Proceedings of the 2023 International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/24/20230424

Exploring the Impact of CO_2 Emissions, GDP, and Health Expenditure on Individual Life Expectancy

Xinyu Chang[1,a,*]

[1] *Information School, University of Washington, Seattle, Washington, United States, 98105*
a. xchang3@uw.edu
**corresponding author*

Abstract: Understanding the link between a person's health spending and their Gross Domestic Product (GDP) can have important effects on policy decisions at both the individual and national level. The research mainly explores the impact of individual CO_2 emissions, GDP, and health expenditure on individual life expectancy. The study, through a method of quantitative data analysis, analyses existing data and explores the relationships between independent variables and dependent variables. The author uses the decision tree model and a linear regression model to predict the impact of these factors on life expectancy. The study aims to understand the trends and patterns in the distribution and outliers of individual GDP/CO_2 Emissions/Life Expectancy/Health Expenditure datasets across 261 countries from 2000 to 2019. The research finds that individual health expenditure has a positive impact on individual life expectancy, while GDP level and CO_2 emissions have a negative impact. The study also explores the greatest impact of these independent variables on an individual's life expectancy and how changing the maximum depth of the decision tree affects model accuracy and feature importance. Overall, the study provides insights into the relationships between these variables and their impact on individual life expectancy, which has important implications for policy and interventions aimed at promoting public health and environmental sustainability.

Keywords: CO_2 emissions, GDP, health expenditure, life expectancy, decision tree model, linear regression model

1. Introduction

Understanding the link between a person's health spending and their Gross Domestic Product (GDP) can have important effects on policy decisions at both the individual and national level. As concerns about the environment and public health continue to grow, people are becoming more interested in the link between economic growth, life expectancy, and CO_2 emissions. By looking at how these things affect each other, the author can learn more about how economic growth and environmental sustainability affect public health. This can help people make policy decisions that promote both economic growth and better health outcomes.

The author uses quantitative data analysis to look at how CO_2 emissions, GDP, and spending on health affect a person's life expectancy. The machine learning model is used most of the time in research to figure out and predict how independent variables and dependent variables are related to

DOI: 10.54254/2754-1169/24/20230424

each other. The main use of the decision tree model and linear regression model is to help people have a better understanding of the contents of the research.

The first one is the trends and patterns in the distribution and outliers of individual GDP/CO_2 Emissions/Life Expectancy/Health Expenditure datasets across 261 countries from 2000 to 2019. The second one is the effect of the individual's health expenditures on their life expectancy. The third one is the impact of personal GDP, GDP level, and CO_2 emissions on each person's life expectancy. Fourth, the paper explores the greatest impact of the independent variables (GDP, CO_2, and health expenditure) on an individual's life expectancy. The effect of changing the maximum depth of the decision tree on the model's accuracy and feature importance.

It is important for both individual health planning and planning for the health of a population to be able to predict life expectancy based on things like CO_2 emissions, GDP, and health spending. By looking at how these factors affect life expectancy, the author can better understand the effects of different policy interventions meant to improve health outcomes and make better decisions about how to spend money on health promotion and disease prevention. Using the decision model to make these predictions can give researchers valuable information about how these factors affect each other and help researchers come up with better ways to improve health and quality of life.

2. Literature Review

In recent years, people have become more interested in studying the link between carbon dioxide (CO_2) emissions, gross domestic product (GDP), health spending, and individual life expectancy [1]. Many studies have looked at how these things affect the health of people and the environment [2]. Studies have shown that high levels of CO_2 emissions have a negative impact on the environment and human health [1]. Increased CO_2 emissions contribute to global warming, which in turn affects air quality, water supplies, and the number of natural disasters [1]. Exposure to CO_2 emissions that pollute the air has been linked to a number of health problems, such as lung and heart diseases, cancer, and dying too soon [2].

Researchers have also found that GDP has a big effect on how long people live [3]. High economic growth is linked to better healthcare infrastructure, more people having access to healthcare services, and better health outcomes [3–4]. But rapid economic growth can also lead to more pollution and damage to the environment, which may hurt public health in the long run.

Health expenditure is another important factor that influences individual life expectancy [4]. Adequate investment in healthcare systems and services is necessary to ensure that individuals have access to quality medical care [4]. Studies have shown that countries with higher health expenditure per capita generally have longer life expectancies [4].

Overall, research suggests that there is a complex interplay between CO_2 emissions, GDP, health expenditure, and individual life expectancy. Understanding these relationships is critical for developing effective policies and interventions that promote sustainable economic growth while also protecting public health and the environment [5].

3. Methodology

3.1. Data Cleaning

For cleaning the data, the researchers read and cleaned data from multiple CSV files, merged them, and then saved the cleaned data to a new CSV file. The cleaned data is set up as a Pandas DataFrame with columns for country name, year, life expectancy, GDP, health spending, and CO_2 emissions. To clean the data, the researcher would need to follow these steps:

First, prepare the environment with the required dependencies, such as Pandas. Second, download the CSV files containing the data related to life expectancy, GDP, health expenditure, and CO_2

DOI: 10.54254/2754-1169/24/20230424

emissions. Second, save the script as a Python file and run it from the command-line interface. Third, the script will call the clean_data() function, which consists of two helper functions: load_data() and merge_data(). Fourth, the load_data() function takes the name of the CSV file and the name of the column containing the data values as input. It loads the data from the CSV file, reshapes it using Pandas, and returns a cleaned Pandas DataFrame object. Fifth, the merge_data() function takes a list of Pandas DataFrame objects as input and merges them into a single DataFrame by joining on the country name, country code, and year columns. Sixth, the clean_data() function calls the load_data() function four times with the CSV files containing the data related to life expectancy, GDP, health expenditure, and CO_2 emissions. It then calls the merge_data() function to merge the resulting four DataFrames into a single DataFrame. Finally, it saves the cleaned data to a new CSV file called "clean_data.csv". After running the script, the author can open the "clean_data.csv" file and analyse the cleaned data. The author can use various Pandas functions to manipulate and summarise the data, such as describe(), groupby(), and pivot_table().

3.2. Data Analyzing

For analysing the data, the researcher loads the data from a CSV file and calls the necessary functions to plot histograms, normal distributions, boxplots, and histograms with outliers removed for specified columns in the Pandas DataFrame. To analyse the data, the researcher would need to follow these steps:

First, set up the environment with things like Pandas, NumPy, Matplotlib, and SciPy that are needed. Second, download the CSV file containing the cleaned data. Third, save the script as a Python file and run it from the command-line interface. Fourth, the script will call the main() function, which loads the cleaned data from the CSV file into a Pandas DataFrame. Fifth, the plot_distributions() function loops over the specified columns in the DataFrame, plots histograms and normal distributions for each column, and shows the resulting plots. Sixth, the no_outliers_plot_distribution() function takes a Pandas DataFrame and the name of a column as input. It removes the outliers from the column, calculates the mean and standard deviation of the filtered data, plots a histogram and the normal distribution of the filtered data, and shows the resulting plots. Seventh, the plot_boxplot() function takes a Pandas DataFrame and the name of a column as input. It plots a box plot of the column and shows the resulting plot. Eighth, after running the script, the author can analyse the plots to answer various research questions related to the challenge goals of promoting global health and sustainable development.

3.3. Linear Regression Model

To create the linear regression model, the researcher used the Statsmodel package and the OLS (Ordinary Least Squares) regression model to estimate the relationship between each dependent variable and independent variable. The author also predicted the life expectancy under each dependent variable condition and calculated the root mean square error (RMSE) as well as the adjusted root mean square error of each model to test the model. The author used the test_train_split function in the sklearn package to train and test the model. To create the regression models to make predictions and estimations, the following steps were implemented:

First, load in cleaned data and read the data in each regression model. Second, create a new column to store the prior GDP and calculate the GDP growth rate for each country. Take a GDP growth rate higher than 0.02 to be a high GDP level annotated as 1, and a low GDP level (lower than 0.02) annotated as 0. Third, train the linear regression models based on the independent variables and dependent variables stated in the question. Fourth, predict each country's life expectancy based on dependent variables or variables. Fifth, use the Sklearn train_test_split function to split the dataset

DOI: 10.54254/2754-1169/24/20230424

into 20% and 80% models. The last step is calculating and printing the root mean squared error and adjusting the root mean square error based on training and test data.

3.4. Decision Tree Model

To create the decision tree model, the researcher used the Scikit-Learn (Sklearn) package, (Decision tree) classification model to estimate the relationship between each dependent variable and independent variable. The author involved training and testing a decision tree model to classify life expectancy into different age groups based on several independent variables, such as GDP, CO_2 emissions, and health expenditure. To create the decision tree models to make predictions and estimations, the following steps were implemented:

First, load the cleaned dataset from a CSV file using Pandas. Second, classify the life expectancy column into different age groups using the classify_lifeexp function. This will create a new DataFrame with the LifeExp column classified into different age groups. Third, filter the data to include only years after 2010 using DataFrame filtering. Fourth, call the decision_tree function to train and test a decision tree model with the following arguments:

a. The DataFrame contains the classified LifeExp column and independent variables.
b. A list of independent variable names (GDP, CO_2, and Health_Expenditure).
c. The dependent variable name (LifeExp).

The decision_tree function will perform the following steps:

a. Split the data into training and test sets using the train_test_split function.
b. Create two decision tree classifiers, one with a max depth of 3 and one without.
c. Train the classifiers on the training data.
d. Use the trained classifiers to predict the labels of the test data.
e. Print the accuracy score of the predictions.
f. Get the feature names and target names for plotting the decision trees and feature importance graphs.
g. Plot the decision tree of the classifier with a max depth of 3 using the plot_tree function.
h. Calculate and print the feature importance of both classifiers.
i. Calculate the accuracy scores and feature importances of classifiers with different max depths.
j. Plot the accuracy scores and feature importances using the plot_depth_score and plot_feature_importances functions, respectively.

3.5. Data Visualizations

For the data visualisation part, the researcher loads a clean data set from a CSV file and uses different functions to create scatter plots, country-specific scatter plots, and heatmaps to visualise correlations between columns in the pandas DataFrame. To visualise the data, the author would need to follow these steps:

First, set up the environment with things like Pandas, Plotly, Seaborn, and Matplotlib that are needed. Second, load the data from the "clean_data.csv" file into a pandas DataFrame using the pd.read_csv() function. Third, call the create_scatter_plot() function with the DataFrame object and the x and y columns as arguments to create a scatter plot with a trendline for the specified x and y columns. The function uses the px. scatter () function from the plotly.express library to create the plot. Fourth, call the create_country_scatter_plot_to() function with the DataFrame object as an argument to create a scatter plot with trendlines for the Life Expectancy vs. Health Expenditure relationship for four countries (the United States, China, Germany, and South Africa) in the DataFrame. The function filters the DataFrame using the isin() method to select the rows corresponding to the four countries, and then uses the px.scatter() function from the plotly.express

DOI: 10.54254/2754-1169/24/20230424

library to create the plot. Fifth, call the plot_corr_heatmap() function with the DataFrame object, a list of columns to drop (in this case, only the "Year" column), a flag indicating whether to display the correlation coefficients on the heatmap, the name of the matplotlib colormap to use, and the title of the heatmap as arguments to create a heatmap to visualise the correlation matrix of the specified columns in the DataFrame. The function uses the drop () method to drop the specified columns from the DataFrame, calculates the correlation matrix using the corr() method, and then uses the sns.heatmap() function from the Seaborn library to create the heatmap.

4. Results and Analysis

4.1. The trends and Patterns in the Distribution and Outliers of GDP/CO_2/Life Expectancy/Health Expenditure Datasets across 261 Countries from 2000 to 2019

First, by using the frequency distribution graph (Figure 1), the paper explores the distribution of individual life expectancy, individual GDP, individual health expenditure, and individual CO_2 emissions data.

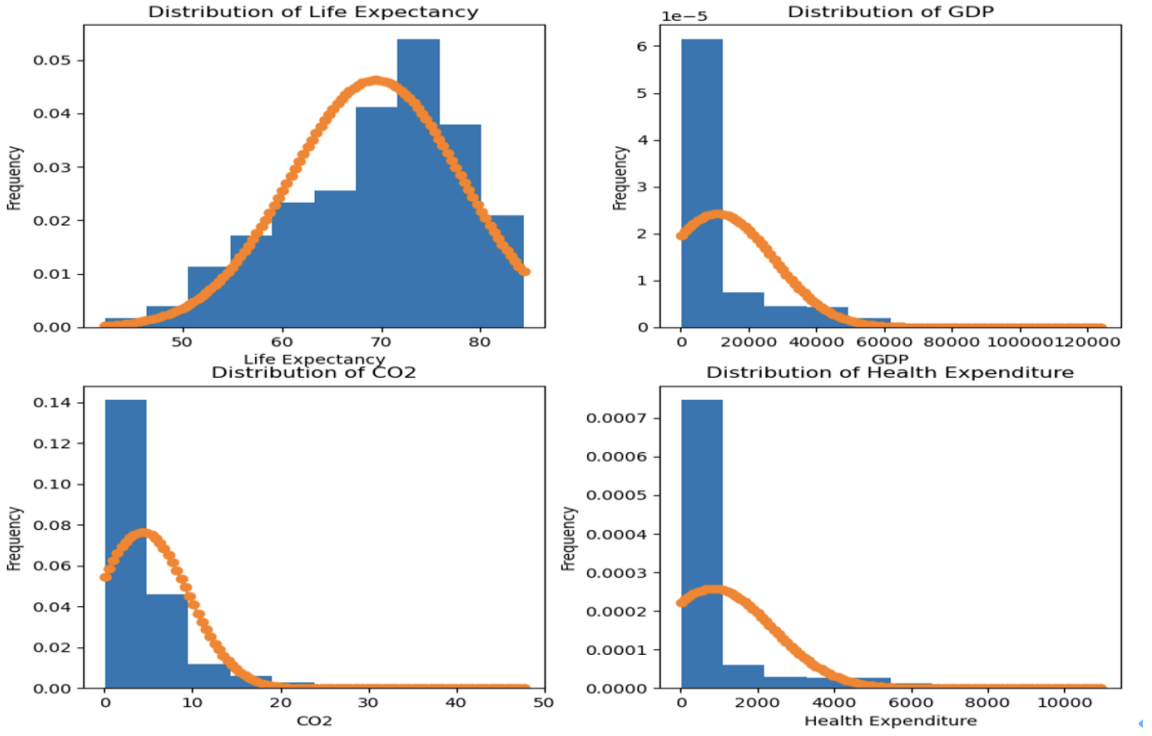

Figure 1: The distribution of individual life expectancy, individual GDP, individual health expenditure, and individual CO2 emissions without removing the outliers (Original).

By analysing the graph (Figure 1), the author was surprised to find that life expectancy, individual GDP, individual health expenditures, and individual CO_2 emissions all belong to skewed distributions. The data on life expectancy belongs to the left-skewed distribution, while individual GDP, individual health expenditures, and individual CO_2 emissions belong to the right-skewed distribution. According to the data on the world's average life expectancy [6], the world's average life expectancy is 72.27 years old. Compared with the data obtained through the study of data distribution, it shows that the world's average life expectancy has increased. Second, by using box plots and frequency distribution plots (Figures 2 and 3), the paper explores outliers in life expectancy, personal health expenditures, and data distributions after outliers are removed.

DOI: 10.54254/2754-1169/24/20230424

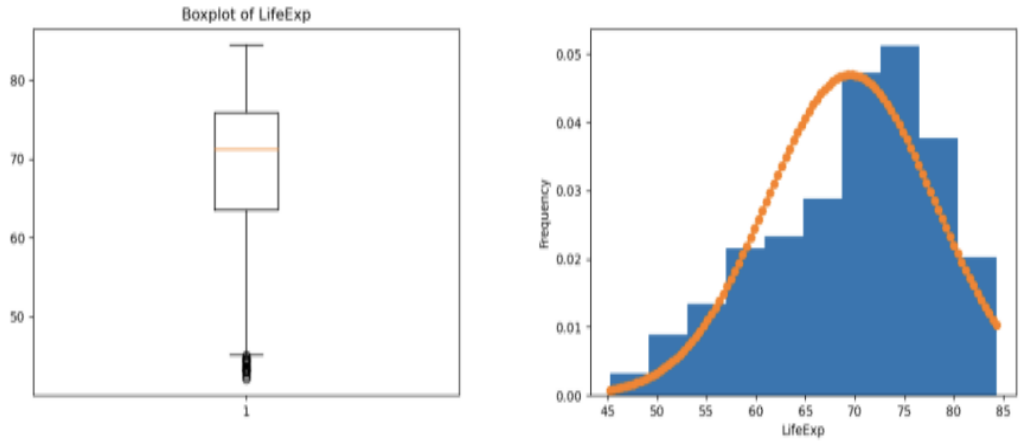

Figure 2: The distribution of the life expectancy after remove the outliers (Original).

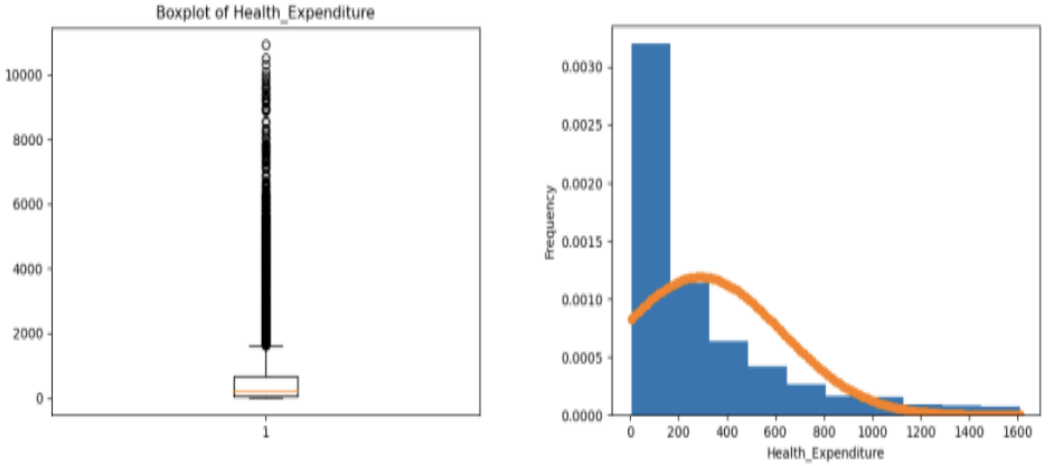

Figure 3: The distribution of the health expenditure after remove the outliers (Original).

For the second question's research on life expectancy and health spending, the author counted and analysed outliers in the data on the life cycle and health spending.Through the boxplot of life expectancy data, the author found that the life expectancy data are mainly concentrated in the early 60s to the end of the 70s, and the outliers are concentrated in those under 50 years old. According to the research [7], the backwardness of the country's development and poor medical and health conditions lead to a low life expectancy. These outliers correspond to the low life expectancies of relatively backward countries in our database.

Through the box plot of the health expenditure data (Figures 2 and 3), the author also found that the health expenditure data is mainly concentrated in the range of 0 USD to 2000 USD, and a large number of outliers are above 2000 USD to 10000 USD. These outliers come from the high cost of living in the developed countries of the world, which reveals the inhomogeneity of world health development. In conclusion, by removing outliers and counting the distribution of life expectancy data, the author found that life expectancy ranges from 0 USD to 2000 USD and is mainly concentrated in the 0-400 USD range.

DOI: 10.54254/2754-1169/24/20230424

4.2. The Effect of an Individual's Health Expenditure on the Individual's Life Expectancy

4.2.1. The Relationship between Health Expenditure and Life Expectancy for Four Countries from 2000 to 2019

By using the scatterplot, this research explores the relationship between health expenditure and life expectancy for four countries: the United States, China, Germany, and South Africa from 2000 to 2019 (Figure 4).

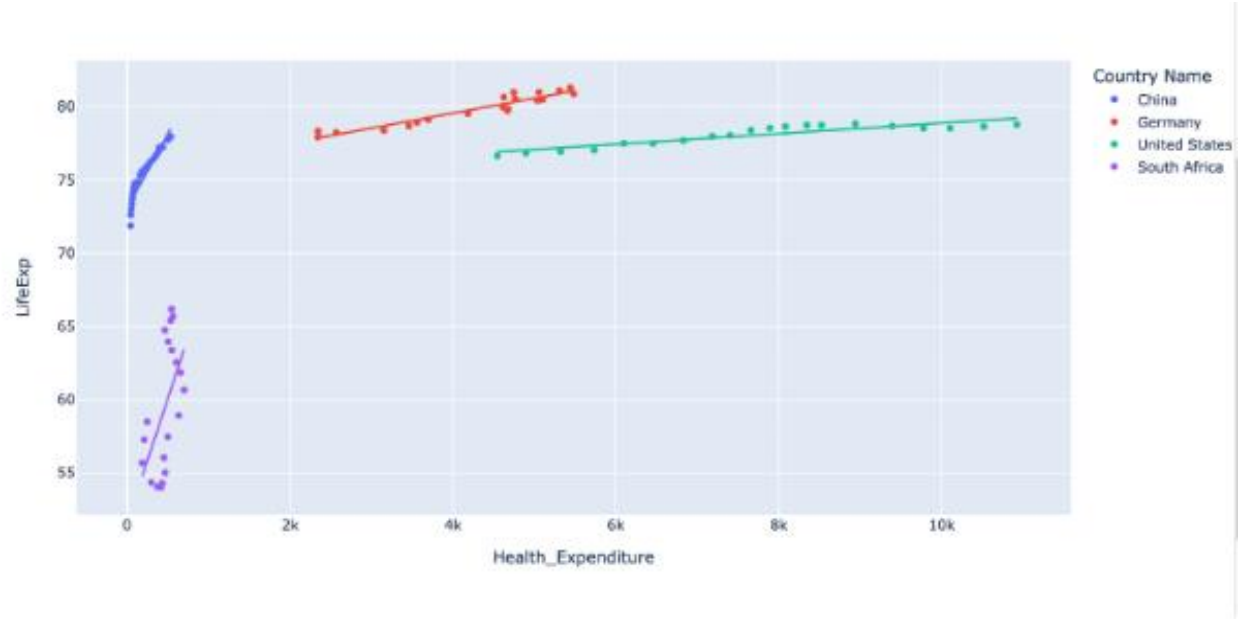

Figure 4: The relationship between life expectancy and health expenditure for four countries from 2000 to 2019 (original).

Researchers, policymakers, and public health officials are interested in the link between health care spending and life expectancy.The United States, China, Germany, and South Africa are all countries with different health systems and levels of economic development, making them potentially interesting case studies for investigating this relationship. Above is a scatterplot depicting the relationship between health expenditure and life expectancy for four representative nations (the United States, China, Germany, and South Africa). A scatterplot of four countries on one graph enables analysis and comparison of the relationship between health expenditure and life expectancy in the four nations.

In all four countries, there is a strong link between how much money is spent on health care and how long people live. This conclusion is consistent with prior research demonstrating a positive correlation between health expenditures and health outcomes, including life expectancy [8]. This conclusion shows how important it is to put healthcare investments at the top of the list in order to improve health outcomes and increase life expectancy. In comparison to other nations, Germany has the greatest healthcare expenditures and life expectancy (about 76–83 years old), while South Africa has the lowest (about 55–63 years old). Nevertheless, it is important to notice that scatterplots may not give adequate information to draw conclusions about the relationship between variables; further research, such as linear regression modelling, is necessary to validate and quantify the association.

DOI: 10.54254/2754-1169/24/20230424

4.2.2. The Relationship between Health Expenditure and Life Expectancy for 261 Countries from 2000 to 2019

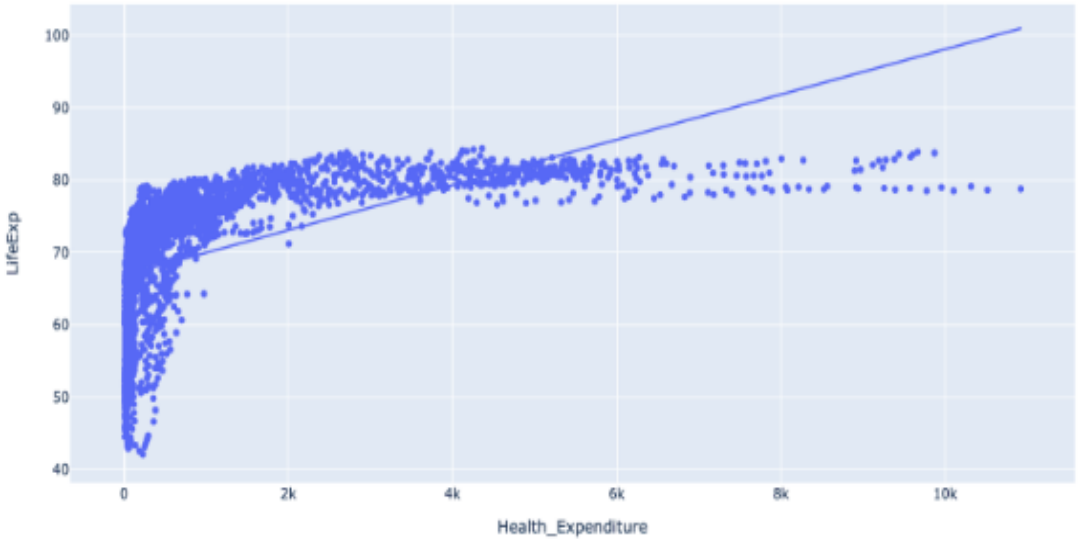

Figure 5: The relationship between health expenditure and the life expectancy for 261 countries from 2000 to 2019 (Original).

The above scatterplot (Figure 5) between personal health expenditure and life expectancy demonstrates that there is no clear linear relationship between the two variables, a finding that differs from the data visualisation in the upper section. The main reason why this phenomenon may occur is that the scope of research the author has chosen is different. In the above part, focus on these four typical countries as research objects. However, in this part, the author extended the scope of the research objects to a range of 261 countries. In addition, according to the previous analysis, the author can conclude that there are some outliers in the data (the health expenditure of the developed countries is much higher than that of other developing countries), which causes the uneven distribution of the scatterplot. The scatterplot in this part indicates that there is no discernible pattern or trend based only on the scatterplot. There may be a correlation between personal health expenditures and life expectancy, but it may not be straightforward or clear. Also, the regression line is distorted and unrealistic. In the figure, when the life cost exceeds $10,000, the life expectancy is as high as 100 years old, which is very unrealistic and rare in reality.

To investigate if an individual's health spending has a major impact on his or her life expectancy. The author can estimate the degree and direction of the association between personal life expenditures and life expectancy using a linear regression model.

4.2.3. The Relationship between Health Expenditure and Life Expectancy for 261 Countries from 2000 to 2019

Scatterplots are useful exploratory tools, but machine learning models are necessary for rigorous analysis. The above scatterplots provide limited information about the strength and direction of the relationship between the independence and dependence variables; the author cannot determine the statistical significance or identify nonlinearities. Thus, the author decided to use the regression model for a more in-depth and comprehensive analysis by quantifying the relationship and identifying interactions so that the author could make predictions based on the relationship.

DOI: 10.54254/2754-1169/24/20230424

Table 1: The linear regression model results between the independent variables for life expectancy and the dependent variable for health expenditure (Original).

	Coef	Std err	t	P>|t|	[0.025 0.975]
Intercept	66.8389	0.121	553.060	0.000	[66.602 67.076]
Health_Expenditure	0.0031	6.87e-05	45.473	0.000	[0.003 0.003]

From the linear regression model above (Table 1), the paper can conclude that there is a positive correlation between individual health expenditure and individual life expectancy. Thus, an individual's health expenditure has a positive effect on their life expectancy. One unit increase in life expenditure will lead to a 0.0031-year increase in life expectancy. When people do not spend any money on their health, they are expected to live to be about 66 years old. The statistical significance of the results was assessed using a p-value and a t-test. The results of this model have statistical significance since the p value is smaller than the alpha the author used. The t test for this model is 2.581 based on the degree of freedom 4520 and alpha 0.05. The t value for health expenditure is 45.473. The t value of the intercept is 553.06. Both the coefficient and intercept t values are bigger than the t test value, suggesting that both of them are statistically significant. The null hypothesis can be rejected based on these results.

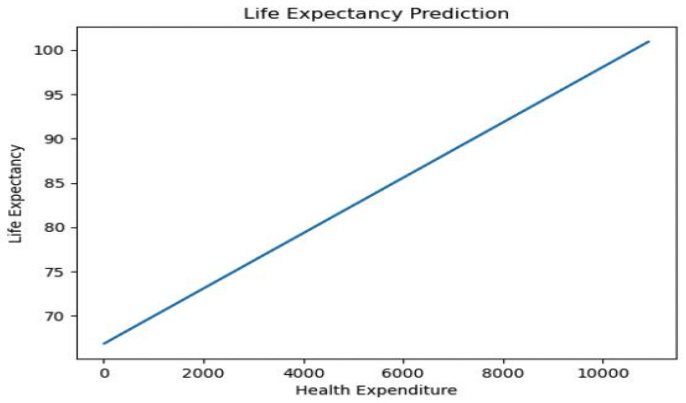

Figure 6: The relationship between the health expenditure and the life expectancy (Original).

Based on the linear regression model, the author created a line graph (Figure 6). According to the above line graph, there is a positive correlation between health expenditures and life expectancy. In particular, the graph demonstrates that spending around $1000 on health is related to an average life expectancy of 69 years. When health expenditures rise, so does life expectancy, showing that people who are willing to spend more on their health may attain a longer life span. Notably, the graph depicts a maximum life expectancy of 100 years, which is an expected constraint given that no amount of health expenditure can guarantee indefinite longevity. The author split the dataset into eighty and twenty percent. The adjusted root mean square error of this model is 0.169112, indicating that the model can well predict life expectancy based on health expenditure. The root mean square root errors for training and testing data are 7.072511261847639 and 7.413254432192041, respectively. The difference between the two root mean square errors is not significant, so the model's accuracy is reliable enough to predict life expectancy. Although the model reveals an association between health spending and life expectancy, it does not account for other factors that may impact life expectancy, such as genetics, lifestyle factors, and environmental factors, which may limit the generalizability of the results.

DOI: 10.54254/2754-1169/24/20230424

4.3. The Impact of Personal GDP, GDP Level and CO_2 Emissions on People's Life Expectancy

4.3.1. The Impact of Individual CO_2 Emissions and Individual GDP on the Life Expectancy of 261 Countries from 2000 to 2019

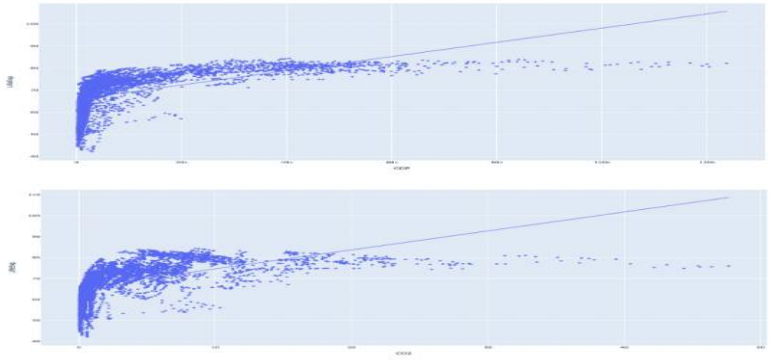

Figure 7: The relationship between CO_2 emissions, GDP and the life expectancy for 261 countries from 2000 to 2019 (original).

Personal GDP and CO_2 emissions are also two important factors that may impact each person's life expectancy. To see how these variables are related, the author can use scatterplots (Figure 7) to show the data in a visual way. The scatterplot between personal GDP and life expectancy shows that there is no obvious linear relationship between the two variables. Besides, as was said in the paper, the regression line in these scatterplots is off and doesn't make sense. In the first figure, when the GDP value exceeds 80k, the life expectancy is as high as 90 years old, which is very unrealistic and rare in reality. This means that a simple glance at the scatterplot cannot reveal any discernible pattern or trend. However, there may still be a relationship between personal GDP and life expectancy, but it may not be a simple or direct one. Similar to this, the scatterplot demonstrates that there is no obvious linear relationship between personal CO_2 emissions and life expectancy. Most importantly, when an author studies how two variables (GDP and CO_2) affect life expectancy, they can't just look at how one variable (GDP or CO_2) affects life expectancy on its own. It needs to consider the two variables together and how these two variables together affect life expectancy.

To explore the question of how personal GDP and CO_2 emissions impact each person's life expectancy, the author can use a linear regression model. The author will be able to figure out the strength and direction of the link between personal GDP, CO_2 emissions, and life expectancy by using this model.

4.3.2. The Impact of Individual CO_2 Emissions and Individual GDP on the Life Expectancy of 261 Countries from 2000 to 2019

Scatterplots are useful for figuring out what's going on, but for a thorough analysis, it needs machine learning models. Even though the above scatterplots only show a small amount of information about the strength and direction of the relationship between the independent and dependent variables, the author cannot determine statistical significance or find nonlinearities. Thus, the author decided to use the regression model for a more in-depth and comprehensive analysis by quantifying the relationship and identifying interactions so that the author could make predictions based on the relationship.

There are two measures to determine if the GDP is high or low in one region. The first is to convert GDP per capita to GDP level. That is, creating GDP growth rates and taking the growth rate higher than 2 percent as a high GDP region and the growth rate lower than 2 percent as a low GDP region.

DOI: 10.54254/2754-1169/24/20230424

The other method is directly looking at the GDP per capita, and the higher the GDP per capita is, the better the GDP in that region.

Table 2: The linear regression model results between the independent variables for life expectancy and the dependent variables for GDP level and CO_2 emissions (Original).

	coef	Std err	t	P>|t|	[0.025 0.975]
Intercept	65.5928	0.160	409.111	0.000	[65.278 65.907]
GDP_Level	-0.0273	0.224	-0.122	0.903	[-0.466 0.411]
CO_2	0.9046	0.021	43.926	0.000	[0.864 0.945]

The paper initially included GDP level, CO_2 emissions, and life expectancy in the statistical model (Table 2). However, the p-value for the lower level GDP coefficient was found to be 0.903, which is greater than the alpha value of 0.05. This suggests that the author cannot reject the null hypothesis and that the coefficient is not statistically significant. Consequently, this model cannot be used to predict life expectancy, and the relationship between GDP, CO_2 emissions, and life expectancy should be approached with caution. However, it is important to note that this model is not entirely useless. It indicates that there is no significant relationship between GDP level and life expectancy, which may be due to the fact that a good GDP growth rate does not necessarily translate to high income (GDP) for individuals in a given region. Thus, it may be more reasonable to use GDP per capita values directly in the model to analyse this relationship.

Table 3: The linear regression model results between the independent variables for life expectancy and the dependent variables for GDP and CO_2 emissions (original).

	coef	Std err	t	P>|t|	[0.025 0.975]
Intercept	65.1501	0.128	507.521	0.000	[64.898 65.402]
GDP	0.0002	8.07e-06	29.153	0.000	[0.000 0.000]
CO_2	0.4079	0.025	16.034	0.000	[0.358 0.458]

Based on this model (Table 3), the author can conclude that there is a positive relationship between CO_2 emissions, GDP per capita, and life expectancy. The CO_2 emissions coefficient of 0.4079 suggests that a one-unit increase in CO_2 emissions is associated with a 0.4079 increase in life expectancy. In the same way, the GDP per capita coefficient of 0.0002 suggests that a one-dollar increase in GDP per capita is linked to a 0.0002 increase in life expectancy. The intercept of the regression at 65.1501 suggests that in the absence of CO_2 emissions and GDP, people are expected to live up to 65 years. The p value is zero for every coefficient and interception. The t test for this model is 2.581 based on the degree of freedom 4520 and alpha 0.05. The t value for CO_2 emissions is 16.034. The t value for GDP is 29.153. The t value of the intercept is 507.521. Since the p value is less than the alpha the author used in this model and the t value for the coefficient and intercept is bigger than the t test value. The values in this model are statistically significant. The author split the dataset into eighty and twenty percent. The adjusted root mean square error of this model is 0.15679917, indicating that the model can well predict life expectancy based on health expenditure. The root mean square root errors for training and testing data are 6.639395757097121 and

DOI: 10.54254/2754-1169/24/20230424

6.56691455560923, respectively. The difference between the two root mean square errors is not significant, so the model's accuracy is reliable enough to predict life expectancy.

However, it is important to note that this model may have limitations. For example, it only considers the relationship between CO_2 emissions, GDP per capita, and life expectancy and does not account for other factors that may affect life expectancy, such as healthcare, education, and social policies. Additionally, the model assumes a linear relationship between the variables, which may not always hold in reality. Therefore, the conclusions drawn from this model should be interpreted in the context of its limitations.

4.4. The Impact of Independent Variables (GDP, CO_2, and Health Expenditure) on an Individual's Life Expectancy and the Effect of Changing the Max Depth of the Decision Tree on Its Accuracy and Feature Importance

4.4.1. The Relationship between Individual GDP, Individual CO_2 Emissions, and Individual Health Expenditure and the Individual Health Expectancy of 261 Countries from 2000 to 2019

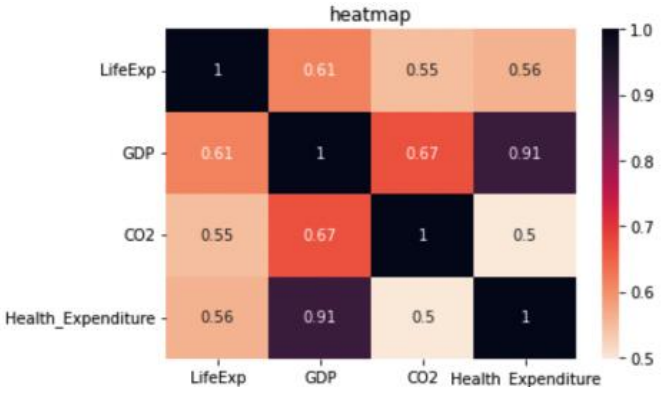

Figure 8: The relationships between the independent variables for life expectancy and the dependent variables for GDP, health expenditure and CO_2 emissions (Original)

This analysis utilised a heat map (Figure 8) with life expectancy, GDP, CO_2 emissions, and health expenditure as its x and y axes, forming a 4x4 data matrix. This heat map reveals that the variable having the largest influence on life expectancy is individual GDP, with a value of 0.61. Hence, there is a substantial positive correlation between GDP and life expectancy; as GDP rises, so does life expectancy. The other factors in the heat map have a lesser effect on life expectancy. Surprisingly, there is also a positive correlation between CO_2 emissions and life expectancy (0.55), although it is smaller than the correlation between GDP and life expectancy. Health expenditure has a value of 0.56, indicating a positive association between health expenditure and life expectancy; however, this relationship is weaker than the one between GDP and life expectancy.

Even though a heat map can give an overview of the relationship between things like a person's life expectancy, GDP, CO_2 emissions, and health care costs, it may not be able to give an exact or precise assessment of how all of these things affect a person's life expectancy. In other words, the heat map can only give a basic idea of how strongly these variables are related. It can't look at how three variables affect life expectancy all at once. Hence, a decision tree model may provide a more precise estimate of a person's life expectancy depending on their CO_2 emissions, GDP, and health expenditure. The decision tree approach may identify the most significant factors and their interactions, allowing for a more accurate estimation of an individual's life expectancy. In addition,

DOI: 10.54254/2754-1169/24/20230424

the decision tree model handles missing data more efficiently than a heat map or other graphical depiction, which is particularly important when working with actual data.

4.4.2. The Impact of the Individual CO_2 Emissions, Individual GDP, and Individual Health Expenditure on the Dependent Variable of Individual Life Expectancy

The above heat map is a useful exploratory tool, but machine learning models are necessary for rigorous analysis. Although the above scatterplots provide limited information about the strength and direction of the relationship between the independence and dependence variables, the author cannot determine the statistical significance or identify nonlinearities. Thus, the author decided to use the decision tree model for a more in-depth and comprehensive analysis by quantifying the relationship and identifying interactions so that the author can make predictions or classifications based on the relationship.

First of all, the first step in creating a decision tree model is to determine the maximum depth of the model and convert our continuous variable into a discrete variable for classification. For example, in the question the author explored, the author classified the dependent continuous variable life expectancy into the following groups: younger than 50 years old; 50–60 years old; 60–70 years old; 70–80 years old; older than 80 years old.

When the author determines the maximum depth of the model, the author first creates two decision tree classifiers, one with a maximum depth of 3 (with an accuracy score of 74.14%) and one without limiting the maximum depth (with an accuracy score of 83.49%). But the author found that when the maximum depth of the model is set to 3, the features of importance for GDP, CO_2, and health expenditure are 0.0, 0.37, and 0.63, respectively. It can be found that the feature importance of GDP is only 0.0. On the other hand, when the maximum depth of the model I created is not set, the feature importance of GDP, CO_2, and health expenditure is 0.14, 0.37, and 0.49, respectively. Through this interesting phenomenon, the author realises that the possible reasons for this phenomenon are: First, the impact of the GDP itself on life expectancy is very small, causing the result to be only 0.0. Next, the max depth of the initially established model is only 3 layers, and the accuracy of 3 layers (with an accuracy score of 74.14%) is lower than the accuracy of the model without limiting the max depth (with an accuracy score of 83.49%).

Therefore, based on this conjecture, the author conducted further verification. The author set the maximum depth of the model to "12" to ensure that the maximum depth of the model set is interpreted without caution, so it avoids overfitting the training data. At the same time, the model can eliminate the deviation because the number of maximum depth layers is high enough. Therefore, the author used this idea to do the visualisation to explore the impact of the max depth of the model on the features and importance of variables and what value the max depth of the model should use to ensure that the accuracy of the model is as high as possible.

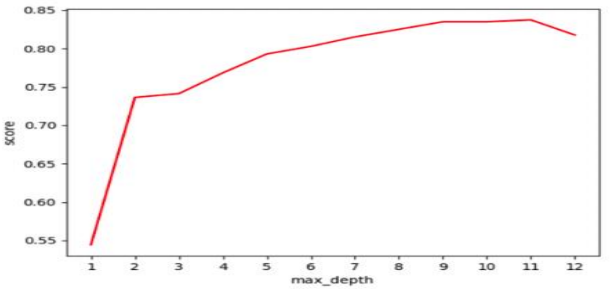

Figure 9: The relationship between the max depth of the decision tree model and the accuracy score of the model (Original).

DOI: 10.54254/2754-1169/24/20230424

In order to better express the relationship that the maximum depth of the decision tree model has a significant impact on the accuracy of the model, the author shows a linear graph (Figure 9) above representing the accuracy score of the model and how accuracy changes with maximum depth changes.

Through the above image of the model, the relationship between the maximum depth of the model and its accuracy can be observed. When the maximum depth of the model is 1, the corresponding model accuracy is about 55%. And when the max-depth of the model changes from 1 layer to 2-3 layers, the accuracy of the model has a significant increase, rising to about 74%. But it is worth noting that the accuracy of the model rises to about 83% when it has about 9–10 layers.

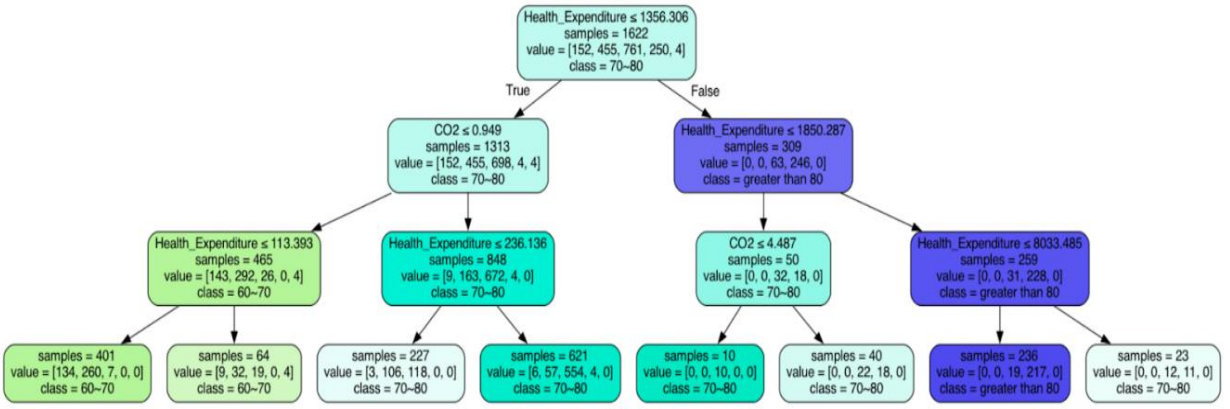

Figure 10: The decision tree model results between the independent variables for life expectancy and the dependent variables for GDP, health expenditure and CO_2 emissions (Original).

From the standpoint that the report image can be readily navigated, the decision tree model above (Figure 10) is the model when the model max depth is set to "3" so that the data in the model can be seen more clearly. The decision tree analysis in this code suggests that the most important factor impacting life expectancy is health expenditure. If health expenditure is less than or equal to 1356.306, then the tree splits further based on the CO_2 emissions. If CO_2 emissions are less than or equal to 0.949, the model then splits again based on the health expenditure variable. If health expenditure is less than or equal to 113.393, then the predicted life expectancy is in the 60-70 age range. If health expenditure is greater than 113.393, then the predicted life expectancy is in the 70–80 age range. On the other hand, if CO_2 emissions are greater than 0.949, then the predicted life expectancy is in the 70–80 age range, regardless of health expenditure.

Interestingly, the decision tree model suggests that personal GDP does not have a significant impact on life expectancy. This is evident as the model does not split based on the personal GDP variable. Furthermore, the model suggests that increasing CO_2 emissions may actually have a positive impact on life expectancy. However, it is important to note that the increase in life expectancy is marginal (0.4742 years for each unit increase in CO_2 emissions), and other factors such as health expenditure still have a larger impact on life expectancy. Overall, this decision tree model can provide insights into the complex relationship between different variables and their impact on life expectancy.

DOI: 10.54254/2754-1169/24/20230424

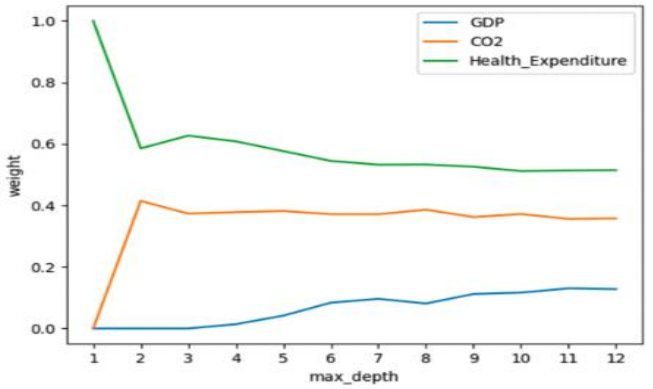

Figure 11: The relationship between the max depth of the decision tree model and the feature importance of three dependent variables (Original).

Additionally, in order to better express the relationship that the max depth of the decision tree model has a significant impact on the feature importance of independent variables, the author shows a linear graph above representing the feature importance of the three independent variables changing with max depth changes.

Through the above image, the author can see that the impact of individual GDP on the life expenditures of individuals is very small. This above line graph also backs up the conclusion from the decision tree graph that the author can't see much data about how GDP affects how much people spend on their lives. Besides, in the line graph, the author can clearly see that although GDP's feature importance has always been stable, compared with the other two variables, its feature importance has always been the smallest. When the maximum depth is 1 to 5, its feature importance is always less than 0.1. Until the maximum depth is 9, its feature importance will reach about 0.15. In contrast, the importance of CO_2 emissions is higher than that of GDP. It is worth noting that when the model's max-depth ranges from 1 to 2, its feature importance increases from 0.0 to 0.4, a dramatic increase. When the model's maximum depth reaches 3, it gradually tends to be stable. Among all independent variables, health expenditure has the greatest impact on life expectancy. Unlike the other two independent variables (CO_2 emissions and GDP), the feature importance of health expenditure shows a downward trend when the model max-depth is 1 to 2, decreasing from 1.0 to 0.6. However, with the increase in maximum depth, it also maintains the highest feature importance. It is noticeable that the feature importance of the three independent variables remains relatively stable after the maximum depth of the decision tree model is "9".

5. Conclusion

The decision tree model the author established can help the author understand the relationship between life expectancy and independent variables (GDP, CO_2, and health expenditure). Based on the model the author has built, the author can draw the following conclusions:

First, of the three independent variables, a person's health care spending has the biggest effect on his or her life expectancy. This is followed by a person's CO_2 emissions and then a person's GDP, which proves the same thing that the linear regression model did in the previous problems.Moreover, the author also found that the maximum depth of the decision tree model has a significant impact on the accuracy of the model and the feature importance of independent variables. The results suggest that when the max depth of the model is about 9 or 10, it is more accurate than the model with a max depth of 3, and the feature importance of GDP is higher than the result of the max depth of 3.

DOI: 10.54254/2754-1169/24/20230424

Despite the benefits of the decision tree model, it has some limitations. It requires a relatively large dataset, and the model may overfit the training data. In addition, decision trees are vulnerable to instability, and slight changes in the data can lead to completely different models. Moreover, decision trees are susceptible to bias towards dominant classes and may not be suitable for complex datasets. Finally, while the model can identify correlations between variables, it cannot prove causality. Therefore, it is essential to consider these limitations when using decision trees in real-world applications.

In the end, the research tells us a lot about the factors that affect life expectancy and how they vary between countries, regions, and years. The information might be used by policymakers to build targeted initiatives that could enhance health outcomes, but they also need to consider other aspects when making judgements. When adopting the study's results as a basis for their own work, researchers should be cognizant of the study's limitations.

References

[1] *Zhang, Y., & Shao, S. (2021). The impact of carbon dioxide emissions on global health: a systematic review. International Journal of Environmental Research and Public Health, 18(1), 116.*

[2] *Ezzati, M., & Kammen, D. M. (2002). The health impacts of exposure to indoor air pollution from solid fuels in developing countries: knowledge, gaps, and data needs. Environmental Health Perspectives, 110(11), 1057-1068.*

[3] *Bloom, D. E., Canning, D., & Sevilla, J. (2004). The effect of health on economic growth: a production function approach. World Development, 32(1), 1-13.*

[4] *World Health Organization. (2018). World health statistics 2018: monitoring health for the SDGs. World Health Organization.*

[5] *OECD. (2019). Health at a Glance 2019: OECD indicators. OECD Publishing.*

[6] *Earth - Place Explorer. (n.d.). Data Commons. Retrieved March 6, 2023, from https://datacommons.org/place/Earth?utm_medium=explore&mprop=lifeExpectancy&popt=Person&hl=en*

[7] *Roser, M., Ortiz, E., & Ritchie, H. (n.d.). Life Expectancy. Our World in Data. Retrieved March 6, 2023, from https://ourworldindata.org/life-expectancy#citation*

[8] *Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., ... & Zurayk, H. (2019). Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. The Lancet, 376(9756), 1923-1958.*