Analysis of Life Expectancy as a function of Health and Environmental Conditions

Aanchal Narendran
Computer Science and Engineering
PES University
Bangalore, India
aanchalnarendran@gmail.com

Arushi Kumar
Computer Science and Engineering
PES University
Bangalore, India
arushi32001@gmail.com

Ayush Godbole
Computer Science and Engineering
PES University
Bangalore, India
godbole.ayush@gmail.com

Ayush Kapasi
Computer Science and Engineering
PES University
Bangalore, India
kapasiayush@gmail.com

***Abstract*—The main objective of this project is to aid in understanding the impact of various medical statistics and environmental conditions on the Life Expectancy of the citizens across a multitude of countries. These countries are either classified as developing and developed countries. We aim to develop a classification strategy that breaks away from this binary classification. The primary aim of the project is to use regression analysis to help in the prediction of Life Expectancy based on a few key features obtained via Exploratory Data Analysis and Feature Selection. The data used in this project has been collected from the World Health Organisation and includes various attributes such as Health Expenditure, Air Pollution, Unsafe Sanitation, GDP and so on. The project additionally aims to explore life expectancy as a function of time.**

***Index Terms*—Life expectancy, Linear Regression, Time Series Analysis, Feature Selection**

I. Introduction

Life Expectancy is a statistical measure used by Health Organisations around the globe to estimate how long an organism born in a country under specific demographics is likely to live. In this project, the aim is to predict the life expectancy of Humans across various countries around the globe based on a multitude of lifestyle, health and environmental factors. The World Health Organisation often estimates Life Expectancy based on country, age, sex and income group. Over the past few decades, leaps in the medical domain and increased expenditure in medical infrastructure has lead to an increase in the estimate of life expectancy.

This paper aims to predict the life expectancy of various countries around the globe without restricting its scope to prominent and famous countries. The objective is to avoid bias and to build a more generalised model that can predict life expectancy for a multitude of nations. The project also aims to harness the use of classification and clustering algorithms to help improve the categorization of a country. For instance, there are a significant number of countries that fall in the developing category but are better than their peers. Incorrect classification could result in a more conservative or liberal estimate of the countries life expectancy.

The goal of this project is to use regression analysis to predict the life expectancy of a certain country. The regression model is built on a few key factors, determined via Exploratory Data Analysis and Feature Selection to constrain the impact of human bias. This project also aims to explore life expectancy as a function of time using various Time Series Analysis models, contingent on the auto-correlation test. The secondary goal of the project is to enhance prediction by using clustering and classification algorithms.

II. Literature Review

A. Impact of pollution on Life Expectancy

The objective of this paper [1] was to gauge the impact of various air pollutants, such as greenhouse gases and particulate matter, on the life expectancy of the citizens of Romania using Machine Learning algorithms. The author focuses on understanding the longstanding impacts of these metrics from a socio-economic standpoint and assumes the negative impact on the environment to be negligible.

The author focuses on using correlation and linear regression to help in establishing the significance of the papers' finds. The paper mentions 3 main hypotheses considered by the author which focus on correlating and flagging the influence of the aforementioned air pollutants on the impact of Life Expectancy at birth and for senior citizens. The dataset chosen is collated from the National Institute of Statistics. It contains time series data about the various attributes considered from 2000 - 2015.

The results of the authors' hypotheses are verified using a hypothesis test for hypotheses 1 and 2, which talk about the correlation of air pollutants on the life expectancy at 65 and at birth, respectively. The author shows that both of these tests are partially verified and there exists a negative correlation between the two attributes. The author uses a linear regression model to show the influence of air pollutants on

life expectancy. The author shows that there is a significant relationship between the attributes

The author limits themselves to the impact of air pollution on life expectancy. It leaves space for researchers to assess the effect of other pollutants, such as water. Moreover, decreasing quantities of pollutants may not directly impact the increased life expectancy due to spurious correlation. In this case, the hidden variable is health expenditure. Between 2000 to 2015, the increase in health expenditure and the growth of the medical field could have contributed to this increase in life expectancy. In our project, we aim to address this by merging two datasets. This data accounts for both environmental conditions and key statistics such as health expenditure and GDP.

B. Life Expectancy as a function of Health status in EMR

The objective of this paper [2] is to estimate the health production function in the East Mediterranean region. Performed using econometric methods, it focuses on determining the relationship between life expectancy and various socio-economic factors from 21 countries using data from 1995 to 2007. The authors chose to use the Grossman theoretical model to build their model.

The Grossman model assumes that a human beings health is like capital goods. Over time the amount of goods decreases. However, it could be increased, to an extent, by investing in Medical care. The model assumes that at some point, the human reaches a minimum threshold. Below this threshold, the model assumes that the human cannot survive. Medical care is not limited to Hospital Expenditures, but also includes, Nutritional Intake, Income, Education and various other lifestyle traits. The model includes Social, Economical and Environmental attributes. The attributes considered were Income, Health Expenditure, Food Availability, Employment, Education, Immunization, Urbanisation and Carbon Dioxide exposure.

The author showed via the regression coefficients that income had an impact on life expectancy. An increase of 1% in average income resulted in a 0.05% increase in life expectancy. The author could conclusively show that the food production rate has an impact on life expectancy. However, the impact of Health Expenditure and Employment Ratio was inconclusive for different genders. The impact of Carbon Dioxide was rendered insignificant post-analysis.

The author admits that multicollinearity may have impacted variables such as income and employment ratio. Taking this into account, we have limited ourselves to GDP for our model. Moreover, the author generates a generalised model for the entire East Mediterranean region. It does not factor in the knowledge that certain countries in this region are extremely well-developed and wealthy while a few are war-torn. This could have resulted in a model with conservative or liberal estimates. The current projects aim to address this by using clustering and building optimised models for countries that are alike in terms of the attributes considered even though they are geographically apart.

C. Life expectancy as function of Health Expenditure cross-country

This paper [3] aimed to analyse the dynamic nature of the Healthcare domain owing to a constant increase in Healthcare expenditure. This paper views the Healthcare field as a function, whose inputs are infrastructure and expenditure, with its output being a change in life expectancy of the citizens. The paper goes one step above highlights the difference in a models significance based on the country.

The author applies panel data analysis grouped by the geographical location and income level of the country. The dataset used covers data from 1995 - 2010 for different countries. The project uses a fixed effects model for analysis. Statistically, a Fixed Effects Model is one where the parameters are fixed or not random variables. The author uses a Fixed Effects Model since the data is longitudinal, time-series data, i.e there exists time series data divided by countries within a fixed period. The Fixed Effects Model helps analyse the independent and dependent variables when spilt over an entity (here Country).

The author shows that developed countries tend to have a higher life expectancy than developing countries. However, there are cases where are a geographical region is developed, yet there is a variation in the life expectancy (The European Region). The author highlights the improvement of life expectancy among various Asian countries over the years. The author classifies the countries into 4 main groups based on income. Based on this classification, the author observes high values of R-Squared, which support the models' analysis.

The author supports the need to recreate this project on bigger datasets that consider other factors that directly impact the health status and consequentially, the life expectancy of the country. Despite Health Expenditure being a good metric to predict, there exist various Social, Economical and Environmental aspects that dictate the life expectancy of the citizens of a country. The current project addresses this issue by analysing a merged dataset that includes attributes from all the domains.

D. Analysis of Life Expectancy between developed and developing countries

The goal of this paper [4] is to use various regression models to establish a relation between various Socioeconomic factors and Life Expectancy. The paper focuses on classifying the countries into two non-overlapping divisions,i.e Developed and Developing countries. The comparative analysis of these models and the features seeks to show their effect on predicting Life Expectancy. The paper aims to predict Life Expectancy for a few developed (United States of America, United Kingdom, France, Germany and Australia) and developing countries (India, Russia, South Africa, Brazil and China).

The author performed Exploratory Data Analysis by handling the missing values and performing correlation analysis by using heatmaps. The author chose to implement three different regressors, i.e linear regressor, decision tree regressor and random forest regressor. The linear regression model was implemented using the Ordinary Least Squares estimate. The

author implemented all of these regressors using the Scikit-learn library of Python.

Based on the R-squared value, the author chose to use the Random forest regressor for analysis. It gave training and test R-squared values of 0.99 and 0.95. The author used feature selection to select attributes from the dataset. The author uses visualisation analysis to show and conclude the initial hypothesis, i.e the life expectancy of developed countries is better than the developing countries.

Although the author analysis the models, there is no explicit information about the performance of the models against each other. Apart from accuracy and error, the Author limits finding other metrics such as F-Score, Time taken and storage. The current project aims to address this by considering a different set of benchmarking metrics. The author does not consider the impact of various environmental conditions in the prediction of life expectancy. The current problem aims to address this by merging datasets with relevant data to aid in prediction.

The above resources aided us in isolating and formulation our problem statement which will be addressed in detail in the upcoming sections. Prior to this, a deep dive into the current dataset would assist in setting the stage for easier and better comprehension of the problem statement.

III. Dataset

The dataset for this project was created by integrating data from two different datasets publicly available on Kaggle [5] [6]. Both of these datasets contain information about various Health, Social, Economical and Environmental attributes. With 50 independent attributes and 1 dependent attribute, The project has ample feature space to navigate and test multiple hypotheses. The project contains data about 164 countries between the 2000 - 2015 period.

Due to a large number of attributes, a few of them were handpicked based on the literature review and domain knowledge for exploratory data analysis. Some of these attributes are GDP, Health Expenditure, Income, Schooling, Unsafe Water Sources, Unsafe Sanitation, Air Pollution levels, Diet and the Status of the country. However, these attributes are not set in stone and are subject to change based on correlation analysis and feature selection methods used during the Exploratory Data Analysis and Implementation phase of the project, respectively.

It is worth noting that there was a significant lack of available data in a few cases. Most of these missing values were in the early 2000s, where data collection was not as important a task. Moreover, a considerable amount of developing nations did not possess the means to collect and store data. The project makes use of various imputation methods to assess and handle any missing records.

IV. Data Preprocessing

This phase of the project was split into four main stages. The standard procedure is to perform Data Cleaning and follow it with Data Integration. However, in the case of this project, Data Integration was performed before Data Cleaning. The reason behind the swap is because the project is using two datasets. We also perform correlation analysis to understand the attributes better. The project was developed using Google Colaboratory.

A. Data Integration

As mentioned in the literature review, previous projects consider the health status, Socioeconomic or Environmental conditions. The main goal of the project is to analyse the impact of the conditions on Life Expectancy. The project started with this dataset [5] which includes information about a few Socioeconomic aspects, Health expenditure and the count of patients infected with certain diseases in a country. However, it did not contain information about environmental aspects such as Unsafe sanitation, Unsafe water resource, air pollution and so on. Due to these reasons, the initial dataset was integrated with another dataset [6] that contained information about these environmental conditions.

The two datasets were initially analysed for compatibility. Post this, they were merged using Pandas. The merge was performed based on the year and country. Due to the presence of redundant columns, a few attributes of the dataset were analysed and dropped.

B. Data Cleaning

Data cleaning is a crucial phase in the life cycle of a project. Data cleaning refers to the measures of handling incomplete, incorrect or redundant data. It is an important stage since unclean data can impact the model during development.

The first step was to clean and make the column names more uniform. Few of the columns had spaces as separators between words, while some had underscores. This discrepancy makes it harder to process the column names further down the line. Hence, the space separators in the column names have been replaced by underscores. The data is tested for duplicate values. The redundant columns have been dropped during integration.

The third step is to handle missing values. As mentioned previously, this data was collated from over 163 countries between 2000 to 2015. During the early 2000s, a considerable number of developing countries lacked the resources to collect and maintain data. Moreover, there was a considerable size of citizens from developing nations who lived in isolated and rural parts of the nation. The lack of proper connection made it harder for organisations to collect data. It resulted in several columns and rows possessing Null values. Manual and automatic imputation was performed. Due to the necessity of values from every year for time series analysis, the null values were imputed instead of being dropped.

C. Correlation Analysis

Correlation helps us in drawing an association relationship between two variables. However, correlation does not imply causation. Often, there are cases of spurious correlation wherein a latent variable forces two uncorrelated variables to appear correlated. The most common measure of correlation is Pearson's correlation. Pearson's correlation coefficient helps us identify a linear relationship between two continuous random variables. The range of this coefficient is [-1,1]. The higher the value of the coefficient, the stronger is the correlation.

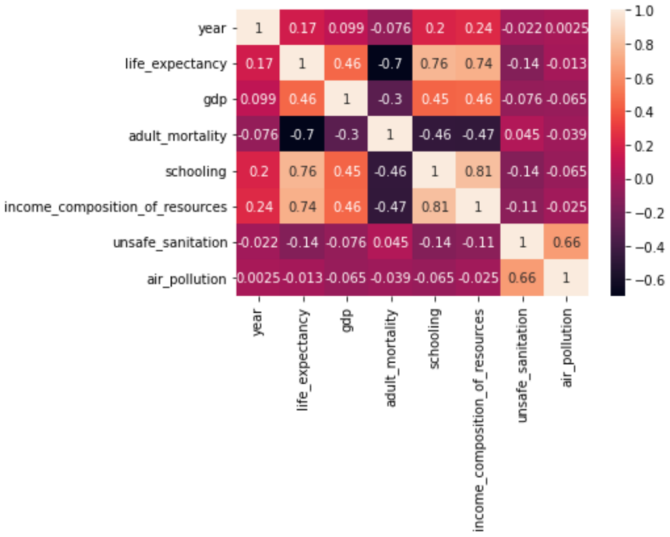

Fig. 1. Correlation Heat-map of certain variables

In the above figure 1, we see that schooling and income have a strong, positive correlation with life expectancy. Similarly, we notice that adult mortality has a strong, negative correlation with life expectancy. However, some factors such as pollution and GDP have a lower correlation coefficient. It does not imply that there is no correlation between these attributes. It is because Pearson's correlation coefficient is incapable of modelling non-linear relationships.

D. Visualisation

Visualisation helps us analyse the general trends of the data. It maps the skewness and spread of the data. It aids in building an ordered relation between measures of central tendency (mean, median and mode). It helps us track outliers far more effectively than textual processing of the data. In this project, we chose to visualise key attributes using histograms to assess the distribution.

The figure below 2 helps in analysing the spread of the data and the measures of central tendency. By default, we notice that the data has a negative skew. It implies that the mean is lesser than the median, which is lesser than the mode. We notice that the graph is Unimodal. The negative skew is likely caused by the fact that between 2000 to 2015, there has been an increase in Health Expenditure. Coupled with the amount of research in the medical domain has led to prolonged lives. Hence, shifting the median value which causes a negative skew.

Fig. 2. Histogram of Life Expectancy

V. Problem Statement and Approach

An in-depth analysis is required to understand the complexities of predicting Life Expectancy cross-country. The majority of the attributes do not share a linear relationship with a life expectancy, depicted by the correlation heat-map. It leaves us with space to analyse the impact of non-linearly dependent attributes, along with the linearly dependent attributes.

The first phase of implementation would involve building a feature selection algorithm. The reason is that with over 51 attributes in a dataset, the model must predict based on the best attributes. Feature selection models, such as Ridge Regression or Decision Trees, would be implemented. Lasso Regression is easier to implement. However, Ridge regression is a better choice since Lasso Regression completely invalidates features, whereas ridge regression reduces their contribution.

The main portion of this project is, tentatively, split into 4 main models. A simple linear regression model, for attributes that show a strong, linear correlation with life expectancy. The second is a polynomial or multiple linear regression model. The model will help in analysing the impact of variables that do not share a linear relationship with life expectancy.

The third model is a clustering model. The clustering model helps build regression models for a cluster of countries that share demographics. These are Schooling, Income, Adult Mortality and so on. It helps us generate a more generalised model as opposed to a specific model, which might give a conservative or liberal estimate. The final model is a time series analysis model to map the life expectancy as a function of time. The exact model will be decided post performing Durbin Watson to analyse auto-collinearity.

REFERENCES

[1] G. M. Toma, "Public health management: life expectancy and air pollution", in Conf. International Conference on Business Excellence,2017

[2] M. Bayati, R. Akbarian, Z. Kavosi, "Determinants of Life Expectancy in Eastern Mediterranean Region:A Health Production Function", International Journal of Health Policy and Management,pp. 57 - 61, 2013

[3] E. Jaba, C. Balan, I. Robu, "The relationship between life expectancy at birth and health expenditures estimated by a cross-country and time-series analysis", Procedia Economics and Finance, vol. 15, pp. 108 - 114, 2014

[4] S. S. Meshram, "Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning ", in conf. IEEE Bombay Section Signature Conference, 2020

[5] K. Rajarshi, "Life Expectancy (WHO)", 2018

[6] A. Verma, "Worldwide deaths by country/risk factors",2021