



Swansea University
Prifysgol Abertawe

Predicting COVID-19 Mortality Using Machine Learning and Regression Analysis Based on Nutrition Factors

Abdullah Sharaf

Supervisor: Dr. Alma Rahat

14/10/2024

Problem Statement

The COVID-19 pandemic has severely impacted over 200 countries, COVID-19 has evolved into a major healthcare industry concern as well as a public health emergency. The World Health Organization 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.

Project Methodology

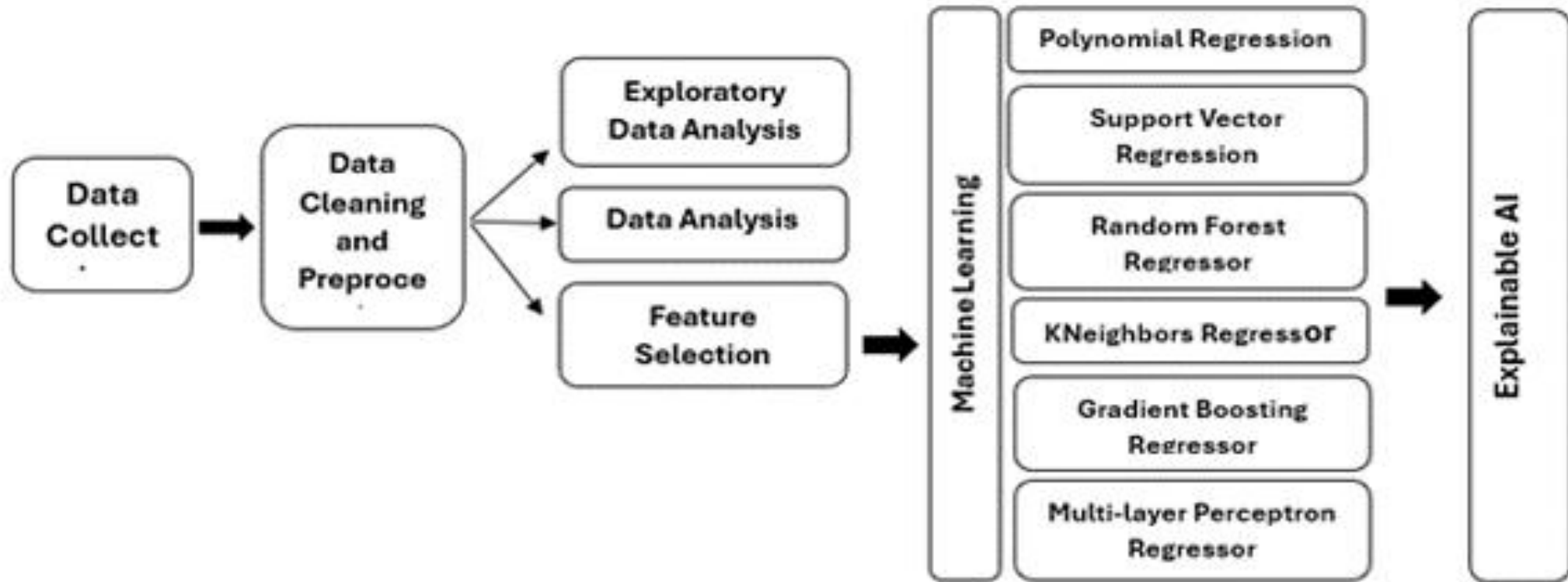


Figure 3.1: Methodology

Summary of Cross-sectional Study Data Analysis Part

Introduction

Methodology(Regression Analysis)

Key finding

recommendation

INTRODUCTION

Project Objective The primary objective of this data analysis project is to identify the **most significant factors**—such as **nutrition**—that influence **COVID-19 Case mortality** rates across different countries.

Public Health Goals: Use data-driven insights to refine **public health decision-making** strategies, improving the effectiveness of **change behaviors at schools**.

Hypotheses:

H_0 : Nutrition does not influence COVID-19 case mortality rates $\beta_1=0$.

H_1 : Nutrition influences COVID-19 case mortality rates $\beta_1 \neq 0$

DATA COLLECTION

Dataset	Open Source	Features
Nutrition.csv	Kaggle – COVID-19 Healthy Diet Dataset	Country, 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, Vegetable Oils (2020) , Case Fatality Rate (2021)

DATA Concept

- Load Data
- Check Data Type
- Wrangling Data(Missing , Duplicate, Outliers)
- Scaled Data(Standardization)
- Exploratory Data Analysis (EDA)
- Data Analysis

Data Exploration

Data Types

Data columns (total 21 columns):

#	Column	Non-Null Count	Dtype
0	Country	164 non-null	object
1	Animal fats	164 non-null	float64
2	Animal Products	164 non-null	float64
3	Cereals - Excluding Beer	164 non-null	float64
4	Eggs	164 non-null	float64
5	Fish, Seafood	164 non-null	float64
6	Fruits - Excluding Wine	164 non-null	float64
7	Meat	164 non-null	float64
8	Milk - Excluding Butter	164 non-null	float64
9	Miscellaneous	164 non-null	float64
10	Offals	164 non-null	float64
11	Oilcrops	164 non-null	float64
12	Pulses	164 non-null	float64
13	Spices	164 non-null	float64
14	Starchy Roots	164 non-null	float64
15	Stimulants	164 non-null	float64
16	Sugar & Sweeteners	164 non-null	float64
17	Treenuts	164 non-null	float64
18	Vegetable Oils	164 non-null	float64
19	Vegetal Products	164 non-null	float64
20	Death rate	164 non-null	float64

dtypes: float64(20), object(1)

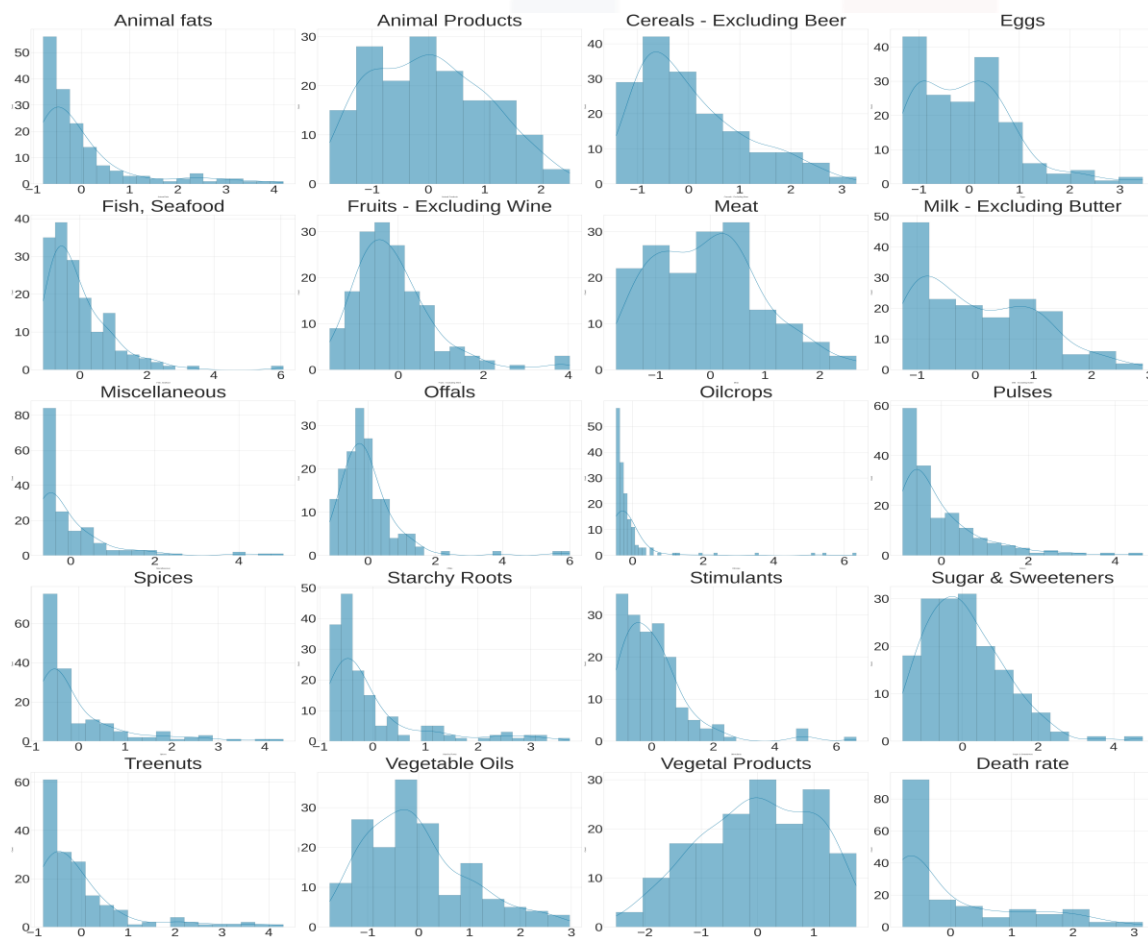
memory usage: 28.2+ KB

Statistical Summery

	Animal fats	Animal Products	Cereals - Excluding Beer	Eggs	Spices	Starchy Roots	Stimulants	Sugar & Sweeteners	Treenuts	Vegetable Oils	Vegetal Products	Death rate
count	164.0	164.0	164.0	164.0	164.0	164.0	164.0	164.0	164.0	164.0	164.0	164.0
mean	0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0
std	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
min	-1.0	-2.0	-1.0	-1.0	-1.0	-1.0	-1.0	-2.0	-1.0	-2.0	-3.0	-1.0
25%	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0
50%	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-1.0
75%	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0
max	4.0	3.0	3.0	3.0	4.0	4.0	7.0	5.0	4.0	3.0	2.0	3.0

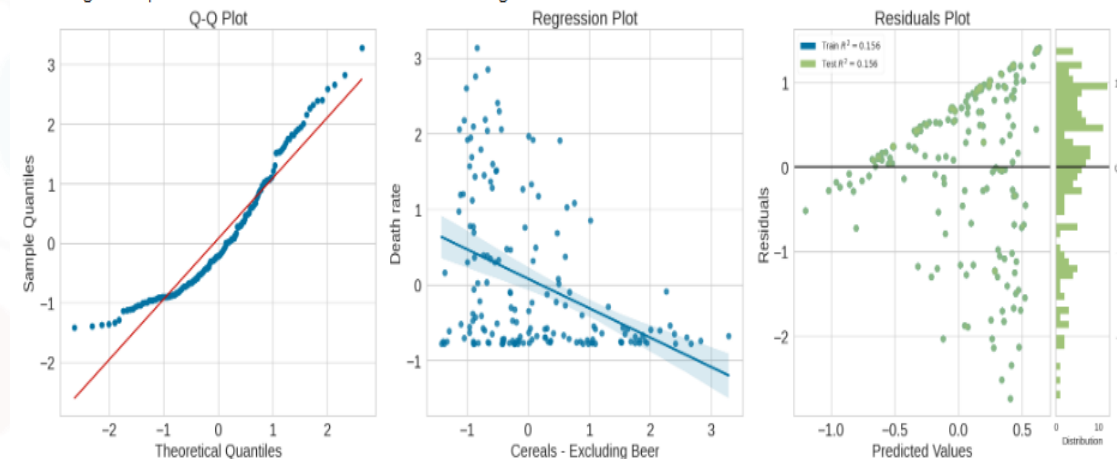
Non-Linear Problem

Nutrition Distribution

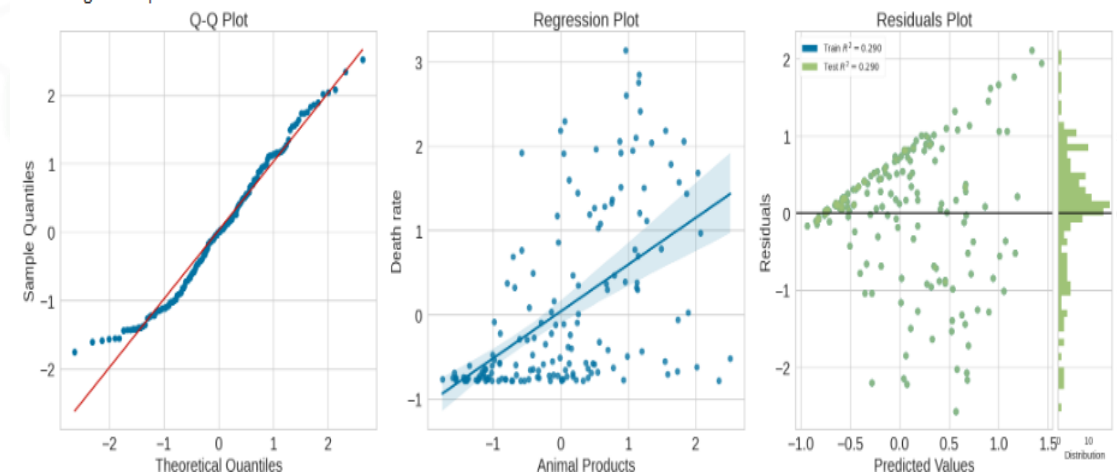


Normality Test

Checking assumptions for feature: Cereals - Excluding Beer

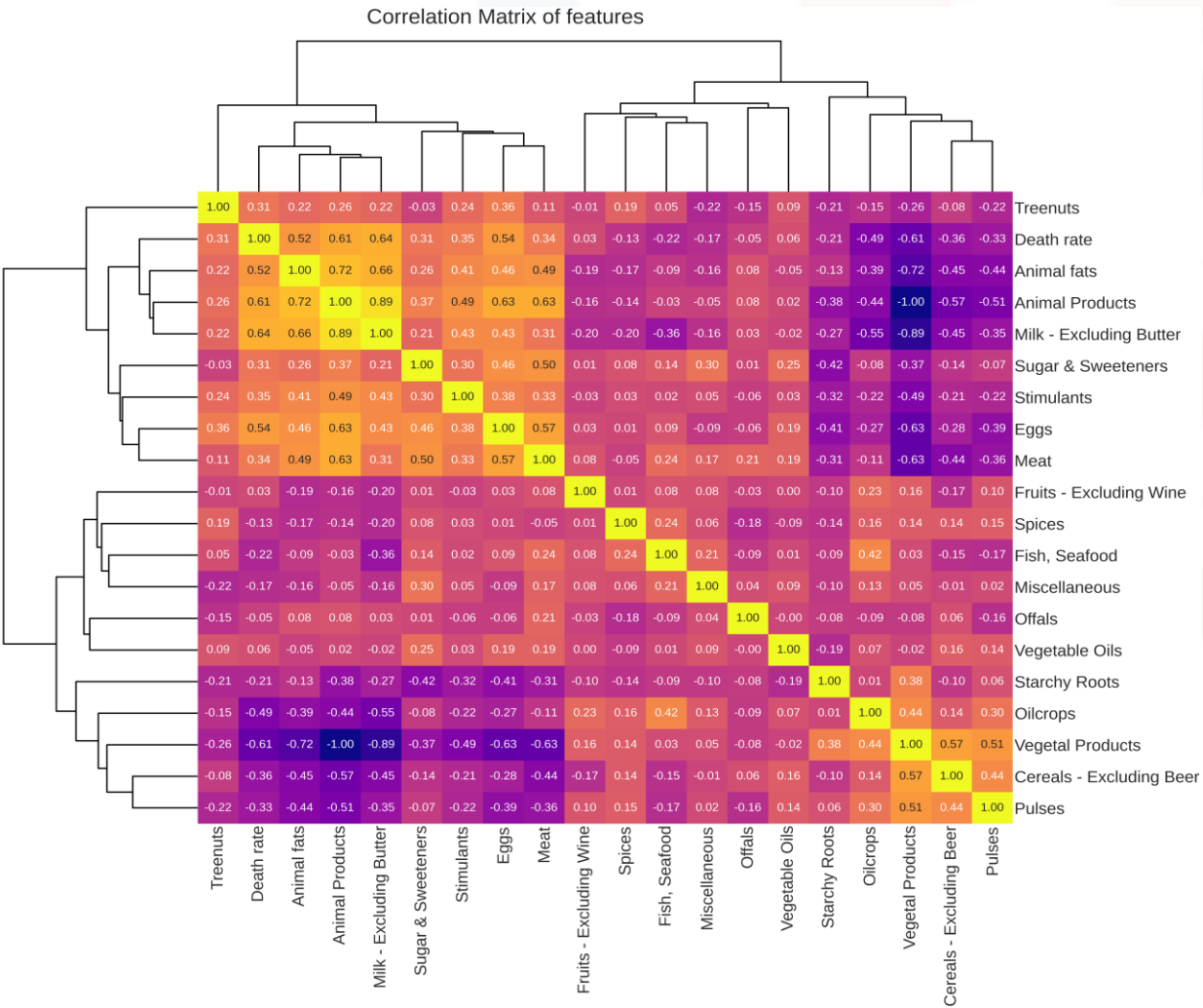


Checking assumptions for feature: Animal Products

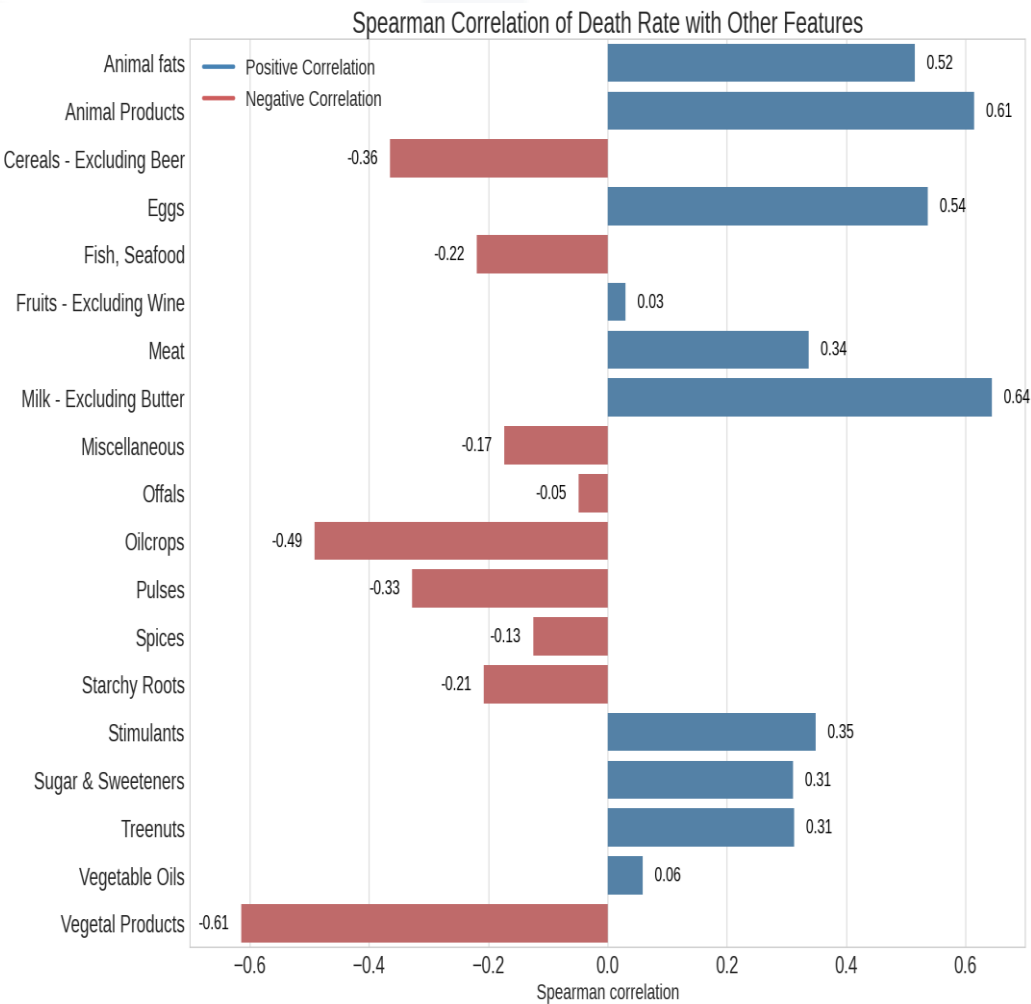


Data Analysis

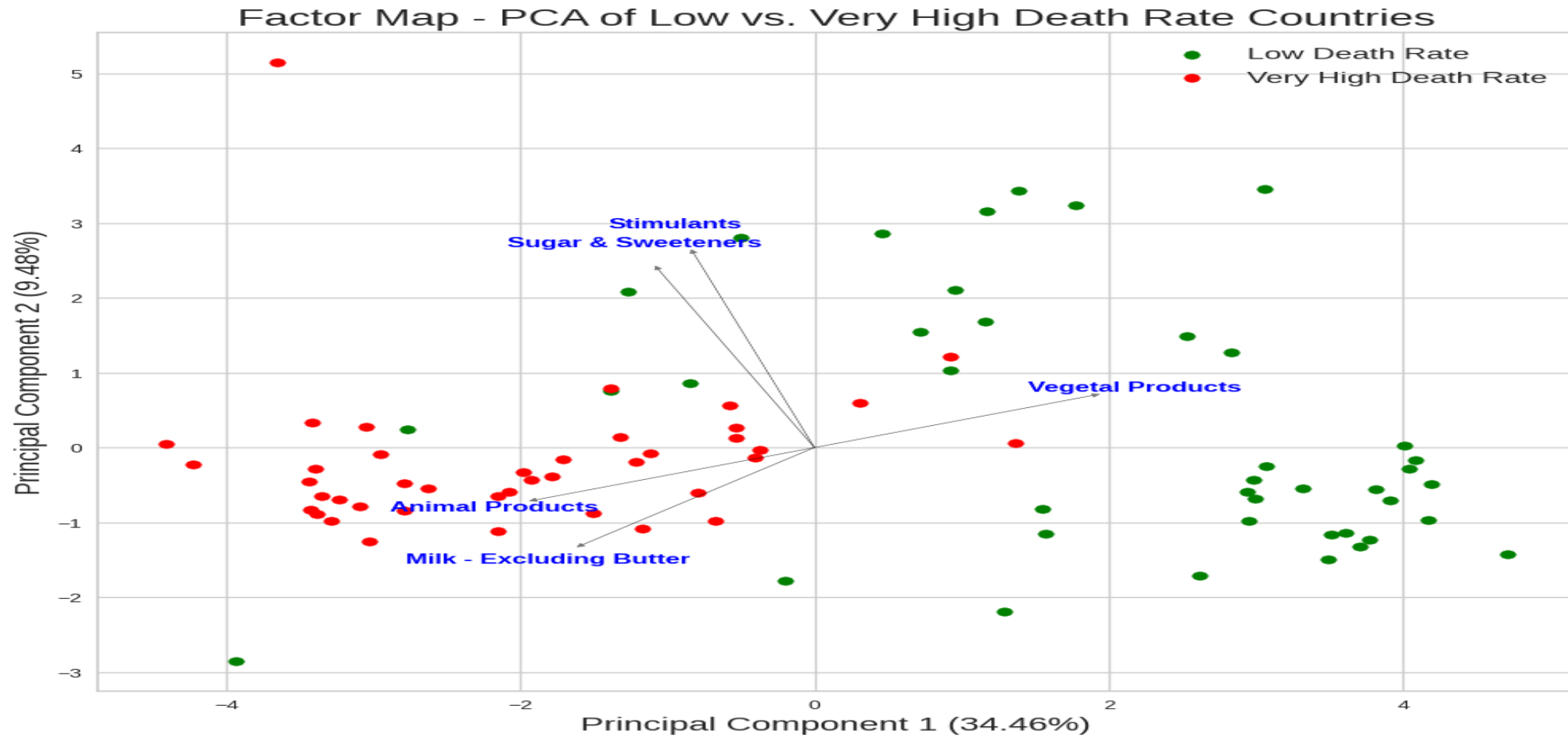
Spearman Correlation matrix



Case Death rate VS Nutrition



Factor Map - PCA



Non-Parametric Test (Spearman, Kendall's Tau)

Feature	Spearman Correlation	Spearman p-value	Kendall's Tau	Kendall p-value
Animal Products	0.614041	2.264509e-18	0.433793	1.769582e-16
Milk - Excl. Butter	0.643527	1.500912e-20	0.452816	8.114807e-18
Vegetal Products	-0.614095	2.244723e-18	-0.433793	1.769582e-16

Confidence Interval: 99% ($\alpha = 0.01$)

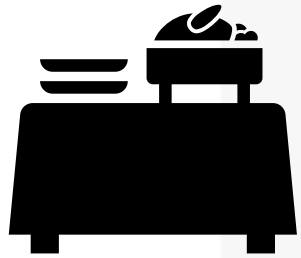
Based on the analysis, we identified that the most influential features affecting the COVID-19 case death rate were statistically significant. As a result, we **rejected the null hypothesis** and **accepted the alternative hypothesis**.

Alternative Hypothesis (N1):

Nutritional factors have a significant influence on COVID-19 case mortality rates.

$$\beta_1 \neq 0$$

Key finding













Higher consumption of **Animal products and Milk** linked to increased mortality, while a balanced diet with sufficient **plant-based** foods is crucial for reducing health risks.

Animal Products: Aquatic Animals, Others; Aquatic Plants; Bovine Meat; Butter, Ghee; Cephalopods; Cream; Crustaceans;

Vegetal Products: Apples and products; Bananas; Barley and products; Beans.

Recommendation

Primary Schools (Behavioral Change)

Intervention	Behavior or Aspect Targeted	Method of Influence / Notes
Difficult-to-open animal product packaging. easy-to-open packaging 	Natural tendency to save effort 	Kids tend to choose the easier-to-open option; leveraging 'productive laziness' in favor of healthy choices.
Appealing names like ..Hero Salad" + cartoon images 	Identity and imagination (I'm a hero = I eat veggies) 	Connecting vegetarian food to the child's identity as a lovable hero through appealing names.
Reward cards (stickers, stars) 	Positive reinforcement after desired behavior 	Kids see an immediate outcome of their good choice, motivating them to repeat the behavior.
Device that makes a playful sound only for choosing a vegetarian dish 	Instant gratification and auditory response 	Building a positive group habit and linking it to team spirit and group pride.
Replace high-fat milk with low-fat milk	Default bias and availability 	

Encourage behavior change starting from an early age to support long-term healthy habits.



Data Science Part

Summary of Cross-sectional Study

Data Science Part

Introduction

Methodology

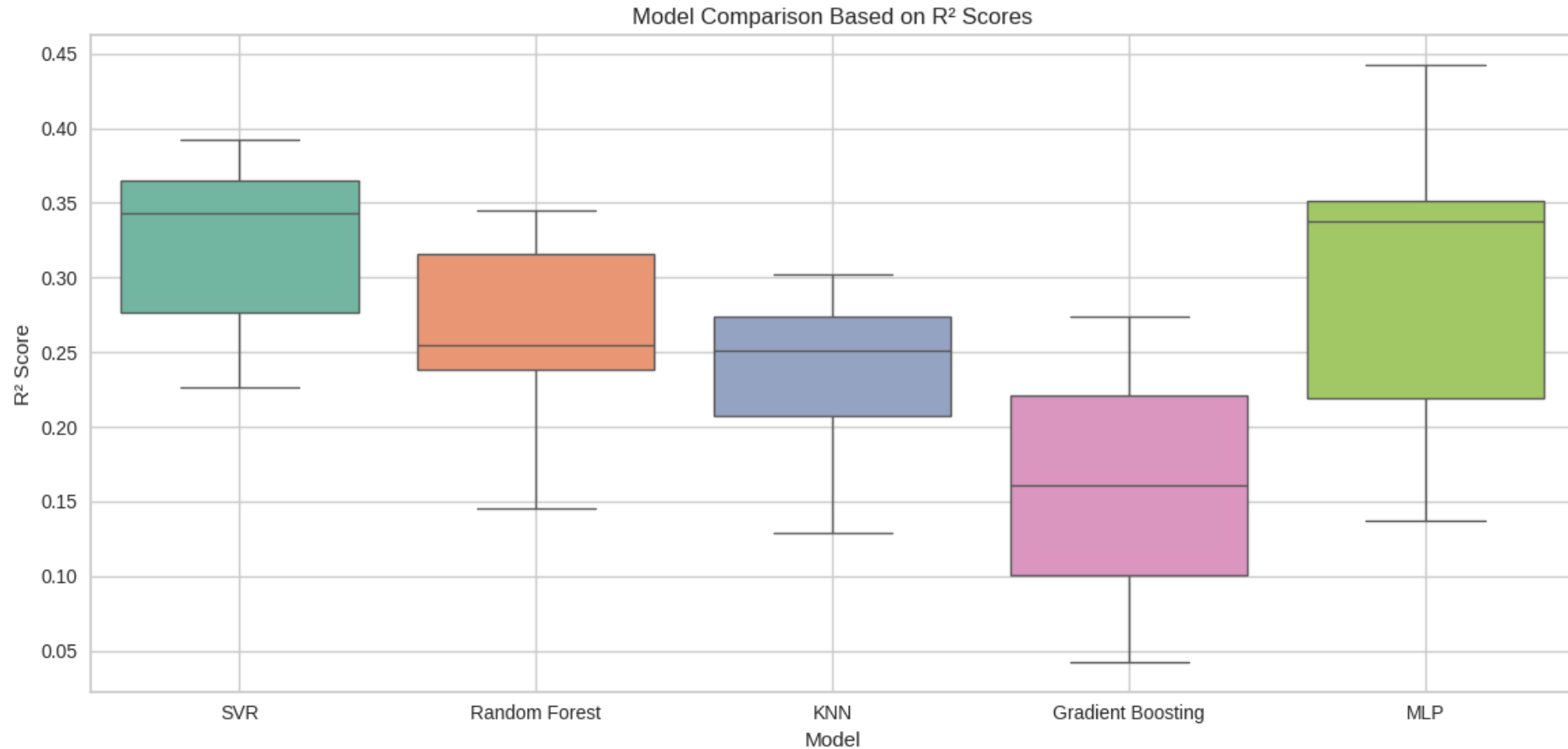
Key finding

Methodology

- Load Data
- Check Data Type
- Wrangling Data(Missing , Duplicate, Outliers)
- Train-Test Split(test 20%, training 80%)
- Scaled Data(Standardization)
- Feature Selection ()
- Machine Learning (Supervised Regression ML)
- Explainable AI (Lime)

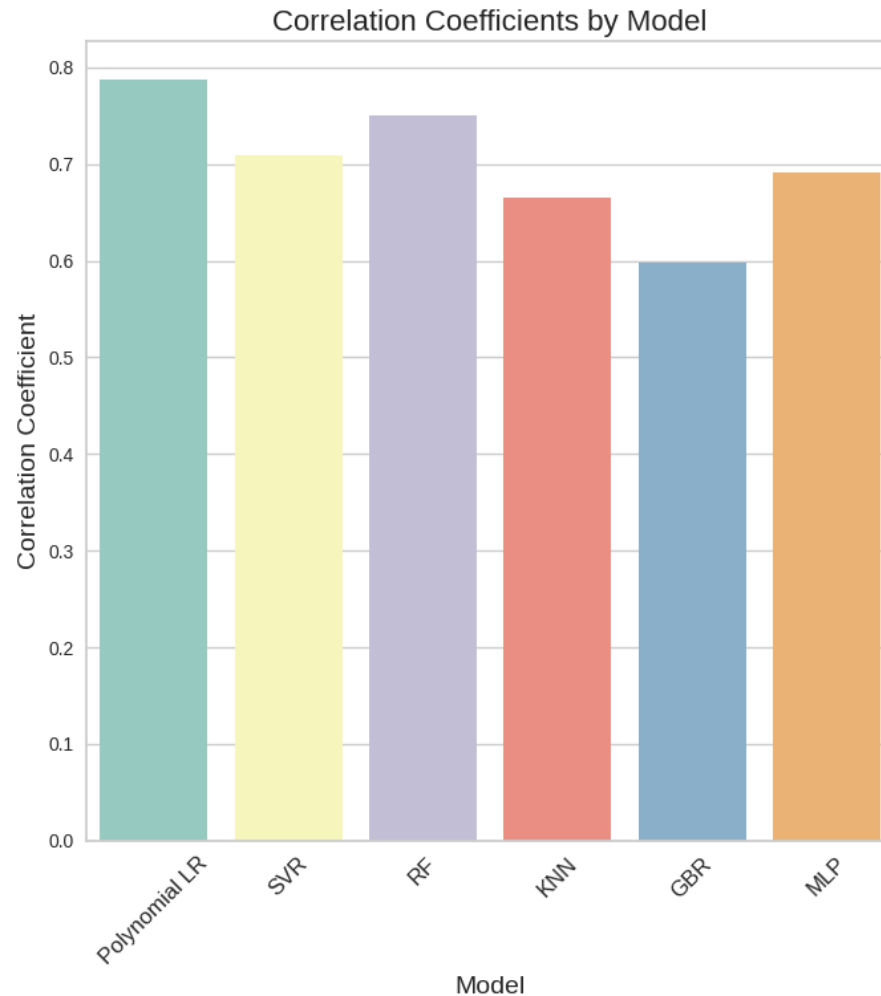
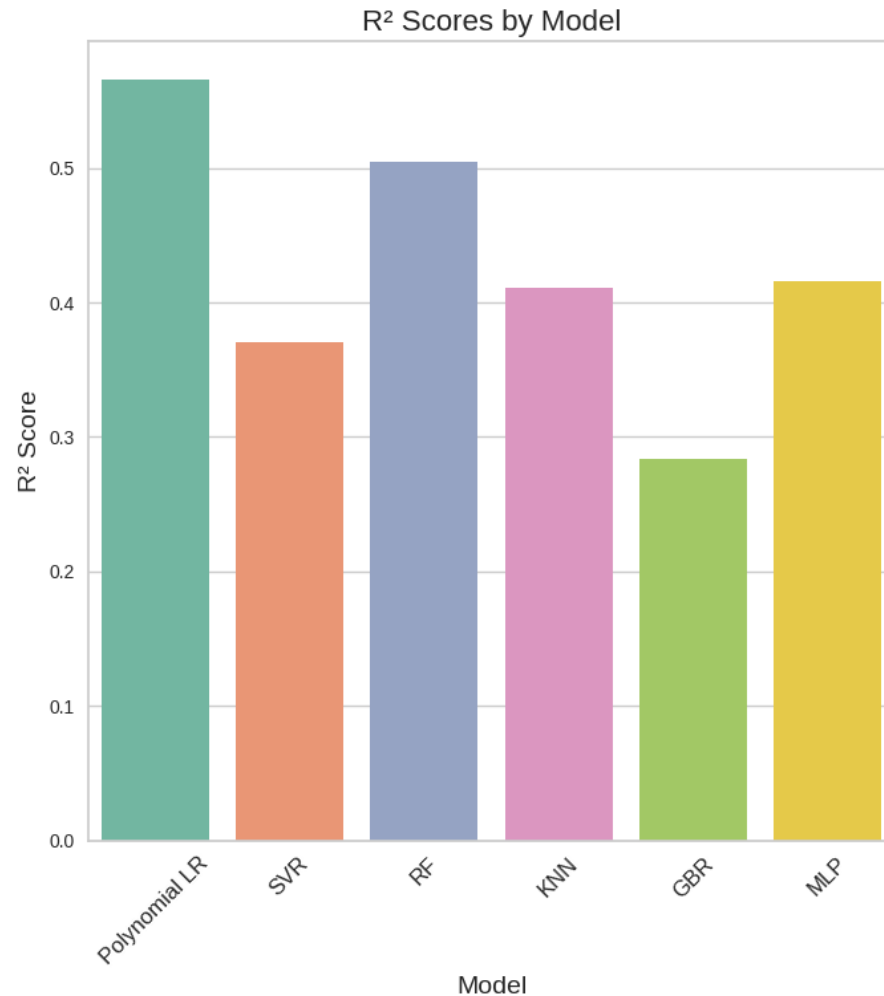
Machine Learning with default Hyperparameters

Result of training Data

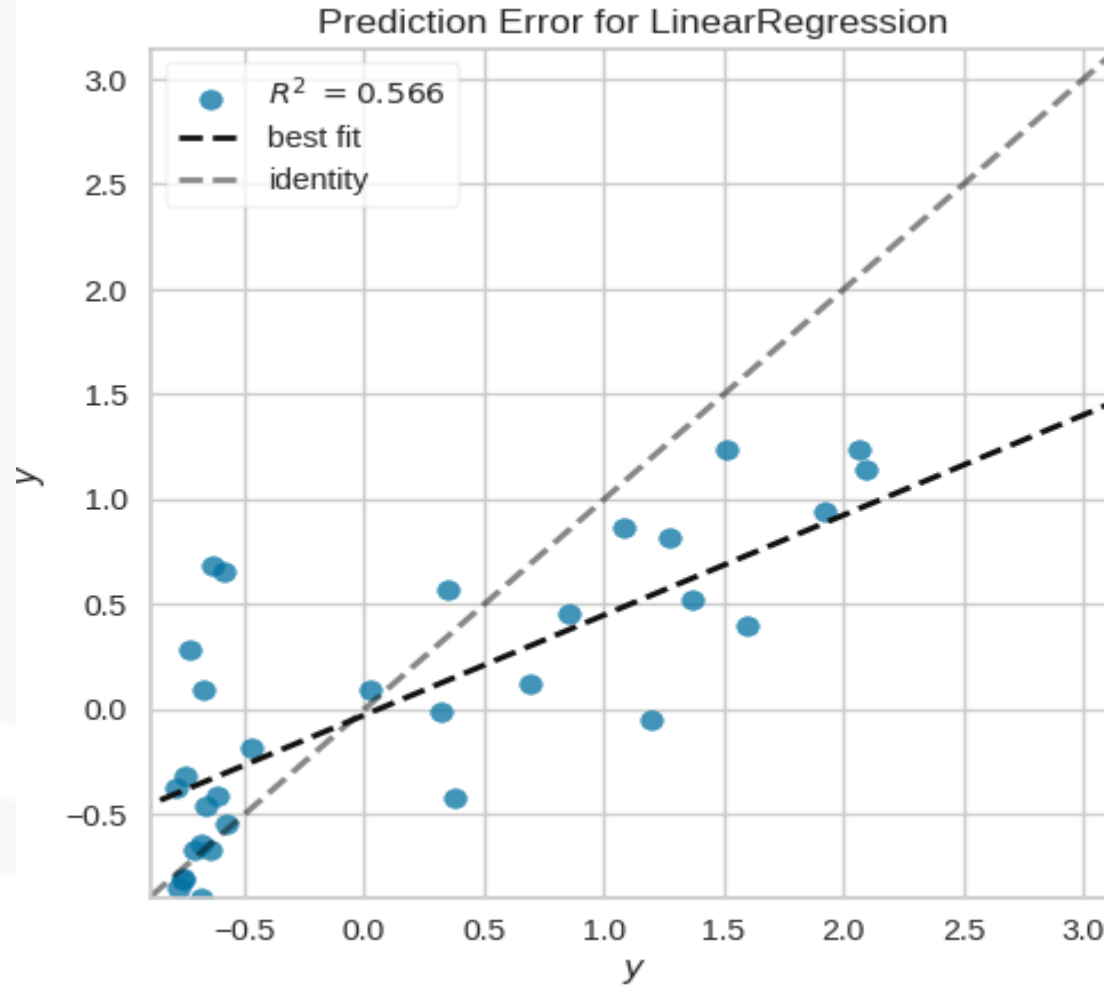


Machine Learning with hyperparameter tuning

Result of Testing Data



The Best Model Polynomial Linear regression Result Of Testing Data



Explainable AI

(High Death Effected Country)

Prediction_local [1.22172904]

Right: 1.3330868040274693

Country: Belgium

Predicted value

-8.58 (min)  1.54 (max)
1.33 (max)

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

negative

positive

Milk - Excluding Butt...	0.28
Animal Products > 0.76	0.28
Animal fats > 0.14	0.25
Vegetal Products <= ...	0.24
Cereals - Excluding Be...	0.13
-0.29 < Fish, Seafood ...	0.08
Miscellaneous <= -0.61	0.07
Pulses <= -0.67	0.06
-0.62 < Spices <= -0.42	0.06
-0.76 < Fruits - Exclud...	0.05

(Low Death Effected Country)

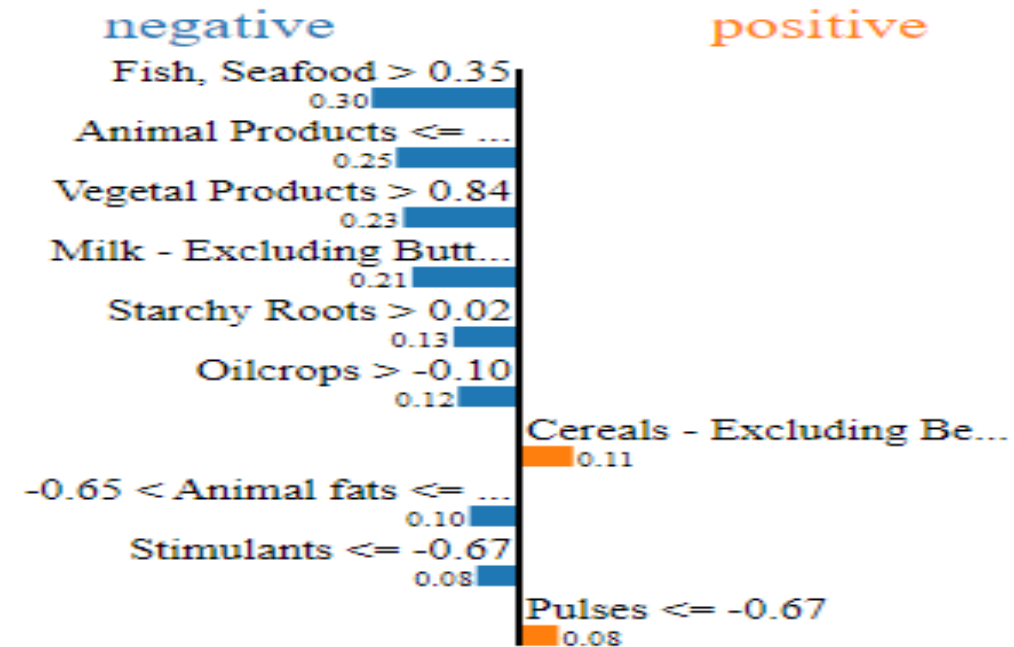
Prediction_local [-0.81452202]

Right: -0.6705301018960877

Country: 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





The End
