**Life Expectancy Prediction**

CSCI B-565 Data Mining

Final Project Report

**By**

***Laxmi Soujanya Taduru - ltaduru@iu.edu***

***Srichandana Chakilam - schakil@iu.edu***

Under the Supervision of

**Dr. Yuzhen Ye**

Chair, Computer Science

Professor of Informatics and Computer Science

Indiana University Bloomington

**ABSTRACT**

Life expectancy gauges the overall health of a community, thereby giving the scope of improving several factors that affect it. In the past few years, there has been a huge improvement in the health sector which has impacted the average age of death of the population. This also resulted in the overall deterioration of the human mortality rate, especially in the developing countries.

However, the factors affecting this were not completely analyzed. The effect of alcohol consumption, adult mortality rate, immunization like Hepatitis, Diphtheria and Measles, GDP of a country, and literacy rate have a larger role than expected.​

The goal of this project is to predict life expectancy based on different factors like immunization, mortality, economic and other health related factors.​

**Keywords**: Dataset, Features, Preprocessing, Data Visualization, Target Variable, Models, Accuracy score, r2 score.

**TABLE OF CONTENTS**

[1. INTRODUCTION 1](#_Toc121704265)

[2. METHODS 2](#_Toc121704269)

[3. RESULTS 10](#_Toc121704270)

[4. DISCUSSION 13](#_Toc121704272)

[5. REFERENCES 14](#_Toc121704273)

### 

### **INTRODUCTION**

### Many studies in the past have taken multiple factors into consideration to understand and analyze the life expectancy of people in a country but overlooked the importance of human development index, immunization, mortality, economic and other health related factors.​

### This project contributes to analyze and understand the factors that influence the life span of the population of several countries and proposing different algorithms and comparing their efficiency and performance in a Kaggle dataset.​ Given its predictor variables, the models implemented in the project also predict the life expectancy values.

### ***1.1 Dataset:***

The dataset used in the project has been made available by World Health Organization (WHO) for health analysis purposes. The data contains information related to life-expectancy, and its factors that has been collected from 193 countries and maintained in WHO repository website. The corresponding economic data has been collected from United Nations website. It contains data from the year 2000-2015.

This dataset is available on Kaggle. It consists of 2938 items and 22 columns, with ‘life expectancy’ being the target variable and 21 predictor variables.

### **METHODS**

The components of the project are:

1. Exploratory Data Analysis
2. Data Preprocessing
3. Data Visualization
4. Model implementation
   1. ***Exploratory Data Analysis:***

After loading the data, it is necessary to understand the structure and features of the data. Hence, we performed data analysis as the first step in this project. In this critical step, simple operations have been implemented on the data. For instance, the method – info() provides details about the features’ datatypes, count of non-null columns alongside feature names as shown below.

Table

Description automatically generated

Figure 1: Dataset Information

We have also identified the missing values in the dataset.

Table

Description automatically generated

Figure 2: Missing values per each column

* 1. ***Data Preprocessing:***

In this step, the raw dataset will be transformed into the data that can be fed into the models we implement. Some of the important preprocessing steps include:

1. Changing column names for easy access by using underscores and removing extra spaces.

Text, letter

Description automatically generated

Table

Description automatically generated

Figure 3.2.1: Renaming columns

1. Filling the null values that we identified during EDA, with the mean values of its corresponding attribute.

Table

Description automatically generated

Figure 3.2.2: Columns showing no null values after filling with mean values.

1. Converting the categorical variables into numerical variables using encoding techniques.

* There are two object-based attributes (‘Status’, ‘Country’) in the dataset. To apply the training models, these attributes must be handled.
* The column- *‘Status’* is a categorical attribute with only two values (‘Developing’ and ‘Developed’), it can be replaced with values ‘0’ and ‘1’.
* The column - *‘Country’* is a nominal attribute with 192 values, it can be handled using Feature Encoding; It converts categorical data into dummy or indicator variables that can be used to feed the regressor models.
* A similar manipulation technique has been used on the column - *‘Year’*.

Table

Description automatically generatedFigure 3.2.3: Categorical columns after being encoded into numerical variables.

1. Normalizing the data.

* The scale of data for some features may be significantly different from those of others, which may harm the performance of our models.
* Using Min-max scaling, we can limit the data values of a column to a specific range using each column’s minimum and maximum value.
* This has been applied to all the numeric attributes in our dataset.

Text

Description automatically generated

Figure 3.2.4: Normalization of all numerical features in the dataset.

* 1. ***Data Visualization:***

During this step, the preprocessed or formatted data is represented using visual elements like bar plots, histograms etc. A few observations have been made in the project

1. Using correlation analysis, the variables that impact the predictor variable can be found. ​
2. Life expectancy has positive correlation with schooling and Income composition of resources, thinness\_1to19\_years and thinness\_5to9\_years.
3. Life expectancy has negative correlation with adult mortality and HIV/AIDS.

Chart

Description automatically generated

Figure 3.3.1: Correlation heatmap of the dataset.

1. The features like schooling and Income composition of resources, thinness\_1to19\_years and thinness\_5to9\_years, adult mortality, HIV/AIDS may be good predictors of Life Expectancy.

Chart, bar chart

Description automatically generated

Figure 3.3.2: Bar plot showing the life expectancy for developing and developed countries.

* 1. ***Model Implementation:***

Regression is a technique for determining the relationship between independent variables or characteristics and a dependent variable or result. Once the link between the independent and dependent variables has been estimated, outcomes can be predicted. Following models have been implemented in this project:

1. Linear Regression
2. Ridge Regression
3. Decision Tree Regression
4. AdaBoost – Adaptive Boosting
5. XGBoost – Extreme Gradient Boosting
   * 1. *Linear Regression:*

Linear Regression is used to model the relationship between predictor and target variables.The target variable must be continuous in order to apply this model. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is an explanatory variable, and the other is a dependent variable.

* + 1. *Ridge Regression:*

Ridge regression can be used when there is multicollinearity in the dataset or when the number of predictor variables is higher than the number of observations. This method performs L2 regularization.

* + 1. *Decision Tree Regression:*

Numerical values can be handled by decision trees. It gradually builds an associated decision tree while dividing a dataset into smaller and smaller sections. This model regresses the data using true or false answers to certain questions.

* + 1. *AdaBoost Regression:*

AdaBoost stands for Adaptive boosting, is a meta-algorithm, which enables performance enhancement when combined with other algorithms. It starts by fitting one regressor on the original dataset, and then it fits subsequent copies of the regressor on the same dataset with the weights of the instances being changed in accordance with the error of the most recent prediction. As a result, successive regressors concentrate more on challenging cases.

* + 1. *XGBoost Regression:*

XGBoost stands for “Extreme Gradient Boosting”. It can be used directly for **regression predictive modeling**. The model depends on the primary learners which are not good at the remainder. But when all the predictions are combined, they sum up to form final good predictions.

### **RESULTS**

The models that are implemented in the model are ranked based on their r2 and accuracy scores. Below are the training and testing scores of each model. R2 score is an accuracy evaluation technique that represents the proportion of the variance for a dependent variable which is determined by an independent variable(s) in a regression model.

* 1. *Linear Regression:*

Text

Description automatically generated

*Figure 3.5.1: Train and test accuracy scores of Linear Regressor model.*

Overall, the accuracy for the model is 95.26% and can predict the life expectancy value for a given input.

* 1. *Ridge Regression:*

Text

Description automatically generated

Figure 3.5.2: Train and test accuracy scores of Ridge Regressor model.

The accuracy of the Ridge regressor for the given dataset is 95.42% and can predict the life expectancy value for a given input.

* 1. *Decision Tree Regression:*

**Graphical user interface, text

Description automatically generated**

Figure 3.4.3: Train and test accuracy scores of Decision Tree Regressor model.

The accuracy of the Decision Tree regressor for the given dataset is 91.83%.

* 1. *AdaBoost Regression:*

Graphical user interface, text, application

Description automatically generated

Figure 3.5.4: Train and test accuracy scores of AdaBoost Regressor model.

The accuracy of the AdaBoost regressor for the given dataset is 89.70%.

* 1. *XGBoost Regression:*

Graphical user interface, text

Description automatically generated

Figure 3.5.5: Train and test accuracy scores of XGBoost Regressor model.

The accuracy of the AdaBoost regressor for the given dataset is 89.65%.

**Table

Description automatically generated**

Figure 3.5.6: Test accuracy scores of all models.

Table

Description automatically generated

Figure 3.5.7: R2 Test accuracy scores of all models.

The above table of accuracy scores help us understand the performance of individual models when test data is applied.

### **DISCUSSION**

*Data limitations*: Our dataset consisted of roughly 3000 items which is a fair amount of data according to health sector perspective. However, it is not considered to be a lot of data for training regressor models. Although we have tried to split the data into test and train data by 20% and 80% respectively, more data would have been desirable for training the model, in order to increase the accuracy and reduce overfitting.

Another challenge is that there are some outliers present in the dataset. Since the size of the data is not very large, these outliers were not removed, which would adversely affect while training the models and hamper the performance. It is also possible that these outliers are not completely noisy to be eliminated.

*Interpretation of the output*: As part of this project, we have performed exploratory data analysis, pre-processed the data for it to be trainable to models and visualized it for analysis. Although the models – AdaBoost and XGBoost are extremely powerful and outperform the other models, the Linear regressor model and Ridge regressor works better as the current dataset is not large.

*Conclusions*: We choose to understand the best model that predicts the output target variable – ‘Life expectancy’ when different predictor variable values are fed into each model. All the models implemented in this project seem to predict the life expectancy given its independent variables. Using evaluating criteria, we could clearly see better performance being shown by Ridge and Linear regressors competing, albeit the presence of powerful regressors like Adaptive Boosting or XGBoost.

### **REFERENCES**

* WHO statistical information system [online database]. Life expectancy at birth. Geneva: World Health Organization; 2016

(<https://www.who.int/whosis/whostat2006DefinitionsAndMetadata.pdf>, accessed 4 January 2016).

* Global Health Observatory data repository. World Health Statistics [online database]. Geneva: World Health Organization; 2013 (<https://apps.who.int/gho/data/node.main.1?lang=en>, accessed 4 January 2016).
* World Bank. Data [online database] Washington (DC): World Bank; 2013 (<https://databank.worldbank.org/source/world-development-indicators>, accessed 27 June 2019).
* Klenk J, Keil U, Jaensch A, Christiansen MC, Nagel G. Changes in life expectancy 1950–2010: contributions from age-and disease-specific mortality in selected countries. Popul Health Metr. 2016;14(1):20–30.
* Ho JY, Hendi AS. Recent trends in life expectancy across high income countries: retrospective observational study. BMJ. 2018;362. pmid:30111634
* Kinsella K, Velkoff V. Life expectancy and changing mortality. Aging Clin Exp Res. 2002;5(14):322–32
* Sinkon Nayak, Manjusha Pandey, Siddharth S. Rautaray. A Proposal for Life Expectancy Analysis using Machine Learning Techniques (<https://www.researchgate.net/publication/365662993_A_Proposal_for_Life_Expectancy_Analysis_using_Machine_Learning_Techniques>)
* <https://www.researchgate.net/publication/356235215_Life_Expectancy_Prediction_Analysis_using_ML>
* Dataset - <https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>
* AdaBoost Regression - <https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>
* Linear Regression - <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>
* Decision Tree Regression - <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>
* Ridge Regression - <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>
* XGBoost Regression - <https://xgboost.readthedocs.io/en/stable/python/python_api.html>
* Feature Engineering - <https://www.projectpro.io/article/8-feature-engineering-techniques-for-machine-learning/423>