



42048 Studio 3: Innovation - Autumn 2023
Assessment task 3: Research paper

LIFE EXPECTANCY RESEARCH FOR THE DEVELOPMENT & IMPROVEMENT OF MEDICAL PACKAGES

By: Group 4

Kristen Nguyen – 13815105
Huong Ly Nguyen – 14232286
Yuwen Sheng – 14224582
Kaushal Kumar Gandla – 14200923
Mingrui Song – 14202473
Chentao Hu – 11589254

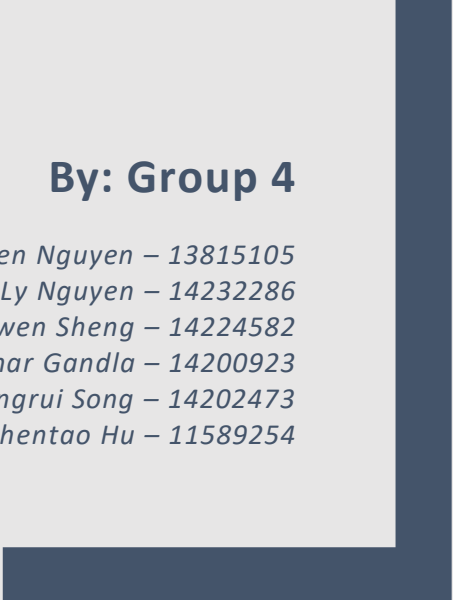


Table of Contents

INTRODUCTION	2
BUSINESS CONTEXT	2
PROJECT ORIGINS AND MOTIVATION	2
PROJECT CONTRIBUTIONS.....	3
HOW THE PAPER IS ORGANISED	3
SCOPING THE PROJECT	4
PROJECT AIMS.....	4
PROJECT OBJECTIVES & DELIVERABLES	4
STAKEHOLDER'S ANALYSIS.....	5
SUCCESS CRITERIA.....	6
RECOMMENDATIONS FOR DIFFERENT SECTORS.....	7
METHODOLOGY	9
METHODOLOGY FRAMEWORK	9
DATA CLEANING.....	9
DATA UNDERSTANDING (EDA).....	13
MODEL SELECTION	18
EXPERIMENTS	19
DATA PARTITIONING	19
DECISION TREE	19
<i>Modelling and performance measurement - Decision Tree</i>	19
RANDOM FOREST	21
<i>Modelling and performance measurement - Random Forest</i>	21
LINEAR REGRESSION.....	23
<i>Modelling and performance measurement - Linear Forest</i>	23
MODEL COMPARISON	24
MODEL TUNING.....	25
DISCUSSION AND CONCLUSIONS (LY)	26
DISCUSSION.....	27
LIMITATIONS	27
FUTURE RECOMMENDATIONS	27
REFERENCE	29
APPENDIX	30
APPENDIX 1: ATTRIBUTE DESCRIPTION.....	30
APPENDIX 2: CONTRIBUTION TABLE	31
APPENDIX 3: CODING LINK	31

Introduction

Business context

In the digital 21st century world, the analysis of data is becoming paramount in the functions of everyday life. It has an innate quality to tell a story and present key insights to stakeholders, allowing them to make informed decisions at all levels of business and government. As datasets become larger due to increasing technological adoption by the population, deeper analysis and insights can thus be extracted.

Our firm performed modelling of life expectancy and health data provided by the World Health Organisation (WHO) and generated key metrics, focusing on the top factors that reduce life expectancy. This is a fundamental piece of information that will be priceless for the healthcare industry, insurance industry, as well as businesses and society in general. By giving this data to healthcare providers, they can tailor their resource allocation to align with the most common causes of early death, as well as focus on preventative measures for patients such as advanced screening. Insurance companies will also be a core consumer of this data analysis, as they need to use the insights for their actuarial models, basing pricing off our life expectancy data. Insurance companies need to know which factors contribute the most to lower life expectancy, so that they can alter their policy pricing accordingly. Finally, the last stakeholder group we are focusing on are Businesses (employers), as several employers offer health packages as part of their employee benefits program.

By understanding the leading causes of lower life expectancy as well as the most common afflictions, employers can thus model their healthcare packages to align with the data to optimise costs and provide the most benefit. Ultimately, every dataset has a story and a purpose. Our project aims to uncover the story behind individual health and empower our stakeholders through machine learning and big data.

Project origins and motivation

The origins of our project lie in the recognition that data analytics is crucial to successful patient engagement strategies and can inform how marketers engage with individuals and specific groups of patients. We sought to empower healthcare providers, insurance companies, and businesses with data that is useful in tailoring their resource allocation, actuarial models, and health packages.

Our motivation was to utilise machine learning and big data to uncover the story behind individual health, ultimately optimising costs and improving the overall quality of life. Through our research, we aimed to enable stakeholders to use data to align their policies with the most common causes of early death and focus on preventative measures. By understanding the leading causes of lower life expectancy and common afflictions, healthcare providers can tailor their resource allocation to align with the most common causes of early death, and insurance companies can adjust their policy pricing accordingly. Employers can also model their healthcare packages to align with the data to optimise costs and provide the most

benefit. Our study findings demonstrate the value of data analytics in healthcare and its potential to revolutionise the industry.

Project contributions

The purpose of this project is to provide the hospital product department with insights into the features that affect life expectancy. By understanding these features, the department can tailor health check packs for people from different areas to identify issues that impact life expectancy. Ultimately, this project aims to enhance life expectancy.

To gain insight into life expectancy, this project will train a model to predict life expectancy based on various factors. Hospitals can then use the model to make predictions for life expectancy in their local area. By using the trained model. A featured importance can be provided. For medical-related features such as HIV, hepatitis, thinness, and vaccination, hospitals can design health check packs accordingly. For non-medical features such as adult mortality and infant death, hospitals can work with the government to identify the root causes and come up with suggestions to enhance them. Economic factors such as GDP, schooling, and income composition may be challenging to change. Hospitals may consider these factors when designing their health check package and make life expectancy prediction.

How the paper is organised

This paper is organized into five parts, each corresponding to a phase in the project lifecycle. The structure of the paper is outlined as follows:

Structure	CRISP-DM stage
Introduction	Business context
	Main problem to solve
	Motivation
	Contributions
Scoping the project	Project aims
	Objectives and deliverables
	Stakeholder analysis
	Success criteria
Methodology	Methodology framework
	Data preparation
	Model selection
Experiments	Data partitioning
	Modelling and performance measurement
	Model comparison
	Model tuning
Discussion and Conclusion	Deployment

Table 1: Paper Structure

Scoping the project

Project aims

In order to improve the healthcare standards as well as business performance, it is crucial to understand what factors affect life expectancy and how they affect it. There are a multitude of factors that affect life expectancy in the form of illnesses or vices, so it is paramount to understand the impacts of each.

Our aim in this project is to leverage the power of machine learning to extract insights regarding the factors that affect life expectancy and develop a model framework for clients such as hospitals, insurance agencies, and businesses, allowing them to optimize their internal metrics based on the data insights provided. By packaging the analysis in a machine learning model, we can produce a portable, repeatable, and flexible tool that can be used in real-world settings to assist the prediction process.

Project Objectives & Deliverables

Key objectives required to reach the delivery milestone and satisfy the project aims. These are:

- Develop predictor models using Machine Learning (ML) algorithms to study factors affecting an individual's life expectancy.
- Evaluate predictor models based on results and select the best one according to project aims.
- Conduct cost benefit analysis to determine cost effectiveness of implementation.
- Deploy and incorporate the selected machine learning model in real-world setting to assist with prediction processes.

The project deliverables will integrate the defined research and technical objectives into four outputs to address the business problem, listed as below:

- **Proposal document:** A formal document that lists the involved stakeholders, methodology applied in the research and the timelines for project delivery.
- **Exploratory Data Analysis (EDA):** The EDA will involve summarising the key statistics, trends, and interpretation of patterns between life expectancy and customer demographics and factors that affect life expectancy.
- **Python Prototype:** The prediction model will be implemented and evaluated using Python. The prototype will involve data exploration, pre-processing, feature engineering, and the application of three machine learning algorithms to develop a predictive model.
- **Research document:** The showcase of the prediction's results in Python along with discussions of the research insights and recommendations for further research.

Stakeholder's analysis

A healthcare project involves six main stakeholders, which are patients, healthcare providers, hospital administrators, insurance companies, public health authorities, and regulatory bodies.

- Patients are the primary beneficiaries of healthcare projects, and they rely on healthcare providers to diagnose and treat their medical conditions.
- Healthcare providers including physicians, nurses, and other medical professionals. They are responsible for providing medical care to patients. They must adhere to ethical and professional standards and maintain patient confidentiality while ensuring that patients receive high-quality care and follow-up care.
- Hospital administrators are responsible for managing and overseeing the operations of hospitals, clinics, and other healthcare facilities. They ensure that these facilities are adequately staffed, equipped, and funded to provide high-quality care. They also work to improve patient safety, patient satisfaction, and the overall efficiency of healthcare operations.
- Insurance companies provide coverage for healthcare services to individuals and employers. They negotiate contracts with healthcare providers to ensure that their customers have access to medical care at a reasonable cost. Insurance companies also work to manage healthcare costs by encouraging preventive care, managing chronic conditions, and reducing unnecessary medical procedures.
- Public health authorities are government agencies responsible for promoting and protecting the health of the public. They work to prevent the spread of infectious diseases, promote healthy behaviours, and monitor public health trends. They also provide education and resources to individuals and communities to help them make informed decisions about their health.
- Regulatory bodies are government agencies responsible for ensuring that healthcare providers and facilities comply with regulations and laws. They monitor healthcare quality, safety, and ethics. They also license and certify healthcare professionals and facilities, investigate complaints and incidents of malpractice, and enforce laws and regulations related to healthcare. They work to protect patients from harm and ensure that healthcare services are delivered safely and effectively.

Table 2 below shows how will each stakeholder be involved in the project.

Stakeholder	Stake/ Interest	Influence	Importance	Contribution	Information Needs	Strategy for Engagement
Patients	Quality care, affordability, accessibility, privacy, and security of health information	Low to Medium	High	Feedback, participation in research, adherence to treatment plans	Treatment options, cost of care, health information privacy and security, insurance coverage	Empowerment through education and information, patient-centred approach, patient feedback mechanisms
Healthcare Providers	Ethical and professional standards, patient	High	High	Adherence to standards, collaboration	Patient medical history, treatment plans, clinical guidelines	Training and education, recognition of achievements, open

	satisfaction, high-quality care			with other providers		communication channels
Hospital Administrators	Efficient hospital operations, patient safety, patient satisfaction	High	High	Management of budgets, policies and procedures development, resource allocation	Performance metrics, patient satisfaction surveys, financial reports	Open communication, collaboration with staff and patients, transparency in decision-making
Insurance Companies	Reasonable healthcare costs, preventive care, chronic condition management	High	Medium to High	Contract negotiation with healthcare providers, management of healthcare costs	Health plan coverage, claim processing, cost-sharing	Collaboration with healthcare providers, promotion of preventive care, transparency in cost-sharing
Public Health Authorities	Public health promotion, disease prevention, healthy behaviours	High	High	Health education, research, policy development, community engagement	Health statistics, disease surveillance, research findings	Collaboration with healthcare providers, promotion of healthy behaviours, participation in public health campaigns
Regulatory Bodies	Healthcare regulation and compliance, patient safety	High	High	Licensing and certification of healthcare providers and facilities, investigation of malpractice claims	Compliance with regulations and laws, adherence to ethical standards	Clear communication of regulations and guidelines, cooperation with healthcare providers and facilities, transparency in regulatory processes

Table 2: Stakeholder's involvement in the project

Success Criteria

- **Mean Squared Error:** The average difference between predicted values and actual values of the model generated should be lower as possible to indicate a good model.
- **R-Squared score:** To indicate the best fit of the model, the score should be closer to One, as the proportion of attributes in the life expectancy data can be explained by the generated model.
- **Accuracy:** As for a classification model where the model is predicting life expectancy, the accuracy rate should be 95% or more.
- **Important features:** Identifying the key features from the data, which would impact the model for life expectancy as these features are important to help the machine learning model to make predictions.

Recommendations for different sectors

Recent research, for instance, applied algorithms that use machine learning to forecast the risk of someone dying over a five-year period based on a range of parameters such as gender, age, smoking habits, and BMI (Badea et al., 2017). Researchers discovered that machine learning programs predicted mortality risk more accurately than conventional statistical techniques, highlighting the promise of new approaches to enhancing the quality of healthcare.

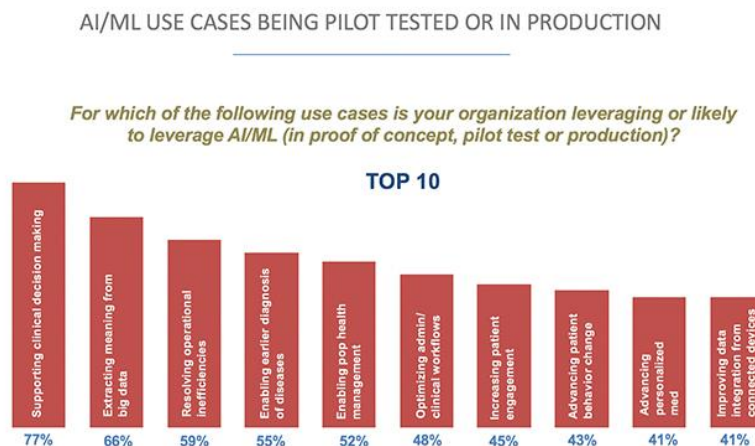


Figure 1: Top 10 places where ML is used in health sectors

Machine learning techniques will be used to enhance health care plans in addition to forecasting life expectancy. The algorithms used can predict which therapies are most successful for sorts of illnesses by analysing data on previous treatments and outcomes, enabling health plans to be customized to individual requirements.

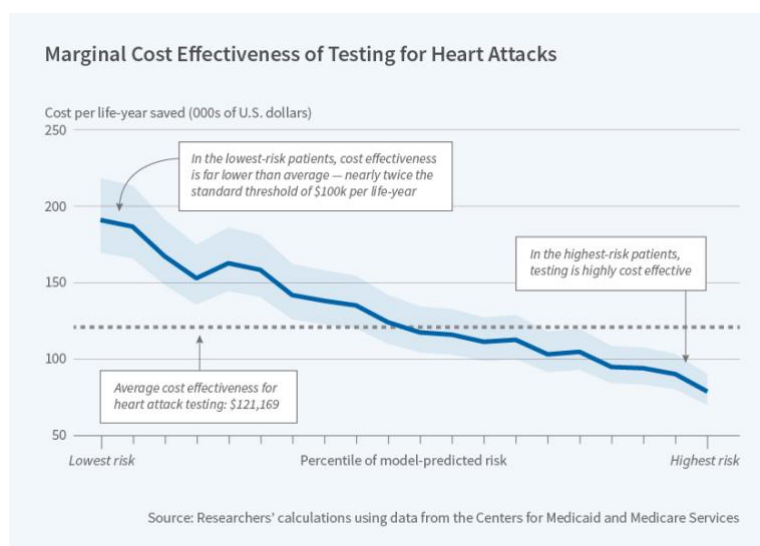


Figure 2: Reduction cost for testing

Economic status, as well as income, are factors related to health, connected life expectancy, the standard of life, and health concerns, with many diseases being more frequent among low socioeconomic status persons. Individuals from low-income families are also more vulnerable

to the impacts of unhealthy lifestyles. Income inequality and health imbalance are inextricably related ("Income, health, and social welfare policies," 2020). However, with the help of machine learning algorithms, we can provide everyone the cost-effective and more reliable treatment plans which will ultimately improve economic growth.

The insurance sector may use machine learning algorithms to analyse massive amounts of existing data and assist underwriters in focusing on the most valued clients. The result will allow corporations to estimate the sorts of insurance and policy plans that new consumers will purchase, as well as the number of false requests for reimbursement submissions.

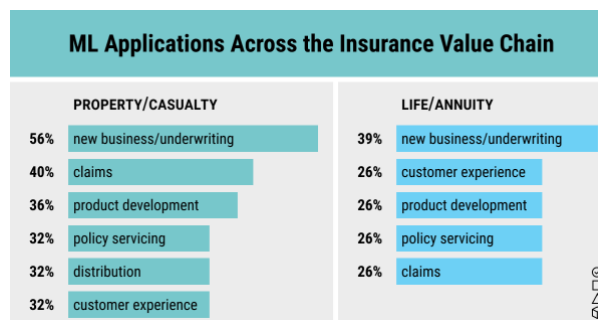


Figure 3: Insurance Value chain

According to McKinsey, automation will affect twenty-five percent of the insurance industry by 2025. Many aspects of the business, from handling claims to policy termination, may be automated. ML algorithms are quite useful when it deals with automation (Daivi, 2023).

Recently many companies started using ML to improve their product's efficiency and methods of testing, as Algorithms based on machine learning finish years' worth of work in a matter of seconds. Companies like PFIZER, JANSSEN PHARMACEUTICA, SANOFI, NOVARTIS, and BAYER, which are major players in their industry, have started using ML for better outcomes. Moreover, many start-ups have implemented the same strategy to improve Life expectancy, the main goal, and discover new medicines with the help of machine learning.

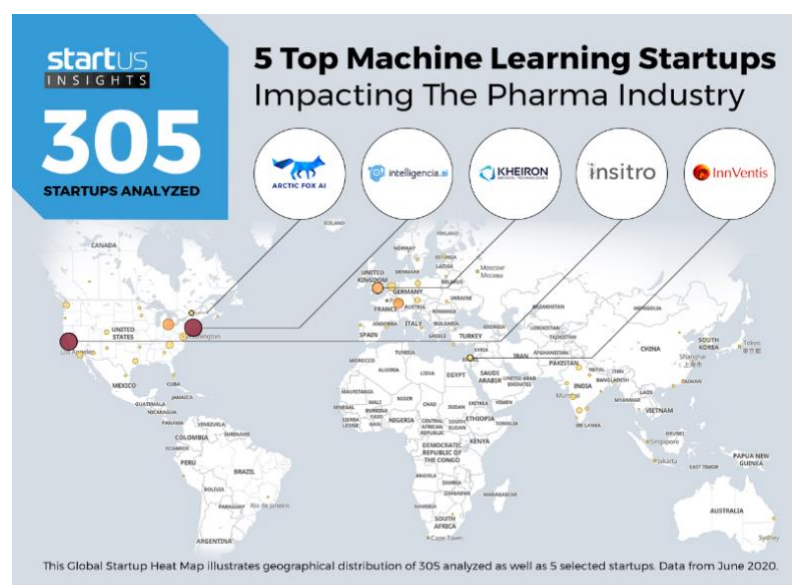


Figure 4: Initiative of Pharma companies to use ML

Pfizer established cooperation with Concreto HealthAI in 2019 to promote Precision Oncology research utilizing real-world data and Machine learning. "Pfizer believes real-world data has enormous potential to guide how we create and distribute medicines that will enhance patient outcomes," said Chris Boshoff, Ph.D., Chief Development Officer in Pfizer's Global Product Development division (Kantify, 2021).

Methodology

Methodology framework

In this project, we will be using the CRISP-DM methodology framework, which is a widely recognised and applied method for guiding data mining projects, including those in the healthcare industry. The acronym CRISP-DM consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Chapman et.al, 2000). This project utilises the dataset of life expectancy and health data taken from the Kaggle website. Additionally, in this paper, the Business Understanding phase identifies the research question and the intended outcomes of the project. In the Data Understanding section, the sources of data, their quality, and their appropriateness for the research question are explored. The Data Preparation phase involves cleaning and transforming the data for analysis. The Modelling phase entails building three models, including a decision tree, random forest, and linear regression, then assessing their accuracy. In the Evaluation phase, the models are evaluated and validated against a set of performance metrics. The CRISP-DM framework provides a structured and systematic approach to the projects, which helps ensure that the results are robust and useful to stakeholders.

Data cleaning

In the process of data cleaning, the initial step is to detect any missing data in the dataset. This can be accomplished by using various tools, one of which involves utilizing the `isnull().sum()` function and a heatmap. The result of both function is as showing in Figures 5 and 6 below.

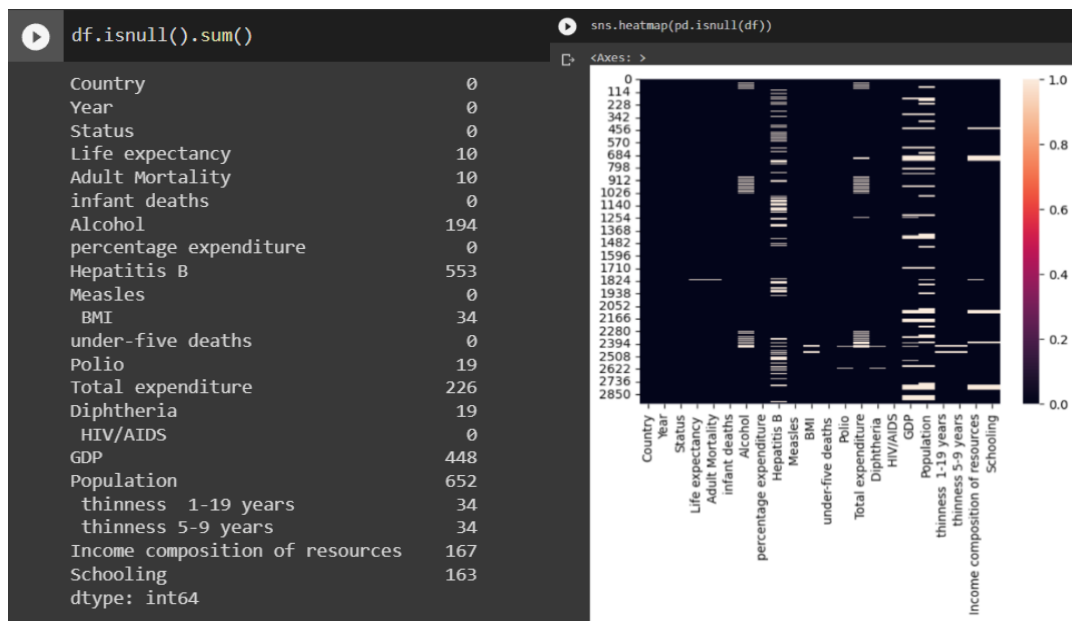


Figure 5-6: Codes & Result of null checking

From the result, there are plenty of null values in the dataset, some of them are continues. One clear example for this is the white block showing in the GDP column with in the heatmap.

The next step involves selecting a method for handling missing values, specifically the approach to be used to replace null values. This decision requires careful consideration as it can have significant implications for the accuracy and validity of subsequent analyses. Based on the result of EDA, there is a clear trend or pattern for each country, therefore, interpolation of time series data by country can be a reasonable approach. However, due to the continues of the missing data, there is a chance where the data for some of the country is missing for the same years. If so, interpolation may not be effective. In this case, imputing missing values by year is a reasonable alternative.

After deciding on the filling method, two for loops and fillna() are used to fill the missing values. And running the isnull().sum() function again after the filling is done has determined that there are no more null values in the data group. The detail of the code is as follows (Figure 7).

```
#replace the null value with mean of each year
imputed_data = []
for year in list(df.Year.unique()):
    year_data = df[df.Year == year].copy()
    for col in list(year_data.columns)[3:]:
        year_data[col] = year_data[col].fillna(year_data[col].dropna().mean()).copy()
    imputed_data.append(year_data)
df = pd.concat(imputed_data).copy()

#again show if there is any null left
df.isnull().sum()
```

Figure 7: Codes for filling in the null value

After addressing missing values, it is important to handle outliers to improve the accuracy of the model. However, before addressing outliers, any string variables in the dataset need to be removed.

```
# Select only the non object data to do a box plot for outlier analysis
outlier = df.copy()
# The time related data and the number is temporarily removed from the set as there is no need for this process
outlier.drop(['Country','Status'],axis=1,inplace=True)
# Determine the remaining column name in one serie
colName = outlier.columns

# plot box plot for all none object variables to see which set of data have an outlier
plt.figure(figsize=(25,20))
for i in range(0,colName.shape[0]):
    plt.subplot(4, 6, i+1)
    plt.boxplot(outlier[colName[i]],autorange=True)
    plt.title(colName[i])
```

Figure 8: Code for outlier check

To accomplish this, a new data group can be established by using the copy() function. Then, any variables with an object data type, such as 'Country' and 'Status', can be removed using the drop() function. After removing the string variables, a boxplot can be generated to visualize the remaining data and identify any outliers. Below shows the function and the result of the boxplot.

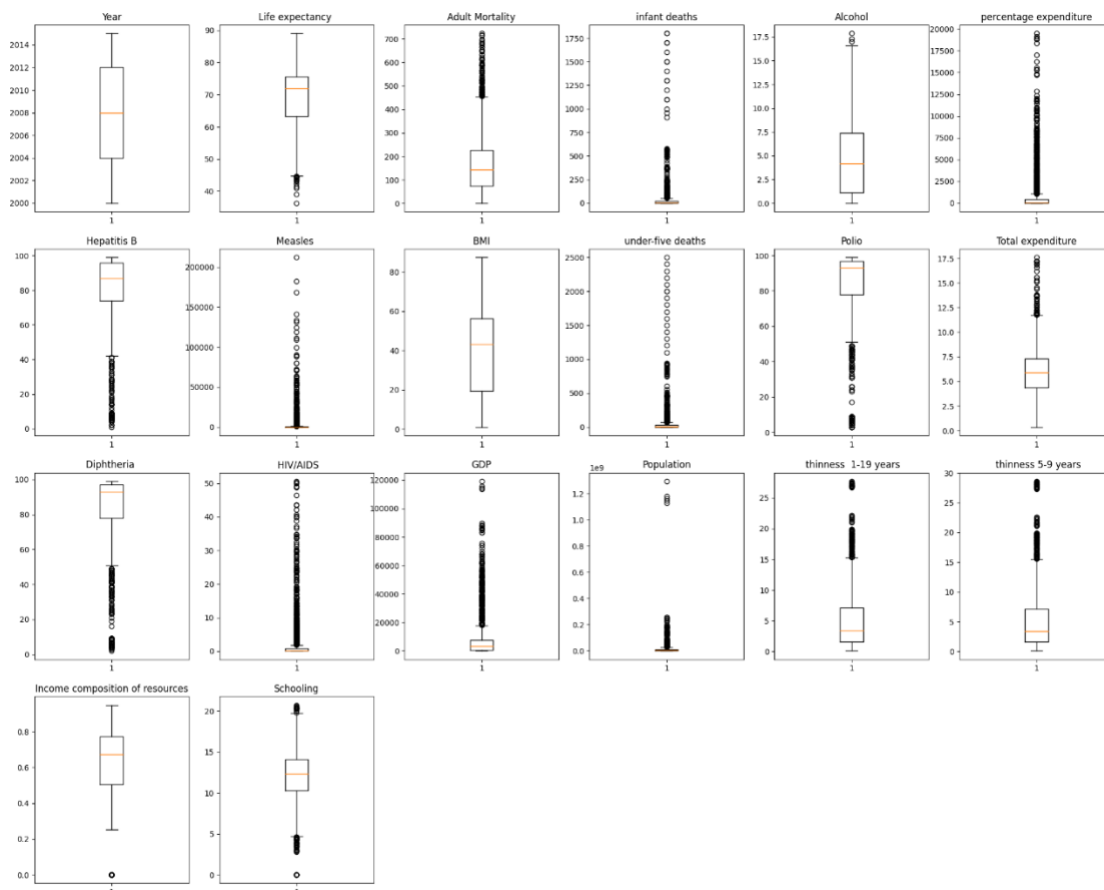


Figure 9: Outlier checking result

From the result, it can be observed that within the dataset most of the data has outliers and the amount of it is staggering.

In order to understand more specifically the proportion of outliers in each set of data, a function called outlier_count() is created, and a for loop is used to import the data set into this function. The codes and result are shown in Figure 10-11 below.

```
[ ] def outlier_count(col, data=df):
    print(15*'-' + col + 15*'-' )
    q75, q25 = np.percentile(data[col], [75, 25])
    iqr = q75 - q25
    min_val = q25 - (iqr*1.5)
    max_val = q75 + (iqr*1.5)
    outlier_count = len(np.where((data[col] > max_val) | (data[col] < min_val))[0])
    outlier_percent = round(outlier_count/len(data[col])*100, 2)
    print('Number of outliers: {}'.format(outlier_count))
    print('Percent of data that is outlier: {}'.format(outlier_percent))

outlier.drop(['Year'],axis=1,inplace=True)
for col in outlier:
    outlier_count(col)
```

```
-----Life expectancy -----
Number of outliers: 17
Percent of data that is outlier: 0.58%
-----Adult Mortality-----
Number of outliers: 86
Percent of data that is outlier: 2.93%
-----infant deaths-----
Number of outliers: 315
Percent of data that is outlier: 10.72%
-----Alcohol-----
Number of outliers: 3
Percent of data that is outlier: 0.1%
-----percentage expenditure-----
Number of outliers: 389
Percent of data that is outlier: 13.24%
-----Hepatitis B-----
Number of outliers: 222
Percent of data that is outlier: 7.56%
-----Measles -----
Number of outliers: 542
Percent of data that is outlier: 18.45%
----- BMI -----
Number of outliers: 0
Percent of data that is outlier: 0.0%
-----under-five deaths -----
Number of outliers: 394
Percent of data that is outlier: 13.41%

-----Polio-----
Number of outliers: 279
Percent of data that is outlier: 9.5%
-----Total expenditure-----
Number of outliers: 51
Percent of data that is outlier: 1.74%
-----Diphtheria -----
Number of outliers: 298
Percent of data that is outlier: 10.14%
----- HIV/AIDS-----
Number of outliers: 542
Percent of data that is outlier: 18.45%
-----GDP-----
Number of outliers: 300
Percent of data that is outlier: 10.21%
-----Population-----
Number of outliers: 203
Percent of data that is outlier: 6.91%
----- thinness 1-19 years-----
Number of outliers: 100
Percent of data that is outlier: 3.4%
----- thinness 5-9 years-----
Number of outliers: 99
Percent of data that is outlier: 3.37%
-----Income composition of resources-----
Number of outliers: 130
Percent of data that is outlier: 4.42%
-----Schooling-----
Number of outliers: 77
Percent of data that is outlier: 2.62%
```

Figure 10-11: Detailed outlier check and the result

After knowing the proportion of outliers in each column of data and the approximate distribution, you can use the winsorize() function to clean up the outliers in each column of data. Details of the codes are showcased in Figure 12.

```
[ ] lifeExpectancy = winsorize(outlier['Life expectancy '],(0.01,0))
adultMortality = winsorize(outlier['Adult Mortality'],(0,0.04))
infantDeaths = winsorize(outlier['infant deaths'],(0,0.15))
alcohol = winsorize(outlier['Alcohol'],(0,0.0025))
percentageExp = winsorize(outlier['percentage expenditure'],(0,0.135))
hepatitisB = winsorize(outlier['Hepatitis B'],(0.1,0))
measles = winsorize(outlier['Measles '],(0,0.19))
bmi = winsorize(outlier[' BMI '],(0,0))
under5Deaths = winsorize(outlier['under-five deaths '],(0,0.15))
polio = winsorize(outlier['Polio'],(0.1,0))
totalExp = winsorize(outlier['Total expenditure'],(0,0.02))
diphtheria = winsorize(outlier['Diphtheria '],(0.105,0))
hiv = winsorize(outlier[' HIV/AIDS'],(0,0.185))
gdp = winsorize(outlier['GDP'],(0,0.105))
population = winsorize(outlier['Population'],(0,0.07))
thinness10to19 = winsorize(outlier[' thinness 1-19 years'],(0,0.035))
thinness5to9 = winsorize(outlier[' thinness 5-9 years'],(0,0.035))
incomComp = winsorize(outlier['Income composition of resources'],(0.05,0))
schooling = winsorize(outlier['Schooling'],(0.025,0.005))
```

Figure 12: Code of fixed outlier

Finally, create a new empty set called data, and import each column of data, including the previously removed two columns containing object data, and the data cleaning work is complete.

Data Understanding (EDA)

EDA is the first step after data cleaning, involving using summary statistics and visualisations to analyse data in a univariate and bivariate sense. It helps identify relationships between life expectancy and other independent attributes, detect relationships among other attributes, observe trends, and test hypotheses.

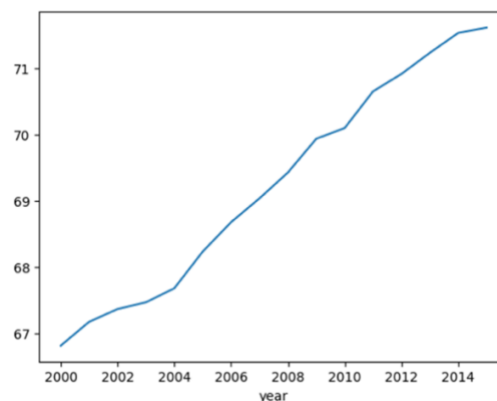


Figure 13: Life Expectancy Trending by Year

According to the line chart in Figure 13, the average life expectancy worldwide increased from 65 to 74 from 2000 to 2014.

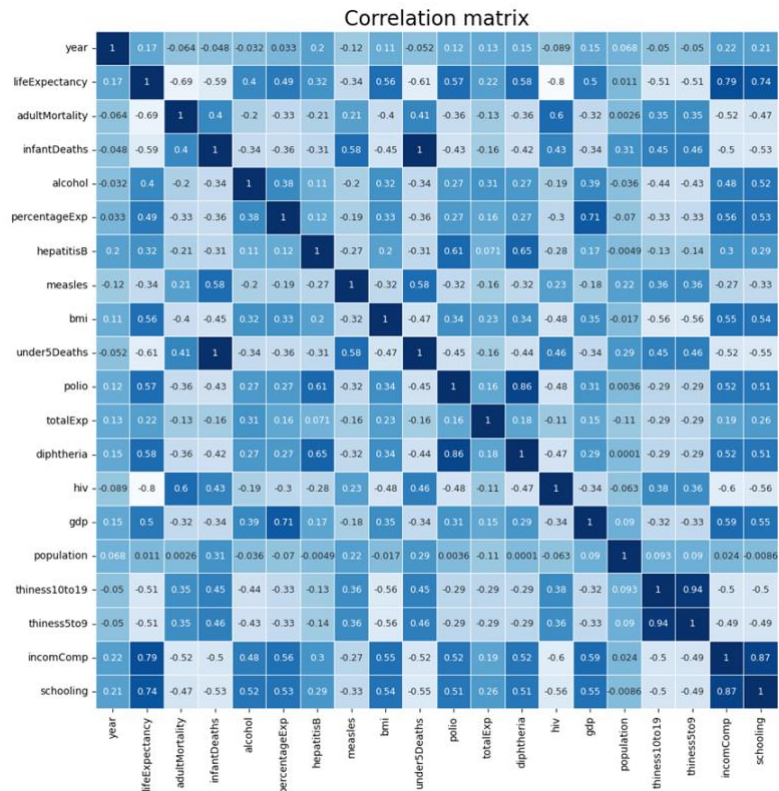


Figure 14: Correlation Matrix

Figure 14 demonstrates a heatmap, which was used to explore the correlation between features and life expectancy. The results of the heatmap showed positive correlations between life expectancy and income composition as well as schooling. Similarly, polio vaccine rate and diphtheria vaccine rate along with Hepatitis B vaccine rate are highly correlated. On the other hand, HIV had a significant negative correlation with life expectancy.

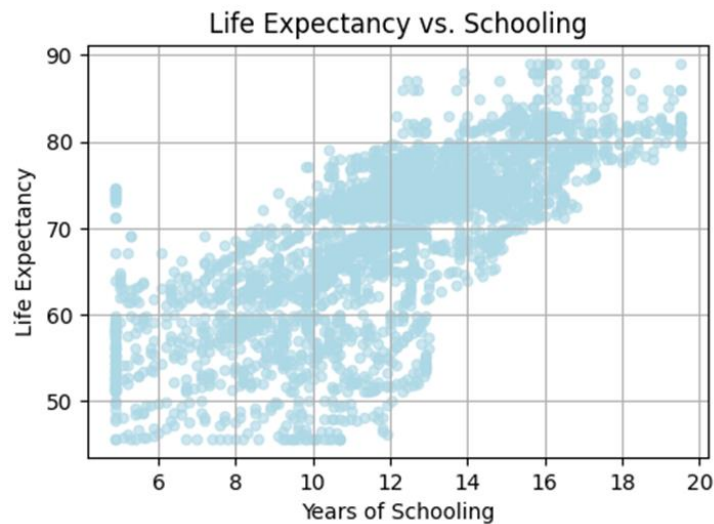


Figure 15: Correlation between Life Expectancy and Schooling

As per the scatterplot analysis in Figure 15, there is a strong positive correlation between life expectancy and years of schooling, indicating that individuals with higher levels of education tend to live longer. This relationship between education and life expectancy can be attributed

to several factors, such as improved access to healthcare, better decision-making abilities, and a higher standard of living, leading to improved health and longevity.

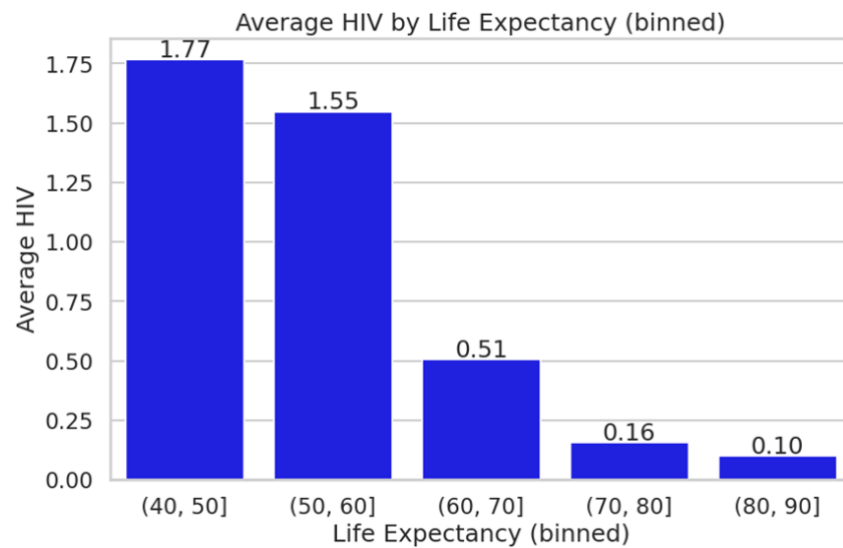


Figure 16: Average HIV Death Rate by Life Expectancy Binned Group

When comparing the average HIV death rate with life expectancy (Figure 16), it was observed that individuals between the ages 40-60 have high HIV death rate than those aged 60 and above.

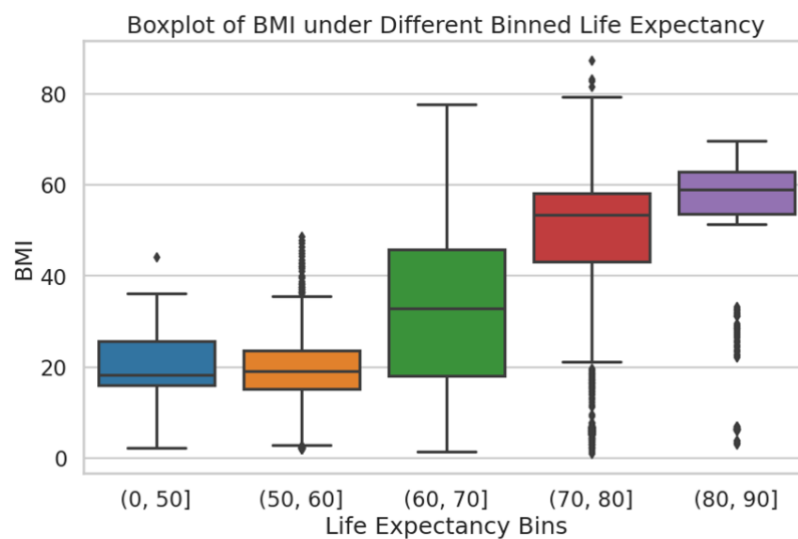


Figure 17: BMI Distribution on Each Life Expectancy Group

The NSW health department defines a healthy BMI range as 18.5 to 24.99 (NSW Health, 2021). As a result, the box plot in Figure 17 showed that the BMI was higher in the higher life expectancy range. However, since BMI is an average of the entire population, the BMI of elderly individuals might be no longer considered healthy. Countries with higher life expectancy tended to have a higher average BMI, as they had a higher proportion of elderly individuals in their population.

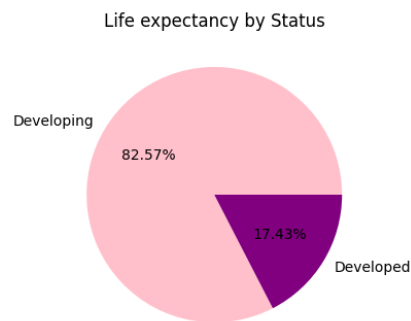
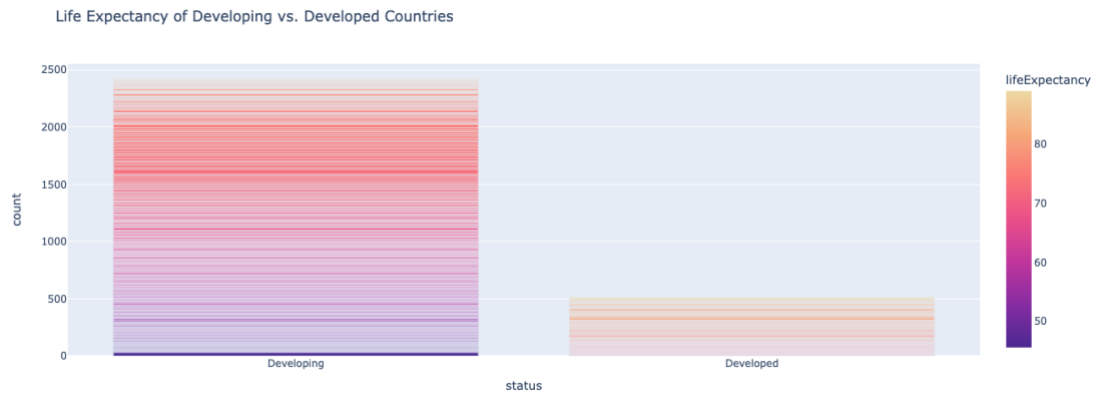


Figure 18-19: Life Expectancy by status

Figure 18 - 19 visually displays the comparison between developed and developing countries. As shown in the pie chart, the number of developing countries dominate developed countries by nearly 5 times. However, developed countries have higher life expectancy range of 80 years old or older, whereas most of the population in developing countries have life expectancy from 60-70 years old.

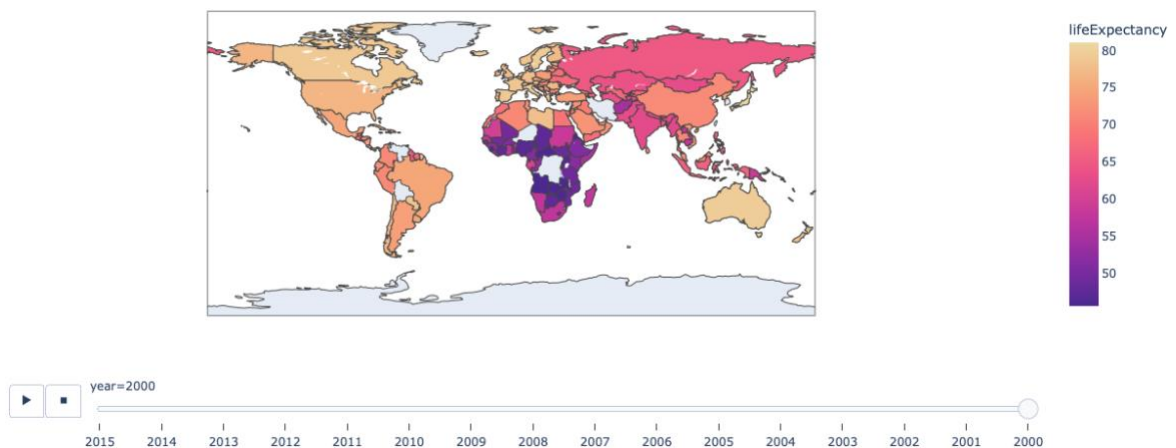


Figure 20: Life expectancy by region in 2000

Figure 20 illustrates a world map that presents the disparities in life expectancy across different regions of the globe in the year 2000. Notably, the map reveals that the highest life expectancy age recorded was 80 years old, while the lowest was 50 years old. Regions such as North and South America as well as Oceania recorded the highest life expectancy rates, with individuals living up to 75 years and above. In contrast, Africa emerged as the region with the

lowest life expectancy, with individuals mostly living between 50-70 years old. Asia has a middle range of life expectancy, varying from 60-75 years old.

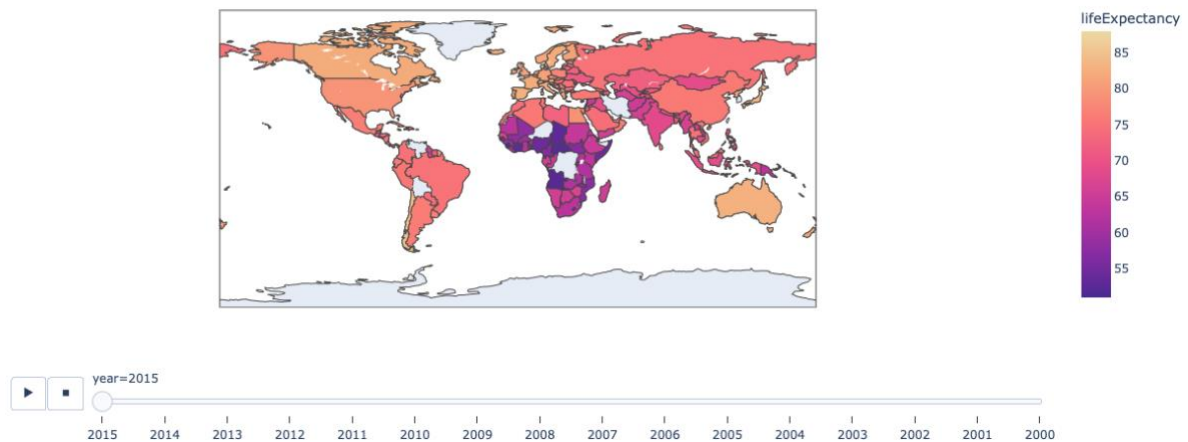


Figure 21 showcases a remarkable global increase of 5 years in life expectancy records, which now range from 55-85 years old. Regions such as North and South America, as well as Oceania, maintained their position as having the highest life expectancy rates, with a slight upward shift of 1-2 years in each country. Furthermore, there was a significant increase in life expectancy rates across Africa, with individuals living between 55-80 years old, indicating a noteworthy improvement. Additionally, Asia has experienced the highest increase in life expectancy rates, with Russia alone surging from 65 to 75 years old, reflecting a remarkable 5-year increase.

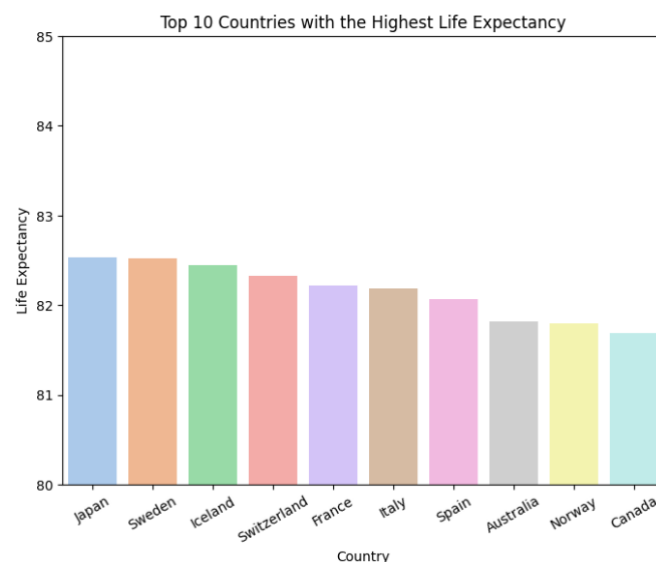


Figure 22: Top 10 countries with the highest Life Expectancy

Figure 22 displays the top 10 countries in the world with the highest life expectancy rates. Coincidentally, all these countries belong to the developed group, with only slight variations between each country's life expectancy rates. Japan emerged as the country with the highest life expectancy worldwide, with an average life expectancy of approximately 82.5 years old, closely followed by Sweden, which recorded a similar life expectancy rate. Canada, on the other hand, placed at the bottom of the chart, with citizens living on average until 81.8 years

old. Despite the differences in life expectancy rates between these countries, developed countries generally have higher life expectancy rates due to better access to healthcare, education, and resources.



Figure 23: Life Expectancy vs. Income Composition of resources

The scatterplot in Figure 23 reveals a positive correlation between life expectancy rates and the ratio of income composition of resources, indicating that individuals with higher income levels tend to live longer. This positive correlation also suggests that the more an individual earns, the higher the likelihood of accessing quality healthcare, education, and resources, which can positively impact their life expectancy.

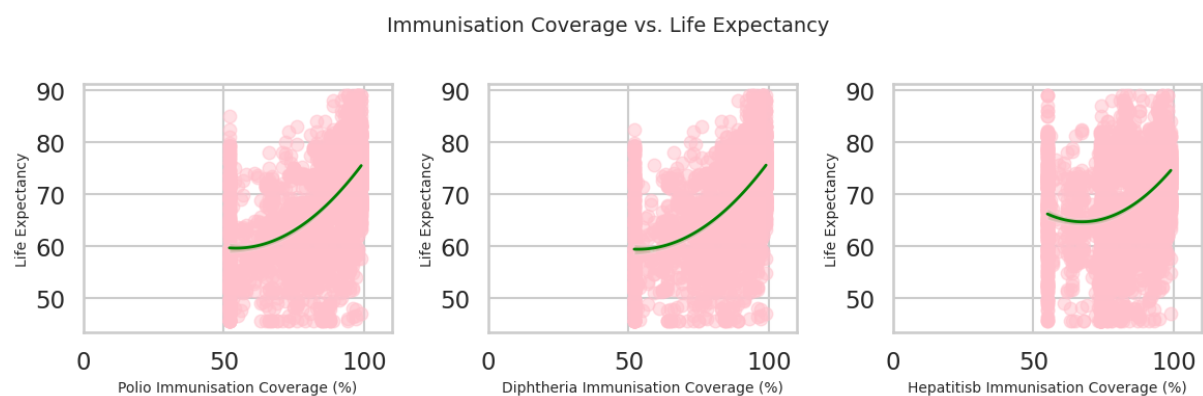


Figure 24: Immunisation Coverage vs. Life Expectancy

Similarly, Figure 24 displays a positive correlation between the vaccine coverage rates for polio, diphtheria, and Hepatitis B, and life expectancy rates. The scatterplots reveal a consistent pattern of variables across all charts, suggesting a strong relationship between vaccine coverage rates and life expectancy rates. This finding implies that individuals who receive vaccines against these diseases are more likely to live longer and enjoy better health outcomes.

Model Selection

The machine learning models that can be used for data analysis are very diverse. Therefore, further analysis is needed to determine which model is the best choice. For this reason, a total of three machine learning models were selected for this project, namely decision trees, random forests, and linear regression.

Decision trees are a tree-like model used for classification and regression tasks. Each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or a numeric value. Linear regression predicts a continuous output variable based on one or more input variables. It estimates the coefficients of the linear equation that best fits the data and can be used for forecasting and trend analysis. Random forest is an ensemble algorithm that combines multiple decision trees to improve the accuracy and stability of the predictions. It can handle high-dimensional datasets with complex relationships between variables. These three machine learning algorithms are essential tools for data scientists and analysts to make accurate predictions and gain insights from complex data.

The experimental procedures for implementing machine learning models usually involve data collection, data pre-processing, feature engineering, model selection, training, evaluation, and deployment.

The previous steps focus on cleaning the data to ensure accuracy and reliability, including dealing with missing values and outlier. next, this paper will analyse the existing data.

Experiments

Data partitioning

To split the dataset, the input and output variables were identified. Life expectancy was chosen as the output variable (y), while the 'country' variable was removed from the analysis due to its string values. Additionally, the 'status' variable was transformed into a numerical value for model training, with the remaining variables were used as input variables (x). The dataset was then randomly split into training and testing sets, with 80% of both x and y sets used for training, and the remaining 20% used for testing.

Decision Tree

Modelling and performance measurement - Decision Tree

Decision trees can be used in a wide variety of situations, including variable selection, assessing the relative importance of variables, handling of missing values, prediction, and data manipulation (Song & Lu, 2015). The creation of a decision tree requires the identification of dependent and independent variables. As we know from the business problem mentioned earlier, the purpose of this analysis is to find the factor that has the greatest impact on life

expectancy, so life expectancy is the dependent variable, and all other factors are independent variables.

Because of the need to compare the accuracy between different models, this analysis uses a comparison of Mean Squared Error and R square to make a comparison.

First, a decision tree model is built first. To build a decision tree model, a decision tree classifier object is created and stored in the variable `clf`, and the model is trained using `x_train` and `y_train`. Figure 25 shows the process of creating a decision tree and calculating MSE and R-square.

```
[ ] # Feature name
    feature_names = x.columns

# Creating decision tree object
clf = tree.DecisionTreeClassifier(criterion = 'entropy')

[ ] #Create decision tree
    np.nan_to_num(x_train, copy=False)
    np.nan_to_num(y_train, copy=False)
    np.nan_to_num(x_test, copy=False)
    np.nan_to_num(y_test, copy=False)
    clf_tree = clf.fit(x_train, y_train.astype('int'))
    score = clf_tree.score(x_test, y_test.astype('int'))
    print("Models are scored on the test set: \n", score)
    predict_y = clf.predict(x_test)
    print("Prediction results for the test set samples: \n", predict_y)
    predict_y1 = clf.predict_proba(x_test)
    print("Predict the probability of a sample being a certain label: \n", predict_y1)
    clf_dot = tree.export_graphviz(clf_tree,
                                   out_file= None,
                                   # feature_names= feature_names,
                                   # class_names= target_names,
                                   filled= True,
                                   rounded= True)

[ ] # Calculating the mean squared error (MSE) and R-squared (R2) metrics
    y_pred = clf.predict(x_test)
    mse_tree = mean_squared_error(y_test, y_pred)
    r2_tree = r2_score(y_test, y_pred)

# Print the MSE and R2 values
print("Mean Squared Error: ", mse_tree)
print("R-Squared: ", r2_tree)

Mean Squared Error:  7.873950541786187
R-Squared:  0.9138658320054202
```

Figure 25: Create Decision Tree and Validation Result

The MSE measures the error between the predicted and actual values, so the smaller the value the better. The MSE is calculated to be 7.87, which is an acceptable range. The R-squared value is between 0 and 1, the closer to 0, the worse the fit, and the closer to 1, the better the model fit. The R- square of this decision tree is 0.914, which indicates that this decision tree model fits the data very well.

```
[ ] print("\nFeature importance is: ")
    info = [*zip(feature_names, clf.feature_importances_)]
    max=0.0
    r=()
    for cell in info:
        print(cell)
        if cell[1]>max:
            max=cell[1]
            r=cell
    print("Most important features:",r[0])

Feature importance is:
('year', 0.05388730818478698)
('status', 0.003802575315122314)
('adultMortality', 0.19822625574783118)
('infantDeaths', 0.02366991697650891)
('alcohol', 0.061893615005416104)
('percentageExp', 0.03140863043839775)
('hepatitisB', 0.013079402298571878)
('measles', 0.020507763857964587)
('bmi', 0.04864193090251175)
('under5Deaths', 0.017267427051355744)
('polio', 0.026076928755283234)
('totalExp', 0.03761234694860306)
('diphtheria', 0.015001832288159088)
('hiv', 0.1173557296361851)
('gdp', 0.02425239533701377)
('population', 0.022892930627193796)
('thinness10to19', 0.034686428998877535)
('thinness5to9', 0.03569811828841351)
('incomComp', 0.1562764341885017)
('schooling', 0.05776202915330216)
Most important features: adultMortality
```

Figure 26: Feature Importance based on Decision Tree Result

Figure 26 presents the result of the decision tree, and each factor has a different degree of importance. The most important feature is adultMortality, and the combination of other data shows that incomComp and HIV are also important factors that affect life expectancy.

Random Forest

Modelling and performance measurement - Random Forest

Random forest was a popular model used to solve classification and regression problems. This model has high accuracy and specializes in training for datasets with numerous features (Smith, 2021). Additionally, this model can identify the importance of features, which was essential information for designing the health check pack.

To validate the model performance, a combination of Mean Squared Error and R-squared were used. The Mean Squared Error compared the average squared difference between predicted and actual values. And R-squared error was a measurement to evaluate the proportion of variance in the target variable that can be explained by the independent variables. Which is the indication of how well the model fits the data (Smith, 2021).

In this case, the MSE of 3.418 indicated the model accurately predicted the target variable. The R-squared of 0.9626 suggested that the model explained 96.26% of the variance in the target variable, which expressed a good fit (Figure 27).

```
1 #random forest
2 rf = RandomForestRegressor(random_state=0)
3 rf.fit(x_train_st,y_train)
4 y_pred = rf.predict (x_test_st)
5 mse_forest = mean_squared_error(y_test,y_pred)
6 r2_forest = r2_score(y_test,y_pred)
7 print("Mean Squared Error:",mse_forest)
8 print("R-Squared:",r2_forest)
```

Mean Squared Error: 3.418380293815029
R-Squared: 0.9626058938350872

Figure 27: Random Forest Model Coding and Validation Result

According to the random forest model, an importance feature ranking can be generated to analysis the importance of each input variable. Based on the result in figure 28, HIV was identified as the most important feature, scoring three times more than the second most important feature. When considering the health check pack, it is recommended to focus on the features that can be examined by the hospital, among the top 10 important features. In this case, HIV, BMI, and 'Prevalence of thinness' should be included in the health check pack. As 'Prevalence of thinness' can be discovered from BMI, it is sufficient to include HIV and BMI for the health check.

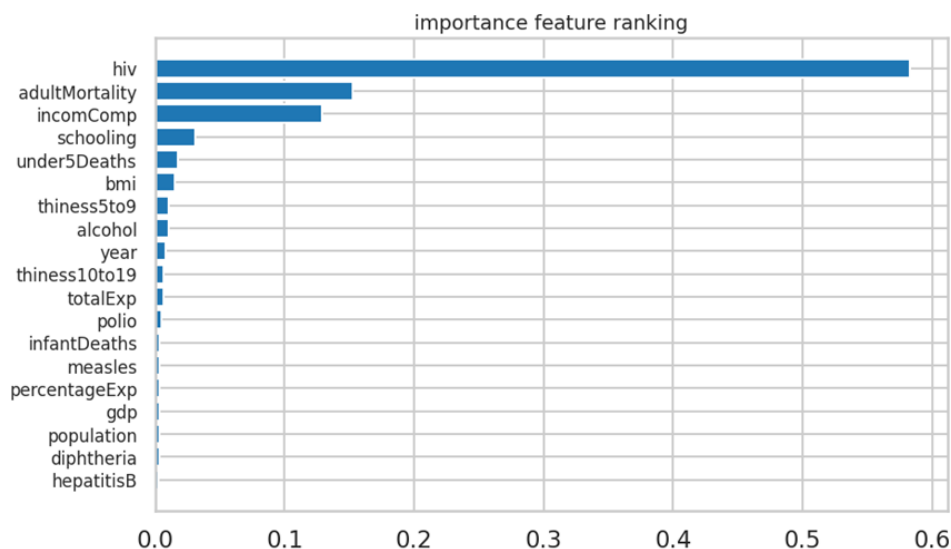


Figure 28: Feature Importance Rank by Random Forest Model

Linear Regression

Modelling and performance measurement - Linear Forest

Based on numerous characteristics, we are using the regression model to estimate life expectancy. The data is initially pre-processed by employing the StandardScaler tool obtained from the sklearn processing module to standardize the variables that are independent. The model is then trained by the Linear Regression algorithm from sklearn. linear_model package on the processed data used for training.

For evaluating the predictive power of the model of linear regression, the R-squared (R2) & Mean Squared Error (MSE) meters are generated on the test data using the sklearn metrics module's mean_squared_error and r2_score operations. The cross_val_score code is obtained from the sklearn model selection module and is also used to generate cross-validation mean squared errors and R-squared scores. Both the R2 & MSE scores for testing and cross-validation results are displayed in Table 3.

Metrics	Values
Metric Squared Error	13.2950
R-Squared	0.85456
CV Mean Squared Error	12.82023
CV R-Squared	0.85547

Table 3: Results of Linear regression model

As per the obtained results from the linear regression model, the Mean Square Error (MSE) is 13.2950, and the R-squared (R2) score is 0.8545, so this shows that the model accuracy is about 85.52% of the given dataset. The Average difference across the predicted score and the actual score is calculated by the MSE, so the lesser the Scores the better efficiency of the Machine learning model.

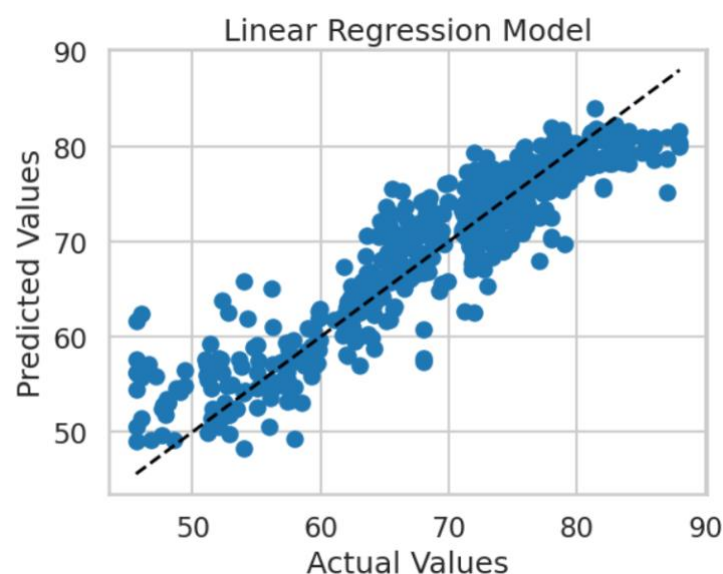


Figure 29: Predicted values vs Actual values

Figure 29 demonstrates predicted values vs actual values to display the predicted and the actual values on the y-axis, and x-axis respectively. By showing how close the predicted are to the actual values, we can understand how this linear regression model is performing on the data used.

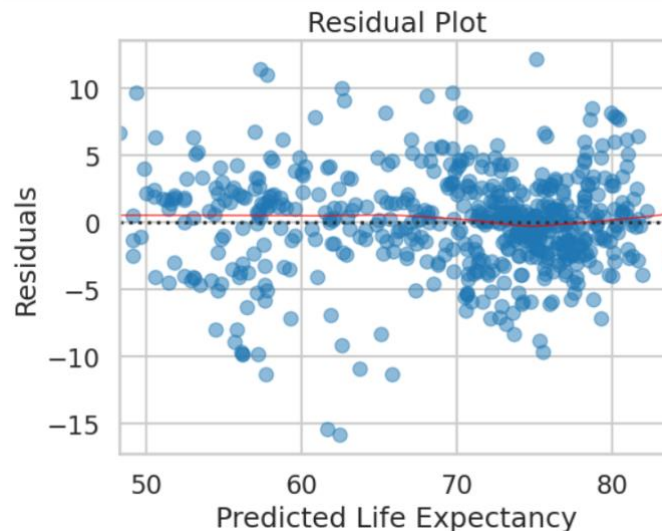


Figure 30: Residual plot

The residual plot in Figure 30 is to check if the linear regression model assumptions are met.

$$\text{Residual} = \text{actual values} - \text{predictive values}$$

As per Figure 30 generated by the model, linear regression is not accurate because some of the residuals are scattered far away from the horizontal line.

Model comparison

Life expectancy prediction is a regression problem, and the evaluation of the model's performance can be achieved by using appropriate indicators. MSE and R2 are widely used to compare the performance of different regression models. In this study, MSE and R2 were chosen as the indicators for evaluating the performance of the models. Among the compared models, the Random Forest model showed less prediction variance and the best fit to the data, making it the most suitable model for predicting life expectancy (Figure 31).

	Model_name	mean_squared_error	r2_score
0	Random_forest	3.418380	0.962606
1	Decision_tree	9.156548	0.899835
2	Linear_regression	13.295038	0.854564

Figure 31: Model Comparison by Mean Squared Error and R Square Error

Model Tuning

After defining the most suitable model for the project, a hyper-parameter optimization process was conducted to tune the learning algorithms automatically. Grid search was used to tune the Random Forest model by identifying the best combination of hyperparameters, such as the number of estimators, max depth, minimum sample split, and minimum sample leaf. The performance metric used in this case was MSE. For each combination of hyperparameters, the grid search was computed, and the MSE was tested to find the best performance combination of hyperparameters (Bergstra & Bengio, 2012). The results in Figure 32 showed that the model required a minimum of 30 layers of trees with a minimum number of samples per leaf and split. The model also used a total of 150 decision trees to make decisions.

```
1 rf = RandomForestRegressor()
2
3 param_grid = {
4     'n_estimators': [50, 100, 150, 200],
5     'max_depth': [10, 20, 30, None],
6     'min_samples_split': [2, 5, 10],
7     'min_samples_leaf': [1, 2, 4]
8 }
9
10 scorer = make_scorer(mean_squared_error, greater_is_better=False)
11
12 grid_search = GridSearchCV(rf, param_grid=param_grid, scoring=scorer, cv=5)
13
14 grid_search.fit(x_train_st, y_train)
15
16 # Print the best hyperparameters
17 print("Best Hyperparameters: ", grid_search.best_params_)

Best Hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
```

Figure 32: Grid Search for Model Tuning

After optimizing the hyperparameters and re-training the Random Forest model, there was a noticeable improvement in its performance. In Figure 33, the Mean Squared Error was reduced from 3.41 to 3.35, indicating higher accuracy on model prediction. Additionally, the R-squared value increased from 0.962 to 0.963, indicating that the model explained 96.3% of the variability in the target variable.

```
mean_squared_error_after HOP: 3.3559201797686096
R-Squared_after HOP: 0.9632891531377313
```

Figure 33: Model Validation after Model Tuning – Random Forest

By comparing the predict result with the testing output, the prediction was 99.1% accurate (Figure 34). This result met the success criteria.

```
[45] 1 pred_com=pd.DataFrame(y_pred,y_test)
      2 pred_com['accuracy']= 1 - abs(y_pred - y_test)/y_test
      3 pred_accuracy = pred_com['accuracy'].mean()
      4 print('model_prediction_accuracy:',pred_accuracy)

model_prediction_accuracy: 0.9911900472225657
```

Figure 34: Prediction Accuracy – Random Forest

As the Random Forest model has more than 30 layers, it may be helpful to provide a visualised tree for the first two layers for the purpose to better understand how the model is working (Figure 35).

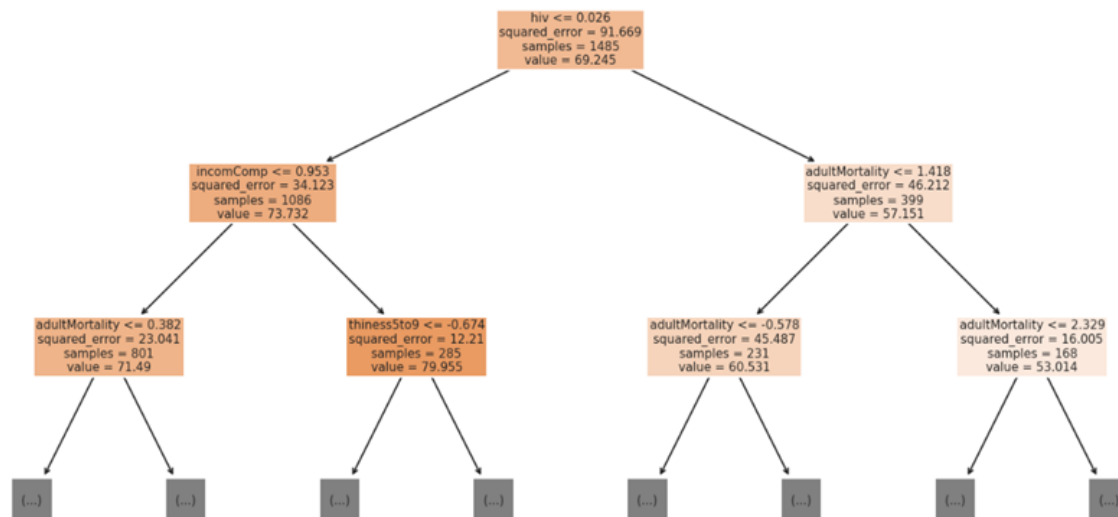


Figure 35: Random Forest Model Visualisation

Discussion and conclusions

To conclude, machine learning techniques were utilised in this paper to examine the elements that influence life expectancy. With data analytics, the project aims to empower stakeholders in the healthcare industry to make informed decisions with insightful information. Our obtained results suggest that the most significant determinants influencing life expectancy are adult mortality, income composition, BMI and HIV. These findings have practical benefits for businesses, insurance providers, and healthcare organisations. Healthcare providers can better allocate their resources according to the primary causes of early mortality, insurance companies can adjust their pricing in accordance with the risk factors identified, and businesses can model their healthcare packages based on the data insights provided.

The project successfully achieved the defined objectives and deliverables. We developed a predictor model using machine learning algorithms, conducted a cost-benefit analysis to determine cost-effectiveness, and incorporated the machine learning model in a real-world setting to assist with prediction processes. The proposal document, exploratory data analysis, Python prototype, and research document were all delivered as promised.

The decision tree and random forest models used in this project produced valuable insights and information about the factors that affect life expectancy. We were able to determine the most crucial features that should be included in a health check pack thanks to the feature ranking that the random forest model gave us to examine the significance of each input variable.

In conclusion, the project emphasises the significance of data analytics in the healthcare sector and its potential to transform the sector. Healthcare providers, insurance providers, and businesses can optimise their internal metrics, raise the general quality of life, and minimise healthcare costs by understanding the main reasons why life expectancy is declining and common illnesses.

Discussion

This project has provided valuable insights for stakeholders in the healthcare industry, insurance companies, and businesses. The analysis has revealed the key factors that affect life expectancy, including adult mortality, income composition, and HIV. This information can be used by healthcare providers to tailor their resource allocation and focus on preventative measures for patients. Insurance companies can use these insights for their actuarial models, basing pricing on our life expectancy data. Employers can tailor their healthcare benefits to match the most common causes of early mortality using the information gathered.

Limitations

There are several restrictions that apply to our analysis. First, the chosen dataset for this project only contains data from a limited number of countries, which means it does not entirely represent the whole world's population. Besides, the quality of the data utilised can have an impact on the accuracy of the machine learning models. Specifically, the accuracy of the generated models and the reliability of the findings might be impacted by the World Health Organization's data, which could be incomplete or out-of-date and contain errors or inconsistencies. Lastly, the project's models were relatively basic, which might limit the ability to capture the data complexity, and although the decision tree and random forest models are effective at forecasting outcomes, they might be challenging to interpret and pass on the model's results to stakeholders.

Future Recommendations

To develop the machine learning model's accuracy, it is crucial to ensure that the quality of data employed is high-quality by getting the data from reliable sources, verifying it before use, and updating it. Besides, future studies could investigate how additional elements such as air pollution affect life expectancy. The components that are impacting life expectancy can be more fully comprehended if other variables, such as environmental or lifestyle factors, are considered. Additionally, there are more available sophisticated models that could be used in the future, such as deep learning algorithms, which might be better for identifying complex patterns in the data. Moreover, to increase the model's understandability to the stakeholders, making the model easier to interpret is essential. This might be accomplished by using models that are more easily understandable or offering more beginner-friendly visualisations or explanations of the model's outcomes. Furthermore, testing the model on datasets from different nations or populations could help determine its generalisability and effectiveness in different contexts. Lastly, the model's impact might be maximised by cooperation with healthcare providers and other stakeholders in developing a thorough data-driven approach.

to improving health outcomes. This might involve developing focused interventions based on the model's findings or integrating the model's predictions into the processes used to make healthcare decisions.

In general, this study emphasises the value of applying data analytics to healthcare to enhance patient outcomes and lower costs. With the knowledge gathered from this project, tailored interventions and policies can be created to address the major variables that determine life expectancy. Machine learning algorithms will be more and more crucial as datasets continue to expand in size and complexity for uncovering the story of individual health and providing stakeholders with data-driven insights.

Reference

- Badea, L., Onu, M., Wu, T., Roceanu, A., & Bajenaru, O. (2017). Exploring the reproducibility of functional connectivity alterations in Parkinson's disease. *PLOS ONE*, 12(11), e0188196. <https://doi.org/10.1371/journal.pone.0188196>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb), 281-305.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T.P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.
- Daivi. (2023, April 24). *Machine learning in insurance: Applications, use cases, and projects*. ProjectPro. <https://www.projectpro.io/article/machine-learning-in-insurance/774>
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17. <https://doi.org/10.1016/j.csbj.2014.11.005>
- Income, health, and social welfare policies*. (2020, March). The Lancet Public Health. [https://doi.org/10.1016/S2468-2667\(20\)30034-7](https://doi.org/10.1016/S2468-2667(20)30034-7)
- Kantify. (2021, May 4). 5 Pharma companies using Artificial Intelligence. Kantify. <https://www.kantify.com/insights/5-pharma-companies-using-artificial-intelligence>
- Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, 104, 398–412. <https://doi.org/10.1016/j.neuroimage.2014.10.002>
- NSW Health. (2021, January 13). Measuring Body Mass Index. <https://www.health.nsw.gov.au/heal/Pages/bmi.aspx>
- Smith, J. (2021). Understanding mean squared error and R-squared. *Journal of Machine Learning Research*, 22(1), 145-156.
- Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, 27(2), 130–135. <https://doi.org/10.11919/j.issn.1002-0829.215044>
- Wang, J., Siddicky, S. F., Oliver, T. E., Dohm, M., Barnes, C. W., & Mannen, E. M. (2019). Biomechanical Changes Following Knee Arthroplasty During Sit-To-Stand Transfers: Systematic Review. *Journal of Arthroplasty*, 34(10), 2494–2501. <https://doi.org/10.1016/j.arth.2019.05.028>

Appendix

Appendix 1: Attribute description

Attribute	Description
Country	Name of the country in which the indicators are from
Year	The calendar year that the data is recorded (ranging from 2000 to 2015)
Status	Whether a country is 'Developing' or 'Developed' by WHO standards
Life expectancy	The life expectancy of people in years for a particular country and year
Adult Mortality	The adult Mortality rate for both gender (the probability of death between the age of 15 and 60 per 1000 population)
Infant deaths	Number of infant deaths per 1000 population
Alcohol	Recording of alcohol consumption per capita (15 years old and above) in liters of pure alcohol
Percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product (GDP) per capita (%)
Hepatitis B	Percentage of Hepatitis B (HepB) immunisation coverage among 1-year-olds (%)
Measles	Number of Measles cases reported per 1000 population
BMI	Average Body Mass Index (BMI) of a country's total population
Under-five deaths	Number of children under-five deaths per 1000 population
Polio	Percentage of Polio immunisation coverage among 1-year-olds (%)
Total expenditure	Percentage of general government expenditure on health with total government expenditure (%)
Diphtheria	Percentage diphtheria-tetanus toxoid and pertussis (DTP3) immunisation coverage among 1-year-olds (%)
HIV/ AIDS	Number of people under 5 who die due to HIV/AIDS per 1000 births (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Population	Population of the country
Thinness 1-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9 (%)
Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Number of years of Schooling (years)

Appendix 2: Contribution table

Student name	Contribution
Kristen Nguyen	Business context, project aims, objectives and deliverables, data understanding (EDA), research paper formatting and reviewing
Chentao Hu	Project contribution, how paper is organised, data understanding (EDA), data partitioning, random forest model, model comparison, model tuning
Huong Ly Nguyen	Project origins and motivation, Methodology framework, Discussion and Conclusions, Attribute description
Kaushal Gandla	Success criteria, linear regression model
Mingrui Song	Stakeholder analysis, data preparation
Yuwen Sheng	Model selection, decision tree model

Appendix 3: Coding link

<https://colab.research.google.com/drive/1UZ-y77D3dVCAO410Wmp3XkOpvCFTzswf?usp=sharing>