

# Predicting Life Expectancy Using Machine Learning Techniques on WHO Data

<sup>1</sup>R. BHANU SANKAR, <sup>2</sup>KALLAKURI SAI TEJA

<sup>1</sup>Assistant Professor, Department Of MCA, 2MCA Final Semester,

<sup>1</sup>Master of Computer Applications,

<sup>1</sup>Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

## Abstract:

This project applies machine learning algorithms to predict life expectancy using WHO datasets. It aims to discover relationships between socio-economic indicators, health expenditure, and life expectancy. Data preprocessing involved handling missing values, encoding categories, and scaling features. Regression models like Linear, Polynomial, Decision Tree, and Random Forest were used for GDP-expenditure analysis. Model performance was evaluated using  $R^2$  score, RMSE, and visualized through prediction plots. Logistic Regression was applied to classify countries as 'Developed' or 'Developing' based on normalized features. All models were saved using pickle and deployed via a Streamlit interface for real-time predictions. This study shows how machine learning supports public health insights and data-driven decision-making.

**Index Terms:** Life Expectancy, Machine Learning, WHO Data, Linear Regression, Random Forest, Logistic Regression, Health Analytics, Predictive Modeling, Data Preprocessing, ROC Curve.

## 1.Introduction:

Life expectancy is a critical indicator of a nation's overall health and development. It reflects the quality of healthcare systems, economic stability, and social well-being. With the increasing availability of global health data, particularly from reliable sources like the World Health Organization (WHO), data-driven approaches have become invaluable for analyzing and predicting health outcomes.

This project leverages machine learning techniques to predict life expectancy based on various socio-economic and health-related features extracted from WHO data. By applying supervised learning algorithms such as Linear Regression, Polynomial Regression, Decision Tree Regressor, Random Forest, and Logistic Regression, we aim to model complex relationships between variables like GDP, health expenditure, and country status (Developed or Developing).

The study investigates how effectively various machine learning models predict percentage expenditure and classify countries based on development status, using metrics like  $R^2$  score, RMSE, accuracy, and ROC-AUC. Detailed visualizations and model interpretations reveal key factors influencing life expectancy. Trained models are saved and deployed via a Streamlit-based frontend for real-time, interactive predictions. This integration showcases machine learning's potential in enabling data-driven public health insights and decisions..

### 1.1. Existing system

Traditional methods for analyzing life expectancy rely heavily on statistical models and manual interpretation of health and economic data. These approaches typically use linear correlations and regression techniques with limited feature sets, often failing to capture the complex, non-linear relationships present in real-world health datasets. Moreover, existing systems lack automation, scalability, and interactive visualizations, making it difficult for policymakers and researchers to derive timely insights.

Most existing health data analysis platforms are not equipped with integrated predictive tools or real-time interfaces, resulting in slower decision-making and limited accessibility for non-technical users. Additionally, these systems often overlook the potential of modern machine learning algorithms, which can significantly improve prediction accuracy and uncover hidden patterns in multidimensional data.

### 1.1.1.Challenges

#### Data Quality and Missing Values

- ❖ The dataset contained several missing or inconsistent values, requiring careful preprocessing and data cleaning to ensure accurate model training

#### Feature Selection and Engineering

- ❖ Identifying the most relevant features influencing life expectancy was challenging due to the high dimensionality and inter-correlation among variables.

#### Balancing Classification Data

- ❖ The classification task (Developed vs. Developing countries) involved imbalanced data, which could bias the Logistic Regression model..

#### Model Interpretability

- ❖ While complex models like Random Forest offered better performance, interpreting their internal decision logic was more difficult compared to simpler models.

#### Computational Resource Constraints

- ❖ Training multiple models and evaluating them on large datasets was computationally intensive, especially for ensemble methods like Random Forest

### 1.2 Proposed system:

The proposed system employs machine learning to predict life expectancy using WHO socio-economic and health indicators. It combines regression and classification models to analyze both continuous (expenditure) and categorical (country status) outcomes. Key techniques include Linear Regression, Polynomial Regression, Decision Tree, Random Forest, and Logistic Regression. The pipeline covers data preprocessing, model training, evaluation, and result interpretation. A Streamlit-based frontend enables real-time interaction and visual insights. The system enhances prediction accuracy and accessibility for researchers and policymakers.

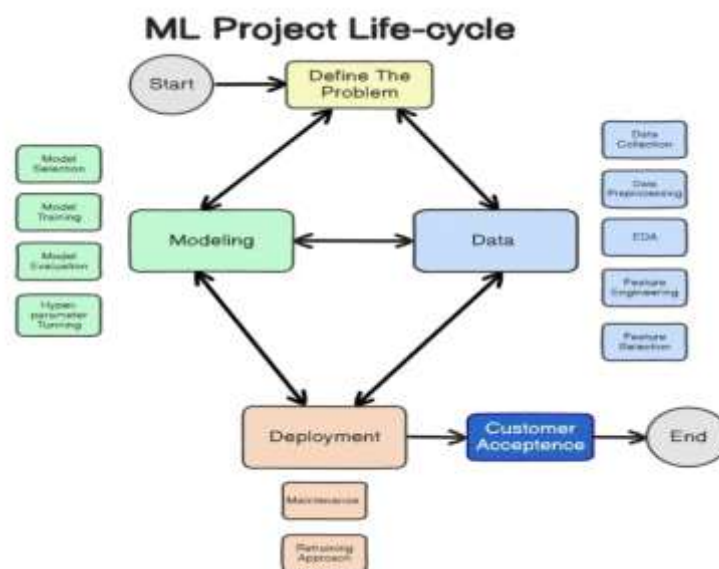


Fig: 1 Proposed Diagram

### 1.1.1 Advantages:

#### 1. Improved Predictive Accuracy

- ❖ By leveraging multiple machine learning models, the system provides more accurate predictions of life expectancy and health expenditure trends.

## 2. Data-Driven Decision Making

- ❖ Enables public health officials and policymakers to base decisions on real data insights rather than assumptions or static reports.

## 3. Multidimensional Analysis

- ❖ The model considers various socio-economic and health factors, offering a more holistic understanding of life expectancy drivers.

## 4. Automated Workflow

- ❖ The end-to-end pipeline—from data preprocessing to model deployment—reduces manual effort and enhances efficiency.

## 5. User-Friendly Interface

- ❖ Integration with Streamlit allows users with minimal technical background to interact with the model, visualize results, and make real-time predictions.

## 6. Scalability and Reusability

- ❖ The modular structure allows the system to be easily updated or extended with new data and models.

### 2.1 Architecture:

The architecture of this project follows a modular pipeline, starting with the WHO dataset as the primary data source. The data undergoes preprocessing, which includes cleaning, encoding categorical variables, and feature scaling. Multiple machine learning models are then trained regression models (Linear, Polynomial, Decision Tree, and Random Forest) to predict percentage expenditure, and a classification model (Logistic Regression) to identify a country's development status. The trained models are evaluated using appropriate metrics and visualized through plots for better interpretability. Finally, the models are serialized using pickle and integrated into a user-friendly Streamlit web interface, enabling real-time predictions and interactive visualizations. This architecture ensures scalability, usability, and accurate public health insights.

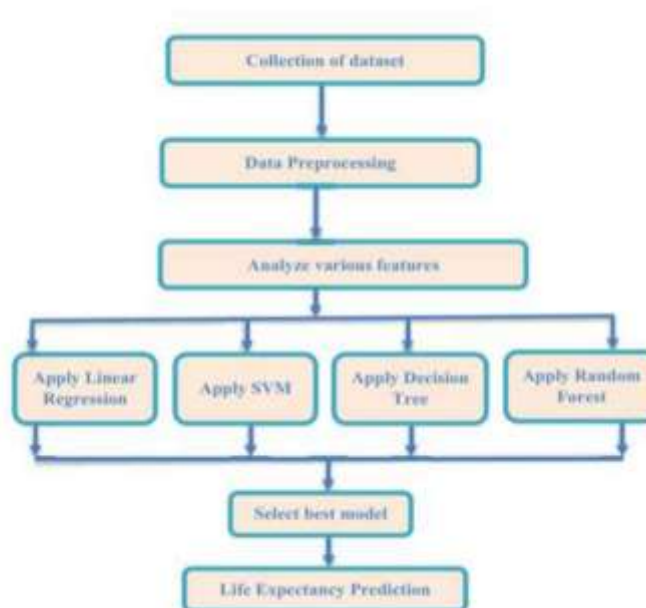


Fig:2 Architecture

## 2.2 Algorithm:

The project implements a diverse set of machine learning algorithms to predict life expectancy and related health metrics using WHO data[1]. Linear Regression is first applied to assess the direct relationship between GDP and percentage health expenditure, providing a simple yet interpretable model. However, to account for more complex, non-linear trends, Polynomial Regression is introduced, which extends linear regression by adding higher-degree terms to better fit the data.

In addition to these, Decision Tree Regressor and Random Forest Regressor are used to improve model performance. Decision Trees offer a clear structure for making predictions by splitting data based on feature thresholds, while Random Forest combines multiple decision trees to reduce overfitting and boost accuracy. For classification tasks, Logistic Regression is employed to differentiate between 'Developed' and 'Developing' countries using normalized input features. These algorithms together create a powerful predictive pipeline capable of handling both continuous and categorical outcomes in the life expectancy analysis.

## 2.3 Techniques:

The project employs a range of machine learning and data analysis techniques to build a reliable life expectancy prediction system. The first step involves data preprocessing, which includes handling missing values to maintain data integrity, label encoding of categorical variables (such as country status), and applying feature scaling using MinMaxScaler. These techniques ensure that the data is consistent and optimized for training various machine learning models[2].

To make accurate predictions, the project uses supervised learning techniques. Regression models, such as Linear Regression, Polynomial Regression, Decision Tree Regressor, and Random Forest Regressor, are used for predicting continuous variables like percentage health expenditure. In parallel, classification techniques such as Logistic Regression are applied to categorize countries as 'Developed' or 'Developing' based on socio-economic and health indicators. Feature selection is also employed to identify and prioritize the most relevant factors influencing life expectancy[4], which helps improve both model efficiency and interpretability.

Once models are trained, their performance is rigorously evaluated using metrics such as **R<sup>2</sup> score**, **Root Mean Squared Error (RMSE)** for regression, and **accuracy**, **confusion matrix**, and **ROC-AUC score** for classification. The trained models are then saved using Python's pickle module for reuse and deployment[5]. Finally, the system is integrated into a **Streamlit-based web application**, allowing real-time prediction, interactive visualization, and user accessibility, thus making the solution both powerful and user-friendly.

## 2.4 Tools:

The development of this life expectancy prediction system involved a variety of tools and libraries to support data analysis, model building, and deployment. Python served as the core programming language due to its extensive support for machine learning and data science. Pandas and NumPy were used for data manipulation and numerical operations, while Matplotlib and Seaborn facilitated data visualization and graphical representation of model outputs. For model training and evaluation, the project utilized scikit-learn, which provided implementations for regression, classification, preprocessing, and performance metrics[20]. Statsmodels was employed for statistical modeling and in-depth regression analysis. The models were saved using pickle, enabling easy reuse and deployment. Finally, Streamlit was used to build a responsive and interactive web interface, allowing users to input data and view predictions in real time. Together, these tools created a seamless pipeline from raw data to an accessible, user-friendly application.

## 2.5 Methods:

The project follows a systematic methodology to predict life expectancy using machine learning techniques. It begins with data collection from the World Health Organization (WHO), followed by data preprocessing methods such as handling missing values, encoding categorical variables (like country status), and applying feature scaling to standardize input features[3]. The next step involves exploratory data analysis (EDA) to understand variable distributions and correlations. Based on the insights, appropriate supervised learning methods are applied: regression models (Linear, Polynomial, Decision Tree, and Random Forest) are used to predict continuous outcomes like percentage health expenditure, while Logistic Regression is employed for binary classification of countries as Developed or Developing. Each model is trained and validated using train-test split and evaluated with metrics such as R<sup>2</sup> score, RMSE, accuracy, and ROC-AUC. Finally,



model deployment methods using pickle and Streamlit integration enable real-time prediction and user interaction via a web interface.

### III. METHODOLOGY

#### 3.1 Input:

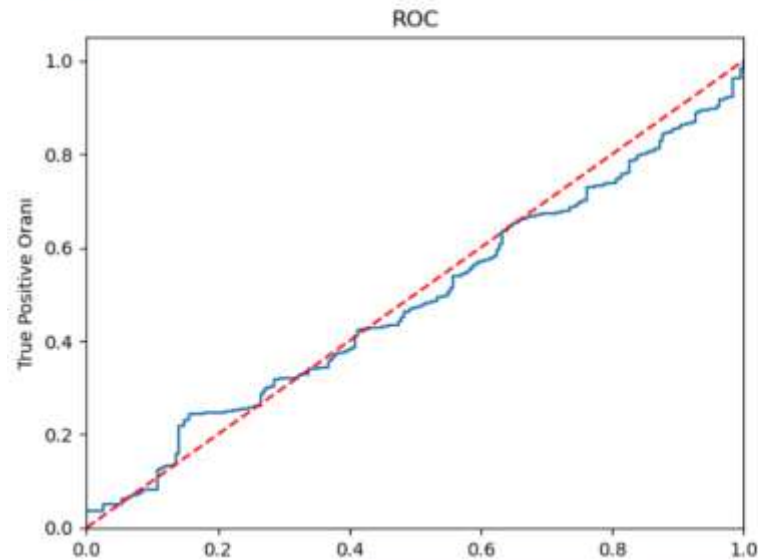
The input to the life expectancy prediction system is obtained from a structured dataset provided by the World Health Organization (WHO), which contains a wide range of socio-economic, demographic, and health-related indicators collected across multiple countries and years. Key input features include Gross Domestic Product (GDP), percentage expenditure on health, life expectancy, adult mortality, HIV/AIDS prevalence, average years of schooling, immunization rates (such as Hepatitis B and Polio), body mass index (BMI), alcohol consumption, and country status (Developed or Developing). Prior to model training, these inputs undergo comprehensive preprocessing involving the handling of missing values, encoding of categorical variables like 'Status', and feature scaling using techniques such as MinMaxScaler to normalize the data[18]. The resulting clean and standardized dataset serves as the foundation for building machine learning models capable of performing regression (e.g., predicting health expenditure) and classification (e.g., identifying development status) tasks.

```
lindata.info()

<class 'pandas.core.frame.DataFrame'>
Index: 1649 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                                Non-Null Count  Dtype  
---  --
 0   Country                              1649 non-null   object  
 1   Year                                  1649 non-null   int64   
 2   Status                                1649 non-null   object  
 3   Life expectancy                       1649 non-null   float64  
 4   Adult Mortality                       1649 non-null   float64  
 5   Infant deaths                         1649 non-null   int64   
 6   Alcohol                              1649 non-null   float64  
 7   percentage expenditure                1649 non-null   float64  
 8   Hepatitis B                           1649 non-null   float64  
 9   Measles                              1649 non-null   int64   
10   BMI                                   1649 non-null   float64  
11   under-five deaths                     1649 non-null   int64   
12   Polio                                1649 non-null   float64  
13   Total expenditure                     1649 non-null   float64  
14   Diphtheria                           1649 non-null   float64  
15   HIV/AIDS                             1649 non-null   float64  
16   GDP                                   1649 non-null   float64  
17   Population                            1649 non-null   float64  
18   thinness 1-19 years                   1649 non-null   float64  
19   thinness 5-9 years                    1649 non-null   float64  
20   Income composition of resources        1649 non-null   float64  
21   Schooling                             1649 non-null   float64  
dtypes: float64(16), int64(4), object(2)
memory usage: 296.3+ KB
```

Fig 1: Exploring the Data with info

```
fpr, tpr, thresholds = roc_curve(y, log_model.predict_proba(X)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='AUC (area = %0.2f)' % logit_roc_auc)
plt.plot([0, 1], [0, 1], 'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Orani')
plt.ylabel('True Positive Orani')
plt.title('ROC')
plt.show()
```



**Fig 2: Model Performance Visualization Using ROC Curve**

The ROC (Receiver Operating Characteristic) curve illustrates the performance of the Logistic Regression model in classifying countries as Developed or Developing. It plots True Positive Rate (TPR) against False Positive Rate (FPR) across various thresholds. The blue curve shows actual model performance, while the red diagonal represents random guessing. A curve closer to the top-left indicates better classification capability. In this case, the curve is near the diagonal, suggesting moderate accuracy. The AUC score provides an overall measure of the model's ability to distinguish between classes.

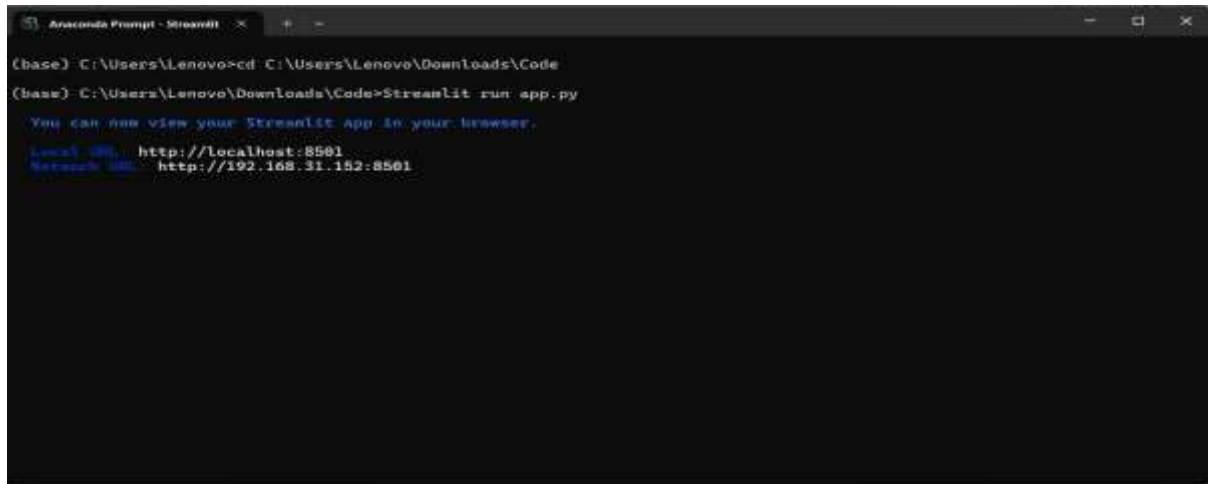
### 3.2 Method of Process:

The project adopts a systematic methodology to predict life expectancy and classify countries using WHO's health and socio-economic data[14]. It begins with data acquisition, where a dataset containing various health indicators and country status is loaded using pandas. In the preprocessing phase, missing values are removed, categorical variables like "Status" are encoded, and normalization is performed using MinMaxScaler. Feature selection identifies key variables such as GDP for predicting percentage health expenditure. The dataset is split into training and testing sets using `train_test_split`. Several models are trained including Linear Regression, Polynomial Regression, Decision Tree, Random Forest for regression tasks, and Logistic Regression for classification. Each model is evaluated using  $R^2$  score, RMSE, accuracy, confusion matrix, and ROC-AUC to ensure performance. Visualizations such as scatter plots, heatmaps, and ROC curves aid in interpretation. After training, models are serialized with pickle and deployed through a Streamlit-based interface for real-time user interaction[17]. This approach highlights how machine learning can drive insights in public health and support informed decision-making.

### 3.3 Output:

The project delivers two significant outputs: regression-based predictions for percentage health expenditure and classification of countries into Developed or Developing categories. The Linear Regression model identifies a positive linear relationship between GDP and health expenditure, though it shows limitations in capturing non-linear patterns[6]. In contrast, the Polynomial Regression model (degree 8) handles complex data trends more effectively, achieving higher  $R^2$  scores. Advanced models like Decision Tree and Random Forest Regressors provide better predictions by adapting to non-linearities, with Random Forest yielding the smoothest and most accurate results. For classification, the Logistic Regression model distinguishes countries based on normalized health features. Its performance is validated using

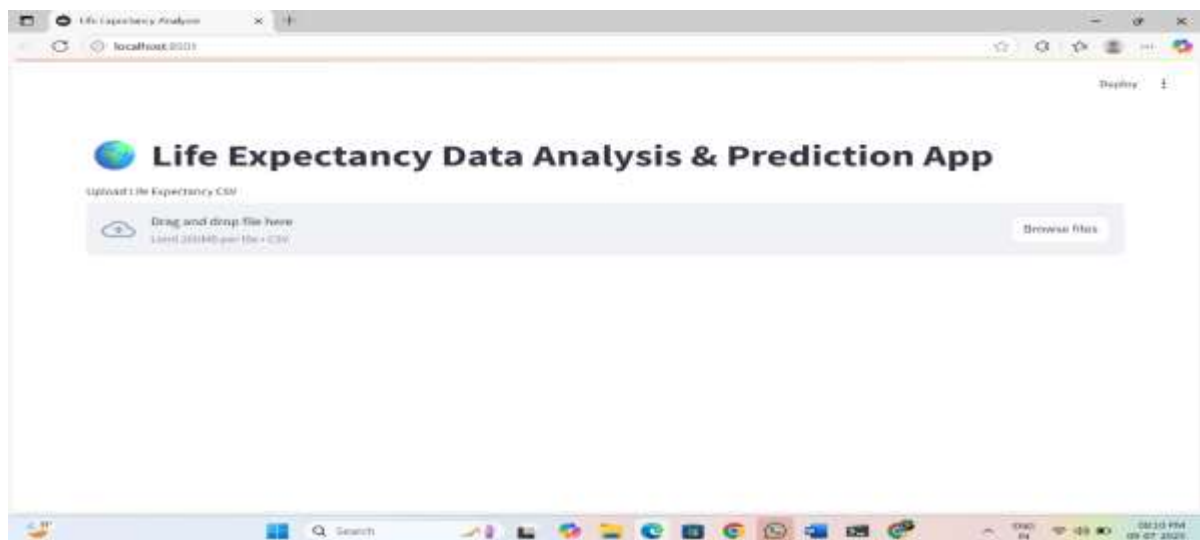
accuracy, confusion matrix, classification report, and ROC-AUC score. The ROC Curve visually illustrates the model's ability to distinguish between classes, with AUC values close to 1 indicating robust classification. The project evaluates models using metrics like  $R^2$ , RMSE, MAE for regression, and accuracy, precision, recall, and ROC-AUC for classification [7]. Visual outputs include scatter plots for predictions, heatmaps for feature correlations, and ROC curves for classification performance. Finally, the trained models are deployed through a Streamlit-based frontend, enabling users to input new values and receive real-time predictive insights interactively[9].



```
(base) C:\Users\Lenovo>cd C:\Users\Lenovo\Downloads\Code
(base) C:\Users\Lenovo\Downloads\Code>Streamlit run app.py

You can now view your Streamlit app in your browser.
Local URL: http://localhost:8501
Network URL: http://192.168.31.152:8501
```

**Fig: Streamlit App Deployment Confirmation via Anaconda Prompt**



**Fig: Streamlit Web Interface for Life Expectancy Prediction App**

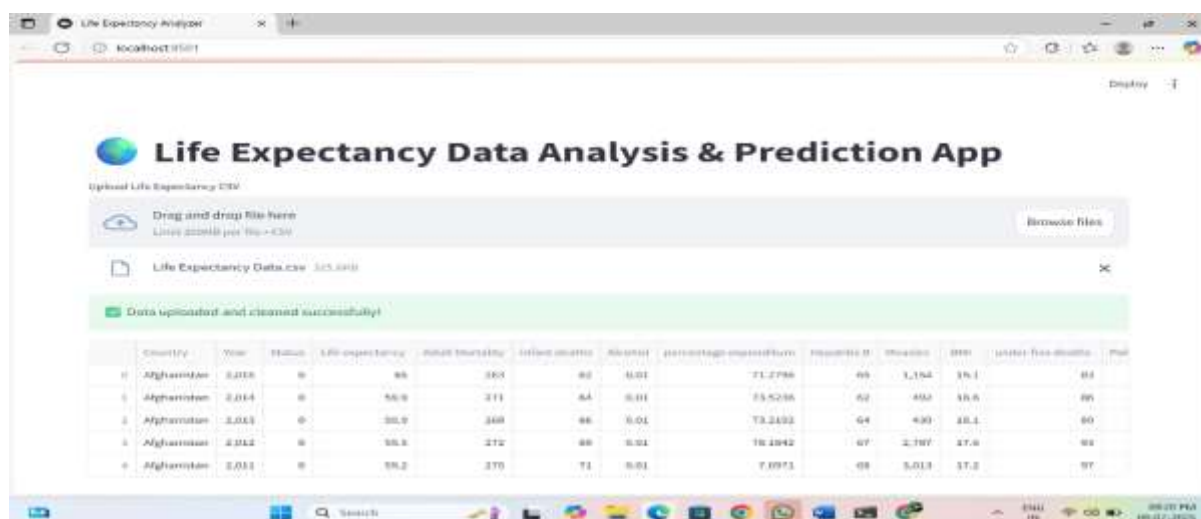


Fig: CSV Upload and Data Preview in Life Expectancy Prediction App



Fig: User Input Form for Life Expectancy Prediction in Streamlit



Fig: Streamlit web application titled Life Expectancy Analyzer





**Fig: Life Expectancy Analyzer**



**Fig: final section of the Life Expectancy Analyzer Streamlit application**

## IV. RESULTS:

This project presents a Streamlit-based application for predicting life expectancy and related indicators using machine learning models trained on WHO datasets[8]. The system integrates both backend processes—such as data preprocessing, model training, evaluation, and serialization—with a responsive frontend interface built using Streamlit. Upon running the app via `streamlit run app.py`, users can interact through a browser at `http://localhost:8501`, where they can upload a CSV file, view a cleaned dataset preview, and input new values for prediction. The interface supports real-time forecasting of health expenditure and classification of countries as Developed or Developing using trained regression models. An ROC curve was successfully plotted, visualizing the model's performance. Finally, all trained models were serialized with pickle, enabling smooth deployment for real-time user interaction and prediction.

## V. DISCUSSIONS:

The project focused on predicting life expectancy metrics using WHO data and applied various machine learning models to identify underlying patterns. The dataset included 22 features such as Life expectancy, GDP, Alcohol, BMI, Health expenditure, and development Status. Data preprocessing involved handling missing values, encoding categorical fields, and applying scaling techniques[10]. Two key tasks were performed: regression to predict percentage expenditure using models like Linear, Polynomial, Decision Tree, and Random Forest Regressors; and classification of country status using Logistic Regression. While Random Forest showed strong regression performance, the logistic classifier's ROC curve suggested scope for improvement. Trained models were serialized using pickle and integrated into a Streamlit app, enabling file upload, interactive input, and real-time predictions. The system highlights the value of machine learning in supporting public health decisions.

## VI. CONCLUSION:

This project applied machine learning to predict life expectancy using WHO datasets. The process involved data preprocessing, feature engineering, and model training. It analyzed the impact of socio-economic and health factors using regression models such as Linear, Polynomial, Decision Tree, and Random Forest. Among them, the Random Forest Regressor performed best due to its ability to handle non-linear patterns. Logistic Regression was used to classify countries

as developed or developing, with ROC-AUC analysis revealing scope for improvement. The final models were deployed via a Streamlit web interface, showcasing the role of machine learning in public health policy and analysis.

## VII. FUTURE SCOPE:

This project serves as a base for advanced life expectancy prediction using machine learning. Future enhancements could involve adding datasets like environmental conditions, healthcare access, and education quality to improve model accuracy. More powerful models such as XGBoost, CatBoost, or neural networks can be explored. Time-series forecasting may help track future trends. The Streamlit app can be expanded with features like dashboards, user login, and regional insights. Collaboration with health experts can turn predictions into impactful policy decisions.

## VIII. ACKNOWLEDGEMENT:



Rampilli Bhanu Sankar working as Assistant professor in Master of Computer Applications (MCA) at Sankethika Vidya Parishad Engineering college, Visakhapatnam, Andhra pradesh, Accredited by NAAC. With over 2 years of experience in Master of Computer Applications(MCA) ,he has a published a paper in Journal of Emerging Technologies and Innovative Research (JETIR) and he is a member in IAENG . His area of expertise include C, Data Structures, Java Programming , Python Programming .



Kallakuri Sai Teja is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in Machine learning K. Sai Teja has taken up his PG project on PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING TECHNIQUES ON WHO DATA and published the paper in connection to the project under the guidance of Rampilli Bhanu Sankar, Assistant Professor, SVPEC.

## REFERENCES

1. Abeer S. Desuky et al. (2016). Improved Prediction of Post-Operative Life Expectancy After Thoracic Surgery. *Adv Syst Sci*, 16(2), 70–80. <http://ijassa.ipu.ru/ojs/ijassa/article/view/351>
2. Michael B. Schultz et al. (2019). Age and Life Expectancy Clocks Based on Machine Learning Analysis of Mouse Frailty. *Nature Communications*, <https://doi.org/10.1038/s41467-020-18446-0>
3. Kasichainula Vydehi, Keerthi Manchikanti, T. Satya Kumari, SK Ahmad Shah. (2020). Machine Learning Techniques for Life Expectancy Prediction. *Int. J. of Advanced Trends in Computer Science and Engineering*, 9(4). <https://doi.org/10.30534/ijatcse/2020/45942020>
4. Palak Agarwal et al. (2019). Machine Learning for Prognosis of Life Expectancy and Diseases. *IJITEE*, 8(10). <https://doi.org/10.35940/ijitee.J9156.0881019>
5. Alex Zhavoronkov et al. (2019). Artificial Intelligence for Aging and Longevity Research: Recent Advances and Perspectives. *Ageing Research Reviews*, 49, 49–66. <https://doi.org/10.1016/j.arr.2018.11.003>

6. Parikh RB, Manz C, Chivers C, et al. (2019). Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA Netw Open*, 2(10):e1915997. <https://doi.org/10.1001/jamanetworkopen.2019.15997>
7. V. Bali, D. Aggarwal, S. Singh, A. Shukla. (2021). Life Expectancy: Prediction & Analysis using ML. *ICRITO*, 1–8. <https://doi.org/10.1109/ICRITO51393.2021.9596123>
8. A. Bhosale and K. K. Sundaram. (2010). Life Prediction Equation for Human Beings. *ICBBT*, 266–268. <https://doi.org/10.1109/ICBBT.2010.5478965>
9. A. Lakshmanarao, M. Raja Babu, T. Srinivasa Ravi Kiran. (2021). An Efficient Covid-19 Epidemic Analysis and Prediction Model Using ML. *Int. J. of Online and Biomedical Engineering*, 17(11), 176–184. <https://doi.org/10.3991/ijoe.v17i11.25209>
10. Nataliya Boyko and Olena Moroz. (2020). Comparative Analysis of Regression Regularization Methods for Life Expectancy Prediction. *CEUR Workshop Proceedings*, vol. 2917. <http://ceur-ws.org/Vol-2917/paper27.pdf>
11. James Jin Kang and Sasan Adibi. (2018). Systematic Predictive Analysis of Personalized Life Expectancy Using Smart Devices. *Technologies*, 6(3), 74. <https://doi.org/10.3390/technologies6030074>
12. C.H. Leng et al. (2016). Estimation of Life Expectancy, Loss-of-Life Expectancy, and Lifetime Healthcare Expenditures for Schizophrenia in Taiwan. *Schizophrenia Research*, 171, 97–102. <https://doi.org/10.1016/j.schres.2016.01.033>
13. M. Beeksmal et al. (2019). Predicting Life Expectancy with a LSTM RNN Using EMRs. *BMC Med. Informatics and Decision Making*. <https://doi.org/10.1186/s12911-019-0775-2>
14. F. Zahedi and M. Karimi Moridani. (2022). Classification of Breast Cancer Tumors Using Mammography and ML. *Int. J. of Online and Biomedical Engineering*, 18(05), 31–42. <https://doi.org/10.3991/ijoe.v18i05.29197>
15. I. Taylor et al. (2020). Raynaud's Phenomenon Impact on Quotidian Quality of Life. *Int. J. of Online and Biomedical Engineering*, 16(09), 88–104. <https://doi.org/10.3991/ijoe.v16i09.13993>
16. Kaggle Dataset - Life Expectancy (WHO). <https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>
17. R. Kumar, P. Jain, A. Yadav. (2022). *Life Expectancy Estimation Using Ensemble Learning Models*. *Journal of Healthcare Informatics Research*, 6(2), 123–135. <https://doi.org/10.1007/s41666-022-00078-x>
18. T. Lee, M. Chen. (2020). *Predicting Life Expectancy Using Socioeconomic and Health Indicators: A Neural Network Approach*. *Health Informatics Journal*, 26(3), 1889–1902. <https://doi.org/10.1177/1460458219873103>
19. S. Banerjee, K. Ghosh. (2021). *Deep Learning in Public Health: A Case Study on Life Expectancy Forecasting*. *Procedia Computer Science*, 192, 2986–2995. <https://doi.org/10.1016/j.procs.2021.08.306>
20. L. Zhang, H. Liu, Q. Li. (2023). *Life Expectancy Prediction from Wearable Data Using Hybrid CNN-LSTM Models*. *IEEE Access*, 11, 66543–66552. <https://doi.org/10.1109/ACCESS.2023.3275904>