Factors Influencing Human Life Expectancy

Analysis, Design, and Implementation Report

Submitted in partial requirements for the degree of MSc Applied Data Science

School of Computing, Engineering and Digital Technologies

Department of Computing and Games

Teesside University

Middlesbrough TS1 3BA

Date: 31st Aug 2021

SUBMITTED BY

Mohana Kamanooru

A0223038@live.tees.ac.uk
School of Computing, Engineering &
Digital Technologies
TEESSIDE UNIVERSITY

SUPERVISOR

Dr Zia Ush Shamszaman

Z.Shamszaman@tees.ac.uk
Senior Lecturer in Computer Science
Department of Computing and Games
TEESSIDE UNIVERSITY

Acknowledgement

This research would not have been feasible without the help and guidance of everyone around me and in my thoughts. Firstly, I want to express my gratitude to my supervisor Dr Zia Ush Shamszaman, my module leader Dr Alessandro Di Stefano, Dr Yar Muhammad, and the rest of my Teesside University acquaintances. Also, I want to thank all the staff at Teesside University for their incredible and timely support, and finally, I would like to take this opportunity to thank my family for their unwavering support and belief in me and my potential more than anybody else.

Abstract

Life Expectancy (LE) is the most significant statistic for assessing population health. It is defined as the average number of years a human being may presume to live after birth. It is a significant analytical indicator for measuring the socio - economic growth of countries or regions. Growing living standards, improved lifestyles, education, and more availability to high-quality health care are all factors that might contribute to a rise in LE at birth. Furthermore, greater LE leads to a rise in population and, as a result, better human capital investment returns for those who live longer. As a result, more human capital contributes to higher GDP per capita.

Therefore, it is essential to explore and create a unique application in which an effective and appropriate machine learning model is constructed and developed to assist in determining whether certain factors impact the longevity and, if so, how we may increase life expectancy.

From this research, we could conclude that there is an essential link between protein consumption and LE, according to the data. Protein is vital for energy consumption, physiological activities, and immune functions. As a result, nations such as Japan, Italy, Switzerland, Spain, Singapore, Australia, Iceland, and the Netherlands have longer life expectancies. No comprehensive research of vaccines, illnesses, or LE could be done because of a lack of data. Injuries, long- and short-term health issues, non-communicable diseases, and mental health all directly influence death rates, making life expectancy inversely proportional, according to the available statistics. Several machine learning approaches and the analyses and observations gathered throughout the research were used to create machine learning models. As a result, we created an 85 per cent accurate machine learning model using the K Nearest Neighbor method.

Table of Contents

Table of Figures

Table of Tables

Acronyms

LE	Life Expectancy
GDP	Gross Domestic Product
BMI	Body Mass Index
CD	Communicable Diseases
NCD	Non-Communicable Diseases
HIV	Human Immunodeficiency Virus
BCG	Bacille Calmette Guerin
WHO	World Health Organisation
GHO	Global Health Observatory
LR	Logistic Regression
KNN	K Nearest Neighbor
SVM	Support Vector Machine
MAE	Mean Average Error
MSE	Mean Square Error
RMSE	Root Mean Square Error
RBF	Radial Basis Kernel Function
PCA	Principal Component Analysis
EDA	Exploratory Data Analysis
ML	Machine Learning

1. Introduction

The most important statistic for measuring population health is **Life Expectancy** (LE), the average number of years a person could survive. It is an essential synthetic indicator for measuring a country's or region's economic and social progress. (Beeksma *et al.*, 2019). Life expectancy is extensively examined as part of the composition of demographic statistics for nations throughout the world, and it is used to measure mortality experiences and compare them over time and across geographic locations. (Vlatka Bilas, 2014). Achieving higher LE is crucial to have a healthier lifestyle, and it is often challenging to identify what factors influence a person's healthy lifestyle. Human rights, political and civic, and economic, social, and cultural rights must be promoted to achieve health and development. Progress and health are linked in two ways. Health is a vital component of development, which is the result of enhancing health and wellbeing. (Beeksma *et al.*, 2019)

LE increased substantially since the 17th and 18th centuries. In the early nineteenth century, life expectancy began to grow in early industrialised countries, but it remained low in the world. As a result, there was a massive gap in how health was distributed across the globe. In the developed regions, health is usually excellent, whereas, in the developing countries, health is consistently low. (Cutler, Deaton and Lleras-Muney, 2006). Considerable reduction in global inequality is observed in recent times. Territories with the most incredible life expectancy in the 18th century have the lowest life expectancy globally. Many countries that were formerly afflicted by ill health are making rapid progress. (Preston, 2003). The worldwide average LE has more than doubled since 1900, reaching more than 70 years. LE disparities continue to exist across and within countries. The Central African Republic has the lowest life expectancy in 2019, at 53 years, while Japan has 30 years. (Riley, 2005)

Why is LE important, and why should we improve? Increases in LE at birth can be related to distinct reasons, including growing living standards, improved lifestyles, education, and increased access to high-quality health care. In addition, higher LE results in population increase and higher human capital investment returns for individuals who live longer. As a result, more human capital help to raise Gross domestic product (GDP) per capita. (Miladinov, 2020).

Often it is challenging to identify the factors influencing the longevity of a person. Several factors that influence could be outlined with socioeconomic status. It includes factors like employment, education, household income, and financial health; the quality of the mental and physical health system and people's awareness and accessibility of those services; health-related lifestyles like tobacco consumption, higher intake of alcohol, undernourishment, and inactive, sluggish lifestyles; and social, genetic, and environmental factors like overcrowding, contaminated water consumption, and unhealthy sanitation. (*Department of Health | Tier 1—Life expectancy and wellbeing—1.19 Life expectancy at birth*, no date). In England, from 2013 to 2015, neonatal males might expect to survive until they were 80 years old, while neonatal girls could expect to live until they were 83. On the other hand, these young people are expected to have at least 20% of their life in bad health. On average, boys born between 2013 and 2015 would have a healthy 63 years and girls with an additional year, 64years. (*What affects an area's healthy life expectancy? - Office for National Statistics*, no date)

2. Research Question

We research and develop a novel application in which an effective and suitable machine learning model is designed and developed, so the model can aid to identify if any factors affect longevity and, if so, how can we improve life expectancy.

2.1. Research Focus

In this study, firstly, we focus on dietary consumption and its relation to the LE. Do people who consume a vegetarian diet have higher longevity compared to people who consume meat? Secondly, we analyse and identify which countries have higher LE and what factors might influence achieving higher LE. Is there any similarity amongst the group? Finally, it is evident that illness, diseases are indirectly proportional to human longevity. In addition, we try to analyse the available data and see if there is any new correlation that can be identified concerning immunisation, distinct types of communicable diseases, non-communicable illnesses, and a person's longevity.

3. Literature Review

Mortality study in social science has a long history, beginning with a focus on the problems of urbanisation and the misery of urban populations. Aaron's research on longevity and mortality began with life expectancy at birth and progressed to total mortality. (Antonovsky, no date) This approach began when infectious diseases were widespread, and it was accompanied by the implementation of public health measures to minimise mortality. The leading causes of death have moved over time to chronic illnesses, which develop over a lifetime rather than a few days, and have different origins than contagious infections. Current population assessments of death rates and variations give insight into the Population's changing well-being and its diverse subgroups and alter causes of death and strategies to reduce excess mortality. (Crimmins and Zhang, 2019). LE is a more intuitive measure of death than mortality rates and is a valuable and essential summary measure of mortality. (Klenk *et al.*, 2007) Life expectancy is a more intuitive measure of death than mortality rates and is a valuable and vital summary measure of mortality. (Klenk *et al.*, 2007) (Miladinov, 2020).

Several studies and models have been developed and implemented to observe and understand the influence of LE on socioeconomic and social factors. The research conducted by Boucekkine in 2003 indicated that the rise in education and literacy caused an increase in the LE, and the improvement in human capital has pushed the Population's growth rate to higher limits by the end of the Industrial revolution. Adult mortality at the dawn of the 17th and the rise of the 18th centuries has significantly increased by 70%. (Boucekkine, De La Croix and Licandro, 2003) (Miladinov, 2020). Mortality is also affected by the economic conditions, food supply and other living standards shelter, living space. Evidence has been collected and analysed by Shamuel H. Preston in 2003, which indicates that national LE is strongly related to the national per head income. (Preston, 2003)

Global life expectancy is increasing every year in every country around the world. According to estimates, LE at birth increased in a curved pattern from around 28.5 years in 1800 to 66.6 years in 2001. Before the 1920s, about 30 nations began making persistent improvements in survival, and worldwide life expectancy grew steadily until 1913, but the difference between the most excellent and lowest regional life expectancies widened dramatically, culminating about 1950. During much of the twentieth century, gains were quick and generally shared, until around 1990, when the consequences of HIV/AIDS, in particular, widened the disparity between nations and areas with the lowest and greatest life expectancy. (Riley, 2005). An increase in LE increases the Country's Population until the onset of demographic transition, reducing per capita income but increasing per capita income after the transition. In recent decades most countries have made progress towards a demographic transition.

Vaccination offers benefits that extend beyond preventing specific ailments in humans. They enable society and countries to harvest a diversified and abundant harvest. It is a moral obligation for the international community to reduce child mortality globally by supporting widespread access to efficient, effective, safe vaccines since it is a human right for everyone to live a better and brighter life. (Andre, 2008) Vaccination is cost-effective and satisfies the requirement to care for society's most vulnerable people. If applied now, the average causal effect of increases in life expectancy on per capita income is expected to be beneficial to health advances and mortality reductions. When calculating the economic benefits of increased life expectancy, the fundamental effect of increased LE causing the shift should be included as it is one of the crucial influencing factors. (Cervellati, 2009) (Berry, 2021). Air pollution (fine particulate) exposure is also a cause of adult mortality. Exposure to these pollutants causes a 0.3years span loss in overall adult mortality in Taiwan. (Chen, Chen and Yang, 2019)

Multiple socioeconomic variables, health, healthcare system-related factors, illness load, and complicated interconnections influence life expectancy. (Girum, Muktar and Shegaze, 2018). Also, schooling, population changes at various stages of the demographic evolution influence the LE, as mentioned by Cervellati, Matteo in "The effect of life expectancy on education and population dynamics" Empirical Economics journal published in 2015. Extrinsic changes to their surroundings like Available diet, vaccination, hygiene, sanitation, and disease prevalence, were primarily responsible for the increase in LE advancements. (Strulik and Vollmer, 2013) (Herzer, 2017).

Correspondingly, as we progress in achieving LE, it also calls for a contribution towards social equality. Even in less-developed countries like Brazil or India, mortality is now more evenly distributed than income in a developed welfare state. The historical transition of lifetime inequality from a significant source of social inequality to a minor source of inequality is well established. This change reflects the unique method in which success in lowering mortality has occurred. It has saved many individuals from death throughout childhood and maturity, but not in old age. (Peltzman, 2009)

In the aftermath of the 17th and 18th Centuries, mortality started to decline in England, both by applying innovative ideas about personal health and public administration to health and incidentally through increased productivity, which allowed for higher living standards and suitable nutrition housing sanitation. Changes in public health infrastructure and personal behaviour were dependent on ideas about the germ hypothesis of illness. Similarly, towards the middle of the twentieth century, information about the health implications of smoking had a significant impact on behaviour and health. Recently, important life-saving technological advances in medical procedures and evolving medicines have significantly lowered heart-related mortality. There have also been significant health advances that have primarily benefited poorer nations. Health is determined by institutional competence and political desire to apply known technology in wealthy and developing nations, neither of which is an inherent result of growing incomes. Rather than a causal link from higher income to better health, the lower wages of unwell individuals explain much of the association between income and health within nations. (Cutler, Deaton and Lleras-Muney, 2006)

Veganism is a rigorous type of vegetarianism that has grown in popularity in recent years. Plant-based diets have been linked to increased longevity and better health. According to research on a vegan diet, plant-based diets are linked to more significant health but not necessarily reduced death rates (Norman and Klaus, 2020). The specific processes by which vegan diets promote health are unknown, although they are most likely complex. In lifespan research, the reasons for and quality of a vegan diet should be evaluated.

Similarly, Low meat consumption doesn't necessarily result in longer LE in humans. (Singh, Sabaté and Fraser, 2003), Furthermore, it is believed that women live longer than men, but it has only been proven to be a myth by the study conducted in three Nordic Populations over men and women aged 75 and above. (Heikkinen *et al.*, 2016) (Dicker *et al.*, 2018). The evidence above shows an increase in life expectancy, studies are being carried out to determine the maximum end of life for human beings while recovering, and the resilience is withdrawn. Ageing is a multi-determinant trait, but a substantial genetic component influences survival to extreme ages. The deregulation of immune responses that occurs as people get older is thought to play a role in human morbidity and mortality.

On the other hand, some genetic factors of effective ageing may be found in polymorphisms for immune system genes that regulate immunological responses. We examined the significant impacts of single loci and multi-locus interactions to test the hypothesis that the adenosine deaminase (ADA) and tumour necrosis factor-alpha (TNF-a) genes may affect human life expectancy. (Napolioni *et al.*, 2011). There is robust evidence that human longevity is heritable, and significant effort is being put into discovering genes linked to longer lives. (Beekman *et al.*, 2013) (Rootzén and Zholud, 2017).

Demographic research has indicated a steady decrease in old-age mortality and an increase in the maximum age of death, suggesting that human lifespan may be extended through time. It also shows that increases in survival with age tend to diminish after age 100 and that the world's oldest person's age at death has not grown since the 1990s, using worldwide demographic data. These findings imply that humans' maximum lifespan is set and subject to natural limitations. With lifespan observations in many animal species, these findings are flexible and may be enhanced by genetic or pharmacological intervention, which has led to speculation that species-specific genetic restrictions may not constrain longevity.

(Dong, Milholland and Vijg, 2016)(Herzer, 2017). As mentioned earlier and the journal published by Robert J. Pignolo in "Exceptional Human Longevity", the evidence suggests that extraordinary longevity is complex, including various combinations of genes, environment, resilience, and luck, all of which are impacted culture and geography. (Pignolo, 2019). end-of-life criticality is an intrinsic biological property of an organism independent of stress factors and represents a fundamental or absolute limit to human lifespan. (Pyrkov *et al.*, 2021)

4. Data Description

From all the above literature review, we consider the key features that have been identified to answer the current research question, if any.

4.1.Identifying components affecting Life Expectancy

The geographical location of the person, i.e., the **Country** information, **Population** which gives the human capital information, **GDP,** which is a measure of economic size and economic health, **Economic status** of the Country as rich and developing countries have their significance in defining the mortality inequality, **Literacy Rate** which provides information on better-educated citizens with better schooling. This information is categorised under the Country. Similarly, the diet information if the person consumes **plant-based** food, **seafood** or **non-vegetarian** food provides the information about protein intake and way of life of the person. We further analyse this information to see if there is any impact or relation to their LE. Also, **alcohol** consumption, **smoking** and **physical exercise** provide in-depth information about the person lifestyle. All these attributes are person-related, so we categorise them into **People Group**.

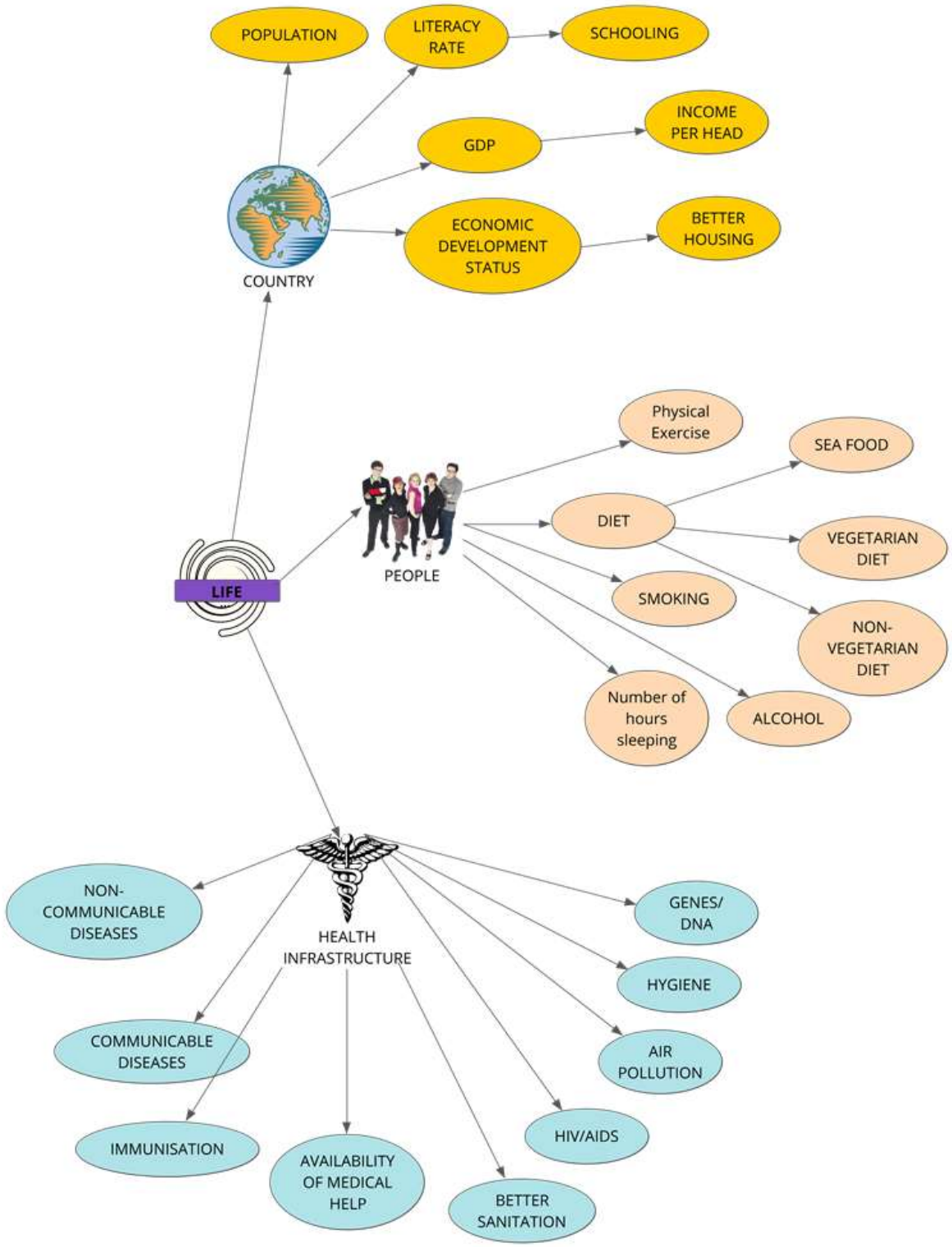


Figure 1 Factors affecting Life expectancy

Lastly, we define the medical and health-related features into Health Group. **Genes** and DNA information of the person, levels of **hygiene** and **sanitation** conditions also hold a substantial role in

influencing the LE, as explained in the literature above. In addition, environmental conditions like **air pollution** and environmental cleanliness also impact the LE of a person. Similarly, all communicable diseases (CD) and non-communicable diseases (NCD) and significant illnesses or diseases negatively correlate with longevity. It is self-explanatory that pandemics, epidemics, and other viral outbreaks would decrease the average lifespan of the community. All the influencing factors have been pictured as below.

INDEX	GROUP	FEATURES
1	COUNTRY	Country
2		Population
3		GDP
4		Household Income
5		Literacy Rate
6		Schooling
7		Development Status
8	PERSON	Vegetarian Food Consumption
9		Sea Food Consumption
10		Meat Consumption
11		Other protein-based diets
12		Alcohol Consumption
13		Smoking/Tobacco Consumption
14		Physical Exercise
15	HEALTH/ MEDICAL	Genes/DNA
16		Hygiene
17		Sanitation
18		HIV/AIDS
19		Immunisation
20		Air Pollution
21		Communicable Diseases
22		Non-Communicable diseases
23		Availability of Medical Help
24		Medical Infrastructure

Table 1 Factors affecting Life expectancy

4.2. Data Collection

A significant number of studies and research have been carried out in the past in forecasting life expectancy, understanding the demographic transition, and identifying the global changes concerning the increase in LE, including demographic characteristics, income composition, and mortality rates. There have been few recent studies on factors affecting life expectancy. As per the research focus, we also gather information on protein consumption and gather possible information of the countries with LE higher than 80 years and identify the similarities. Therefore possibly, our current research might be able to answer or analyse new trends, if any. In addition, this study would look at features such as vaccination, economics, social factors, mortality, and other health issues. Because the observations in this dataset come from many countries, it shall be possible for a country to discover the predictive factor contributing to a decreased LE value.

The initiative is reliant on data accuracy. The World Health Organization's (WHO) Global Health Observatory (GHO) data repository keeps track of health status and any other relevant parameters for

all nations. The data sets are publicly accessible for analysing health data. Also, additional data has been collected from Our World in Data, OECD and Data world Bank. These are openly available data sources.

In addition to the above-identified features, **adult mortality, child mortality, BMI** (since the physical exercise data is not available countrywide), the average **retirement** age of a person, **child malnutrition**, availability of **essential medication**, **Diphtheria, polio, hepatitis B, measles** (which are significant illnesses relating to immunisation and living standards). Similarly, mental health is another critical impact factor affecting LE. Since there is no direct measure for mental health, **world happiness rank** data and **suicide** information (for which stress is a cause, such as the pressure of life settlement, the pressure of higher education) are considered. In terms of a person's financial stability, **household income, expenditure** and concerning the country's **climatic conditions**, weather data is added to the list of information to be collected. Below is the data collected from WHO, with the relevant GHO indicator in the next column.

FROM	TO	DATA COLLECTED	GHO INDICATOR
1975	2016	BMI	Mean body mass index trends, age-standardised (kg/m²)
2014	2019	Child Mortality	Child mortality levels
2000	2016	Child Malnutrition	Anaemia in children < 5 years
1960	2016	Cholera	Number of reported deaths
2007	2013	Essential Medicines	Median availability of selected generic medicines
2000	2009	Alcohol Consumption	Recorded alcohol per capita consumption, 2000-2009
2010	2019		Recorded alcohol per capita consumption from 2010
2000	2019	HIV	Number of deaths due to HIV/AIDS
1980	2019	BCG	BCG Immunization coverage
1980	2019	Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3)
1989	2019	Hepatitis B	Hepatitis B (HepB3)
2000	2019	Measles	Measles, 2nd dose (MCV2)
1980	2019	Polio	Polio (Pol3)
2000	2019	Suicides	Suicide rate estimates, age-standardized
2000	2016	Adult Mortality	Adult mortality
2000	2019	Non Comm Diseases	Total NCD Mortality
2000	2019	Tuberculosis	TB Mortality
2000	2019	Env Pollution	Mortality from environmental pollution

Table 2 Data collected from GHO

Data collected from, Our World in Data, OECD and Data world bank are as below.

1950	**2019**	**Life expectancy**	**from our world in data website**
2000	2019	Population	Total Population
2014	2017	Meat and poultry Consumption	Per capita meat consumption by type, 2017

1961	2017	Egg Consumption	per capita egg consumption
1961	2017	milk Consumption	per capita milk consumption
2000	2018	Medical Expenditure	Current health expenditure
2000	2018	retirement	OECD.org

Table 3 Data collected from Our World in Data, OECD, Data World Bank

Data has been gathered for 215 countries globally, and research focuses on the data collected from the year 2000 to 2016. Since the features data is not updated/available for all the countries even until 2016, only the most demonstrative vital factors were chosen from all the categories of health-related factors. Below is the list of features ignored because of data unavailability.

INDEX	FEATURES
1	Happiness rank of the country
2	Household income
3	Household Expenditure
4	Literacy Rate
5	Schooling
6	Smoking /Tobacco Consumption
7	Vegetarian Food Consumption
8	Genes/DNA
9	Hygiene
10	Sanitation
11	Stress in general
12	Pressure - of life settlement
13	Pressure - of higher Education
14	Essential Medicine

Table 4 Unselected Features for research because of data unavailability

The several data files have been combined into a single dataset. A brief observation revealed some missing numbers in the data. We further noticed no obvious problems because the datasets originated from WHO. The collected data files have been initially pre-processed and merged using Power BI, and the dataset has been exported from Microsoft Power BI. The data model follows star schema with one-one and one-many relationships. The data table from power BI software is as shown below.

Country	Year	Life expectancy	BMI	ChildMalnutrition	Alcohol	Adult Mortality	ChildMortality	Population	Egg
Albania	2011	76.91	26.20	22.90	5.03	103	868	2905195	
Albania	2012	77.25	26.30	23.10	4.43	103	868	2900401	
Albania	2013	77.55	26.40	23.60	4.28	100	868	2895092	
Algeria	2014	75.88	25.40	33.90	0.54	98	60319	38923687	
Algeria	2015	76.09	25.50	33.90	0.55	96	60319	39728025	
Bahrain	2001	74.64	25.50	25.60	1.89	98	345	697545	
Bahrain	2003	75.00	25.40	24.90	2.04	92	345	778708	
Bahrain	2004	75.17	25.30	24.70	1.98	94	345	829844	
Bahrain	2005	75.33	25.30	24.40	1.92	83	345	889164	
Bahrain	2006	75.48	25.20	24.10	1.94	89	345	958418	
Bahrain	2007	75.63	25.10	23.70	1.83	80	345	1035919	
Bahrain	2009	75.91	25.00	23.00	1.95	69	345	1185076	
Bahrain	2012	76.34	24.90	22.30	1.72	66	345	1299943	
Bahrain	2013	76.48	24.90	22.30	1.65	63	345	1315029	
Bahrain	2014	76.62	24.90	22.40	1.57	62	345	1336075	
Bahrain	2015	76.76	24.80	22.50	1.53	59	345	1371851	
Bahrain	2016	76.90	24.80	22.70	1.39	57	345	1425791	
Barbados	2004	77.81	26.80	22.50	7.43	116	101	275284	
Belarus	2007	69.11	26.20	19.80	13.91	230	761	9560953	
Belarus	2008	69.62	26.20	18.90	14.58	228	761	9527985	
Belarus	2009	70.19	26.30	18.10	13.98	225	761	9506765	
Belarus	2010	70.81	26.30	17.50	14.43	226	761	9490583	
Belarus	2011	71.44	26.40	17.10	14.50	227	761	9473172	
Belarus	2013	72.66	26.50	16.80	12.58	183	761	9465997	
Total		**210,820.68**	**80,785.50**	**116,721.20**	**15,218.17**	**592884**	**11373174**	**99542479312**	

Figure 2 Preview of Consolidated data

4.3. Dataset Description

In this research, we collect data for this study from the year 2000 through the year 2020. However, the data is only available till 2019. (except for dietary consumption and BMI, available until 2017 and 2016). As a result, we considered data from 2000 to 2016, a total of 16 years, to analyse correct data without imputing or deleting data. The collected raw data consists of 2856 rows and 30 columns. The overview of the dataset is presented below.

Overview

Dataset Statistics

Number of Variables	30
Number of Rows	2856
Missing Cells	12691
Missing Cells (%)	14.8%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	833.0 KB
Average Row Size in Memory	298.7 B

Variable Types

Categorical	1
Numerical	29

Figure 3 Overview of Raw Data

5. Exploratory Analysis, Design, and Implementation

The collected raw data set has been analysed using Jupyter notebook and python programming. The values in the dataset have null and missing values.

```
# The number of rows and columns in the dataset
dataset.shape
```

```
(2856, 30)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2856 entries, 0 to 2855
Data columns (total 30 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Country               2856 non-null   object
 1   Year                  2856 non-null   int64
 2   Life expectancy       2567 non-null   float64
 3   BMI                   2856 non-null   float64
 4   ChildMalnutrition     2822 non-null   float64
 5   Cholera               781 non-null    float64
 6   Alcohol               2800 non-null   float64
 7   HIV                   2040 non-null   float64
 8   BCG                   2344 non-null   float64
 9   Adult Mortality       2856 non-null   int64
 10  ChildMortality        2856 non-null   int64
 11  Population            2528 non-null   float64
 12  Eggs Consumption      2339 non-null   float64
 13  Bovine Meat           2339 non-null   float64
 14  Mutton & Goat meat    2339 non-null   float64
 15  Other Meat            2339 non-null   float64
 16  Pig Meat              2273 non-null   float64
 17  Poultry Meat          2339 non-null   float64
 18  Milk Consumption      2339 non-null   float64
 19  Fish and Seafood      2339 non-null   float64
 20  Medical Expenditure   2458 non-null   float64
 21  Retirement Age        782 non-null    float64
 22  Diphtheria            2837 non-null   float64
 23  Suicides              2856 non-null   float64
 24  NCD                   2856 non-null   float64
 25  Env Pollution         2856 non-null   float64
 26  Hepatitis8            2288 non-null   float64
 27  Measles               1629 non-null   float64
 28  Polio                 2820 non-null   float64
 29  Tuberculosis          2799 non-null   float64
dtypes: float64(26), int64(3), object(1)
```

Figure 4 Data size and data types information

5.1. EDA and Design

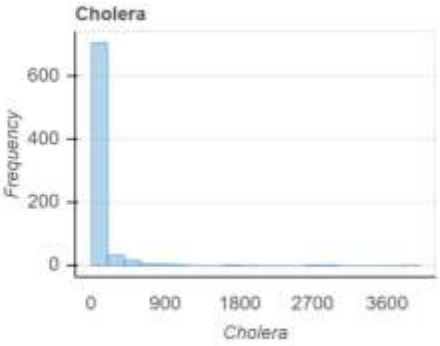
The dataset consists of 30 variables, one categorical variable, "Country", one time-series variable "year", and 28 numeric variables. The data mostly looks accurate with initial observations but not complete as there are some null and missing values in "Cholera"," Retirement Age", and "Measles", etc. Data collected is consistent, and data sources are completely reliable. As mentioned earlier, the data has been collected from WHO, GHO, Our World in Data, OECD and Data world Bank, Where data has been appropriately recorded. Though some data is not updated to the latest the year 2021, it is understandable as data collection and update is time-consuming on a worldwide scale. Also, Data is genuine, trustable and is collected from open source. Also, captured data is easily mapped with numeric scales, and description was given for all the variables. As the data came from different sources and different file formats, merging the data into one file was challenging.

The dataset consists of the "year" column, which is time-series data. So, the datatype has been changed to datetime. While investigating missing values, "Cholera" and "Retirement Age" have more than 50% missing data. "Measles" has about 43% missing information. Hence, we discarded studying these features in our current research.

Cholera
numerical

Hide Details

Distinct Count	189	Mean	83.6914
Unique (%)	24.2%	Minimum	0
Missing	2075	Maximum	3990
Missing (%)	72.7%	Zeros	291
Infinite	0	Zeros (%)	10.2%
Infinite (%)	0.0%	Negatives	0
Memory Size	12.2 KB	Negatives (%)	0.0%



Stats | Plots

Quantile Statistics

Minimum	0
5-th Percentile	0
Q1	0
Median	4
Q3	38
95-th Percentile	390
Maximum	3990
Range	3990
IQR	38

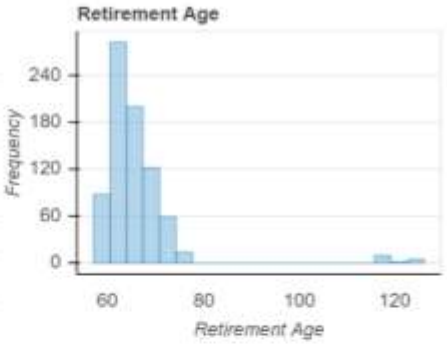
Descriptive Statistics

Mean	83.6914
Standard Deviation	301.2323
Variance	90740.9213
Sum	65363
Skewness	7.7296
Kurtosis	73.0748
Coefficient of Variation	3.5993

Retirement Age
numerical

Hide Details

Distinct Count	776	Mean	66.1256
Unique (%)	99.2%	Minimum	57.0353
Missing	2074	Maximum	126.1264
Missing (%)	72.6%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	12.2 KB	Negatives (%)	0.0%



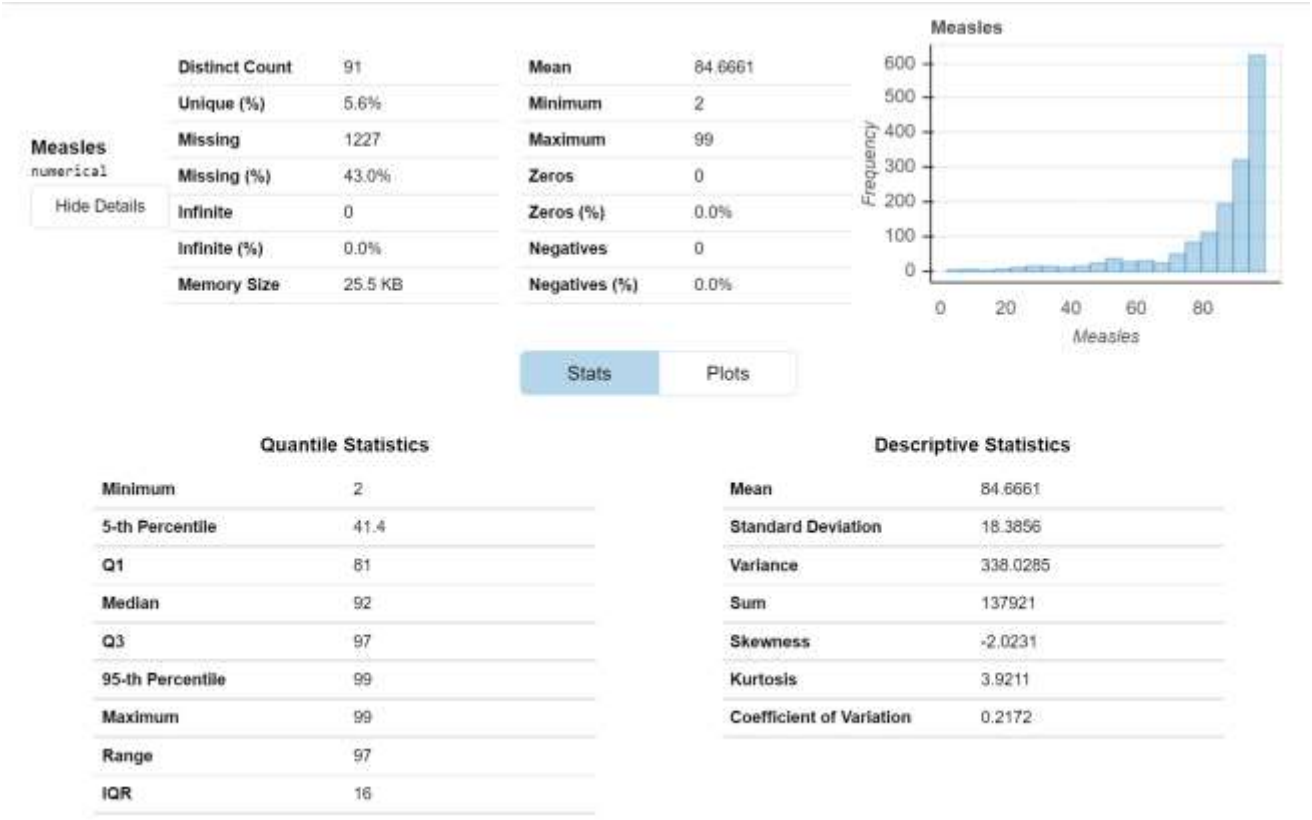
Stats | Plots

Quantile Statistics

Minimum	57.0353
5-th Percentile	59.1888
Q1	61.876
Median	64.2499
Q3	67.8664
95-th Percentile	73.3676
Maximum	126.1264
Range	69.0911
IQR	5.9904

Descriptive Statistics

Mean	66.1256
Standard Deviation	9.0432
Variance	81.779
Sum	51710.2301
Skewness	4.75
Kurtosis	26.1903
Coefficient of Variation	0.1368



Also, there is a 10% (289 rows) null value Life expectancy Column. Our research focuses on LE so, imputing the values introduces bias which is not desirable. Hence, we delete these null rows from our dataset, and we have 2567 rows and 27 columns. Except "HIV", "BCG" and "Hepatitis B" has missing values less than 25%, and these are continuous numerical variable. Henceforward, we replace the missing values in these columns with mean. And other missing values are less than 10%, so we delete the null rows, and the dataset now consists of 2074 rows with 27 columns.

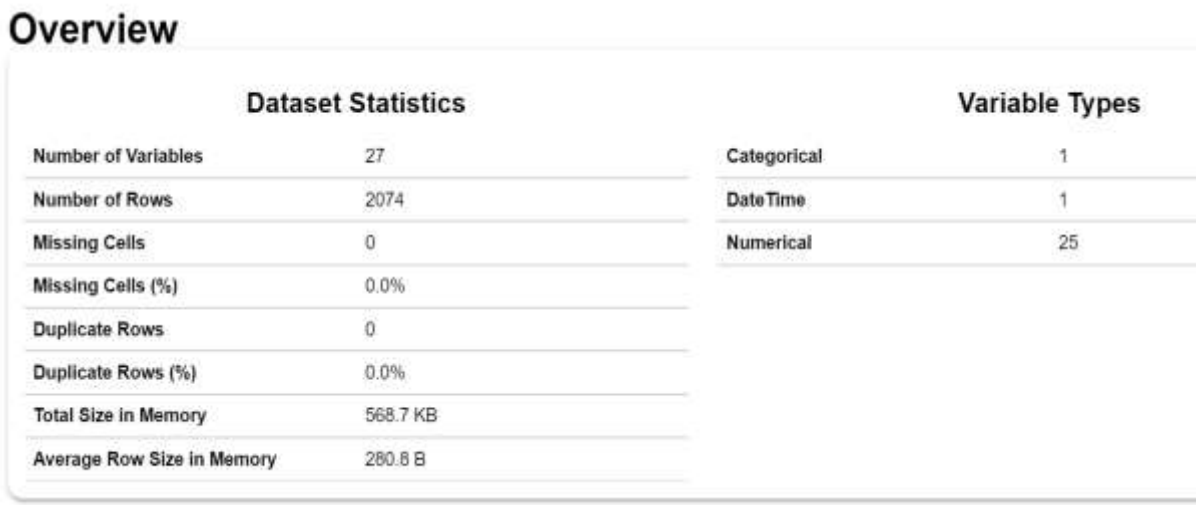


Figure 5 Overview of data after pre-processing

After performing EDA and processing the data, the data set has 2074 rows and 27 columns, categorical, numerical, and datetime variables and 0% missing values and columns.

5.2. Implementation

5.2.1. Data Encoding

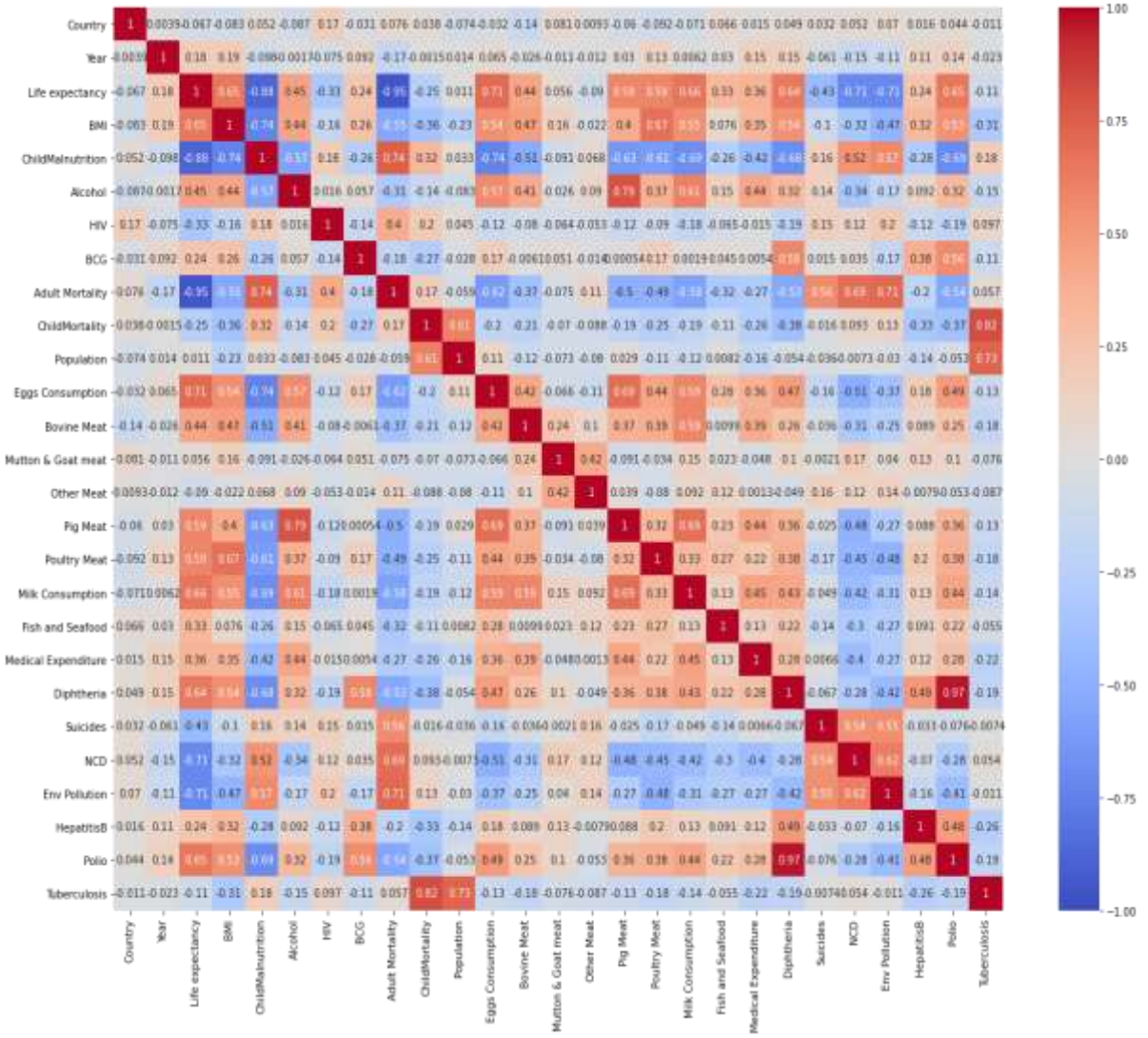
Since the finalised dataset consists of numerical and categorical variables, we perform data encoding to convert categorical variables to numeric using Label Encoder. Label encoding is the process of translating labels into the numeric form so that machines may read them. Machine learning algorithms can then make better decisions and define the correlation between the features and how those labels should be used.

5.2.2. Feature Selection

Features could be selected by following any one of the methods. Filter, wrapper, or embedded methods. We are using the Pearson correlation technique under the filter method to select the highly positively correlated related features (> 0.25) and highly negatively correlated features (< -0.25)

Life expectancy	1.000000
Eggs Consumption	0.708943
Milk Consumption	0.659305
BMI	0.654637
Polio	0.649462
Diphtheria	0.641525
Pig Meat	0.591232
Poultry Meat	0.590431
Alcohol	0.450114
Bovine Meat	0.439087
Medical Expenditure	0.364535
Fish and Seafood	0.330970
HepatitisB	0.244216
BCG	0.235935
Year	0.183378
Mutton & Goat meat	0.056031
Population	0.010906
Country	-0.067061
Other Meat	-0.090164
Tuberculosis	-0.106261
ChildMortality	-0.250402
HIV	-0.326044
Suicides	-0.429654
NCD	-0.706556
Env Pollution	-0.708975
ChildMalnutrition	-0.882725
Adult Mortality	-0.948039

Figure 6 Correlation between features in the dataset



5.2.3. PCA and Feature Scaling

Because there are just 19 highly associated variables in the dataset, it is not very dimensional. As a result, PCA or dimensionality reduction are not required in our present dataset. The data recorded in each column has a huge difference in magnitude, as shown in the first few rows of the dataset. We standardised features by removing the mean and scaling to unit variance.

	BMI	ChildMalnutrition	Alcohol	HIV	Adult Mortality	ChildMortality	Eggs Consumption	Bovine Meat	Pig Meat	Poultry Meat	Milk Consumption	Fish and Seafood	Medical Expenditure
0	26.2	22.9	5.03	100.0	103	868	7.72	21.24	11.03	13.41	301.27	5.86	4.795327
1	26.3	23.1	4.43	100.0	103	868	12.69	22.40	11.04	12.76	299.85	4.97	5.055262
2	26.4	23.6	4.28	100.0	100	868	12.45	22.50	10.88	13.23	303.72	4.87	5.385599
3	25.4	33.9	0.54	200.0	96	60319	7.93	5.43	0.00	6.86	151.06	4.40	6.547214
4	25.5	33.9	0.55	200.0	96	60319	8.65	5.35	0.00	6.64	125.37	4.16	6.978492

Figure 7 Data preview before feature scaling

After normalising the values, the data is split into training and testing sets in a 70-30 ratio using the test_train_split method from sklearn.model_selection package. The split data was improperly balanced. Hence, the data were resampled using the oversampling technique to obtain properly balanced test and train datasets.

6. Model Selection

The research objective is now explicit, and the data needed to use machine learning techniques has been processed and is available. We use classification models to investigate the factors that influence longer life expectancy. Recognition, comprehending, and arranging concepts and objects into predetermined groups. Machine learning algorithms classify future datasets by using pre-categorised training datasets and a range of machine learning methods. Among the most efficient and popular classification algorithms, we have chosen to work on Logistic Regression, Naive Bayes, K-Nearest Neighbor and Support Vector Machines algorithms.

6.1. Logistic Regression

The Life Expectancy dataset after processing contains 19 features with 2074 observations. Therefore, the number of observations is greater than the number of features and is suitable for applying Logistic Regression (LR). Since LR applies regularisation by default, this model may **not** perform comparatively well for the dataset with complex **non-linear relationships**. As the features in our dataset have a complex non-linear correlation with LE, we increase the value of C, the inverse regularisation strength, to build a better model (at C=14) with better accuracy of 38%.

```
from sklearn.linear_model import LogisticRegression

#defining default values to store high accuracy and values of C at higher ccuracy
HIGH_ACCURACY =0
MAX_C = 1

# Looping for values of c from 1 till 20 to check the behaviour and accuracy of the model.
for c in range(1,20):

    #defining linear model with default parameters and C=c
    model = LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=c,
                               fit_intercept=True, intercept_scaling=1, class_weight=None,
                               random_state=None, solver='lbfgs', max_iter=100, multi_class='auto',
                               verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)

    #Fitting the model for the test and train sets
    model = model.fit(X_train,y_train)

    #Predicting the test values with the built model
    pred_y = model.predict(X_test)

    #printing the accuracy on the console
    print("C = {} ,  Accuracy = {}".format(c, model.score(X_test, y_test)))

    #if accuracy is high then store it in the variables
    if(HIGH_ACCURACY < model.score(X_test, y_test)):
        HIGH_ACCURACY = model.score(X_test, y_test)
        MAX_C = c


#Printing the the highest accuracy achieved and the respective C value
print("**************************************")
print("C = {} ,  Accuracy = {}".format(maxc, HIGH_ACCURACY))
```

```
C = 1 ,  Accuracy = 0.3467094703049759
C = 2 ,  Accuracy = 0.36436597110754415
C = 3 ,  Accuracy = 0.3611556982343499
C = 4 ,  Accuracy = 0.3611556982343499
C = 5 ,  Accuracy = 0.36276083467094705
C = 6 ,  Accuracy = 0.36597110754414125
C = 7 ,  Accuracy = 0.36757624398073835
C = 8 ,  Accuracy = 0.36757624398073835
C = 9 ,  Accuracy = 0.36918138041733545
C = 10 ,  Accuracy = 0.3723916532905297
C = 11 ,  Accuracy = 0.3739967897271268
C = 12 ,  Accuracy = 0.36918138041733545
C = 13 ,  Accuracy = 0.36597110754414125
C = 14 ,  Accuracy = 0.38202247191011235
C = 15 ,  Accuracy = 0.3723916532905297
C = 16 ,  Accuracy = 0.36918138041733545
C = 17 ,  Accuracy = 0.3739967897271268
C = 18 ,  Accuracy = 0.36757624398073835
C = 19 ,  Accuracy = 0.36757624398073835
```

Figure 8 Logistic Regression Model

```
LOGISTIC REGRESSION , C=14
------------------------------
Accuracy : 0.38202247191011235
MAE : 1.16
MSE : 3.83
RMSE  : 1.958232
R2_SCORE  : 0.960430
F1_SCORE  : 0.340216
```

Figure 9 Logistic Regression Performance Metrics

6.2. Naïve Bayes

Naïve Bayes is an eager learning algorithm widely used for classification purposes and can be applied for multiclass prediction and performs better than linear regression algorithms. This model assumes **class conditional independence,** which is more suitable for the current life expectancy dataset. We build models using **Gaussian Model** and **Bernoulli Model**. We choose to apply Gaussian Model as the features in our model are normalised using the StandardScalar technique. This model is suitable for our current dataset. Compared to Gaussian, Bernoulli Model is not very suitable for the application as our data is not a binomial model; hence, it might result in lower accuracy when applied.

```
# GAUSSIAN NAIVE BAYES

#IMPORTING GAUSSIAN NAIVE BAYES PACKAGE
from sklearn.naive_bayes import GaussianNB

# DEFINING MODEL
gnb = GaussianNB()

# TRAINING THE DEFINED MODEL AND PREDICTING THE TARGET WITH TRAINED MODEL
y_pred = gnb.fit(X_train, y_train).predict(X_test)

# CALCULATING MEAN ABSOLUTE ERROR
MAE = mean_absolute_error(y_test, y_pred)

# CALCULATING THE MEAN SQUARED ERROR
MSE = mean_squared_error(y_test, y_pred)

# CALCULATING ROOT MEAN SQUARED ERROR
RMSE = sqrt(MSE)

# CALCULATING R2_SCORE
R2_SCORE=r2_score(y_test, y_pred)

# CALCULATING F1_SCORE
F1_SCORE = f1_score(y_test, y_pred, average='macro')

print('GAUSSIAN NAIVE BAYES ')
print('------------------------------')
print('Accuracy : {}'.format(gnb.score(X_test, y_test)))
print('MAE : {}'.format(round(MAE, 2)))
print('MSE : {}'.format(round(MSE, 2)))
print('RMSE  : %f' % RMSE)
print('R2_SCORE  : %f' % R2_SCORE)
print('F1_SCORE  : %f' % F1_SCORE)
```

```
# BernoulliNB Naive Bayes
from sklearn.naive_bayes import BernoulliNB

# DEFINING MODEL
bnb = BernoulliNB()

# TRAINING THE DEFINED MODEL AND PREDICTING THE TARGET WITH TRAINED MODEL
model = bnb.fit(X_train, y_train).predict(X_test)

# CALCULATING MEAN ABSOLUTE ERROR
MAE = mean_absolute_error(y_test, y_pred)

# CALCULATING THE MEAN SQUARED ERROR
MSE = mean_squared_error(y_test, y_pred)

# CALCULATING ROOT MEAN SQUARED ERROR
RMSE = sqrt(MSE)

# CALCULATING R2_SCORE
R2_SCORE=r2_score(y_test, y_pred)

# CALCULATING F1_SCORE
F1_SCORE = f1_score(y_test, y_pred, average='macro')

print('BERNOULLI NAIVE BAYES ')
print('------------------------------')
print('Accuracy : {}'.format(gnb.score(X_test, y_test)))
print('MAE : {}'.format(round(MAE, 2)))
print('MSE : {}'.format(round(MSE, 2)))
print('RMSE  : %f' % RMSE)
print('R2_SCORE  : %f' % R2_SCORE)
print('F1_SCORE  : %f' % F1_SCORE)
```

Figure 10 Gaussian and Bernoulli Naive Bayes Models

```
GAUSSIAN NAIVE BAYES
------------------------------
Accuracy : 0.30818619582664525
MAE : 1.63
MSE : 5.96
RMSE  : 2.440956
R2_SCORE  : 0.938516
F1_SCORE  : 0.249275
```

Figure 11 Gaussian Naive Bayes Performance Metrics

```
BERNOULLI NAIVE BAYES
------------------------------
Accuracy : 0.22150882825040127
MAE : 1.63
MSE : 5.96
RMSE  : 2.440956
R2_SCORE  : 0.938516
F1_SCORE  : 0.249275
```

Figure 12 Bernoulli Naive Bayes Performance Metrics

6.3. K-Nearest Neighbors

KNN is a simple yet effective machine learning algorithm. Unlike Naïve Bayes Algorithm, KNN is a lazy learning algorithm and does not make any assumptions of the data distribution in the dataset. This model predicts the target based on the similarity measures and is well suitable for multi-modal classes. We have built the KNN model using default parameters and predicted the target for n=5. Predicting for n=3 or n=4 gives higher accuracy but could result in overfitting as we consider only a few nearest neighbours to determine the target value. This model resulted in an accuracy of 85%, as demonstrated below.

```
from sklearn.neighbors import KNeighborsClassifier

#defining default values to store high accuracy and values of n at higher ccuracy
HIGH_ACCURACY =0
MAX_N = 1

# Looping for values of n from 5 till 20 to check the behaviour and accuracy of the model.
for n in range(5,20):

    #defining model with default parameters and n_neighbors = n
    knn = KNeighborsClassifier(n_neighbors = n)

    #Fitting the model for the test and train sets
    knn.fit(X_train,y_train)

     #Predicting the test values with the built model
    y_pred = knn.predict(X_test)

    #printing the accuracy on the console
    print('n = {} , Accuracy = {}'.format(n, knn.score(X_test, y_test)))

     #if accuracy is high then store it in the variables
    if(HIGH_ACCURACY < knn.score(X_test, y_test)):
        HIGH_ACCURACY = knn.score(X_test, y_test)
        MAX_N= n


# print('KNeighborsClassifier')
#Printing the the highest accuracy achieved and the respective C value
print("************************************")
print("Maximum Accuracy is achieved at n = {} ,  Accuracy = {}".format(MAX_N, HIGH_ACCURACY))
```

Figure 13 K Nearest Neighbor Model

```
n = 5 , Accuracy = 0.8507223113964687
n = 6 , Accuracy = 0.797752808988764
n = 7 , Accuracy = 0.7736757624398074
n = 8 , Accuracy = 0.7223113964686998
n = 9 , Accuracy = 0.6934189406099518
n = 10 , Accuracy = 0.666131621187801
n = 11 , Accuracy = 0.6324237560192616
n = 12 , Accuracy = 0.5858747993579454
n = 13 , Accuracy = 0.565008025682183
n = 14 , Accuracy = 0.5280898876404494
n = 15 , Accuracy = 0.5136436597110754
n = 16 , Accuracy = 0.4767255216693419
n = 17 , Accuracy = 0.47191011235955055
n = 18 , Accuracy = 0.47351524879614765
n = 19 , Accuracy = 0.4510433386837881
```

```
KNeighborsClassifier
------------------------------
Accuracy : 0.8507223113964687
MAE : 0.17
MSE : 0.21
RMSE  : 0.458555
R2_SCORE  : 0.997830
F1_SCORE  : 0.892837
```

Figure 14 KNN Performance Metrics

6.4. Support Vector Machine

SVM is a robust, powerful machine learning algorithm that classifies data using the hyperplane concept, separating with maximum margin. It is a well-known, most efficient ML algorithm for data with non-regularity, i.e., unknown distribution, and can easily be overfitted. This model makes use of kernel functions to perform non-linear partitioning on the current life expectancy dataset. We build the model using all the four kernel functions **linear, radial basis, polynomial** and **sigmoid**. Also, the other tuning **hyperparameter C** and kernel coefficient **gamma**, used for soft margin classification. Lower the value of c, wider the margins and higher the violations when data is classified.

```python
#Importing the necessary packages and libaries
from sklearn import svm

print('SUPPORT VECTOR CLASSIFICATION ')

# LIST OF KERNALS
kernal = ["linear","rbf",'poly','sigmoid']

#LOOPING THROUGH THE LIST OF KERNALS
for k in kernal:

    print('------------------------------')

    # LOOPING THROUGH THE PARAMETER g TO TUNE OUR MODEL ACCORDINGLY
    for g in ["auto" ,"scale"]:
        HIGH_ACC = 0
        Max_C = 1

        # LOOPING THROUGH VALUE OF C FROM 1 TO 15
        # higher value of c gives l2 penality --> overfitting
        for c in range(1,15):

            # DEFINING THE MODEL
            model = svm.SVC(C=c, kernel=k, gamma= g, decision_function_shape='ovo')

            # TRAINING THE CREATED MODEL
            model = model.fit(X_train,y_train)

            # PREDICTING THE TATRGET WITH TEST VALUES
            pred_y = model.predict(X_test)

            # STORING THE HIGH ACCURACY AND CORRESPONDING C AND g VALUES
            if HIGH_ACC < model.score(X_test,y_test):
                HIGH_ACC = model.score(X_test,y_test)
                Max_C = c

        # PRINTING THE HIGH ACCURACY FOR ALL THE KERNAL WITH DIFF PARAMETER COMBINATIONS
        print("Kernal = {} , C = {} , gamma = {} - MaxAccuracy = {}".format(k,Max_C,g,HIGH_ACC))
```

Figure 15 Support Vector Machine Model

```
SUPPORT VECTOR CLASSIFICATION
------------------------------
Kernal = linear , C = 10 , gamma = auto - MaxAccuracy = 0.6773675762439807
Kernal = linear , C = 10 , gamma = scale - MaxAccuracy = 0.6773675762439807
------------------------------
Kernal = rbf , C = 14 , gamma = auto - MaxAccuracy = 0.8154093097913323
Kernal = rbf , C = 14 , gamma = scale - MaxAccuracy = 0.7479935794542536
------------------------------
Kernal = poly , C = 14 , gamma = auto - MaxAccuracy = 0.7897271268057785
Kernal = poly , C = 14 , gamma = scale - MaxAccuracy = 0.6821829855537721
------------------------------
Kernal = sigmoid , C = 4 , gamma = auto - MaxAccuracy = 0.22632423756019263
Kernal = sigmoid , C = 5 , gamma = scale - MaxAccuracy = 0.28892455858747995
```

Figure 16 SVM Highest accuracy with different Kernels

We built the classification machine learning model using SVM and identified the maximum accuracy point with respective kernel coefficient values and hyperparameter C. The results are displayed above. The performance, respective metric values are calculated for all the built models, and their accuracies are as below.

```
Linear SVC
-------------------------------
Accuracy : 0.6773675762439807
MAE : 0.4
MSE : 0.58
RMSE  : 0.763325
R2_SCORE  : 0.993987
F1_SCORE  : 0.703947
```

```
Radial Basis Function - SVC
-------------------------------
Accuracy : 0.8154093097913323
MAE : 0.21
MSE : 0.29
RMSE  : 0.541978
R2_SCORE  : 0.996969
F1_SCORE  : 0.812503
```

```
Polynomial Kernal Function - SVC
-------------------------------
Accuracy : 0.7897271268057785
MAE : 0.21
MSE : 0.29
RMSE  : 0.541978
R2_SCORE  : 0.996969
F1_SCORE  : 0.812503
```

```
Sigmoid Function - SVC
-------------------------------
Accuracy : 0.28892455858747995
MAE : 0.21
MSE : 0.29
RMSE  : 0.541978
R2_SCORE  : 0.996969
F1_SCORE  : 0.812503
```

Figure 17 SVM Performance Metrics for different Kernels

The Models achieved different accuracies for different kernel functions, and the model with comparatively higher accuracy is the Radial Basis Kernel Function with an accuracy of 81%

7. Evaluation

We have successfully developed machine learning models with different classification algorithms. Accuracy is not the only factor to analyse and decide the appropriate and most efficient model. The accuracy of the model can sometimes be misleading when the model tuning has overfitted the test results. This overfitting might achieve high accuracy for test data but gives a poor performance in real-time data prediction. Hence, we make use of performance metrics to decide the best model among the developed ML models. Many performance factors are already defined to evaluate the model like Confusion matrix, Accuracy of the model, Precision and Recall, Specificity, F1 score, Precision-Recall or PR curve, ROC (Receiver Operating Characteristics) curve and PR vs ROC curve.

In this research, we have chosen to calculate Mean Average Error (**MAE**), Mean Squared Error (**MSE**), Root Mean Squared Error (**RMSE**), R2 Score and F1 Score to evaluate the model performance. The best fit model must have a lower MAE (where the mean average error is not very high, implying the prediction is closer to the actual value), sometimes this can be misleading as the error could be positive or negative, so we calculate MSE and RMSE.

Also, the model with lower MSE and lower RMSE are preferred, like MAE. Similarly, the **R2 Score** is the proportion of target variance and is closely related to MAE, and the higher R2 Score better fit model could be. Precision and recall have a harmonic mean. This factor considers both, therefore the more significant the **F1 Score**, the better.

MODEL	ACCURACY	MAE	MSE	RMSE	R2_SCORE	F1_SCORE
GAUSSIAN NAIVE BAYES	31%	1.63	5.96	2.440956	0.938516	0.249275
BERNOULLI NAIVE BAYES	22%	2.95	23.42	4.839478	0.758323	0.153571
LOGISTIC REGRESSION	38%	1.16	3.83	1.958232	0.96043	0.340216
KNN CLASSIFIER	85%	0.17	0.21	0.458555	0.99783	0.892837
LINEAR SVC	68%	0.4	0.58	0.763325	0.993987	0.703947
RADIAL BASIS FUNCTION SVC	82%	0.21	0.29	0.541978	0.996969	0.812503
POLYNOMIAL SVC	79%	0.21	0.29	0.541978	0.996969	0.812503
SIGMOID SVC	29%	0.21	0.29	0.541978	0.996969	0.812503

Figure 18 Performance Metrics for various developed Models

The KNN model is developed by considering 5 nearest neighbours, where considering less than 5 gives higher accuracy but may result in overfitting, and choosing a value greater than 5 may result in a drop in accuracy and F1 Score, so it is often impossible to say which model outperforms and is the best; this always depends on the assumptions, dataset, and requirements set by the research. Performance metrics are only a measure, and the best fit model always depends on the type of dataset it has been built. Hence, there are no right and wrong models. Also, model fitting is a pretty straightforward process, but appropriately tuning them and choosing among them is the significant problem of applied machine learning. The current life expectancy data set KNN model has the highest F1 Score, R2 Score, lowest values for errors MSE, RMSE and high accuracy of 85% from the helpful guiding measure. Though SVM is proven to be the most effective and robust model sometimes, it can be outperformed by the KNN model. (KURAMOCHI and KARYPIS, 2011).

8. Conclusion

8.1. Discussion

Consumption of protein is analysed by the quantity and variety of protein-rich foods consumed by different countries. We have categorised the features showing the intake of Eggs, Meat, Poultry, Milk, and seafood into groups. All these groups have shown that people with higher longevity have consumed higher proteins than other countries. Proteins are essential components of bones, blood, skin, cartilage, and muscles, and they help in energy consumption, physiological activities, and

immunological functions, resulting in a better likelihood of recovery from diseases. As a result, a longer life expectancy could be expected.

Similarly, other mentioned food types in this category contain different types of protein. There are different types of protein available in a different type of foods such as eggs contain Ovalbumin (54%), Beef / Bovine meat products contain nearly 55 percent myofibrillar proteins, which are mostly actin and myosin, 30–34 percent sarcoplasmic proteins, which are mostly enzymes and myoglobin, and connective tissue, which is about 15 percent collagen and elastin fibres embedded in mucopolysaccharides, and connective tissue, which is about 15 percent collagen and elastin fibres embedded in mucopolysaccharides. As a result, we were unable to make any conclusions based on the protein type.

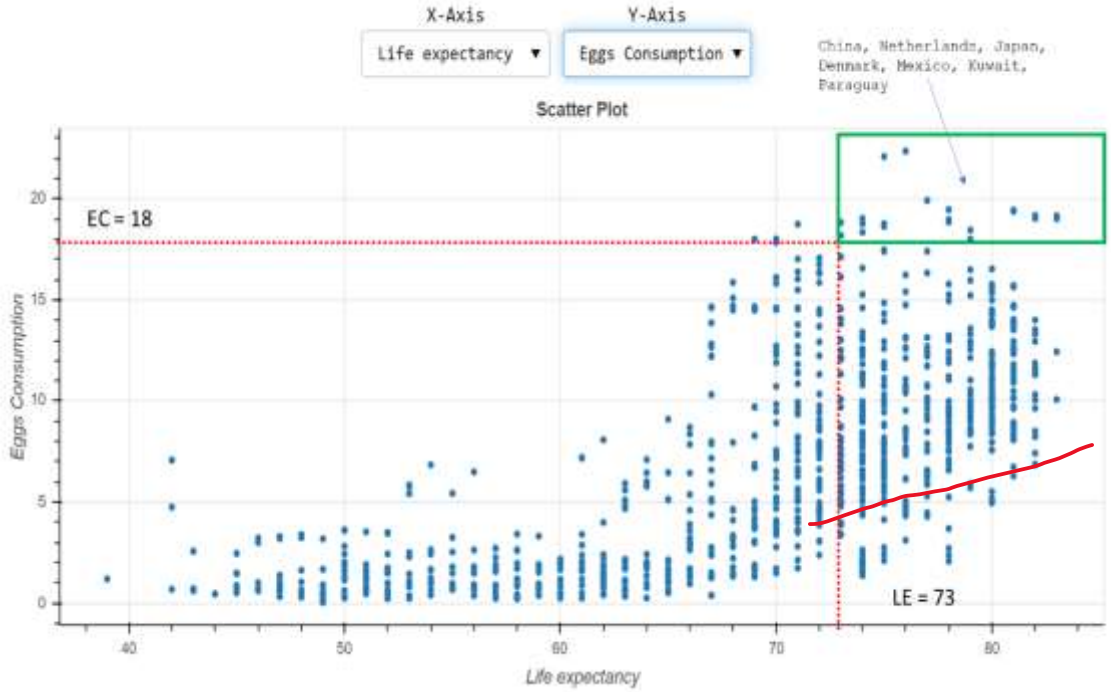


Figure 19 Egg Consumption Vs Life Expectancy

Looking at the scatter plot between Life expectancy (LE in years) and Egg consumptions (Per capita consumption of eggs kilograms per year). It can be observed that after the average life expectancy (LE = 72.6 (73 approx.)) there is a linear increase in the consumption of protein from the egg. **China, Netherlands, Japan, Denmark, Mexico, Kuwait,** and **Paraguay** are at the top end of the consumption plot shown in Figure 19. Similarly, in Figure 20, Consumption of pork is comparatively high in **Poland, Montenegro, Spain, Germany, Austria,** and **Luxembourg**

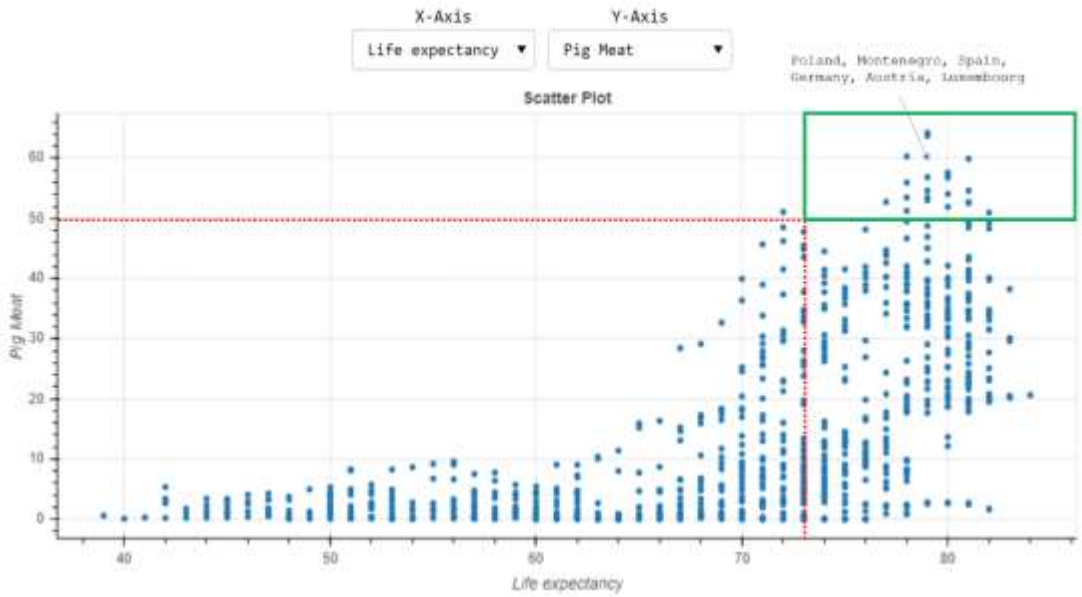


Figure 20 Pork / Pig Meat Consumption Vs Life Expectancy

Similar data behaviour is observed for Beef/ Bovine meat consumption, Poultry meat and milk consumption. **Argentina, Australia, Brazil, New Zealand,** and **Uruguay** have beef more than 35 per kilogram per year per capita. **Kuwait, Barbados, Israel, Jamaica, United Arab Emirates, Trinidad**, and **Tobago** consumes high poultry. The data could be cross-validated with the availability of live cattle and the import and export data of meat for these countries.

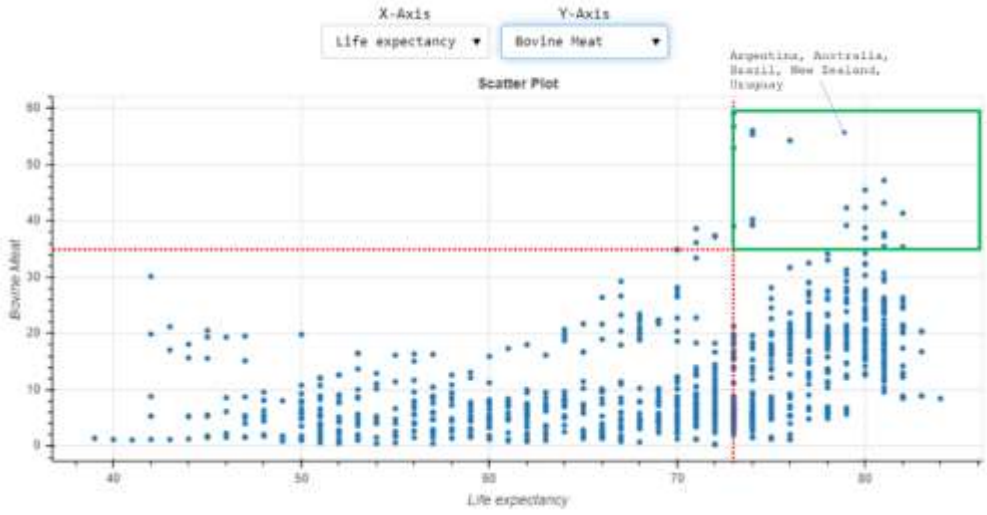


Figure 21 Beef Consumption Vs Life Expectancy

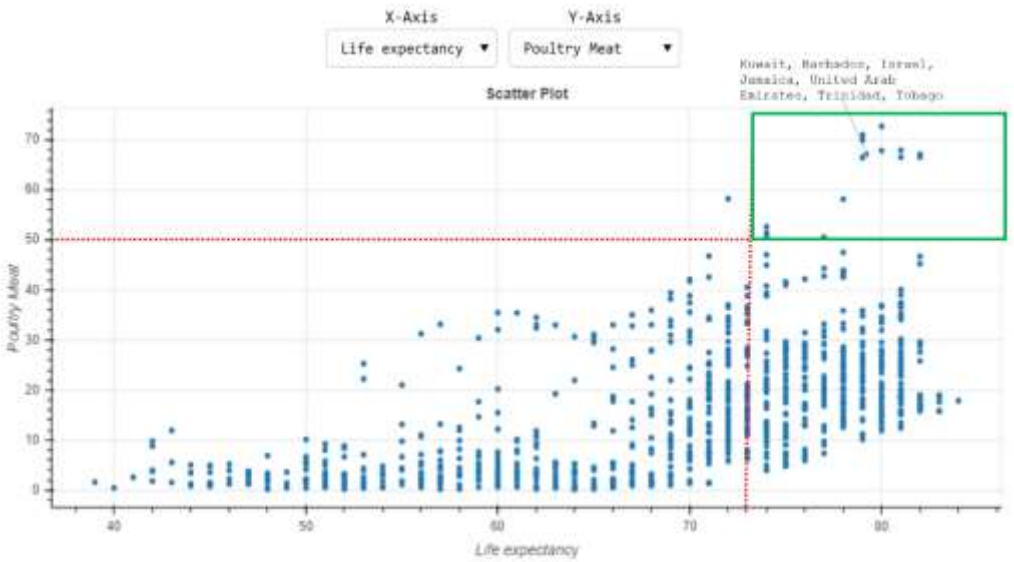


Figure 22 Poultry Meat Consumption Vs Life Expectancy

Albania, Netherlands, Montenegro, Sweden, Estonia, Lithuania, Switzerland, Denmark, Finland, Greece, Ireland, and **Luxembourg** are high in milk and milk product consumption, excluding butter and related products. The relationship between milk consumption and LE is linear, like other meat consumption plots. Many countries use milk and milk-related products in good quantity, but only a few countries have high LE.

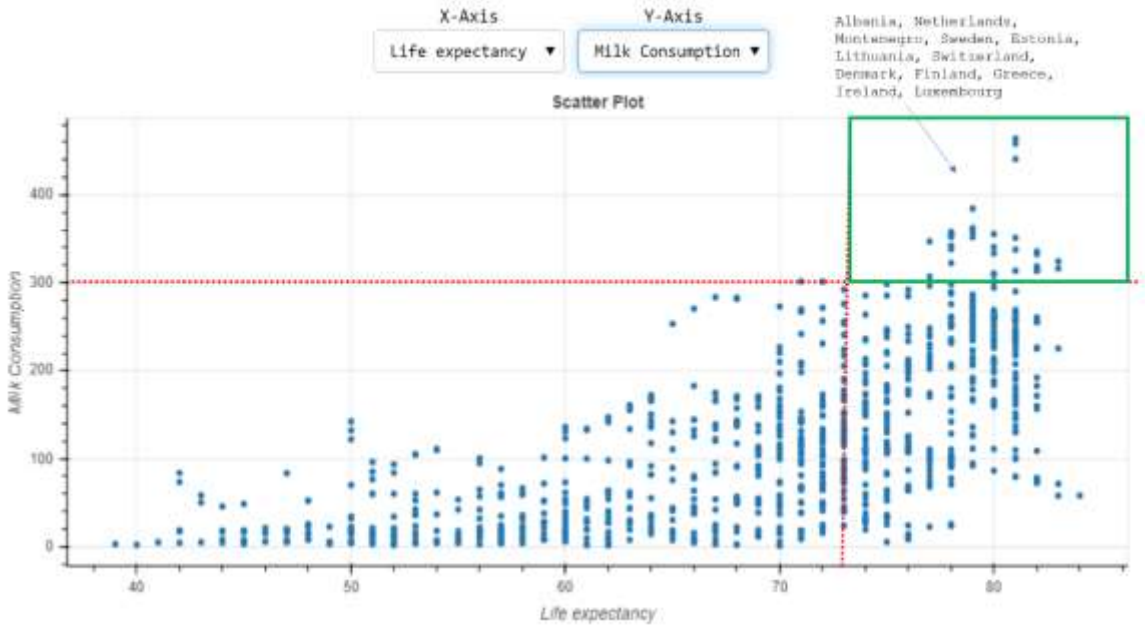


Figure 23 Milk Consumption Vs Life Expectancy

Iceland and **Maldives** being islands and have abundant availability of seafood compared to live cattle and land grown food products. Fish and seafood is the primary source of diet, and it strongly correlates with LE. Other countries **Malaysia, Japan, Norway,** and **Portugal,** also have availability of seafood with consumption per capita between 50 to 100 kilograms per year.

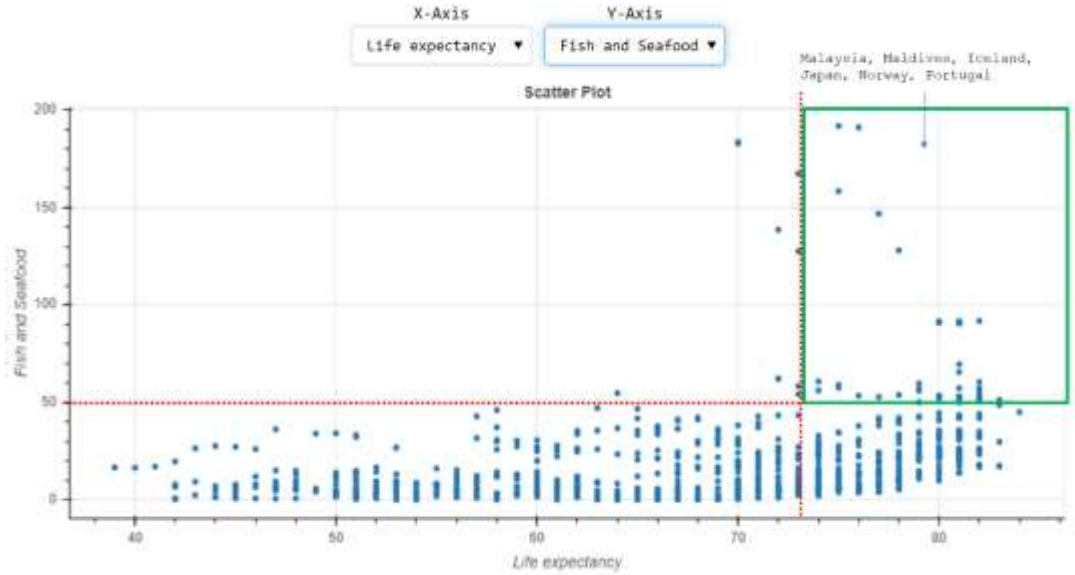


Figure 24 Fish and Seafood Consumption Vs Life Expectancy

Infectious illnesses continue to account for a substantial number of fatalities in low-income nations, emphasising health disparities primarily induced by economic inequalities. Vaccination can help to decrease healthcare expenditures and inequalities. Controlling, eliminating, or eradicating diseases may save communities and governments billions of pounds. Vaccines are seen as critical in the fight against bioterrorism. (Andre, 2008). They can help fight antibiotic resistance in some infections. Influenza vaccination may also help to prevent non-communicable illnesses like ischemic heart disease. Long, healthy lives are increasingly considered a requirement for riches, and money encourages health. Vaccines are thus effective instruments for reducing income gaps and health inequities.

The correlation between BCG Immunisation coverage and LE is 0.2, which shows no strong correlation between these two features; hence this has not been included in building our model. Other immunisation data has not been collected because of data unavailability for the year 2000 till 2016, by observing the illnesses (Non-Communicable Diseases, Mental health (measured by the number of suicides), HIV, Diphtheria and polio has been considered in building our model. Hepatitis B and Tuberculosis have been deleted from the dataset because of a weak correlation with LE. Other illnesses like Cholera and Measles have been removed from the dataset as there was more than 50% of missing data. The Pearson correlation matrix below demonstrates a negative correlation between illnesses and life expectancy.

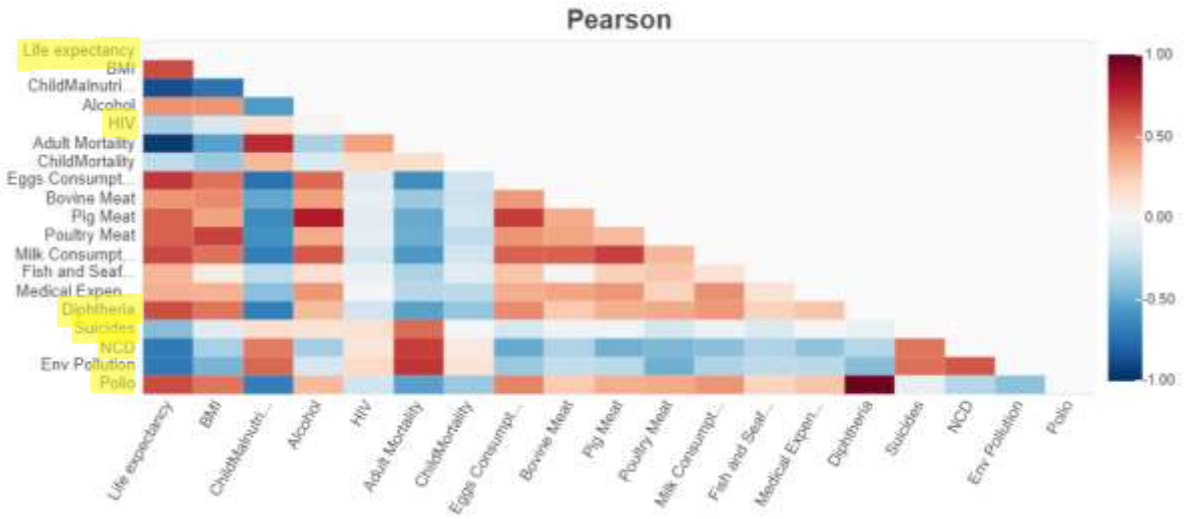


Figure 25 Correlation between Immunisation, Illnesses and Life Expectancy

8.2. Conclusions

How could protein consumption affect the life span? This research aimed to understand the factors impacting the life expectancy of a human. And we also focused on countries with LE more than average LE= 72.6 years and identified similarities, if any. The data in this research shows that there is a strong correlation between protein intake and LE. Protein aids in energy consumption and physiological functions, and immunological functions. Hence a higher life expectancy is observed in countries like Japan, Italy, Switzerland, Spain, Singapore, Australia, Iceland, and the Netherlands. Due to a lack of data, no in-depth study of vaccinations, diseases, or LE could be conducted. However, the available data shows that illnesses, long- and short-term health problems, non-communicable diseases, and mental health all directly impact death rates, making life expectancy inversely proportional.

We built machine learning models using several machine learning methods and the analyses and observations collected throughout the research. We calculated Mean Average Error, Mean Squared Error, and Root Mean Squared Error, R2 Score, and F1 Score to assess the model's performance. Performance measurements are simply a statistic, and the best-fit model is always dependent on the dataset from which it was created. Therefore, we could develop an efficient machine learning model using the K Nearest Neighbor algorithm with an accuracy of 85%.

8.3. Future Work

This study may be expanded to include other dependent characteristics such as vegetarian food consumption, immunisation, and vaccination data from various nations and their influence on Life Expectancy. Weather, climatic circumstances, and a person's regular physical activity are all factors to consider. The indicated data were not included in this study due to a lack of availability, although they may have a strong correlation and provide avenues for further research. Because the study's findings suggest that increased protein consumption is linked to a longer life expectancy, this dataset may be combined with ketogenic diet-related datasets for additional analysis and research. The ketogenic diet (or keto diet) is commonly thought to be a low-carb, high-fat diet with several health advantages. (Paoli, 2014)Many studies have shown that eating this way can help you lose weight and improve your health. Alzheimer's disease Diabetes and epilepsy may all benefit from ketogenic diets.(Weber, Aminazdeh-Gohari and Kofler, 2018).

In addition to the research findings, I want to provide the established machine learning model as an API service using the Flask Framework so that other parties may use it for their research and studies. The created API might be helpful in predictive analysis and the formulation and planning of policies linked to Life Expectancy.

9. Legal Ethical and Professional Issues

The research data for this project has been gathered from open source (WHO, GHO, Our World in Data, OECD and Data world Bank), which is provided for research and other analytical purposes. This project does not involve any personal or third-party information and respects all the principles under the Research Ethics of Teesside University.

10. References

Andre, F. E. (2008) 'Policy and practice Vaccination and reduction of disease and inequity', *Bulletin of the World Health Organization*, 86(2).

Antonovsky, A. (no date) *SOCIAL CLASS, LIFE EXPECTANCY AND OVERALL MORTALITY*.

Beekman, M. *et al.* (2013) 'Genome-wide linkage analysis for human longevity: Genetics of healthy aging study', *Aging Cell*, 12(2), pp. 184–193. doi: 10.1111/acel.12039.

Beeksma, M. *et al.* (2019) 'Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records'. doi: 10.1186/s12911-019-0775-2.

Berry, S. (2021) 'The Most Important Thing That Ever Happened: Big, Bad Data and the Doubling of Human Life Expectancy', *Journal of Planning History*. doi: 10.1177/15385132211013797.

Boucekkine, R., De La Croix, D. and Licandro, O. (2003) *Early Mortality Declines at the Dawn of Modern Growth*, *Source: The Scandinavian Journal of Economics*.

Cervellati, M. (2009) *Life Expectancy and Economic Growth: The Role of the Demographic Transition*.

Chen, C. C., Chen, P. S. and Yang, C. Y. (2019)' Relationship between fine particulate air pollution exposure and human adult life expectancy in Taiwan', *Journal of Toxicology and Environmental Health - Part A: Current Issues*, 82(14), pp. 826–832. doi: 10.1080/15287394.2019.1658386.

Crimmins, E. M. and Zhang, Y. S. (2019) 'Aging Populations, Mortality, and Life Expectancy', *Annual Review of Sociology*, 45, pp. 69–89. doi: 10.1146/annurev-soc-073117-041351.

Cutler, D., Deaton, A. and Lleras-Muney, A. (2006) 'The determinants of mortality', *Journal of Economic Perspectives*, 20(3), pp. 97–120. doi: 10.1257/jep.20.3.97.

Department of Health | Tier 1—Life expectancy and wellbeing—1.19 Life expectancy at birth (no date). Available at: https://www1.health.gov.au/internet/publications/publishing.nsf/Content/oatsih-hpf-2012-toc~tier1~life-exp-wellb~119 (Accessed: 26 June 2021).

Dicker, D. *et al.* (2018) 'Global, regional, and national age-sex-specific mortality and life expectancy, 1950-2017: A systematic analysis for the Global Burden of Disease Study 2017', *The Lancet*, 392(10159), pp. 1684–1735. doi: 10.1016/S0140-6736(18)31891-9.

Dong, X., Milholland, B. and Vijg, J. (2016) 'Evidence for a limit to human lifespan', *Nature Publishing*

Group, 538. doi: 10.1038/nature19793.

Girum, T., Muktar, E. and Shegaze, M. (2018) 'Determinants of life expectancy in low and medium human development index countries', *Medical Studies*, 34(3), pp. 218–225. doi: 10.5114/ms.2018.78685.

Heikkinen, E. *et al.* (2016) 'Survival and its predictors from age 75 to 85 in men and women belonging to cohorts with marked survival differences to age 75: a comparative study in three Nordic populations', *Aging Clinical and Experimental Research*, 28(3), pp. 541–550. doi: 10.1007/s40520-015-0418-0.

Herzer, D. (2017) 'Life expectancy and human capital: long-run evidence from a panel of countries', *Applied Economics Letters*, 24(17), pp. 1247–1250. doi: 10.1080/13504851.2016.1270405.

Klenk, J. *et al.* (2007) 'Increasing life expectancy in Germany: Quantitative contributions from changes in age- and disease-specific mortality', *European Journal of Public Health*, 17(6), pp. 587–592. doi: 10.1093/eurpub/ckm024.

KURAMOCHI, M. and KARYPIS, G. (2011) 'GENE CLASSIFICATION USING EXPRESSION PROFILES: A FEASIBILITY STUDY', *http://dx.doi.org/10.1142/S0218213005002302*, 14(4), pp. 641–660. doi: 10.1142/S0218213005002302.

Miladinov, G. (2020) 'Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries', *Genus*, 76(1), pp. 1–20. doi: 10.1186/s41118-019-0071-0.

Napolioni, V. *et al.* (2011) 'Age- and gender-specific epistasis between ADA and TNF-α influences human life-expectancy', *Cytokine*, 56(2), pp. 481–488. doi: 10.1016/j.cyto.2011.07.023.

Norman, K. and Klaus, S. (2020) 'Veganism, aging and longevity: New insight into old concepts', *Current Opinion in Clinical Nutrition and Metabolic Care*. Lippincott Williams and Wilkins, pp. 145–150. doi: 10.1097/MCO.0000000000000625.

Paoli, A. (2014) 'Ketogenic Diet for Obesity: Friend or Foe?', *International Journal of Environmental Research and Public Health*, 11(2), p. 2092. doi: 10.3390/IJERPH110202092.

Peltzman, S. (2009) 'Mortality inequality', *Journal of Economic Perspectives*, 23(4), pp. 175–190. doi: 10.1257/jep.23.4.175.

Pignolo, R. J. (2019) 'Exceptional Human Longevity', *Mayo Clin Proc*, 94(1), pp. 110–124. doi: 10.1016/j.mayocp.2018.10.005.

Preston, S. H. (2003) 'The changing relation between mortality and level of economic development. 1975.', *Bulletin of the World Health Organization*, 81(11), pp. 833–841. doi: 10.2307/2173509.

Pyrkov, T. V. *et al.* (2021) 'Longitudinal analysis of blood markers reveals progressive loss of resilience and predicts human lifespan limit', *Nature Communications*, 12(1), p. 2765. doi: 10.1038/s41467-021-23014-1.

Riley, J. C. (2005)' Estimates of regional and global life expectancy, 1800-2001', *Population and Development Review*, 31(3), pp. 537–543. doi: 10.1111/j.1728-4457.2005.00083.x.

Rootzén, H. and Zholud, D. (2017) 'Human life is unlimited-but short', 20, pp. 713–728. doi: 10.1007/s10687-017-0305-5.

Singh, P. N., Sabaté, J. and Fraser, G. E. (2003) 'Does low meat consumption increase life expectancy in humans?', in *American Journal of Clinical Nutrition*. American Society for Nutrition, pp. 526–558. doi: 10.1093/ajcn/78.3.526s.

Strulik, H. and Vollmer, S. (2013) 'Long-run trends of human aging and longevity', *Journal of Population Economics*, 26(4), pp. 1303–1323. doi: 10.1007/s00148-012-0459-z.

Vlatka Bilas, S. F. and M. B. (no date) 'Determinant factors of life expectancy at birth in the European union countries', *Collegium Antropologicum*.

Weber, D. D., Aminazdeh-Gohari, S. and Kofler, B. (2018) 'Ketogenic diet in cancer therapy', *Aging (Albany NY)*, 10(2), p. 164. doi: 10.18632/AGING.101382.

What affects an area's healthy life expectancy? - Office for National Statistics (no date). Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/articles/whataffectsanareashealthylifeexpectancy/2017-06-28 (Accessed: 26 June 2021).