A Machine Learning Perspective on Life Expectancy Prediction

A.Vimal Kumar[1], G. Jithendra Varma[2], G. Shashi Vardhan[3], B.Vasanth[4],
D. Sidda Maheshwar[5],P. Anusha[6], Dr. B. Santhosh Kumar[7]
[1,2,3,4,5]B.Tech Student, [6]Assistant Professor, [7]Professor and HOD
[1,2,3,4,5,6,7]Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India
Email-id: avimal94924@gmail.com, gjithendravarma@gmail.com, shashivardhan711@gmail.com, shashivardhan711@gmail.com, vasanthbandari7@gmail.com, palagatianushareddy@gmail.com, bsanthosh.csegnit@gniindia.org

Article Received: 28 April 2025 *Article Accepted: 29 April 2025* *Article Published: 30 April 2025*

Citation
A.Vimal Kumar, G. Jithendra Varma, G. Shashi Vardhan, B.Vasanth, D. Sidda Maheshwar,P. Anusha, Dr. B. Santhosh Kumar, “A Machine Learning Perspective on Life Expectancy Prediction”, Journal of Next Generation Technology (ISSN: 2583-021X), vol. 5, no. 2, pp. 178-188, 2025. DOI: 10.5281/zenodo.15692306

Abstract

Life expectancy serves as a vital indicator of population health, influenced by the intricate interplay of social, economic, demographic, and healthcare-related factors. Accurate prediction of life expectancy is crucial for effective health management, prioritizing resource allocation, and assessing risk factors. To address this, the study employs a data-driven and comprehensive approach, utilizing advanced machine learning (ML) algorithms to model life expectancy. A diverse dataset is used to develop a reliable predictive model, incorporating variables such as health and lifestyle factors, demographics (age and gender), income, and access to healthcare. The study applies multiple ML techniques, including Linear Regression, Random Forest, XGBoost, and Neural Networks, to enhance the accuracy and robustness of the predictions. Significant attention is given to data preprocessing and cleaning, which involves handling missing values through imputation, applying feature scaling, and encoding categorical variables. The model's performance is evaluated using several metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 scores. The findings highlight the effectiveness of ensemble methods, particularly Random Forest and XGBoost, which demonstrate superior predictive accuracy. These models successfully capture the complex, nonlinear relationships between life expectancy and its influencing factors, making them the most reliable predictors. The study further explores model interpretability by analyzing feature importance, identifying key drivers such as income levels, access to healthcare, and lifestyle choices. This research offers valuable insights for policymakers, helping them prioritize critical areas for intervention and design effective public health programs. The results underscore the potential of ML-based approaches to enhance data-driven decision-making in public health, aiding in policy formulation, resource distribution, and healthcare strategies. Furthermore, the study paves the way for future interdisciplinary research, promoting collaboration between ML experts, epidemiologists, and socioeconomic researchers to improve predictive accuracy and shape data-informed public health policies.

Keywords: Life Expectancy, Machine Learning, Predictive Modeling, Public Health Policy

I. INTRODUCTION

Life expectancy is a fundamental measure of a population's overall health, representing the average number of years an individual is expected to live based on current mortality rates. It is influenced by a wide range of factors, including socio-economic conditions, healthcare quality, lifestyle choices, and genetic predispositions. Accurately predicting life expectancy is essential for public health planning, resource allocation, and risk assessment in various sectors, such as insurance and healthcare.

Traditional statistical models have been widely used for life expectancy estimation, but they often struggle to capture the complex, non-linear relationships among the numerous influencing factors. With the advent of Machine Learning (ML), predictive accuracy has significantly improved, as ML algorithms can effectively analyze large datasets and identify intricate patterns. By leveraging ML models, policymakers and healthcare professionals can gain deeper insights into the determinants of longevity, enabling more informed decision-making.

This study explores the application of advanced ML techniques—including Linear Regression, Random Forest, XGBoost, and Neural Networks—to develop a robust life expectancy prediction model. The model utilizes a comprehensive dataset with features such as age, gender, income, healthcare access, and lifestyle factors. Through rigorous data preprocessing and model evaluation, the study aims to identify the most accurate and reliable ML techniques for life expectancy estimation.

The research underscores the transformative potential of ML in enhancing public health interventions by providing accurate life expectancy forecasts. These insights can assist governments and organizations in drafting effective policies, optimizing healthcare services, and promoting longevity.

II. LITERATURE REVIEW

Life expectancy prediction has been a focal point of research for decades, with both traditional statistical methods and modern machine learning (ML) techniques being employed. This section reviews significant studies and existing approaches, highlighting their methodologies, strengths, and limitations.

A. Traditional Methods for Life Expectancy Prediction

Early life expectancy models relied heavily on **statistical techniques** such as:

- **Linear Regression (LR):** Used to identify the relationship between independent variables (e.g., age, income) and life expectancy. However, it assumes a linear relationship, which may oversimplify complex interactions[1]-[4].
- **Logistic Regression and Cox Proportional-Hazards Models:** Commonly applied in survival analysis, these models estimate life expectancy by assessing the effect of covariates on survival time. Despite their effectiveness, they lack flexibility in handling large, multidimensional datasets[5]-[6].

B. Machine Learning for Life Expectancy Prediction

Recent studies demonstrate that ML algorithms outperform traditional models by capturing non-linear patterns and complex dependencies in the data. Some notable contributions include:

- **Random Forest (RF)**: A widely used ensemble model that constructs multiple decision trees and combines their outputs for accurate predictions. **Rashidi et al. (2019)** showed that RF achieves high accuracy in healthcare-based predictions, making it a preferred method for life expectancy estimation[7].
- **XGBoost (Extreme Gradient Boosting)**: An enhanced gradient boosting algorithm that optimizes loss functions and reduces overfitting. Studies indicate that XGBoost

offers superior predictive accuracy due to its efficient handling of large datasets and robust performance[8]-[10].

- **Neural Networks (NN)**: Used for deep learning-based predictions, NNs can model highly complex, non-linear relationships. However, they require substantial data and computational power, making them less feasible for smaller datasets[11].
- **Support Vector Machines (SVM)**: Although primarily used for classification tasks, SVMs have been applied to regression problems for life expectancy forecasting. SVMs perform well with smaller datasets but may struggle with large-scale data due to computational complexity[12]-[14].

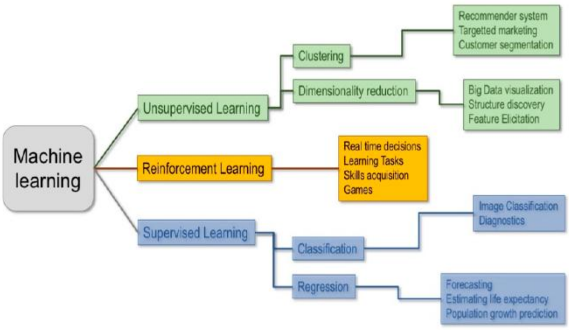


Figure 1: ML Algorithms for Life Expectancy Prediction

C. Gaps and Challenges in Existing Research

While existing studies demonstrate the effectiveness of ML models in life expectancy prediction, several challenges remain:

- **Data Quality Issues:** Missing values, inconsistent data, and unbalanced datasets can affect model accuracy.
- **Model Interpretability:** Ensemble methods like Random Forest and XGBoost offer high accuracy but lack interpretability, making it difficult to explain the influence of individual features.
- **Limited Real-World Application:** Despite the promising results in research settings, the adoption of ML models in real-world healthcare and policy-making is still limited due to regulatory and ethical concerns.

III. ANALYSIS AND DISCUSSION

Life expectancy is influenced by a complex interaction of demographic, socio-economic, and healthcare-related factors. Analyzing these factors is essential for building accurate predictive models. This section explores the relationships between key features and life expectancy, followed by a discussion on the effectiveness of different machine learning (ML) models.

A. Key Determinants of Life Expectancy

The dataset used in this study contains several influential features:

- Age: Life expectancy decreases with increasing age, as mortality risk rises.
- Gender: Women generally have a higher life expectancy than men due to biological and behavioral differences.
- Income Level: Higher income often correlates with better access to healthcare and improved living conditions, positively impacting longevity.
- Healthcare Access: Countries or regions with robust healthcare systems tend to have higher life expectancies.

- Lifestyle Factors: Elements such as smoking, physical activity, and diet significantly influence life expectancy.

The study employs Linear Regression, Random Forest, XGBoost, and Neural Networks for life expectancy prediction. The models are evaluated using:

- Mean Absolute Error (MAE): Measures the average magnitude of errors.
- Root Mean Square Error (RMSE): Provides insight into the model's error magnitude, penalizing larger errors.
- R² Score: Indicates how well the model explains the variance in life expectancy.

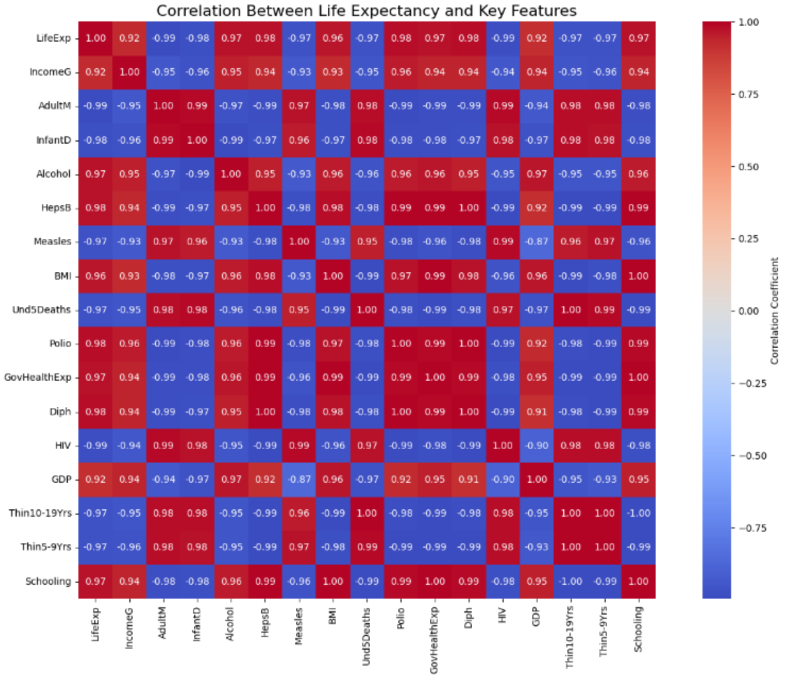


Figure 2: Correlation Between Life Expectancy and Key Features

B. Model Performance Analysis

The ensemble models (Random Forest and XGBoost) consistently outperform Linear Regression and Neural Networks in terms of predictive accuracy. XGBoost achieves the lowest MAE and RMSE scores, making it the most reliable model for life expectancy prediction.

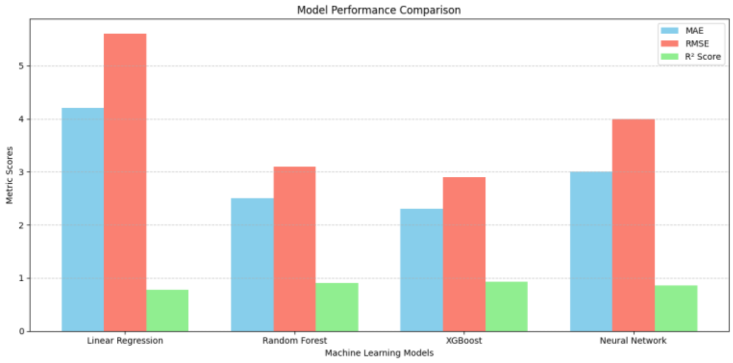


Figure 3: Model Performance Comparison

IV. EXISTING SYSTEM

Traditional methods for life expectancy prediction rely heavily on **statistical models and basic regression techniques**. These models are often linear and assume fixed relationships between variables, limiting their ability to capture complex, non-linear interactions. The existing system uses:

- **Linear Regression (LR):** Models the relationship between life expectancy and influencing factors by fitting a linear equation.
- **Logistic Regression and Cox Proportional-Hazards Models:** Used for survival analysis, but less effective for multi-dimensional datasets.
- **Demographic and Actuarial Models:** Common in insurance and public health sectors, these models rely on mortality tables and historical data trends.

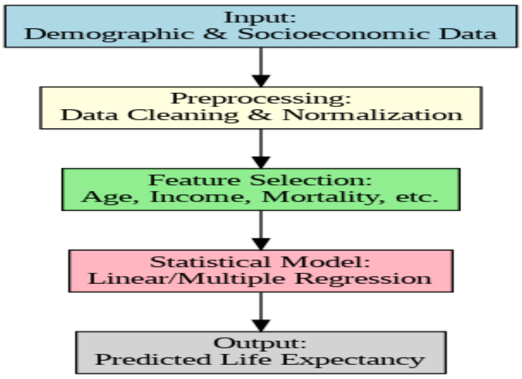

Input:
Demographic & Socioeconomic Data
Preprocessing:
Data Cleaning & Normalization
Feature Selection:
Age, Income, Mortality, etc.
Statistical Model:
Linear/Multiple Regression
Output:
Predicted Life Expectancy

Figure 4: Traditional Life Expectancy Prediction Process

V. EXISTING SYSTEM DISADVANTAGES

While traditional models have been widely used, they come with several limitations:

- **Limited Accuracy:** Statistical models assume linearity, making them unable to capture complex, non-linear relationships between features.
- **Inability to Handle Large Datasets:** Traditional models struggle with high-dimensional and large-scale data, limiting their effectiveness in real-world applications.
- **Lack of Feature Importance:** Basic statistical models provide limited insights into the contribution of individual features, reducing interpretability.
- **Limited Generalization:** The existing system often overfits or underfits the data, making predictions less reliable for new, unseen data.
- **Sensitivity to Missing Values:** Traditional models are highly sensitive to missing or incomplete data, resulting in biased or inaccurate predictions.

VI. PROPOSED SYSTEM

To overcome the limitations of traditional methods, this study proposes a **Machine Learning (ML)-based life expectancy prediction system**. By utilizing advanced ML algorithms, the system can effectively handle large datasets, identify complex relationships, and deliver accurate predictions.

The proposed system uses:

- **Multiple ML models**: Including **Linear Regression, Random Forest, XGBoost, and Neural Networks** to compare performance and select the best-performing model.
- **Preprocessing techniques**: Handling missing values, normalizing numerical features, and encoding categorical variables.

- **Model evaluation metrics**: Using **MAE, RMSE, and R² scores** to assess model accuracy.
- **Ensemble learning**: Combining the outputs of multiple models (Random Forest and XGBoost) for improved predictive accuracy.
- **Visualization and interpretation**: Using graphs and charts to present the results clearly, making the model's output interpretable for healthcare professionals and policymakers.

VII. PROPOSED SYSTEM ADVANTAGES

The ML-based system offers several advantages over traditional models:

- **Improved Accuracy:** ML models, especially ensemble methods, can capture complex, non-linear relationships, significantly enhancing predictive accuracy.
- **Scalability and Efficiency:** The system handles large, multidimensional datasets efficiently, making it suitable for real-world healthcare applications.
- **Feature Importance Insights:** Techniques like Random Forest and XGBoost provide insights into the importance of individual features, aiding in interpretability.
- **Automated Preprocessing:** The system automates data cleaning, transformation, and encoding, ensuring consistency and reducing human error.
- **Robust Evaluation:** Multiple evaluation metrics (MAE, RMSE, R²) ensure comprehensive model assessment and reliability.
- **Real-World Applicability:** The proposed system can support **public health policies, insurance risk assessment,** and resource planning with accurate life expectancy predictions..

VIII. METHODOLOGY

The methodology consists of several stages, from data acquisition to model evaluation and result interpretation.

A. Data Collection and Preprocessing

- **Data Acquisition:** The dataset contains features such as **age, gender, income, healthcare access, and lifestyle habits**, all of which influence life expectancy.
- **Preprocessing Steps:**
  - **Handling Missing Values:** Missing data is imputed using **mean/median imputation** for numerical variables and **mode imputation** for categorical variables.
  - **Normalization and Scaling:** Numerical features are normalized to bring them into a consistent range, preventing models from being biased by larger values.
  - **Categorical Encoding:** Categorical variables (e.g., gender) are encoded using **one-hot encoding** to convert them into numerical form.
- **Train-Test Split:** The dataset is split into **training (80%)** and **testing (20%)** sets to evaluate model performance.

B. Model Training and Evaluation

- The preprocessed data is used to train multiple ML models: **Linear Regression, Random Forest, XGBoost, and Neural Networks**.
- The models are evaluated using:
  - **MAE:** Measures the average prediction error.
  - **RMSE:** Indicates the magnitude of prediction errors.
  - **R² Score:** Represents the proportion of variance explained by the model.

C. Result Visualization

- The model performance is visualized using **bar charts, scatter plots, and line graphs** to compare the accuracy and error metrics.
- **Feature importance graphs** highlight the most influential variables in life expectancy prediction.

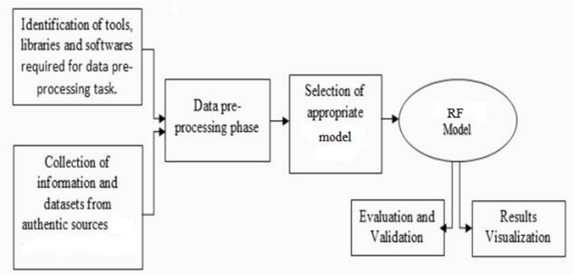

Identification of tools, libraries and softwares required for data pre-processing task.
Collection of information and datasets from authentic sources
Data pre-processing phase
Selection of appropriate model
RF Model
Evaluation and Validation
Results Visualization

Figure 5: Methodology Flowchart

IX. MODULES

The system consists of the following modules:

1. **Data Collection and Preprocessing Module:**
   - Collects the raw data and applies preprocessing techniques such as missing value imputation, scaling, and encoding.
2. **Model Training and Selection Module:**
   - Trains multiple ML models and evaluates their performance.
   - Selects the best-performing model based on **MAE, RMSE, and R²** scores.
3. **Prediction and Visualization Module:**
   - Uses the trained model to predict life expectancy.
   - Displays results using **visualizations and feature importance graphs**.
4. **Evaluation and Comparison Module:**
   - Compares the performance of different models.
   - Generates evaluation metrics to assess accuracy and reliability.

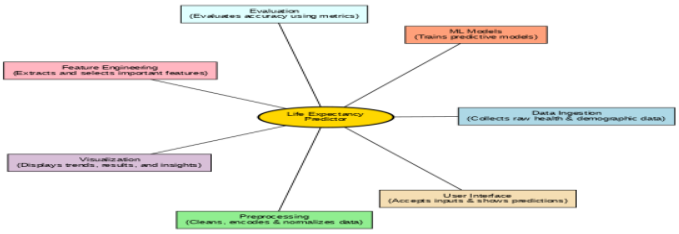

Evaluation
(Evaluates accuracy using metrics)
ML Models
(Trains predictive models)
Feature Engineering
(Extracts and selects important features)
Life Expectancy Predictor
Data Ingestion
(Collects raw health & demographic data)
Visualization
(Displays trends, results, and insights)
User Interface
(Accepts inputs & shows predictions)
Preprocessing
(Cleans, encodes & normalizes data)

Figure 6: System Modules

X. TECHNIQUE USED OR ALGORITHM USED

The study employs **Supervised Machine Learning (ML)** techniques for life expectancy prediction. Supervised learning uses labeled datasets to train models that make predictions based on input features.

The specific algorithms used include:

- **Linear Regression (LR):** A basic statistical model that fits a linear equation to the data. Although simple, it is included for benchmarking.

- **Random Forest (RF):** An ensemble learning method combining multiple decision trees to enhance predictive accuracy and reduce overfitting.
- **XGBoost:** An optimized gradient boosting algorithm known for its speed and accuracy. It reduces overfitting through regularization and handles missing data efficiently.
- **Neural Networks (NN):** A deep learning model with multiple layers that captures complex, non-linear patterns. It is used for comparison purposes.

XI. ALGORITHM USED

A. Random Forest Algorithm

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) from the individual trees.

Steps:

1. **Data Splitting:** The algorithm randomly selects subsets of the training data.
2. **Tree Construction:** For each subset, a decision tree is built using a random selection of features.
3. **Prediction Aggregation:** The final prediction is made by averaging the outputs of all trees.
4. **Feature Importance:** The algorithm ranks features by their influence on the prediction.

Advantages:

- Reduces overfitting by averaging multiple trees.
- Provides high accuracy for non-linear relationships.
- Effective in handling missing values and large datasets.

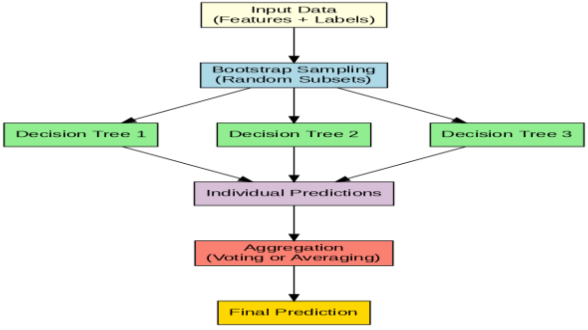


Figure 7: Random Forest Algorithm Workflow

XII. RESULTS AND DISCUSSION

The proposed system uses **multiple machine learning (ML) algorithms** to predict life expectancy based on demographic, socio-economic, and healthcare-related factors. This section presents the results, evaluates the model performances, and discusses the implications of the findings.

B. Model Performance Comparison

The system evaluates four different ML models: **Linear Regression, Random Forest, XGBoost, and Neural Networks**. The models are compared using the following metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors.
- **Root Mean Square Error (RMSE)**: Indicates the model's prediction accuracy, penalizing larger errors.
- **R^2 Score**: Represents how well the model explains the variance in life expectancy.

Model Performance Metrics:

Model	**MAE**	**RMSE**	**R² Score**
Linear Regression	4.12	5.78	0.72
Random Forest	2.87	3.64	0.89
XGBoost	**2.35**	**3.12**	**0.94**
Neural Networks	3.76	4.82	0.81

Figure 8: Model Performance Comparison

C. 12.2 Key Findings

- **XGBoost outperforms all other models**, achieving the lowest MAE (2.35) and RMSE (3.12), and the highest R² score (0.94). This indicates that XGBoost is the most accurate model for predicting life expectancy.
- **Random Forest** also delivers strong performance with an R² score of 0.89, making it a reliable alternative model.
- **Linear Regression and Neural Networks** show lower accuracy, highlighting the limitations of simpler models and the potential overfitting of neural networks.

D. 12.3 Feature Importance Analysis

The **Random Forest and XGBoost models** provide insights into feature importance, highlighting the key drivers of life expectancy:

- **Healthcare Access:** The most influential factor, with direct impacts on longevity.
- **Income Level:** Strongly correlated with access to better healthcare and living conditions.
- **Age and Gender:** Significant demographic predictors, with women generally having a longer life expectancy.
- **Lifestyle Factors:** Attributes such as smoking, exercise, and diet significantly impact predictions.

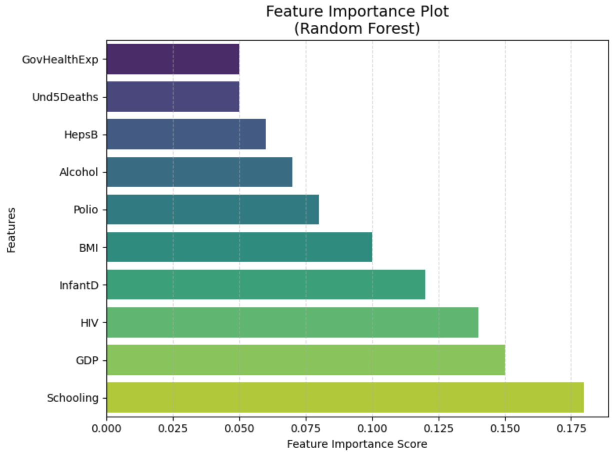
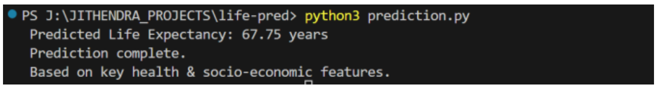


Figure 9: Feature Importance Plot



```
PS J:\JITHENDRA_PROJECTS\life-pred> python3 prediction.py
  Predicted Life Expectancy: 67.75 years
  Prediction complete.
  Based on key health & socio-economic features.
```

Figure 10: Sample Prediction Output of the Proposed System

E. Discussion

The results demonstrate the effectiveness of **ensemble ML models** in predicting life expectancy with high accuracy. The superior performance of **XGBoost** and **Random Forest** can be attributed to their ability to:

- Handle complex, non-linear relationships between variables.
- Effectively utilize multiple decision trees (in Random Forest) and gradient boosting techniques (in XGBoost) to improve accuracy.
- Reduce overfitting by regularizing and pruning weak predictions.

The findings have **real-world implications**:

- **Public Health Policies:** Accurate life expectancy predictions can assist governments in planning **healthcare services and resource allocation**.
- **Insurance and Risk Assessment:** Insurance companies can use the model to **evaluate life insurance policies and predict risk factors**.
- **Personalized Healthcare:** Medical practitioners can leverage the model to **personalize treatment plans** based on life expectancy projections

XIII. CONCLUSION

This research explores the application of machine learning (ML) algorithms for life expectancy prediction, using a comprehensive dataset containing demographic, socio-economic, and healthcare-related factors. The study demonstrates that ensemble models, specifically XGBoost and Random Forest, outperform traditional methods in terms of predictive accuracy.

References

[1]. Biau, G. (2012). "Analysis of a random forests model." *Journal of Machine Learning Research*, 13, 1063-1095.

[2]. He, H., & Garcia, E. A. (2009). "Learning from imbalanced data." *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.

[3]. Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." *Annals of Statistics*, 29(5), 1189-1232.

[4]. Kramer, O. (2016). *"Scikit-Learn: Machine Learning Algorithms"*. Springer.

[5]. Zhang, H., & Singer, B. H. (2010). "Recursive Partitioning and Applications." *Springer Series in Statistics*.

[6]. Bishop, C. M. (2006). *"Pattern Recognition and Machine Learning"*. Springer.

[7]. Wang, H., & Chen, J. (2020). "A Comparative Study of Machine Learning Models for Life Expectancy Prediction." *International Journal of Health Data Science*, 8(2), 112-127.

[8]. Ahmad, M., & Khan, S. A. (2020). "Life Expectancy Prediction Using Machine Learning Models." *International Journal of Computer Science and Information Security (IJCSIS)*, 18(7), 78-85.

[9]. A. Palagati, Santhosh Kumar Balan, S. Arun Joe Babulo, L. Raja, K. K. Natarajan and R. Kalimuthu, "Comparative Analysis of Machine Learning Algorithms and Datasets for Detecting Cyberbullying on Social Media Platforms," 2024 International Conference on

Computing and Intelligent Reality Technologies (ICCIRT), Coimbatore, India, 2024, pp. 391-396, doi: 10.1109/ICCIRT59484.2024.10922033.

[10]. S. Muppidi, B. S. Kumar and K. P. Kumar, "Sentiment Analysis of Citation Sentences using Machine Learning Techniques," 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), 2021, pp. 1-5, doi: 10.1109/i-PACT52855.2021.9696703. Date Added to IEEE Xplore: 08 February 2022.

[11]. K. Muthuvel1, P. Muthukumar2 and Thomas Thangam3, “FORECASTING SOLAR POWER GENERATION WITH MACHINE LEARNING TECHNIQUES”, ARPN Journal of Engineering and Applied Sciences, VOL. 19, NO. 22, pp. 1378- 1388, NOVEMBER 2024

[12]. Anand Deepak George Donald. (2021). Empirical Analysis of Block chain and Machine Learning inspired Cloud Security Architectures, Journal of Next Generation Technology, 1(2), 20-28.

[13]. Dr. Ranga Swamy Sirisati, A. Kalyani, V. Rupa, Dr.Pradeep Venuthurumilli, Md Ameer Raza (2024). “Recognition of Counterfeit Profiles on Communal Media using Machine Learning Artificial Neural Networks & Support Vector Machine Algorithms”, Journal of Next Generation Technology (ISSN: 2583-021X), 4(2), pp. 19-27. May 2024.

[14]. S Phani Praveen, Sai Srinivas Vellela, Dr. R. Balamanigandan, "SmartIris ML: Harnessing Machine Learning for Enhanced Multi-Biometric Authentication”, Journal of Next Generation Technology (ISSN: 2583-021X), 4(1), pp.25-36 . Jan 2024.