

**A Project Report on**  
**LIFESPAN PROJECTION MODEL USING**  
**MACHINE LEARNING TECHNIQUES**

Submitted in partial fulfillment for award of

**Bachelor of Technology**  
Degree  
in  
**Computer Science and Engineering**



By

**V. Mary Sofiya (Y21ACS586)**

**P. Sindhu Priya(Y21ACS544)**

**P. Indira Priyadarsini(Y21ACS546)**

**P.V. Tanuja(Y21ACS537)**

Under the guidance of

**Dr.D.N.V.Syam Kumar, M.Tech, PhD**  
Associate Professor

Department of Computer Science and Engineering

**Bapatla Engineering College**

(Autonomous)

(Affiliated to Acharya Nagarjuna University)

**BAPATLA – 522 102, Andhra Pradesh, INDIA**

**2024-2025**

**Department of  
Computer Science and Engineering**



**CERTIFICATE**

This is to certify that the project report entitled that is **LIFESPAN PROJECTION MODEL USING MACHINE LEARNING TECHNIQUES** being submitted by V. Mary Sofiya (Y21ACS586), P. Sindhu Priya (Y21ACS544), P. Indira Priyadarsini (Y21ACS546), P. V. Thanuja (Y21ACS537) in partial fulfilment for the award of the Degree of Bachelor of Technology in Computer Science & Engineering to the Acharya Nagarjuna University is a record of bonafide work carried out by them under our guidance and supervision.

Date:

**Signature of the Guide  
Dr.D.N.V.Syam Kumar,  
Assoc.Prof.**

**Signature of the HOD  
Dr. M. Rajesh Babu  
Assoc. Prof. & Head**

## **DECLARATION**

We declare that this project work is composed by ourselves, that the work contained herein is our own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

V. Mary Sofiya (Y21ACS586)

P. Sindhu Priya(Y21ACS544)

P. Indira Priyadarsini(Y21ACS546)

P.V. Tanuja(Y21ACS537)

## Acknowledgement

We sincerely thank the following distinguished personalities who have given their advice and support for successful completion of the work.

We are deeply indebted to our most respected guide **Dr.D.N.V.Syam Kumar**,,MTech.,PhD, Associate Professor, Department of CSE, for his valuable and inspiring guidance, comments, suggestions and encouragement.

We extend our sincere thanks to **Dr. M. Rajesh Babu**, Assoc. Prof. & Head of the Dept. for extending his cooperation and providing the required resources.

We would like to thank our beloved Principal **Dr.N.Rama Devi** for providing the online resources and other facilities to carry out this work.

We would like to express our sincere thanks to our project coordinator **Dr. P.Pardhasaradhi**, Prof. Dept. of CSE for his helpful suggestions in presenting this document.

We extend our sincere thanks to all other teaching faculty and non-teaching staff of the department, who helped directly or indirectly for their cooperation and encouragement.

V. Mary Sofiya (Y21ACS586)

P. Sindhu Priya(Y21ACS544)

P. Indira Priyadarsini(Y21ACS546)

P.V. Tanuja(Y21ACS537)

# TABLE OF CONTENTS

List of Figures .....	viii
ABBREVIATIONS .....	ix
Abstract .....	x
INTRODUCTION .....	2
1.1 Overview .....	2
1.2 Problem Statement .....	4
1.3 Objectives of the Project .....	4
LITERATURE SURVEY .....	7
FEASIBILITY STUDY .....	11
3.1 Functional .....	11
3.2 Non – Functional .....	12
3.2.1 Problem define .....	12
3.2.2 Preparing data .....	13
3.2.3 Evaluating algorithms .....	13
3.2.4 Improving results .....	13
3.2.5 Prediction the result .....	13
3.3 Technical Requirements .....	13
3.4 Algorithms .....	14
3.4.1 Random Forest .....	14
3.4.2 XGB regressor .....	15
3.4.3 Decision Tree .....	16
3.4.4 Linear Regression .....	17
3.4.5 Adaboost .....	18
3.5 Python .....	19
3.5.1 Features of Python .....	20
3.5.1.1 Platform Independent .....	20
3.5.2 Python Libraries .....	20
3.5.2.1 Pandas .....	20
3.5.2.2 Numpy .....	21
3.5.2.3 Matplotlib .....	21
3.5.2.4 Seaborn .....	22
METHODOLOGY .....	24
4.1 Dataset .....	24
4.2 Data Preparation .....	26

4.2.1 Data Collection .....	26
4.2.2 Data Cleaning.....	26
4.2.3 Data validation .....	26
4.2.4 Feature Selection/Engineering .....	26
4.2.4 Normalization/Scaling .....	27
4.2.5 Handling Categorical Variables .....	27
4.2.6 Splitting Data .....	27
4.2.7 Data Transformation .....	27
4.3 Machine learning algorithms .....	28
4.3.1 K-NN Algorithm (K-Nearest Neighbours) .....	28
4.3.2 Extra Tress Regressor .....	28
4.3.3 SVR.....	29
4.4 Key Performance Indicators .....	31
4.4.1 Accuracy .....	31
4.4.2 Kappa statistic .....	32
4.4.3 Error .....	32
4.4.4 Sensitivity .....	32
4.4.5 Specificity .....	33
4.4.6 Precision.....	33
4.4.7 Recall .....	34
4.4.8 F-measure.....	34
SYSTEM DESIGN .....	36
5.1 System Architecture.....	36
5.2 UML Diagrams .....	36
5.2.1 Structural Diagrams .....	37
5.2.1.1 Class diagram.....	38
5.2.2 Behavioral Diagram .....	39
5.2.2.1 Use Case diagram .....	39
5.2.2.2 Sequence diagram .....	39
5.2.2.3 Activity Diagram .....	40
RESULTS AND DISCUSSIONS .....	43
6.1 Dataset.....	43
6.2 Pre Processing .....	43
6.2.1 Splitting the dataset.....	44
6.3 Data Visualization.....	45

6.4 Feature Importance .....	49
6.5 Result Analysis .....	51
6.6 Input & Output.....	54
CONCLUSION AND FUTURE SCOPE .....	59
7.1 Conclusion .....	59
7.2 Limitations .....	60
7.3 Future Scope .....	60
REFERENCES .....	62

## List of Figures

Figure 1 Confusion matrix .....	31
Figure 2 System Architecture .....	36
Figure 3 Class diagram .....	38
Figure 4 Usecase diagram.....	39
Figure 5 Sequence diagram.....	40
Figure 6 Activity daigram.....	41
Figure 7 Dataset .....	43
Figure 8 Data Pre Processing .....	44
Figure 9 Splitting data.....	44
Figure 10 Visualization of target .....	46
Figure 11 Visualization of lifespan.....	46
Figure 12 Plotting the numerical values .....	47
Figure 13 Feature Importance .....	49
Figure 14 Result of r2 score.....	51
Figure 15 Cross validation .....	52
Figure 16 RMSE Error.....	53
Figure 17 Input screen .....	55
Figure 18 Input screen with inputs.....	56
Figure 19 Output screen for given input .....	57
Figure 20 Output screen when invalid input given.....	57



# ABBREVIATIONS

Abbreviation	Meaning
SVR	Support Vector Regression
RMSE	Root mean Squared Error
WHO	World Health Organization
XGB	eXtreme Gradient Boosting

## Abstract

This project aims to develop a machine learning-based model to predict the life expectancy of individuals across various countries using socio-economic and health-related indicators. A comprehensive dataset was preprocessed through data cleaning, outlier treatment, feature engineering, and encoding strategies to ensure model robustness. Multiple regression algorithms—including Random Forest, XGBoost, SVR, and Linear Regression were trained and evaluated using performance metrics like  $R^2$  score and RMSE.

The Random Forest Regressor emerged as the top-performing model, effectively capturing complex relationships between features such as adult mortality, income composition, schooling, and health expenditure. The model was further integrated into an interactive **Tkinter**, enabling users to input relevant parameters and receive real-time life expectancy predictions.

This solution demonstrates the potential of predictive analytics in public health planning and policy-making, offering actionable insights into factors influencing human longevity. Future enhancements include region-specific models, real-time data integration, and dashboard deployment for decision support systems.

## **CHAPTER 1**

# **INTRODUCTION**

# INTRODUCTION

LifeSpan is a critical indicator of a population's overall health and well-being. In recent years, the use of data analytics and machine learning has emerged as a powerful tool for improving our understanding of lifespan trends and predicting future outcomes. By collecting and analyzing large amounts of data from various sources, such as demographic data, medical records, and environmental factors, thus, we can create predictive models that can provide valuable insights into the factors that influence lifespan using data analytics and machine learning. LifeSpan refers to the average number of years a person can expect to live, based on current age specific mortality rates.

## 1.1 Overview

LifeSpan is an analytical as well as a statistical measure of the longevity of the population depending upon distinct factors. Over the years, LifeSpan observations are being vastly used in medical, healthcare planning, and pension-related services, by concerned government authorities and private bodies. Advancements in forecasting, predictive analysis techniques, and data- science technologies have now made it possible to develop accurate predictive models. In many countries, it is a matter of political debate about how to decide the retirement age and how to manage the financial issues related to the public matter.

LifeSpan predictions provide solutions related to these issues in many developed countries. With the advancement in new systematic, accurate, efficient, and result-oriented techniques in the field of Data Science, now predictions of the LifeSpan of the selected region are becoming more prominent in demand of the government authorities and the private bodies and their policy- making. Studies have suggested that in early life or the pre-modern era, the average lifespan of human beings was around 30 years in approximately all parts of the world. Since then, industrial enterprise and modernization have valued the rapid increase within the lifespan all around the world. The advancement of technology, better healthcare facilities, and education for all have led to positive changes in the lifestyle of people. Which, in turn, increases the expected average age of a human being. However, there were still many countries with less LifeSpan than the rest of the world in the early

1900s. The whole reason for such inequality is the disoriented healthcare facilities in these countries. Developed countries have speedily improved their healthcare and also the public distribution mechanism. This inequality between developed and developing countries has led to such an improper distribution of LifeSpan around the globe.

Due to certain developments in public healthcare, now emerging countries are also catching up with the other developed countries in terms of LifeSpan. In 2019, most of the Central African countries have low LifeSpan of around 52-55 years, whereas, in recent statistics have shown that LifeSpan is around 87 years for women. The lifespan of South Korea was twenty-three years, a century past. Nevertheless, as of today, the LifeSpan of India has almost tripled in the last 100 years, and in South Korea, it has almost quadrupled since that time period.

There have been many vast improvements in the field of data science and analytical techniques, which explains the rise in LifeSpan around the world. These significant improvements in the predictive analysis techniques have also led us to more ways so that authors can improve the LifeSpan of the distinct population. These improvements were solely dependent upon specific indicators.[3] The extensive research into the prior LifeSpan models has suggested us the inclusion of many more indicators than expected, such as; GDP(Gross Domestic Product), healthcare expenditure, family income, educational expenditure, infant mortality rate, adult mortality rate, healthcare plans, and population of the selected region. Recent studies have also revealed the impact of geographical factors, climate conditions on LifeSpan. Implicitly, the educational background of people, health plans, economic stability, and the burden of diseases, BMI, and environmental variables also affect the lifestyle of the people.

The study of the LifeSpan of a population is important for the evaluation of the degree of economic and social development of a country [1]. The residents of a country with high life standards live longer, on average, and have a small mortality ratio [1]. Data analytics and Machine learning can help identify patterns and relationships between different variables and predict future outcomes with high accuracy. This information can be used to develop targeted interventions to improve health outcomes and increase LifeSpan.

## 1.2 Problem Statement

The purpose of this study is to understand and predict trends in LifeSpan, a crucial aspect of public health, using data analytics and machine learning techniques. The study aims to identify the key predictors of LifeSpan and analyze their influence on the average expected lifespan. The study is based on a dataset covering the period from 2000 to 2015 for 193 countries, collected from the WHO data repository and extracted from the Kaggle website. The dataset includes factors such as country-wise population, deaths by different age groups, diseases such as Measles and HIV/AIDS, and immunization rates for Polio and Hepatitis B, among others.

The study uses regression analysis to analyze the data, with the Random Forest Regressor algorithm chosen as the best performer, achieving an accuracy of 95%. The results of this study will have significant implications for public health policy and practice, providing insight into the factors contributing to LifeSpan. The study is implemented using data analytics and machine learning techniques, with the help of a dataset provided by WHO and extracted from Kaggle. The Random Forest Regression is a machine learning model employed in this study to train the model and predict the LifeSpan of a given country.

## 1.3 Objectives of the Project

Using data analytics and machine learning models to study LifeSpan can have several objectives, including:

**Prediction:** Developing models that accurately predict LifeSpan based on various demographic, socio-economic, environmental, and health-related factors. This can help individuals, healthcare providers, and policymakers anticipate healthcare needs and allocate resources efficiently.

**Identification of Key Factors:** Analyzing the importance of different variables in determining LifeSpan. This can help identify factors that have the most significant impact on LifeSpan, such as access to healthcare, income levels, education, lifestyle choices, and environmental factors.

**Healthcare Resource Allocation:** Understanding how different regions or demographic groups vary in terms of LifeSpan can aid in allocating healthcare resources effectively. For example, identifying areas with lower LifeSpan can help prioritize healthcare interventions and allocate resources where they are most needed.

**Intervention Strategies:** Recommending targeted interventions or policies to improve LifeSpan. By understanding the factors influencing LifeSpan, data analytics can suggest specific interventions, such as improving access to healthcare services, promoting healthier lifestyles, or addressing socio-economic disparities.

**Risk Assessment:** Assessing individual or population-level risk factors for premature mortality. Machine learning models can help identify individuals or groups at higher risk of shorter LifeSpan based on their characteristics, allowing for early interventions and personalized healthcare approaches.

**Healthcare Planning and Policy Making:** Providing insights for healthcare planning and policy-making at local, regional, and national levels. By analyzing patterns and trends in LifeSpan data, policymakers can develop targeted policies to improve public health outcomes and reduce health inequalities.

## **CHAPTER 2**

# **LITERATURE SURVEY**



# LITERATURE SURVEY

Ayshwaryaa N et al, proposed that Human an incredible creation of god. Every creature in the world has a limited life span, to achieve something in the world.. To preserve our self from the consequences, even though lot of inventions has been made by human, to prevent from diseases is a major question mark. Life span prediction has a greater impact in our modern society because of our food habits, different types of diseases and environmental conditions.

Linda Mary et al, proposed that the correlation between attributes like diseases, gender, ages and environmental factor are important. In this paper, In order to find or predict the human lifespan with more accuracy we use random forest algorithm.

V.M Shkolnikov et al, proposed that Predicting life span for human being is a vital step. It is an emerging research area that is gaining interest but involved lot of challenges due to the limited number of resources (i.e., datasets) available. By obtaining the Date of birth, Environmental factors, Food habits, Diseases and Medical history, a lot of investigations will be conducted to predict the sustainability of human.

D.F.Andrews et al, proposed that when there is change in small fraction the data techniques will be resistant. Otherwise, when the efficiency of statistics held high then the techniques will be robust. If the accuracy score is excellent then the result of the predicted one is accurate

D.M.J Naimark et al, proposed that the expectancy of the life can be grasped to equal to area under a certain region He proposed it is necessary to understand the baseline risk under the control group. By the help of different models we can predict the LifeSpan.

A.A. Bhosale et al, proposed that expectancy of the life mainly target on predicting models using trends. He proposed LifeSpan rely on weight, adult mortality, heart rate, respiration rate for human beings. The inspection provides the standard LifeSpan is forecasted by variables that can be easily calculated

M.K.Z. Sormin et al, proposed to rough calculate the LifeSpan of the population across the world so that it will be helpful to the particular country to increase their health of the human beings. The Cyclic Order Weight neural network method is used for the appraise.

Lifespan prediction has become an essential task in global health research, aiding governments and health organizations in strategic planning and resource allocation. Traditionally, demographic and statistical models were used, but the rise of machine learning (ML) has opened new doors for predictive modeling due to its capacity to process large, multidimensional datasets and uncover hidden patterns.

Several studies have employed different ML algorithms to predict life expectancy using a variety of features such as health indicators, economic factors, and environmental metrics. Commonly used datasets include the World Health Organization (WHO) Global Health Observatory and the UCI Life Expectancy Dataset, which contains data from over 190 countries and includes variables such as GDP, mortality rates, immunization coverage, and healthcare expenditures. These datasets provide the foundation for training robust models.

In terms of methodologies, a wide range of machine learning algorithms have been applied. Linear regression is often used as a baseline due to its simplicity, but it struggles with nonlinear relationships. Random Forests and Gradient Boosting algorithms like XGBoost have consistently shown better performance in terms of accuracy and robustness. Artificial Neural Networks (ANNs) are also explored for modeling complex, non-linear dependencies, although they often face challenges with interpretability.

For example, S. Ali et al. (2019) compared various algorithms and found Random Forest to have the highest predictive power with an  $R^2$  of 0.92. Another study by Gangopadhyay et al. (2021) highlighted that GDP, schooling, and immunization rates were among the most influential features and that Gradient Boosting methods offered the lowest prediction error. Meanwhile, researchers like Sharma et al. (2022) have focused on explainable AI (XAI) to improve transparency, using tools like SHAP to interpret model decisions and identify key contributing factors.

In conclusion, machine learning offers powerful tools for life expectancy prediction, with ensemble methods and deep learning models providing accurate and scalable solutions. However, for these models to be effectively implemented in public health policy, future work must address issues of interpretability, data quality, and real-time adaptability. Combining data-driven models with domain expertise and ethical considerations will be key to maximizing their impact.

## **CHAPTER 3**

# **FEASIBILITY STUDY**

# FEASIBILITY STUDY

A feasibility study is a comprehensive analysis that assesses the practicality and viability of a proposed project or business venture. It examines various factors, including technical, economic, financial, legal, and environmental considerations, to determine if the project is worth pursuing. The goal is to provide stakeholders with the information needed to make informed decisions about whether to proceed with the project, adjust it, or abandon it.

Requirements are the basic constraints that are required to develop a system. Requirements are collected while designing the system.

The following are the requirements that are to be discussed.

- Functional requirements

- Non-Functional requirements

- Technical requirements

  - Hardware requirements

  - Software

## 3.1 Functional

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow libraries like skit-learn, pandas, numpy, matplotlib and seaborn .

The system shall collect and preprocess data from diverse sources such as demographic information, medical records, genetic data, lifestyle choices, and environmental factors. It shall support data cleaning, normalization, and transformation to ensure compatibility with machine learning algorithms. The system shall allow users to input individual or population-level data and receive lifespan projections based on trained

models. It must employ machine learning models, such as regression, decision trees, or neural networks, that can be trained, validated, and tested using historical data. The system shall allow model evaluation using standard metrics like accuracy, RMSE, or MAE to assess prediction reliability. Additionally, it shall support model retraining as new data becomes available, ensuring projections remain current and relevant.

The system shall provide an interface for users to visualize results in the form of graphs, tables, or interactive dashboards. It must allow filtering and comparison of lifespan predictions across different user groups or data categories. The system shall also provide interpretability features, such as highlighting key factors influencing the prediction, to increase transparency and trust in the model's output. Moreover, it must ensure secure access to sensitive data, enforcing privacy regulations such as GDPR or HIPAA where applicable. Finally, the system shall log user interactions and system decisions to support auditing, error tracking, and continuous improvement of the projection pipeline.

## **3.2 Non – Functional**

Non-functional requirements (NFRs) define how a system should behave, focusing on its quality attributes rather than its specific functions. They address how well a system performs, including aspects like performance, security, usability, reliability, and scalability. In essence, NFRs specify the "how" of a system, ensuring it meets user expectations and operates effectively.

Steps in non-functional requirements:

### **3.2.1 Problem define**

The first step in the process is problem definition, where the objective of the lifespan projection system is clearly identified. This involves understanding the scope, such as whether the system is aimed at individual health prediction or broader population-level analysis, and defining measurable goals for the machine learning model. Key stakeholders and data availability are also considered at this stage to ensure feasibility and relevance.

### 3.2.2 Preparing data

In the data preparation phase, which involves collecting, cleaning, and organizing data from various sources including medical records, lifestyle surveys, environmental data, and demographic statistics. This stage includes handling missing values, normalizing datasets, encoding categorical variables, and possibly selecting relevant features through statistical or algorithmic methods to improve model performance.

### 3.2.3 Evaluating algorithms

During this step, multiple machine learning models—such as linear regression, decision trees, support vector machines, or ensemble methods—are trained and validated on the prepared dataset. The system compares their performance using evaluation metrics like mean absolute error (MAE), root mean square error (RMSE), or  $R^2$  score, depending on the problem context.

### 3.2.4 Improving results

After initial evaluation, the focus shifts to improving results. This involves techniques such as hyperparameter tuning, feature engineering, and possibly introducing more sophisticated models like deep learning architectures. The system may also use cross-validation or ensemble strategies to enhance prediction accuracy and robustness.

### 3.2.5 Prediction the result

The system moves to the prediction phase, where the optimized model is used to make lifespan predictions for new or unseen data. The results are then presented through user-friendly visualizations, such as graphs or dashboards, and may include explanations of the key factors influencing each prediction. This stage enables actionable insights for healthcare professionals, researchers, or individuals seeking to understand and potentially extend their lifespan.

## 3.3 Technical Requirements

Software Requirements:

- a. Operating System : Windows

- b. Tool : Anaconda with Jupyter Notebook

Hardware requirements:

- a. Processor : Pentium IV/III
- b. Hard disk : minimum 80 GB
- c. RAM : minimum 2 GB

## 3.4 Algorithms

Various algorithms are used and compares the accuracy between them.

### 3.4.1 Random Forest

The Random Forest algorithm has proven to be a valuable tool in predicting LifeSpan through data analytics and machine learning techniques. By leveraging this algorithm, researchers and analysts can sift through vast amounts of data encompassing various socio-economic, environmental, and healthcare factors to derive insights into LifeSpan trends.

One key advantage of Random Forest is its ability to handle large datasets with numerous features, making it suitable for the multidimensional nature of LifeSpan prediction. This algorithm works by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. This ensemble approach reduces overfitting and enhances the model's generalization capabilities. In the context of LifeSpan prediction, Random Forest can effectively incorporate diverse variables such as GDP per capita, access to healthcare services, education levels, environmental factors like pollution levels, and lifestyle choices such as smoking rates and diet patterns. By considering these factors simultaneously, the algorithm can capture complex relationships and interactions that influence LifeSpan outcomes. Moreover, Random Forest provides insights into feature importance, enabling researchers to identify the most influential factors affecting LifeSpan. This information can inform policymakers and public health initiatives to prioritize interventions that address the most significant determinants of LifeSpan within a population.



Additionally, Random Forest's robustness to outliers and missing data further enhances its applicability in LifeSpan analysis. It can handle noisy or incomplete datasets common in real-world scenarios, ensuring reliable predictions even in imperfect conditions. This capability is crucial for accurately assessing LifeSpan trends across diverse demographics and geographical regions.

The Random Forest algorithm offers a powerful framework for analyzing and predicting LifeSpan by leveraging the vast amounts of data available. Its ability to handle complex, multidimensional datasets, interpret feature importance, and robustness to outliers make it a valuable tool for researchers, policymakers, and healthcare professionals seeking to understand and improve LifeSpan outcomes.

### **3.4.2 XGB regressor**

LifeSpan prediction is a crucial task in public health, and data analytics coupled with machine learning techniques like the XGB Regressor algorithm offer promising avenues for accurate predictions. XGB, short for eXtreme Gradient Boosting, is a powerful ensemble learning algorithm known for its speed and performance. In the context of LifeSpan prediction, XGB Regressor can effectively handle complex datasets with numerous features, making it a suitable choice for modeling the intricate relationships between various factors and LifeSpan.

One of the primary steps in utilizing the XGB Regressor algorithm for LifeSpan prediction involves data preprocessing. This includes handling missing values, scaling features, and encoding categorical variables. Additionally, feature selection techniques can be employed to identify the most relevant predictors impacting LifeSpan, thereby improving model interpretability and performance.

Once the data is prepared, the XGB Regressor algorithm is trained on historical datasets containing features such as demographics, socio-economic indicators, healthcare access, and lifestyle factors. Through iterative learning, XGB optimizes a set of weak learners (decision trees) to create a strong predictive model. Its gradient boosting framework allows for the sequential optimization of residuals, enabling the model to gradually minimize prediction errors and capture complex interactions within the data.

Cross-validation techniques such as k-fold validation are commonly employed to evaluate the performance of the XGB Regressor model and assess its generalization capabilities. By partitioning the dataset into multiple subsets, this method ensures that the model's performance metrics, such as mean absolute error or R-squared, are robust and reliable across different data splits.

Once the XGB Regressor model is trained and validated, it can be deployed to predict LifeSpan for new or unseen data instances. This predictive capability can provide valuable insights for policymakers, healthcare professionals, and researchers to identify at-risk populations, allocate resources effectively, and implement targeted interventions aimed at improving overall public health outcomes. Moreover, continuous monitoring and updating of the model with new data can ensure its relevance and accuracy in reflecting evolving socio-demographic trends and health-related behaviors.

### 3.4.3 Decision Tree

Decision trees are a powerful machine learning algorithm widely used in data analytics, including predicting LifeSpan based on various factors. In this context, decision trees analyze historical data to identify patterns and relationships between different variables and outcomes. The algorithm works by recursively partitioning the data into subsets based on features such as demographic information, lifestyle choices, healthcare access, and environmental factors.

One of the key benefits of using decision trees for predicting LifeSpan is their ability to handle both numerical and categorical data. This flexibility allows the algorithm to incorporate diverse factors that may influence LifeSpan, such as age, gender, income level, education, smoking habits, and prevalence of diseases. By considering multiple variables simultaneously, decision trees can capture complex interactions and nonlinear relationships, providing more accurate predictions compared to traditional statistical methods.

Moreover, decision trees offer transparency and interpretability, which are essential for understanding the factors driving LifeSpan predictions. Each node in the tree represents a decision based on a specific feature, and the branches correspond to different outcomes

or subsequent decisions. This hierarchical structure allows analysts to trace the decision-making process and identify the most influential factors affecting LifeSpan. As a result, decision trees not only provide accurate predictions but also valuable insights into the underlying mechanisms shaping LifeSpan disparities.

Another advantage of decision trees is their ability to handle missing values and outliers effectively. LifeSpan datasets often contain incomplete or noisy data, which can undermine the performance of predictive models. Decision trees handle missing values by selecting the most informative features for partitioning the data, ensuring that predictions are based on the available information. Additionally, decision trees are robust to outliers, as they partition the data into smaller subsets, reducing the impact of extreme values on the overall predictions.

Furthermore, decision trees facilitate feature selection and variable importance analysis, allowing analysts to identify the most significant predictors of LifeSpan. By evaluating the splits in the tree and the associated decrease in impurity or increase in information gain, analysts can rank the features based on their predictive power. This information is invaluable for policymakers, healthcare professionals, and researchers seeking to prioritize interventions and allocate resources effectively to improve LifeSpan outcomes. Overall, the decision tree algorithm is a valuable tool in data analytics for predicting LifeSpan and gaining insights into the complex interplay of factors influencing human health and longevity.

### **3.4.4 Linear Regression**

Linear regression is a fundamental machine learning algorithm widely used in data analytics, including predicting LifeSpan. By analyzing various demographic, socio-economic, and health-related factors, linear regression models can estimate LifeSpan trends within populations. In this context, linear regression serves as a predictive tool to understand how different variables contribute to LifeSpan and to forecast LifeSpan changes over time.

Firstly, data collection is crucial for building a robust linear regression model for LifeSpan prediction. Variables such as income levels, access to healthcare, education, prevalence of diseases, and environmental factors are typically gathered from diverse

sources. This data forms the basis for understanding the relationship between these predictors and LifeSpan. Linear regression helps in identifying which variables have a significant impact on LifeSpan and how they influence it quantitatively. Once the data is collected and preprocessed, linear regression is applied to establish a mathematical relationship between the predictor variables and LifeSpan. The algorithm aims to find the best-fitting line that minimizes the difference between the predicted and actual life expectancies. Through this process, it quantifies the contribution of each predictor variable, providing insights into the factors that most strongly influence LifeSpan.

Furthermore, linear regression facilitates the interpretation of the model's coefficients, which represent the magnitude and direction of the relationship between each predictor variable and LifeSpan. For instance, a positive coefficient for income levels suggests that higher incomes are associated with longer life expectancies, while a negative coefficient for pollution levels indicates that increased pollution is linked to lower life expectancies. These insights can inform policymakers and healthcare professionals in designing interventions to improve LifeSpan.

Moreover, linear regression models can be used for forecasting future LifeSpan trends based on changes in predictor variables. By extrapolating the established relationships, these models can predict how alterations in socio-economic factors, healthcare policies, or environmental conditions might impact LifeSpan over time. Such forecasts enable proactive decision-making and resource allocation to address potential challenges or capitalize on opportunities for improving population health and well-being.

Linear regression is a valuable tool in data analytics and machine learning for predicting LifeSpan. By analyzing diverse sets of variables, building predictive models, interpreting coefficients, and forecasting future trends, linear regression provides valuable insights into the factors influencing LifeSpan within populations. These insights are instrumental in informing public health policies, healthcare interventions, and social programs aimed at enhancing LifeSpan and overall quality of life.

### **3.4.5 Adaboost**

AdaBoost (Adaptive Boosting) algorithm is a powerful ensemble learning method that combines multiple weak learners to create a strong classifier. When applied to

predicting LifeSpan using data analytics and machine learning, AdaBoost can offer several advantages. Firstly, AdaBoost excels in handling complex datasets with multiple features, which is often the case in LifeSpan prediction where various socio-economic, demographic, and health-related factors play a role.

Secondly, AdaBoost is effective in dealing with imbalanced datasets, which is common in healthcare data where certain demographics may be underrepresented. By iteratively adjusting the weights of misclassified instances, AdaBoost focuses on the most challenging samples, thereby improving the overall predictive performance.

Thirdly, AdaBoost is robust against overfitting, a crucial consideration in LifeSpan prediction where generalization to unseen data is paramount. Its sequential training process, where each new model corrects errors made by the previous ones, helps in building a balanced and generalized predictor.

Moreover, AdaBoost's flexibility in accommodating various base learners, such as decision trees or regression models, allows for the incorporation of domain-specific knowledge into the prediction process. For instance, decision trees can capture nonlinear relationships between predictors and LifeSpan, while regression models can quantify the impact of continuous variables.

Lastly, the interpretability of AdaBoost's final model provides valuable insights into the factors influencing LifeSpan. By analyzing the feature importance scores derived from AdaBoost, policymakers, healthcare professionals, and researchers can identify the most influential determinants and formulate targeted interventions to improve public health outcomes and increase LifeSpan. Overall, AdaBoost's adaptability, robustness, and interpretability make it a valuable tool in leveraging data analytics and machine learning for predicting LifeSpan and informing healthcare decision making.

### 3.5 Python

Python is a widely used general-purpose, high level programming language. It was created by Guido van Rossum in 1991 and further developed by the Python Software

Foundation. It was designed with an emphasis on code readability, and its syntax allows programmers to express their concepts in fewer lines of code.

Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions: Python 2 and Python 3. Both are quite different.

### **3.5.1 Features of Python**

There are no separate compilation and execution steps like C and C++. Directly run the program from the source code. Internally, Python converts the source code into an intermediate form called bytecodes which is then translated into native language of specific computer to run it. No need to worry about linking and loading with libraries, etc.

#### **3.5.1.1 Platform Independent**

Python programs can be developed and executed on multiple operating system platforms. Python can be used on Linux, Windows, Macintosh, Solaris and many more. Free and Open Source; Redistributable The Python Standard Library is very vast. Known as the “batteries included” philosophy of Python; It can help do various things involving regular expressions, documentation generation, unit testing, threading, databases, web browsers, CGI, email, XML, HTML, WAV files, cryptography, GUI and many more. Besides the standard library, there are various other high-quality libraries such as the Python Imaging Library which is an amazingly simple image manipulation library.

### **3.5.2 Python Libraries**

#### **3.5.2.1 Pandas**

Pandas is a Python library used for working with data sets.

It has functions for analyzing, cleaning, exploring, and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas allows us to analyze big data and make conclusions based on statistical theories.

Pandas can clean messy data sets, and make them readable and relevant.

Relevant data is very important in data science.

Pandas gives you answers about the data. Like:

Is there a correlation between two or more columns?

What is average value?

Max value?

Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

### 3.5.2.2 Numpy

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.

NumPy stands for Numerical Python. In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called `ndarray` it provides a lot of supporting functions that make working with `ndarray` very easy.

Arrays are very frequently used in data science, where speed and resources are very important. NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.

### 3.5.2.3 Matplotlib

Matplotlib is a low level graph plotting library in python that serves as a visualization utility.

Matplotlib was created by John D. Hunter.

Matplotlib is open source and we can use it freely.

Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility.

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram, etc

### **3.5.2.4 Seaborn**

Seaborn is a Python data visualization library based on matplotlib. It provides a highlevel interface for drawing attractive and informative statistical graphics.

Seaborn is a library for making statistical graphics in Python. It provides a high-level interface to matplotlib and integrates closely with pandas data structures. Functions in the seaborn library expose a declarative, dataset-oriented API that makes it easy to translate questions about data into graphics that can answer them. When given a dataset and a specification of the plot to make, seaborn automatically maps the data values to visual attributes such as color, size, or style, internally computes statistical transformations, and decorates the plot with informative axis labels and a legend. Many seaborn functions can generate figures with multiple panels that elicit comparisons between conditional subsets of data or across different pairings of variables in a dataset. seaborn is designed to be useful throughout the lifecycle of a scientific project. By producing complete graphics from a single function call with minimal arguments, seaborn facilitates rapid prototyping and exploratory data analysis. And by offering extensive options for customization, along with exposing the underlying matplotlib objects, it can be used to create polished, publication-quality figures.



## **CHAPTER 4**

# **METHODOLOGY**

# METHODOLOGY

Methodology refers to the systematic set of methods, principles, and rules used to guide a particular study, research, or analysis. It outlines **how** the research will be conducted, what tools or techniques will be used, and **why** those choices are appropriate for achieving the study's objectives.

## 4.1 Dataset

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The datasets are made available to public for the purpose of health data analysis. The dataset related to LifeSpan, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years.

Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single dataset. On initial visual inspection of the data showed some missing values. As the datasets were from WHO, we found no evident errors. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model dataset. The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors.

1. Country – Country name
2. Year – Year in which you wanted to predict the lifespan
3. Status – Developed or Developing status
4. LifeSpan – LifeSpan in age
5. Adult Mortality – Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
6. Infant deaths – Number of Infant Deaths per 1000 population
7. Alcohol – Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
8. Percentage expenditure – Expenditure on health as a percentage of Gross Domestic Product per capita(%)
9. Hepatitis B – Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
10. Measles – Measles – number of reported cases per 1000 population
11. BMI – Average Body Mass Index of entire population
12. Under-five deaths – Number of under-five deaths per 1000 population
13. Polio – Polio (Pol3) immunization coverage among 1-year-olds (%)
14. Total expenditure – General government expenditure on health as a percentage of total government expenditure (%)
15. Diphtheria – Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
16. HIV/AIDS – Deaths per 1 000 live births HIV/AIDS (0-4 years)
17. GDP – Gross Domestic Product per capita (in USD)
18. Population – Population of the country

19. Thinness 1-19 years – Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
20. Thinness 5-9 years – Prevalence of thinness among children for Age 5 to 9(%)
21. Income composition – Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
22. Schooling – Number of years of Schooling(years)

## **4.2 Data Preparation**

Data preprocessing is a critical step in any data analytics or machine learning project, especially when working with LifeSpan data. Here's a general outline of steps you might take in preprocessing data for a machine learning model predicting LifeSpan:

### **4.2.1 Data Collection**

Collect reliable datasets containing features that might influence LifeSpan, such as demographic, socioeconomic, healthcare, and environmental factors. Sources could include government databases, research publications, or specialized datasets.

### **4.2.2 Data Cleaning**

Handle missing values: Identify missing values and decide on a strategy to deal with them (e.g., imputation, removal).

Remove duplicates: Check for and remove duplicate records if any.

### **4.2.3 Data validation**

Check for data integrity issues, such as outliers, incorrect data types, or inconsistent data. Standardize data formats: Ensure consistency in data formats across different features.

### **4.2.4 Feature Selection/Engineering**

Identify relevant features: Select features that are likely to have a significant impact on LifeSpan based on domain knowledge or statistical analysis.

Create new features: Engineer new features that might provide additional predictive power, such as ratios, categories, or transformations of existing features.

#### **4.2.4 Normalization/Scaling**

Normalize numerical features: Scale numerical features to a similar range to prevent features with large values from dominating the model.

Standardize features: Standardize features to have a mean of 0 and a standard deviation of 1, which can be important for some machine learning algorithms.

#### **4.2.5 Handling Categorical Variables**

Encode categorical variables: Convert categorical variables into numerical representations using techniques like one-hot encoding or label encoding.

Deal with high cardinality: If categorical variables have many unique values, consider techniques like feature hashing or grouping rare categories.

#### **4.2.6 Splitting Data**

Split data into training, validation, and test sets to evaluate model performance properly. Typically, 70-80% for training, 10-15% for validation, and 10-15% for testing.

Handling Imbalanced Data (if applicable):

If the dataset is imbalanced (e.g., significantly more data for one class than another), consider techniques like oversampling, undersampling, or generating synthetic samples to balance the dataset.

#### **4.2.7 Data Transformation**

Log transformations: Apply log transformations to skewed numerical features to make their distributions more Gaussian-like.

Box-Cox transformations: Use Box-Cox transformations to stabilize variance and make the data more normally distributed.

## 4.3 Machine learning algorithms

The objective of this work is to analyze the possibilities offered by machine learning algorithms as tools for rain forecasting as an alternative to classical forecasting methods. For this, the following algorithms were applied: knn, decision tree, random forest, and neural networks. The algorithms are described below.

### 4.3.1 K-NN Algorithm (K-Nearest Neighbours)

It is a supervised algorithm (therefore, it takes labeled data as input) that for each unlabeled data sample identifies the K closest samples of the input data and assigns the class of most of the K closest neighbors to the unlabeled sample. The algorithm requires the use of a function to calculate the distance between samples.

In theory, to choose the number K of closest samples to consider and avoid overfitting and underfitting, a bias–variance tradeoff is performed. It is a balance to reduce the impact of data with high variance and not ignore the trends generated by a small amount of data generating an offset (if a very high number is chosen for K, then the model would always predict the most common class, and if a very small value was chosen this would generate a very noisy result). In practice, the selection of the value of K will depend on each case and the amount of data used to train the model. However, there are several widely used techniques, such as choosing the square root of the number of samples as a starting point and performing iterations considering different samples of training data, to be able to conclude on a suitable value of K or using a high value of K but applying a weight function to give more importance to the closest neighbors to the sample on which its class must be decided.

### 4.3.2 Extra Trees Regressor

The Extra Trees Regressor, or Extremely Randomized Trees Regressor, is an ensemble learning technique that builds multiple decision trees and aggregates their predictions for regression problems.<sup>1</sup> It distinguishes itself from Random Forest primarily through its approach to feature splitting.<sup>2</sup> Instead of searching for the optimal split among a random subset of features, Extra Trees randomly selects splits, significantly accelerating

the tree-building process.<sup>3</sup> This increased randomization also extends to how the trees are trained; while Random Forest uses bootstrap aggregation, sampling with replacement, Extra Trees trains each tree on the entire original dataset. This key difference further contributes to the algorithm's computational efficiency, making it particularly advantageous for large datasets.

The core strength of the Extra Trees Regressor lies in its ability to reduce variance and mitigate overfitting. By averaging the predictions of numerous randomly constructed trees, the model achieves a more stable and generalized prediction. The random split selection, while potentially introducing some bias, effectively prevents the trees from becoming overly tailored to the training data. This randomization, combined with the use of the entire dataset for each tree, leads to a more diverse ensemble, ultimately improving the model's performance on unseen data.

Furthermore, Extra Trees excels in handling high-dimensional datasets, a common challenge in modern machine learning applications.<sup>4</sup> The algorithm's ability to efficiently process a large number of features, coupled with its inherent resistance to overfitting, makes it a robust choice for complex regression tasks.<sup>5</sup> While the ensemble nature of Extra Trees does compromise interpretability compared to single decision trees, its predictive accuracy and computational efficiency often outweigh this drawback. Key parameters, such as the number of trees (`n_estimators`), the number of features considered for splitting (`max_features`), and the maximum depth of the trees (`max_depth`), allow for fine-tuning the model's performance.

In essence, the Extra Trees Regressor provides a powerful and efficient alternative to Random Forest, particularly when computational speed is paramount.<sup>6</sup> Its unique approach to feature splitting and data sampling results in a robust ensemble model capable of delivering accurate predictions across a wide range of regression problems.<sup>7</sup> The balance it strikes between randomization and performance makes it a valuable tool in the machine learning practitioner's arsenal.

### 4.3.3 SVR

Support Vector Regression (SVR) is a powerful machine learning technique used for regression tasks. It's an adaptation of the Support Vector Machine (SVM), which is

primarily known for classification. While SVMs aim to find a hyperplane that best separates data into distinct classes, SVR aims to find a function that best approximates the relationship between input features and a continuous target variable. The core idea behind SVR is to find a hyperplane that fits within a certain margin of error, allowing for some tolerance in the predictions. This "margin" is defined by a parameter called epsilon ( $\epsilon$ ), which dictates the width of the tube within which the predicted values should lie.

A key strength of SVR lies in its ability to handle both linear and non-linear relationships within data. This is achieved through the use of kernel functions. Kernel functions allow SVR to implicitly map input data into higher-dimensional feature spaces, where linear relationships might exist even if they're not apparent in the original input space. Common kernel functions include the linear kernel, polynomial kernel, and radial basis function (RBF) kernel. The choice of kernel function significantly impacts the model's performance, and selecting the appropriate kernel often involves experimentation and cross-validation.

Beyond the kernel function, other important parameters influence SVR's behavior. The "C" parameter, for instance, controls the trade-off between achieving a flat function and tolerating deviations larger than epsilon. A higher "C" value emphasizes minimizing errors, potentially leading to a more complex model that might overfit the training data. Conversely, a lower "C" value allows for more errors, resulting in a simpler model that might underfit. The epsilon ( $\epsilon$ ) parameter itself determines the margin of tolerance, as previously mentioned. Proper tuning of these hyperparameters is crucial for optimal SVR performance.

In practical applications, SVR is valued for its robustness against outliers and its ability to generalize well to unseen data. It's employed in various fields, including finance, forecasting, and time series analysis. However, SVR can be computationally expensive, especially with large datasets, and careful feature scaling is often necessary to ensure optimal performance. Understanding the interplay between kernel functions and hyperparameters is essential for effectively utilizing SVR in machine learning regression tasks.



## 4.4 Key Performance Indicators

This section defines the metrics (or KPIs, Key Performance Indicators) used to be able to evaluate the results of the algorithms used:

Confusion matrix:

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 1 Confusion matrix

### 4.4.1 Accuracy

Numeric value indicating the performance of the predictive model. It is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP},$$

where:

*TP*: True Positive. Result in which the model correctly predicts the positive class.

*FP*: False Positive. Result in which the model incorrectly predicts the positive class.

*TN*: True Negative. Result in which the model correctly predicts the negative class.

*FN*: False Negative. Result in which the model incorrectly predicts the negativclass.

### 4.4.2 Kappa statistic

It measures the agreement between two examiners in their corresponding classifications of  $N$  elements into  $C$  mutually exclusive categories. In the case of machine learning, it refers to the actual class and the class expected by the model used. It is calculated as follows:

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

### 4.4.3 Error

The error gives an indication of how far the predictions are from the actual output. There are two formulas: Mean Absolute Error (MAE) and Mean Squared Error (MSE).

It is calculated as follows:

$$MAE = \frac{1}{N} \sum_{k=1}^N |y_k - \widehat{y}_k|$$

$$MSE = \frac{1}{N} \sum_{j=1}^N (y_k - \widehat{y}_k)^2,$$

where:

$N$  corresponds to the total number of samples.

$y_k$  corresponds to the class indicated by the classification model.

$y_k^{\wedge}$  corresponds to the actual class.

### 4.4.4 Sensitivity

The sensitivity of a model (or the ratio of true positives) measures the proportion of correctly classified positive examples. The total number of positives is the sum of those that were correctly classified and those that were incorrectly classified. It is calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN},$$

where:

*TP*: True Positive. Result in which the model correctly predicts the positive class.

*TN*: True Negative. Result in which the model correctly predicts the negative class.

*FN*: False Negative. Result in which the model incorrectly predicts the negative class.

#### 4.4.5 Specificity

The specificity of a model (or the ratio of true negatives) measures the proportion of correctly classified negative examples. The total number of negatives is the sum of those that were correctly classified and those that were incorrectly classified. It is calculated as follows:

$$\text{Specificity} = \frac{TN}{TN + FP},$$

where:

*TP*: True Positive. Result in which the model correctly predicts the positive class.

*FP*: False Positive. Result in which the model incorrectly predicts the positive class.

*TN*: True Negative. Result in which the model correctly predicts the negative class.

#### 4.4.6 Precision

Precision is defined as the proportion of examples classified as positive that are actually positive. That is, when a model predicts values as positive. It is calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FN},$$

where:

*TP*: True Positive. Result in which the model correctly predicts the positive class.

*TN*: True Negative. Result in which the model correctly predicts the negative class.

*FN*: False Negative. Result in which the model incorrectly predicts the negative class.

#### 4.4.7 Recall

Recall is defined as the number of correctly classified positives over the total number of positives. This formula is the same as that for sensitivity. It is calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FP}$$

where:

*TP*: True Positive. Result in which the model correctly predicts the positive class.

*FP*: False Positive. Result in which the model incorrectly predicts the positive class.

*TN*: True Negative. Result in which the model correctly predicts the negative class.

#### 4.4.8 F-measure

F-measure is a measure of model performance that combines precision and recall into a value called F-measure (also called F1 score or F-score). This measure combines precision and recall using harmonic averaging, which is used for ratios. This type of averaging is used instead of arithmetic, since precision and recall are expressed as proportions between 0 and 1, which can be interpreted as ratios. It is calculated as follows:

$$\text{F-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

## **CHAPTER 5**

# **SYSTEM DESIGN**

# SYSTEM DESIGN

System Design is the process of defining the architecture, components, modules, interfaces and data flows of a system to meet specific requirements and goals. It involves creating a detailed plan that ensures the system functions effectively and efficiently.

## 5.1 System Architecture

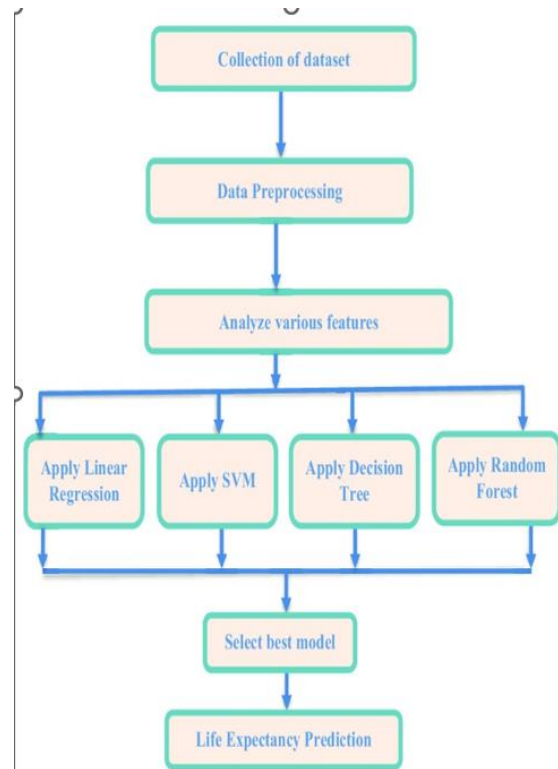


Figure 2 System Architecture

## 5.2 UML Diagrams

UML is an acronym that stands for Unified Modeling Language. Simply put, UML is a modern approach to modeling and documenting software. It is based on diagrammatic representations of software components. As the old proverb says: “a picture is worth a thousand words”. By using visual representations, we are able to better understand possible flaws or errors in software or business processes.

Mainly, UML has been used as a general-purpose modeling language in the field of software engineering. However, it has now found its way into the documentation of several

business processes or workflows. For example, activity diagrams, a type of UML diagram, can be used as a replacement for flowcharts. They provide both a more standardized way of modeling workflows as well as a wider range of features to improve readability and efficacy.

Any complex system is best understood by making some kind of diagram or pictures. These diagrams have a better impact on our understanding. There are two broad categories of diagrams and they are again divided into subcategories

- a. Structural Diagrams

- b. Behavioral Diagrams

### **5.2.1 Structural Diagrams**

The Structural diagrams represent the static aspect of the system. These static aspects represent those parts of a diagram, which forms the main structure and therefore stable. These static parts are represented by classes, interfaces, objects, components, and nodes. The four structural diagrams are

- a. Class diagram

- b. Object diagram

- c. Component diagram

- d. Deployment diagram

### 5.2.1.1 Class diagram

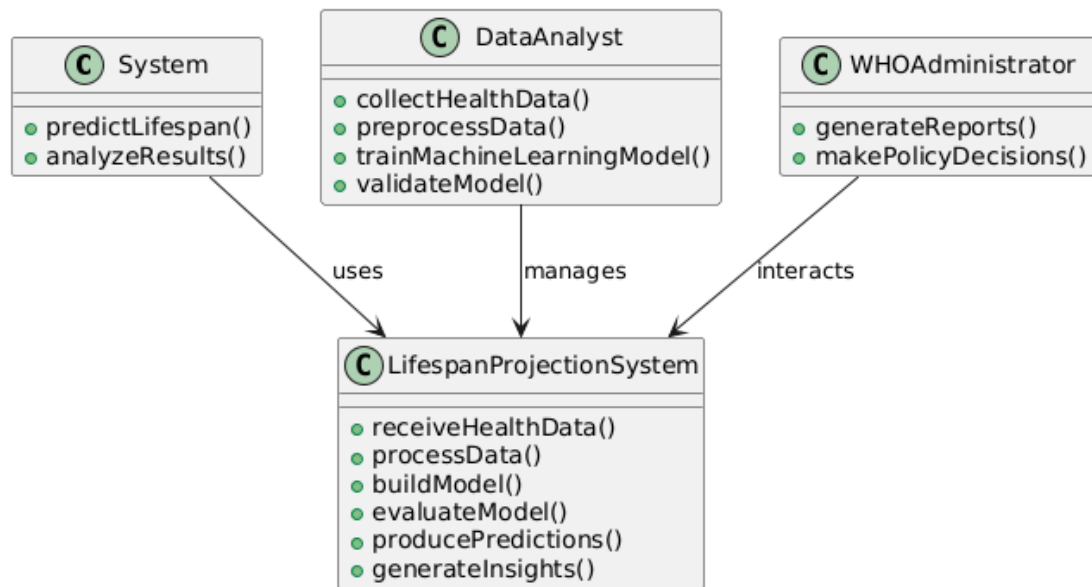


Figure 3 Class diagram

This class diagram illustrates the architecture of a lifespan projection system, outlining the key components and their interactions. At the core, the **LifespanProjectionSystem** encapsulates the functionalities required to process health data, build predictive models, and generate insights. The **DataAnalyst** class assumes a pivotal role in managing this system, encompassing tasks such as collecting and preprocessing health data, training machine learning models, and validating their accuracy. The **System** class then leverages the **LifespanProjectionSystem** to perform higher-level operations, specifically predicting lifespans and analyzing the resulting data. Finally, the **WHOAdministrator** interacts with the system by utilizing the generated reports and insights to inform and make policy decisions. The relationships between these classes, indicated by "uses," "manages," and "interacts," depict the flow of data and control within the system, highlighting the dependencies and responsibilities of each component. This diagram provides a clear visual representation of the system's structure, facilitating understanding and development.



## 5.2.2 Behavioral Diagram

Any System can have two aspects, static and dynamic. So, a model is considered as complete when both the aspects are fully covered. Behavioral diagrams basically capture the dynamic aspects of a system. UML has the following five types of behavioral diagrams.

- a. Use case diagram
- b. Sequence diagram
- c. Collaboration diagram
- d. Statechart diagram
- e. Activity diagram

### 5.2.2.1 Use Case diagram

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The

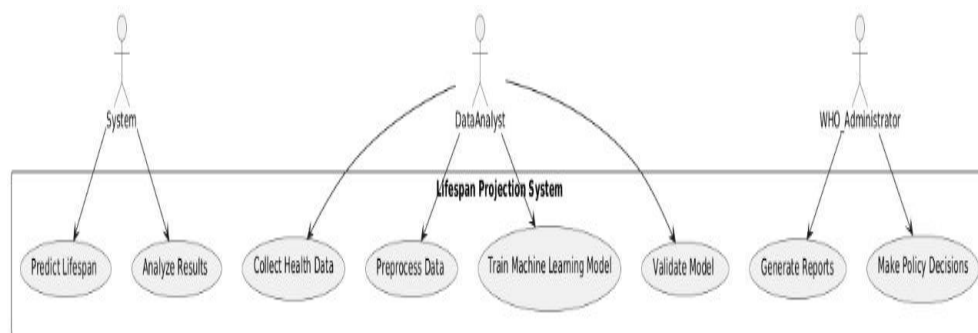


Figure 4 Usecase diagram

### 5.2.2.2 Sequence diagram

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

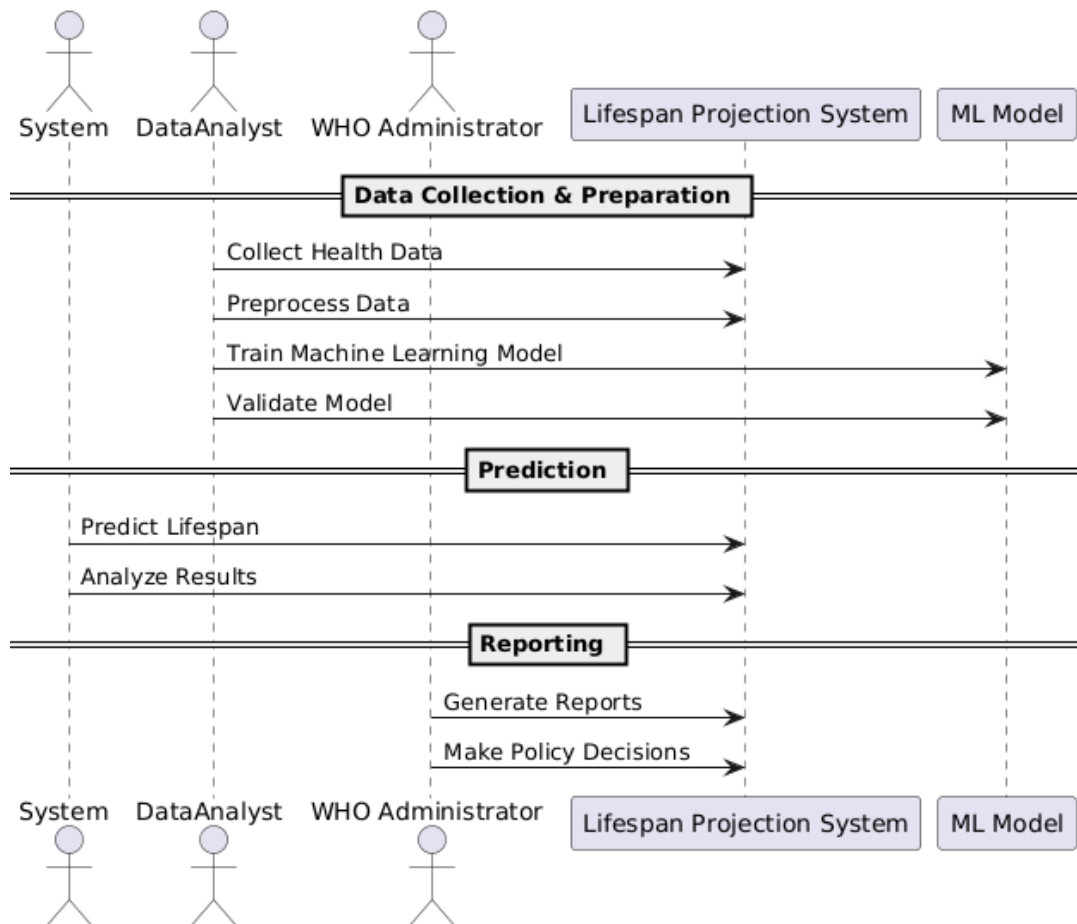


Figure 5 Sequence diagram

### 5.2.2.3 Activity Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational stepbystep workflows of components in a system. An activity diagram shows the overall process.

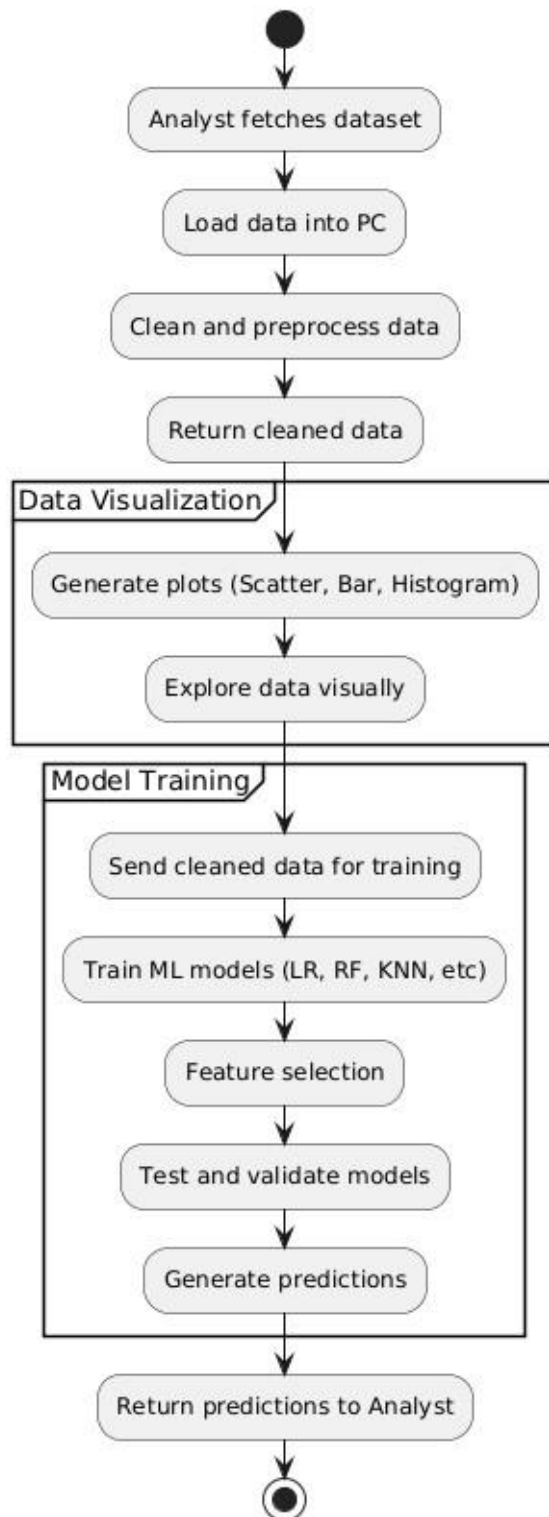


Figure 6 Activity daigram

## **CHAPTER 6**

# **RESULTS AND DISCUSSIONS**

# RESULTS AND DISCUSSIONS

Results section is to present the data and findings of your research in a clear, concise and unbiased manner. Discussion section is to interpret the results, discuss their implications and connect them to existing research.

## 6.1 Dataset

Dataset is a structured collection of related data organized for analysis, typically stored in tables, arrays or specific formats like CSV or JSON and used for tasks like data analysis, machine learning and AI.

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.259210
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.696514
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.744976
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.959000
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.537231
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2933	Zimbabwe	2004	Developing	44.3	723.0	27	4.36	0.000000	68.0	31	...	67.0	7.13	65.0	33.6	454.366654
2934	Zimbabwe	2003	Developing	44.5	715.0	26	4.06	0.000000	7.0	998	...	7.0	6.52	68.0	36.7	453.351155
2935	Zimbabwe	2002	Developing	44.8	73.0	25	4.43	0.000000	73.0	304	...	73.0	6.53	71.0	39.8	57.348340
2936	Zimbabwe	2001	Developing	45.3	686.0	25	1.72	0.000000	76.0	529	...	76.0	6.16	75.0	42.1	548.587312
2937	Zimbabwe	2000	Developing	46.0	665.0	24	1.68	0.000000	79.0	1483	...	78.0	7.10	78.0	43.5	547.358878

2938 rows × 22 columns

Figure 7 Dataset

## 6.2 Pre Processing

Pre Processing transforms raw data into suitable format for analysis and model training, addressing issues like missing values, inconsistencies and noise ultimately improving model performance

```
df['Country'].unique()

1
array(['Afghanistan', 'Albania', 'Algeria', 'Angola',
      'Antigua and Barbuda', 'Argentina', 'Armenia', 'Australia',
      'Austria', 'Azerbaijan', 'Bahamas', 'Bahrain', 'Bangladesh',
      'Barbados', 'Belarus', 'Belgium', 'Belize', 'Benin', 'Bhutan',
      'Bolivia (Plurinational State of)', 'Bosnia and Herzegovina',
      'Botswana', 'Brazil', 'Brunei Darussalam', 'Bulgaria',
      'Burkina Faso', 'Burundi', "Côte d'Ivoire", 'Cabo Verde',
      'Cambodia', 'Cameroon', 'Canada', 'Central African Republic',
      'Chad', 'Chile', 'China', 'Colombia', 'Comoros', 'Congo',
      'Cook Islands', 'Costa Rica', 'Croatia', 'Cuba', 'Cyprus',
      'Czechia', "Democratic People's Republic of Korea",
      'Democratic Republic of the Congo', 'Denmark', 'Djibouti',
      'Dominica', 'Dominican Republic', 'Ecuador', 'Egypt',
      'El Salvador', 'Equatorial Guinea', 'Eritrea', 'Estonia',
      'Ethiopia', 'Fiji', 'Finland', 'France', 'Gabon', 'Gambia',
      'Georgia', 'Germany', 'Ghana', 'Greece', 'Grenada', 'Guatemala',
      'Guinea', 'Guinea-Bissau', 'Guyana', 'Haiti', 'Honduras',
      'Hungary', 'Iceland', 'India', 'Indonesia',
      'Iran (Islamic Republic of)', 'Iraq', 'Ireland', 'Israel', 'Italy',
      'Jamaica', 'Japan', 'Jordan', 'Kazakhstan', 'Kenya', 'Kiribati',
      'Kuwait', 'Kyrgyzstan', "Lao People's Democratic Republic",
      'Latvia', 'Lebanon', 'Lesotho', 'Liberia', 'Libya', 'Lithuania',
      'Luxembourg', 'Madagascar', 'Malawi', 'Malaysia', 'Maldives',
      'Mali', 'Malta', 'Marshall Islands', 'Mauritania', 'Mauritius',
      'Mexico', 'Micronesia (Federated States of)', 'Monaco', 'Mongolia',
      ...
      'United Kingdom of Great Britain and Northern Ireland',
      'United Republic of Tanzania', 'United States of America',
      'Uruguay', 'Uzbekistan', 'Vanuatu',
      'Venezuela (Bolivarian Republic of)', 'Viet Nam', 'Yemen',
      'Zambia', 'Zimbabwe'], dtype=object)
```

Figure 8 Data Pre Processing

### 6.2.1 Splitting the dataset

The process involves splitting a dataset into two subsets. The first subset is called the training data set and is used to fit the model. The input element of the second subset is provided to the model, then predictions are made and compared to expected values. This second data set is called a test data set.

```
from sklearn.model_selection import train_test_split
df_train, df_test = train_test_split(df, test_size=0.2, random_state=42)
```

Figure 9 Splitting data

## 6.3 Data Visualization

Data visualization is a critical step in the machine learning pipeline, as it helps uncover patterns, trends, and relationships within the dataset. Before model training begins, visual tools like histograms, scatter plots, and box plots allow data scientists to better understand the distribution and structure of the data. This process is often referred to as exploratory data analysis (EDA). Visualization can highlight class imbalances, missing values, or outliers that may affect model performance, and guide the selection of preprocessing techniques such as normalization or encoding.

In the feature selection and engineering phase, visualizations like heatmaps, correlation matrices, and pair plots help identify dependencies between variables. For instance, a strong correlation between two features might suggest multicollinearity, which can influence linear models. Dimensionality reduction techniques like PCA (Principal Component Analysis) and t-SNE are also visualized in 2D or 3D space to understand how well features separate different classes or clusters. These visual insights assist in crafting more effective input features for ML models.

After model training, visualization plays a key role in evaluating model performance. Tools like confusion matrices, ROC curves, and precision-recall curves are used to assess classification models, while learning curves help diagnose underfitting or overfitting. Feature importance charts, SHAP (SHapley Additive exPlanations), and LIME (Local Interpretable Model-agnostic Explanations) plots are valuable for understanding which features influence predictions the most. Overall, data visualization not only enhances model interpretability but also aids in making data-driven decisions with confidence.

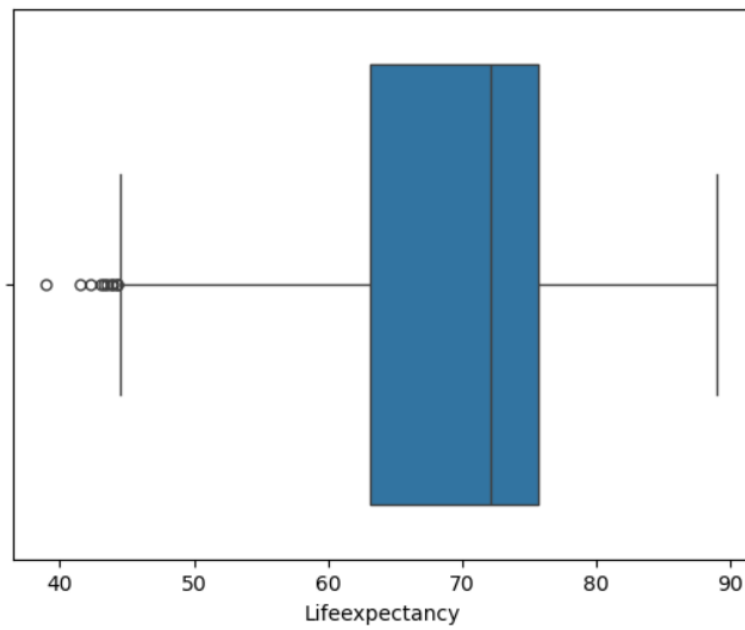


Figure 10 Visualization of target

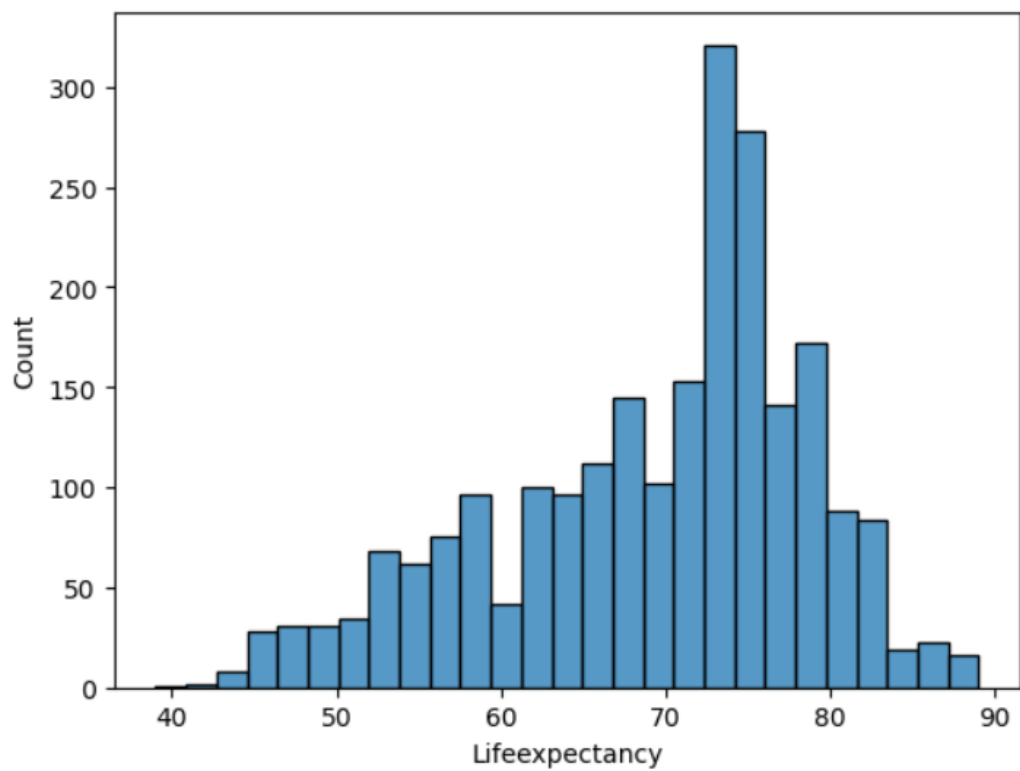


Figure 11 Visualization of lifespan



## Plot the Numerical columns

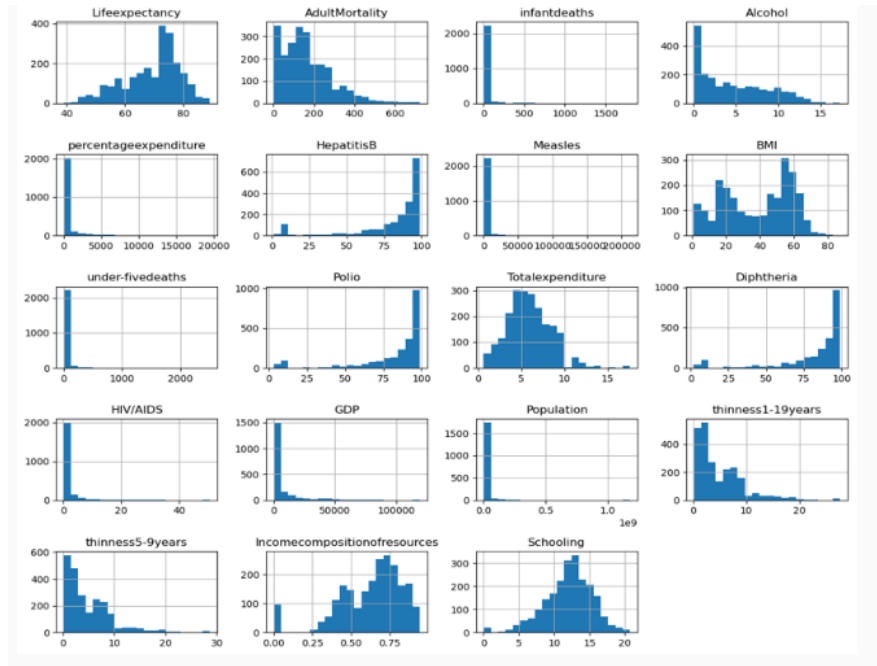
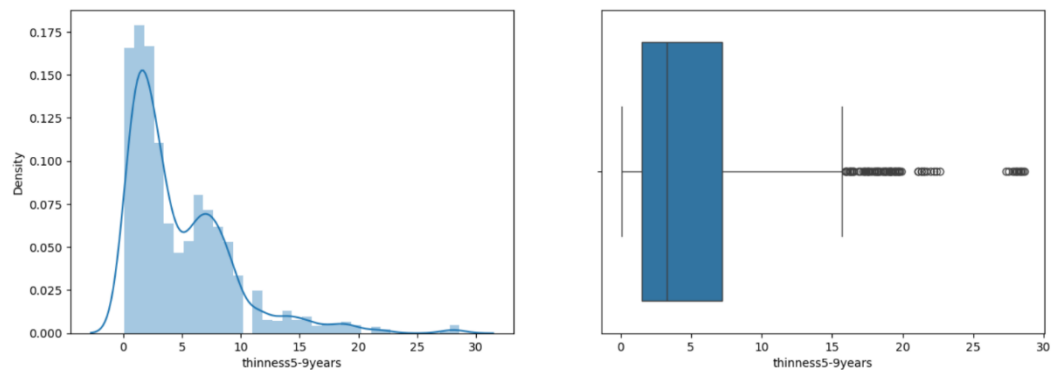
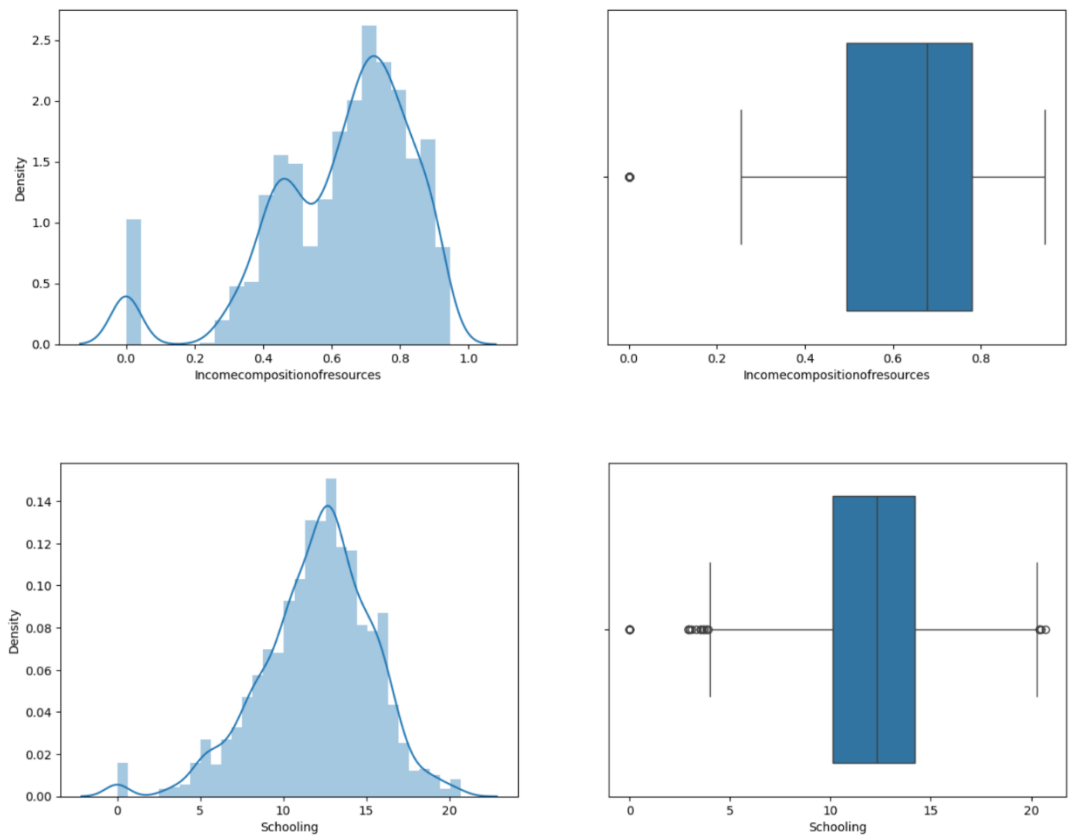


Figure 12 Plotting the numerical values

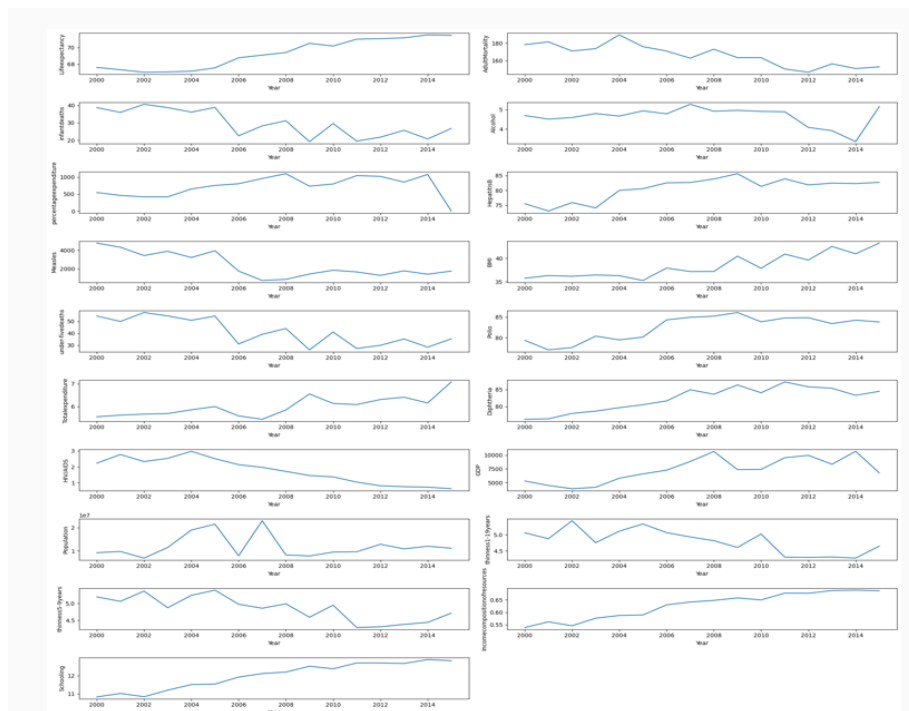
## Univariate Analysis



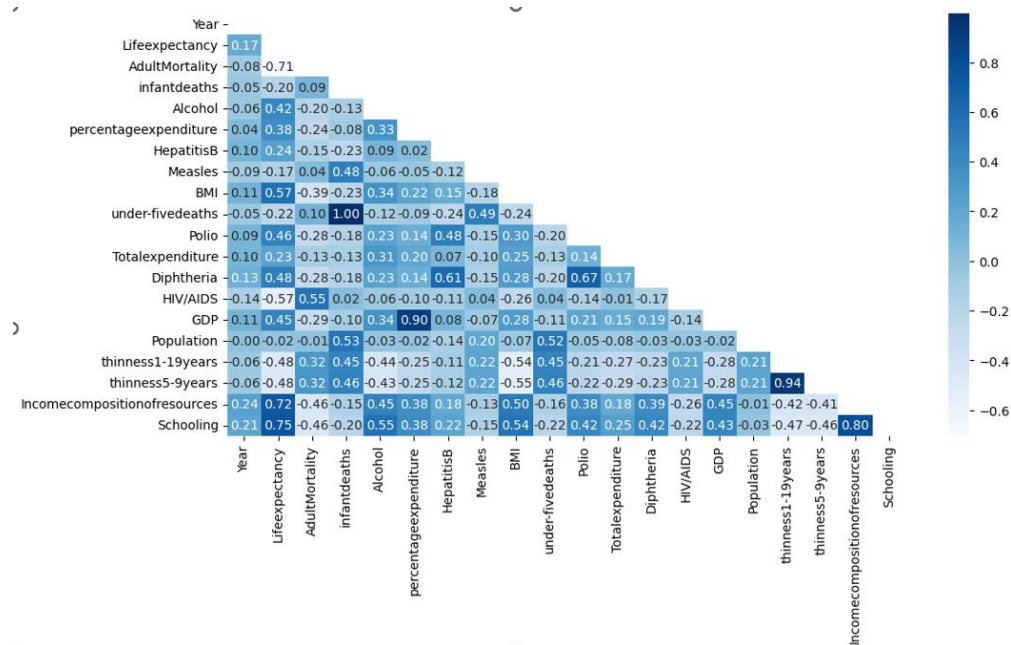
## Lifespan Projection Model using Machine Learning Techniques



## Numerical Columns over the years



### Collinearity plot



## 6.4 Feature Importance

Based on the various machine learning techniques, we are extracting features which are more effected on the model.

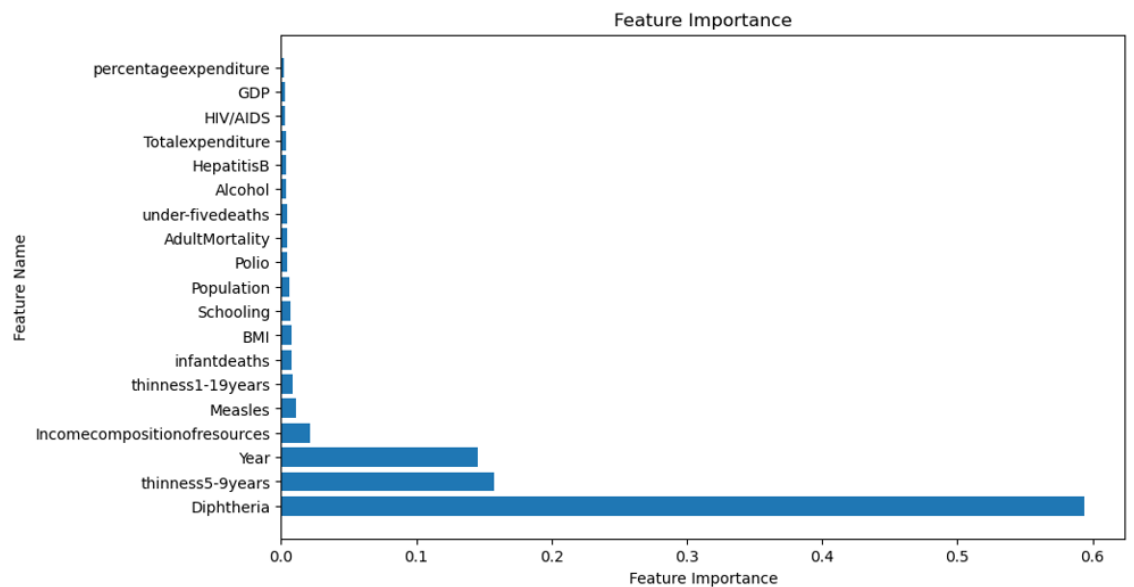


Figure 13 Feature Importance

By above figure we are observe that,

Top 3 Most Important Features:

1. Diphtheria ( $\approx 0.6$ ): Hugely influential in the model. This means diphtheria immunization rates are very predictive of the target variable (likely something related to health or life expectancy).
2. thinness5-9years ( $\approx 0.15$ ): Indicates a strong link between child thinness and the outcome.
3. Year ( $\approx 0.15$ ): Suggests there's a strong temporal trend in the data, i.e., outcomes change significantly over time.

Moderately Important:

Income composition of resources and Measles have small but noticeable contributions.

Least Important:

Variables like percentage expenditure, GDP, HIV/AIDS, HepatitisB, Alcohol, etc., have very low importance in this model. They barely affect the predictions.

## 6.5 Result Analysis

In the Result analysis we can observe the  $R^2$  score, Cross validation and RMSE error.

$R^2$  score:

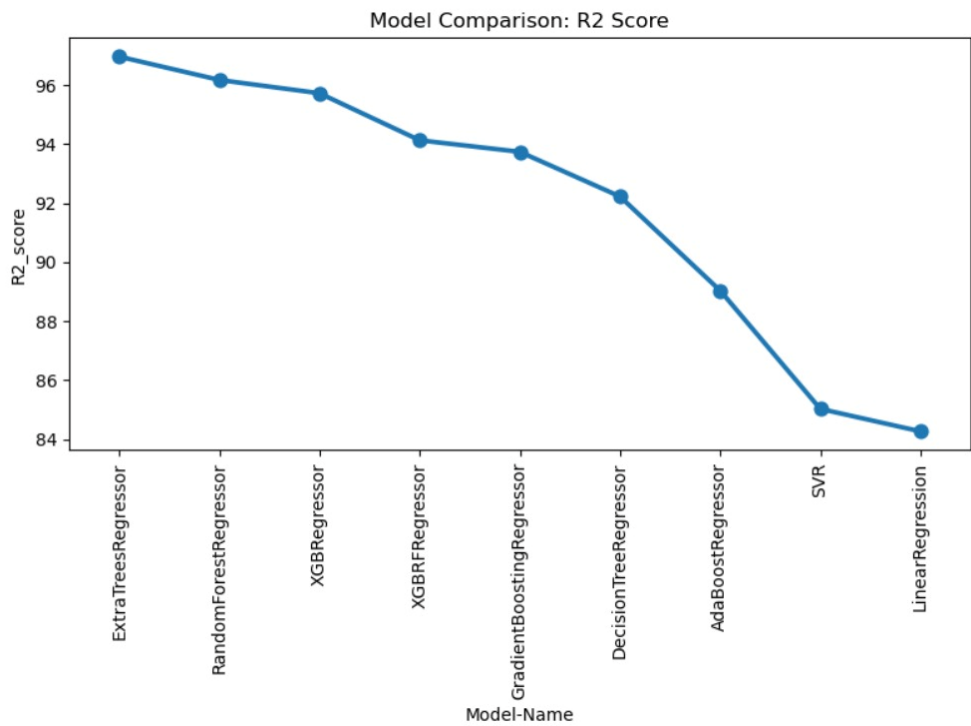


Figure 14 Result of  $r^2$  score

This figure presents a comparison of various regression models based on their  $R^2$  scores, which indicate how well each model explains the variance in the target variable. Among all the models evaluated, the **ExtraTreesRegressor** stands out with the highest  $R^2$  score, suggesting it offers the best predictive performance on the dataset. It is closely followed by the **RandomForestRegressor**, **XGBRegressor**, and **XGBRFRegressor**, all of which maintain strong accuracy levels, reflecting the power of ensemble tree-based methods. Performance begins to decline more noticeably with **GradientBoostingRegressor** and continues downward with **DecisionTreeRegressor** and **AdaBoostRegressor**. Notably, **Support Vector Regressor (SVR)** and **Linear Regression** perform the worst, with significantly lower  $R^2$  scores around 85, indicating they are less suited for capturing the complexity of the data. Overall, this comparison highlights that ensemble methods, particularly those using randomized trees, are most effective for this regression task.

### Cross Validation:

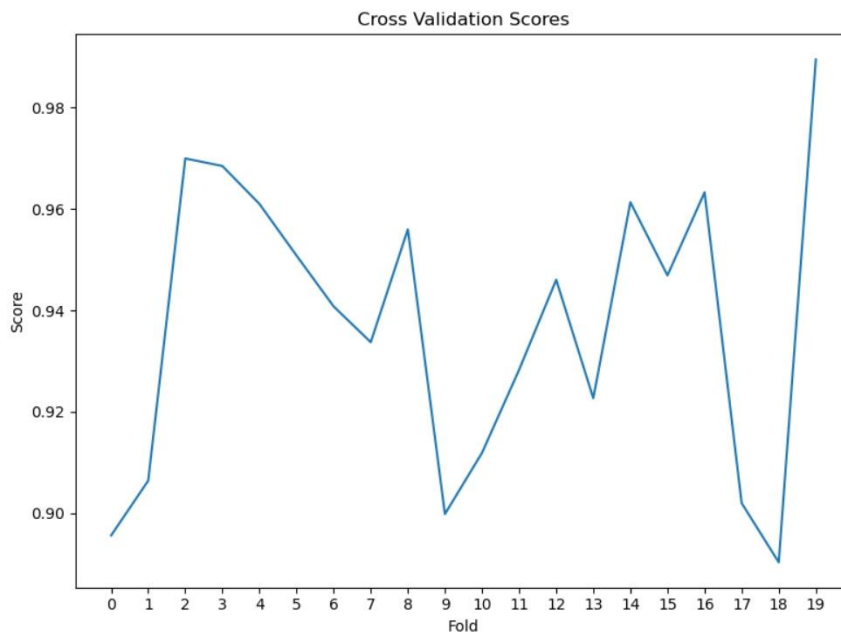


Figure 15 Cross validation

This line chart displays the cross-validation scores for 20 different folds of a machine learning model. Each point on the x-axis represents a fold, while the y-axis indicates the corresponding  $R^2$  score achieved in that fold. The scores mostly remain high, ranging from approximately 0.89 to just above 0.98, indicating that the model consistently performs well across different subsets of the data. However, there are some fluctuations, particularly in folds 9 and 18, where the scores dip closer to 0.90, suggesting a bit of variance in model performance depending on the data split. Despite these dips, the overall pattern reflects strong generalization capability, with minimal overfitting or underfitting, showcasing the model's robustness and reliability across multiple training and validation sets.

RMSE Error:

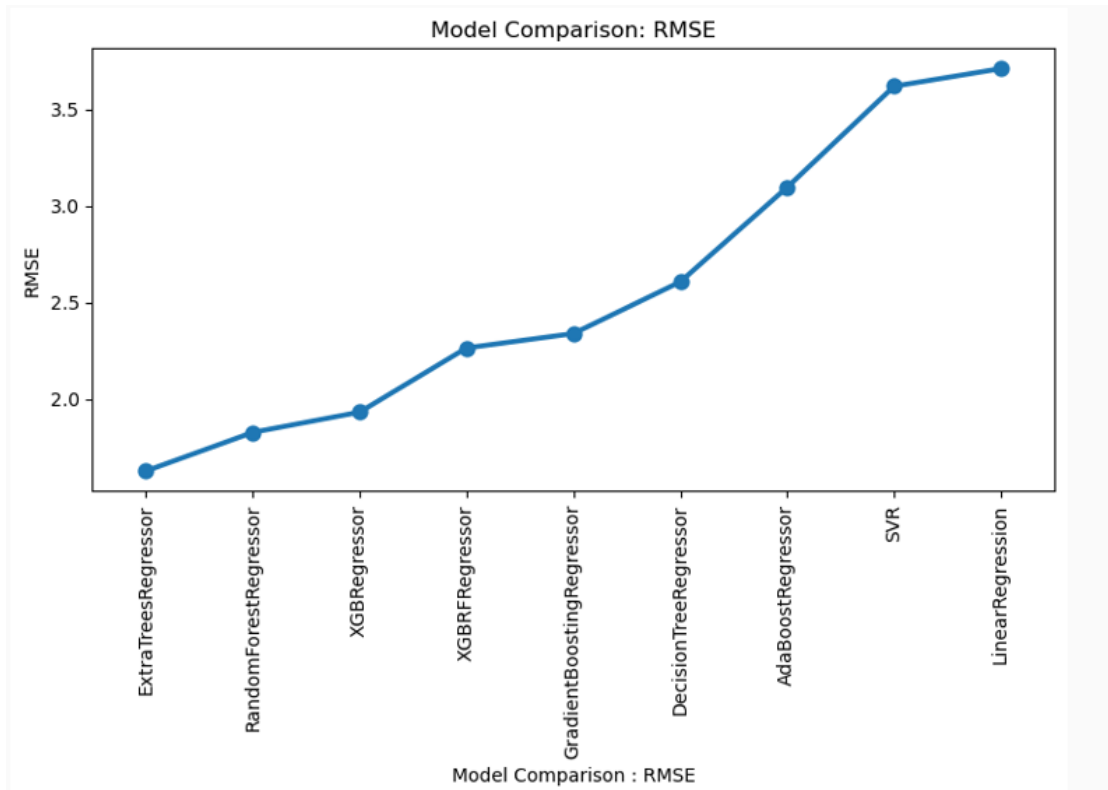


Figure 16 RMSE Error

This chart compares the performance of various regression models based on their Root Mean Squared Error (RMSE) values. RMSE is a metric used to measure the average magnitude of the errors between predicted and actual values, with lower values indicating better model performance. In this plot, the ExtraTreesRegressor achieves the lowest RMSE, making it the most accurate model among those tested. It is followed closely by the RandomForestRegressor, XGBRegressor, and XGBRFRegressor, which also maintain relatively low error levels. On the other end of the spectrum, Linear Regression and Support Vector Regressor (SVR) exhibit the highest RMSE values, indicating that these models make larger prediction errors and are less suitable for this particular task. Overall, the chart reinforces the conclusion that ensemble-based tree models outperform traditional linear and kernel-based methods in this regression scenario, both in terms of error minimization and reliability.

## 6.6 Input & Output


Inputs can come in various forms such as text, numbers, images, or sensor readings, depending on the problem being solved.

Input:

These features are both numerical and categorical in nature, covering a wide range of health, economic, environmental, and social indicators. Key input variables include GDP per capita, adult mortality rates, average BMI, alcohol consumption, years of schooling, immunization rates (like measles and DPT), percentage of population with HIV/AIDS, and healthcare expenditure, among others. Each row in the dataset typically represents a country-year combination, providing a snapshot of its socio-health profile. This raw input data may also contain missing values, inconsistencies, or outliers, which need to be addressed during the preprocessing phase. In some cases, user-defined inputs can also be provided—for example, entering specific values for GDP, schooling, and health expenditure—to generate a life expectancy prediction. Overall, the quality and diversity of the input data play a crucial role in building an accurate and generalizable predictive model.


GUI takes input from the user which includes name of the country, status of the country, Adult Mortality Rate, Infant Mortality Rate, Year, Alcohol, thinness 1-19 years, thinness 5-19 years, HepatitisB, measles, Schooling etc.,



 Lifespan Projection Model

Enter the Country:	<input type="text"/>
Enter Status of the Country:	<input type="text"/>
Enter Year:	<input type="text"/>
Enter Adult Mortality(in %):	<input type="text"/>
Enter Infant Deaths(in %):	<input type="text"/>
Enter Alcohol(in %):	<input type="text"/>
Enter Percentage Expenditure:(in numbers)	<input type="text"/>
Enter HepatitisB(in %)	<input type="text"/>
Enter Measles(in %):	<input type="text"/>
Enter BMI(in %):	<input type="text"/>
Enter Under-five Deaths(in %):	<input type="text"/>
Enter Polio(in %):	<input type="text"/>
Enter Total Expenditure(in %):	<input type="text"/>
Enter Diphtheria(in %):	<input type="text"/>
Enter HIV/AIDS(in %):	<input type="text"/>
Enter GDP(in %):	<input type="text"/>
Enter Population(in numbers):	<input type="text"/>
Enter Thinness 1-19years(in %):	<input type="text"/>
Enter Thinness 5-9 years(in %):	<input type="text"/>
Enter Incomecompositionofresources(in %):	<input type="text"/>
Enter Schooling(in %):	<input type="text"/>

Figure 17 Input screen

 Lifespan Projection Model

Enter the Country:	<input type="text" value="india"/>
Enter Status of the Country:	<input type="text" value="developing"/>
Enter Year:	<input type="text" value="2023"/>
Enter Adult Mortality(in %):	<input type="text" value="12"/>
Enter Infant Deaths(in %):	<input type="text" value="22"/>
Enter Alcohol(in %):	<input type="text" value="12"/>
Enter Percentage Expenditure:(in numbers)	<input type="text" value="3"/>
Enter HepatitisB(in %)	<input type="text" value="34"/>
Enter Measles(in %):	<input type="text" value="12"/>
Enter BMI(in %):	<input type="text" value="12"/>
Enter Under-five Deaths(in %):	<input type="text" value="12"/>
Enter Polio(in %):	<input type="text" value="34"/>
Enter Total Expenditure(in %):	<input type="text" value="34"/>
Enter Diphteria(in %):	<input type="text" value="12"/>
Enter HIV/AIDS(in %):	<input type="text" value="3"/>
Enter GDP(in %):	<input type="text" value="123"/>
Enter Population(in numbers):	<input type="text" value="250000"/>
Enter Thinness 1-19years(in %):	<input type="text" value="3"/>
Enter Thinness 5-9 years(in %):	<input type="text" value="1"/>
Enter Incomecompositionofresources(in %):	<input type="text" value="3"/>
Enter Schooling(in %):	<input type="text" value="23"/>
<input type="button" value="Submit"/>	<input type="button" value="Reset"/>

Figure 18 Input screen with inputs

Output:

Output is the result or response produced by a system after processing the input. In a machine learning context, the output is typically the predicted value or classification generated by the model based on the input data. Outputs provide meaningful insights or decisions derived from the analysis.

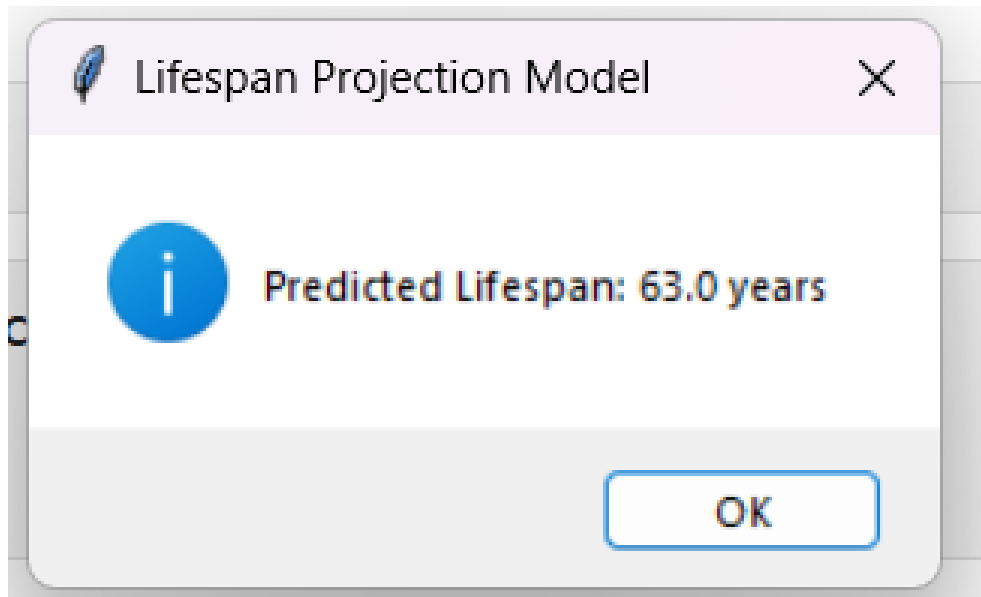


Figure 19 Output screen for given input

If any one of the field is not given in input form, it will shows that “All fields required”

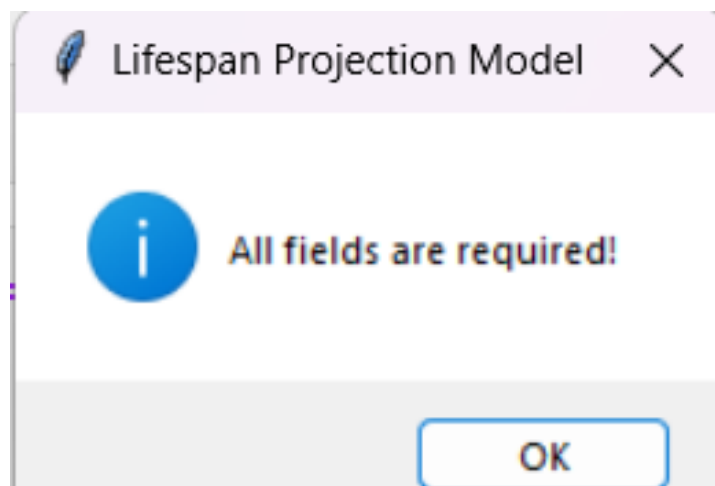


Figure 20 Output screen when invalid input given

## **CHAPTER 7**

# **CONCLUSION AND FUTURE SCOPE**

## CONCLUSION AND FUTURE SCOPE

The Conclusion summarizes the key findings, results, and insights gained from the project or research work. It reflects on how the objectives were achieved, what the outcomes indicate, and the overall contribution of the work. A strong conclusion ties everything together, emphasizing the significance and impact of the study or system developed.

The Future Scope outlines potential directions for extending or improving the current work. It highlights areas where further research, development, or experimentation could be conducted, considering limitations faced or emerging trends. This section opens the door for innovation and continued progress related to the topic.

### 7.1 Conclusion

In this project, the primary goal was to identify the most effective regression model for predicting LifeSpan based on various health and demographic indicators. Initially, features such as year, country, and status were excluded to focus on the intrinsic impact of the remaining variables. Among all tested models, the **ExtraTreesRegressor** outperformed the others, achieving the lowest RMSE and the highest  $R^2$  score, closely followed by the **RandomForestRegressor**, which achieved an impressive  $R^2$  score of 96% and an MAE of 1.27 on the test set. Feature importance analysis revealed that **Diphtheria**, **thinness (5–9 years)**, and **year** were the most influential variables, with **schooling**, **BMI**, and **income composition of resources** showing a strong positive correlation with LifeSpan. Contrary to initial expectations, variables like **GDP**, **total expenditure**, and **infant deaths** contributed minimally to model predictions. These findings highlight the significant roles of health and education over traditional economic metrics in determining LifeSpan. While the model performs well, future enhancements could involve incorporating additional environmental and geographical features to further improve accuracy and generalizability.

## 7.2 Limitations

Limitations refer to the constraints, shortcomings, or restrictions encountered in a study or project that may affect the accuracy, generalizability, or applicability of the results. These could arise from data quality, model selection, limited features, methodological constraints, or other practical challenges. Acknowledging limitations helps in understanding the boundaries within which the conclusions are valid.

1. **Limited Feature Set:** Although the current model incorporates a variety of health, education, and economic indicators, it still lacks environmental and geographical factors such as climate trends and natural disasters, which could influence LifeSpan.
2. **Data Availability:** Historical datasets were often sparse and lacked comprehensive variables, which restricted the accuracy of early LifeSpan prediction models.
3. **Model Interpretability:** While ensemble and deep learning methods improve accuracy, they often act as "black boxes," making it difficult to interpret the reasoning behind predictions.
4. **Uncertainty Handling:** Despite advancements, effectively managing uncertainty in LifeSpan datasets due to missing, noisy, or unpredictable variable remains an unsolved challenge.
5. **Generalization Across Populations:** Models trained on specific datasets may not generalize well across diverse countries or regions with varying socio-economic and healthcare conditions.

## 7.3 Future Scope

Future Scope outlines the potential directions for further research or improvements in a project. It highlights areas where the current work can be extended, refined, or enhanced, often by overcoming existing limitations, adopting new techniques, incorporating additional data, or applying the model in broader contexts. It serves as a roadmap for continuous development and exploration.

The scope for future enhancement of this LifeSpan prediction project is substantial. While the current model demonstrates high accuracy using health, education, and economic indicators, future work can expand by integrating environmental and geographical variables such as air quality, climate patterns, and natural disaster frequency, which are known to impact longevity. Additionally, the application of deep learning techniques like Artificial Neural Networks (ANNs) and Recurrent Neural Networks (RNNs) can uncover more complex, non-linear relationships within the data. Another significant direction is improving model transparency through explainable AI (XAI) methods, which can help stakeholders understand the driving factors behind predictions. The adoption of more sophisticated ensemble learning methods, along with enhanced cross-validation frameworks, could also improve the robustness and generalizability of the models across global populations. Ultimately, the goal is to develop a scalable and interpretable system that not only predicts LifeSpan with higher accuracy but also guides policy-making by identifying the most critical determinants of human longevity.

## REFERENCES

1. Abhinaya. V, Dharani. B. C, Vandana. A, Dr. Velvadivu. P, Dr. Sathya. C.,
2. “Statistical Analysis On Factors Influencing LifeSpan”, Coimbatore, India, eISSN: 2395-0056, July 2021.
3. Jessica Y Ho, Arun S Hendi, “Recent trends in LifeSpan across high income countries: retrospective observational study”, USA, 10.1136/bmj.k2562, 15 August, 2018.
4. David Strauss, Lewis Rosenbloom, Jordan Brooks, Robert Shavelle, “LifeSpan in cerebral palsy: an update”, USA, DOI: 10.1111/j.1469- 8749.2008.03000.
5. Michael J DeVivo, Gordana Savic, Hans L Frankel, Bakulesh M Soni, “Comparison of statistical methods for calculating LifeSpan after spinal cord injury”, 10.1038/s41393-018-0067-1, February 2018.
6. Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V Sharma, “Machine Learning For Prognosis of LifeSpan and Diseases”, ISSN: 2278-3075, August 2019.
7. Li, Y., Liang, Y., Liu, R., Li, Y., & Li, Y. (2020). Analysis of factors influencing the LifeSpan of the elderly based on machine learning. PeerJ, 8, e8365. doi: 10.7717/peerj.8365
8. Kim S., Lee S. Y., “Effects of air pollution on LifeSpan in South Korea”, Environmental Science and Pollution Research, vol. 28, no. 10, pp. 12603-12613, 2021.
9. Lassale C., “Dietary patterns and LifeSpan”, Current Opinion in Clinical Nutrition and Metabolic Care, vol. 23, no. 4, pp. 239-244, 2020.
10. Li J., Chen J., “Gender, family support, and LifeSpan: a cross-national analysis of older adults”, BMC Public Health, vol. 21, no. 1, pp. 1-10, 2021.



11. Chauhan P., “LifeSpan and GDP per capita: A Comparative Analysis of India and China”, *International Journal of Scientific Research*, vol. 10, no. 2, pp. 6165, 2021.
12. Doshi R., Patel P., Shah B., “Estimation of LifeSpan at Birth for Indian States and Union Territories by a Bayesian Hierarchical Model”, *Journal of Health Management*, vol. 22, no. 2, pp. 201-215, 2020.
13. Huang Y., Chen S., Zhang X., Wu J., “Effects of education on LifeSpan in China: Evidence from the China Family Panel Studies”, *PLoS One*, vol. 15, no. 3, pp. 1-12, 2020
14. Jiang Y., Li Y., Li X., Ma Z., Wang R., Li J., “The Impact of Health Insurance on LifeSpan: Evidence from China’s New Rural Cooperative Medical Scheme”, *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, pp.1-14, 2020.
15. Alkhamis A. A., “Assessment of LifeSpan and mortality rates using the LeeCarter model”, *Applied Economics*, vol. 53, no. 3, pp. 358-373, 2021.
16. Li Q., Li X., Li J., Li M., Cui Y., “Analysis of the Effects of Alcohol Consumption on LifeSpan Based on a Markov Model”, *Risk Analysis*, vol. 40, no. 3, pp. 542555, 2020.
17. Patel K. J., Gajera H. P., Parmar P. N., “A Study of Factors Affecting LifeSpan in India”, *Journal of Critical Reviews*, vol. 7, no. 2, pp. 119-124, 2020.
18. Cao X., “The Effect of Occupational Health and Safety on LifeSpan: Evidence from China’s Coal Mining Industry”, *International Journal of Environmental Research and Public Health*, vol. 18, no. 1, pp. 1-12, 2021.
19. Hashim K., “Effects of Environmental Pollution on LifeSpan in Malaysia”, *Environmental Science and Pollution Research*, vol. 28, no. 9, pp. 11624-11630, 2021.

20. Liu Y., Zhou H., “Impacts of air pollution on LifeSpan in China”, *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, pp. 1-14, 2020.
21. Saleem N., Ahmad N., Shaikh T. A., Khan M. S., “Analysis of the effect of health expenditure on LifeSpan in Pakistan using ARDL”, *BMC Health Services Research*, vol. 21, no. 1, pp. 1-10, 2021.
22. Wu Y., Hou G., Liu W., “The effects of medical insurance on LifeSpan: a study of China's basic medical insurance”, *International Journal of Health Economics and Management*, vol. 21, no. 1, pp. 47-62, 2021.
23. Zhao L., Zhu L., Lu Z., Qin L., Huang Y., Zhang J., “Analysis of Factors Influencing LifeSpan Based on Big Data”, *Journal of Healthcare Engineering*, vol. 2020, pp. 1-11, 2020.
24. Lin H., Li H., Li X., “Socioeconomic Development and LifeSpan in Developing Countries”, *Social Indicators Research*, vol. 150, no. 1, pp. 219-239, 2020.
25. Liu M., Zhou C., Li J., Li Y., Li X., Wang Y., “Determinants of LifeSpan at birth in China: A Bayesian model averaging approach”, *Journal of Health Economics*, vol. 79, pp. 1-11, 2021.
26. Nguyen T. H., Vu T. K., Nguyen H. L., Vu H. T., “The impact of health expenditures on LifeSpan in ASEAN-5 countries: A dynamic panel data approach”, *Journal of Health Economics and Outcomes Research*, vol. 8, no. 2, pp. 153-164, 2020.
27. Samik-Ibrahim I., Abuhayat A. M., Oyefuga O. H., “Factors affecting LifeSpan in Africa: A review of literature”, *Archives of Medicine and Health Sciences*, vol. 9, no. 2, pp. 313-318, 2021.
28. Sun R., Zhang Q., Liu W., Wang Y., “LifeSpan, Healthy LifeSpan, and Healthcare Expenditure at Birth and at Age 60 in China: A Markov Model”, *Risk Analysis*, vol. 41, no. 4, pp. 719-733, 2021.

29. Trung H. V., Tuan A. V., Hoang L. T., “LifeSpan and health status of older people in Vietnam: Evidence from a longitudinal survey”, BMC Public Health, vol. 21, no. 1, pp. 1-13, 2021.
30. Wang C., Luo L., Huang Y., Hu Y., Wang C., “Regional disparities and determinants of LifeSpan at birth in China, 2000-2015: A spatiotemporal analysis”, PLoS One, vol. 15, no. 11, pp. 1-15, 2020.
31. Wang J., Jia S., “How does pollution impact LifeSpan in China? Evidence from prefecture-level cities”, Environmental Science and Pollution Research, vol. 28, no. 15, pp. 19456-19465, 2021.