



Aluno:

Alessandra de Assis Barbosa

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

**PREVISÃO DA EXPECTATIVA DE VIDA DA  
POPULAÇÃO UTILIZANDO ALGORITMOS DE  
MACHINE LEARNING A PARTIR DE  
INFORMAÇÕES SOCIOECONÔMICAS**

Belo Horizonte  
2022

# *Mas o que é a expectativa de vida?*

Expressão usada para indicar o número médio de anos que cada indivíduo provavelmente viverá caso sejam mantidas as mesmas condições vivenciadas no momento do nascimento.



# *Por que este projeto?*



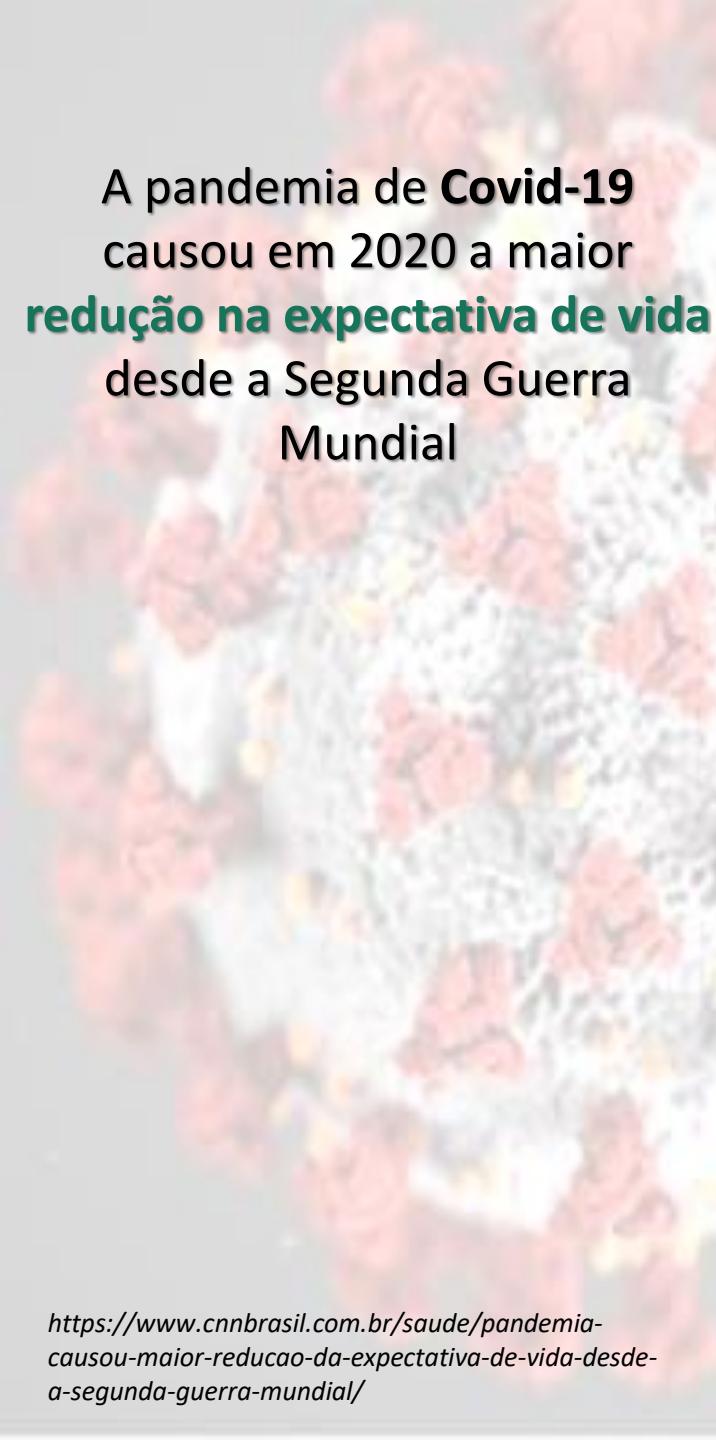
Trabalho em organização  
nao governamental que se  
ocupa da **saúde, qualidade**  
**de vida e direitos dos**  
**aposentados e idosos.**

E' muito comum que  
exista uma **preocupação**  
com a **diminuição da**  
**expectativa de vida e**  
sobre que **iniciativa ou**  
**projetos investir.**

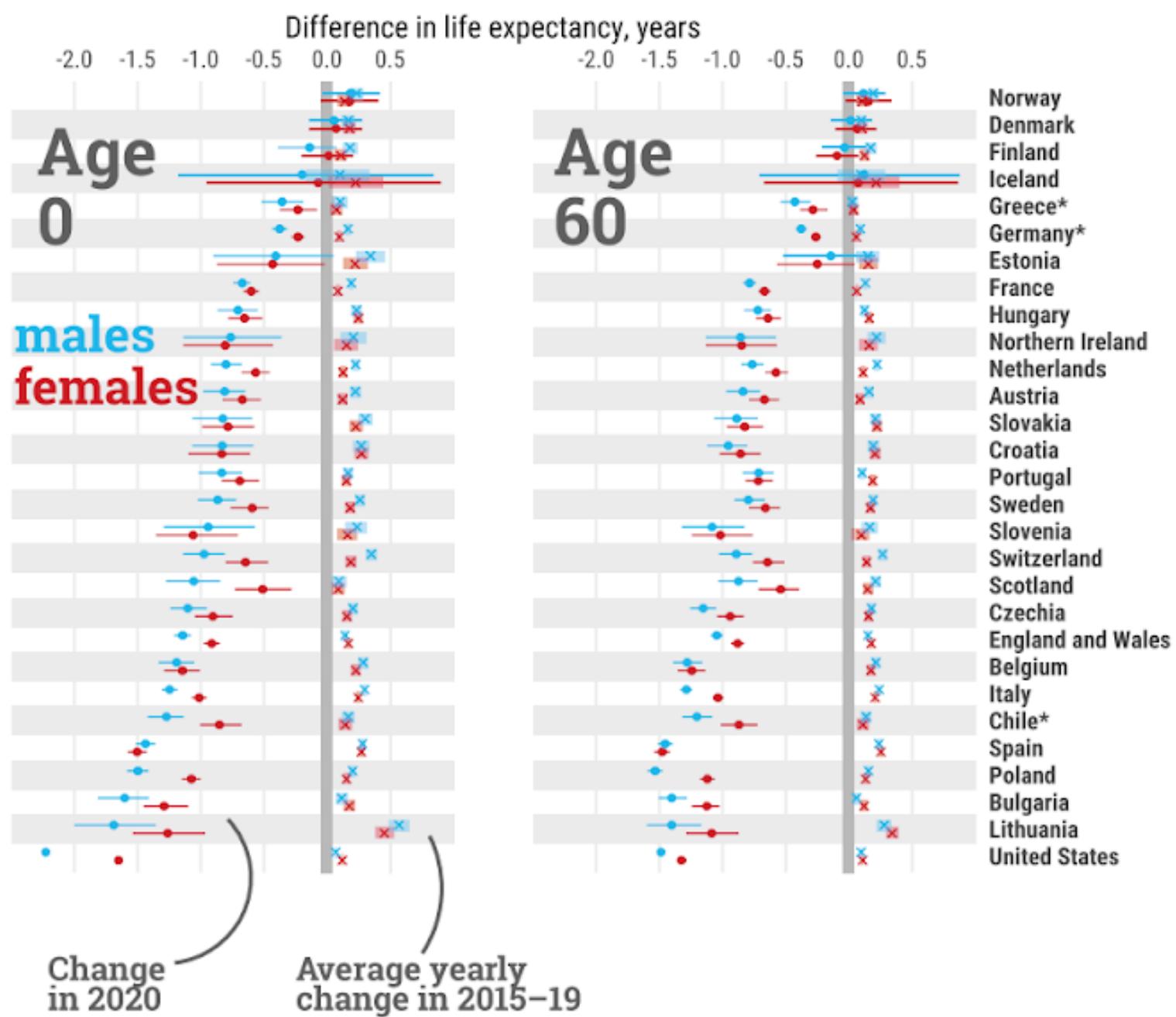
Outras organizações e outras associações também acompanham o estado de saúde dos países e utilizam **tecnicas estatisticas** para estudar a expectativa de vida e o **impacto das pandemias na vida da população.**

Tanto que si é visto durante a **pandemia iniciada no ano 2019 a prospectiva de um abaixamento da estimativa de vida mundial desencadeando uma corrida na tomada de providências para reduzir os danos.**



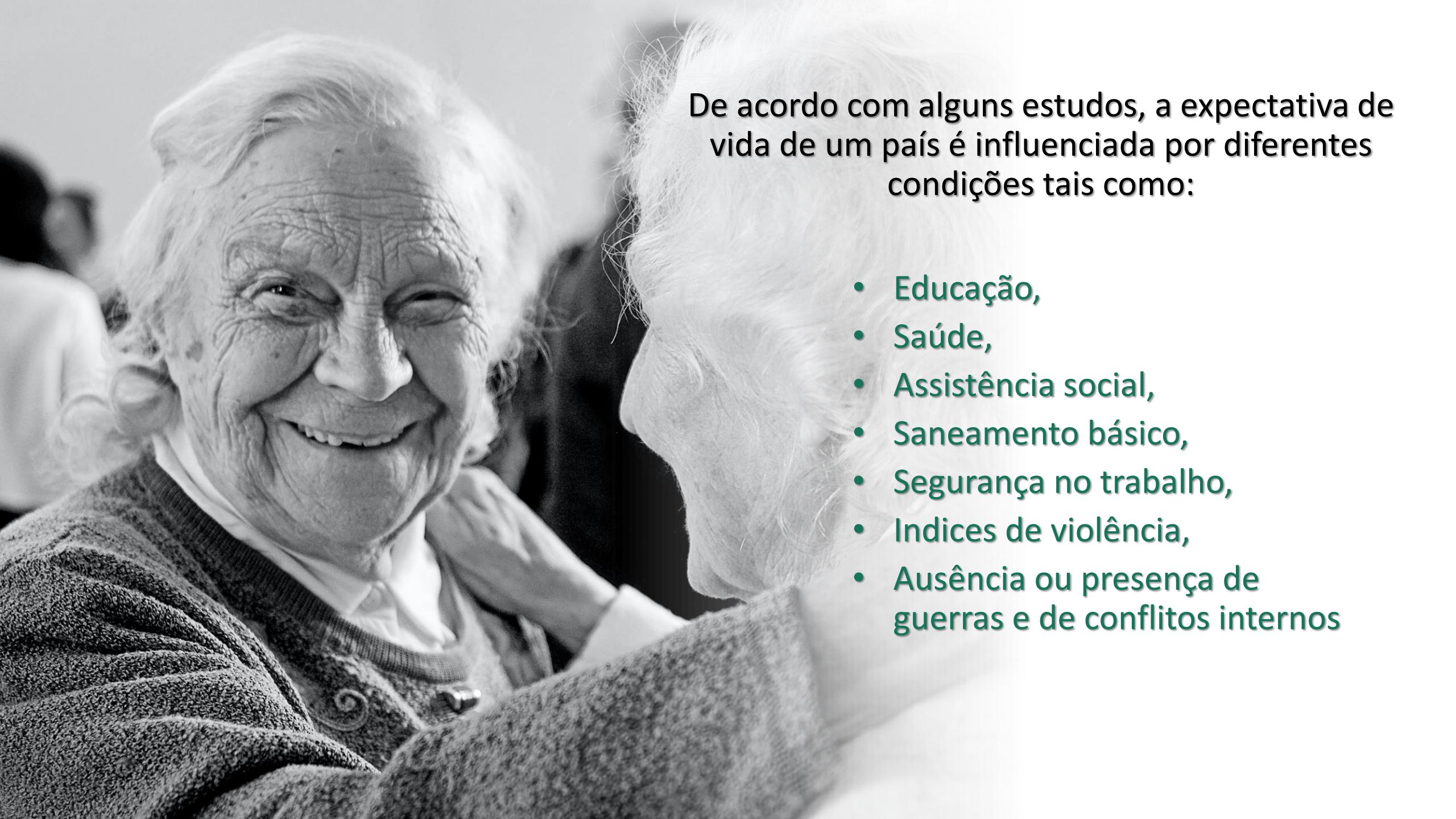


A pandemia de Covid-19 causou em 2020 a maior redução na expectativa de vida desde a Segunda Guerra Mundial



<https://www.cnnbrasil.com.br/saude/pandemia-causou-maior-reducao-da-expectativa-de-vida-desde-a-segunda-guerra-mundial/>

Figure 2 Average per-year change in life expectancy at birth (age 0) and age 60 years, by country and sex, from 2015 to 2019, and total change from 2019 to 2020. Estimates for females (red), males (blue), average per-year changes from 2015 to 2019 are depicted by the symbol (x), dots depict the to-



De acordo com alguns estudos, a expectativa de vida de um país é influenciada por diferentes condições tais como:

- Educação,
- Saúde,
- Assistência social,
- Saneamento básico,
- Segurança no trabalho,
- Índices de violência,
- Ausência ou presença de guerras e de conflitos internos

# *Objetivos do trabalho*

O objetivo deste trabalho é realizar uma análise exploratória que permitirá de individuar os fatores correlacionados a redução ou aumento da expectativa de vida e identificar um modelo matemático que possa ser usado para a previsão da expectativa de vida

O objetivo maior deste projeto não é somente prever a expectativa de vida nos países mas  
identificar fatores diretamente vinculados



# Além disso esse trabalho também 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? ?
- Países mais poluídos apresentam uma expectativa de vida menor? ?

## *Ferramentas utilizadas*



 python™

 scikit  
**learn**

 jupyter

 seaborn



# *Etapas do Trabalho*

---

**Coleta dos dados**

---

**Processamento / Tratamento dos dados**

---

**Análise e Exploração dos Dados**

---

**Criação de Modelos de Machine Learning**

---

**Interpretação dos Resultados**

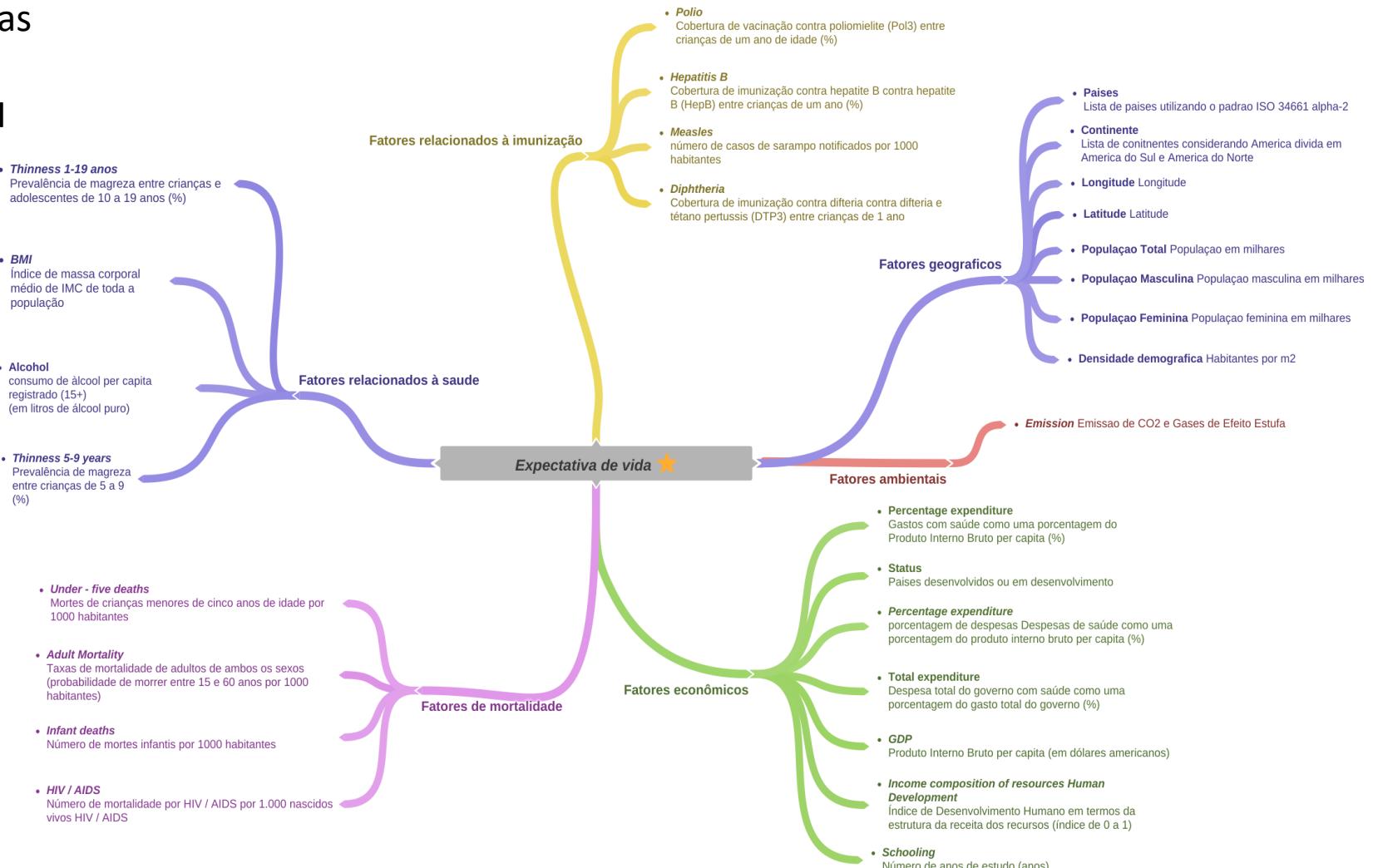
# Coleta de Dados



❑ **Fontes:** órgão mundial da saúde e das nações unidas, dados do IBGE, organização Our World in Data e API de geolocalização.

❑ **Grandes áreas:** economia, ambientais, demográficos, mortalidade, imunização, saúde

❑ **Período:** 2000 à 2015



# Processamento Tratamento dos dados



## ➤ Criação de features

```
#female percent  
df1['perc_female']=df1['pop_female']/df1['pop_total']  
  
# emission per population tax  
df1['emission_pop']=df1['emission']/df1['pop_total']
```

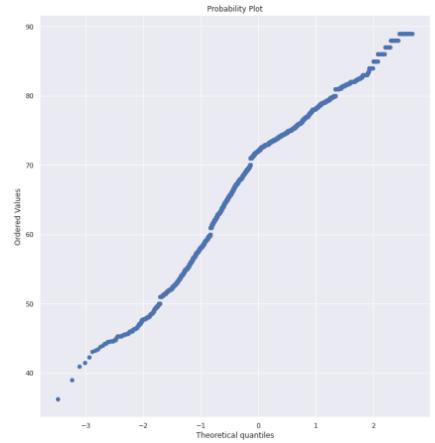
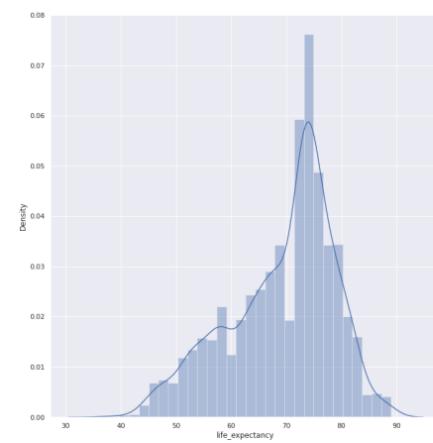
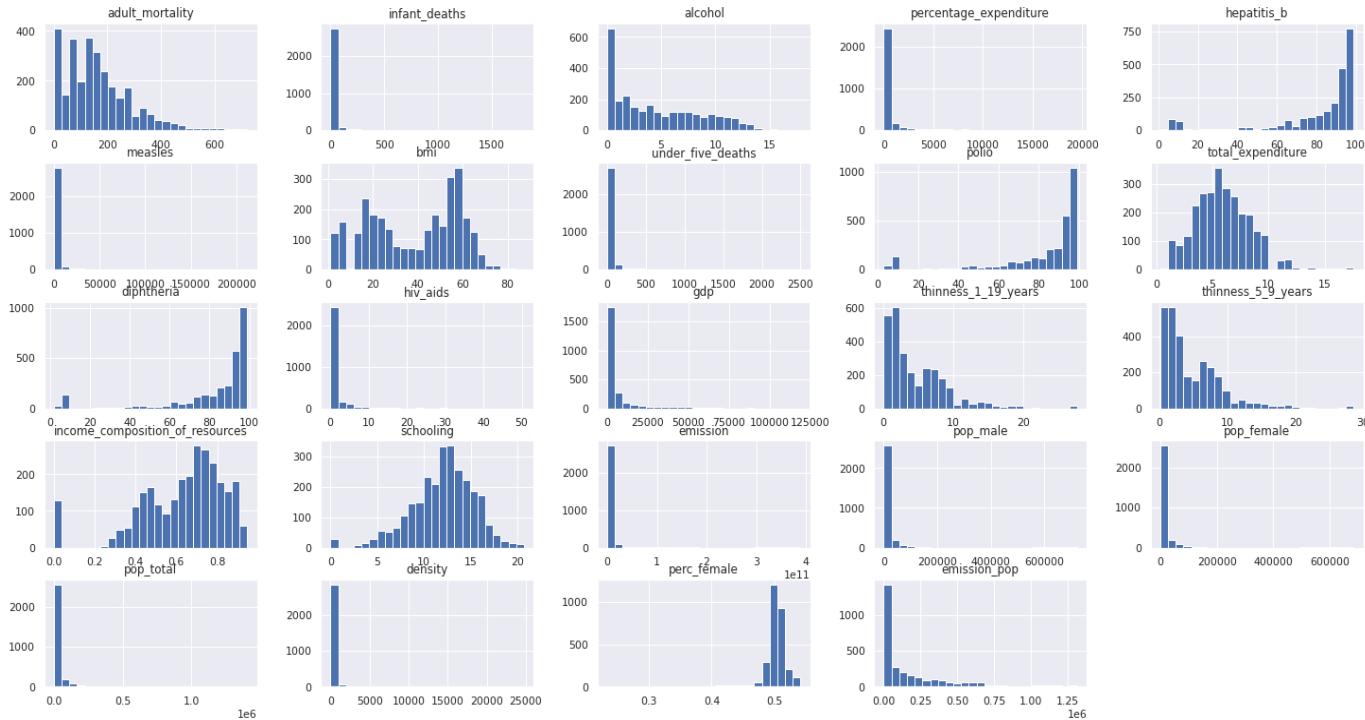
## ➤ Informação do Database

```
RangeIndex: 2938 entries, 0 to 2937  
Data columns (total 32 columns):  
 #   Column           Non-Null Count  Dtype    
 ---  --    
 0   country          2938 non-null   object   
 1   year              2938 non-null   int64   
 2   status             2938 non-null   object   
 3   life_expectancy  2928 non-null   float64  
 4   adult_mortality  2928 non-null   float64  
 5   infant_deaths    2938 non-null   int64   
 ..  ...  ..  ..  ..
```

# Processamento Tratamento dos dados



## ➤ Estatística descritiva



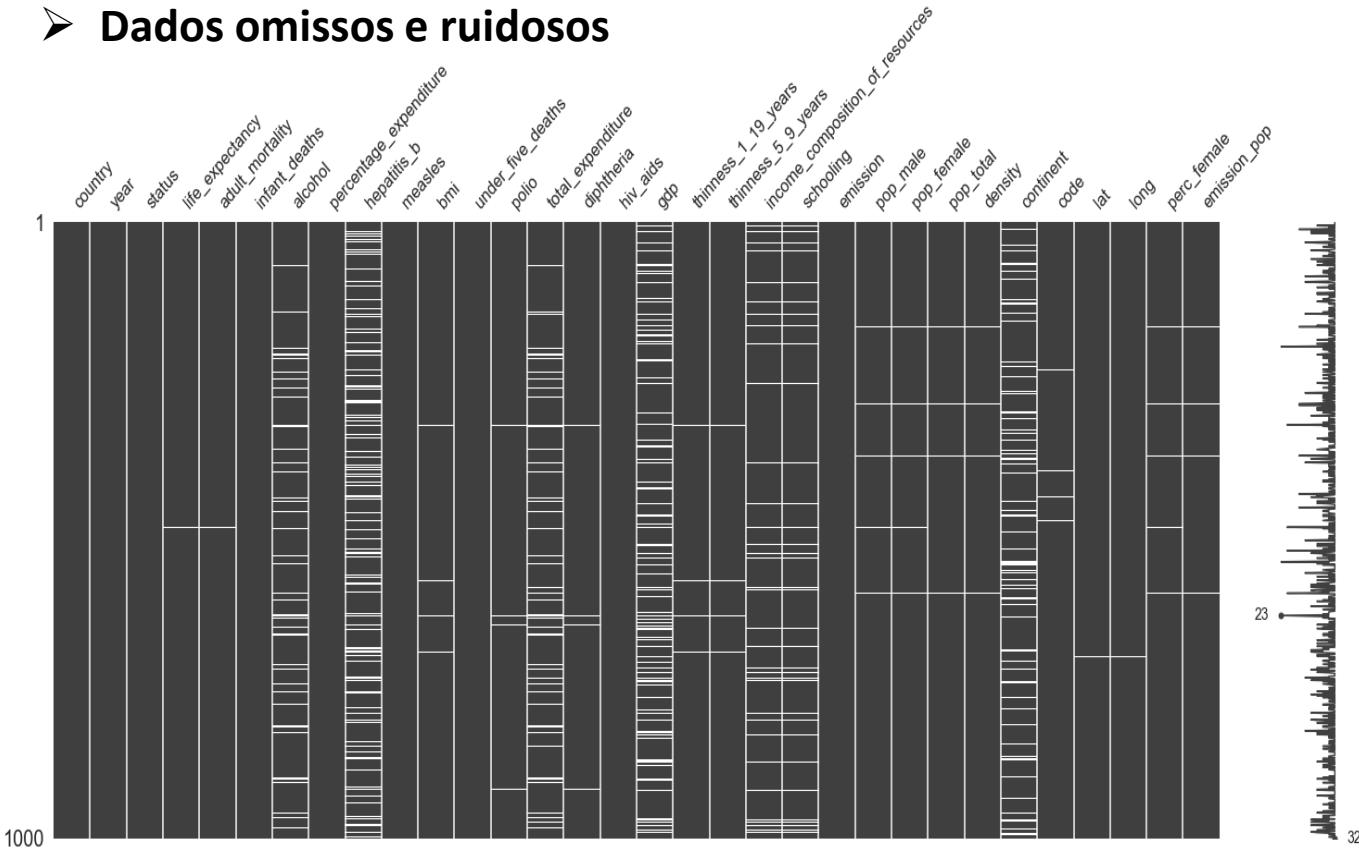
```
cat_attributes.apply( lambda x: x.unique().shape[0] )
```

```
: country      193
status        2
continent     6
code         188
dtype: int64
```

# Processamento Tratamento dos dados



## ➤ Dados omissos e ruidosos



- ✓ Nova coleta de dados
- ✓ Estimação dos valores utilizando algoritmo KNN Input
- ✓ Eliminação de linhas do DB com dados faltantes

# Processamento Tratamento dos dados

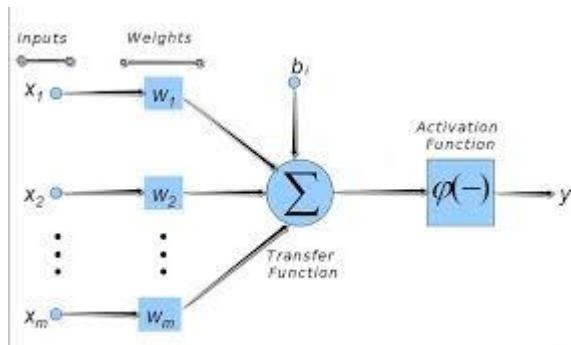


	Nan	Nan %
hepatitis_b	553	18.82
gdp	448	15.25
continent	338	11.50
total_expenditure	226	7.69
alcohol	194	6.60
income_composition_of_resources	167	5.68
schooling	163	5.55
thinness_1_19_years	34	1.16
bmi	34	1.16
thinness_5_9_years	34	1.16
perc_female	26	0.88
pop_female	26	0.88
pop_male	26	0.88
polio	19	0.65
diphtheria	19	0.65
lat	18	0.61
emission_pop	18	0.61
pop_total	16	0.54
density	16	0.54
code	16	0.54
long	16	0.54
adult_mortality	10	0.34
life_expectancy	10	0.34

**PIB(GDP)** - Decidiu-se por coletar novamente o dado de uma outra fonte mais confiável e por substituir os dados anteriores.

**% Despesa com a saude (Percentage expenditure)** - Decidiu-se por eliminar a feature dado que a variável total\_expenditure pode ser considerada como uma sua derivante.

**Alcohol, Despesa com a saùde, Indice de desenvolvimento, Escolaridade, Hepatitis B -**  
Utilizamos o algoritmo KNN (K-Nearest Neighbor) que tem como objetivo determinar a qual grupo uma determinada variavel vai pertencer com base nas amostras vizinhas.

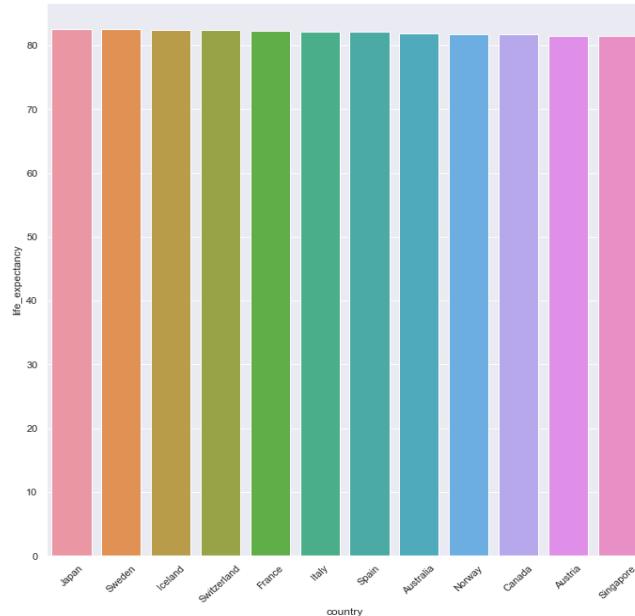
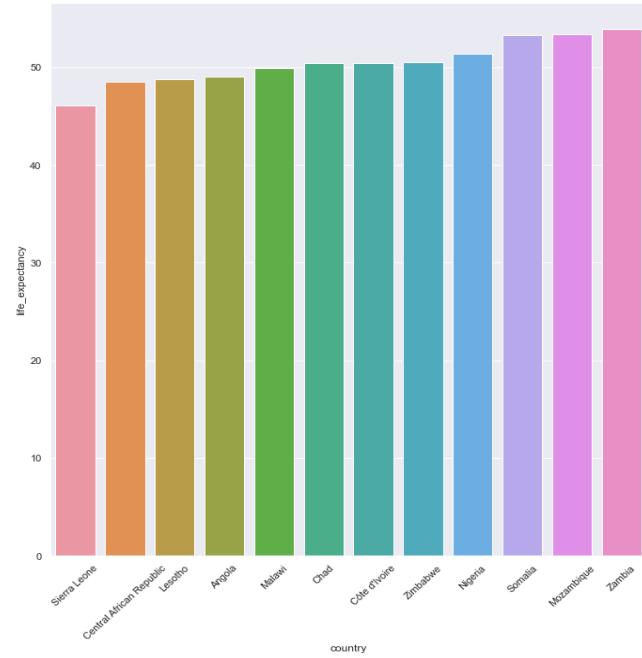


NaN com algoritmo KNN

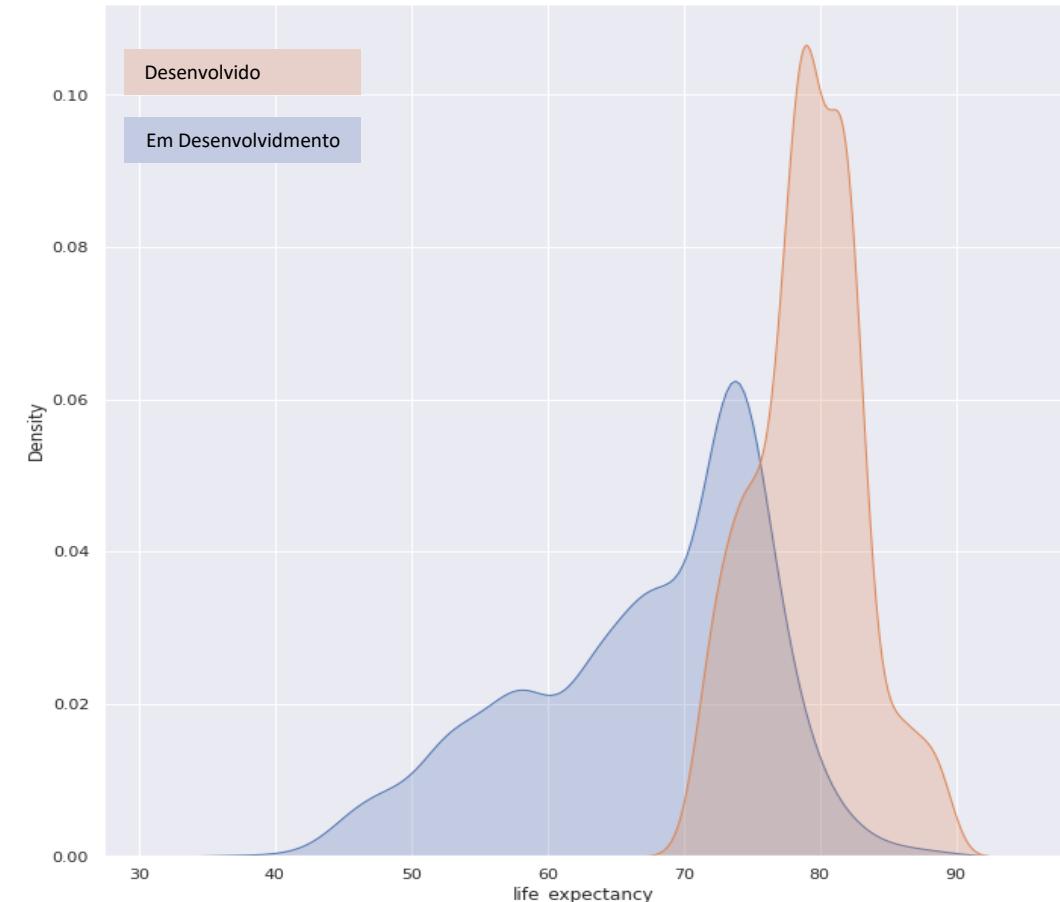
```
[1]: knn_imputer = KNNImputer(n_neighbors=5, weights="uniform", metric='nan_euclidean')
df1['hepatitis_b'] = knn_imputer.fit_transform(df1[['hepatitis_b']])
df1['alcohol'] = knn_imputer.fit_transform(df1[['alcohol']])
df1['total_expenditure'] = knn_imputer.fit_transform(df1[['total_expenditure']])
df1['income_composition_of_resources'] = knn_imputer.fit_transform(df1[['income_composition_of_resources']])
df1['schooling'] = knn_imputer.fit_transform(df1[['schooling']])
```

# Análise e Exploração dos Dados

Países com expectativa de vida mais baixa e mais alta



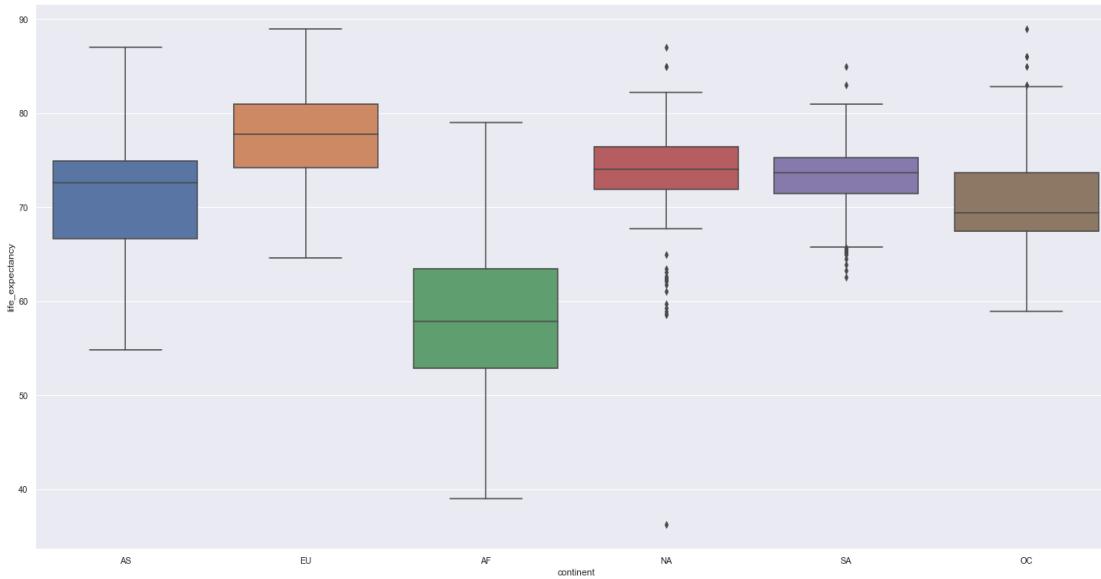
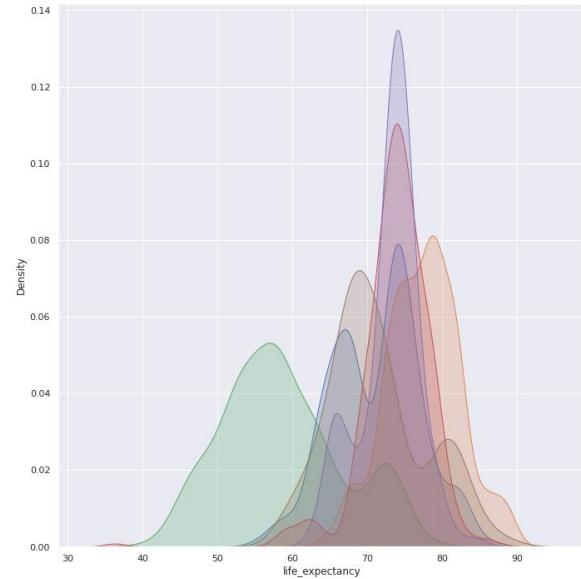
Expectativa de vida nos países desenvolvidos e em desenvolvimento



# Análise e Exploração dos Dados



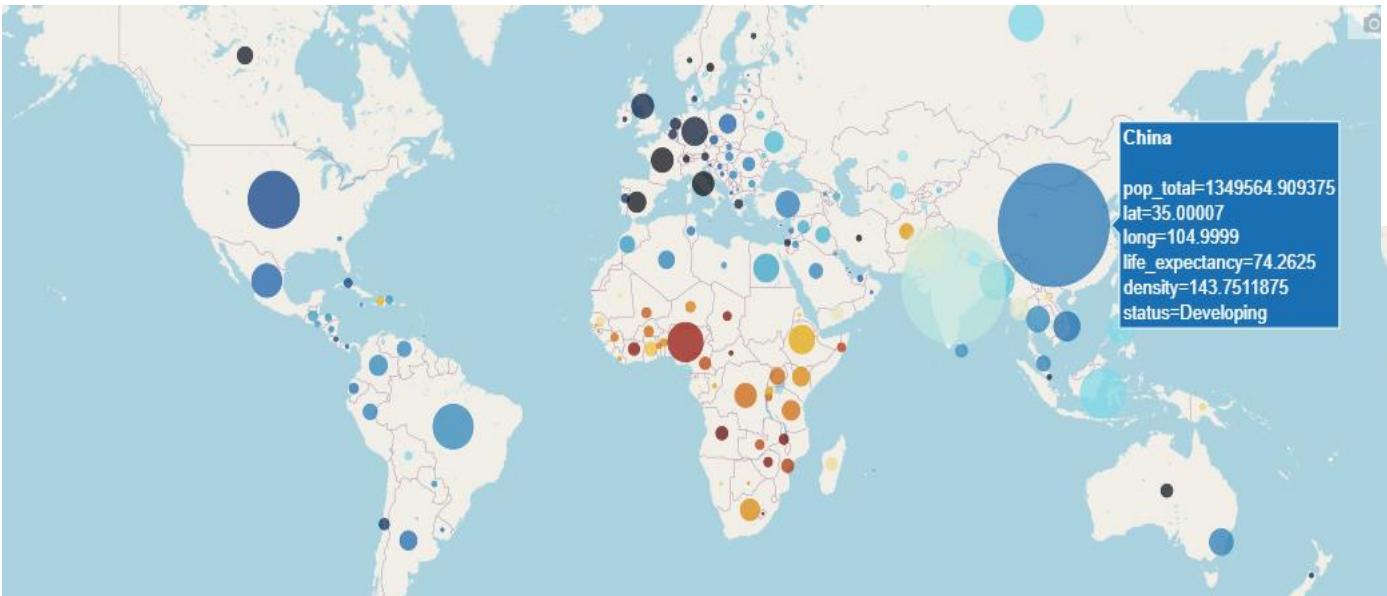
Como é a variação da expectativa de vida dentro dos continentes?



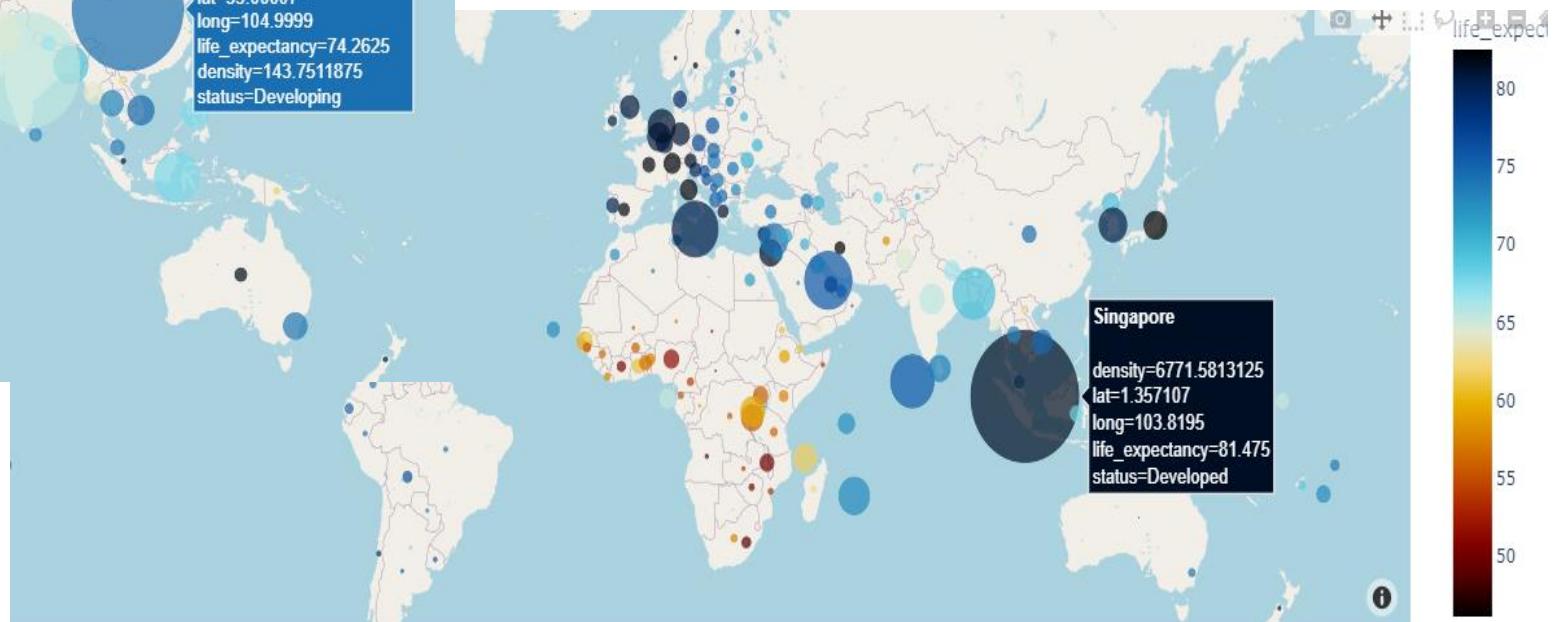
# Análise e Exploração dos Dados



Países densamente povoados ou altamente populosos tendem a ter menor expectativa de vida?



Países mais populosos

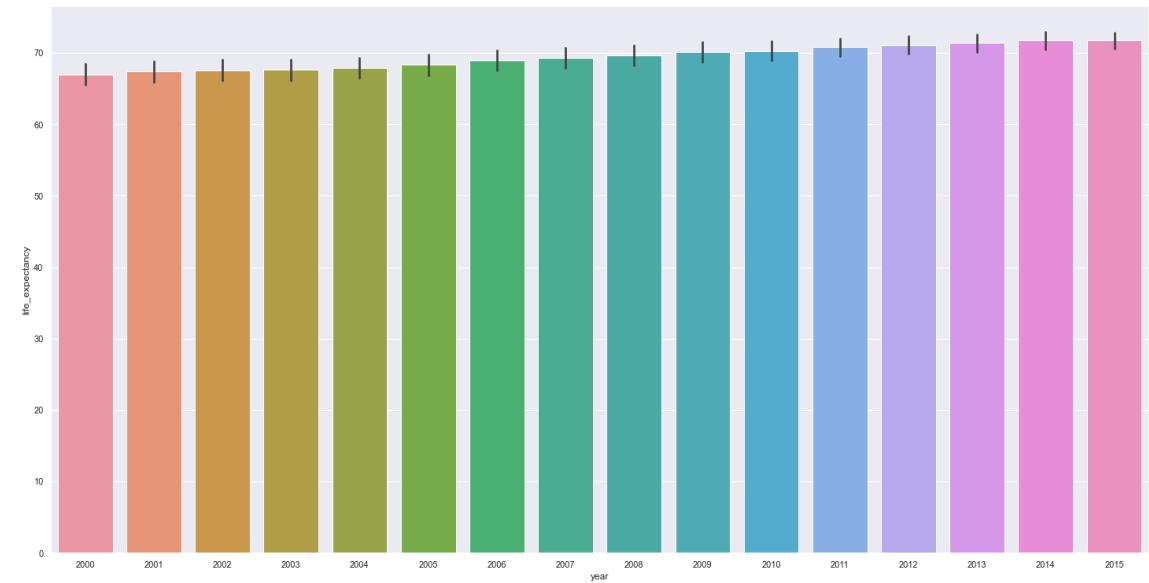
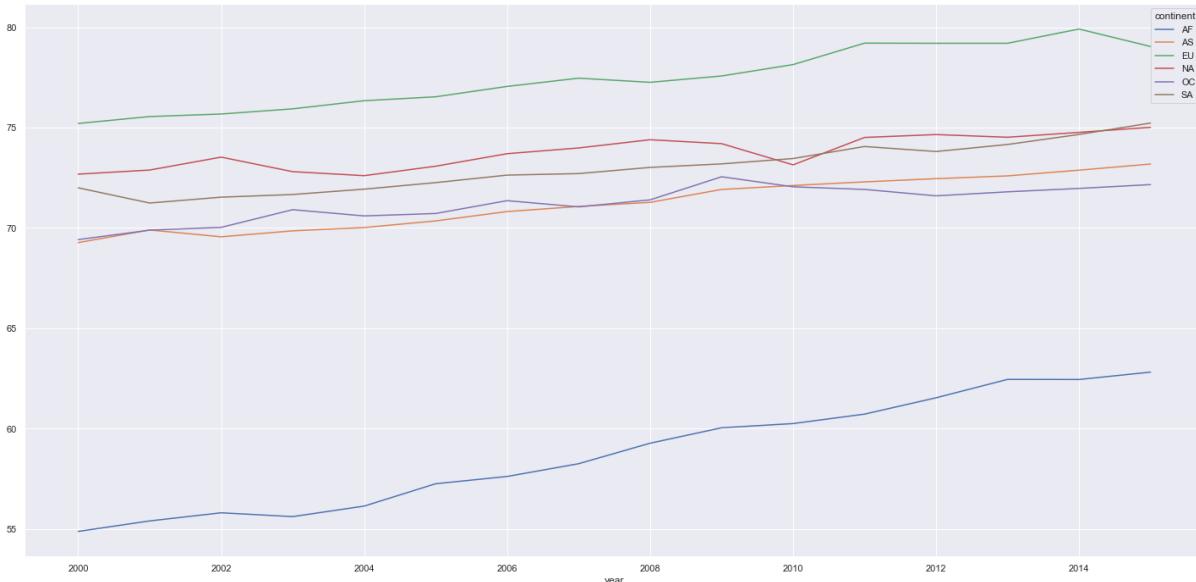


Países mais povoados

# Análise e Exploração dos Dados



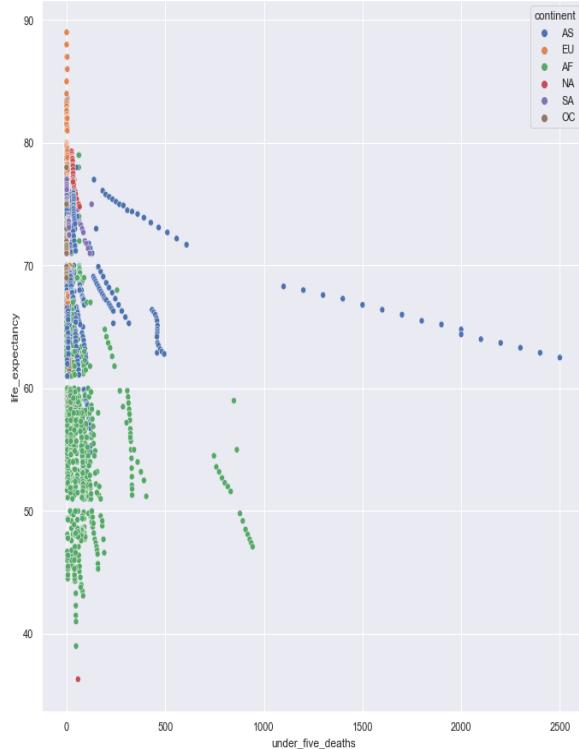
Como foi a evolução da expectativa de vida ao longo dos anos?



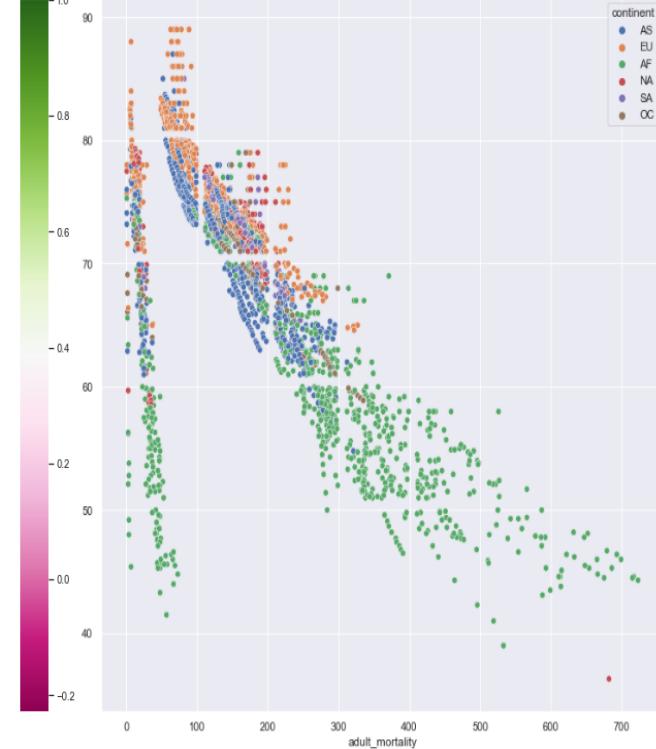
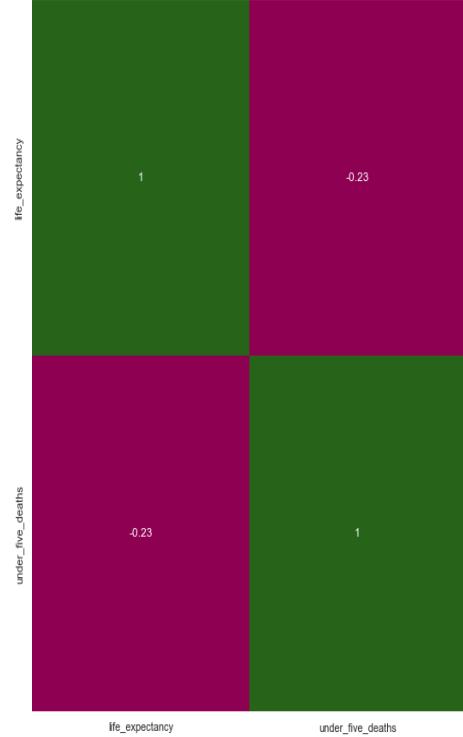
# Análise e Exploração dos Dados



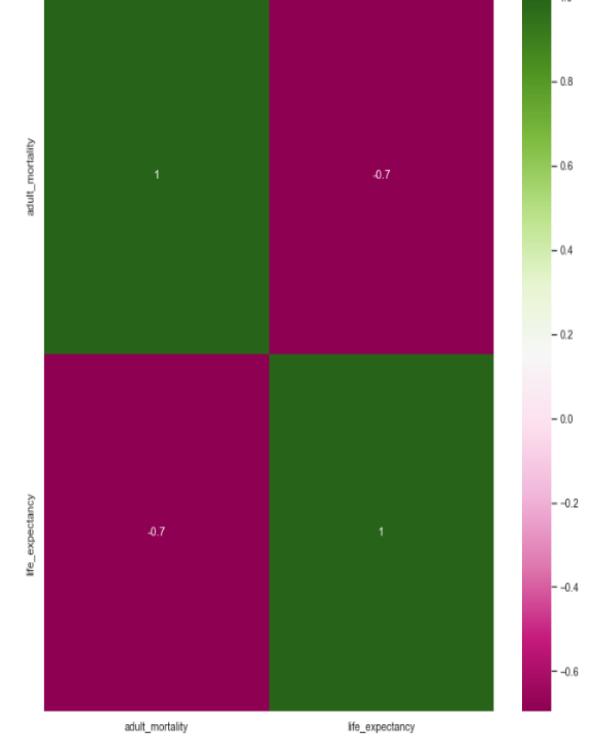
Como as taxas de mortalidade de bebês e adultos afetam a expectativa de vida?



Entre 5 anos de vida



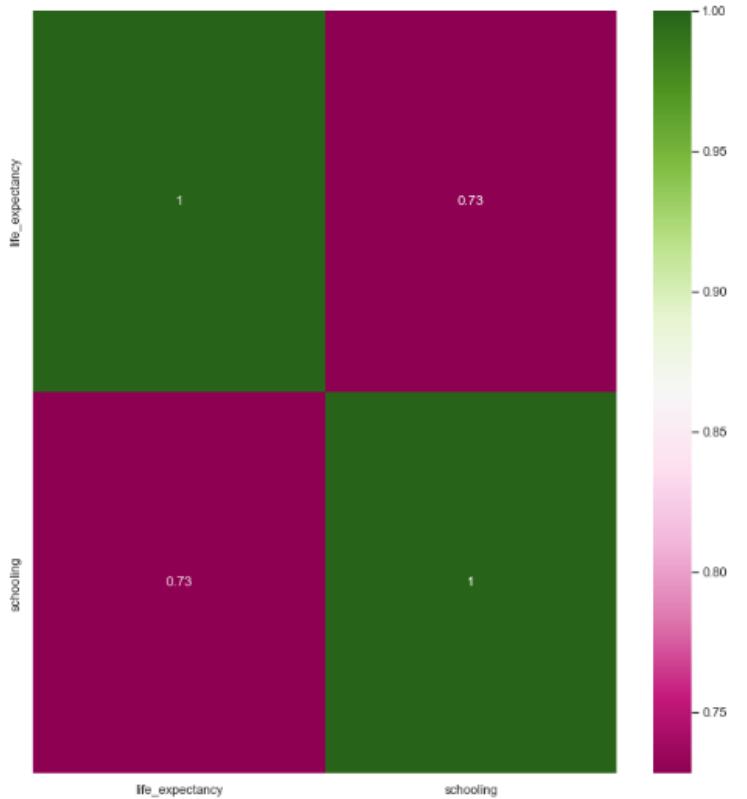
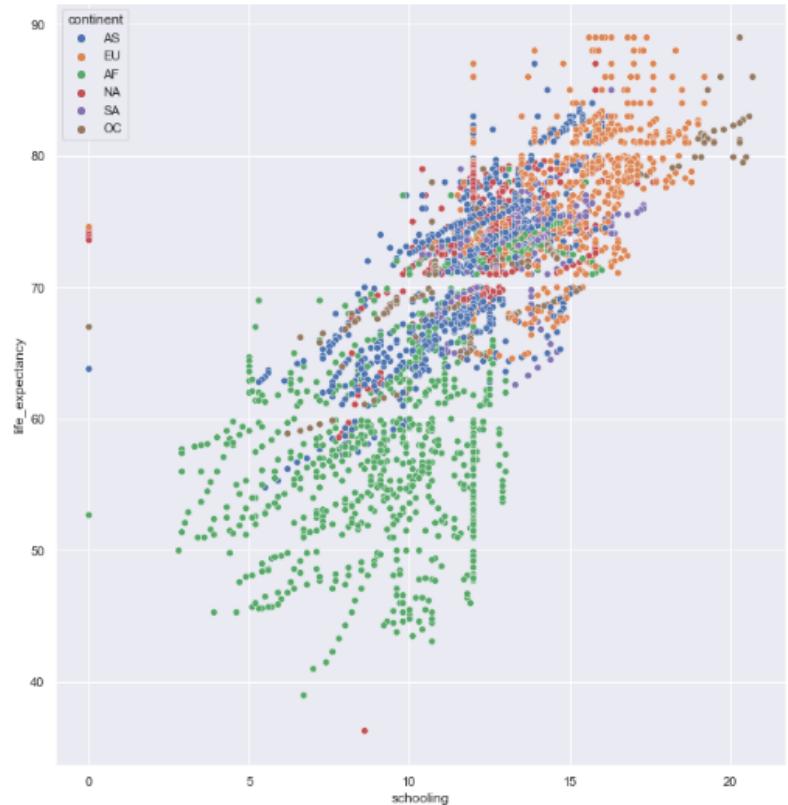
Entre 15 e 49 anos de vida



# Análise e Exploração dos Dados



Qual é o impacto da escolaridade na expectativa de vida?



Anos de estudo

# Análise e Exploração dos Dados



Qual é o impacto da cobertura de imunização na expectativa de vida?



Hepatite, polio, difteria - Cobertura de vacinação entre crianças de um ano de idade (%)

Sarampo - Número de casos de sarampo notificados por 1000 habitantes

# Criação de Modelos de Machine Learning



## Pré-processamento dos dados

### Robust Scaler

Utiliza a diferença interquartilica para rescalar variáveis com larga amplitude e outliers

- emission
- gdp
- infant\_deaths
- percentage\_expenditure
- measles
- under\_five\_deaths
- pop\_male
- pop\_female
- pop\_total
- density
- perc\_female
- emission\_pop

### order encoding

- year

### label encoding

- status
- continent
- code
- country

### Transformação logarítmica da variável resposta

- life-expectancy

# Criação de Modelos de Machine Learning



## Separação da base de dados em treino e teste

- primeiros 13 anos como treino (relativo a 81% dos dados) e os demais 3 anos como teste

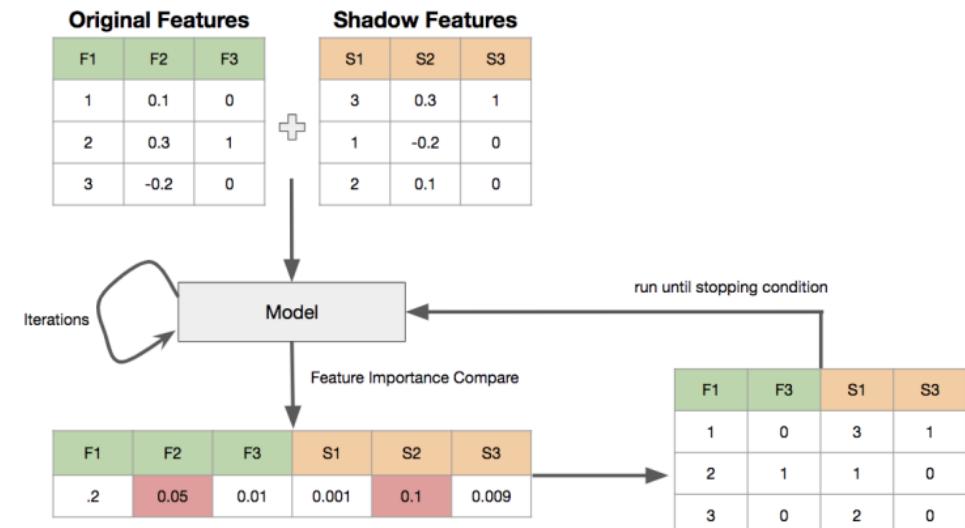
```
Training Min Date: 1  
Training Max Date: 13  
  
Test Min Date: 14  
Test Max Date: 16
```

## Seleção das variáveis

### Algoritmo Boruta - seleção por subset com shadows

#### Variáveis não inclusas no modelo:

- Sarampo
- Status
- Código
- Latitude
- Longitude
- População total
- Percentual mulheres



# Criação de Modelos de Machine Learning



$$MAE = \frac{1}{n} \sum \left| y - \hat{y} \right|$$

Divide by the total number of data points  
Sum of  
Actual output value  
Predicted output value  
The absolute value of the residual

$$MAPE = \frac{100\%}{n} \sum \left| \frac{\hat{y} - y}{y} \right|$$

Multiplying by 100% converts to percentage  
The residual  
Each residual is scaled against the actual value

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

- **Modelo de Regressão Linear (Modelo Base)**

Score Train: 0.8390243732916304  
Score Test: 0.8043404439389107

Model Name	MAE	MAPE	RMSE
Linear Regression	2.85	0.04	3.73

'income_composition_of_resources	0.05992
'infant_deaths	0.02676
'under_five_deaths	-0.02633
'gdp	0.01424
'continent	0.01292
'schooling	0.01000
'hiv_aids	-0.00884
'pop_male	0.00366
'pop_female	-0.00223
'density	0.00204

- **Modelo de Regressão Linear – Lasso**

Score Train: 0.8206015610639548  
Score Test: 0.7982589525916054

Model Name	MAE	MAPE	RMSE
Linear Regression - Lasso	2.83	0.04	3.74

'schooling	0.01324
'hiv_aids	-0.00870
'gdp	0.00728
'continent	0.00508
'pop_male	0.00244
'density	0.00178
'under_five_deaths	-0.00170
'bmi	0.00085
'diphtheria	0.00075
'polio	0.00049

- **Modelo de Regressão Linear – Ridge**

Score Train: 0.8390243730879412  
Score Test: 0.8043392371916586

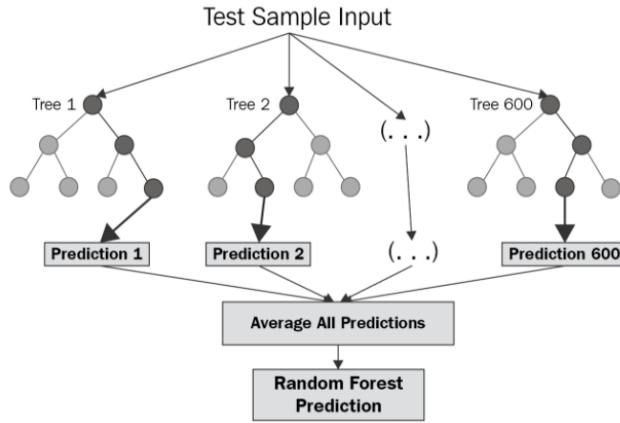
Model Name	MAE	MAPE	RMSE
Linear Regression - Ridge	2.85	0.04	3.73

'income_composition_of_resources	0.05852
'infant_deaths	0.02665
'under_five_deaths	-0.02622
'gdp	0.01426
'continent	0.01291
'schooling	0.01005
'hiv_aids	-0.00884
'pop_male	0.00350
'pop_female	-0.00204
'density	0.00204

# Criação de Modelos de Machine Learning



## • Modelo de Random Forest



```
# model
rf = RandomForestRegressor( bootstrap= True,
                            criterion='mse',
                            min_samples_leaf= 1,
                            min_samples_split= 2,
                            n_estimators=100,
                            n_jobs=1,
                            ).fit( x_train, y_train )
```

```
# prediction
yhat_rf = rf.predict( x_test )
```

```
# print( 'Score Train: {}'.format( rf.score(x_train,y_train) ) )
print( 'Score Test: {}'.format( rf.score(x_test,y_test) ) )
```

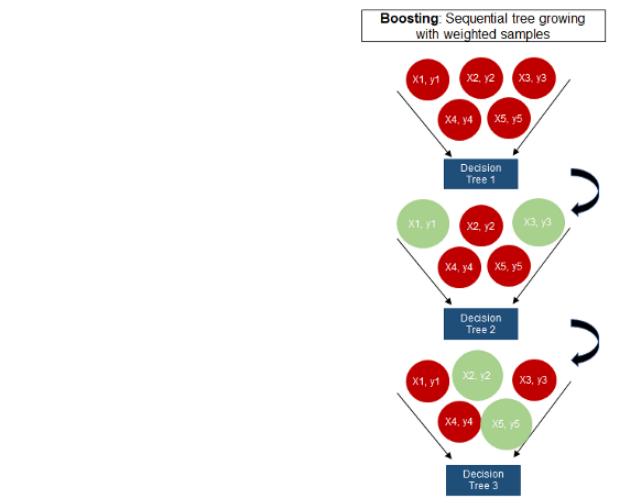
```
# performance
rf_result = ml_error( 'Random Forest Regressor', np.expm1( y_test ), np.expm1( yhat_rf ) )
rf_result
```

```
Score Train: 0.9954453690320629
Score Test: 0.918272800101161
```

5]:

	Model Name	MAE	MAPE	RMSE
0	Random Forest Regressor	1.49	0.02	2.31

## • Modelo XG Boost



```
# model
model_xgb = xgb.XGBRegressor( objective='reg:squarederror',
                               n_estimators=100,
                               max_depth=5,
                               subsample=0.7
                               ).fit( x_train, y_train )
```

```
# prediction
yhat_xgb = model_xgb.predict( x_test )
```

```
# performance
xgb_result = ml_error( 'XGBoost Regressor', np.expm1( y_test ), np.expm1( yhat_xgb ) )
xgb_result
```

7]:

	Model Name	MAE	MAPE	RMSE
0	XGBoost Regressor	1.74	0.03	2.49

# Interpretação dos Resultados



## ➤ Cross validation – Métrica real (K-fold igual a 3)

	Model Name	MAE CV	MAPE CV	RMSE CV
0	Random Forest Regressor	1.63 +/- 0.092	0.03 +/- 0.001	2.49 +/- 0.104
0	XGBoost Regressor	1.77 +/- 0.023	0.03 +/- 0.0	2.56 +/- 0.015
0	Linear Regression - Lasso	3.13 +/- 0.134	0.05 +/- 0.003	4.1 +/- 0.157
0	Linear Regression	3.17 +/- 0.208	0.05 +/- 0.004	4.12 +/- 0.178
0	Linear Regression - Ridge	3.17 +/- 0.207	0.05 +/- 0.004	4.12 +/- 0.178

# Interpretação dos Resultados



## ➤ Fine Tuning - Random Search

```
KFold Number: 1
{'n_estimators': 218, 'max_features': 'auto', 'max_depth': 78, 'min_samples_split': 2, 'min_sample

KFold Number: 3

KFold Number: 2

KFold Number: 1
```

Out[174]:

	Model Name	MAE CV	MAPE CV	RMSE CV
0	model_RandomForest	2.11 +/- 0.051	0.03 +/- 0.001	3.12 +/- 0.079
0	model_RandomForest	1.66 +/- 0.024	0.03 +/- 0.001	2.46 +/- 0.027
0	model_RandomForest	1.63 +/- 0.081	0.03 +/- 0.001	2.49 +/- 0.111
0	model_RandomForest	1.56 +/- 0.036	0.02 +/- 0.001	2.37 +/- 0.068
0	model RandomForest	1.68 +/- 0.064	0.03 +/- 0.001	2.54 +/- 0.096

## ➤ Modelo Final – Random Forest

```
# model
rf = RandomForestRegressor(n_estimators= 1345,
                           max_features= 'sqrt',
                           max_depth= 110,
                           min_samples_split= 2,
                           min_samples_leaf= 1,
                           bootstrap= False
                           ).fit( x_train, y_train )

# prediction
yhat_rf = rf.predict( x_test )

print( 'Score Test: {}'.format( rf.score(x_test,y_test) ) )

# performance
rf_result = ml_error( 'Random Forest Regressor', np.expm1( y_test ), np.expm1( yhat_rf ) )
rf_result
```

Score Test: 0.9360534676411818

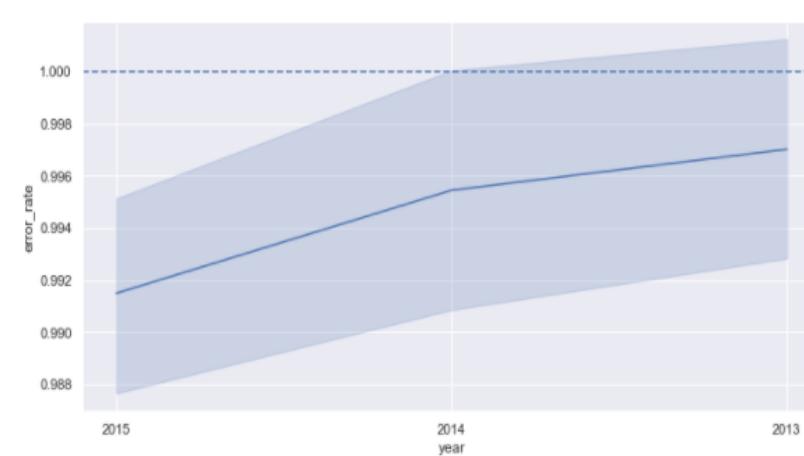
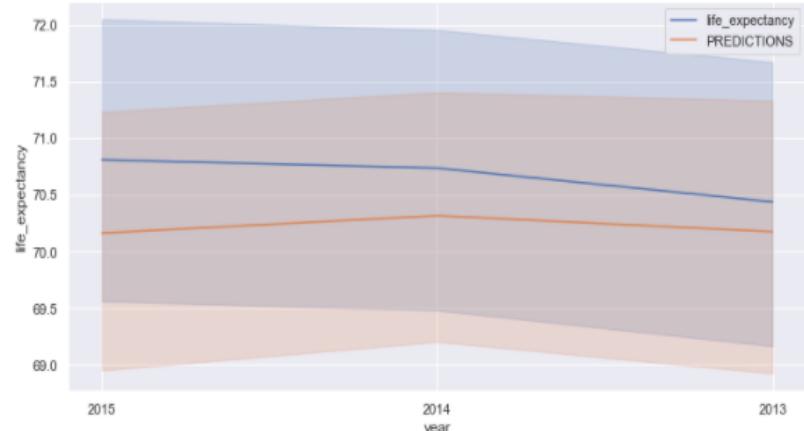
!]:

	Model Name	MAE	MAPE	RMSE
0	Random Forest Regressor	1.35	0.02	2.10

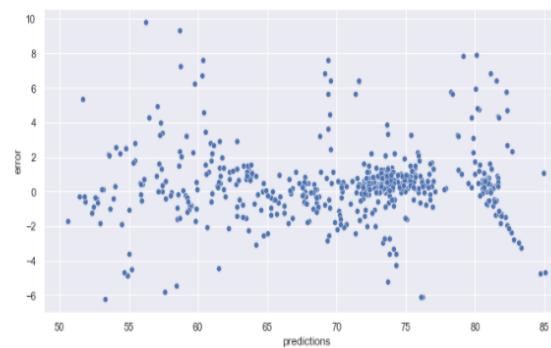
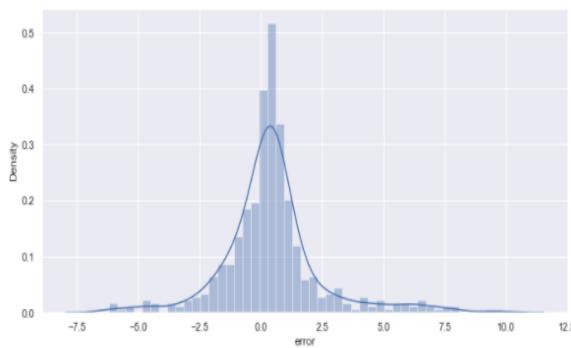
# Interpretação dos Resultados



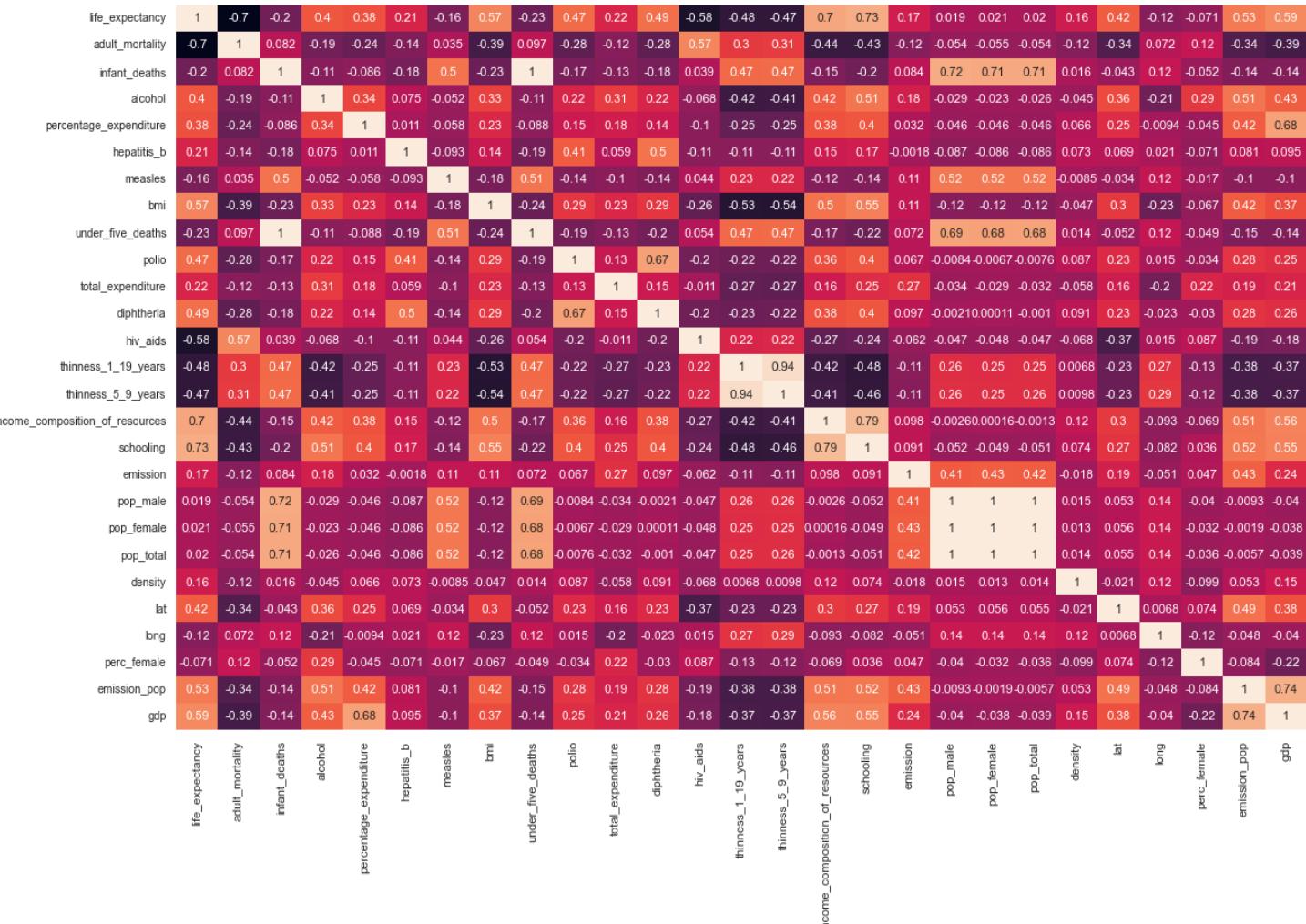
## ➤ Modelo final - Random Forest



## ➤ Anàlise de resíduos



# Conclusão



## Modelo final

Random Forest erro médio absoluto de 1,35 anos

## Variáveis não relevantes para o modelo:

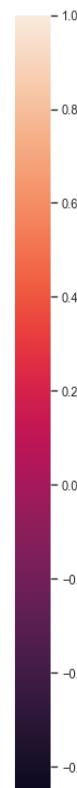
- Sarampo
- Status
- População total
- Percentual de mulheres

## Correlações mais pertinentes:

- Anos de estudo
- Mortalidade adulta
- Magreza ou mal nutrição
- HIV- Aids
- Morte com menos de 5 anos
- Imunização Difteria e Polio

## Outras considerações :

- Continente com maior necessidade de investimentos: Africano





Aluno:

Alessandra de Assis Barbosa

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

**Fim**

**Obrigada**

Belo Horizonte  
2022