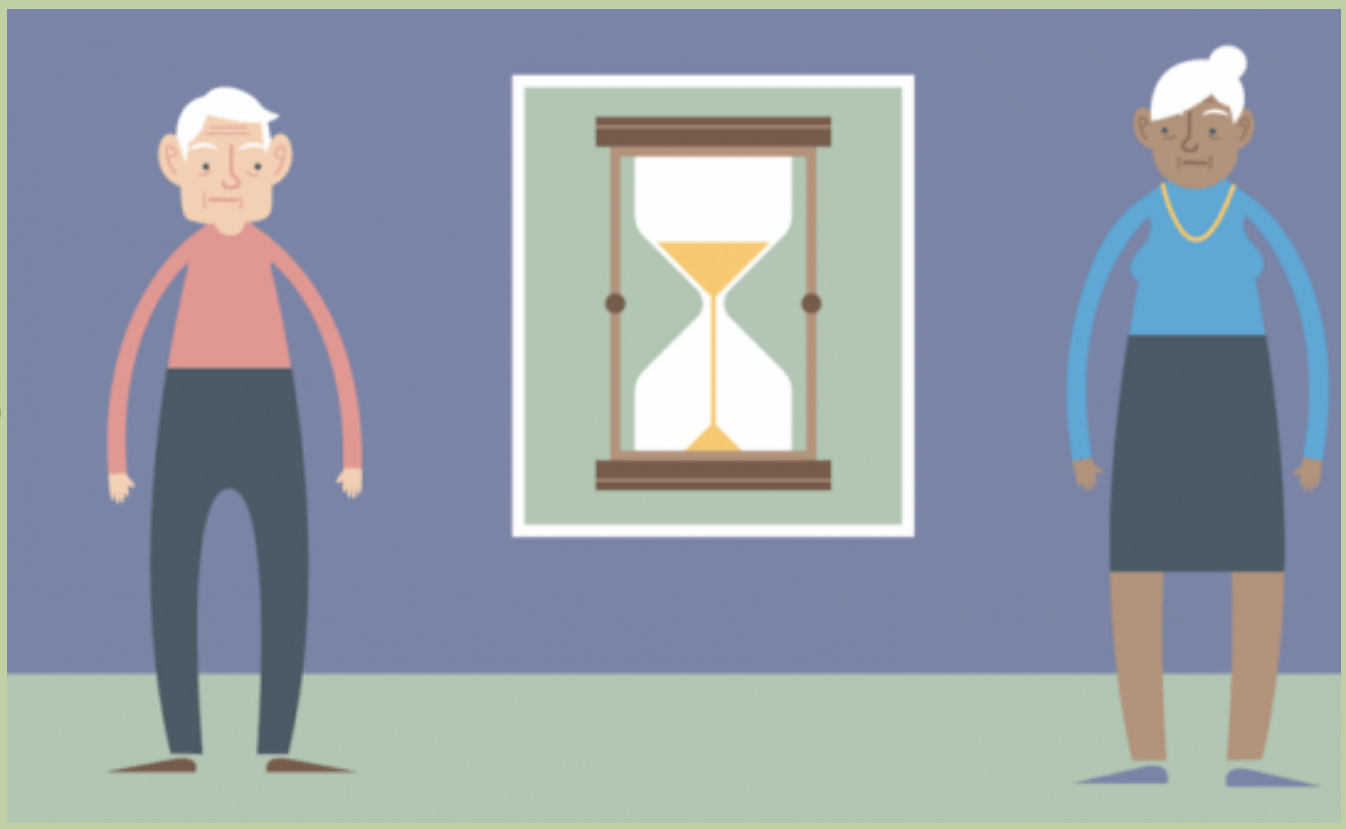
CHRIST(Deemed to be) University
Pune Lavasa campus

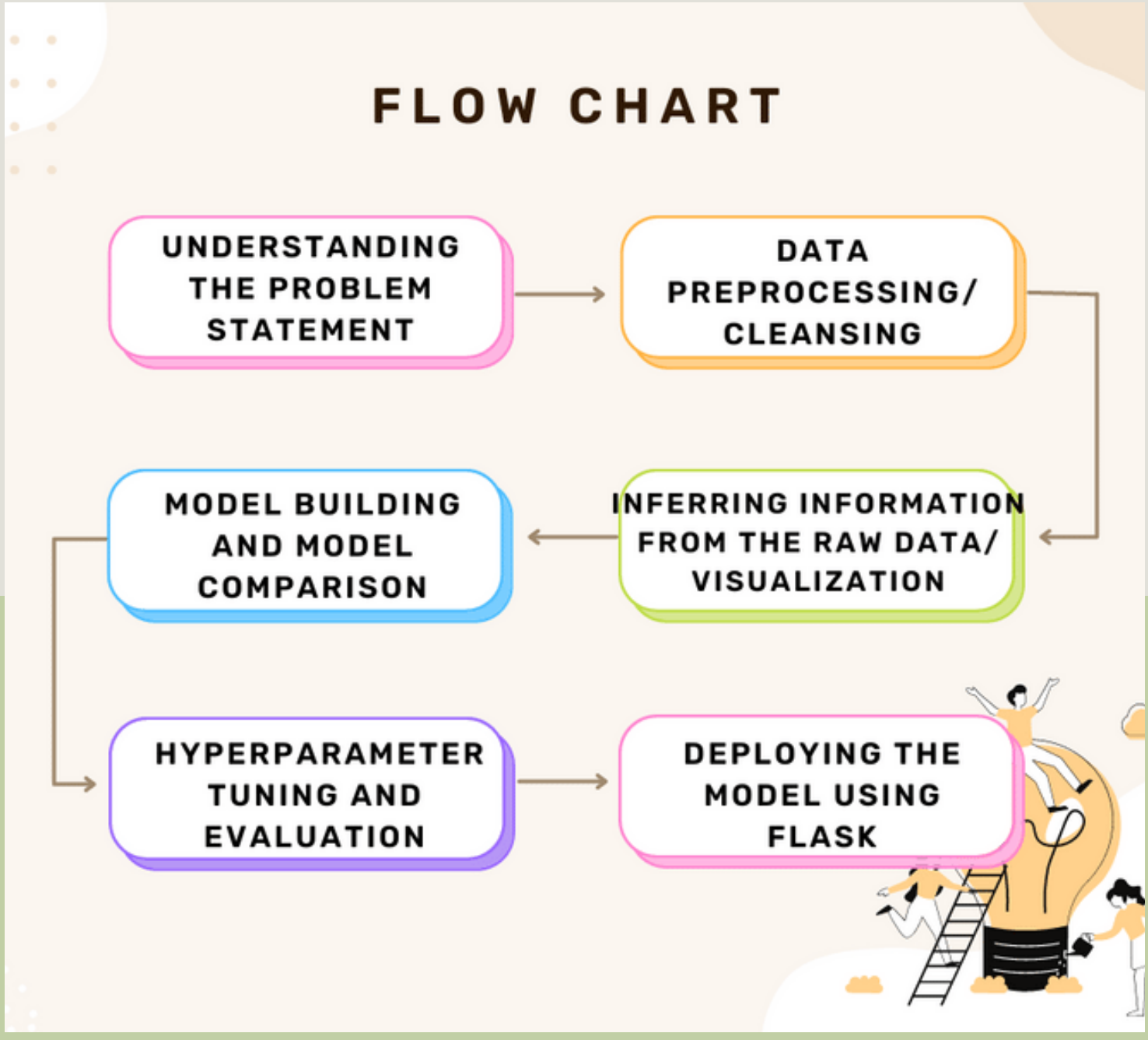
LIFE EXPECTANCY ANALYSIS

HACKVERSE'22

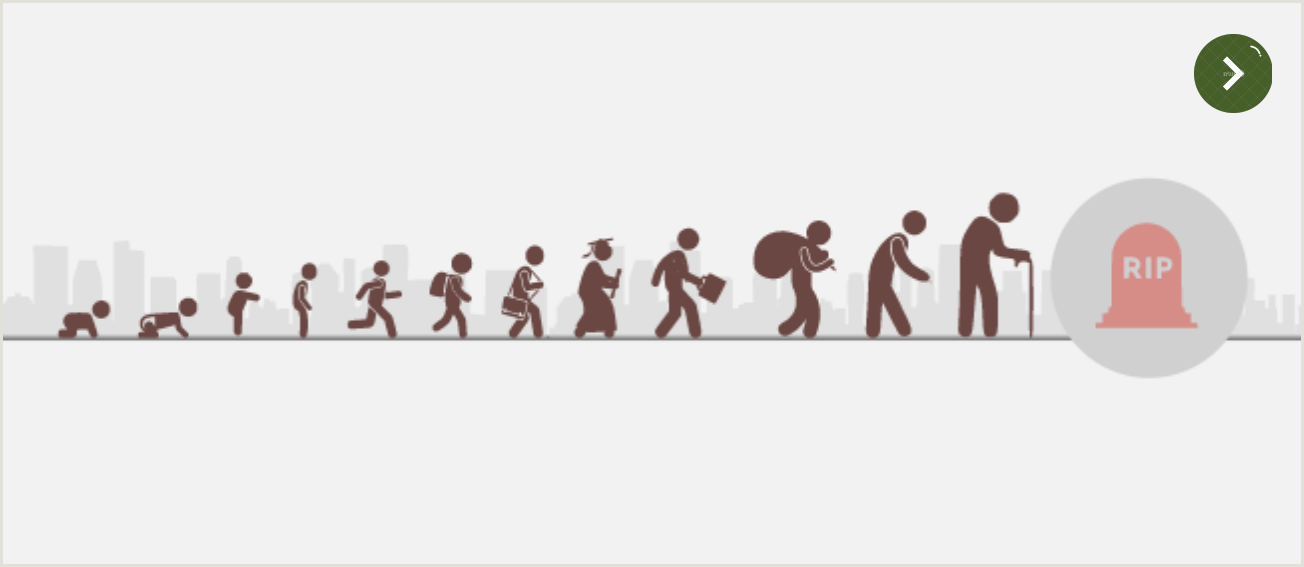
Done By
Shri Rajarajeshwari Adholokam

INDEX





Picture 1. Flow Chart



Picture 2. Life expectancy

PROBLEM STATEMENT

Various factors affect the life expectancy of different countries. And various studies have been done to understand these different factors that affect the life expectancy, But most of these studies failed to consider the effect of immunization like Hepatitis B, Polio, Diphtheria etc and Human Development Index.

So we are trying to do a study about life expectancy using the dataset of different countries from the year 2000 to 2015. We will be focusing on factors that were not taken into account in the past, like, immunization factors, mortality factors, economic factors, social factors, and other health related factors as well.

AIM

Life Expectancy Prediction: Based on features predict life expectancy of various countries

OVERVIEW

This study will help in giving important inference to different countries in order to let them know which area they should focus more on to improve the life expectancy of its population

APPROACH

We started with a proper Flow Chart and different works were delegated according to the strength of different members.The data given is not the ideal data to build a model and thus it is necessary to know and learn about the data. Most of the work time has to be given to manipulate with data and use Trial and error method. After Data Preprocessing we used the data in different models.About 10 models were taken and the best 4 was selected for hyperparameter tuning. PCA was also done to the dataset to check how it improves the accuracy. Various other trial and error methods were done to reach into the optimum model which was deployed using Flask. A Picture (Drishyam) tells a better story than words so we focused more on Visualization when it came to solving various case questions and also to get other important inferences from the dataset.

Picture 3. When life begins

MODELS USED

- Random Forest Regressor
- AdaBoost Regressor
- DecisionTree Regressor
- MLP Regressor
- XGBoost
- Ridge Regressor
- Elastic Net
- LIGHGBM
- Linear Support Vector Machine

TOOLS USED

- Python
- Pycharm
- Jupyter Notebook
- Excel Workbook
- PowerBI
- Kaggle
- Flask
- Plotly
- Prezi
- Canva

DATASET

The 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.

-**Country**: Country Name
-**Year**: Year of the data
-**Status**: Developed or Developing status
-**Life Expectancy**: Life Expectancy in age
-**Adult Mortality**: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
-**Infant Deaths**: Number of Infant Deaths per 1000 population
- **Alcohol**: Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
-**Percentage expenditure**: Expenditure on health as a percentage of Gross Domestic Product per capita(%)
-**Hepatitis B**: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
-**Measles**: Measles - number of reported cases per 1000 population
-**BMI**: Average Body Mass Index of entire population
-under-five deaths: Number of under-five deaths per 1000 population
-**Polio**: Polio (Pol3) immunization coverage among 1-year-olds (%)
-**Total expenditure**: General government expenditure on health as a percentage of total government expenditure (%)
-**Diphtheria**: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
-**HIV/AIDS**: Deaths per 1 000 live births HIV/AIDS (0-4 years)
-**GDP**: Gross Domestic Product per capita (in USD)
-**Population**: Population of the country
-**thinness 1-19 years**: Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
-**thinness 5-9 years**: Prevalence of thinness among children for Age 5 to 9(%)
-**Income composition of resources**: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
-**Schooling**: Number of years of Schooling(years)

DATA PREPROCESSING

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning. For this analysis, the data was collected from WHO and United Nations website with the help of Deeksha Russell and Duan Wang.

STEPS

- Statistical analysis
- Null value handling
  1. Mean imputation
  2. Dropping null values
  3. Forward fill
- Standardization
- PCA

OVERVIEW

We had to do a lot of trial and error to take a decision about how to handle the missing values and outliers in our dataset.
We explored different techniques like :

- Deleting (quick and dirty)
- Imputation :- Mean, ForwardFill, BackwardFill etc.

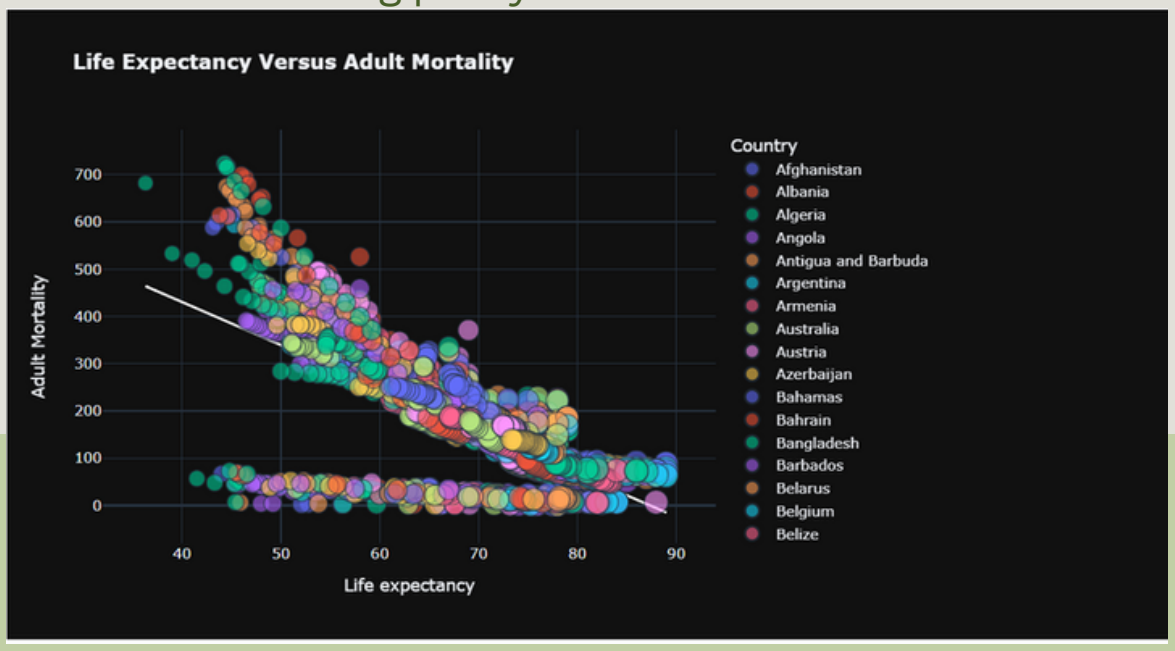
While we tried to clean the data by deleting or dropping that was resulting in
-Significant loss of data (~40%)

So we decided to do Imputation to handle our missing values.

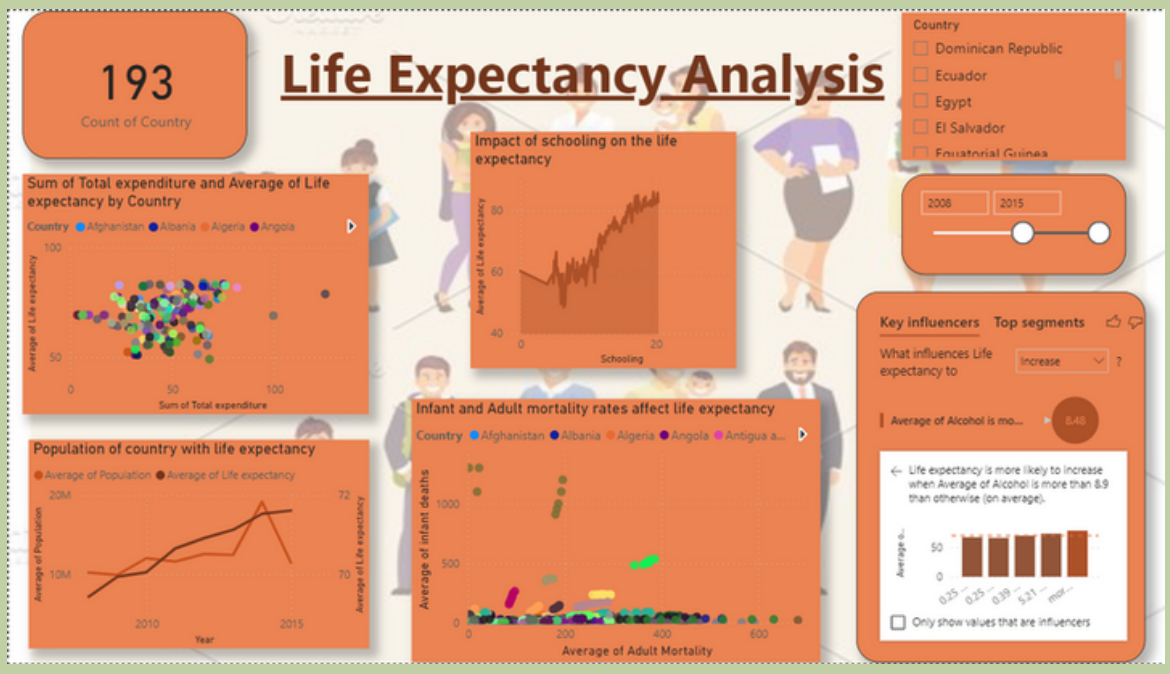
DATA VISUALIZATION

Data visualization is essential to assist businesses in quickly identifying data trends, which would otherwise be a hassle. The pictorial representation of data sets allows analysts to visualize concepts and new patterns. With the increasing surge in data every day, making sense of the quintillion bytes of data is impossible without Data Proliferation, which includes data visualization.

Visualization using plotly



Visualization using PowerBI



We were able to understand and infer more about the data using different visualization technique.

MODELING

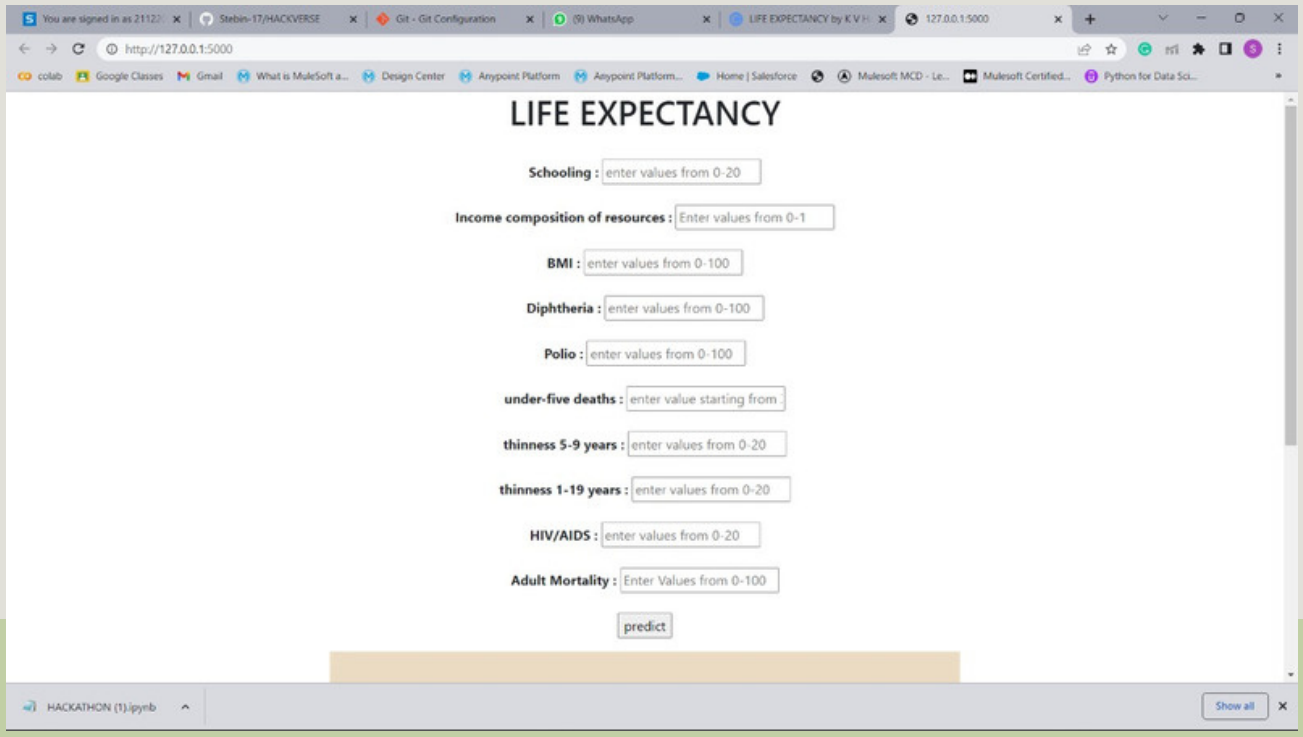
	ALGORITHM	R2_SQUARE
2	RandomForest	0.964152
5	XGBOOST	0.947282
3	MLPRegressor	0.927596
1	DecisionTree	0.857766
7	ADABOOSTREGRESSOR	0.843097
0	Ridge	0.824917
8	ELASTICNET	0.812277
6	LIGHGBM	0.566719
4	LinearSupportVectorMachine	0.512065

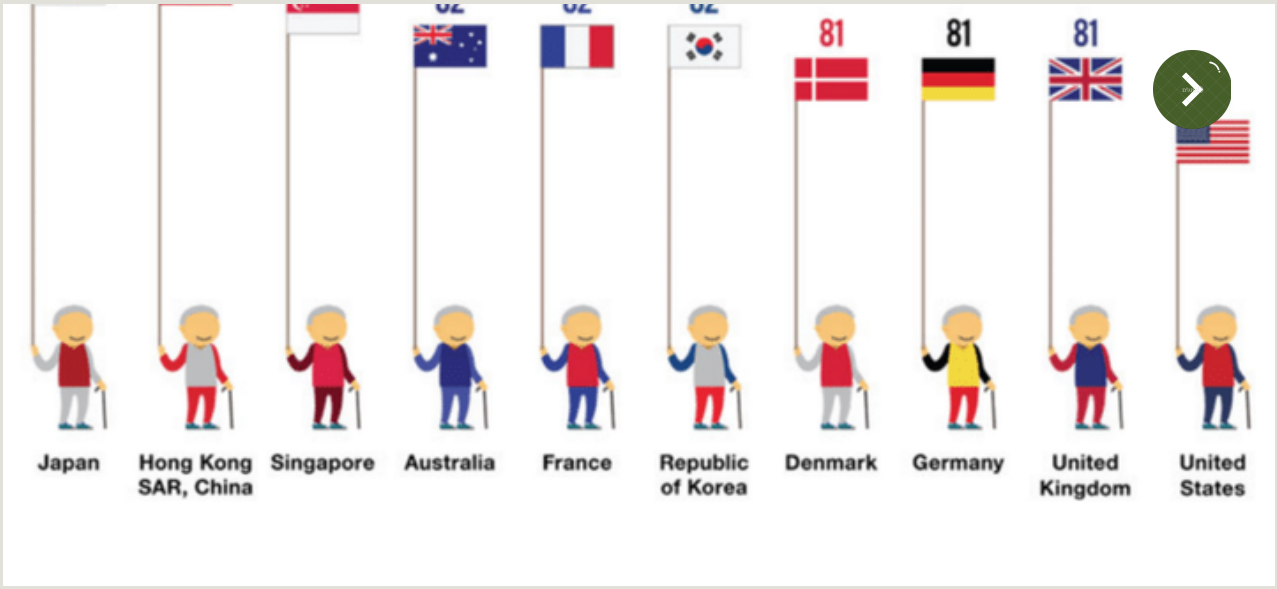
	ALGORITHM	R2_SQUARE
3	RandomForest	0.960020
0	XGBOOST	0.943703
1	ADABOOST	0.895432
2	DecisionTree	0.800857

Multiple models were built to compare the accuracies and then Model which yielded high accuracy was selected and GridSearchCV was used to tune the Hyperparameters to get an optimum model. The obtained model was used for deployment.

MODEL DEPLOYMENT

Model deployment was done with the help of Flask





CONCLUSION

Life expectancy is the average number of years that a person is expected to live. It is a measure that summarizes the mortality of a country, allowing us to compare it by generations and analyze trends. Its interpretation and meaning is even richer and can provide us with key information on the level of development of a country's welfare state.In fact, this indicator is so important for describing population conditions that, together with the education index and the Gross Domestic Product (GDP) index, it forms the Human Development Index.There is no better indicator of a country's social development than having a long and healthy life.Life expectancy expansion is a result of, among other things, improvements in nutrition, health and, above all, a decrease in mortality.

We used the the dataset to get as many inference we can and also to answer all the given key problems in our problem statement.Through the process we learned a lot about Life Expectancy and various factors that affect Life expectancy.Now these information can be used each countries to improve various sectors of their country which will in turn improve their Health care and Life Expectancy