

Big Data Analytics for Life Expectancy Prediction

1st Priya Shrestha

Faculty of Science and Engineering
Herald College Kathmandu
Kathmandu, Nepal

2nd Riddhi Maskey

Faculty of Science and Engineering
Herald College Kathmandu
Kathmandu, Nepal

Abstract—Life expectancy refers to the overall health of the community and life-span of the community. Life expectancy is used as a vital indicator of overall health and socio-economic wellness in any country. Predicting life expectancy is a challenging task as it is dependent on multiple factors. This report attempts to demonstrate the life expectancy trend and life expectancy prediction using big data analysis. The proposed study used the Kaggle life expectancy data set collected by WHO between 2000 and 2015 from 195 countries around the world. The dataset consists of 22 columns and 2938 rows which includes adult mortality, infant death, population, schooling, GDP, percentage expenditure, various vaccination, thinness 1-19 and 5-9 years, country, status, year, income composition of resources, etc. PySpark was used for the handling and processing of the large dataset and predictive models such as multi-linear regression (MLR) and random forest are built to analyze and identify the key determinants of the life expectancy.

Index Terms—Life expectancy, Big data analysis, PySpark, Linear Regression, Random Forest, Predictive model, WHO.

I. BACKGROUND OF THE STUDY

Life expectancy refers to the average expected life span of individual within a community which is influenced by various socio-economic as well as health-related factors. It is one of the major indicator of wellness of the population. [1] This study uses a dataset for life expectancy, which is collected from the World Health Organization (WHO) repository and accessed through Kaggle source, which includes the life expectancy of 193 countries between the time frame of 2000 to 2015. The dataset includes 21 variables independent variable and one dependent variable - life expectancy.

The understanding of life expectancy trends and its future prediction is done using the big data analysis and various machine learning techniques. Collecting and analyzing the key determinants of life expectancy, this study builds predictive models such as Multiple Linear Regression (MLR) and Random Forest (RF) and ultimately identifying the best-fit model to predict global life expectancy.

A. Problem Statement

The accurate prediction of life expectancy is a challenging task, as it involves various independent variables that influence life expectancy.

- The data includes misleading data, missing data, null values as well as irrelevant data as well.
- Difficulty in identifying the key influential variable or factor from wide range of variables.

- It is a challenging task to build a best fit predictive model which generalizes perfectly across the 193 countries.

B. Aims and Objectives

The main purpose of this study is to identify the life expectancy trend and predict the global life expectancy based on various independent variable using big data analysis.

- To use PySpark for data collection and pre-processing.
- To identify the key factors affecting life expectancy using independent factors.
- To visualize relationships between the independent and dependent variables.
- To develop and evaluate predictive models with high accuracy.

C. Contributions of the Work Connected with Methodology

- Data Collection and Pre-processing using PySpark.
- Pre-processing includes handling of null values with the median imputation and dropping the columns which has lesser influence on life expectancy
- EDA using heatmap, scatter plot and some statistical graphs.
- Build predictive models Multiple Linear Regression and Random Forest.
- Compared models to identify the best fit for life expectancy prediction.

II. RELATED WORK

Many research works have been done on life expectancy prediction using big data analysis and machine learning techniques. In-depth study of existing studies and research conducted in related work was done to explore and gain insight into the trends of life expectations and the models used.

The main purpose of all existing studies was to perform the prediction of life expectancy using related independent factors of dataset of 193 countries [2].

The majority of the studies have explored and applied the multiple regression techniques to forecast the life expectancy trend. Random Forest is one of the popular predictive models used on the life expectancy dataset to identify the key influential factor or determinants for predicting the life expectancy trend [1]. Further research has explored the life expectancy trends and predictions using the statistical methods, which includes conducting the statistical analysis on the key determinants of life expectancy using Multiple Linear Regression

(MLR) [3]. Similarly, the study by [4] designed MLR as predictive model for the prediction of life expectancy at birth while also identifying the major socio-economic factors. These study do not integrate the large data analysis and focuses on the structured regression model.

Hence, to enhance the accuracy in the prediction, the recent studies have applied the machine learning models such as Artificial Neural Network (ANN) while also utilizing MLR. ANN was used to predict the life expectancy of Bangladesh using the GDP and population volume as key determinants. This study shows the better accuracy of machine learning model than the traditional statistic models [5]. Another crucial factor for improving the life expectancy prediction has been the feature selection. According to the study by [6] showed the impact of selecting the key influential features on the performance of the regression models. The study found out adult mortality, schooling and the BMI as the key determinants. While identifying the key influential features, the idea of knowing how to measure and identify the impact of various independent variables on dependent variable using quantitative analysis is important. The traditional studies often rely on the average life expectancy as key indicator which fails to capture variation across different population segments, leading to misleading analysis. The study by [7] suggested the use of percentile-based indicators gives more accurate view of the factors influencing the life expectancy. The study proposed by [9] suggested that the mortality rate the life expectancy is calculated and has introduces machine learning for life expectancy analysis to increase awareness of the suddent life expectancy changes.

While the above studies significantly contributed to life expectancy modeling, most either focused solely on traditional regression or limited machine learning techniques. Our work differentiates itself by adopting a comprehensive big data analytics approach using PySpark, allowing large-scale data processing. Moreover, we integrate multiple socio-economic, health, environmental, and educational factors into one predictive framework. We also place strong emphasis on data visualization through advanced techniques like dashboards and heatmaps, enabling multidimensional exploration of the data, and validating our model rigorously using multiple performance metrics (MAE, MSE, RMSE, R^2 , Adjusted R^2). Thus, our study offers a more scalable, holistic, and actionable analysis of global life expectancy trends.

III. METHODOLOGY

The dataset of life expectancy was used to predict, visualize, and evaluate various models and use it for analyzing the life expectancy trend and predicting the life expectancy of people globally using big data analysis. The main objective is to analyze the relationship between the dependent variable (life expectancy) with 21 other independent variables using PySpark in Google Colab. The following steps are used to build the prediction model and enhance the life expectancy prediction performance.

A. Proposed Methodology

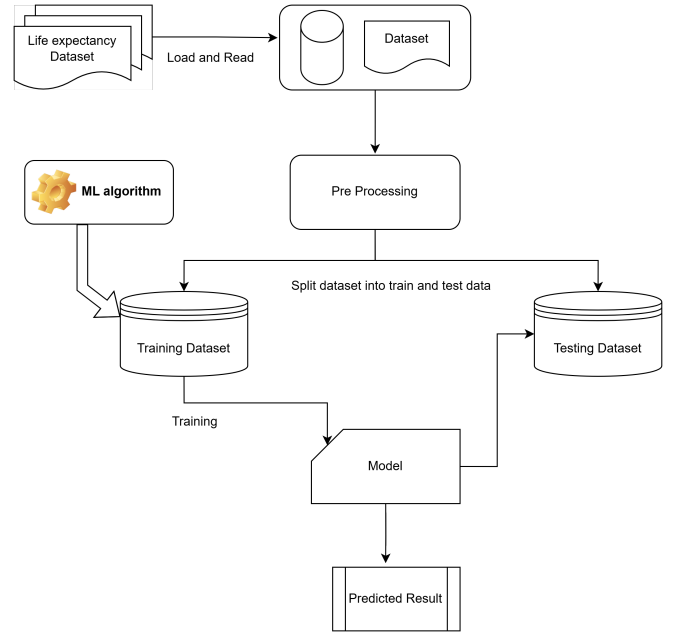


Fig. 1. System Architecture

This methodology utilizes machine learning and big data analytics techniques to predict life expectancy, considering various socio-economic and health-related factors. The model is developed through several steps using suitable techniques. The dataset selection and collection were done after exploring a wide range of datasets on Kaggle. The collected dataset was then loaded and read in the Google Colab using PySpark. On the next step, data preprocessing is done by identifying the mean, median from the datasets statistics and checking for null and missing values. The null values were then handled by the median imputation. On the next stage, the model training is done using Multiple Linear Regression (MLR) and Random Forest (RF) techniques, and then the model is sent for testing. Here, the evaluation and prediction takes place, and the final result is predicted.

B. Data Understanding and Pre-processing

The initial phase of our analysis began by importing and loading the dataset. Then all the necessary information about the dataset is studied, which helped us to understand its structures and the type of data it contains. It involved finding out the shape of the dataset, columns, datatypes, as well as statistical information of the dataset. Following this, the focus was on processing the data, which involves these steps like: data cleaning, visualization, and large-scale analysis to understand the information better. During the pre-processing, lags to the datasets are implemented, and in the same way, data cleaning is done by detecting the null values and replacing them by median imputation. Here is the descriptive data of the dataset.

summary	Adult Mortality	Infant deaths	Alcohol percentage expenditure	Hepatitis B	Measles	BMI	Polio
count	2938	2938	2744	2938	2385	2938	2919
mean	164.7954408761681	38.30394026422257	4.60280507772075	738.2522954533823	80.30804612259291	2419.592239518704	38.32214655647373
stddev	124.2920790034231	117.9265013133996	4.85242658755558	1987.914858016134	25.870015593818063	11467.27248923461	20.844033582562083
min	1	0	0.01	0.0	1	0	1.0
max	723	1000	17.87	19479.91361	99	212383	87.3

Fig. 2. Descriptive Dataset

Diphtheria	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Life expectancy
count	2919	2904	2904	2771
mean	82.32408359027065	4.8397038567493205	4.870316804407711	0.6275510645976166
stddev	23.7169120685726	4.420194947144322	4.508882086983007	0.2109035551515931
min	2	0.1	0.1	0.0
max	99	27.7	28.6	0.948

Fig. 3. Descriptive Dataset

C. Training Dataset

The dataset is split into training and testing data in the ratio of 70:20. Training step includes training the model to learn the pattern and trend and identifying the relationship between the variables. Two predictive models are build (MLR and RF) to analyze which one performs better.

Descriptive Statistic of training data:

summary	life_expectancy
count	2112
mean	69.42410037878793
stddev	9.429343399961155
min	36.3
max	89.0

Fig. 4. Performance in training data

D. Model

For modeling the dataset, two predictive models are used MLR and RF, which will help to find out the relationship of the dependent variable with the independent variables.

1) *Multiple Linear Regression*: Multiple Linear Regression is a statistical method that is used to estimate the outcome of a dependent variable by examining the influence of two or more independent variables. It is also known as the extended model version of the simple linear regression model, which predicts the value of a single variable by improving accuracy by using the knowledge of the multiple factors. This technique helps to analyze how much each independent variable influences the overall model's variability and the outcome. Usually, multiple regression is used for linear relationships between the variables, but it can also be used for non-linear relationships, as both dependent and independent variables do not establish a straight line. However, in the case of non-linear regression, which is complex as its execution is difficult and the trial and error method, which is developed through assumption, is used for the accurate modeling. [10]

2) *Random Forest*: Random Forest is a learning method that is made up of a collection of decision trees that uses the random samples of data to make predictions. Each Tree of the forest is unique because each tree is trained on a different random subset of the data. When creating each tree, the algorithm randomly. the selection of the subset of features to split on is done randomly by the algorithm when creating each tree, which adds diversity among the trees. Each tree in the forest makes its own prediction based on the data it was trained with. The final prediction is done by calculating the average of all the predictions made by all the trees. This method helps to minimize the overfitting, which leads to improving the accuracy and reliability of the model's predictions. [11]

E. Testing Dataset

After the training the dataset, the testing of the datasets is performed, the described model is represented as:

Descriptive Statistic of testing data:

summary	life_expectancy
count	826
mean	68.7504842615012
stddev	9.699071171168331
min	44.6
max	89.0

Fig. 5. Performance on testing data

IV. RESULT AND DISCUSSION

Should cover the experimental setup, discussion of the findings, and analysis of the findings with the below mentioned subtitle:

A. Read In and Explore the Data

The "Life Expectancy" dataset was collected from the WHO across 193 countries and sourced from Kaggle. The dataset was then imported and loading into Google Collab using PySpark. Instead of traditional data analysis tool like Pandas, PySpark has been used as it to handle data processing as it is beneficial for small to large dataset unlike Pandas. In the initial inspection of the dataset using pyspark, the shape of the dataset is (2938, 22) and printSchema() showed the mixed data types in the dataset.

B. Data Analysis

- To understand the dataset for the data cleaning and pre-processing, first of all the detailed and descriptive statistical information of the data was retrieved. The statistic showed mean, standard deviation and range of life expectancy, missing values etc.

```
df.describe() #Show the descriptive statistics of the dataset and display the results
```

	Country	Year	Status	Life expectancy	Adult Mortality	Infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	GPI
count	201	201	2	2028	2028	2028	2028	2028	2028	2028	2028
mean	2007	2007	0.5	72.0	2028	2028	2028	2028	2028	2028	2028
std	1.0	1.0	0.0	10.0	2028	2028	2028	2028	2028	2028	2028
min	2007	2007	0.0	40.0	2028	2028	2028	2028	2028	2028	2028
max	2017	2017	1.0	90.0	2028	2028	2028	2028	2028	2028	2028

Fig. 6. Dataset Statistic

C. Data Visualization

Visualization techniques like Heatmaps and Scatter plots were used to analyze the relationship between the independent variables with dependent variable (Life expectancy). Furthermore, box-plot for the 'status' vs 'Life expectancy' was used to identify the outliers present and a histogram was used to see the frequency of life expectancy.

1) *Histogram*: The figure shows the histogram representing the distribution of life expectancy on the x-axis and its value (frequency ranging from 0 to 400) on the y-axis in the dataset across the different age groups (40-90 years) along with the kernel density estimate (KDE) curve, which helps in visualizing the overall distribution trend. From the image, we can conclude that the life expectancy of most of the individuals on average is around 65 -80 years, with the peak around 70-75 years, as bars are tallest here. There are also fewer individuals with lower life expectancies under 50 years (frequency below 50) and higher life expectancies above 80 - 90 years (frequency below 100) as shown by the shorter bars on both ends. The chart shows the right-skewed distribution, which means the majority of the values are mostly concentrated in the range of 70 - 80 years, living longer lives, while a minority of the values are stretching towards the left, indicating they have significantly lower life expectancy. Overall, it shows that the majority of the people are likely to live into their 70s in this dataset.

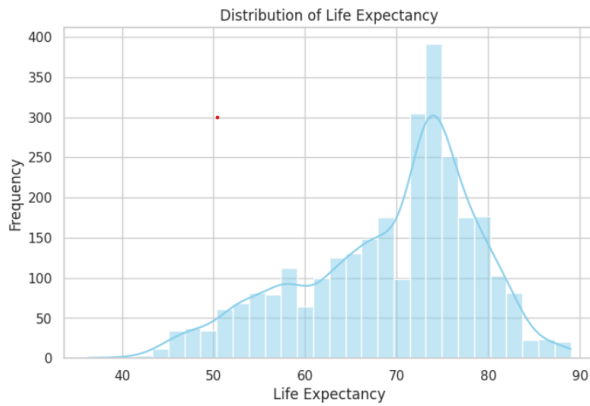


Fig. 7. Histogram of Life Expectancy Distribution

2) *Scatter Plot*: The scatterplot shows the relationship between the target variable (Life expectancy) and other features. Some features, like 'life expectancy', 'country', and 'year' are excluded when plotting the features with life expectancy. The scatterplot is used to pinpoint the

correlations and identify the trends between each existing feature and life expectancy. A rise in the trend means positive correlation, a decline means negative correlation, while a scatter pattern indicates no clear relationship (without linear trends). According to the plot, the indicators (adult mortality, infant deaths, under-five deaths, and HIV/AIDS) that have a strong negative relationship with life expectancy show that higher mortality rates are linked with low life expectancy. Where positive relationships exist with factors like schooling, income, and vaccination rates (like polio and diphtheria), associated with high life expectancy. The scatter ones such as alcohol and BMI, suggest no linear relationships. These visualizations highlight which features might have more impact on life expectancy.

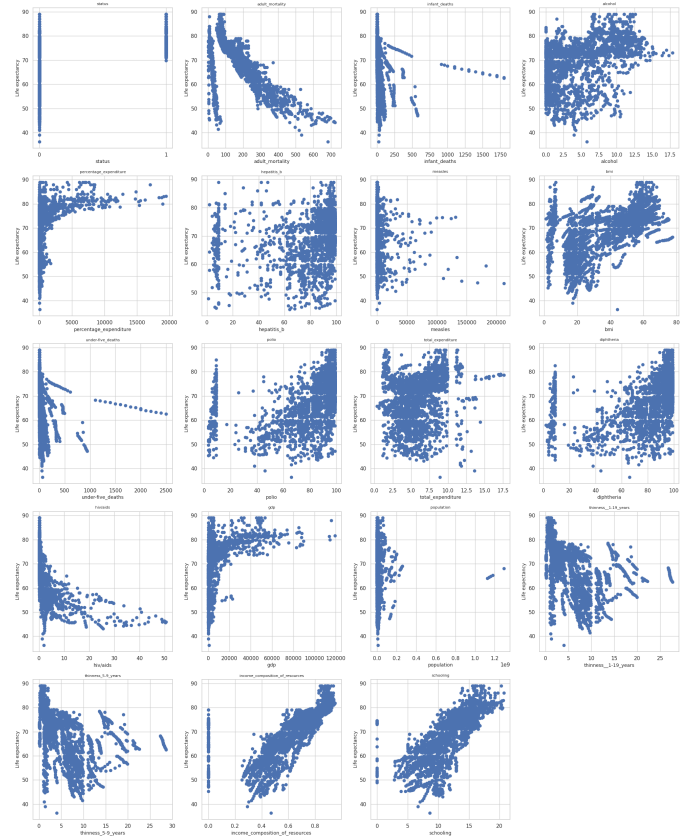


Fig. 8. Scatter Plot

3) *Box-Plot*: The distribution of life expectancy is done between developed and developing countries, which range from 40 - 90 years. The median of developed countries is high, which is around 80 years, with a narrow box pinpointing the more consistent outcomes indicating higher average life expectancy, whereas the median of developing countries is low, which is around 70 years, along with a wider box, which indicates the greater variability in life expectancy showing lower average life expectancy. There are a number of outliers present at the lower ends of the developing countries, which show the

differences between them. The box plot represents the high and consistent life expectancy of developed countries as compared to developing countries, which is highly likely due to their living conditions, access to healthcare facilities, infrastructure, and some socio-economic factors as compared to the developing countries.

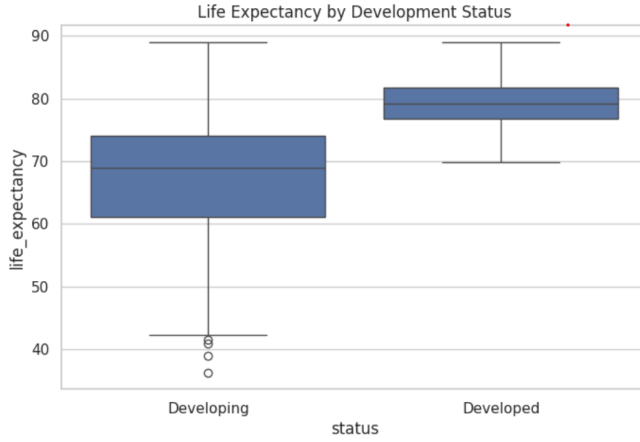


Fig. 9. Box Plot

4) *Correlation Heatmap*: The visualization shows the correlation heatmap representing the strength of the relationship between pairs of features in a dataset. The color coding indicates the correlation values between the variables. The red shade shows the strong positive correlation with the values closer to +1, the blue shade shows the strong negative correlation closer to -1, and the white or neutral shades show little or no correlation with the value around 0. The red shades, which follow the diagonal line, indicate that all the features are perfectly correlated with itself. From the graph we can say that there is a strong positive relationship between schooling and income_composition_of_resources (0.80) features, while strong negative relationship between life_expectancy and adult_mortality (-0.70) features. This heatmap provides a clear overview of how the variables are associated with each other and helps in analysis for feature selection.

D. Feature Selection

Feature selection was done based on the correlation analysis using correlation heatmap. Following analysis was made with the correlation heatmap:

1) Strong Correlation with Life Expectancy:

- Schooling (0.75) and Income composition of resources (0.72) showed the strong correlation with the life expectancy.
- While, adult mortality (-0.70) and HIV/AIDs (-0.56) showed strong negative correlation with the life expectancy.

2) Weak Correlation with Life Expectancy:

- Population (-0.02) has almost no correlation with life expectancy, while year, total expenditure and

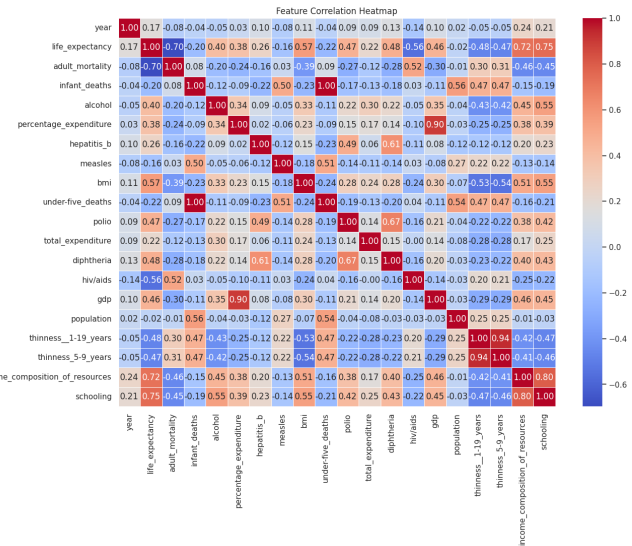


Fig. 10. Correlation Heatmap

measures have very weak and minor correlation with life expectancy.

3) *Drop Features with minimal influence*: Based on the above analysis using correlation heatmap, following features were dropped:

- Population
- Total Expenditure
- Year
- Total Expenditure
- Thinness 5-9 as it was redundant with Thinness 1-19 years
- Under-five deaths as it was perfectly correlated with infants death
- Measles
- Percentage expenditure
- Alcohol
- Country

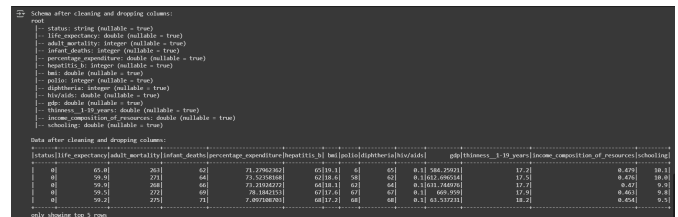


Fig. 11. Dataset after dropping columns

E. Cleaning Data

Data cleaning is an important process to ensure the quality of the data, so that it does not affect the performance of the model. It is done by searching for the presence of null data as well as duplicate data in the dataset.

- The inspection of the duplicate data was identified by grouping the data by columns and counting the number

```

null_counts = df.select([sum(df[c].isNull().cast("int")).alias(c) for c in df.columns])
null_counts.show()

```

	country	year	status	life_expectancy	adult_mortality	infant_deaths	alcohol	percentage_expenditure	hepatitis_b	measles	hiv	under-five_deaths
0	0	0	0	10	10	0	194	0	553	0	34	0

Fig. 12. Before Handling Null Values

```

from pyspark.sql.functions import col, sum, when

# Count nulls in each column
df_imputed.select(
    sum(when(col(c).isNull(), 1), otherwise(0)).alias(c)
    for c in df_imputed.columns
).show()

```

	country	year	status	life_expectancy	adult_mortality	infant_deaths	alcohol	percentage_expenditure	hepatitis_b	measles	hiv	under-five_deaths
0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 13. After Handling Null Values

of duplicated records. This inspection confirmed that now duplicated rows were found.

- In the data cleaning process, missing values were checked, as a result, a total of 2563 null values were found. This was handled by the median imputation to ensure the accuracy of the data.
- Categorical data like 'status' was encoded.

```

+-----+
| status |
+-----+
| 0      |
| 1      |
+-----+

```

Fig. 14. Handling Categorical Data

- The column names were cleaned by removing extra spaces and special characters using a standardized format, and the columns were renamed for consistency.

```

df = df.withColumnRenamed(' ', '_') for c in df.columns)
print(df.columns)

['country', 'year', 'status', 'life_expectancy', 'adult_mortality', 'infant_deaths', 'alcohol', 'percentage_expenditure', 'hepatitis_b', 'measles', 'hiv', 'under-five_deaths', 'total_expenditure']

```

Fig. 15. Cleaned Column names

F. Model Development

- **Vector Assembler:** For modeling vector assembler was used to combine the feature columns into a single vector column.
- **Split Training and Testing Data:** Then, the data was split into 70 - 30 ratio for training and testing, respectively. Data split into a 70%-30% ratio.

```

Descriptive Statistic of training data:
+-----+-----+
| summary | life_expectancy |
+-----+-----+
| count   | 2112            |
| mean    | 69.424100378793 |
| stddev  | 9.429343399961155 |
| min     | 36.3            |
| max     | 89.0            |
+-----+-----+

Descriptive Statistic of testing data:
+-----+-----+
| summary | life_expectancy |
+-----+-----+
| count   | 826             |
| mean    | 68.7504842615012 |
| stddev  | 9.699071171168331 |
| min     | 44.6            |
| max     | 89.0            |
+-----+-----+

```

Fig. 16. Training and Testing dataset

Models Used:

- Multiple Linear Regression (MLR)
- Random Forest Regression

G. Evaluation Metrics

Several matrices were used to evaluate the performance of the regression models, such as:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (R^2)
- Adjusted R-squared (Adj R^2)

These metrics were used to assess how well the models fit the data and their prediction accuracy on unseen test data.

H. Choosing the Best Model

The Best Model for predicting the life expectancy is chosen based on the evaluation metrics and the performance of the model for both Multiple Linear Regression (MLR) and Random Forest (RF) models.

1) Multiple Linear Regression:

Linear Regression Metrics Evaluation

Metric	Value
Mean Absolute Error (MAE)	3.1590
Mean Squared Error (MSE)	18.5168
Root Mean Squared Error (RMSE)	4.3031
R^2 Score	0.8029
Adjusted R^2 Score	0.8000

2) Random Forest:

Random Forest (RF) Metrics Evaluation

Metric	Value
Mean Absolute Error(MAE)	2.0003888720436325
Mean Squared Error(MSE)	7.708318404966506
Root Mean Squared Error(RMSE)	2.776385853041055
R ² Score	0.9179600377145903
Adjusted R ² Score	0.9167491157620382

After modeling the data:

- The Random Forest (RF) demonstrated a better performance on predicting the life expectancy than the Multiple Linear Regression (MLR). The significant lower values of MSE and MAE than those of MLR shows the better accurate prediction with fewer errors. The RMSE of 2.88, which shows the error is lower in the RF model. Furthermore, R2 and Adjusted R2 are higher in RF, which means RF explains the variation better than making the RF model reliable for prediction. We can conclude that the RF model produces more accurate predictions and it learns the pattern of the data better. Hence, the Random Forest (RF) has outperformed the Multiple Linear Regression model, making it an ideal choice for the prediction of life expectancy.

```

----- MLR (Multiple Linear Regression) Results -----
MAE 3.1590295187183224
MSE 18.51683197338581
RMSE 4.30311886582114
R2 0.8029245657829778
Adj R2: 0.8000157032853096

----- RF (Random Forest) Results -----
MAE: 2.0003888720436325
MSE: 7.708318404966506
RMSE: 2.776385853041055
R2: 0.9179600377145903
Adjusted R2: 0.9167491157620382

```

Fig. 17. Evaluation (MLR Vs RF)

V. CONCLUSION

The goal of this study is to build, visualize, and evaluate the models for predicting the global Life Expectancy trends using machine learning techniques and big data analytics using the Life Expectancy dataset from Kaggle. The Multiple Linear Regression and the Random Forest model were used for the life expectancy prediction, where the target variable was life expectancy, with other independent variables.

The study demonstrates that for the prediction of the life expectancy, random forest outperformed the multiple linear regression. The clear interpretability was provided by linear regression, while random forest offered better predictive evaluation on the dataset.

Life expectancy is the crucial factor to determine the wellness of an individual and community, which is influenced by several socio-economic as well as health-related factors. Big data analysis is used to analyze the pattern of life expectancy trends and build a perfect predictive model with higher accuracy.

REFERENCES

- [1] Deshpande, R. and Uttarkar, V., "Life Expectancy using Data Analytics," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, ISSN: 2321-9653, Volume 11, Issue IV, April 2023.
- [2] D. A. S. S. Vikram Bali, *Life Expectancy: Prediction & Analysis using ML*, in *Proceedings of Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2021, pp. 1–8.
- [3] V. Abhinaya, B. C. Dharani, A. Vandana, P. Dr. Velvadivu, and C. Dr. Sathya, *Statistical Analysis On Factors Influencing Life Expectancy*, *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, July 2021.
- [4] K. A. Alsalem et al., *Predicting Life Expectancy at Birth*, in *Proceedings of the International Conference on Computer and Information Sciences*, 2020.
- [5] M. A. Rubi et al., *Life Expectancy Prediction Based on GDP and Population Size of Bangladesh using Multiple Linear Regression and ANN Model*, in *Proceedings of the International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2021.
- [6] M. Biltawi et al., *The Impact of Feature Selection on the Regression Task for Life Expectancy Prediction*, in *Proceedings of the International Conference on Emerging Trends in Computing and Engineering Applications*, 2022.
- [7] V. Galabov, *How to Measure the Impact of Factors Affecting Life Expectancy*, in *Proceedings of the Metrology and Metrology Assurance Symposium*, 2019.
- [8] He, X. et al., "Analysis on Relevant Factors Affecting Life Expectancy," *IEEE*, Dalian, China, 2022.
- [9] S. Nayak et al., *A Proposal for Life Expectancy Analysis using Machine Learning Techniques*, in *Proceedings of the International Conference on Smart Electronics and Communication*, 2022.
- [10] S. Taylor, "Multiple Linear Regression," *Corporate Finance Institute*, 2025. [Online]. Available: <https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/>. [Accessed 9 May 2025].
- [11] GeeksforGeeks, "Random Forest Algorithm in Machine Learning," *GeeksforGeeks*, 16 January 2025. [Online]. Available: <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>. [Accessed 9 May 2025].