**Life Expectancy: A Data-Driven Exploration and Prediction**

**Milestone: Implementation of Selected Models**

Group 13

Sai Varun Kumar Namburi

Prajwal Srinivas

[namburi.sai@northeastern.edu](mailto:namburi.sai@northeastern.edu)

[srinivas.pra@northeastern.edu](mailto:srinivas.pra@northeastern.edu)

**Percentage of Effort Contributed by Student1: 50%**

**Percentage of Effort Contributed by Student2: 50%**

**Signature of Student 1: Sai Varun**

**Signature of Student 2: Prajwal**

**Date of Submission: 2nd Apr 2023**

**Data Source:** This dataset is taken from kaggle.com

<https://www.kaggle.com/code/philbowman212/life-expectancy-exploratory-data-analysis/data>

**Data Description:**

The World Health Organization's Global Health Observatory (GHO) maintains records of all countries' health statuses and related factors. The data concerning life expectancy and health factors for 193 countries were obtained from the WHO's Global Health Observatory data repository. It was noted that over the past 15 years, there has been significant progress in the health sector, leading to a significant improvement in human mortality rates, particularly in developing nations compared to the last 30 years. In this project, data from the years 2000 to 2015 for 193 countries were selected for further analysis. In this dataset, we have 22 columns as below

(Country, Year, Status, Life expectancy, Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources, Schooling)

**Implementation of Selected Model:**

We have decided to use Linear Regression, a Mixed Effect Model, and a Neural Network

**Linear Regression:**

Linear Regression is a simple and widely used statistical model for predicting a continuous outcome variable based on one or more predictor variables. It assumes a linear relationship between the outcome variable and the predictors, which can be limiting in cases where the relationship is more complex.

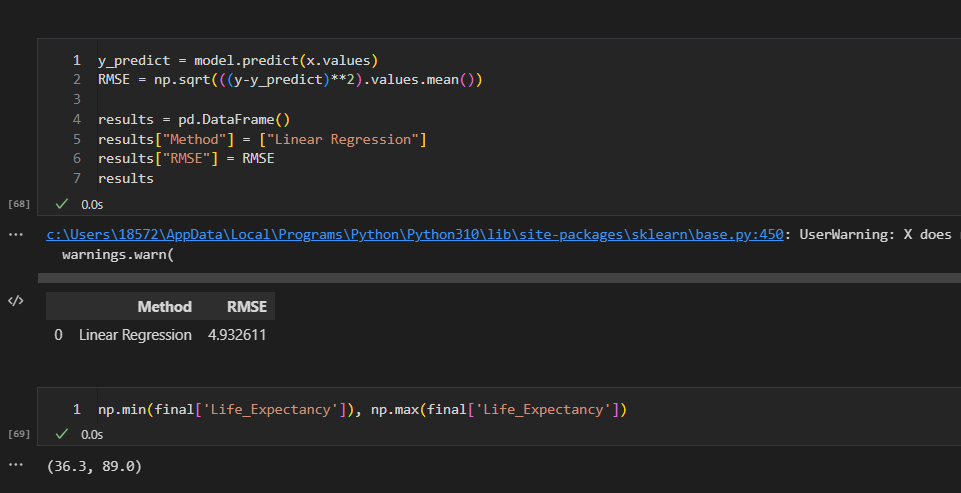
In this model, life expectancy would be the dependent variable or the outcome variable, and the predictor variables would be the independent variables. The goal is to find a linear relationship between the outcome variable and the predictor variables.

To build a linear regression model for life expectancy, you would need a dataset that includes observations of individuals along with their ages, sex, education, income, health behaviors, and life expectancy. You would use this dataset to estimate the linear regression equation coefficients and make predictions for new individuals.

It is important to note that linear regression assumes that the relationship between the outcome variable and the predictor variables is linear. Additionally, it assumes that the errors or residuals are normally distributed and have constant variance. Therefore, you should assess the assumptions of the linear regression model before using it to make predictions.

Graphical user interface, text, application

Description automatically generated



**Mixed Effect Model:**

A mixed-effects model can be a useful approach to analyze life expectancy data when there are repeated measurements on the same individuals, or when individuals are nested within groups (e.g., cities, countries, or families).

In a mixed-effects model, there are both fixed effects and random effects. Fixed effects are the variables that have a consistent effect on the outcome variable across all levels of the factor. Random effects are variables that have an effect on the outcome variable that varies across different levels of the factor.

For example, if you are analyzing life expectancy data for individuals in different cities, the city can be a random effect because life expectancy may differ across different cities, but the effect of the city on life expectancy may not be the same for all individuals.

To build a mixed-effects model for life expectancy, you would need to identify the fixed and random effects and include them in the model. The fixed effects can include variables such as age, sex, education, and income. The random effects can include variables such as city, family, or other group identifiers.

Mixed-effects models have the advantage of accounting for the correlation between observations within the same individual or group, which can lead to more accurate estimates of the effects of the predictor variables on the outcome variable. However, the interpretation of the model coefficients may be more complex compared to a simple linear regression model.

It is important to note that mixed-effects models assume that the random effects are normally distributed and have a constant variance across the different levels of the factor. Therefore, you should assess the assumptions of the mixed-effects model before using it to make predictions.

Text

Description automatically generated with medium confidence

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

**Neural Networks Model:**

A neural network can be a powerful approach for predicting life expectancy based on a wide range of predictor variables. Neural networks are a type of machine learning model that can learn complex non-linear relationships between input and output variables.

To build a neural network model for life expectancy, you would need a dataset that includes observations of individuals along with their ages, sex, education, income, health behaviors, and life expectancy. You would use this dataset to train the neural network to predict life expectancy based on the predictor variables.

Text

Description automatically generated with medium confidence

The neural network would consist of layers of interconnected nodes, with each node representing a mathematical function that combines the input variables to produce an output. The weights between the nodes are adjusted during the training process to minimize the difference between the predicted and actual life expectancy values.

One advantage of neural networks is that they can capture complex interactions and patterns in the data that may be difficult to capture using traditional statistical models such as linear regression or mixed-effects models. However, they can be prone to overfitting if the model is too complex or if there is not enough data to support the complexity of the model.

It is important to note that neural networks require a large amount of data and computational resources to train effectively. Additionally, the model architecture and hyperparameters need to be carefully chosen and tuned to avoid overfitting and achieve good performance on the test data.

Text

Description automatically generated

Text

Description automatically generated

**Logistic Regression:**

Logistic regression is typically used to model binary outcomes, such as the presence or absence of a disease, and is not well-suited for predicting life expectancy, which is a continuous variable. However, logistic regression could be used to predict a binary outcome related to life expectancy, such as the likelihood of an individual surviving to a certain age.

For example, you could use logistic regression to model the probability of an individual surviving to a certain age based on predictor variables such as age, sex, education, income, and health behaviors. In this case, the outcome variable would be binary, with 1 indicating survival to the specified age and 0 indicating death before that age.

To build a logistic regression model for this problem, you would need a dataset that includes observations of individuals along with their ages, sex, education, income, health behaviors, and survival status (1 or 0) at the specified age. You would use this dataset to estimate the coefficients of the logistic regression equation and to make predictions for new individuals.

It is important to note that logistic regression assumes that the relationship between the predictor variables and the outcome variable is linear on the logit scale. Additionally, it assumes that the errors or residuals are independent and have a constant variance. Therefore, you should assess the assumptions of the logistic regression model before using it to make predictions.

Text

Description automatically generated