**Forecasting life expectancy through machine learning model**

**PROJECT REPORT**

**JULY 2021**

Estimator of the Life Expectancy (WEB APP)

Renzo Bejarano Varela | Arron Nguyen | Callum Linnegan | Carmen Jiaman Li | Tealiie Le | Nicklaus Ng

CONTEXT:

One of the most used measure of a population’s overall health is through its life expectancy. By examining causes, patterns, and trends in death, it is possible to help explain differences and changes in the health of a population. Therefore, life expectancy data is important as it contributes to the evaluation of health strategies, planning and policymaking.

As expected, there has been a multitude of data regarding factors affecting life expectancy. These data usually only considered demographic variables, income composition and mortality rates.

OBJECTIVE:

Using our group’s diverse field of specialisations, we have sought to collate a more rounded database which considers 19 social, economic, mortality and health-related indicators in which we considered to be essential for the calculation of life expectancy.

Furthermore, using this updated dataset, we will create a machine learning model in an attempt to forecast a more accurate prediction of life expectancy based on these indicators.

3.ETL

3.1DATA SOURCES

& Extraction, Transformation, Loading (ELT) PROCESS:

To supplement the dataset with social, economic, mortality and health-related indicators, we have gathered additional data from the World bank Organisation. <https://data.worldbank.org/>

**(Steps by step process on joining data)**

THE DATABASE:

3.2DATA WRANGLING

The original dataset was very tall, with one single row per country-year-indicator combination. For machine learning, the dataframe needs to be reshaped to have features (development indicators) as column headers and the countries as the rows. As a result, the main data wrangling steps were unstacking the data to structure it in a workable state.  
  
While the datasets are fairly clean, there are missing data - for several years for some metrics, and large gaps for specific counties.

3.3

4. MACHINE LEARNING MODEL /APP:

We’ll use linear regression / Gradient boosting to predict life expectancy (a continuous variable) from a number of independent variables (development indicators).  
  
Linear regression takes the form y i = β0 + β 1X 1 + ... + β iX i where y is the dependent variable (in this case, Life Expectancy) which we will predict X are the regressors (development indicators) β is the intercept, and are the regression coefficients.

5.POTENTIAL LIMITATIONS:

-time

-more models

THE SOCIAL, ECONOMIC, MORTALITY AND HEALTH-RELATED INDICATORS:

1. Life expectancy at birth, total (years)

Average number of years a newborn is expected to live if mortality patterns at the time of birth remains constant in the future.

High mortality in younger age groups significantly lowers the life expectancy at birth. However, if a person survives a childhood of high mortality, they might live much longer than the expected age.

1. Access to electricity (% of population)

Energy is necessary for creating the conditions for economic growth.

Access to electricity is particularly crucial to human development as electricity is indispensable for certain basic activities, such as lighting, refrigeration, and the running of medical appliances.

Energy is important for improving a populations standard of living. But electricity generation also can be detrimental to the environment and people’s health. The damage depends largely on how electricity is generated. For example, burning coal releases twice as much carbon dioxide - a major contributor to global warming and pollution.

1. Cause of death, by non-communicable diseases (% of total)

Data on cause of death is compiled by the WHO, based mainly on data from the International Statistical Classification of Diseases and Related Health Problems, 10th revision. The data has been carefully analysed to consider incomplete coverage of vital registration and the likely differences in cause of death patterns that would be expected in under covered and often poorer subpopulations. It also considers cardiovascular diseases, cancer, injuries, and general ill-defined categories.

1. Current health expenditure (% of GDP)

Current Health Expenditure describes the share of spending on health in each country relative to the size of its economy. It includes expenditures corresponding to the final consumption of health care goods and services and excludes investment, exports, and intermediate consumption.

A higher health expenditure generally means a more developed economy and a heathier population.

1. Diabetes prevalence (% of population ages 20 to 79)

Long-term effects of diabetes include damage to large and small blood vessels, which leads to heart attack, strokes, and problems with vital organs.

The prevalence of diabetes within a population is an indication of the general health within the population. A high prevalence in diabetes might see a decrease in life expectancy, especially within less developed countries, with population without proper access to healthcare.

1. Government expenditure on education, total (% of GDP)

Government expenditure has an indirect effect on health levels. A higher Government expenditure on education was linked to higher completion rates and lower levels of poverty.

This generally translates into more affluent and healthier communities, due to their access to health care services.

1. Hospital beds (per 1,000 people)

The amount of available hospital beds serves as a measure for a populations access to inpatient services. It should only be used as a general indication, as inpatient services required for individual countries depends on several factors - such as demographic issues and the burden of diseases. Therefore, 2 beds per 1000 people may be sufficient for one country, however the same figure may be inadequate for another country.

1. Immunisation, HepB3 (% of one-year-old children)

The percentage of one-year-olds who have received three doses of hepatitis B vaccine each year. Immunization is an essential component for reducing under-five mortality. Immunization coverage estimates are used to monitor coverage of immunization services and to guide disease eradication and elimination efforts. It is a good indicator of health system performance.

1. Immunization, measles (% of children ages 12-23 months)

Immunization against measles is an essential component for reducing under-five mortality. Immunization coverage estimates are used to monitor coverage of immunization services and to guide disease eradication and elimination efforts. It is a good indicator of health system performance.

It is estimated that between 2000 – 2018, measles vaccinations prevented 23.2 million deaths. As a highly contagious disease, measles was the cause of 140,000 deaths of mostly children under the age of five in 2018 alone.

1. Inflation, consumer prices (annual %)

Inflation measures the increase in prices of goods and services. As inflation affects the cost of foods such as grains, eggs, fish, and meat. It becomes an important indicator for the ability of the population to afford nutritious food needed for a healthy balance diet.

Access to food, especially those needed for a healthy balance diet will inevitably lead to healthier lifestyle and a greater life expectancy.

1. Mobile cellular subscriptions (per 100 people)

Access to mobile cellular subscriptions includes both prepaid and post-paid phone services.

Cellular subscriptions are used as an indicator for development and a nations quality of economic infrastructure. Communication is seen as an important tool for both foreign and domestic investors. Therefore, a higher level of Mobile cellular subscription, should therefore translate into a more developed economy.

1. Mortality caused by road traffic injury (per 100,000 population)

Road traffic injuries and deaths are a major global health problem. Traffic crashes are currently the leading cause of death for children and young adults in the world. There is a strong correlation between the risk of road traffic death and the income levels of a country.

It is disproportionately high among low- and middle-income countries. This could be due to a government’s inability to provide safe road infrastructure, or a lack of access for the population to afford safe, modern cars.

1. Mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene (per 100,000 population)

Unsafe drinking water, sanitation and lack of hygiene are important causes of death. Most cases of diarrheal deaths in the world are a result of unsafe water, sanitation, and hygiene. These diarrheal diseases could be easily prevented if adequate sanitation and clean water standards are provided.

Therefore, high mortality rates caused by a lack of hygiene is a strong indicator of poverty and low-socioeconomic statuses.

1. Number of infant deaths

In 2019 alone, 7.4 million children, youths and adolescents died mostly of preventable or treatable diseases.

Early childhood mortality is synonymous with developing countries due to the lack of access to food, clean water, and sanitation. Furthermore, populations that belong to developing countries experience more diseases due to the lack of vaccines, adequate medical treatment, and facilities.

1. Population, total

Drastic increases in population impacts the populations’ access to resources. Especially in developing countries, large populations places pressures on a country’s economy and its government’s ability to deliver vital infrastructures and services needed for higher life expectancy. Generally, with some exceptions, a country with a smaller population should expect a higher life expectancy in contrasts to larger populations.

1. Proportion of people living below 50 percent of median income (%)

The percentage of people in the population who live in households whose per capita income or consumption is below half of the median income or consumption per capita.

The higher the proportion of the population living under 50% of the median income, indicates a possibility of an uneven distribution of income within the population. A majority of the population living under the median income may highlight a significant proportion of the population’s inability to afford basic necessities such as food, water and essential medical supplies, which are essential for a higher standard of living.

1. Real interest rate (%)

Real interest rate is the lending interest rate adjusted for inflation.

Whilst a higher interest rate might indicate that disposable income is readily available for a large proportion of the population. A high interest rate could also indicate that there is too much money in the market. (1930’s Germany). However, a decrease in interest rate indicates a contraction within the economy. This would affect the populations’ ability to purchase essential goods and services associated with higher life expectancy.

1. Total alcohol consumption per capita (liters of pure alcohol, projected estimates, 15+ years of age)

Total alcohol per capita consumption is defined as the total amount of pure alcohol consumed per person over a calendar year.

Alcohol consumption is a causal factor of more than 200 diseases and injury conditions such as alcoholism, cancers, cardiovascular diseases as well as injuries from violence and road accidents.

Higher levels of alcohol consumption might indicate adverse societal issues and higher rates of alcohol related illnesses, which would impact life expectancy within a region.

1. Unemployment, total (% of total labour force) (modelled ILO estimate)

Unemployment refers to the share of the labour force that is without work but available for and seeking employment.

High and sustained unemployment indicates serious inefficiencies in resource allocation. A high and sustained unemployment rate, especially in developing countries would most likely hinder a large proportion of the population’s ability to afford necessities such as food, water, and medical supplies, adversely affecting life expectancy rate.

FINDINGS:

In this case we have sought to collate a more rounded dataset(from 1999-2019) which considers 19 social, economic, mortality and health-related indicators in which we considered to be essential for the calculation of life expectancy