**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**ALESSANDRA DE ASSIS BARBOSA**

**PREVISÃO DA EXPECTATIVA DE VIDA DA POPULACAO UTILIZANDO ALGORITMOS DE MACHINE LEARNING**

Belo Horizonte

2022

**ALESSANDRA DE ASSIS BARBOSA**

**PREVISÃO DA EXPECTATIVA DE VIDA DA POPULACAO UTILIZANDO ALGORITMOS DE MACHINE LEARNING**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte

2022

Sommario

[1. Introdução 4](#_Toc89357667)

[1.1. Contextualização 5](#_Toc89357668)

[1.2. O problema proposto 7](#_Toc89357669)

[1.3. Objetivos 7](#_Toc89357670)

[2. Coleta de Dados 8](#_Toc89357671)

[2.1. Dataset 01\_ Expectativa de vida 11](#_Toc89357672)

[2.2. Dataset 02\_ Dados de emissão de CO2 e de Gases de Efeito Estufa 12](#_Toc89357673)

[2.3. Dataset 03\_ Dados Demograficos 13](#_Toc89357674)

[2.4. APIs de geolocalizaçao e nomes e codificaçao dos paises 13](#_Toc89357675)

[2.4.1. Data geolocation API Nominatim 13](#_Toc89357676)

[2.4.2. Data codigos dos paises e continentes 14](#_Toc89357677)

[3. Processamento/Tratamento de Dados 16](#_Toc89357678)

[4. Análise e Exploração dos Dados 17](#_Toc89357679)

[5. Criação de Modelos de Machine Learning 18](#_Toc89357680)

[6. Interpretação dos Resultados 19](#_Toc89357681)

[7. Apresentação dos Resultados 20](#_Toc89357682)

[8. Links 21](#_Toc89357683)

[REFERÊNCIAS 22](#_Toc89357684)

[APÊNDICE 23](#_Toc89357685)

[Figura 1 - Life Expectancy from 1955, from www.worldometers.info 5](#_heading=h.lnxbz9)

[Figura 2 - Diferença em anos da expectativa de vida 6](#_heading=h.35nkun2)

[Figura 3 - Fatores usados no projeto](about:blank) **Errore. Il segnalibro non è definito.**

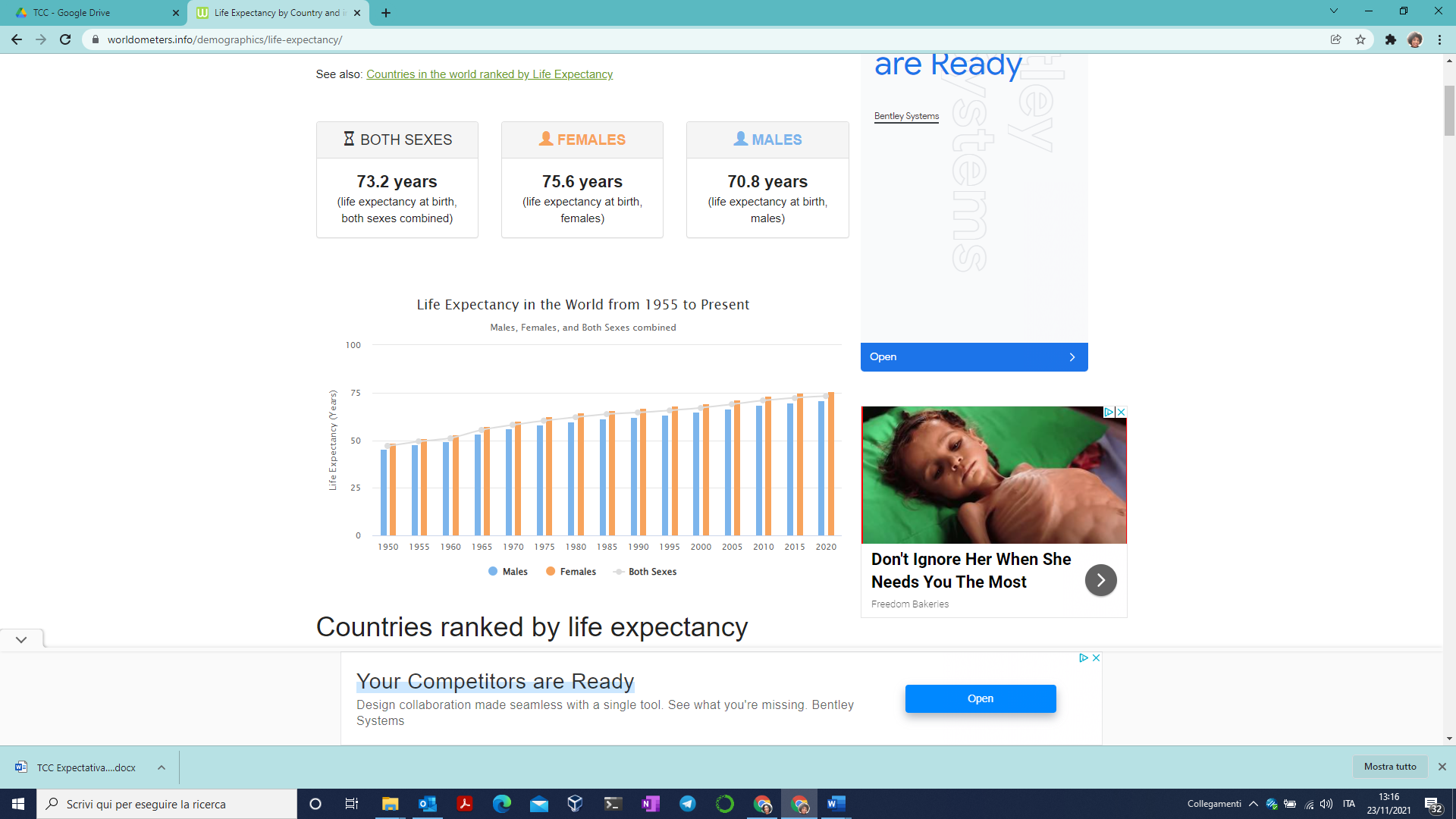
# Introdução

A Expectativa de vida, também chamada de esperança de vida, è a expressão usada para indicar o número médio de anos que cada individuo provavelmente viverá caso sejam mantidas as mesmas condições vivenciadas no momento do nascimento. Em particular, a expectativa de vida ao nascer indica quantos anos em média um recém-nascido está destinado a viver.

De acordo com alguns estudos a expectativa de vida de um país é influenciada por diferentes condições tais como a qualidade de vida que um país possui, fatores como educação, saúde, assistência social, saneamento básico, segurança no trabalho, índices de violência, ausência ou presença de guerras e de conflitos internos. Ao longo da história essas condiçoes nem sempre foram as mesmas variando de acordo com as condições de vida apresentadas, diminuindo em épocas de conflitos, pestes e pandemias e aumentando de acordo com a elevação dos padrões de vida da população.

A Revolução Industrial è um exemplo disto. Iniciada a partir do século XVIII, foi um marco na evolução das taxas de expectativa de vida em todo o mundo, uma vez que o progresso da medicina, da qualidade nutricional e o surgimento de infraestruturas de saneamento básico e higiene foram responsáveis pela redução acentuada das taxas de mortalidade mundiais, o que provocou um grande crescimento demográfico e, consequentemente, o aumento da expectativa de vida no mundo todo.

Outro fator muito estudado è a diferença na expectativa de vida entre mulheres e homens que tambèm aumentou nos últimos dois séculos.[[1]](#footnote-1) Segundo Warren Farrell, esse aumento seria causado pelo maior estresse exercido pela sociedade industrializada sobre o gênero masculino.[[2]](#footnote-2)



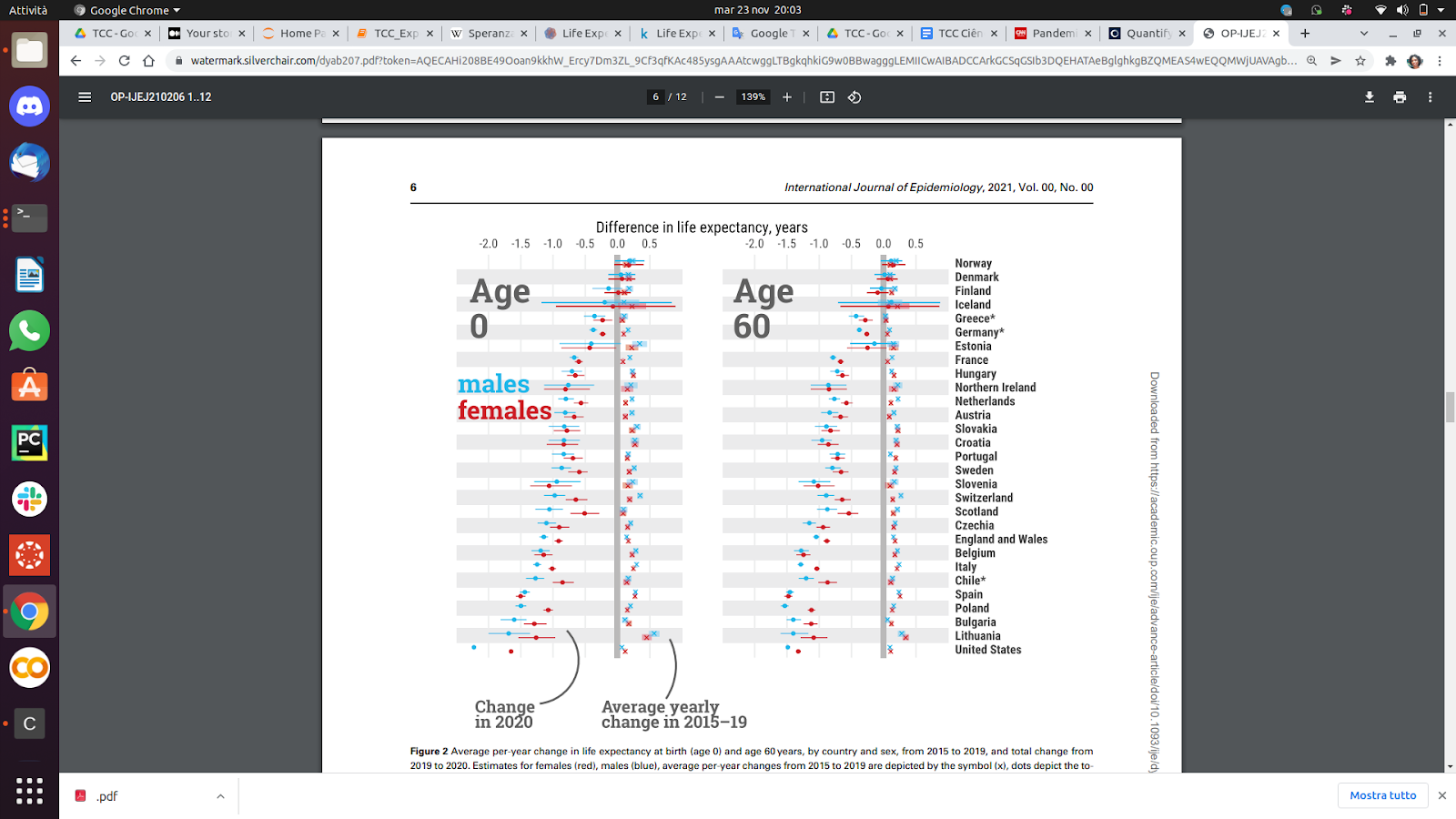
**Figura 1 - Life Expectancy from 1955, from www.worldometers.info**

De acordo com um outro estudo realizado por pesquisadores da Southern California University, liderados por Eileen M. Crimmins, 30 por cento do mesmo fenômeno é atribuível a doenças relacionadas ao fumo, enquanto a maior parte às doenças cardiovasculares. Basicamente, segundo os autores, a melhora nas condições gerais de saúde teria trazido à tona o maior risco de eventos cardiovasculares determinados no gênero masculino, tanto do ponto de vista biológico quanto genético. [[3]](#footnote-3)

# Contextualização

Organizações como Organização Mundial da Saúde (OMS) e outras associações acompanham o estado de saúde dos países, bem como muitos outros fatores relacionados. Eles são responsáveis por monitorar o risco para a saúde pública, promovendo a saúde humana e coordenando as respostas às emergências. Orgaos como OMS dependem muito de algoritmos estatísticos para estudar a expectativa de vida e o impacto das pandemias ao longo da vida, tanto si e visto durante a pandemia iniciada no ano 2019 la prospectiva de um abaixamento da estimativa de vida mundial desencadeando uma corrida na tomada de providências para reduzir os danos.

A pandemia de Covid-19 causou em 2020 a maior redução na expectativa de vida desde a Segunda Guerra Mundial, de acordo com um estudo publicado pela Universidade de Oxford, esse estudo mostra através de técnicas demográficas e bayesian probability qual efeito da pandemia na redução da expectativa de vida.



**Figura 2 - Diferença em anos da expectativa de vida**

Isso significa dizer que a expectativa de vida, embora venha progressivamente aumentando, não é livre de morbidade, de forma que a carga de doença para condições crônicas é crescente (SCHRAMM et al., 2004), num cenário de rápido processo de envelhecimento (WONG; CARVALHO, 2006).

# O problema proposto

Embora tenha havido muitos estudos realizados no passado sobre fatores que afetam a expectativa de vida, considerando variáveis demográficas, composição de renda e taxas de mortalidade, verificou-se que o efeito da imunização não foi levado tanto em consideração no passado. Além disso, algumas das pesquisas anteriores foram feitas considerando a regressão linear múltipla com base no conjunto de dados de um ano para todos os países. Portanto, isso dá motivação para resolver ambos os fatores declarados anteriormente, formulando um modelo de regressão baseado no modelo de efeitos mistos, considerando os dados de um período de 2000 a 2015 para todos os países onde imunizações importantes como hepatite B, poliomielite e difteria também serão consideradas.

Este estudo se propoe a utilizar modelos de machine learning para prever a expectativa de vida dos paises e se concentrarà em fatores de imunização, fatores de mortalidade, fatores econômicos, fatores geograficos, fatores ambientais e outros fatores relacionados à saúde nos diferentes países.

Nessa direção, esse estudo aponta na utilizaçao de tecnicas de machine learning de forma a prever com acuracia a expectativa de vida de um paìs a fim de propor cenarios futuros che individuem os fatores que fortemente atuam no aumento da expectativa de vida atraves de tecnicas de features selection e com isso endereçar os investimentos publicos e privados na promoçao de iniciativas e de incentivo ao envelhecimento saudavel e ativo dos cidadaos.

# Objetivos

O objetivo deste trabalho è realizar uma análise exploratória que permitirá de individuar os fatores correlacionados a redução ou aumento expectativa de vida , identificar um modelo matemático que possa ser usado para a previsão e a predição da expectativa de vida, comparar a performance de vários modelos de machine learning e a realização de cenários que possam auxiliar na tomada de decisões de investimentos público e privado.

**Além disso esse trabalho tambèm visa responder as seguintes hipótese:**

* Os vários fatores de previsão escolhidos inicialmente realmente afetam a expectativa de vida?
* Quais são as variáveis de previsão que realmente afetam a expectativa de vida?
* Um país com expectativa de vida menor (<65) deve aumentar seus gastos com saúde para melhorar sua expectativa de vida média?
* Como as taxas de mortalidade de bebês e adultos afetam a expectativa de vida?
* A expectativa de vida tem correlação positiva ou negativa com hábitos alimentares, estilo de vida, exercícios, fumo, bebida alcoólica etc.
* Qual é o impacto da escolaridade na expectativa de vida dos humanos?
* A expectativa de vida tem uma relação positiva ou negativa com o consumo de álcool?
* Países densamente povoados ou altamente populosos tendem a ter menor expectativa de vida?
* Qual é o impacto da cobertura de imunização na expectativa de vida?
* Paises mais inquinados representam uma expectativa de vida menor?

O objetivo maior deste projeto nao è somente predizer a expectativa de vida nos paises mas identificar fatores diretamente vinculados**.**

# Ferramentas utilizadas

Para a realizaçao deste projeto foi usado como ferramenta o jupyter notebook no qual foram instaladas extensoes. As extensões do Jupyter Notebook são complementos simples que estendem a funcionalidade básica do ambiente do notebook e o codigo pip para instalaçao è:

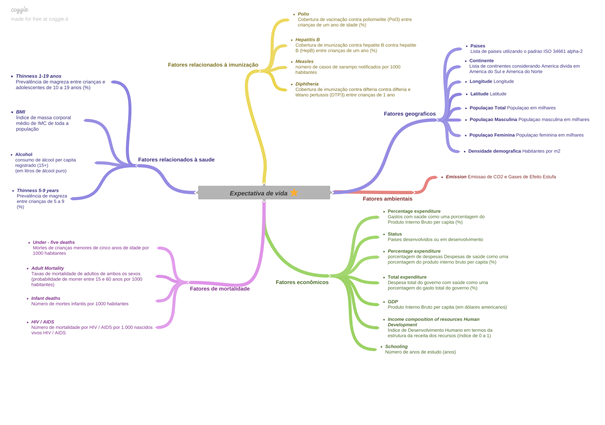
*pip install jupyter\_contrib\_nbextensions e jupyter contrib nbextension install --user* e as

principais extesoes ativadas foram: (xxxx) .

A linguagem de programaçao escolhida foi o Python mas foram instaladas uma serie de bibliotecas e suas relativas versoes que podem ser cnsultadas no documento requirements.txt (Anexo1)

# Coleta de Dados

As informaçoes utilizados nesste projeto sao dados dos anos 2010 à 2015 e foram coletados em diferentes fontes: orgao mundial da saude e das naçoes unidas, dados do IBGE, organizaçao Our World in Data e API de geolocalizaçao. Sao dados de 193 paìses che foram subdividos em seis grandes areas: Fatores relacionados a Economia, fatores ambientais, fatores demograficos, fatores de mortalidade, fatores relacionados a imunizaçao e fatores relacionados a saùde. A figura abaixo ajuda a descrever essa subdivisao.



**Figura SEQ Figura \\* ARABIC 3 - Fatores usados no projeto**

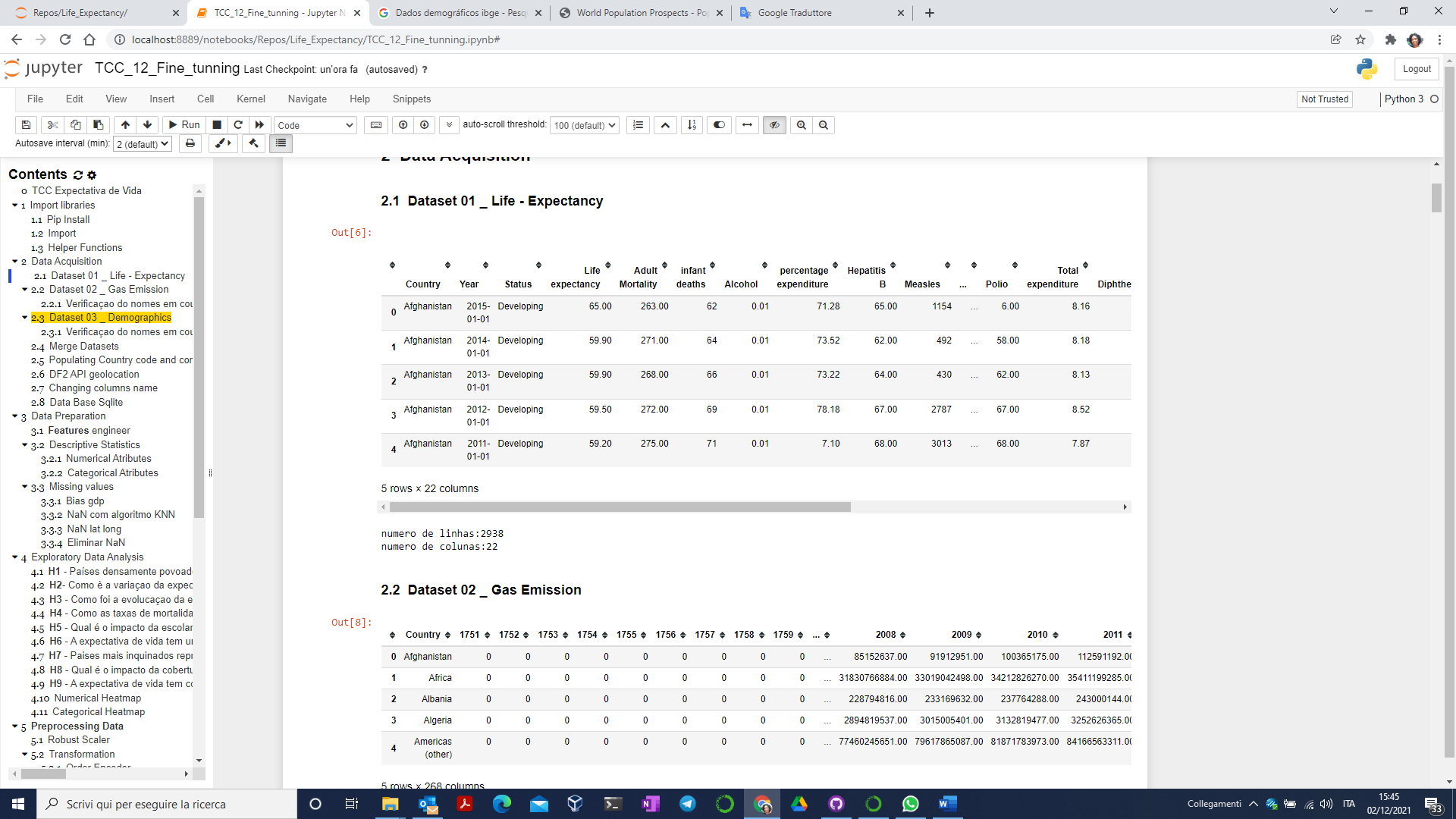
|  |  |  |  |
| --- | --- | --- | --- |
| **Fatores** | **Nome da coluna** | **Descrição** | **Variavel** |
|  | Country | Nome do Pais | categorica |
| Year | Ano de misuraçao | data |
| Life Expectancy | Anos da expectativa de vida da populaçao nascente | numerica |
| Saude | Thinness 5-9 anos | Prevalência de magreza entre crianças de 5 a 9 anos (%) numerica | numerica |
| Thinness 1-19 anos | Prevalência de magreza entre crianças e adolescentes de 10 a 19 anos | numerica |
| BMI | Índice de massa corporal médio de IMC de toda a população | numerica |
| Alcohol | consumo de àlcool per capita registrado em litros de álcool puro | numerica |
| mortalidade | Under - five deaths | Mortes de crianças menores de cinco anos de idade por 1000 habitantes | numerica |
| Adult Mortality | Taxas de mortalidade de adultos de ambos os sexos (probabilidade de morrer entre 15 e 60 anos por 1000 habitantes) | numerica |
| Infant deaths | Número de mortes infantis por 1000 habitantes | numerica |
| HIV / AIDS | Número de mortalidade por HIV / AIDS por 1.000 nascidos vivos HIV / AIDS | numerica |
| econômicos | Percentage expenditure | Gastos com saúde como uma porcentagem do  Produto Interno Bruto per capita (%) | numerica |
| Status | Paises desenvolvidos ou em desenvolvimento | categorica |
| Percentage expenditure | porcentagem de despesas Despesas de saúde como uma porcentagem do produto interno bruto per capita (%) | numerica |
| Total expenditure | Despesa total do governo com saúde como uma porcentagem do gasto total do governo (%) | numerica |
| GDP | Produto Interno Bruto per capita (em dólares americanos) | numerica |
| Income composition of resources Human Development | Índice de Desenvolvimento Humano em termos da estrutura da receita dos recursos (índice de 0 a 1) | numerica |
| Schooling | Número de anos de estudo (anos) | numerica |
| imunização | Polio | Cobertura de vacinação contra poliomielite (Pol3) entre crianças de um ano de idade (%) | numerica |
| Hepatitis B | Cobertura de imunização contra hepatite B contra hepatite B (HepB) entre crianças de um ano (%) | numerica |
| Measles | número de casos de sarampo notificados por 1000 habitantes | numerica |
| Diphtheria | Cobertura de imunização contra difteria contra difteria e tétano pertussis (DTP3) entre crianças de 1 ano | numerica |
| geograficos | Code | Codigo dos paises utilizando o padrao ISO 34661 alpha-2 | categorica |
| Continente | Lista de conitnentes considerando America divida em America do Sul e America do Norte | categorica |
| Long | Longitude | numerica |
| Lat | Latitude | numerica |
| pop\_total | Populaçao total em milhares | numerica |
| pop\_male | Populaçao masculina em milhares | numerica |
| pop\_female | Populaçao feminina em milhares | numerica |
| density | Densidade demografica Habitantes por m2 | numerica |
| Fatores ambientais | Emission | Emissao de CO2 e Gases de Efeito Estufa | numerica |

# Dataset 01\_ Expectativa de vida

O conjunto de dados relacionado à expectativa de vida e fatores de saúde para 193 países foi coletado do repositório de dados do Observatório Global de Saúde (GHO) da Organização Mundial da Saúde (OMS) que acompanha o estado de saúde, bem como muitos outros fatores relacionados para todos os países e seus dados econômicos correspondentes foram coletados do site das Nações Unidas e foram reunidos em um unico dataframe pelos cientistas Deeksha Russell e Duan Wang sendo disponibilizados na plataforma Kaggle. <https://www.kaggle.com/kumarajarshi/life-expectancy-who>

Observou-se que, nos últimos 15 anos, houve um grande desenvolvimento no setor da saúde, resultando na melhoria das taxas de mortalidade humana, especialmente nas nações em desenvolvimento, em comparação com os últimos 30 anos.

Os arquivos de dados individuais foram mesclados em um único conjunto de dados. O arquivo final mesclado (conjunto de dados final) consiste em 22 colunas e 2938 linhas, o que significa 20 variáveis de previsão.



# Dataset 02\_ Dados de emissão de CO2 e de Gases de Efeito Estufa

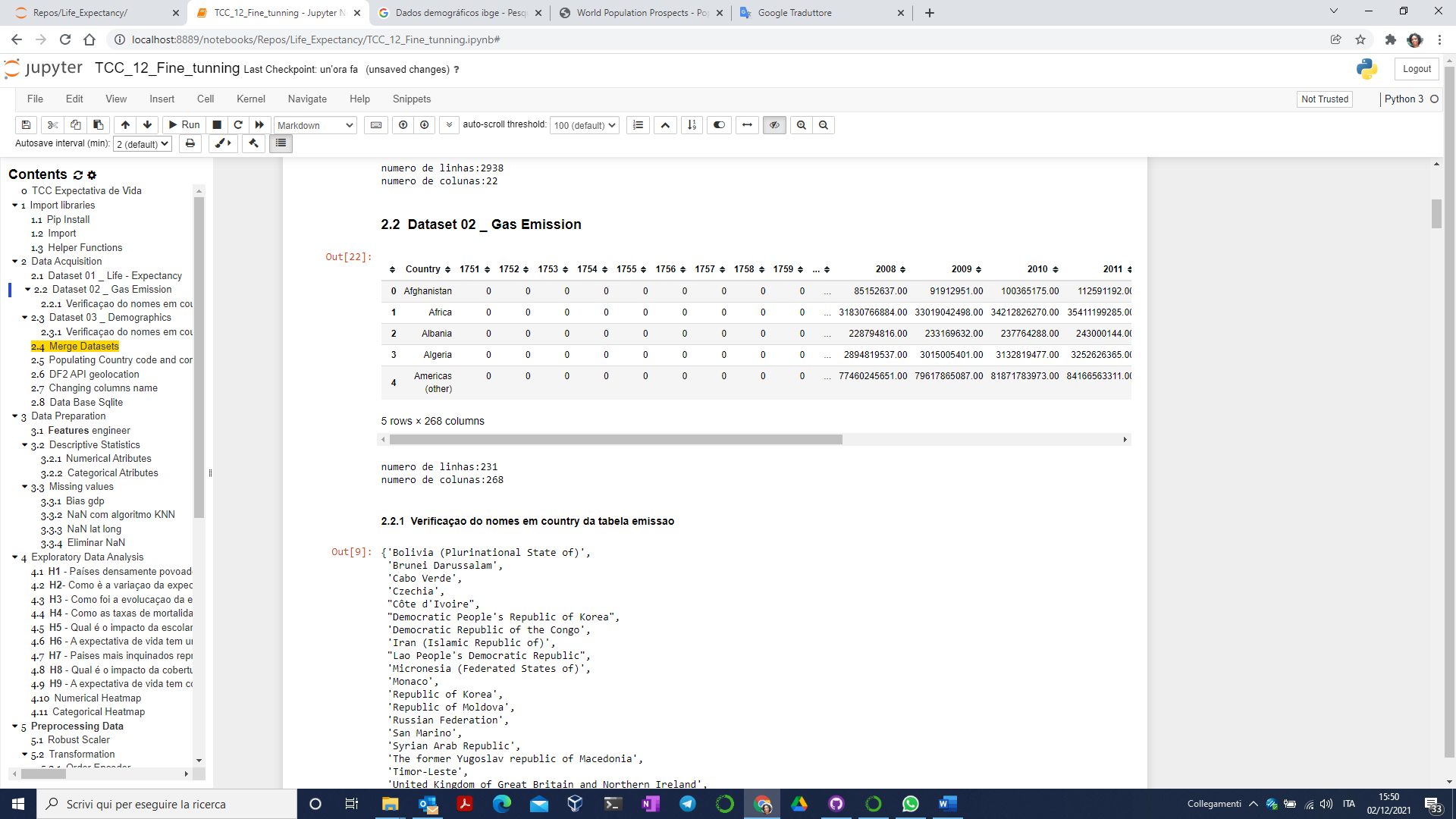
As grandes concentrações de poluentes encontrados na atmosfera e no meio ambiente dao origem as infecções não só do trato respiratório de nós humanos, como a outras doenças e catastrofes naturais.

O dióxido de carbono é um dos produtos químicos liberados na atmosfera, e a absorção térmica deste e outros gases contribuintes para a poluição, interferem diretamente no equilíbrio térmico do Planeta e Saúde Pública. Estes gases interferem de forma grave na estabilidade do efeito estufa, e com isso deixa o meio ambiente e a saúde em constante calamidade desencadeando novas doenças e elevadas temperaturas climáticas.

Algumas analises mostram que a emissão do gás dióxido de carbono (CO2) em nível global/ou praticamente em todos os países, independente de suas economias e evolução industrial, esse gás é produzido das mais variadas formas: sendo desde queima de combustíveis derivados do petróleo, indústrias e queimadas. Pensando nisto se decidiu incluir esses dados no projeto .

O conjunto de dados de emissões de CO2 e Gases de Efeito Estufa utilizados neste projeto é uma coleção de indicadores-chave mantidos e disponibilizados por Our World in Data (<https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions>).

Sao dados atualizados regularmente e que inclui informaçoes sobre as emissões de dióxido de carbono, Metano, Óxido Nitroso, Hexafluoreto de Enxofre e duas famílias de gases, Hidrofluorcarbono e Perfluorcarbono medidos em milhoes de toneladas métricas de CO2e por ano (MtCO2e).

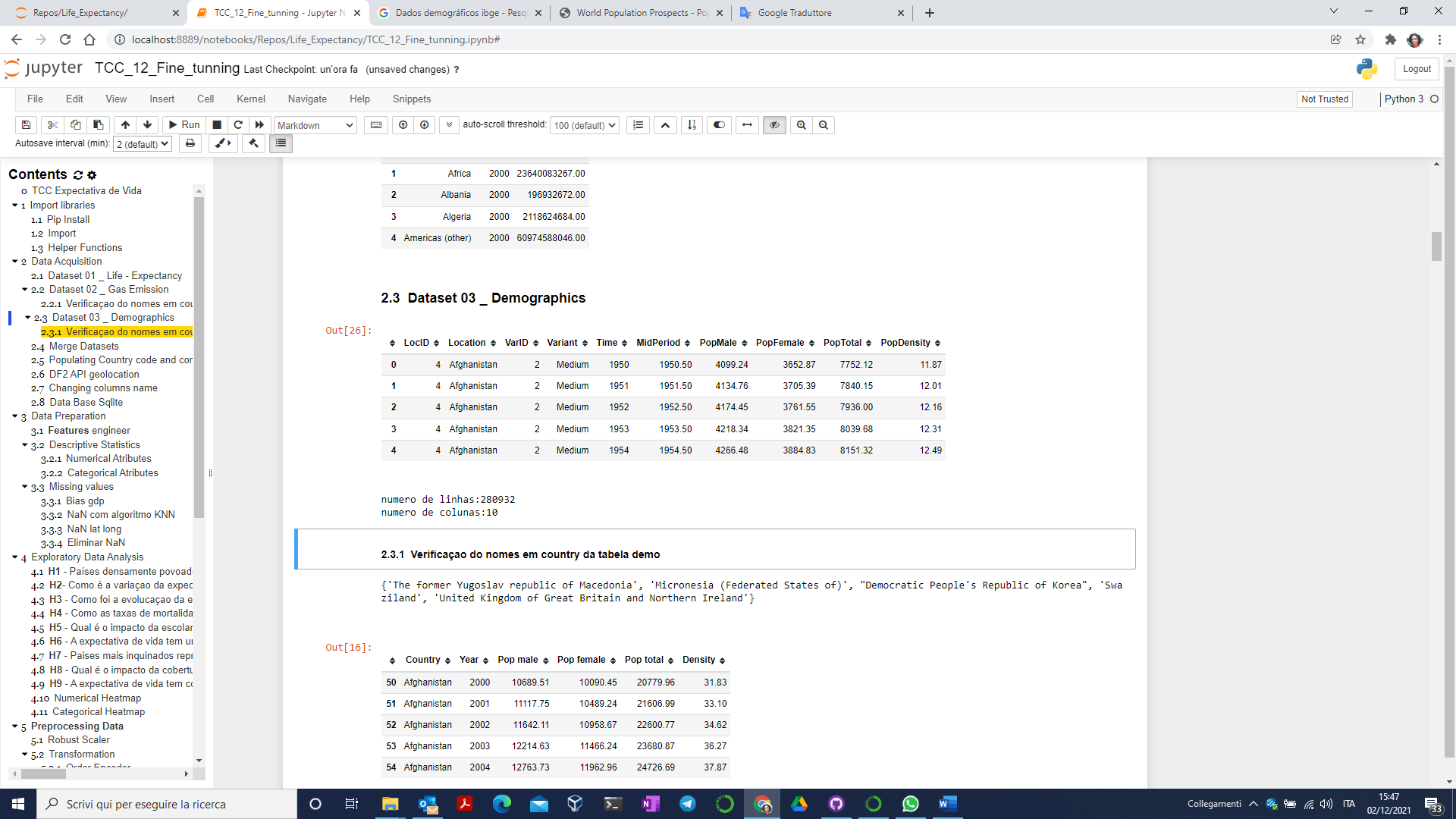


# Dataset 03\_ Dados Demograficos

Com esse dataset foram obtidos mais dados que segmentam a população dos 193 países selecionados. Foram coletadas as seguintes informaçoes:

* Populaçao Total do pais
* Populaçao por genere
* Densidade demografica

Essas informaçoes foram coletadas do site do Departamento de Assuntos Econômicos e Sociais (<https://population.un.org/wpp/Download/Standard/Population/>) das Naçoes Unidas.



# APIs de geolocalizaçao e nomes e codificaçao dos paises

# Data geolocation API Nominatim

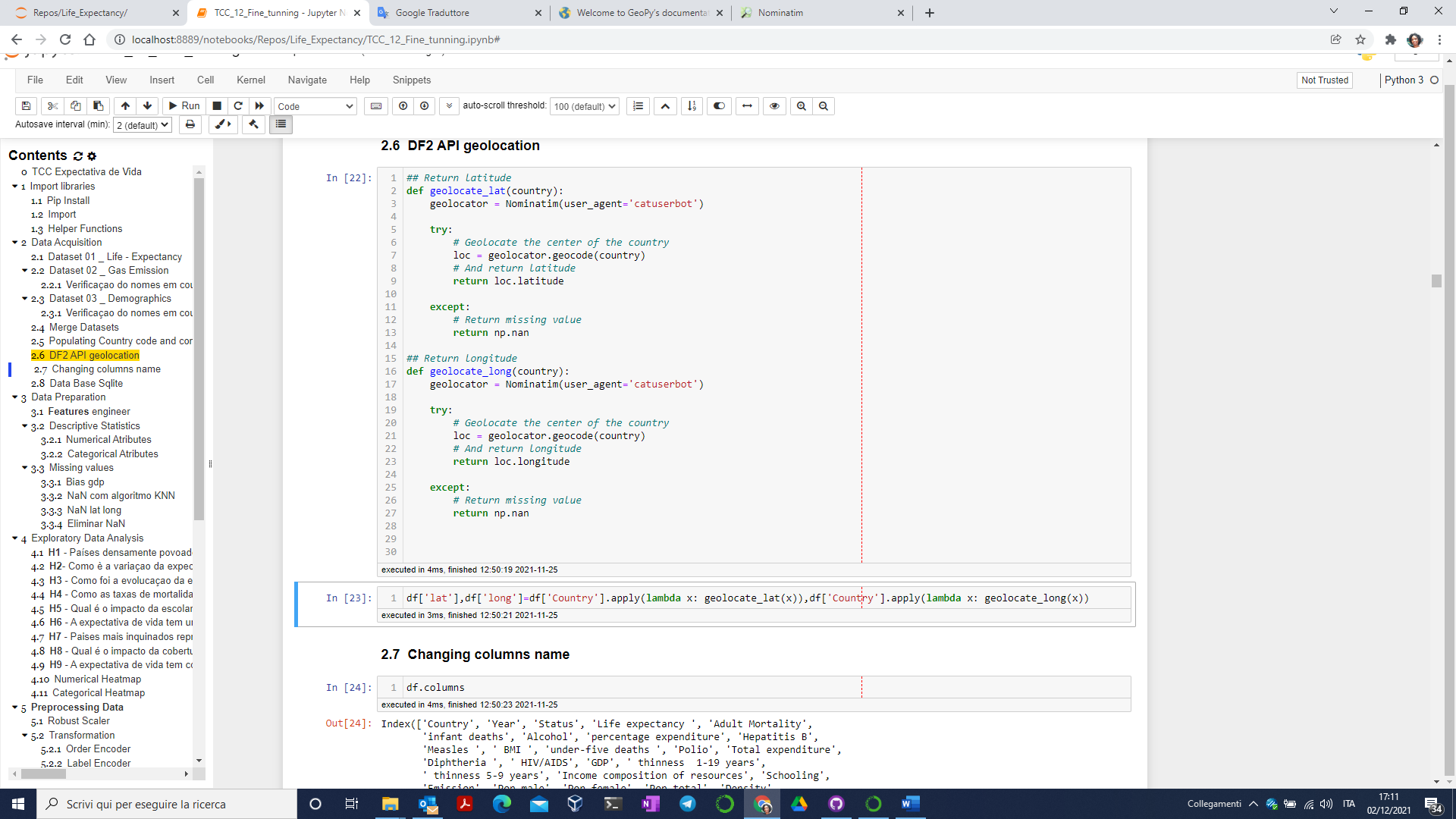
API Geolocation è um RESTful API (Application programming interface) que fornece informações de geolocalização para um endereço IP.

Uma API REST (também conhecida como API RESTful) é uma interface de programação de aplicativo (API ou API da web) que está em conformidade com as restrições do estilo arquitetônico REST e permite a interação com serviços da web RESTful. REST significa transferência de estado representacional e é um conjunto de restrições arquitetônicas criado pelo cientista da computação Roy Fielding.

Neste projeto foi usado uma API Restful denominada Nominatim que fornece serviço atraves da instalaçao da biblioteca Geopy (<https://geopy.readthedocs.io/en/stable/>).

Nominatim (do latim, 'por nome') é uma ferramenta para pesquisar dados por nome e endereço (geocodificação) e para gerar geocodificação reversa.

A intençao da utilizaçao dessa API era individuar os dados de latitute e longitude de cada pais com o objetivo de realizar uma analise grafica de mapatura .



# Data codigos dos paises e continentes[[4]](#endnote-1)

Neste projeto foi utilizado o pycountry\_convert que è uma Extensão do pacote Python pycountry que fornece funções de entre nomes de países ISO, códigos de países e nomes de continentes.

As funçoes utilizadas neste projetos foram country\_alpha2\_to\_continent\_code e country\_name\_to\_country\_alpha2 que devolvem informaçoes de codigos de paises e continentes.

O Padrão Internacional para códigos de país e códigos para suas subdivisões è a ISO 3166-1 alpha-2 standard to encode the country names. O objetivo da ISO 3166 é definir códigos de duas letras e / ou números reconhecidos internacionalmente que podemos usar quando nos referimos a países e suas subdivisões.

Ja a funçao country\_alpha2\_to\_continent\_code converte os codigos dos paises em nome de continentes.

# Processamento/Tratamento de Dados

~~Nessa seção você deve deixar registrado todo o processamento e tratamento feitos sobre os dados obtidos. É importante que você informe a quantidade de registros obtidos, a quantidade de registros duplicados ou com informações ausentes, que tratamento você deu para cada problema encontrado em seus datasets, etc. Você deve descrever cada passo de forma minuciosa, de forma que outra pessoa consiga reproduzir o seu processamento/tratamento de forma precisa. Justifique as decisões tomadas no tratamento dos dados. Por exemplo: para os valores ausentes para o campo X eu decidi preenchê-los utilizando o cálculo da média aritmética pelo motivo ..., e então justifique sua decisão.~~

Essa fase do projeto foi dividida em varias seçoes de forma a descrever com cura os dados obtidos.

# Criação de features

A criação de features uma das tecnicas de feature engineering envolve derivar novas features das existentes. Isso pode ser feito por operações matemáticas simples, como agregações para obter a média, mediana, modo, soma ou diferença e até mesmo o produto de dois valores. Nesse projeto foram derivadas 2 novas features: percentual de mulheres e emissao de gas por tamanho da populaçao. A ideia era alinear a taxa de emissao em cada pais e testar a hipotese que paises com um percentual maior de mulheres tem uma expectativa de vida maior.

Graphical user interface, text, application, email

Description automatically generated

# Informação do Database

O database final contem 32 colunas e 2938 registros .Graphical user interface, text, application, email

Description automatically generated

# Graphical user interface, application, table Description automatically generated

# A tabela info mostra tambèm o tipo de dado de cada coluna e o numero de registros nao nulos O tratamento dos dados nao nulos serà abordado com mais detalhe posteriormente. Análise e Exploração dos Dados

Atualmente, as expectativas de vida nos países desenvolvidos tendem a ser maiores que as dos países subdesenvolvidos, pois o desenvolvimento econômico desses países, na maioria das vezes, é acompanhado de altos padrões de vida, o que eleva a expectativa de vida da população. Isso, infelizmente, não é visto em países em que a vulnerabilidade social e econômica é elevada. Essa realidade pode ser notada nas taxas de expectativa de vida\* de países desenvolvidos e subdesenvolvidos do ano de 2015. Enquanto as expectativas de vida em países como Japão (89,79 anos), Suíça (82,50 anos), França (81,75 anos), Espanha (81,57 anos) e Alemanha (80,57 anos), todos países desenvolvidos, ficaram acima de 80 anos, nos países subdesenvolvidos, como Nigéria (53,02 anos), Moçambique (52,94) e Afeganistão (50,87), elas sequer chegaram aos 60 anos.

# 5. Criação de Modelos de Machine Learning

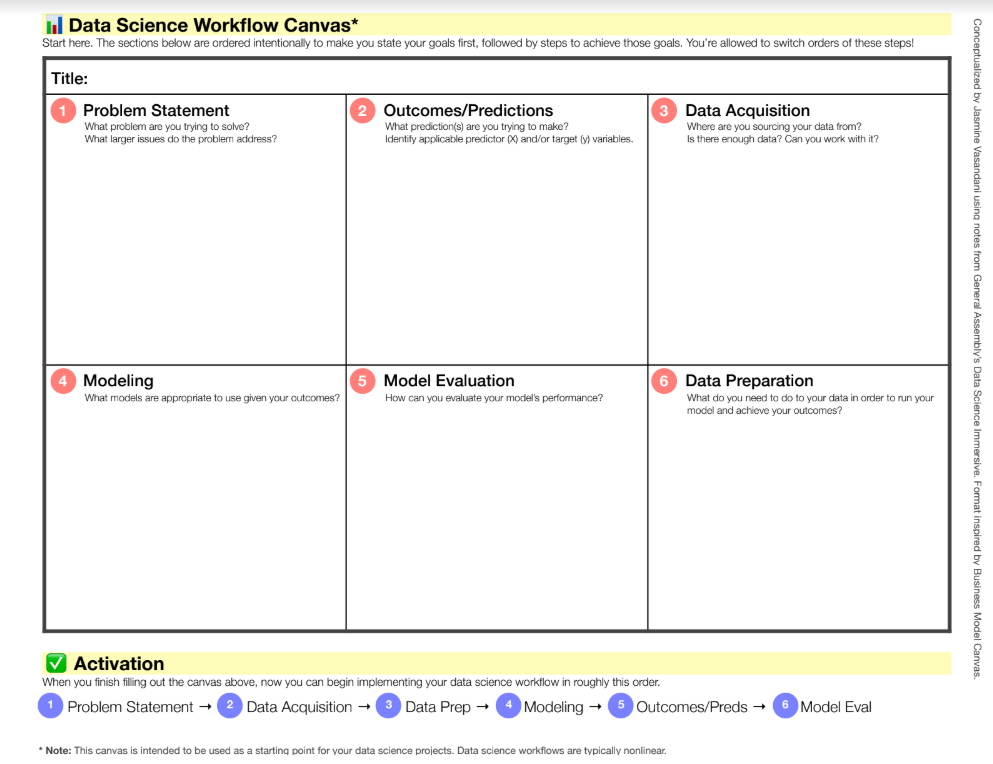
Conforme o documento de instruções para o TCC, essa etapa é obrigatória. Nessa seção você irá descrever as ferramentas e algoritmos utilizados. Se você utilizou o Knime, coloque aqui um print dos seus modelos e a descrição detalhada do workflow de cada modelo. Caso você tenha escrito scripts em Python ou R, por exemplo, coloque aqui apenas os trechos do código que você considera extremamente importantes para entendimento do seu trabalho. Explique as *features* utilizadas, faça a comparação entre diferentes algoritmos/modelos, justifique a escolha por determinado modelo, os parâmetros utilizados, etc. Por fim, salienta-se que embora você possa utilizar o KNIME para testar protótipos do seu modelo de ML, encorajamos você a fazer seus modelos em Python ou R.

# 6. Interpretação dos Resultados

Nessa seção você deve interpretar os resultados obtidos na análise e exploração de dados e também interpretar os resultados da aplicação dos algoritmos de Machine Learning, descobrindo insights importantes para responder o problema proposto.

# 7. Apresentação dos Resultados

Nessa seção você deve apresentar os resultados obtidos. Apresente gráficos, *dashboards*, conte a sua história de forma bastante criativa. Aqui você pode utilizar os modelos de Canvas propostos por Dourard (clique [aqui](https://www.louisdorard.com/machine-learning-canvas)) ou por Vasandani (clique [aqui](https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0)).



# 8. Links

Aqui você deve disponibilizar os links para o vídeo com sua apresentação de 5 minutos e para o repositório contendo os dados utilizados no projeto, scripts criados, etc.

Link para o vídeo: youtube.com/...

Link para o repositório: github.com/...

# REFERÊNCIAS

SILVA, Thamires Olimpia. "O que é expectativa de vida?"; Brasil Escola. Disponível em: https://brasilescola.uol.com.br/o-que-e/geografia/o-que-e-expectativa-vida.htm. Acesso em 23 de novembro de 2021.

David Robson, [*Why do women live longer than men?*](https://www.bbc.com/future/article/20151001-why-women-live-longer-than-men), su *www.bbc.com*. URL consultado il 20 março 2021

Warren Farrell, [*Warren Farrell - The Myth Of Male Power: Why Men Are The Disposable Sex*](http://archive.org/details/warren_farrell_the_myth_of_male_power_why_men_are_the_disposable_sex), 1993. URL consultado il 20 março 2021.

[*Perché le donne vivono più degli uomini*](https://www.lescienze.it/news/2015/07/08/news/differenze_genere_aspettiativa_vita-2681727/), su *Le Scienze*. URL consultado il 21 junho 2021.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

# APÊNDICE

**Programação/Scripts**

Cole aqui seus scripts em Python e/ou R.

**Gráficos**

Cole aqui workflows (KNIME), gráficos e figuras que você tenha gerado e não colocou no texto principal.

**Tabelas**

Cole aqui tabelas de dados que você tenha gerado e não colocou no texto principal.

1. David Robson, *[Why do women live longer than men?](https://www.bbc.com/future/article/20151001-why-women-live-longer-than-men)*, su *www.bbc.com*. URL consultado il 20 março 2021 [↑](#footnote-ref-1)
2. Warren Farrell, [*Warren Farrell - The Myth Of Male Power: Why Men Are The Disposable Sex*](http://archive.org/details/warren_farrell_the_myth_of_male_power_why_men_are_the_disposable_sex), 1993. URL consultado il 20 março 2021. [↑](#footnote-ref-2)
3. *[Perché le donne vivono più degli uomini](https://www.lescienze.it/news/2015/07/08/news/differenze_genere_aspettiativa_vita-2681727/)*, su *Le Scienze*. URL consultado il 21 junho 2021. [↑](#footnote-ref-3)
4. <https://pycountry-convert.readthedocs.io/_/downloads/en/latest/pdf/>

   <https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2> [↑](#endnote-ref-1)