DETERMINANTES DAS CAUSAS DE ÓBITOS NO BRASIL (2015-2020): UMA ANÁLISE COMPARATIVA ENTRE ÓBITOS E NASCIMENTOS UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Moisés Moreira[1]

[1]Instituto de Ciências Exatas e Tecnológicas – Universidade Federal de Viçosa, Campus Rio Paranaíba (UFVCRP)
Caixa Postal 22 – Rio Paranaíba– MG – Brasil

2

{moises.ribeiro}@ufv.br

Abstract. *This study investigates the comparison between death rates and birth rates in Brazil from 2015 to 2020, using machine learning algorithms such as Random Forest and Naive Bayes. The results highlight the effectiveness of Random Forest in capturing complex patterns, achieving 100% accuracy on the full dataset, but suffering from overfitting in cross-validation. In contrast, Naive Bayes shows more stable performance with 90% accuracy, demonstrating robustness for generalization. The analysis suggests that while Random Forest is useful for identifying patterns related to mortality and socio-economic factors, Naive Bayes might offer better consistency for long-term mortality prediction.*

Resumo. *Este estudo investiga a comparação entre as taxas de óbitos e nascimentos no Brasil de 2015 a 2020, utilizando algoritmos de aprendizado de máquina como Random Forest e Naive Bayes. Os resultados destacam a eficácia do Random Forest em capturar padrões complexos, com 100% de acurácia no conjunto completo de dados, mas com tendência a overfitting na validação cruzada. Por outro lado, o Naive Bayes apresenta desempenho mais estável, com 90% de acurácia, demonstrando robustez para generalização. A análise sugere que, embora o Random Forest seja útil para identificar padrões de mortalidade, o Naive Bayes oferece maior consistência na previsão de mortalidade a longo prazo. Todo o projeto pode ser acessado pelo GitHub.*

1. Introdução

Mudanças na renda, educação, acesso a serviços públicos e na organização socioeconômica impactam significativamente a qualidade de vida, influenciando padrões de mortalidade por estratos sociais [Wood and Carvalho 1994]. A esperança de vida ao nascer aumentou globalmente no século XX, com avanços expressivos em países em desenvolvimento nas últimas décadas [Vaupel 2010]. No Brasil, entre 1980 e 2009, a mortalidade infantil caiu de 69,12 para 22,57 óbitos por mil nascidos vivos, enquanto a esperança de vida subiu de 62,57 para 73,17 anos [IBGE – Instituto Brasileiro de Geografia e Estatística 2010]. Essa redução se deve, em grande parte, à melhoria do saneamento básico e ao combate a doenças infeciosas [Prata 1992].

Após 1980, a mortalidade também diminuiu entre os idosos, contribuindo para maior longevidade [Campos and Rodrigues 2004]. Essas mudanças refletem a transição epidemiológica, marcada pela redução de doenças transmissíveis graças a avanços em saúde pública, como vacinas, antibióticos e saneamento [Omran 1971]. Além disso, a compreensão da mortalidade desloca os óbitos para idades mais avançadas, evidenciando a importância de estudar a mortalidade adulta [Queiroz et al. 2017].

No Brasil, as desigualdades socioeconômicas desempenham papel central no declínio da mortalidade, com grupos mais favorecidos apresentando mais baixas de óbitos [Cutler et al. 2006]. Apesar de avanços, como a redução do coeficiente de Gini de 0,61 para 0,54 entre 1990 e 2009, desigualdades regionais persistem, especialmente no Norte e Nordeste [Andrade et al. 2013]. Estudos sobre mortalidade infantil são amplos, mas a mortalidade adulta ainda é pouco explorada, especialmente quanto aos fatores socioeconômicos e regionais [Queiroz et al. 2017].

Compreender os determinantes da mortalidade adulta nas microrregiões brasileiras é essencial para orientar políticas públicas. Fatores como infraestrutura, saúde e contexto socioeconômico atuam de forma integrada, exigindo abordagens analíticas mais avançadas. Além disso, a análise comparativa entre óbitos e os nascimentos é crucial para compreender a dinâmica demográfica e suas implicações socioeconômicas. Métodos de aprendizado de máquina podem oferecer insights valiosos sobre essas interações, sendo ainda pouco aplicados no Brasil.

O objetivo deste estudo é investigar realizar uma análise comparativa entre o número de óbitos e nascimentos no Brasil durante os anos de 2015 à 2020, avaliando o desempenho de algoritmos como Ranom Forest e Naive Bayes. Os resultados ajudaram a identificar padrões e propor políticas que reduzam a mortalidade e melhores a qualidade de vida.

2. Fatores Determinantes e Tendências de Mortalidade no Brasil

A mortalidade é influenciada por uma combinação de fatores socioeconômicos, biológicos e comportamentais. Esses fatores operam em correntes causais, afetando as tendências de mortalidade ao longo do tempo [Kunst et al. 1999]. Fatore individuais, como a educação do país, e domésticos, como a renda e o a cesso a serviços, impactam diretamente a saúde e a mortalidade infantil [Mosley and Chen 1984]. O Programa Saúde da Família tem sido um importante fator de redução da mortalidade no Brasil, promovendo ações preventivas e de proteção à saúde [Sousa and Hamann 2009].

Para mortalidade adulta, fatores como condições macroeconômicas e desigualdade de renda também desempenham papéis significativos. Regiões urbanas, embora com mais oportunidades de emprego e acesso à saúde, enfrentam desafios como violência e doenças crônicas [WHO – World Health Organization 2010]. No Brasil, embora tenha ocorrido uma queda geral na mortalidade, ainda persistem desigualdades regionais e socioeconômicas, com áreas menos urbanizaas apresentando mortalidade adulta relativamente baixa [Santana et al. 2015].

Observa-se na Figura 1 que microrregiões da região Norte do Brasil, que estão localizadas em estados considerados menos desenvolvidos, apresentam taxas de mortalidade adulta surpreendentemente baixas, semelhantes às observadas nas microrregiões do

Sul do país. Exemplos incluem as microrregiões de Japurá, Jaruá, Parintins, Purus e Madeira no Amazonas, Brasiléia no Acre e Óbitos no Pará. Essas regiões, embora em estados com menor desenvolvimento, apresentam características com baixa urbanização e infraestrutura deficiente, mas com taxas de mortalidade adulta mais baixas do que esperado [Santana et al. 2015].

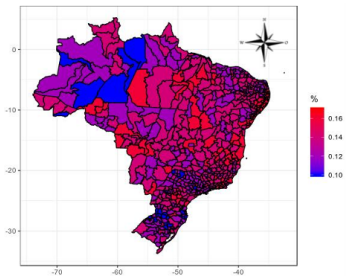

Figura 1. Probabilidade de morte adulta corrigida – Microrregiões, Brasil – 2010 [Schmertmann and Gonzaga 2018]

Além disso, os índices de nascimento também são determinantes importantes nas tendências de mortalidade. A queda nas taxas de natalidade tem sido observada no Brasil, refletindo mudanças socioeconômicas, como maior escolarização e acesso a métodos contraceptivos. Essa redução nas taxas de natalidade está associada a uma melhoria nas condições de vida, mas também implica em desafios para o sistema de saúde, como a necessidade de maior atenção à saúde da população envelhecida.

3. Material e Métodos

3.1. Conjunto de Dados

Os dados foram oriundos do Sistema de Informação de Mortalidade, do Cadastro Nacional de Estabelecimento de Saúde (CNES) e do Sistema de Informação da Atenção Básica (SIAB) organizados pelo Ministério da Saúde por meio do Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS) durante o período de 2015 até 2020.

Para obter os dados de óbitos segui o critério de pesquisa abaixo, salvando-os em um único arquivo (.csv), onde as linhas correspondem ao ano e as colunas ao CID-10:

- Estatísticas Vitais > Mortalidade - desde 1996 pela CID-10 > Mortalidade Geral > Brasil por Região e Unidade da Federação > Linha (capítulo-CID-10) / Coluna (Ano do Óbito) / Conteúdo (Óbitos p/ocorrência) / Período (2015 a 2019).

- Estatísticas Vitais > Dados Preliminares de 2020 > Mortalidade Geral > Brasil por Região e Unidade da Federação > Linha (Capítulo-CID-10) / Coluna (Ano do óbito) / Conteúdo (Óbitos p/ocorrência) / Período (2020).

Para complementar, serão comparados os totais de óbitos com os de nascimentos, também para o período de 2015 a 2020. Os dados de nascimento foram coletados no site do DATASUS seguindo o seguinte critério:

- Estatísticas Vitais > Nascidos vivos - desde 1994 > Nascidos Vivos > Brasil por Região e Unidade da Federação > Linha (Ano do nascimento) / Coluna (não

ativa) / Conteúdo (Nascim p/ocorrênc) / Período (2015 a 2019).

- Estatísticas Vitais > Dados Preliminares de 2020 > Nascidos Vivos > Brasil por Região e Unidade da Federação > Linha (Ano do nascimento) / Coluna (não ativa) / Conteúdo (Nascim p/ocorrênc) / Período (2020).

3.2. Variáveis Utilizadas

A Tabela 1 apresenta as variáveis que foram utilizadas no modelo, as fontes de onde foram retiradas e uma descrição referente a forma como as mesmas foram operacionalizadas.

A Coluna "Ano do Óbito"indica os anos que foram escolhidos para análise (2015 a 2020), e as Colunas (Cap I, Cap II...Cap XX) indicam as colunas que representam o Capítulo CID-10 e os Óbitos p/Ocorrência.

Ano do Óbito	Cap I	Cap II	Cap III	Cap IV	Cap V	Cap VI	Cap VII	Cap VIII	Cap IX	Cap X	Cap XI	Cap XII	Cap XIII	Cap XIV	Cap XV	Cap XVI	Cap XVII	Cap XVIII	Cap XX	Total
2015	55022	209780	6506	76235	12558	34721	21	147	349642	149541	64202	4970	5385	36549	1896	22162	10989	71713	152136	1264175
2016	57188	215217	6878	78075	12674	36870	20	173	362091	158041	66044	5874	5787	39367	1814	21049	10882	75869	155861	1309774
2017	54874	221821	6622	79662	12858	38786	19	179	358882	155620	66052	6100	5912	40470	1874	21458	10995	71822	158657	1312663
2018	54679	227920	6601	81365	13697	41035	21	169	357770	155191	67316	6273	6153	43428	1862	20764	11156	70505	150814	1316719
2019	56666	235301	7068	83485	14526	45235	23	206	364132	162005	68770	7152	6506	47566	1726	20354	11308	74972	142800	1349801
2020	264666	227519	6622	91055	17188	44881	25	158	354093	150374	66131	6815	6060	44994	1989	18815	9536	97436	144382	1552739

Tabela 1. Variáveis utilizadas no modelo

3.3. Métodos de Aprendizado de Máquina

O Aprendizado de Máquina, subcampo da Inteligência Artificial, tem ganhado destaque nas últimas décadas [Arpino et al. 2018]. Diferentemente da regressão paramétrica, o aprendizado de máquina não impõe modelos fixos, permitindo que os algoritmos identifiquem automaticamente relações e interações entre variáveis independentes. Assim, problemas como colinearidade e violações de pressupostos tornam-se menos relevantes dependendo do algoritmo utilizado [De Rose and Pallara 1997][Billari et al. 2006].

As técnicas de aprendizado de máquina dividem-se em duas categorias principais: aprendizado supervisionado e não supervisionado, , Figura 2. O aprendizado não supervisionado foca na identificação de padrões e redução de dados, sendo amplamente aplicado em problemas de agrupamento e classificação. Já o aprendizado supervisionado é voltado à modelagem preditiva, utilizando conjuntos de dados rotulados para estimar relações entre variáveis socioeconômicas e resultados de interesse, como a probabilidade de morte adulta [Kuhn and Johnson 2013].

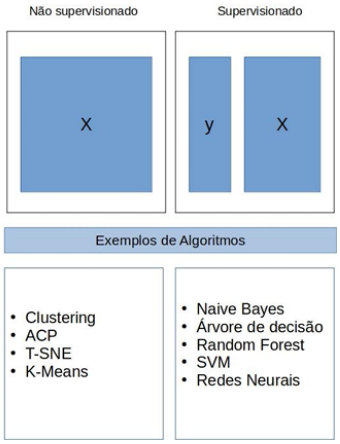

Figura 2. Exemplos de algoritmos de aprendizado supervisionado e não supervisionado [Jiang et al. 2017]

Os algoritmos supervisionados são iterativos e flexíveis, ajustando-se automaticamente a relações complexas e não lineares. Eles são especialmente úteis em grandes conjuntos de dados, permitindo não apenas prever resultados, mas também explorar como variáveis independentes se relacionam com variáveis dependentes. Entre os algoritmos supervisionados analisados no estudo estão Random Forest e o Naive Bayes.

4. Pré-processamento dos dados

Nesta seção serão apresentados alguns métodos de pré-processamento de dados que foram utilizados neste trabalho a fim de entender melhor o conjunto de dados trabalhado.

4.1. Estatísticas descritivas para entender a distribuição dos dados

A Figura 3 apresenta estatísticas descritivas dos dados referentes ao número de óbitos registrados entre os anos de 2015 e 2020. Para cada ano, são fornecidos os seguintes parâmetros estatísticos:

1. **Count:** O número de registros disponíveis para cada ano, que permanece constante em 20 para todos os períodos analisados.
2. **Mean:** A média dos óbitos registrados em cada ano. Observa-se que a média varia ao longo dos anos, iniciando em aproximadamente 126 mil óbitos em 2015, reduzindo gradualmente até cerca de 131 mil em 2017, e apresentando um aumento significativo nos anos de 2019 (cerca de 135 mil) e 2020 (aproximadamente 155 mil).
3. **Std (Desvio padrão):** Mede a dispersão dos dados em relação à média. Valores mais altos, como os observados em 2015 (282.077) e 2020 (348.357), indicam maior variabilidade nos registros de óbitos para esses anos.
4. **Mínimo (Min):** Representa o menor número de óbitos registrados em um único caso em cada ano, que se mantém relativamente estável entre 2 mil e 2.500.
5. **1º Quartil (25%):** O valor abaixo do qual 25% dos dados se encontram. Esse valor cresce progressivamente ao longo do período, indicando um aumento geral nos registros.

6. **Mediana (50%):** O valor central dos dados, onde 50% dos registros estão abaixo e 50% estão acima. A mediana segue um padrão de aumento ao longo dos anos, evidenciando o crescimento gradual no número de óbitos.
7. **3º Quartil (75%):** O valor abaixo do qual 75% dos dados se encontram, também mostrando crescimento ao longo dos anos.
8. **Máximo (Max):** O maior número de óbitos registrados em um único caso, que cresce de 1.264.175 em 2015 para 1.552.739 em 2020.

```
# Estatísticas descritivas para entender a distribuição dos dados
print(obitos.describe())
```

```
Ano do Óbito          2015          2016          2017          2018  \
count         2.000000e+01  2.000000e+01  2.000000e+01  2.000000e+01
mean          1.264175e+05  1.309774e+05  1.312663e+05  1.316719e+05
std           2.820778e+05  2.922092e+05  2.927856e+05  2.935558e+05
min           2.100000e+01  2.000000e+01  1.900000e+01  2.100000e+01
25%           6.225750e+03  6.627000e+03  6.491500e+03  6.519000e+03
50%           3.563500e+04  3.811850e+04  3.962800e+04  4.223150e+04
75%           9.456150e+04  9.752150e+04  9.865150e+04  9.872725e+04
max           1.264175e+06  1.309774e+06  1.312663e+06  1.316719e+06

Ano do Óbito          2019          2020
count         2.000000e+01  2.000000e+01
mean          1.349801e+05  1.552739e+05
std           3.006927e+05  3.438574e+05
min           2.300000e+01  2.500000e+01
25%           7.131000e+03  6.766750e+03
50%           4.640050e+04  4.493750e+04
75%           9.831375e+04  1.458800e+05
max           1.349801e+06  1.552739e+06
```

Figura 3. Estatísticas descritivas para entender a distribuição dos dados

Essas estatísticas descritivas ajudam a identificar padrões importantes nos dados antes da modelagem. A presença de valores máximos muito superiores às médias pode indicar outliers que necessitam de análise para possível remoção ou ajuste. O aumento no desvio padrão ao longo dos anos sugere maior dispersão nos dados, podendo exigir normalização para evitar impactos desproporcionais. Além disso, o crescimento consistente da mediana e dos quartis superiores indica uma tendência de aumento no número de óbitos, relevante para análises futuras.

4.2. Detecção de *outliers*

A Figura 4 apresenta um gráfico de caixas (boxplot) que mostra a distribuição dos óbitos por capítulo (ou categorias) em diferentes anos (2015 a 2020). Cada linha representa os dados de um ano, e o boxplot fornece uma visão clara dos seguintes aspectos estatísticos:

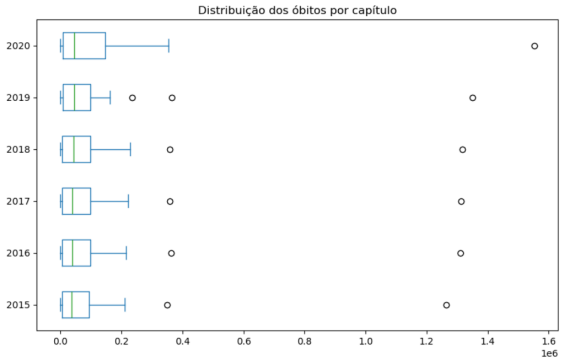

Figura 4. Gráfico de boxplot para detectar outliers

Podemos observar que há outliers em todos os anos, com valores muito altos (próximos a 1,2 milhão e 1,5 milhão). Isso indica que algumas categorias ou regiões tiveram números de óbitos excepcionalmente altos em comparação com o restante.

A mediana (linha dentro da caixa) parece aumentar ao longo dos anos, especialmente em 2020, o que sugere crescimento nos óbitos registrados.

O tamanho da caixa (IQR) é relativamente estável ao longo dos anos, indicando que a dispersão dos dados na maioria dos registros não mudou muito.

Assim, o gráfico reforça que há grande desigualdade nos números de óbitos entre diferentes capítulos ou regiões, com alguns registros extremamente altos (outliers). O aumento na mediana em 2020 pode estar associado a um evento significativo, como a pandemia de COVID-19.

4.3. Cálculo da correlação entre as variáveis

Uma matriz de correlação é uma medida estatística que varia de -1 a 1. Sendo **1** correlação perfeita positiva (os valores aumentam ou diminuem de forma idêntica). **0:** nenhuma correlação (não há relação entre os valores) e **-1:** Correlação perfeita negatica (quando um valor aumenta, o outro diminui de forma idêntica).

A Figura 5 representa uma matriz de correlação que mede o grau de associação entre os óbitos registrados entre os anos de (2015 a 2020). Cada célula na matriz indica o coeficiente de correlação entre dois anos específicos.

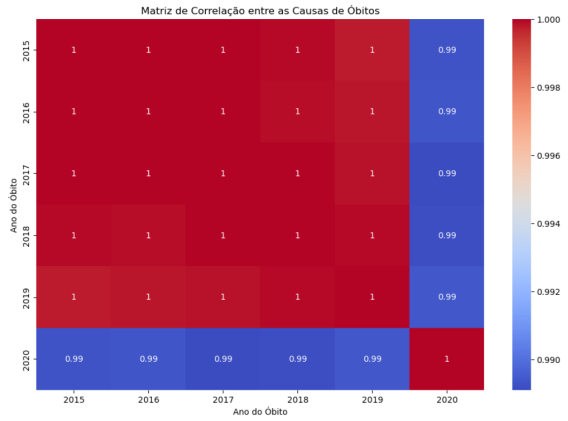

Figura 5. Cálculo da correlação entre as variáveis

Interpretação da matriz:

1. **Diagonal principal:** Os valores da diagonal (de 2015 com 2015, 2016 com 2016, etc) são sempre iguais a 1, porque representam a correlação de um ano consigo mesmo.
2. **Fora da diagonal:** Os próximos a 1 indicam uma forte correlação positiva entre os óbitos de diferentes anos. Por exemplo, os anos de 2015 a 2019 têm uma correlação de 1 entre si, indicando que os padrões de óbitos foram extremamente semelhantes nesses anos.
3. **Ano de 2020:** A correlação de 2020 com os outros anos (2015 a 2019) é menor, cerca de 0,99, o que ainda é uma correlação forte, mas sugere uma leve mudança

nos padrões de óbitos em 2020. Isso pode ser explicado por fatores excepcionais ocorridos em 2020, como a pandemia de COVID-19, que impactou significamente o número de óbitos.

Assim, a matriz indica que os padrões de óbitos foram altamente consistentes entre 2015 e 2019, mas em 2020 uma pequena variação nos dados, provavelmente devido a eventos externos que afetaram as causas ou distribuição dos óbitos.

4.4. Decomposição da série temporal para análise de tendências e sazonalidades

A Figura 6 mostra a decomposição de uma série temporal de óbitos para o CID (Classificação Internacional de Doenças) identificado o CID ”I”. A decomposição foi realizada com o objetivo de analisar componentes de série temporal: tendência, sazonalidade e resíduos.

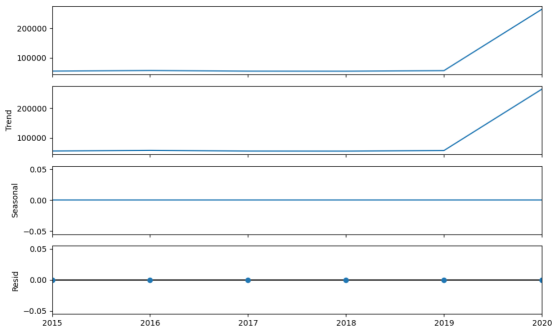

Figura 6. Decomposição da série temporal para análise de tendências e sazonalidades

A seguir será detalhado cada parte da imagem:

1. **Trend (Tendência):** Este gráfico mostra a tendência geral dos óbitos ao longo do tempo. Observa-se que a tendência é relativamente estável entre 2015 e 2019, mas um aumento significativo em 2020. Isso sugere que houve um aumento no número de óbitos relacionados ao CID ”I”em 2020, possivelmente devido a um evento específico (como a pandemia de COVID-19).
2. **Seasonal (Sazonalidade):** A componente sazonal representa padrões repetitivos que ocorrem em intervalos regulares de tempo (como flutuações mensais ou anuais). No gráfico, a sazonalidade está praticamente inexistente (valores próximos de 0), indicandi que não há um padrão sazonal relevante nessa série temporal.
3. **Resid (Resíduos):** Os resíduos representam as variações que podem ser explicadas pela tendência ou pela sazonalidade. No gráfico, os resíduos estão praticamente nulos ao longo de todo período analisado, o que indica que a série é bem explicada pela tendência.

Com isso, este gráfico reforça que os óbitos relacionados ao CID ”I”apresentam um comportamento consistente ao longo do tempo, com uma tendência de aumento acentuado em 2020. Não há evidências de sazonalidade ou grandes variações inexplicáveis (resíduos). Esse padrão pode ser atribuído a um evento excepcional, como a pandemia de COVID-19, que impactou diretamente doenças classificadas sob o CID ”I”(possivelmente doenças cardiovasculares).

4.5. Ocorrências de Nascimentos x Óbitos por Ano

O gráfico ilustrado na Figura 7 ilustra a relação entre nascimentos e óbitos no Brasil ao longo de seis anos, de 2015 a 2020. Os dados evidenciam uma tendência clara: o número de nascimentos supera consistentemente o número de óbitos em todos os anos analisados. No entanto, é possível observar uma ligeira redução no número de nascimentos ao longo do período, enquanto os óbitos apresentam um comportamento mais estável, com leve aumento nos anos mais recentes. O declínio gradual no número de nascimentos pode refletir mudanças sociais, econômicas e culturais, como a transição demográfica que o Brasil atravessa, caracterizada pela redução das taxas de natalidade devido ao maior acesso a métodos contraceptivos, educação, e planejamento familiar. Por outro lado, o aumento leve, mas consistente, dos óbitos, culminando em um pico em 2020, pode estar associado a fatores como o envelhecimento populacional e o impacto de crises sanitárias.

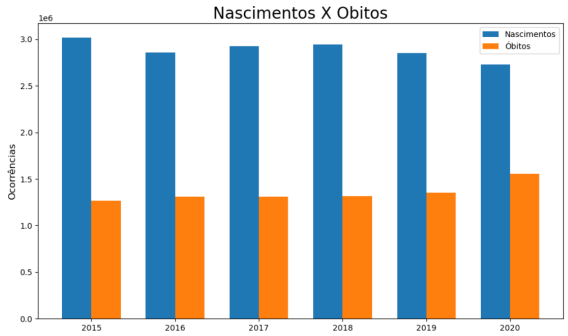

Figura 7. Ocorrências de Nascimentos x Óbitos por Ano

5. Principais métricas utilizadas

A partir da matriz de confusão, como ilustrado pela Tabela 2, é possível calcular as principais medidas que serão usadas para comparação dos algoritmos.

Tabela 2. Exemplo de matriz de confusão para um problema de classificação

	Observado	
	Classe 1	**Classe 2**
Classe 1	Verdadeiro Positivo (a)	Falso Positivo (b)
Classe 2	Falso Negativo (c)	Verdadeiro Negativo (d)

Fonte: Elaboração própria.

As linhas são classes preditas e as colunas são classes reais e os valores da diagonal principal são os valores que o modelo gerou como prognósticos corretos. Ou seja, os valores do primeiro quadrante (a) significam que o ponto observado do dado era da Classe 1 e o valor predito também foi da Classe 1, e os valores do último quadrante (d) seguem a mesma lógica, porém com a classe 2.

Os valores de *b* indicam que o ponto observado era da Classe 2, porém o modelo fez a predição que era da Classe 1, enquanto que os valores de *c* significam que os pontos observados eram da Classe 1 e o modelo classificou como sendo da Classe 2.

A partir dos valores da matriz de confusão foram calculadas algumas métricas. São elas:

$$Acurácia = \frac{(a+d)}{(a+b+c+d)} \tag{1}$$

$$Precisão = \frac{a}{(a+b)} \tag{2}$$

$$Sensibilidade = \frac{a}{(a+c)} \tag{3}$$

$$F1 = \frac{2 \times Precisão \times Sensibilidade}{Precisão + Sensibilidade} \tag{4}$$

A Acurácia (Equação 1) determina os números de predições feitas corretamente pelo modelo sobre todas as predições feitas. Já a medida de Precisão (Equação 2) nos diz qual a proporção de observações da Classe 1 que foram classificadas como tal que realmente eram da Classe 1 e Sensibilidade (Equação 3) calcula a proporção de observações que realmente eram da Classe 1 que foram diagnosticados pelo algoritmo como sendo da Classe 1, ou seja, é a proporção de verdadeiros positivos [Awad and Khanna 2015]. F1 é uma medida da média ponderada das métricas de Precisão e Sensibilidade. Portanto, essa pontuação leva em conta tanto os falsos positivos quanto os falsos negativos [Awad and Khanna 2015].

6. Funcionamento dos Algoritmos

6.1. Naive Bayes

O N*aive Bayes* é um método que examina a distribuição condicional das variáveis preditoras em cada classe, assumindo que os preditores são independentes entre si [Pan et al. 2017]. Baseado no Teorema de Bayes, ao algoritmo busca responder à seguinte pergunta: "Qual a probabilidade de a variável desfecho *Y* pertencer a uma classe *Cl*, dado os preditores X observados ?"[Kuhn and Johnson 2013]. Para isso, o modelo estima $P(Y = Cl \mid X_n)$, ou seja, a probabilidade de *Y* ser uma classe *Cl*, dado o valor de um preditor *X_n*. Na análise apresentada, avalia-se, por exemplo, qual a probabilidade de pertencer à classe P0 ou P50 com base em uma variável preditora. A classe prevista será aquela com a maior probabilidade condicional associada. Dessa forma, a essência do modelo está no cálculo das probabilidades condicionais e incondicionais dos preditores [Kuhn and Johnson 2013].

6.2. Random Forest

O algoritmo *Random Forest* constrói múltiplas árvores de decisão a partir de amostras reamostradas dos dados e utiliza subconjuntos aleatórios de variáveis em cada divisão, introduzindo aleatoriedade para evitar o sobreajuste [Kuhn and Johnson 2013]. As previsões finais são combinadas a partir de todas as árvores. O principal parâmetro do modelo é o **mtry**, que representa o número de preditores selecionados aleatoriamente em cada divisão, geralmente definido como k $\approx \sqrt{p}$, sendo *p* o total de preditores. O critério de divisão utilizado é o **Índice de Gini**, que avalia a pureza dos nós filhos e

ajuda a selecionar divisões ótimas, além de fornecer uma medida geral da importância das variáveis. No ajuste do modelo neste estudo, foram utilizadas 500 árvores com *mtry=3* [Pan et al. 2017]. A abordagem avalia a relevância das variáveis para identificar fatores que influenciam a mortalidade adulta, considerando dados relacionados à saúde, saneamento, educação, renda e condições sociais [Kuhn and Johnson 2013].

7. Resultados e Discussão

O desempenho geral dos modelos no código foi avaliado por meio de múltiplas métricas e técnicas. A seguir será detalhado cada uma delas com base nos resultados obtidos:

7.1. Random Forest

Ao analisar o desempenho do algoritmo *Random Forest* no conjunto de dados completo, observa-se uma acurácia de 100% na classificação dos dados. A Precisão, Recall e F1-Score para todas as classes (Baixo, Médio, Alto) foram de 1.00. Isso sugere que o modelo conseguiu classificar todas as amostras corretamente no conjunto completo de dados.

Como pode-se observar na Figura 8, todas as previsões estão na diagonal principal da matriz de confusão, indicando que o modelo não cometeu erros nesse conjunto.

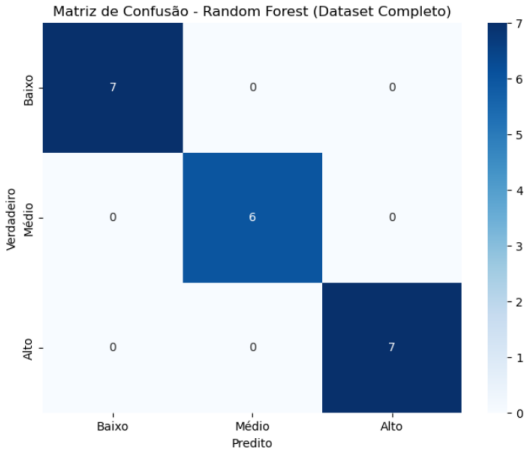

Figura 8. Matriz de Confusão *Random Forest*

7.2. Naive Bayes

Analisando o desempenho do algoritmo *Naive Bayes* no conjunto de dados completo, observa-se uma acurácia de 95% na classificação dos dados.

A Classe Alto, observa-se um Recall menor (0.86), indicando que o modelo deixou de identificar corretamente algumas amostras dessa classe. Para a Classe Médio, a Precisão menor (0.86), indica que algumas amostras de outras classes foram incorretamente classificadas como Médio. Porém, apesar disso, os valores de precisão, recall e F1-Score são bastante altos.

Ao observar a matriz de confusão indicada pela Figura 9, observa-se algumas predições incorretas estão fora da diagonal principal, especialmente para a classe Alto.

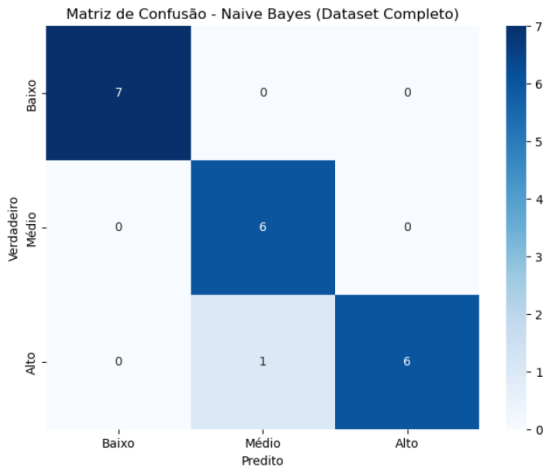

Figura 9. Matriz de Confusão *Naive Bayes*

7.3. Validação Cruzada (5-Folds)

Realizou-se uma validação cruzada com 5-folds a fim de avaliar a generalização do modelo, verificando como ele se comporta em diferentes divisões de treino e teste, minimizando o risco de *overfitting*.

Para o algoritmo **Random Forest** a média da acurácia foi de 75% com um desvio padrão de 22%. O desempenho apresentado é inconsistente, com alta variação entre os folds. Isso indica que o modelo está sofrendo com overfitting no conjunto completo de dados.

Para o algoritmo *Naive Bayes* a média da acurácia é de 90% apresentando um desvio padrão de 12%. O desempenho do modelo é mais estável do que o *Random Forest*. A menor variância sugere que o modelo é mais robusto, ainda que menos flexível.

Por fim, a importância das variáveis revela quais anos tiveram maior influência no desempenho do *Random Forest*:

- 2018 (18.6%) e 2019 (18.2%) foram os anos mais relevantes.
- 2020 (13.9%) foi o ano menos relevante, o que pode identificar menor variabilidade nos dados desse período.

8. Conclusão

Em conclusão, a escolha do modelo de aprendizado de máquina deve ser orientada pelo objetivo final da análise. O *Random Forest* é particularmente eficaz em capturar padrões complexos e interações entre as variáveis, mas pode ser propenso a overfitting se não forem feitos ajustes adequados, como a definição de parâmetros de regularização.Por outro lado, o *Naive Bayes* tende a ser mais robusto em termos de generalização, oferecendo um desempenho superior em tarefas de validação cruzada, especialmente quando se trata de dados com alta dimensionalidade ou distribuições simples. No contexto das questões de pesquisa propostas sobre mortalidade no Brasil, o modelo *Random Forest* seria útil para identificar padrões complexos de mortalidade associados a fatores sociodemográficos e doenças específicas, como as cardiovasculares e respiratórias.

No entanto, para uma análise de previsão de mortalidade a longo prazo, seria interessante explorar modelos mais robustos e generalizáveis, como o *Naive Bayes*, que poderiam oferecer um desempenho mais consistente com dados novos. Em relação às perguntas de pesquisa, os resultados indicam que, embora seja possível realizar uma análise exploratória de mortalidade, seria necessário ajustar o modelo e incluir variáveis adicionais, como localização geográfica, fatores de risco e categorias específicas de CID, para uma compreensão mais completa dos padrões e tendências de mortalidade no Brasil ao longo dos anos.

9. References

Referências

[Andrade et al. 2013] Andrade, M. V. et al. (2013). Desigualdade socioeconômica no acesso aos serviços de saúde no brasil: um estudo comparativo entre as regiões brasileiras em 1998 e 2008. *Economia Aplicada*, 17(4):623–645.

[Arpino et al. 2018] Arpino, B., Le Moglie, M., and Mencarini, L. (2018). Machine-learning techniques for family demography: an application of random forests to the analysis of divorce determinants in germany. In *Annual Meeting PAA, 83.*, Denver. PAA.

[Awad and Khanna 2015] Awad, M. and Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress Open, New York, NY.

[Billari et al. 2006] Billari, F. C., Fürnkranz, J., and Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: a machine learning approach. *European Journal of Population*, 22(1):37–65.

[Campos and Rodrigues 2004] Campos, N. O. B. and Rodrigues, R. N. (2004). Ritmo de declínio nas taxas de mortalidade dos idosos nos estados do sudeste, 1980-2000. *Revista Brasileira de Estudos Populacionais*, 21(2):323–342.

[Cutler et al. 2006] Cutler, D. M., Deaton, A. S., and Lleras-Muney, A. (2006). *The determinants of mortality*. National Bureau of Economic Research, Cambridge.

[De Rose and Pallara 1997] De Rose, A. and Pallara, A. (1997). Survival trees: an alternative non-parametric multivariate technique for life history analysis. *European Journal of Population*, 13(3):223–241.

[IBGE – Instituto Brasileiro de Geografia e Estatística 2010] IBGE – Instituto Brasileiro de Geografia e Estatística (2010). *Observações sobre a evolução da mortalidade no Brasil: o passado, o presente e perspectivas*. IBGE, Rio de Janeiro, RJ. 2010c.

[Jiang et al. 2017] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4):230–243.

[Kuhn and Johnson 2013] Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York, NY.

[Kunst et al. 1999] Kunst, A. E., Wolleswinkel-Van Den Bosch, J. H., and Mackenbach, J. P. (1999). Medical demography on the netherlands: recent advances, future challenges. In

Van Wissen, L. J. G. and Dykstra, P. A., editors, *Population issues: an interdisciplinary focus*. Kluwer Academic/Plenum Publishers, New York, NY.

[Mosley and Chen 1984] Mosley, W. H. and Chen, L. (1984). An analytical framework for the study of child survival in developing countries. *Population and Development Review*, 10(Supl.):25–45.

[Omran 1971] Omran, A. R. (1971). The epidemiologic transition: a theory of the epidemiology of population change. *The Milbank Memorial Fund Quarterly*, 49(4):509–538.

[Pan et al. 2017] Pan, I., Cusimano, A., Ludwig, A., Heisler, E., Carroll, C., Rouland, B., and Freund, K. (2017). Machine learning for social services: a study of prenatal case management in illinois. *American Journal of Public Health*, 107(6):938–944.

[Prata 1992] Prata, P. R. (1992). A transição epidemiológica no brasil. *Cadernos de Saúde Pública*, 8(2):168–175.

[Queiroz et al. 2017] Queiroz, B. L. et al. (2017). Adult mortality differentials and regional development at the local level in brazil, 1980-2010. In *Annual Meeting of the Population Association of America*, Chicago. PAA. [S. l.].

[Santana et al. 2015] Santana, P. et al. (2015). Mortality, material deprivation and urbanization: exploring the social patterns of a metropolitan area. *International Journal for Equity in Health*, 14(1):1–13.

[Schmertmann and Gonzaga 2018] Schmertmann, C. P. and Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, 55(4):1363–1388.

[Sousa and Hamann 2009] Sousa, M. F. and Hamann, E. M. (2009). Programa saúde da família no brasil: uma agenda incompleta? *Ciência & Saúde Coletiva*, 14(1):1325–1335.

[Vaupel 2010] Vaupel, J. W. (2010). Biodemography of human ageing. *Nature*, 464(7288):536–542.

[WHO – World Health Organization 2010] WHO – World Health Organization (2010). Urbanization and health. *Bulletin World Health Organization*, 88(4):241–320.

[Wood and Carvalho 1994] Wood, C. H. and Carvalho, J. A. M. d. (1994). *A demografia da desigualdade no Brasil*. IPEA, Brasília, DF.