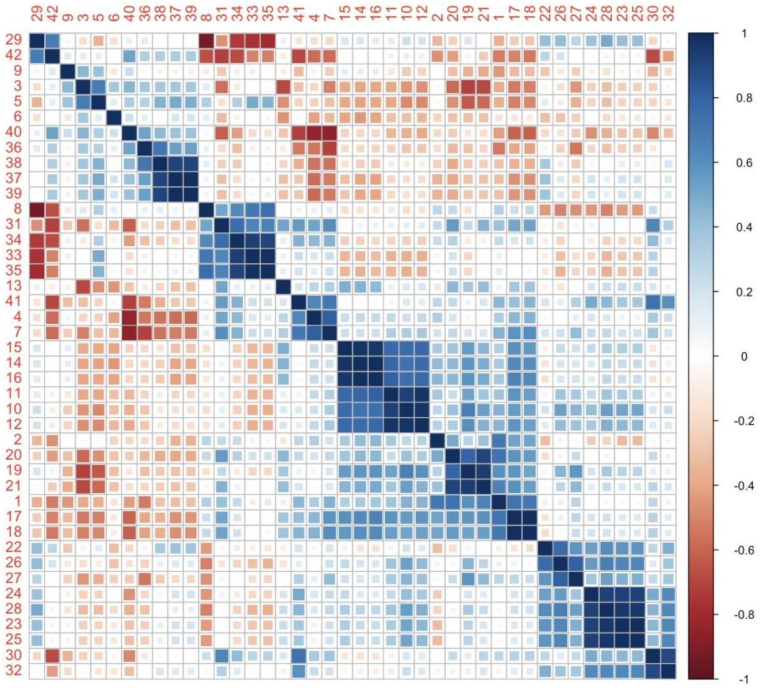

Figure 3: The correlogram of the Spearman correlation coefficients of the selected variables.

Secondly, an Elastic net model (a regularized linear regression model) was implemented for the predictor subset selection. The hyperparameters of the model, namely alpha and lambda were tuned using ten-fold repeated cross-validation, minimizing the root mean squared error. The variable importance of the model was calculated using normalized ranked regression coefficients which are listed below in TABLE 05.

The RMSE achieved for incidence data by the final model was 0.73 for alpha 1.0 and lambda 0.007 for the Elastic Net regression model.

3.2 Analysis of Factors Affecting Total Deaths per Million Using Conventional Dimension-Reduction Techniques

The same variable selection and modelling approach was carried out for the total number of deaths per million. The first ten variables which showed a significant correlation with total deaths per million, in order of the absolute value of the correlation coefficient are shown in TABLE 06.

The scree plot of the PCA carried out in a similar manner as the previous analysis is illustrated below (FIGURE 05).

The variables identified by the PCA model to have the highest explanatory power and the variables with the highest importance in the Elastic Net model are listed below in TABLE 07 and TABLE 08.

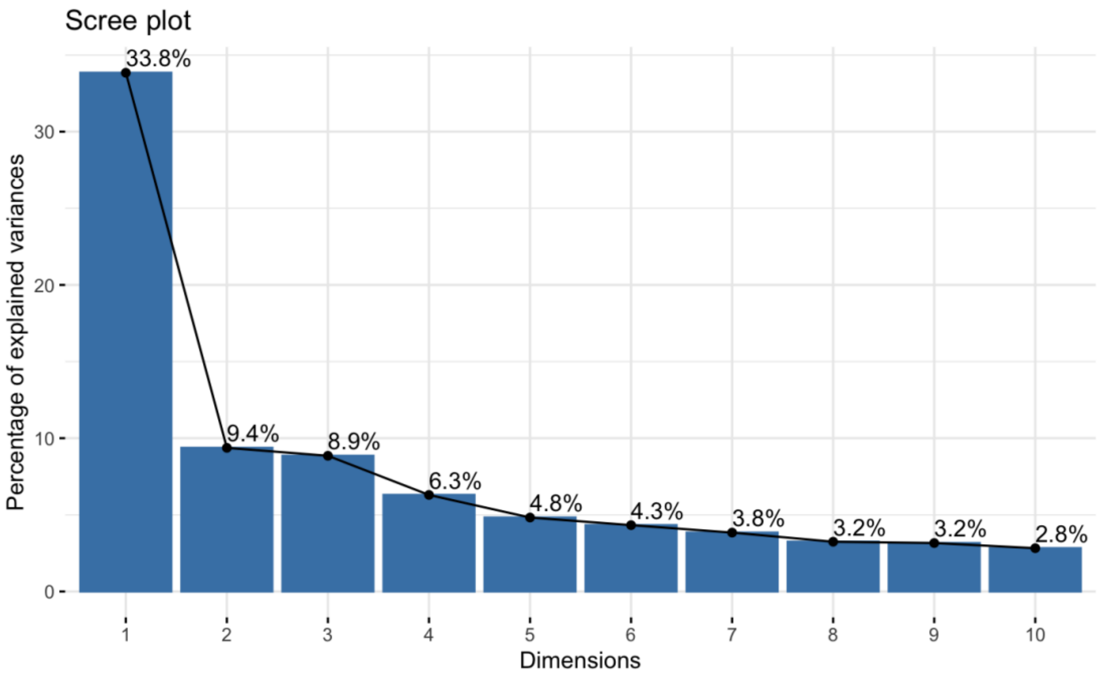

Figure 4: The scree plot representing the percentage of variance explained in each dimension.

Table 5: Variable importance of the Elastic net model (for the analysis of total cases per million).

Variable of the Elastic model	**Importance**
Prevalence of anemia among children (% of children ages 6-59 months)	100.00
Prevalence of anemia among non-pregnant women (% of women ages 15-49)	72.12
Prevalence of anemia among pregnant women (%)	43.75
Prevalence of current tobacco use, females (% of female adults)	37.61
Diabetes prevalence (% of population ages 20 to 79)	22.90
Tuberculosis treatment success rate (% of new cases)	20.39
Unemployment, total (% of total labor force)	19.76
Total alcohol consumption per capita, female (liters of pure alcohol, projected estimates, female 15+ years of age)	18.70
Immunization, measles second dose (% of children by the nationally recommended age)	13.20

Table 6: Variables demonstrating the highest correlation with total deaths per million.

Variable number	Variable name	Correlation coefficient	p value
17	School enrollment, secondary (% gross)	0.525	<0.001
18	School enrollment, secondary, male (% gross)	0.515	<0.001
20	Unemployment, male (% of male labor force)	0.494	<0.001
42	temp	-0.463	<0.001
16	School enrollment, primary, male (% gross)	0.445	<0.001
14	School enrollment, primary (% gross)	0.428	<0.001
21	Unemployment, total (% of total labor force)	0.397	<0.001
11	Primary completion rate, male (% of relevant age group)	0.360	<0.001
37	Prevalence of anemia among non-pregnant women (% of women ages 15-49)	-0.358	<0.001
39	Prevalence of anemia among women of reproductive age (% of women ages 15-49)	-0.352	<0.001

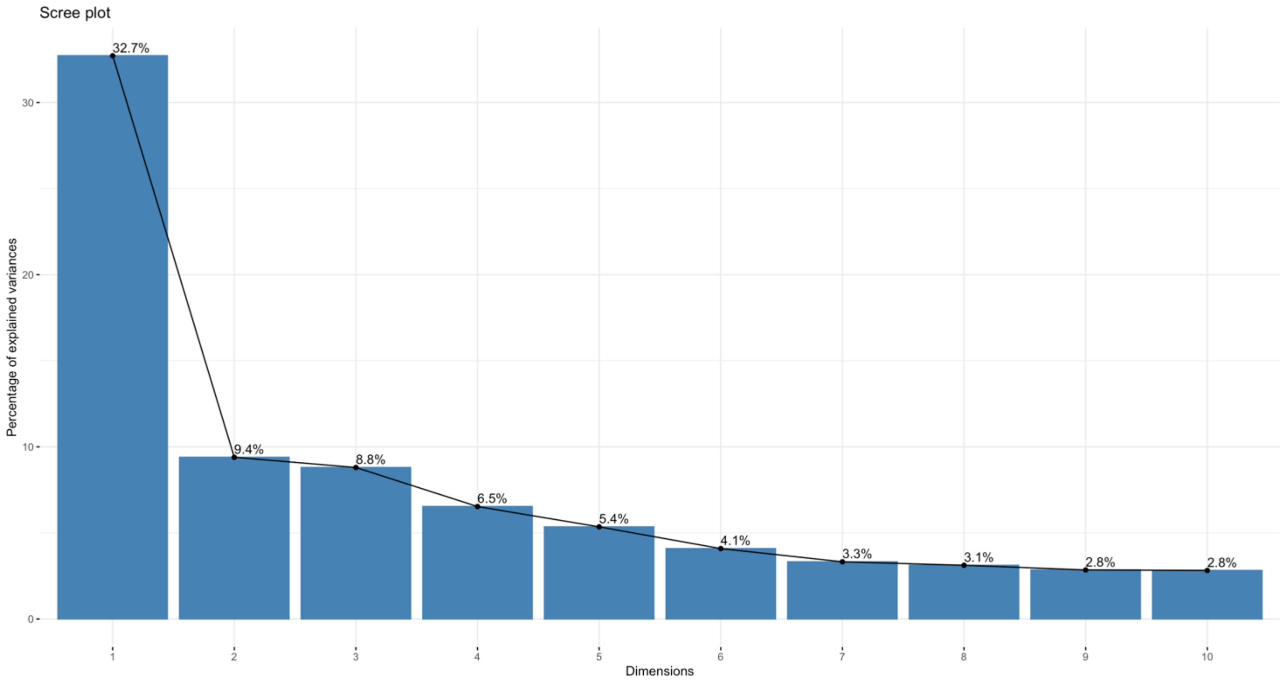

Figure 5: A scree plot representing the percentage of variance explained in each dimension (for the analysis of total deaths per million).

Table 7: Variables with the highest strength of contribution to the significant dimensions (for the analysis of total deaths per million).

Variable of the PCA model - For the total deaths per million	**Contribution %**
Prevalence of anemia among pregnant women (%)	6.61
Prevalence of anemia among children (% of children ages 6-59 months)	6.56
Prevalence of anemia among women of reproductive age (% of women ages 15-49)	6.49
Prevalence of anemia among non-pregnant women (% of women ages 15-49)	6.43
Immunization, Hib3 (% of children ages 12-23 months)	5.06
Immunization, DPT (% of children ages 12-23 months)	4.94
Tuberculosis case detection rate (%, all forms)	4.72
School enrollment, secondary (% gross)	3.97
School enrollment, secondary, male (% gross)	3.94
Incidence of tuberculosis (per 100,000 people)	3.58

Table 8: Variable importance of the Elastic net model (for the analysis of total deaths per million).

Variable of the Elastic model	**Importance**
Prevalence of anemia among children (% of children ages 6-59 months)	100
Prevalence of anemia among pregnant women (%)	95.63
Total alcohol consumption per capita, female (liters of pure alcohol, projected estimates, female 15+ years of age)	86.54
Prevalence of current tobacco use, females (% of female adults)	82.70
Immunization, DPT (% of children ages 12-23 months)	80.86
Unemployment, female (% of female labor force)	72.55
Tuberculosis treatment success rate (% of new cases)	53.56
Diabetes prevalence (% of population ages 20 to 79)	41.16
Prevalence of undernourishment (% of population)	23.44
Immunization, BCG (% of one-year-old children)	17.25

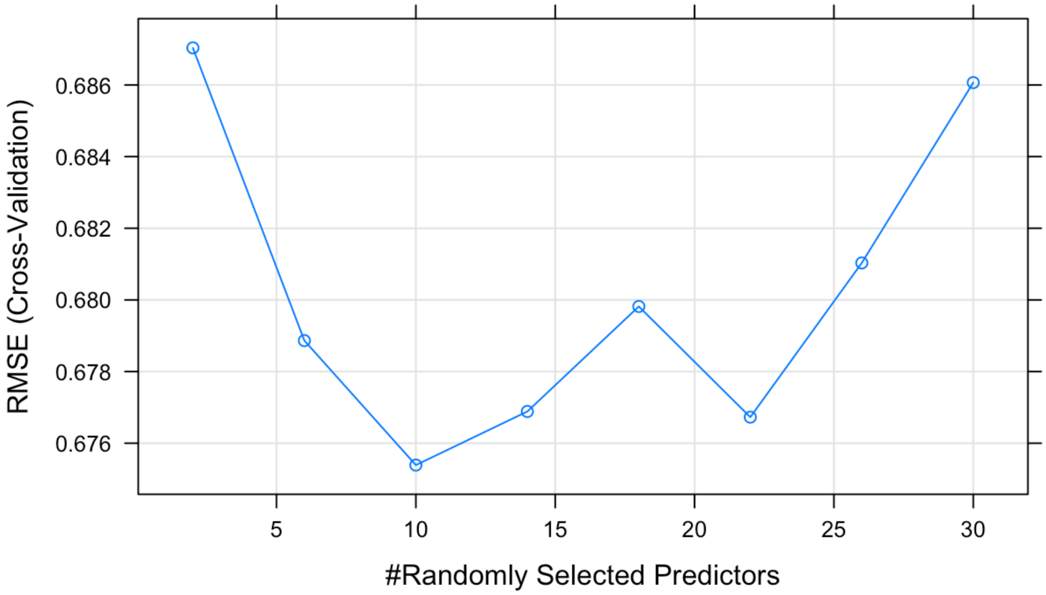

Figure 6: Cross-validation results for COVID-19 cases – Random Forests model.

For mortality data, the average RMSE was 0.76 for alpha 0.22 and lambda 0.12 for the Elastic Net regression model.

3.3 Comparison to Bootstrap-Aggregating Decision Trees and Xgboost Algorithm

Random Forest models: The Random Forests model was able to achieve an RMSE of 0.67 for 10 randomly selected variables for COVID-19 incidence and an RMSE of 0.72 for 6 randomly selected variables for COVID-19 mortality. The cross-validation RMSE for different random predictor numbers are shown in FIGURE 6 and FIGURE 7.

The relative importance of the predictors of the Random Forests models is summarized in TABLE 9.

XG Boost models: The XGBoost model could achieve an average RMSE of 0.69 across all the cross-validation folds for the final hyperparameter set for the incidence data and average RMSE of 0.70 for mortality. The parameter tuning plots are visualized in FIGURE 8 and FIGURE 9.

The relative importance of the predictors of the XGBoost models is summarized in TABLE 10.

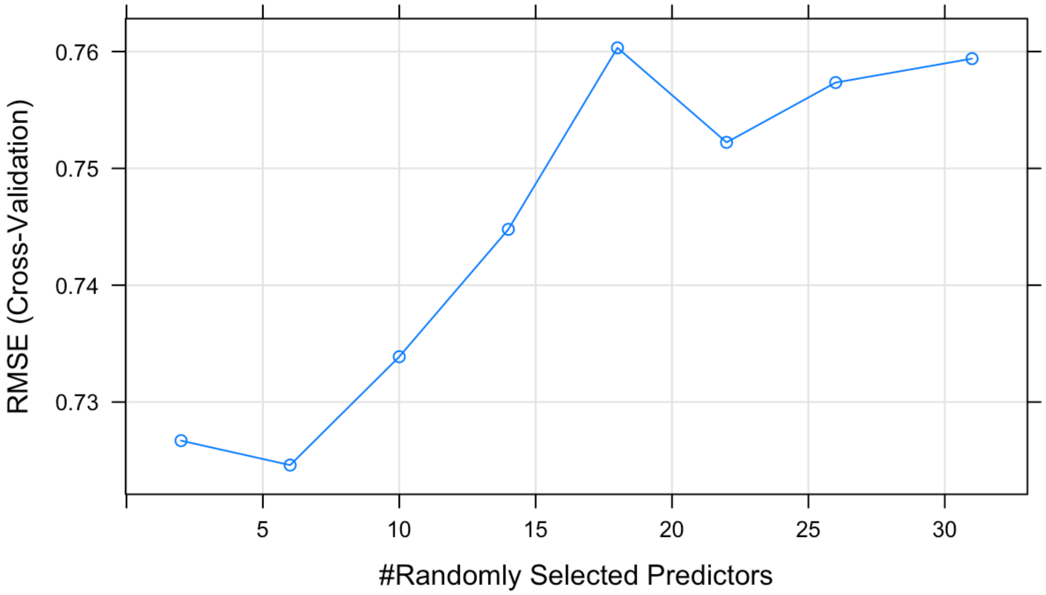

Figure 7: Cross-validation results for COVID-19 deaths – Random Forests model.

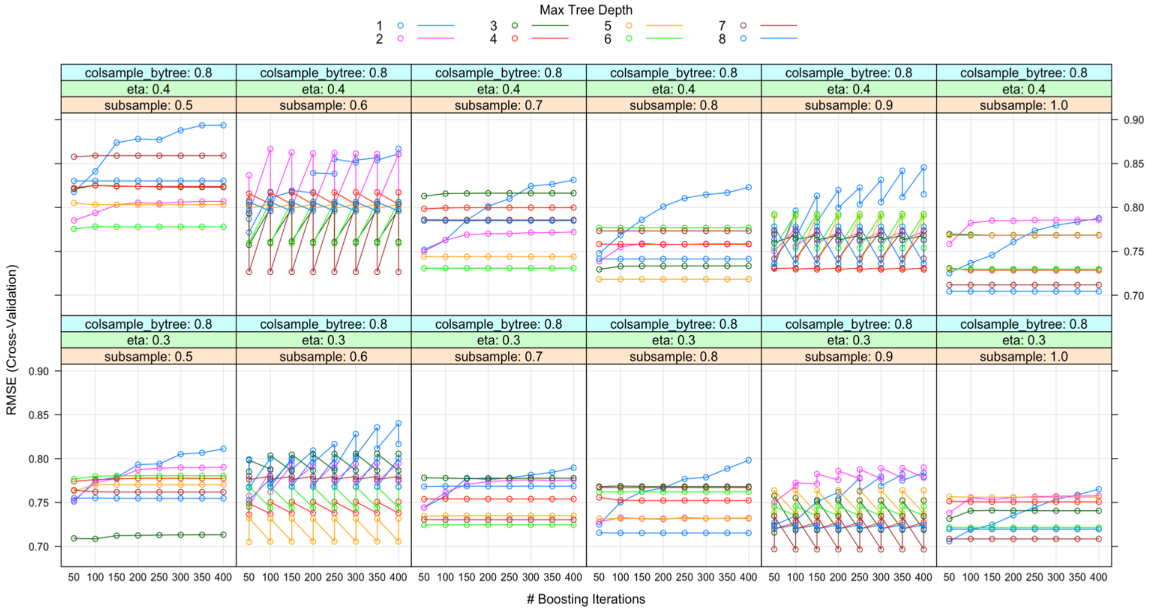

Figure 8: Cross-validation results for COVID-19 cases – XGBoost model.

Table 9: Variable importance of the Random Forest (for the analysis of total cases and deaths per million).

Variable – Random Forest model for cases	**Importance**	**Variable – Random Forest model for deaths**	**Importance**
Prevalence of anemia among children (% of children ages 6-59 months)	100	Prevalence of anemia among children (% of children ages 6-59 months)	100.00
Prevalence of anemia among pregnant women (%)	57.80	Prevalence of anemia among pregnant women (%)	92.88
Tuberculosis death rate (per 100,000 people)	39.93	Prevalence of anemia among women of reproductive age (% of women ages 15-49)	70.23
Prevalence of current tobacco use, females (% of female adults)	34.78	Prevalence of anemia among non-pregnant women (% of women ages 15-49)	55.84
Prevalence of anemia among women of reproductive age (% of women ages 15-49)	29.93	Immunization, DPT (% of children ages 12-23 months)	39.33

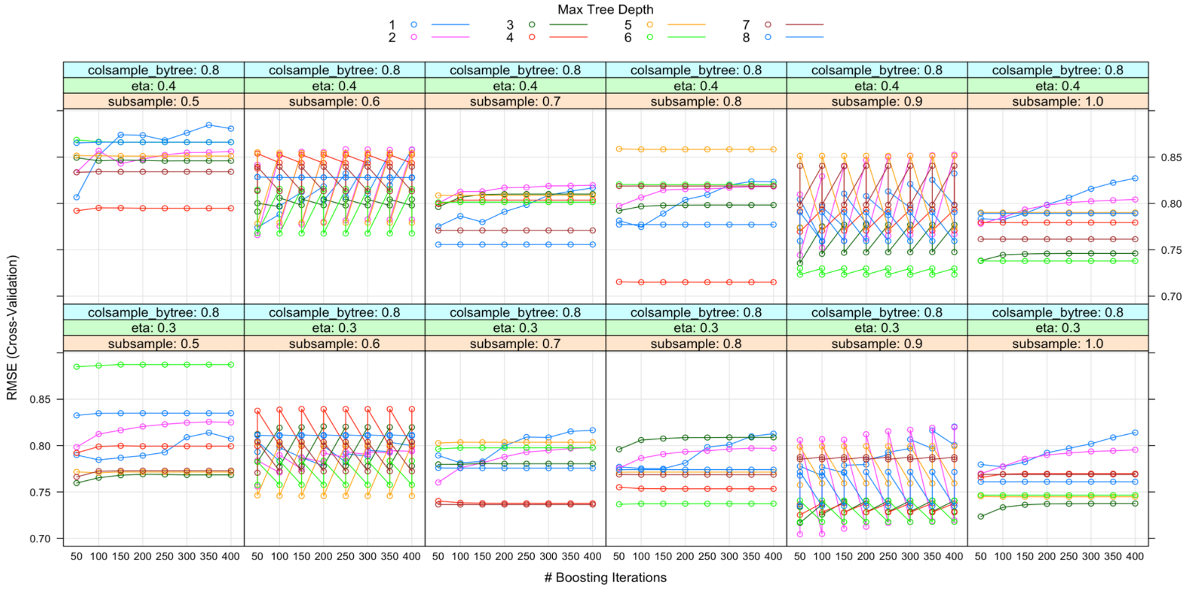

Figure 9: Cross-validation results for COVID-19 deaths – XGBoost model.

Table 10: Variable importance of the XGBoost (for the analysis of total cases and deaths per million).

Variable – XGBoost model for cases	**Importance**	**Variable – XGBoost model for deaths**	**Importance**
Prevalence of anemia among pregnant women (%)	100	Prevalence of anemia among children (% of children ages 6-59 months)	100.00
Prevalence of anemia among children (% of children ages 6-59 months)	63.77	Prevalence of anemia among pregnant women (%)	36.59
Unemployment, total (% of total labor force)	36.23	Prevalence of anemia among women of reproductive age (% of women ages 15-49)	32.10
Prevalence of current tobacco use, females (% of female adults)	30.15	Prevalence of anemia among non-pregnant women (% of women ages 15-49)	23.84
Tuberculosis death rate (per 100,000 people)	22.89	Immunization, DPT (% of children ages 12-23 months)	22.29

4. DISCUSSION

It is evident by observing the correlogram that a significant amount of multicollinearity exists in the explored variable space. Therefore, considering statistically significant correlations of those individual factors with the total incidence and mortality is likely to introduce bias, due to the possibility of multiple interactions. Furthermore, it is possible for some spurious correlations to emerge when exploring a large predictor space.

In order to minimize this bias and to detect the individual variables likely to exert an independent impact on COVID-19 incidence and mortality, two different conventional dimension reduction and feature subset selection methods, which operate in fundamentally different ways were utilized, and were compared with decision-tree based models. The intersect variable sets of the multiple models were considered to encompass the most impactful predictors of COVID-19 incidence and mortality.

The Random Forest model and the XGboost model outperformed the PCA and Elastic Regression models in both COVID-19 incidence and mortality prediction in terms of root mean squared value and proved beneficial in identifying the most important predictor variables.

Of these predictors, it was surprisingly noted that the prevalence of anaemia had a negative correlation with the COVID-19 incidence and mortality, which was detected independently by all the models. Interestingly, this finding was evident across multiple age groups.

A study by Dinevari et al. concluded that anaemia was associated with poor outcomes evidenced by significantly high mortality, ventilator requirement and ICU admissions in COVID-19 patients

[15]. However, a significant number of confounding variables such as old age, and the prevalence of other co-morbidities were present among the anaemic group. Although the final multivariate logistic regression model was adjusted for the significant variables in the univariate models, this approach is not guaranteed to eliminate the problem of multicollinearity. Gaetano et al. identified anaemia as a common manifestation of COVID-19 and stated that although anaemia does not directly influence mortality, it usually affects elderly, frail patients and can negatively influence their quality of life [16].

The studies done previously have compared anaemic and non-anaemic individuals of the same community, at the time of COVID-19 infection whereas, our study explored the association of pre-existing anaemia and the incidence and mortality of COVID-19. The prevalence used in this study is the aggregated value across three years by nearest-neighbour imputation, before the COVID-19 pandemic (over 2017, 2018 and 2019) for each country. Further studies into the mechanism in which pre-existing anaemia reduces the incidence of COVID-19 may be a way forward in understanding this complex disease. A possibility may be put forward that the risk of thrombosis is reduced in patients with anaemia and hence the chain of events initiated after infection with COVID-19 may be impeded.

In addition, it was also found that the percentage of children given DTP immunization showed a negative linear relationship with COVID-19 mortality. This finding is in consensus with multiple studies which describe cross-reactive immunity from these vaccines as a possible mechanism [6], [17], [18]. A study by Monereo-Sanchez et al. reports that patients, who received diphtheria and tetanus vaccines were less likely to develop severe COVID-19 infection [6]. A molecular-level analysis was done by Reche et al., which found that there was significant cross-reactivity between diphtheria, tetanus and pertussis combination vaccine [17]. This was also supported by a study by Mysore et al [18].

5. CONCLUSION

This study highlights the negative association of anaemia seen across multiple age groups and DTP immunization with the incidence and mortality due to COVID-19. Both the Random Forest model and XGBoost model outperformed conventional regularized regression methods in predicting the COVID-19 incidence and mortality.

References

[1] https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum

[2] Nikoloski Z, Alqunaibet AM, Alfawaz RA, Almudarra SS, Herbst, et al., COVID-19 and Non-communicable Diseases: Evidence From a Systematic Literature Review. BMC Public Health. 2021;21:1068.

[3] Otero JA, Figuero LSB, Mattín MG, Martín IU, Morais PC, et al. The Nutritional Status of the Elderly Patient Infected With COVID-19: The Forgotten Risk Factor?. Current Medical Research and Opinion. 2021; 37: 549–554.

[4] Li T, Zhang Y, Gong C, Wang J, Liu B, et al. Prevalence of Malnutrition and Analysis of Related Factors in Elderly Patients With COVID-19 in Wuhan, China. European Journal of Clinical Nutrition. 2020;74: 871–875.

[5] Malki Z, Atlam El-S, Hassanien AE, Dagnew G, Elhosseini MA, et al. Association Between Weather Data and COVID-19 Pandemic Predicting Mortality Rate: Machine Learning Approaches. Chaos Solitons Fractals. 2020;138:110137.

[6] Monereo-Sánchez J, Luykx JJ , Pinzón-Espinosa J, Richard G, Motazedi E, et al. Diphtheria and Tetanus Vaccination History Is Associated With Lower Odds of COVID-19 Hospitalization. Frontiers in immunology. 2021;12: 3934.

[7] https://databank.worldbank.org/metadataglossary/health-nutrition-and-population-statistics/series

[8] https://ourworldindata.org/covid-stringency-index

[9] Pearson RK, Neuvo Y, Astola J, Gabbouj M. Generalized Hampel Filters. Eurasip Journal on Advances in Signal Processing. vol. 2016; 1: 87.

[10] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B: Statistical Methodology. 2005;67: 301–320.

[11] L. Breiman. Random Forests. Machine Learning. 2001; 45: 5–32.

[12] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA. 2016: 785–794

[13] Wold S, Esbensen K, Geladi P. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems. 1987; 2: 37–52.

[14] https://www.sciencedirect.com/science/article/pii/S1877705817341474

[15] Dinevari MF, Somi MH, Majd ES, Farhangi M A, Nikniaz Z. Anemia Predicts Poor Outcomes of COVID-19 in Hospitalized Patients: A Prospective Study in Iran. BMC Infectious Diseases. 2021;21:170.

[16] Bergamaschi G, Andreis FBd, Aronico N, Lenti MV, Barteselli C, et al. Anemia in Patients With COVID-19: Pathogenesis and Clinical Significance. Clinical and Experimental Medicine. 2021;21: 239–246.

[17] Reche P. A. Potential Cross-Reactive Immunity to SARS-CoV-2 From Common Human Pathogens and Vaccines. Frontiers in Immunology. 2020;11: 586984.

[18] 18. Mysore V, Cullere X, Settles ML, Ji X, Kattan MW, et al. Protective Heterologous T Cell Immunity in COVID-19 Induced by the Trivalent MMR and Tdap Vaccine Antigens. Med (New York, N.Y.). 2021;2:1050-1071.e7.