Project Documentation

Project Title: Predicting Life Expectancy using Machine Learning

Intern Name: Khwaja Bilkhis

Index

1. Introduction

1.1 Overview

A typical Regression Machine Learning project leverages historical data to predict insights into the future. This problem statement is aimed at predicting Life Expectancy rate of a country given various features.

1.2 Purpose

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 systems and some specific disease related deaths that happened in the country are given.

2. Literature Survey

2.1 Existing Problem

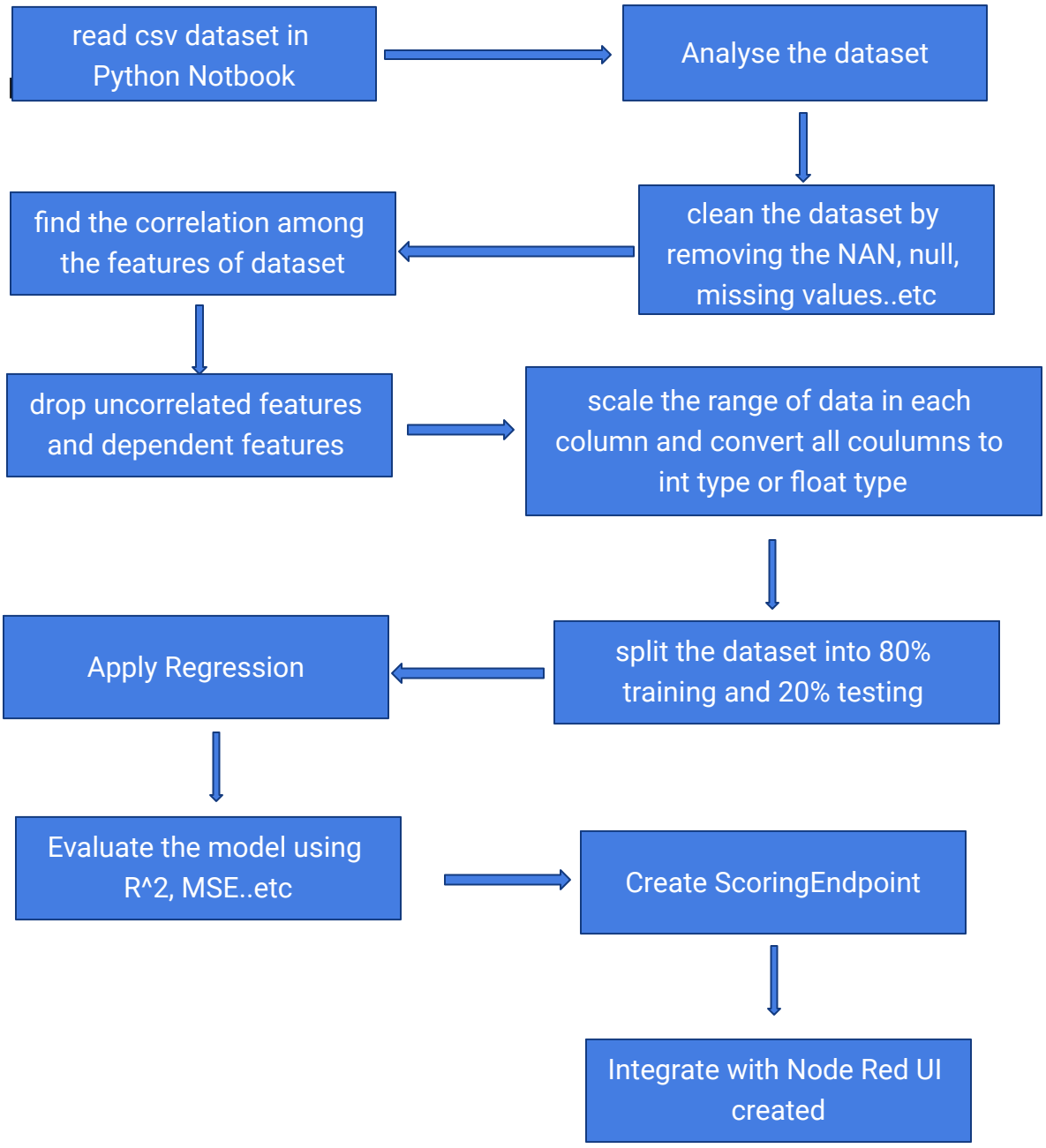
Although there have been many studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that the effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on data set of one year for all the countries.

2.2 Proposed Solution

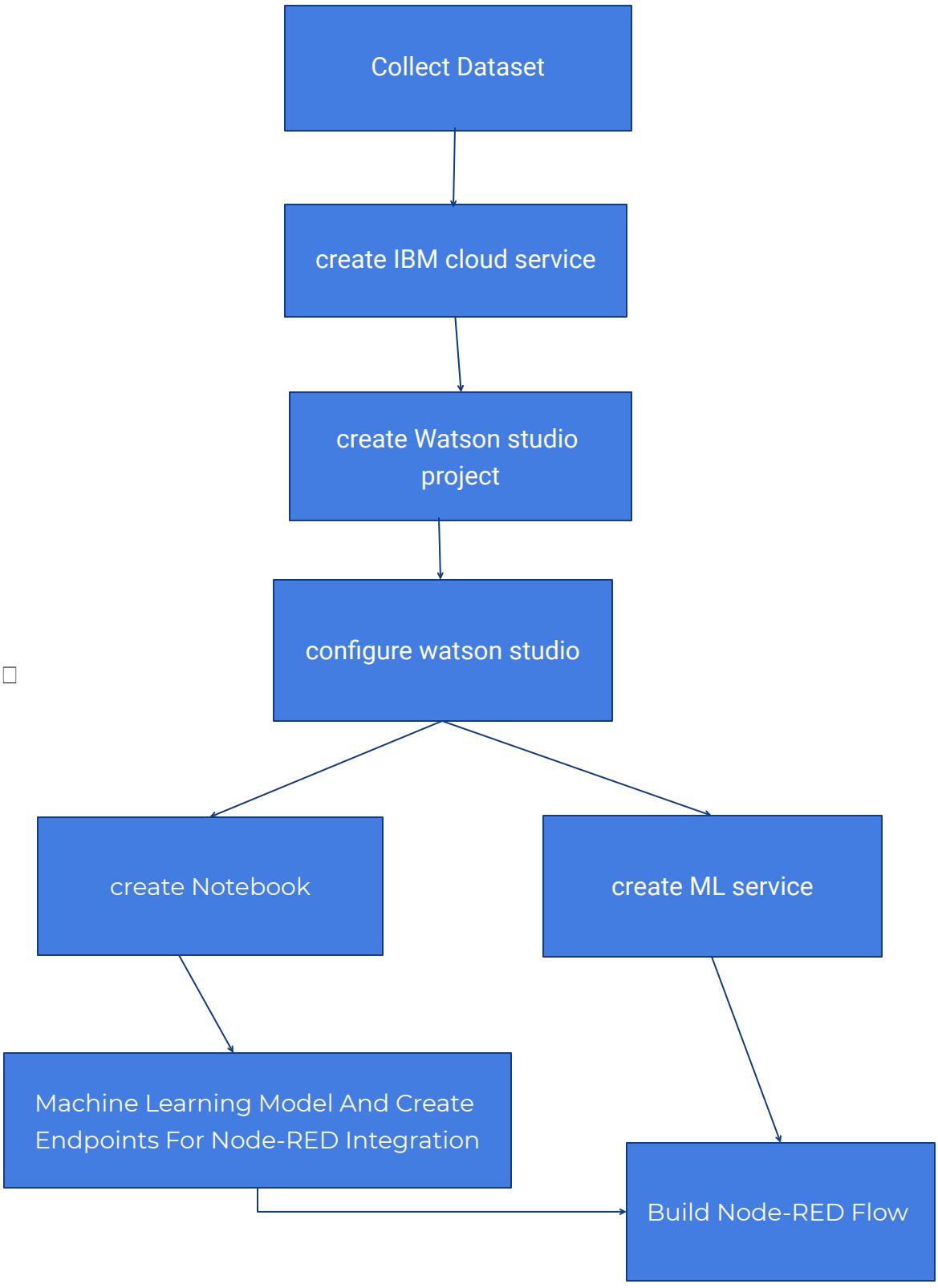
This model resolves both the factors stated above by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this model will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its

3. Theoretical analysis

3.1 Block Diagram



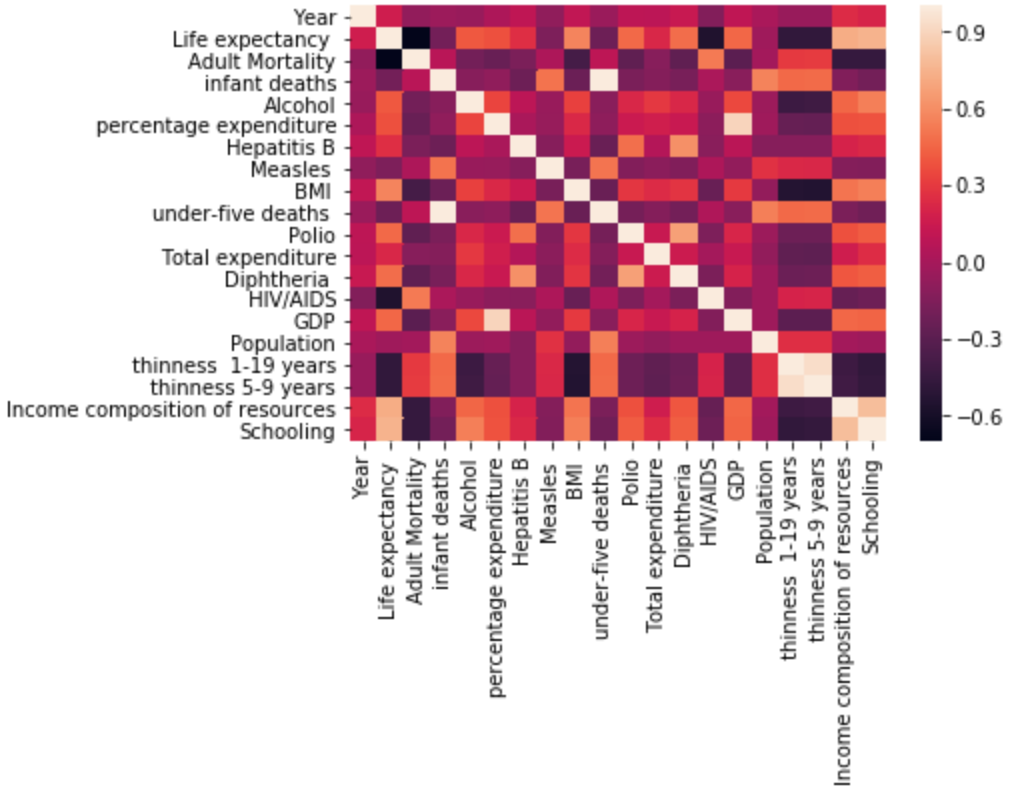
3.2 Hardware/Software Designing



4.Experimental Investigations

	Year	Life expectancy	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
count	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000	2938.000000	2904.000000	2938.000000	2919.000000	2712.00000	2919.000000	2938.000000	2490.000000	2.286000e+03	2904.000000	2904.000000	2771.000000	2775.000000
mean	2007.518720	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461	2419.592240	38.321247	42.035739	82.550188	5.938019	82.324084	1.742103	7483.158469	1.275338e+07	4.839704	4.870317	0.627551	11.992793
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016	11467.272489	20.044034	160.445548	23.428046	2.49832	23.716912	5.077785	14270.169342	6.101210e+07	4.420195	4.508882	0.210904	3.358920
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	0.000000	3.000000	0.37000	2.000000	0.100000	1.681350	3.400000e+01	0.100000	0.100000	0.000000	0.000000
25%	2004.000000	63.100000	74.000000	0.000000	0.877500	4.685343	77.000000	0.000000	19.300000	0.000000	78.000000	4.26000	78.000000	0.100000	463.935626	1.957932e+05	1.600000	1.500000	0.493000	10.100000
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000	17.000000	43.500000	4.000000	93.000000	5.75500	93.000000	0.100000	1766.947595	1.386542e+06	3.300000	3.300000	0.677000	12.300000
75%	2012.000000	75.700000	228.000000	22.000000	7.702500	441.534144	97.000000	360.250000	56.200000	28.000000	97.000000	7.49250	97.000000	0.800000	5910.806335	7.420359e+06	7.200000	7.200000	0.779000	14.300000
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000	87.300000	2500.000000	99.000000	17.60000	99.000000	50.600000	119172.741800	1.293859e+09	27.700000	28.600000	0.948000	20.700000

Heat Map



5. Result

The obtained Regression model has following coefficients for every feature considered

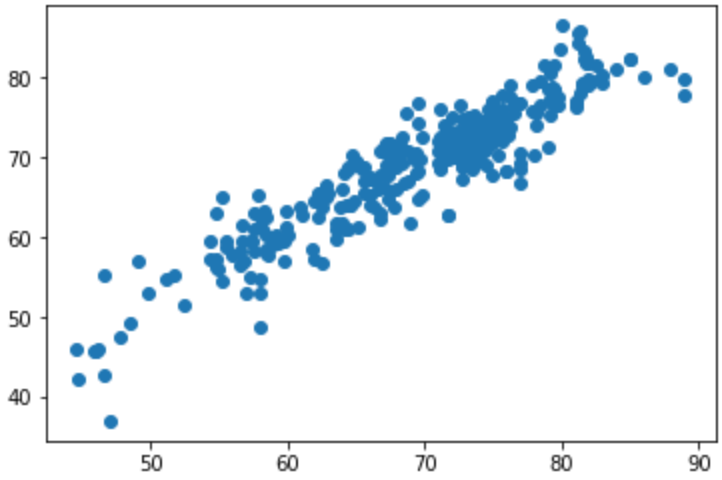
	Coefficient
Adult Mortality	-1.626953e-02
infant deaths	8.389445e-02
Alcohol	-7.817120e-02
percentage expenditure	4.462377e-04
BMI	3.177180e-02
under-five deaths	-6.386235e-02
Polio	7.904515e-03

Total expenditure	7.659089e-02
Diphtheria	1.583841e-02
Hepatitis B	-7.400443e-03
HIV/AIDS	-4.430984e-01
GDP	9.786084e-06
Population	-3.302396e-11
thinness 1-19 years	-6.452201e-02
Income composition of resources	1.023292e+01
Schooling	8.880559e-01

Model Evaluation

MAE: 2.6021747066702856
MSE: 11.403838201944144
RMSE: 3.3769569440465395
R^2: 0.852399041924826

Actual Vs Predicted Result



6. Advantages and Disadvantages

Advantages

- Immunization and human development index are considered
- Utilizes past data to predict future values
- Assists countries figure out factors affecting the health of their population

Disadvantages

- Results can never be 100% accurate.
- There might be other factors which are affecting Life Expectancy and not included in this model.
- There might be a more complex and hybrid model which can require data set larger than current ones and can give more accurate results.

7. Applications

- Timely recognition of the right moment to start Advance Care Planning
- We can see how different economics impact health

8. Conclusion

The project tries to create a model based on data provided by the World Health Organization (WHO) to evaluate the life expectancy of human beings.The data offers a time frame from 2000 to 2015. The data originates from here:
https://www.kaggle.com/kumarajarshi/life-expectancy-who/data.
Here, Linear Regression algorithm is used in building a model to predict the life expectancy. In order to build this model IBM Cloud Services were used. In IBM Cloud Services, Watson Studio was used to create a python notebook which contained the main python code. Then, Node-RED flow was created where the UI for the model was designed and finally ML services were integrated with Node-RED.

9. Future Scope

- Increase R^2 Score of the model by using Cross Validation
- Insert feature 'Country' in the Regression Model
- Use multiple data sets to build Regression model

10.Bibliography

- https://www.kaggle.com/kumarajarshi/life-expectancy-who
- https://bookdown.org/caoying4work/watsonstudio-workshop/jn.html
- https://www.ibm.com/watson/products-services
- https://developer.ibm.com/technologies/machine-learning/series/learning-path-machine-learning-for-developers/

Appendix

A. Source Code

```
import types
import pandas as pd
from botocore.client import Config
import ibm_boto3
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
%matplotlib inline

def __iter__(self): return 0

# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.
# You might want to remove those credentials before you share the notebook.
client_cf4cd54841054407ab5e5bdef2945f2e = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='4v6P7_fYbOwBhyA99cccTi4WiEbQ11pT8DiaGBv41OW-',
    ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3.eu-geo.objectstorage.service.networklayer.com')

body = client_cf4cd54841054407ab5e5bdef2945f2e.get_object(Bucket='khwaja-donotdelete-pr-jnh11fyjb65fq7',Key='Life Expectancy Data.csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

who = pd.read_csv(body)
who.head()
##Cleaning data
df=who[['Life expectancy ', 'Adult Mortality',
       'infant deaths', 'Alcohol', 'percentage expenditure',
        ' BMI ', 'under-five deaths ', 'Polio', 'Total expenditure',
       'Diphtheria ', 'Hepatitis B',' HIV/AIDS', 'GDP', 'Population',
```

```
      ' thinness  1-19 years',
      'Income composition of resources', 'Schooling']]
df = df.dropna()
df = df.dropna(how='all',axis=1)
##Applying Regression
X = df.drop('Life expectancy ', axis=1)
y = df['Life expectancy ']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,random_state=10)
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train,y_train)
##Modal Evaluation
from sklearn import metrics
predictions = lm.predict(X_test)
plt.scatter(y_test,predictions)

print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
print('R^2:', metrics.r2_score(y_test, predictions))
##Scoring Endpoint Creation
!pip install watson-machine-learning-client
from watson_machine_learning_client import WatsonMachineLearningAPIClient
wml_credentials={
  "apikey": "VroV2wwlrqW9tQl-2v2dveXXR3lvEZJvNClIyX-Xfoy8",
  "instance_id": "94cb1802-512d-4244-a010-e5714229dd3a",
  "url": "https://eu-gb.ml.cloud.ibm.com"
}
client = WatsonMachineLearningAPIClient( wml_credentials )
model_props = {client.repository.ModelMetaNames.AUTHOR_NAME: "KB",

            client.repository.ModelMetaNames.NAME: "LifeExpectancy"}
model_artifact =client.repository.store_model(lm, meta_props=model_props)
published_model_uid = client.repository.get_model_uid(model_artifact)
```

```
scoring_endpoint = client.deployments.get_scoring_url(deployment)
```