Master Dissertation CSCM20

Predicting COVID-19 Mortality Using Machine Learning and Regression Analysis Based on Multiple Health-Related Factors

---

by:

Student: Abdullah Basher Sharaf (2139533)

Supervisor: Dr.Alma Rahat

A thesis submitted in fulfilment of the requirements for the Master of Data Science

of the

Department of Computer Science,
Faculty of Science and Engineering,
Swansea University

September 30, 2024

Declaration of Authorship

I, Abdullah Basher Sharaf, declare that this thesis titled, "Predicting COVID-19 Mortality Using Machine Learning and Regression Analysis Based on Multiple Health-Related Factors" and the work presented in it are my own. I confirm that:

- This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.
- This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.
- I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.
- The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed: 

Date: 28/09/2024

Abstract

The global impact of the COVID-19 pandemic has significantly changed different aspects of society, including health systems, economies, and daily routines. As communities grapple with the repercussions, the effects on various areas of life are still emerging. The road to recovery requires more than just a return to the way things were; it demands a deliberate rethinking of how to rebuild in the face of potential future global crises. The primary goal is to explore the connection between factors that impact human health and COVID-19 mortality. The research takes into account dietary categories such as plant-based, animal-based, and seafood; demographic data like age, gender, and population density; and lifestyle factors such as physical activity, obesity, diabetes prevalence, alcohol consumption, smoking habits, and tuberculosis rates. It also takes into account environmental factors, such as pollution levels and the population's exposure to risk factors. Multiple machine learning regression techniques were evaluated, including Polynomial Regression, Support Vector Regressor, Random Forest Regressor, KNeighbors Regressor, Gradient Boosting Regressor, and Multilayer Perceptron of Artificial Neural Networks. The research demonstrated a correlation between these variables and mortality, with the polynomial regression model exhibiting a good performance. The model attained a R^2 score of 0.5303 and a correlation coefficient of 0.7688 in the nutrition domain, demonstrating its enhanced capacity to predict COVID-19 mortality relative to other fields. Furthermore, we analysed the model outcomes utilising Explainable AI techniques such as LIME, which enhanced the transparency and clarity of the predictions.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Dr.Alma Rahat for their continuous support, invaluable guidance, and insightful advice throughout the course of my research and thesis writing. Your encouragement and expertise have been crucial in shaping this work, A special thank you to my family and friends, whose unwavering support, patience, and understanding have been my constant source of strength. To my parents, your belief in me has been an inspiration. I am forever grateful for your love and encouragement. To everyone who contributed to this journey in one way or another, I extend my heartfelt appreciation, thank you all.

Contents

Chapter 1

Introduction

Globally, COVID-19 has evolved into a major healthcare industry concern as well as a public health emergency. The World Health Organisation estimates that COVID-19 has caused over 7 million deaths, or about 0.09% of the 8 billion global population, so stressing the extreme damage the epidemic has done on human life. Additionally, more than 704 million infections account for nearly 9% of the world's population, indicating that the virus has had a significant impact on a large portion of society [27]. The COVID-19 pandemic has severely impacted over 200 countries, resulting in exponential growth in infection and mortality rates [1].

Figure 1.1 illustrates the mortality rate across the world.

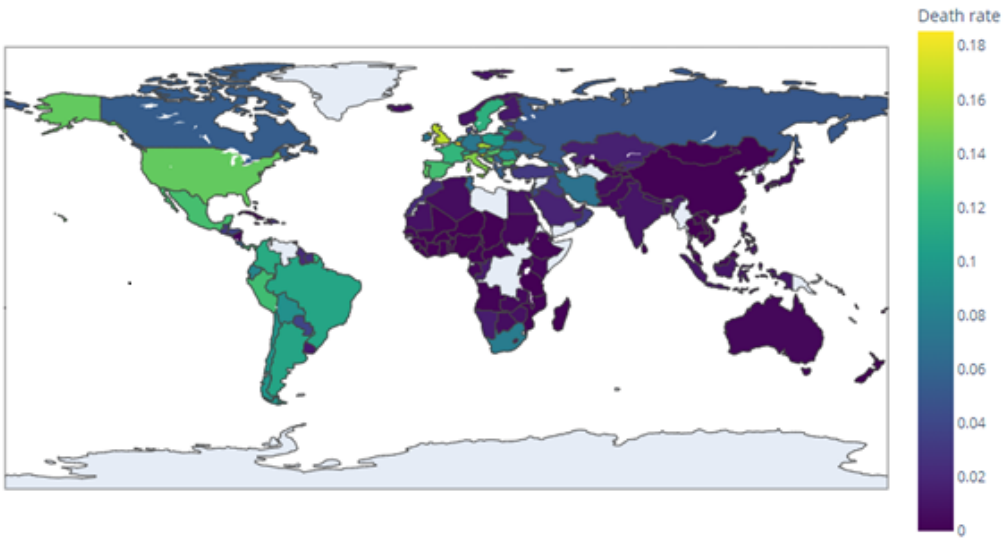

Figure 1.1: World Map of COVID-19 Mortality Rates: June-2021

It attracted immediate interest from health researchers all around as well as from the general public. Different spheres of life were alarmed by the absence of a suitable strategy for infection control and the ambiguity about the new virus's behavior. Driven by the necessity of informed strategies in controlling such a new epidemic, governments has thus progressively turned to artificial intelligence to negotiate this crisis. Several machine learning models have been fitted to the data for COVID-19 outcomes in various affected countries to forecast and monitor the cases and death rates. As a result, the utilization of artificial intelligence has emerged as a promising strategy for enhancing predictive analytics in tracking disease outcomes, and optimizing healthcare resource allocation during such crises [29].

This study aims to provide insights into predicting COVID-19 mortality rates through advanced machine learning models [18, 22]. The research explores the influence of various independent variables, including food, demographics, lifestyle and health indicators, and environmental factors, on death rates in different countries. The thesis begins by establishing the necessity for this research, emphasising its aim to address deficiencies in existing literature by analysing the correlations between these factors and COVID-19 mortality. It provides pertinent contextual information. The methodology section delineates the procedures utilised in data collection from diverse sources. It outlines preprocessing techniques for managing missing and duplicate data, standardisation, and the implementation of feature selection methods. Several machine learning models are used in the study, including Polynomial Regression (PL), Support Vector Regression (SVR), Random Forest Regression (RFR), KNeighbors Regressor (KNR), Gradient Boosting Regression (GBR), and Multi-Layer Perceptron (MLP). The main focus is on hyperparameter optimisation to make the models work better. The research employs Explainable AI (XAI) methodologies, to enhance transparency and foster trust in model predictions by rendering the outcomes comprehensible [15]. Finally, the thesis notes the limits of the research including difficulties with model accuracy and data availability restrictions. It also suggests future study paths, including improving model accuracy and broadening the feature set to support public health predictive analytics.

1.1 Motivation

- **Improved Public Health Approaches:** The ability to customise awareness and guidance initiatives to meet the specific needs of populations is facilitated by an understanding of the relationship between the mortality rate associated with COVID-19 and the varying prevalence of different factors across multiple life domains globally. This targeted approach enhances the efficacy of mortality reduction efforts and promotes informed behavioural changes.

- **Explaining Mortality Trends Using XAI:** Through the use of explainable artificial intelligence and machine learning forecasting analysis, this research contributes to a better understanding of the problem in some countries by predicting the reasons for the rise or fall in mortality rates. This approach aids in addressing the issue more accurately and effectively, taking into account the specific circumstances of the country.

- **Data Science:** This research enhances the significance of data science within the realm of public health by advancing predictive modeling techniques to include intricate and interconnected variables. This improved modeling approach not only deepens our comprehension of the dynamics of pandemics but also bolsters readiness for prospective outbreaks, facilitating swifter and more efficient responses.

1.2 Key Contributions of the Study

- Initially, the current study fills a significant gap in the literature by creating predictive models for COVID-19 mortality rates that take into account a wide range of independent variables linked to human health, including food with (19) features, (6) lifestyle and health-related features, (9) demographic-related features, and (6) environmental factors, each of which receives individual attention. Previous studies [6, 33] frequently concentrated on a restricted number of factors from a variety of disciplines and subsequently addressed them collectively, such as poverty, medical income, population density, high blood pressure, high cholesterol, unhealthy food outlets, urban population, smoking, and individuals aged 15 and older. This research showd that factors related to food, lifestyle, and health, which are the most significant contributors.

- Secondly, the study was successful in making COVID-19 mortality predictions using a polynomial regression machine learning model, achieving a correlation coefficient of 0.7688 with nutrition domain features. In contrast, previous research [6] achieved a correlation coefficient of 0.60 by using support vector regression.

- Finally, the study used XAI techniques that make COVID-19 mortality predictions more clear, accurate, and simple to understand. It also looked at how independent factors affect predictions in several countries, such as the UK and Belgium. This idea fixed a big problem with earlier research [6, 12, 33, 37] the predictions didn't give enough details about prediction reasons. This ensures that the predictions are not only accurate but also understandable, thereby enhancing their usefulness for policymakers and healthcare professionals to understand weak points in any country they start to target.

Chapter 2

Literature Review

2.1 Background:

2.1.1 Immune System Response to a Pandemic

The immune system [19, 28] plays a vital role in safeguarding the body against viral infections, especially those that lead to extensive outbreaks. During a pandemic, the immune system's response is complex and may occasionally lead to an exaggerated reaction. This severe inflammatory response can significantly affect the body, exacerbating the severity of the condition. In the instance of SARS-CoV-2 infection, the virus can circumvent the immune system's defences, resulting in detrimental effects on the body's organs. Comprehending the immune response to pandemics is essential for formulating effective prevention and treatment strategies, encompassing the development of accurate diagnostic methods and novel therapeutic interventions [10].

2.1.2 Nutrition Factors Influencing Immune Function

This essay comprehensively examines the diverse impacts of a plant-based diet in contrast to an animal-based diet on the immune system. It also examines the capacity of both diets to substantially diminish immune system response concerning COVID-19. The variations in nutritional constituents and their corresponding effects on immune function are analysed [21, 34].

2.1.3 Demographic Factors Influencing Immune Function

Variations in immune system functionality are evident across different demographic groups, with factors such as age, gender, ethnicity, and geographic location substantially

influencing immune responses and disease severity. Research [11, 25, 30] indicates that variations in the clinical presentation of SARS-CoV-2 infection by gender are often correlated with age and pre-existing health conditions; however, certain studies have emphasised gender-specific disparities in the severity of COVID-19 that are independent of these variables. Moreover, age significantly impacts immune function, as individuals experience various stages of immune development from infancy to old age. The evolutionary significance of these differences lies in the fact that immune responses, despite being energetically taxing, are essential for survival.

2.1.4 Lifestyle and heath Factors Influencing Immune Function

The lifestyle is fundamentally rooted in the fundamental needs of human life and has a substantial impact on and regulates physiological functions. The consumption of alcoholic beverages and tobacco use is a common feature of many contemporary lifestyles, which also prioritise physical activity [38]. In general, the primary objective of engaging in sports is to improve physical health and flexibility, which in turn increases the strength of the immune system and reduces the susceptibility to illness.

Research studies [20, 23] have identified a significant impact of smoking on the immune system's functionality and demonstrated that smokers exhibit altered T cell and neutrophil behaviour, as well as decreased levels of numerous pro-inflammatory cytokines and chemokines. Furthermore, cigarette smoke contains a variety of harmful substances that can impede the immune system's effectiveness. Research [13] has demonstrated that immune function is significantly impacted by excessive weight, particularly obesity .

Diabetes is recognized for its influence on the immune system, resulting in numerous deficiencies that elevate susceptibility to infection [7]. The primary cause of weakened immune systems on a global scale is undernutrition. For an extended period, the correlation between undernutrition and an elevated susceptibility to infections has been extensively documented [32].

2.1.5 Environment Factors Influencing Human Health

The respiratory system serves as the body's primary defense against the rapidly changing external environment. Recent observations [8, 9] indicates that the air we breathe contains elevated levels of gaseous pollutants, dust, and smoke in environments with increased pollution, such as smog. The serious health issues associated with air pollution are significantly exacerbated by the congestion and emissions from vehicles and industries in urban areas, which leads to an increased prevalence of respiratory problems. These diseases are not transmitted through personal contact; rather, they are the consequence of

vehicle emissions. Initially, respiratory issues may appear to be minor; however, prolonged exposure can result in life-threatening conditions.

2.2 Applications using Machine Learning

The attempts to reduce the COVID-19 epidemic effects have benefited much from artificial intelligence and machine learning [24]. These technologies have been used in a number of ways, such as to prioritise patient care, find COVID-19 cases and deaths, and predict how the virus will spread [3]. In particular, the development of predictive models that detect and categorise virus instances has been significantly influenced by supervised machine learning. These models can accurately classify new cases based on similar patterns by training on historical clinical and laboratory data [26]. Additionally, scientists have used supervised machine learning to predict the severity and mortality risk among COVID-19 patients by analyzing pre-existing medical conditions and laboratory results [5].

2.2.1 First Regression Problem:

The global consequences of the COVID-19 pandemic have been unparalleled, necessitating thorough investigations into the determinants of its transmission and severity, The study [6] examines the relationship between environmental, socioeconomic, and health-related factors and COVID-19 outcomes in Guilford County, North Carolina. The research emphasises the potential contribution of food accessibility, income levels, population density, and prevalent health issues to the spread and severity of the virus. This study is a component of a more comprehensive effort to understand the local dynamics of pandemic transmission, with the potential to inform public health strategies. The study emphasises the importance of these factors in understanding the effects of COVID-19 and the possibility that the pandemic's adverse effects could be exacerbated by the unequal access to food and the diverse health conditions of different demographics. leverages machine learning algorithms to reveal complex, non-linear relationships that can forecast mortality and infection rates.

Methodology:

Using GIS technology such as Moran's I and geographically weighted regression (GWR) to visually represent and measure the geographic spread of COVID-19 cases and deaths. These methods assist in pinpointing areas with elevated rates of infection or mortality and their correlation with the arrangement of food establishments and socioeconomic variables. The study implemented an assortment of machine learning models, including Linear

Regression, Multi-output Linear Regression, Random Forest Regression, K-Nearest Neighbour Regression, and Support Vector Regression. These models were used to predict the number of COVID-19 cases and fatalities in relation to independent variables, thereby enabling a thorough analysis of both linear and non-linear relationships. The dataset was divided into two subsets: training (80%) and testing (20%). In order to assess the models' ability to predict cases and fatalities during the COVID-19 pandemic, they were evaluated using metrics such as Root Mean Square Error and Correlation Coefficient.

Result:

The machine learning models show greater predictive accuracy for COVID-19 deaths, with a correlation coefficient of 0.60 for SVR compared to 0.446 for linear regression in cases Figure 2.1. This indicates that although the independent variables may not strongly predict infection rates, they have a more significant impact on determining the severity of outcomes, especially mortality.

Correlation Coefficient		
Models	**CVID-19 Cases**	**COVID-19 Deaths**
Linear regression for multioutput regression	0.446	0.508
K-nearest neighbors for multioutput regression	−0.085	0.466
Random forest for multioutput regression	0.137	0.239
Support Vector Regression	0.290	0.601

Figure 2.1: Correlation Coefficient, copied from [6]

Limitations:

The study is limited by the accessibility and thoroughness of data, as only 107 out of 118 census tracts were considered. This constraint indicates that the results may not accurately reflect the entire county, and incorporating more extensive data could improve the study. Furthermore, upcoming research could examine the application of feature selection in machine learning models to enhance predictive precision, as well as examine the impact of environmental elements, such as air quality and climate.

2.2.2 Second Regression Problem:

The research [33] aims to identify the underlying factors responsible for the disparities in COVID-19 mortality and infection rates among various nations. Utilising a blend of multivariate regression analysis and least absolute shrinkage and selection operator (LASSO) regression, the research scrutinises the impact of diverse demographic, socioeconomic, and healthcare system elements on the outcomes of COVID-9, such as population ages 70 and above, population density, urban population, hospital beds, vaccination, measles, smoking prevalence, extreme poverty rate, cardiovascular diseases, total COVID-19 tests, to identify the elements that are impacting the rates of COVID-19 outcomes.

Methodology:

This analysis is based on the "Our World in Data COVID-19 Dataset," which has been meticulously compiled by the University of Oxford in collaboration with the World Bank's World Development Indicators. Information from 186 countries is included in this dataset, with observations recorded through June 13, 2020. The objective of this analysis is to compare and pinpoint notable disparities across various variables among nations that have encountered differing outcomes related to COVID-19. The analytical techniques employed in this study include T-tests to identify differences between groups, a correlation matrix to examine relationships among various factors, as well as Ordinary Least Squares (OLS) and Least Absolute Shrinkage and Selection Operator (LASSO) regression methods to evaluate the effects of independent variables on dependent variables. Notably, LASSO regression is particularly advantageous for model selection and for addressing issues of multicollinearity.

Result:

The research indicates that there is a notable correlation between a higher COVID-19 death rate and an older population, as well as the number of hospital beds per 1,000 population. Specifically, an increase of 1% in the population aged 70 years and above is linked to a significant rise in the Case Fatality Rate. Furthermore closely connected to the urban population and the number of tests carried out per thousand individuals are the number of cases per million people. This implies that countries with higher degrees of urbanisation and denser populations as well as higher testing rates usually show more COVID-19 cases.

Limitations:

The research is constrained by its use of cross-sectional data, which may not entirely capture the evolving nature of the pandemic. Subsequent studies could investigate longitudinal data to analyze changes over time and include additional factors such as government interventions, cultural influences, and real-time data updates. Broadening the scope of the analysis to encompass more nations and more up-to-date data could also offer a more thorough comprehension of the factors impacting COVID-19 outcomes worldwide.

2.2.3 Third Regression Problem:

The study [12] is being conducted in Turkey, Using neural network models to predict the spread of COVID-19 and the mortality rate. The study's primary goal is to develop a model that can accurately predict the rates of infection and death associated with COVID-19 using data from Turkey and other countries. The goal of this predictive capability is to facilitate effective resource allocation and formulation of public health decisions during the pandemic. We employ specific independent variables like the population of the country, the total number of days, and the number of infections.

Methodology:

The study employed COVID-19 data sourced from the Ministry of Health of the Republic of Turkey and the World Health Organization, encompassing the countries of Turkey, Italy, Spain, France, the United States, and Iran. A multilayer perceptron feed-forward back-propagation consisting of 15 neurons in the hidden layer was constructed, utilizing the Levenberg-Marquardt algorithm for the purpose of training. The dataset was partitioned into 70% designated for training, 20% for validation, and 10% for testing.

Result:

The training phase of the ANN model yielded an R-value of 0.9995 (Figure 2.2). These results indicate that the model exhibits a high degree of precision in its ability to predict the mortality rate of COVID-19 in Turkey.

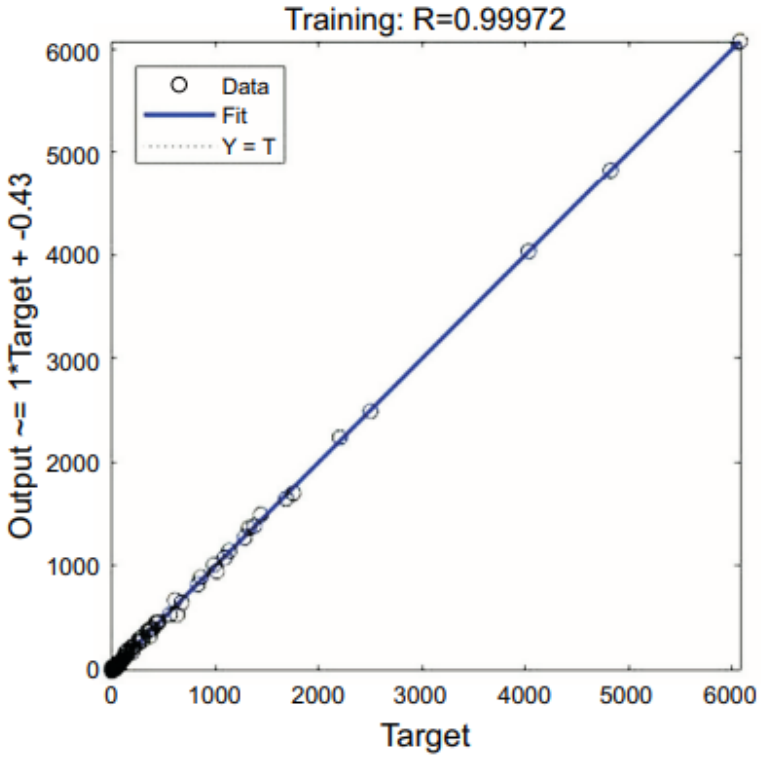

Figure 2.2: Training state, copied from [12]

Limitations:

The research is restricted by its focus on a specific group of six nations, which may restrict its ability to be generalised to other countries.

2.2.4 Fourth Regression Problem:

The study [37] is centered on the use of machine learning methods to predict COVID-19 trends, including confirmations, recoveries, and deaths, over a specified time frame in India in order to assess their ability to generate precise forecasts, which can be valuable for public health preparation and response. The study employs independent variables such as the country's population size, the total duration in days, and the infection count.

Methodology:

The World Health Organisation (WHO) has made the COVID-19 dataset for India available for use in the study. Furthermore, it integrates machine learning techniques, such as Support Vector Regression and Polynomial Regression, which are integrated within linear regression. Additionally, the research employs ARIMA, a well-established method

for forecasting time series data. In order to forecast future values within the series, this approach integrates autoregressive models and moving averages.

Result:

The RMSE of 31,465.7 was satisfactorily achieved by the Polynomial Regression model. Nevertheless, the ARIMA model outperformed it, achieving the lowest RMSE of 27,233 (Figure 2.3). Conversely, the Support Vector Regression model demonstrated inadequate performance, resulting in skewed results and a high RMSE of 775,124.

TABLE VII. RMSE FOR COMPARED MODELS

Stack rank	Model name	RMSE	Remarks
2	PR	31465.7	Second best
3	SVR	775124	Ignored as skewed results
1	ARIMA	27233	Best model

Figure 2.3: RMSE values of models, copied from [37]

Limitations:

The research is limited by its reliance on historical data from a designated period, which may not account for changes in the pandemic's trends over time. Future studies might explore additional machine learning models, such as ensemble methods or deep learning approaches.

Chapter 3

Methodology

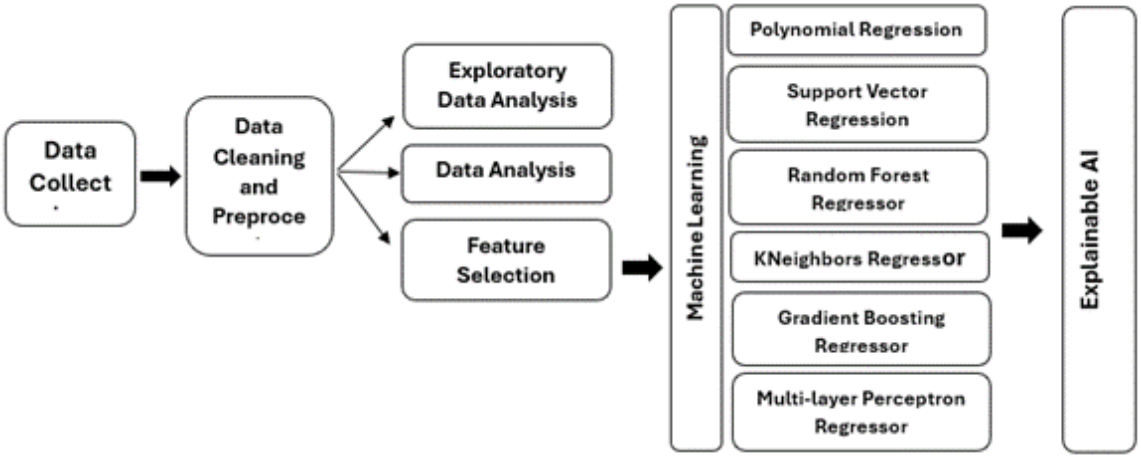

Figure 3.1: Methodology

3.1 Data Collection

The COVID-19 healthy diet dataset from Kaggle serves as the food data source for this analysis. The dependent features include a range of dietary components by country, specifically focusing on animal products, animal fats, cereals-excluding beer, eggs, fish, seafood, fruits-excluding wine, meat, milk-excluding butter, miscellaneous items, offals, oilcrops, pulses, spices, starchy roots, stimulants, sugar & sweeteners, treenuts, vegetal products, and vegetable Oils. The study's target feature is the death rate as of February 6, 2021.

The lifestyle and health data source for this analysis includes FAOSTAT, Our World in Data, WHO, and the World Bank Group, all from the year 2020. The dependent features examined are country, obesity, alcoholic beverages, diabetes prevalence, smoking, physical

activity, and tuberculosis. The target feature for this study is the death rate, as reported in the Kaggle dataset on February 6, 2021.
he demographic data source for this analysis includes Our World in Data and the World Bank Group, both from the year 2020. The dependent features considered are country, total population aged 0-14, total population aged 15-64, female population aged 15-64, male population aged 15-64, total population aged 65 and older, female population aged 65 and older, male population aged 65 and older, population density, and median age. The target feature for this study is the death rate, as reported in the Kaggle dataset on February 6, 2021.
The environmental data source for this analysis includes Our World in Data and the World Bank Group, both from the year 2020. The dependent features examined are country, PM2.5 concentration (MC), CO2 emissions, greenhouse gas emissions, methane emissions, nitrogen oxide emissions, and forest area. The target feature for this study is the death rate, as reported in the Kaggle dataset on February 6, 2021.

3.2 Data Cleaning And Preprocessing

- **Load and Explore Dataset:** Load one of four datasets (Food, demography, Lifestyle And Health, or Environment).
- **Handling missing values:** It is essential, particularly in the preparation of data for analysis or machine learning, remove them [2].
- **Handling Duplicate Data:** Assisting in the process of data cleansing contributes to the enhancement of data integrity. Therefore, we take measures to prevent the presence of duplicate values.
- **Train-Test Split:** The dataset will be divided into separate training and testing sets, with 80% of the data allocated for training and 20% for testing. This division will enable us to standardize each subset individually.
- **Standardization:** The process of data standardisation [4] offers significant advantages to SVR and ANN. Such standardisation plays a crucial role in reducing bias and skewness, particularly for models that depend on distance metrics and gradient-based optimisation techniques. Conversely, feature scaling has less impact on tree-based models like RFR and GBR because they base their decisions on the relative order of features rather than their size. Additionally, KNR often finds standardisation useful in classification situations to ensure accurate distance measurements. On the other hand, it may not be as useful in regression tasks. Standardisation can also help polynomial regression, especially when working with features that have higher

degrees, because it helps keep the numbers stable and makes optimisation processes more effective.

3.3 Exploratory Data Analysis

- **Descriptive Statistics:** Understanding the spread of data (25th, 50th, and 75th percentiles), identifying any absent values, and ensuring proper standardization of the data are crucial preprocessing steps, as well as determining the minimum and maximum values.
- **Distribution:** Data that displays symmetry or follows a normal distribution typically implies the presence of linear relationships, rendering linear regression models suitable. Conversely, when the data assumes intricate shapes, such as bimodal distributions, heavy tails, or skewness, it is probable that the relationships are non-linear.
- **Checking Linearity:** In statistical analyses, both linear and non-linear relationships between variables can be discerned using scatter plots, residual plots, and quantile-quantile plots. A scatter plot showing a straight-line pattern between an independent feature and a target suggests linearity, which is further affirmed by residual plots displaying random scatter around zero. Non-linear relationships, on the other hand, show up as cyclic patterns, curved lines, or clusters in scatter plots, and variable spreads or non-random patterns in residuals, which could be signs of variation or missing components in the model. quantile-quantile plots are indicated non-linear by deviations from a straight line or by the presence of outliers, which affects the validity of tests that assume normality. These diagnostic tools are essential for confirming hypotheses and directing the necessary model modifications to accurately capture the underlying dynamics of the data.

- **Correlation Analysis**
  - **Correlation Matrix Heatmap with Clustering:** We employed the spearman rank correlation method to investigate non-linear relationships. We made the decision not to eliminate countries with exceptionally high or low death rates as outliers as they provide important insights. Clustering is a useful way to look at and understand the connections between the things in your dataset by putting together the ones that are most closely linked. Correlation range

from -1 to 1, with -1 indicating a strong negative relationship and 0 indicating no relationship. A correlation coefficient of 1 indicating a strong positive relationship.

- **Bar Chart of Death Rate Correlation with Various Dataset Features:** We intend to examine the spearman rank correlation among independent variables within different categories and the mortality rate, with the specific aim of detecting positive and negative correlations. Through this approach,, thus illustrating direct and inverse associations in more focus than previous heatmap.

3.4 Data Analysis

- **Mortality rate Map:** Before we can understand how COVID-19 deaths will affect the world until June 2021, we need to look at how death rates vary between countries and regions. This first look will lead us to wonder why there are big differences in death rates, even between countries that are close to each other. It will also give us a starting point for more research into the real reasons behind these differences.

- **Death rate Categories:** To facilitate the comparison of countries by their mortality rates, we will categorize them into four groups depend on death rate for analytical purposes. This will be accomplished by employing the box-plot percentile and the Central Limit Theorem to guarantee that our sample size is sufficiently large ($X \geq 30$) [16, 35]. By dividing the data into quarterlies, we can classify countries with low death rates as Q1, median values as Q2, high rates as Q3, and very high death rates as Q4. This classification will enable us to effectively compare countries with low death rates to those with very high rates, to create a clear contrast between the extremes, which can reveal the most significant differences in factors affecting mortality rates.

- **Factor Map of Feature Importance via PCA:** Principal Component Analysis (PCA) [36] is used. In this case, factor maps do more than one thing: they make datasets simpler while keeping important data, which makes it easier to understand and see large sets of variables. PCA takes the original variables and turns them into a smaller group of independent parts that can represent most of the variation in the data. We can see deeper patterns and connections between the variables by displaying these parts, often through factor maps. PCA helps find the most important differences between countries with high and low death rates, which lets us study the main reasons for these differences in more detail.

3.5 Feature Selection

- **Variance Threshold:** This is a method for selecting features that is quite straightforward and mainly utilized for removing features with low variance. We will use the variance threshold [14] result for mutual information.
- **Mutual Information:** This method evaluates the relationship between each characteristic and the intended outcome [39], effectively capturing non-linear relationships through the use of Support Vector Regression with an RBF kernel to streamline the process. We subsequently used the Mutual Information (MI) analysis results as the input for feature selection methods, such as recursive feature elimination (RFE) and extra-tree regression.
- **Recursive Feature Elimination:** This is a method for selecting features that involves iteratively removing the least important features from the dataset and building the model [31], We applied the Random Forest Regression training algorithm and then optimised the selection of the optimal number of features.
- **Feature Selection Using ExtraTreesRegressor and SelectFromModel:** This method utilizes the ExtraTreesRegressor model to determine the most important features [17], The ExtraTreesRegressor gives the chosen features the highest importance scores based on how the SelectFromModel class rates them. This method works especially well for non-linear regression assignments because it makes sure that the most important variables are found and added to the model.

3.6 Machine Learning (Default hyperparameters)

We trained all machine learning models using standardised data with 5-fold cross-validation. The machine learning models used all independent features or feature selections (RFE, ExtraTrees) without KPCA, as well as with KPCA. The models included Polynomial Regression, which is non-linear; Support Vector Regression, which can be linear or non-linear depending on the kernel used; Random Forest Regressor, which is non-linear; KNeighborsRegressor, which is non-linear; Gradient Boosting Regressor, which is non-linear; and Multi-layer Perceptron Regressor, which is also non-linear.

3.7 Machine Learning (Hyperparameters tuning):

Utilizing standardized training data with machine learning models that employ GridSearchCV or RandomisedSearchCV (1000 iterations) and proper cross-validation for hy-

perparameter tuning, with or without KPCA, and feature selection (RFE, ExtracTrees), After tuning and training, we evaluated the optimal model on separate testing data.

- **Polynomial Regression :** The model used GridSearchCV for optimization, utilizing the following hyperparameters:

Polynomial Regression
poly_degree: 1, 2, 3, 4
poly_interaction_only:True, False
poly_include_bias: True, False
kpca_n_components:None, 2, 3, 4, 5, 6, 7
kpca_kernel: linear, poly, rbf, sigmoid

Figure 3.2: PL hyperparameters

- **Support Vector Regression:** The model used RandomizedSearchCV with 1000 iteration for optimization, utilizing the following hyperparameters:

Support Vector Regression
svr_kernel: None, linear, rbf, poly, sigmoid
Svr_epsilon: range(0.00, 0.101, 0.001(step))
svr_c: range (0, 151, 1(step))
kpca_n_components: None, 2, 3, 4, 5, 6, 7
kpca_kernel: None, linear, poly, rbf, sigmoid

Figure 3.3: SVR hyperparameters

- **Random Forest Regressor:** The model used RandomizedSearchCV with 1000 iteration for optimization, utilizing the following hyperparameters:

Random Forest Regressor
rf_n_estimators: 50, 100, 150, 200, 250, 300, 400, 500
rf_max_depth: None, 3, 6, 9, 12, 15
rf_min_samples_leaf: 1, 2, 3, 4, 5
rf_min_samples_split: 2, 4, 6, 8, 10
rf_max_leaf_nodes: None, 5, 10, 15, 20
kpca_n_components: None, 2, 3, 4, 5, 6, 7
kpca_kernel: None, linear, poly, rbf, sigmoid

Figure 3.4: RFR hyperparameters

- **KNeighborsRegressor:** The model used GridSearchCV for optimization, utilizing the following hyperparameters:

KNeighborsRegressor
knn_n_neighbors: 3, 5, 7, 10, 15
knn_weights: uniform, distance
knn_algorithm: auto, ball_tree, kd_tree, brute
knn_leaf_size: 20, 30, 40, 50
knn_p: 1, 2
kpca_n_components: None, 2, 3, 4, 5, 6, 7
kpca_kernel: None, linear, poly, rbf, sigmoid

Figure 3.5: KNR hyperparameters

- **Gradient Boosting Regressor:** The model used RandomizedSearchCV with 1000 iteration for optimization, utilizing the following hyperparameters:

Gradient Boosting Regressor
gbr_n_estimators: 50, 100, 150, 200
gbr_max_depth: 3, 6, 9
gbr_min_samples_split: 2, 4, 6
gbr_min_samples_leaf: 1, 2, 3
gbr_learning_rate: 0.01, 0.05, 0.1
gbr_max_features: None, sqrt, log2
kpca_n_components: None, 2, 3, 4, 5, 6, 7
kpca_kernel: None, linear, poly, rbf, sigmoid

Figure 3.6: GBR hyperparameters

- **Multi-layer Perceptron Regressor:** The model RandomizedSearchCV with 1000 iteration for optimization, utilizing the following hyperparameters:

Multi-layer Perceptron Regressor
mlp_hidden_layer_sizes: (50,), (100,), (50, 50), (100, 50).
mlp_activation: relu, tanh, logistic
mlp_solver: adam, lbfgs
mlp_alpha: 0.0001, 0.001, 0.01
mlp_learning_rate_init: 0.001, 0.01, 0.1
mlp_learning_rate: constant, adaptive
kpca_n_components: None, 2, 3, 4, 5, 6, 7
kpca_kernel: None, linear, poly, rbf, sigmoid

Figure 3.7: MLP hyperparameters

3.8 Explainable Artificial Intelligence

Local Interpretable Model-agnostic Explanations (LIME): This method effectively clarifies the outcomes of machine learning models, especially in situations where comprehending the rationale behind each prediction is essential. These methods analyse the prediction process for each instance, illustrating the influence of each input data feature on specific predictions. This study employs LIME to examine the determinants of mortality rates across different countries, focusing on those with the highest and lowest rates to identify key factors influencing these outcomes.

Chapter 4

Results and Findings

4.1 Food

4.1.1 Exploratory Data Analysis

- **Distribution Analysis:** Looking at the histograms in Figure 4.1, it's clear that some features have a right-skewed distribution with thick tails. This means that linear regression models might not be the best way to show the relationships. For instance, both the spices and sugar & sweeteners features demonstrate a pronounced right skew, with the majority of data points clustered at the lower end and a few values extending into significantly higher ranges. The vegetable oils feature displays a comparable distribution, characterised by a concentration of data on the left and an extended tail on the right. In addition, the stimulants feature exhibits right-skewed behaviour with heavy tails, indicating a preference for non-linear modelling approaches.

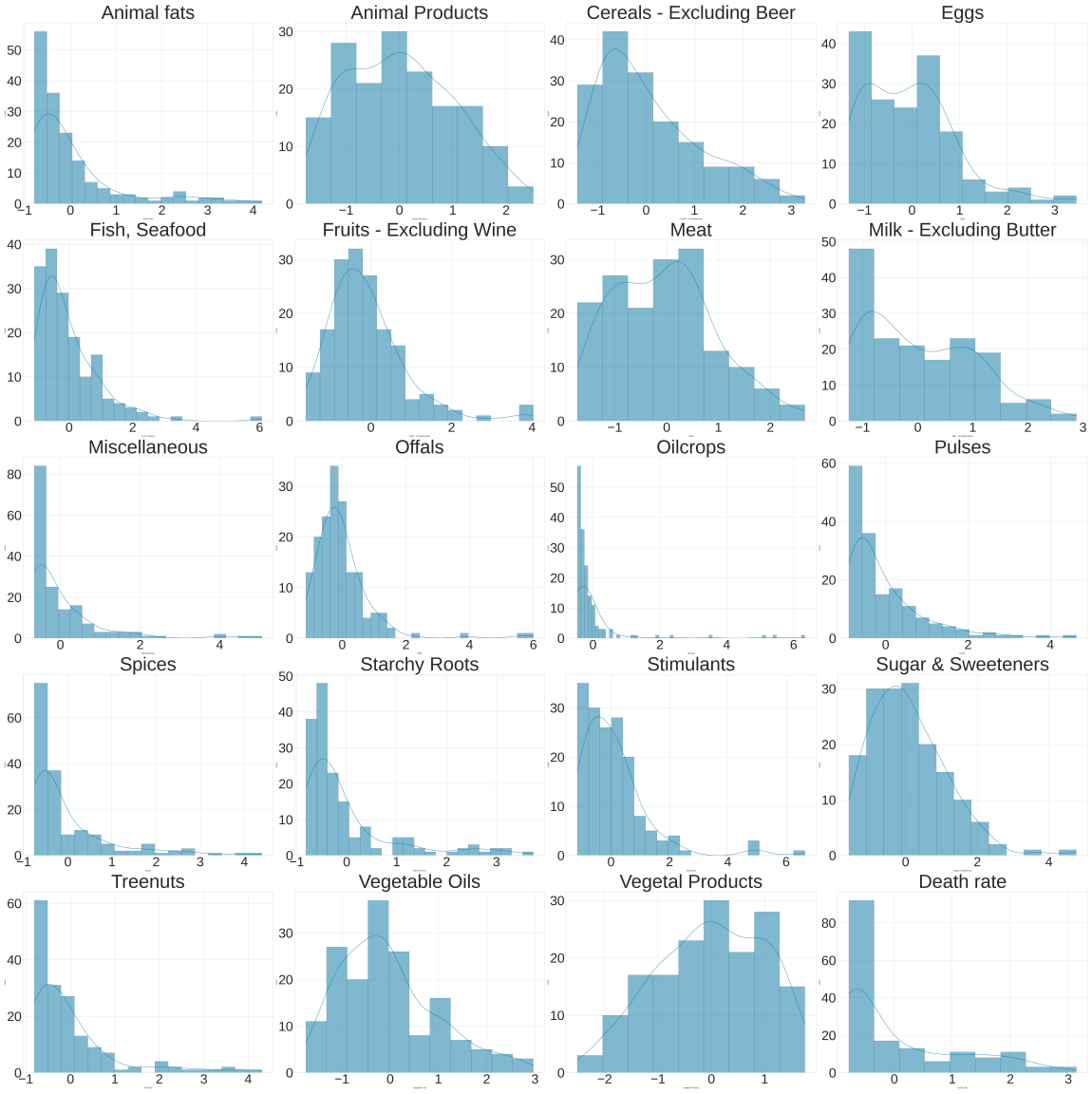

Figure 4.1: Food Features Histogram Plot

- **Correlation Analysis:** Figure 4.2 demonstrates a significant positive correlation between the death rate and the consumption of milk excluding butter (0.64), animal products (0.61), eggs (0.54), and animal fats (0.52). This suggests a robust association between elevated intake of these food items and an increased risk of mortality. Meat (0.34), stimulants (0.35), treenuts, and sugar & sweet (0.31) exhibit moderate positive correlations, suggesting a comparable yet somewhat weaker relationship. The consumption of vegetal products (-0.61), oil crops (-0.49), cereals excluding beer (-0.36), and pulses (-0.33) exhibits strong to moderate negative correlations, indicating an association between increased intake of these plant-based foods and reduced mortality rates. This indicates a distinct correlation between specific animal-derived

products and a heightened risk of mortality, whereas plant-based foods appear to offer a protective benefit.

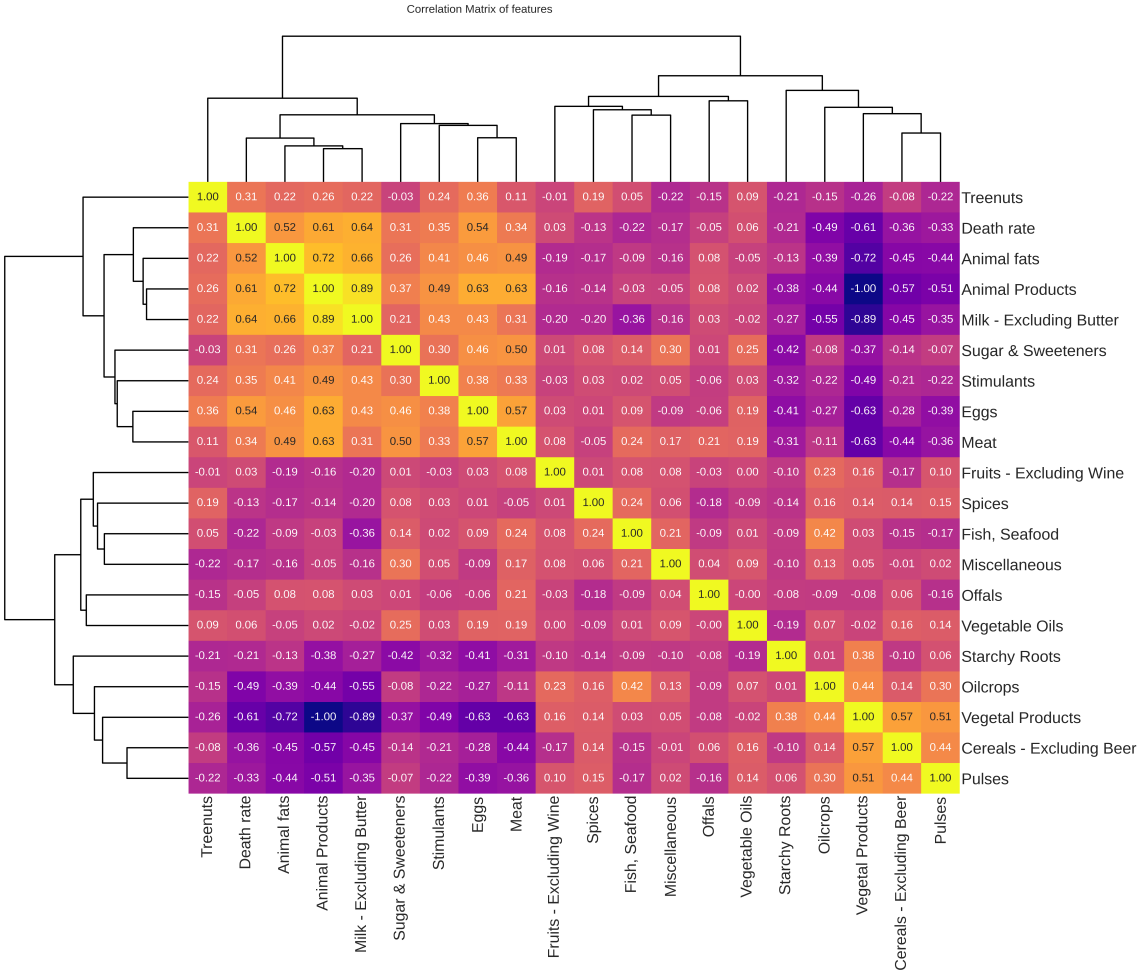

Figure 4.2: Food Heatmap Correlation Matrix

4.1.2 Data Analysis

- **Death rate Categories:** Based on specific cut-off points obtained from box-plot percentiles, we have grouped the 164 countries into four tiers of mortality rates: 'Low', 'Medium', 'High', and 'Very High'. Each tier consists of 41 countries, ensuring an equitable and statistically meaningful dataset for examination.

- **Factory Map:** This PCA diagram (Figure 4.3) illustrates the significant correlation between mortality rates and dietary patterns, emphasising the correlation between the consumption of specific food categories and health outcomes, particularly during pandemics. The potential risk factor is suggested by the close relationship between high mortality rates and increased consumption of animal products, including milk.

This is readily apparent from the red points, which denote countries with elevated mortality rates that are closely associated with these food categories. In contrast, the protective advantages of plant-based diets are emphasised by the correlation between low mortality rates and vegetal products. Stimulants and, sugar & sweeteners are centrally located on the factor map, which suggests that they may play a complex role in the potential affect on mortality that is not yet fully understood. This stance encourages additional investigation into the potential health consequences of stimulants and, sugar & sweeteners consumption, particularly in the context of their critical role during pandemics.

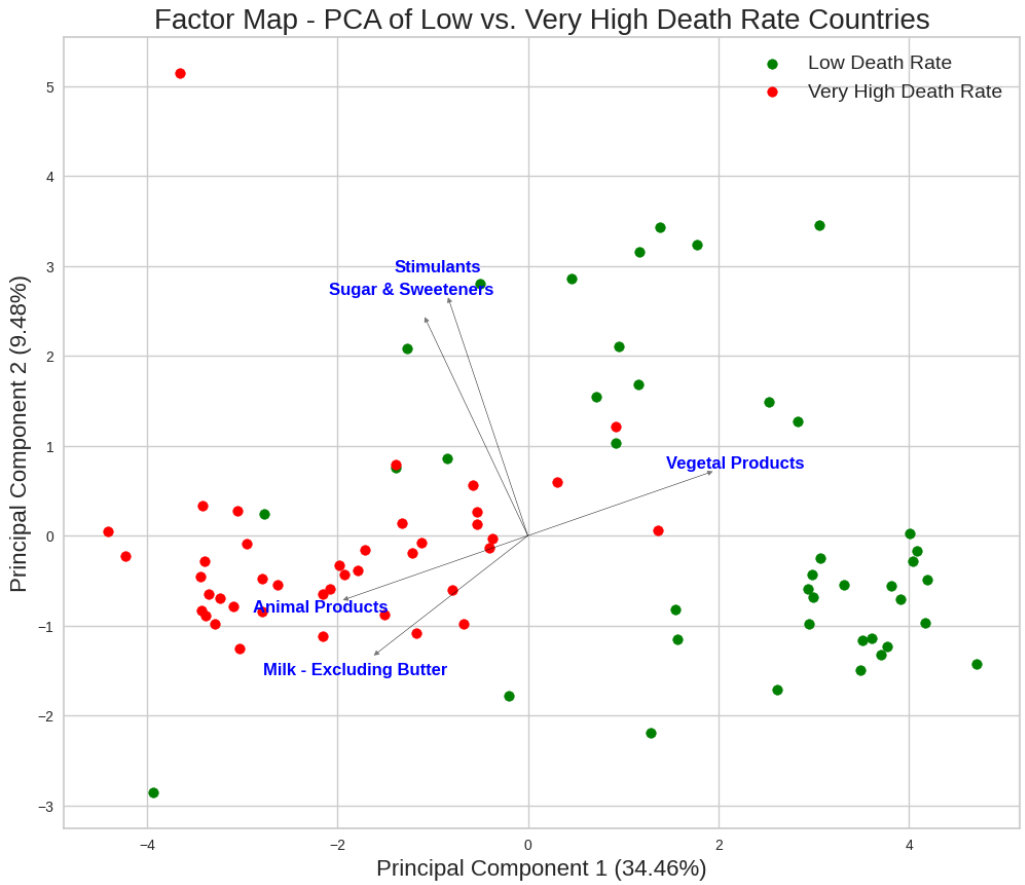

Figure 4.3: Compare Top vs Lowest Countries Features Importance

4.1.3 Machine Learning performance:

- **R2 scores:** It is evident that the Polynomial Regression model has the highest capacity to predict the mortality rate, as evidenced by the R2 scores in Figures (4.4, 4.5). The R2 score of 0.5303 is particularly impressive. This suggests that the model explains approximately 53.03% of the variation in the mortality rate data, thereby

establishing its superiority over other models. The model's high R2 score indicates a more precise fit to the data, capturing a greater proportion of the underlying patterns that influence mortality rates during the COVID-19 pandemic.

Models	R2 Score
Linear Regression for Polynomial Regression	0.5303
Support Vector Regression	0.3641
Random Forest Regressor	0.4609
K-Neighbors Regressor	0.4110
Gradient Boosting Regressor	0.1770
Multi-Layer Perceptron	0.4085

Figure 4.4: R-Square Score

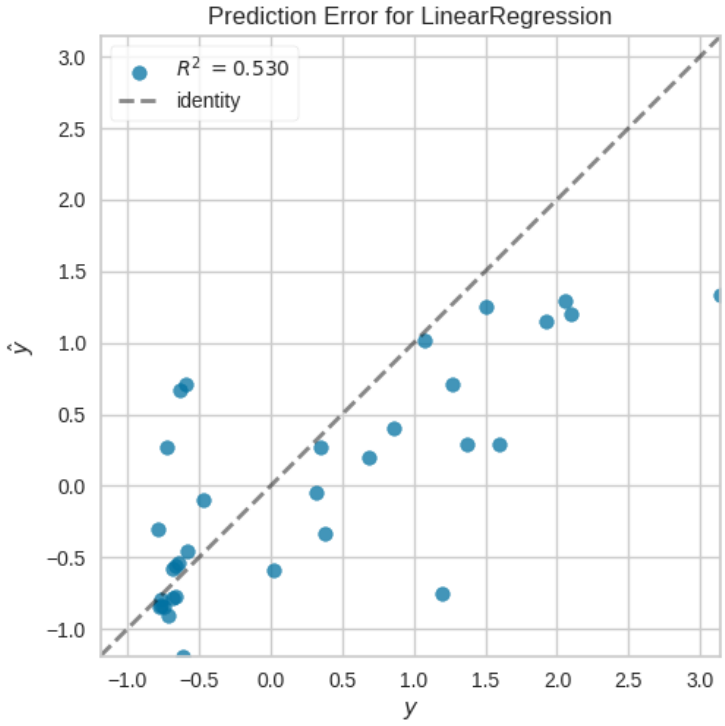

Figure 4.5: polynomial regression prediction

Correlation coefficient: The Polynomial Regression model is the leading model depicted in the chart, with a correlation coefficient of 0.7688 (Figure 4.6). This high correlation coefficient indicates that the model is more effective in capturing the

relationship between the input variables and the output than the SVR model used in the study [6], which has a correlation coefficient of 0.60 (Figure 2.1). By employing linear regression on polynomial features, this approach enhances the model's capacity to manage intricate, non-linear patterns in the data. As a result, the predictions are more precise than those of the other models.

Models	Correlation Coefficient
Linear Regression for Polynomial Regression	0.7688
Support Vector Regression	0.7076
Random Forest Regressor	0.7015
K-Neighbors Regressor	0.6645
Gradient Boosting Regressor	0.5059
Multi-Layer Perceptron	0.6952

Figure 4.6: Correlation Coefficient

4.1.4 Explainable Artificial Intelligence Using LIME

- **Countries with the Highest COVID-19 Mortality Rates:** In this analysis, we are comparing the primary projections for Belgium and the United Kingdom (UK). Each of these predictions is based on a multitude of critical factors. The model predicts a value of 1.22 in Belgium, but the real value is 1.33 (Figure 4.7-4.8). In the same way, the model predicts a value of 1.0 for the UK but the real value is 1.31 (Figure 4.9-4.10).

  Belgium sees a similar trend, though slightly different. animal products also adds 0.28 with a feature value of 0.96, and milk—excluding butter increases the prediction by 0.28 with a feature value of 1.05 as well. With an exceptionally high feature value of 2.35, animal Fats adds 0.25. Even with a raw value of -0.96, vegetal products adds 0.24 to the prediction, making a positive contribution. Fish and seafood also boost the prediction by 0.08, with a feature value of -0.19.

  With a feature value of 1.19 and an increase in projection of 0.32, the category milk—excluding butter has a notably positive impact in the UK. Comparably, animal fats add 0.25 with a feature value of 0.22 and "animal products" contribute 0.27 with a high consumption feature value of 1.15. Furthermore, vegetal products has a negative feature value of -1.16 but adds 0.28 to the projection. Fish and seafood also boost the prediction by 0.04, with a feature value of -0.23.

  A decrease in the consumption of fish and seafood and plant-based foods is correlated with an increase in the forecast in both cases, though the latter's influence is not as great. Not to mention the expected health benefits of milk, animal products, and

animal fats. In Belgium, the greater influence of products derived from animals is particularly apparent. Overall, the main variables in the projections for both countries remain to be milk, animal products, and low consumption of plant-based foods.

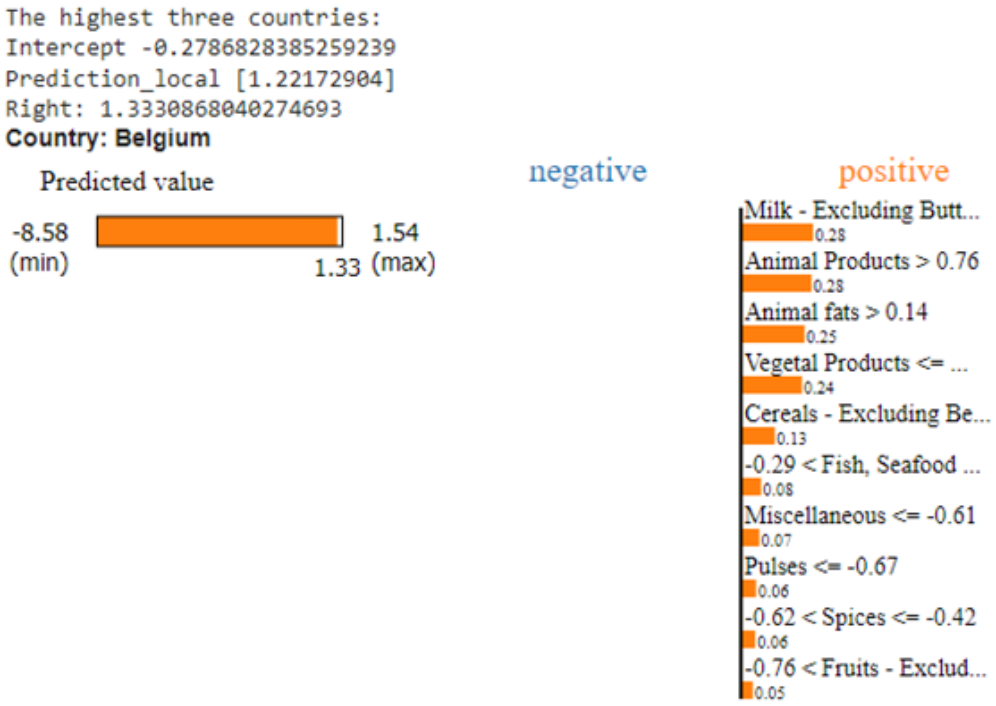

Figure 4.7: X AI regression Predication: Lime (Belgium)

Feature	Value
Milk - Excluding Butter	1.05
Animal Products	0.96
Animal fats	2.35
Vegetal Products	-0.96
Cereals - Excluding Beer	-0.84
Fish, Seafood	-0.19
Miscellaneous	-0.65
Pulses	-0.71
Spices	-0.44
Fruits - Excluding Wine	-0.55

Figure 4.8: Belgium

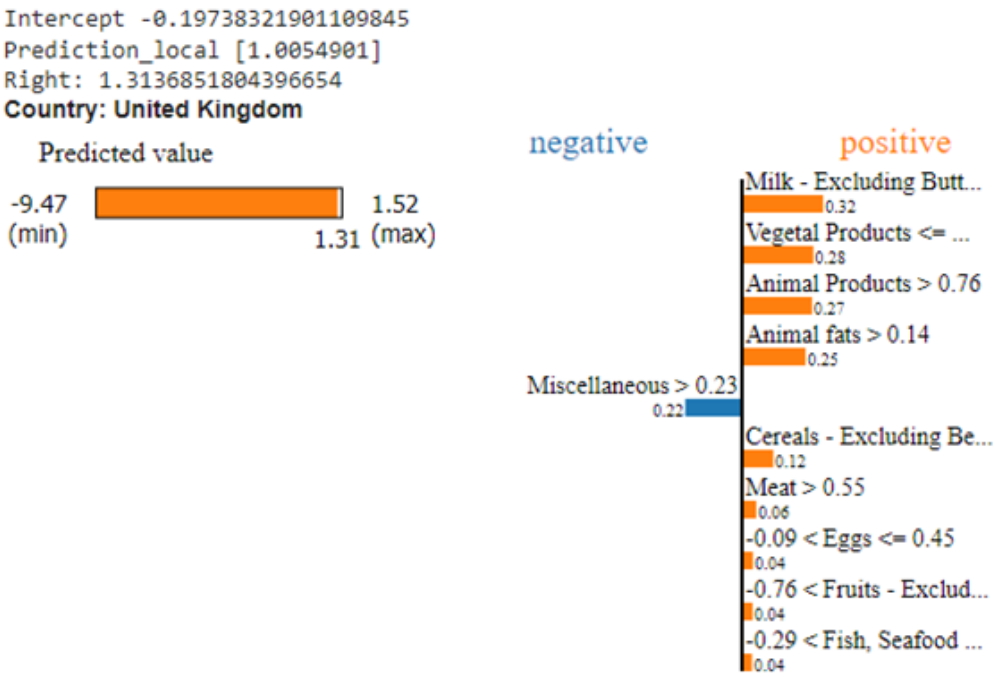

Figure 4.9: X AI regression Predication: Lime (United Kingdom)

Feature	Value
Milk - Excluding Butter	1.19
Vegetal Products	-1.16
Animal fats	0.22
Animal Products	1.15
Miscellaneous	0.38
Cereals - Excluding Beer	-0.86
Sugar & Sweeteners	-0.18
Oilcrops	-0.34
Fruits - Excluding Wine	-0.31
Fish, Seafood	-0.23

Figure 4.10: Feature values (United Kingdom)

- **Countries with the Lowest COVID-19 Mortality Rates:** In Vanuatu, the model predicts a mortality rate of -0.814, compared to the actual rate of -0.670 (Figure 4.11-4.12). For Saint Kitts and Nevis, the model foresees a mortality rate of -0.594 in comparison to the actual rate of -0.438 (Figure 4.13-4.14).
  Vanuatu experiences positive effects from fish and seafood, which lowers the projection by 0.30 with a feature value of 0.65. vegetal products improve the forecast by 0.23, with a feature value of 0.93. mil-excluding butter also contributes positively, decreasing the forecast by 0.21 with a feature value of -1.02. However, animal prod-

ucts and animal fats have negative impacts with feature values of -0.93 and -0.51, reducing the forecast by 0.25 and 0.10, respectively.

In Saint Kitts and Nevis, favourable factors such as fish and seafood, with a feature value of 1.67, significantly contribute to a decrease in the mortality projection by - 0.29. Furthermore, the feature value of -0.89 indicates a decrease in milk consumption, which reduces the prediction by 0.25. Negative elements also influence the model's prediction. cereals – excluding beer, despite a negative feature value of -0.90, increasing mortality by 0.09. However, animal products, with a feature value of 0.33, have a positive impact on the forecast by 0.07.

In brief, the study emphasizes that in both Saint Kitts and Nevis and Vanuatu, diets rich in fish and seafood play a significant role in lowering mortality estimates, demonstrating their favorable effect on health outcomes. Furthermore, plant-based foods, particularly in Vanuatu, also have a favourable impact on mortality projections. Conversely, animal-based products and fats have a detrimental effect on mortality predictions in both regions.

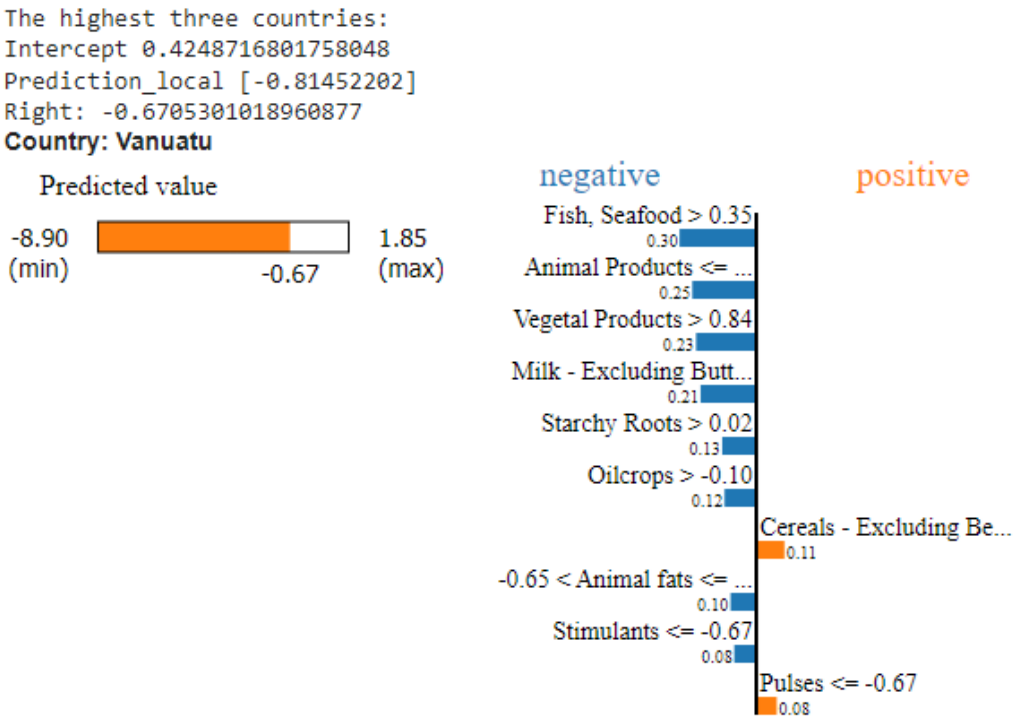

Figure 4.11: X AI regression Predication: Lime (Vanuatu)

Feature	Value
Fish, Seafood	0.65
Animal Products	-0.93
Vegetal Products	0.93
Milk - Excluding Butter	-1.02
Starchy Roots	1.21
Oilcrops	6.34
Cereals - Excluding Beer	-0.88
Animal fats	-0.51
Stimulants	-0.99
Pulses	-0.89

Figure 4.12: Feature values (Vanuatu)

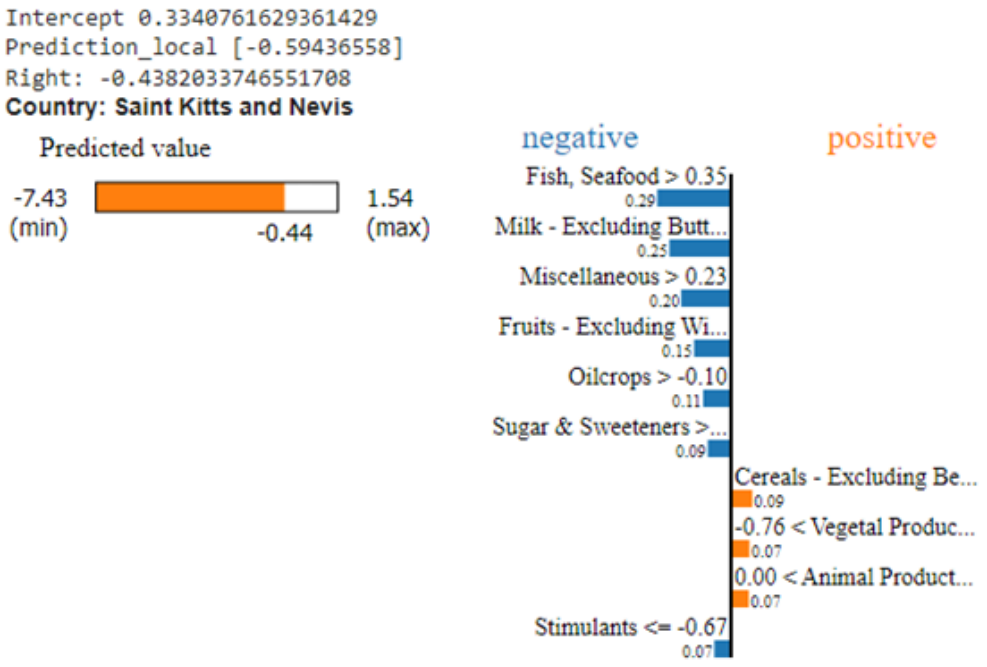

Figure 4.13: X AI regression Predication: Lime (Saint Kitts and Nevis)

Feature	Value
Fish, Seafood	1.67
Milk - Excluding Butter	-0.89
Miscellaneous	1.75
Fruits - Excluding Wine	0.47
Oilcrops	0.06
Sugar & Sweeteners	2.55
Cereals - Excluding Beer	-0.90
Vegetal Products	-0.33
Animal Products	0.33
Stimulants	-0.72

Figure 4.14: Saint Kitts and Nevis

In Belgium and the UK, higher milk consumption is associated with higher mortality estimates. Lower milk consumption, on the other hand, is associated with better health outcomes and lower mortality estimates in Saint Kitts, Nevis, and Vanuatu. This suggests that the health effects of milk may differ based on the amount and fat content consumed, necessitating additional research. In addition, eating animal fats and products usually results in higher mortality rates. Fish and seafood are good for your health in Vanuatu, Saint Kitts and Nevis, but not so much in Belgium and the UK. Plant-based foods typically lower mortality rates in countries like Vanuatu, but insufficient consumption of these foods may raise mortality rates in Belgium and the UK, highlighting the possible health risks of inadequate plant-based diets.

4.2 Demography

4.2.1 Exploratory Data Analysis

- **Distribution Analysis:** It is evident from the histograms (Figure 4.15) that several of the characteristics have distributions that are right-skewed and have heavy tails. This implies that discovering the underlying relationships may not be the best use case for linear regression models. The "65+ Total," "65+ Female," and "65+ Male" attributes, for example, all show a noticeable right skew, with most data points concentrated on the lower end and a small number of values extending into much higher ranges. The characteristics of "population density" and "death rate" also show a right-skewed distribution, with a long tail extending to the right and most of the data concentrated on the lower end. Moreover, the distributions shown by the "15-64 Total," "15-64 Female," and "15-64 Male" characteristics show that, even

in cases where linear models are not as strongly skewed, care should be taken. The "0-14 Total" and "Median Age" characteristics, on the other hand, seem to have a more balanced distribution, though they still exhibit some skewness. The histograms' patterns indicate that non-linear modelling techniques or transformations might be required in order to accurately depict the relationships in this demographic data.

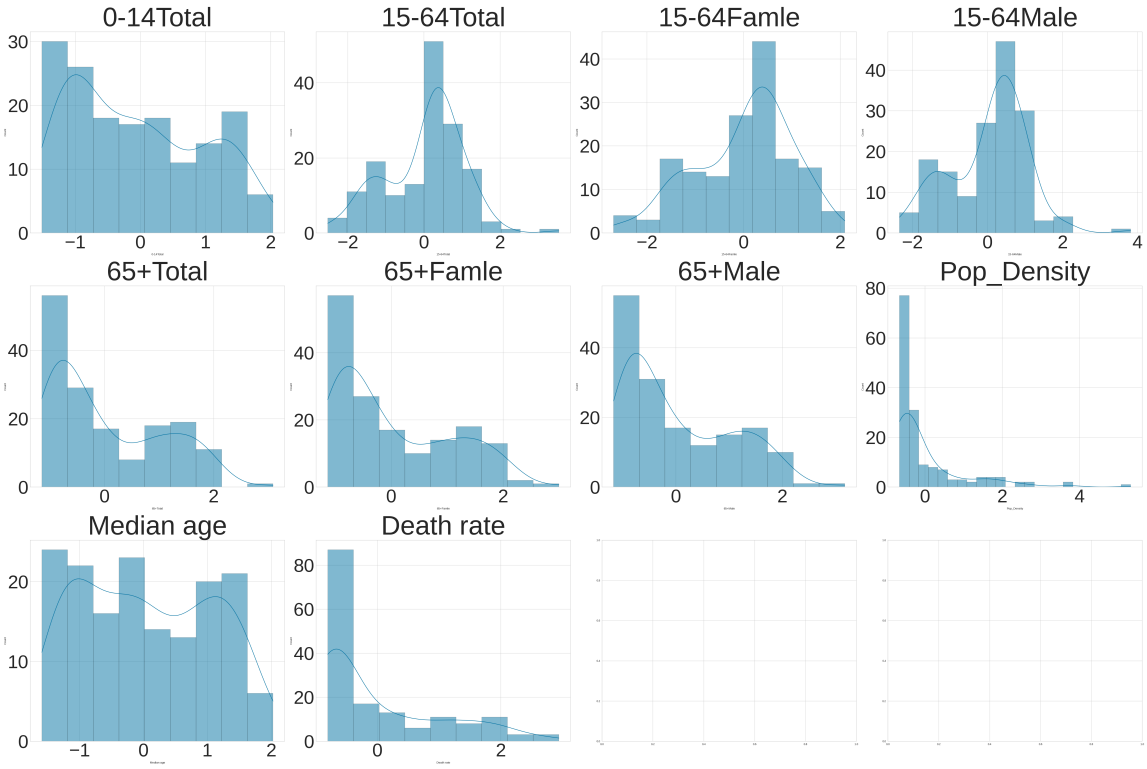

Figure 4.15: Demography Features Histogram Plot

- **Correlation Analysis:** In populations with higher death rates, certain demographic characteristics demonstrate clear spearman rank correlations that underscore their impact on mortality (Figure 4.16). For instance, the percentage of children aged 0–14 is strongly inversely associated with death rates (-0.62), indicating that a younger population may act as a protective factor against high mortality. Conversely, the percentage of individuals aged 65 and over, especially females, exhibits a strong positive correlation (0.64), implying that an older population is more susceptible to higher death rates. This pattern holds true for both males (0.62) and the overall elderly population (0.63), underscoring the heightened risk for older adults in these circumstances. Interestingly, there is a strong positive correlation (0.58) between population density and death rates, indicating that higher densities may have higher death rates for various reasons, including healthcare accessibility or the spread of disease. The population's median age is another important factor; it has a strong

positive correlation (0.64) with death rates, meaning that the mortality rate rises in tandem with the population's median age.

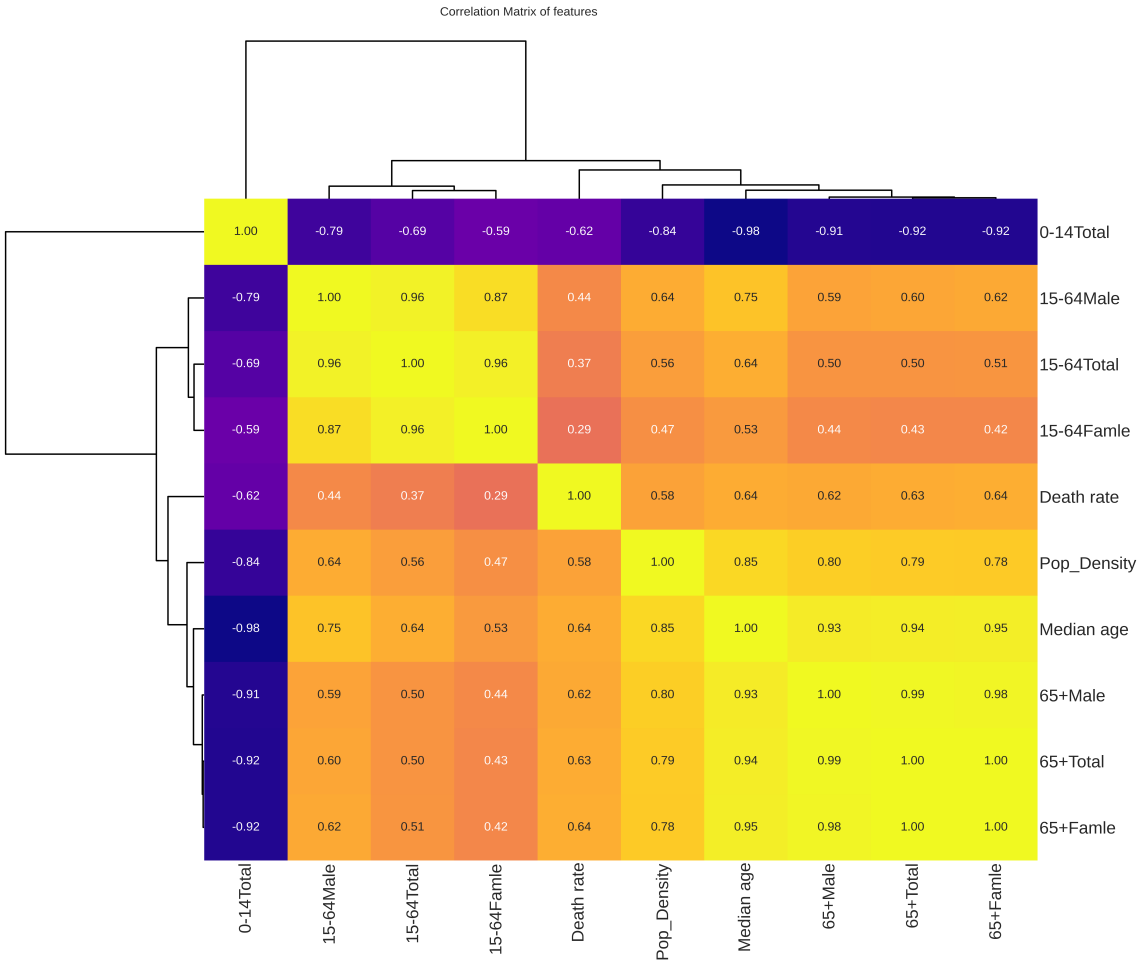

Figure 4.16: Demography Heatmap Correlation Matrix

4.2.2 Data Analysis

- **Death rate Categories:** We determined specific cut-off points based on box-plot percentiles. Based on the rate of mortality, we have divided the 159 nations into four tiers: "low," "medium," "high," and "very high." These tiers consist of 42 countries in the 'Low' classification, 37 countries in the 'Medium' classification, 40 countries in the 'High' classification, and 40 countries in the 'Very High' classification.
- **Factory Map:** The PCA plot (Figure 4.17) highlights age-related patterns by distributing the nations based on their death rates and demographic characteristics. A correlation between lower mortality rates and younger populations is suggested by the plot, which primarily places countries with low death rates (green points) on the right side, closely associated with younger age groups, especially the 0-14

category. Conversely, nations with extremely high death rates (red dots) tend to be more dispersed, but mostly focus on the left, where older age groups—especially those 65 and older—have a stronger influence. The male and female age groups between 15 and 64 are closest to the centre, which suggests a moderate but significant impact on death rates.

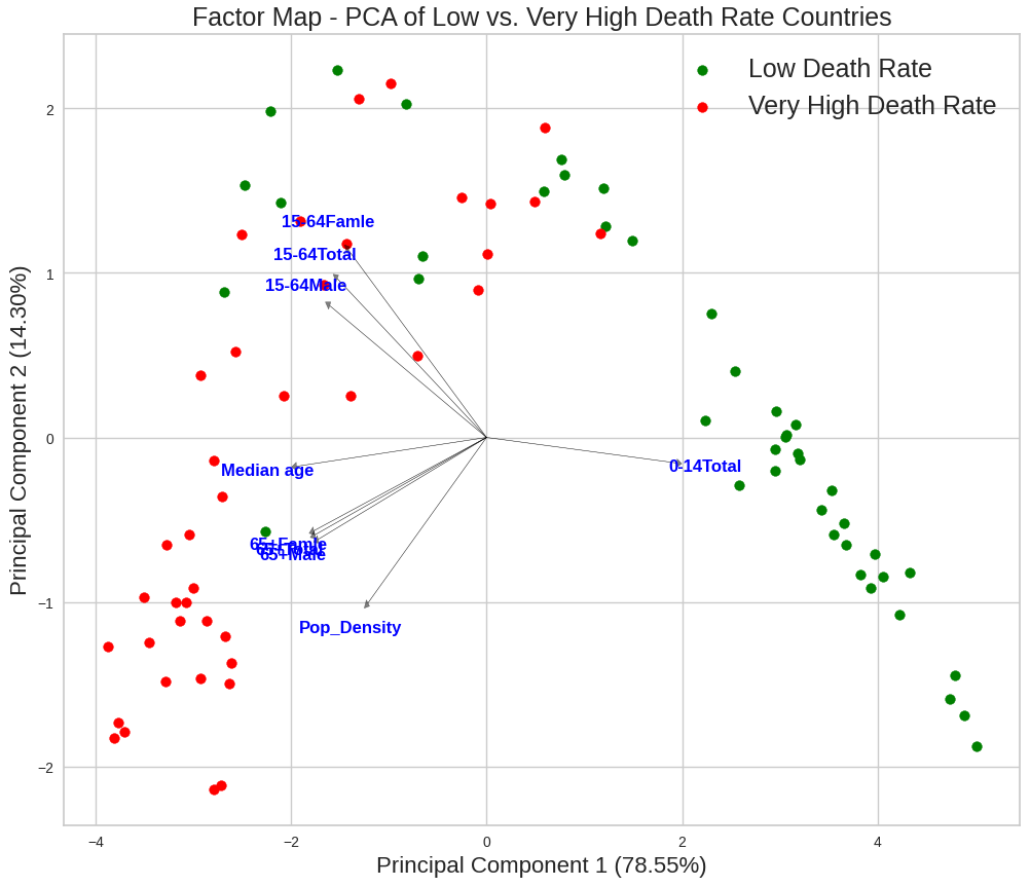

Figure 4.17: Compare Top vs Lowest Countries Features Importance

4.2.3 Machine Learning performance:

- **R2 scores:** In (Figure 4.18-4.19) With an R2 score of 0.1567, the gradient boosting regression model performs better and is thought to account for 15.67% of the variability in the target variable. However, even though the model outperforms other models, the comparatively low R2 score highlights the model's shortcomings in fully capturing the variance in the dataset.

Models	R2 Score
Polynomial Linear Regression	0.0736
Support Vector Regression	- 0.1441
Random Forest Regressor	0.0422
K-Neighbors Regressor	0.0683
Gradient Boosting Regressor	0.1567
Multilayer Perceptron	0.0314

Figure 4.18: R-Square Score

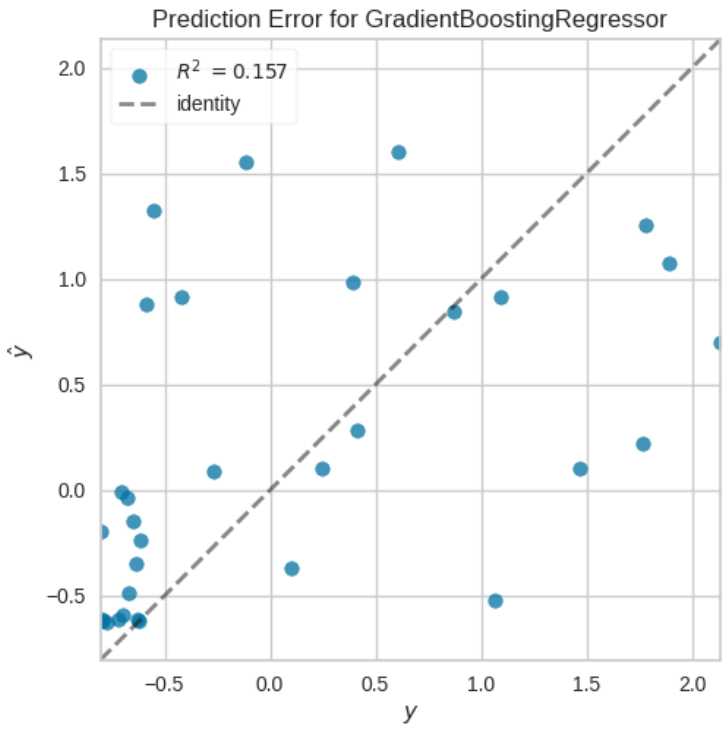

Figure 4.19: Gradient Boosting Regressor

Correlation Coefficient: The top-performing model displayed in the Figure 4.20 is the Polynomial Linear Regression, boasting a correlation coefficient of 0.5258. This figure signifies the model's median effectiveness in capturing the relationship between the input variables and the output. Through the application of linear regression to polynomial features.

Models	Correlation Coefficient
Polynomial Linear Regression	0.5258
Support Vector Regression	0.4043
Random Forest Regressor	0.5142
K-Neighbors Regressor	0.4860
Gradient Boosting Regressor	0.4983
Multilayer Perceptron	0.5126

Figure 4.20: Correlation Coefficient

4.2.4 Explainable Artificial Intelligence Using LIME

- **Countries with the Highest COVID-19 Mortality Rates:** The projected value of 0.038 for Belgium deviates significantly from the original value of 1.4351, as can be seen by comparing the original data and the model's projections (Figure 4.21-4.22). The primary determinants of this prediction are age groups, gender, and population density; all three have magnitudes less than 0.04. These discrepancies may not exactly match the initial value in the PCA factor map or the heatmap correlation matrix due to potential biases or differences in the model.

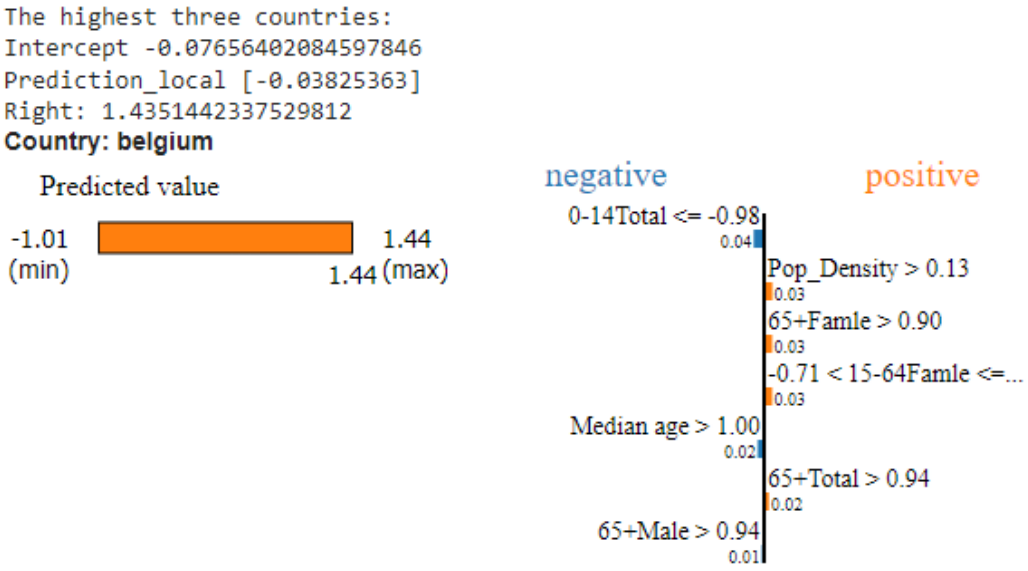

Figure 4.21: X AI regression Predication: Lime (Belgium)

Feature	Value
0-14Total	-1.05
Pop_Density	1.66
65+Famle	1.46
15-64Famle	0.03
Median age	1.28
65+Total	1.52
65+Male	1.58

Figure 4.22: Feature values (Belgium)

- **Countries with the Lowest COVID-19 Mortality Rates:** In this case, after analysing the original data and the model's forecasts for Cambodia, it is clear that the right value of -0.40 differs greatly from the forecasted value of -0.12 (Figure 4.23-4.24). This forecast is based on factors such as population density and gender age groups, all of which have a magnitude less than 0.03. These differences may not fully coincide with the initial value in the heatmap correlation matrix or the PCA factor map because of possible biases or discrepancies in the model.

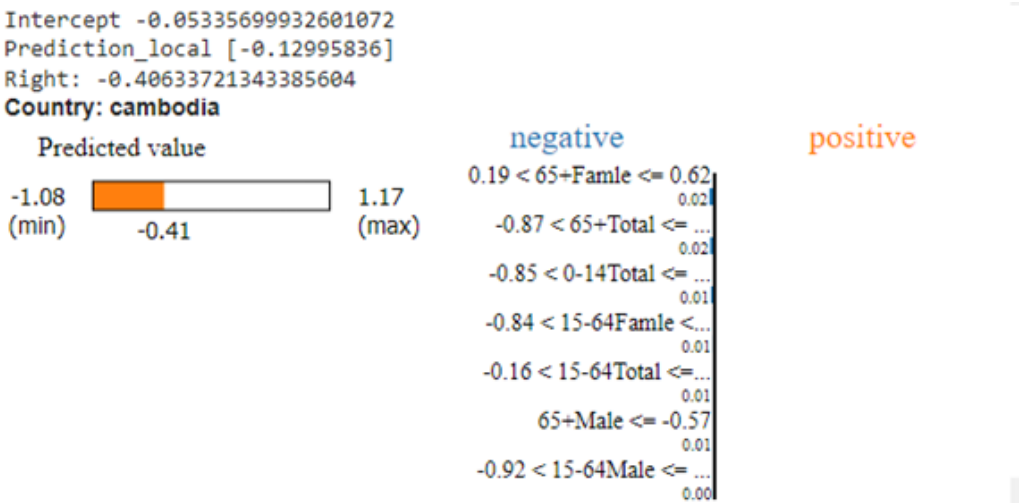

Figure 4.23: X AI regression Predication: Lime (Cambodia)

Feature	Value
65+Famle	0.48
65+Total	-0.54
0-14Total	-0.60
15-64Famle	-0.67
15-64Total	0.18
65+Male	-0.60
15-64Male	-0.42

Figure 4.24: Feature values (Cambodia)

XAI focused on nations like Belgium and Cambodia while utilising lime to evaluate the difference between the actual and projected values produced by the machine learning model. This illustrates how the model's poor predictive ability results from its inability to appropriately account for the influence of particular factors such as age groups. For example, the model gives little weight to important demographic factors such as females over 65, even though the dataset has a strong correlation. This leads to a decrease in prediction accuracy.

4.3 Lifestyle And Health

4.3.1 Exploratory Data Analysis

- **Distribution Analysis:** The study of the histograms (Figure 4.25) shows a number of features that have right-skewed distributions with heavy tails. This suggests that linear regression models might have trouble capturing the underlying relationships. For instance, the "Diabetes Prevalence," "Undernourished," and "Death Rate" features present strong right skewness, with the majority of data points concentrated at the lower end and a few values extending into significantly higher ranges. Similarly, the "Alcoholic Beverages," "Tuberculosis," and "Smoking" features also exhibit right-skewed distributions, albeit with varying degrees of skewness. These features concentrate most of their data on the lower end, with long tails extending to the right, potentially introducing heteroscedasticity and impacting the performance of linear models. Conversely, features like "obesity" and "physical activity" demonstrate distributions that, while still skewed, are closer to a normal distribution. These may be more suitable for linear modeling than the more heavily skewed features. These observed patterns across the histograms suggest that non-linear modelling approaches.

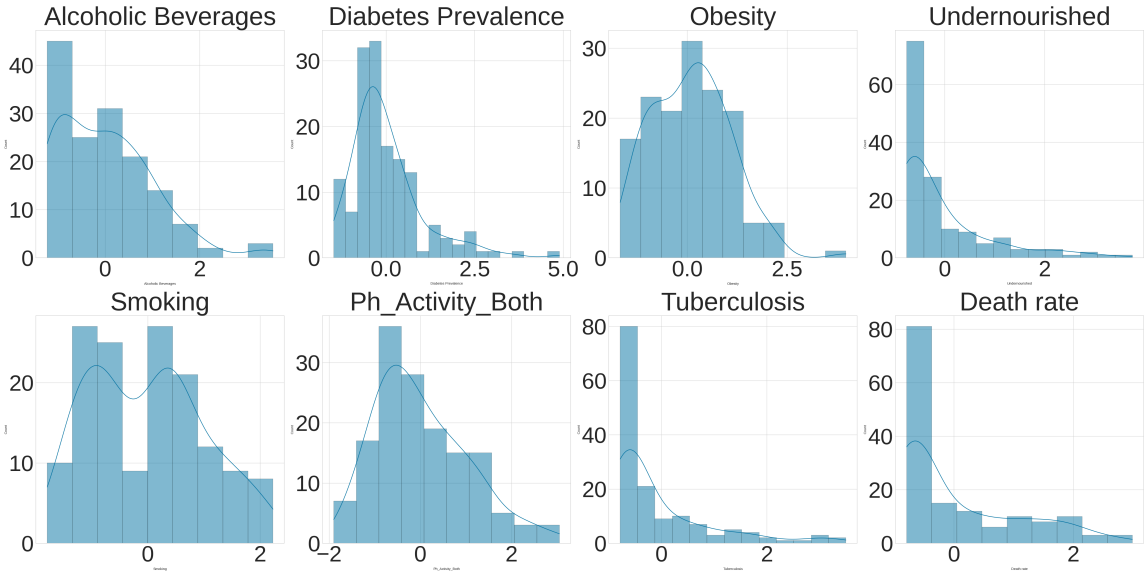

Figure 4.25: Demography Features Histogram Plot

- **Correlation Analysis** In populations with higher rates of mortality, certain lifestyle factors demonstrate distinct spearman rank correlations (Figure 4.26) that emphasize their impact on death. For instance, "Obesity" has a strong positive correlation with mortality rates (0.48), indicating that higher levels of obesity are closely linked to increased death. Similarly, the consumption of "Alcoholic Beverages" also displays a significant positive correlation (0.45) with mortality rates, suggesting that alcohol use is another important factor contributing to higher death rates. Intriguingly, the "Undernourished" feature shows a strong negative correlation with mortality rates (-0.56), implying that in populations where undernourishment is more common, the death rate tends to be lower. This could be due to the inverse relationship between undernourishment and other factors like obesity that are more directly linked to higher mortality. "Smoking" is positively correlated with mortality rates 0.31, indicating that smoking is associated with increased mortality. Nonetheless, although noteworthy, the influence of these variables is not as strong as that of obesity and alcohol intake. The negative value of -0.53 found in the correlation between "Tuberculosis" and death rates could be attributed to complex socioeconomic or healthcare-related factors.

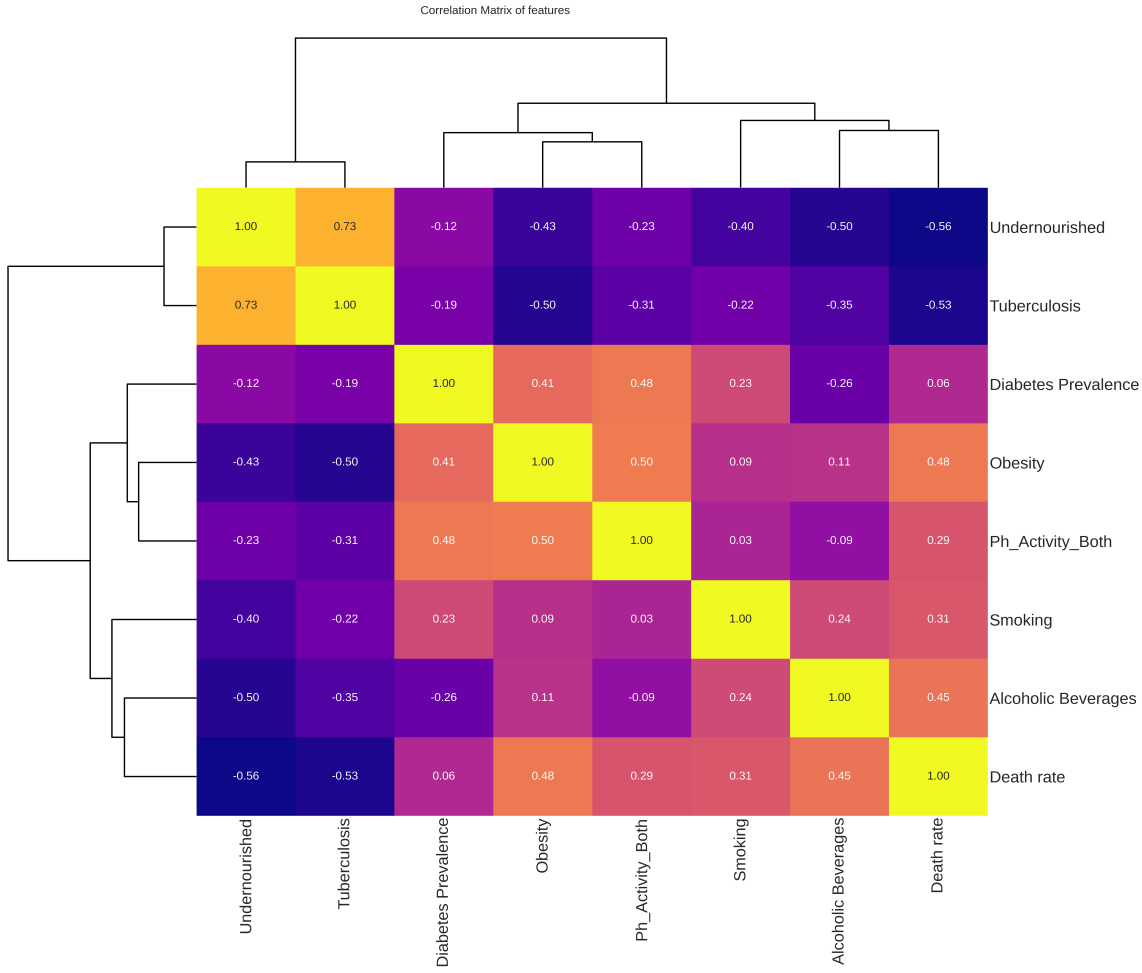

Figure 4.26: Lifestyle And Health Heatmap Correlation Matrix

4.3.2 Data Analysis

- **Death rate Categories:** We have classified the 148 nations into four tiers of mortality rates—'low,"medium,' 'high,' 'and'very high'—utilizing box-plot percentiles . The breakdown is as follows: 35 nations in the 'Low' tier, 39 nations in the 'Medium' tier, 37 nations in the 'High' tier, and 37 nations in the 'Very High' tier.

- **Factory Map:** The PCA biplot (Figure 4.27) illustrates the correlation between higher smoking rates and increased mortality. Likewise, alcohol consumption seems to be associated with significantly higher death rates, with the direction for alcoholic beverages and smoking pointing toward the red points. We demonstrate the important roles that disease management and healthcare quality play in this association. Moreover, a more robust association is observed between the prevalence of obesity and diabetes and nations with exceptionally high death rates, highlighting

the noteworthy influence of these conditions on mortality. On the other hand, undernourishment and tuberculosis are strongly linked to nations with low death rates (green points). This surprising connection may highlight the complex interactions between mortality and health indicators.

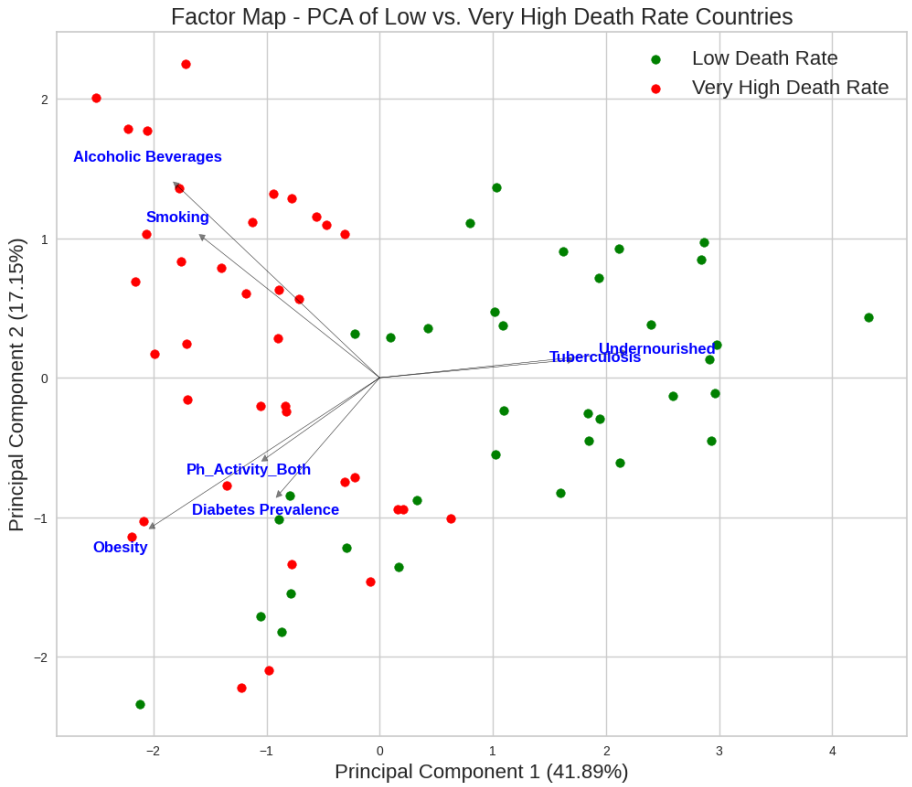

Figure 4.27: Compare Top vs Lowest Countries Features Importance

4.3.3 Machine Learning performance:

- **R2 scores:** The PL model has the highest R2 score of 0.4631, making it the best model for predicting the death rate in figures (4.28–4.29). This indicates that the model explains about 46.31% of the variation in the data, making it the best model among those analysed. However, the KNR model almost matches the PL performance, albeit with a marginally lower R2 score of 0.4618. However, each model's comparatively low R2 values indicate that they had difficulty illuminating the underlying patterns in the data. This could indicate that the models were unable to account for other significant factors, or that the data was extremely complex or noisy.

Models	R2 Score
Polynomial Linear Regression	0.4631
Support Vector Regression	0.3783
Random Forest Regressor	0.3655
K-Neighbors Regressor	0.4618
Gradient Boosting Regressor	0.3026
Multilayer Perceptron	0.3623

Figure 4.28: R-Square Score

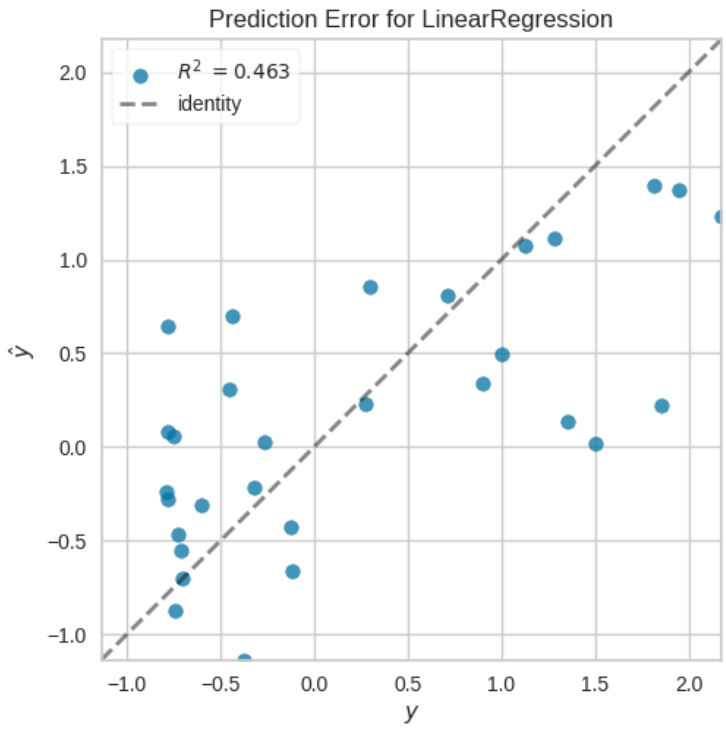

Figure 4.29: Polynomial Linear Regression

Correlation Coefficient: With a correlation coefficient of 0.683, the KNR is the most successful model shown in Figure 4.30. This high number suggests that the model is the most dependable option in this analysis since it accurately depicts the relationship between the input variables and the output. The PL model, with a correlation coefficient of 0.682, comes in right behind. This model also shows a high level of skill in identifying the relationship between the input and output variables.

Models	Correlation Coefficient
Polynomial Linear Regression	0.6825
Support Vector Regression	0.6464
Random Forest Regressor	0.6065
K-Neighbors Regressor	0.6836
Gradient Boosting Regressor	0.5897
Multilayer Perceptron	0.6063

Figure 4.30: Correlation Coefficient

4.3.4 Explainable Artificial Intelligence Using LIME

- **Countries with the Highest COVID-19 Mortality Rates:** The forecasting models for Belgium and the United Kingdom are affected by a variety of factors, including health and lifestyle characteristics. In Belgium, the observed mortality rate is 0.753, while the model's forecast is 0.5703(Figures 4.31-4.32), For the United Kingdom, the forecast value is 0.5188, whereas the true model value is 0.6943 (Figures 4.33-4.34).

  In term of Belgium, "Alcoholic Beverages" emerges as a significant factor in this model, with a substantial positive impact of 0.51 with a median feature value of 0.91. Remarkably, "Obesity," with a feature value of almost zero (0.02), has a marginally negative effect and lowers the prediction. Despite having negative feature values of -0.73 and -0.94, respectively, "Tuberculosis" and "Diabetes Prevalence," which have impacts of 0.20 and 0.08, respectively, are other significant factors.

  Additionally, "smoking" and "undernourished" have a small impact on the prediction, with feature values of 0.41 and -0.76, respectively. Drinking alcohol is clearly the main factor influencing the prediction.

  In the United Kingdom, "Alcoholic Beverages" has a positive feature value of 0.83 and an impact of 0.47. With an impact of 0.21 of "tuberculosis," it nevertheless makes a positive contribution to the forecast despite having a negative feature value of -0.73. The feature value for "obesity" is higher at 0.67, adding 0.07 to the prediction. "Smoking" has a minor impact of -0.06 and a feature value of -0.47, indicating a slightly negative influence. Other factors that have little to no effect on the prediction are "Physical Activity" and "Diabetes Prevalence," both of which have tiny positive impacts of 0.04 and less.

  The highest-ranking high-impact factor in Belgium and the UK is "Alcoholic Beverages," which significantly influences the forecasts. Despite having negative feature values, some health factors, such as "Tuberculosis" and "Diabetes Prevalence," nevertheless have positive effects in both models, demonstrating the complex relationship

between feature values and their effects.

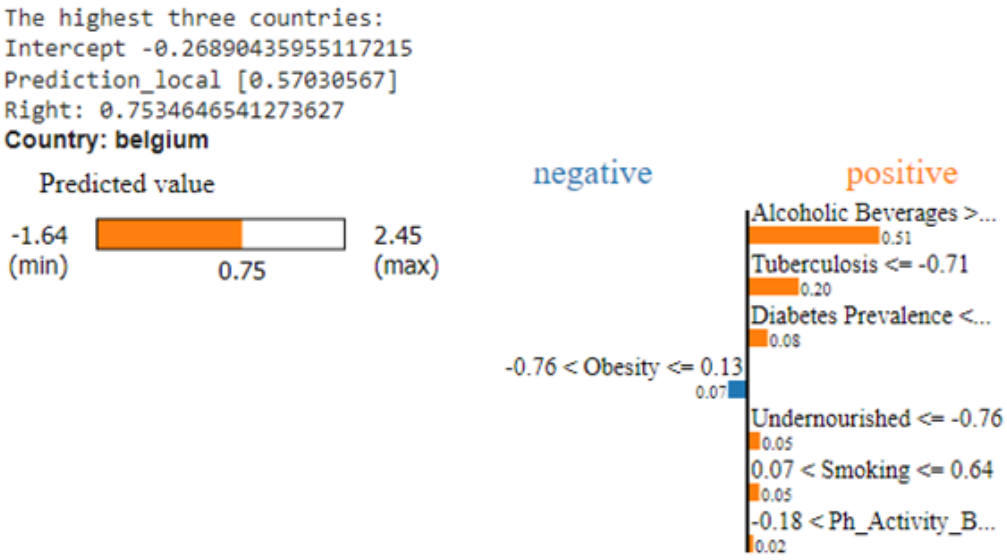

Figure 4.31: X AI regression Predication: Lime (Belgium)

Feature	Value
Alcoholic Beverages	0.91
Tuberculosis	-0.73
Diabetes Prevalence	-0.94
Obesity	0.02
Undernourished	-0.76
Smoking	0.41
Ph_Activity_Both	0.03

Figure 4.32: Feature values(Belgium)

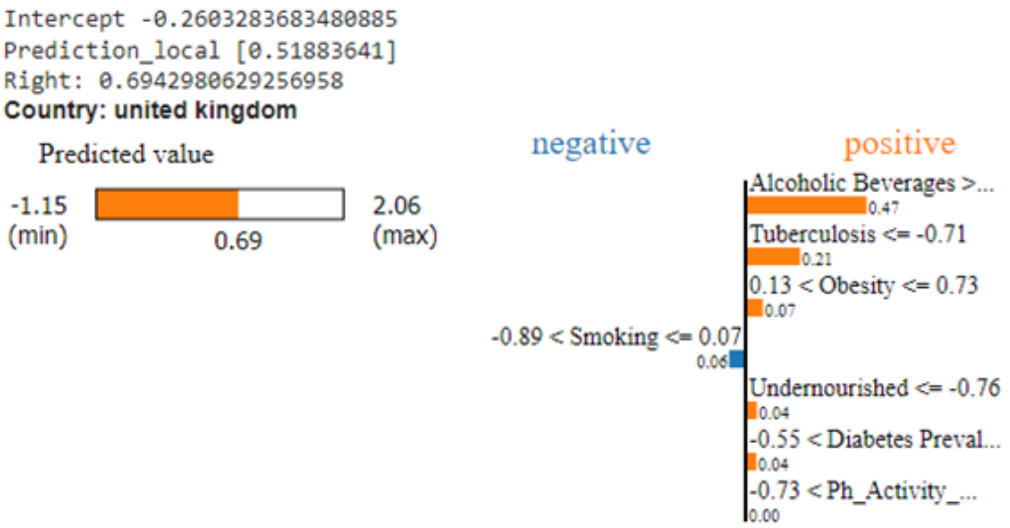

Figure 4.33: X AI regression Predication: Lime (United Kingdom)

Feature	Value
Alcoholic Beverages	0.83
Tuberculosis	-0.73
Obesity	0.67
Smoking	-0.47
Undernourished	-0.76
Diabetes Prevalence	-0.36
Ph_Activity_Both	-0.50

Figure 4.34: Feature values (United Kingdom)

Countries with the Lowest COVID-19 Mortality Rates: The mortality rate forecasted by the model for Vanuatu is -0.4630, as opposed to the observed rate of -0.2374 (Figures 4.35-4.36). Similarly, in Cambodia, the model's projected death rate of -0.7554 differs from the actual rate of -0.643 (Figures 4.37-4.38).

In terms of Vanuatu, the presence of "Alcoholic Beverages" with a feature value of -0.98 significantly reduces the forecast by 0.38. On the other hand, "Tuberculosis," with a feature value of -0.52, raises the forecast by 0.13. Given the high feature value of 1.63, it is paradoxical that "Diabetes Prevalence" actually lowers the prediction by -0.14. The features "Undernourished" and "Obesity" show varied impacts, with a value of -0.39, slightly increases the prediction by 0.09, while the "Obesity" with a value of 0.12, decreases it by 0.09. Likewise, the feature value of -0.21 for "Smoking" results in a decrease of 0.10 in the prediction. This highlights how sensitive smoking consumption is.

In Cambodia, "Tuberculosis," with a high feature value of 1.21, significantly lowers

the prediction by 0.35. "Undernourished," with a 0.34 feature value, has a -0.33 positive impact on the prediction. By lowering the prediction by -0.29, "Obesity," on the other hand, provides a significant protective effect with a feature value of -1.48. Another positive contribution comes from "Physical Activity," which has a negative feature value of -1.09 and lowers the prediction by -0.07. Other features, such as "Alcoholic Beverages" and "Smoking," have small positive feature values that slightly increase the prediction by 0.07 and 0.03, respectively, indicating lesser but still notable effects.

Examining the impact of health and lifestyle factors on mortality projections in Vanuatu and Cambodia reveals distinct patterns, in Vanuatu, the effects of smoking and obesity are less severe, but decreased alcohol consumption significantly lowers mortality risks. On the other hand, low rates of undernourished and tuberculosis modestly raise mortality estimates in Vanuatu, suggesting intricate interactions with additional variables that raise the mortality estimate only a little. In Cambodia, modest undernourishment lowers the projection, while a higher tuberculosis rate unexpectedly lowers mortality. Although the impact of alcohol use is clear, more investigation is required to completely understand the consequences of undernourished and tuberculosis.

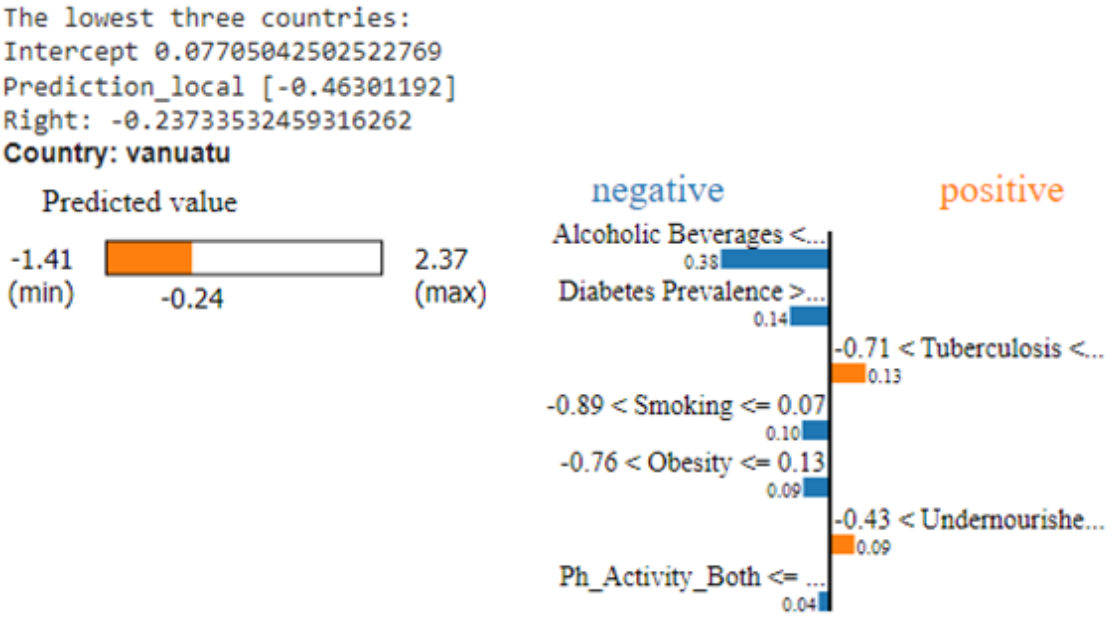

Figure 4.35: X AI regression Predication: Lime (Vanuatu)

Feature	Value
Alcoholic Beverages	-0.98
Diabetes Prevalence	1.63
Tuberculosis	-0.52
Smoking	-0.21
Obesity	0.12
Undernourished	-0.39
Ph_Activity_Both	-1.53

Figure 4.36: Feature Values(Vanuatu)

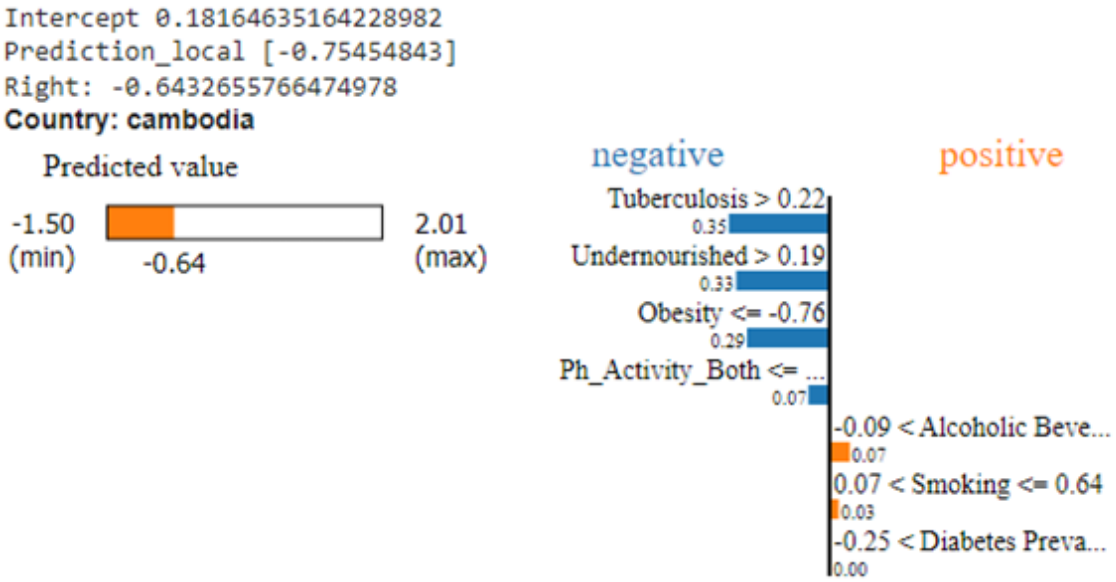

Figure 4.37: X AI regression Predication: Lime (Combodia)

Feature	Value
Tuberculosis	1.21
Undernourished	0.34
Obesity	-1.48
Ph_Activity_Both	-1.09
Alcoholic Beverages	0.09
Smoking	0.16
Diabetes Prevalence	-0.15

Figure 4.38: Feature Values(Combodia)

It could be argued that alcohol has significantly impacted mortality rates in certain

countries when consumption is high, as evidenced by increased mortality in Belgium and the United Kingdom due to "alcoholic beverages." In contrast, lower alcohol consumption in countries with lower mortality rates, such as Vanuatu, has been shown to have a protective effect by reducing mortality predictions. Although less so, the effects of smoking and obesity are also significant. Regarding undernourishment and tuberculosis, their outcomes are unexpected. While "undernourished" and "tuberculosis" are generally linked to higher health risks, in certain nations, such as Cambodia, they unexpectedly acted as a protective factor, partially mitigating the predictions. This necessitates more research to comprehend the intricate relationships between these variables and other model variables as well as how these factors affect outcomes in various contexts.

4.4 Environment

4.4.1 Exploratory Data Analysis

- **Distribution Analysis:** There are different levels of asymmetry in the frequency distributions depicted in the histograms (Figure 4.39), which may affect how well the linear regression model fits the data. Metrics like "CO2 Emissions," "Greenhouse Gas Emissions," and "Methane Emissions," for example, show signs of right-skewness, suggesting that the data may need to be adjusted to satisfy the linearity and normality assumptions. The distributions of "PM2.5_MC" and "Forest (% of land area)" are comparatively more balanced, but they still exhibit some skewness. Moreover, the observed skewness in "Death Rate" and additional variables may present difficulties, including heteroscedasticity, indicating the possible requirement for non-linear models or particular transformations for precise analysis.

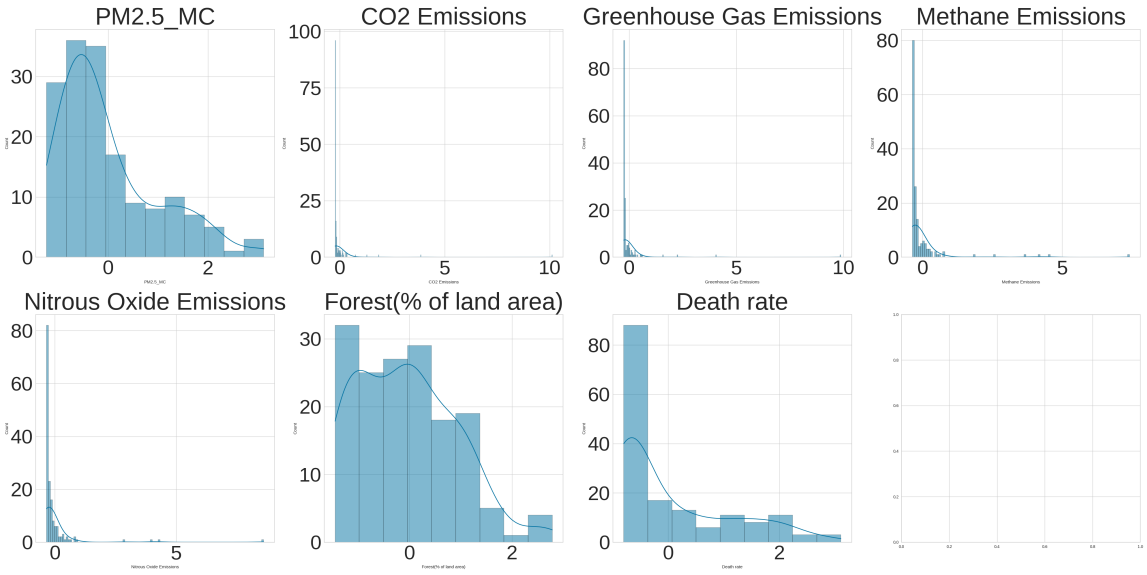

Figure 4.39: Environment Features Histogram Plot

- **Correlation Analysis** In the figure 4.40, the heatmap provides an insightful analysis of the correlation between environmental factors and death rates. PM2.5_MC has a strong negative correlation (-0.38) with the death rate, according to the statistics. We need to conduct further research on this potentially complex or surprising relationship. Conversely, CO2 emissions have a modest positive correlation (0.36) with the death rate; hence, mortality rates rise as CO2 emissions rise, presumably in response to the more general consequences of pollution on health and climate change. Furthermore, greenhouse gas emissions show a positive correlation (0.21), helping to further link higher emissions with rising death rates. On the other hand, the correlations between methane emissions and nitrogen oxide emissions with death rates, which range from roughly 0.02 to 0.04, are negligible. There is a very weak correlation (0.02) between forest coverage and death rates. These findings highlight the need to address carbon emissions and air pollution in order to improve public health outcomes.

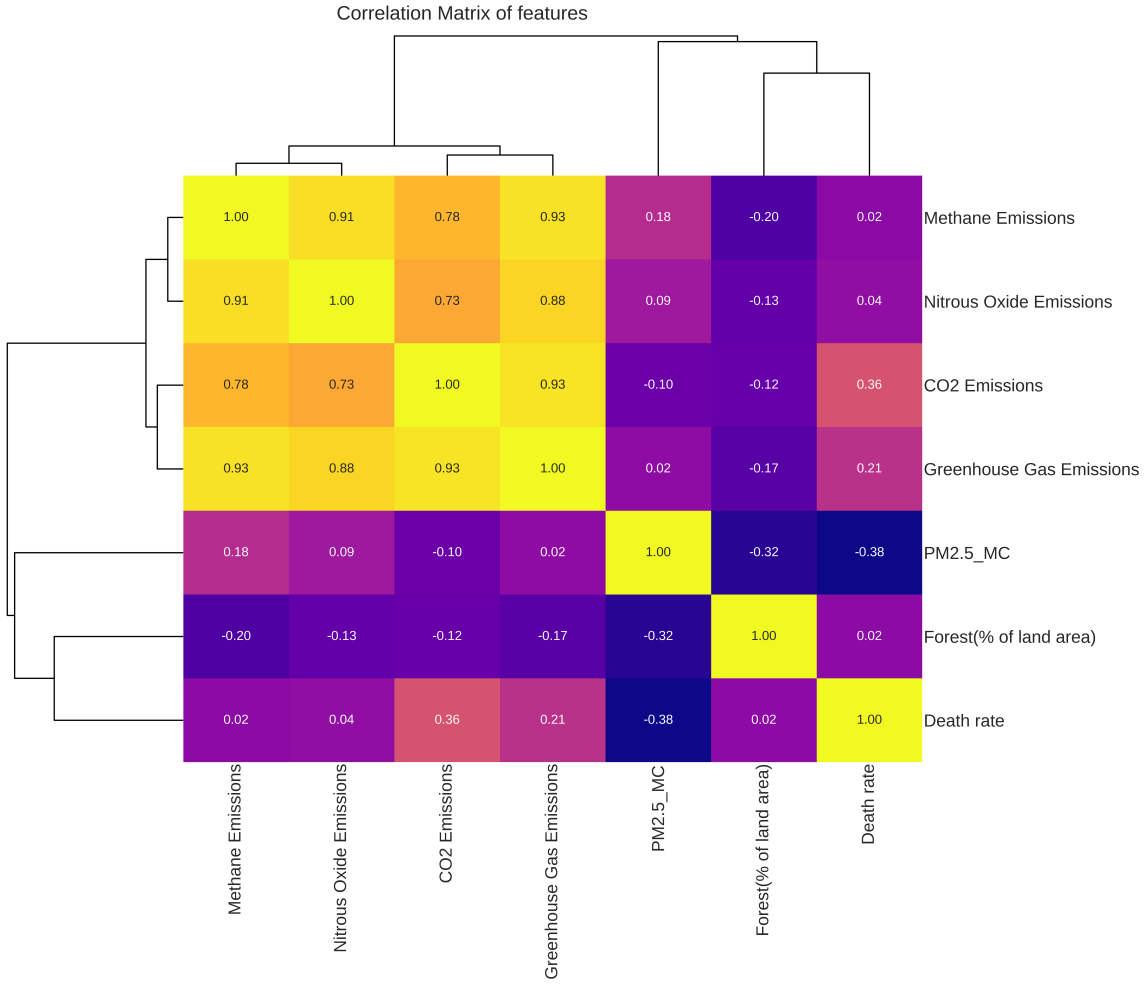

Figure 4.40: Environment Heatmap Correlation Matrix

4.4.2 Data Analysis

- **Death rate Categories:** We have classified the 160 nations into four tiers of mortality rates—'low,"medium,' 'high', 'and'very high'—utilizing box-plot percentiles . Specifically, there are 39 countries in the 'low' tier, 41 countries in the 'medium' tier, 40 countries in the 'high' tier, and 40 countries in the 'very high' tier.

- **Factory Map:** The chart 4.41 illustrates a seemingly contradictory association in which nations with higher PM-2.5 lower mortality related to green points and forest cover demonstrate higher mortality rates related to red points. We can attribute this inconsistency to variables that obscure the correlation, thereby aligning and reinforcing our earlier conclusions.

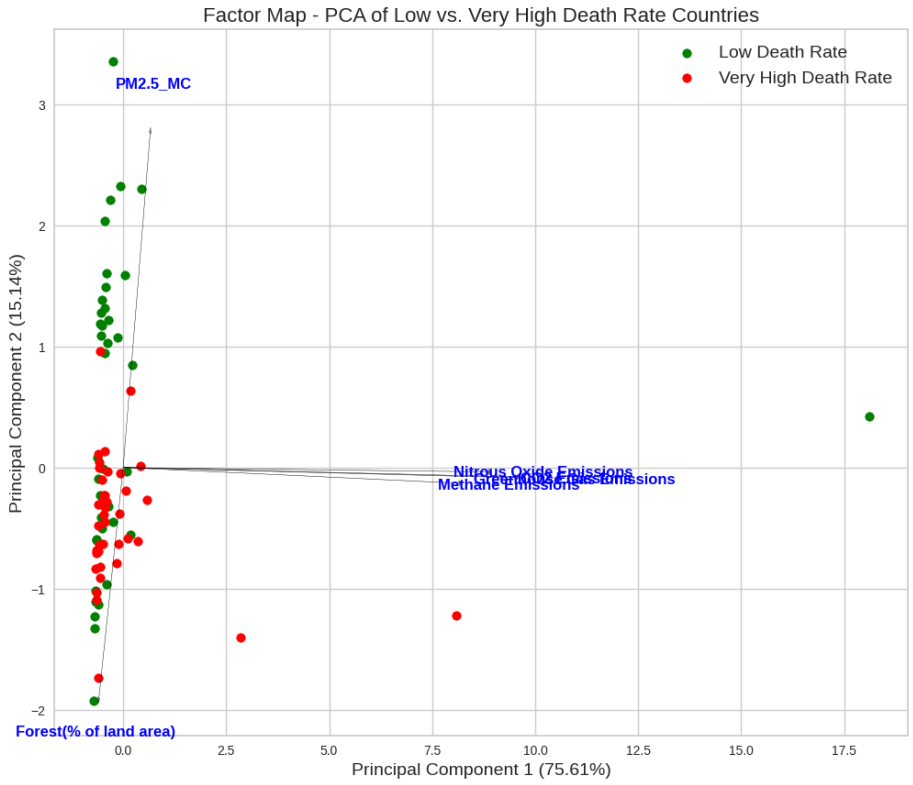

Figure 4.41: Compare Top vs Lowest Countries Features Importance

Note: Our environment feature ends here, as it does incorporate machine learning models. The selection process didn't yield any noteworthy features required to proceed with ML and XAI applications.

4.5 Study Limitations and Future Research

A significant limitation of this study is the reliance on publicly available datasets, which may not encompass all relevant variables across different countries. This reliance may lead to analytical shortcomings, especially in countries where the collection or updating of data on nutrition, lifestyles, demographics, and environmental factors is inconsistent. Furthermore, the study lacks detailed nutritional data, encompassing macronutrient (fats, proteins, carbohydrates) and micronutrient (vitamins, minerals) content, which could considerably influence COVID-19 outcomes. The omission of these factors may hinder a comprehensive understanding of the impact of diet on the severity and mortality of COVID-19. The study is constrained by its dependence on historical data from a specific timeframe, which may not reflect alterations in the pandemic's trends over time. This study exclusively projected mortality across different countries; future research should also prioritise predictions concerning infection and recovery rates.

Chapter 5

Conclusion

Examining the relationships between several elements influencing human health—including diet, demographics, lifestyle and health practices, and environment—and how they affect COVID-19 mortality rates in different countries was the aim of this study. We investigated the complex interactions leading to variations in mortality rates during the epidemic using advanced machine learning approaches, regression analysis, and explainable artificial intelligence .

Principal component analysis and graphical depictions of the correlation matrix have shown clear significant correlations between dietary patterns and COVID-19 death rates. Higher death rates (e.g., the UK, Belgium) among nations whose consumption of animal products including milk and animal fats indicated that these diets might raise COVID-19 death rate. On the other hand, nations where people ate more plant-based foods had lower death rates (e.g., Saint Kitts and Nevis, Vanuatu) suggesting possible protective impacts. The XAI models affirm these conclusions. From a machine learning standpoint, the Polynomial Regression model—which boasts a correlation coefficient of 0.7688—has shown to be the most successful. This shows that the model is useful for comprehending these interactions and guiding next studies since it is efficient in spotting trends between dietary elements and COVID-19 death.

In terms of demography, the heatmap correlation matrix and PCA clearly indicated a connection between mortality risks and older age brackets, with a slightly higher spearman rank correlation for females. Nevertheless, the XAI failed to effectively demonstrate this due to forecasting challenges within the machine learning model. This suggests that age by itself could not be sufficient to fully understand the relationship with mortality rate during pandamic. Among all the models, the Gradient Boosting Regressor model gets the best R2 score 0.1567. It still finds it difficult to fully capture the variance of the dataset notwithstanding this. Conversely, especially in modelling complex, non-linear patterns, the Polynomial Linear Regression model stands out with a correlation coefficient of 0.5258, in-

dicating its median ability to capture the relationship between input and output variables. Studies have shown that alcohol intake is a major determinant of lifestyle and health; it is also regularly connected to higher death rates. XAI analysis revealed in Belgium and the United Kingdom a link between alcohol intake and rising death rates as well as obesity and smoking to a less degree. The PCA factory map and heatmap correlation matrix support even more the important influence of alcohol consumption, obesity, and smoking. Undernourishment and tuberculosis with mortality have a negative correlation according the heatmap correlation matrix. In some cases, such Vanuatu and Cambodia, despite the usual association with health hazards, it surprisingly performed a protective role in the XAI analysis and helped to somewhat raise the prediction when the values of undernourishment and tuberculosis are small. Furthermore showing a strong link between nations with low mortality rates, undernourishment, and tuberculosis is the PCA factory map. The complexity of this problem underlines the need of more research. Regarding machine learning, the K-Neighbors Regressor shines in correlation coefficient with a value of 0.6836 while the Polynomial Linear Regression model shows the best R2 score of 0.4631. These results show top performances in their corresponding measures.

The environmental factors under investigation, particularly in relation to the COVID-19 epidemic, do not appear to have any clear or consistent influence on mortality rates. PCA factor analysis and heatmap correlation matrices also supported the findings. While the feature selection process for ML models or XAI environmental factors did not reveal any significant features to select, this suggests that their importance in predicting mortality during the epidemic is relatively low.
During the epidemic, nutrition and lifestyle/health practices overall had the biggest impact on COVID-19 death prediction. These domains demonstrated excellent model performance and strong correlations. Comparatively, environmental elements had little effect on the model, and demographics showed modest correlation efficacy in the model, so it had less predictive power. This paper emphasises the complex nature of pandemic response and the possibility of targeted dietary and lifestyle changes to lower death risks.

Bibliography

[1] Naeem Abas, Esmat Kalair, Saad Dilshad, and Nasrullah Khan. Impact of covid-19 pandemic on community lifelines. *Continuity & Resilience Review*, 4(1):94–123, 2022. Emerald Publishing Limited.

[2] Alan C Acock. What to do about missing values. 2012. American Psychological Association.

[3] Norah Alballa and Isra Al-Turaiki. Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in medicine unlocked*, 24:100564, 2021. Elsevier.

[4] Peshawa Jamal Muhammad Ali, Rezhna Hassan Faraj, Erbil Koya, Peshawa J Muhammad Ali, and Rezhna H Faraj. Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, 1(1):1–6, 2014. Mach Learn Tech.

[5] Sumayh S Aljameel, Irfan Ullah Khan, Nida Aslam, Malak Aljabri, and Eman S Alsulmi. Machine learning-based model to predict the disease severity and outcome in covid-19 patients. *Scientific programming*, 2021(1):5587188, 2021. Wiley.

[6] Abrar Almalki, Balakrishna Gokaraju, Yaa Acquaah, and Anish Turlapaty. Regression analysis for covid-19 infections and deaths based on food access and health issues. In *Healthcare*, volume 10, page 324. MDPI, 2022. MDPI.

[7] Afiat Berbudi, Nofri Rahmadika, Adi I Tjahjadi, and Rovina Ruslami. Type 2 diabetes and its impact on the immune system. *Current diabetes reviews*, 16(5):442–449, 2020. Bentham Science.

[8] Valeria Bernardo, Xavier Fageda, and Ricardo Flores-Fillol. Pollution and congestion in urban areas: The effects of low emission zones. *Economics of Transportation*, 26: 100221, 2021. Elsevier.

[9] Joseph D Brain. The respiratory tract and the environment. *Environmental health perspectives*, 20:113–126, 1977. Environmental health perspectives.

[10] Feng-Yee Chang, Hsiang-Cheng Chen, Pei-Jer Chen, Mei-Shang Ho, Shie-Liang Hsieh, Jung-Chung Lin, Fu-Tong Liu, and Huey-Kang Sytwu. Immunologic aspects of characteristics, diagnosis, and treatment of coronavirus disease 2019 (covid-19). *Journal of biomedical science*, 27:1–13, 2020. Springer.

[11] Tiziana Ciarambino, Ombretta Para, and Mauro Giordano. Immune system and covid-19 by sex differences and age. *Women's Health*, 17:17455065211022262, 2021. Women's Health.

[12] Andaç B Çolak. Prediction of infection and death ratio of covid-19 virus in turkey by using artificial neural network (ann). *Coronaviruses*, 2(1):106–112, 2021. Bentham Science.

[13] Fátima Pérez De Heredia, Sonia Gómez-Martínez, and Ascensión Marcos. Obesity, inflammation and the immune system. *Proceedings of the Nutrition Society*, 71(2): 332–338, 2012. Cambridge.

[14] Doreswamy, Mohammad Kazim Hooshmand, and Ibrahim Gad. Feature selection approach using ensemble learning for network anomaly detection. *CAAI Transactions on Intelligence Technology*, 5(4):283–293, 2020. Wiley.

[15] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023. ACM New York.

[16] Saeed Ghahramani. *Fundamentals of probability*. CRC Press, 2024. CRC Press.

[17] Waleej Haider, Muhammad Nadeem, Sallar Khan, Haris Ahmed, Asad Abbasi, and Zainab Anwar. Computing in humanity: To predict the human behaviors over social media. *Webology*, 19(3), 2022. Webology.

[18] Tammy Jiang, Jaimie L Gradus, and Anthony J Rosellini. Supervised machine learning: a brief primer. *Behavior therapy*, 51(5):675–687, 2020. Elsevier.

[19] Christoph SN Klose and David Artis. Innate lymphoid cells control signaling circuits to regulate tissue-specific immunity. *Cell research*, 30(6):475–491, 2020. Springer.

[20] J Lee, V Taneja, and Robert Vassallo. Cigarette smoking and inflammation: cellular and molecular mechanisms. *Journal of dental research*, 91(2):142–149, 2012. SAGE.

[21] S Lockyer. Effects of diets, foods and nutrients on immunity: Implications for covid-19? *Nutrition Bulletin*, 45(4):456–473, 2020. Wiley.

[22] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020. International Journal of Science.

[23] H Mehta, K Nazzal, and RT Sadikot. Cigarette smoking and innate immunity. *Inflammation Research*, 57:497–503, 2008. Springer.

[24] David Mhlanga. The role of artificial intelligence and machine learning amid the covid-19 pandemic: What lessons are we learning on 4ir and the sustainable development goals. *International Journal of Environmental Research and Public Health*, 19(3):1879, 2022. MDPI.

[25] Denis A Mogilenko, Irina Shchukina, and Maxim N Artyomov. Immune ageing at single-cell resolution. *Nature Reviews Immunology*, 22(8):484–498, 2022. Nature.

[26] LJ Muhammad, Ebrahem A Algehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty, and Ibrahim Alh Mohammed. Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. *SN computer science*, 2(1):1–13, 2021. Springer.

[27] Luiza Camelia Nechita, Mariana Daniela Ignat, Alexia Anastasia Stefania Balta, Raisa Eloise Barbu, Liliana Baroiu, Doina Carina Voinescu, Aurel Nechita, Mihaela Debita, Camelia Busila, and Ioana Anca Stefanopol. The impact of cardiovascular antecedents on the prognosis of covid-19 critically ill patients. *Journal of Clinical Medicine*, 13(12):3518, 2024. MDPI.

[28] Jacqueline Parkin and Bryony Cohen. An overview of the immune system. *The Lancet*, 357(9270):1777–1789, 2001. Elsevier.

[29] Dinah V Parums. infectious disease surveillance using artificial intelligence (ai) and its role in epidemic and pandemic preparedness. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, 29:e941209–1, 2023. Medical Science Monitor.

[30] Bart G Pijls, Shahab Jolani, Anique Atherley, Raissa T Derckx, Janna IR Dijkstra, Gregor HL Franssen, Stevie Hendriks, Anke Richters, Annemarie Venemans-Jellema, Saurabh Zalpuri, et al. Demographic risk factors for covid-19 infection, severity, icu admission and death: a meta-analysis of 59 studies. *BMJ open*, 11(1):e044640, 2021. British Medical Journal.

[31] Bharat Richhariya, Muhammad Tanveer, Ashraf Haroon Rashid, Alzheimer's Disease Neuroimaging Initiative, et al. Diagnosis of alzheimer's disease using universum support vector machine based recursive feature elimination (usvm-rfe). *Biomedical Signal Processing and Control*, 59:101903, 2020. Elsevier.

[32] K Saha, R Mehta, RC Misra, DS Chaudhury, and SN Ray. Undernutrition and immunity: smallpox vaccination in chronically starved, undernourished subjects and its immunologic evaluation. *Scandinavian Journal of Immunology*, 6(6-7):581–589, 1977. Wiley .

[33] Shafiun Nahin Shimul, Fariha Kadir, and Muhammad Ihsan-Ul-Kabir. Factors associated with coronavirus (covid-19) deaths and infections: A cross country evidence. *medRxiv*, pages 2020–11, 2020. medRxiv.

[34] Carlotta Suardi, Emanuela Cazzaniga, Stephanie Graci, Dario Dongo, and Paola Palestini. Link between viral infections, immune system, inflammation and diet. *International Journal of Environmental Research and Public Health*, 18(5):2455, 2021. International Journal of Environmental Research and Public Health.

[35] Shaun Turney. Central limit theorem formula, definition & examples. *Scribbr*, 2022. Scribbr.

[36] Patrizio Vanella, Ugofilippo Basellini, and Berit Lange. Assessing excess mortality in times of pandemics based on principal component analysis of weekly mortality data—the case of covid-19. *Genus*, 77:1–36, 2021. Springer.

[37] Anuradha Yenkikar and C Narendra Babu. Spread analysis and prediction of covid-19 in india using machine learning. In *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, pages 421–426. IEEE, 2023. IEEE.

[38] Rui Zhong, Lingxia Chen, Qiong Zhang, Binbin Li, Yanfang Qiu, Wei Wang, Dongyi Tan, and Yanhui Zou. Which factors, smoking, drinking alcohol, betel quid chewing, or underlying diseases, are more likely to influence the severity of covid-19? *Frontiers in Physiology*, 11:623498, 2021. Frontiers in Physiology.

[39] Hongfang Zhou, Xiqian Wang, and Rourou Zhu. Feature selection based on mutual information with correlation coefficient. *Applied intelligence*, 52(5):5457–5474, 2022. Springer.

Chapter 6

Appendices

6.1 Additional Results

6.1.1 Food Groups Explainable AI Countries

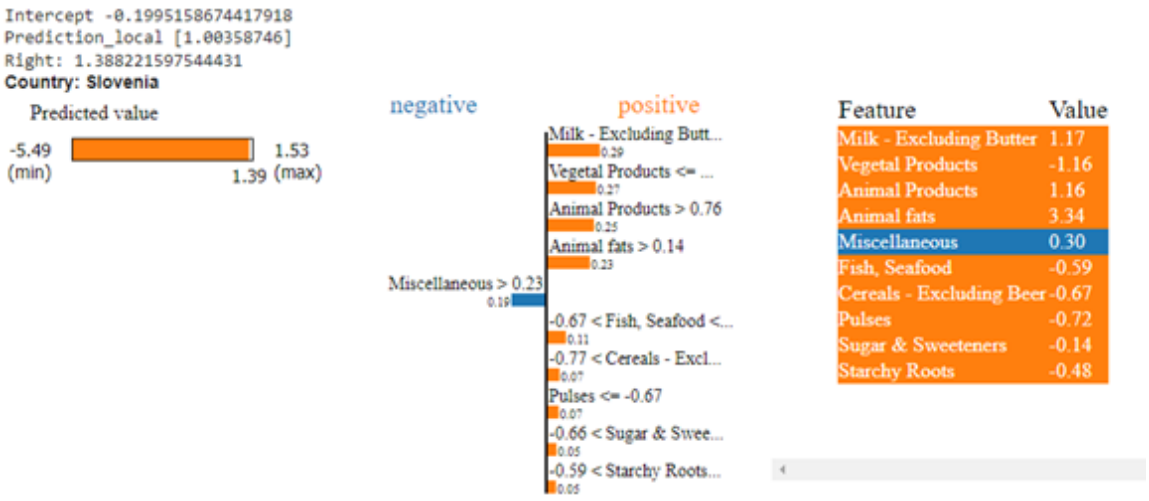

Figure 6.1: XAI (Slovenia)

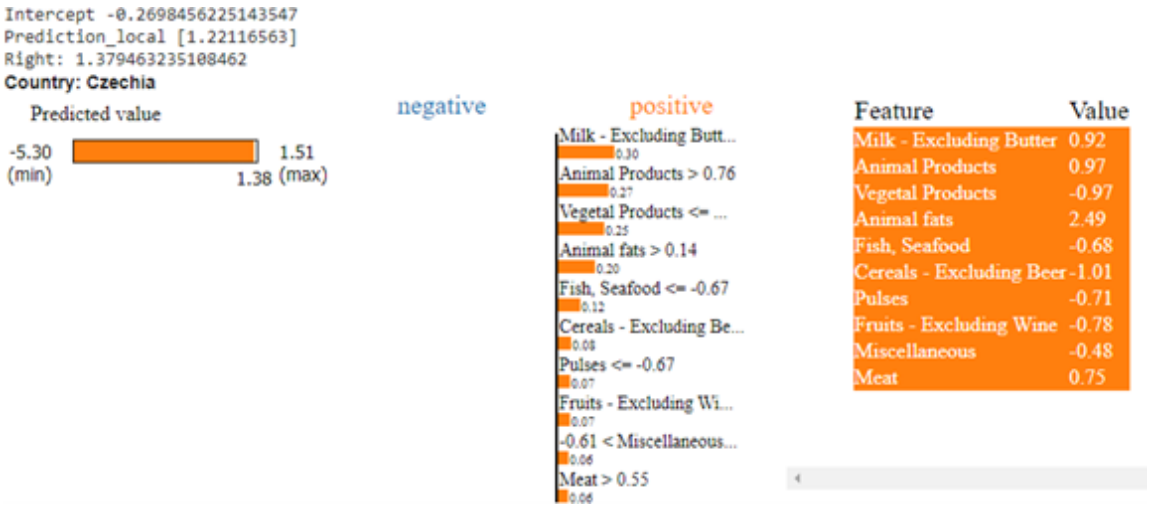

Figure 6.2: XAI (Czechia)

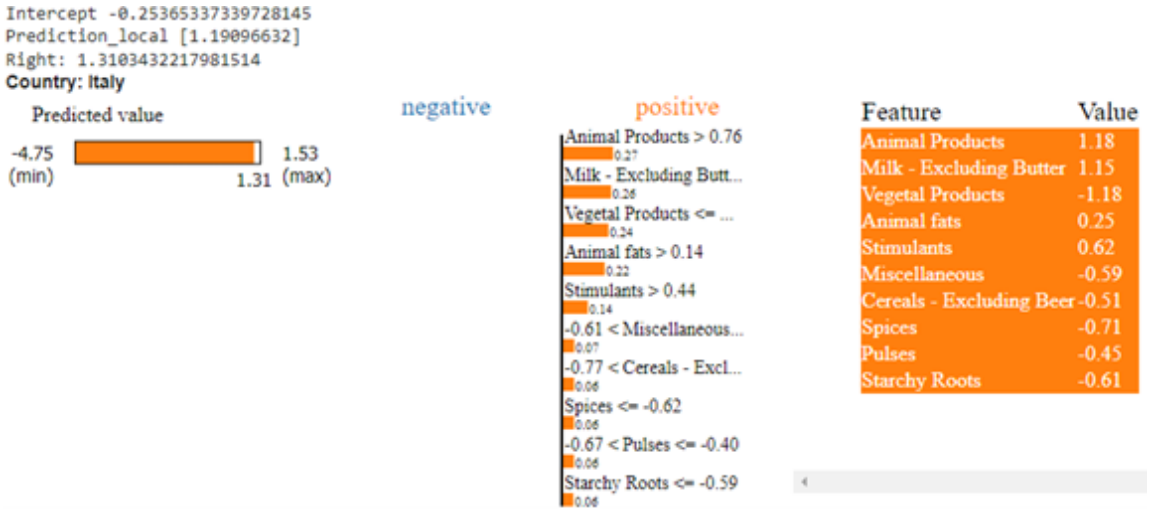

Figure 6.3: XAI (Italy)

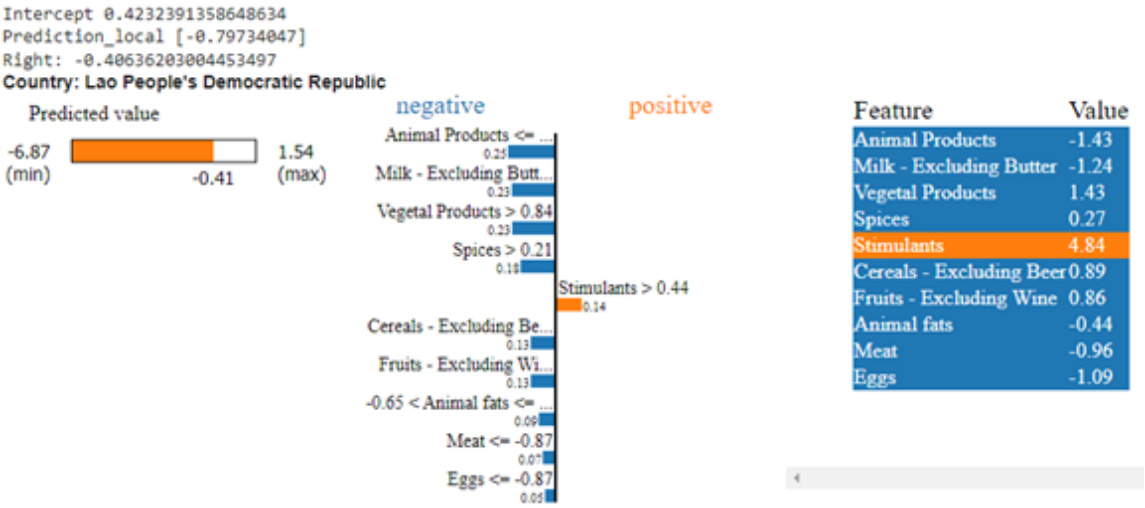

Figure 6.4: XAI (Lao People's Democratic Republic)

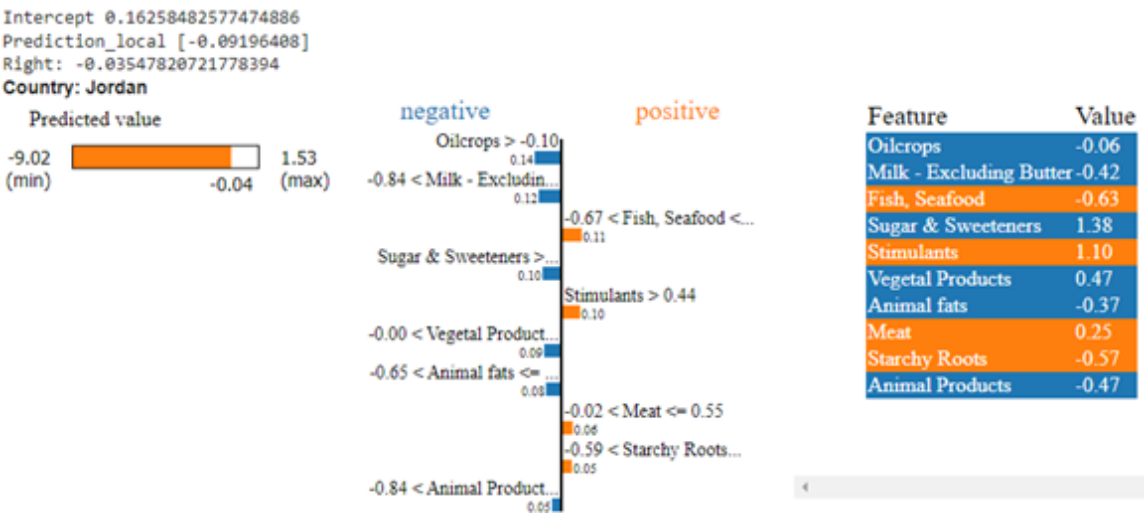

Figure 6.5: XAI (Jordan)

6.1.2 Lifestyle And Health Groups Explainable AI Countries

Figure 6.6: XAI (slovenia)

Figure 6.7: XAI (Italy)

Figure 6.8: XAI (United States)