School of Computing, Engineering and Digital Technologies

Department of Computing and Games

Teesside University

Middlesbrough TS1 3BA



**Factors Influencing Human Life Expectancy**

*Use as appropriate*:

An academic research paper for possible submission to <*name of your chosen conference or journal*> OR

Analysis, Design, and Implementation Report

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

**Date:**

**Mohana Kamanooru**

**A0223038**

**Supervisor**

**Dr Zia Ush Shamszaman**

**Teesside University**

Table of Contents

[Acknowledgement 2](#_Toc76767332)

[Abstract 3](#_Toc76767333)

[Introduction 4](#_Toc76767334)

[Research Question 5](#_Toc76767335)

[Literature Review 5](#_Toc76767336)

[Data Description 7](#_Toc76767337)

[Exploratory Analysis, Design, and Implementation 12](#_Toc76767338)

[Model Selection 15](#_Toc76767339)

[Evaluation 19](#_Toc76767340)

[Conclusion 20](#_Toc76767341)

[Legal Ethical and Professional Issues 25](#_Toc76767342)

[References 25](#_Toc76767343)

[Acronyms 27](#_Toc76767344)

[Table Of Figures 27](#_Toc76767345)

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 person may expect to live after birth. It is a crucial synthetic indicator for assessing a country's or region's economic and social development. 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 that we investigate 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 longevity and, if so, how we may increase life expectancy. From this research we could conclude that there is a substantial link between protein consumption and LE, according to the data. Protein is important 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 to a lack of data. Injuries, long- and short-term health issues, non-communicable diseases, and mental health all have a direct influence on death rates, making life expectancy inversely proportional, according to the available statistics. Several machine learning approaches, as well as the analyses and observations gathered throughout the research, were used to create machine learning models. As a result, we were able to create an 85 percent accurate machine learning model using the K Nearest Neighbor method.

# Introduction

The most important statistic for measuring population health is **Life Expectancy** (LE), which can be defined as the average count of years a person could live after birth. 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 widely 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, both 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)

Since the Age of Enlightenment, life expectancy has risen dramatically. Life expectancy began to rise in early industrialised countries in the early nineteenth century, but it remained low in the rest of the globe. As a result, there was a massive disparity in how health was spread across the world. In the developed regions, health is usually excellent, whereas, in the developing countries, health is consistently low. (Cutler, Deaton and Lleras-Muney, 2006). Global inequality has reduced in recent decades. The countries with the most incredible life expectancy in 1800 have the lowest life expectancy on the globe. Many countries that were once plagued by poor health are fast catching up. (Preston, 2003). The worldwide average life expectancy 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 a life expectancy of 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 hence 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, which includes employment, income, education, and financial well-being; the quality of the health system and people's ability to access it; health behaviours such as smoking, excessive alcohol consumption, poor nutrition, and lack of exercise, social, genetic, and environmental factors such as overcrowded housing, lack of clean drinking water and adequate sanitation. (*Department of Health | Tier 1—Life expectancy and wellbeing—1.19 Life expectancy at birth*, no date). In England from 2013 to 2015, new born males might expect to survive until they were 80 years old, while new born girls could expect to live until they were 83. On the other hand, these young people are expected to spend at least 20% of their life in bad health. On average, boys born between 2013 and 2015 would have 63 years of good health, while girls would have 64 years. (*What affects an area's healthy life expectancy? - Office for National Statistics*, no date)

# 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 are affecting longevity and, if so, how can we improve life expectancy.

## 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.

# 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 during an age 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 infectious diseases. Current population assessments of mortality trends 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). Life expectancy 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 end of the 17th century and the beginning of the 18th century 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 about 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 gap 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, which reduces per capita income, but this increases per capita income after the transition. In recent decades most countries have made progress towards a demographic transition.

Vaccination has advantages that go beyond the prevention of specific illnesses in people. They allow society and nations to reap a diverse and bountiful crop. Reducing global child mortality by allowing universal access to proven-effective safe vaccinations is a moral duty for the international community, since it is a human right for everyone to enjoy a better and richer life. (Andre, 2008) Vaccination is cost-effective and satisfies the requirement to care for society's most vulnerable people. The average causal effect of improvements in life expectancy on per capita income is likely to be positive for health innovations and mortality reductions if applied today. When calculating the economic benefits of increased life expectancy, the causal effect of increased life expectancy 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 dynamics at various stages of the demographic transition 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 from death in old age. (Peltzman, 2009)

In the aftermath of the Enlightenment, mortality in England began to decline, both directly through the application of innovative ideas about personal health and public administration to health and indirectly through increased productivity, which allowed for higher standards of living, better nutrition, better housing, and better 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 new pharmaceuticals have significantly impacted lowered cardiovascular disease 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 conducted 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. To test the hypothesis that the adenosine deaminase (ADA) and tumour necrosis factor-alpha (TNF-a) genes may impact human life expectancy, we looked at the significant effects of single loci and multi-locus interactions. (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)

# Data Description



## Identifying components affecting Life Expectancy

From all the above literature review, any key features have been identified to answer the current research question. 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**.

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 play a significant role in determining 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.

Diagram, schematic

Description automatically generated

Figure 1 Factors affecting Life expectancy

|  |  |  |
| --- | --- | --- |
| 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

## 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 research would look at aspects such as vaccination, mortality, economics, social factors, and other health-related issues. Because the observations in this dataset come from many countries, it shall be possible for a country to identify the predictive factor contributing to a decreased life expectancy value.

The initiative is reliant on data accuracy. The World Health Organization's ([WHO](https://www.who.int/)) Global Health Observatory ([GHO](https://www.who.int/data/gho)) data repository keeps track of health status and any other relevant parameters for all nations. The data sets are made accessible for the aim of analysing health data. Also, additional data has been collected from [Our World in Data](https://ourworldindata.org/life-expectancy), [OECD](https://www.oecd.org/) and [Data world Bank](https://data.worldbank.org/). These are publicly 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](https://ourworldindata.org/grapher/per-capita-meat-type?country=CHN~USA~IND~ARG~PRT~ETH~JPN~GBR~BRA) |
| 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 representative critical 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 Selected Features for research

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.

`Table

Description automatically generated

Figure 2 Preview of Consolidated data

## Dataset Description

In this research, we intend to collect the data until 2020, but the data has been available only until 2019, except for dietary consumption and BMI, available until 2017 and 2016. Hence, to analyse the accurate data without imputing or deleting the data, we have considered the data from 2000 to 2016, a span of 16 years. The collected raw data consists of 2856 rows and 30 columns. The overview of the dataset is presented below.

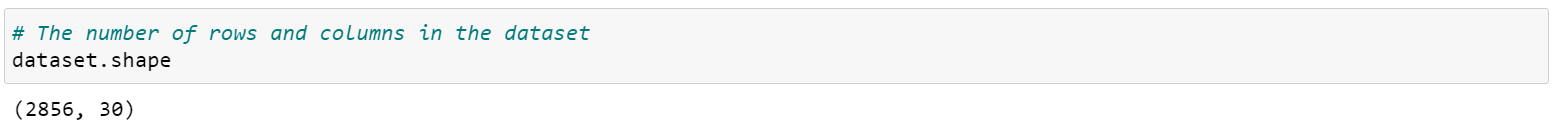
Table

Description automatically generated

Figure 3 Overview of Raw Data

# 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.



Table

Description automatically generated

Figure 4 Data size and data types information

## 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](https://www.who.int/), [GHO](https://www.who.int/data/gho), [Our World in Data](https://ourworldindata.org/life-expectancy), [OECD](https://www.oecd.org/) and [Data world Bank](https://data.worldbank.org/), 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.

Table

Description automatically generated

Chart

Description automatically generated with medium confidenceTable

Description automatically generated

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.

Table

Description automatically generated

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.

## Implementation

### 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.

### 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)

A picture containing text, screenshot

Description automatically generated A close-up of a page

Description automatically generated with low confidence

Figure 6 Correlation between features in the dataset

### 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.

Table

Description automatically generated

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.

# Model Selection

The research objective is now explicit, and the data needed to use machine learning techniques has been processed and is available. We used classification models to investigate the factors that influence longer life expectancy. Recognition, comprehending, and arranging concepts and objects into predetermined groups or "sub-populations" is the o of classification. 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.

## 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%.

Text

Description automatically generatedText, table

Description automatically generated with medium confidence

Figure 8 Logistic Regression Model

Text, letter

Description automatically generated

Figure 9 Logistic Regression Performance Metrics

## 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.

Text

Description automatically generated Text

Description automatically generated

Figure 10 Gaussian and Bernoulli Naive Bayes Models

Text, letter

Description automatically generated

Figure 11 Gaussian Naive Bayes Performance Metrics

Text, letter

Description automatically generated

Figure 12 Bernoulli Naive Bayes Performance Metrics

## K-Nearest Neighbors

KNN is a simple yet effective machine learning algorithm. Unlike Naïve Bayes Algorithm, KNN is 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 few nearest neighbours to determine the target value. This model resulted in accuracy of 85% as pictured below.

Text

Description automatically generated

Figure 13 K Nearest Neighbor Model

Text

Description automatically generated with medium confidence Text, letter

Description automatically generated

Figure 14 KNN Performance Metrics

## 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.

Graphical user interface, text, application, email

Description automatically generated

Figure 15 Support Vector Machine Model

Text

Description automatically generated with medium confidence

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 and the respective metric values are calculated for all the built models, and their accuracies are as below.

Text, letter

Description automatically generated Text

Description automatically generated

Text

Description automatically generated Text, letter

Description automatically generated

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%

# 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

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. 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)

# Conclusion

## 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%), Ovotransferrin (12%) and Ovomucoid (11%) per 100g, Beef / Bovine meat products contain myofibrillar proteins (50–55%, mostly myosin and actin), sarcoplasmic proteins (30–34%, primarily enzymes and myoglobin) and connective tissue (10–15%, mostly collagen and elastin fibres embedded in mucopolysaccharides). Hence, we could not draw any combined conclusion depending on the type of protein.

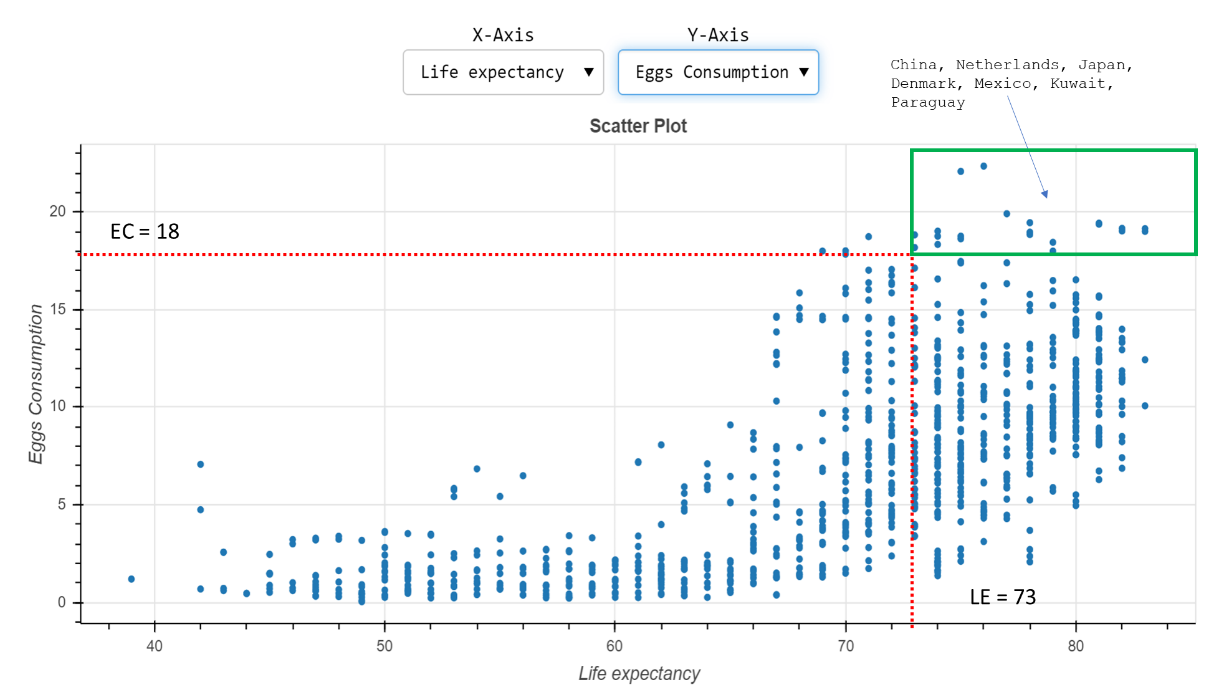




Figure 19 Egg Consumption Vs Life Expectancy

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**

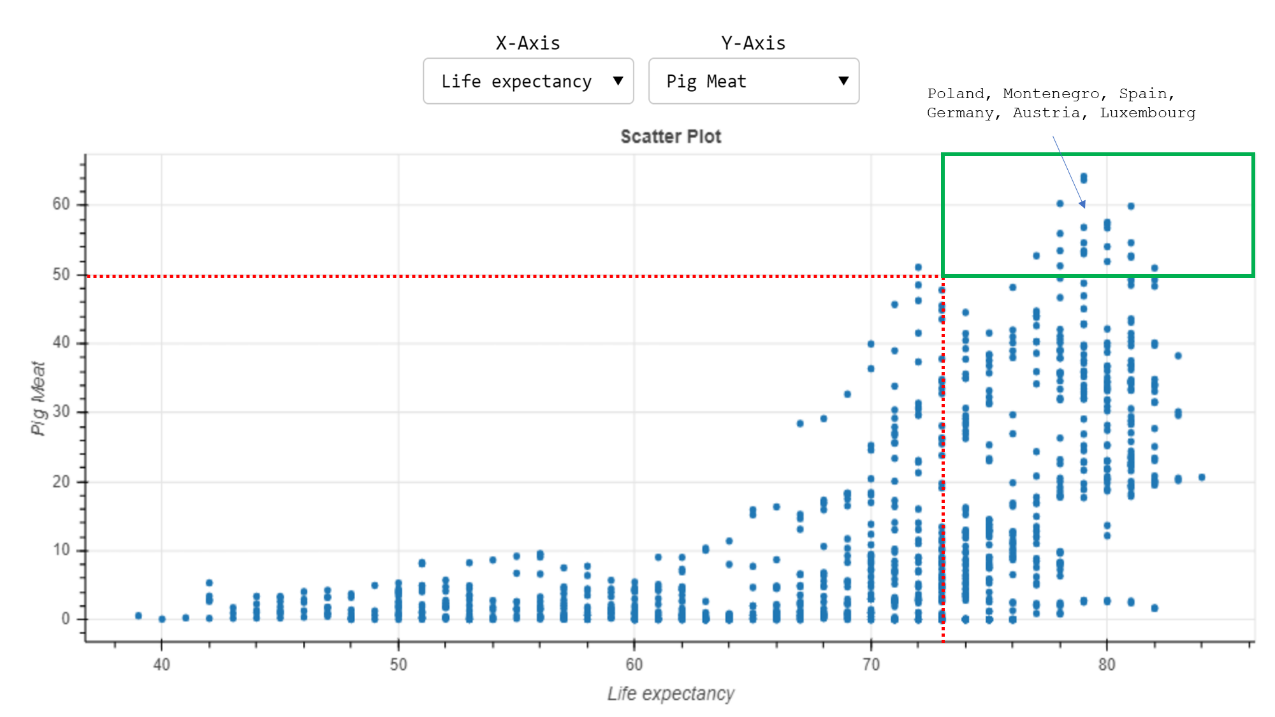


Figure 20 Pork / Pig Meat Consumption Vs Life Expectancy

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.

Chart

Description automatically generated

Figure 21 Beef Consumption Vs Life Expectancy

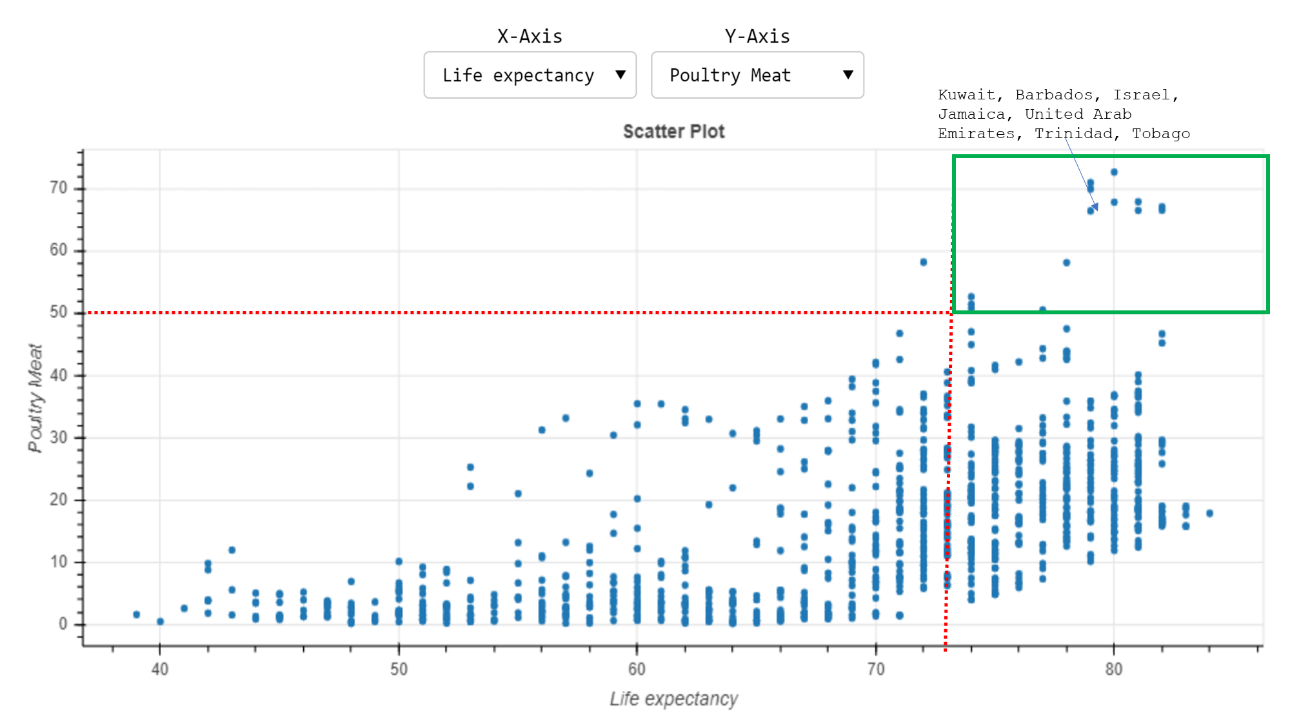


Figure 22 Poultry Meat Consumption Vs Life Expectancy

**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.

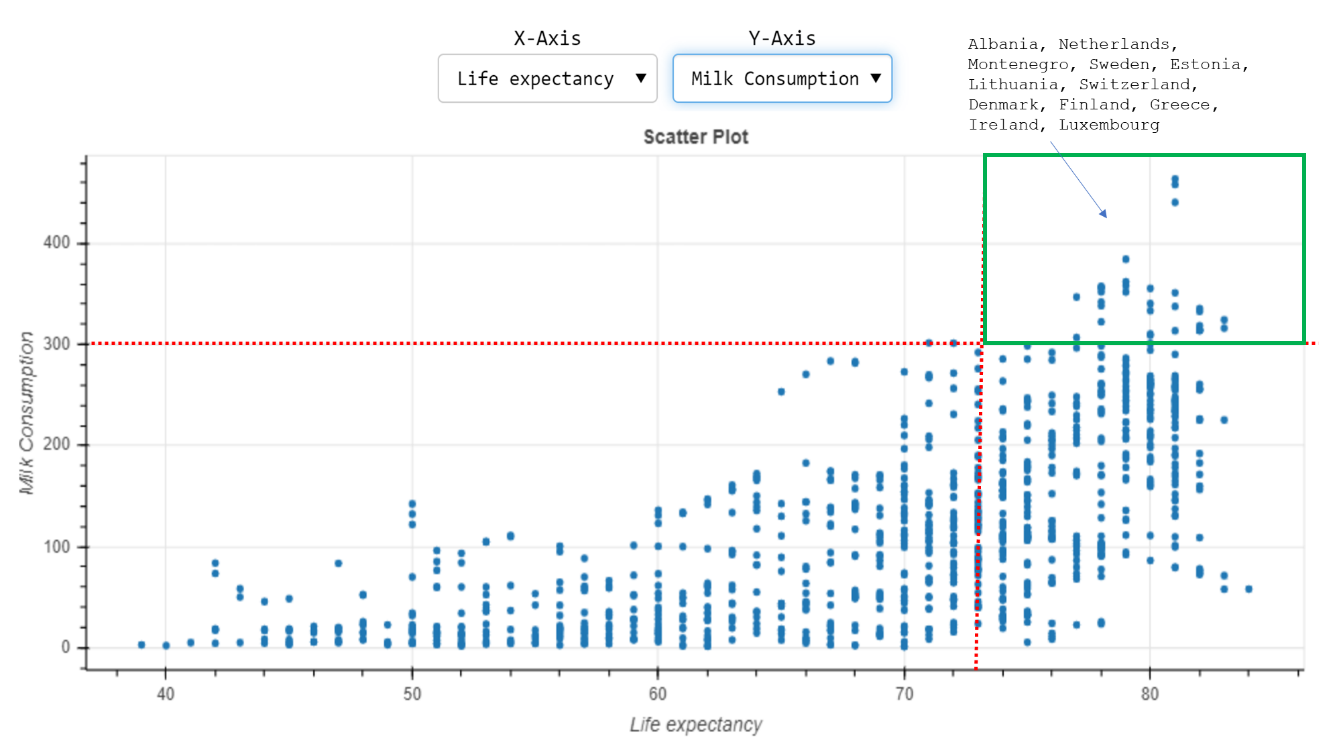


Figure 23 Milk Consumption Vs Life Expectancy

**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.

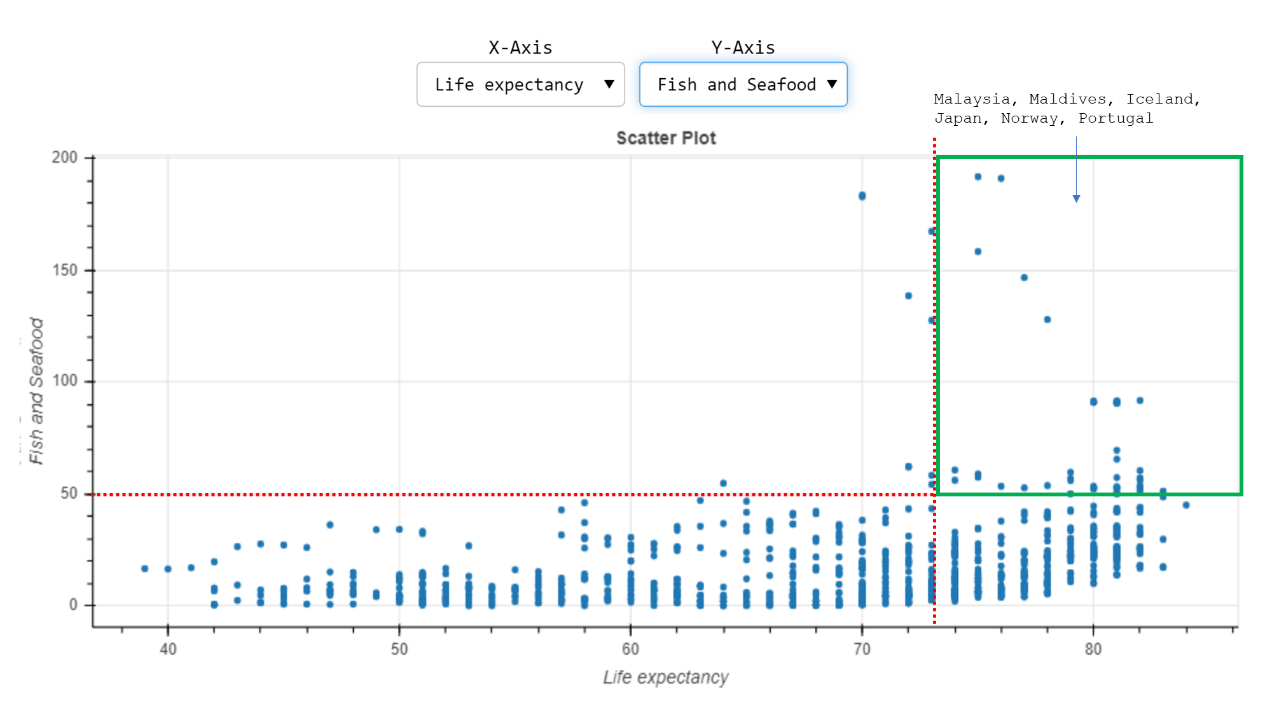


Figure 24 Fish and Seafood Consumption Vs Life Expectancy

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 health-care 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 noncommunicable 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.

Chart, bar chart

Description automatically generated



Figure 25 Correlation between Immunisation, Illnesses and Life Expectancy

## Conclusions

This research aimed to understand the factors impacting the life expectancy of a human. How could protein consumption affect the life span? 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 by using several machine learning methods and the analyses and observations collected throughout the research. To assess the model's performance, we calculated Mean Average Error, Mean Squared Error, Root Mean Squared Error, R2 Score, and F1 Score. 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%.

## Future Work

1. Non-availability of data until 2021 and for other vital factors like vegetarian consumption,
2. No dataset for the keto diet
3. Immunisation and vaccination information and time constrain working on the project
4. Mental health and happiness

# Legal Ethical and Professional Issues

The research data for this project has been gathered from open source ([WHO](https://www.who.int/), [GHO](https://www.who.int/data/gho), [Our World in Data](https://ourworldindata.org/life-expectancy), [OECD](https://www.oecd.org/) and [Data world Bank](https://data.worldbank.org/)), 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.

# 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.

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*.

*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).

# 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 |

# Table Of Figures

[Figure 1 Factors affecting Life expectancy 8](#_Toc76764846)

[Figure 2 Preview of Consolidated data 11](#_Toc76764847)

[Figure 3 Overview of Raw Data 11](#_Toc76764848)

[Figure 4 Data size and data types information 12](#_Toc76764849)

[Figure 5 Overview of data after pre-processing 13](#_Toc76764850)

[Figure 6 Correlation between features in the dataset 14](#_Toc76764851)

[Figure 7 Data preview before feature scaling 14](#_Toc76764852)

[Figure 8 Logistic Regression Model 15](#_Toc76764853)

[Figure 9 Logistic Regression Performance Metrics 15](#_Toc76764854)

[Figure 10 Gaussian and Bernoulli Naive Bayes Models 16](#_Toc76764855)

[Figure 11 Gaussian Naive Bayes Performance Metrics 16](#_Toc76764856)

[Figure 12 Bernoulli Naive Bayes Performance Metrics 17](#_Toc76764857)

[Figure 13 K Nearest Neighbor Model 17](#_Toc76764858)

[Figure 14 KNN Performance Metrics 18](#_Toc76764859)

[Figure 15 Support Vector Machine Model 18](#_Toc76764860)

[Figure 16 SVM Highest accuracy with different Kernels 19](#_Toc76764861)

[Figure 17 SVM Performance Metrics for different Kernels 19](#_Toc76764862)

[Figure 18 Performance Metrics for various developed Models 20](#_Toc76764863)

[Figure 19 Egg Consumption Vs Life Expectancy 21](#_Toc76764864)

[Figure 20 Pork / Pig Meat Consumption Vs Life Expectancy 21](#_Toc76764865)

[Figure 21 Beef Consumption Vs Life Expectancy 22](#_Toc76764866)

[Figure 22 Poultry Meat Consumption Vs Life Expectancy 22](#_Toc76764867)

[Figure 23 Milk Consumption Vs Life Expectancy 23](#_Toc76764868)

[Figure 24 Fish and Seafood Consumption Vs Life Expectancy 23](#_Toc76764869)

[Figure 25 Correlation between Immunisation, Illnesses and Life Expectancy 24](#_Toc76764870)

[Table 1 Factors affecting Life expectancy 8](#_Toc76764871)

[Table 2 Data collected from GHO 9](#_Toc76764872)

[Table 3 Data collected from Oue World in Data, OECD, Data World Bank 10](#_Toc76764873)

[Table 4 Selected Features for research 10](#_Toc76764874)