Report on-

***LIFE EXPECTANCY CALCULATION USING MACHINE LEARNING***

By- Chhavi

**TABLE OF CONTENTS**

C

H

A

P

T

E

R

1

~~:~~

I

N

T

R

O

D

U

C

T

I

O

N

#

E

C

T

I

O

N

1

1

V

E

R

V

I

E

W

#

E

C

T

I

O

N

.

2

P

U

R

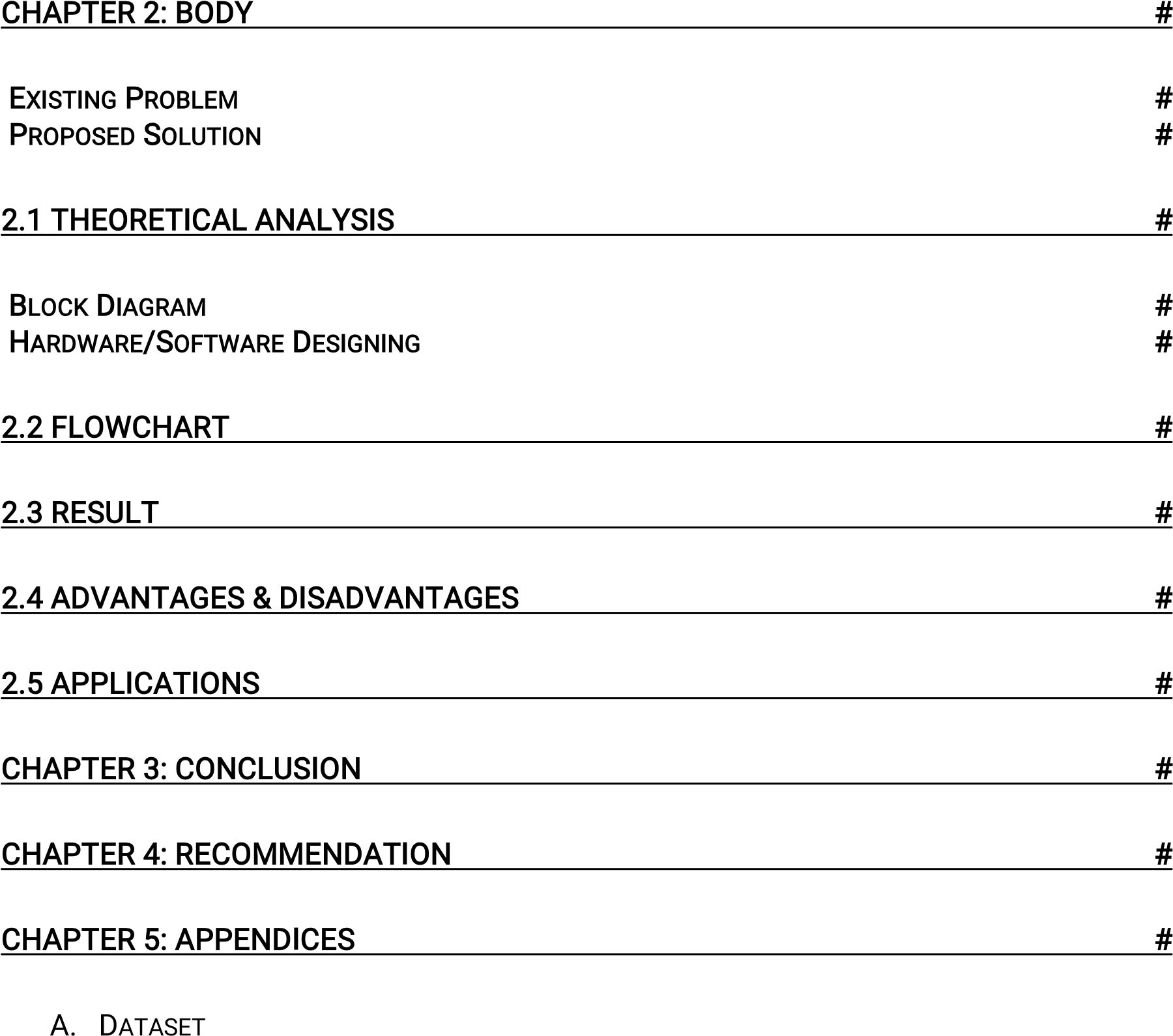
P

O

S

E

#



Life expectancy prediction using machine learning is the process of using algorithms and statistical models to analyse various factors that influence a person's lifespan and predict how long they are likely to live. These factors include demographic data, medical history, lifestyle choices, and environmental factors. The goal of life expectancy prediction is to improve healthcare and assist in making informed decisions related to public health policy, individual health, and financial planning. Machine learning models such as Random Forest, Support Vector Machines (SVMs), and Neural Networks are commonly used to predict life expectancy based on a wide range of input features. However, it's important to note that these models may not be accurate for all individuals and should be used in combination with expert medical advice.

# **1. INTRODUCTION**

# 1.1 OVERVIEW:

# Human an incredible creation of god. Every creature in the world has a limited life span, to achieve something in the world. We have a limited life span to survive in the current 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. Investigations about the life span of vertebrates have been made, except the human (HOMO SAPIENS). It is an emerging research area that is gaining interest but involved lot of challenges due to the limited amount of resources (i.e., datasets) available.

# In our proposed system the life span of human is predicted by analysis of human. By obtaining the Environmental factors, Food habits, Diseases and Medical history, a lot of investigations will be conducted to predict the sustainability of human. By the machine learning algorithms and data analytics, We can prognosticate and examine the life span of the individual human being and we can use different classification algorithms for this prediction to accomplish higher accuracy. Here, we use python and machine learning in order to calculate life expectancy of human beings.

## 1.2 PURPOSE

The purpose of building a life expectancy prediction model using machine learning is to provide insights into factors that influence life expectancy and to predict an individual's life expectancy based on their demographic, lifestyle, and health factors. This can be useful for medical professionals, insurance companies, and individuals to better understand health risks and make informed decisions about lifestyle changes, preventive measures, and medical treatments. Additionally, governments and public health organizations can use life expectancy prediction models to identify population groups that are at higher risk for premature mortality and to design targeted interventions to improve public health outcomes.

# EXISTING PROBLEM

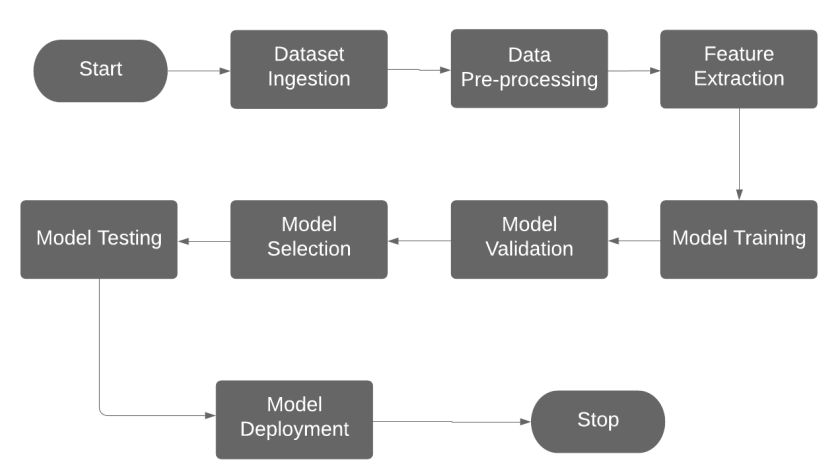
Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. Life expectancy is a measure that is often used to gauge the overall health of a community. Life expectancy at birth measures health status across all age groups. Shifts in life expectancy are often used to describe trends in mortality.

# PROPOSED SOLUTION

To design a machine learning model using regression techniques that can predict the life expectancy of a country given the various factors that affect the health of its citizens. Using this model, we can assess the extent of effect these factors have on the average life expectancy. This can give us an idea about the changes that need to be brought about in order to improve the health conditions of people around the world.

2.1 **THEORETICAL ANALYSIS**

# BLOCK DIAGRAM



HARDWARE/SOFTWARE DESIGNING

Hardware Requirements:

Any working laptop/PC with minimum 2.2 GHz processor and at least 8GB of memory with an Internet connection.

Software Requirements:

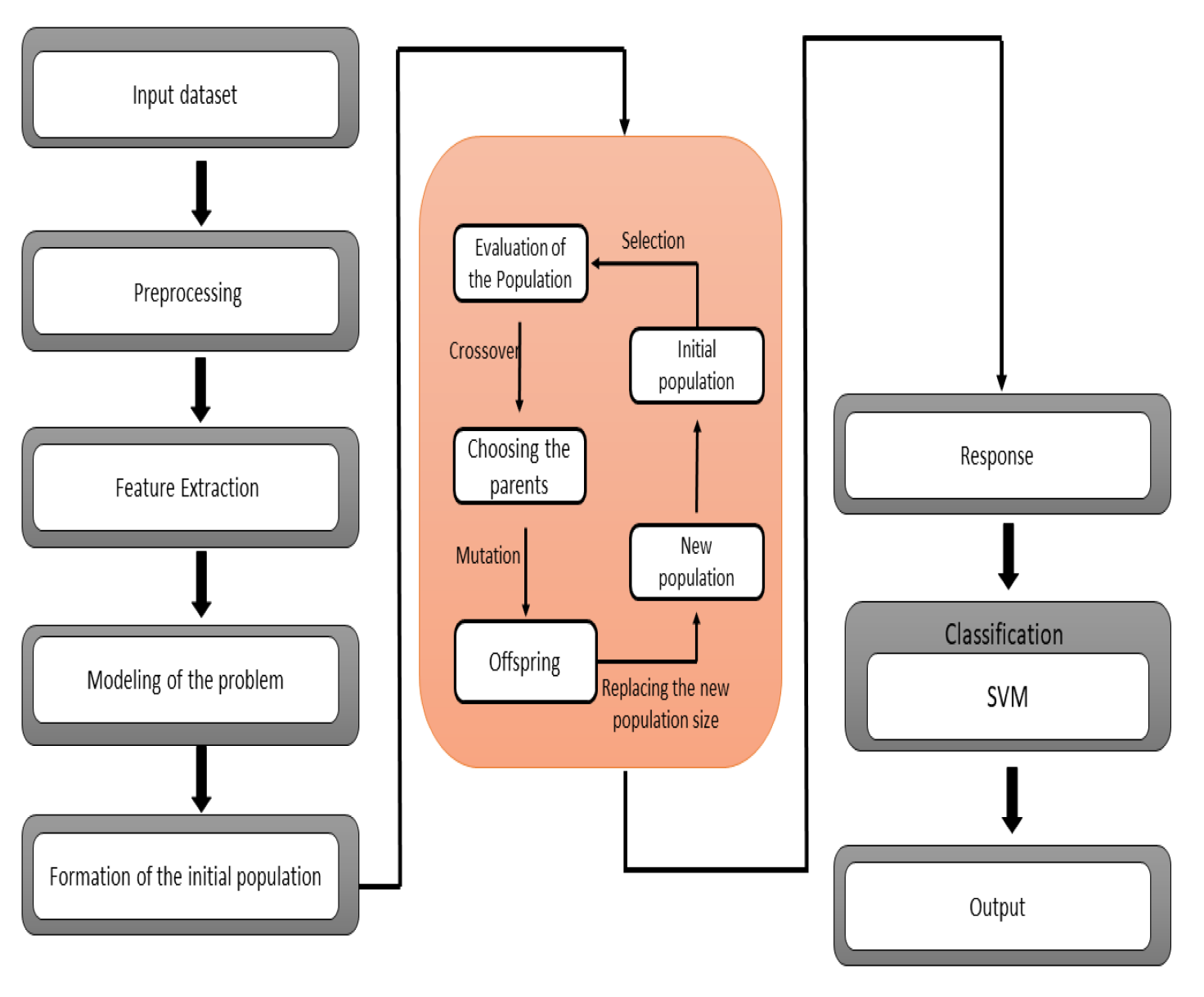
Python, Jupyter notebook, Kaggle

## **Procedure**

## The system begins with installation of anaconda software. This process is followed by launching Jupyter notebook which helps to import the certain necessary packages i.e., pandas, NumPy, sklearn etc. After importing all the packages, various machine learning is implemented for identifying an algorithm with high accuracy. In this proposed system, we analysed the lifespan among human beings based on some of the health and environmental factors. In this work, we also analyse the life expectation of individual people. The lifespan expectancy of each and every human being was analysed with the help of given data and shown as a result. In our proposed system we are using Random Forest, Decision tree, KNN, Gradient Boosting Algorithms. In our proposed system. Finally, we obtain a better accuracy with the help of random forest algorithms through which better result will be obtained comparatively with other algorithms.

## 

## 2.2 **FLOWCHART**



**RESULT**

IV. RESULT & DISCUSSIONS

In the implementation part initially, authors have

created a profile report using the pandas\_profiling

library of python. It shows that a lot of data in GDP and

population features for a lot of countries is missing.

Imputation is not the best method for handling missing

data in this particular dataset, if imputation is used then

we are taking information from other countries and

putting it in for a different That information would be

inaccurate. So, dropping those rows with missing

population & GDP is the only option. There might be a bit

of loss of data. Other features with missing values, such as

BMI, thinness, and hepatitis B, will be filled in with 0s. If

those features have no values, 0 is a safe assumption that

won't skew the data or cause problems with data

evaluation.

IV. RESULT & DISCUSSIONS

In the implementation part initially, authors have

created a profile report using the pandas\_profiling

library of python. It shows that a lot of data in GDP and

population features for a lot of countries is missing.

Imputation is not the best method for handling missing

data in this particular dataset, if imputation is used then

we are taking information from other countries and

putting it in for a different That information would be

inaccurate. So, dropping those rows with missing

population & GDP is the only option. There might be a bit

of loss of data. Other features with missing values, such as

BMI, thinness, and hepatitis B, will be filled in with 0s. If

those features have no values, 0 is a safe assumption that

won't skew the data or cause problems with data

evaluation.

In the implementation part initially, authors have

created a profile report using the pandas\_profiling

library of python. It shows that a lot of data in GDP and

population features for a lot of countries is missing.

Imputation is not the best method for handling missing

data in this particular dataset, if imputation is used then

we are taking information from other countries and

putting it in for a different That information would be

inaccurate. So, dropping those rows with missing

population & GDP is the only option. There might be a bit

of loss of data. Other features with missing values, such as

BMI, thinness, and hepatitis B, will be filled in with 0s. If

those features have no values, 0 is a safe assumption that

won't ske

In the implementation part initially, authors have

created a profile report using the pandas\_profiling

library of python. It shows that a lot of data in GDP and

population features for a lot of countries is missing.

Imputation is not the best method for handling missing

data in this particular dataset, if imputation is used then

we are taking information from other countries and

putting it in for a different That information would be

inaccurate. So, dropping those rows with missing

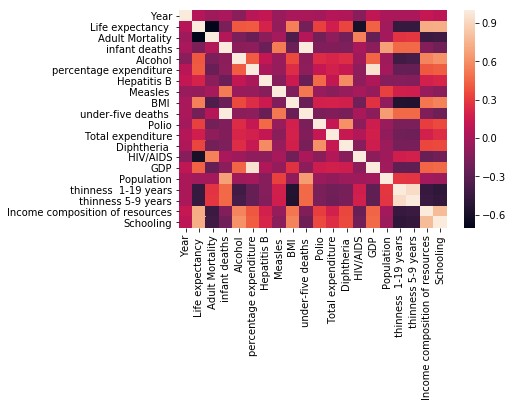
population & GDP is the only option. There might be a bit

of loss of data. Other features with missing values, such as

BMI, thinness, and hepatitis B, will be filled in with 0s. If

those features have no values, 0 is a safe assumption that

won't ske



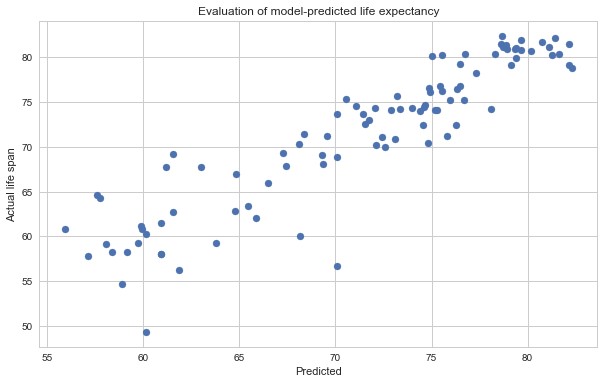


Figure4:PredictionPage

2.4 **ADVANTAGES & DISADVANTAGES**

# ADVANTAGES

* It helps determine which factors matter most, which it can ignore, and how those factors interact with each other.
* It provides a powerful statistical method that allows a business to examine the relationship between two or more variables of interest.

# DISADVANTAGES

* It is assumed that the cause and effect relationship between the variables remains unchanged. This assumption may not always hold good and hence estimation of the values of a variable made on the basis of the regression equation may lead to erroneous and misleading results.
* The functional relationship that is established between any two or more variables on the basis of some limited data may not hold good if more and more data are taken into consideration  It cannot be used in case of qualitative phenomenon.

## 2.5 APPLICATIONS

* Life expectancy provides a useful measure of average life spans and gives insights into uncertainty about age at death.
* It shows how patterns of change in life expectancy are described by trajectories of mortality improvements over age and time.
* Life expectancy gives an idea of the health issues prevalent in particular country and the healthcare system available to combat these issues.

## 3**CONCLUSION**

We implemented a machine learning model using regression analysis to predict the life expectancy of a particular country. We used JUPYTER NOTEBOOK to work on this project. The dataset we used related the life expectancy and health factors for 193 countries from the WHO data repository website and its corresponding economic data was collected from United Nation website. The following factors were applied to predict the life expectancy: Year, Status, Adult Mortality, Infant Deaths, Alcohol, Percentage Expenditure, Hepatitis B,

Measles, BMI, Under-Five Deaths, Polio, Total Expenditure, Diphtheria, HIV/AIDS, GDP, Population, Thinness 1-19 years, Thinness 5-9 years, Income Composition of Resources, Schooling.

## 4: **RECOMMENDATION**

In order to extend this work, we can increase the size of the dataset and observe the effect on the results. We can also apply other Machine Learning models in order to obtain higher accuracy.

Based on the analysis of the different machine learning algorithms used in life expectancy prediction, the following recommendations can be made:

1. Use a combination of machine learning algorithms: Combining different machine learning algorithms can improve the accuracy of life expectancy prediction. For example, combining linear regression with decision trees or random forests can improve the accuracy of life expectancy prediction.
2. Use relevant and high-quality data: The quality of the data used for training the machine learning model significantly affects the accuracy of life expectancy prediction. Therefore, it is important to use relevant and high quality data for training the model.
3. Perform feature selection: Feature selection is crucial in ensuring that the machine learning model is trained with the most relevant features. Correlation analysis and statistical analysis can be used to perform feature selection.
4. Consider multiple factors: Life expectancy prediction should consider multiple factors such as age, gender, and socio-economic status. By considering multiple factors, the accuracy of life expectancy prediction can be improved.
5. Validate the model: It is important to validate the machine learning model before deploying it for life expectancy prediction. Cross-validation and hold-out validation can be used to validate the model.
6. In conclusion, life expectancy prediction using machine learning algorithms has the potential to improve our understanding of demographic trends and health status. By using relevant and high-quality data, performing feature selection, and considering multiple factors, the accuracy of life expectancy prediction can be improved. Additionally, it is important to validate the model, continuously update it, and consider ethical implications.

LIMITATIONS & FUTURE SCOPE

Availability of all health-related data, education, and

economic expenditure stats have made possible the proper

and error-free estimation of life expectancy models.

Earlier life expectancy models were dependent on very few

variables and due to the unavailability of advanced data

exploratory and validation techniques, the trained model

was not so much accurate. Now, in recent studies, various

milestones have been achieved in this field. For example,

the ensemble of different base models, which includes the

results of individual base models, the inclusion of ANN,

and RNN techniques in solving such tasks. So, the

accuracy and addition of decisive variables into the final

life expectancy model will be far most two significant

areas of concern in the further research part. The inclusion

of newly suggested variables from many kinds of research,

such as weather-related trends and effects of natural

disasters is still a matterof concern and debate in recent

times. However, it will be great a challenge for other

authors to access this challenge into the dataset of life

expectancy. But, even having such advancements in

technology and data science forecasting accompanying

such uncertainty into th

LIMITATIONS & FUTURE SCOPE

Availability of all health-related data, education, and

economic expenditure stats have made possible the proper

and error-free estimation of life expectancy models.

Earlier life expectancy models were dependent on very few

variables and due to the unavailability of advanced data

exploratory and validation techniques, the trained model

was not so much accurate. Now, in recent studies, various

milestones have been achieved in this field. For example,

the ensemble of different base models, which includes the

results of individual base models, the inclusion of ANN,

and RNN techniques in solving such tasks. So, the

accuracy and addition of decisive variables into the final

life expectancy model will be far most two significant

areas of concern in the further research part. The inclusion

of newly suggested variables from many kinds of research,

such as weather-related trends and effects of natural

disasters is still a matterof concern and debate in recent

times. However, it will be great a challenge for other

authors to access this challenge into the dataset of life

expectancy. But, even having such advancements in

technology and data science forecasting accompanying

such uncertainty into th

LIMITATIONS & FUTURE SCOPE

Availability of all health-related data, education, and

economic expenditure stats have made possible the proper

and error-free estimation of life expectancy models.

Earlier life expectancy models were dependent on very few

variables and due to the unavailability of advanced data

exploratory and validation techniques, the trained model

was not so much accurate. Now, in recent studies, various

milestones have been achieved in this field. For example,

the ensemble of different base models, which includes the

results of individual base models, the inclusion of ANN,

and RNN techniques in solving such tasks. So, the

accuracy and addition of decisive variables into the final

life expectancy model will be far most two significant

areas of concern in the further research part. The inclusion

of newly suggested variables from many kinds of research,

such as weather-related trends and effects of natural

disasters is still a matterof concern and debate in recent

times. However, it will be great a challenge for other

authors to access this challenge into the dataset of life

expectancy. But, even having such advancements in

technology and data science forecasting accompanying

such uncertainty into th

5. **LIMITATIONS AND FUTURE SCOPE**

Availability of all health-related data, education, and economic expenditure stats have made possible the proper and error free estimation of life expectancy models. Earlier life expectancy models were dependent on very few variables and due to the unavailability of advanced data exploratory and validation techniques, the trained model was not much accurate. Now, in recent studies many milestones have been achieved in this field.

The inclusion of newly suggested variables from many kinds of research, such as weather-related trends and effects of natural disasters is still a matter of concern and debate in recent times. But, even having such advancements in technology and data science forecasting accompanying such uncertainty into the dataset is still quite not achieved yet.

The main aim of future work in this field will be the optimization of results with the inclusion of a few more variables by not affecting the overall performance, complexity and accuracy of the trained model.

6. **APPENDICES**

## APPENDIX

A. Dataset: https://www.kaggle.com/kumarajarshi/life-expectancywho