Coimbra
Business School
Politécnico de Coimbra

# Unidade Curricular de Data Mining & Machine Learning
# Mestrado em Análise de Dados e Sistemas de Apoio à Decisão
# Ano Letivo 2024-2025

## Análise de fatores que influenciaram a saúde até 2020


Autores

**Catarina Auxiliar Nº 2021134297**

**Diogo Machado Nº 2020153309**

**João Leal Nº 2021130506**


Elaborado em

**08/01/2025**

# Análise de fatores que influenciaram a saúde até 2020

## Índice de figuras

# Lista de siglas e acrónimos

| | |
|---|---|
| **AD** | Árvore de Decisão |
| **CRISP-DM** | *Cross Industry Standard Process for Data Mining* |
| **CSV** | *Coma Separated Values* |
| **DBSCAN** | *Density-Based Spatial Clustering of Applications with Noise* |
| **OMS** | Organização Mundial da Saúde |
| **SDG** | *Sustainable Development Goals* |
| **WHR** | *World Health Report* |

# Análise de fatores que influenciaram a saúde até 2020

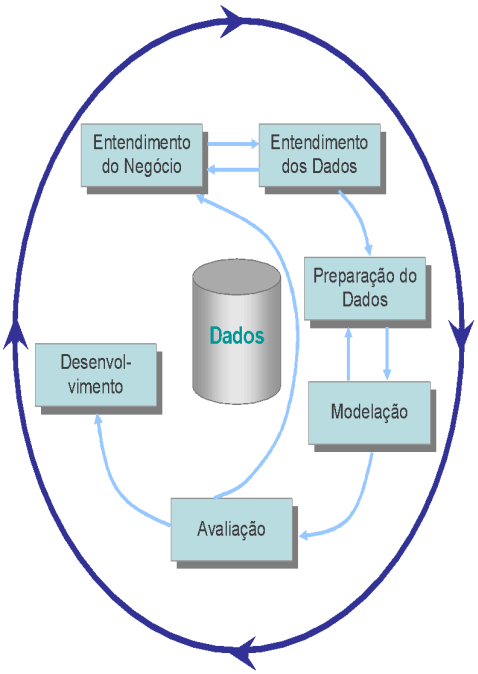
# Introdução

A saúde é um conceito amplo que envolve o equilíbrio e o bem-estar físico, mental e social. É um fator que tem impacto sobre a qualidade de vida dos indivíduos e existem vários fatores que influenciam esse equilíbrio, entre eles os hábitos de vida individuais, a poluição e até as condições do sistema de saúde dos países. É um assunto que afeta todos os seres vivos, logo é necessário ter um entendimento abrangente sobre o tema, de forma a perceber de que forma se pode melhorar as condições de saúde individual e coletiva. Deste modo, a Organização Mundial da Saúde (OMS) foi criada com o objetivo de direcionar e coordenar a saúde internacional dentro do sistema das Nações Unidas.

No âmbito da unidade curricular de Data Mining & Machine Learning do Mestrado em Análise de Dados e Sistemas de Apoio à Decisão do Instituto Superior de Contabilidade e Administração de Coimbra foi desenvolvido este projeto, com o objetivo principal de explorar padrões, relações e tendências em dados de saúde globais. Mais especificamente, os objetivos deste projeto são: **1)** Analisar quais são os fatores predominantes que afetam a saúde na população no mundo; **2)** Perceber se os fatores que vão ser analisados são comparativos entre eles; **3)** Analisar de uma forma holística o estado da saúde no mundo.

Além da presente introdução, este relatório está organizado em 5 seções principais, de acordo com a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), que identifica as diferentes fases na implantação de um projeto de mineração de dados. Sendo assim, as fases do projeto são: 1. Entendimento do tema, 2. Estudo dos dados; 3. Preparação dos dados; 4. Modelação; e 5. Avaliação. A figura 1 representa as fases do ciclo de vida da metodologia CRISP-DM.

Figura 1: Fases do ciclo de vida da metodologia CRISP-DM



# 1 Entendimento do tema

## 1.1 Introdução ao Tema

O ***World Health Report*** (WHR) é um relatório de saúde anual que compila os indicadores que influenciam o nível de saúde entre os seus 194 Estados Membros. Este relatório é desenvolvido pela OMS e o objetivo é sumarizar padrões na esperança média de vida e causas de morte, bem como analisar o avanço das metas de desenvolvimento sustentável relacionadas com a saúde (*Sustainable Development Goals* (SDG)).

Desde a sua primeira edição em 1995, o relatório tem desempenhado um papel essencial ao destacar os principais desafios de saúde pública, propor estratégias baseadas em evidências e incentivar debates entre formuladores de políticas, profissionais de saúde e a sociedade civil.

Cada edição do relatório concentra-se em um tema central, como desigualdades em saúde, sistemas de saúde, doenças infecciosas emergentes, mudanças climáticas e saúde mental. Esses temas refletem os problemas mais urgentes enfrentados pela humanidade e servem como um guia para governos e organizações internacionais na formulação de políticas e alocação de recursos.

Por meio de dados estatísticos, análises de políticas e estudos de caso, o *WHR* não apenas descreve o estado da saúde no mundo, mas também oferece recomendações práticas para melhorar a qualidade de vida e reduzir disparidades. Este documento é um reflexo do compromisso da OMS em promover o direito universal à saúde e criar um mundo onde todos tenham acesso a serviços de saúde de qualidade.

## 1.2 Avaliação da Situação Atual

O principal objetivo deste trabalho é prever a variável dependente “Probabilidade de falecer por doenças cardiovasculares, cancerígenas, etc”, através das demais variáveis independentes selecionadas, sendo que para isso iremos utilizar um algoritmo de regressão linear múltipla, conforme a figura - modelo genérico de regressão.

Com o resultado desta técnica, pretende-se também avaliar a relação da variável.

## 1.3 Definição dos Objetivos do Data Mining

O principal objetivo deste trabalho é prever a variável dependente “Probabilidade de falecer por doenças cardiovasculares, cancerígenas, etc”, através das demais variáveis independentes selecionadas, sendo que para isso vai ser utilizado um algoritmo de regressão linear múltipla, conforme a figura 2 - modelo genérico de regressão.

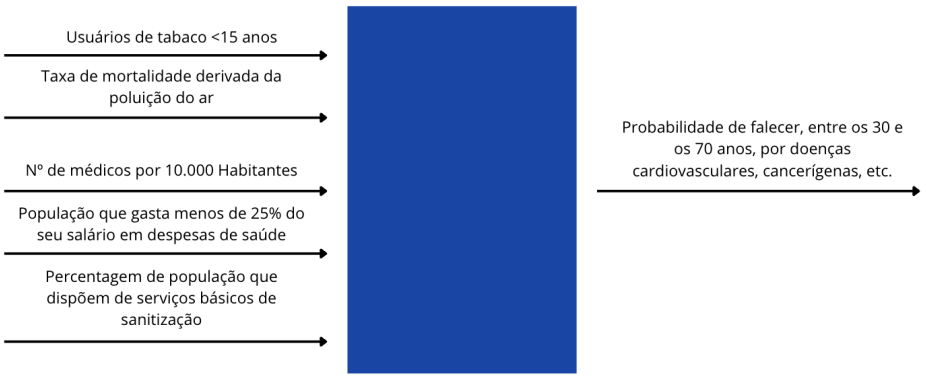



Figura 2: Modelo genérico de regressão

Esta análise procura identificar as condições que mais afetam a população global. Pretende-se compreender, não apenas os fatores que influenciam negativamente a saúde, mas também as estratégias implementadas pelos serviços de saúde para mitigar os seus efeitos.

Para a concretização deste estudo, será utilizado um conjunto de dados (*dataset*) que compila informações provenientes de vários outros *datasets* complementares. Estes *datasets* incluem indicadores sobre taxas de mortalidade, prevalência de doenças, acesso a cuidados de saúde e fatores socioeconómicos, permitindo uma análise abrangente e detalhada.

A investigação será conduzida com o auxílio da linguagem de programação Python, uma ferramenta amplamente utilizada na área da ciência de dados devido à sua versatilidade e eficiência. Python permitirá não apenas a manipulação e análise dos dados, mas também a criação de visualizações gráficas que facilitem a interpretação dos resultados. Os *outputs* gerados servirão de base para sustentar as conclusões obtidas e poderão ser utilizados para propor soluções ou estratégias que contribuam para melhorar a saúde global.

Ao longo do estudo, serão seguidas boas práticas de análise de dados, assegurando a precisão, a transparência e a reprodutibilidade dos resultados. Assim, esta investigação visa não só compreender os fatores que influenciam a saúde, mas também oferecer contributos que possam apoiar a tomada de decisão no contexto das políticas de saúde pública.

## 1.4 Produzir o Plano do Projeto

Este projeto começa por procurar primeiro um conjunto de dados globais relacionados à saúde extraído por meio da plataforma *Kaggle*, com o objetivo de encontrar o conjunto de dados mais completo. Considerando os objetivos do projeto, escolhemos o *World Health Statistics* 2020, uma base de dados composta por vários conjuntos de dados, dos quais encontramos os mais relevantes sobre cancro, número de médicos por 10.000 habitantes, qualidade do ar, tabagismo e população 25 . A primeira fase levou cerca de duas semanas para ser preparada, onde também foi criada uma proposta para o projeto.

Na segunda fase, que teve a duração de três semanas, foram trabalhados os conjuntos de dados. Esta fase iniciou-se através da concatenação dos vários *datasets* para começar a limpeza. A limpeza é essencial para obter um conjunto de dados mais organizado e fácil de visualizar, por isso, primeiro foram encurtados os nomes dos indicadores para conferir uma leitura mais rápida. A coluna *First Tooltip* foi dividida em três partes para separar o valor dos intervalos de confiança estimados. Em seguida, foram identificados os valores nulos, e posteriormente foram feitas inspeções detalhadas e visualizações para descrever esses valores e começar a análise dos dados.

# Análise de fatores que influenciaram a saúde até 2020

A terceira fase concentrou-se na aplicação das técnicas de modelação, revisão do modelo e revisão do processo e por fim, avaliação do trabalho. Esta fase tomou cerca de 1 semana, onde foram feitas as árvores de decisão e *clustering* através de DBSCAN e devida análise.

# 2 Estudo dos dados

## 2.1 Recolha dos dados iniciais

Os dados foram retirados da plataforma *Kaggle*, onde o publicador utilizou como fonte os dados da OMS. Foram escolhidos 6 indicadores com um certo nível de relação entre eles e foram descarregados os seus ficheiros em formato CSV (*Coma Separated Values*), com um total de tamanho de 1,2 MB, no idioma inglês.

## 2.2 Descrição dos dados

O *dataset* contém 27123 linhas e 9 colunas após a extração e concatenação dos 6 *datasets*, referente a informações de 195 países durante os anos de 2000 a 2020. As linhas são distribuídas de acordo com o ano e com o nome dos países, sendo que nem todos os países participaram na pesquisa em todos os anos, explicando a existência de valores nulos. Nas colunas encontram-se os indicadores, ou variáveis da pesquisa. As variáveis "*Location*", "*Indicator*", "*Dim1*" e "*Dim2*" são variáveis qualitativas, sendo que as restantes são variáveis quantitativas. Segue-se em detalhe a explicação de cada atributo, ou variável.

1. "*Location*": Nome de cada país;
2. "*Period*": Ano dos dados (De 2000 a 2020);
3. "*Indicator*": Apresenta os indicadores que afetam a saúde mundial;
4. "*Dim1*": Refere-se ao género dos indivíduos da pesquisa;
5. "*First Tooltip*": Percentagem relacionada com o indicador analisado;
6. "*Dim2*": Refere-se aos órgãos que são lesados devido à poluição do ar;
7. "*Point Estimate*": Ponto médio relacionado com o indicador analisado;
8. "*Lower Bound*": Valor mínimo da percentagem média;
9. "*Upper Bound*": Valor máximo da percentagem média.

A figura 3 representa uma amostra aleatória dos dados recolhidos antes da limpeza dos mesmos.

Amostra aleatória dos dados do dataset

| | Location | Period | Indicator | Dim1 | First Tooltip | Dim2 |
|---|---|---|---|---|---|---|
| **16637** | Saudi Arabia | 2004 | Population using at least basic sanitation services (%) | Total | 98.620000 | nan |
| **1894** | Peru | 2015 | Probability (%) of dying between age 30 and exact age 70 from any of cardiovascular disease, cancer, diabetes, or chronic respiratory disease | Male | 14.300000 | nan |
| **13829** | Kyrgyzstan | 2016 | Population using at least basic sanitation services (%) | Total | 96.520000 | nan |
| **4877** | France | 2016 | Ambient and household air pollution attributable death rate (per 100 000 population) | Female | 3.35 [1.15-6.38] | Lower respiratory infections |
| **13065** | Honduras | 2007 | Population using at least basic sanitation services (%) | Urban | 79.750000 | nan |
| **24689** | Iceland | 2005 | Age-standardized prevalence of current tobacco smoking among persons aged 15 years and older | Female | 22.800000 | nan |
| **3255** | Belarus | 2016 | Ambient and household air pollution attributable death rate (per 100 000 population) | Both sexes | 1.67 [0.86-2.63] | Lower respiratory infections |
| **19239** | Croatia | 2008 | Medical doctors (per 10,000) | nan | 27.110000 | nan |
| **26606** | Switzerland | 2005 | Age-standardized prevalence of current tobacco smoking among persons aged 15 years and older | Female | 23.400000 | nan |
| **9904** | Bangladesh | 2012 | Population using at least basic sanitation services (%) | Total | 41.860000 | nan |

Figura 3: Dataset antes da realização da limpeza

A figura 4 representa uma amostra aleatória de dados do *dataset* após a limpeza do mesmo.

Amostra aleatória dos dados do dataset

| | Location | Period | Indicator | Dim1 | First Tooltip | Point Estimate | Lower Bound | Upper Bound |
|---|---|---|---|---|---|---|---|---|
| **25206** | Lithuania | 2018 | Tobacco usage | Both sexes | 27.100000 | 27.100000 | nan | nan |
| **22579** | Republic of Moldova | 2001 | Health expenditure | Rural | 3.150000 | 3.150000 | nan | nan |
| **5680** | Italy | 2016 | Household air pollution death | Male | 21.83 [14.9-29.04] | 21.830000 | 14.900000 | 29.040000 |
| **25791** | Oman | 2010 | Tobacco usage | Both sexes | 9.600000 | 9.600000 | nan | nan |
| **13562** | Jamaica | 2015 | Basic sanitation | Total | 86.950000 | 86.950000 | nan | nan |
| **17084** | South Africa | 2014 | Basic sanitation | Urban | 75.390000 | 75.390000 | nan | nan |
| **18000** | Tuvalu | 2008 | Basic sanitation | Rural | 79.480000 | 79.480000 | nan | nan |
| **25484** | Morocco | 2016 | Tobacco usage | Female | 0.900000 | 0.900000 | nan | nan |
| **22843** | Tajikistan | 1999 | Health expenditure | Rural | 3.570000 | 3.570000 | nan | nan |
| **10613** | Burundi | 2001 | Basic sanitation | Urban | 41.250000 | 41.250000 | nan | nan |

Figura 4: Dataset após a limpeza de dados

## 2.3 Exploração dos dados

Numa instância inicial, foi feita uma visualização gráfica que representasse a dimensão dos dados existentes, bem como as zonas geográficas que apresentam maior existência de dados. A figura 5 representa um mapa de calor que contém a soma de todos os dados.

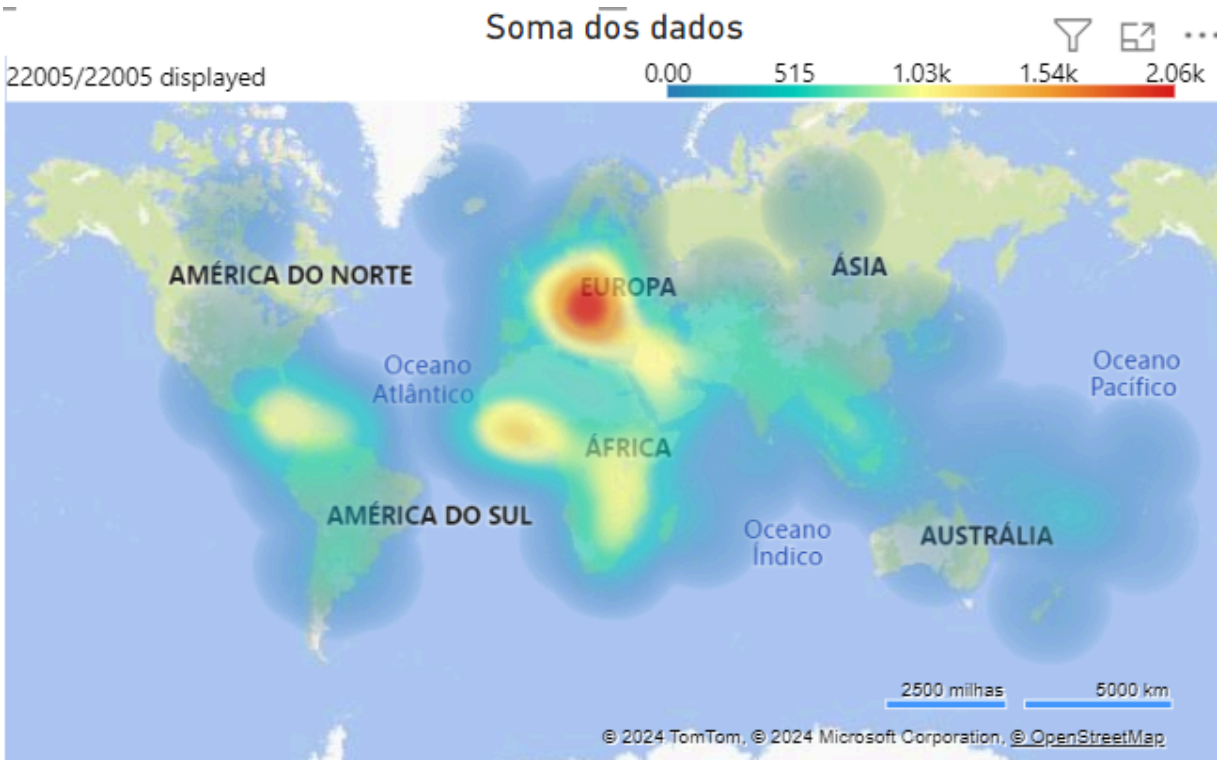



Figura 5: Mapa que identifica a origem dos dados

# Análise de fatores que influenciaram a saúde até 2020

Após a análise do mapa, conclui-se que no período considerado, existem dados para todos os continentes. O que se destaca é a Europa, que contém uma mancha mais quente, indicando que esse é o continente com mais dados estatísticos.

Além disso, uma vez que foram considerados vários anos para o estudo, foi feita uma tabela que representasse quais os anos com mais dados, bem como o peso de cada um. A figura 6 representa a contagem desses dados.

| Period | Contagem de Percentage |
|---|---|
| 2016 | 8278 |
| 2010 | 1801 |
| 2005 | 1716 |
| 2015 | 1688 |
| 2014 | 1174 |
| 2013 | 1161 |
| 2017 | 1073 |
| 2009 | 761 |
| 2008 | 760 |
| 2004 | 756 |
| 2007 | 751 |
| 2011 | 740 |
| 2012 | 740 |
| 2006 | 712 |
| 2002 | 705 |
| 2003 | 697 |
| 2001 | 666 |
| 2018 | 514 |
| **Total** | **24693** |

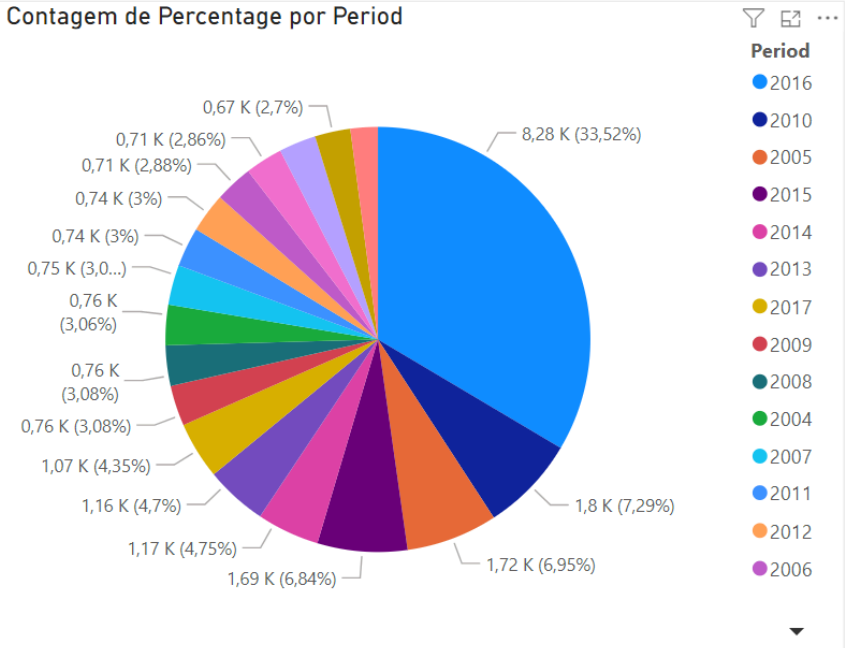



Figura 6: Contagem de dados por ano

Considerando a figura 6, conclui-se que 2016 foi o ano em que houve maior registo de dados estatísticos para os vários indicadores, com um total de 8.278 registos, em um total de 24.693, indicando que 33,52% dos valores correspondem a esse ano.

Posteriormente foi feita uma análise para perceber quais foram os 5 países com melhor nível de saúde e os 5 piores. Para isso foi feita uma normalização dos dados em *Python* por dois motivos: 1. Alguns indicadores representam um indício positivo quanto mais alto for o seu valor, enquanto que outros têm impacto negativo (Exº: O nº de médicos é positivo e a utilização do tabaco é negativo); e 2. Algumas variáveis estão expressas em percentagem e outras não. Desta forma é possível fazer uma comparação justa entre variáveis. Esta foi feita através do método *Min-Max Scaling*, onde os valores variam entre 0 e 1. Após essa normalização foram criadas 2 tabelas.

A figura 7 representa o TOP 5 de países com melhor índice de saúde, já a figura 8 representa o oposto.

| | Location | Normalized PE |
|---|---|---|
| **112** | Monaco | 0.286089 |
| **148** | San Marino | 0.274308 |
| **126** | Niue | 0.244955 |
| **130** | Palau | 0.215591 |
| **107** | Marshall Islands | 0.214362 |

Figura 7: Top 5 países com melhor índice de saúde.

| | Location | Normalized PE |
|---|---|---|
| **59** | Ethiopia | 0.044282 |
| **24** | Brunei Darussalam | 0.055935 |
| **38** | Congo | 0.056006 |
| **67** | Ghana | 0.058542 |
| **89** | Kenya | 0.061430 |

Figura 8: top 5 países com o pior índice de saúde

Após a análise das figuras 7 e 8 pode concluir-se que os países que obtiveram os melhores índices de saúde no período em estudo foram Mónaco, San Marino, Niue, Palau e Ilhas Marshall, localizados na Europa e Oceania com resultados acima de 0.2, enquanto que os que obtiveram os piores resultados foram Etiópia, Brunei Darussalam, Congo, Gana e Kenya, situados em África e Ásia e com resultados abaixo dos 0.07.

Para medir a relação entre os indicadores foi feita uma matriz de correlação, que é uma tabela que mostra os coeficientes de correlação entre variáveis. Os valores variam entre -1 e 1, onde os **valores menores que 0** indicam **correlação negativa** (quando os valores de uma variável aumentam, os da outra diminuem); os **valores maiores que 0** indicam **correlação positiva** (quando os valores de uma variável aumentam, os da outra também sobem); e valores **iguais a 0** representa **ausência de correlação** entre variáveis.

A Figura 9 representa a correlação entre os principais atributos. Ilustra como os valores selecionados variam em conjunto, indicando tendências de aumento ou diminuição simultâneos. No entanto, é importante destacar que essa análise não comprova a existência de uma relação causal entre os atributos.

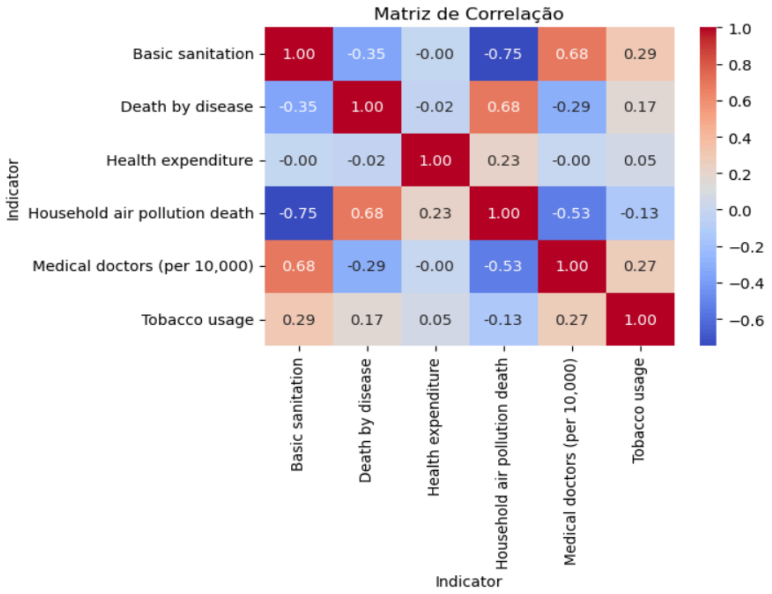



**Figura 9:** A correlação entre os principais atributos

Após a análise da matriz de correlação é possível perceber que existe uma distribuição distinta de correlação entre variáveis, existindo 7 correlações positivas, 6 negativas e 2 ausências de correlação. O valor mais elevado é de 0.68 para 4 variáveis: *Household air pollution death* com *Death by disease*; e *Medical doctors* com *Basic sanitization*. O valor mais baixo é de -0.75 para as variáveis *Household air pollution death* e *Basic sanitization*.

Esta análise valida que esta matriz não comprova a existência de uma relação causal entre atributos, como se pode verificar por exemplo em *'Medical doctors'* e *'Basic sanitation'*, pois de forma lógica sabe-se que a existência de saneamento básico não tem relação direta com a existência de mais médicos, mas o facto de os valores se acompanharem dão-se devido a fatores externos como o facto de os países com melhores condições geralmente terem maiores índices de crescimento nestas duas variáveis.

## 2.4 Verificação da qualidade dos dados

Numa primeira abordagem para verificar a qualidade dos dados, foi feita a verificação de valores ausentes detectados pela função df.isnull().sum() no *Python*, apresentado na Figura 10.

```
df.isnull().sum()
```

```
Location              0
Period                0
Indicator             0
Dim1               2506
First Tooltip         0
Dim2              20535
Point Estimate        0
Lower Bound       20535
Upper Bound       20535
dtype: int64
```

**Figura 10:** Verificação da qualidade dos dados

# Análise de fatores que influenciaram a saúde até 2020

De forma a não distorcer a análise, não foi feito o preenchimento dos dados ausentes uma vez que o estudo envolve vários anos e países, pelo que não faz sentido substituir esses valores.

```
df['Point Estimate'].mean()
```

```
38.39889894185746
```

```
df['Lower Bound'].mean()
```

```
21.702120218579235
```

```
df['Upper Bound'].mean()
```

```
33.36624620522161
```

**Figura 11:** Cálculo das médias

Outra abordagem é o cálculo das médias dos atributos relevantes.

Para finalizar esta análise foram feitos histogramas para cada atributo, onde o eixo do X representa a frequência desses valores. Assim, a figura 12 apresenta 6 histogramas para os indicadores em estudo.

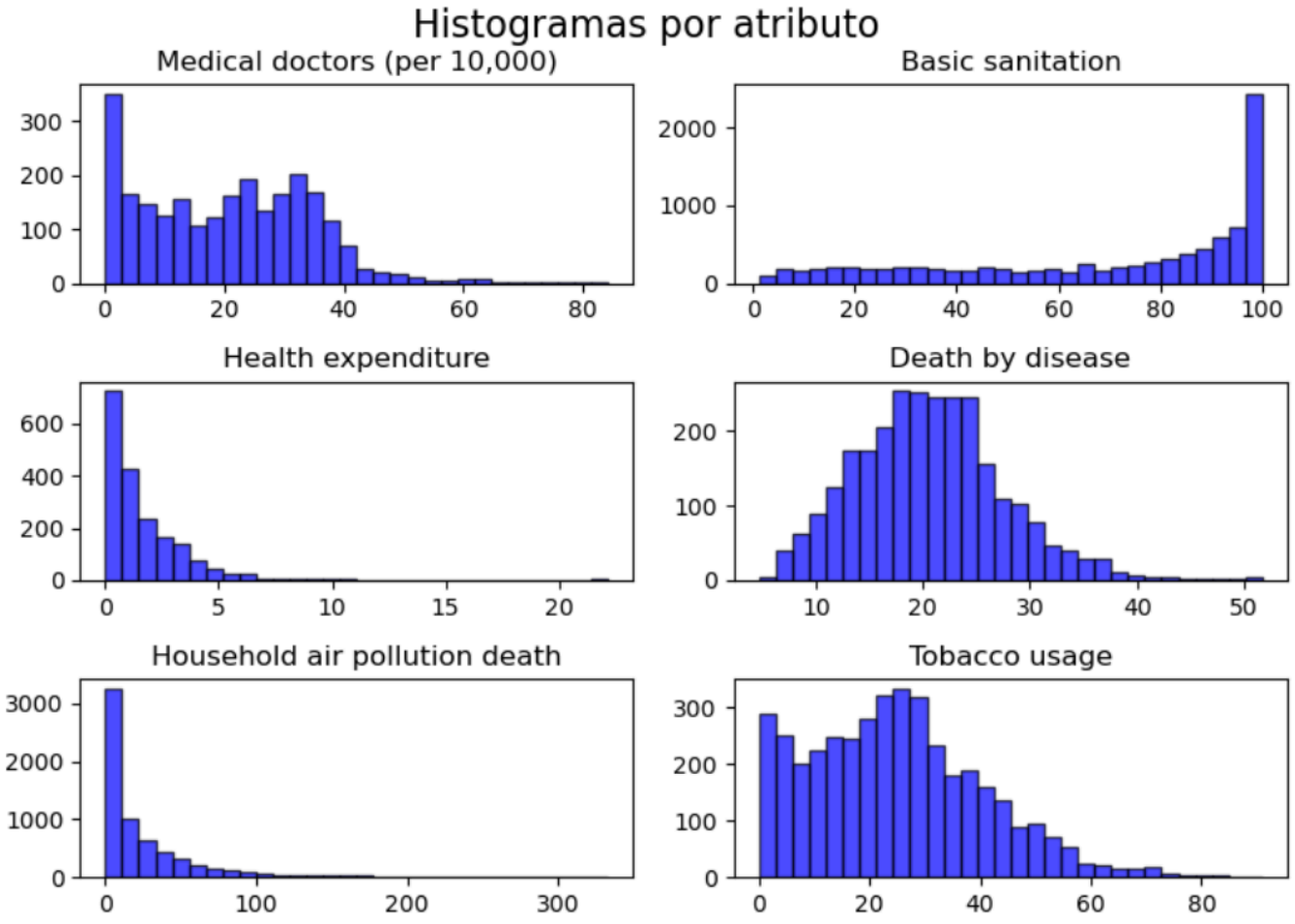



Figura 12: Histogramas relativos aos indicadores de saúde do dataset

# Análise de fatores que influenciaram a saúde até 2020

Cada histograma mostra a distribuição dos valores para os indicadores de saúde no dataset. Deste modo, foi constatada a seguinte análise: No primeiro histograma (*Medical Doctors*), os países com poucos médicos estão mais concentrados à esquerda, ou seja, poucos países têm valores altos; no segundo histograma (*Basic Sanitation*), a maioria dos países têm bom acesso ao saneamento básico, pois, concentra-se mais do lado direito; no terceiro histograma (*Health expenditure*), a maioria dos países gasta pouco dinheiro em despesas de saúde; no quarto histograma (*Death by disease*), existe uma concentração central, o que indica um número médio de mortalidade para a maioria dos países; no quinto histograma (*Household air pollution death*), a maioria dos países tem valores baixos, apesar de alguns países apresentarem valores mais elevados; no sexto histograma (*Tobacco usage*), os países apresentam uma variação moderada, com alguns valores altos e outros baixos.

Assim, conclui-se que os histogramas Medical doctors (per 10,000), Health expenditure, Household air pollution e Tobacco usage seguem distribuição assimétrica à esquerda; O histograma Basic sanitation segue distribuição assimétrica à direita; e o histograma death by disease segue distribuição normal.

# 3 Preparação dos dados

## 3.1 Seleção dos dados

Os dados e as variáveis selecionadas para a análise dos fatores que influenciaram a saúde mundial entre 2000 e 2020 foram os seguintes:

| Variáveis em Português | Variáveis em Inglês | Base de Dados Kaggle |
|---|---|---|
| Usuários de tabaco <15 anos | Prevalence of current tobacco use among persons aged 15 years and older | tobaccoAge15.csv |
| Taxa de mortalidade derivada da poluição do ar | Ambient and household air pollution attributable death rate per 100,00 population and the same data with age-standardized | airPollutionDeathRate.csv |
| Nº de médicos por 10.000 Habitantes | Medical doctors per 10,000 population | medicalDoctors.csv |
| População que gasta menos de 25% do seu salário em despesas de saúde | Population with household expenditures on health greater than 25% of total household expenditure or income | population25%SDG3.8.2.csv |
| Percentagem de população que dispõem de serviços básicos de sanitização | Population using at least basic sanitation services | atLeastBasicSanitizationServices.csv |
| Probabilidade de falecer, entre os 30 e os 70 anos, por doenças cardiovasculares, cancerígenas, etc. | Probability of dying between the age of 30 and exact age of 70 from any of the cardiovascular disease, cancer, diabetes, or chronic respiratory disease. | 30-70cancerChdEtc.csv |

Figura 13: As variáveis que influenciam a saúde mundial

## 3.2 Limpeza dos dados

Após a seleção dos dados, compreendeu-se que certas colunas não forneciam certos dados significativos na comparação entre países. Deste modo, foram removidas as colunas “*First Tooltip*” e “*DIM2*” do *dataset*.

# 4 Modelação

## 4.1 Seleção da técnica de modelação

A seleção da técnica de modelação é importante para que se consiga realizar uma análise mais adequada das variáveis. Uma vez que anteriormente foi decidido manter os valores nulos, uma boa opção para fazer a modelação dos dados é o uso da Árvore de Decisão (AD). Além disso o uso de uma AD é apropriado quando existe um grande número de registos, pois esta reduz os mesmos em conjuntos sucessivamente menores aplicando uma sequência de regras de decisão simples. Por outro lado, é importante reconhecer padrões nos dados devido ao facto de também ser importante compreender se há tendências entre zonas geográficas. Deste modo foi selecionado o modelo de *Clustering*, e mais especificamente o *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*).

Uma vez que a variável de saída é a probabilidade de falecer de doenças cancerígenas entre os 30 e os 70 anos, foi utilizada como variável-alvo (eixo y).

Para a aplicação dos dois modelos foi utilizado o *software Python*.

## 4.2 Geração do desenho de testes

Para criar a árvore de decisão, utilizamos a linguagem *Python* e importamos algumas bibliotecas necessárias como scikit-learn e o matplotlib. As figuras 14 e 15 representam o processo de criação da árvore.

```python
# Criar uma nova coluna para identificar se o indicador é "Death by disease"
df['Target'] = (df['Indicator'] == 'Death by disease').astype(int)

# Pivotar os dados para criar colunas para cada indicador usando 'Point Estimate'
pivoted_data = df.pivot_table(
    values='Point Estimate',
    index=['Location', 'Period'],
    columns='Indicator',
    aggfunc='mean'
).reset_index()

# Adicionar a variável-alvo (Target) ao conjunto de dados pivotado
final_data = pd.merge(pivoted_data, df[['Location', 'Period', 'Target']].drop_duplicates(), on=['Location', 'Period'])

# Exibir as primeiras linhas para inspeção
final_data.head()
```

Figura 14: Código utilizado para formatar os dados para modelação

```python
# Preencher valores ausentes com 0 (assumindo ausência de dados como 0 impacto)
final_data.fillna(0, inplace=True)

# Separar as variáveis preditoras (X) e a variável alvo (y)
X = final_data.drop(columns=['Location', 'Period', 'Target'])
y = final_data['Target']

# Dividir os dados em conjuntos de treino e teste
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Treinar o modelo de árvore de decisão
from sklearn.tree import DecisionTreeClassifier
clf = DecisionTreeClassifier(random_state=42)
clf.fit(X_train, y_train)

# Fazer previsões
y_pred = clf.predict(X_test)

# Avaliar o modelo
from sklearn.metrics import classification_report, accuracy_score

accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred)

accuracy, report
```

Figura 15: Código utilizado para criar a Árvore de Decisão

De seguida, para enriquecermos a nossa análise, aplicamos o método clustering *DBSCAN* (conforme indica a figura 16) que detecta clusters de forma não linear (útil para datasets com valores nulos). Realizámos os seguintes passos para aplicar o *DBSCAN*:

1. Preenchemos os valores nulos com o valor 0;
2. Normalizamos os dados para evitar que diferentes escalas prejudiquem o algoritmo;
3. Configuramos os parâmetros do *DBSCAN*: eps é a distância máxima entre dois pontos para serem considerados vizinhos; *min_samples* é o número mínimo de pontos para formar um *cluster*;
4. Executamos o *DBSCAN* e analisamos os clusters que foram detectados.

```python
from sklearn.cluster import DBSCAN
from sklearn.preprocessing import StandardScaler
import numpy as np

# Preencher valores nulos com 0 (pode ser ajustado para outro método, como a média)
data_for_clustering = final_data.drop(columns=['Location', 'Period', 'Target']).fillna(0)

# Normalizar os dados para clustering
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_for_clustering)

# Aplicar o algoritmo DBSCAN
dbscan = DBSCAN(eps=1.5, min_samples=5)  # Ajustar parâmetros conforme necessário
clusters = dbscan.fit_predict(data_scaled)

# Adicionar os clusters ao dataset original
final_data['Cluster'] = clusters

# Contar o número de pontos por cluster
cluster_counts = final_data['Cluster'].value_counts()

# Mostrar os primeiros resultados e a distribuição dos clusters
final_data[['Location', 'Cluster']].head(), cluster_counts
```

Figura 16: Código utilizado para fazer o clustering DBSCAN

## 4.3 Construção do modelo

Como podemos observar, a figura 17 representa a Árvore de Decisão que elaboramos numa primeira instância.

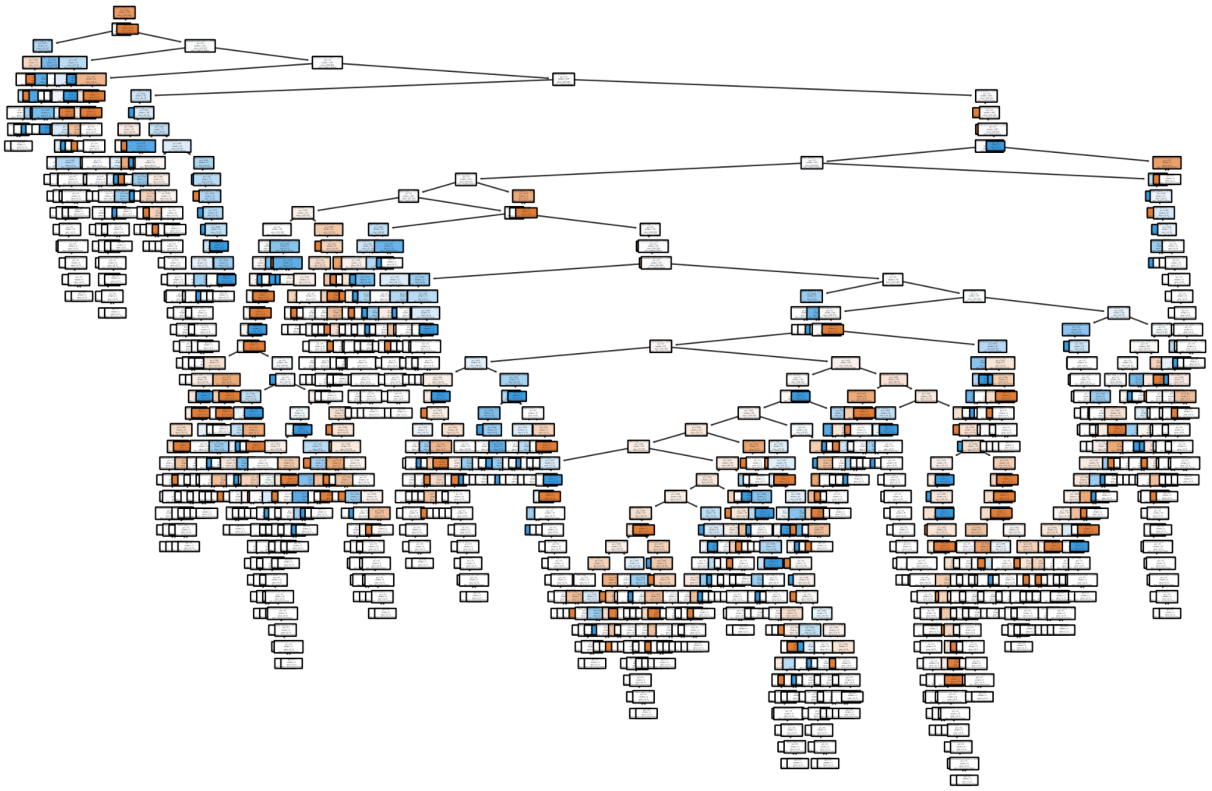


Figura 17: Árvore de Decisão completa

# Análise de fatores que influenciaram a saúde até 2020

Como podemos analisar a nossa Árvore de Decisão é complexa e profunda devido ao grande número de divisões. Isto pode indicar que existe um sobreajuste (*overfitting*). Como tal, limitamos a profundidade máxima da árvore (*max_depth*) para torná-la mais interpretável e reduzir o risco de sobreajuste.

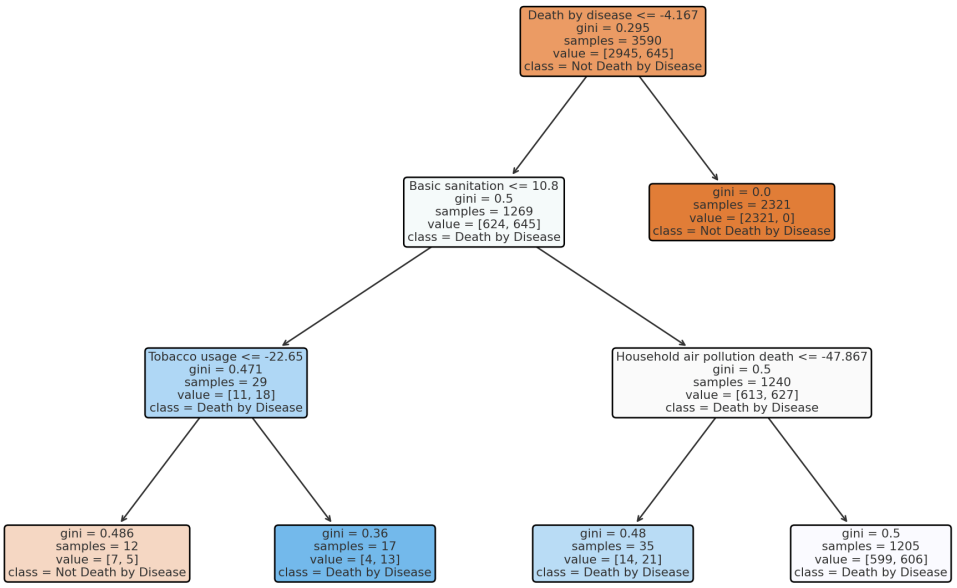



Figura 18: Árvore de decisão simplificada

Deste modo, após limitarmos a profundidade a três níveis realizamos as seguintes observações:

1. Divisões principais: a primeira divisão é com base na variável *Death By Disease* (valor numérico do indicador); subsequentemente, as outras variáveis como *Basic Sanitation* e *Household Air Pollution Death* são utilizadas;

2. Número de amostras por nó: cada nó mostra quantas amostras possui e a proporção entre as classes (*Not Death By Disease* e *Death By Disease*);

3. Cálculo de *Gini*: O índice de *Gini* mede a pureza de cada nó (quanto menor for o número, melhor).

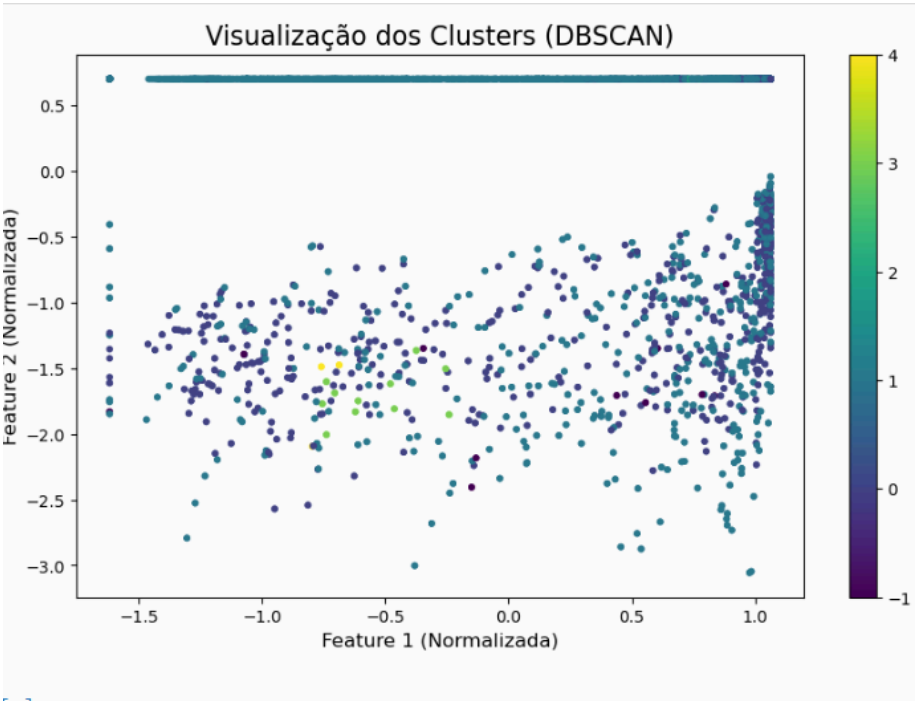



Figura 19: Visualização dos Clusters

Após a execução do *DBSCAN* obtivemos os seguintes resultados:

- *Cluster 0*: Contém 5094 amostras (a maior parte dos dados foi agrupada aqui);
- *Cluster -1*: Representa 29 amostras classificadas como ruído (não pertencem a nenhum cluster);
- *Cluster 1*: Contém apenas 6 amostras.

Após obtermos a nossa imagem gráfica do *DBSCAN* realizamos as seguintes observações:

1. A maioria dos dados foram agrupados no *Cluster* 0, o que indica que a densidade dos dados é alta e homogénea;
2. O *Cluster* -1 contém outliers ou dados dispersos que não atendem aos critérios de densidade dos algoritmos;
3. Os resultados indicam que os parâmetros (*eps* e *min_samples*) podem precisar de ajustes para detetar mais *clusters*.

## 4.4 Revisão do modelo

A revisão do modelo foi realizada com base nos critérios de avaliação de sucesso previamente definidos para a mineração de dados. O objetivo foi assegurar que o modelo atende aos parâmetros de qualidade e desempenho esperados, considerando as variáveis analisadas.

# Análise de fatores que influenciaram a saúde até 2020

Resultados e qualidades do modelo:

Os resultados da avaliação indicam que o desempenho médio do modelo, representado pelo indicador *Normalized PE*, foi de 0,8008 numa escala de 0 a 1. Este desempenho demonstra um alinhamento moderado do modelo com os objetivos estabelecidos. Entretanto, a correlação fraca e negativa (-0,2262) entre o indicador *Normalized PE* e a variável-alvo sugere a necessidade de refinamento para melhorar a previsibilidade e a aderência ao objetivo final.

Principais Qualidades Identificadas:

- Alguns grupos apresentaram desempenhos ideais (*Normalized PE* = 1), indicando cenários específicos onde o modelo é altamente eficaz;
- Os segmentos hierarquizados evidenciam que o contexto geográfico e demográfico afeta significativamente os resultados do modelo.

Interpretação e Ajustes Necessários:

- Embora alguns grupos tenham atingido o desempenho ideal, outros apresentaram valores mais baixos, revelando inconsistências que podem ser exploradas para ajustes;
- Recomenda-se verificar a lógica e plausibilidade dos resultados em segmentos com baixa correlação com a variável-alvo, como nos cenários onde *Target* = 1.

Ajustes Propostos:

- Refinar os parâmetros do modelo, considerando diferenças demográficas e geográficas, para alinhar os resultados ao objetivo final;
- Implementar estratégias de balanceamento das classes, uma vez que a variável-alvo (*Target*) possui um número desproporcional de observações em 0;
- Testar novos parâmetros ou técnicas de modelagem que possam aumentar a correlação entre o *Normalized PE* e o *Target*.

Sendo assim, a análise revelou bons resultados em segmentos específicos, mas há oportunidades claras de melhoria no modelo como um todo, nomeadamente fazer ajustes nos parâmetros e ampliar a análise da plausibilidade dos resultados para garantir uma implementação robusta.

# 5 Avaliação

## 5.1 Avaliação de resultados

A avaliação dos resultados gerados pelo modelo revelou insights valiosos para os objetivos estabelecidos no início do nosso trabalho.

O modelo identificou relações significativas entre os indicadores-chave de saúde, como a mortalidade, fatores socioeconômicos e demográficos. A importância destes fatores varia conforme o contexto geográfico e temporal e destaca padrões relevantes em determinadas regiões ou períodos. Os resultados indicam que os fatores analisados possuem graus variados de correlação que indica interdependências que podem ser exploradas para compreender melhor as influências mútuas. Esta análise relevou comparações robustas entre dimensões como o sexo, a idade e a localização geográfica. O modelo forneceu uma perspectiva ampla sobre as condições globais de saúde, o que permitiu identificarmos áreas prioritárias para intervenção ou melhoria, especialmente em regiões de baixa performance em termos de saúde pública.

Foi criado um ranking para priorizar os fatores mais relevantes para a saúde global. Os fatores socioeconómicos emergiram como os mais significativos, seguidos por condições ambientais e o acesso aos serviços de saúde.

A análise inicial revelou que o modelo apresentou uma eficácia geral de 70,1%, o que significa que, em média, cerca de 70% das amostras do teste foram classificadas de forma correta. Mas ao analisar o relatório de classificação, verificou-se um desempenho desequilibrado entre as classes, para a classe 0 (amostras que não correspondem a mortes por doença), o modelo alcançou uma precisão de 81%, um recall de 83% e um valor F1-score de 82%, refletindo um bom equilíbrio entre a capacidade de identificar corretamente estas amostras. Por outro lado, para a classe 1 (amostras que correspondem a mortes por doença), o desempenho foi significativamente inferior, com uma precisão de apenas 10%, um recall de 9% e um F1-score igualmente reduzido (10%).

Assim, observou-se um desequilíbrio no conjunto de dados, evidenciado pela distribuição das amostras nas classes: 1269 amostras pertenciam à classe 0, enquanto apenas 270 pertenciam à classe 1. Este desequilíbrio contribuiu para a distorção do modelo a favor da classe maioritária, resultando numa menor capacidade de identificação das amostras da classe minoritária.

A aprovação dos modelos reflete a adequação aos objetivos do projeto, mas a aplicação em cenários reais e o acompanhamento contínuo serão cruciais para validar e expandir os resultados.

## 5.2 Revisão do processo

Tendo em conta o trabalho realizado, foi concluído que de facto existem algumas falhas no processo. Na fase de entendimento do tema, teria sido ideal se houvesse um foco mais específico, uma vez que a saúde é um tema bastante geral que pode ter várias abordagens e torna-se complicado estabelecer uma visão simplificada do que significa. Houve ainda uma certa dificuldade em fazer uma análise imparcial dos dados, sobretudo pelo facto do período considerado ser extenso, tendo em conta que muitos países começaram a fornecer dados a partir do século XXI e que ainda existe muita falta de dados estatísticos para vários países, sobretudo nos de terceiro mundo. Neste sentido teria sido útil recolher mais dados para tentar combater essa tendência.

## 5.3 Determinar os próximos passos

Com base nos resultados obtidos pela árvore de decisão e pelo *clustering DBSCAN*, elaboramos as seguintes recomendações que podem ser feitas para alinhar os *insights* do modelo com os objetivos da saúde mundial:

A árvore de decisão revelou que fatores como saneamento básico e controlo da poluição doméstica estão diretamente associados à mortalidade por doenças. É crucial que governos e organizações priorizem investimentos nessas áreas para reduzir a mortalidade.

O clustering *DBSCAN* identificou outliers que representam grupos com comportamento distinto em relação à mortalidade. Esses grupos devem ser analisados detalhadamente para entender as causas específicas e desenvolver intervenções personalizadas.

O modelo de árvore de decisão foi eficiente em identificar divisões claras, mas pode ser complementado com modelos mais robustos, como *Random Forests*, para capturar relações mais complexas e não-lineares. Além disso, ajustar os parâmetros do *DBSCAN* pode revelar mais granularidade nos dados.

Implementar sistemas de monitorização para acompanhar o impacto de políticas públicas em tempo real e implementar estratégias conforme necessário.

Os insights obtidos abrem possibilidades para projetos futuros, como a análise de impacto de fatores económicos, educação e infraestrutura na mortalidade por doenças. Esses projetos podem complementar os objetivos iniciais e expandir as oportunidades de ação.

Estas recomendações não apenas reforçam os objetivos do nosso modelo, mas também criam uma direção clara para futuras iniciativas na melhoria da saúde pública e gestão dos recursos associados à saúde.

# 6 Referências

Belfo, F. (2020). *Apresentação Resumida e Adaptada do Modelo CRISP-DM.* Unidade Curricular de Data Mining & Machine Learning, ISCAC.

Belfo, F. (2024-2025). *DM&ML 05 2024-2025 Árvores de Decisão (parte 1+2+3)*. Unidade Curricular de Data Mining & Machine Learning, ISCAC.

Belfo, F. (2022). *Modelo CRISP-DM*. Unidade Curricular de Data Mining & Machine Learning, ISCAC.

Leite, J. (n.d.). *Preparação dos Dados*. Unidade Curricular de Complementos de Estatística para a Ciência dos Dados, ISCAC.

World Health Organization (WHO). (n.d.). *World Health Organization*. Retrieved January 8, 2025, from https://www.who.int/

Zeus. (n.d.). *WHO World Health Statistics 2020 - Complete*. Retrieved January 8, 2025, from https://www.kaggle.com/datasets/utkarshxy/who-worldhealth-statistics-2020-complete?select=adolescentBirthRate.csv