9/1/2022

Proposal

Life Expectancy

Ebowusim Michael

STUDENT NUMBER - 4050630

Table of Content

Introduction

I have decided to take on a project for the open programme. The following document contains information about the project, stating my understanding of the data, the goal of the project as well as graphical representations of information I found after exploring the dataset.

Also give some insight on how the project might impact society in both positive and negative ways. This document will also include the proposal in regards to achieving the project goal as well as the constraints and risks I may face that could impact the quality of the project.

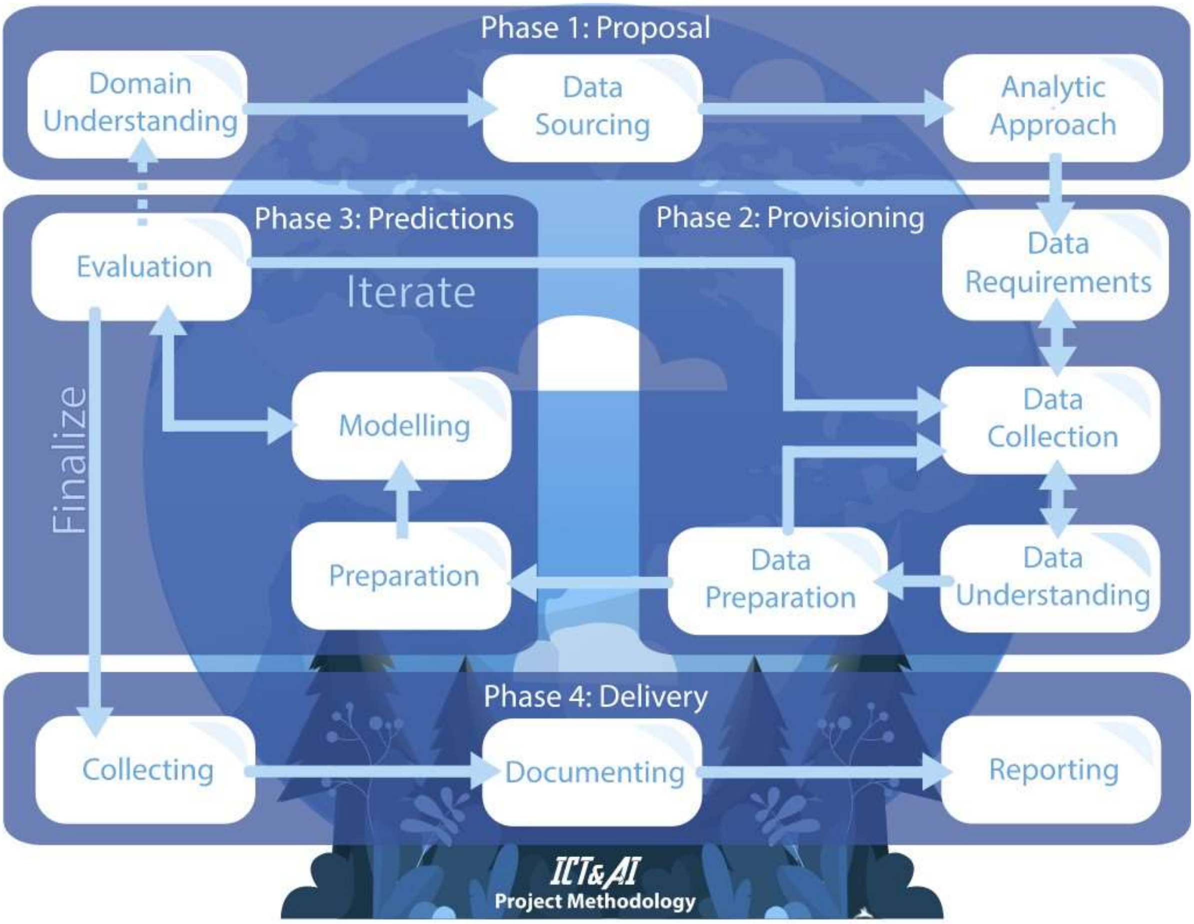
Business Understanding

The dataset related to life expectancy, 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.

Project Methodology

The project is going to be split into 4. The 4 phases are as follows: Phase 1: Proposal, Phase 2: Provisioning, Phase 3: Predictions, Phase 4: Delivery.



Proposal

In the proposal part I will be working on the initial project proposal and the exploratory data analysis in which the dataset will be explored and also try to find sensible trends in the data that could be used for further process.

Provisioning

In the provisioning phase, there will be a deeper dive into the data trying to expand EDA and prepare the data for modeling. This phase also includes understanding and expanding the data.

Predictions

In the prediction phase I will create a machine learning model and train it with the data that I have at hand to yield a prediction that justifies the project . Different machine learning models will be tested to make sure I can deliver the best possible results.

Delivery

In the delivery phase, I will present the completed project and report to the teachers so as to get feedback.

Project Goal

Since the observations this dataset are based on different countries, it will be easier for a count ry to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently i mprove the life expectancy of its population. Find a set of features that affect Life Expectancy.

The goal in this challenge is to find the factors that affect the life expectancy. The target variabl e is Life_Expectancy.

Stakeholders

- Nick Welman (Semester coach & Internal Stakeholder)
- Konings Hans (Semester coach & Internal Stakeholder)
- Kuijpers Nico(Teacher & Internal Stakeholder)
- Huisman Jose(Teacher & Internal Stakeholder)
- Michael Ebowusim(Developer & Internal Stakeholder)

How are the stakeholders affected?

All Stakeholders listed to produce this technology require its functionality. Effort has been given by each of the stakeholders to make sure the project is known and perceived.

Communication

Communication between the semester coach, teachers and developer will beheld mainly online due to the current situation regarding the pandemic using Microsoft teams as the main platform.

Data Sourcing & Storage

Data Sourcing:

The dataset which I am working with at this was provided by The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status. The datasets are made available to public for the purpose of health data analysis

Data Storage:

For data storage, the data will be stored locally on my PC.

Privacy

This technology does not register personal data, due to the data not having any columns about individual data about people e.g Patient name, date of birth e.t.c.

Data analysis(EDA)

To understand the columns I will be dealing with, below is the data description of these columns.

- **Country** - the country in which the indicators are from (i.e. United States of America or Congo)
- **Year** - the calendar year the indicators are from (ranging from 2000 to 2015)
- **Status**- whether a country is considered to be 'Developing' or 'Developed' by WHO standards
- **life_expectancy**- the life expectancy of people in years for a particular country and year. This is also the target/dependent variable.
- **adult_mortality** - the adult mortality rate per 1000 population (i.e. number of people dying between 15 and 60 years per 1000 population); if the rate is 263 then that means 263 people will die out of 1000 between the ages of 15 and 60; another way to think of this is that the chance an individual will die between 15 and 60 is 26.3%
- **infant_deaths**- number of infant deaths per 1000 population; similar to above, but for infants
- **alcohol** - a country's alcohol consumption rate measured as liters of pure alcohol consumption per capita
- **percentage_expenditure** - expenditure on health as a percentage of Gross Domestic Product (gdp)
- **hepatitis_b** - number of 1 year olds with Hepatitis B immunization over all 1 year olds in population

- **measles** - number of reported Measles cases per 1000 population
- **bmi** - average Body Mass Index (BMI) of a country's total population
- **under-five_deaths** - number of people under the age of five deaths per 1000 population
- **polio** - number of 1 year olds with Polio immunization over the number of all 1 year olds in population
- **total_expenditure** - government expenditure on health as a percentage of total government expenditure
- **diphtheria** - Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1 year olds
- **hiv/aids** - deaths per 1000 live births caused by HIV/AIDS for people under 5; number of people under 5 who die due to HIV/AIDS per 1000 births
- **gdp** - Gross Domestic Product per capita
- **population** - population of a country

- **thinness_1-19_years** - rate of thinness among people aged 10-19 (Note: variable should be renamed to thinness_10-19_years to more accurately represent the variable)
- **thinness_5-9_years** - rate of thinness among people aged 5-9
- **income_composition_of_resources** - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- **schooling** - average number of years of schooling of a population

Key Questions

The dataset aims to answer the following key questions:

- Does various predicting factors which has been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting the life expectancy?

- Should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan?

- How does Infant and Adult mortality rates affect life expectancy?

- Does Life Expectancy has positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol etc.

- What is the impact of schooling on the lifespan of humans?

- Does Life Expectancy have positive or negative relationship with drinking alcohol?

- Do densely populated countries tend to have lower life expectancy?

- What is the impact of Immunization coverage on life Expectancy?

Modelling

On the topic of modelling, I will have discussions with the teacher and then conclude on what algorithms to use. For now, I will just list the one I think:

- Decision Tree
- Support-vector machine
- Linear Regression
- Random Forest