**PREDICTING LIFE EXPECTANCY USING MACHINE LEARNING**

**CONTENTS**

1 INTRODUCTION

1.1 Overview

1.2 Purpose

2 LITERATURE SURVEY

2.1 Existing problem

2.2 Proposed solution

3 THEORITICAL ANALYSIS

3.1 Block diagram

3.2 Hardware / Software designing

4 EXPERIMENTAL INVESTIGATIONS

5 FLOWCHART

6 RESULT

7 ADVANTAGES & DISADVANTAGES

8 APPLICATIONS

9 CONCLUSION

10 FUTURE SCOPE

11 BIBILOGRAPHY

APPENDIX

A. Source code

**1.Introduction**

Life expectancy plays an important role when decisions about the final phase of life need to be made. Good prognostication for example helps to determine the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning. Advance Care Planning (ACP) is the process during which patients make decisions about the health care they wish to receive in the future, in case the patient loses the capacity of making decisions or communicating about them . Successful ACP enhances the quality of life and death for palliative patients, by providing timely palliative care and documenting preferences regarding resuscitation and euthanasia, among other things .

Accurate prognosis of life expectancy is essential for general practitioners (GPs) to decide when to introduce the topic of ACP to the patient, and it is a key determinant in end-of-life decisions . Increasing the accuracy of prognoses has the potential to benefit patients in various ways by enabling more consistent ACP, earlier and better anticipation on palliative needs, and preventing excessive treatment. This study focuses on automatic life expectancy prediction based on medical records.

**1.1Overview**

We define the task to solve as follows: predict the life expectancy (in number of months) of a patient at a certain moment in time, given the patient’s medical history up to that moment. In order to learn the task automatically from data, we trained an LSTM model on medical records of deceased patients with a recorded date of death, in which the month of death functions as the target to be predicted. We optimized the model architecture and feature set, and tested the performance of several models. The following sections describe:

* the dataset;
* the train-validation-test split;
* our methods for creating the input data for the model;
* our methods for determining the model architecture;
* our methods for feature selection;
* the evaluation protocol.

The project tries to create a model based on data provided by the World Health Organization (WHO) to evaluate the life expectancy for different countries in years. The data offers a timeframe from 2000 to 2015. The data output algorithms have been used to test if they can maintain their accuracy in predicting the life expectancy for data they haven’t been trained.

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. This problem statement provides a way to predict average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country are given.

**1.2 Purpose**

The purpose of this study is to train machine learning models using a dataset containing data from over 800 medical examinations, in order to investigate what information can be learned regarding life expectancy. The models are trained as binary classifiers and four different target features are used. The features used are limited to features understandable to the author, who is not a medical professional. Results point to significant features primarily being smoking and body mass index and secondarily alcohol consumption, physical activity and coffee consumption.

In this project I will use machine learning and AutoAI to determine the key factors driving life expectancy in the countries of the world throughout history.

Using these factors, one can predict the life expectancy of a population using health, social, and economic variables. It would be a stretch to use this information to predict the life expectancy of an individual because there are many more life variables involved than the variables presented in this project. This project and data analysis is most useful at predicting life expectancy at the population level.

**2.Literature survey**

**2.1Existing Problem**

Life expectancy plays an important role when decisions about the final phase of life need to be made. Good prognostication for example helps to determine the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning. Advance Care Planning (ACP) is the process during which patients make decisions about the health care they wish to receive in the future, in case the patient loses the capacity of making decisions or communicating about them . Successful ACP enhances the quality of life and death for palliative patients, by providing timely palliative care and documenting preferences regarding resuscitation and euthanasia, among other things .

Accurate prognosis of life expectancy is essential for general practitioners (GPs) to decide when to introduce the topic of ACP to the patient, and it is a key determinant in end-of-life decisions. Increasing the accuracy of prognoses has the potential to benefit patients in various ways by enabling more consistent ACP, earlier and better anticipation on palliative needs, and preventing excessive treatment. This study focuses on automatic life expectancy prediction based on medical records.

**2.2 Proposed Solution**

The project tries to create a model based on data provided by the World Health Organization (WHO) to evaluate the life expectancy for different countries in years. The data offers a timeframe from 2000 to 2015. The data originates from here: https://www.kaggle.com/kumarajarshi/life-expectancy-who/data The output algorithms have been used to test if they can maintain their accuracy in predicting the life expectancy for data they haven’t been trained.

**3.Theoritical Analysis**

**3.1Block Diagram**

**3.2Hardware/Software designing**

**3.2.1IBM Cloud Platform**

IBM cloud computing is a set of [cloud computing](https://en.wikipedia.org/wiki/Cloud_computing) services for business offered by the information technology company IBM. IBM Cloud includes infrastructure as a service (IaaS), software as a service (SaaS) and platform as a service (PaaS) offered through public, private and hybrid cloud delivery models, in addition to the components that make up those clouds.

IBM offers three hardware platforms for cloud computing.[1] These platforms offer built-in support for [virtualization](https://en.wikipedia.org/wiki/Virtualization). For [virtualization](https://en.wikipedia.org/wiki/Virtualization) IBM offers IBM [Websphere](https://en.wikipedia.org/wiki/Websphere) application infrastructure that supports programming models and open standards for virtualization.

The management layer of the IBM cloud framework includes IBM [Tivoli](https://en.wikipedia.org/wiki/Tivoli_Software) middleware. Management tools provide capabilities to regulate images with automated [provisioning](https://en.wikipedia.org/wiki/Provisioning) and de-provisioning, monitor operations and meter usage while tracking costs and allocating billing. The last layer of the framework provides integrated workload tools. Workloads for cloud computing are services or instances of code that can be executed to meet specific business needs. IBM offers tools for cloud based collaboration, development and test, application development, analytics, business-to-business integration, and security.

**3.2.2IBM Node-Red Service**

### Browser-based flow editing

Node-RED provides a browser-based flow editor that makes it easy to wire together flows using the wide range of nodes in the palette. Flows can be then deployed to the runtime in a single-click.

JavaScript functions can be created within the editor using a rich text editor.A built-in library allows you to save useful functions, templates or flows for re-use.

**Built on Node.js**

The light-weight runtime is built on Node.js, taking full advantage of its event-driven, non-blocking model. This makes it ideal to run at the edge of the network on low-cost hardware such as the Raspberry Pi as well as in the cloud.

With over 225,000 modules in Node's package repository, it is easy to extend the range of palette nodes to add new capabilities.

### Social Development

The flows created in Node-RED are stored using JSON which can be easily imported and exported for sharing with others.

An online flow library allows you to share your best flows with the world.

**3.2.3IBM Watson Service**

Watson was created as a [question answering](https://en.wikipedia.org/wiki/Question_answering) (QA) computing system that IBM built to apply advanced [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing), [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), [knowledge representation](https://en.wikipedia.org/wiki/Knowledge_representation), [automated reasoning](https://en.wikipedia.org/wiki/Automated_reasoning), and [machine learning](https://en.wikipedia.org/wiki/Machine_learning) technologies to the field of [open domain question answering](https://en.wikipedia.org/wiki/Open_domain_question_answering).

The key difference between QA technology and document search is that document search takes a keyword query and returns a list of documents, ranked in order of relevance to the query (often based on popularity and page ranking), while QA technology takes a question expressed in natural language, seeks to understand it in much greater detail, and returns a precise answer to the question.

When created, IBM stated that, "more than 100 different techniques are used to analyze natural language, identify sources, find and generate hypotheses, find and score evidence, and merge and rank hypotheses.

In recent years, the Watson capabilities have been extended and the way in which Watson works has been changed to take advantage of new deployment models (Watson on IBM Cloud) and evolved machine learning capabilities and optimised hardware available to developers and researchers. It is no longer purely a [question answering](https://en.wikipedia.org/wiki/Question_answering) (QA) computing system designed from Q&A pairs but can now 'see', 'hear', 'read', 'talk', 'taste', 'interpret', 'learn' and 'recommend'.

* **Software**

Watson uses IBM's software and the Apache [UIMA](https://en.wikipedia.org/wiki/UIMA) (Unstructured Information Management Architecture) framework implementation. The system was written in various languages, including [Java](https://en.wikipedia.org/wiki/Java_(programming_language)), [C++](https://en.wikipedia.org/wiki/C%2B%2B), and [Prolog](https://en.wikipedia.org/wiki/Prolog), and runs on the [SUSE Linux Enterprise Server](https://en.wikipedia.org/wiki/SUSE_Linux_Enterprise_Server) 11.

* **Hardware**

The system is workload-optimized, integrating massively parallel POWER7 processors and built on IBM’s DeepQA technology, which it uses to generate hypotheses, gather massive evidence, and analyze data. Watson employs a cluster of ninety IBM Power 750 servers, each of which uses a 3.5GHz POWER7 eight-core processor, with four threads per core. In total, the system has 2,880 POWER7 processor threads and 16 terab ytes of RAM.

* **Data**

The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles and literally works. Watson also used databases, taxonomies and ontologies including DBPedia, WordNet and Yago. The IBM team provided Watson with millions of documents, including dictionaries, encyclopedias and other reference material that it could use to build its knowledge.

**3.2.4IBM Watson Studio**

IBM Watson Studio helps data scientists and analysts prepare data and build models at scale across any cloud. With its open, flexible multicloud architecture, Watson Studio provides capabilities that empower businesses to simplify enterprise data science and AI:

* Automate AI lifecycle management with AutoAI.
* Visually prepare and build models with IBM SPSS Modeler.
* Build models using images with IBM Watson Visual Recognition and texts with IBM Watson Natural Language Classifier.
* Deploy and run models through one-click integration with IBM Watson Machine Learning.
* Manage and monitor models through integration with IBM Watson Open Scale.

**Hardware:**

* A laptop with at least 4GB RAM
* A 2GB GPU

**Software:**

* **IDE**- Spyder, Jupyter
* **Scientific Computation Library** – Pandas
* **Visualization Libraries** – Matplotlib , Seaborn
* **Algorithmic Libraries** – Scikit-Learn , Stats models
* **Dependencies** – Data from internet

**4 Experimental Investigations**

**4.1**

Increasing age is a risk factor for many diseases; therefore developing pharmacological interventions that slow down ageing and consequently postpone the onset of many age-related diseases is highly desirable. In this work we analyse data from the Drug Age database, which contains chemical compounds and their effect on the lifespan of model organisms. Predictive models were built using the machine learning method random forests to predict whether or not a chemical compound will increase Caenorhabditis elegans' lifespan, using as features Gene Ontology (GO) terms annotated for proteins targeted by the compounds and chemical descriptors calculated from each compound's chemical structure. The model with the best predictive accuracy used both biological and chemical features, achieving a prediction accuracy of 80%. The top 20 most important GO terms include those related to mitochondrial processes, to enzymatic and immunological processes, and terms related to metabolic and transport processes. We applied our best model to predict compounds which are more likely to increase C. elegans' lifespan in the DGIdb database, where the effect of the compounds on an organism's lifespan is unknown. The top hit compounds can be broadly divided into four groups: compounds affecting mitochondria, compounds for cancer treatment, anti-inflammatories, and compounds for gonadotropinreleasing hormone therapies.

**4.2 Prediction using AutoAI/ Without Python**

In this project we do prediction by using AutoAI it is the app in IBM Cloud service which is used to define prediction by many algorithm within a step without writing a code in it.

Two types of algorithm applied in it :-

* Extra Trees Regression
* Decision Tree Regression

AutoAI have the ability to choose best algorithm for data by visualization.

It can chose, Linear Regression and Logistic Regression etc.

In this we used Node-Red application also, it is used for the make Flowgraph and calculate the prediction easily if your Flowgraph is correct.

* **AutoAI**

**Automated Artificial Intelligence (AutoAI)** is a variation of the [automated machine learning](https://en.wikipedia.org/wiki/Automated_machine_learning), or AutoML, technology, which extends the automation of model building towards automation of the full life cycle of a machine learning model. It applies intelligent automation to the task of building [predictive](https://en.wikipedia.org/wiki/Predictive_modelling) machine learning models by preparing data for training, identifying the best type of model for the given data, then choosing the features, or columns of data, that best support the problem the model is solving. Finally, automation tests a variety of tuning options to reach the best result as it generates, then ranks, model-candidate pipelines. The best performing pipelines can be put into production to process new data, and deliver predictions based on the model training. Automated artificial intelligence can also be applied to making sure the model does not have inherent bias and automating the tasks for continuous improvement of the model.

The Automated Machine Learning and Data Science Team (AMLDS),[2] a small team within IBM Research, which was formed to “apply techniques from [Artificial Intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence) (AI), [Machine Learning](https://en.wikipedia.org/wiki/Machine_learning) (ML), and [data management](https://en.wikipedia.org/wiki/Data_management) to accelerate and optimize the creation of machine learning and data science workflows,” is credited with advancing the development of AutoAI.

Usecase of AutoAi:-

A typical use case for AutoAI would be training a model to predict how customers might respond to a sales incentive. The model is first trained with actual data on how customers responded to the promotion. Presented with new data, the model can provide a prediction of how a new customer might respond, with a confidence score for the prediction. Prior to AutoML, data scientists had to build these predictive models by hand, testing various combinations of algorithms, then testing to see how predictions compared to actual results. Where AutoML automated some of the process of preparing the data for training, applying algorithms to process the data and then further optimizing the results, AutoAI provides greater intelligent automation that allows for testing significantly more combinations of factors to generate model candidate pipelines that more accurately reflect and address the problem being solved. Once built, a model can be tested for bias and updated to improve performance.

The AutoAI have do their prediction in 5 steps:-

1. Provide data in .csv file.
2. Prepare data.

* Featured type detection.
* Missing values imputation.
* Feature encoding and scaling.

1. Select model type.

Selection of best algorithm for the data.

1. Generate and Rank model pipelines.

* Hyper-Parameter optimization(HPO).
* Optimized feature engineering.

1. Save and Deploy a model.

* **Node-Red**

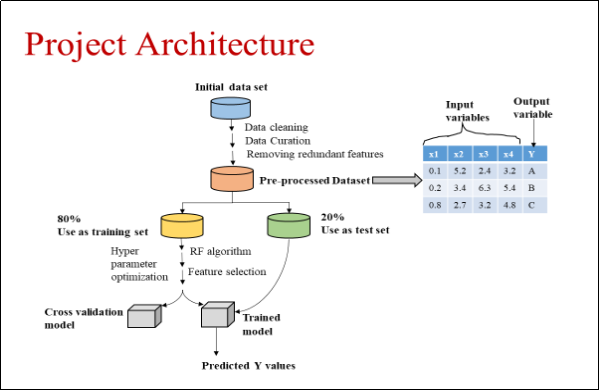
Node-Red is used to make Flowgraph.

**Node-RED** is a [flow-based](https://en.wikipedia.org/wiki/Flow-based_programming) development tool for [visual programming](https://en.wikipedia.org/wiki/Visual_programming_language) developed originally by [IBM](https://en.wikipedia.org/wiki/IBM) for wiring together hardware devices, [APIs](https://en.wikipedia.org/wiki/Application_programming_interface) and [online services](https://en.wikipedia.org/wiki/Online_services) as part of the [Internet of Things](https://en.wikipedia.org/wiki/Internet_of_Things).

Node-RED provides a [web browser](https://en.wikipedia.org/wiki/Web_browser)-based flow editor, which can be used to create [JavaScript](https://en.wikipedia.org/wiki/JavaScript) functions. Elements of applications can be saved or shared for re-use. The runtime is built on [Node.js](https://en.wikipedia.org/wiki/Node.js). The flows created in Node-RED are stored using JSON. MQTT nodes can make properly configures TLS connections.

Node-Red is a Visual tool for wiring the internet of Things.

**5.Flowchart**



**6.Result**

The user should enter the credentials required for prediction . The avg life expectancy prediction will be displayed on the screen.

**7.Advantages & Disadvantages**

**Advantages of Machine learning**

**1. Easily identifies trends and patterns**

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviours and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

**2. No human intervention needed (automation)**

With ML, you don’t need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus softwires; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

**3. Continuous Improvement**

As ml models gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

**4. Handling multi-dimensional and multi-variety data**

Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

**5. Wide Applications**

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

**Disadvantages of Machine Learning**

With all those advantages to its powerfulness and popularity, Machine Learning isn’t perfect. The following factors serve to limit it:

**1. Data Acquisition**

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

**2. Time and Resources**

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

**3. Interpretation of Results**

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

**4. High error-susceptibility**

**ML** is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

**8.Applications**

Existing health mobile app can integrate PLE feature as additional services (b) as both functions require the same physiological data to transfer to a monitoring center for PLE calculation. Applications Along with existing heath applications such as fitness tracking, chronic disease monitoring and real-time patient monitoring, the PLE application can be useful for users to improve their lifestyle and exercise by planning goals on a short and long-term basis. For example, the current PLE outcome of 85 years will be adjusted when the user changes their attributes such as smoking cessation, reducing alcohol consumption, commencing regular exercise, or modifying dietary plans.

**9.Conclusion**

Prognostication of life expectancy is difficult for humans. Our research shows that machine learning and natural language processing techniques offer a feasible and promising approach to predicting life expectancy. The research has potential for real-life applications, such as supporting timely recognition of the right moment to start Advance Care Planning. After, the code we got the life ecpectancy prediction nearly accuracy with some error.

**10.Future Scope**

Few works have been done to provide an individually customized life expectancy prediction. We have reviewed existing works and techniques in the prediction of human LE, and reached a conclusion that it is feasible to predict a PLE for individuals using evolving technologies and devices such as big data, AI, machine learning techniques, and PHDs, wearables and mobile health monitoring devices. We also identified that the collection of data will be a huge challenge due to the privacy and government policy considerations, which will require collaboration of various bodies in the health industry. The interworking of a heterogeneous health network is also a challenge for data collection. Despite these challenges, we showed a possibility of a PLE prediction by proposing an approach of data collection and application by smartphone, with which users can enter their information to access the cloud server to obtain their own PLE. No attempt has been made to create this novel idea of using smartphone integrating cloud servers for real-time data entry. We investigated obstacles and barriers that can be resolved by future works described below. Previous works have described a five year LE prediction, however it is not oriented as a personalized prediction but rather utilizes a median model-predicted probability of 5-year survival of patients who are either sick or healthy. It is proposed that this can be extended to a lifetime prediction by using big data to generate a generic data, which can be used to create a PLE based on training data as a future solution. Building a generic database will take a considerable amount of time for data collection and analysis, taken from birth to death for various demographic groups to be useful and accurate in representing each attribute classifications. Whilst current applications attempt to show PLE for smartphone users, they are complicated and difficult to collect technical data requested by the questionnaire, as users are unlikely to be able to provide these data themselves. This can be resolved by connecting the app to the central cloud server with the mHealth networks which provide other health related applications. A centralized cloud server plays a key role in collecting, processing, and creating meaningful value using big data, which forms the input of the solution as well as creates generic data against each user’s PLE requirements. Service providers shall envisage challenges and hurdles to obtain consent of personal health ‘of heterogeneous health networks across developed countries. This will lead to classification of data based on processing big data and each group’s traits, which can be used as personalized threshold ranges; When this has been completed in a cloud, it can be connected to a smart device app that can provide questionnaires developed by health specialists and collect answers to customize the user’s PLE; Optimization of the generic groups’ data is done by developing an algorithm using machine learning for continuously building and optimizing the user’s generic data. As the proposed solution requires processing and transmitting health information of users, information security is a key aspect to consider such as privacy as well as ethical requirements recommenced by regulation bodies, such as the Australian national health and medical research . The scope of security and ethical requirements need to be clearly defined and specified for future work as challenges are expected to build a centralized database with incorporation of health networks. For example, North America, Asia, and Europe may have their own unique requirements to satisfy in dealing with health data with different health research guidelines.

**Bibilography**

1. <https://www.kaggle.com/kumarajarshi/life-expectancy-who>
2. <https://docs.anaconda..com/a.naconda/user-guide/tasks/integration/spyder/> Official documentation of Anaconda Distribution.”
3. https:///www.edureka.co/blog/machine-learning-applications/
4. <https://my15.digitalexperience.ibm.com/b73a5759-c6a6-4033-ab6b-d9d4f9a6d65b/dxsites/151914d1-03d2-48fe-97d9-d21166848e65/>
5. <https://cloud.ibm.com/login>
6. <https://www.ibm.com/cloud/get-started>
7. <https://developer.ibm.com/tutorials/how-to-create-a-node-red-starter-application/>
8. <https://www.ibm.com/watson/products-services>
9. <https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/>
10. <https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html>
11. <https://developer.ibm.com/tutorials/watson-studio-auto-ai/>
12. https://bookdown.org/caoying4work/watsonstudio-workshop/auto.html#add-asset-as-auto-ai
13. <https://github.com/yashcool9268>
14. https://node-red-tdpfv.mybluemix.net/ui/#!/0?socketid=-d7eZW1fRpE5fW7nAAAB

Sample Code

#!/usr/bin/env python

# coding: utf-8

# # importing libraries

#In[1]:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# # reading file

# In[2]:

proj=pd.read\_csv(r'C:\Users\acer\Desktop\INTERNSHIP PROJECT\Life Expectancy Data.csv')

proj.head()

proj.size

#In[3]:

proj.count('rows')

# In[4]:

proj.isnull().sum()

# In[5]:

for country in country\_list:

proj.loc[proj['Country'] == country,fill\_list] = proj.loc[proj['Country'] == country,fill\_list].interpolate()

proj.dropna(inplace=True)

# In[6]:

proj.isna().sum()

# In[7]:

col\_dict = {'Life expectancy ':1 , 'Adult Mortality':2 ,

'Alcohol':3 , 'percentage expenditure': 4, 'Hepatitis B': 5,

'Measles ' : 6, ' BMI ': 7, 'under-five deaths ' : 8, 'Polio' : 9, 'Total expenditure' :10,

'Diphtheria ':11, ' HIV/AIDS':12, 'GDP':13, 'Population' :14,

' thinness 1-19 years' :15, ' thinness 5-9 years' :16,

'Income composition of resources' : 17, 'Schooling' :18, 'infant deaths':19}

# # visualisation

# In[8]:

#plots all columns against index

proj.plot()

# In[9]:

col\_dict = {'Life expectancy ':1 , 'Adult Mortality':2 ,

'Alcohol':3 , 'percentage expenditure': 4, 'Hepatitis B': 5,

'Measles ' : 6, ' BMI ': 7, 'under-five deaths ' : 8, 'Polio' : 9, 'Total expenditure' :10,

'Diphtheria ':11, ' HIV/AIDS':12, 'GDP':13, 'Population' :14,

' thinness 1-19 years' :15, ' thinness 5-9 years' :16,

'Income composition of resources' : 17, 'Schooling' :18, 'infant deaths':19}

# In[10]:

plt.figure(figsize=(20,30))

for variable, i in col\_dict.items():

plt.subplot(5,4,i)

plt.hist(proj[variable],color="grey")

plt.title(variable)

plt.show()

# In[11]:

plt.figure(figsize=(20,30))

for variable, i in col\_dict.items():

plt.subplot(6,4,i)

plt.boxplot(proj[variable])

plt.title(variable)

plt.show()

# In[12]:

for variable in col\_dict.keys():

q75, q25 = np.percentile(proj[variable], [75 ,25])

iqr = q75 - q25

min\_val = q25 - (iqr\*1.5)

max\_val = q75 + (iqr\*1.5)

print("Number of outliers and percentage of it in {} : {} and {}".format(variable,

len((np.where((proj[variable] > max\_val) | (proj[variable] < min\_val))[0])),

len((np.where((proj[variable] > max\_val) | (proj[variable] < min\_val))[0]))\*100/1987))

# In[13]:

plt.bar(proj.groupby('Status')['Status'].count().index,proj.groupby('Status')['Life expectancy '].mean())

plt.xlabel("Status",fontsize=12)

plt.ylabel("Avg Life expectancy",fontsize=12)

plt.title("Life expectancy w.r.t Status")

# In[14]:

le\_country = proj.groupby('Country')['Life expectancy '].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='grey')

plt.title("Life expectancy w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Life expectancy",fontsize=35)

# In[15]:

le\_country = proj.groupby('Country')['GDP'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Average GDP w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg GDP",fontsize=35)

# In[16]:

le\_country = proj.groupby('Country')['Adult Mortality'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Adult Mortality w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Adult Mortality",fontsize=35)

# In[17]:

le\_country = proj.groupby('Country')['Alcohol'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Alcohol w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Alcohol Comsumption",fontsize=35)

# In[18]:

le\_country = proj.groupby('Country')[' HIV/AIDS'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("HIV w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg HIV cases",fontsize=35)

# In[19]:

le\_country = proj.groupby('Country')['Income composition of resources'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Income Composition of Resources w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg income composition of resourses",fontsize=35)

# In[20]:

le\_country = proj.groupby('Country')['infant deaths'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Infant Deaths w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Infant Deaths",fontsize=35)

# In[21]:

le\_country = proj.groupby('Country')['Measles '].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Measles w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Measles cases",fontsize=35)

# In[22]:

le\_country = proj.groupby('Country')['Polio'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Polio w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Polio Cases",fontsize=35)

# In[23]:

le\_country = proj.groupby('Country')['Population'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Population w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Population",fontsize=35)

# In[24]:

le\_country = proj.groupby('Country')[' thinness 1-19 years'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title(" Thinness 1to19 years w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Thinness 1 to 19 Years",fontsize=35)

# In[25]:

le\_country = proj.groupby('Country')[' thinness 5-9 years'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title("Thinness 5 to 9 years w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg thinness 5 to 9 years ",fontsize=35)

# In[26]:

le\_country = proj.groupby('Country')['Total expenditure'].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title(" Total Expenditure w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg Total Expenditure",fontsize=35)

# In[27]:

le\_country = proj.groupby('Country')['under-five deaths '].mean().sort\_values()

le\_country.plot(kind='bar', figsize=(50,15), fontsize=25, color='red')

plt.title(" Under five Deaths w.r.t Country",fontsize=40)

plt.xlabel("Country",fontsize=35)

plt.ylabel("Avg under 5 deaths",fontsize=35)

# In[28]:

plt.bar(proj.groupby('Year')['Year'].count().index,proj.groupby('Year')['Life expectancy '].mean(),color='green',alpha=0.65)

plt.xlabel("Year",fontsize=12)

plt.ylabel("Avg Life\_Expectancy",fontsize=12)

plt.title("Life\_Expectancy w.r.t Year")

# In[29]:

plt.figure(figsize=(9,6))

corr= proj.corr()

sns.heatmap(corr)

# In[30]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["Adult Mortality"], color='green')

plt.title("Life Expectancy vs Adult Mortality")

plt.show()

# In[31]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["infant deaths"], color='green')

plt.title("Life expectancy vs Infant deaths")

# In[32]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["Alcohol"])

plt.title("Life expectancy vs Alcohol")

# In[33]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["Total expenditure"])

plt.title("Life expectancy vs Total expenditure")

# In[34]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["Diphtheria "])

plt.title("Life expectancy vs Diphtheria")

# In[35]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj[" HIV/AIDS"])

plt.title("Life expectancy vs HIV/AIDS")

# In[36]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["GDP"])

plt.title("Life expectancy vs GDP")

# In[37]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["Population"])

plt.title("Life expectancy vs Population")

# In[38]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj[" thinness 1-19 years"])

plt.title("Life expectancy vs thinness 1-19 years")

# In[39]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj[" thinness 5-9 years"])

plt.title("Life expectancy vs thinness 5-9 years")

# In[40]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["Income composition of resources"])

plt.title("Life expectancy vs Income composition of resources")

# In[41]:

plt.figure(figsize=(4,4))

plt.scatter(proj["Life expectancy "], proj["Schooling"])

plt.title("Life expectancy vs Schooling")

# In[42]:

round(proj[['Status','Life expectancy ']].groupby(['Status']).mean(),2)

# In[43]:

feature\_df = proj[['Status','Life expectancy ','Income composition of resources',' HIV/AIDS','Adult Mortality']]

feature\_df = pd.concat([feature\_df,pd.get\_dummies(feature\_df['Status'],drop\_first=True)],axis=1)

final = feature\_df.drop('Status',axis=1)

final.sample(10)

# In[44]:

proj = pd.concat([proj,pd.get\_dummies(proj['Status'],drop\_first=True)],axis=1)

proj = proj.drop('Status',axis=1)

proj.sample(10)

# In[45]:

X=proj[['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']]

Y=proj['Life expectancy ']

# In[46]:

from sklearn.model\_selection import train\_test\_split

# In[47]:

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.3, random\_state= 42)

# In[48]:

print(X\_train.shape)

print(Y\_train.shape)

print(X\_test.shape)

print(Y\_test.shape)

# In[49]:

from sklearn.linear\_model import LinearRegression

# In[50]:

from sklearn.metrics import r2\_score, mean\_squared\_error, mean\_absolute\_error

# In[51]:

model = LinearRegression(fit\_intercept=True, normalize=True).fit(X\_train, Y\_train)

predictions= model.predict(X\_test)

# In[52]:

print(predictions)

# In[53]:

len(predictions)

# In[54]:

predictions.shape

# In[55]:

r2\_score(predictions, Y\_test)

# In[56]:

mean\_squared\_error(predictions, Y\_test)

# In[57]:

mean\_absolute\_error(predictions, Y\_test)



**OUTPUT**

(Node-red UI URL):<https://node-red-qojfu.mybluemix.net/ui/#!/0?socketid=jib-FBJyvXf8RbCWAAAF>