STRATEGIC THINKIG

CAPSTONE PROJECT

LIFE EXPECTANCY MACHINE LEARNING PREDICTION MODELS

ADEWALE TAOFIQ ARIWO-OLA
STUDENT NO: SBS23043

Table of Contents

ABSTRACT

This study explores the application of machine learning models to predict life expectancy based on various set of features such as Infant deaths, adults' mortality, alcohol consumption, incidents of HIV, GDP per capital, schooling that are presented in a data set downloaded from dataset repo sitory Kaggle. The dataset utilized for this research comprises a comprehensive collection of heal th-related attributes obtained from diverse populations of different countries. Exploratory Data A nalysis and feature engineering techniques are applied to enhance model performance, and exten sive preprocessing is conducted to handle imbalance class in target variable and outliers. The mo dels are trained and validated using a robust set of evaluation metrics, including mean squared er ror, accuracy, Precision, F1-Score and Recall. Results demonstrate the efficiency of machine lear ning models in accurately predicting life expectancy, showcasing the potential for personalized h ealth interventions and resource allocation. Additionally, interpretability analyses are conducted t o elucidate the key features influencing the predictions, providing valuable insights for healthcar e professionals and policymakers. The study also discusses challenges associated with implement ing life expectancy prediction models in real-world healthcare settings. In conclusion, this resear ch contributes to the growing body of knowledge in the field of predictive healthcare analytics, e mphasizing the promising role of machine learning in advancing our understanding of life expect ancy determinants.

INTRODUCTION

The aim of the capstone project is to apply skills obtained during my learning for Higher Diploma Data Analytics for Business to evaluate a possibility to predict a chosen target variable based on other independent variables. I will use data exploratory analysis (EDA) and statistical methods to understand patterns and trends of a dataset and apply a predictor machine learning model on the processed data.

The data "Life Expectancy data" used for the initial phase of the capstone was downloaded from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who and has been adjusted into a new dataset "Life Expectancy-Data-Updated"- https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated. Data contains life expectancy, health, immunization, and economic and demographic information about 179 countries from 2000-2015 years with 21 variables and 2864 rows. The database has one variable that categorizes countries into two groups: Developed vs Developing countries. Life expectancy is a statistical measure that represents the average number of years a person is expected to live based on various factors such as their birth year, gender, and other demographic characteristics. It is typically expressed as an average number of years and is often used as an indicator of the overall health and quality of life in a particular country or region.

The capstone project is executed using the CRISP DM methodology. As expected, changes happened through some stages of the project as I need to re-assess and re-evaluate some of these stages and come up with alternative means of achieving the project goals.

Life expectancy is a key summary measure of the health and wellbeing of a population. A nation's life expectancy reflects its social and economic conditions and the quality of its public health infrastructure, among other factors. Monumental improvements in life expectancy have been the predominant trend for higher income, developed countries over the course of the 20th and 21st centuries. (J*essica Y Ho and Arun S Hendi, 2018)*

The United Nations Committee for Development Policy review the list of Low Developed Countries every three years and make recommendations on the inclusion and graduation of eligible countries using the following criteria.

1. Income: Measured by the Gross National Income (GNI) per capita

GNI per capita provides information on the income status and the overall level of resources available to a country.

2. Human Assets: Measured by the Human Asset Index (HAI)

The HAI is a measure of level of human capital. Low levels of human assets indicate major structural impediments to sustainable development and a lower HAI represents a lower development of human capital. (Fig. 1)

3. Economic and Environmental Vulnerability: Measured by the Economic and Environmental Vulnerability Index (EVI)

The EVI is a measure of structural vulnerability to economic and environmental shocks. High vulnerability indicates major structural impediments to sustainable development. A higher EVI represents a higher economic vulnerability. (Fig.2)

Fig.1: Human Asset Index (HAI)

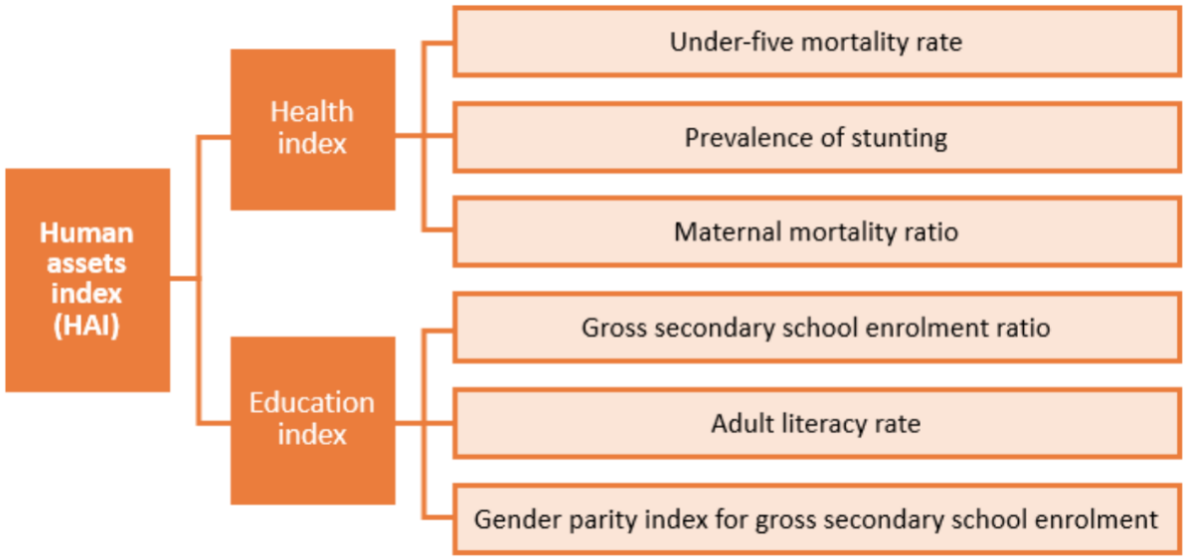
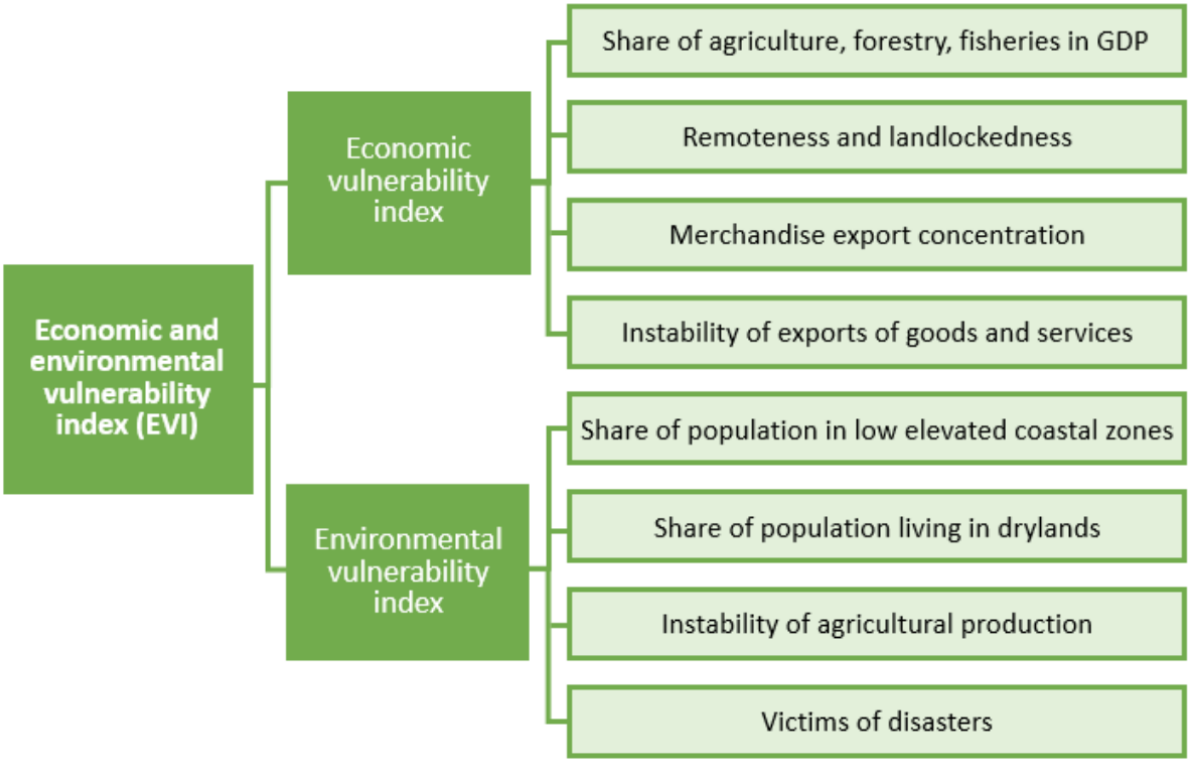


Fig.2: Economic and Environmental Vulnerability Index (EVI)



Source: UN DESA (https://www.un.org/development/desa/dpad/least-developed-country-category/ldc-criteria.html?target=human-assets)

HYPOTHESIS

Can socioeconomic, health, environmental and geographical location criteria be used to predict if a populations life expectancy is low or high?

GENERAL GOAL

The goal of this project is to analyze the dataset and generate a synthetic data to be used to construct a machine learning model that can predict life expectancy in regions of countries around the world based on factors that are represented in the dataset.

SUCCESS CRITERIA/INDICATORS

I was able to establish correlation between life expectancy and chosen socioeconomic, geographical location and health conditions.

Random Forest Model was used to predict if a populations life expectancy is low or high with an outcome of 100% accuracy.

Independent features used for predicting country status for the machine learning model were extracted based on their importance.

SHAP (Shapley Additive exPlanations) was used to interpret the machine learning model used and understand the impact of individual features on the model's prediction.

DATA SOURCE

The data was obtained from a data set repository Kaggle "Life Expectancy-Data-Updated"- https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

TECHNOLOGIES USED

There are certain sets of Python libraries that is collectively labeled the Scientific Stacks. This stack comprises of, among others, the following:

NumPy: NumPy provides a multidimensional array object to store homogenous or heterogenous data; it also provides optimized functions/methods to operate on this array object.

Matplotlib: This is the most popular plotting and visualization library for Python, providing both 2D and 3D visualization capabilities.

Pandas: Pandas builds on NumPy and provides richer classes for the management and analysis of time series and tabular data; it is tightly integrated with matplotlib for plotting and PyTable for data storage and retrieval. (Python for Finance. Analyze Big Financial Data, Yves Saarland, 2014.Pg 16 & 17)

Seaborn is a Python library built on top of matplotlib and allows you to easily produce prettier (and more complex) visualization (Data Science from Scratch. Joel Grus,2015. Pg.83)

Also, we are importing machine learning library for Python such as Scikit-learn, often called sklearn. provides simple and efficient tools for data analysis and modelling, including various machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more.

SciPy is an open-source library for mathematics, science, and engineering. It is built on top pf NumPy library and provides additional functionality for wide range of scientific and engineering applications.

Machine Learning models were built using Decision Tree and Random Forest Classifier algorithms which builds multiple decision trees and combines their predictions to improve the accuracy and reduce overfitting.

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is a useful method generally employed to understand data set by summarizing main characteristics of the data set and often visualizing by plotting.

Making informative visualizations (sometimes called plots) is one of the most important tasks in data analysis. It may be a part of the exploratory process-for example, to help identify outliers or needed for data transformations, or as a way of generating ideas for models. (Wes McKinney, 2017)

Python has become a component of the much broader Jupyter open-source project, which provides a productive environment for interactive and exploratory computing. Its oldest and simplest mode is as an enhanced Python shell designed to accelerate the writing, testing and debugging of Python code. IPython system can be used through Jupyter Notebook, an interactive web-based code “notebook” offering support for dozens of programming languages. The IPython shell and Jupyter notebooks are especially useful for data exploration and visualization. (Wes McKinney, 2017)

All data analysis and visualization for completing this project are generated using Python.

Python contains lots of data science libraries, frameworks, modules, and toolkits that efficiently implement the most common as well as least common data science algorithm and techniques. (Joel Grus, 2015)

DATA UNDERSTANDING

The dataset has 2864 entries in 21 columns of which float64(11), int64(8), object (2). There are neither missing values nor duplicated rows in the dataset.

DATA PREPARATION AND PROCESSING

In this phase, the dataset will be converted into meaningful and useful information. This involves series of operation that manipulates, analyze, and transform data to extract insight, make decisions or generate reports. For the dataset, some aspects of the data processing phase used are data cleaning, data transformation, data analysis and data visualization.

To prepare our dataset for insightful analysis and importing it to a machine learning model, the following steps were taken:

- Number of variables (columns) was reduced from 21 to 19 by dropping features ‘Year* and ‘Countries’ not significant to the project goal.
- Features in the target variable (Life Expectancy) columns were categorized into ‘Low = below 64’ or ‘High = over 64’. The 2020 average normal retirement age across OECD countries for an individual with a full career and who entered the labor market at age 22 was equal to 63.4 years for women and 64.2 years for men.
- A new column was created for the target variable ‘Life expectancy category’ and text classification of the target variable was replaced with 0 and 1 for each category of low and high respectively.
- The dataset was analyzed for duplicates, but there were duplicated rows identified.
- Data in all columns, apart from "Life expectancy" was converted to floats to ensure it could be used with a wide range of models without any compatibility issue.
- Data was shuffled and scaled using Standardization algorithm StandardScaler. Data standardization is an important technique in data preprocessing. The goal of data standardization is to transform the numeric variables so that each variable has zero mean and unit variance.

BASIC DATA EXPLORATORATION

In this phase, I am gauging the overall data. The volume of data and the types of columns present in the data.

This initial assessment of the data is done to identify which columns are Quantitative, Categorical or Qualitative.

All data analysis and visualization are generated using Python.

Python contains lots of data science libraries, frameworks, modules, and toolkits that efficiently implement the most common as well as least common data science algorithm and techniques. (Joel Grus, 2015)

This step helps to start the column rejection process. I am looking at each column carefully and ask, does this column affect the values of the Target variable? In this case I am asking the question does this column affect the life expectancy of a population? If the answer is a clear "No", then I will remove the column immediately from the data, otherwise I am keeping the column for further analysis.

There are four commands which are used for Basic data exploration in Python.

head() : This helps to see a few sample rows of the data

info() : This provides the summarized information of the data

describe() : This provides the descriptive statistical details of the data

nunique(): This helps us to identify if a column is categorical or continuous

Basic data exploratory analysis shows the data contains 21 columns of variables and 2864 rows of observations with categorical and numerical datapoints. There are no missing values in all columns and rows are not duplicated in the dataset.

Fig 3: Features in columns and data types.

```
RangeIndex: 2864 entries, 0 to 2863
Data columns (total 21 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Country                         2864 non-null   object
 1   Region                          2864 non-null   object
 2   Year                            2864 non-null   int64
 3   Infant_deaths                   2864 non-null   float64
 4   Under_five_deaths               2864 non-null   float64
 5   Adult_mortality                 2864 non-null   float64
 6   Alcohol_consumption             2864 non-null   float64
 7   Hepatitis_B                     2864 non-null   int64
 8   Measles                         2864 non-null   int64
 9   BMI                             2864 non-null   float64
 10  Polio                           2864 non-null   int64
 11  Diphtheria                      2864 non-null   int64
 12  Incidents_HIV                   2864 non-null   float64
 13  GDP_per_capita                  2864 non-null   int64
 14  Population_mln                  2864 non-null   float64
 15  Thinness_ten_nineteen_years     2864 non-null   float64
 16  Thinness_five_nine_years        2864 non-null   float64
 17  Schooling                       2864 non-null   float64
 18  Economy_status_Developed        2864 non-null   int64
 19  Economy_status_Developing       2864 non-null   int64
 20  Life_expectancy                 2864 non-null   float64
dtypes: float64(11), int64(8), object(2)
```

Using the nunique() command to find unique values for each column to understand which column is categorical and which one is continuous. Typically if the number of unique values are less than 20 then the variable is likely to be a category and numbers greater than 20 are likely to represent continuous variables.

Fig 4: Unique value for each column

Country	179
Region	9
Year	16
Infant_deaths	847
Under_five_deaths	1035
Adult_mortality	2850
Alcohol_consumption	1164
Hepatitis_B	80
Measles	87
BMI	120
Polio	77
Diphtheria	80
Incidents_HIV	393
GDP_per_capita	2564
Population_mln	1803
Thinness_ten_nineteen_years	200
Thinness_five_nine_years	207
Schooling	130
Economy_status_Developed	2
Economy_status_Developing	2
Life_expectancy	396

LOOKING AT THE DESCRIPTIVE STATISTICS OF THE DATASET

Descriptive measures that indicate where the center or most typical value of a data set lies are called measure of central tendency or more simply, measure of center. Measure of centers are often called averages. There are 3 most important measure of center: the mean, median and mode. (Neil A. Weiss (Introductory Statistics, 2017, pg.118)

In Fig 3, we can see the mean of the data showing average of each of the numerical values, Also, standard deviation of each column showing how values varies. This is a measure of dispersion from the average value.

Also displayed are the lower 25% values, average 50% values and upper 75% values representing first quartile (Q1), second quartile (Q2) and thirst quartile (Q3) respectively.

The standard deviation measures variation by indicating how far, on average, the observation is from the mean. For a data set with a large amount of variation, the observation will, on average, be far from the mean; so, the standard deviation will be large. For a data set with a small amount of variation, the observation will, on average, be close to the mean; so, the standard deviation will be small (Neil A. Weiss, Introductory Statistics, 2017 pg. 129)

There are inconsistencies in the min and max values in the data. The max and min values of one feature are significantly larger than the other features. These features are standardized using StandardScaler.

StandardScaler is a processing technique used in machine learning and statistics to standardize the features in a dataset. StandardScaler in scikit-learn ensures that for each feature the mean is 0 and variance is 1 (Andreas C. Müller & Sarah Guido, 2017.pg133)

Fig 5: Descriptive Statistics of dataset

	count	mean	std	min	25%	50%	75%	max
Year	2864.0	2007.500000	4.610577	2000.000	2003.75000	2007.5000	2011.250000	2015.0000
Infant_deaths	2864.0	30.363792	27.538117	1.800	8.10000	19.6000	47.350000	138.1000
Under_five_deaths	2864.0	42.938268	44.569974	2.300	9.67500	23.1000	66.000000	224.9000
Adult_mortality	2864.0	192.251775	114.910281	49.384	106.91025	163.8415	246.791375	719.3605
Alcohol_consumption	2864.0	4.820882	3.981949	0.000	1.20000	4.0200	7.777500	17.8700
Hepatitis_B	2864.0	84.292598	15.995511	12.000	78.00000	89.0000	96.000000	99.0000
Measles	2864.0	77.344972	18.659693	10.000	64.00000	83.0000	93.000000	99.0000
BMI	2864.0	25.032926	2.193905	19.800	23.20000	25.5000	26.400000	32.1000
Polio	2864.0	86.499651	15.080365	8.000	81.00000	93.0000	97.000000	99.0000
Diphtheria	2864.0	86.271648	15.534225	16.000	81.00000	93.0000	97.000000	99.0000
Incidents_HIV	2864.0	0.894288	2.381389	0.010	0.08000	0.1500	0.460000	21.6800
GDP_per_capita	2864.0	11540.924930	16934.788931	148.000	1415.75000	4217.0000	12557.000000	112418.0000
Population_mln	2864.0	36.675915	136.485867	0.080	2.09750	7.8500	23.687500	1379.8600
Thinness_ten_nineteen_years	2864.0	4.865852	4.438234	0.100	1.60000	3.3000	7.200000	27.7000
Thinness_five_nine_years	2864.0	4.899825	4.525217	0.100	1.60000	3.4000	7.300000	28.6000
Schooling	2864.0	7.632123	3.171556	1.100	5.10000	7.8000	10.300000	14.1000
Economy_status_Developed	2864.0	0.206704	0.405012	0.000	0.00000	0.0000	0.000000	1.0000
Economy_status_Developing	2864.0	0.793296	0.405012	0.000	1.00000	1.0000	1.000000	1.0000
Life_expectancy	2864.0	68.856075	9.405608	39.400	62.70000	71.4000	75.400000	83.8000

DEFINING THE PROBLEM STATEMENT

I am creating a categorical machine model which can predict life expectancy of a population in regions of countries around the world.

I started by grouping values in the target variable "life expectancy" into low vs high. The threshold was defined as 64. This is based on the world's average retirement age according to the OECD (Organization for Economic Cooperation and Development).

Predictors: Country, Region, Year, Infant deaths, Under-five deaths, Adult mortality, Alcohol co nsumption, Hepatitis B, Measles, BMI, Polio, Diphtheria, Incidents HIV, GDP_per_capita, Popul ation, Thinness_ten_nineteen_years, Thinness_five_nine_years, Schooling, Economy_status_De veloped, Economy_status_Developing,

Target Variable: Life Expectancy

REMOVING COLUMNS NOT USEFUL FROM THE DATA

I called out all the columns in the data set and deleted those columns I thought were not significant to obtaining the goals of the project. Deleted columns are typically qualitative.

Deleted columns are Country and Year.

I then created a new data frame for useful columns. These columns are the newly created predictive columns for target variable - Life Expectancy category.

Predictors: Region, Infant deaths, Under-five deaths, Adult mortality, Alcohol consumption, He patitis B, Measles, BMI, Polio, Diphtheria, Incidents HIV, GDP_per_capita, Population, Thinnes s_ten_nineteen_years, Thinness_five_nine_years, Schooling, Economy_status_Developed, Econ omy_status_Developing,

Target Variable: Life Expectancy category

Low Life Expectancy Category = 0

High Life Expectancy Category = 1

Categorical values in target column (Life Expectancy category) are replaced with numerical values. Low is replaced with 0, and high is replaced with 1.

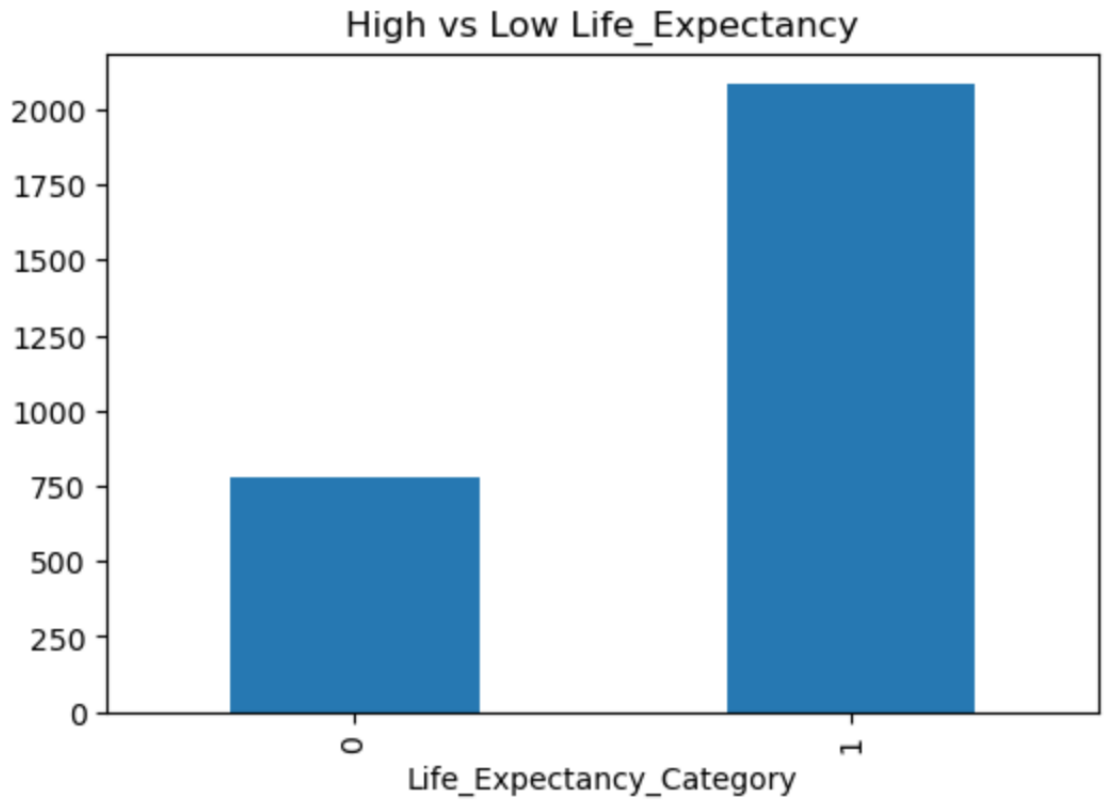
DISTRIBUTION OF TARGET VARIABLE

I had a look at the target variable’s distribution of each class to make sure there is class balance.

This is important because if target variable's distribution is too skewed, the class imbalance will impact the machine learning algorithm ability to learn all the classes, thus, the predictive modeling will not be possible.

Bell curve is desirable but slightly positive skew or negative skew is also fine for performing Classification.

Fig 6: Uni-variate plots



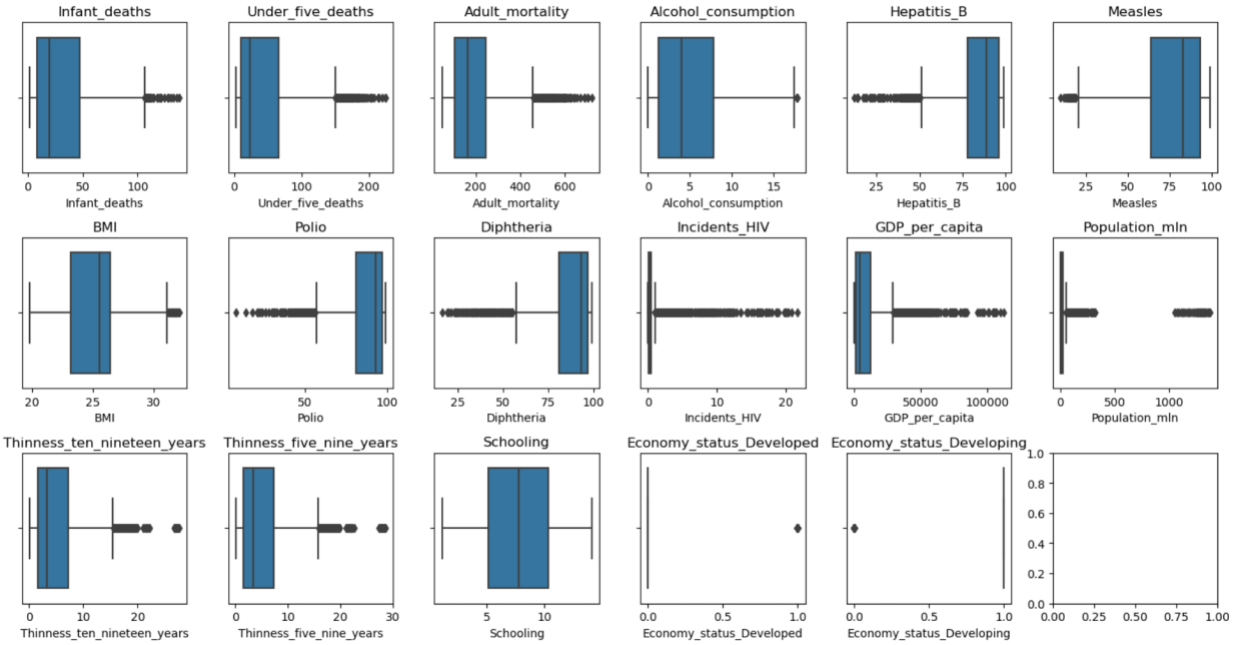
The above graph shows the distribution of the target variables. One class is highly bias to another class.

OUTLIERS

As part of the exploratory data analysis, I plotted a boxplot for all independent variables in the dataset to view data points that significantly differ from the rest of the dataset. These odd datapoints are called outliers and can lead to trouble for scaling techniques. (Andreas C. Müller & Sarah Guido, 2017.pg133)

Below boxplot shows outliers in all the features which are handled using StandardScaler technique during machine learning phase.

Fig 7: Boxplot of independent variables

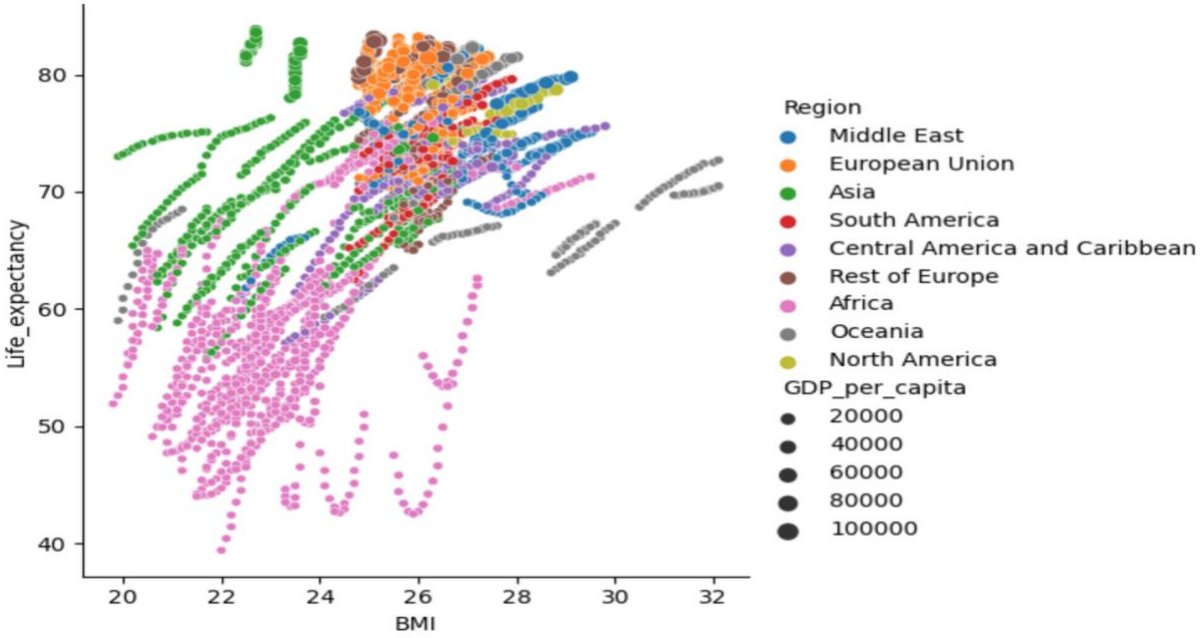


VISUAL EXPLORATORY DATA ANALYSIS

Making informative visualizations (sometimes called plots) is one of the most important tasks in data analysis. It may be a part of the exploratory process-for example, to help identify outliers or needed for data transformations, or as a way of generating ideas for models. (Wes McKinney (Python for Data Analysis, Second edition, 2017, pg.19)

To get more insight into the dataset, I visualized how some of the independent variables in the original dataset correlate with life expectancy. For example, the below plot depicts how life expectancy correlates with BMI, regions, and GDP per capital.

Fig 8. Life expectancy correlation with BMI, Region and GDP per capital.



Furthermore, From the selected useful columns, I visualize distribution of all the categorical predictor variables in the new dataset using bar plots. Categorical variable in the data can be spotted by looking at the unique values in them. Typically, a categorical variable contains less than 20 Unique values and there is repetition of values, which means the data can be grouped by those unique values.

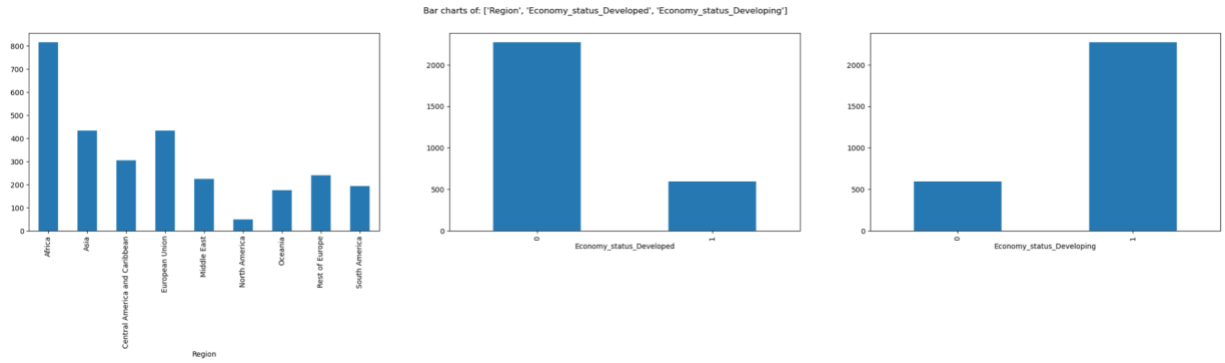
Based on the Basic Data Exploration above, I have spotted three categorical predictors in the data.

Categorical Predictors: Region, Economy_status_Developed and Economy_status_Developing.

Continuous variables columns: Infant deaths, Under-five deaths, Adult mortality, Alcohol consumption, Hepatitis B, Measles, BMI, Polio, Diphtheria, Incidents HIV, GDP_per_capita, Population, Thinness_ten_nineteen_years, Thinness_five_nine_years, Schooling.

I use bar charts to see how the data is distributed for the above listed categorical columns.

Fig 9: Bar chat distribution of categorical columns



BAR CHARTS INTERPRETATION

These bar charts represent the frequencies of each category in the Y-axis and the category names in the X-axis.

In the ideal bar chart, each category has comparable frequency. Hence, there are enough rows for each category in the data for the ML algorithm to learn.

If there is a column which shows too skewed distribution where there is only one dominant bar, and the other categories are present in very low numbers. These kinds of columns may not be very helpful in machine learning, and I can confirm this in the correlation analysis section and take a final call to select or reject the column.

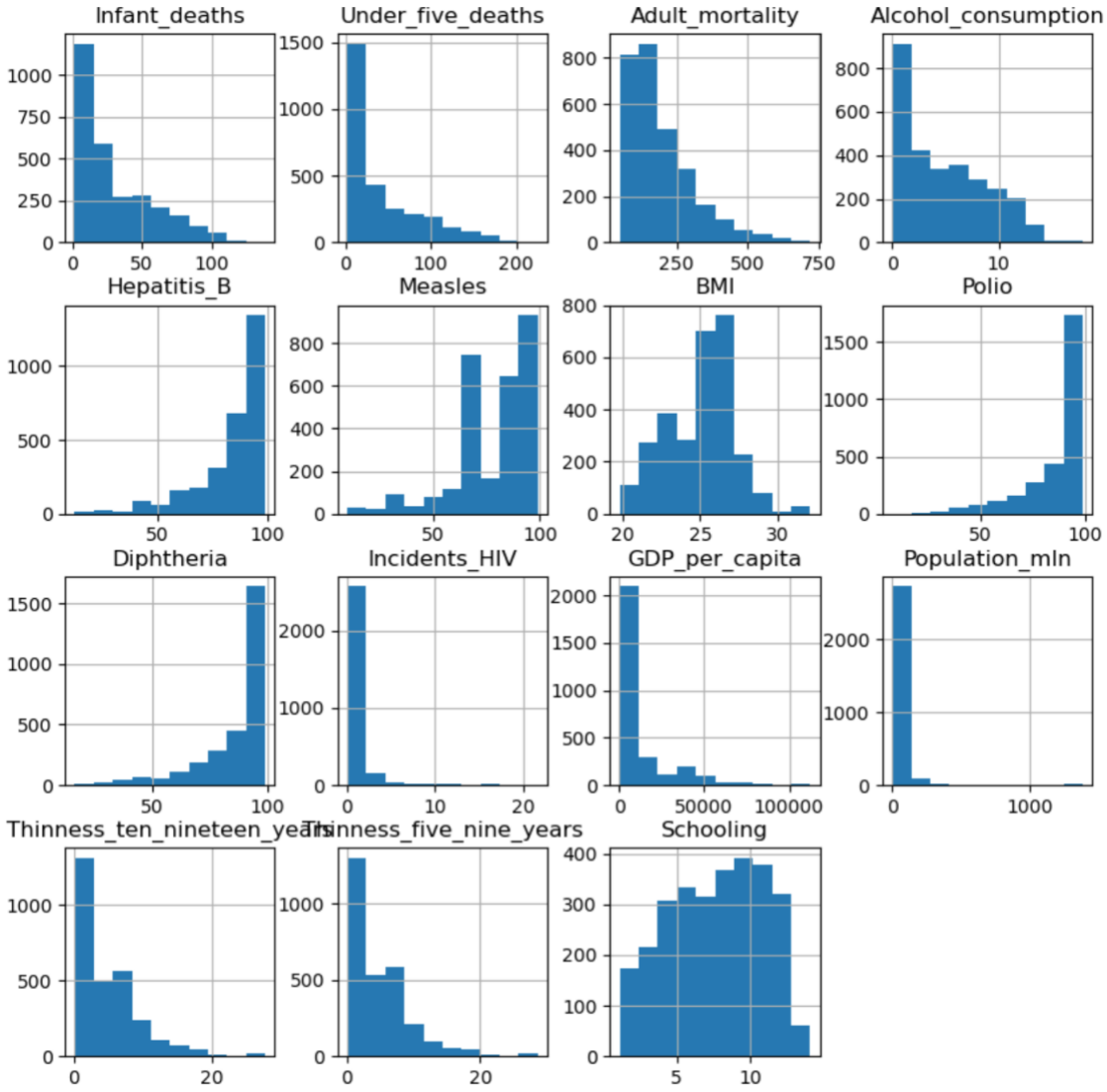
In this dataset, the categorical column ‘Region’ has satisfactory distribution to be considered for machine learning. Columns ‘Economy_status_Developed and Economy_status_Developing’ have distributions where there is only one dominant bar, and the other categories are present in very low numbers.

Selected Categorical Variables: All the categorical variables are selected for further analysis.

Region, Economy_status_Developed, Economy_status_Developing.

HISTOGRAM TO SEE HOW THE DATA IS DISTRIBUTED FOR CONTINUOUS VARIABLES COLUMNS.

Fig 10: Histogram distribution for Continuous variables



HISTOGRAM INTERPRETATION

Histograms shows the data distribution for a single continuous variable.

The X-axis shows the range of values and the Y-axis represent the number of values in that range.

The ideal outcome for histogram is a bell curve or slightly skewed bell curve. If there is too much skewness, then outlier treatment should be done, and the column should be re-examined. If the problem with skewness and outliers can't be solve. The column is rejected.

The distribution is good for selected continuous variables: Infant deaths, Under-five deaths, Adult mortality, Alcohol consumption, Hepatitis B, Measles, BMI, Polio, Diphtheria, Incidents HIV, GDP_per_capita, Population, Thinness_ten_nineteen_years, Thinness_five_nine_years, Schooling.

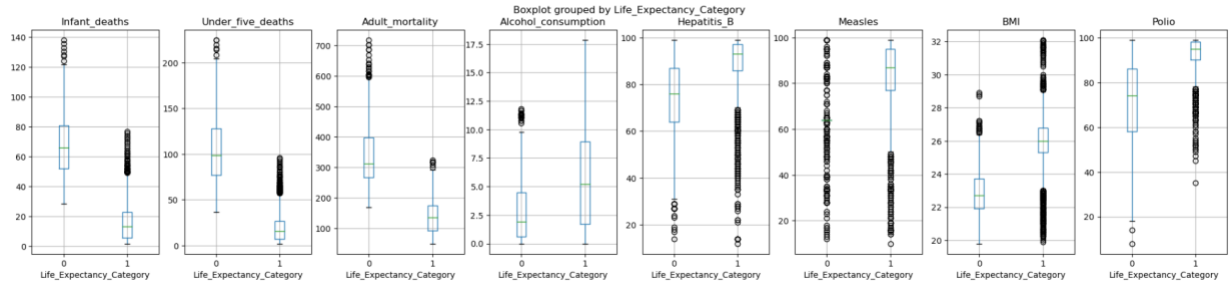
CORRELATION MATRICES

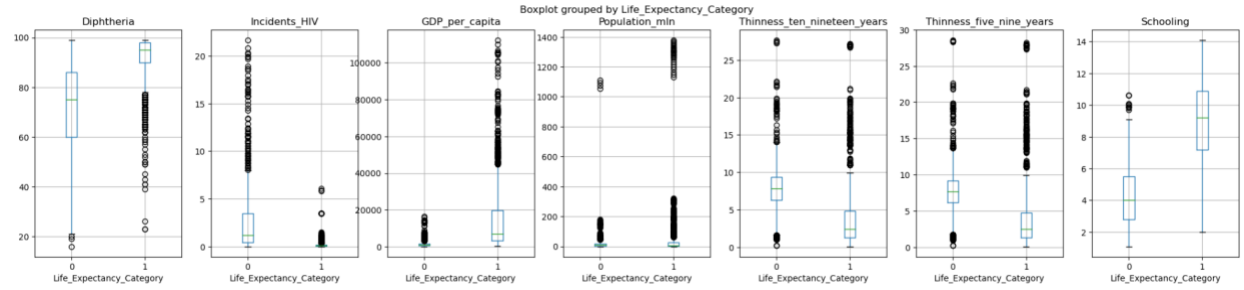
I visualize the relationship between the Target variable (Life expectancy category) and each of the predictors to get a better understating of data.

RELATIONSHIP EXPLORATION: CATEGORICAL VS CONTINUOUS

The target variable (Life_Expectancy_Category) is Categorical and the predictor variable is Continuous. I analyze the relation using bar plots/Boxplots and measure the strength of relation using ANOVA test.

Fig 11: Box plots for each continuous predictor against the Target Variable





BOX-PLOTS INTERPRETATION

These plots give an idea about the data distribution of continuous predictor in the Y-axis for each of the category in the X-Axis.

If the distribution looks similar for each category (Boxes are in the same line), that means the continuous variable has NO effect on the target variable. Hence, the variables are not correlated to each other.

In the charts the boxes are in different lines. This mean the data distribution is different (the boxes are not in same line) for each category of attrition. It hints that these variables are correlated with attrition.

ANOVA TEST

Analysis of variance (ANOVA) is performed to check if there is any relationship between the given continuous and categorical variable.

Assumption(H0): There is NO relation between the given variables (i.e. The average(mean) values of the numeric Target variable is the same for all the groups in the categorical Predictor variable)

In Python Jupyter notebook ANOVA test result is printed for Probability of H0 being true, If the ANOVA P-Value is <0.05, that means we reject H0.

ANOVA RESULTS

Infant_deaths is correlated with Life_Expectancy_Category | P-Value: 0.0

Under_five_deaths is correlated with Life_Expectancy_Category | P-Value: 0.0
Adult_mortality is correlated with Life_Expectancy_Category | P-Value: 0.0
Alcohol_consumption is correlated with Life_Expectancy_Category | P-Value: 2.415836524221 6077e-60.
Hepatitis_B is correlated with Life_Expectancy_Category | P-Value: 2.7362962953964935e-123.
Measles is correlated with Life_Expectancy_Category | P-Value: 2.1010780706175882e-149
BMI is correlated with Life_Expectancy_Category | P-Value: 8.12210810213787e-305.
Polio is correlated with Life_Expectancy_Category | P-Value: 2.638864619157098e-301.
Diphtheria is correlated with Life_Expectancy_Category | P-Value: 5.787394456243716e-280.
Incidents_HIV is correlated with Life_Expectancy_Category | P-Value: 6.502549144832511e-16 0.
GDP_per_capita is correlated with Life_Expectancy_Category | P-Value: 9.51279416370404e-8 9.
Population_mln is correlated with Life_Expectancy_Category | P-Value: 0.00177300743771278 17
Thinness_ten_nineteen_years is correlated with Life_Expectancy_Category | P-Value: 3.861912 8264664017e-122.
Thinness_five_nine_years is correlated with Life_Expectancy_Category | P-Value: 3.630697025 176205e-108.
Schooling is correlated with Life_Expectancy_Category | P-Value: 0.0

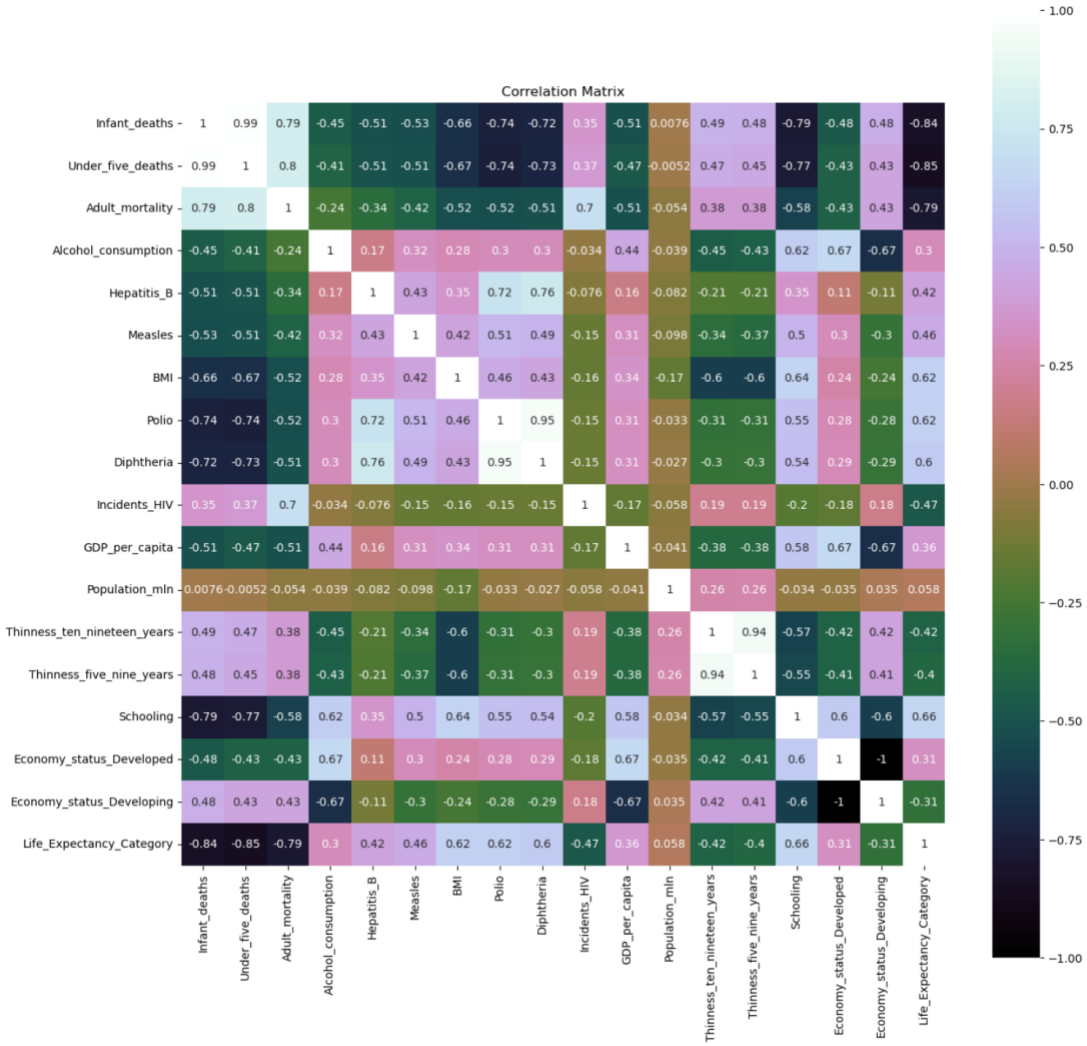
The results of ANOVA confirm our visual analysis using box plots above.

CORRELATION MATRICES PLOT

Correlation matrix was used to identify pattern, relationship, and dependencies between the variables.

As infant death, adult mortality, and incident of HIV decreases, and life expectancy increases, this indicates a perfect negative correlation.

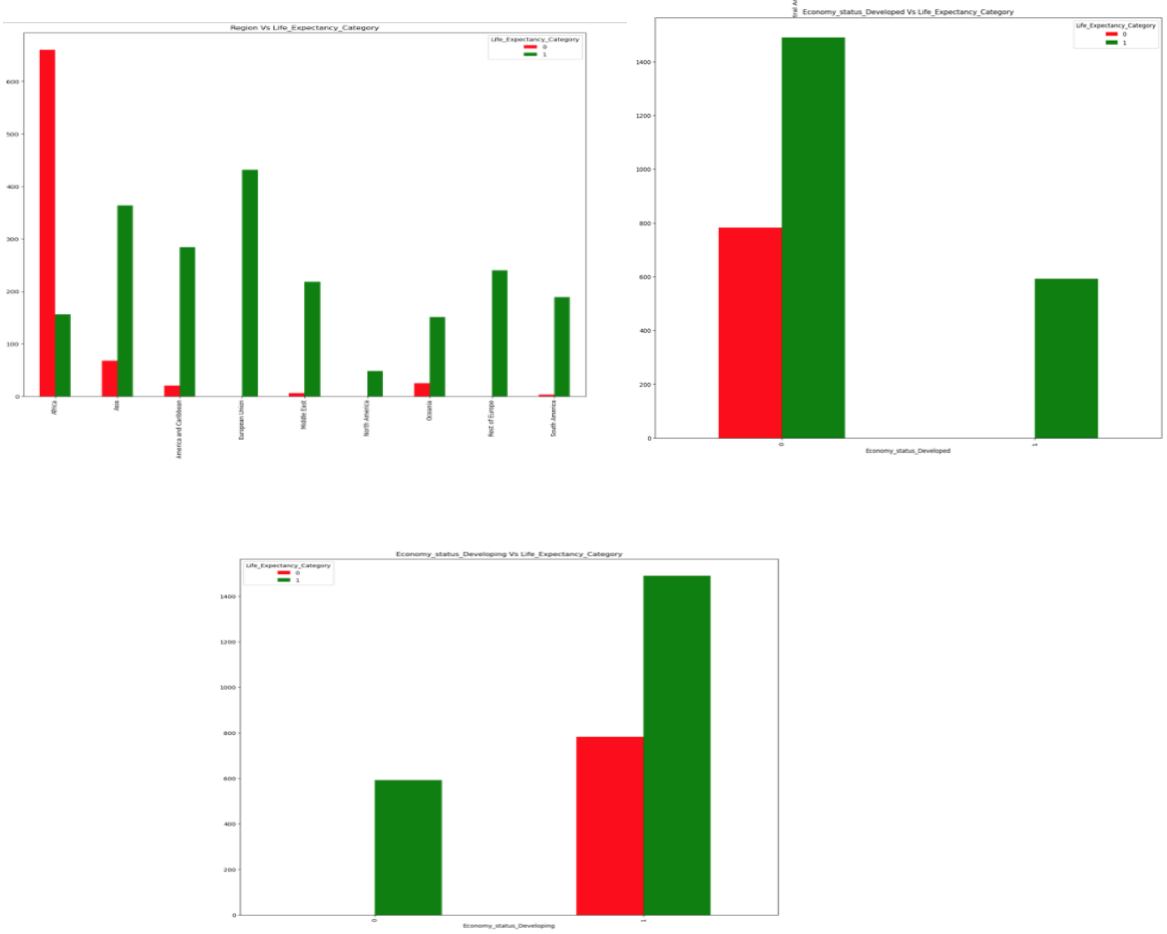
Fig 12: Correlation Matrix



RELATIONSHIP EXPLORATION: CATEGORICAL VS CATEGORICAL

I explore the correlation visually of the categorical target variable and the categorical predictors using grouped bar plots and confirm the results statistically using Chi-square test.

Fig 13: Bar plot for each categorical predictor against the Target Variable



GROUPED BAR CHARTS INTERPRETATION

These grouped bar charts show the frequency in the Y-Axis and the category in the X-Axis. If the ratio of bars is similar across all categories, then the two columns are not correlated.

We confirm this analysis in below section by using Chi-Square Tests.

CHI-SQUARE TEST

Chi-Square test is conducted to check the correlation between two categorical variables.

Assumption(H0): The two columns are NOT related to each other Result of Chi-Sq Test: The Probability of H0 being True.

In Python Jupyter notebook Chi-Square Test result is printed for Probability of H0 being true, If the Chi-Square P-Value is <0.05, that means we reject H0.

CHI-SQUARE RESULTS

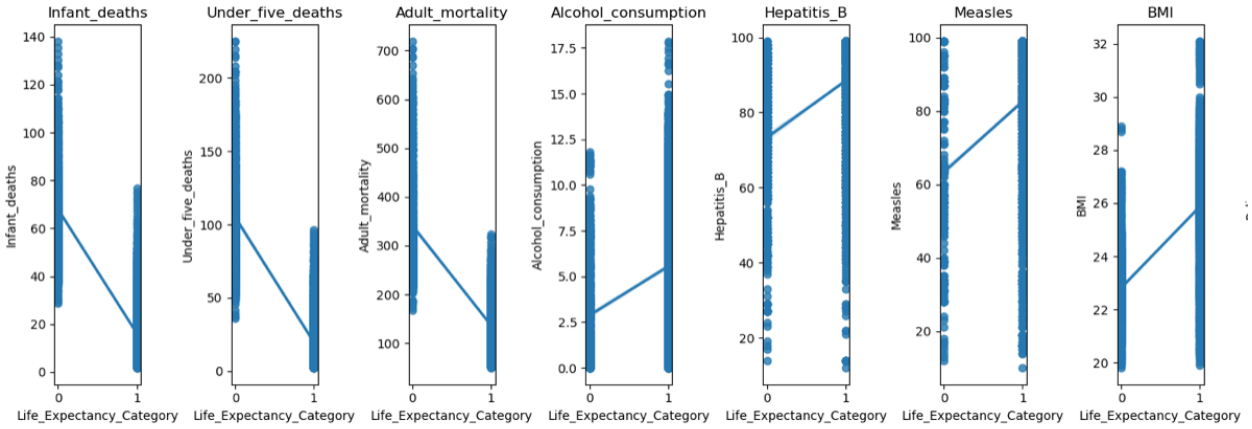
Region is correlated with Life_Expectancy_Category | P-Value: 0.0
Economy_status_Developed is correlated with Life_Expectancy_Category | P-Value: 1.5453129678739983e-62.
Economy_status_Developing is correlated with Life_Expectancy_Category | P-Value: 1.5453129678739983e-62.

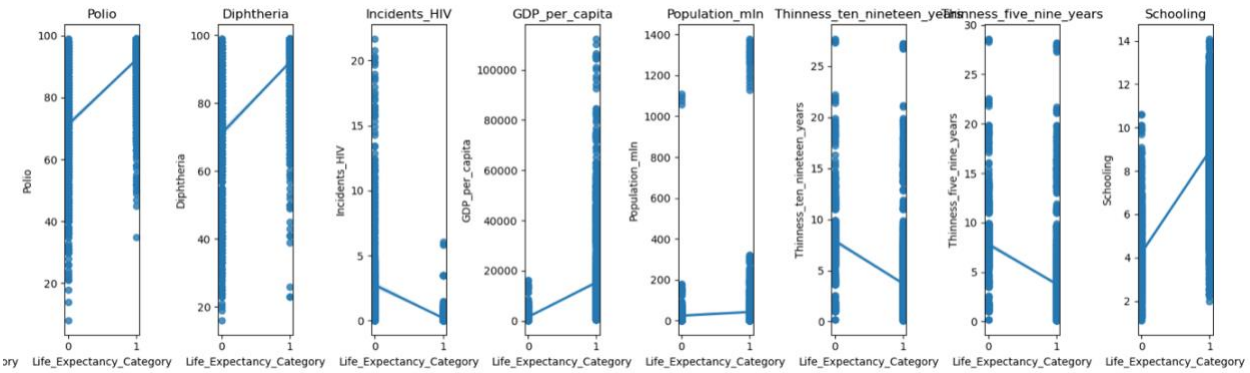
Final columns for machine learning will be selected based on the results for ANOVA and Chi-Square tests.

REGRESSION PLOT

To corroborate results from ANOVA and Chi-Square test statistical analysis, I used plots correlations between each continues predictors against the target variable.

Fig 14: Regression plot for each continuous predictor against the Target Variable





As expected, Life expectancy category increases GDP per capital and schooling while it decreases with infant deaths, adult mortality, and incidents of HIV.

However, there is an interesting observation from the regression plot. The life expectancy category increases with alcohol consumption, BMI and diseases like Hepatitis B, Measles, Polio, and Diphtheria. This could be due to advancement in medicines and the availability of pharmaceutical drugs and vaccines.

FURTHER DATA EXPLORATORY ANALYSIS

Based on the above tests, the following features are selected for the final columns for machine learning.

Region, Infant deaths, Under-five deaths, Adult mortality, Alcohol consumption, Hepatitis B, Measles, BMI, Polio, Diphtheria, Incidents HIV, GDP_per_capita, Population, Thinness_ten_nineteen_years, Thinness_five_nine_years, Schooling.

I am dropping the categorical variables 'Economy_status_Developed and 'Economy_status_Developing as the datapoints are 0 and 1 and could impact the machine learning algorithm.

CONVERTING THE NOMINAL VARIABLE TO NUMERIC USING GET_DUMMIES()

Categorical columns with text values in the data set can be converted to numerical values using get_dummies. From the below figure we can see that the categorical column Region has been split by putting different regions in columns.

Fig 15: Region split into columns.

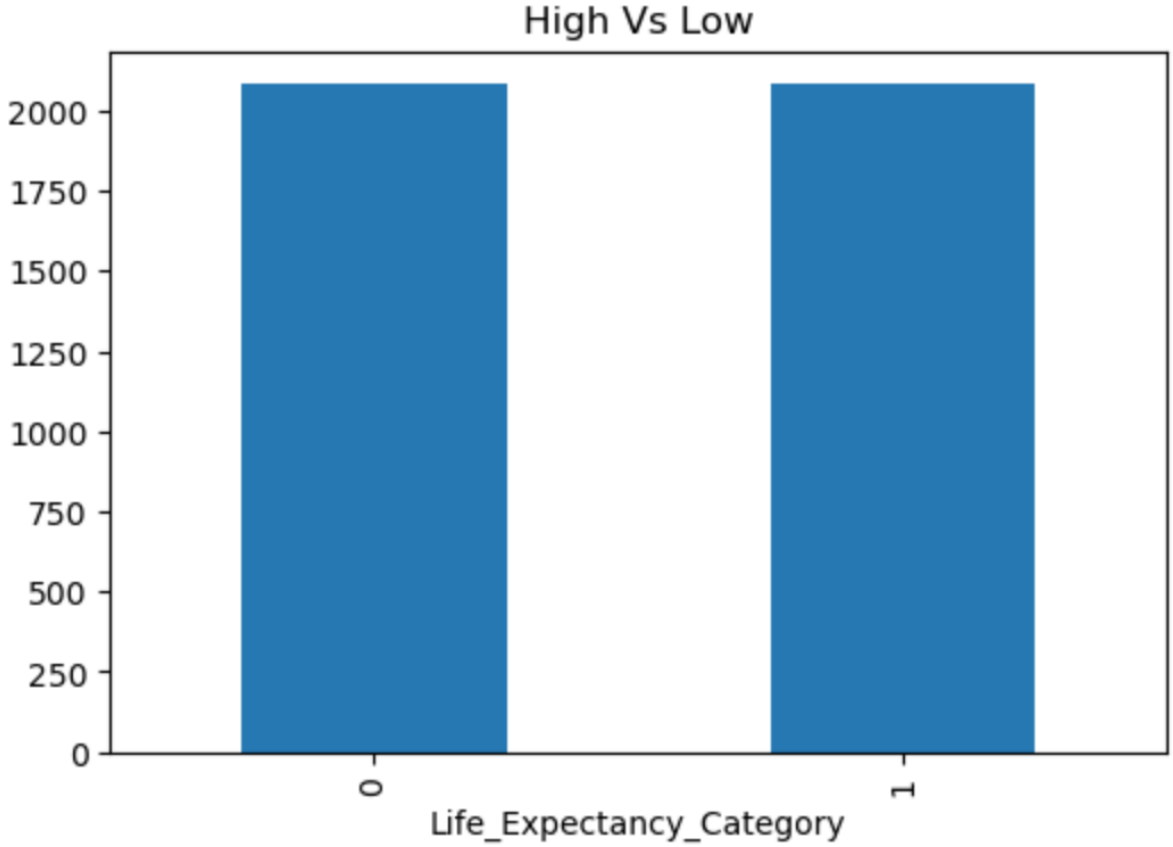
Region_Central America and Caribbean	Region_European Union	Region_Middle East	Region_North America	Region_Oceania	Region_Rest of Europe	Region_South America
0	0	1	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	1
0	0	1	0	0	0	0

SMOTE: (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE)

Because there is class imbalance in the distribution of target variable 'Life Expectancy Category’'

I am using SMOTE method to balance the class distribution.

Fig 16: Balanced class distribution of target variable



SMOTE was applied to solve the class balance by creating synthetic examples of the minority class by randomly selecting one of its neighbors and generating a new example along the line joining the two points.

After applying SMOTE, I now combine the oversampled minority class with the majority class to create a new balanced dataset.

I then shuffled the balanced dataset to remove any biases introduced by the oversampling technique.

STANDARDIZATION OF SELECTED FEATURES FROM DATASET

Data standardization is an important technique in data preprocessing. The goal of data standardization is to transform the numeric variables so that each variable has zero mean and unit variance.

The data for selected predictive features was split into input and output across splits 30/70 (30 % used for testing and 70% used for training).

The independent features are then scored based on their importance to the machine learning.

Fig17: Features Importance

Variables	Score
Infant_deaths	10353.651754
Under_five_deaths	10276.406080
Region_Africa	6730.745684
Adult_mortality	6569.778811
Schooling	4573.589874
BMI	3683.843670
Polio	2435.095172
Diphtheria	2198.200743
Measles	1594.802402
Thinness_ten_nineteen_years	1150.630937
GDP_per_capita	1127.798680
Thinness_five_nine_years	1025.257723
Hepatitis_B	1024.442703
Incidents_HIV	855.473216
Alcohol_consumption	634.965459
Region_European Union	544.843636
Region_Rest of Europe	271.140065
Region_Middle East	192.235754
Region_Central America and Caribbean	182.348151

MODELLING PHASE

Four models: Random Forest and Decision Tree, Support Vector Model, and Logistic Regression were selected for the binary classification.

Binary classification is a special case of distinguishing between exactly two classes. In Binary classification we often speak of one class being the positive class and the other class being the negative class. (Andreas C. Müller & Sarah Guido, 2017)

The models were evaluated using the score methods.

Accuracy is the number of correct predictions (TP and TN) divided by the number of all samples (all entire of the confusion matrix summed up)

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall measures how many of the positive samples are captured by the positive predictions. Recall is used as performance metric when we need to identify all positive samples, that is when it's important to avoid false negatives.

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-Score or f-measure is a harmonic mean of precision and recall. This particular variant is also known as the f1-score as it takes precision and recall into account. It can be better measure than accuracy on imbalanced binary classification dataset. (Andreas C. Müller & Sarah Guido, 2017).

$$F1 = 2.\frac{\text{Precision. Recall}}{\text{Precision + recall}}$$

Fig 18: Score evaluation of models

	model	accuracy	precision	recall	f1-score
0	Logistic_Regression	0.99	1.00	0.98	0.99
1	SVC	0.98	0.99	0.98	0.99
2	Random_Forest	0.99	0.99	0.99	0.99
3	Decision_Tree	0.99	0.99	0.99	0.99

From the above table, all the classifier models performed excellently. Random Forest and Decision Tress selected for machine learning models as they are widely used for classification models.

RANDOM FOREST CLASSIFIER.

A Random Forest model was build using a Gaussian Classifier n_estimator

For our dataset, training accuracy 100% and testing accuracy 99% were achieved.

Training Accuracy: 1.0
Testing Accuracy: 0.9928
[[598 3]
 [6 643]]

Fig 19: Confusion Matrix model.

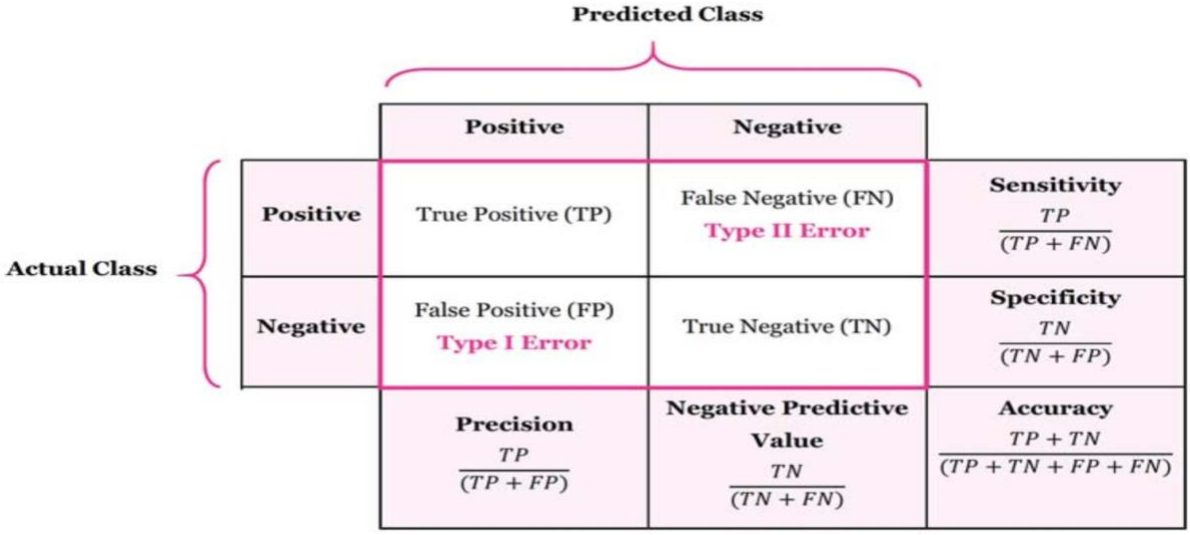
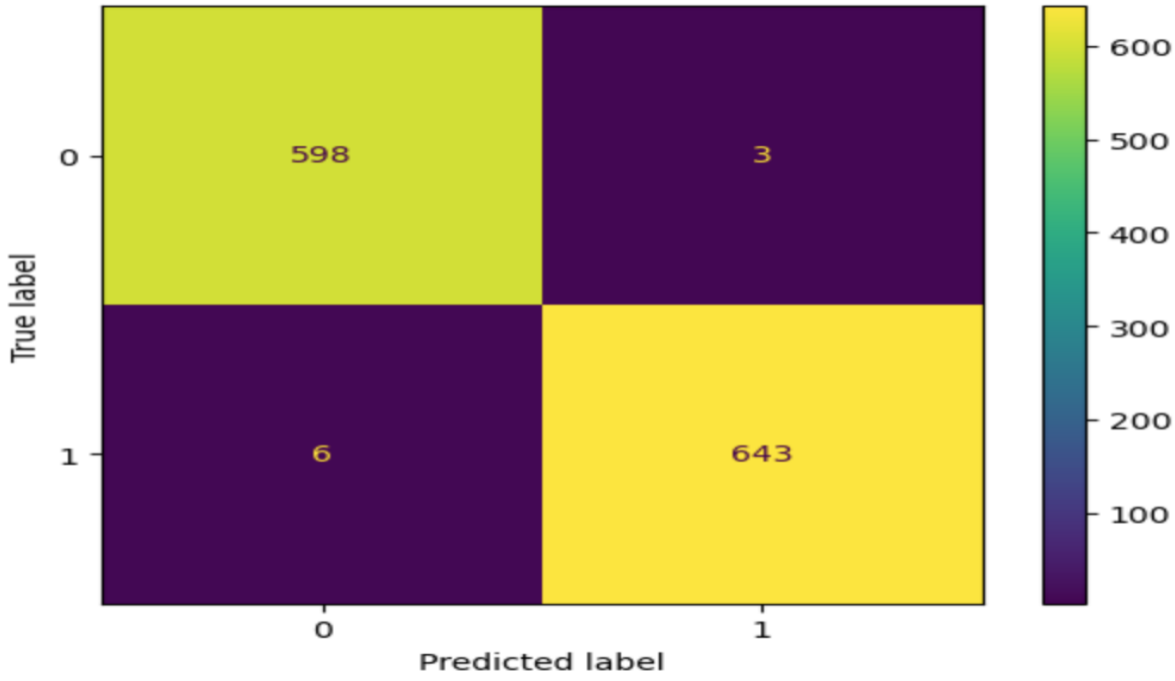


Fig 20: Confusion Matrix for Random Forest



From the confusion matrix table, 598 (True Positive) correctly predicted as low life expectancy. 643 (True Negative) correctly predicted as high life expectancy from the dataset.

3 (False Positive) incorrectly predicted as low life expectancy while 6 (False Negative) were incorrectly classified as high life expectancy.

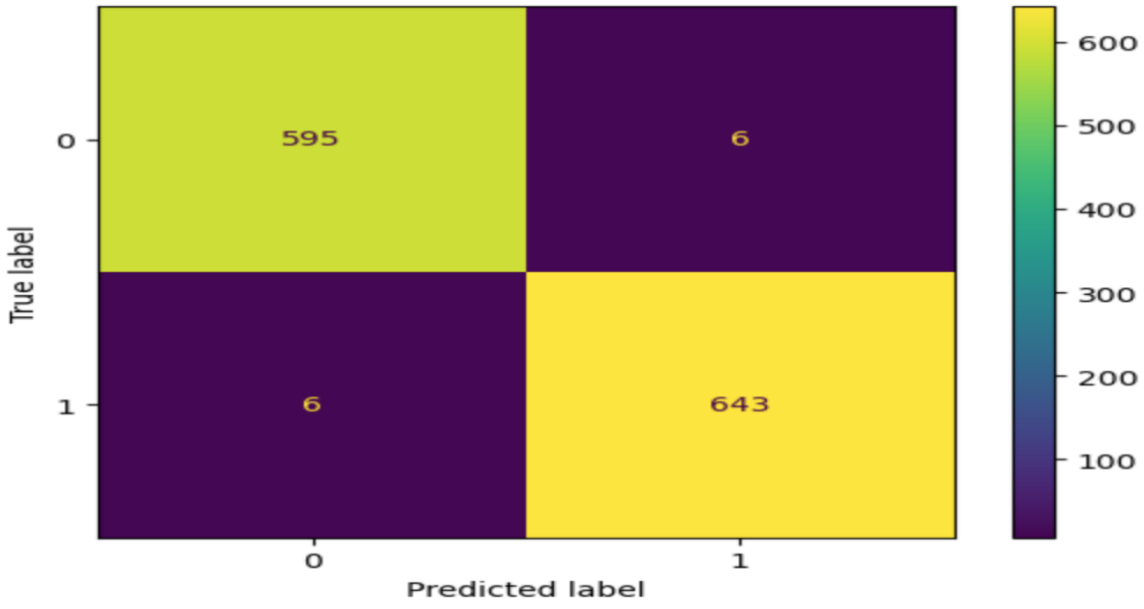
DECISION TREE

Running 10-Fold Cross validation on Decision Tree algorithm by passing full data X and y because the K-fold will split the data and automatically choose train/test.

F1 values for 10-fold Cross Validation: 0.9891768666650156

For our dataset, training accuracy 100% and testing accuracy 99% were achieved.
Training Accuracy: 1.0
Testing Accuracy: 0.9904
[[595 6]
[6 643]]

Fig 21: Confusion Matrix for Decision Tree



From the confusion matrix table, 595 (True Positive) correctly predicted as low life expectancy. 643 (True Negative) correctly predicted as high life expectancy from the dataset.

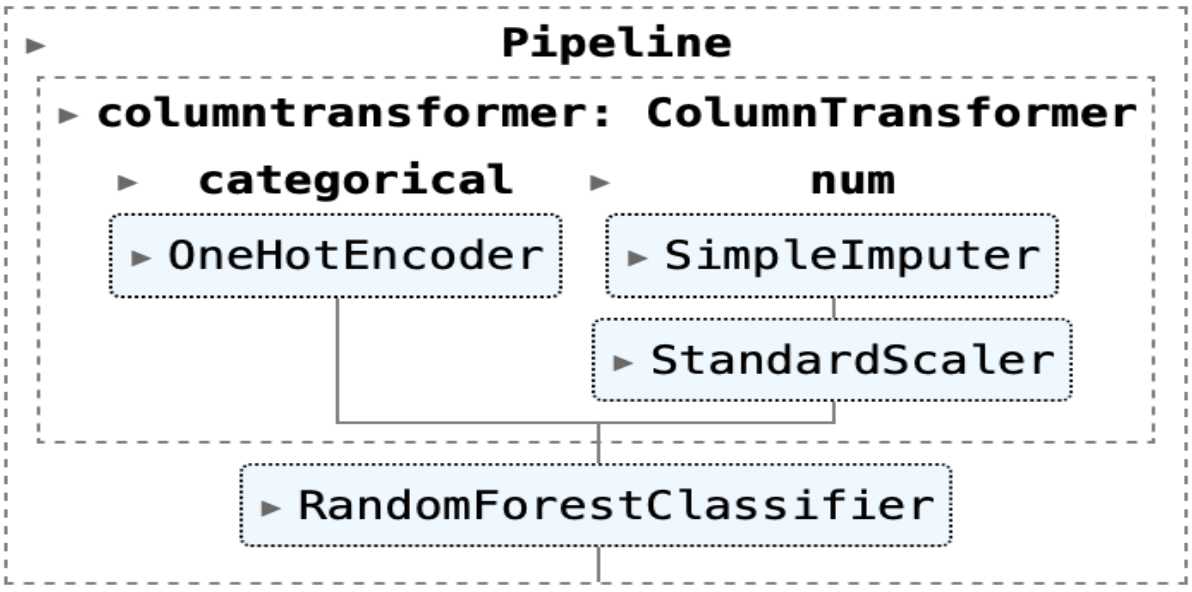
6 (False Positive) incorrectly predicted as low life expectancy while 6 (False Negative) were incorrectly classified as high life expectancy.

PIPELINE AND SHAP

The Pipeline class is a class that allows ‘gluing’ together multiple processing steps into a single scikit-learn estimator. The Pipeline class itself has fit, predict, and score methods and behave just like any other model in scikit-learn. The most common use case of the Pipeline class is in chaining preprocessing steps (Like scaling of data) together with a supervised model like a classifier (Andreas C. Müller & Sarah Guido, 2017.pg308)

I applied Pipeline on the dataset created in df2 for processing using standard scaler to standardize the dataset and one hot encoder to transform categorical columns. The processed data was then trained and tested on Random Forest machine learning model.

Fig 22: Pipeline for processing



The diagram can be clicked on in Python to see details of each step.
MODEL SCORE 100.0%

RANDOMIZED SEARCH CV

I defined a parameter in Pipeline to search over and construct a search CV from the pipeline and parameter grid. For each split in the cross validation, the StandardScaler is refit with only the training splits and no information is leaked from the test split into the parameter search.

Cross Validation Score 100.0%

SHAP (SHAPLEY ADDITIVE EXPLANATIONS)

SHAP is a method that provides insights into the decision-making process of machine learning models and help us understand the factors that contribute to their predictions (https://medium.com/Python-in-plain-english/machine-learning-interpretability-in-finance-investigating-shap-and-lime-baaa9f96684c)

I used SHAP to explain the Random Forest model built in pipeline using dataset 'Life-Expectancy-WHO-Updated'. I used the predictors features to predict whether a population life expectancy is low or high.

Based on the best estimator in the pipeline, I created the explainer in explaining the prediction of the Random Forest Classifier.

Fig 23: Features used by the machine learning algorithm in making predictions.

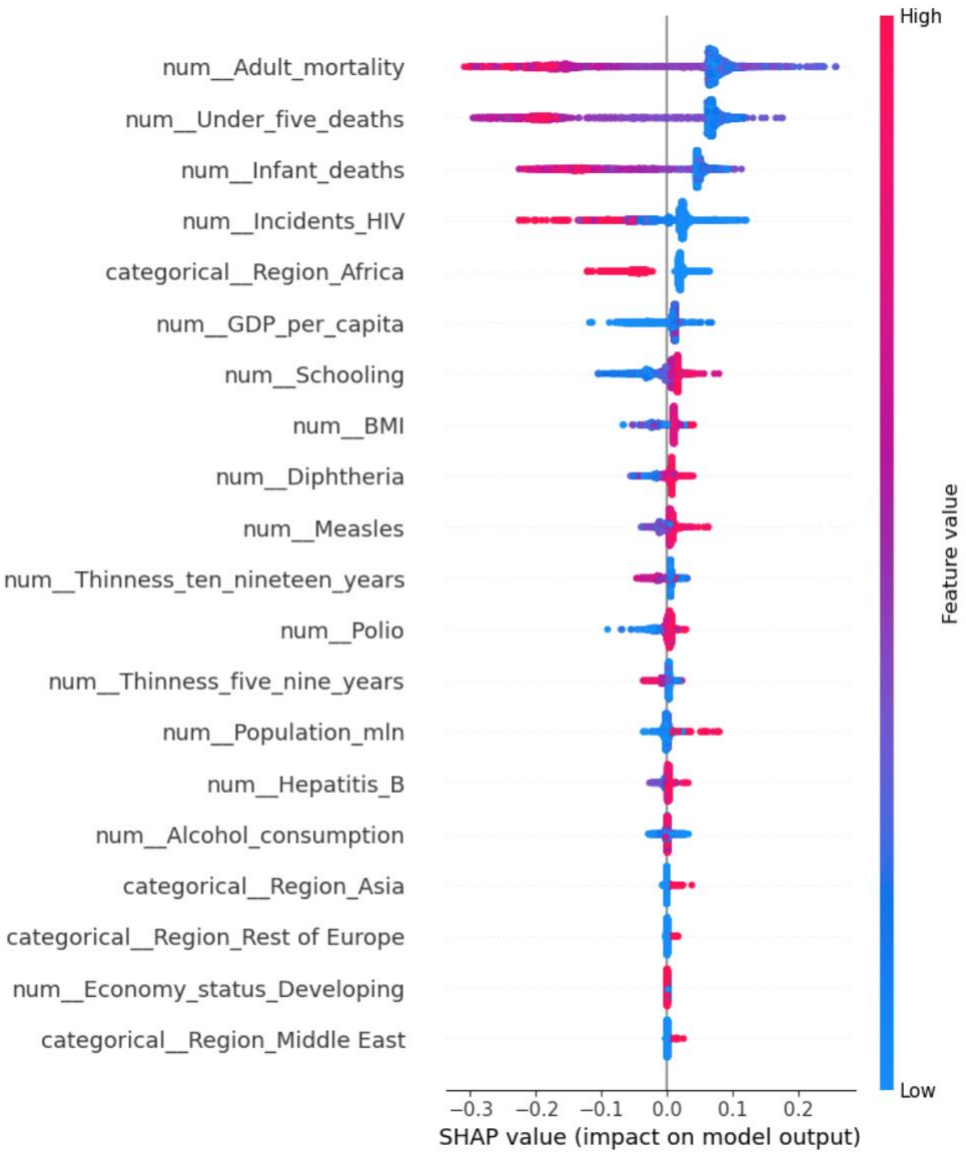
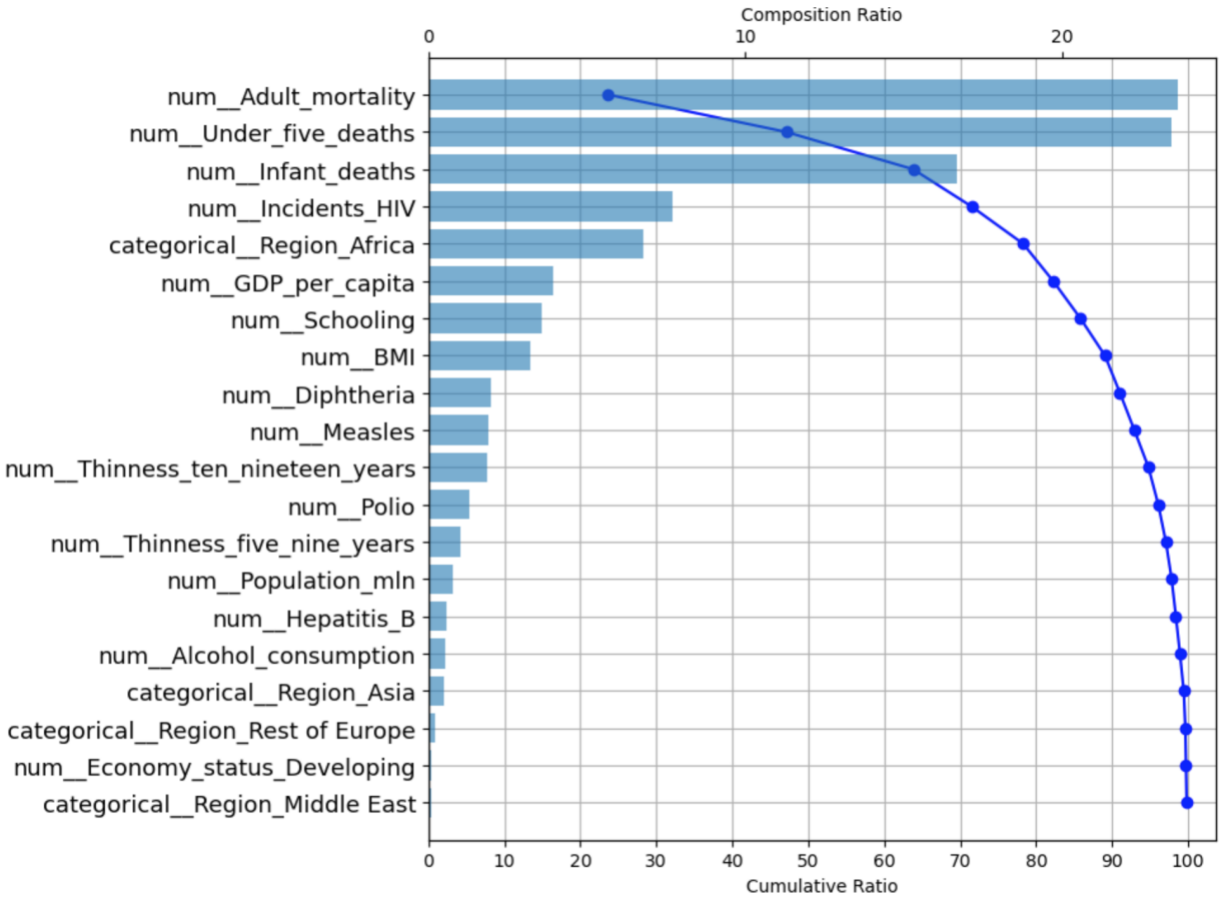


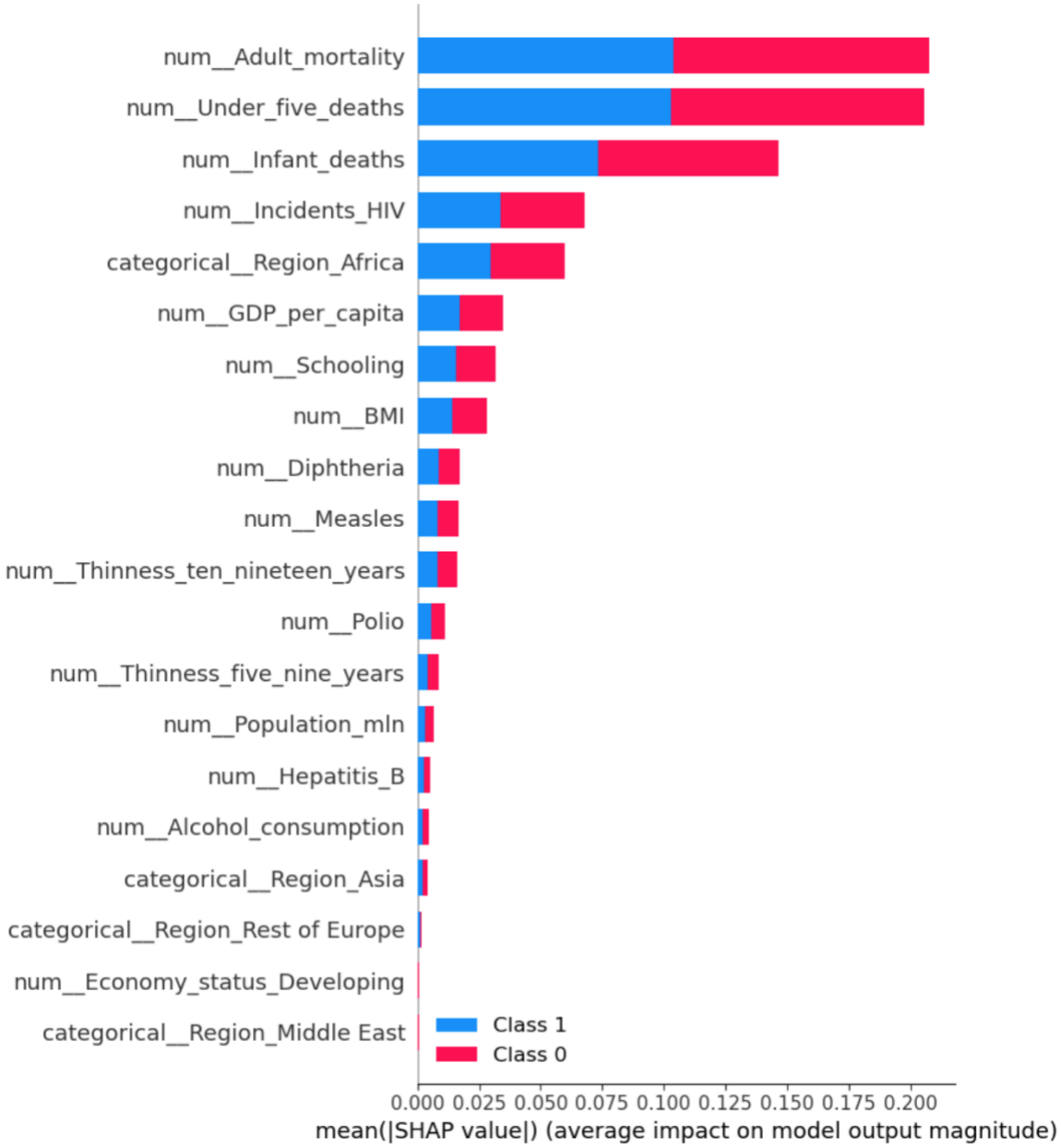
Fig 23 shows all the features used by the machine learning algorithm in making prediction and a SHAP value is assigned to each feature based on how the feature impacts model output.

Fig 24: Waterfall plot



SHAP waterfall plot was created to display the contributions of individual features to the difference between the Random Forest classifier model's output for a specific instance and the expected model output which is usually the average output for the entire dataset.

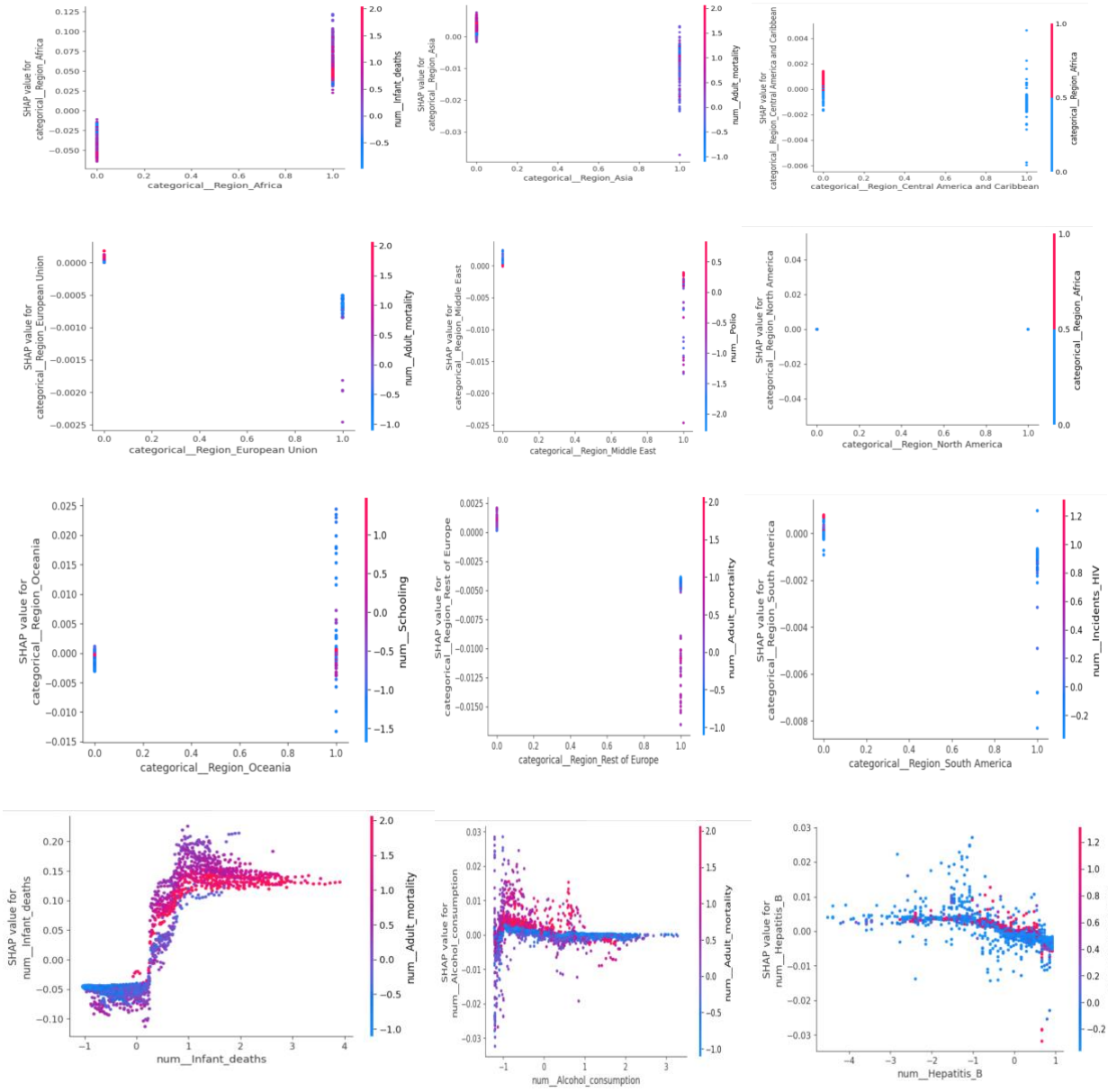
Fig 25: SHAP Summary bar plots

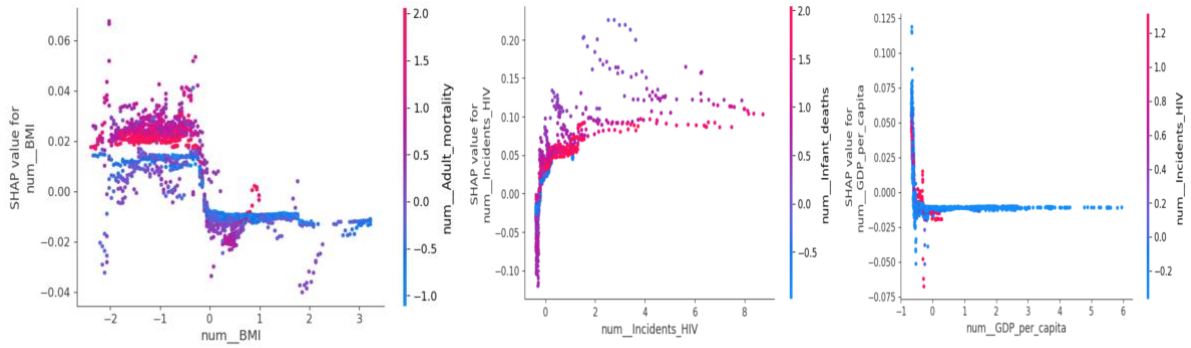


I created a summary bar plot from SHAP library to provide a visual summary of the features importance of the Random Forest model used in prediction across all instances in the dataset.

Furthermore, dependence plots were created to visualize the relationship between features and the Random Forest model output. The plots show how the output of the model change as the value of a particular feature vary.

Fig 26: SHAP Dependency plots





SAVE AND LOAD MODEL

The Random Forest machine learning model was saved using Python module ‘Pickle’. The machine leaning model is saved to file ‘LifeExpectancyModel’ and can be loaded later for predictions.

CHALLENGES ENCOUNTERED AND STRATEGIES USED TO OVERCOME THEM

1. Inconsistencies were identified in the min, max values, and outliers in the dataset. The max and min values of one feature are significantly larger than the other features. I standardized all the features to the same scale using StandardScaler.
2. Class imbalance was observed in the distribution of target variable 'Life Expectancy Category’. SMOTE method was used to balance the class distribution.
3. Random Forest classifier model can be considered as ‘black box’ model, making it challenging to interpret the decision-making process. SHAP was used in understanding the rationale behind individual prediction.

CONCLUSIONS

In this report, the focus of the analysis on the dataset Life expectancy-WHO-updated which contains features that are necessary to establish whether a country's population life expectancy is low or high. I selected socio economic, health, environmental and regional features in relation to the project goals within the dataset to determine a country's populations life expectancy. These features were used to predict the proportions of the population expected to have low or high life expectancy from the dataset. Using the Decision Tree Classifier, I was able to predict with accuracy of 100% the numbers of the population with low vs high life expectancy.

As part of the project, I ran a few statistical models and data visualizations methods. I was able to get a clear correlation between life expectancy and independent features such as regions, GDP per capital, schooling, incidents of HIV, populations, and several other diseases.

The project was proceeded to re-evaluation stage as per CRISP DM model. Some of the earlier steps and methods were revisited and a new approach was used to building a Random Forest model. Random Forest Classifier was able to predict with accuracy of 100% whether a region's populations life expectancy is low or high. I used the Shapley Additive Explanations (SHAP) to understand decision-making process of the Random Forest model in making predictions and how each independent factor contributes to the predictions.

REFERENCES

J*essica Y Ho and Arun S Hendi, (Recent trends in life expectancy across high income countries: Retrospective observation study, BMJ 2018:* doi: https://doi.org/10.1136/bmj.k2562)

UN DESA (https://www.un.org/development/desa/dpad/least-developed-country-category/ldc-criteria.html?target=human-assets)

Joel Grus (Data Science from scratch. First Principle with Python, 2015 pg. 10)

Andreas C. Müller & Sarah Guido (Introduction to Machine Learning with Python, 2017. pg.25-26, Pg.70, pg.83, pg.133, pg. 282-284, pg.308)

Wes McKinney. Python for Data Analysis. Data Wrangling with Pandas, NumPy, and IPython. Second edition, 2017, pg.19, pg.263)

https://medium.com/Python-in-plain-english/machine-learning-interpretability-in-finance-investigating-shap-and-lime-baaa9f96684c