Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 10, No. 3, 2022, 345-360.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Upgrading Human Development Index to Control Pandemic Mortality Rates: a Data Mining Approach to COVID-19

Saba Sareminia[1,2*]

1. Department of Industrial and Systems Engineering, Isfahan University of Technology, Isfahan, Iran
2. Center for optimization and intelligent decision making in healthcare systems (COID-Health), Isfahan University of Technology, Isfahan, Iran.

Article Info

Article History:
Received 09 January 2022
Revised 23 March 2022
Accepted 04 May 2022

DOI:10.22044/jadm.2022.11503.2307

Keywords:
Coronavirus Disease (COVID-19), Pandemics, Ensemble Data Mining Methods, Human Development Index.

**Corresponding author: s.sareminia@iut.ac.ir (S. Sareminia).*

Abstract

In recent years, the occurrence of various pandemics (COVID-19, SARS, etc.) and their widespread impacts on human life have led researchers to focus on their pathology and epidemiology components. One of the most significant inconveniences of these epidemics is the human mortality rate, which has had highly social adverse effects. This work, in addition to the major attributes that affect the COVID-19 mortality rate (health factors, people-health status, and climate) and the social and economic components of the social studies. These components have been extracted from the countries' human development index (HDI), and the effect of the level of social development on the mortality rate has been investigated using ensemble data mining methods. The results obtained indicate that the level of community ***education*** has the highest effect on the disease mortality rate. In a way, the extent of its effect is much higher than the environmental factors such as the ***air temperature*** and ***regional health*** factors, and the community ***welfare***. The impact of this factor is probably due to "the ability of knowledge-based societies on managing the crises", "their attention to health advisories", "a lower involvement of rumors", and consequently, "a lower incidence of mental health problems". This work shows the impact of education on reducing the severity of the crisis in the communities and opens a new window in terms of the cultural and social factors in the interpretation of the medical data. Furthermore, according to the results and comparing different types of single and ensemble data mining methods, the application of ensemble data mining methods VOTE, etc.) in terms of classification accuracy and prediction error has the best result.

1. Introduction

Due to the prevalence of viral diseases in recent years, the social effects of epidemic mortality have become one of the most significant concerns of human societies. As the pandemic spreads globally, the efforts to support healthcare systems overwhelmed by the COVID-19 patients have become a public health priority. Despite these efforts, more than 1,900,000 lives have been lost from this virus at the time of writing [1]. The rate of new cases and deaths in the countries is very different, despite having similar policies such as social distancing. Some of this variation may relate to the social contextual factors and the adaptation of social recommendations to the culture of countries (e.g. physical distancing, washing hands frequently, wearing a facemask) [2, 3]. These cultural differences, along with social inequalities in access to health care, have made them more vulnerable [4]. But what factors can affect the mortality rate of this disease?

According to the previous studies, various factors can influence coronavirus disease mortality. Some

researchers have focused on the personal factors such as physical condition, age, and underlying disease [5-7], mental and personality traits [8, 9], environmental and climatic conditions of the region [10, 11], economic conditions, welfare, and the level of public health [7, 8, 12, 13] and type of culture and level of education [12]. Most research works in this area have focused more on the medical factors and climatic conditions, and have examined the effectiveness of one or more components individually. In none of them, the medical factors along with the environmental and social factors have been studied simultaneously to find the superiority of these factors in shaping the mortality rate.

Accordingly, the variety of components identified in this issue is very wide and increasing. Therefore, identifying more effective component(s) and finding hidden patterns between them have special importance to be included in the process of reducing the mortality rate of this disease.

The purpose of this work is to identify the causative factors, identify the factor(s) with the highest impact level, and the hidden pattern of factors in shaping the mortality rate of this disease in different countries by investigating the death rate in the last 11 months in all countries. The factors considered in this work are climate, level of welfare, and education.

One of the most acceptable indicators in the social, medical, and wealth fields is the Human Development Index (HDI), which is calculated and reported every year for each country [14, 15]. Therefore, this indicator was used, and to achieve the research purpose, the data mining techniques were utilized, and the hidden pattern between the average mortality rate and the mentioned factors was identified. The basic dataset is a joint between four datasets from different sources about the mortality rate in different countries that have been compiled since the beginning of the epidemic [1], the climatic characteristics of the countries [16], and the HDI of the countries[1] [17].

This paper is divided into five sections. The current introduction section is followed by the literature review, which provides an analytical review of the research works conducted in the field of "factors related to coronavirus disease mortality rate" and categorizes the identified influencing factors. The third and fourth sections are the data mining process and results; in these sections, after describing the research methodology and explaining the dataset, the outcomes obtained from the data mining process are compared and discussed. Finally, in the last section, some of the extracted results and future works that come out of this data mining project are presented.

2. Literature Review

The first cases of COVID-19 were diagnosed in December 2019 in China as an acute respiratory syndrome caused by the coronavirus (SARS-CoV-2). The disease was reported to the World Health Organization (WHO) on December 31, 2020, and in March 2020, the global epidemic was officially declared [13].

With the prevalence of coronavirus disease and its widespread impact on various aspects of life, numerous studies have been conducted in this field. A number of these studies have examined the disease mortality rate and the factors affecting it. Upon a closer inspection, these factors can be classified into three groups: "individual", "environmental and climatic", and "economic, social, and cultural" factors.

The ***first group*** of factors related to the coronavirus disease mortality rate; are those that are related to people individually such as physical features and health status (age, gender, lifestyle, and underlying medical problem), personality, and mental characteristics (trust and social dependence, depression), and genetic factors (race) .These types of factors are known to affect the mortality rate of many diseases [18, 19].

The ***second group*** of factors is the environmental and climatic factors that have been identified as the air temperature and the amount of air pollution in the region. The ***third category*** is the characteristics of the studied community in terms of the living and economic status, health status, the amount of clinical care provided, level of education, and culture. The identified factors in each category can be seen in Figure 1. In the following, how to study each one of these factors is considered in different research works.

In one study, Zaveri and Chouhan show the impacts of "the community's age distribution" on mortality rate diversity in the European and South-East Asian countries by age structure, especially the percentage share of child and youth population.

[1] The human development index (HDI Index) of countries has been used to determine the level of welfare, education, and general hygiene of a society.

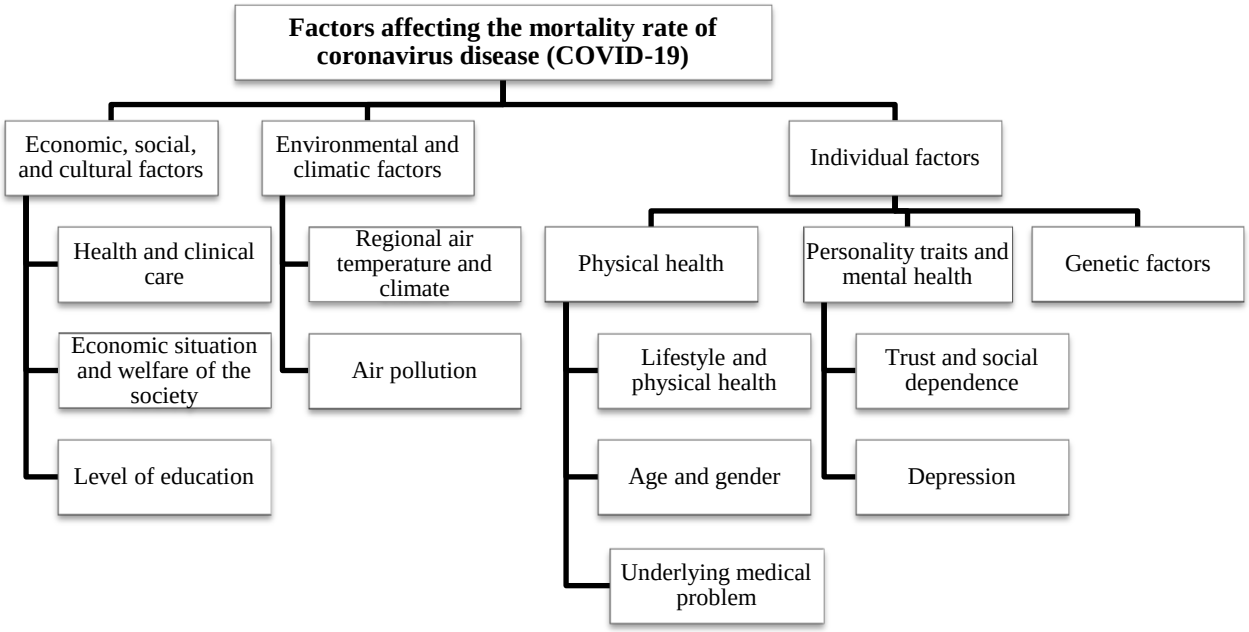

Figure 1. Classification of effective factors on the coronavirus disease (COVID-19) mortality.

They revealed that the COVID-19 mortality rates are substantially higher in highly developed European countries compared to the South-East Asian countries due to having a higher average age and lower child and youth population; therefore, age is an important factor in the disease mortality rate [5]. In this study, the composite Z scoring method has been used. In another study, which was conducted in Mexico, the factors such as the high population average age having underlying diseases (diabetes, obesity, immunosuppression, and chronic renal insufficiency) have been identified as the most important factors in increasing the risk of disease mortality by several classification methods[6].In these studies, only the physical individual criteria have been considered, and the other factors have not been analyzed. Both papers have limited their studies to one or more specific communities (the first study focused only on the fatality rate in Europe and South-East Asia, and the second study limited its investigation to Mexico).

Harris *et al*. have investigated some ethical, neighboring, and welfare factors correlating with the COVID-19 deaths in London during the initial pandemic within the UK. This study shows that a higher COVID-19 mortality rate continues to be associated with Asian and Black ethnic groups, very large households, socio-economic disadvantage, and less from younger age groups. In other words, wealth or deprivation, age, and ethnicity are the critical risk factors associated with higher mortality rates from COVID-19 [7].

This study, although has focused on the social and economic components, in addition to the physical individual components, most of its outcomes have been on the statistical analysis of the spatial patterns of the disease. The result of another study in the United States has revealed that socio-economic factors have an important role in the coronavirus disease (COVID-19) prevalence and mortality. A lower education level, a higher proportion of black residents, poverty rates, and median income are associated with the COVID-19 mortality rates.

The research methodology in this paper is Hierarchical linear mixed models [12]. Elgar and his coworkers noted that income inequality and social capital (group affiliations, civic responsibility, trust, and confidence in public institutions) had affected the COVID-19 deaths in 84 countries. This study shows that the societies that are economically more unequal and underdeveloped in certain cultural and social dimensions have experienced more deaths so far during the COVID-19 pandemic. The results of this study show that societies with high social trust and group affiliation have higher mortality rates, and this may be due to the discrepancy between the behavioral style of these societies and the physical distance policies [8]. The last two papers have concentrated on the social and economic components but have limited their studies to one specific community too.

In another study, by analyzing the Twitter social network data in three different time intervals, during the pandemic, a set of feelings and opinions about this disease in 6 categories were extracted, and a significant relationship between

the patients' emotional characteristics and the mortality rates has been identified [9].

Table 1. Factors influencing Coronavirus disease mortality according to literature review.

Ref.	Title	Factors influencing Coronavirus disease mortality rate	The statistical population	Research method	Factor type						
					Environmental and climatic factors	Individual factors			Economic-social and cultural factors		
						Physically	Mentally	Genetically	National Health	Culture and education	Economic
[5]	Assessing the average age of the population at risk for COVID-19 fatalities	Age	European and South-East Asian countries	Composite Z score technique		◆					
[6]	Develop a model to predict the necessary actions for COVID-19 patients: hospitalizations, mortality, and need for an ICU or ventilator	Age, immunosuppression, chronic renal insufficiency, obesity, diabetes	Mexico	Datamining methods (logistic regression, support vector machines, random forests, and gradient-boosted decision trees)		◆					
[7]	Exploring ethical, neighboring, and welfare factor correlates of COVID-19 deaths in London using a difference across spatial boundary method	Age, wealth, deprivation, ethnicity	United Kingdom	Statistical analysis of spatial patterns		◆			◆		◆
[12]	Socio-economic status and COVID-19–related cases and fatalities	Lower education level, poverty, and a higher proportion of black residents.	United States	Hierarchical linear mixed models				◆		◆	◆
[8]	Investigating the effect of social capital, trust, and income inequality on COVID-19 deaths in 84 countries by time-series analysis	Economic inequality, social trust	All of the world	Time-series analysis			◆				◆
[9]	Monitoring dynamics of emotions during COVID-19 using social network (Twitter) data	Patients' emotional characteristics	All of the world	Natural language processing and social network analysis techniques			◆				
[20]	Exploring the efficiency of IoT (application, architecture, technology, and security) during the COVID-19 pandemic	Internet of things in health care systems	All of the world	Qualitative methods					◆		◆
[10]	Factors associated with outbreak and mortality from COVID-19 in autonomous communities of Spain.	Age, air temperature	Spain	Spearman correlation coefficient and multiple regression analysis	◆	◆					
[11]	Exploring the relationship between development density and COVID-19 morbidity and mortality rates: evidence from 1,165 metropolitan counties in the United States	Urban density	United States-Metropolises	Multi-level linear modeling	◆						
[13]	Investigating the impact of the local care environment and social characteristics on hospital mortality rate from COVID-19 in France	Clinical care	France	Poisson regression					◆		

Natural language processing and social network analysis techniques were applied in this research work and concentrated on the mental and emotional dynamics during the pandemics. In Spain, by examining the epidemiological, demographic, environmental, and health services variables of the region, some factors such as age

and air temperature have been identified as the most important factors influencing the COVID-19 mortality rate by using the Spearman correlation coefficient and multiple regression analysis [10]. In a study conducted in France on the COVID-19 patients, some factors such as social or wealth factors are ineffective to predict the patient mortality rate. However, the results of this study show that lower activity in the for-profit private sector, a higher capacity for primary intensive care, and a lower density of general practitioners lead to higher mortality [13]. In this research work, the Poisson regression was used and both of these research works were conducted in the European Countries.

Hamidi and his coworkers noted the importance of urban congestion in metropolitan areas in the United States by using Multi-Level Linear Modeling (MLM). They proposed that large metropolitan areas were the most vulnerable to the pandemic outbreak, but dense counties had a significantly lower rate of COVID-19 mortality [11]. Finally, Azan *et al*. have considered the importance of using the Internet of Things in reducing the adversity of the disease and thus reducing the mortality rate of patients [20]. However, this concept is completely dependent on the economic situation of society. In this research work, unlike others, qualitative methods have been used. As shown, the factors that affect the mortality rate of this disease are high. A summary of these studies according to the presented classification of factors (Figure 1) can be seen in Table 1. According to the research literature, while the majority of research has focused on the medical and individual components, most studies have examined the effect of one or more components on the COVID mortality rate. The majority of the studies are limited to one or more geographical areas. No research work has been done to simultaneously examine the environmental criteria alongside the social, economic, and cultural attributes. Therefore, due to the unknown nature of the problem, which is likely to be repeated, the concurrent attention to different components and identifying the most effective ones can help the countries manage such crises. In addition, most studies use one or more statistical methods to obtain results; however, in this research work, using single and ensemble data mining models, while using the best model (with higher performance) in identifying the effective components, the performance of each model was also examined.

3. Method

The reference model used during the development of this work was a standard process for data mining, which is known as CRISP-DM. The CRISP-DM methodology provides a structured approach for planning a data mining project and is a six-phase hierarchical process, divided into the following steps: business (subject) understanding, data understanding, data preparation, modeling, and evaluation and deployment [21, 22]. The sequence of the phases is not strict, and moving back and forth between the phases is always required.

The first phase focuses on understanding the subject and business requirements and transforming this knowledge into a data mining scenario and a rudimentary plan designed to achieve the objectives. The data understanding phase continues with the initial data collection. The first insight into data representation is obtained at this stage, and the data mining process is explained to extract the hidden knowledge. The data preparation phase covers all the activities required to clean and prepare the final dataset from the initial raw data according to any of the defined scenarios. In the modeling phase, various modeling techniques are applied, and their parameters are optimized. At the evaluation phase, a model (or models) was (were) built that appears to have a high performance from a data analysis perspective, and finally, depending on the project's purposes, the deployment phase can be implemented [23].

Figure 2 describes the methodology of this research work in detail and the application of all these steps in the context of this work.

To analyze and explore the collected data and to induce the data mining models, the chosen data mining software was Rapid miner. According to the purpose of the research work, the data mining process is followed by two scenarios: the first scenario seeks to identify the most important components affecting the mortality rate, and the second scenario is designed to identify the best compatible model for predicting the fatality. The scenarios and applied models can be seen in Figure 3.

In the first scenario, the label attribute (average mortality rate) is considered categorized, and the focus is on the samples that are in the lower quarter of the mortality rate. To achieve this goal, the average mortality rate was transformed from numerical to nominal by discretizing into four clusters, and consequently, this nominal attribute was converted into four binaries, and the lower

quarter of the mortality rate was determined as the label.

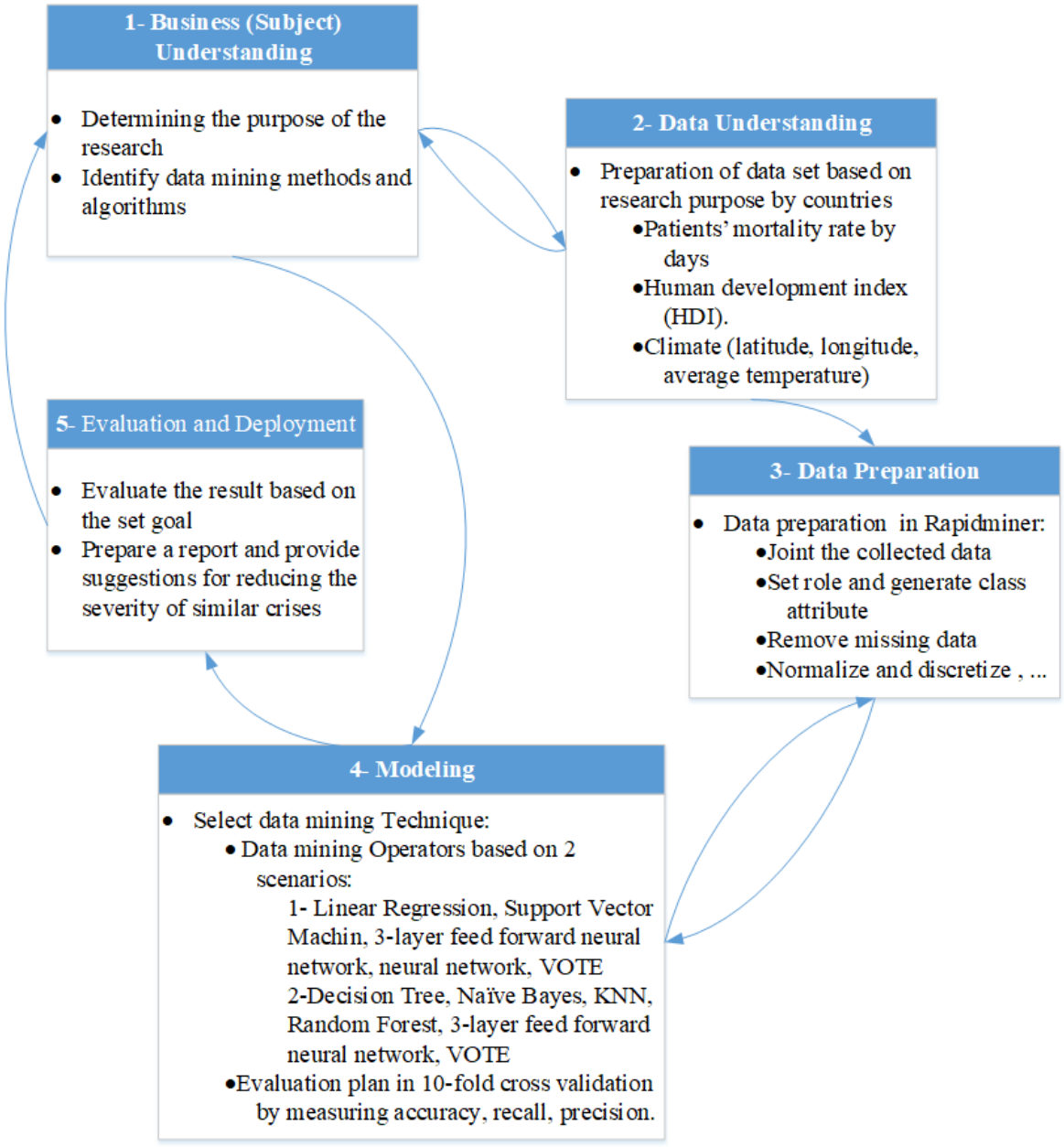

Figure 2. Research methodology based on CRISP-DM.

To analyze and explore the collected data and to induce the data mining models, the chosen data mining software was Rapid Miner. According to the purpose of the research work, the data mining process was followed by two scenarios: the ***first scenario*** was designed to identify the best compatible model for predicting fatality, and the ***second scenario*** seeks to identify the most important components affecting the mortality rate. In both scenarios, after data preparation, the compatible classification and prediction models (decision tree, Naïve Bayes, Random Forest (RF), K-nearest neighbor, naïve Bayes and 3-layer feed-foreword neural network, Support Vector Machine (SVM), linear regression, VOTE, and forward selection have been utilized. The scenarios and applied models can be seen in Figure 3.

In the first scenario, the label attribute (average mortality rate) is considered categorized, and the focus is on the samples that are in the lower quarter of the mortality rate. To achieve this goal, the average mortality rate was transformed from numerical to nominal by discretizing into four clusters, and consequently, this nominal attribute has been converted into four binaries, and the lower quarter of the mortality rate has been determined as the label.

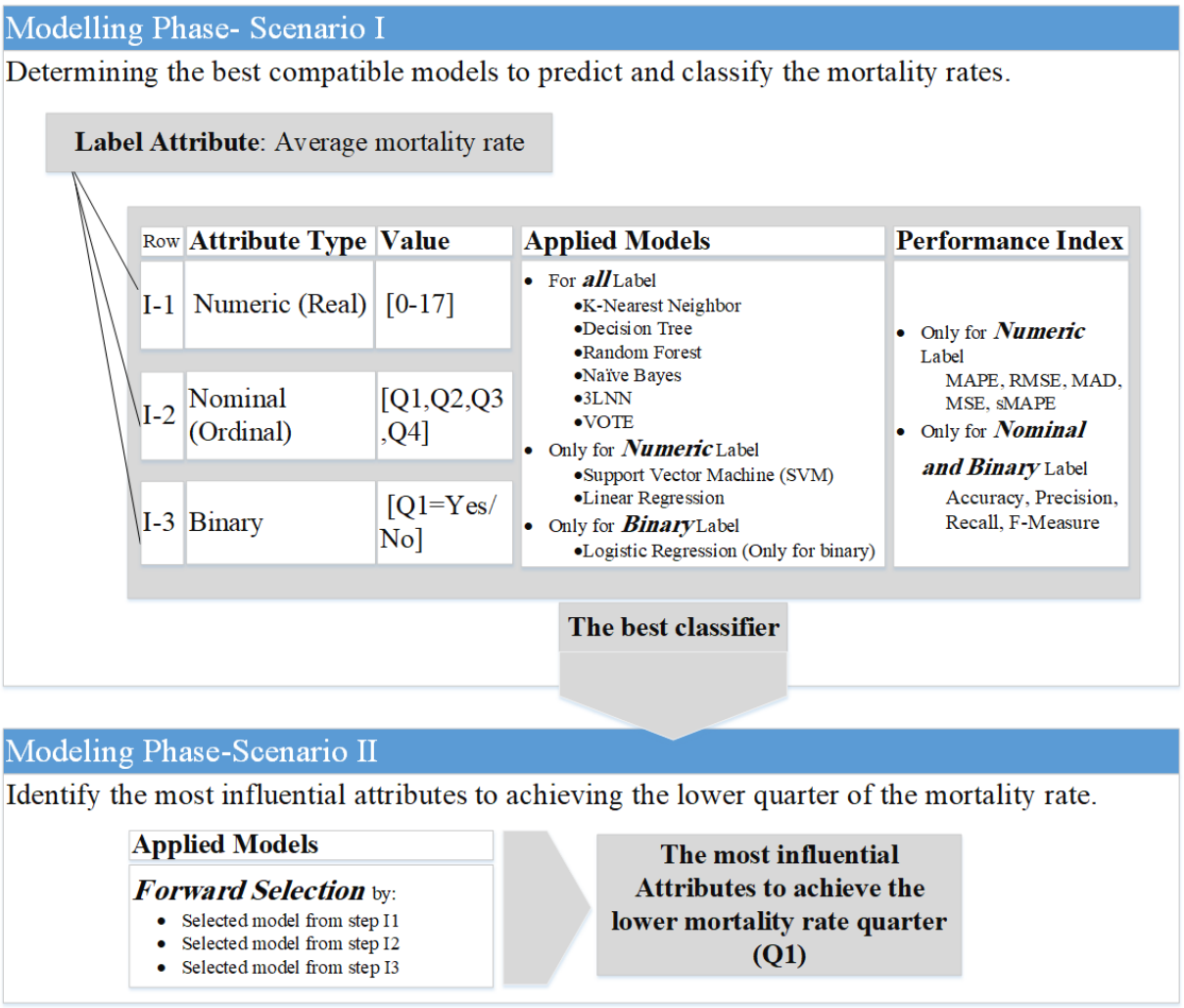

Figure 3. Research data mining scenarios and expected outcomes.

4. Result

4.1. Subject and data understanding

The recent experiences with the Coronavirus disease (COVID-19) show that most people infected with this virus will experience a mild to moderate respiratory illness and recover without requiring special treatment but older people and those with underlying medical problems are at a higher risk of death [1]. This disease has shown dissimilar mortality rates in different parts of the world, and this issue has been considered by the scientific community as to what factors affect the mortality rate variety.

By reviewing the research literature, various factors affect the mortality rate of this disease, which have been considered in some cases. This work addresses some of these factors with global indicators across the world, and the desired dataset is created by connecting the following datasets (Table 2). The first dataset was obtained from the World Health Organization website, which shows the latest news on the daily mortality rate of this disease in different countries. The next one is from the United Nations Development Programme, which publishes HDI by country each year. HDI is a cumulative measure of a country's prosperity in terms of various dimensions of human development. The main purpose of HDI is to express the importance of "human ability and skills" along with "economic growth" in the development and progress of a country. These indicators are in three dimensions: “a long and healthy life”, “being knowledgeable”, and “having an appropriate standard of living”. They have been calculated from the normalization of the geometric mean of these dimensions achieved.

The “***health dimension***” is determined by life expectancy at birth, and the “***education dimension***” is assessed by the mean of years of schooling for adults aged 25 years and more and expected years of schooling for children of school entering age. The “***standard of living dimension***” is determined by the gross national income (GNI) per capita. As one can see in Figure 4, the score

calculation process for the three HDI dimensions is presented [17].

Table 2. Specifications and resources of the dataset used to set the research dataset.

Dataset name	Ref.	Record	Attributes
Coronavirus disease (COVID-19) mortality rate in the last 11 months by day	[1]	After Aggregation (on month and country) 214	• Date • Country code • Case • Death • Mortality rate (11 attributes for any month) • WHO_region • New_cases • New_deaths
HDI by countries	[17]	196	• Country • HDI rank • HDI • SDG3-Life expectancy at birth • SDG4.3-Expected years of schooling • SDG4.6-Mean years of schooling • SDG8.5-Gross national income (GNI) per capita • GNI per capita rank minus HDI rank
Average temperature by countries (1991-2016)	[16]	192	• Country • Average yearly temperature (degrees Celsius)
Countries' Longitude and latitude	[24]	245	• Country code • Country name • Longitude • latitude

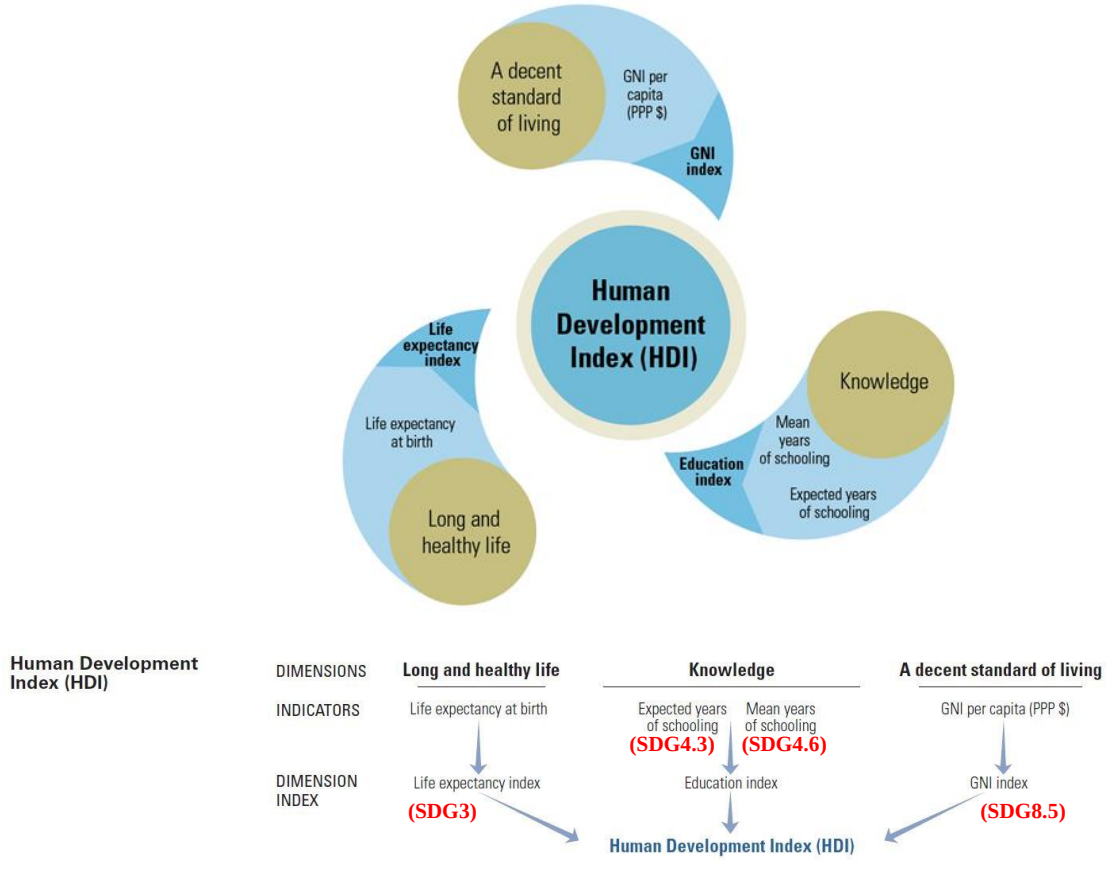

Figure 4- HDI factors [17].

The mentioned datasets were joined (right joint base on the first dataset) based on the "country name", and finally, a dataset with 215 records and 27 attributes has been created as follows (Table 3).

4.2. Data preparation

Data preparation is the process of cleaning and transforming the gathered data before modeling and analysis. It is an important step before processing and often involves transforming data,

making corrections to data (remove noises and missing data), and combining datasets to enrich data. The dataset in this work is created from the joint of four datasets, and this issue has caused inconsistencies in the final dataset.

Table 3. Dataset attributes.

Row	Attribute	Type	Specification
1	Country name	Nominal	Name of country in any dataset that is used as ID to join 4 datasets
2	Average temperature	Real	The average temperature of any country (1991-2-16)
3	ContinentExp	Polynomial	Continent name (Asia, Africa, Europe, America, Oceania)
4	Latitude	Integer	Latitude and longitude of any country
5	Longitude	Integer	
6	MT January	Real	The average mortality rate of each country in January
7	MT February	Real	The average mortality rate of each country in February
8	MT March	Real	The average mortality rate of each country in March
9	MT Q1	Real	The average mortality rate of each country in Quarter 1
10	MT April	Real	The average mortality rate of each country in April
11	MT May	Real	The average mortality rate of each country in May
12	MT June	Real	The average mortality rate of each country in June
13	MT Q2	Real	The average mortality rate of each country in Quarter 2
14	MT July	Real	The average mortality rate of each country in July
15	MT August	Real	The average mortality rate of each country in August
16	MT September	Real	The average mortality rate of each country in September
17	MT Q3	Real	The average mortality rate of each country in Quarter 2
18	MT October	Real	The average mortality rate of each country in October
19	MT November	Real	The average mortality rate of each country in November
20	MT Total Avg	Real/ Polynomial/Binary	(Class attribute) The average mortality rate of each country in 11 month
21	HDI rank	Integer	HDI rank for any country
22	HD	Real	HDI (0-1)
23	SDG3-Life expectancy at birth	Real	The number of years a newborn infant could expect to live if prevailing patterns of age-specific mortality rates at the time of birth stay the same throughout the infant's life (years: 0,100)
24	SDG4.3-Expected years of schooling	Real	Number of years of schooling that a child of school entrance age can expect to receive if prevailing patterns of age-specific enrolment rates persist throughout the child's life (0,23)
25	SDG4.6-Mean years of schooling	Real	The average number of years of education received by people ages 25 and older, converted from educational attainment levels using official durations of each level (0,15)
26	SDG8.5-Gross national income (GNI) per capita	Real	Gross national income (GNI) per capita (0,111000)
27	GNI per capita rank minus HDI rank	Integer	GNI per capita rank minus HDI rank (-100,100)

The preparation data process has been done in Rapid miner, and is as follows:

- Remove records with missing data in average temperature, latitude, and longitude attributes.
- Normalize any regular and numeric attribute to more accurately investigate the effects of the attribute on the class (label) attribute.
- Discretize the class (label) attribute by binning it into 4 categories and focusing on the first quartile (for some of the defined scenarios).

4.3. Data Modeling

Classification and prediction is a vastly used technique in health care [25-28]. Here, several classification and prediction models have been utilized to determine which attribute(s) have (have) the greatest impact on the coronavirus disease (COVID-19) mortality rate all over the world. Two scenarios are followed in this paper:

Scenario I: As mentioned before, in this scenario, while data pre-processing, the label attribute (average mortality rate) sets in three data types. The numerical (a real value from zero to 17), nominal (a nominal value; Q1, Q2, Q3, and Q4 according to the quartiles), and binary (Q1=Yes/No) attributes are defined.

Consequently, according to the label type, some consistent models have been applied and evaluated in predicting and classifying.

Scenario II: In this Scenario, the forward selection operator is applied based on the compatible superior models (from the scenario I), which are identified in each type of label. The more effective components in achieving the lower quartile are identified in this scenario.

In both scenarios, after data preparation, the compatible classification and prediction models (decision tree, Naïve Bayes, Random Forest (RF), K-nearest neighbor, Naïve Bayes and 3-layer feed-foreword neural network, Support Vector

Machin (SVM), linear regression, VOTE, and forward selection have been utilized.

According to the low number of records, the K-fold cross-validation method by stratified sampling is used. Cross-validation is a sampling method used resampling to improve the machine learning models and to avoid overfitting and underfitting on a limited data sample. The procedure has a parameter (k) that refers to the number of sections that a given data sample is to be fraction into. In the following, while describing the applied models, the results are given separately.

4.3.1. Decision tree

The J48 Decision tree algorithm partitions the dataset based on the best attribute according to their gain ratio. At each iteration, the attribute with the maximum gain ratio is chosen as the splitting attribute. The decision tree classification models are easy to interpret and are known to have comparable performance to other classification models, especially in healthcare data mining [29-32]. Given that the data mining algorithms have parameters for calculations, choosing the optimal parameter according to the data is very important. Therefore, in this research work, the parameters of the selected algorithms have been optimized; the optimal criteria of the decision tree in this research work is "criterion: Accuracy and maximal depth: 44" which has provided the highest "accuracy: 82.19%". The result of implementing this algorithm is depicted in Figure 5. The rule extracted from the decision tree indicates that the feature "SDG4.3-Expected years of schooling" has the most gained information in predicting the mortality rate, and after two layers of repetition of this attribute, the welfare of the country "SDG8.5-Gross national income (GNI) per capita" is important in determining the mortality rate. According to this rule and contrary to the results of the previous research works [10], the impact of temperature in increasing the mortality rate has not been strengthened in this rule.

4.3.2. Random forest (RF)

RF is a classifier with k (optimized parameter) separate decision trees. This model is also very considerable in the statistical research related to healthcare, and high accuracy has been reported for it [33, 34]34]. The result of this classifier is the same as the decision tree. In this research work, the optimal parameters are: "Number of trees: 18, Criterion: Accuracy and Maximal depth: 4", which have provided the highest "accuracy: 81.16%".

4.3.2. K-nearest neighbor (KNN)

The KNN classification is a lazy learner by "memorizing" the training dataset. In the KNN models, a new tuple is classified by a majority vote of its *k* neighbors, and therefore, the tuple is assigned to the class most prevalent among its *k* nearest neighbors. This algorithm has been used alternately in the medical field data analysis [35-37]. The KNN algorithm has been run using different values of the *k* parameter (in this study from k = 1 to k = 100), and the best results obtained in this research work when k = 12, with the highest "accuracy: 82.19%".

4.3.4. Naïve Bayes

The naive Bayes classifier is lazy; it is a probabilistic model based on the Bayes theorem. In this classifier, the class value and the attributes assume independence and based on the probability, the class value of a new instance is predicted. The research work has shown that the naive Bayes classifiers have comparable performance to other classification algorithms such as decision trees and neural networks besides; they produce high-performance models and can deal with large datasets [34, 36, 38, 39]. The accuracy of this classifier in this paper is 76.32%.

4.3.4. Support vector machine (SVM)

The SVM model is a supervised classification model based on machine learning and statistical dimensional theory. The important point in this algorithm is the optimal kernel selection to increase the model accuracy [40]. This model is also very considerable in statistical research related to healthcare and high accuracy is reported for it [41-43]. In this work, the algorithm was implemented with different kernels (Dot, Radial, Polynomial, Neural, and Anova), and finally, the best kernel (Radial) was selected. According to the result of this classifier, as shown in Figure 6, "SDG4.3-Expected years of schooling" has the highest weight in mortality rate prediction (weight = 20.577), and after that, "GNI per capita rank minus HDI rank" with a weight of 14.382 has the highest impact on this disease mortality.

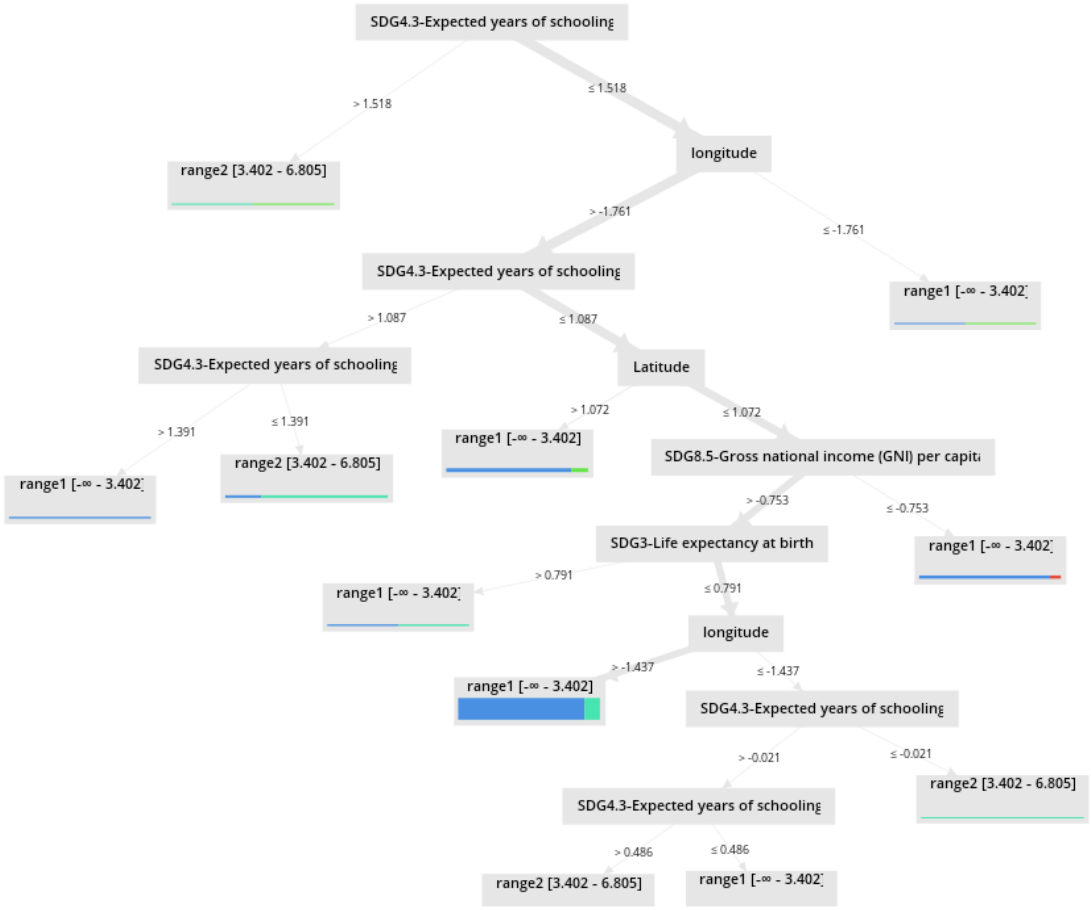

Figure 5. Predicting mortality rate class based on decision tree algorithm.

4.3.5. Linear regression

A correlation coefficient can be used to measure the intensity of the relationship between the dependent variable (mortality rate) and the independent one (HDI index, temperature, latitude, longitude …). The closer the correlation coefficient is to 1 or -1, the stronger the intensity of the linear relationship between the independent and dependent variables. Of course, if the correlation coefficient is close to 1, the direction of change of both variables is the same, which is called direct relation, and if the correlation coefficient is close to -1, the direction of change of variables will be inverse to each other, and we call it the inverse relationship. This algorithm is also very considerable in the statistical research related to healthcare, and high accuracy has been reported for it [41, 44]. In this work, a linear regression algorithm with a T-Test has been used. For $\alpha = 0.05$, the p-values of the coefficients of the independent variables are shown in Table 4 and Equation 1.

$$\begin{aligned} Total\_Avg &= 0.435 * Latitude \\ &-0.373 * Longitude \\ &+0.710 * SDG\,3 \\ &+0.925 * SDG\,4.3 \\ &-0.839 * SDG\,4.6 \\ &+1.793 \end{aligned} \quad (1)$$

According to the obtained results, "***SDG4.3 (expected year of schooling***)" has a 92% effect in the positive direction, and "***SDG4.6 (mean years of schooling)***" has an effect of 84% in the negative direction of the average mortality rate. The p-value = 1-2% also reinforces this issue. It means that a higher level of people's education in a country will lead to lower mortality rates in this disease.

4.3.6. 3-layer feed-forward neural network

The multi-layer feed-forward artificial neural network is trained with stochastic gradient downturn using back-propagation. The network can contain a large number of hidden layers consisting of neurons with "Tanh", "Rectifier", "Maxout", etc. activation functions. Based on the previous research, a 3-layer network has presented more accurate results [45-47]; in this paper, 3-

layer feed-forward neural network has been used, and the optimal parameters in the first scenario are "Activation: Tanh, L1: 0.1, L2:0" with accuracy: 82.34%, and in the second scenario are "Activation: Rectifier, L1: 0.3, L2:0" with RMSE: 0.105.

4.3.7. VOTE

The VOTE method is a boosting algorithm that is an ensemble meta-algorithm in supervised learning, and a family of machine learning algorithms that improve the weak learners to the strong ones. In the case of the VOTE classification task, all the models in the sub-process of the VOTE operator generate a classification or prediction model. For the prediction of a new example, the VOTE operator applies all the classification/prediction models and assigns the most predicted class with maximum votes to the unknown tuple. In both of this paper's scenarios, this algorithm has presented the best result.

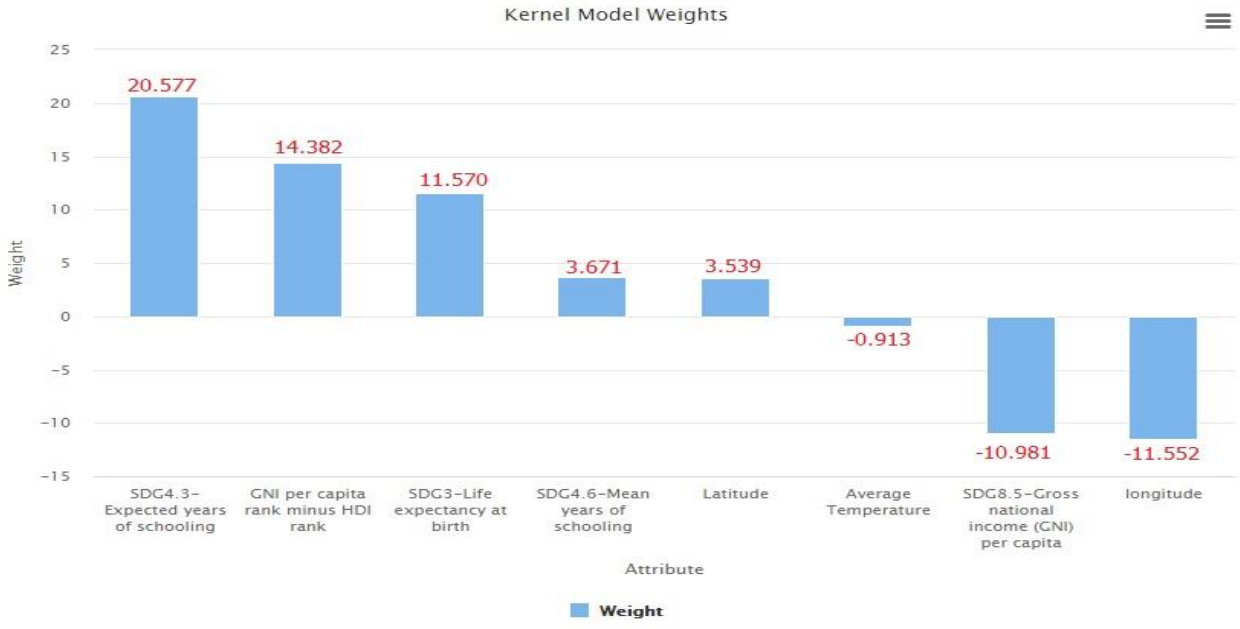

Figure 6. Weight of any attribute in mortality rate prediction based on the SVM algorithm.

Table 4. Obtained results from linear regression with t-test.

Coefficient	Attribute	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
0.434875	Latitude	0.16977	0.201333	0.985751	2.561553	0.011264	**
-0.37269	longitude	0.156579	-0.16996	0.998092	-2.38019	0.018379	**
0.709986	SDG3	0.478846	0.158502	0.914634	1.482704	0.139952	
0.925113	SDG4.3	0.396621	0.288024	0.88435	2.332485	0.020813	**
-0.83929	SDG4.6	0.345624	-0.32576	0.575169	-2.42832	0.016181	**
1.793283	(Intercept)	0.19084	NaN	NaN	9.39678	0	****

4.4. Evaluation and deployment

As mentioned earlier, two scenarios are followed in this work, and the results obtained can be described from two perspectives:

1- ***Comparing the performance of different data mining models***: based on the results of the first scenario, it is worth mentioning that the ensemble model (VOTE) has provided the highest accuracy and F-measure and the lowest prediction error (RMSE and MAPE). Therefore, the application of the ensemble data mining method in this work has provided more accurate and better results.

2- ***Identifying the factors affecting the disease mortality rate***: the purpose of this work is to identify the most important factors influencing the variability of the Coronavirus disease (COVID-19) mortality rates around the world. Based on the results of this work and during the application of various data mining methods, the following hypothesis was reinforced:

"The level of *education* of the community has the highest effect on the disease mortality rate in a way that the extent of its effect is much higher than the environmental factors such as *air temperature* and *regional health* factors and the level of community *welfare*."

Table 5. Comparison of classification methods (First scenario).

	Row	DM model	Optimized parameter	Accuracy	Classification error	Mean recall	Mean precision
Discrete Labels	1	K-NN	K=12	82.19% ± 3.95	19.11% ± 3.595	99.33%	82.32%
	2	Random forest (10 fold)	• Number of trees: 18 • Criterion: accuracy • Maximal depth: 4	81.16% ±4.20	18.44% ± 4.20	98.66%	85.96%
	3	Decision tree	• Criterion: Accuracy • Maximal Depth: 44	82.19% ± 3.95	19.11% ± 3.95	97.99%	83.91%
	4	Naïve Bayes	-	76.32±6.99	23.68 ± 6.99	91.28%	83.95%
	5	3-layer feed-forward neural network	• Activation: Tanh • L1: 0.1 • L2:0	82.34% ± 2.11	17.66% ± 2.11	100%	85.96%
	6	Vote (KNN, decision tree, naïve Bayes, 3LFFNN)		82.34% +/- 2.11	17.66% +/- 2.11	100%	85.96%

4.4.1. Analysis of second scenario's results

In the second scenario, some prediction methods such as linear regression, Support Vector Machine (SVM), neural network, and 3-layer feed-forward neural network have been used. The performance indicators of these methods that are calculated in this work are Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and symmetric Mean Absolute Percentage Error (sMAPE).

According to the results obtained (Table 6), in this scenario, 3-layer feed-forward neural network has provided the lowest RMSE: 10.5% and VOTE has provided an acceptable RMSE: 10.7% and the lowest MAPE: 21.9%; hence, the best data mining method in this scenario is an ensemble method (VOTE).

Table 6. Comparison of classification methods (second scenario).

	Row	DM model	Optimized parameter	MSE	MAE	RMSE	sMAPE
Continuous Label	1	Linear regression	• Feature selection: M5 prime • Min tolerance: 0.7	0.014	0.078	0.118	0.704
	2	Support vector machine	• Kernel: Radial	0.016	0.085	0.126	0.803
	3	3-layer feed-forward neural network	• Activation: Rectifier • L1: 0.3 • L2:0	0.013	0.076	0.114	0.654
	4	Neural network	• Training Cycle:192 • Learning Rate: 0.01 • Momentum: 0.8	0.014	0.077	0.118	0.671
	5	Vote (SVM, NN, 3LFFNN)		0.013	0.079	0.114	0.219

Finally, the results of this work can be described from two perspectives:

1- ***Comparing the performance of different data mining models*:** Based on the results of the first and second scenarios, it is worth mentioning that the 3-layer feed-foreword neural network and ensemble model (VOTE) have provided the highest accuracy and recall in the first scenario and the lowest prediction error (RMSE and sMAPE) in the second one. Therefore, the application of the ensemble data mining method in this work has provided more accurate and better results.

2- ***Identifying the factors affecting the disease mortality rate*:** the purpose of this work was to identify the most important factors influencing the variability of the Coronavirus disease (COVID-19) mortality rates around the world. Based on the results of this work and during the application of various data mining methods, the following hypothesis was reinforced: "The level of ***education*** of the community has the highest effect on the disease mortality rate in a way that the extent of its effect is much higher than the environmental factors such as the ***air temperature*** and ***regional health*** factors and the level of community ***welfare***."

5. Discussion

Due to the prevalence of viral diseases in recent years, the social effects of epidemic mortality have become one of the most important concerns of human societies. Therefore, identifying the preventive actions to reduce the probability of mortality can be considered one of the ways to manage such crises. Based on the World Health Organization reports, it is easy to recognize that the mortality rate of this disease is very diverse in different countries, and the recent studies have stated various reasons to justify this distinction. In this research work, during a data mining process on the information of COVID-19 mortality rate in different countries, the macro-components at the country level and the extent of their impact on the corona mortality rate have been considered. These components can be classified into three general categories: temperature components, geographical

components, and human development level components.

During the implementation of the data mining process, some classifications (decision tree, random forest (RF), naïve Bayes, K-NN, and 3-layer feed-foreword neural network) and prediction (linear regression, Support Vector Machine (SVM), neural network, and 3-layer feed-foreword neural network) methods have been used, and in addition to single methods, the ensemble method (VOTE) has also been applied. The results of the research work can be mentioned as follows:

- The rule extracted from the decision tree and random forest indicates that the feature "***SDG4.3-Expected years of schooling***" has the most gained information in predicting the mortality rate. According to this rule, and contrary to the results of the previous research works [10], the impact of temperature on increasing the mortality rate has not been strengthened.
- According to the result of the Support Vector Machine (SVM) classifier, "***SDG4.3-Expected years of schooling***" has the highest weight in the mortality rate prediction (weight = 20.577).
- The rule extracted from regression indicates that "***SDG4.3-Expected years of schooling***" has a 92% effect in the positive direction on the average mortality rate of a country.
- Based on the classification and prediction methods result, ***the ensemble model (VOTE)*** has provided the highest accuracy and recall in the first scenario and the lowest prediction error (RMSE and sMAPE) in the second one.

Thereupon, during the application of various data mining methods, the following hypothesis was reinforced: "The level of ***education*** of the community has the highest effect on the disease mortality in such a way that the extent of its effect is much higher than the environmental factors such as the ***air temperature*** and ***regional health*** factors and the level of community ***welfare***."

The impact of this factor on the mortality rate is probably due to these reasons:

- Knowledge-based and scientific societies, even with low welfare, can manage crises more than the others.
- Paying attention to the health advisories in such communities is higher than the others because they are knowledge-based and realistic.
- In knowledgeable societies, involvement in the rumors is lower, and so the incidence of mental health problems is lower too.

The results of this work open a new window in terms of the cultural and social factors in the interpretation of medical data. On the other hand, considering the prediction of such diseases in the not too distant future, paying attention to the education of individuals in communities, improving the level of culture and education can have a significant impact on reducing the severity of the crisis in the communities.

6. Author Statement

6.1. Ethical approval

I hereby accept the terms of the Public Health Journal ethical codes. The author did not collect any personal information but only used aggregate secondary data from public websites mentioned in the manuscript.

6.2. Funding

There was no funding source.

6.3. Competing interests

None declared.

Conflicts of interest

I and my co-authors have no conflicts of interest to declare.

7. References

[1] World Health Organization. "https://www.who.int/." (accessed.

[2] I. Gilles *et al.*, "Trust in medical organizations predicts pandemic (H1N1) 2009 vaccination behavior and perceived efficacy of protection measures in the Swiss public," *European Journal of Epidemiology*, vol. 26, pp. 203-210, 2011, DOI: 10.1007/s10654-011-9577-2.

[3] W. van der Weerd, D. R. Timmermans, D. J. Beaujean, J. Oudhoff, and J. E. Van Steenbergen, "Monitoring the level of government trust, risk perception and intention of the general public to adopt protective measures during the influenza A (H1N1) pandemic in The Netherlands," *BMC Public Health*, vol. 11, p. 575, 2011, DOI: 10.1186/1471-2458-11-575.

[4] A. Takian, M. M. Kiani, and K. Khanjankhani, "COVID-19 and the need to prioritize health equity and social determinants of health," *International Journal of Public Health*, vol. in the press, 2020.

[5] A. Zaveri and P. Chouhan, "Are child and youth population at lower risk of COVID-19 fatalities? Evidence from South-East Asian and European countries," *Children and Youth Services Review*, vol. 119, p. 105360, 2020.

[6] S. W.-. Betech, C. G. Cassandras, and I. C. Paschalidis, "Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: Hospitalizations, mortality, and the need for an ICU or ventilator," *International Journal of Medical Informatics*, vol. 142, p. 104258, 2020.

[7] R. Harris, "Exploring the neighborhood-level correlates of Covid-19 deaths in London using a difference across spatial boundaries method," *Health & Place*, vol. 66, p. 102446, 2020.

[8] F. J. Elgar, A. Stefaniak, and M. J. A. Wohl, "The trouble with trust: Time-series analysis of social capital, income inequality, and COVID-19 deaths in 84 countries," *Social Science & Medicine*, vol. 263, p. 113365, 2020.

[9] S. Kaur, P. Kaul, and P. MoradianZadeh, "Monitoring the Dynamics of Emotions during COVID-19 Using Twitter Data," *Procedia Computer Science*, vol. 177, pp. 423–430, 2020.

[10] A. M. Figueiredo, A. Daponte-Codina, Daniela, C. M. Marculino, P. V. Rodrigo, d. L. Kenio Costa, and G.-G. Eugenia, "Factores asociados a la incidencia y la mortalidad por COVID-19 en las comunidades autónomasFactors associated with the incidence and mortality from COVID-19 in the autonomous communities of Spain," *Gaceta Sanitaria*, vol. In Press, 2020.

[11] S. Hamidi, R. Ewing, and S. Sabouri, "Longitudinal analyses of the relationship between development density and the COVID-19 morbidity and mortality rates: Early evidence from 1,165 metropolitan counties in the United States," *Health & Place*, vol. 64, p. 102378, 2020.

[12] R. B. Hawkins, E. J. Charles, and J. H. Mehaffey, "Socio-economic status and COVID-19–related cases and fatalities," *Public Health*, vol. 189, pp. 29-134, 2020.

[13] J.-D. Zeitoun, M. Faron, and J. H. Lefèvre, "Impact of the local care environment and social characteristics on aggregated hospital fatality rate from COVID-19 in France: a nationwide observational study," *Public Health*, vol. 189, pp. 104-109, 2020.

[14] G. J. Bamber, S. Ryan, and N. Wailes, "Globalization, employment relations, and human resources indicators in ten developed market economies: international data sets," *The International Journal of Human Resource Management*, vol. 15, no. 8, pp. 1481-1516, 2004, DOI: https://doi.org/10.1080/0958519042000257968.

[15] J. Susnik and P. v. d. Zaang, "Correlation and causation between the UN Human Development Index and national and personal wealth and resource exploitation," *Economic Research-Ekonomska Istraživanja*, vol. 30, no. 1, pp. 1705-1723, 2017, DOI: 10.1080/1331677X.2017.1383175.

[16] World Bank Group. "Climate Change Knowledge Portal." https://climateknowledgeportal.worldbank.org/download-data (accessed.

[17] United Nations Development Programme. "http://hdr.undp.org/en/content/human-development-index-hdi." (accessed.

[18] O. Oladimeji and O. Oladimeji, "Detecting Breast Cancer through Blood Analysis Data using Classification Algorithms," *Journal of AI & Data Mining*, vol. 9, no. 3, pp. 359-351, 2021.

[19] M. Salehi, J. Razmara, and Sh. Lotfi, "Development of an Ensemble Multi-stage Machine for Prediction of Breast Cancer Survivability," *Journal of AI & Data Mining*, vol. 8, no. 3, pp. 378-371, 2020.

[20] H. Azan *et al.*, "IoMT amid COVID-19 pandemic: Application, architecture, technology, and security," *Journal of Network and Computer Applications*, p. 102886, 2020.

[21] B. Tesfaye, S. Atique, and N. Elias, "Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining approach," *Computer Methods and Programs in Biomedicine* vol. 140, pp. 45-51, 2017.

[22] M. Eliot, L. Azzoni, and C. Firnhaber, "Tree-Based Methods for Discovery of Association between Flow Cytometry Data and Clinical Endpoints," *Adv Bioinformatics*, 2009.

[23] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, and C. Shearer, *CRISP-DM 1.0: Step-by-step data mining guide*. 2000.

[24] Dataset Publishing Language. "Dataset Publishing Language, countries.csv." https://developers.google.com/public-data/docs/canonical/countries_csv (accessed.

[25] Young MoonChae, Seung HeeHo, Kyoung WonCho, Dong HaLee, and Sun HaJi, "Data mining approach to policy analysis in a health insurance domain," *International Journal of Medical Informatics*, vol. 62, no. 2-3, pp. 103-111

2001.

[26] B. Samwaysdos Santos, M. Teresinh ArnsSteiner, A. TrojanFenerich, A. Henrique, and P. Lima, "Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018," *Computers & Industrial Engineering*, vol. 138, p. 106120, 2019.

[27] B. Robson and S. Boray, "Studies in the use of data mining, prediction algorithms, and a universal exchange and inference language in the analysis of socioeconomic health data," *Computers in Biology and Medicine*, vol. 112, p. 103369, 2019.

[28] T. Käkilehto, S. kaSalo, and M. Larmas, "Data mining of clinical oral health documents for analysis of the longevity of different restorative materials in Finland," *International Journal of Medical Informatics*, vol. 78, no. 12, pp. e68-e74, 209.

[29] K. Eyasu, W. Jimma, and T. Tadesse, "Developing a Prototype Knowledge-Based System for Diagnosis and

Treatment of Diabetes Using Data Mining Techniques," *Ethiopian Journal of health sciences*, vol. 30, no. 1, pp. 115-124, 2020, DOI: https://doi.org/10.1016/j.artmed.2017.06.003.

[30] L. Schroeder, M. R. Veronez, E. M. de Souza, D. Brum, L. Gonzaga, and V. F. Rofatto, "Respiratory diseases, malaria, and leishmaniasis: Temporal and spatial association with fire occurrences from knowledge discovery and data mining," *International Journal of Environmental Research and Public Health*, vol. 17, no. 10, 2020.

[31] C. Neto, M. Brito, V. Lopes, H. Peixoto, A. Abelha, and J. Machado, "Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients," *Entropy*, vol. 21, no. 12, 2019.

[32] W. Meng, Ou, W., S. Chandwani, X. Chen, W. Black, and Z. Cai, "Temporal phenotyping by mining healthcare data to derive lines of therapy for cancer," *Journal of Biomedical Informatics*, vol. 100, 2019, DOI: 10.1016/j.jbi.2019.103335.

[33] H. Rawashdeh *et al.*, "Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage, " *Computational Biology and Chemistry*, 2020, DOI: https://doi.org/10.1016/j.compbiolchem.2020.107233.

[34] T. R. Stella Mary and S. Sebastian, "Predicting heart ailment in patients with a varying number of features using data mining techniques," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, pp. 2675-2681, 2019.

[35] T. Srivastava, A. Bhatnagar, J. Jayapradha, and M. Prakash, "Diabetes detection and monitoring using data mining and machine learning," *International Journal of Advanced Science and Technology*, vol. 29, pp. 1889-1897, 2020.

[36] A. S. Albahri, R. A. Hamid, J. Alwan, and Z. T. Al-qays, "Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review," *Journal of Medical Systems*, vol. 44, no. 7, 2020.

[37] T. Mohana Priya and M. Punithavalli, "An efficient data mining techniques - Multi-objective KNN algorithm to predict breast cancer," *International Journal of Recent Technology and Engineering*, vol. 8, pp. 986-990, 2019.

[38] J. Diz, G. Marreiros, and A. Freitas, "Applying Data Mining Techniques to Improve Breast Cancer Diagnosis," *Journal of Medical Systems*, vol. 40, no. 9, 2016.

[39] M. Manjusree and K. A. Sateesh Kumar, "Diabetes prediction using data mining classification techniques," *International Journal of Recent Technology and Engineering*, vol. 8, no. 3, pp. 5901-5905, 2019.

[40] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation and signal processing," *Advances in neural information processing systems*, vol. 9, pp. 281-287, 1996.

[41] T. Fusco, Bi, Y., H. Wang, and F. Browne, "approach for prediction modeling of schistosomiasis disease vectors: Epidemic disease prediction modeling," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 6, pp. 1159-1178, 2020.

[42] S. Geeitha and M. Thangamani, "A cognizant study of machine learning in predicting cervical cancer at various levels-a data mining concept," *International Journal on Emerging Technologies*, vol. 11, no. 1, pp. 23-28, 2020.

[43] H. Ayatollahi, L. Gholamhosseini, and M. Salehi, "Predicting coronary artery disease: A comparison between two data mining algorithms," *BMC Public Health*, vol. 19, no. 1, 2019, DOI: 10.1186/s12889-019-6721-5.

[44] A. Dela Cruz Galapon, "An assessment: Respiratory analysis using data mining method - A decision support system," *Test Engineering and Management*, vol. 83, pp. 4824-4829, 2020.

[45] Yulong. Bai, Lihong. Tang, Manhong. Fan, Xiaoyan. Ma, and Y. Yang, "Fuzzy First-Order Transition-Rules-Trained Hybrid Forecasting System for Short-Term Wind Speed Forecasts," *Energies*, vol. 13, no. 3332, 2020, DOI: 10.3390/en13133332.

[46] S. Simsek, U. Kursuncu, E. Kibis, M. AnisAbdellatif, and A. Dag, "A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival," *Expert Systems with Applications*, vol. 139, 2020.

[47] Li-Hong. Tang, Yu-Long. Bai, Jie. Yang, and Y.-N. Lu, "A hybrid prediction method based on empirical mode decomposition and multiple model fusion for chaotic time series," *Chaos, Solitons, and Fractals*, vol. 141, no. 110366, 2020, DOI: https://doi.org/10.1016/j.chaos.2020.110366.

کنترل نرخ مرگ و میر همه‌گیری‌ها با ارتقای شاخص توسعه انسانی: رویکرد داده کاوی در تحلیل داده-های COVID-19

صبا صارمی نیا [۱،۲]*

[۱] دانشکده مهندسی صنایع و سیستم ها؛ دانشگاه صنعتی اصفهان، اصفهان، ایران.

[۲] گروه پژوهشی بهینه سازی و تصمیم گیری هوشمند در سیستم های سلامت، دانشگاه صنعتی اصفهان، اصفهان، ایران.

ارسال ۲۰۲۲/۰۱/۰۹؛ بازنگری ۲۰۲۲/۰۳/۲۳؛ پذیرش ۲۰۲۲/۰۵/۰۴

چکیده:

در سال‌های اخیر، وقوع همه‌گیری‌ها (سارس و کوید۱۹) و تأثیرات گسترده آن‌ها بر زندگی انسان، سبب تمرکز محققان بر آسیب‌شناسی و شناخت مولفه‌های اپیدمیولوژیک آن‌ها شده‌است. یکی از مهم‌ترین چالش‌ها در این خصوص، نرخ مرگ ومیر بیماری‌ها است که اثرات نامطلوب اجتماعی به بار می‌آورد. لذا شناسایی و اولویت‌بندی عوامل موثر بر مرگ ومیر آن‌ها از اهمیت ویژه‌ای برخودار است. در این پژوهش، با استفاده از انواع مدل‌های داده-کاوی (منفرد و ترکیبی) ضمن تمرکز بر عوامل شناسایی شده اثرگذار بر نرخ مرگ و میر این بیماری (بهداشتی، وضعیت سلامت و آب وهوا) عوامل اجتماعی، اقتصادی، رفاهی، بهداشتی و فرهنگی نیز مدنظر قرار گرفته‌اند. جهت تعیین این معیارها از «شاخص توسعه انسانی» که بیانگر سطح توسعه‌ی اجتماعی هرکشور است و سالانه توسط سازمان ملل متحد انتشار می‌یابد؛ استفاده‌شده‌است. نتایج به‌دست‌آمده نشان می‌دهد که «شاخص سطح تحصیلات» جامعه بیشترین تأثیر را بر میزان مرگ و میر بیماری دارد؛ به نحوی که تأثیر آن بیش از عوامل محیطی (دمای هوا و عوامل بهداشتی و رفاه جامعه) می‌باشد و این ضریب تاثیر ناشی از «توانایی جوامع دانش‌بنیان در مدیریت بحران‌ها»، «توجه آنها به توصیه‌های بهداشتی»، «دخالت کمتر شایعات» و در نتیجه «کمتر بودن بروز مشکلات روانی» است. نتایج این پژوهش دریچه جدیدی را در تفسیر داده‌های پزشکی از منظر فرهنگی و اجتماعی می‌گشاید و «آموزش» را به عنوان مهمترین عامل در کاهش شدت بحران‌ها نشان می دهد. همچنین با توجه به نتایج و مقایسه انواع روش‌های داده‌کاوی، استفاده از روش‌های داده‌کاوی ترکیبی مانند VOTE از عملکرد بهتری در مقایسه با سایر روش‌ها برخوردارند.

کلمات کلیدی: کرونا ویروس (کوید ۱۹)، همه گیری، روش‌های ترکیبی داده کاوی، شاخص توسعه انسانی.