UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

DAVI DOS SANTOS FREITAS

ANÁLISE E PREDIÇÃO DE MORTALIDADE INFANTIL UTILIZANDO MODELOS DE APRENDIZADO DE MÁQUINA

QUIXADÁ

2023

DAVI DOS SANTOS FREITAS

ANÁLISE E PREDIÇÃO DE MORTALIDADE INFANTIL UTILIZANDO MODELOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Orientador: Prof. Dr. Davi Romero de Vasconcelos.

QUIXADÁ

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

F936a Freitas, Davi dos Santos.
Análise e predição de mortalidade infantil utilizando modelos de aprendizado de máquina / Davi dos Santos Freitas. – 2023.
55 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2023.
Orientação: Prof. Dr. Davi Romero de Vasconcelos.

1. Aprendizado do computador. 2. Mortalidade Infantil. 3. Análise lógica de dados. I. Título.

CDD 005.1

DAVI DOS SANTOS FREITAS

ANÁLISE E PREDIÇÃO DE MORTALIDADE INFANTIL UTILIZANDO MODELOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Aprovada em: _____/_____/_____.

BANCA EXAMINADORA

Prof. Dr. Davi Romero de Vasconcelos (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Criston Pereira de Souza
Universidade Federal do Ceará (UFC)

Prof. Dr. Regis Pires Magalhães
Universidade Federal do Ceará (UFC)

RESUMO

A mortalidade infantil é um sério problema social que ainda é discutido globalmente. O seu índice é definido como número de crianças que morrem antes de completar 1 ano de vida para cada 1.000 nascimento. Dado a sua urgência, a OMS (Organização Mundial de Saúde) tem como meta até 2030 encerrar as mortes evitáveis de recém-nascidos até os 5 anos. Apesar da redução desse indicador, muitos países ainda sofrem com a sua ocorrência. Com o avanço da tecnologia, é possível utilizar o aprendizado de máquina para realizar predições dessa problemática. Este presente trabalho visa realizar um estudo dos conjuntos de dados do SIM e do SINASC disponibilizados pelo DATASUS, dessa forma realizando uma análise geral dos dados com métodos estatísticos para identificar as principais variáveis ligadas a mortalidade infantil. Além disso, é desenvolvido modelos de aprendizado de máquina para predizer a mortalidade infantil. Como os dados são desbalanceados, foram testadas duas técnicas de reamostragem: SMOTE e Random Undersampling

Palavras-chave: aprendizado do computador; mortalidade infantil; análise lógica de dados.

ABSTRACT

Infant mortality is a serious social problem that is still being discussed globally. Its rate is defined as the number of children who die before their first birthday for every 1,000 births. Given its urgency, the WHO (World Health Organization) has set a goal of ending preventable deaths of newborns by the age of 5 by 2030. Despite the reduction in this indicator, many countries still suffer from its occurrence. With advances in technology, it is possible to use machine learning to make predictions about this problem. The aim of this work is to study the SIM and SINASC data sets made available by DATASUS, thus carrying out a general analysis of the data using statistical methods to identify the main variables linked to infant mortality. Machine learning models are also developed to predict infant mortality. As the data is unbalanced, two resampling techniques were tested: SMOTE and Random Undersampling.

Keywords: computer learning; infant mortality; logical data analysis.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

A mortalidade infantil é definida como o número de crianças que morrem antes de completar 1 ano de vida para cada 1.000 nascimentos (OECD, 2023). Portanto, é um indicador social de extrema relevância, uma vez que fornece dados não somente relativos à saúde materna e infantil, mas também atua como um indicador da qualidade de saúde geral de uma sociedade (CDC, 2023).

De maneira global, os índices de mortalidade infantil apresentam um declínio bastante expressivo desde 1990. O número total de óbitos antes de alcançar os 5 anos reduziu de 12,9 milhões em 1990 para 5 milhões em 2021, um declínio de 59% (WHO, 2023). Portanto, observa-se que a tendência desse problema social é de permanecer em redução até que possivelmente os seus índices sejam completamente zerados. Para a Organização Mundial de Saúde (OMS), a nova meta até 2030 trata de encerrar as mortes evitáveis de recém-nascidos, bem como de crianças menores de 5 anos. Entende-se como morte evitável toda morte que é prevenível, seja total ou parcialmente, mediante ações dos serviços de saúde em um determinado local e época (MALTA *et al.*, 2007). Espera-se, portanto, uma colaboração mútua entre todos os países no objetivo de reduzir as mortes neonatais.

As principais causas globais para mortes infantis são doenças infecciosas — como infecção respiratória aguda —, anomalias congênitas, trauma e asfixia perinatal (WHO, 2023). O acesso a serviços básicos de saúde, como cuidados pós-natais adequados, dieta adequada, aplicação de vacinas quando solicitado pelas autoridades de saúde, e outras medidas preventivas, auxiliam a melhorar a qualidade de vida das crianças (WHO, 2023). Para crianças menores de 1 ano, é importante considerar como a mãe cuida do bebê antes dele nascer, pois isso pode afetar a saúde do recém-nascido.

Embora o Brasil tenha reduzido a taxa de mortalidade infantil em todas as regiões do território nacional para crianças de até 1 ano — um comportamento convergente aos outros países do mundo —, ainda há traços de desigualdades tanto intra quanto inter-regionais que interferem na persistência dessa problemática, sobretudo por afetar a saúde da criança (SAÚDE, 2021).

Em 2010, o Brasil registrou uma taxa de mortalidade infantil de 16,0 a cada mil nascidos vivos. No Nordeste, por exemplo, com uma taxa de 19,0 a cada mil nascidos vivos, apesar da redução deste indicador, os locais com piores condições de vida apresentaram um aumento no risco de morte infantil em comparação com aqueles com melhores condições

(SAÚDE, 2021). Enquanto isso, em 2019, a região Sul foi uma das regiões que apresentou os menores índices de mortalidade (10,1 a cada mil nascidos), a maioria dos valores inferiores à média nacional (13,4) com poucas ocorrências que ultrapassem o valor médio (KREUTZ; SANTOS, 2022). Para o mesmo ano, o nordeste apresentou um índice de 15,2, acima da média nacional.

O progresso tecnológico na ciência de dados (data science em inglês) permitiu um notável aperfeiçoamento na busca e extração de informações para fins analíticos, de tal forma que o esforço empregado sobre um conjunto de dados resulta em um conjunto de benefícios. De acordo com Medeiros *et al.* (2020), em um estudo sobre a utilização da ciência de dados para a área de negócios (DSB, do inglês *data science for business*) foram identificados quatro elementos benéficos principais: qualidade de dados, inteligência analítica, capacidades dinâmicas e avanços competitivos.

Em relação aos dois primeiros mencionados, esses elementos representam dois aspectos encontrados na área de ciência de dados, conforme apresentado pelo pesquisador Cao (2017), sendo: a melhoria na qualidade de dados e análise profunda, o aprendizado e descoberta. Esses aspectos descritos por Cao são reforçados quando estudados por Weihs e Ickstadt (2018), que apresentam o impacto da estatística na ciência de dados como disciplina relevante para fornecer ferramentas e métodos cujos efeitos aprofundam o conhecimento a respeito de um conjunto de dados.

Neste sentido, avanço tecnológico e as ferramentas em constante evolução de ciência de dados provocam uma fusão de outras áreas do conhecimento tanto no meio acadêmico quanto no mercado de trabalho. Essa fusão inclui ciência de dados na medicina, de tal forma que o processo de digitalização do sistema de saúde alterou o modo como a medicina e seus estudos clínicos são conduzidos (SANCHEZ-PINTO *et al.*, 2018). Dessa maneira, entende-se a relevância da ciência de dados como modelo de pesquisa e estudo, além melhorar o trabalho de outras áreas do conhecimento.

Por conseguinte, ao referir-se à mortalidade infantil, há uma relação entre as áreas de medicina e estatística, visto que este tema aborda o conhecimento vigente da estatística descritiva, bem como os aspectos estudados pela medicina a fim de que haja uma interpretação e compreensão científica dos dados coletados e analisados.

Em Singha *et al.* (2016), os autores dedicaram o trabalho na finalidade de abordar a problemática da “mortalidade infantil neonatal em razão de acesso inadequado a cuidados

médicos básicos". Para isso, é proposto por eles uma análise de dados para diferentes atributos maternos durante a gravidez que inclui a identificação de um padrão estatístico no objetivo de desenvolver um modelo baseado em aprendizado de máquina (ML, do inglês *machine learning*).

À vista disso, este trabalho objetiva explorar os dados referentes à mortalidade infantil da República Federativa do Brasil de modo a investigar, mediante análise de causalidade, os principais fatores que mantêm ou agravam os índices de mortalidade no país. Assim, o trabalho apresenta as causas para ajudar e encontrar soluções. Os dados são coletados através dos conjuntos de dados públicos disponíveis pelo site do DATASUS, denominados SIM — Sistema de Informação sobre Mortalidade — e SINASC — Sistema de Informação sobre Nascidos Vivos. Assim sendo, espera-se que este trabalho contribua para analistas de dados que estudam e produzem trabalhos sobre a mortalidade infantil.

1.1 Objetivos

Desenvolver uma análise geral de fatores que influenciam na identificação de mortalidade infantil a partir das bases de dados do SIM (Sistema de Informação sobre Mortalidade) e SINASC (Sistema de Informação de Nascidos Vivos).

1.1.1 Objetivos específicos

Os objetivos específicos foram guiados a partir das seguintes perguntas norteadores:

- **Pergunta 1:** Quais são as características para identificar melhor a mortalidade infantil?
    1. Como analisar e descobrir as principais variáveis que influenciam a mortalidade infantil?
    2. Quais fatores contribuem mais para o aumento da mortalidade?
        2.1 A idade da mãe pode influenciar na mortalidade?
            2.1.1 Se sim, qual idade apresenta maior taxa de filhos vivos e mortos?
        2.2 Índices de qualidade de vida de uma região influencia na mortalidade infantil?
- **Pergunta 2:** Quais as principais anomalias estão associadas com mortalidade e quais fatores as influenciam?
- **Pergunta 3:** Qual a situação clínica das crianças ao nascer pela escala Apgar?
    1. O tipo de parto pode afetar na morte da criança?

Dessa forma, segue a seguinte lista de objetivos:

a) Relacionar as bases de dados a fim de identificar atributos relevantes;
b) Fazer análises estatísticas para visualização de informações;
c) Desenvolver modelo de aprendizado de máquina para classificação;
d) Identificar os principais fatores dos índices de mortalidade;

O trabalho foi organizado da seguinte maneira. O Capítulo 2 aborda toda a fundamentação teórica para compreensão do estudo. Em seguida, no Capítulo 3 são apresentados todos os trabalhos relacionados. No Capítulo 4 contém os procedimentos metodológicos e suas respectivas descrições e o cronograma da pesquisa. O Capítulo 5 apresenta os resultados alcançados. Por último, o Capítulo 6 apresenta as conclusões do presente trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os principais conceitos para o desenvolvimento deste trabalho. A Seção 2.1 são expostos conceitos relacionados a *Machine Learning*. Já na Seção 2.2 aborda o conceito de Aprendizado Supervisionado. Na Seção 2.3, são apresentados os Algoritmos de Classificação. As Seções 2.4, 2.5 e 2.6 são destinadas a *Ensemble Methods*, XGBoost e *Näive Bayes*, respectivamente. Para a Seção 2.7, é discorrido as Métricas de Avaliação Para Modelos de Aprendizado de Máquina. Por fim, as Seções 2.7, 2.8 e 2.9 são destinadas à Desbalanceamento de dados, Mineração de Dados e Engenharia de Atributos.

2.1 Aprendizado de Máquina (*Machine Learning*)

Segundo Misilmani e Naous (2019), aprendizado de máquina consiste em técnicas capazes de simular o aprendizado em um software sem ser explicitamente programado e, assim que expostos a novos dados, seu desempenho é otimizada. Assim, entende-se que o conjunto de conceitos contidos na área de aprendizado de máquina pertence a um subconjunto da área da inteligência artificial. Ademais, aprendizado de máquina atua ativa e diretamente nas áreas de estatística e análise de dados (HARRINGTON, 2012).

Pode-se definir um problema de aprendizado como um problema de aprimorar alguma medida de desempenho ao executar alguma tarefa por meio de algum tipo de treinamento (JORDAN; MITCHELL, 2015). Para cada problema de aprendizado, métricas podem ser definidas da mesma forma que o tipo de treinamento para que assim seja identificado a configuração mais adequada.

O processo de aprendizado de máquina pode ser dividido em dois estágios segundo Liu *et al.* (2021):

1. Treinamento do modelo, isto é, o processo de treinamento do modelo de ML no objetivo de encontrar parâmetros que identificam a relação entre variáveis x e y com uma ótima acurácia. Assim, é utilizado uma amostra de treino, em seguida é aplicado uma função de perda para ser calculado a diferença entre os dados para teste e aqueles que foram previstos. Entende-se, portanto, que o objetivo desse estágio é reduzir a função de perda.
2. O segundo estágio é denominado modelo de inferência ou predição. Após o modelo ser treinado, para determinada entrada $\vec{x}$ o valor de saída por ser calculado com a seguinte fórmula $y = f_\theta(\vec{x}_i)$.fórmula. Com a inferência (predição) realizada, o desempenho pode

ser avaliada a partir de certas medidas, como a acurácia e a precisão em caso de problemas de classificação.

2.2 Aprendizado Supervisionado (Supervised Learning)

O aprendizado supervisionado é composto de métodos que, segundo Rokach e Maimon (2010), tentam descobrir o relacionamento entre as variáveis independentes e as dependentes, desse modo a representação a partir de um modelo. O termo supervisionado remete a ideia da presença de um supervisor que, a partir de rótulos, executa a atividade de associação com os dados de treinamento. Além disso, os algoritmos para esse tipo de aprendizado podem também classificar dados não-rotulados em razão do tipo de treinamento aplicado (CUNNINGHAM *et al.*, 2008).

Além disso, é importante que reconheça os dois tipos principais de aprendizagem: classificadores e regressores. Os modelos de classificação realizam uma atividade de mapeamento das entradas para transformá-las em classes pré-definidas; enquanto aqueles relacionados a regressão predizem um certo conteúdo a partir de suas respectivas características. As alternativas para esses classificadores são diversas, entre elas encontra-se *support vector machines*, árvores de decisão, assim por diante (ROKACH; MAIMON, 2010).

2.3 Algoritmos de Classificação

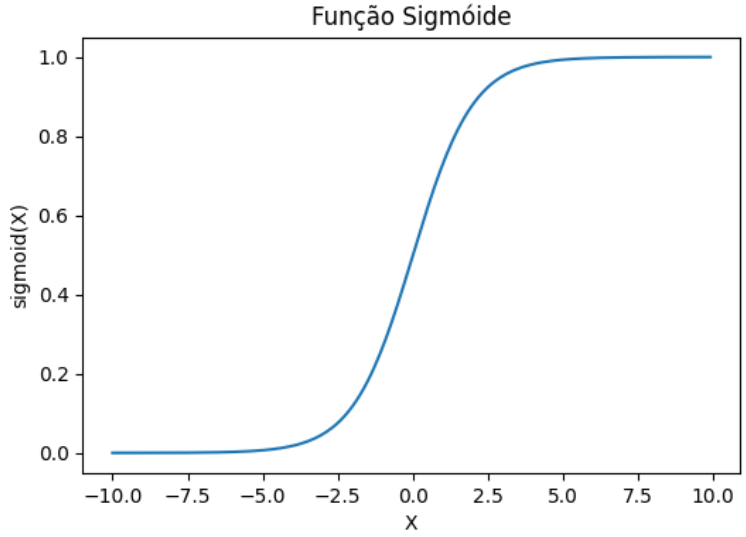
Nesta seção, primeiramente o conteúdo apresenta os algoritmos de classificação selecionados, bem como uma breve introdução a respeito de seus métodos de operação.

2.3.1 Regressão Logística (Logistic Regression)

O modelo de Regressão Logística (LR, do inglês *Logistic Regression*) é utilizado quando o foco da pesquisa é determinar se determinado evento ocorreu ou não, sendo bastante apropriado em problemas que envolvem estados de doença ou saúde e tomadas de decisão. O LR não assume relação linear entre as variáveis dependentes — valores categóricos — e independentes (BOATENG; ABAYE, 2019).

O LR aplica, para cada preditor, um coeficiente cujo valor trata-se de uma medida para a relação entre a variável dependente e independente. Dessa maneira, a variável dependente receberá dois tipos de valores para indicar se determinado evento ocorrerá ou não. Os valores

Figura 1 – Função sigmóide



Fonte: Elaborado pelo autor

para variáveis dependentes binárias são definidas a partir dos valores 0 e 1 e os valores previstos são probabilísticos (BOATENG; ABAYE, 2019). Para ser definida a relação entre os dois tipos de variáveis, o presente modelo utiliza a curva logística, caracterizada pelo formato de S (curva sigmoide) definida pela seguinte equação:

$$\phi(x) = \frac{1}{1 - e^{-x}} \tag{2.1}$$

Para que os valores previstos se mantenham no intervalo $[0,1]$, primeiramente a probabilidade é redefinida como *odds*, isto é, a razão entre a probabilidade de um evento ocorrer e não ocorrer (BOATENG; ABAYE, 2019). Após isso, os valores dos *odds* são convertidos novamente através do seguinte cálculo:

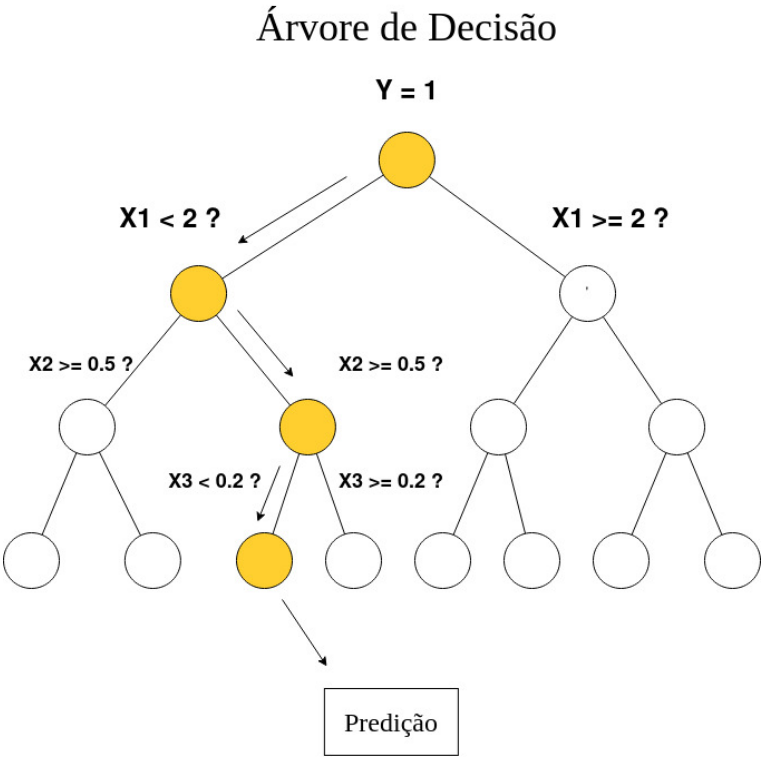
$$P(x) = \frac{odds(x)}{1 + odds(x)} \tag{2.2}$$

A Figura 1 apresenta um exemplo da função sigmoide com um limiar em 0,5. Conclui-se, portanto, que os valores reais são transformados de maneira que sejam inseridos no intervalo $[0,1]$ a partir de uma interceptação em $\phi(x) = 0.5$.

2.3.2 Árvore de Decisão (Decision Tree)

De acordo com Kotsiantis (2013), o modelo de árvore de decisão é um modelo sequencial com sequência de testes combinados logicamente os quais um atributo numérico é

Figura 2 – Árvore de Decisão



Fonte: Elaborado pelo autor

comparado com um valor limiar; e valor nominal é comparado com um conjunto de valores. Por essa lógica, as regras de decisão presentes nesse modelo tornam a sua interpretação mais fácil em comparação ao sistema de pesos numéricos de conexões em uma rede neural.

Uma árvore de decisão é formada por nós e por raiz, e os passos mais importantes para construir um modelo são *splitting*, *stopping*, *prunning* (dividir, parar e podar em português) (SONG; YING, 2015).

Em relação aos nós, há três tipos: (a) o nó raiz — ou nó de decisão — que representa uma escolha lógica que produzirá a divisão de dois ou mais subconjuntos mutuamente exclusivos; (b) nós internos que representam próximas escolhas disponíveis em algum ponto da árvore; (c) nós-folhas cuja representação são os resultados das combinações de eventos ou decisões (SONG; YING, 2015).

As raízes representam saídas do nó raiz e nós internos que atingem os nós-folhas que contém regras de decisão, estas que podem ser representadas por estruturas condicionais (SONG; YING, 2015). Dessa maneira, supondo um exemplo em que se tem uma condição p e q, caso a primeira seja verdadeira, uma saída i será gerada em um nó à esquerda. Já o contrário, a condição q resultará em uma saída j para outro nó, dessa vez à direita da árvore. Esse processo segue sucessivamente até que se alcance o nó-folha.

É possível observar este comportamento na Figura 2. Inicialmente, há um teste aplicado para a classe $y = 1$. No primeiro teste, verifica-se se $x_1 < 2$ ou $x_1 >= 2$. Como a primeira condição é verdadeira, logo a saída é gerada no nó à esquerda. Em seguida, um novo teste é executado para saber se $x_2 >= 0,5$ ou $x_2 < 0.5$. Sendo a primeira condição falsa, a saída dessa vez é gerada no nó à direita. Um último teste é executado e a partir dele é alcançado a predição.

Em relação aos passos executados numa árvore de decisão, tem-se a seguinte descrição sobre os processos segundo Song e Ying (2015):

a) O processo de divisão (*splitting*) ocorre entre as entradas relacionadas com as variáveis de destino, dessa forma separando nós pais de nós filhos. As entradas mais importantes são identificadas e o processo de divisão prossegue até os nós mais internos;
b) Já o passo de parada (*stopping*) é essencial para que um modelo de árvore não seja construído de modo que sua estrutura não atinja altos níveis de complexidade, uma vez que um modelo mais complexo indica uma menor confiabilidade para predizer informações. Dessa forma, aplicam-se regras de parada no momento de construção do modelo, como número mínimo de registros numa folha, profundidade de qualquer nó-folha até o nó raiz, assim em diante;
c) A etapa de podar (*prunning*) é definida como o processo de seleção do tamanho ótimo para a estrutura da árvore ao remover os nós que fornecem a menor quantidade de informação adicional. O modo mais comum de executar esse processo é considerando a proporção de registros com erros de previsão;

2.4 Ensemble Methods

Os *ensemble methods* (métodos de conjunto em português) refere-se a combinação de diferentes classificadores ou regressores em um metaclassificador que apresenta uma taxa de desempenho de generalização melhor do que cada classificador individualmente. Nesse sentido, supondo um cenário em que 10 previsões foram medidas em classificadores diferentes, os *ensemble methods* permitem que essas previsões sejam combinadas de maneira que uma previsão seja gerada com uma acurácia e robustez muito maior do que os classificadores (RASCHKA; MIRJALILI, 2019).

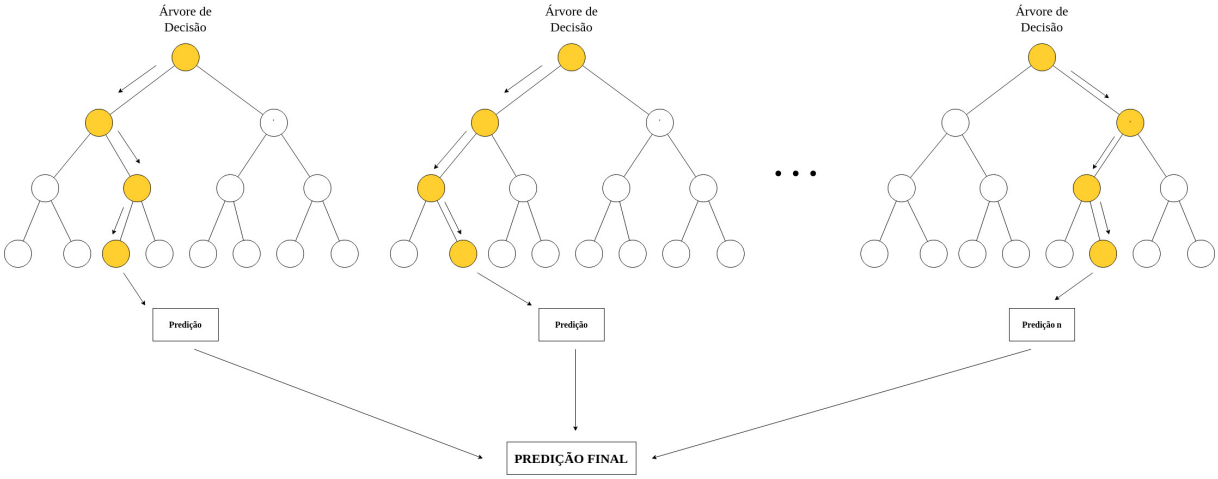
2.4.1 *Agregação Bootstrap (Bootstrap Aggregating)*

Bootstrap Aggregating, do qual deriva o acrônimo Bagging, consiste em uma técnica no qual o conjunto de dados é aleatoriamente dividido em um conjunto de teste T e de treino $£$. Cada classificador é criado a partir de um conjunto de treinos $£$ diferentes que corresponde a uma amostra do *dataset* original (BREIMAN, 1996)

Para um problema de classificação, tem-se que um preditor $\varphi(x,£)$ prever um rótulo de classe $j \in \{1,\ldots,J\}$ e, ao combinar os classificadores, o preditor que alcança um bom resultado individualmente, ao ser agregado, pode ser transformado em um preditor quase ideal (BREIMAN, 1996).

2.4.2 *Floresta Aleatória (Random Forest)*

Figura 3 – Floresta Aleatória



Fonte: Elaborado pelo autor

O modelo de floresta aleatória (RF, do inglês *random forest*) consiste em uma combinação de algoritmos de aprendizado de máquina, especificamente árvores de decisão. Ao combiná-los, os resultados de cada um dos classificadores resultará no valor final que representa o resultado do modelo (LIU *et al.*, 2012). Matematicamente, seguindo o conceito Breiman (2001), RF é um classificador que consiste em uma coleção de classificadores de árvore descrito por $\{h(x,\Theta_k), k = 1,\ldots\}$ em que $\{\Theta_k\}$ são *vectors* aleatórios distribuídos identicamente em que cada árvore lança um voto para a classe mais popular na entrada x.

No processo de construção do RF, cada árvore é alimentada com um novo conjunto de treinamento ao usar seleção aleatória de atributos a partir de métodos *bagging*. Além disso, O

modelo RF contém alta acurácia e, além de tolerar *outliers*, nunca possui *overfitting*.

Na Figura 3, observa-se que este modelo utiliza *n* árvores de decisão que operam separadamente. Cada uma destas árvores retorna um valor de predição. Por fim, o resultado do modelo corresponde a uma combinação de todas as predições realizadas pelas árvores de decisão.

2.5 XGBoost

O *Extreme Gradient Boosting Trees*, também conhecido como XGBoost, é um algoritmo de aprendizado de máquina que se baseia no gradiente descendente e em árvores de decisão. Por ter como base o *gradient boosting*, este algoritmo é utilizado para resolver vários tipos de problema, como classificação e regressão, por exemplo. Assim, cada árvore de decisão utiliza a aplicação do gradiente descendente para gerar um resultado de predição (CHEN; GUESTRIN, 2016).

A maior particularidade do XGBoost está na sua alta escalabilidade em trabalhar com informações com algoritmos otimizados. A utilização de computação distribuída ou paralela permite que o processo de aprendizado seja mais rápido do que os classificadores mais populares (CHEN; GUESTRIN, 2016).

2.6 Näive Bayes

O classificador Näive Bayes é um modelo de classificação que parte do princípio que as features são independentes dada uma classe (RISH *et al.*, 2001). Dessa forma, dado um *vector* de *features* $X = (X_1, \ldots, X_n)$, uma hipótese H e C sendo a classe; para os problemas de classificação é necessário determinar a probabilidade da hipótese H para a amostra observável X, isto é, $P(H|X)$. Simplificando, o objetivo é descobrir a probabilidade da amostra X pertencer à classe C. Nesse sentido, $P(H|X)$ pode ser expresso em termos de probabilidade $P(H)$, $P(X|H)$ e $P(X)$ como:

$$P(H|X) = \frac{P(X|H) \times P(H)}{P(X)} \tag{2.3}$$

O funcionamento do classificador Näive Bayes pode ser compreendido a partir do seguinte fluxo segundo Han *et al.* (2012):

1. Sendo T o conjunto de treinamento, há k classes $C_1, \ldots, C_k$ e, para cada amostra de T, cada uma é representado pelo *vector* $X = (X_1, \ldots, X_n)$, sendo n a quantidade de atributos, $A_1, \ldots, An$.
2. Para cada amostra de X, o classificador prever se X pertence à classe C_k a partir da maior probabilidade *posteriori*, consequentemente encontra-se a classe que maximiza $P(C_i|X)$. Pelo teorema de Bayes, portanto, tem-se:

$$P(C_i|X) = \frac{P(X|C_i) \times P(C_i)}{P(X)} \tag{2.4}$$

3. Como P(X) é igual para todas as classes, logo somente $P(X|C_i) \times P(C_i)$ é maximizada;
4. Para computar todos $P(X|C_i)$ dado um conjunto com muitos atributos, assume-se que os atributos são independentes em relação entre eles para uma determinada classe C_i. Expressa-se matematicamente como:

$$P(X|C_i) \approx \prod_{k=1}^{n} P(x_k|C_i) \tag{2.5}$$

5. Para predizer se a amostra de X pertence a uma classe, então avalia-se $P(X|C_i) \times P(C_i)$ para a classe C_i. Assim, $X \in C_i$, se e somente se, a classe maximiza $P(X|C_i) \times P(C_i)$.

2.7 Métricas de Avaliação para Modelos de Aprendizado de Máquina

Conforme Wardhani *et al.* (2019), dados desbalanceados geram problemas de classificação, uma vez que, caso um determinado modelo seja treinado utilizando esse conjunto de dados, o classificador beneficiará o grupo mais frequente em detrimento de outros, assim sendo necessário o seu balanceamento. Sejam os dados desbalanceados ou não, o modelo precisa ser avaliado utilizando métricas.

Para esta seção, o conteúdo apresenta a descrição das métricas de avaliação utilizadas principalmente para modelos de classificação. É de importante interesse compreender que cada uma dessas métricas é influenciada diretamente por uma matriz de confusão cujo conceito apresenta os valores de *True Positives* (TP), *False Positives* (FP), *True Negatives* (TN), *False Negatives* (FN).

Entre as principais métricas de avaliação para o tipo de modelo citado estão elas: Acurácia (*Accuracy*), Precisão (*Precision*), Revocação (*Recall*), *F1 score* e AUCROC.

Essas quatro métricas serão definidas conforme apresentado por Dalianis (2018) a partir da seguinte matriz de confusão:

Quadro 1 – Matriz de Confusão

		Valor predito	
		Positivo	Negativo
Real	Positivo	42	12
	Negativo	15	36

Fonte: Elaborado pelo autor

2.7.1 Acurácia (Accuracy)

Essa métrica é uma medida definida como a proporção entre instâncias calculadas corretamente, sejam elas verdadeiras ou falsas, e todas as outras instâncias, isto é, uma média aritmética ponderada. Assim, a acurácia possibilita indicar quantos exemplos foram corretamente classificados. Ela é calculada da seguinte forma:

$$Accuracy : A = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

Para a matriz de confusão do Quadro 1, temos:

$$A = \frac{42 + 36}{42 + 36 + 15 + 12} \Rightarrow \frac{78}{105} \therefore A \approx 0,74 \tag{2.7}$$

2.7.2 Precisão (Precision)

A precisão fornece a quantidade de exemplos classificados como positivos e que, portanto, são verdadeiramente positivos. O seu valor é calculado através da razão entre a quantidade de verdadeiros positivos e soma de verdadeiros positivos e falsos positivos. Segue a fórmula abaixo:

$$Precision : P = \frac{TP}{TP + FP} \tag{2.8}$$

Para a matriz de confusão do Quadro 1, temos:

$$P = \frac{42}{42 + 15} \Rightarrow \frac{42}{57} \therefore P \approx 0,74 \tag{2.9}$$

2.7.3 Revocação (Recall)

A métrica de revocação informa a capacidade do modelo em prever exemplos classificados positivamente. Portanto, trata-se da proporção de verdadeiros positivos entre todos os exemplos classificados realmente como positivos. Esse valor é calculado da seguinte maneira:

$$Recall : R = \frac{TP}{TP + FN} \tag{2.10}$$

Para a matriz de confusão do Quadro 1, temos:

$$R = \frac{42}{42 + 12} \Rightarrow \frac{42}{54} \therefore R \approx 0,78 \tag{2.11}$$

2.7.4 F1 Score

O FI Score é uma métrica particular, uma vez que é utilizada quando há presença de classes desbalanceadas. Dessa forma, ela é a média harmônica entre a revocação e a precisão, calculada da seguinte forma:

$$F_1 score : F_1 = 2 \times \frac{P \times R}{P + R} \tag{2.12}$$

Para a matriz de confusão do Quadro 1, temos:

$$F_1 = 2 \times \frac{0,74 \times 0,78}{0,74 + 0,78} \Rightarrow \frac{0,5772}{1,52} \therefore F_1 \approx 0,76 \tag{2.13}$$

2.7.5 AUCROC

A área sob a curva é uma métrica calculada a partir da curva ROC. A curva ROC é utilizada para avaliar o desempenho do modelo em problemas de classificação binária, isto é, que contém dois tipos de elemento. Ao utilizar a taxa de verdadeiros positivos e falsos positivos, é apresentado o desempenho do modelo para distinguir as classes. A AUCROC serve como um resumo da curva ROC ao apresentar o resultado em um único valor probabilístico.

2.8 Desbalanceamento de dados

O desbalanceamento de dados é um tipo dos problemas mais discutidos nas últimas décadas no campo do aprendizado de máquina. Em situações reais é comum que a frequência de cada classe seja bastante diferente entre si, dessa forma dificultando os modelos de aprendizado a trabalharem com essas informações, uma vez nesse cenário os modelos predizem a classe dominante com maior facilidade (KRAWCZYK, 2016). Dessa forma, uma série de técnicas foram desenvolvidas para lidar com esse tipo de problema, entre elas o SMOTE e o *Random Undersampling*.

2.8.1 SMOTE

O *Synthetic Minority Oversampling Technique* (SMOTE) é um algoritmo de pré-processamento para dados desbalanceados aplicando a ideia de *oversampling*, isto é, aumentar o número de amostras da classe minoritária. Para isso, o SMOTE cria dados sintéticos ao invés de apenas replicar dados da amostra (FERNÁNDEZ *et al.*, 2018).

2.8.2 Random Undersampling

A técnica de subamostragem aleatória (*Random Undersampling*) é uma técnica para dados lidar com dados desbalanceados cujo princípio é remover instâncias da classe majoritária aleatoriamente. Esse tipo de abordagem, apesar de reduzir a diferença de frequência entre as classes, ocasiona na perda de informações importantes para o treinamento dos modelos de aprendizado de máquina (MISHRA, 2017).

2.9 Mineração de dados

Segundo Mikut e Reischl (2011), mineração de dados é um passo na descoberta de conhecimento a partir de banco de dados (KDD, do inglês *knowledge discovery from databases*) que aplica uma análise de dados para encontrar padrões. Assim, o KDD é um procedimento não-trivial, o que é coerente com a compreensão mais recente sobre mineração de dados, uma vez que o conceito mais amplo de KDD é utilizado como sinônimo para essa área.

De acordo com Agarwal (2013), mineração de dados é um estudo entre computação e estatística cujo objetivo é descobrir padrões e informações. Dessa forma, o objetivo principal

é extrair informações de um conjunto de dados e modelá-las em uma estrutura ideal para um determinado uso. Além disso, sua atividade processual em relação à extração de dados não é intuitiva, especialmente quando trabalhada em grandes bases de dados contendo informações até então desconhecidas.

O conhecimento obtido pós-extração, segundo Algarni (2016), é valioso e afeta significativamente as decisões, como mostra a EDM (*Educational Data Mining* ou mineração de dados educacionais) na qual a extração em sistemas educacionais pode aprimorar técnicas de aprendizado e ensino.

A mineração de dados é composta por um conjunto de técnicas que, em síntese, podem ser definidas como processos de identificação de padrões e tendências em um conjunto de dados contendo um enorme volume de informações Osman (2019). Com o progresso do campo da mineração de dados, diversas técnicas foram desenvolvidas visando proporcionar maneiras variadas de tratar e transformar os dados conforme as necessidades identificadas pelo cientista de dados; tais como: associação, classificação, clusterização, árvore de decisão, predição, redes neurais, assim por diante. Portanto, cada técnica é usada para resolver um determinado problema e contém regras específicas.

Osman (2019) descreve cada uma das técnicas supracitadas em seu trabalho. A seguir a descrição de cada uma delas:

a) **Associação:** procura de ocorrências a partir de uma conexão de atributos;
b) **Classificação:** derivação de dados partir de um atributo específico;
c) **Clusterização:** agrupamento de dados baseado em semelhança;
d) **Árvore de decisão:** relacionamento de dados a partir de regras;
e) **Predição:** validação de informações a partir de uma classificação;
f) **Redes neurais:** procura de padrões a partir de um modelo neural;

Desse modo, entende-se que cada técnica deve ser aplicada para a resolução de um problema em específico, assim como contém suas especificações de utilização. Osman (2019) descreve cada uma das técnicas supracitadas em seu trabalho.

2.9.1 Associação

Também conhecida como técnica de relação, consiste em descobrir padrões com base no relacionamento entre variáveis. Ao serem relacionadas, as variáveis podem indicar a frequência entre diferentes itens e indicar aquele com maior frequência. No mundo do mercado,

por exemplo, pode auxiliar vendedores a identificar o comportamento de clientes com base no seu histórico de compras.

2.9.2 Classificação

Como o nome sugere, a técnica consiste em classificar dados em diferentes grupos ou classes de modo a alcançar uma melhor acurácia no modelo de predição, assim como uma análise mais precisa no conjunto de dados. Neste sentido, os atributos identificados em uma base de dados podem categorizar informações de modo que sejam aplicadas para um objetivo específico.

2.9.3 Clusterização

O processo envolvendo a clusterização realiza uma etapa de análise de atributos na finalidade de identificar dados similares entre si. Essa técnica também pode ser denominada de segmentação pelo fato desse método segmentar os dados em categorias.

2.9.4 Árvore de decisão

A técnica comporta-se a partir de um critério de seleção em que auxilia na escolha de seleção de dados específicos. Um dado gera duas ou mais folhas de resposta que, por sua vez, geram folhas extras para apoiar a classificação. No fim, os dados podem ser classificados e uma predição pode ser realizada com base na estrutura da árvore.

2.9.5 Predição

Esse método é utilizado junto a outras técnicas de mineração, uma vez que predição inclui atividades como análise de tendência, classificação, relação e casamento de padrão. Assim, é possível utilizar a predição a fim de validar se uma determinada informação será definida em uma especificação de classificação ou não.

2.9.6 Redes Neurais

Esse método consiste em um processo automatizado até um determinado momento, uma vez que não é esperado tanto conhecimento a respeito do trabalho ou banco de dados. Para uma rede neural funcionar de maneira adequada é necessário saber a forma de conexão dos

nós, o número de unidades de processamento e a condição de parada do treinamento. Uma das utilizações mais comuns de redes neurais está em aprendizado não-supervisionado.

2.10 Engenharia de atributos

De acordo com Zheng e Casari (2018), a engenharia de atributos (do inglês *feature engineering*) refere-se a extração de atributos a partir de um conjunto de dados não tratados no objetivo de transformá-los de modo que se tornem adequados para o modelo de aprendizado de máquina.

Toda a atividade que remete a engenharia de atributos é conduzida por um cientista de dados mediante um trabalho manual por tentativa e erro baseado em sua experiência e conhecimento que comprove o seu domínio (NARGESIAN *et al.*, 2017). No entanto, há como utilizar essa atividade automatizada por abordagens já desenvolvidas pelos cientistas por meio de pesquisa guiada usando medidas apoiadas em heurísticas de qualidade (FAN *et al.*, 2010). Dessa maneira, o trabalho executado sobre os dados na finalidade de manipular os seus atributos mesmo em meios objetivando a sua automação ainda é composto de um esforço manual e a longo prazo.

Os autores Dong e Liu (2018) abordam diversos conceitos para o termo engenharia de atributos a partir de diversos aspectos que o compõe, sendo eles:

a) **Transformação de atributo:** do inglês *feature transformation* é a construção de novos atributos a partir daqueles já existentes na base de dados;
b) **Geração de atributo:** *generation feature* trata-se da criação de novos atributos, contudo não sendo necessariamente resultado da transformação de atributo;
c) **Seleção de feature:** *feature selection* é a seleção de um pequeno grupo de atributos a partir de um conjunto maior. Essa atividade fornece uma maior confiabilidade ao trabalhar com certos algoritmos em modelos de aprendizado de máquina;
d) **Análise e avaliação de atributo:** *feature analysis and evaluation* geralmente é incluso no *feature selection*, uma vez que nesse processo ocorre um grupo de conceitos relacionados na medição de utilidade dos atributos;
e) **Metodologia automática geral de engenharia de atributos:** *general automatic feature engineering methodology* é definido como um método de gerar atributos automaticamente em larga escala e selecionar o conjunto mais efetivo.

f) **Aplicações de engenharia de atributos:** também chamado de *feature engineering applications* foca em outras atividades de análise a partir dos atributos de uma base de dados.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados os trabalhos que contribuíram para o desenvolvimento desta pesquisa.

3.1 Contextual, maternal, and infant factors in preventable infant deaths: a statewide ecological and cross-sectional study in Rio Grande do SUL, Brazil (KREUTZ; SANTOS, 2022)

As autoras dessa pesquisa afirmam inicialmente que encerrar com as mortes evitáveis de recém-nascidos e crianças abaixo de 5 anos até 2030 é uma das metas do Grupo das Nações Unidas para Desenvolvimento; dessa maneira, indicando que a mortalidade infantil ainda persiste como um problema de forte preocupação globalmente. O Brasil nos últimos anos reduziu significativamente os índices de mortalidade em diversas regiões, contudo as taxas ainda persistem e é de alta relevância identificar seus fatores.

A partir das informações dispostas sobre os índices de mortalidade infantil no Brasil em suas cinco regiões, esta pesquisa objetiva validar a hipótese de que as causas evitáveis da maioria das mortes infantis estão relacionadas as características da criança ao nascer, maternais e contextuais para a região Sul do país.

As autoras conduziram um estudo ecológico descritivo transversal em todo o estado do Rio Grande do Sul. Para isso, foi utilizado para aquisição de informações gerais a base de dados do SIM por conter informações relacionadas a dados socioeconômicos, local de residência, tipo de morte e condições e causas da morte. Além disso, separaram as mortes evitáveis em três categorias: causa evitável, causa definida por doença e outras causas; e, para classificar as mortes, utilizaram informações extraídas do Certificados de Morte e Certificados de Nascido Vivo.

Os resultados das pesquisas analisou dados de 141.568 nascidos vivos e 1425 mortos menores de 1 ano. Metade das mortes ocorreram na primeira semana de vida; cerca de 79% correspondiam a mortes evitáveis. A taxa de mortes evitáveis foi maior para crianças do sexo masculino, filhos de mães adolescentes e entre crianças de mães sem parceiro. Além disso, para as causas de morte neonatal evitável as mais frequentes foram: síndrome respiratória aguda, septicemia bacteriana não especificada, feto e recém-nascido afetado por desordens maternal, hipertensiva, por corioamnionite e por ruptura prematura das membranas. Cerca de um quarto

das mortes eram de recém-nascidos com alguma malformação — sendo a mais frequente a do sistema cardiovascular — ou síndrome genética.

Em convergência ao objetivo do trabalho proposto, o artigo realiza um estudo estatístico descritivo a respeito dos índices de mortalidade infantil e explorando seus fatores, contudo utilizando apenas a base de dados SIM e limitando-se ao estado do Rio Grande do Sul. Assim, o trabalho a ser desenvolvido expande o seu campo de atuação para outras regiões, além de combinar as informações com a base de dados do SINASC para o ano de 2020.

3.2 Using Predictive Classifiers to Prevent Infant Mortality in the Brazilian Northeast (RAMOS *et al.*, 2017)

Os autores introduzem a pesquisa relatando que a mortalidade infantil trata-se de um problema que afeta todos os países do mundo, principalmente aqueles definidos como subdesenvolvidos. Apesar da redução da taxa de mortalidade infantil no Brasil nos últimos 22 anos — cerca de 77% —, esse problema social ainda é considerado alarmante para os indicadores de sistema de saúde brasileiros. Por isso, o sistema de saúde utiliza um *framework* governamental chamado GISSA[1], pertencente a um programa do governo federal chamado *Stork Network*, cuja principal tarefa é fornecer alerta a respeito da saúde de recém-nascidos e mães grávidas a fim fornecer auxílios em processos de tomada de decisão.

Nesse sentido, o trabalho apresenta o sistema de análise inteligente denominado LAIS pertencente ao framework GISSA que, através da mineração de dados (DM, do inglês *data mining*), gera alerta sobre riscos de morte de recém-nascidos com base em métodos probabilísticos. A partir disso, a pesquisa demonstra o modelo de predição desenvolvido para mortalidade infantil na região nordestina do Brasil por ser a região mais sensível a este problema.

No artigo, os autores apresentam os cinco passos para o trabalho de meta-aprendizado do *framework* GISSA. No primeiro passo há uma seleção de algoritmos para avaliação e seus hiperparâmetros, sendo eles divididos em 6 grupos: árvore de decisão, algoritmos baseados na Teoria de Bayes, redes neurais, métodos Kernel, classificadores elementares e algoritmo baseado em regra. No segundo passo, há a preparação dos dados a partir das bases de dados do SINASC e do SIM dos anos 2013 e 2014, ambos disponíveis no portal DATASUS, em busca das *features* melhor correlacionadas com o tema. Em seguida, os dois próximos passos se concentram na construção iterativa dos modelos e na seleção da melhor abordagem. Por fim, a configuração do

[1] Disponível em: https://lariisasaudedigital.com/gissa/

melhor algoritmo para o sistema LAIS para ser utilizado em produção.

Como resultado, ao aplicar técnicas de ciência de dados para a construção de um modelo de predição para a região nordeste do Brasil e utilizando o algoritmo de classificação Naïve Bayes — que apresentou a melhor acurácia em relação aos outros algoritmos —, o sistema apresentado entrega um serviço capaz de fornecer a probabilidade de uma criança no nordeste brasileiro sobreviver até o primeiro ano de vida. Através disso, unido ao GISSA o LAIS consegue utilizar do sistema de alerta para realizar medições de risco para cada recém-nascido na região.

O presente trabalho, semelhante a esse artigo, utiliza as mesmas bases de dados disponíveis pelo portal DATASUS de modo a elaborar uma pesquisa com o foco em identificar os principais fatores dos índices de mortalidade infantil através da análise de dados; dessa forma explorando de maneira outras *features* descartadas para o objetivo do artigo apresentado por Ramos *et al.* (2017).

3.3 Data Mining and Risk Analysis Supporting Decision in Brazilian Public Health Systems (VALTER *et al.*, 2019)

Para este artigo, os autores apontam a importância do processo de análise de dados para a criação de modelos de tomada de decisão na finalidade de reduzir as taxas de mortalidade materna, neonatal e infantil. Dessa maneira, através do processo de mineração de dados é possível desenvolver um modelo de classificação de risco de morte de mortalidade ao utilizar técnicas de aprendizado de máquina.

Nesse sentido, o objetivo do trabalho em discussão é contribuir com a redução das taxas de mortalidade materna, neonatal e infantil a partir do fornecimento de um serviço web o qual fornece diversos modelos de predição ordenados a partir da disponibilidade de informação. Soma-se a isso a apresentação de uma prova de conceito (*proof of concept* em inglês) na finalidade de demonstrar o uso desse serviço no portal GISSA cuja definição fora apresentada por Ramos *et al.* (2017).

Em relação ao processo metodológico, o fluxo do projeto parte desde o processo de coleta de dados utilizando mineração de dados até a implantação (*deployment* em inglês), portanto tem-se a seguinte linha de atividades: aplicação da metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) para mineração de dados no GISSA; compreensão do domínio de negócios do GISSA e a produção de um dicionário para as bases de dados do SINASC e SIM; preparação e limpeza dos dados para preparação de três conjuntos de dados para

índices de mortalidade materna, neonatal e infantil; aplicação de modelagem de dados a partir de um pré-processamento; avaliação de classificadores supervisionados; e, por fim, implantação do serviço.

O produto final dos esforços aplicados neste trabalho resultou na escolha do melhor algoritmo como modelo de predição para cada um dos três conjuntos de dados construídos em uma das atividades metodológicas anteriores. Dessa maneira, entre os algoritmos de classificação binária escolhidos o *Random Forest* foi aquele com o melhor desempenho para os 3 conjuntos de dados. Para os indicadores maternos, neonatais e infantis, respectivamente, o classificador em destaque apresentou acurácias médias de 97,50%, 93,90% e 99,73%. Conclui-se, portanto, que o processo de mineração de dados no GISSA para a construção de um modelo de predição a partir de um classificador binário demonstrou alta capacidade de generalização para os dados em estudo, além de alcançar uma ótima taxa de acurácia para os conjuntos de dados.

O artigo utiliza a atividade de mineração de dados no sistema GISSA para a extração de informações, assim como detalha as *features* utilizadas para a construção dos 3 conjuntos de dados detalhados na seção de metodologia. Neste sentido, a pesquisa contribui diretamente para o presente trabalho uma vez que fornece uma base mais direcionada para a exploração de dados a fim de extrair informações úteis para a identificação dos principais fatores para os índices de mortalidade.

3.4 Análise comparativa

Essa seção é destinada ao quadro comparativo do trabalho proposto e os relacionados. O Quadro 1 compara os trabalhos a partir dos seguintes pontos: envolvimento do tema Mortalidade Infantil; uso de *machine learning*; *data mining*; análise de causalidade; SIM e SINASC como fonte dos dados da pesquisa.

Todos os trabalhos relacionados abordam o tema da mortalidade infantil como parte central para a definição de seus objetivos. Além disso, entre as bases de dados presentes no DATASUS, o SIM é utilizado em todas as pesquisas, enquanto o SINASC apenas não é utilizado pelo trabalho de Kreutz e Santos (2022).

Os trabalhos que utilizam de técnicas de ML para a resolução de problemas de classificação a partir das bases de dados do DATASUS produzem um trabalho relevante em relação à pré-processamento de dados e engenharia de *features* cujos efeitos produzem um apoio para o desenvolvimento do presente trabalho.

Quadro 2 – Quadro comparativo de trabalhos relacionados

Trabalho	**Algoritmos**	**Conjunto de dados**	**Problema**	**Métricas**
Ramos *et al.* (2017)	ID3, MLP, RF, PART, BN, NB, KNN, VP	SIM, SINASC	Predição de nascido vivo	Precisão, Recall, F-Meas., AUROC
Valter *et al.* (2019)	NB, DT, RF	SIM, SINASC	Predição de risco de morte em estágios de gestação	AUROC, Acurácia
Kreutz e Santos (2022)	Sem abordagem de aprendizado de máquina	SIM, SINASC	Identificação de fatores de mortalidade infantil em casos previsíveis	Estudo descritivo
Este trabalho	LR, DT, RF, NB, SVM	SIM, SINASC	Predição de crianças mortas antes do primeiro ano de vida	Precisão, Recall, F1-Score, Acurácia, AUROC

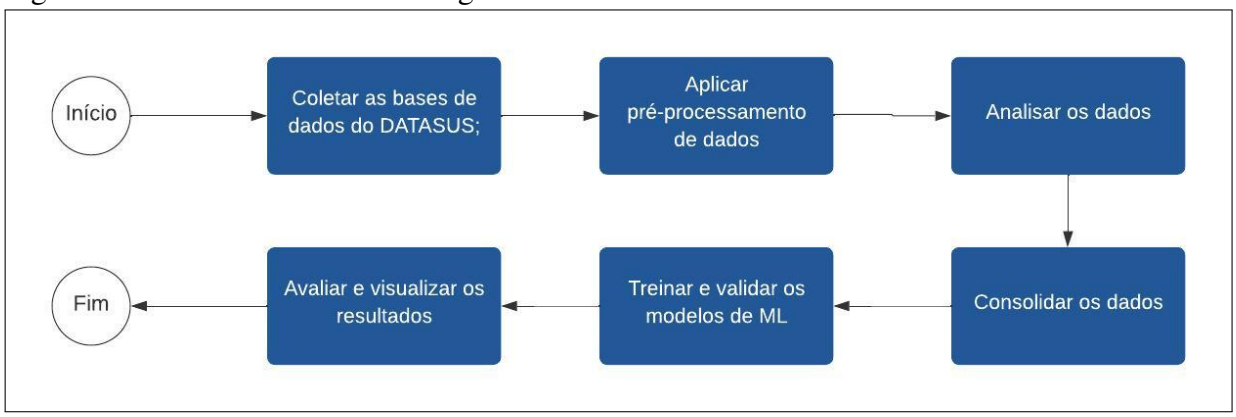
Fonte: Elaborado pela autora

Em relação ao trabalho de Kreutz e Santos (2022), o estudo estatístico descritivo e a produção de uma análise de causalidade apoia o trabalho proposto, uma vez que seu método de estudo colabora com a expansão de uma nova pesquisa para outras regiões ao mesmo tempo que novas análises podem ser realizadas visando a exploração de cenários alternativos.

4 METODOLOGIA

Este capítulo trata das atividades metodológicas a serem executadas para alcançar o objetivo final do trabalho, tal como representado pelo fluxograma abaixo:

Figura 4 – Procedimento Metodológico



Fonte: Elaborado pelo autor

4.1 Coleta dos conjuntos de dados do DATASUS

Os dados coletados para este estudo dizem respeito a nascimentos e óbitos ocorridos no ano de 2020. Essas informações são disponibilizadas pelo DATASUS[1], o qual é um departamento de tecnologia do SUS (Sistema Único de Saúde). Para a coleta dos registros de nascimentos, a base utilizada é o SINASC (Sistema de Informação de Nascidos Vivos) e, para óbitos, a base é o SIM (Sistema de Informação sobre Mortalidade).

Como dados auxiliares, foram coletados informações a respeito dos índices de desenvolvimento humano (IDH) e população dos municípios a fim de apoiar as análises sobre os conjuntos de dados e, posteriormente, verificar a sua importância no treino dos modelos de aprendizado de máquina.

O SIM contém ao todo 1.556.824 linhas e 87 colunas (atributos), enquanto o SINASC é composto por 2.730.145 linhas e 61 colunas. Esse conjunto de informações é utilizado para realizar a análise geral dos dados nas Seções 5.1 e 5.2.

Em relação aos dados para aprendizado de máquina, por questões de tempo de execução, as informações são filtradas para dados correspondentes ao estado do Ceará. Além disso, todos os óbitos registrados a partir de 1 ano de vida foram excluídos. Assim, é possível retornar cerca de 8.930 instâncias do SIM.

[1] Disponível em: https://opendatasus.saude.gov.br/

Em razão da utilização de dois conjuntos de bases distintos, a união dessas informações é necessária. Em trabalhos como Silva *et al.* (2017), a ligação é executada mediante um atributo denominado Número da Declaração de Nascido Vivo (numerodn), descrito no arquivo de estrutura fornecido pelo site do DATASUS. Contudo, pela inexistência desse atributo nas bases atuais, uma alternativa de vínculo precisou ser utilizada.

Segundo a LGPD (Lei Geral de Proteção de Dados Pessoais), somente órgãos de pesquisa podem ter acesso a dados pessoais e conforme uma determinada prática de segurança prevista em algum regulamento (BRASIL, 2019). Assim, deduz se o número de declaração de nascido vivo não esteja disponível de maneira pública por ser considerado um dado pessoal.

Assim, a combinação melhor encontrada é formada pelos campos "Código do município de residência", "Data de nascimento", "Raça", "Sexo", "Idade da mãe", "Tipo de parto" e "Tipo de gravidez". No Quadro 3, é apresentado a quantidade de valores ausentes após a união dos dados.

Quadro 3 – Valores percentuais de campos ausentes nos campos utilizados para unir os dados do SIM e SINASC

Campo	Percentual de valor ausente
Código do município de residência	0,00%
Data de nascimento	0,00%
Raça	2,59%
Sexo	0,00%
Idade da mãe	0,00%
Tipo de parto	0,00%
Tipo de gravidez	0,00%

Fonte: Elaborado pelo autor

4.1.1 Criação do novo conjunto de dados

Apesar dessa combinação, das 8.930 instâncias encontradas no SIM, apenas 3.829 conseguiram ser ligadas com a base do SINASC por conterem as mesmas chaves, logo havendo um retorno de aproximadamente 43% dos dados totais. Unindo as informações ligadas com os dados restantes do SINASC, o conjunto final é composto por 681.688 instâncias.

Para rotular as informações como óbito infantil, foi criado o campo "Vivo" que, para as instâncias presentes tanto no SIM quanto no SINASC, receberam o valor *False*, enquanto as outras instâncias receberam o valor *True*. Além disso, foi identificado um excesso de campos

com valores ausentes, assim sendo, foram removidos todos os campos com mais de 50% dos valores ausentes. Campos desnecessários, como "Versão do sistema" e semelhantes, foram também removidos. A quantidade de colunas no final é de 65, contando com dados de IDH e população dos municípios.

4.2 Aplicação de pré-processamento de dados

As atividades que envolvem análise de dados requerem uma etapa chamada pré-processamento dos dados antes que qualquer estudo seja elaborado. Dessa forma, é crucial preparar e adequar os dados, uma vez que somente após o término desta etapa as informações apresentarão uma maior consistência e qualidade, para adequar-se ao estudo em questão.

Para trabalhar somente com dados de recém-nascidos, foram removidos campos cuja descrição não é informada no documento oficial do conjunto de dados no DATASUS. Ademais, a coluna referente a idade teve de ser trabalhada para a criação da coluna "Tipo de Idade", uma vez que o registro de idade no conjunto de dados engloba indivíduos que viveram minutos, horas, meses e anos. O campo é composto de três dígitos, sendo o primeiro referente a unidade de idade (se 1 = minuto, se 2 = hora, se 3 = mês, se 4 = ano, se = 5 idade maior que 100 anos) e os outros dois é a quantidade de unidades, logo o registro 434 indica que o indivíduo morreu aos 34 anos. Dessa forma, para facilitar a identificação de recém-nascidos, a coluna IDADETIPO indica a unidade de idade de cada registro.

Em relação aos códigos CID contido no conjunto de dados, a coluna "Descrição da Causa Base" é para fornecer uma identificação clara da descrição de anomalia ou quaisquer doenças relacionadas.

Apesar da remoção dos dados ausentes, há ainda campos com dados não preenchidos. Dessa maneira, foi aplicado a técnica de preenchimento com base no valor da mediana para os campos numéricos. Neste sentido, foi aplicado a técnica *MinMaxScaler* para que os valores dos campos pertençam ao intervalo entre 0 e 1.

4.3 Análise dos dados

Nesta atividade, utilizar-se-á conhecimento estatístico e ferramentas de manipulação de informações visando análises mais sofisticadas. Como apoio, os documentos oficiais relacionados às descrições dos atributos nos conjuntos de dados fornecerão informações relevantes para

uma compreensão mais aprofundada.

Compõe a análise um estudo de correlação entre os atributos identificados na etapa anterior para validar hipóteses e suposições de quais informações mais impactam nas informações de mortalidade e, consequentemente, na predição. Em seguida, investiga-se como essas informações se relacionam e influenciam direta ou indiretamente os conjuntos de dados escolhidos, fortalecendo a compreensão do material.

O processo de análise de dados incluirá representações visuais — imagética ou gráfica — para identificar padrões e auxiliar os tomadores de decisão na compreensão de conceitos relevantes para seus trabalhos.

Assim, a interpretação dos dados dependerá do conhecimento sobre a mortalidade infantil e dos termos específicos encontrados nos dicionários correspondentes aos dados.

4.4 Consolidação dos dados

A utilização do conhecimento estatístico no que concerne ao processo analítico desta etapa produzirá uma série de materiais imagéticos e gráficos que comporão o conjunto final de informações. Portanto, o produto dessa atividade, oriunda das duas etapas anteriormente supracitadas, compõe o conjunto de interpretações individuais para cada análise realizada a partir dos dados de modo a gerar um resultado consolidado das informações.

4.5 Treino e validação dos modelos de ML

Os modelos de ML são utilizados para classificar, a partir dos dados coletados, quais crianças foram vítimas de mortalidade infantil. Com isso, o treinamento produz modelos que reconhecem quais padrões de informação são mais úteis para esta identificação.

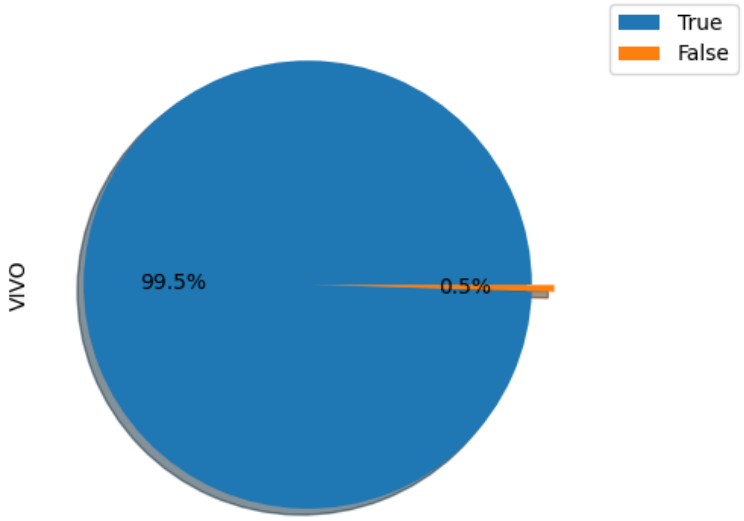
Para treinamento e validação dos modelos, é utilizado a linguagem *Python* com o apoio das seguintes bibliotecas: *Scikit-learn*, *ImbLearn*, *Pandas* e *MatplotLib*.

Antes de escolher os atributos para os modelos de aprendizado, usa-se os *feature selection algorithms* (algoritmos de seleção de atributos em inglês) para identificar atributos que devem ser aplicados nos modelos e avaliá-los. Os algoritmos de seleção de atributos são: variance threshold, f-classif, chi2, mutual info classif, extra trees, logistic regression e linear SVC.

No início do processo de treinamento, os dados são divididos em 80% para dados de

treinamento e 20% para teste. O conjunto de dados para o estado do Ceará somam 649.389 registros, sendo que 99,5% das instâncias correspondem a classe positiva e apenas 0,5% correspondem a classe negativa. Essa desbalanceamento é apresentado na Figura 5.

Figura 5 – Desbalanceamento das classes



Fonte: Elaborado pelo autor

Para contornar este problema, foram selecionados duas técnicas de reamostragem (em inglês, *resampling*): SMOTE e subamostragem aleatória. A aplicação dessas técnicas é aplicada somente aos dados de treinamento, haja vista que, se aplicado em todo o conjunto de dados, informações podem aparecer tanto nos dados de treinamento quanto nos de teste, resultando em um modelo com baixa capacidade de generalização.

Para encontrar a melhor combinação de hiperparâmetros, é utilizado o *GrindSearchCV* para cada um dos modelos escolhidos, utilizando os dados de treinamento. Além disso, a métrica para função de perda escolhida para este algoritmo foi a precisão, que calcula a precisão geral da predição. Para cada modelo, é configurado alguns possíveis valores para os seus respectivos hiperparâmetros.

Em seguida, cada um dos modelos são treinados utilizando as técnicas de balanceamento selecionadas, a fim de encontrar aquela que apresenta o melhor desempenho para trabalhar com o desbalanceamento.

Para cada um dos modelos, as métricas capturadas são: acurácia, precisão, sensibilidade, f1-score e AUROC. São considerados apenas 2 casas decimais para os valores de resultado de desempenho.

4.6 Avaliação dos dados

O presente trabalho valida a análise dos conjuntos de dados para identificar as variáveis que afetam os índices de mortalidade infantil, de modo que este trabalho colabore com essas outras pesquisas. Adicionalmente, os fatores abrangerão aspectos relacionados à medicina e à sociedade, combinando o trabalho com várias áreas de estudo.

5 RESULTADOS

Este capítulo apresenta a obtenção de resultados a partir das atividades metodológicas. Para isso, o conjunto de ferramentas foram: Python, bibliotecas *sklearn* para tudo referente a aprendizado de máquina, *matplotlib* para apresentação gráfica e pandas para manipulação de *dataframes*. O ambiente de execução foi o serviço em nuvem do Google Colab[1].

Para a base do SINASC, há os seguintes resultados:

1. É identificado que mães de 31 anos apresentam a maior quantidade de filhos vivos e mortos, o que pode ocasionar em maiores ricos de óbitos infantis;
2. É identificado que as anomalias dos recém-nascidos estão ligados genética, ambiente e a qualidade nutricional;
3. Crianças com ou sem anomalia apresentam uma boa vitalidade ao nascer, mas os casos de risco de morte estão ligadas ao parto prematuro, anomalias congênitas e asfixia perinatal;
4. Dados socioeconômicos, como o IDH, podem ajudar a identificar uma menor taxa de mortalidade infantil para uma certa localidade, mas avaliar a relação ao nível de unidade federativa apresenta diferenciais a serem melhor explorados;
5. O nível de escolaridade da mãe pode acarretar maior ocorrência de óbito infantil;
6. O tipo de parto cesáreo pode estar ligado a maior ocorrência de óbito infantil;

Para a base do SIM, temos os seguintes resultados:

1. 68,6% dos recém-nascidos nascem com peso abaixo de 2,5 kg, enquanto 1,8% nascem com peso acima de 4,0 kg;
2. As doenças mais ligadas aos óbitos infantis são vinculadas a sepse bacteriana, além de problemas relacionados ao parto prematuro e asfixia ao nascer;

Para os algoritmos de aprendizado de máquina, temos os seguintes resultados:

1. Naive Bayes apresentou os piores resultados, apesar de ter sido o modelo com maior valor para a métrica de sensibilidade (*recall*);
2. Random Forest, utilizando o método de reamostragem de subamostragem aleatória, apresentou os resultados mais equilibrados para as métricas de sensibilidade e AUCROC;

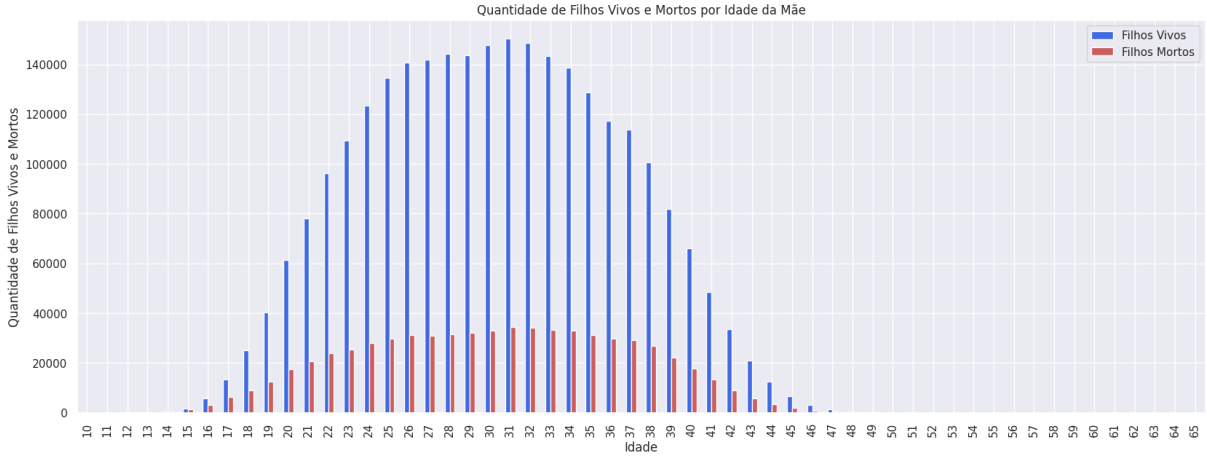
[1] https://colab.google/

5.1 Análises na base de dados SINASC

Na Figura 6, observa-se um gráfico de barras em que no eixo y há a representação da quantidade de filhos vivos — barra azul — e mortos — barra vermelha — para cada idade das mães, estas representadas horizontalmente pelas informações no eixo x. No gráfico, observa-se o comportamento semelhante a uma distribuição gaussiana, em que mães da idade entre 25 a 34 anos apresentaram a maior quantidade de filhos vivos, enquanto aquelas entre 26 a 35 anos contém a maior quantidade de filhos mortos. Nesse gráfico, temos que a média de idade é de 30,237 anos com um desvio padrão de 6,216. Mães de 31 anos apresentam tanto a maior quantidade de filhos vivos quanto a de mortos.

Conforme Lima (2010) em sua pesquisa realizada em 2010, mulheres abaixo de 25 anos e a partir de 35 anos apresentaram maiores ocorrências de óbito infantil, seja por fatores comportamentais, socioeconômicos e biológicos. Em relação aos subitens *2.1.1* (*A idade da mãe pode influenciar na mortalidade?*) e *2.1* (*A idade da mãe pode influenciar na mortalidade*) do item 2 da **Pergunta 1**, entende-se que a idade da mãe pode acarretar uma maior quantidade de filhos vivos e mortos. Para o perfil traçado nesse gráfico em 2020, isso ocorre principalmente por volta dos 31 anos.

Figura 6 – Quantidade de Filhos Vivos e Mortos por Idade da Mãe



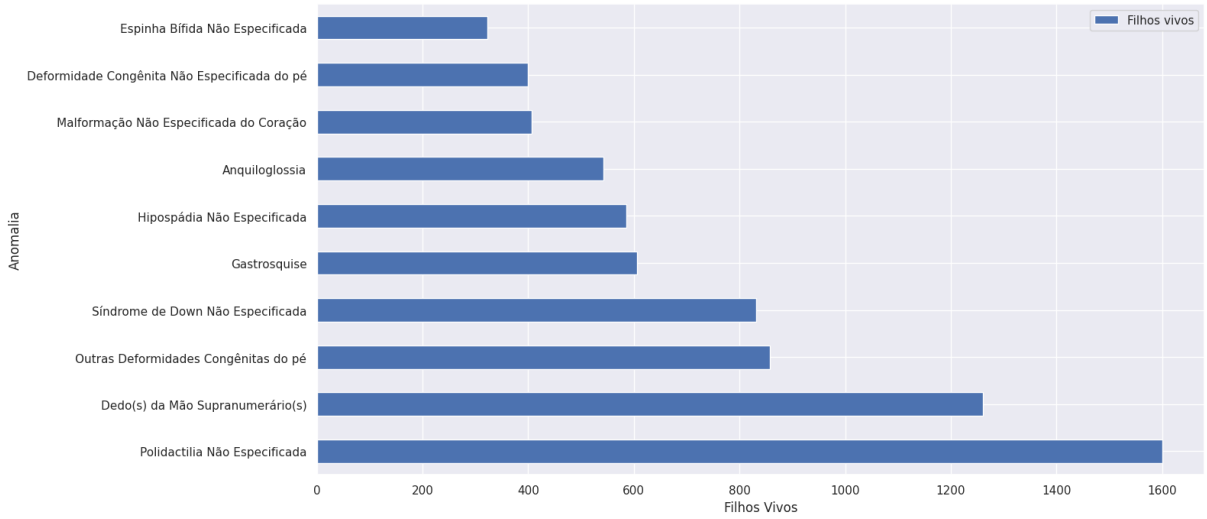
Fonte: Elaborado pelo autor

Para a Figura 7, é apresentado um gráfico de barras que relaciona a quantidade de filhos vivos por anomalias registradas. Neste cenário, a ocorrência mais comum de anomalia é a de Polidactilia Não Especificada e Dedo(s) da Mão Suranumerário(s). Por se tratarem de anomalias congênitas, muito de suas causas desconhecidas, mas é sabido que infecções, a genética e alguns fatores ambientes podem aumentar a sua ocorrência (MSD, 2022a). Além

disso, esse tipo de anomalia também é ligado com as condições nutricionais durante o período de gestação. A hipospádia, por exemplo, a quinta anomalia mais frequente no gráfico, além de fatores genéticos, é discutível que o contato com substâncias no ambiente durante pela mãe durante período de gestação podem afetar a sua ocorrência.

Desse modo, para a **Pergunta 2** — *Quais as principais anomalias estão associadas com mortalidade e quais fatores as influenciam?* —, deduz que as anomalias mais frequentes nos recém-nascidos podem estar relacionada não somente a fatores genéticos, mas também ao ambiente e a qualidade nutricional.

Figura 7 – Filhos Vivos X Anomalias (2020)



Fonte: Elaborado pelo autor

No Quadro 4, há dados relacionando recém-nascidos com e sem anomalia e a média de Apgar para os primeiros 5 minutos de vida. O Apgar é uma escala utilizada para medir a saúde de vida do recém-nascido ao primeiro minuto de vida e aos 5 minutos (OLIVEIRA *et al.*, 2012). Uma pontuação entre 7-10 indica uma boa saúde sem problemas no futuro, enquanto abaixo de 7 indica um alerta para as condições clínicas.

Observa-se que os recém-nascidos com anomalia apresentam um índice de Apgar menor comparado ao outro grupo, com uma média de 8,62.

Ao observar o gráfico na Figura 8, é apresentado a relação entre o peso ao nascer e a pontuação na escala Apgar 5. Entende-se que os recém-nascidos com pontuações menores na escala Apgar 5 estão associados a condições de baixo peso em relação àqueles com pontuações maiores. Uma pontuação de Apgar abaixo de 5 para os primeiros 5 minutos de vida é uma medida que indica um alto risco ao nascer (SAÚDE, 2012). Como fatores para essa baixa pontuação, têm-se fatores como asfixia perinatal, infecções congênitas, presença de corioamnionite e parto

prematuro (parto que acontece antes de 37 semanas de gestação), por exemplo (PERSSON *et al.*, 2018). A presença de asfixia e corioamnionite podem ter suas relações relacionadas a óbitos infantis na Figura 17, que será discutido posteriormente.

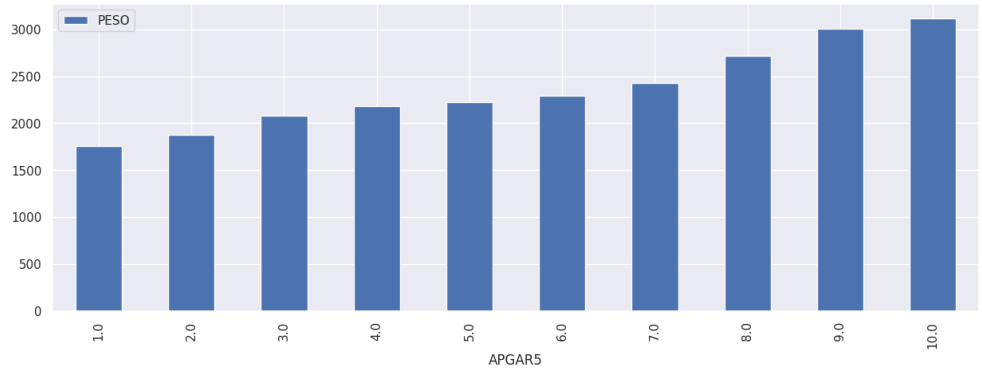
Essas informações nos ajudam a compreender melhor o questionamento da **Pergunta 3** — *Qual a situação clínica das crianças ao nascer pela escala Apgar.*

Quadro 4 – Média da pontuação Apgar para os primeiros 5 minutos de vida para recém-nascidos com e sem anomalia

Presença de anomalia	**Média da pontuação Apgar 5**
Anomalia	8,62
Sem anomalia	9,36

Fonte: Elaborado pelo autor

Figura 8 – Relação Peso e Apgar 5



Fonte: Elaborado pelo autor

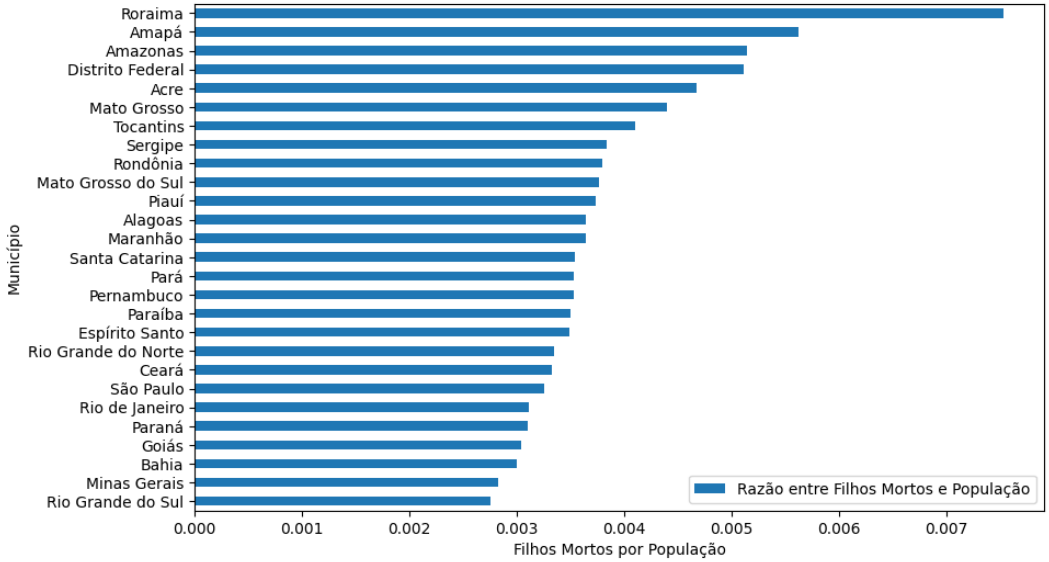
Em relação a Figura 9, é apresentado a razão entre número de filhos mortos por estado. Dos 26 estados, observa-se que o estado de Roraima apresenta a maior quantidade de filhos mortos em relação a sua população, seguido por Amapá e Amazonas. Além disso, observa-se que os estados da região Norte apresentam os maiores valores desta taxa, logo sendo uma das regiões cuja taxa de mortes precisa ser melhor averiguada. Ao observar o gráfico da Figura 10, alguns estados que compõe a região norte com alta taxa filhos mortos, como visto no gráfico anterior, apresentam os menores índices de IDH de renda. Contudo, observa-se que o Distrito Federal, mesmo apresentando alto IDH, também conteve uma taxa alta de filhos mortos por Estado. O Rio Grande do Sul, que contém um valor baixo para essa taxa, apresenta um IDH elevado.

Pode-se afirmar que avaliar a relação de IDH e unidade federativa dificulta uma análise melhor a respeito do impacto desse índice socioeconômico com a mortalidade infantil.

Desse modo, os dados citados induzem aplicar uma investigação em um nível mais específico, como municipal, para que essa relação possa ser melhor estudada.

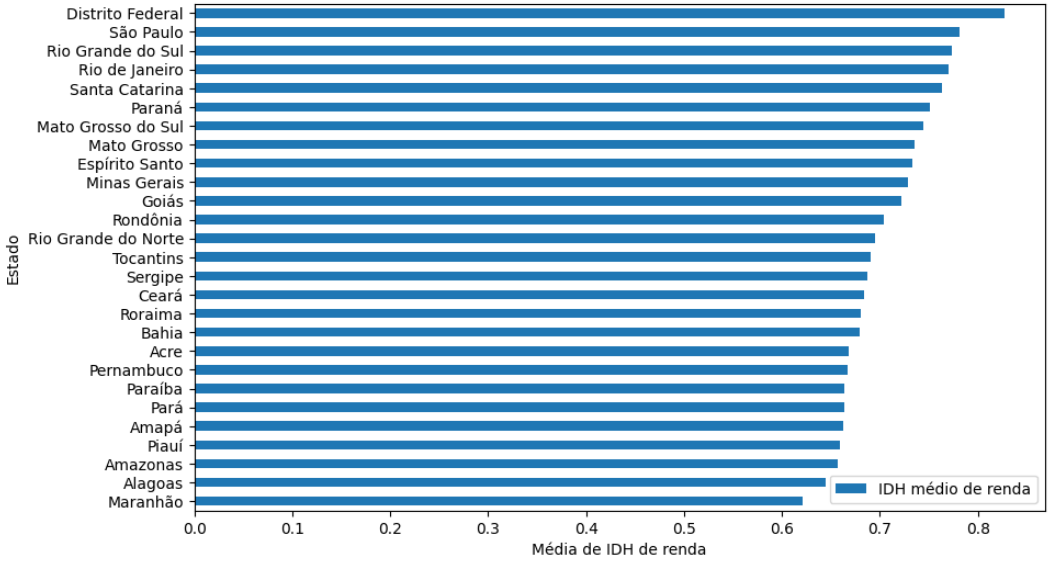
Assim, indicadores socioeconômicos, como o IDH, como mostrados por Martins *et al.* (2018), a medida que aumentam, tendem a reduzir as taxas de mortalidade infantil. Contudo, quando analisados por unidades da federação, eles apresentam diferencias quando analisados. Isso nos ajuda a compreender a resposta para o item *2.2* do item 2 da **Pergunta 1** — *Índices de qualidade de vida de uma região influencia na mortalidade infantil?*.

Figura 9 – Taxa de Filhos Mortos por Estado



Fonte: Elaborado pelo autor

Figura 10 – Nível de IDH médio de renda por estado

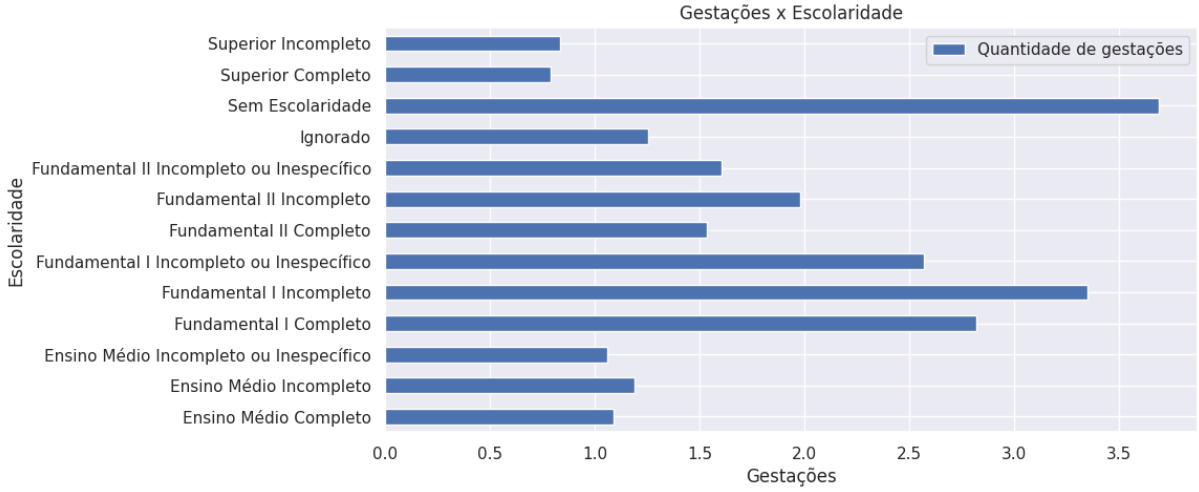


Fonte: Elaborado pelo autor

Nos gráficos contidos na Figura 11, Figura 12 e Figura 13, há a relação entre escolaridade e média de gestações, filhos vivos e mortos, respectivamente. É possível observar que as mães com o menor grau de escolaridade apresentam a maior média do número de gestações, consequentemente afetando na quantidade de crianças vivas e mortas. Podemos afirmar, dessarte, que as mulheres com menor nível de escolaridade concentram tanto a maior taxa de nascidos vivos quanto a de mortes infantis.

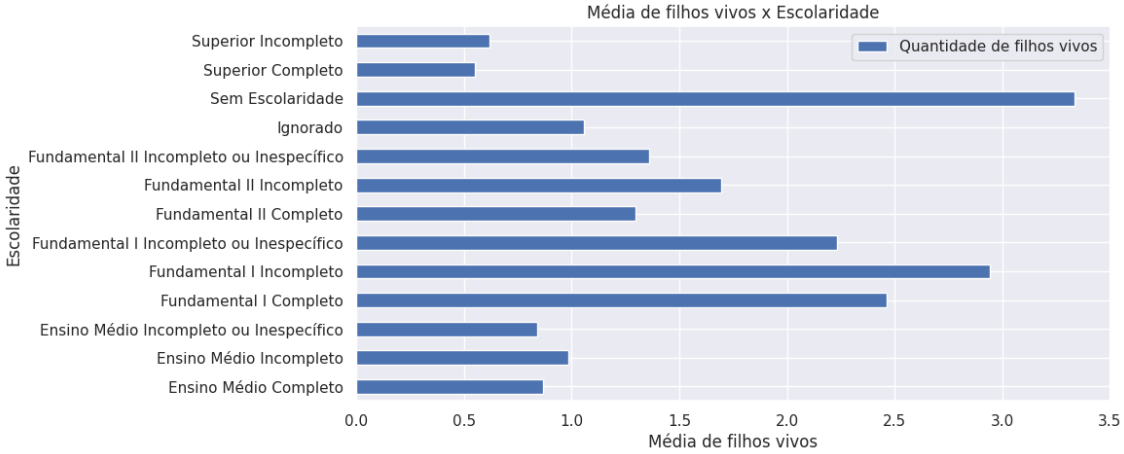
Quanto menor a escolaridade, maiores os riscos para a ocorrência de um óbito infantil. Conforme o Saúde (2012), mães com baixa escolaridade apresentam menor acesso aos serviços de saúde de melhor qualidade, provocando em condições inadequadas durante o período de gestação e após o parto.

Figura 11 – Número de Gestações x Escolaridade



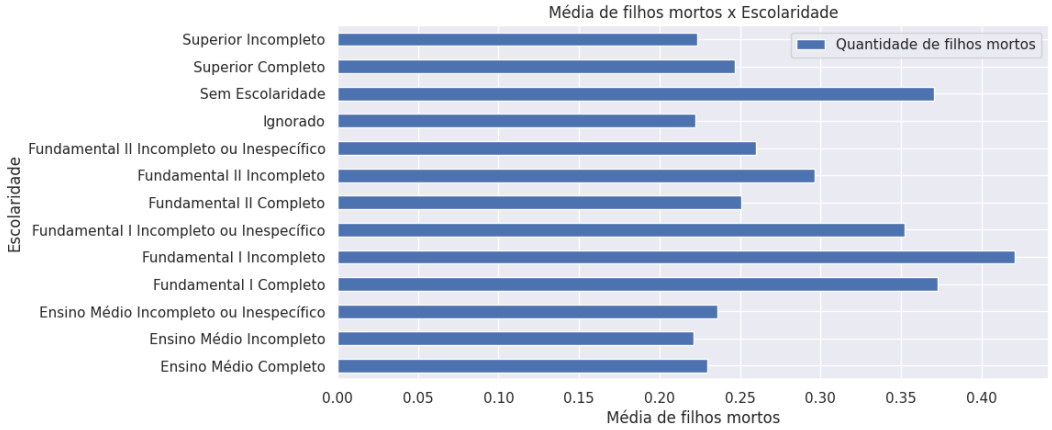
Fonte: Elaborado pelo autor

Figura 12 – Média do Número de Filhos Vivos x Escolaridade



Fonte: Elaborado pelo autor

Figura 13 – Média do Número de Filhos Mortos x Escolaridade

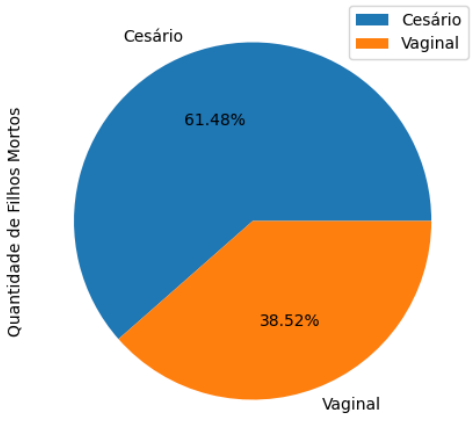


Fonte: Elaborado pelo autor

Em relação ao tipo de parto e a quantidade de filhos mortos, é possível na Figura 14 observar que a maioria das mortes está relacionada ao parto do tipo cesário, cobrindo cerca de 61,48% dos óbitos. Enquanto isso, o parto do tipo vaginal está relacionado com 38,52% dos óbitos.

Segundo Rocha *et al.* (2023), houve um crescimento da realização de partos cesáreos nas últimas décadas, sendo o Brasil um dos países que mais adota essa prática. A sua pesquisa mostra que os partos cesáreos sem indicação estão associados ao aumento da mortalidade infantil, enquanto estes mesmos partos sob recomendação apresentam uma redução dos índices de óbitos. Neste sentido, nota-se que o tipo de parto pode indicar uma possível maior probabilidade de ocorrência de óbito infantil, o que a torna uma variável importante para o processo de predição. Isso apoia a resposta ao item 1 da **Pergunta 3** — *O tipo de parto pode afetar na morte da criança?*.

Figura 14 – Quantidade de filhos mortos em relação ao parto



Fonte: Elaborado pelo autor

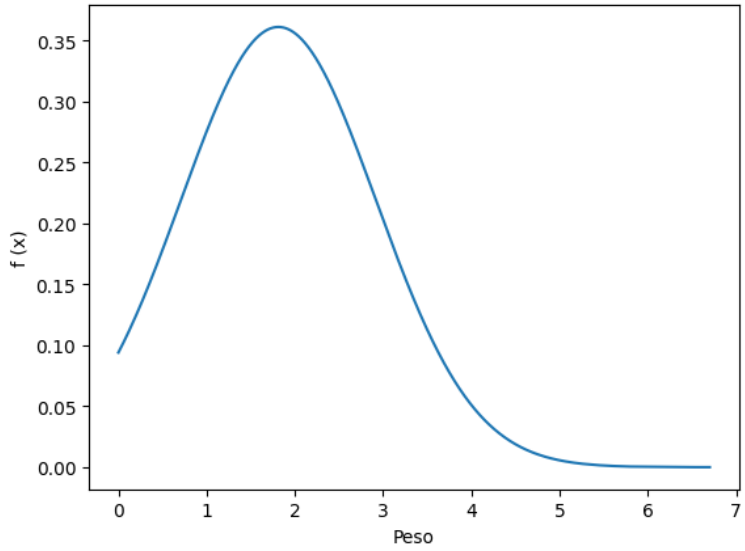
5.2 Análises na base de dados SIM

O parâmetro Peso ao Nascer é utilizado para avaliar as condições de saúde do recém-nascido. Caso o peso ao nascer esteja abaixo 2.500 g (< 2.500 g), a condição é associada a maior mortalidade e morbidade neonatal e infantil. Por outro lado, caso o peso esteja acima de 4.000 g (> 4.000 g), sendo uma macrossomia fetal, a condição de saúde é relacionada às situações como asfixia neonatal, parto prematuro, aspiração mecônico, entre outros. Na vida adulta, isso pode levar a doenças crônicas não transmissíveis (TOURINHO; REIS, 2012).

Com o gráfico da Figura 15, é possível observar o comportamento da distribuição dos dados que aparenta ser uma distribuição normal. Contudo, é necessário realizar um teste de hipótese para validar se os dados são normalmente distribuídos. Assim, define-se a hipótese nula (H_0) e a hipótese alternativa (H_1):

- H_0: os dados são normalmente distribuídos
- H_1: os dados não são normalmente distribuídos

Figura 15 – Distribuição de Peso ao Nascer



Fonte: Elaborado pelo autor

Para este teste, utiliza-se um nível de significância α de 5%, ou seja, $\alpha = 0,05$. Isso significa que a probabilidade de rejeição da hipótese nula quando ela é verdadeira é de 5%.

Aplicando um teste K-S (ou teste Kolmogorov-Smirnov) fornecido pela biblioteca *scipy*, temos que para esse conjunto de dados o valor do *p-value* corresponde a 0,0 com uma distância D entre as distribuições de 0,63, aproximadamente. Como o valor do *p-value* é menor do que 0,05, logo rejeitamos a hipótese nula. Os dados não são normalmente distribuídos.

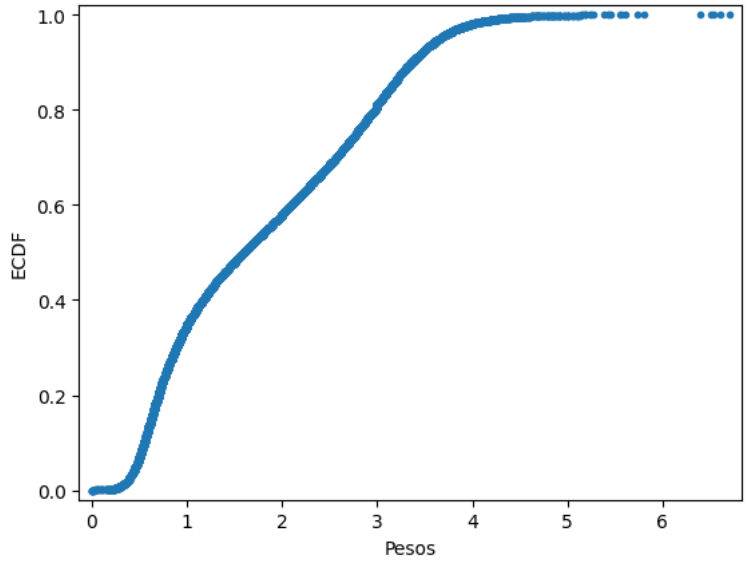
Para melhor representá-los, utiliza-se uma distribuição empírica, como apresentada na Figura 16, pois assim entende-se melhor como as probabilidades dos dados estão distribuídas dentro de uma faixa. A partir dessa distribuição, podem ser calculadas as probabilidades para os recém-nascidos com o peso abaixo de 2,5 kg e acima de 4 kg. Portanto, desejamos encontrar $P(x \leq 2,5)$ e $P(x > 4,0)$.

Com o apoio da biblioteca *statsmodels*, conseguimos calcular que:

$$P(x < 2,5) = 0,686 \quad \text{e} \quad P(x > 4,0) = 0.018 \tag{5.1}$$

Neste sentido, concluímos que 68,6% dos dados marcam recém-nascidos com um peso ao nascer abaixo de 2,5 kg, enquanto 1,8% dos dados marcam recém-nascidos com peso acima de 4,0 kg. Pode-se deduzir que o peso dos recém-nascidos no ano de 2020 está ligado a uma maior ocorrência de mortalidade infantil, conforme explicado por Tourinho e Reis (2012).

Figura 16 – Distribuição Acumulativa Empírica de Peso ao Nascer



Fonte: Elaborado pelo autor

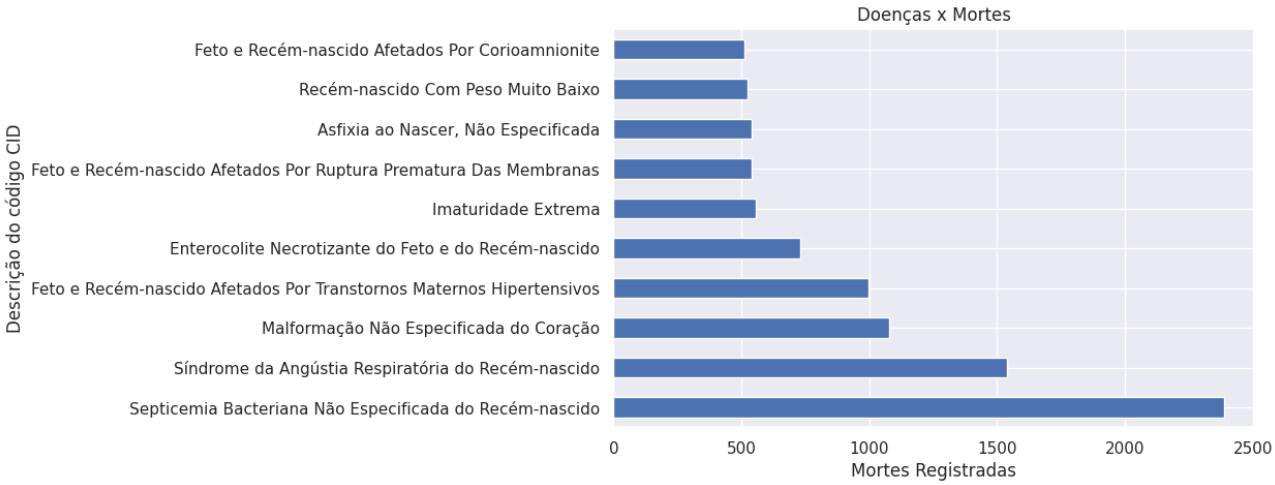
Na Figura 17 é possível analisar a relação entre as mortes registradas e as doenças causa-base das mortes. Seguindo as informações apresentadas por Tourinho e Reis (2012) e disponibilizadas em Manual MSD, pode-se interpretar o gráfico com maior facilidade, assim correlacionando os dados capturados caso seja apropriado e lógico.

No gráfico observa-se que, entre as 10 causas listadas, Septicemia Bacteriana Não Especificada apresenta-se como maior causa das mortes em recém-nascidos, seguida de Síndrome da Angústia Respiratória e Malformação Não Especificada do Coração.

O dado Recém-nascido Com Peso Muito Baixo surge logicamente como causa de mortes, uma vez que o baixo peso é uma condição de saúde associado com a maior propensão de ocorrer sepse — seja ela bacteriana, viral ou fúngica —, casando com a ocorrência da septicemia bacteriana (MSD, 2022b). Além disso, o baixo peso ainda é associado com a imaturidade extrema, também presente no gráfico.

Como apresentado por Tourinho e Reis (2012), recém-nascidos com o peso ao nascer superior a 4.000 g estão mais associados a asfixia neonatal, ruptura prematura das membranas (RPM) e parto prematuro, por exemplo. Tais problemáticas são mais explícitas no gráfico a partir de Asfixia ao Nascer, Não Especificada; e Feto e Recém-nascido Afetados Por Ruptura Prematura das Membranas. Ainda é associável com parto prematuro o dado de Enterocolite Necrotizante do Feto e do Recém-nascido, haja vista que indivíduos prematuros aumentam a possibilidade de ocorrência dessa problemática. Já em relação aos dados de Feto e Recém-nascido Afetados Por Corioamnionite, associamos a sua ocorrência com RPM e parto prematuro por serem seus fatores de risco.

Figura 17 – Doenças x Mortes



Fonte: Elaborado pelo autor

5.3 Identificação de variáveis

Seguindo as análises realizadas nas seções 5.1 e 5.2, para o item 1 da **Pergunta 1** — *Como analisar e descobrir as principais variáveis que influenciam a mortalidade infantil?* —, foram identificadas as seguintes variáveis relacionadas com mortalidade infantil: “Apgar 5”, “Semanas de gestação”, “idh”, “Nível de escolaridade da mãe”, “Idade da mãe”, “Tipo de parto”, “Peso ao nascer” e “Anomalias”.

O “Nível de escolaridade da mãe”, “Idade da mãe” estão relacionados com condições socioeconômicas que permitem uma exploração de seus fatores que acarretam ocorrência de óbito infantil. A variável “Tipo de parto”, “Peso ao nascer” e “Anomalias” estão ligadas principalmente a questões de saúde pública, esta que por sua vez que pode afetar mais diretamente na saúde da criança.

5.4 Ligação entre as bases de dados

Como informado na Seção 4.1,o conjunto de dados ligados é utilizado para o aprendizado de máquina. O retorno das instâncias presentes tanto no SINASC quanto do SIM foi de aproximadamente 43% para o conjunto de atributos selecionados para a ligação. Em razão disso, foi feito um curto experimento com as bases de 2019, 2018 e 2017 para identificar se o mesmo conjunto de atributos conseguiria retornar uma quantidade semelhante de informações. A ligação é feita com os dados filtrados para o estado do Ceará utilizando a mesma combinação de chaves. O resultado é encontrado no Quadro 5.

É possível observar uma pequena variação entre os anos de 2020, 2019 e 2018, mas 2017 foi o ano em que nenhuma instância conseguiu ser ligada, isto é, não foram encontrados registros em comum nas duas bases. Por não pertencer ao foco deste trabalho, não foram encontradas os motivos e soluções para esse comportamento.

Quadro 5 – Porcentagem de instâncias ligadas entre as bases SINASC e SIM

Ano das bases	Porcentagem de ligação de instâncias
2020	43,00%
2019	45,13%
2018	42,07%
2017	0,00%

Fonte: Elaborado pelo autor

5.5 Aprendizado de Máquina

O objetivo dos classificadores é de predizer o óbito infantil a partir das informações de nascimento unido a dados socioeconômicos e estatísticos. Dessa maneira, as métricas utilizadas para este trabalho são: acurácia, precisão, sensibilidade (*recall*) e AUCROC. Em relação ao tema, como há uma preferência em garantir a taxa de verdeiros positivos e reduzir falsos negativos, a

sensibilidade, o f1-score e o AUCROC serão mais importantes para a análise.

Como base para todos os classificadores, o modelo *Dummy Classifier* é executado utilizando a estratégia da classe mais frequente para gerar as predições. Assim, este modelo serve de base para entender uma tendência dos classificadores sem haver uma reamostagem nos dados desbalanceados.

Entre os algoritmos de seleção de atributos, aquele baseado no meta-classificador *SelectFromModel* fornecido pela biblioteca *scikit-learn* com o estimador regressão logística selecionou o melhor conjunto de informações para realizar a predição. Os atributos selecionados e suas descrições são apresentadas no Quadro 6. Os resultados dos classificadores podem ser observados no Quadro 7.

Quadro 6 – Atributos escolhidos pelo algoritmo de seleção de atributos

n.º	**atributo**	**descrição**
1	APGAR1	Valor do Apgar para o primeiro minuto de vida
2	APGAR5	Valor do Apgar para os 5 primeiros minutos de vida
3	CODESTAB	Código do estabelecimento de saúde onde ocorreu o nascimento
4	DTDECLARAC	Data da declaração
5	CODOCUPMAE	Código de ocupação da mãe conforme tabela do CBO (Código Brasileiro de Ocupações)
6	ESCMAE	Nível de escolaridade da mãe em anos
7	GESTACAO	Semanas de gestação
8	IDADEMAE	Idade da mãe
9	PESO	Peso da criança ao nascer
10	QTDPARTCES	Quantidade de partos cesários
11	IDANOMAL	Identificação de anomalia
12	RACACOR	Raça/etnia da criança
13	SEXO	Sexo da criança
14	SEMAGESTAC	Número de semanas de gestação
15	STCESPARTO	Identificação de cesárea antes do início do trabalho de parto
16	PARTO	Tipo de parto
17	QTDFILMORT	Quantidade de filhos mortos
18	QTDFILVIVO	Quantidade de filhos vivos
19	IDHM	Média do IDH geral de um município
20	IDHMRENDA	Média do IDH de renda de um município

Fonte: Elaborado pelo autor

Quadro 7 – Resultados dos classificadores

Classificador	Método de Reamostragem	Acurácia	Precisão	Sensibilidade	F1-Sore	AUCROC
Dummy Classifier		100,00%	0,00%	0,00%	0,00%	5,00%
Logistic Regression	SMOTE	88,70%	2,90%	68,00%	5,50%	86,20%
Logistic Regression	Random Undersampling	87,90%	2,60%	67,00%	5,10%	85,50%
Decision Tree	SMOTE	89,80%	2,90%	61,20%	5,50%	78,90%
Decision Tree	Random Undersampling	73,80%	1,30%	71,80%	2,60%	75,50%
Random Forest	SMOTE	97,00%	9,00%	56,30%	15,50%	86,20%
Random Forest	Random Undersampling	87,10%	2,70%	72,80%	5,10%	86,30%
Naive-Bayes	SMOTE	35,10%	0,60%	84,50%	1,20%	66,30%
Naive-Bayes	Random Undersampling	36,00%	0,60%	84,50%	1,30%	60,70%
XGBoost	SMOTE	99,00%	20,70%	38,80%	27,00%	86,20%
XGBoost	Random Undersampling	82,70%	2,00%	70,90%	3,80%	85,20%

Fonte: Elaborado pelo autor

Para o método SMOTE, o algoritmo que apresenta as melhores métrica é o Random Forest com 9,00% de precisão, 56,30% de sensibilidade e 86,20% de AUCROC. Porém, ressalta que, em questão de sensibilidade, o algoritmo de Regressão Logística apresentou o maior valor (68,00%). Já para a técnica de subamostragem aleatória, o classificador Random Forest também apresenta os melhores resultados gerais, com 2,70% de precisão, 72,80% de sensibilidade e 86,30% de AUCROC.

Para o modelo de Regressão Logística, as abordagens de reamostragem apresentaram resultados semelhantes, com uma diferença de 0,3% para a métrica de precisão, 1% para a sensibilidade e 0,7% para o AUCROC.

O classificador de Árvore de Decisão apresenta os melhores resultados para a métrica de sensibilidade quando utilizado a técnica de subamostragem aleatória, enquanto as taxas de precisão mais altas são encontradas para técnica SMOTE. Contudo, a métrica AUCROC informa que sua predição para as classes positivas e negativas é inferior aos outros classificadores, com exceção do Naive Bayes.

O classificador XGBoost apresenta resultados mais equilibrados para as métricas de precisão de sensibilidade em comparação aos outros classificadores. Para a técnica de subamostragem aleatória, obteve a melhor taxa de sensibilidade (70,90%), reduzindo a quantidade de falsos negativos.

Entre os modelos, o Naive Bayes foi o que apresentou os piores resultados. Mesmo após a aplicação das técnicas de reamostragem, este modelo apresentou um desempenho inferior ao classificador de Árvore de Decisão. Apesar da sensibilidade ter sido superior aos restantes (84,50%), modelo apresentou as taxas de precisão mais baixas (0,60%), o que indica haver um aumento de falsos positivos e uma redução dos falsos negativos. Além disso, os valores obtidos pela AUCROC indica que o modelo apresenta uma maior dificuldade em identificar os valores positivos e negativos entre as classes. A abordagem com este modelo

É notório que os algoritmos que apresentam uma melhor taxa de precisão consequentemente contém uma baixa taxa de sensibilidade, o inverso também é observado nos classificadores treinados.

Entre as técnicas de reamostagem aplicadas, verifica-se que a técnica subamostragem aleatória apresenta uma melhora para a métrica de sensibilidade, enquanto a técnica SMOTE aumentou levemente a taxa de precisão para alguns dos classificadores. É importante ressaltar que a técnica de subamostragem aleatória, por basear-se na remoção de instâncias da classe majoritária, causa uma perda significativa de informações, afetando diretamente em seu desempenho.

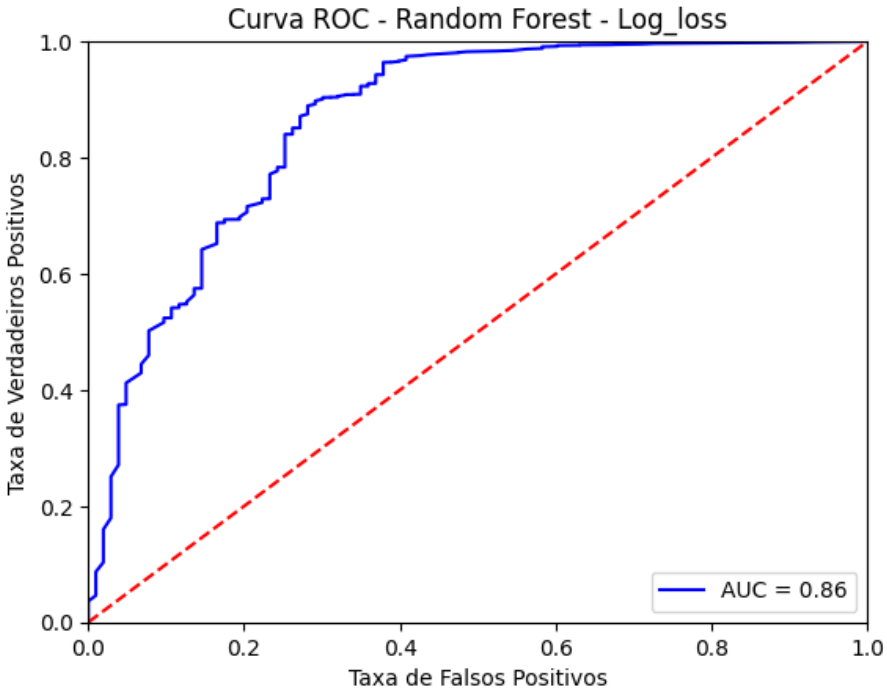
A partir do experimento, observa-se que o classificador com as melhores métricas para predizer óbito infantil é o Random Forest através da técnica de subamostragem aleatória, com uma taxa de sensibilidade de 72,80% e um AUCROC de 86,30%. Contudo, é importante ressaltar que a aplicação da técnica SMOTE permitiu um aumento de sua precisão. É possível observar o comportamento da curva ROC desse modelo através da Figura 18.

A linha em vermelho representa uma bissetriz. Caso a curva se converge para uma bissetriz, representando uma área igual a 0,5, isso indica que o modelo não consegue distinguir bem uma classe da outra. No caso do Random Forest, o valor da área se aproxima mais de 1, uma vez que o modelo apresenta uma boa distinção entre as classes.

Os resultados do seguinte trabalho são afetados pela dificuldade em ligar as bases de dados a partir de um atributo identificador. Dessa forma, uma quantidade de amostras é significativamente perdida a partir da utilização de uma combinação para realizar esta operação. Com esta perda de informação, uma a baixa variabilidade de amostras para a classe positiva (óbito infantil) afeta no desbalanceamento e, consequentemente, no desempenho dos modelos.

Além disso, por utilizar informações apenas do ano de 2020, os exemplos das amostras contemplam uma variabilidade menor caso fossem utilizadas informações de anos anteriores em conjunto.

Figura 18 – Curva ROC para o classificador Random Forest utilizando *random undersampling* como técnica de reamostragem



Fonte: Elaborado pelo autor

6 CONCLUSÕES

Este trabalho apresentou métodos para treinamento e modelagem de algoritmos de aprendizado de máquina em prol da predição de óbitos infantis. Além disso, foi explorada a relação entre certos dados socioeconômicos e estatísticos (IDH).

A partir dos conjuntos de dados coletados no DATASUS, um novo conjunto de dados foi criado a partir da ligação entre elas com dados de óbito infantil do estado do Ceará. A partir disso, é possível identificar a dificuldade em ligar as bases, uma vez que não há um atributo em comum na estrutura atual que facilite essa operação.

O presente trabalho ainda apresentou a utilização de algoritmos de seleção de atributos para auxiliar na identificação de informações que melhor contribuem para o treinamento dos modelos de aprendizado de máquina. A partir disso, é apresentada a dificuldade em tratar de problemas de classificação com dados desbalanceados. Para isso, são utilizadas duas técnicas de reamostragem: SMOTE e subamostragem aleatória. Entre elas, a subamostragem aleatória apresenta uma maneira melhor para aprimorar a métrica de sensibilidade, permitindo uma melhor identificação de verdadeiros positivos. Já a técnica SMOTE, apesar do aumento da precisão, apresenta uma menor taxa de identificação de verdadeiros positivos.

A contribuição deste trabalho pode ser utilizada para encontrar maneiras mais eficazes de realizar a ligação entre as bases do SIM e do SINASC e maneiras mais eficazes de pré-processamento e técnicas de reamostragem de forma que os desempenhos dos modelos sejam aprimorados. Explorar este problema para estado ou regiões mais específicas também colaboraria para entender melhor o comportamento não só da problemática da mortalidade infantil, mas também dos modelos de aprendizado de máquina.

Uma contribuição para um trabalho futuro está relacionado a atividade de ligação entre os conjuntos de dados, uma vez que a falta de um atributo comum prejudica a captura de registros importantes para a pesquisa. Desse modo, um possível trabalho futuro poderia investigar a razão desse comportamento e sugerir soluções.

REFERÊNCIAS

AGARWAL, S. Data mining: Data mining concepts and techniques. In: **2013 International Conference on Machine Intelligence and Research Advancement**. Katra, India: IEEE, 2013. p. 203–207.

ALGARNI, A. Data mining in education. **International Journal of Advanced Computer Science and Applications**, Science and Information (SAI) Organization Limited, [S.l], v. 7, n. 6, 2016.

BOATENG, E. Y.; ABAYE, D. A. A review of the logistic regression model with emphasis on medical research. **Journal of data analysis and information processing**, Scientific Research Publishing, [S.l], v. 7, n. 4, p. 190–207, 2019.

BRASIL. **Lei Geral de Proteção de Dados Pessoais (LGPD)**. 2019. Disponível em https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 05 dez. 2023.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, p. 123–140, 1996. Disponível em: https://doi.org/10.1007/BF00058655. Acesso em: 12 abr. 2023.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: 12 abr. 2023.

CAO, L. Data science: A comprehensive overview. **ACM Journals**, New York, NY, USA, v. 50, n. 43, p. 1–42, jul 2017.

CDC. **Infant mortality**. 2023. Disponível em: https://www.cdc.gov/reproductivehealth/maternalinfanthealth/infantmortality.htm. Acesso em: 01 abr. 2023.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [*S. l.*: *s. n.*], 2016. p. 785–794.

CUNNINGHAM, P.; CORD, M.; DELANY, S. J. **Supervised Learning**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. 21–49 p. ISBN 978-3-540-75171-7. Disponível em: https://doi.org/10.1007/978-3-540-75171-7_2. Acesso em: 05 dez. 2023.

DALIANIS, H. **Evaluation Metrics and Evaluation**. Cham: Springer International Publishing, 2018. 45-53 p. ISBN 978-3-319-78503-5. Disponível em: https://doi.org/10.1007/978-3-319-78503-5_6. Acesso em: 11 abr. 2023.

DONG, G.; LIU, H. **Feature engineering for machine learning and data analytics**. NW Boca Raton: CRC Press, 2018.

FAN, W.; ZHONG, E.; PENG, J.; VERSCHEURE, O.; ZHANG, K.; REN, J.; YAN, R.; YANG, Q. Generalized and heuristic-free feature construction for improved accuracy. **Proceedings of the 2010 SIAM International Conference on Data Mining (SDM)**, [S.l], p. 629–640, 2010.

FERNÁNDEZ, A.; GARCIA, S.; HERRERA, F.; CHAWLA, N. V. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. **Journal of artificial intelligence research**, [S.l], v. 61, p. 863–905, 2018.

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques, third edition**. Waltham, Mass.: Morgan Kaufmann Publishers, 2012. ISBN 0123814790.

HARRINGTON, P. **Machine learning in action**. 3 Lewis Street Greenwich, CT, United States: Manning Publications Co., 2012.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255–260, 2015. Disponível em: https://www.science.org/doi/abs/10.1126/science.aaa8415. Acesso em: 11 abr. 2023.

KOTSIANTIS, S. B. Decision trees: a recent overview. **Artificial Intelligence Review**, Springer, v. 39, p. 261–283, 4 2013. Disponível em: https://doi.org/10.1007/s10462-011-9272-4. Acesso em: 12 abr. 2023.

KRAWCZYK, B. Learning from imbalanced data: open challenges and future directions. **Progress in Artificial Intelligence**, Springer, v. 5, n. 4, p. 221–232, 2016. Disponível em: https://doi.org/10.1007/s13748-016-0094-0. Acesso em: 4 nov. 2023.

KREUTZ, I. M.; SANTOS, I. S. Contextual, maternal, and infant factors in preventable infant deaths: a statewide ecological and cross-sectional study in rio grande do sul, brazil. **BMC Public Health**, [S.l], v. 23, n. 87, jan 2022.

LIMA, L. C. d. Idade materna e mortalidade infantil: efeitos nulos, biológicos ou socioeconômicos? **Revista Brasileira de Estudos de População**, SciELO Brasil, v. 27, p. 211–226, 2010. Disponível em: https://doi.org/10.1590/S0102-30982010000100012. Acesso em: 1 dez. 2023.

LIU, B.; DING, M.; SHAHAM, S.; RAHAYU, W.; FAROKHI, F.; LIN, Z. When machine learning meets privacy: A survey and outlook. **ACM Comput. Surv.**, Association for Computing Machinery, New York, v. 54, n. 2, mar 2021. ISSN 0360-0300. Disponível em: https://doi.org/10.1145/3436755. Acesso em: 11 abr. 2023.

LIU, Y.; WANG, Y.; ZHANG, J. New machine learning algorithm: Random forest. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

MALTA, D. C.; DUARTE, E. C.; ALMEIDA, M. F. d.; DIAS, M. A. d. S.; NETO, O. L. d. M.; MOURA, L. d.; FERRAZ, W.; SOUZA, M. d. F. M. d. Lista de causas de mortes evitáveis por intervenções do sistema único de saúde do brasil. **Epidemiol. Serv. Saúde**, Brasília, v. 16, n. 4, p. 233–244, dez 2007. Disponível em: http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S1679-49742007000400002&lng=pt&nrm=iso. Acesso em: 20 jun. 2023.

MARTINS, P. C. R.; PONTES, E. R. J. C.; HIGA, L. T. Convergência entre as taxas de mortalidade infantil e os índices de desenvolvimento humano no brasil no período de 2000 a 2010. **Interações (Campo Grande)**, SciELO Brasil, v. 19, p. 291–303, 2018. Disponível em: https://doi.org/10.20435/inter.v19i2.1552. Acesso em: 4 nov. 2023.

MEDEIROS, M. M. de; HOPPEN, N.; MAÇADA, A. C. G. Data science for business: benefits, challenges and opportunities. **The Bottom Line**, Bradford, v. 33, n. 2, p. 149–163, mar 2020.

MIKUT, R.; REISCHL, M. Data mining tools. **WIREs Data Mining and Knowledge Discovery**, v. 1, n. 5, p. 431–443, 2011. Disponível em: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.24. Acesso em: 12 abr. 2023.

MISHRA, S. Handling imbalanced data: Smote vs. random undersampling. **Int. Res. J. Eng. Technol**, [S.l], v. 4, n. 8, p. 317–320, 2017.

MISILMANI, H. M. E.; NAOUS, T. Machine learning in antenna design: An overview on machine learning concept and algorithms. In: **2019 International Conference on High Performance Computing Simulation (HPCS)**. Dublin, Ireland: IEEE, 2019. p. 600–607.

MSD, M. **Considerações gerais sobre defeitos congênitos**. 2022. Disponível em: https://www.msdmanuals.com/pt-br/casa/problemas-de-sa%C3%BAde-infantil/considera%C3%A7%C3%B5es-gerais-sobre-defeitos-cong%C3%AAnitos/considera%C3%A7%C3%B5es-gerais-sobre-defeitos-cong%C3%AAnitos. Acesso em: 05 dez. 2023.

MSD, M. **Sepse no recém-nascido**. 2022. Disponível em: https://www.msdmanuals.com/pt-br/casa/problemas-de-sa%C3%BAde-infantil/infec%C3%A7%C3%B5es-em-rec%C3%A9m-nascidos/sepse-no-rec%C3%A9m-nascido. Acesso em: 05 dez. 2023.

NARGESIAN, F.; SAMULOWITZ, H.; KHURANA, U.; KHALIL, E.; TURAGA, S. D. In: **Learning Feature Engineering for Classification**. Toronto, Canada: IJCAI, 2017. p. 2529–2535.

OECD. **Infant mortality rates**. 2023. Disponível em: https://data.oecd.org/healthstat/infant-mortality-rates.htm. Acesso em: 01 abr. 2023.

OLIVEIRA, T. G. d.; FREIRE, P. V.; MOREIRA, F. T.; MORAES, J. d. S. B. d.; ARRELARO, R. C.; ROSSI, S.; RICARDI, V. A.; JULIANO, Y.; NOVO, N. F.; BERTAGNON, J. R. D. Escore de apgar e mortalidade neonatal em um hospital localizado na zona sul do município de são paulo. **Einstein (São Paulo)**, SciELO Brasil, São Paulo, v. 10, p. 22–28, 2012. Disponível em: https://doi.org/10.1590/S1679-45082012000100006. Acesso em: 4 nov. 2023.

OSMAN, A. S. Data mining techniques. **International Journal of Data Science Research**, [S.l], v. 2, jul 2019.

PERSSON, M.; RAZAZ, N.; TEDROFF, K.; JOSEPH, K.; CNATTINGIUS, S. Five and 10 minute apgar scores and risks of cerebral palsy and epilepsy: population based cohort study in sweden. **Bmj**, British Medical Journal Publishing Group, v. 360, 2018.

RAMOS, R.; SILVA, C.; MOREIRA, M. W. L.; RODRIGUES, J. J. P. C.; OLIVEIRA, M.; MONTEIRO, O. Using predictive classifiers to prevent infant mortality in the brazilian northeast. In: **2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)**. Dalian, China: IEEE, 2017. p. 1–6.

RASCHKA, S.; MIRJALILI, V. **Python Machine Learning**: Machine learning and deep learning with python, scikit-learn, and tensorflow 2. 3. ed. Birmingham, UK: Packt Publishing, 2019.

RISH, I. *et al.* An empirical study of the naive bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. [*S. l.*: *s. n.*], 2001. v. 3, n. 22, p. 41–46.

ROCHA, A. S.; PAIXAO, E. S.; ALVES, F. J. O.; FALCÃO, I. R.; SILVA, N. J.; TEIXEIRA, C. S.; ORTELAN, N.; FIACCONE, R. L.; RODRIGUES, L. C.; ICHIHARA, M. Y. *et al.* Cesarean sections and early-term births according to robson classification: a population-based study with more than 17 million births in brazil. **BMC Pregnancy and Childbirth**, Springer, [S.l], v. 23, n. 1, p. 562, 2023.

ROKACH, L.; MAIMON, O. **Supervised Learning**. Boston, MA: Springer US, 2010. 133–147 p. ISBN 978-0-387-09823-4. Disponível em: https://doi.org/10.1007/978-0-387-09823-4_8. Acesso em: 11 abr. 2023.

SANCHEZ-PINTO, L. N.; LUO, Y.; CHURPEK, M. M. Big data and data science in critical care. **Chest**, v. 154, n. 5, p. 1239–1248, mai 2018.

SAÚDE, S. da Vigilância em. **Mortalidade Infantil no Brasil**. Brasília, 2021. v. 52, n. 37.

SAÚDE, S. de A. **Atenção à Saúde do Recém-Nascido**. Brasília, 2012. v. 1, n. 2.

SILVA, C.; ALVES, J.; BRAGA, O.; JúNIOR, J.; ANDRADE, L.; OLIVEIRA, A. Usando o classificador naive bayes para geração de alertas de risco de Óbito infantil. **Revista Eletrônica de Sistemas de Informação**, [S.l], v. 16, 08 2017.

SINGHA, A. K.; PHUKAN, D.; BHASIN, S.; SANTHANAM, R. Application of machine learning in analysis of infant mortality and its factors. **Work Pap**, [S.l], p. 1–5, fev 2016.

SONG, Y.-Y.; YING, L. Decision tree methods: applications for classification and prediction. **Shanghai archives of psychiatry**, Shanghai Mental Health Center, v. 27, n. 2, p. 130, 2015.

TOURINHO, A. B.; REIS, M. L. B. D. S. Peso ao nascer: uma abordagem nutricional. **Revista da Faculdade de Ciências Médicas de Sorocaba**, [S.l], p. 19–30, 2012. Disponível em: https://revistas.pucsp.br/index.php/RFCMS/article/view/35830. Acesso em: 5 nov. 2023.

VALTER, R.; SANTIAGO, S.; RAMOS, R.; OLIVEIRA, M.; ANDRADE, L. O. M.; BARRETO, I. C. d. H. C. Data mining and risk analysis supporting decision in brazilian public health systems. In: **2019 IEEE International Conference on E-health Networking, Application Services (HealthCom)**. Bogota, Colombia: IEEE, 2019. p. 1–6.

WARDHANI, N. W. S.; ROCHAYANI, M. Y.; IRIANY, A.; SULISTYONO, A. D.; LESTANTYO, P. Cross-validation metrics for evaluating classification performance on imbalanced data. In: **2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)**. Tangerang, Indonesia: IEEE, 2019. p. 14–18.

WEIHS, C.; ICKSTADT, K. Data science: the impact of statistics. **International Journal of Data Science and Analytics**, [S.l], v. 6, p. 189–194, fev 2018.

WHO. **Child mortality and causes of death**. 2023. Disponível em: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/child-mortality-and-causes-of-death. Acesso em: 03 abr. 2023.

ZHENG, A.; CASARI, A. **Feature engineering for machine learning**: principles and techniques for data scientists. Santa Rosa: O'Reilly Media, Inc., 2018.