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 life expectancy trends and life expectancy prediction using big data analysis. The proposed study used the Kaggle life expectancy dataset 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 Multiple Linear Regression (MLR) and Random Forest (RF) are built to analyze and identify the key determinants of life expectancy. The performance matrix was evaluated and based on that RF achieved an R^2 of 91.8%, while MLR achieved an R^2 of 80.29%. This demonstrates that RF outperformed the MLR model in prediction accuracy.**

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

I. Background of the study

Life expectancy is the average expected life span of an individual within a community, which is one of the major indicators of community's wellness, influenced by various factors [1]. This study uses a life expectancy dataset, 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 independent variables and one dependent variable, i.e. life expectancy.

Understanding of life expectancy trends and its future prediction is done using 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), ultimately identifying the best-fit model to predict global life expectancy.

A. Problem Statement

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 trends and predict the global life expectancy based on various independent variables 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 several prior studies was to predict life expectancy using statistical and machine learning techniques, focusing on basic demographic factors and mortality rates. Due to this, many other influential factors were often overlooked such as schooling, healthcare access and socioeconomic factors. The need for multi-factor models have emphasized by recent works as they better reflect the complexities of longevity prediction [2].

Majority of studies have explored and applied multiple regression techniques to forecast life expectancy trends. Random

Forest is one of the popular predictive models used in the life expectancy dataset to identify the key influential factor or determinants predicting life expectancy trends [1]. Further research has explored life expectancy trends and predictions using statistical methods, conducting statistical analysis on 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 socioeconomic factors. These studies did not integrate large data analysis and focus on the structured regression model.

Hence, to enhance the accuracy in the prediction, the recent studies have applied machine learning models such as Artificial Neural Network (ANN) while also utilizing MLR. ANN was used to predict the life expectancy of Bangladesh using GDP and population volume as key determinants. This study shows better accuracy of machine learning model than traditional statistic models [5]. Another crucial factor for improving life expectancy prediction has been 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. Traditional studies often rely on the average life expectancy as a key indicator, which fails to capture variation across different population segments, leading to misleading analysis. The study by [7] suggested that the use of percentile-based indicators, gives more accurate view of the factors influencing life expectancy. The study proposed by [9] suggested that the mortality rate is calculated and introduced machine learning for life expectancy analysis to increase awareness on sudden change in life expectancy rate.

Multiple linear regression (MLR) models were used, considering 18 independent variables to investigate the various determinants influencing life expectancy. The study [8] revealed disease prevalence, GDP and education as the major factors influencing life expectancy as. Their model achieved overall 3.818 Root Mean Square Error (RMSE), confirming that different models are required for countries with doifferent status due to differing influencing factors.

While the above studies significantly contributed to life expectancy modeling, they either focused solely on traditional regression or limited machine learning techniques. Our work differentiates itself by adopting 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 life expectancy dataset was used to predict, visualize, and evaluate various models and use it for analyzing the life expectancy trends 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

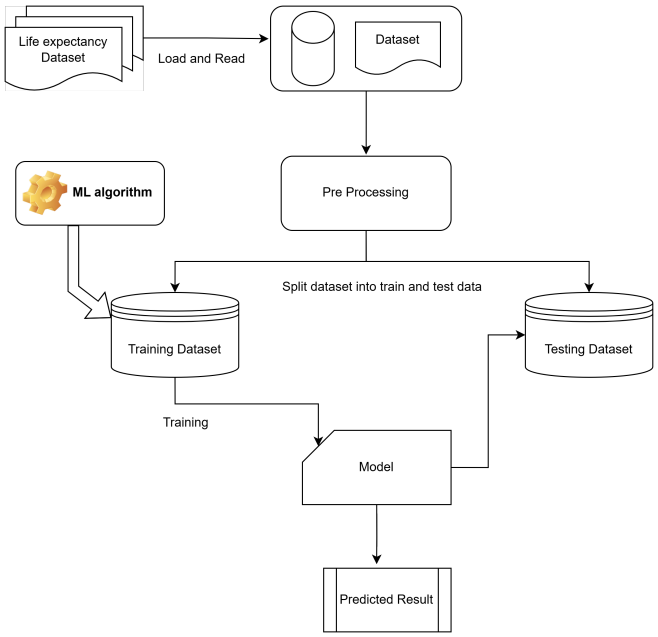


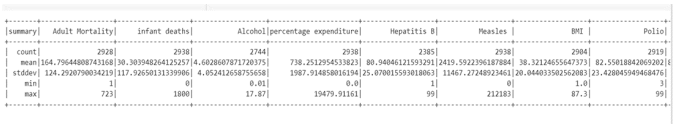
Fig. 1. System Architecture

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

B. Data Understanding and Pre-processing

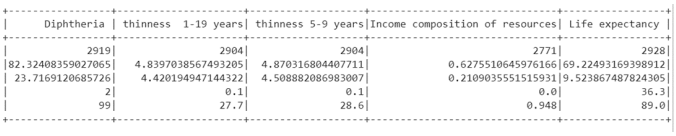
The initial phase of our analysis began with importing and loading the data set. Then, all the necessary information

about the dataset is studied, which helped us to understand its structures and the type of data it contains. It involved finding the shape of the dataset, columns, datatype, as well as statistical information about the dataset. Following this, the focus was on processing the data, which involves these steps such as data cleaning, visualization, and large-scale analysis to better understand the information. During preprocessing, 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. Median imputation was chosen to replace the null because median is less affected by outliers and is a more robust measure of central tendency, while doing better representation even for skewed distributions [13]. In the dataset,various factors such as GDP, mortality and population had a larger standard deviation and a smaller mean, indicating outliers and skewness. Here are the descriptive data for the dataset.



summary	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	Polio
count	2928	2938	2744	2938	2385	2938	2904	2919
mean	164.79644808743168	30.303948264125257	4.602660787172075	738.2512954553823	80.94046121593291	2419.592239618784	38.32124655647373	82.55018042069202
stddev	124.29207900342191	117.92650131339906	4.052412658755658	1987.9148580816194	25.070015593018063	11467.272489234661	20.044033502562083	23.428045949468476
min	1	0	0.01	0.0	1	0	1.0	3
max	723	1800	17.87	19479.91161	99	212183	87.3	99

Fig. 2. Descriptive Dataset

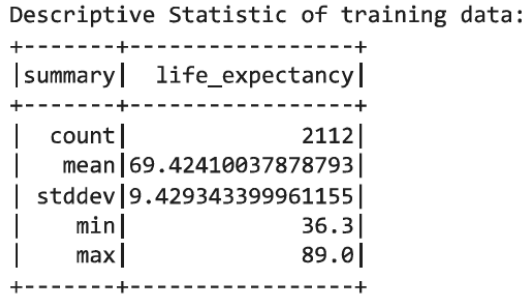


Diphtheria	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Life expectancy
2919	2904	2904	2771	2928
82.324083590270 65	4.8397038567493205	4.870316804407711	0.6275510645976166	69.224931693989 12
23.7169120685726	4.420194947144322	4.508882086983007	0.2109035551515931	9.523867487824305
2	0.1	0.1	0.0	36.3
99	27.7	28.6	0.948	89.0

Fig. 3. Descriptive Dataset

C. Training Dataset

The dataset is split into training and testing data in a ratio of 70:30. 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] The formula of the multiple linear regression is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \qquad (1)$$

In this equation, Y represents the predicted outcome of the dependent variable, while $X_1, X_2, \ldots, X_n$ represents the independent variables.β_0 is the intercept, which indicates the expected value of Y where all the predictor (independent variables) values are set to 0. The $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of each independent variable, which show how much influence it has on Y. The residual or error term is denoted by ϵ. [11]

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. [12] Each tree in the forest makes its own prediction based on the data it was trained with. The final prediction is made 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. and it has advantages over traditional decision trees as this model has the capability to handle the high-dimensional data and identify complex relationships. [14]

E. Testing Dataset

After the training the dataset, the testing of the dataset 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

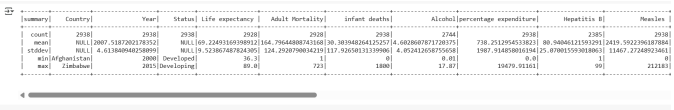
Experimental Setup: First of all, the dataset was chosen from Kaggle, imported them and Exploratory Data Analysis (EDA) was done to know about the data structure, patterns, and relationship among the features. An inspection was also done to check if the data is cleaned or not; if not data cleaning process was executed along with the data visualization for a better view and understanding of the data. After cleaning the data, the model was selected for predicting the life expectancy. How experimental setup was for this study is explained in brief below.

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.



summary	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles
count	2938	2938	2938	2928	2928	2938	2744	2938	2385	2938
mean	NULL	2007.5187202178352	NULL	69.22493169398912	164.79644808743168	30.303948264125257	4.602860787172075	738.2512954533823	80.94046121593291	2419.592239618784
stddev	NULL	4.613840944258099	NULL	9.523867487824305	124.29207900342219	117.92650133339906	4.052412658755658	1987.9148580816194	25.070015593018063	11467.272489234616
min	Afghanistan	2000	Developed	36.3	1	0	0.01	0.0	1	0
max	Zimbabwe	2015	Developing	89.0	723	1800	17.87	19479.91161	99	212183

Fig. 6. Dataset Statistic

The statistic showed mean, standard deviation and range of life expectancy, missing values etc.

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 left-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.

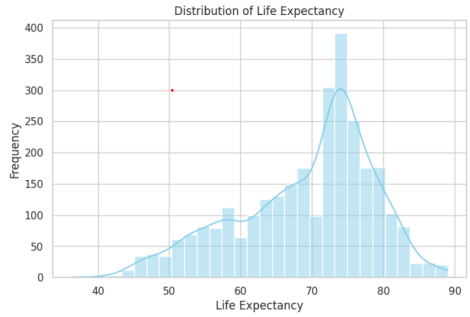


Fig. 7. Histogram of Life Expectancy Distribution

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.

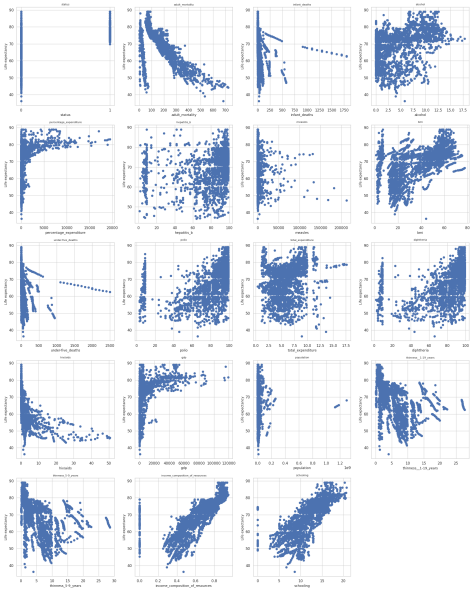


Fig. 8. Scatter Plot

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 socioeconomic factors as compared to the developing countries.

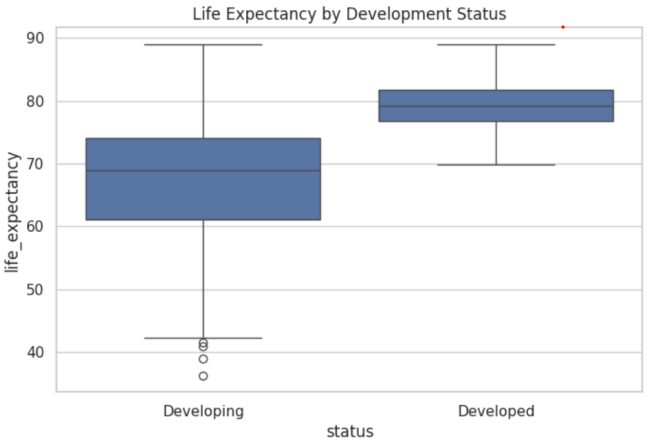


Fig. 9. Box Plot

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.

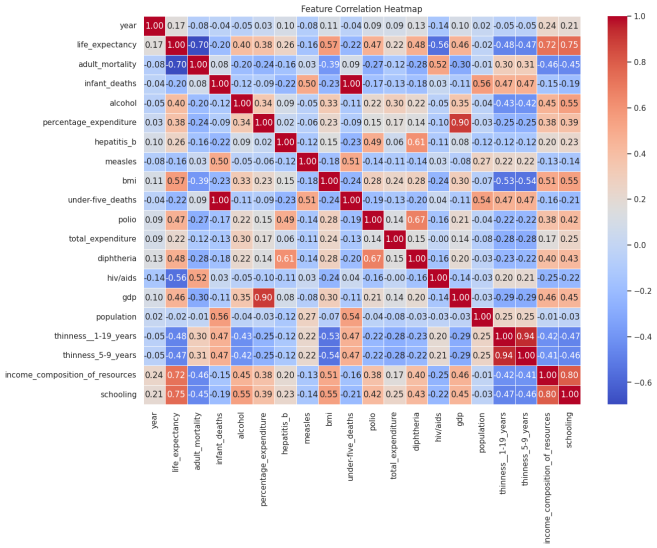


Fig. 10. Correlation Heatmap

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 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

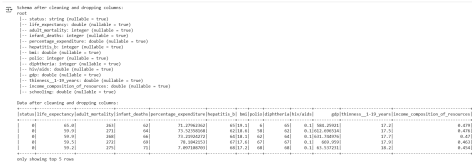


Fig. 11. Dataset after dropping columns

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 duplicate data was identified by grouping the data by columns and counting the number 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 median imputation to ensure the accuracy of the data. The median imputation was chosen because, as per the summary statistic, it was found that the data are highly deviated, for some independent variables such as GDP and population, the standard deviation was much higher than mean, suggesting that the data has high variation (outliers) and might be skewed. As median is less sensitive to outliers, median imputation was done.

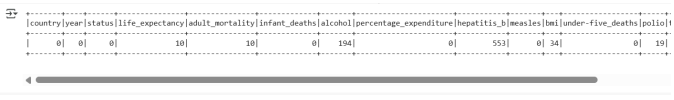


Fig. 12. Before Handling Null Values

- Categorical data like 'status' was encoded.

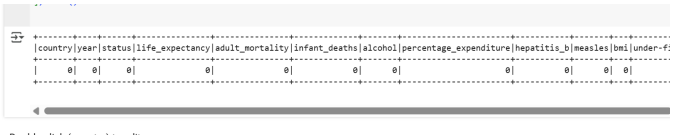


Fig. 13. After Handling Null Values

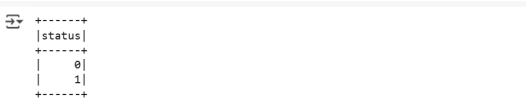


Fig. 14. Handling Categorical Data

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

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. Vector Assembler collects the values from a certain input column, packs them into a vector, and stores them in a new output column for each row in the dataframe. The original values remain the same, as it only organizes the values into a suitable format for ML models. The data preparation process is sped up with its use, which makes the data ready for machine learning tasks easier. [15]
- **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.

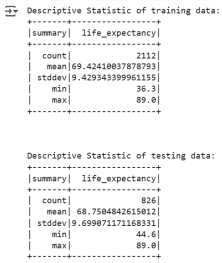


Fig. 16. Training and Testing dataset

- **Models Used:**
  - Multiple Linear Regression (MLR)
  - Random Forest Regression

Discussion of the finding: Till now, we have explored the data, understood its pattern, cleaned the dataset by confirming

if the null values and duplicated data were present, and if yes, then the null values were removed by implementing median imputation. For further analysis, evaluation of both models' performance was carried out.

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 better performance on predicting life expectancy than the Multiple Linear Regression (MLR). The significant lower values of MSE and MAE than those of MLR show 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 that 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: 3.1590
MSE: 18.5168
RMSE: 4.3031
R2: 0.8029 (80.29%)
Adjusted R2: 0.8000 (80.00%)

----- RF (Random Forest) Results -----
MAE: 2.0004
MSE: 7.7083
RMSE: 2.7764
R2: 0.9180 (91.80%)
Adjusted R2: 0.9167 (91.67%)
```

Fig. 17. Evaluation (MLR Vs RF)

Analysis of the findings: The research shows that the distribution of life expectancy is left-skewed, as the frequency of high life expectancy is rising towards the higher ages. No linear relationships were found between life expectancy and variables like BMI, alcohol. Due to healthcare access and living conditions, developed countries show more consistent and higher life expectancy as compare to the developing countries, some outliers was identified on it.The strongly positive relationship between schooling, income composition of resources, with life expectancy, and the negative relationship between adult mortality, HIV/AIDS, which are socioeconomic indicators, supports what public health experts already know, that a person's health is strongly affected by both social and economic conditions.From the observation, we can say that among the multiple linear regression and random forest regression models, which were used in the life expectancy prediction, random forest regression is superior as it performs better across all the evaluation metrics.and alcohol variables indicate little to no correlation with life expectancy. Some limitations, like issues with the data quality, excluding a few features that didn't seem important, should be kept in mind when understanding the results of this study. To improve the prediction accuracy for further studies, some additional variables can be explored or the use of different modeling techniques.

V. Conclusion

The goal of this study is to build, visualize, and evaluate the models for predicting global Life Expectancy trends using machine learning techniques and big data analytics using life expectancy dataset from Kaggle. Multiple Linear Regression and 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 life expectancy, random forest outperformed 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 the community, which is influenced

by several socio-economic and 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] R. Bevans, "Multiple Linear Regression — A Quick Guide (Examples)," *Scribbr*, 20 February 2020. [Online]. Available: https://www.scribbr.com/statistics/multiple-linear-regression/. [Accessed 10 May 2025].

[12] 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].

[13] MITSDE, "Data Imputation Techniques: Handling Missing Data in Machine Learning," *MIT School of Distance Education Blog*, [Online]. Available: https://blog.mitsde.com/data-imputation-techniques-handling-missing-data-in-machine-learning/. [Accessed 10 May 2025].

[14] Sumbatilinda, "RANDOM FORESTS REGRESSION BY EXAMPLE," *Medium*, 26 March, 2024. [Online]. Available: https://medium.com/@sumbatilinda/random-forests-regression-by-example-1baa062506f5. [Accessed 11 May 2025].

[15] SparkCodeHub, "Feature Engineering: VectorAssembler in PySpark," *SparkCodeHub*. [Online]. Available: https://www.sparkcodehub.com/pyspark/mllib/vector-assembler. [Accessed: May 12, 2025].