Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Dietary, comorbidity, and geo-economic data fusion for explainable COVID-19 mortality prediction

Milena Trajanoska\textsuperscript{a,\*}, Risto Trajanov\textsuperscript{a}, Tome Eftimov\textsuperscript{b}

\textsuperscript{a} Faculty of Computer Science and Engineering, Ss. Cyril and Methodius, University - Skopje, 1000, Macedonia
\textsuperscript{b} Computer Systems Department, Jožef Stefan Institute, Ljubljana 1000, Slovenia

\textbf{A R T I C L E   I N F O}

\textbf{Keywords:}
COVID-19 mortality prediction
Data fusion
Dietary habits
Geo-economic factors
Comorbidity

\textbf{A B S T R A C T}

Many factors significantly influence the outcomes of infectious diseases such as COVID-19. A significant focus needs to be put on dietary habits as environmental factors since it has been deemed that imbalanced diets contribute to chronic diseases. However, not enough effort has been made in order to assess these relations. So far, studies in the field have shown that comorbid conditions influence the severity of COVID-19 symptoms in infected patients. Furthermore, COVID-19 has exhibited seasonal patterns in its spread; therefore, considering weather-related factors in the analysis of the mortality rates might introduce a more relevant explanation of the disease’s progression. In this work, we provide an explainable analysis of the global risk factors for COVID-19 mortality on a national scale, considering dietary habits fused with data on past comorbidity prevalence and environmental factors such as seasonally averaged temperature geolocation, economic and development indices, undernourished and obesity rates. The innovation in this paper lies in the explainability of the obtained results and is equally essential in the data fusion methods and the broad context considered in the analysis. Apart from a country’s age and gender distribution, which has already been proven to influence COVID-19 mortality rates, our empirical analysis shows that countries with imbalanced dietary habits generally tend to have higher COVID-19 mortality predictions. Ultimately, we show that the fusion of the dietary data set with the geo-economic variables provides more accurate modeling of the country-wise COVID-19 mortality rates with respect to considering only dietary habits, proving the hypothesis that fusing factors from different contexts contribute to a better descriptive analysis of the COVID-19 mortality rates.

1. Introduction

The novel coronavirus SARS-CoV-2 caused an outbreak of the COVID-19 disease in December 2019 that originated in the Hubei Province of the People’s Republic of China. It has become a major global concern since the World Health Organization declared it a pandemic in March 2020 (WHO, Health Topics, 2020). Currently, there have been 225,680,357 confirmed cases of COVID-19 globally, with 4,644,740 deaths (WHO, 2021). The severity of the situation led many countries to develop different preventive measures to reduce the spread of the virus, including social distancing and entire lockdown.

In these challenging times, understanding the leading factors of the pandemic’s progress is crucial for countries to aid their healthcare systems in controlling and lowering case fatality rates due to COVID-19. Identifying essential contributors in the situation may serve as a guide for focusing on the significant risk factors and approaching the pandemic more successfully by prioritizing high-risk situations.

Contributing to the problem is that countries worldwide have experienced wide variations in COVID-19 mortality rates. Significant effort has been made to discover and explain the aspects that contribute to such outcomes. One study examined sociodemographic, structural, and environmental sources and concluded that differences in country-level risk factors significantly represent COVID-19 mortality-rate differences across countries (Kranjac & Kranjac, 2020). Additionally, another study identified 24 potential risk factors affecting the COVID-19 mortality rate by examining the variables’ univariate relationships with COVID-19 mortality rates across 39 countries (Pan et al., 2020). Out of all the variables they had selected, the COVID-19 case fatality rate was best predicted by the time to implement social distancing measures. By comparing demographic data from China, France, Germany, Italy, the Netherlands, South Korea, Spain, Switzerland, and the United States, studies have shown that the adjustment for the age distribution of the

\textsuperscript{\*} Corresponding author.
\textbf{E-mail addresses:} trajanoska28@gmail.com, milena.trajanoska@students.finki.ukim.mk (M. Trajanoska).

https://doi.org/10.1016/j.eswa.2022.118377
Received 5 December 2021; Received in revised form 8 July 2022; Accepted 1 August 2022
Available online 5 August 2022
0957-4174/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
examined cases explains 66% of the variation across countries (Sudharsanan et al., 2020). A similar study focused on analyzing the 10 US states with the highest deaths per population due to COVID-19 until May 2020. They analyzed 30 risk factors connected to COVID-19 mortality. They showed that the most critical risk factors are temperature, neonatal and under-5 mortality rates, the percentage of under-5 deaths due to acute respiratory infections and diarrhea, and tuberculosis incidence (Siddiqui et al., 2021). Considering a wider context of 93 countries and comparing risk factors such as aging, underlying chronic diseases, and social determinants such as poverty and overcrowding, it was discovered that countries with a high prevalence of population risk factors such as AD, lung cancer, asthma, and COPD had higher COVID-19 case mortality rates (Hashim & Khan, 2020). Moreover, based on data from 79 countries, population-related risk factors such as the proportion of the population over 70 years old and medical resources such as the number of hospital beds have been identified as the main factors influencing the COVID-19 mortality in each country included in the experiment (Cui et al., 2021).

Some of the research has focused on explaining these variations as a consequence of the prevalence of comorbidities in patients with severe symptoms (Li et al., 2020; Imam et al., 2020; Li et al., 2020; Liu et al., 2020; Wang et al., 2020; Wang et al., 2020), other research focuses on sociodemographic, structural, government, and environmental characteristics as drivers of high variations (Kranjac & Kranjac, 2020; Mathur et al., 2020; Ma et al., 2020). Concerning the population-related factors, obesity was found to be the most impactful (Kamary et al., 2021; Rajkumar, 2021; Dietz & Santos-Burgoa, 2020).

On an individual basis, studies (Imam et al., 2020; Caramelo et al., 2020; Bertsimas et al., 2020; Gebhard et al., 2020; Yan et al., 2020; Barda et al., 2020; Morales et al., 2021) have shown that older patients have an increased risk of death, with age being the most influential factor in the analysis. Moreover, the study (Gebhard et al., 2020) emphasizes the importance of gender in the incidence and case fatality of the disease, suggesting that gender differences in lifestyle behaviors may contribute to the observed mortality gap rates.

Further research has considered diet one of the most critical factors influencing human health. Recent findings in the field have pointed out nutrition habits as a significant causality in COVID-19 outcomes (Kamary et al., 2021; Rajkumar, 2021; Chesnut et al., 2021; García-Ordás et al., 2020; Greene et al., 2021; Eiser, 2021; Butler & Barrientos, 2020; Jayawardena & Misra, 2020; Richardson & Lovegrove, 2021; Cena & Chieppa, 2020). Although nutrition is not a treatment for COVID-19, it has a major impact on the development of some chronic diseases which have been shown to cause more severe symptoms in COVID-19 infected patients.

Studies have been made that aim to assess the impact of nutrition, obesity, and other population-related factors on the COVID-19 mortality rates on a national scale. The study (Kamary et al., 2021) analyzed these factors using statistically marginalized two-part models. It concluded that populations that consume more meat, vegetable products, sugar and sweeteners, sugar crops, animal fats, and animal products were associated with more death and fewer recoveries in patients.

Besides dietary habits, mental disorders have also affected people’s ability to practice healthy behavior. Furthermore, mental illnesses have been positively associated with increased chronic diseases (Chapman et al., 2008). For this purpose, analyzing the impact of mental disorders in predicting the outcomes of COVID-19 was considered important, specifically, the effect of depression was found to have a positive correlation with COVID-19 mortality rates (Rajkumar, 2021; Clouston et al., 2021).

Additionally, the pandemic was deemed to exhibit seasonal patterns in its spread, therefore analysis of environmental variables such as weather and humidity (Malki et al., 2020; Quilodrán et al., 2021; Quilodrán et al., 2020) have contributed to a better understanding of the situation. In general, regions with higher temperatures were associated with a lower number of infection cases, making temperature an influential factor in the analysis and inference process connected to COVID-19.

The purpose of our study is to provide an explainable analysis of the global risk factors for COVID-19 mortality on a national scale, taking into consideration dietary habits enriched with data on past comorbidity prevalence, environmental factors such as seasonally averaged temperature, geolocation, economic and development indices, undernourished and obesity rates. In contrast to prior studies that only focus on analyzing one aspect of the situation, we consider a broad context for developing explainable modeling of the COVID-19 country-wise mortality rates. For that purpose, we have gathered data from various areas, initially consisting of multiple factors representing country-wise nutritional habits, enriched with past comorbidity prevalence, different economic and development factors, environmental and ecological variables.

The motivation for this study is to investigate the joint impact of different factors (taking into consideration dietary habits enriched with data on past comorbidity prevalence, environmental factors such as seasonally averaged temperature, geolocation, economic and development indices, undernourished and obesity rates) to the prediction of COVID-19 mortality rates. The COVID-19 pandemic is still progressing and making research such as this one is contributing to providing knowledge for a better understanding of what is happening and what kind of measures can be taken in the future. All previously published papers are focused on investigating a specific group of factors in relation to the COVID-19 pandemic by using classical statistical methods (hypothesis testing) and correlation analysis in most cases performed by looking at linear correlation. However, our study is a fusion of different groups of variables, we are not investigating them in isolation, rather we examine them by developing an explainable ML pipeline that can further provide explanations of the joint effects of the variables from different groups.

The contributions of our experimental analysis follow:

1. We have developed an explainable ML pipeline for COVID-19 mortality prediction by fusing dietary, comorbidity, environmental, and economic data. The pipeline consists of two components integrated to investigate global risk factors for COVID-19 mortality on a national scale. The first one uses supervised learning, whereby fusing different feature portfolios (i.e., risk factors) the COVID-19 mortality rate has been predicted. The second one uses unsupervised learning to find clusters of countries that have similar food consumption patterns and their relation to the COVID-19 mortality rate. For both components, sensitivity analysis has been performed by investigating a portfolio of different modeling algorithms to select the best one to be included in the pipeline.

2. We have shown that fusing factors from different contexts, rather than considering only a single domain of contributing factors, provides a more accurate prediction of country-wise COVID-19 mortality rates.

3. We have provided both a global analysis of the impact of each factor on the COVID-19 mortality prediction and a local explainable analysis of the most important factors in each country.

4. We have shown through cluster analysis that countries that have similar dietary habits, also tend to have similar COVID-19 mortality rates.

5. With regard to the outcomes, some of the known results about most contributing factors such as obesity, and unhealthy dietary patterns have been proved. However, it was also investigated how other factors together with dietary patterns contribute to the end prediction results. It was shown that the food consumption of several food groups has a bigger impact on the COVID-19 mortality prediction (some food groups can be suggested to be consumed such as fish and some should be decreased since they have a negative influence). Moreover, we have shown that geographic characteristics of a country as well as population development have a significant effect.
on the prediction of COVID-19 mortality. These results can further be used by medical doctors and dietitians since they are contributing to providing knowledge to be used for improving the COVID-19 pandemic situation.

2. Methods

This section provides a description of the methods and algorithms used to obtain this study’s results.

2.1. The Machine Learning pipeline

The Machine Learning pipeline begins with a data fusion for the three separate data sets which are explained in detail in the following subsection. The fused data set was used to perform an explainable analysis for each country, and the results were presented for two specific countries. Feature selection was performed on the separate data sets and on the full data set, to obtain the most influential features for the COVID-19 mortality rates.

The selected features were then used in the cluster and regression analysis. The results from the cluster analysis are represented using explainable feature maps, based on the self-organizing maps algorithm (Kohonen, 1997), showing which features have the most impact on the creation of certain clusters.

The regression analysis incorporates the selected features as predictors for the COVID-19 mortality rates. The results from the analysis are represented with feature importance plots using Shapley values (Cohen et al., 2005) to measure the influence of each feature on the mortality rates.
response variable. Extreme gradient boosting (Chen & Guestrin, 2016) was used as the algorithm for performing the regression analysis.

Finally, the prediction distributions of the mortality rates for each country are presented. Separate graphs are created for the data subsets: dietary, comorbidity, country development, and the full data set.

The described workflow is depicted in Fig. 1. The data sets and the result data are represented with blue rectangles, and the actions and algorithms are represented with rounded orange rectangles.

2.2. Data set description

Our goal was to obtain a rich explanation of the variations in country-wise COVID-19 mortality rates. For that reason, a more comprehensive context was considered, consisting initially of multiple factors representing country-wise nutritional habits, enriched with past comorbidity prevalence, different economic and development factors, environmental, and ecological variables.

The final data set contains data for 154 countries. The maximum number of countries for which we had gathered data was 170, in the dietary data set. However, for some of these countries, data about the COVID-19 mortality rates was missing when conducting the analysis. For that reason, only countries for which we had the COVID-19 mortality rates were included in the study. The attributes were logically combined into three subgroups of data: dietary data, comorbidity data, and country development data.

The data used for conducting the analysis in this article is publicly available (Trajanov et al., 2021b).

2.2.1. Dietary data

The dietary data set consists of food supply data collected from the (FAO)/WHO STAT database (FAO, 1945). The dietary data is divided into food groups. The food groups are organized by the hierarchy defined by Food and Agriculture Organization (FAO)/WHO STAT, resulting in 23 diet-related features. The available features represent fat supply quantity, food supply measured in kg, food supply measured in kcal, and protein supply quantity. The features are measured in percentage of the total country-wise food consumption for fat supply, food supply in kg, food supply in kcal, and protein supply, as separate consumption data.

Analyzing how different nutritional patterns influence the outcomes of COVID-19 might give us the knowledge to identify food consumption patterns that countries with low case fatality rates exhibit. If deemed significant, these dietary patterns may be used to prevent a more severe escalation of the disease by adjusting our diet accordingly.

Statistical measures of this dataset are presented in Table 1. Only statistics about the essential features obtained after performing feature selection are shown.

2.2.2. Comorbidity data

It is widely considered that some chronic diseases influence the severity of symptoms and the outcome in COVID-19-infected patients (Liu et al., 2020; Wang et al., 2020; Wang et al., 2020).

Accordingly, we have considered establishing a relationship between mortalities caused by various diseases and illnesses to achieve a broader and more relevant explanation of COVID-19 mortality rates. For this purpose, we have collected country-wise data on past mortality rates attributed to 17 different groups of illnesses to assess the impact of other comorbid conditions. We have hypothesized that a similar distribution of comorbid conditions exists in the current population and use these factors to estimate the situation.

The comorbidity data set contains features representing the country-wise number of deaths due to different diseases, as organized by the highest level of ICD-10 categorization (ICD-10, 2010). The data set consists of 17 comorbidities.

Statistical measures of this dataset are presented in Table 2. Only statistics about the most important features obtained after performing feature selection are shown.

2.2.3. Geo-economic data

To further enrich the context, we have gathered data to represent country-wise economic and development status by adding HDI scores and GDP values dating from 2016 to the latest available date. In addition, we have considered the percentage of the obese and undernourished population of each country to obtain a more detailed characterization of different lifestyles.

Since countries that are close together often have similar lifestyles and dietary habits, we have included the minimum and maximum latitude and longitude of each country as spatial characteristics to consider the potential influence of the geographic location.

The studies (Malki et al., 2020; Quilodran et al., 2021; Quilodran et al., 2020) have evaluated environmental factors such as temperature and humidity as sole vital factors in their analysis of the COVID-19 pandemic. Considering that temperature was shown to be an influential contributing factor, we have added annual and seasonal temperatures expressed in degrees Celsius, averaged for the past 10 years. Each country’s geographic and temperature data was gathered from Berkeley Earth (Berkeley Earth, 2016) and The World Bank Group (WBG, 2020).

This data is used to provide additional contextual insight and potential factors that may contribute to the outcome of COVID-19 cases in each country. The data set contains 25 attributes that represent the above-mentioned country characteristics.

Statistical measures of this dataset are presented in Table 3. Only statistics about the most important features obtained after performing feature selection are shown.

2.2.4. COVID-19 data

Ultimately, country-wise COVID-19 mortality rates were used as the response variable for modeling dependencies in the data and explainable analysis of the importance and impacts of each feature in the specific tasks. The variable was represented in the percentage of every country’s total population.

Regression and cluster analysis was performed on each of the subgroups of data individually and on the complete data set containing all the attributes. A comparison between the obtained results is given in the sections that follow.

Statistical measures of the dataset are presented in Table 4.

| Feature                  | count | mean | std | min  | 25% | 50% | 75% | max |
|-------------------------|-------|------|-----|------|-----|-----|-----|-----|
| Milk                    | 170   | 0.29 | 0.20| 0.00 | 0.11| 0.26| 0.43| 0.92|
| Excluding Butter        | 170   | 0.15 | 0.12| 0.00 | 0.05| 0.14| 0.23| 0.57|
| Alcoholic Beverages     | 170   | 0.43 | 0.23| 0.00 | 0.24| 0.42| 0.61| 0.96|
| Animal Products         | 170   | 0.13 | 0.12| 0.00 | 0.05| 0.09| 0.18| 1.00|
| Fish, Seafood           | 170   | 0.28 | 0.20| 0.00 | 0.09| 0.27| 0.40| 0.98|
| Eggs                    | 170   | 0.18 | 0.13| 0.00 | 0.11| 0.15| 0.21| 0.93|
| Fat                     | 170   | 0.20 | 0.19| 0.00 | 0.07| 0.14| 0.27| 1.00|

Table 1 Food dataset basic statistics for the most important features. The table represents the main statistical measures for the selected features from the dietary data set. The features represent the consumption of different dietary products measured in the percent of intake for each population. For each feature, the number of recorded values, the mean, the standard deviation, the minimum, the 25-th percentile, the 50-th percentile, the 75-th percentile, and the maximum value are shown.
Table 2
Comorbidity dataset basic statistics for the most important features. The table represents the main statistical measures for the selected features from the co-morbidity data set. The features represent the number of cases for each disease, normalized by the population of each country. For each feature, the number of recorded values, the mean, the standard deviation, the minimum, the 25-th percentile, the 50-th percentile, the 75-th percentile, and the maximum value are shown.

| Features                          | count | mean  | std  | min  | 25 % | 50 % | 75 % | max  |
|-----------------------------------|-------|-------|------|------|------|------|------|------|
| Neoplasms                         | 128   | 34423.74 | 92840.42 | 0.00 | 617.67 | 6350.22 | 25588.79 | 729692.62 |
| Diseases of the musculoskeletal system and connective tissue | 128 | 814.65 | 2131.80 | 0.00 | 26.38 | 136.67 | 514.90 | 15884.67 |
| Mental and behavioural disorders  | 128   | 4037.81 | 14737.34 | 0.00 | 18.86 | 206.92 | 1796.94 | 142567.18 |

Table 3
Geo-economic dataset basic statistics for the most important features. The table represents the main statistical measures for the selected features from the geo-economic data set. The features represent the different environmental, geographic, and population-related features. For each feature, the number of recorded values, the mean, the standard deviation, the minimum, the 25-th percentile, the 50-th percentile, the 75-th percentile, and the maximum value are shown.

| Features                          | count | mean  | std  | min  | 25 % | 50 % | 75 % | max  |
|-----------------------------------|-------|-------|------|------|------|------|------|------|
| Max Latitude                      | 162   | 0.48  | 0.22 | 0.00 | 0.34 | 0.46 | 0.65 | 1.00 |
| Obesity                           | 162   | 0.38  | 0.22 | 0.00 | 0.14 | 0.44 | 0.54 | 1.00 |
| Annual Average Temperature        | 162   | 0.79  | 0.19 | 0.00 | 0.63 | 0.87 | 0.94 | 1.00 |

Table 4
Covid-19 mortality rate basic statistics. The table represents the main statistical measures for the COVID-19 mortality rates measured in percent of the population. For each feature, the number of recorded values, the mean, the standard deviation, the minimum, the 25-th percentile, the 50-th percentile, the 75-th percentile, and the maximum value are shown.

| Features                          | count | mean  | std  | min  | 25 % | 50 % | 75 % | max  |
|-----------------------------------|-------|-------|------|------|------|------|------|------|
| Deaths                            | 154   | 0.04  | 0.05 | 0.00 | 0.00 | 0.01 | 0.07 | 0.19 |

2.3. Feature selection

With the goal of selecting the best combination of parameters for the problem, we have performed feature selection using wrapper methods over the Extreme Gradient Boosting algorithm (Chen & Guestrin, 2016) for regression, combined with hyperparameter space search. The tuning was executed using 5-fold cross-validation (Refaeilzadeh et al., 2016) with negative mean absolute error as an evaluation metric. The negative value of the mean fundamental error was chosen as a metric because the implementation of the feature selection algorithm maximizes the evaluation function. In this case, the maximization of the negative value of the error corresponds to minimizing the actual error. The reason for choosing wrapper methods for this analysis is the large number of features in the data sets. Because of this, it wasn’t easy to perform manual feature selection or feature selection based on correlation due to the possible existence of multicollinearity. The Extreme Gradient Boosting algorithm handles multicollinearity internally, thus allowing for more accurate modeling of essential features by using wrapper feature selection methods.

After choosing the best parameters for each data set, we proceeded with different feature selections on the data sets, using three distinct algorithms. We compared three different algorithms for feature selection: recursive feature elimination with cross-validation using the estimator’s feature importance as feature weights (RFECV) (Chen & Jeong, 2007), recursive feature elimination with cross-validation using Shapley values (RFECV SHAP) (Cohen et al., 2005) as a feature evaluation metric (Lundberg & Lee, 2017) and Boruta search with Shapley values as a feature evaluation metric (Boruta SHAP) (Kursa & Rudnicki, 2010). For each of the algorithms, the mean absolute error was used as a metric for optimization in the cross-validation.

The feature selection was performed separately on the dietary, comorbidity, and geo-economic data set and then on the fused data set containing all the attributes in these subsets.

The best performing method on the datasets was recursive feature elimination using Shapley values as a feature importance metric, which produced the lowest mean absolute error during cross-validation and selected the most relevant features contributing to better model performance. The results from the feature selection methods are displayed in Table 5.

2.4. Explainable analysis

SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value that would be predicted if we did not know any features of the current output. These diagrams show a single order. When the model is non-linear or the input features are not independent, however, the order in which features are added to the expectation matters. The SHAP values arise from averaging the $\phi_i$ values across all possible orderings.

2.5. Cluster analysis with self-organised maps - SOM

A self-organizing map (SOM), or Kohonen Map, is a computational data analysis method that produces non-linear data mappings to lower dimensions. Alternatively, the SOM can be viewed as a clustering algorithm that produces a set of clusters organized on a regular grid. The roots of SOM are in neural computation (see neural networks); it has been used as an abstract model to form ordered maps of brain functions, such as sensory feature maps. Several variants have been proposed, ranging from dynamic models to Bayesian variants. The SOM has been used widely as an engineering tool for data analysis, process monitoring, and information visualization in numerous application areas (Kohonen, 1997).

2.6. Regression analysis with Extreme Gradient Boosting

The regression was performed using the selected features on each

Table 5
Feature selection methods comparison. The table presents the results from each of the three feature selection methods used in the analysis. The analysis was performed on the dietary data set, the comorbidity data set, and the geo-economic data set, separately, and finally on the fused data set. The metrics represent the value of the mean squared error metric. The lowest errors are presented in bold. It is evident that the RFECV SHAP feature selection method achieves the lowest MAE score.

| Dataset   | RFECV       | RFECV SHAP  | Boruta SHAP |
|-----------|-------------|-------------|--------------|
| Dietary   | 0.12        | 0.09        | 0.12         |
| Comorbidity | 0.041      | 0.034       | 0.053        |
| Geo-economic | 0.014   | 0.011       | 0.014        |
| Fused     | 0.0029      | 0.023       | 0.024        |
data set with respect to the COVID-19 mortality rates as the response variable. We have used the Extreme Gradient Boosting algorithm (XGBoost) to perform the analysis since we hypothesized that a non-linear relationship exists between the regressor variables \(x_i\) and the response \(y_i\). In order to eliminate as much bias as possible, we have created the prediction distributions using leave-one-out cross-validation (Wong, 2015) on each data subset and on the fused dataset.

The XGBoost algorithm is used in supervised learning problems. The model of choice for the algorithm is decision tree ensembles. The tree ensemble model consists of classification and regression trees (CART).

In order to learn the model parameters, which in this case are functions \(f_i\) containing the tree structure and leaf scores, the tree boosting algorithm is used. An additive strategy is used to learn the parameters, the learned structure is kept fixed, and one new tree is added at each step. The details of the algorithm can be viewed in the documentation (Chen & Guestrin, 2016).

3. Results and discussion

This study provided an explainable analysis of the most influential factors in the COVID-19 mortality rates. The research was conducted separately on the dietary, comorbidity, and geo-economic subsets and finally on the complete data set. The data which was collected originated from different domains, and some variability was present in the data collection process.

This section presents the obtained results from the feature selection, explainable analysis, regression, and cluster analysis. An explanation for the methods used in this research is found in the Methods section. Feature selection was performed in order to reduce the dimensionality of the data set with respect to the low number of countries and assess the most important factors that influence the COVID-19 mortality rates. The cluster analysis aimed to group the countries based on different dietary patterns, comorbidity prevalence, and country characteristics. The regression analysis aimed to estimate the overall COVID-19 mortality rates for each country and explain the model’s most significant errors. The full analysis was performed on all subsets of the data set, but only the most significant ones are presented in the following subsections.

3.1. Interpreting the influence of the most relevant features of the dietary data set

The first step in the analysis was to evaluate the extent to which dietary habits can explain the country-wise variation of COVID-19 mortality rates. Using the recursive feature elimination algorithm, from a total of 23 features in the dietary data set, only seven have been considered the most influential in predicting COVID-19 mortality (MAE = 0.0267). The chosen features that are the most significant in discovering dietary patterns are Alcoholic Beverages; Eggs, Fruits - Excluding Wine, Animal Products, Milk - Excluding Butter, Fish, Seafood, and Animal fats.

The global influence of the values of each of these features on COVID-19 mortality is presented in Fig. 2. The explanation is generated using Shapley values as a feature importance measure. As evident from the figure, in the countries where the intake of Alcoholic Beverages, Eggs, and Fruits - Excluding Wine, Animal Products, Milk - Excluding Butter, and Animal Fats is high, the estimates of COVID-19 mortality are higher. Conversely, low consumption of the above-mentioned products indicates a lower value of the estimate of COVID-19 mortality. On the other hand, countries with high consumption of Fish, Seafood experience lower mortality. Equivalently, lower consumption of Fish, Seafood cause countries to have significantly higher estimated COVID-19 mortality rates. The results are consistent with other research findings (Kamyari et al., 2021; Rajkumar, 2021; Dietz & Santos-Burgoa, 2020; Chesnut et al., 2021; Garcia-Ordas et al., 2020; Greene et al., 2021).

3.2. Interpreting the influence of the most relevant features of the fusion of dietary and geo-economic data set

The next step in the analysis was to assess the fusion of the dietary data set with the geo-economic data set. For this reason, a different feature selection was performed on the combination of the dietary and the geo-economic data sets which resulted in the highest explained variance (\(R^2 = 0.638\)). The explainable analysis of the selected features is shown in Fig 3.

As noted from the figure, taking the global influence of all the features in conjunction, higher annual temperatures contribute to lower COVID-19-induced deaths, whereas lower temperatures contribute to more deaths. This result is consistent with the findings in previous research (Malki et al., 2020; Quilodran et al., 2021; Quilodran et al.,

**Fig. 2. Feature importance plot for the dietary data set.** Red dots represent high values of the current feature in the used data set. Blue dots represent low values of the current feature in the used data set. Values on the x-axis represent the magnitude and sign of the impact that each value of the feature has on predicting the target variable COVID-19 mortality.
Additionally, Western and northern countries are predicted to have more deaths than most eastern and southern countries, which are predicted to have fewer deaths. More significant consumption of Alcoholic Beverages and Milk - Excluding Butter, correlates with higher COVID-19 mortality rates. Lower consumption of the mentioned products contributes to lower COVID-19 mortality rates.

On the other hand, larger consumption of Oil Crops contributes to a lowering in the predictions of the target variable. Countries that have a larger percentage of obese population on average are predicted to have larger COVID-19 mortality rates. This finding is consistent with other research in the field (Kamyari et al., 2021; Rajkumar, 2021; Dietz & Santos-Burgoa, 2020). Finally, countries with higher values for human development indices in 2016 and 2017 have experienced larger COVID-19 mortality rates than countries with lower values for human development indices in 2016 and 2017, which are negatively correlated with a large number of COVID-19 deaths.

Considering the results until this moment, high consumption of alcoholic beverages and milk are consistently influencing a more significant number of COVID-19 deaths in the separate and combined case. Additionally, larger annual temperatures impact COVID-19 mortality by lowering the estimated prediction, which may be due to the lowering of the spread of the virus in regions with high temperatures. Moreover, countries with a large percentage of obese population may experience a higher risk of COVID-19 mortality than countries with a lower rate of obese population.

3.3. Interpreting the influence of the most relevant features of the complete data fusion

Ultimately, we have decided to consider all of the features we have gathered and identify the global most influential factors in the analysis of COVID-19 mortality. Overall, eight features were selected as significant on the full data set. The feature importance as measured by Shapley values is presented in Fig. 4.
The summary from these features shows that high consumption of Animal fats leads to increased estimates of COVID-19 mortality, and lower consumption levels decrease the COVID-19 mortality. The same is generally observed for Alcoholic Beverages and Fruits – Excluding Wine.

On the other hand, meat’s high consumption levels lead to lower estimates of the response variable than lower levels of meat consumption, which cause a moderate increase in estimates. This may be due to the high protein level in meat, which contributes to lowering the undernourished rate in the population, which in turn implies the immune system.

Generally, it can be seen that unbalanced patterns in food consumption contribute to having larger COVID-19 mortality, and healthy dietary habits lower the predictions of mortality.

Lower frequencies of deaths due to Diseases of the skin and subcutaneous tissue affect the estimated mortality, whereas higher frequencies increase the estimation.

Higher temperature levels are generally linked with lower COVID-19 mortality and moderate to low temperature levels increase the COVID-19 mortality estimates. Additionally, specific geographic locations have significant impacts on predicting the response variable. Western countries, in general, are predicted to have higher COVID-19 mortality rates.

Using the fused data set, it is evident that some very important features on the separated data sets lose their predictive influence with respect to the target. One of the most important features in the geo-economic data set - Obesity, was not selected as important when the features from different aspects were combined. This may be because features that represent healthy and unhealthy dietary habits and comorbid conditions were introduced. Some of these features may have explained the impact of Obesity as an indirect factor, but further research is needed to evaluate this hypothesis.

3.4. Explanatory analysis of the influence of factors for specific countries

This subsection presents the results of the descriptive analysis of the impact of every factor in the data set on certain chosen countries’ COVID-19 mortality rates. Only two countries are presented in this subsection due to the limitation of the study’s length. Explanatory plots of an additional 84 countries are available on our code GitHub repository, linked in the Code availability section. The 10 most important features and their influences are shown in both figures.

**North Macedonia.** The first chosen country was North Macedonia. As it is evident from the feature importance plot in Fig. 5, the features with the largest influence for North Macedonia are mostly related to dietary habits and seasonal weather. More precisely, the most important contributing factor is meat consumption, which is lower than more than 75 % of the other countries. This low meat consumption contributes to estimating higher values of COVID-19 mortalities in North Macedonia, which may be due to the fact that meat is a good source of micro and macronutrients, and low consumption may have an impact on the overall health of the country’s population.

Additionally, lower than average temperatures contribute to higher COVID-19 mortality in the country. Moderate consumption of fruits (around the median of all countries) correlates with larger COVID-19 mortality. On the other hand, the consumption of alcohol is around 25 % percent of the lowest intake compared to the other countries and has an impact on lowering the COVID-19 estimated mortality.

Furthermore, the country’s GDP in 2017 was selected in the top 10 important factors. The GDP value was in the lower 25 % of the countries and indicated an increase in the mortality rate. This might mean that the economic situation also influences how a certain country deals with the pandemic.

Generally, for North Macedonia, most of the crucial factors from a...
total of 60 belong to the dietary data set. We can see that healthy diet patterns influence a reduction of the response variable and accordingly, unhealthy patterns influence an increase of the response variable.

**South Africa.** For South Africa, the most influential feature is eye diseases, adnexa, ear and mastoid processes. The feature importance plot is shown in Fig. 6. South Africa has a large number of deaths due to these diseases. More specifically, it is in the top 25% of countries that experience these many deaths. This factor, as comorbidity, has an influence of increasing the predicted COVID-19 mortality.

Additionally, low consumption of fruits and milk contributes to lowering COVID-19 mortality. South Africa’s consumption of these products is in the lower 25 percentile of countries. Milk contains saturated fats, which may contribute to the development of heart disease, and high sugar consumption has many adverse effects on human health. Accordingly, the low consumption of such products means less risk of these diseases in the South African population.

On the other hand, the intake of alcohol in the country is exceptionally high, influencing an increase in the estimate of COVID-19 mortality. Furthermore, the low consumption of tree nuts contributes to lowering the predicted mortality rate for South Africa.

A vast number of deaths due to diseases of the skin and subcutaneous tissue have an effect of increasing COVID-19 mortality. Diseases of the skin include skin cancer and various diseases caused by radiation, which impact the overall health of the population. The same is true for symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified. The number of deaths attributed to these illnesses is huge in South Africa, and it has an impact on increasing COVID-19 mortality.

Moreover, the high autumn average temperature decreases the mortality rate with a temperature value significantly above average.

As with the previous countries, dietary habits and temperatures greatly influence COVID-19 mortality. Additionally, for South Africa, some of the comorbidities were found significant due to a large number of the country’s population experiencing death due to these conditions.

From the previous analysis of the countries, we can identify that dietary factors and seasonal temperatures are generally regarded as the most influencing factors in modeling COVID-19 mortality rates. Specifically, the most contributing dietary factors represent approximately always very imbalanced consumption patterns. The influence of comorbidities was only significant if the number of deaths due to an illness was very high.

Overall, we could see that high consumption of alcohol, milk, and fruits led to an increase of COVID-19 mortality in the examined situations. Accordingly, lower consumption of the same products led to a decrease in COVID-19 mortality for the specific examined countries.

### 3.5. An in-depth view of country mortality rate with Self-Organized maps

This section presents a visualization of the countries and their mortality rate from COVID-19. We segmented our data into three bins, where we divide them into low, medium, and high mortality rates.

The names of the countries are colored accordingly in green, which is the low mortality rate, yellow, medium mortality rate, and red, high mortality rate. The clustering results are shown in Fig. 7.

The exciting pattern here shows that countries with the same dietary cultures are close together and have similar mortality rates. Also, there are some interesting cases like Italy close to Malta but have very different mortality rates. Another interesting case is Serbia and Croatia, countries that have almost the same diet but different mortality rates. It could be interpreted that food has a big impact on how COVID-19 affects people.

The decision map for the food data set regarding the SOM map is displayed in Fig. 8. The decision map represents the most important feature on which the decision was made for the particular cell. Animal products are the reason for the majority of the cells with medium and high mortality rates, so we can conclude that the intake of animal products correlates with a higher mortality rate from COVID-19.

The fusion of dietary and geo-economic data set resulted in an even more clustered SOM. This data set is enriched with the location of the countries, the annual temperature, the population of the country, and the obesity rate. The dataset is more representable, and the clusters are more interpretable. Countries that are closer together are placed in the same cluster. This map confirms the one before, and it is just enriched with geo-economic variables, and the clusters are more visually separated. These figures are presented in our code Git repository linked in the Code availability section.

### 3.6. Regression analysis concerning COVID-19 mortality rates

Regression was performed for the response variable Covid-19 mortality, using the Extreme Gradient Boosting algorithm for regression, leaving one out cross-validation to reduce the bias in the analysis maximally.

The results from performing regression on the separate and fused data sets are displayed in Table 6. The baseline model represents a constant prediction of the average mortality rate of all countries. It is evident from the results that the combination of dietary data, fused with the geo-economic data, marked bold in the table, with more

---

**Fig. 6. Feature impact analysis for South Africa.** Values on the x-axis represent the magnitude and sign of the impact that each value of the feature has on the prediction of the target variable COVID-19 mortality for the specific country. Red arrows pointing right represent high values of the current feature for the country. Blue arrows pointing left represent low values of the current feature for the country.
minimize loss scores (MSE and MAE) and larger R2 score, models the COVID-19 mortality rates most accurately and explains most of the variance in the response variable. Moreover, the comorbidities data set does not seem to have a significant impact on the COVID-19 mortality rates. However, comorbidities on an individual level are significant predictors of the outcome of the disease. Adding these country-wise factors to the model seems to affect only adding noise negatively.

From Table 6, we can see that the features of comorbidities are worse than the baseline model, which predicts a constant value for every country (i.e., the sum of the mortality rate of all countries/number of countries).

3.6.1. Regression results for the dietary data set

The regression analysis was performed using the global most influential dietary features outlined in the first subsection. The distribution of predictions is displayed in Fig. 9. The projections are sorted in increasing order by the absolute error with respect to the actual values of the targets.

It is interesting to note that for the countries Italy, North Macedonia, and Bosnia and Herzegovina, the model undershoots by predicting a low value of COVID-19 deaths. However, the mortality rates in all of these countries are high. The clustering grouped these countries as having a similar diet. They were clustered together with countries with generally low rates due to COVID-19. The model predicts a low value for the target when considering only the dietary habits as predictors. Cyprus is another country that is close to the above-mentioned, for which the model overshoots with the estimation. Cyprus has a low actual mortality rate and is clustered in a region of primarily low mortality rates. But since the diet is evaluated as similar to the other countries with high mortality rates, the modeling overestimates the actual COVID-19 mortality.

Bolivia and Peru are a part of another cluster where the estimates are significantly lower values than the actual COVID-19 mortality rates. These countries were clustered in a region of similar dietary habits with countries with moderate COVID-19 mortality rates, but the model predicts low mortality rates based on diet patterns.

The opposite situation is actual for Iceland and Finland, for which the estimates are significantly lower values than the actual COVID-19 mortality rates. These countries were clustered according to their food consumption with other countries with mostly low rates of deaths due to COVID-19. But their dietary patterns suggest a relation for predicting high COVID-19 mortality.

This suggests that, although dietary habits have a significant impact on a country’s population lifestyle, there exist additional factors that may have a more considerable influence on the outcome of diseases such as COVID-19. Even though some countries generally practice healthy diet habits, other factors such as social distancing, vaccination rates, and healthcare systems also play a crucial role.
dietary patterns, reality shows that demographic, geographic, and other environmental factors have an important causality in the outcome of COVID-19.

3.6.2. Regression results for the fused data set

Since using only dietary habits to analyze the pandemic situation may produce poor results, we have considered using the global most influential features discovered on the fused data. The error distribution on the fused data set is displayed in Fig. 10.

As it is evident, the model’s error improves when using the fused data set with the selected features, lowering the total error rate and per-instance error rate with respect to only using the dietary data as predictors. The distribution is quite different from that of the dietary data set. Combining these features introduces a better explanation for some countries and the overall error was significantly reduced.

In this case, it is remarkable that the modeling of the COVID-19 mortality for Italy, Peru, New Zealand, and Iceland has improved with respect to modeling the pandemic in these countries with only dietary habits. In contrast, Belgium still has the most significant error in the estimated COVID-19 mortality, and the estimates for Finland have not improved significantly.

We can see that adding more contextual features and enriching the data contributed to a more accurate estimation of the deaths caused by COVID-19 for most countries. However, there are still a lot of unknown factors that influence the situation which in addition may not be controllable. Thus, focusing on the factors that we can modify significantly contributes to effectively handling the disease.

Table 6

XGBoost regression models evaluation. The table presents the results obtained from training XGBoost models on the comorbidity, geo-economic, dietary data sets, the fusion of all data sets, and the fusion of the dietary data set with the geo-economic data set. The baseline represents predicting the mean value of COVID-19 mortality for each country. The best-performing model is outlined in bold font.

| Model                | Number of features | MSE       | MAE        | R²        |
|----------------------|--------------------|-----------|------------|-----------|
| xgboost_comorbidity  | 3                  | 0.00255785| 0.04172805 | 0.04779931|
| xgboost_geo_economic | 3                  | 0.00104544| 0.02220164 | 0.56301999|
| xgboost_diet         | 7                  | 0.00142011| 0.02072212 | 0.41153750|
| xgboost_all          | 8                  | 0.00116109| 0.02081079 | 0.56116695|
| xgboost_diet_geo_economic | 9           | 0.00086702| 0.02083138 | 0.6379698 |
| baseline             | 0                  | 0.00241325| 0.04108247 | 0         |

Fig. 8. Decision map for the SOM clusters of the dietary data set. The squares in the figure correspond to the ones in Fig. 7. These squares represent the most dominant feature in the decision for clustering the countries.
4. Conclusion

This study provides an explainable analysis of the most influential factors in the prediction of COVID-19 mortality rates. The analysis was conducted separately on the dietary, comorbidity, and country development subsets, and finally on the full data set. The collected data originated from different domains and some variability was present in the data collection process.

Ultimately, the study was limited due to the small number of countries (154) used in the overall process. The complete data set consisted of 60 features (excluding the WHO and FAO codes). For this reason, using feature selection methods for identifying the most influential factors in the analysis of the disease was necessary. Leave-one-country-out was used in the regression analysis in order to achieve a maximum lowering of the bias.

Furthermore, we used past mortality cases caused by 17 different diseases for the modeling of country-wise comorbidity prevalence, which might not have accurately represented the current distribution of comorbidities in the population. Since the results of the prediction of the COVID-19 mortality rates using only the comorbidity data set were very close to the baseline results, they may be considered non-influential in the general case.

On a related note, from all of the diseases, we show that obesity has an extremely important role in predicting the COVID-19 mortality rates, shadowing all of the other comorbidities and a lot of the geo-economic factors considered in the study. Obesity has been identified as a crucial comorbidity to COVID-19 in several previous studies (Kamyari et al., 2021; Rajkumar, 2021; Dietz & Santos-Burgoa, 2020).
We have shown that countries that consume more Alcoholic Beverages, Eggs, Fruits - Excluding Wine, Animal Products, Milk - Excluding Butter, and Animal Fats are related to higher COVID-19 mortality rates. On the other hand, countries with high consumption of Fish, Seafood are predicted to have lower mortality rates. These results are consistent with the findings in previous research (Kamyari et al., 2021; Rajkumar, 2021; Dietz & Santos-Burgoa, 2020; Chesnut et al., 2021; García-Ordás et al., 2020; Greene et al., 2021). Additionally, These consumption patterns contributed to explaining a large amount of the variance in the response variable ($R^2 = 0.4115$, $MAE = 0.0267$).

The most relevant features from the data were dietary-based and weather-based features. Concerning weather-based factors, we have shown that countries with higher annual temperatures on average have lower estimated COVID-19 mortality rates. This result is also consistent with the findings in previous literature (Malki et al., 2020; Quilodrán et al., 2021; Quilodrán et al., 2020). On top of that, we provide additional context to the analysis by showing that the geographic location of a country has a significant joint effect with other geo-spatial features on the outcome of COVID-19. Namely, the results outline that western and northern countries are predicted to have more deaths than most eastern and southern countries.

Additionally, our best performing model was the model trained on the fused dietary and geo-economic dataset. From this fusion, the human development indices of the countries in the dataset from the years 2016 and 2017 were identified as one of the most important features in predicting COVID-19 mortality. With this, we further contribute to the existing literature by showing that life expectancy, education, and per capita income, which are constituting factors of the index, have a large importance in determining the outcome of COVID-19 per country.

Countries with similarly balanced diets were clustered nearby using the SOM method. The regions of close countries were generally pure, with respect to the class distribution, meaning that dietary patterns had
a significant influence on the outcome of COVID-19. Fusion with other population-related factors resulted in purer cluster regions. Furthermore, the fusion of the dietary data set with country development and environmental features resulted in more accurate modeling of the COVID-19 deaths (R2 = 0.6376, MAE = 0.0208), than using each of the subsets of data individually. With this, we show that considering more factors from different areas in the process of jointly estimating the impact of COVID-19 mortality prediction yields more accurate results, and some of the features considered important in an isolated context prove to be less important when considering multiple variables.

Conclusively, our findings are consistent with previous literature in the field of COVID-19 mortality prediction. We contribute to the existing literature by considering a broader context of factors that influence the outcomes of COVID-19. Additionally, we provide both a global, joint effect analysis of the most important factors in the overall context and a local joint effect analysis of the most important factors per country. For this purpose, we use two different approaches, including supervised and unsupervised learning, and through our established Machine Learning pipeline, we show that our results are not only consistent with previous research, but also that our results are consistent between the two Machine Learning approaches.

We need to emphasize here that the presented pipeline can further be used when new data becomes available. Some hyperparameter adaptation might be necessary, but the steps will remain the same. Moreover, the pipeline can be used when a new group of features/variables will be available as well. These features can be added to the already used data groups and estimate their joint effects. The results from the pipeline can be assumed as descriptive, which are data-driven leads and contribute to finding the hidden patterns in the progression of the COVID-19 pandemic. The entire Machine Learning pipeline and the obtained results have practical implications for medical doctors and dietitians since they are contributing to providing knowledge to be used for improving the COVID-19 pandemic situation.

CRediT authorship contribution statement

Milena Trajanoska: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization.
Risto Trajanov: Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization.
Toke Eftimov: Conceptualization, Methodology, Validation, Investigation, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All of the used data is available in the linked data repository at https://github.com/risto-trajanov/covid-19-explainable-healthy-diet-data.

Acknowledgements

The work has been supported by the Slovenian Research Agency (research core funding programmes P2-0098); the European Union’s Horizon 2020 research and innovation programme under grant agreement 863059 (PNS-Cloud, Food Nutrition Security) and under grant agreement 101005259 (COMFOCUS).

Code availability

The code for producing the results explained in this research is available at our public code repository (Trajanov et al., 2021a).

References

Bard, N., Riesel, D., Akriv, A., Levy, J., Fiskel, U., Yona, G., & Idag, N. (2020). Developing a COVID-19 mortality risk prediction model when individual-level data are not available. Nature communications, 1-9.
Berkeley Earth, B. E. (2016). Data Overview. Retrieved May, 2020 from Berkeley Earth: http://berkeleyearth.org/data/.
Bertozzi, V. C., Lukan, G., Minneci, L., Noahdani, O., Orfanoudaki, A., Stellato, B., & Group, H. C.-1. (2020). COVID-19 mortality risk assessment: An international multi- center study. PLoS one, e0243622.
Butler, M. J., & Barrientos, R. M. (2020). The impact of nutrition on COVID-19 susceptibility and long-term consequences. Brain, behavior, and immunity, 53-54.
Caramelo, F., Ferreira, N., & Oliveira, B. (2020). Estimation of risk factors for COVID-19 mortality-preliminary results. MedRxiv.
Cena, H., & Cheppa, M. (2020). Coronavirus disease (COVID-19) -SARS-CoV-2 and nutrition: Is infection in Italy suggesting a connection? Frontiers in endocrinology, 11, 944.
Chapman, D. P., Perry, G. S., & Strine, T. W. (2008). The vital link between chronic disease and depressive disorders. PREVENTION OF CHRONIC DISORDERS.
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system, in proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 16). San Francisco, CA, 785-794.
Chen, X. W., & Jeong, J. C. (2007). Enhanced recursive feature elimination, Sixth International Conference on Machine Learning and Applications (ICMLA 2007), 429-435.
Chesnutt, W. M., MacDonald, S., & Wambier, C. G. (2021). Could diet and exercise reduce risk of COVID-19 syndromes? Medical hypotheses, 110502.
Cloutson, S., Luft, B. J., & Sun, E. (2021). History of premorbid depression is a risk factor for COVID-related mortality: Analysis of a retrospective cohort of 1,387 COVID+ patients. medRxiv.
Cohen, S., Reppin, E., & Dror, G. (2005). Feature selection based on the shapley value. other words, 1, 968-9.
Cui, S., Wang, Y., Dai, Q., Huang, Z., & Cheng, T. C. (2021). A two-layer nested heterogeneous ensemble learning predictive method for COVID-19 mortality. Applied Soft Computing, 113, Article 107496.
Dietz, W., & Santos-Burgos, C. (2020). Obesity and its implications for COVID-19 mortality. Obesity, 1005.
Eiser, A. R. (2021). Could dietary factors reduce COVID-19 mortality rates? Moderating the inflammatory state. The Journal of Alternative and Complementary Medicine, 176-178.
FAO, F. a. (1945). Data. Retrieved May, 2020 from FAOSTAT: https://www.fao.org/faostat/en/
data.
García-Olmedo, M. T., Arias, N., Benavides, C., García-Olalla, O., & Benítez-Andrada, J. A. (2020). Evaluation of country dietary habits using machine learning techniques in relation to deaths from COVID-19. Healthcare, 971.
Gebhard, C., Regitz-Zagrosek, V., Neuhäuser, H. K., Morgan, R., & Klein, S. L. (2020). Impact of sex and gender on COVID-19 outcomes in Europe. Biology of sex differences, 1-13.
Greene, M. W., Roberts, A. P., & Fruge, A. D. (2021). Negative association between Mediterranean diet adherence and COVID-19 cases and related deaths in Spain and 25 OECD countries: An ecological study. Frontiers, Nutrition, 74.
Hashim, M. J., A., & Khan, G. (2020). Population risk factors for COVID-19 mortality in 93 countries. Journal of epidemiology and global health, 204-208.
ICD-10, I. S. (2010). ICD-10 Version2010. Retrieved May, 2020 from ICD: https://icd. who.int/browse/10/2010/en.
Imam, Z., Odish, G., Gill, I., O’Connor, D., Armstrong, J., Vanoo, A., & Halalau, A. (2020). Older age and comorbidity are independent mortality predictors in a large cohort of 1,305 COVID-19 patients in Michigan, United States. Journal of internal medicine, 469-476.
Jayawardena, R., & Misra, A. (2020). Balanced diet is a major casualty in COVID-19. Diabetes & metabolic syndrome, 1085.
Kamyani, N., Soptani, A. R., Mahboub, H., & Moghimbeigi, A. (2021). Diet, nutrition, obesity, and their implications for COVID-19 mortality: Development of a marginalized two-part model for semicontinuous data. JMRI pubic health and surveillance.
Kohonen E. Exploration of very large databases by self-organizing maps Vol. 1 1997 IEEE pp. PL1-PL6.
Kranjac, A. W., & Kranjac, D. (2020). Decomposing Differences in Coronavirus disease 2019-related Case-Fatality Rates across Seventeen Nations. Pathogens and Global Health, 100-107.
Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. Journal of statistical software, 36, 1-13.
Li, B., Jia, X., Zhang, T., Zhao, Y., Tian, F., Li, Y., & Li, B. (2020). Comparison of cardiovascular metabolic characteristics and impact on COVID-19 and MERS. European Journal of Preventive Cardiology, 1320-1324.
Li, J., Chen, Z., Nie, Y., Ma, Y., Guo, Q., & Dai, X. (2020). Identification of symptoms prognostic of COVID-19 severity: Multivariate data analysis of a case series in Henan province. Journal of Medical Internet Research.
Liu, H., Chen, S., Liu, M., Nie, H., & Lu, H. (2020). Comorbid chronic diseases are strongly correlated with disease severity among COVID-19 patients: A systematic review and meta-analysis. Aging and disease, 668-678.
Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (p. 30).
Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., & Luo, B. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Science of the total environment, 724, Article 138226.
Malki, Z., Atlam, E. S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., & Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos, Solitons & Fractals, 138, Article 110137.

Mathur, P., Sethi, T., Mathur, A., Maheshwari, K., Cywinski, J. B., Khanna, A. K., & Papay, F. (2020). Explainable machine learning models to understand determinants of COVID-19 mortality in the United States. medRxiv.

Morales, G. R., Monterrubio, S. M., García, J. A., & Ger, P. M. (2021). Explainable Machine Learning Prediction for Mortality of COVID-19 in the Colombian Population. Research Square.

Pan, J., St. Pierre, J. M., Pickering, T. A., Demirjian, N. L., Fields, B. K., Desai, B., & Gholamrezanezhad, A. (2020). Coronavirus Disease 2019 (COVID-19): A Modeling Study of Factors Driving Variation in Case Fatality Rate by Country. International Journal of Environmental Research and Public Health.

Quilodran, C. S., Currat, M., & Montoya-Burgos, J. I. (2020). Climatic factors influence COVID-19 outbreak as revealed by worldwide mortality. MedReiv.

Quilodran, C. S., Currat, M., & Montoya-Burgos, J. I. (2021). Air temperature influences early Covid-19 outbreak as indicated by worldwide mortality. Science of The Total Environment, 792, Article 148312.

Rajkumar, R. P. (2021). Cross-national variations in COVID-19 mortality: The role of diet, obesity and depression. Diseases, 36.

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-Validation In Liu I. & Özsu MT (Eds.). Encyclopedia of Database Systems, 1-7.

Richardson, D. P., & Lovegrove, J. A. (2021). Nutritional status of micronutrients as a possible and modifiable risk factor for COVID-19: A UK perspective. British Journal of Nutrition, 125(6), 678-684.

Siddiqui, S. H., Sarfaraz, A., Rezvi, A., Shaheen, F., Yousafozai, M. T., & Ali, S. A. (2021). Global variation of COVID-19 mortality rates in the initial phase. Osong Public Health and Research Perspectives, 64-72.

Sudharsanan, N., Didzun, O., Barnighausen, T., & Geldsetzer, P. (2020). The contribution of the age distribution of cases to COVID-19 case fatality across countries: A nine-country demographic study. Annals of internal medicine, 714-720.

Trajanov, R., Trajanoska, M., & Efthimov, T. (2021, September 6). covid-19-explainable-healthy-diet. From GitHub: https://github.com/risto-trajanov/covid-19-explainable-healthy-diet.

Trajanov, R., Trajanoska, M., & Efthimov, T. (2021, September 16). covid-19-explainable-healthy-diet-data. From GitHub: https://github.com/risto-trajanov/covid-19-explainable-healthy-diet-data.

Wang, L., Li, J., Guo, S., Xie, N., Yao, L., Cao, Y., & Sun, D. (2020). Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. Science of the Total Environment, 727, Article 138394.

Wang, X., Fang, X., Gai, Z., Wu, X., Gao, X., Min, J., & Wang, F. (2020). Comorbid chronic diseases and acute organ injuries are strongly correlated with disease severity and mortality among COVID-19 patients: A systemic review and meta-analysis. Research, WBG, W. B. (2020, May). Download Data. From Climate Change Knowledge Portal: https://climateknowledgeportal.worldbank.org/download-data.

WHO, W. H. (2020, March 12). Health Topics. From World Health Organization: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic.

WHO, W. H. (2021, June 19). WHO Coronavirus (COVID-19) Dashboard. Retrieved September 16, 2021 from World Health Organization: https://covid19.who.int/.

Wong, T. T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition, 48(9), 2839-2846.

Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., & Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. Nature Machine Intelligence, 283-288.