Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Linkage Probabilístico e Análise Computacional de Fatores Associados à Mortalidade Infantil na Região Sudeste do Brasil

Raquel Larrat Lopes

JUIZ DE FORA
JULHO, 2026

Linkage Probabilístico e Análise Computacional de Fatores Associados à Mortalidade Infantil na Região Sudeste do Brasil

Raquel Larrat Lopes

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Heder Soares Bernardino

JUIZ DE FORA

JULHO, 2026

Linkage Probabilístico e Análise Computacional de Fatores Associados à Mortalidade Infantil na Região Sudeste do Brasil

Raquel Larrat Lopes

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Heder Soares Bernardino
Doutor em Modelagem Computacional

Pedro Henrique Gasparetto Lugão
Mestre em Matemática

Luciana Conceição Dias Campos
Doutora em Engenharia Elétrica

JUIZ DE FORA

16 DE JULHO, 2026

Resumo

A mortalidade infantil constitui um importante indicador das condições de saúde de uma população. No Brasil, os dados sobre nascimentos e óbitos são registrados de forma independente em dois sistemas de informação do SUS: o Sistema de Informações sobre Nascidos Vivos (SINASC) e o Sistema de Informações sobre Mortalidade (SIM). A ausência de um identificador único comum entre essas bases, contudo, dificulta a análise integrada dos fatores associados ao óbito infantil, tornando necessária a aplicação de técnicas de vinculação de registros. Diante disso, este trabalho teve como objetivo identificar fatores associados ao óbito infantil na região Sudeste do Brasil por meio de técnicas computacionais. Para isso, realizou-se o *linkage* probabilístico entre o SIM e o SINASC, organizado em três rodadas sucessivas com flexibilização progressiva dos critérios de similaridade. O procedimento vinculou aproximadamente 82,6% dos óbitos infantis registrados no SIM (30.220 pares). Sobre a base integrada, realizou-se uma análise exploratória que caracterizou o perfil dos registros e revelou gradientes de mortalidade coerentes com o conhecimento epidemiológico. Em seguida, foram aplicados e comparados três modelos de aprendizado de máquina — *Random Forest*, Regressão Logística e XGBoost — para prever a ocorrência de óbito infantil em um cenário de forte desbalanceamento de classes (0,68% de óbitos). O XGBoost apresentou o melhor desempenho, com AUC-PR de 0,3561, F2-score de 0,5039 e sensibilidade de 0,5746. A análise da importância das variáveis, conduzida por meio da importância por ganho, dos valores SHAP e dos coeficientes da Regressão Logística, revelou convergência entre os modelos: o peso ao nascer, a duração da gestação, os índices de Apgar e a presença de anomalia congênita destacaram-se como os fatores mais associados ao óbito infantil, em consonância com o conhecimento clínico da área. Os resultados evidenciam o potencial das técnicas de aprendizado de máquina como ferramentas de apoio à análise da mortalidade infantil e à identificação de grupos de maior risco em saúde pública.

Palavras-chave: Mortalidade infantil. *Linkage* probabilístico. Aprendizado de máquina.

SIM. SINASC.

Abstract

Infant mortality is an important indicator of a population's health conditions. In Brazil, data on births and deaths are recorded independently in two SUS (Unified Health System) information systems: the Live Birth Information System (SINASC) and the Mortality Information System (SIM). The absence of a common unique identifier between these databases, however, hinders the integrated analysis of factors associated with infant death, making the application of record linkage techniques necessary. In this context, this work aimed to identify factors associated with infant death in the Southeast region of Brazil through computational techniques. To this end, probabilistic linkage was performed between SIM and SINASC, organized in three successive rounds with progressive relaxation of the similarity criteria. The procedure linked approximately 82.6% of the infant deaths recorded in SIM (30,220 pairs). On the integrated database, an exploratory data analysis was carried out, characterizing the profile of the records and revealing mortality gradients consistent with epidemiological knowledge. Then, three machine learning models — Random Forest, Logistic Regression, and XGBoost — were applied and compared to predict the occurrence of infant death in a scenario of strong class imbalance (0.68% of deaths). XGBoost achieved the best performance, with an AUC-PR of 0.3561, an F2-score of 0.5039, and a sensitivity of 0.5746. The analysis of variable importance, conducted through gain-based importance, SHAP values, and the coefficients of the Logistic Regression, revealed convergence among the models: birth weight, gestational age, Apgar scores, and the presence of congenital anomalies stood out as the factors most associated with infant death, in line with clinical knowledge in the field. The results highlight the potential of machine learning techniques as tools to support the analysis of infant mortality and the identification of higher-risk groups in public health.

Keywords: Infant mortality. Probabilistic linkage. Machine learning. SIM. SINASC.

Agradecimentos

Aos meus pais, pelo encorajamento e apoio.

Ao professor Heder pela orientação, amizade e principalmente, pela paciência, sem a qual este trabalho não se realizaria.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AUC-PR	*Area Under the Precision-Recall Curve*
AUC-ROC	*Area Under the ROC Curve*
DCC	Departamento de Ciência da Computação
DN	Declaração de Nascido Vivo
SIM	Sistema de Informações sobre Mortalidade
SINASC	Sistema de Informações sobre Nascidos Vivos
UF	Unidade da Federação
UFJF	Universidade Federal de Juiz de Fora

1 Introdução

A mortalidade infantil, definida como o óbito de crianças menores de um ano de idade, constitui um dos mais importantes indicadores de saúde pública, refletindo as condições socioeconômicas de uma população e a qualidade dos serviços de saúde a ela oferecidos (RASELLA et al., 2013). No Brasil, apesar dos avanços observados nas últimas décadas, persistem desafios significativos na redução desse indicador, com marcantes disparidades entre regiões e grupos populacionais (Brasil. Ministério da Saúde, 2021). A região Sudeste, embora concentre alguns dos melhores indicadores socioeconômicos do país, ainda apresenta heterogeneidade interna relevante.

A análise dos fatores associados à mortalidade infantil depende da disponibilidade de informações integradas sobre o nascimento e o óbito de cada criança. No Brasil, esses dados são registrados em dois sistemas distintos: o Sistema de Informações sobre Nascidos Vivos (SINASC), que reúne informações sobre os nascimentos, e o Sistema de Informações sobre Mortalidade (SIM), que registra os óbitos. Embora a Declaração de Nascido Vivo (DN) esteja presente em ambos os sistemas, esse identificador não é disponibilizado nas bases públicas em razão das restrições impostas pela Lei Geral de Proteção de Dados (LGPD), o que inviabiliza a vinculação determinística direta entre os registros.

Essa fragmentação impõe um desafio metodológico para estudos epidemiológicos. A integração das bases por meio de técnicas de *linkage* possibilita reconstruir, de forma indireta, a associação entre cada nascimento e o eventual óbito infantil, ampliando substancialmente o potencial analítico dos dados. Na ausência de um identificador único confiável, recorre-se ao *linkage* probabilístico, que estima a possibilidade de dois registros pertencerem ao mesmo indivíduo a partir da concordância entre múltiplas variáveis compartilhadas pelas bases (FELLEGI; SUNTER, 1969; CHRISTEN, 2012).

Uma vez construída a base integrada, técnicas de aprendizado de máquina podem ser empregadas para investigar a relação entre características maternas, obstétricas e do recém-nascido e a ocorrência do óbito infantil. Esses métodos são capazes de capturar relações não lineares e interações entre variáveis em grandes volumes de dados, além

de fornecer medidas de importância dos preditores que auxiliam na interpretação dos principais fatores associados ao desfecho.

Diante desse contexto, este trabalho tem como objetivo desenvolver e aplicar uma metodologia de *linkage* probabilístico para integração das bases SIM e SINASC da região Sudeste do Brasil e, a partir da base resultante, analisar os fatores associados à mortalidade infantil. O processo de vinculação foi estruturado em três rodadas sucessivas, com flexibilização progressiva dos critérios de comparação, e permitiu vincular aproximadamente 82,6% dos óbitos infantis registrados no SIM. Sobre a base integrada, foram aplicados e comparados três modelos de classificação — *Random Forest*, Regressão Logística e XGBoost — tanto para prever a ocorrência de óbito infantil quanto para identificar os fatores mais associados ao desfecho, considerando o forte desbalanceamento entre as classes característico do problema.

2 Fundamentação Teórica

2.1 *Linkage* de Bases de Dados

O *linkage* de registros consiste no processo de identificar e vincular registros de diferentes bases de dados que se referem à mesma entidade, indivíduo ou evento. Essa técnica é especialmente importante quando as bases analisadas não possuem um identificador único comum que permita a vinculação direta entre os registros. Nesses casos, a associação entre as bases precisa ser realizada por meio da comparação de variáveis disponíveis em ambas as fontes de dados.

Quando existe um identificador único confiável, como um número de documento ou código individual, o processo pode ser realizado de forma determinística, isto é, por correspondência exata desse identificador. No entanto, quando esse identificador está ausente, incompleto ou não disponível, torna-se necessário utilizar métodos probabilísticos. O *linkage* probabilístico estima a possibilidade de dois registros pertencerem à mesma entidade a partir da concordância entre múltiplas variáveis, como datas, características individuais, informações geográficas e outros atributos compartilhados entre as bases (FELLEGI; SUNTER, 1969).

Em bases de grande volume, a comparação direta entre todos os registros de uma base com todos os registros da outra pode ser computacionalmente inviável. Por esse motivo, uma etapa importante do processo de *linkage* é o *blocking*, técnica que restringe o conjunto de comparações a pares de registros que concordam em determinadas variáveis consideradas estruturantes (CHRISTEN, 2012). Essa estratégia reduz o número de combinações analisadas e concentra o cálculo de similaridade nos pares com maior plausibilidade inicial de correspondência.

Além da comparação exata entre variáveis, podem ser utilizadas medidas de similaridade para lidar com divergências de preenchimento. Entre essas medidas, destaca-se a distância de Levenshtein, uma métrica de similaridade textual baseada no número mínimo de operações de edição, como inserção, remoção ou substituição de caracteres, necessárias

para transformar uma string em outra. Segundo Schraagen, medidas de distância de edição são amplamente utilizadas em problemas de *record linkage*, pois permitem lidar com pequenas variações ou erros nos valores comparados (SCHRAAGEN, 2014).

Outro aspecto relevante em processos de vinculação é o controle de ambiguidades. Em determinadas aplicações, adota-se a restrição de pareamento um-para-um (1:1), garantindo que cada registro de uma base seja vinculado a, no máximo, um registro da outra. Essa estratégia contribui para evitar duplicidades e preservar a consistência da base resultante.

2.2 Aprendizado de Máquina em Saúde Pública

O aprendizado de máquina é uma área da inteligência artificial voltada ao desenvolvimento de métodos computacionais capazes de aprender padrões a partir dos dados. Em vez de depender apenas de regras previamente definidas, esses modelos utilizam exemplos disponíveis em uma base de dados para identificar relações entre variáveis e realizar tarefas como classificação, predição ou agrupamento.

Na área da saúde pública, o uso de aprendizado de máquina tem se tornado cada vez mais relevante devido à grande quantidade de dados disponíveis em sistemas de informação. Esses métodos permitem analisar conjuntos de dados extensos e identificar padrões que nem sempre são facilmente observados por meio de análises descritivas tradicionais. No contexto da mortalidade infantil, modelos de aprendizado de máquina podem auxiliar na investigação da relação entre características maternas, obstétricas e do recém-nascido com a ocorrência do óbito antes de um ano de idade.

2.3 *Random Forest*

Entre os algoritmos de aprendizado de máquina supervisionado, destaca-se o *Random Forest*. Esse modelo foi proposto por Breiman e é baseado na combinação de múltiplas árvores de decisão (BREIMAN, 2001). Uma árvore de decisão funciona dividindo os dados em grupos menores a partir de perguntas sucessivas sobre as variáveis, até chegar a uma classificação final. No entanto, uma única árvore pode se ajustar demais aos dados de

treinamento, reduzindo sua capacidade de generalização para novos dados.

O *Random Forest* busca reduzir esse problema ao construir várias árvores de decisão de forma conjunta. Cada árvore é treinada com uma amostra aleatória dos dados e, em cada divisão, apenas um subconjunto das variáveis é considerado. Ao final, o resultado do modelo é definido pela combinação das previsões de todas as árvores. Em problemas de classificação, a decisão final geralmente ocorre por votação majoritária entre as árvores (BREIMAN, 2001).

Uma das principais vantagens do *Random Forest* é sua capacidade de lidar com bases de dados grandes e com muitas variáveis. Além disso, o modelo consegue capturar relações não lineares e interações entre os atributos, sem exigir que essas relações sejam previamente especificadas. Essa característica é relevante em estudos de mortalidade infantil, pois o desfecho pode estar associado a uma combinação de diferentes fatores, como peso ao nascer, idade gestacional, idade materna, tipo de parto, escolaridade da mãe e histórico reprodutivo.

Outro aspecto relevante do *Random Forest* é a possibilidade de analisar a importância das variáveis utilizadas no modelo. Essa análise permite identificar quais atributos contribuíram mais para a classificação dos registros, auxiliando na interpretação dos principais fatores associados ao desfecho investigado (BREIMAN, 2001).

2.4 XGBoost

O XGBoost (*Extreme Gradient Boosting*) é um algoritmo de aprendizado de máquina supervisionado baseado na técnica de *gradient boosting* sobre árvores de decisão, proposto por Chen e Guestrin (CHEN; GUESTRIN, 2016). Assim como o *Random Forest*, o XGBoost combina múltiplas árvores de decisão para produzir suas previsões. Entretanto, a forma como as árvores são construídas difere de maneira fundamental entre os dois métodos.

Enquanto o *Random Forest* constrói árvores de forma independente e combina seus resultados ao final, o XGBoost constrói as árvores de maneira sequencial. Cada nova árvore é ajustada para corrigir os erros cometidos pelas árvores anteriores, de modo que o modelo é aprimorado progressivamente a cada iteração. Esse processo é guiado pela

minimização de uma função de perda, à qual se incorpora um termo de regularização que penaliza a complexidade do modelo, contribuindo para reduzir o sobreajuste (CHEN; GUESTRIN, 2016).

Entre as características que tornaram o XGBoost amplamente utilizado estão sua eficiência computacional e sua escalabilidade para grandes volumes de dados. O algoritmo incorpora otimizações como o processamento paralelo e métodos eficientes para a construção das árvores, como o método baseado em histogramas, que reduz o custo computacional e o consumo de memória. Essas características tornam o modelo particularmente adequado para bases de dados extensas, como as utilizadas em estudos de saúde pública (CHEN; GUESTRIN, 2016).

Assim como o *Random Forest*, o XGBoost também permite avaliar a importância das variáveis utilizadas no modelo, possibilitando identificar quais atributos mais contribuíram para as previsões. Adicionalmente, por se tratar de um modelo baseado em árvores, não é sensível à escala das variáveis, o que dispensa a necessidade de padronização dos atributos.

2.5 Regressão Logística

A Regressão Logística é um modelo estatístico amplamente utilizado em problemas nos quais a variável resposta possui natureza categórica, especialmente quando o desfecho é binário. Diferentemente da regressão linear, que busca estimar diretamente um valor numérico contínuo, a Regressão Logística estima a probabilidade de ocorrência de determinado evento a partir de um conjunto de variáveis explicativas. Esse tipo de abordagem é adequado para situações em que se deseja analisar a ocorrência ou não de um fenômeno, como presença ou ausência de uma condição, sucesso ou fracasso, ou ocorrência ou não de óbito infantil.

O modelo de Regressão Logística está associado ao estudo de desfechos binários e foi discutido em trabalhos clássicos como o de Cox, voltado à análise de sequências binárias (COX, 1958). Em termos gerais, o modelo utiliza uma função logística para transformar uma combinação linear das variáveis explicativas em uma probabilidade situada entre 0 e 1. Dessa forma, a Regressão Logística permite estimar a probabilidade de um registro

pertencer a uma determinada classe, considerando os valores das variáveis observadas.

No contexto da saúde, a Regressão Logística é frequentemente utilizada por sua capacidade de modelar a relação entre fatores de risco e desfechos binários. Segundo Hosmer, Lemeshow e Sturdivant, o modelo é especialmente útil quando o objetivo é avaliar a associação entre uma variável resposta dicotômica e um conjunto de covariáveis (HOSMER; LEMESHOW; STURDIVANT, 2013). Essa característica torna o método apropriado para estudos epidemiológicos, nos quais é comum investigar a relação entre características individuais, clínicas, sociais ou ambientais e a ocorrência de determinado evento de saúde.

Uma das principais vantagens da Regressão Logística é sua interpretabilidade. Enquanto modelos mais complexos, como o *Random Forest* e o XGBoost, podem apresentar maior capacidade de capturar relações não lineares e interações entre variáveis, a Regressão Logística permite uma interpretação mais direta dos coeficientes estimados. Esses coeficientes indicam a direção e a intensidade da associação entre cada variável explicativa e a probabilidade de ocorrência do desfecho, podendo ser convertidos em razões de chances, conhecidas como *odds ratios* (HOSMER; LEMESHOW; STURDIVANT, 2013).

2.6 Pipeline de Pré-processamento

Em problemas de aprendizado de máquina, os dados brutos raramente estão prontos para serem utilizados diretamente pelos algoritmos. É comum que apresentem valores ausentes, variáveis categóricas representadas por texto e variáveis numéricas em escalas muito diferentes entre si. Antes de treinar um modelo, é necessário tratar essas questões por meio de etapas de pré-processamento, como a imputação de valores ausentes, a codificação das variáveis categóricas e, em alguns casos, a padronização das escalas.

Uma forma organizada de aplicar essas etapas é por meio de um *pipeline*. O *pipeline* pode ser compreendido como uma sequência ordenada de transformações, semelhante a uma linha de montagem: os dados entram em uma extremidade, percorrem automaticamente cada etapa de preparação e, ao final, são entregues ao modelo já prontos para o treinamento ou a previsão. Na biblioteca `scikit-learn`, utilizada neste trabalho,

essa estrutura encadeia as etapas de pré-processamento e o próprio modelo em um único objeto.

A principal vantagem do *pipeline* é de natureza metodológica e está relacionada à prevenção do vazamento de informação (*data leakage*). Esse problema ocorre quando informações dos conjuntos de validação ou teste influenciam, ainda que indiretamente, o processo de treinamento do modelo. Por exemplo, se a mediana utilizada para preencher valores ausentes fosse calculada sobre toda a base, o modelo estaria sendo construído com base em informações que deveriam permanecer reservadas para avaliação. Ao utilizar um *pipeline*, garante-se que os parâmetros de cada transformação — como a mediana da imputação ou as categorias da codificação — sejam aprendidos exclusivamente a partir do conjunto de treino e, em seguida, aplicados aos demais conjuntos sem serem recalculados. Esse cuidado assegura uma avaliação mais realista e confiável do desempenho do modelo.

Neste trabalho, o pré-processamento foi composto por três etapas principais. A primeira é a imputação dos valores ausentes, na qual as variáveis numéricas têm seus valores faltantes preenchidos pela mediana e as variáveis ordinais pela moda, ambas calculadas no conjunto de treino. A segunda é a codificação das variáveis categóricas nominais por meio do *One-Hot Encoding*, que transforma cada categoria em uma coluna binária, permitindo que o modelo as interprete numericamente. A terceira é a padronização das variáveis numéricas e ordinais, que ajusta todas a uma escala comum.

Cabe destacar que a etapa de padronização não é necessária para todos os modelos. Modelos baseados em árvores de decisão, como o *Random Forest* e o XGBoost, realizam divisões a partir de pontos de corte e não são sensíveis à escala das variáveis, dispensando essa etapa. A Figura 2.1 ilustra o *pipeline* utilizado nesses modelos.

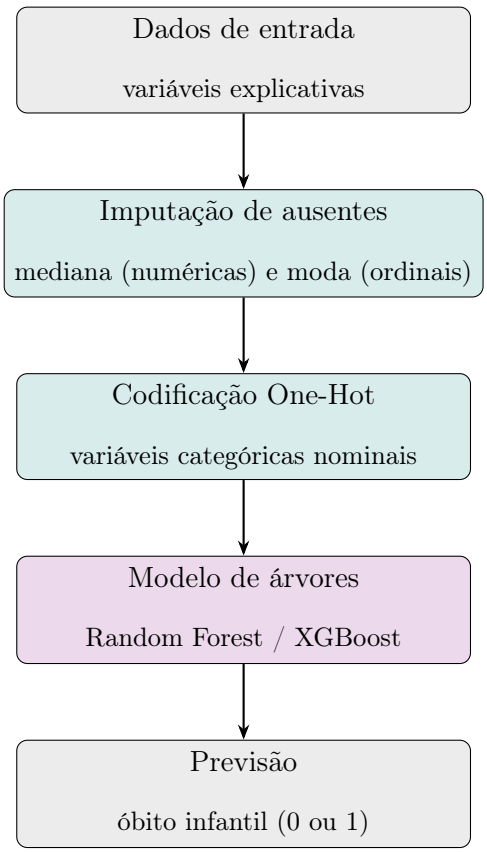

Figura 2.1: *Pipeline* utilizado nos modelos baseados em árvores (*Random Forest* e XGBoost), composto pelas etapas de imputação e codificação, sem padronização.

Já a Regressão Logística, por se basear em uma combinação das variáveis para estimar a probabilidade do desfecho, é sensível à escala, sendo a padronização aplicada apenas nesse modelo. A Figura 2.2 apresenta o *pipeline* correspondente, que inclui a etapa adicional de padronização.

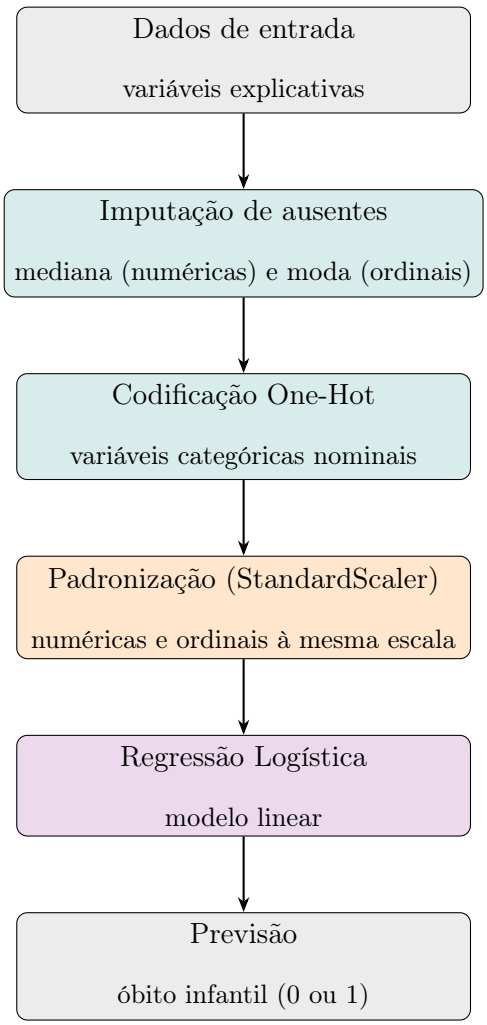

Figura 2.2: *Pipeline* utilizado na Regressão Logística, que inclui a etapa de padronização das variáveis numéricas e ordinais.

Dessa forma, embora os três modelos compartilhem as mesmas etapas iniciais de imputação e codificação, apenas a Regressão Logística incorpora a padronização, refletindo as características de cada algoritmo.

2.7 Desbalanceamento de Classes e Métricas de Avaliação

Em problemas de classificação, especialmente na área da saúde, é comum que uma das classes seja muito mais frequente do que a outra. Esse fenômeno é conhecido como desbalanceamento de classes. Em bases de mortalidade infantil, por exemplo, a quantidade de registros de nascidos vivos que não evoluem para óbito tende a ser muito maior do

que a quantidade de registros associados a óbitos infantis. Segundo Niaz, Shahariar e Patwary, o desbalanceamento de classes constitui um problema relevante em aprendizado de máquina, pois afeta o processo de classificação e exige métodos específicos para seu tratamento (NIAZ; SHAHARIAR; PATWARY, 2022).

O desbalanceamento pode fazer com que o modelo favoreça a classe majoritária, apresentando bom desempenho aparente, mas baixa capacidade de identificar corretamente a classe minoritária. Uma das formas de atenuar esse efeito, sem alterar a distribuição original dos dados, é o ajuste dos pesos das classes durante o treinamento, atribuindo maior peso à classe minoritária, de modo que os erros cometidos sobre ela sejam penalizados mais fortemente. Essa foi a estratégia adotada neste trabalho, por meio de parâmetros específicos de cada algoritmo, como detalhado nas seções de modelagem.

Existem, na literatura, diferentes estratégias para lidar com o desbalanceamento de classes, que podem ser organizadas em dois grandes grupos. O primeiro atua no nível dos dados, alterando a distribuição das classes antes do treinamento, seja por sobreamostragem da classe minoritária — como a simples replicação de exemplos ou a geração de exemplos sintéticos, a exemplo do método SMOTE (*Synthetic Minority Over-sampling Technique*) — seja por subamostragem da classe majoritária, na qual parte dos registros mais frequentes é descartada. O segundo grupo atua no nível do algoritmo, mantendo os dados inalterados e ajustando o processo de aprendizado para penalizar mais fortemente os erros sobre a classe minoritária. A ponderação de classes (*class weighting*) insere-se nessa segunda categoria: em vez de modificar a base, atribui-se um peso maior às instâncias da classe de interesse na função de custo, de modo que os erros cometidos sobre ela tenham maior impacto no ajuste do modelo.

Neste trabalho, optou-se pela ponderação de classes por duas razões principais. Em primeiro lugar, por preservar a distribuição original dos dados, evitando tanto a introdução de exemplos artificiais — que podem não refletir padrões reais — quanto o descarte de informação, indesejável sobretudo diante do número já reduzido de óbitos. Em segundo lugar, por sua simplicidade e pela disponibilidade direta em todos os três algoritmos utilizados, por meio de parâmetros específicos: o `class_weight` no *Random Forest* e na Regressão Logística, e o `scale_pos_weight` no XGBoost. Em todos os casos,

o mecanismo é equivalente, atribuindo maior peso à classe de óbito de forma inversamente proporcional à sua frequência, sem alterar o conjunto de dados.

A avaliação de modelos aplicados a bases desbalanceadas não deve se limitar à acurácia, pois essa métrica pode apresentar interpretação enganosa: em uma base na qual a classe negativa predomina, um classificador que preveja todos os registros como negativos alcançaria acurácia elevada, ainda que não identificasse nenhum caso da classe de interesse. Por esse motivo, foram utilizadas métricas mais informativas, baseadas na matriz de confusão.

A **precisão** informa, entre os registros classificados como óbito, quantos realmente correspondem a óbitos, refletindo a proporção de alarmes verdadeiros. A **sensibilidade** (ou *recall*) indica a proporção de óbitos efetivamente identificados pelo modelo, sendo particularmente relevante no contexto da mortalidade infantil, em que deixar de identificar um caso de risco (falso negativo) tende a ser mais grave do que gerar um alarme falso (falso positivo). O ***F1-score*** combina precisão e sensibilidade por meio de sua média harmônica, atribuindo peso equivalente às duas.

Como, neste trabalho, a identificação dos casos de óbito é prioritária, adotou-se também o ***F2-score*** como métrica principal para a seleção de modelos e limiares. O *F2-score* é uma generalização do *F1-score* que atribui maior peso à sensibilidade do que à precisão, sendo, portanto, mais adequado a situações em que a omissão de casos positivos é especialmente indesejável. Essa escolha reflete o objetivo de favorecer a recuperação do maior número possível de óbitos infantis, ainda que ao custo de algum aumento no número de falsos positivos.

Além das métricas dependentes de um limiar de decisão, foram utilizadas duas métricas que avaliam o desempenho do modelo de forma independente do limiar escolhido. A **AUC-ROC** mede a capacidade geral do modelo de distinguir entre as classes, considerando a relação entre a taxa de verdadeiros positivos e a de falsos positivos. Contudo, em bases fortemente desbalanceadas, a AUC-ROC pode apresentar valores elevados mesmo quando o desempenho sobre a classe minoritária é limitado. Por essa razão, deu-se ênfase à **AUC-PR**, a área sob a curva de precisão e sensibilidade, que avalia diretamente o desempenho na classe de interesse e é considerada mais informativa nesse cenário (NIAZ;

SHAHARIAR; PATWARY, 2022).

2.8 Limiar de Decisão

Os modelos de classificação empregados neste trabalho — *Random Forest*, Regressão Logística e XGBoost — não produzem diretamente uma resposta binária, mas sim uma probabilidade estimada de que cada registro pertença à classe de interesse, situada entre 0 e 1. Para converter essa probabilidade em uma decisão concreta (óbito ou não óbito), é necessário estabelecer um valor de corte, denominado limiar de decisão (*threshold*): registros cuja probabilidade estimada seja maior ou igual ao limiar são classificados como positivos, e os demais, como negativos.

Por convenção, muitas implementações adotam como padrão o limiar de 0,5. Esse valor, contudo, não é uma escolha ótima universal, mas apenas uma referência neutra, adequada sobretudo a bases equilibradas. Em problemas com forte desbalanceamento de classes, como o da mortalidade infantil, o limiar padrão de 0,5 tende a favorecer a classe majoritária, resultando em baixa sensibilidade sobre a classe de interesse. Por essa razão, o limiar passa a ser tratado como um parâmetro de decisão, ajustável de acordo com os objetivos da aplicação.

A escolha do limiar controla diretamente o compromisso entre precisão e sensibilidade. Ao reduzir o limiar, mais registros passam a ser classificados como positivos, o que tende a aumentar a sensibilidade — pois um número maior de óbitos verdadeiros é identificado —, mas também eleva a quantidade de falsos positivos, reduzindo a precisão. Inversamente, ao aumentar o limiar, o modelo torna-se mais exigente para classificar um registro como positivo, elevando a precisão, porém deixando de identificar parte dos casos de interesse e reduzindo a sensibilidade. Não existe, portanto, um limiar universalmente correto: o valor mais adequado depende das consequências práticas de cada tipo de erro no contexto da aplicação.

No caso da mortalidade infantil, deixar de identificar um caso de risco (falso negativo) tende a ser mais grave do que gerar um alarme falso (falso positivo). Esse cenário justifica a adoção de limiares que privilegiem a sensibilidade, ainda que ao custo de alguma perda de precisão, em coerência com a escolha do F2-score como métrica

principal, discutida na seção anterior. Neste trabalho, o limiar de decisão foi selecionado de forma independente para cada modelo, a partir do conjunto de validação, buscando-se o valor que maximiza o F2-score, conforme detalhado nos capítulos de metodologia e experimentos.

2.9 Interpretabilidade de Modelos e Valores SHAP

Modelos de aprendizado de máquina baseados em árvores, como o *Random Forest* e o XGBoost, costumam apresentar bom desempenho preditivo, mas são frequentemente descritos como modelos de "caixa-preta", uma vez que a forma como combinam um grande número de divisões e árvores torna pouco transparente o modo como cada variável influencia uma previsão. Em aplicações de saúde pública, contudo, não basta prever o desfecho: é igualmente importante compreender quais fatores estão associados ao risco e de que maneira contribuem para ele. Por essa razão, métodos de interpretabilidade tornam-se essenciais para extrair conhecimento dos modelos, e não apenas previsões.

Uma forma tradicional de interpretação em modelos baseados em árvores é a importância por ganho, que mede o quanto cada variável contribuiu, em média, para melhorar as divisões das árvores durante o treinamento. Embora útil, essa métrica apresenta limitações. Em primeiro lugar, ela indica apenas o quanto uma variável foi utilizada, sem informar a direção de seu efeito, ou seja, se determinado valor aumenta ou reduz a probabilidade do desfecho. Em segundo lugar, tende a favorecer variáveis de alta cardinalidade, que, ao gerarem muitas colunas após o *One-Hot Encoding*, oferecem ao modelo um número maior de oportunidades de divisão, inflando artificialmente sua importância aparente.

Para superar essas limitações, foram utilizados neste trabalho os valores SHAP (*SHapley Additive exPlanations*), propostos por Lundberg e Lee (LUNDBERG; LEE, 2017). Os valores SHAP têm origem na teoria dos jogos cooperativos, mais especificamente no conceito de valor de Shapley, formulado originalmente para distribuir de forma justa o ganho obtido por um grupo de jogadores entre os seus participantes, de acordo com a contribuição de cada um. Na adaptação para o aprendizado de máquina, cada variável explicativa é tratada como um "jogador", e a "recompensa" a ser distribuída corresponde

à diferença entre a previsão do modelo para um determinado registro e a previsão média sobre todos os registros. Dessa forma, o valor SHAP de uma variável quantifica o quanto ela contribuiu, para mais ou para menos, para afastar aquela previsão específica do valor médio.

A principal vantagem dessa abordagem é fornecer explicações *locais*: para cada registro individual, é possível decompor a previsão na soma das contribuições de cada variável, identificando quais atributos empurraram a estimativa no sentido de maior risco e quais atuaram no sentido oposto. Uma propriedade central dos valores SHAP é a aditividade: a soma das contribuições de todas as variáveis, somada ao valor base (a previsão média), recompõe exatamente a previsão do modelo para aquele registro. Essa característica confere coerência à interpretação, evitando atribuições arbitrárias ou contraditórias.

Além da interpretação individual, os valores SHAP permitem uma visão *global* do modelo. Ao se agregar o valor absoluto das contribuições de cada variável ao longo de muitos registros, obtém-se uma medida de importância que reflete a influência efetiva de cada atributo sobre as previsões, e não apenas a frequência com que foi utilizada nas divisões. Por considerar a contribuição real de cada variável, essa medida é menos suscetível à distorção causada por variáveis de alta cardinalidade do que a importância por ganho, oferecendo um ordenamento de relevância mais fiel ao comportamento do modelo.

O cálculo exato dos valores SHAP é computacionalmente custoso, pois exigiria avaliar a contribuição de cada variável considerando todas as combinações possíveis das demais. Para contornar esse problema em modelos baseados em árvores, Lundberg e Lee propuseram o algoritmo *TreeSHAP*, que explora a estrutura das árvores de decisão para calcular os valores SHAP de forma exata e eficiente, com custo computacional substancialmente menor. Essa implementação é particularmente adequada a modelos como o *Random Forest* e o XGBoost, sendo a utilizada neste trabalho para aprofundar a interpretação dos resultados, de modo complementar à importância por ganho e aos coeficientes da Regressão Logística.

3 Trabalhos Relacionados

A aplicação de técnicas de vinculação de bases e de aprendizado de máquina ao estudo da mortalidade infantil tem sido explorada em diversos trabalhos na literatura, que fornecem o contexto no qual a presente pesquisa se insere.

No que se refere à integração de bases de dados de saúde, Maia, Souza e Mendes analisaram a contribuição do *linkage* entre o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informações sobre Nascidos Vivos (SINASC) para a melhoria das informações da mortalidade infantil em cinco cidades brasileiras, evidenciando a relevância dessa estratégia para qualificar dados relacionados a óbitos infantis (MAIA; SOUZA; MENDES, 2015). Esse trabalho demonstra a viabilidade e a importância da vinculação entre o SIM e o SINASC, abordagem também adotada na presente pesquisa.

Quanto ao uso de aprendizado de máquina na área da saúde, Dos Santos et al. realizaram uma análise bibliométrica sobre a aplicação de técnicas de mineração de dados e aprendizado de máquina em problemas de saúde pública entre 2009 e 2018, apontando a relevância crescente dessas abordagens para a análise de dados na área (SANTOS et al., 2019). Esse panorama evidencia a consolidação dessas técnicas como ferramentas de análise em saúde pública.

Especificamente no estudo da mortalidade infantil, Sanders et al. analisaram fatores associados à mortalidade infantil em Fortaleza, Ceará, utilizando regressão logística para estimar a associação entre características maternas, assistenciais e do recém-nascido e a ocorrência de óbito infantil (SANDERS et al., 2017). Esse estudo ilustra a aplicação da Regressão Logística para investigar fatores associados ao óbito infantil e interpretar a força de associação entre as variáveis e o desfecho, abordagem semelhante à empregada neste trabalho, aqui combinada a modelos baseados em árvores de decisão.

Em conjunto, esses trabalhos evidenciam tanto a pertinência da vinculação entre o SIM e o SINASC quanto a adequação do uso de modelos de aprendizado de máquina e estatísticos para a investigação da mortalidade infantil. A presente pesquisa se diferencia ao integrar essas abordagens, aplicando e comparando três modelos distintos — *Random*

Forest, Regressão Logística e XGBoost — sobre a base integrada SIM-SINASC da região Sudeste.

4 Linkage Probabilístico das Bases SIM-SINASC

4.1 Caracterização Inicial das Bases de Dados

Foram utilizadas as bases do Sistema de Informações sobre Mortalidade (SIM) e do Sistema de Informações sobre Nascidos Vivos (SINASC), ambas referentes à região Sudeste. A base do SIM foi previamente tratada e filtrada para contemplar exclusivamente os registros de mortalidade infantil dos anos de 2018, 2019 e 2020, totalizando 36.594 registros e 87 variáveis. A base do SINASC compreendeu os registros de nascidos vivos dos anos de 2017 a 2020, totalizando 4.456.504 registros e 61 variáveis. A inclusão de nascimentos a partir de 2017 foi necessária para garantir a adequada correspondência temporal entre nascimento e óbito, considerando que os óbitos infantis ocorridos entre 2018 e 2020 poderiam estar associados a nascimentos ocorridos no ano anterior. O processo de *linkage* teve como objetivo identificar, entre os 4.456.504 nascimentos registrados no SINASC, aqueles correspondentes aos 36.594 óbitos infantis registrados no SIM.

4.2 Metodologia Geral do Processo de *Linkage*

Considerando a inexistência de um identificador determinístico comum entre as bases, como o número da Declaração de Nascido Vivo (DN), a vinculação entre o SIM e o SINASC foi conduzida por meio de um procedimento de *linkage* probabilístico, estruturado em três rodadas sucessivas. Cada rodada foi delineada com diferentes graus de restrição e flexibilização dos critérios de comparação, visando equilibrar especificidade e sensibilidade na identificação dos pares.

Apesar das especificidades de cada rodada, todas seguiram uma estrutura metodológica comum, composta pelas seguintes características.

Definição de pares candidatos por meio de *blocking*. Em cada rodada,

inicialmente foi aplicado um critério de *blocking* para restringir o universo de comparações a registros com concordância em variáveis consideradas estruturantes do evento de nascimento. O conjunto de pares candidatos gerado pelo *blocking* permaneceu fixo ao longo das iterações daquela rodada.

Cálculo de similaridade. Para cada par candidato, foi construída uma matriz de similaridade binária, na qual cada variável comparada assumia valor 1 (concordância) ou 0 (discordância), conforme o critério de comparação definido (exato ou flexibilizado).

Estratégia iterativa com redução progressiva do limiar. O pareamento foi conduzido por meio de iterações sucessivas dentro de cada rodada. Inicialmente, exigia-se concordância no maior número possível de variáveis comparadas e, a cada nova iteração, o número mínimo de variáveis concordantes (limiar) era progressivamente reduzido até o limite inferior estabelecido para aquela rodada. Essa abordagem permitiu priorizar, nas primeiras etapas, os pares com maior evidência de correspondência, recuperando posteriormente pares com pequenas inconsistências de preenchimento.

Restrição de pareamento um-para-um (1:1). Em todas as rodadas, foi aplicada a restrição de pareamento um-para-um, permitindo apenas correspondências exclusivas entre registros do SIM e do SINASC. Assim, pares ambíguos, nos quais um mesmo registro apresentava múltiplas possibilidades de vinculação, não eram aceitos na respectiva iteração.

Remoção de registros já vinculados. Os registros vinculados com sucesso em uma determinada iteração eram removidos das iterações e etapas subsequentes, impedindo sua reutilização em limiares mais flexíveis e preservando os vínculos estabelecidos sob maior grau de concordância.

Ao final de cada rodada, os registros ainda não vinculados eram submetidos à rodada seguinte, na qual os critérios de *blocking* e/ou de similaridade podiam ser ajustados com o objetivo de aumentar a sensibilidade do procedimento.

4.3 Primeira Rodada de *Linkage*

A primeira rodada foi estruturada de forma conservadora, priorizando elevada especificidade na identificação dos pares.

4.3.1 Seleção das Variáveis

Foram selecionadas variáveis presentes em ambas as bases com potencial discriminatório para identificação dos indivíduos. As variáveis comparadas incluíram características do recém-nascido — sexo (`SEXO`), peso ao nascer (`PESO`), data de nascimento (`DTNASC`) e município de nascimento (`MUNNASC`); características maternas — idade da mãe (`IDADEMAE`), escolaridade (`ESCMAE`, `ESCMAEAGR1`, `ESCMAE2010`, `SERIESCMAE`) e ocupação (`OCUPMAE` no SIM e `CODOCUPMAE` no SINASC); histórico reprodutivo — número de filhos vivos (`QTDFILVIVO`) e mortos (`QTDFILMORT`); e informações obstétricas — tipo de parto (`PARTO`), tipo de gravidez (`GRAVIDEZ`), duração da gestação (`GESTACAO` e `SEMAGESTAC`) e raça/cor (`RACACOR`). No total, foram consideradas 16 variáveis comparáveis na etapa de cálculo de similaridade.

A seleção dessas variáveis baseou-se em dois critérios principais: disponibilidade simultânea nas duas bases e capacidade discriminatória para identificação do mesmo indivíduo. Em estudos de *linkage* sem identificador único, a correspondência entre registros é inferida a partir da combinação de atributos que, quando analisados em conjunto, apresentam baixa probabilidade de coincidência aleatória. Nesse contexto, variáveis como sexo, data de nascimento, município de nascimento e peso ao nascer descrevem diretamente o recém-nascido e tendem a apresentar alta estabilidade entre os sistemas. Da mesma forma, características maternas e obstétricas, como idade da mãe, escolaridade, ocupação, tipo de parto, tipo de gravidez, idade gestacional e histórico reprodutivo, funcionam como elementos complementares de identificação, pois descrevem o mesmo contexto gestacional e perinatal registrado em ambas as fontes.

A concordância simultânea em múltiplas variáveis dessas dimensões aumenta substancialmente a evidência de que os registros comparados se referem ao mesmo nascimento/óbito infantil, enquanto discordâncias sucessivas reduzem essa plausibilidade. Assim, a utilização conjunta dessas 16 variáveis permitiu construir um critério de similaridade suficientemente informativo para distinguir correspondências verdadeiras de coincidências ocasionais.

4.3.2 Estratégia de *Blocking*

Para reduzir o número de comparações e aumentar a eficiência computacional, foi aplicada a técnica de *blocking*, restringindo as comparações aos registros com concordância exata nas variáveis sexo do recém-nascido (`SEXO`), data de nascimento (`DTNASC`), peso ao nascer (`PESO`) e município de nascimento (`MUNNASC`).

A escolha dessas variáveis para o bloqueio baseou-se em sua alta disponibilidade, consistência entre os sistemas e capacidade discriminatória quando utilizadas em conjunto. Sexo e data de nascimento correspondem a atributos básicos e centrais do registro do recém-nascido, enquanto município de nascimento incorpora a dimensão espacial do evento. O peso ao nascer, por sua vez, acrescenta poder discriminatório adicional, especialmente em situações em que múltiplos nascimentos compartilham a mesma data e o mesmo município. Assim, a concordância simultânea nessas quatro variáveis reduz substancialmente a probabilidade de formação de pares aleatórios, tornando o processo computacionalmente viável sem, naquela primeira rodada, comprometer a especificidade do pareamento. Apenas registros que apresentavam concordância simultânea nessas variáveis foram considerados candidatos à comparação detalhada.

Em termos metodológicos, o *blocking* foi definido de forma intencionalmente conservadora na primeira rodada, priorizando a comparação entre registros com alta plausibilidade inicial de correspondência. Essa estratégia permitiu concentrar a etapa detalhada de cálculo de similaridade em pares mais promissores, reduzindo o volume de combinações possíveis entre as bases.

4.3.3 Estratégia de Flexibilização Progressiva

O pareamento foi realizado de forma iterativa, iniciando-se com a exigência de concordância em todas as 16 variáveis e, progressivamente, reduzindo-se o número mínimo de variáveis concordantes até o limite mínimo de quatro variáveis. Em cada iteração, foram selecionados apenas pares com número de concordâncias maior ou igual ao limiar definido; foi garantida a restrição de pareamento um-para-um (1:1); e registros já vinculados foram removidos das iterações subsequentes. Pares ambíguos, com múltiplas possibilidades de correspondência, foram descartados na respectiva etapa.

O procedimento iniciou-se com a exigência de concordância no conjunto completo das 16 variáveis e, a cada nova iteração, o número mínimo de variáveis concordantes exigido para aceitação do par foi progressivamente reduzido em uma unidade, até o limite inferior de quatro variáveis. Dessa forma, foram realizadas 13 iterações sucessivas, correspondentes aos limiares de 16, 15, 14, até 4 variáveis concordantes.

Para a variável idade materna (`IDADEMAE`), adotou-se critério de concordância com tolerância de até um ano de diferença entre os registros, considerando a possibilidade de pequena discrepância de preenchimento entre as bases. Adicionalmente, foi aplicada regra de coerência temporal, de modo que a idade registrada no SIM não pudesse ser inferior à idade registrada no SINASC, nem superior em mais de um ano, assumindo-se que, nos casos de óbito ocorrido no ano subsequente ao nascimento, a idade da mãe poderia estar acrescida de um ano no registro do SIM.

4.3.4 Resultados da Primeira Rodada

A aplicação da estratégia progressiva resultou na identificação de 25.616 pares únicos entre as bases do SIM e do SINASC. A distribuição dos novos pares por limiar mínimo de variáveis concordantes é apresentada na Tabela 4.1.

Tabela 4.1: Distribuição dos pares identificados na primeira rodada por limiar de variáveis concordantes.

Limiar (variáveis concordantes)	**Novos pares**	**Percentual (%)**
16	2.670	10,42
15	4.805	18,76
14	4.245	16,57
13	3.780	14,76
12	3.156	12,32
11	2.472	9,65
10	2.057	8,03
9	1.288	5,03
8	682	2,66
7	278	1,09
6	102	0,40
5	62	0,24
4	19	0,07
Total	25.616	100

Conforme a Tabela 4.1, observa-se que o maior volume de pareamentos ocorreu nos limiares intermediários (entre 13 e 15 variáveis concordantes), indicando elevada consistência entre os registros das duas bases. À medida que o critério de aceitação foi flexibilizado, verificou-se redução progressiva no número de novos pares identificados. Abaixo de sete variáveis concordantes, o número de novos pareamentos tornou-se residual, sugerindo que a maior parte das correspondências plausíveis foi capturada em níveis mais rigorosos.

Considerando o total de 36.594 óbitos infantis registrados no SIM, a primeira rodada permitiu vincular aproximadamente 70% dos registros (25.616/36.594), evidenciando elevada capacidade de recuperação mesmo sob estratégia inicialmente conservadora. A verificação final confirmou a inexistência de duplicidades, assegurando a integridade do pareamento um-para-um (1:1).

4.4 Segunda Rodada de *Linkage* — Estratégia Flexibilizada

Após a conclusão da primeira rodada de pareamento, os registros remanescentes das bases do SIM e do SINASC foram submetidos a uma segunda etapa de *linkage* com critérios mais flexíveis, visando recuperar possíveis correspondências não identificadas anteriormente. Essa etapa foi aplicada exclusivamente aos registros ainda não vinculados, preservando os pareamentos já estabelecidos e mantendo a restrição de vinculação um-para-um (1:1).

4.4.1 Ajustes na Estratégia de *Blocking*

Na primeira rodada, o bloqueio foi realizado com base nas variáveis sexo (`SEXO`), data de nascimento (`DTNASC`), peso ao nascer (`PESO`) e município de nascimento (`MUNNASC`). Considerando a possibilidade de erros de digitação ou inconsistências no registro do peso ao nascer, essa variável foi removida do critério de bloqueio na segunda rodada. O novo bloqueio passou a considerar apenas sexo do recém-nascido, data de nascimento e município de nascimento. Essa modificação ampliou o número de combinações candidatas, resultando em 385.805 pares candidatos, aumentando a sensibilidade do procedimento.

4.4.2 Flexibilização dos Critérios de Similaridade

Além da alteração no *blocking*, foram introduzidas flexibilizações nos critérios de similaridade. A idade gestacional (`SEMAGESTAC`) passou a admitir variação de até duas semanas. A idade materna (`IDADEMAE`) manteve o critério de diferença máxima de um ano, com verificação adicional de coerência temporal. O peso ao nascer (`PESO`) passou a ser comparado por similaridade textual, com base na distância de Levenshtein, adotando-se limiar de 0,8.

Embora o peso ao nascer seja uma variável numérica, nessa etapa ele foi tratado como texto para permitir a identificação de divergências compatíveis com erros de digitação ou formatação, como omissão, inserção, substituição ou transposição de dígitos. Esse procedimento foi adotado porque discrepâncias desse tipo podem impedir a recuperação de pares verdadeiros quando se utiliza apenas comparação exata. O limiar de 0,8 foi esco-

lhido por representar alta similaridade textual, permitindo tolerância a pequenas variações na grafia do valor registrado, sem admitir diferenças mais amplas que comprometessem a plausibilidade do pareamento. As demais variáveis continuaram sendo comparadas por igualdade exata.

Para tornar concreto o efeito desse limiar, convém observar como a distância de Levenshtein se comporta sobre o peso ao nascer, tipicamente registrado como uma sequência de três a quatro dígitos (por exemplo, "3200"). Erros usuais de digitação correspondem a uma única operação de edição — a substituição de um dígito ("3200" → "3210"), a omissão ou inserção de um dígito ("3200" → "320") ou a transposição de dígitos adjacentes ("3200" → "3020"). Em um valor de quatro dígitos, uma única substituição resulta em similaridade de 0,75, ao passo que a omissão de um dígito eleva essa medida para aproximadamente 0,86, por reduzir o comprimento total comparado. Com o limiar de 0,8 adotado nesta rodada, admitem-se, portanto, divergências compatíveis com a omissão ou a inserção de um dígito, mas rejeitam-se substituições isoladas, que ficam abaixo do corte. Trata-se, assim, de uma flexibilização deliberadamente contida: tolera-se apenas o tipo de erro de digitação mais frequente, preservando-se elevada especificidade na comparação do peso.

4.4.3 Estratégia Iterativa

Mantendo a estrutura metodológica descrita anteriormente, o pareamento na segunda rodada também foi conduzido de forma iterativa. Considerando as 16 variáveis comparáveis utilizadas nesta etapa, o procedimento iniciou-se com exigência de concordância no conjunto completo dessas variáveis e, a cada iteração, o número mínimo de variáveis concordantes foi progressivamente reduzido até o limite inferior de três variáveis.

A definição do limiar mínimo em três variáveis esteve diretamente relacionada ao critério de *blocking* adotado nesta rodada, baseado em sexo, data de nascimento e município de nascimento. Como essas três variáveis já eram necessariamente concordantes entre os pares candidatos, o limite mínimo de três concordâncias permitiu a avaliação de correspondências sustentadas apenas pelas variáveis estruturantes do bloqueio, caracterizando esta etapa como de maior sensibilidade em comparação à primeira rodada. Essa

configuração mais flexível teve como objetivo ampliar a capacidade de recuperação de pares remanescentes que apresentavam inconsistências nas demais variáveis, sem alterar a coerência estrutural garantida pelo *blocking*.

4.4.4 Resultados da Segunda Rodada

A segunda rodada resultou na identificação de 4.268 novos pares, elevando o total acumulado de registros vinculados para 29.884 pares únicos. A distribuição dos novos pares por limiar mínimo de variáveis concordantes é apresentada na Tabela 4.2.

Tabela 4.2: Distribuição dos pares identificados na segunda rodada por limiar de variáveis concordantes.

Limiar (variáveis concordantes)	**Novos pares**	**Percentual (%)**
16	106	2,48
15	278	6,51
14	381	8,93
13	543	12,72
12	564	13,21
11	543	12,72
10	470	11,01
9	372	8,72
8	224	5,25
7	156	3,66
6	120	2,81
5	90	2,11
4	318	7,45
3	103	2,41
Total	4.268	100

Conforme a Tabela 4.2, observa-se que, diferentemente da primeira rodada, caracterizada por forte concentração nos limiares mais elevados, a segunda rodada apresentou

maior dispersão ao longo dos níveis intermediários de concordância, com destaque para os limiares de 11 a 13 variáveis. O maior número de novos pareamentos ocorreu no limiar de 12 variáveis (564 pares), indicando que parte significativa dos registros remanescentes apresentava inconsistências pontuais que impediram sua identificação sob critérios mais restritivos. O incremento relativo foi de aproximadamente 16,7% em relação ao total obtido na primeira rodada (4.268 novos pares sobre 25.616 previamente identificados), evidenciando a relevância da estratégia de flexibilização progressiva.

4.5 Terceira Rodada de *Linkage* — Estratégia de Alta Sensibilidade

Após a aplicação das duas primeiras rodadas de pareamento, os registros remanescentes das bases do SIM e do SINASC foram submetidos a uma terceira etapa de *linkage* com critérios ainda mais flexíveis, com o objetivo de recuperar possíveis correspondências residuais. Assim como na rodada anterior, manteve-se o bloqueio pelas variáveis sexo do recém-nascido (`SEXO`), data de nascimento (`DTNASC`) e município de nascimento (`MUNNASC`). Essa estratégia garantiu a preservação de uma estrutura mínima de coerência entre os registros comparados.

4.5.1 Flexibilização Adicional dos Critérios

Nesta etapa, foi adotada flexibilização adicional no critério de similaridade do peso ao nascer (`PESO`), utilizando a distância de Levenshtein com limiar de 0,7. A redução do limiar, em comparação à rodada anterior, teve como objetivo ampliar a sensibilidade do procedimento para capturar divergências textuais mais pronunciadas, ainda compatíveis com erros de digitação, como transposição, omissão ou inserção de dígitos. O peso ao nascer foi a variável selecionada para essa flexibilização adicional por apresentar maior suscetibilidade a inconsistências de preenchimento numérico, especialmente em registros digitados manualmente.

A redução do limiar de 0,8 para 0,7 amplia o conjunto de divergências toleradas na comparação do peso. Enquanto o limiar de 0,8 rejeitava a substituição isolada de um dígito

(similaridade de 0,75), o limiar de 0,7 passa a aceitá-la em valores de quatro dígitos, bem como a transposição de dígitos adjacentes, ainda coerentes com erros de preenchimento. Essa flexibilização, contudo, não é irrestrita: uma segunda divergência independente no mesmo valor faz a similaridade cair abaixo do corte, de modo que o limiar de 0,7 continua exigindo que o peso registrado nas duas bases seja globalmente muito próximo. Cabe observar, ademais, que o efeito do limiar depende do comprimento do valor comparado: em pesos de apenas três dígitos, uma única substituição produz similaridade de cerca de 0,67, permanecendo abaixo do corte mesmo nesta rodada. Dessa forma, a flexibilização atua de maneira mais permissiva sobre os pesos mais altos, de quatro dígitos, exatamente aqueles em que a comparação textual dispõe de mais caracteres para distinguir uma pequena divergência de uma discordância real.

A idade materna manteve o critério de diferença máxima de um ano, com verificação de coerência temporal, por já representar uma margem de tolerância considerada adequada para acomodar pequenas discrepâncias de registro. As demais variáveis continuaram sendo comparadas por igualdade exata, uma vez que são predominantemente categóricas ou codificadas, de baixa cardinalidade e maior padronização, de modo que flexibilizações adicionais poderiam comprometer a especificidade do pareamento.

O número de pares candidatos gerados nesta rodada foi de 229.205, valor inferior ao observado na segunda rodada, refletindo a redução progressiva do universo de registros remanescentes. Foi estabelecido como critério mínimo a concordância em pelo menos cinco variáveis. Considerando que três variáveis já compunham o *blocking* (sexo, data de nascimento e município de nascimento), a exigência de cinco concordâncias assegurou que, além das variáveis estruturantes do bloqueio, ao menos duas variáveis adicionais apresentassem correspondência, evitando a aceitação de pareamentos excessivamente frágeis nesta etapa de maior sensibilidade.

4.5.2 Resultados da Terceira Rodada

A terceira rodada resultou na identificação de 336 novos pares, distribuídos conforme o limiar mínimo de variáveis concordantes, apresentado na Tabela 4.3.

Tabela 4.3: Distribuição dos pares identificados na terceira rodada por limiar de variáveis concordantes.

Limiar (variáveis concordantes)	**Novos pares**
15	3
14	10
13	21
12	33
11	61
10	72
9	49
8	45
7	26
6	10
5	6
Total	336

Conforme a Tabela 4.3, observa-se que a maior concentração de novos pareamentos ocorreu nos limiares intermediários (10 e 11 variáveis), com 72 e 61 pares, respectivamente. O incremento obtido nesta etapa foi substancialmente inferior ao observado nas rodadas anteriores. A terceira rodada adicionou 336 pares, representando aproximadamente 7,9% em relação à segunda rodada (336/4.268) e aproximadamente 1,1% em relação ao total acumulado ao final da segunda rodada (336/29.884). Ao término do processo, o total acumulado de registros vinculados atingiu 30.220 pares únicos. A verificação final confirmou a inexistência de duplicidades tanto no SIM quanto no SINASC, garantindo a integridade do pareamento um-para-um (1:1).

4.6 Verificação de Unicidade

A checagem final confirmou que cada registro do SIM foi vinculado a, no máximo, um registro do SINASC e vice-versa, não havendo ocorrências de múltiplas correspondências

para um mesmo identificador em nenhuma das bases. Dessa forma, assegurou-se a integridade do pareamento um-para-um (1:1), preservando a correspondência exclusiva entre os registros vinculados.

4.7 Consolidação Geral do Processo de *Linkage*

Considerando as três rodadas realizadas, a primeira rodada identificou 25.616 pares, a segunda adicionou 4.268 pares e a terceira, 336 pares, totalizando 30.220 pares únicos. Tendo como referência os 36.594 óbitos infantis inicialmente presentes na base do SIM, o procedimento de *linkage* permitiu vincular aproximadamente 82,6% dos registros de óbito infantil (30.220/36.594) à base do SINASC.

A análise do ganho incremental ao longo das três rodadas evidencia o comportamento esperado de um procedimento hierárquico de *linkage* probabilístico: a primeira rodada capturou a maior parte dos pares com alto grau de concordância; a segunda recuperou registros com inconsistências pontuais relevantes; e a terceira apresentou ganho marginal reduzido, indicando que os registros remanescentes possuíam maior grau de inconsistência ou ausência de informação compatível entre as bases. O ganho decrescente entre as rodadas sugere que a estratégia progressiva adotada foi adequada e eficiente, maximizando a recuperação de pares plausíveis sem comprometer o controle estrutural do pareamento.

4.8 Limitações do Processo de *Linkage*

Uma limitação relevante do procedimento adotado decorre do fato de o recorte regional ter sido aplicado a ambas as bases *antes* da etapa de *linkage*. Como tanto o SIM quanto o SINASC foram previamente restritos à região Sudeste, o pareamento só é capaz de recuperar pares em que o nascimento e o óbito foram registrados dentro dessa mesma região. Casos em que a criança nasceu no Sudeste, mas cujo óbito foi registrado em outra região do país — ou o inverso — encontram-se, por construção, fora do universo de comparação, não podendo ser recuperados. Dessa forma, parte dos óbitos infantis não vinculados pode não decorrer de falhas do processo de similaridade, mas sim da mobilidade

geográfica das famílias entre o nascimento e o óbito.

Para dimensionar a ordem de grandeza desse fenômeno, dados do Censo Demográfico de 2022 indicam que aproximadamente 19,2 milhões de brasileiros residiam em região distinta daquela em que nasceram, o que evidencia a existência de fluxos migratórios inter-regionais relevantes (Instituto Brasileiro de Geografia e Estatística, 2023). Cabe destacar, contudo, que esse valor reflete a migração acumulada ao longo de toda a vida dos indivíduos, enquanto o *ruído* efetivamente relevante para este trabalho restringe-se à mobilidade ocorrida no primeiro ano de vida da criança, isto é, no intervalo entre o nascimento e o eventual óbito infantil. Essa fração é substancialmente menor, uma vez que a migração de famílias com recém-nascidos em um intervalo tão curto tende a ser pouco frequente. Ainda assim, reconhece-se que uma parcela residual dos pares não recuperados pode estar associada a esse fenômeno, constituindo uma limitação inerente ao recorte regional adotado.

5 Processo Metodológico

Concluído o processo de *linkage* probabilístico descrito no capítulo anterior, obteve-se uma base integrada SIM-SINASC sobre a qual se desenvolveu toda a etapa de modelagem. Este capítulo descreve o processo metodológico comum aos três classificadores avaliados — *Random Forest*, Regressão Logística e XGBoost —, contemplando o tratamento das variáveis explicativas, a divisão dos dados em treino, validação e teste, o *pipeline* de pré-processamento e o procedimento geral de ajuste de hiperparâmetros. As particularidades de cada modelo, bem como os resultados obtidos, são apresentados nos capítulos seguintes. A Figura 5.1 resume as etapas do processo, desde a integração das bases até a avaliação dos modelos e a análise dos fatores associados ao óbito infantil.

5.1 Visão Geral do Fluxo

O processo conduzido neste trabalho pode ser sintetizado em uma sequência de etapas encadeadas. A partir das bases originais do SIM e do SINASC, realizou-se o *linkage* probabilístico, do qual resultou a base integrada. Sobre essa base, cada variável explicativa foi tratada quanto a valores inválidos e ausentes, definindo-se a respectiva estratégia de imputação. Em seguida, os dados foram divididos em conjuntos de treino, validação e teste e submetidos a um *pipeline* de pré-processamento, ajustado exclusivamente sobre o conjunto de treino para evitar o vazamento de informação. Por fim, os três modelos foram ajustados e avaliados, e suas medidas de importância das variáveis foram analisadas. A Figura 5.1 ilustra esse encadeamento.

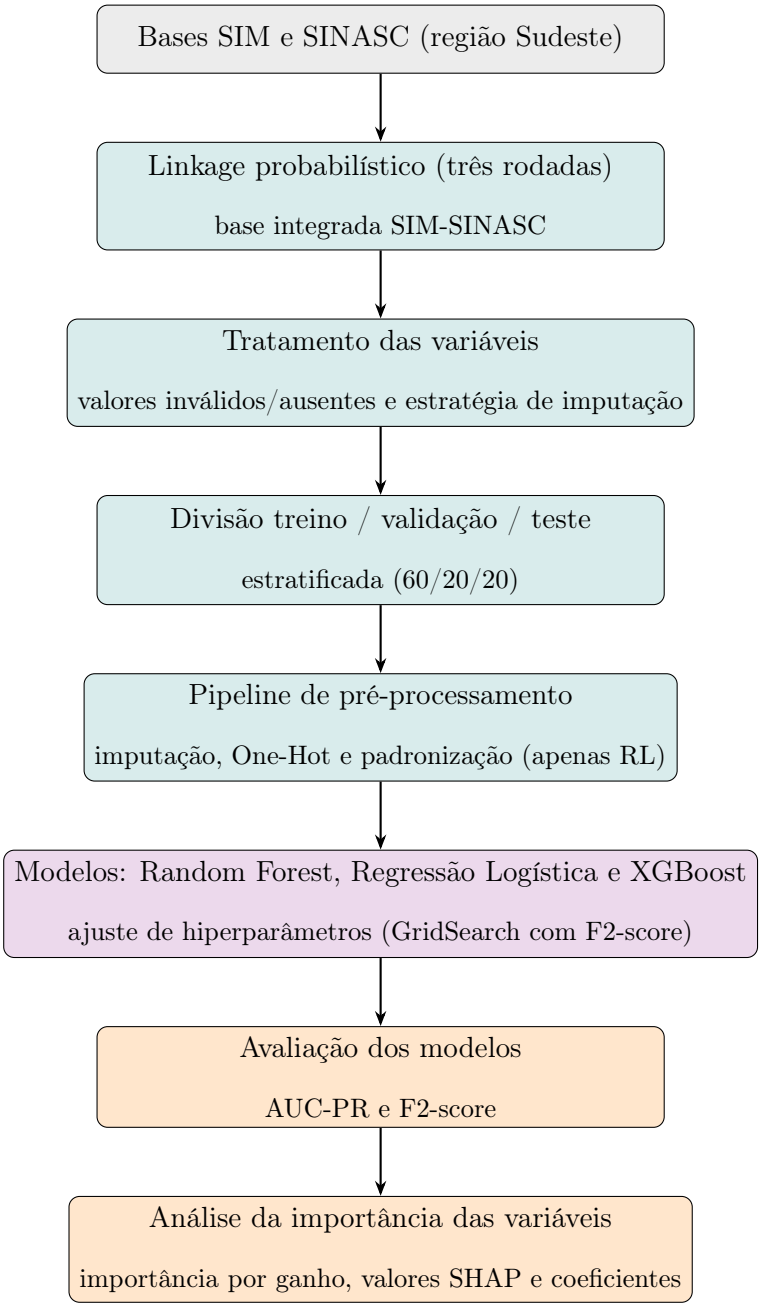

Figura 5.1: Visão geral do processo metodológico, da integração das bases à avaliação dos modelos e à análise das variáveis.

5.2 Tratamento das Variáveis

Nesta seção, descreve-se o tratamento aplicado a cada variável explicativa da base integrada SIM-SINASC, obtida ao final do processo de *linkage* probabilístico descrito anteriormente, contemplando a identificação de valores inválidos ou ausentes e a definição da estratégia de imputação adotada. Cabe esclarecer que, nesta etapa, as variáveis são

apenas tratadas: valores inválidos e códigos de "ignorado" são convertidos em ausentes, e define-se a estratégia de imputação mais adequada a cada caso. A imputação propriamente dita, contudo, não é aplicada aqui, mas sim posteriormente, dentro do *pipeline* de modelagem, ajustado exclusivamente com o conjunto de treino, de modo a evitar o vazamento de informação para os conjuntos de validação e teste.

De modo geral, optou-se pela mediana como estratégia de imputação para as variáveis numéricas, por ser robusta à presença de valores extremos, e pela moda (valor mais frequente) para as variáveis ordinais. Eventuais exceções a essa regra são justificadas individualmente ao longo da seção.

5.2.1 Peso ao nascer (PESO)

A variável `PESO` corresponde ao peso do recém-nascido ao nascer, registrado em gramas. Trata-se de uma variável numérica contínua e de um dos indicadores mais relevantes para a mortalidade infantil, uma vez que o baixo peso ao nascer está associado a maior risco de óbito neonatal.

Tabela 5.1: Descrição e tratamento da variável PESO.

Característica	**Descrição**
Nome original	`PESO`
Tipo	Numérica contínua
Unidade	Gramas
Faixa válida	100 a 7.000 g
Valores ausentes	106 (0,00%)
Estratégia de imputação	Mediana (3.200 g)

Conforme apresentado na Tabela 5.1, a base já se encontrava restrita ao intervalo de 100 a 7.000 gramas, não apresentando valores fora dessa faixa. Foram identificados apenas 106 registros com valor ausente (0,00% do total), correspondentes a campos genuinamente vazios, sem códigos de ignorado ou valores inválidos mascarados.

Conforme ilustra a Figura 5.2, em razão da leve assimetria da distribuição, decorrente da cauda de valores baixos associada a nascimentos prematuros, definiu-se a

mediana como estratégia de imputação dos valores ausentes, por ser uma medida robusta à presença de valores extremos. Vale ressaltar que, nesta etapa, a variável foi apenas tratada e os valores inválidos ou ausentes foram identificados; a imputação propriamente dita não é aplicada aqui, mas sim posteriormente, dentro do *pipeline* de modelagem descrito na Seção 2.6, ajustado exclusivamente com o conjunto de treino, de modo a evitar o vazamento de informação para os conjuntos de validação e teste.

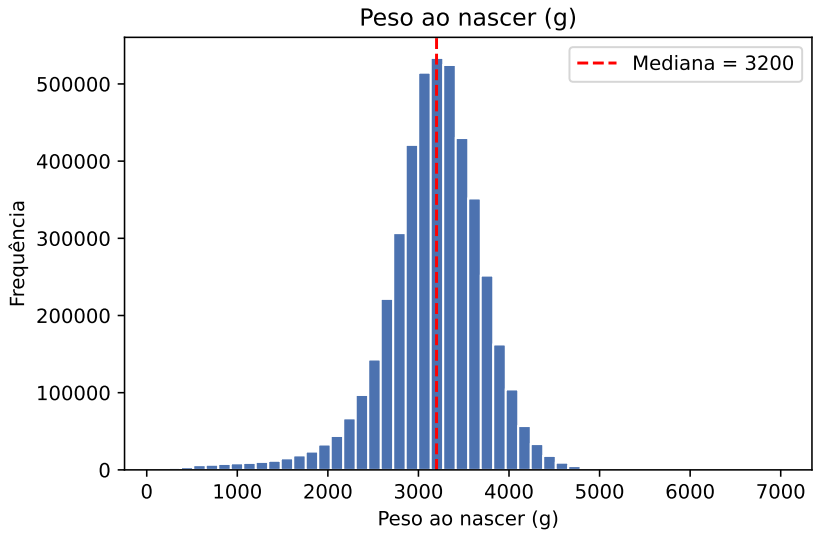

Figura 5.2: Distribuição do peso ao nascer, com a linha tracejada indicando a mediana utilizada na imputação. Nota-se a leve assimetria à esquerda, decorrente da cauda de valores baixos associada a nascimentos prematuros.

5.2.2 Idade da mãe (IDADEMAE)

A variável `IDADEMAE` corresponde à idade da mãe, em anos completos, no momento do nascimento. É uma variável numérica discreta, relevante para o estudo da mortalidade infantil, uma vez que tanto a maternidade muito precoce quanto a tardia estão associadas a riscos gestacionais.

Tabela 5.2: Descrição e tratamento da variável IDADEMAE.

Característica	**Descrição**
Nome original	`IDADEMAE`
Tipo	Numérica discreta
Unidade	Anos
Faixa observada	10 a 64 anos
Código tratado como ausente	99 (ignorado)
Valores tornados ausentes	63
Estratégia de imputação	Mediana (28 anos)

Conforme apresentado na Tabela 5.2, a variável não apresentava valores genuinamente vazios. Foram identificados 63 registros com o valor 99, que corresponde ao código de "ignorado" do SINASC e não a uma idade real, conforme evidenciado pela ausência de qualquer registro entre 65 e 98 anos. Esses valores foram convertidos em ausentes. Os demais valores extremos, tanto os mais baixos (a partir de 10 anos) quanto os mais altos (até 64 anos), foram mantidos, por corresponderem a idades biologicamente plausíveis e epidemiologicamente relevantes.

Dada a baixa assimetria da distribuição, com média (27,8 anos) e mediana (28 anos) praticamente coincidentes, definiu-se a mediana como estratégia de imputação, mantendo a coerência com o tratamento das demais variáveis numéricas. Assim como descrito anteriormente, a imputação não é aplicada nesta etapa, mas posteriormente, dentro do *pipeline* de modelagem, ajustado exclusivamente com o conjunto de treino.

5.2.3 Semanas de gestação (SEMAGESTAC)

A variável `SEMAGESTAC` corresponde à duração da gestação, em semanas completas. É uma variável numérica discreta e um importante preditor de mortalidade infantil, uma vez que a prematuridade está fortemente associada a maior risco de óbito neonatal.

Tabela 5.3: Descrição e tratamento da variável SEMAGESTAC.

Característica	**Descrição**
Nome original	`SEMAGESTAC`
Tipo	Numérica discreta
Unidade	Semanas
Faixa válida	19 a 45 semanas
Total de registros	4.456.504
Valores ausentes	26.640 (0,60%)
Estratégia de imputação	Mediana (39 semanas)

Conforme apresentado na Tabela 5.3, a base já se encontrava restrita ao intervalo de 19 a 45 semanas, não apresentando valores fora dessa faixa nem códigos de ignorado. Foram identificados 26.640 registros com valor ausente (0,60% do total), correspondentes a campos genuinamente vazios.

Conforme ilustra a Figura 5.3, a distribuição da variável é fortemente assimétrica e concentrada em torno do parto a termo (entre 38 e 40 semanas), com uma cauda de menor frequência associada a nascimentos prematuros. Em razão dessa assimetria, definiu-se a mediana (39 semanas) como estratégia de imputação, por ser mais representativa do valor típico e robusta à influência dos valores extremos.

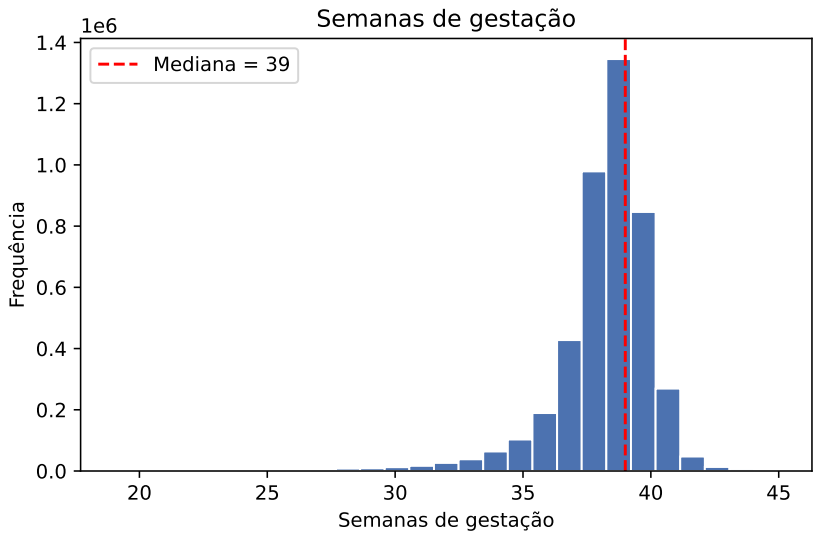

Figura 5.3: Distribuição da duração da gestação em semanas, com a linha tracejada indicando a mediana. Observa-se forte concentração em torno do parto a termo (38 a 40 semanas) e uma cauda associada aos nascimentos prematuros.

5.2.4 Apgar no primeiro minuto (APGAR1)

A variável `APGAR1` corresponde à pontuação do índice de Apgar avaliado no primeiro minuto de vida, que varia de 0 a 10 e mede as condições de vitalidade do recém-nascido. Valores baixos indicam sofrimento ao nascer e estão associados a maior risco de óbito.

Tabela 5.4: Descrição e tratamento da variável APGAR1.

Característica	**Descrição**
Nome original	`APGAR1`
Tipo	Numérica discreta (escala 0–10)
Faixa válida	0 a 10
Código tratado como ausente	99 (ignorado)
Total de registros	4.456.504
Valores vazios na origem	40.410
Valores com código 99	7.171
Total de ausentes	47.581 (1,07%)
Estratégia de imputação	Mediana (9)

Conforme apresentado na Tabela 5.4, a variável apresentava 40.410 registros vazios e 7.171 registros com o valor 99, que corresponde ao código de "ignorado" do SINASC, uma vez que a escala do Apgar varia apenas de 0 a 10. Ambos foram convertidos em ausentes, totalizando 47.581 registros (1,07%). A distribuição é fortemente assimétrica e concentrada nos valores altos (8 e 9), com uma cauda de menor frequência associada a recém-nascidos com baixa vitalidade. Em razão dessa assimetria, definiu-se a mediana (9) como estratégia de imputação.

A Figura 5.4 ilustra essa distribuição.

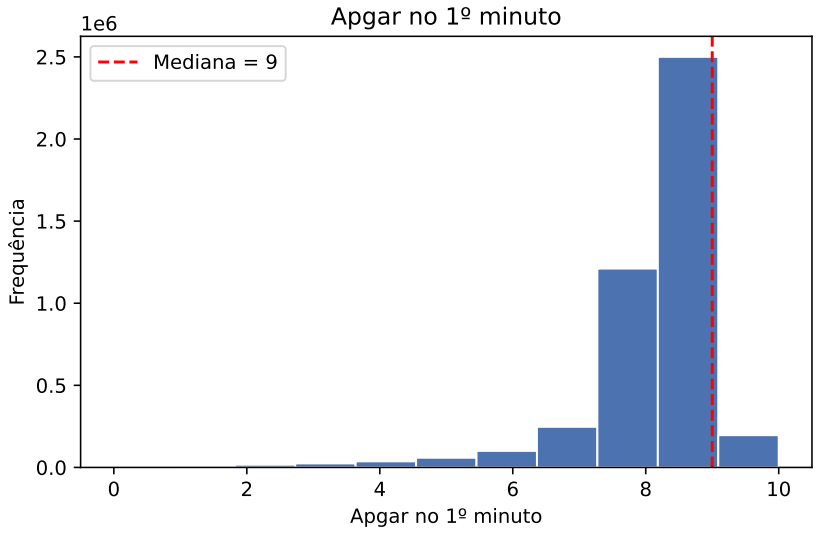

Figura 5.4: Distribuição do índice de Apgar no primeiro minuto, com a linha tracejada indicando a mediana. A distribuição é assimétrica e concentrada nos valores altos (8 e 9), com uma cauda associada a recém-nascidos de baixa vitalidade.

5.2.5 Apgar no quinto minuto (APGAR5)

A variável `APGAR5` corresponde à pontuação do índice de Apgar avaliado no quinto minuto de vida, na mesma escala de 0 a 10. Por ser medido após os primeiros minutos de adaptação do recém-nascido, é considerado um indicador ainda mais relevante de prognóstico do que o Apgar no primeiro minuto.

Tabela 5.5: Descrição e tratamento da variável APGAR5.

Característica	**Descrição**
Nome original	`APGAR5`
Tipo	Numérica discreta (escala 0–10)
Faixa válida	0 a 10
Código tratado como ausente	99 (ignorado)
Total de registros	4.456.504
Valores vazios na origem	38.830
Valores com código 99	5.257
Total de ausentes	44.087 (0,99%)
Estratégia de imputação	Mediana (9)

Conforme apresentado na Tabela 5.5, a variável apresentava 38.830 registros vazios e 5.257 registros com o valor 99, código de "ignorado" do SINASC, que foram convertidos em ausentes, totalizando 44.087 registros (0,99%). A distribuição é ainda mais concentrada nos valores altos do que a do Apgar no primeiro minuto, com a maior parte dos registros nas pontuações 9 e 10, o que é coerente com a recuperação da vitalidade do recém-nascido após os primeiros minutos de vida. Em razão da forte assimetria, definiu-se a mediana (9) como estratégia de imputação.

A Figura 5.5 apresenta essa distribuição.

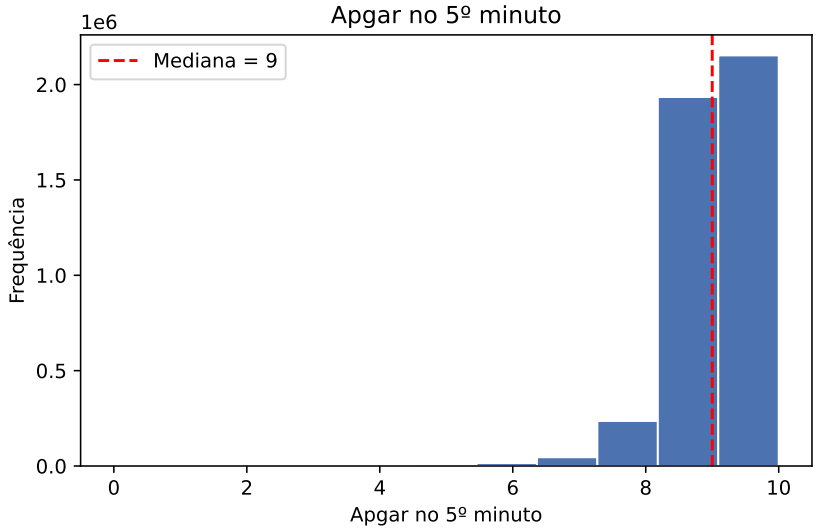

Figura 5.5: Distribuição do índice de Apgar no quinto minuto, com a linha tracejada indicando a mediana. A concentração nos valores altos (9 e 10) é ainda mais acentuada que no primeiro minuto, coerente com a recuperação da vitalidade do recém-nascido.

5.2.6 Escolaridade da mãe (ESCMAE)

A variável `ESCMAE` representa a escolaridade da mãe, em anos de estudo concluídos, codificada de forma ordinal. É uma variável socioeconômica relevante, uma vez que a escolaridade materna está associada ao acesso a cuidados de saúde e a desfechos perinatais.

Tabela 5.6: Códigos e tratamento da variável ESCMAE.

Código	**Significado**	**Registros**
1	Nenhuma escolaridade	5.545
2	1 a 3 anos de estudo	36.856
3	4 a 7 anos de estudo	459.301
4	8 a 11 anos de estudo	2.848.558
5	12 anos ou mais	1.079.886
–	Ausente (vazio)	26.358

Conforme apresentado na Tabela 5.6, a variável apresentava apenas os códigos válidos de 1 a 5, sem ocorrência do código 9 (“ignorado”) nem de valores fora da faixa esperada. Foram identificados 26.358 registros com valor ausente (0,59%), corresponden-

tes a campos vazios. A distribuição concentra-se nos níveis de maior escolaridade, com predominância do código 4 (8 a 11 anos de estudo). Por se tratar de uma variável ordinal, definiu-se a moda (código 4) como estratégia de imputação dos valores ausentes.

5.2.7 Sexo do recém-nascido (SEXO)

A variável `SEXO` indica o sexo do recém-nascido. É uma variável categórica nominal, codificada na base como 1 (masculino) e 2 (feminino). O valor 0 corresponde aos casos de sexo ignorado, registrados pelo SINASC apenas em situações de genitália indefinida.

Tabela 5.7: Códigos e tratamento da variável SEXO.

Código original	**Categoria**	**Registros**
1	Masculino	2.279.861
2	Feminino	2.175.978
0	Ignorado	665

Conforme apresentado na Tabela 5.7, a variável não apresentava valores vazios. Os códigos 1 e 2 foram convertidos, respectivamente, em "MASCULINO" e "FEMININO". Os 665 registros com código 0 (0,01%), correspondentes a sexo indefinido, foram mantidos como uma categoria própria, denominada "IGNORADO", em vez de serem imputados. Optou-se por essa abordagem por se tratar de uma quantidade reduzida de registros e por ser cientificamente mais adequado preservar a informação de indefinição do que atribuir artificialmente um dos sexos. Por ser uma variável nominal, foi posteriormente submetida ao *One-Hot Encoding* no *pipeline* de modelagem.

5.2.8 Tipo de parto (PARTO)

A variável `PARTO` indica o tipo de parto, classificado como vaginal ou cesáreo. É uma variável categórica nominal, relevante por estar associada a condições gestacionais e a indicações clínicas que podem influenciar o desfecho neonatal.

Tabela 5.8: Códigos e tratamento da variável PARTO.

Código original	Categoria	Registros
1	Vaginal	1.839.401
2	Cesáreo	2.615.180
–	Ignorado (ausente)	1.923

Conforme apresentado na Tabela 5.8, a variável apresentava apenas os códigos válidos 1 (vaginal) e 2 (cesáreo), sem ocorrência do código 9 ("ignorado"). Foram identificados 1.923 registros vazios (0,04%), que foram mantidos como uma categoria própria, denominada "IGNORADO", em vez de imputados, dada a quantidade reduzida e a opção por preservar a informação de ausência. Por se tratar de variável nominal, foi posteriormente submetida ao *One-Hot Encoding* no *pipeline* de modelagem.

5.2.9 Raça/cor (RACACOR)

A variável `RACACOR` indica a raça/cor do recém-nascido, informada pela mãe no momento do registro. É uma variável categórica nominal, codificada de 1 a 5, relevante por permitir analisar eventuais desigualdades raciais associadas à mortalidade infantil.

Tabela 5.9: Códigos e tratamento da variável RACACOR.

Código original	Categoria	Registros
1	Branca	1.976.792
2	Preta	361.654
3	Amarela	24.830
4	Parda	2.034.125
5	Indígena	6.070
–	Ignorado (ausente)	53.033

Conforme apresentado na Tabela 5.9, a distribuição concentra-se nas categorias branca e parda, que juntas representam a maior parte dos registros, refletindo a composição racial da população da região Sudeste. Diferentemente de outras variáveis, o formu-

lário do SINASC não prevê o código "ignorado" para a raça/cor, de modo que os 53.033 registros ausentes (1,19%) correspondem a campos não preenchidos. Esses registros foram mantidos como uma categoria própria, denominada "IGNORADO", preservando a informação de ausência. Por ser variável nominal, foi posteriormente submetida ao *One-Hot Encoding* no *pipeline* de modelagem.

5.2.10 Unidade da federação (UF)

A variável `UF` indica a unidade da federação de nascimento. É uma variável categórica nominal. Como a base utilizada restringe-se à região Sudeste, a variável contém apenas os quatro estados que a compõem: São Paulo (SP), Minas Gerais (MG), Rio de Janeiro (RJ) e Espírito Santo (ES).

Tabela 5.10: Distribuição da variável UF.

Estado	**Registros**
São Paulo (SP)	2.359.059
Minas Gerais (MG)	1.026.869
Rio de Janeiro (RJ)	850.828
Espírito Santo (ES)	219.748

Conforme apresentado na Tabela 5.10, a variável não apresentava valores ausentes nem códigos inválidos, contendo apenas as quatro siglas esperadas para a região Sudeste, com predominância do estado de São Paulo. Realizou-se apenas uma padronização preventiva do texto (conversão para maiúsculas e remoção de espaços). Por ser variável nominal, foi posteriormente submetida ao *One-Hot Encoding* no *pipeline* de modelagem.

5.2.11 Idade do pai (IDADEPAI)

A variável `IDADEPAI` corresponde à idade do pai, em anos, no momento do nascimento. Diferentemente das demais variáveis, apresentou uma proporção muito elevada de ausência: 2.271.371 registros (50,97%) não possuíam essa informação.

Tabela 5.11: Descrição e tratamento da variável IDADEPAI.

Característica	**Descrição**
Nome original	`IDADEPAI`
Tipo original	Numérica discreta
Total de registros	4.456.504
Valores ausentes	2.271.371 (50,97%)
Valores informados	2.185.133 (49,03%)
Tratamento adotado	Conversão em indicador binário
Variável resultante	`PAI_INFORMADO` (SIM / NÃO)

Conforme mostra a Tabela 5.11, dos 4.456.504 registros, 2.271.371 (50,97%) não apresentavam a idade do pai. Diante disso, a imputação dos valores ausentes, por mediana ou outra medida, foi considerada inadequada, pois implicaria atribuir um valor estimado a cerca de metade de toda a base, o que introduziria forte distorção na distribuição da variável e risco de viés.

Optou-se, portanto, por transformar a variável em um indicador binário, denominado `PAI_INFORMADO`, que assume o valor "SIM" quando a idade do pai foi informada e "NÃO" quando ausente. Dessa forma, em vez de descartar a variável ou imputar um valor artificial, a própria ausência da informação passou a ser tratada como um dado potencialmente relevante.

Essa decisão mostrou-se pertinente ao se observar a taxa de óbito infantil em cada grupo, apresentada na Tabela 5.12. Entre os 2.271.371 registros sem a idade do pai informada, a taxa de óbito foi de 0,76%, enquanto entre os 2.185.133 registros com a informação a taxa foi de 0,59%. Observa-se, portanto, uma mortalidade infantil aproximadamente 29% maior no grupo em que o pai não foi declarado.

Tabela 5.12: Taxa de óbito infantil segundo a informação da idade do pai.

Situação	**Registros**	**Óbitos**	**Taxa de óbito**
Pai informado (SIM)	2.185.133	12.901	0,59%
Pai não informado (NÃO)	2.271.371	17.319	0,76%

É importante ressaltar que essa associação não implica relação de causa e efeito. A ausência da declaração da idade do pai pode atuar como um marcador de vulnerabilidade social, frequentemente relacionada a situações de menor suporte familiar e socioeconômico, que por sua vez estão associadas a piores desfechos perinatais. Ainda assim, o resultado reforça a adequação de tratar a ausência como informação, em vez de descartá-la.

5.2.12 Consultas pré-natais (CONSPRENAT)

A variável `CONSPRENAT` corresponde ao número de consultas pré-natais realizadas durante a gestação. É uma variável numérica discreta e relevante por refletir o acompanhamento da gestação, sendo o valor zero (ausência de pré-natal) um indicador importante de risco.

Tratamento inicial

A faixa de valores observada foi de 0 a 80 consultas, além do código 99 ("ignorado"). O valor zero foi mantido, por representar a ausência de acompanhamento pré-natal, condição clinicamente relevante. Os 22.272 registros com o código 99 foram convertidos em ausentes. Os demais valores, incluindo os mais elevados, foram preservados, por não haver garantia de que correspondessem a erros de registro, adotando-se uma postura conservadora quanto à exclusão de dados. Após esse tratamento inicial, somando-se os campos vazios, a variável apresentava 52.577 registros ausentes.

Integração com a variável CONSULTAS

A base contém ainda a variável `CONSULTAS`, que, segundo o dicionário de dados do SINASC, representa o número de consultas pré-natais agrupado em faixas: 1 – nenhuma; 2 – de 1 a 3; 3 – de 4 a 6; 4 – 7 e mais; e 9 – ignorado. Trata-se, portanto, da mesma informação da `CONSPRENAT`, em granularidade menor. A Tabela 5.13 apresenta a distribuição de suas categorias.

Tabela 5.13: Códigos da variável CONSULTAS e respectivas frequências.

Código	**Significado**	**Registros**
1	Nenhuma consulta	49.225
2	De 1 a 3 consultas	175.463
3	De 4 a 6 consultas	753.128
4	7 ou mais consultas	3.456.371
9	Ignorado	22.301
–	Ausente (vazio)	16

Por representarem a mesma informação, a manutenção de ambas as variáveis poderia fragmentar a importância atribuída ao acompanhamento pré-natal nos modelos baseados em árvores e comprometer a interpretação dos coeficientes na Regressão Logística. Optou-se, portanto, por manter apenas a `CONSPRENAT`, por sua maior granularidade. Entretanto, em vez de simplesmente descartar a `CONSULTAS`, sua informação foi aproveitada para reduzir os valores ausentes da `CONSPRENAT`, por meio de uma imputação condicional.

A estratégia baseia-se no fato de que, em muitos registros nos quais a `CONSPRENAT` estava ausente, a faixa correspondente da `CONSULTAS` encontrava-se preenchida. Nesses casos, estimou-se um valor plausível para o número de consultas a partir da faixa conhecida, utilizando a mediana da `CONSPRENAT` dentro de cada faixa, calculada apenas sobre os registros em que ambas as variáveis estavam disponíveis. Os valores resultantes são apresentados na Tabela 5.14.

Tabela 5.14: Mapa de imputação condicional: valor estimado de CONSPRENAT por faixa de CONSULTAS.

Faixa (CONSULTAS)	**Mediana da CONSPRENAT**	**Valor imputado**
1 (nenhuma)	0	0
2 (1 a 3)	3	3
3 (4 a 6)	5	5
4 (7 e mais)	9	9

Com a aplicação desse mapeamento, foram recuperados 30.260 registros que, de outra forma, permaneceriam ausentes. A Tabela 5.15 apresenta sua distribuição por faixa. Destaca-se que a maior parte (22.182 registros) pertence à faixa de nenhuma consulta, à qual foi atribuído o valor zero — grupo de especial relevância, por representar a ausência de acompanhamento pré-natal.

Tabela 5.15: Registros recuperados por imputação condicional, segundo a faixa de CONSULTAS.

Faixa (CONSULTAS)	**Registros recuperados**
1 (nenhuma)	22.182
2 (1 a 3)	259
3 (4 a 6)	1.316
4 (7 e mais)	6.503
Total	30.260

Dessa forma, o número de valores ausentes na `CONSPRENAT` foi reduzido de 52.577 para 22.317, correspondendo apenas aos registros em que tanto a `CONSPRENAT` quanto a `CONSULTAS` estavam ausentes. Para esses casos remanescentes, manteve-se a imputação pela mediana (9), aplicada posteriormente no *pipeline* de modelagem. A Tabela 5.16 resume o efeito do tratamento sobre os valores ausentes.

Tabela 5.16: Resumo do tratamento de valores ausentes da CONSPRENAT.

Etapa	**Valores ausentes**
Após tratamento inicial (vazios + código 99)	52.577
Recuperados pela faixa da CONSULTAS	–30.260
Ausentes remanescentes (imputados pela mediana)	22.317

Cabe ressaltar que essa abordagem é mais informativa do que a imputação simples pela mediana global, pois utiliza a informação parcial efetivamente disponível em cada registro — a faixa de consultas — em vez de atribuir um mesmo valor a todos os ausentes, respeitando ainda a relação oficial definida pelo SINASC entre as duas variáveis.

5.2.13 Mês de início do pré-natal (MESPRENAT)

A variável `MESPRENAT` indica o mês de gestação em que a mãe iniciou o acompanhamento pré-natal, variando de 1 (primeiro mês) a 9 (nono mês). É uma variável numérica discreta e um indicador relevante da qualidade do acompanhamento, uma vez que o início precoce do pré-natal está associado a melhores desfechos gestacionais.

Tabela 5.17: Distribuição e tratamento da variável MESPRENAT.

Característica	**Valor**
Nome original	`MESPRENAT`
Tipo	Numérica discreta
Faixa válida	1 a 9 (mês)
Código tratado como ausente	99 (ignorado)
Total de registros	4.456.504
Valores vazios na origem	47.438
Valores com código 99	113.752
Total de ausentes	161.190 (3,62%)
Estratégia de imputação	Mediana (mês 2) + indicador de ausência

Conforme apresentado na Tabela 5.17, a variável apresentava 47.438 registros vazios e 113.752 com o código 99 ("ignorado"), convertidos em ausentes, totalizando 161.190 registros (3,62%). Cabe destacar que o valor 9 é válido, correspondendo ao início no nono mês, não devendo ser confundido com o código 99.

Tratamento dos valores ausentes

A elevada proporção de ausentes e a possibilidade de que essa ausência não fosse aleatória motivaram uma análise cuidadosa da estratégia de imputação. A simples imputação pela mediana (mês 2) atribuiria um início precoce do pré-natal — cenário favorável — a registros cuja informação é desconhecida, podendo mascarar situações de maior risco.

Avaliou-se, inicialmente, uma imputação condicional, estimando o mês de início a partir do número de consultas (`CONSPRENAT`). Contudo, embora exista uma relação clara entre o mês de início e o número de consultas, essa relação não se mostrou suficientemente

consistente no sentido inverso: faixas distintas de número de consultas resultavam em medianas de mês de início semelhantes ou não monotônicas, tornando a estimativa pouco confiável. Por essa razão, a imputação condicional foi descartada.

Optou-se, então, por uma abordagem em duas partes, análoga à adotada para a idade do pai. Os valores ausentes são imputados pela mediana, mas cria-se, adicionalmente, uma variável indicadora binária, `MESPRENAT_INFORMADO`, que registra se o mês de início do pré-natal foi ou não informado. Dessa forma, preserva-se a informação sobre a ausência, permitindo que o modelo avalie seu eventual valor preditivo.

A pertinência dessa decisão é evidenciada pela taxa de óbito infantil em cada grupo, apresentada na Tabela 5.18. Entre os registros sem o mês de início informado, a taxa de óbito foi de 1,65%, ante 0,64% no grupo com a informação — uma mortalidade aproximadamente 2,6 vezes maior. Esse resultado sugere que a ausência dessa informação atua como um marcador de risco, possivelmente associado a gestações com acompanhamento pré-natal precário ou a registros mais incompletos.

Tabela 5.18: Taxa de óbito infantil segundo a informação do mês de início do pré-natal.

Situação	**Registros**	**Óbitos**	**Taxa de óbito**
Mês informado (SIM)	4.295.314	27.562	0,64%
Mês não informado (NÃO)	161.190	2.658	1,65%

5.2.14 Município de nascimento (MUNNASC)

A variável `MUNNASC` indica o município de nascimento. É uma variável categórica nominal de altíssima cardinalidade: a base contém 1.263 municípios distintos. Essa característica exige tratamento específico, pois, ao ser submetida ao *One-Hot Encoding*, cada município gera uma coluna, resultando em mais de mil colunas adicionais.

A análise da distribuição revelou um forte desbalanceamento entre os municípios. Enquanto poucos concentram grande volume de nascimentos — o município mais frequente registra mais de 700 mil nascimentos —, a maior parte apresenta frequência muito baixa: 639 municípios possuem 100 registros ou menos, e 179 possuem apenas um único registro. Essa cauda longa de municípios raros é problemática, pois gera centenas de

colunas com pouquíssimas observações, das quais o modelo não consegue extrair padrões confiáveis, além de inflar artificialmente a importância atribuída à variável nos modelos baseados em árvores.

Para mitigar esse problema, optou-se por agrupar os municípios de baixa frequência em uma única categoria, denominada "OUTROS". Foram mantidos individualmente apenas os municípios com pelo menos 2.000 registros, e os demais foram reunidos na nova categoria. A Tabela 5.19 resume o efeito desse agrupamento.

Tabela 5.19: Efeito do agrupamento dos municípios de baixa frequência.

Característica	**Valor**
Municípios distintos (original)	1.263
Critério de manutenção	$\geq$ 2.000 registros
Municípios mantidos individualmente	292
Categorias após o agrupamento	293 (292 + "OUTROS")
Registros na categoria "OUTROS"	6,3% da base

Com esse tratamento, a cardinalidade da variável foi reduzida de 1.263 para 293 categorias, uma diminuição de aproximadamente 77%, preservando, ainda assim, 93,7% dos registros associados ao seu município de origem. Dessa forma, elimina-se a cauda de municípios raros, que representava majoritariamente ruído, reduzindo o número de colunas geradas pelo *One-Hot Encoding* e atenuando a inflação da importância da variável, sem perda significativa de informação geográfica. Por ser variável nominal, a versão tratada foi posteriormente submetida ao *One-Hot Encoding* no *pipeline* de modelagem.

5.2.15 Local de nascimento (LOCNASC)

A variável `LOCNASC` indica o local de nascimento, originalmente codificada como 1 (hospital), 2 (outro estabelecimento de saúde), 3 (domicílio), 4 (outros) e 9 (ignorado). É uma variável relevante para o estudo da mortalidade infantil, por refletir as condições de assistência ao parto.

A análise revelou que a quase totalidade dos nascimentos ocorreu em hospital (99,3%), com pequenas frações nos demais locais. Apesar desse forte predomínio,

observou-se um gradiente expressivo na taxa de óbito infantil segundo o local: 0,67% nos nascimentos hospitalares, contra cerca de 1,4% a 2,1% nos nascimentos ocorridos fora do hospital — uma mortalidade duas a três vezes maior. Esse padrão é coerente com o menor acesso a cuidados de emergência nesses contextos.

Dada a baixa frequência e a similaridade das taxas entre as categorias de nascimento fora do ambiente hospitalar, optou-se por agrupar a variável de forma binária, distinguindo os nascimentos em hospital dos ocorridos fora dele. Os 183 registros com código ignorado (0,004%) foram tratados como ausentes. A Tabela 5.20 apresenta a distribuição resultante e as respectivas taxas de óbito.

Tabela 5.20: Distribuição e taxa de óbito da variável LOCNASC após binarização.

Categoria	**Registros**	**Óbitos**	**Taxa de óbito**
Hospital	4.426.052	29.667	0,67%
Fora do hospital	30.269	551	1,82%

Por se tratar de variável categórica, a versão binarizada foi posteriormente submetida ao *One-Hot Encoding* no *pipeline* de modelagem.

5.2.16 Situação conjugal da mãe (ESTCIVMAE)

A variável `ESTCIVMAE` indica a situação conjugal da mãe, originalmente codificada como 1 (solteira), 2 (casada), 3 (viúva), 4 (separada judicialmente/divorciada), 5 (união estável) e 9 (ignorada). É uma variável de natureza socioeconômica, relevante por funcionar como possível indicador de suporte social e familiar à gestante.

A análise da taxa de óbito segundo cada categoria revelou um sinal modesto, porém consistente: filhos de mães casadas apresentaram a menor taxa de óbito (0,57%), enquanto as demais categorias situaram-se em torno de 0,70% a 0,79%. O principal contraste, portanto, ocorre entre mães com e sem companheiro estável. Por essa razão, e para evitar a fragmentação do sinal em categorias de baixa frequência, optou-se por agrupar a variável de forma binária: "com companheiro" (casada ou em união estável) e "sem companheiro" (solteira, viúva ou separada). Os 27.165 registros ignorados ou ausentes (0,61%) foram tratados como ausentes.

Tabela 5.21: Distribuição e taxa de óbito da variável ESTCIVMAE após binarização.

Categoria	**Registros**	**Óbitos**	**Taxa de óbito**
Com companheiro	2.360.024	14.121	0,60%
Sem companheiro	2.069.315	15.876	0,77%

Por se tratar de variável categórica, a versão binarizada foi posteriormente submetida ao *One-Hot Encoding* no *pipeline* de modelagem.

5.2.17 Número de perdas fetais e abortos anteriores (QTDFILMORT)

A variável `QTDFILMORT` indica o número de perdas fetais e abortos que a mãe teve anteriormente. É uma variável numérica discreta, de grande relevância clínica, pois o histórico de perdas gestacionais é um fator de risco reconhecido para desfechos adversos em gestações subsequentes.

Tabela 5.22: Distribuição e tratamento da variável QTDFILMORT.

Característica	**Valor**
Nome original	`QTDFILMORT`
Tipo	Numérica discreta
Faixa válida adotada	0 a 20
Total de registros	4.456.504
Valores vazios na origem	57.537
Valores implausíveis (> 20)	9
Total de ausentes	57.546 (1,29%)
Estratégia de imputação	Mediana (0)

Conforme apresentado na Tabela 5.22, a maior parte das mães não possuía perdas anteriores (cerca de 81% dos registros com valor zero), com uma cauda decrescente de frequências. O valor zero foi mantido, por representar a ausência de perdas anteriores. Foram considerados implausíveis os valores superiores a 20, presentes em apenas 9

registros, que foram tratados como ausentes juntamente com os 57.537 campos vazios. Por se tratar de variável numérica fortemente assimétrica, definiu-se a mediana (0) como estratégia de imputação.

A Figura 5.6 ilustra essa distribuição.

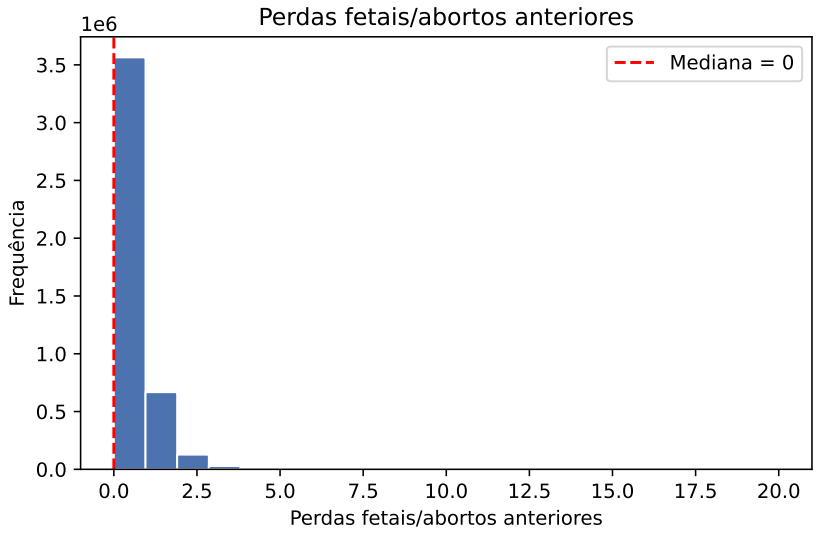

Figura 5.6: Distribuição do número de perdas fetais e abortos anteriores, com a linha tracejada indicando a mediana. A maior parte das mães não possui perdas anteriores (valor zero), com uma cauda decrescente de frequências.

A relevância dessa variável é evidenciada pela relação entre o número de perdas anteriores e a taxa de óbito infantil, apresentada na Tabela 5.23. Observa-se um gradiente crescente e consistente: a taxa de óbito aumenta progressivamente com o número de perdas anteriores, partindo de 0,63% entre as mães sem perdas e atingindo 1,34% entre aquelas com três perdas anteriores — mais que o dobro. Esse padrão é coerente com o conhecimento clínico sobre o histórico obstétrico como fator de risco.

Tabela 5.23: Taxa de óbito infantil segundo o número de perdas anteriores.

Perdas anteriores	**Registros**	**Óbitos**	**Taxa de óbito**
0	3.564.555	22.630	0,63%
1	667.538	5.427	0,81%
2	127.242	1.322	1,04%
3	28.298	378	1,34%

Por ser uma variável numérica, foi posteriormente submetida à imputação no *pipeline* de modelagem.

5.2.18 Anomalia congênita (IDANOMAL)

A variável `IDANOMAL` indica se foi identificada alguma anomalia congênita no recém-nascido, codificada como 1 (sim), 2 (não) e 9 (ignorado). É uma variável categórica de grande relevância clínica, uma vez que as anomalias congênitas estão entre as principais causas de mortalidade infantil.

Tabela 5.24: Distribuição e taxa de óbito da variável IDANOMAL.

Categoria	**Registros**	**Óbitos**	**Taxa de óbito**
Com anomalia (SIM)	46.552	5.546	11,91%
Sem anomalia (NÃO)	4.333.324	24.070	0,56%
Ignorado	76.628	604	0,79%

Conforme apresentado na Tabela 5.24, esta variável revelou o contraste mais expressivo entre todas as analisadas. A taxa de óbito infantil entre os recém-nascidos com anomalia congênita identificada foi de 11,91%, mais de vinte vezes superior à observada entre aqueles sem anomalia (0,56%). Embora represente apenas cerca de 1% dos registros, o grupo com anomalia concentra uma parcela substancial dos óbitos, evidenciando seu elevado poder preditivo. Esse resultado é plenamente coerente com o conhecimento clínico, dado que as anomalias congênitas figuram entre as principais causas de óbito neonatal e infantil.

Os registros com código 9 ("ignorado"), somados aos campos vazios, foram mantidos como uma categoria própria, denominada "IGNORADO". Optou-se por preservá-los como categoria, em vez de imputá-los, uma vez que sua taxa de óbito (0,79%) difere das demais e a sua reclassificação como "sim" ou "não" introduziria uma suposição não sustentada pelos dados. Dessa forma, preserva-se a informação de que a presença ou ausência de anomalia não foi determinada, permitindo que o modelo a trate como uma categoria distinta. Por ser variável nominal, a versão tratada foi posteriormente submetida ao *One-Hot Encoding* no *pipeline* de modelagem.

5.3 Divisão dos Dados e Pré-processamento

Antes da modelagem, a base integrada foi organizada em uma matriz de atributos, representada por X, e em um vetor da variável resposta, representado por y. A matriz X corresponde ao conjunto de variáveis explicativas utilizadas pelos modelos, enquanto o vetor y contém o desfecho associado a cada observação. A variável resposta utilizada foi `OBITO_INFANTIL`, codificada como 1 para registros em que houve óbito infantil e 0 para registros em que não houve óbito. Ao todo, foram utilizadas 4.456.504 observações e 19 variáveis explicativas, cujos tratamentos foram descritos na seção anterior.

Tabela 5.25: Distribuição da variável resposta na base completa.

Classe	**Descrição**	**Quantidade**	**Proporção**
0	Não óbito infantil	4.426.284	99,32%
1	Óbito infantil	30.220	0,68%
	Total	4.456.504	100%

Conforme a Tabela 5.25, a base apresenta forte desbalanceamento entre as classes. A classe de não óbito infantil representa 99,32% dos registros, enquanto a classe de óbito infantil representa apenas 0,68%. Por esse motivo, a acurácia não foi utilizada isoladamente para avaliar os modelos, uma vez que um classificador que previsse todos os registros como não óbito já apresentaria acurácia elevada, apesar de não identificar os casos da classe de interesse.

5.3.1 Separação entre Treino, Validação e Teste

A base foi separada em 60% para treino, 20% para validação e 20% para teste, com estratificação pela variável resposta, preservando a proporção entre óbitos e não óbitos nos três subconjuntos. A mesma divisão dos dados foi utilizada nos três modelos. Esse procedimento é importante para garantir uma comparação justa entre eles, pois assegura que todos sejam treinados e avaliados exatamente sobre os mesmos registros. Dessa forma, eventuais diferenças de desempenho podem ser atribuídas aos próprios modelos, e não a variações na divisão dos dados.

Cada conjunto desempenha um papel distinto no processo. O conjunto de treino é utilizado para ajustar os modelos. O conjunto de validação é utilizado tanto para a seleção dos hiperparâmetros quanto para a escolha do limiar de decisão, etapas realizadas de forma independente, como detalhado nos capítulos seguintes. O conjunto de teste, por sua vez, permanece reservado e não é utilizado em nenhuma etapa de ajuste, sendo empregado exclusivamente na avaliação final do desempenho, de modo a fornecer uma estimativa imparcial da capacidade de generalização dos modelos.

Tabela 5.26: Dimensões dos conjuntos de treino, validação e teste.

Conjunto	**Total de registros**	**Não óbito**	**Óbito**
Treino	2.673.902	2.655.770	18.132
Validação	891.301	885.257	6.044
Teste	891.301	885.257	6.044

Conforme apresentado na Tabela 5.26, a estratificação garantiu que a proporção de óbitos se mantivesse praticamente idêntica nos três conjuntos, em torno de 0,68%. Essa consistência é importante para que a avaliação do modelo reflita o mesmo grau de desbalanceamento em todas as etapas.

5.3.2 Pré-processamento das Variáveis

As variáveis explicativas foram organizadas em três grupos. O pré-processamento foi incorporado ao *pipeline* de cada modelo, conforme discutido na Seção 2.6, de modo que todas as transformações fossem ajustadas exclusivamente com o conjunto de treino, evitando o vazamento de informação.

As variáveis categóricas nominais foram `SEXO_TRATADO`, `PARTO_TRATADO`, `RACACOR_TRATADO`, `UF_TRATADO`, `MUNNASC_TRATADO`, `PAI_INFORMADO_TRATADO`, `MESPRENAT_INFORMADO_TRATADO`, `LOCNASC_TRATADO`, `ESTCIVMAE_TRATADO` e `IDANOMAL_TRATADO`. Essas variáveis foram transformadas por meio de *One-Hot Encoding*, que cria uma coluna binária para cada categoria. Utilizou-se o parâmetro `handle_unknown='ignore'` para lidar com categorias que, eventualmente, estivessem presentes nos conjuntos de validação ou teste, mas ausentes no treinamento. Nesses casos, a categoria desconhecida é codificada com o

valor zero em todas as colunas geradas para a variável, em vez de provocar um erro de execução. Os eventuais valores ausentes nessas variáveis foram representados como uma categoria própria.

As variáveis ordinais foram representadas por `ESCMAE_TRATADO`. Seus valores ausentes foram preenchidos com o valor mais frequente observado no conjunto de treino, utilizando `SimpleImputer(strategy='most_frequent')`.

As variáveis numéricas foram `PESO_TRATADO`, `IDADEMAE_TRATADO`, `SEMAGESTAC_TRATADO`, `APGAR1_TRATADO`, `APGAR5_TRATADO`, `CONSPRENAT_TRATADO`, `MESPRENAT_TRATADO` e `QTDFILMORT_TRATADO`. Seus valores ausentes foram preenchidos pela mediana calculada no conjunto de treino, utilizando `SimpleImputer(strategy='median')`.

No modelo de Regressão Logística, as variáveis ordinais e numéricas foram adicionalmente padronizadas por meio do `StandardScaler`. Essa etapa é necessária para modelos que se baseiam em equações para construir a superfície de separação entre as classes, os quais são sensíveis à escala das variáveis — caso da Regressão Logística. Modelos baseados em árvores de decisão, como o *Random Forest* e o XGBoost, não são sensíveis à escala, pois realizam divisões binárias baseadas em pontos de corte, e não em distâncias ou combinações lineares das variáveis. Por esse motivo, a padronização não traria qualquer efeito sobre seus resultados. Optou-se por aplicá-la apenas onde é efetivamente necessária — na Regressão Logística — mantendo as variáveis em sua escala original nos modelos baseados em árvores, o que preserva a interpretabilidade direta dos valores e a coerência metodológica do pré-processamento de cada modelo.

5.4 Ajuste de Hiperparâmetros: Procedimento Geral

O procedimento de ajuste de hiperparâmetros foi conduzido de forma idêntica para os três modelos, sendo descrito aqui de maneira geral e retomado, nos capítulos seguintes, apenas quanto aos valores específicos avaliados em cada caso.

Para cada modelo, o pré-processamento das variáveis e o algoritmo de classificação foram encadeados em um único *pipeline*, conforme descrito na Seção 2.6. Essa estrutura garante que as transformações de pré-processamento sejam ajustadas exclusivamente com o conjunto de treino, evitando o vazamento de informação para os conjuntos de validação

e teste.

A busca pelos melhores hiperparâmetros foi realizada por meio do `GridSearchCV`. Como a separação entre treino e validação já havia sido definida previamente (Seção 5.3.1), utilizou-se o `PredefinedSplit`, de modo que todas as combinações fossem treinadas no conjunto de treino e avaliadas no conjunto de validação, mantendo a mesma divisão dos dados em todos os modelos.

A função utilizada para pontuar cada combinação de hiperparâmetros baseou-se no F2-score, escolhido por atribuir maior peso à sensibilidade, característica relevante para a identificação dos casos de óbito infantil. Como o F2-score depende de um limiar de decisão, a função calculou esse valor em diferentes limiares (0,3, 0,4, 0,5, 0,6, 0,7, 0,8 e 0,9) e, para cada combinação de hiperparâmetros, considerou o maior F2-score obtido. É importante esclarecer que o limiar não constitui um hiperparâmetro ajustado pelo `GridSearch`: o teste de diferentes limiares serve apenas como critério para pontuar e comparar as combinações. A escolha do limiar de decisão definitivo é realizada posteriormente, de forma independente, sobre o modelo já selecionado, conforme detalhado no capítulo de experimentos.

Para o tratamento do desbalanceamento entre as classes, cada modelo empregou o mecanismo de ponderação equivalente disponível em sua implementação, atribuindo maior peso à classe minoritária, conforme detalhado nas respectivas configurações. Após a seleção dos hiperparâmetros e do limiar, cada modelo foi retreinado sobre a união dos conjuntos de treino e validação e avaliado, por fim, no conjunto de teste.

6 Experimentos Computacionais

Neste capítulo são apresentados os experimentos realizados com os três modelos de classificação — *Random Forest*, Regressão Logística e XGBoost — aplicados à base integrada. Inicialmente, descrevem-se as configurações e os hiperparâmetros selecionados para cada modelo. Em seguida, comparam-se os desempenhos no conjunto de teste e, por fim, apresenta-se uma análise complementar de sensibilidade ao limiar de decisão. O procedimento geral de ajuste, comum aos três modelos, foi descrito na Seção 5.4.

6.1 Configurações dos Modelos

6.1.1 Random Forest

O primeiro modelo utilizado foi o `RandomForestClassifier`, da biblioteca `scikit-learn`. Por se tratar de uma implementação dessa biblioteca, os hiperparâmetros e demais configurações são referenciados pelos nomes utilizados na própria `scikit-learn`. O modelo foi configurado com `random_state = 42` e `class_weight = 'balanced_subsample'`, de modo a reduzir o favorecimento da classe majoritária. Foram testadas 24 combinações de hiperparâmetros, conforme apresentado na Tabela 6.1.

Tabela 6.1: Hiperparâmetros avaliados no Random Forest.

Parâmetro	**Valores avaliados**
`n_estimators`	100 e 200
`max_depth`	15, 20, 25 e 30
`min_samples_leaf`	1, 5 e 10

Conforme o procedimento descrito na Seção 5.4, a melhor configuração encontrada foi `n_estimators = 200`, `max_depth = 30` e `min_samples_leaf = 10`, alcançando F2-score de 0,4738 na validação.

6.1.2 Regressão Logística

O segundo modelo foi uma Regressão Logística implementada por meio do `SGDClassifier`, utilizando a função de perda `log_loss`. Essa configuração permite ajustar um modelo linear probabilístico equivalente à Regressão Logística por meio de gradiente estocástico, alternativa adotada devido ao grande volume de dados, uma vez que o treinamento tradicional apresentou maior custo computacional. O modelo foi configurado com `class_weight = 'balanced'`, `tol = 0,001` e `random_state = 42`, incluindo, no *pipeline*, a etapa de padronização das variáveis, por se tratar de um modelo sensível à escala. Foram testadas 6 combinações de hiperparâmetros, conforme apresentado na Tabela 6.2.

Tabela 6.2: Hiperparâmetros avaliados na Regressão Logística.

Parâmetro	**Valores avaliados**
`alpha`	0,0001, 0,001 e 0,01
`penalty`	L1 e L2

A melhor configuração encontrada foi `alpha = 0,001` com penalidade L2, alcançando F2-score de 0,4812 na validação.

6.1.3 XGBoost

O terceiro modelo utilizado foi o `XGBClassifier`, da biblioteca `xgboost`, que implementa o algoritmo de *gradient boosting* sobre árvores de decisão. O modelo foi configurado com `objective = 'binary:logistic'`, `eval_metric = 'aucpr'`, `tree_method = 'hist'` e `random_state = 42`. O método `hist` foi escolhido por apresentar menor custo computacional e menor consumo de memória, característica relevante diante do grande volume de dados. Para o tratamento do desbalanceamento, utilizou-se o parâmetro `scale_pos_weight`, mecanismo equivalente ao `class_weight` empregado nos modelos do `scikit-learn`. Foram testadas 36 combinações de hiperparâmetros, apresentadas na Tabela 6.3.

Tabela 6.3: Hiperparâmetros avaliados no XGBoost.

Parâmetro	**Valores avaliados**
`n_estimators`	50, 100 e 200
`max_depth`	4, 6 e 8
`learning_rate`	0,05 e 0,1
`scale_pos_weight`	146,5 e 73,2

A faixa de valores do parâmetro `n_estimators` foi definida de modo a manter coerência com a utilizada no *Random Forest*, permitindo uma comparação mais justa entre os modelos. Testes preliminares com números maiores de estimadores (300 e 500) indicaram ganho de desempenho desprezível em relação a 200 árvores, não justificando o custo computacional adicional. A melhor configuração encontrada foi `n_estimators = 200`, `max_depth = 6`, `learning_rate = 0,05` e `scale_pos_weight = 73,2`, alcançando F2-score de 0,5005 na validação.

6.2 Comparação de Desempenho no Teste

Após a seleção dos hiperparâmetros e do limiar de decisão, cada modelo foi retreinado sobre a união dos conjuntos de treino e validação e avaliado no conjunto de teste. Os três modelos apresentaram valores de AUC-ROC próximos de 0,91, indicando capacidade semelhante de discriminação geral entre registros com e sem óbito infantil. Entretanto, devido ao forte desbalanceamento da base, a AUC-PR é mais informativa do que a AUC-ROC para avaliar o desempenho na classe de interesse. Nessa métrica, o XGBoost apresentou o melhor resultado, seguido pelo *Random Forest* e pela Regressão Logística, que apresentaram desempenhos muito próximos entre si.

Tabela 6.4: Comparação das métricas globais dos três modelos no conjunto de teste.

Modelo	**AUC-ROC**	**AUC-PR**
Random Forest	0,9098	0,3211
Regressão Logística	0,9087	0,3167
XGBoost	0,9123	0,3561

Como referência, em uma base com prevalência de óbito infantil de apenas 0,68%, um classificador sem capacidade preditiva apresentaria AUC-PR de aproximadamente 0,0068. Dessa forma, os três modelos apresentaram desempenho muito superior ao acaso, e o XGBoost, com AUC-PR de 0,3561, alcançou um valor cerca de 52 vezes maior que esse patamar de referência. A Figura 6.1 apresenta as curvas Precision-Recall dos três modelos no conjunto de teste.

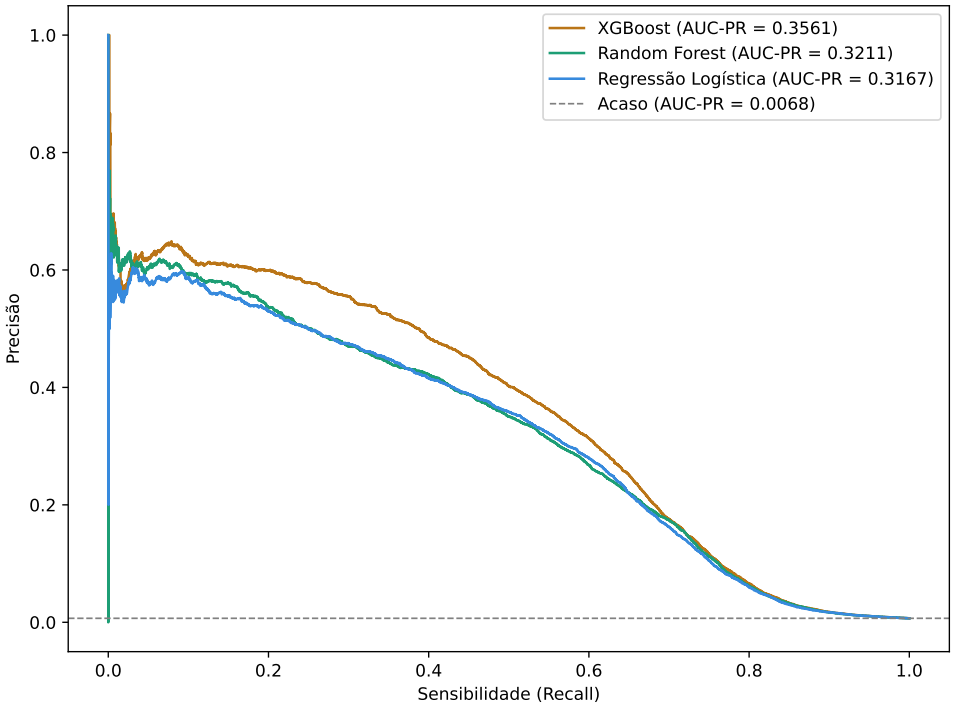

Figura 6.1: Curvas Precision-Recall dos três modelos no conjunto de teste.

Cabe esclarecer que o limiar de decisão foi selecionado de forma independente para cada modelo, sempre com base no maior F2-score obtido no conjunto de validação. Por esse motivo, o *Random Forest* opera no limiar de 0,8, enquanto a Regressão Logística e o XGBoost operam no limiar de 0,9. A Tabela 6.5 apresenta, para cada modelo, o

desempenho no limiar que maximiza seu F2-score, além do limiar de 0,3, utilizado como cenário de máxima sensibilidade, detalhando também os respectivos acertos e erros de classificação.

Tabela 6.5: Comparação de desempenho e dos acertos e erros de classificação dos modelos nos limiares selecionados.

Modelo e limiar	**Precisão**	**Sensibilidade**	**F1-score**	**F2-score**	**VP**	**FP**	**FN**	**VN**
Random Forest 0,3	0,0610	0,8003	0,1134	0,2338	4.837	74.420	1.207	810.837
Regressão Logística 0,3	0,0289	0,8527	0,0558	0,1271	5.154	173.426	890	711.831
XGBoost 0,3	0,0704	0,7950	0,1294	0,2599	4.805	63.442	1.239	821.815
Random Forest 0,8	0,2866	0,5817	0,3841	0,4824	3.516	8.750	2.528	876.507
Regressão Logística 0,9	0,2757	0,6039	0,3786	0,4878	3.650	9.588	2.394	875.669
XGBoost 0,9	0,3378	0,5746	0,4255	0,5039	3.473	6.809	2.571	878.448

Nos limiares que maximizam o F2-score de cada modelo, o XGBoost apresentou o melhor equilíbrio entre sensibilidade e precisão, com o maior F1-score (0,4255) e o maior F2-score (0,5039). A Regressão Logística identificou o maior número de óbitos nesse cenário (3.650 casos, ou 60,4% dos 6.044 óbitos, no limiar de 0,9), seguida pelo *Random Forest* (3.516 casos, 58,2%, no limiar de 0,8) e pelo XGBoost (3.473 casos, 57,5%, no limiar de 0,9). Em contrapartida, o XGBoost gerou o menor número de falsos positivos (6.809), refletindo sua melhor precisão. Observa-se que o XGBoost apresentou, de forma consistente, a melhor relação entre sensibilidade e precisão, o que é coerente com o maior valor de AUC-PR obtido, evidenciando um equilíbrio mais favorável entre a identificação dos óbitos e a geração de alarmes falsos.

6.3 Análise de Sensibilidade ao Limiar de Decisão

Além do limiar que maximiza o F2-score de cada modelo, selecionado a partir do conjunto de validação, realizou-se uma análise complementar do efeito da escolha do limiar sobre o comportamento dos modelos. Diferentemente da comparação principal apresentada na Seção 6.2, esta análise não decorre de uma seleção baseada em desempenho na validação, mas tem caráter ilustrativo, evidenciando o compromisso entre sensibilidade e precisão. Para isso, contrasta-se o limiar ótimo de cada modelo com o limiar de 0,3, voltado à

máxima identificação dos casos de risco, cujos valores foram apresentados na Tabela 6.5.

De modo geral, a redução do limiar aumenta a sensibilidade e a quantidade de verdadeiros positivos, mas também eleva expressivamente o número de falsos positivos, ao passo que limiares mais altos aumentam a precisão e reduzem os falsos positivos, deixando de identificar uma parcela maior dos óbitos infantis.

No limiar de 0,3, a Regressão Logística apresentou a maior sensibilidade, identificando 5.154 dos 6.044 óbitos infantis, seguida pelo *Random Forest* (4.837 casos) e pelo XGBoost (4.805 casos). Contudo, a quantidade de falsos positivos diferiu drasticamente entre os modelos: a Regressão Logística gerou 173.426 falsos positivos, o *Random Forest* gerou 74.420 e o XGBoost, 63.442. Assim, embora a Regressão Logística tenha identificado o maior número de óbitos nesse limiar, o fez ao custo de um número muito elevado de falsos positivos, ao passo que o XGBoost obteve sensibilidade próxima com quantidade substancialmente menor de alarmes falsos.

Esse contraste evidencia que, mesmo em um cenário de máxima sensibilidade, o XGBoost mantém uma relação mais favorável entre a identificação dos óbitos e a geração de alarmes falsos. A escolha final do limiar, no entanto, depende das consequências práticas dos falsos positivos e falsos negativos na aplicação pretendida: em contextos de saúde pública, deixar de identificar um caso de risco (falso negativo) tende a ser mais grave do que gerar um alarme falso (falso positivo), o que justifica a priorização da sensibilidade adotada ao longo deste trabalho.

7 Análise das Importâncias das Variáveis

Além do desempenho preditivo, um dos objetivos deste trabalho é identificar os fatores mais associados à ocorrência de óbito infantil. Para isso, analisam-se neste capítulo as medidas de importância das variáveis fornecidas por cada modelo: a importância por ganho nos modelos baseados em árvores, os valores SHAP como forma complementar de interpretação e os coeficientes da Regressão Logística. Ao final, os resultados são confrontados com análises descritivas da própria base e com a literatura da área, de modo a discutir a plausibilidade clínica dos achados.

7.1 Importância nos Modelos Baseados em Árvores

Nos modelos baseados em árvores, a importância das variáveis é obtida a partir de sua contribuição para as divisões das árvores durante o treinamento — pela redução média de impureza, no *Random Forest*, e pelo ganho relativo, no XGBoost. Em ambos os casos, a análise foi realizada sobre o modelo final, treinado sobre a união dos conjuntos de treino e validação, refletindo a estrutura interna do modelo e não uma medida calculada sobre o conjunto de teste. A importância de cada categoria das variáveis nominais foi somada de volta à variável original, permitindo comparar as 19 variáveis de forma direta.

7.1.1 Random Forest

A Figura 7.1 e a Tabela 7.1 apresentam a importância das variáveis no *Random Forest*.

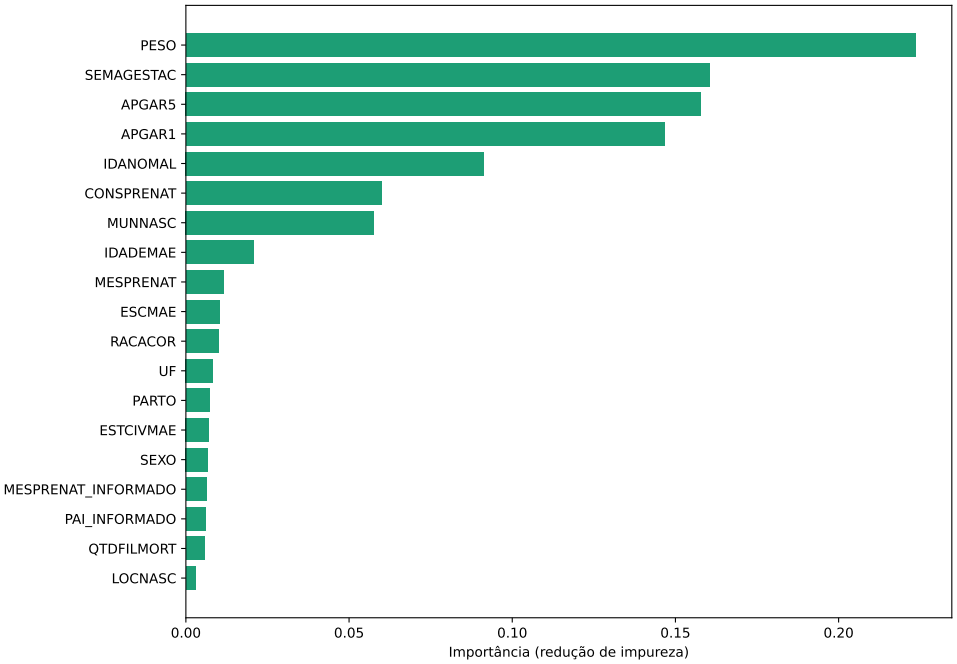

Figura 7.1: Importância das variáveis no modelo Random Forest.

Tabela 7.1: Importância relativa das variáveis no modelo Random Forest.

Variável	**Importância (%)**
Peso ao nascer (PESO)	22,35
Semanas de gestação (SEMAGESTAC)	16,05
Apgar no 5º minuto (APGAR5)	15,79
Apgar no 1º minuto (APGAR1)	14,67
Anomalia congênita (IDANOMAL)	9,14
Consultas pré-natais (CONSPRENAT)	5,99
Município de nascimento (MUNNASC)	5,77
Idade da mãe (IDADEMAE)	2,07
Mês de início do pré-natal (MESPRENAT)	1,15
Escolaridade da mãe (ESCMAE)	1,02
Raça/cor (RACACOR)	1,00
Unidade da federação (UF)	0,83
Tipo de parto (PARTO)	0,72
Situação conjugal (ESTCIVMAE)	0,71
Sexo (SEXO)	0,68
Mês de início do pré-natal informado (MESPRENAT_INFORMADO)	0,63
Pai informado (PAI_INFORMADO)	0,59
Perdas anteriores (QTDFILMORT)	0,57
Local de nascimento (LOCNASC)	0,29

No *Random Forest*, as variáveis de natureza clínica destacaram-se como as mais importantes. As quatro variáveis mais influentes foram o peso ao nascer (22,35%), a duração da gestação (16,05%) e os índices de Apgar no quinto e no primeiro minuto (15,79% e 14,67%), seguidas pela presença de anomalia congênita (9,14%). Esse padrão é fortemente coerente com o conhecimento clínico sobre os principais determinantes da mortalidade infantil.

Um aspecto que merece destaque é o comportamento da variável município de nascimento (`MUNNASC`). Embora se trate de uma variável de alta cardinalidade, que gera um grande número de colunas após o *One-Hot Encoding*, sua importância no *Random*

Forest foi modesta, de apenas 5,77%. Esse comportamento decorre de uma característica central do algoritmo: como cada divisão considera apenas um subconjunto aleatório das variáveis, limita-se a possibilidade de que uma única variável de alta cardinalidade domine o modelo. Dessa forma, no *Random Forest* a importância atribuída ao município reflete de maneira mais contida sua contribuição efetiva, permitindo que as variáveis clínicas emerjam como as mais relevantes. Esse efeito de contenção contrasta com o observado no XGBoost, discutido a seguir.

7.1.2 XGBoost

A Figura 7.2 e a Tabela 7.2 apresentam a importância das variáveis no XGBoost.

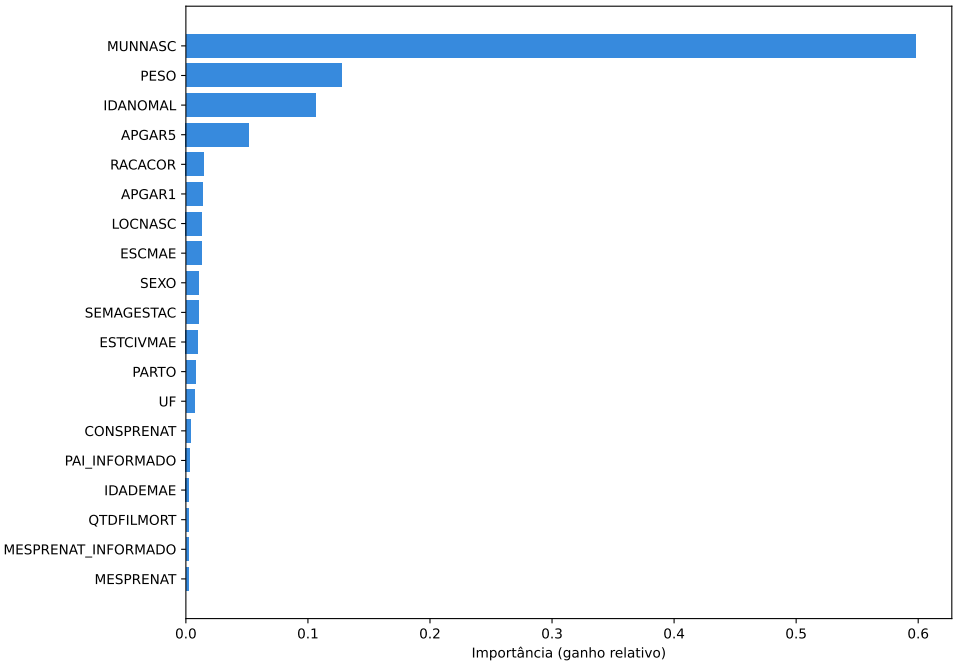

Figura 7.2: Importância das variáveis no modelo XGBoost, considerando todas as variáveis explicativas.

Tabela 7.2: Importância relativa das variáveis no modelo XGBoost.

Variável	**Importância (%)**
Município de nascimento (MUNNASC)	59,77
Peso ao nascer (PESO)	12,74
Anomalia congênita (IDANOMAL)	10,61
Apgar no 5º minuto (APGAR5)	5,18
Raça/cor (RACACOR)	1,49
Apgar no 1º minuto (APGAR1)	1,38
Local de nascimento (LOCNASC)	1,31
Escolaridade da mãe (ESCMAE)	1,29
Sexo (SEXO)	1,03
Semanas de gestação (SEMAGESTAC)	1,03
Situação conjugal (ESTCIVMAE)	0,99
Tipo de parto (PARTO)	0,82
Unidade da federação (UF)	0,73
Consultas pré-natais (CONSPRENAT)	0,41
Pai informado (PAI_INFORMADO)	0,33
Idade da mãe (IDADEMAE)	0,25
Perdas anteriores (QTDFILMORT)	0,22
Mês de início do pré-natal informado (MESPRENAT_INFORMADO)	0,21
Mês de início do pré-natal (MESPRENAT)	0,20

Conforme a Figura 7.2 e a Tabela 7.2, a variável `MUNNASC` concentrou aproximadamente 59,8% da importância total. Esse resultado, no entanto, deve ser interpretado com cautela. Ainda que o município tenha sido agrupado durante o tratamento, reduzindo-se de mais de mil para algumas centenas de categorias, ele permanece como uma variável de cardinalidade elevada, traduzindo-se em um grande número de colunas após o *One-Hot Encoding*. Cada uma dessas colunas oferece ao modelo uma oportunidade adicional de realizar divisões, de modo que a soma de suas importâncias tende a ser inflada. Assim, o predomínio do município reflete, em boa parte, sua cardinalidade, e não necessariamente um poder preditivo superior ao das demais variáveis. Diferentemente do *Random Forest*,

o XGBoost não restringe as variáveis candidatas em cada divisão, o que o torna mais suscetível a esse efeito.

Por esse motivo, para analisar com maior clareza a contribuição das variáveis de natureza clínica e socioeconômica, foi gerada uma segunda versão do gráfico, excluindo a variável `MUNNASC`, apresentada na Figura 7.3.

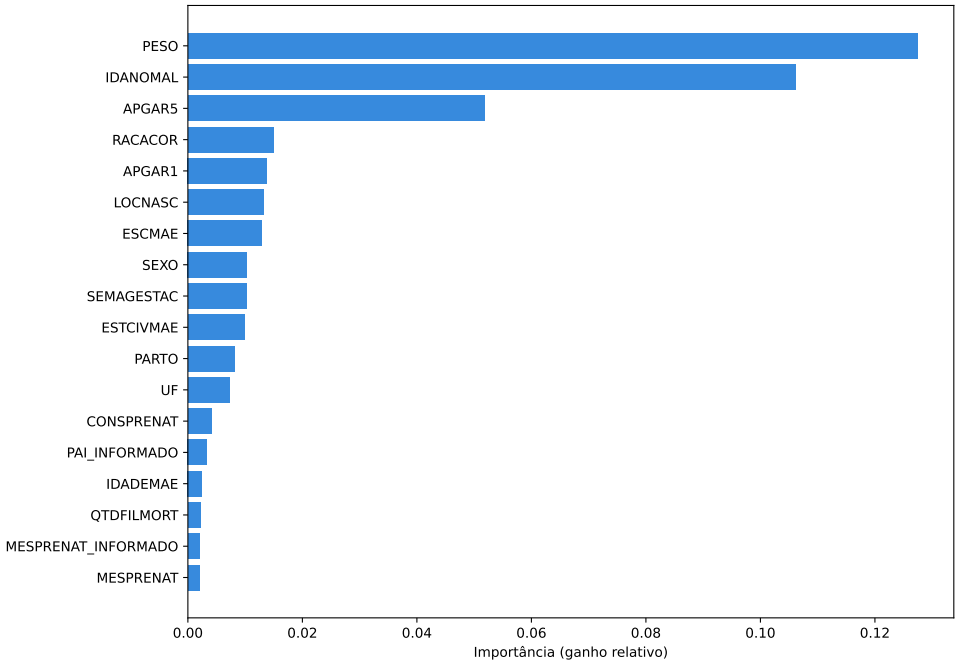

Figura 7.3: Importância das variáveis no modelo XGBoost, excluindo o município de nascimento.

Desconsiderando o município de nascimento, as variáveis mais influentes foram o peso ao nascer (`PESO`) e a presença de anomalia congênita (`IDANOMAL`), seguidas pelo índice de Apgar no quinto minuto (`APGAR5`). Esse resultado é fortemente coerente com o conhecimento clínico sobre mortalidade infantil: o baixo peso ao nascer e as anomalias congênitas estão entre os principais fatores de risco para o óbito infantil. O fato de o modelo ter atribuído maior importância justamente a essas variáveis sugere que as relações aprendidas são clinicamente plausíveis.

Cabe ressaltar que a importância por ganho indica quais variáveis foram mais influentes, mas não informa a direção do efeito. Essa limitação é abordada na análise dos valores SHAP, apresentada na seção seguinte.

7.2 Interpretação por Valores SHAP

Para superar a principal limitação da importância por ganho — que não informa a direção do efeito de cada variável — e atenuar a distorção causada por variáveis de alta cardinalidade, foram calculados os valores SHAP (*SHapley Additive exPlanations*), propostos por Lundberg e Lee (LUNDBERG; LEE, 2017). Diferentemente da importância por ganho, os valores SHAP quantificam a contribuição efetiva de cada variável para cada previsão individual, permitindo avaliar tanto a magnitude quanto a direção de seu efeito. Em ambos os modelos utilizou-se a implementação *TreeSHAP*, otimizada para modelos baseados em árvores.

7.2.1 Random Forest

Em razão do custo computacional do *TreeSHAP* para o *Random Forest*, os valores foram calculados sobre uma amostra de 500 registros do conjunto de teste. A Figura 7.4 apresenta a importância média das variáveis segundo o valor absoluto dos valores SHAP.

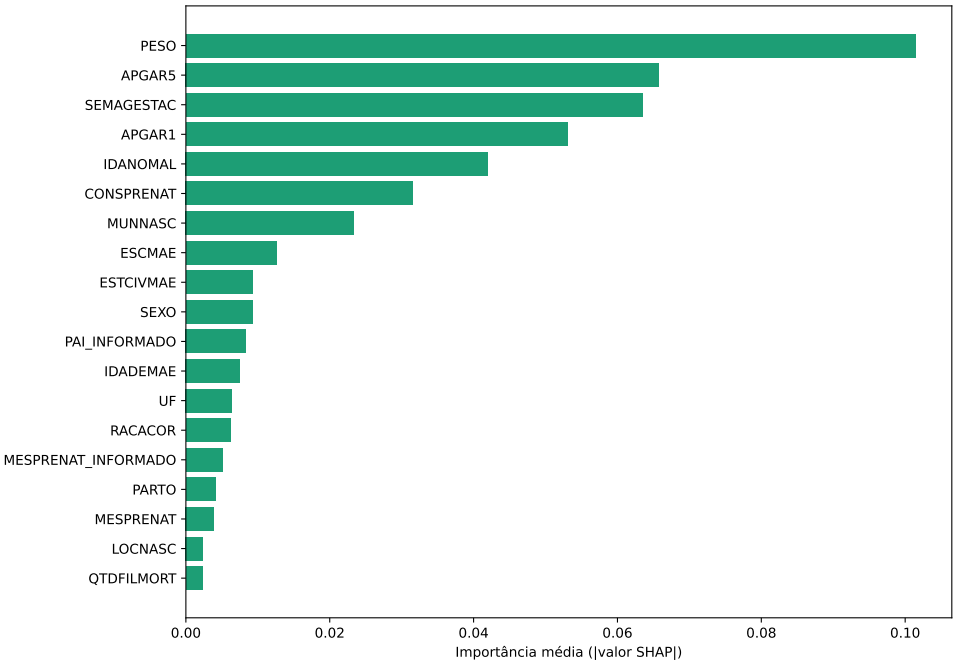

Figura 7.4: Importância média das variáveis no Random Forest segundo os valores SHAP.

Os valores SHAP confirmam o predomínio das variáveis clínicas já observado na importância por redução de impureza. O peso ao nascer (`PESO`) destacou-se como a

variável mais influente, seguido pelo índice de Apgar no quinto minuto (`APGAR5`), pela duração da gestação (`SEMAGESTAC`) e pelo índice de Apgar no primeiro minuto (`APGAR1`). A presença de anomalia congênita (`IDANOMAL`) também figurou entre as variáveis mais relevantes. A variável município de nascimento (`MUNNASC`) apresentou contribuição reduzida segundo os valores SHAP, situando-se entre as de menor influência, em coerência com a análise da importância por redução de impureza e reforçando que o *Random Forest* é menos suscetível à inflação da importância de variáveis de alta cardinalidade.

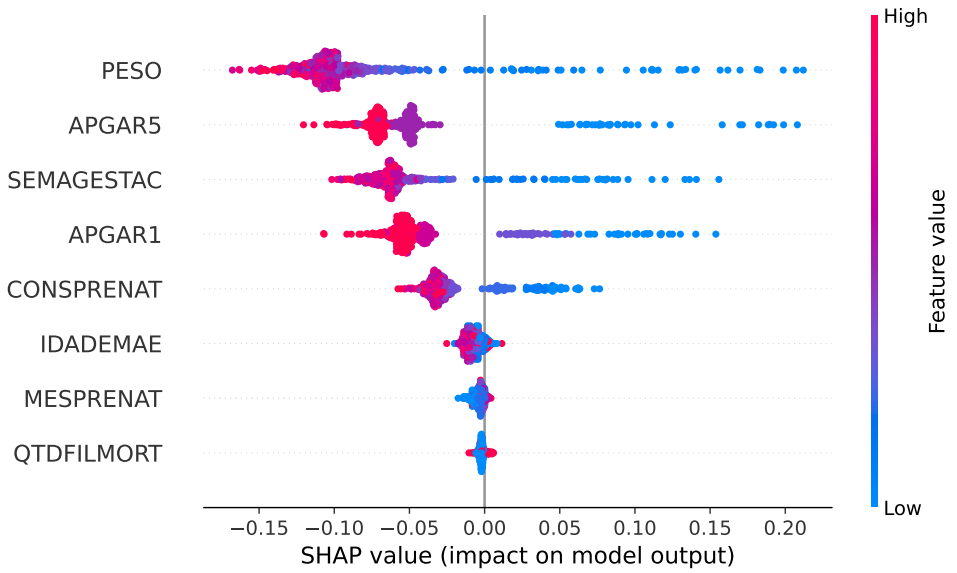

Figura 7.5: Distribuição dos valores SHAP para as variáveis numéricas no Random Forest.

A Figura 7.5 apresenta a distribuição dos valores SHAP para as variáveis numéricas. Os padrões observados são consistentes com o conhecimento clínico: valores baixos de peso ao nascer, dos índices de Apgar e da duração da gestação deslocam-se para a direita, contribuindo para o aumento da probabilidade de óbito, enquanto valores altos atuam no sentido oposto.

7.2.2 XGBoost

Para o XGBoost, os valores SHAP foram calculados sobre uma amostra de 2.000 registros do conjunto de teste. A Figura 7.6 apresenta a importância média das variáveis segundo o valor absoluto dos valores SHAP.

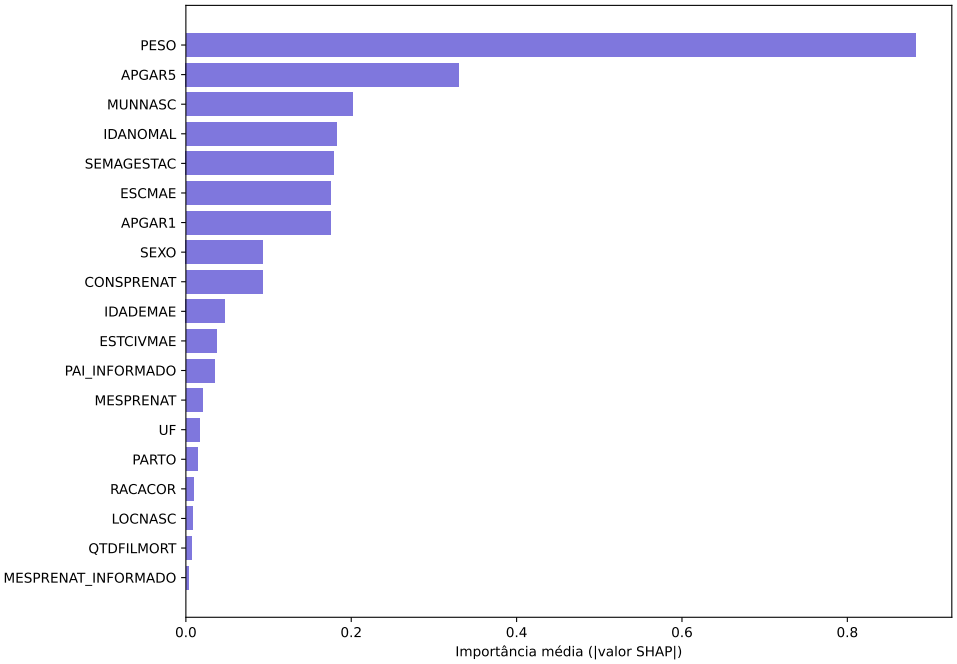

Figura 7.6: Importância média das variáveis no XGBoost segundo os valores SHAP.

Os valores SHAP revelaram um ordenamento de importância mais coerente do ponto de vista clínico do que o obtido pela importância por ganho. O peso ao nascer (`PESO`) destacou-se como a variável mais influente, com larga margem sobre as demais, seguido pelo índice de Apgar no quinto minuto (`APGAR5`). Um aspecto particularmente relevante diz respeito à variável município de nascimento (`MUNNASC`): enquanto na importância por ganho ela concentrava aproximadamente 59,8% da importância total, ocupando a primeira posição, na análise SHAP aparece apenas em terceiro lugar, com contribuição inferior à do peso ao nascer e do Apgar. Esse resultado confirma que a elevada importância atribuída ao município pela métrica de ganho decorria, em grande parte, de sua alta cardinalidade, e não de um poder preditivo real superior. Ao quantificar a contribuição efetiva de cada variável, o SHAP atenua essa distorção e evidencia o predomínio dos fatores clínicos.

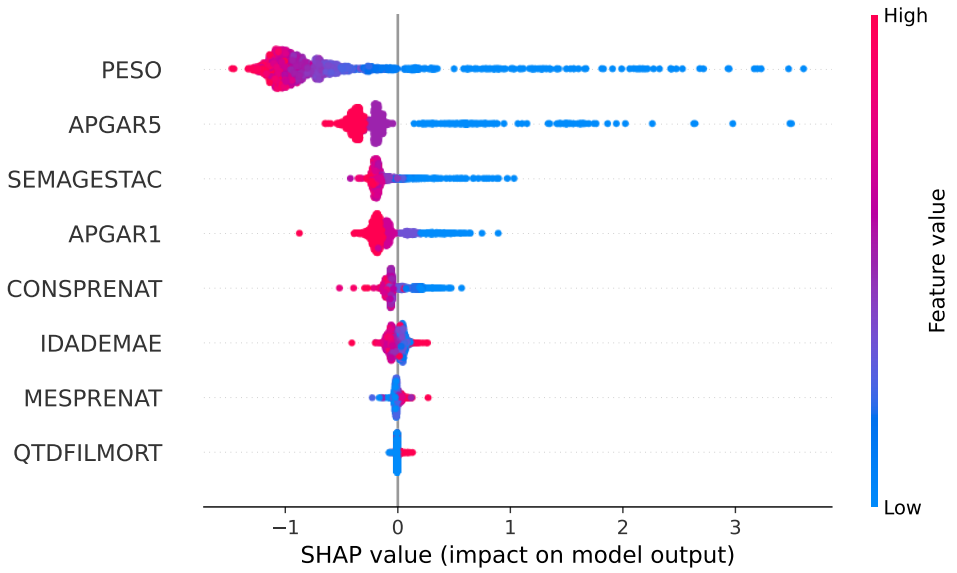

Figura 7.7: Distribuição dos valores SHAP para as variáveis numéricas no XGBoost.

A Figura 7.7 apresenta a distribuição dos valores SHAP para as variáveis numéricas, na qual cada ponto representa um registro. A posição horizontal indica se a variável contribuiu para aumentar (à direita) ou reduzir (à esquerda) a probabilidade de óbito, e a cor representa o valor da variável (vermelho para valores altos, azul para baixos). Observa-se que, para o peso ao nascer, valores baixos (azuis) deslocam-se para a direita, aumentando a probabilidade de óbito, enquanto valores altos (vermelhos) a reduzem, confirmando que o baixo peso ao nascer é fator de risco. Padrão semelhante ocorre para os índices de Apgar e para a duração da gestação, nos quais valores mais baixos contribuem para o aumento da probabilidade de óbito. Esses resultados são consistentes com o conhecimento clínico e reforçam a validade das relações aprendidas pelo modelo.

7.3 Coeficientes da Regressão Logística

Diferentemente dos modelos baseados em árvores, que fornecem a importância das variáveis por ganho, a Regressão Logística fornece coeficientes, que indicam tanto a magnitude quanto a direção do efeito de cada variável sobre a probabilidade de óbito infantil. Coeficientes negativos indicam associação com a redução do risco, enquanto coeficientes positivos indicam associação com o aumento do risco.

Cabe esclarecer que, para a Regressão Logística, não foram calculados os valores

SHAP, ao contrário do que se fez para o *Random Forest* e o XGBoost. Isso se justifica porque os valores SHAP são especialmente úteis para interpretar modelos de natureza "caixa-preta", cuja relação entre as variáveis e a previsão não é diretamente legível. A Regressão Logística, por ser um modelo linear, já é intrinsecamente interpretável: seus coeficientes fornecem diretamente a magnitude e a direção do efeito de cada variável, dispensando um método externo de atribuição. De fato, para modelos lineares, a contribuição dada pelos valores SHAP é equivalente à informação já expressa pelos próprios coeficientes, de modo que seu cálculo seria redundante. Por essa razão, a interpretação da Regressão Logística é conduzida diretamente a partir dos coeficientes estimados, apresentados a seguir.

7.3.1 Variáveis numéricas e ordinais

Para as variáveis numéricas e ordinais, que foram padronizadas e, portanto, possuem coeficientes em uma escala comum, é possível comparar diretamente as magnitudes. A Tabela 7.3 apresenta esses coeficientes, acompanhados das respectivas razões de chance (*odds ratios*), obtidas pela exponenciação dos coeficientes (e^{β}). A razão de chances é a medida usualmente adotada na literatura epidemiológica, por permitir uma interpretação mais direta do efeito de cada variável: valores inferiores a 1 indicam associação com a redução da chance de óbito, enquanto valores superiores a 1 indicam associação com o seu aumento.

Tabela 7.3: Coeficientes padronizados e razões de chance das variáveis numéricas e ordinais.

Variável	**Coeficiente**	**Razão de chances**
Peso ao nascer (PESO)	$-0,4922$	0,6113
Apgar no 5º minuto (APGAR5)	$-0,3505$	0,7043
Semanas de gestação (SEMAGESTAC)	$-0,2776$	0,7576
Escolaridade da mãe (ESCMAE)	$-0,2355$	0,7902
Apgar no 1º minuto (APGAR1)	$-0,2268$	0,7971
Consultas pré-natais (CONSPRENAT)	$-0,0928$	0,9114
Idade da mãe (IDADEMAE)	$-0,0162$	0,9839
Perdas anteriores (QTDFILMORT)	$+0,0074$	1,0074
Mês de início do pré-natal (MESPRENAT)	$+0,0499$	1,0512

O peso ao nascer apresentou o maior coeficiente em valor absoluto ($-0,4922$), seguido pelos índices de Apgar e pela duração da gestação. O sinal negativo dessas variáveis indica que maiores valores estão associados a menor probabilidade de óbito, o que é consistente com o conhecimento clínico. A escolaridade da mãe e o número de consultas pré-natais também apresentaram efeito protetor. As duas variáveis com coeficiente positivo foram o número de perdas anteriores (`QTDFILMORT`) e o mês de início do pré-natal (`MESPRENAT`). Ambos os sinais são clinicamente coerentes: um maior número de perdas gestacionais anteriores e um início mais tardio do acompanhamento pré-natal estão associados a maior risco de óbito.

7.3.2 Variáveis categóricas

Para as variáveis categóricas com três ou mais categorias, transformadas por *One-Hot Encoding*, os coeficientes devem ser interpretados por comparação entre as categorias de uma mesma variável, e não isoladamente, como no caso da raça/cor e da presença de anomalia congênita. Já as variáveis de natureza binária — como o sexo, o local de nascimento (em hospital ou fora dele) e a situação conjugal (com ou sem companheiro) — correspondem, essencialmente, a um único atributo booleano, de modo que suas duas

categorias expressam apenas os dois estados possíveis de uma mesma condição. Nesses casos, os coeficientes das duas categorias refletem, de forma complementar, o mesmo contraste, e sua interpretação recai sobre a diferença entre os dois estados. A Tabela 7.4 apresenta os coeficientes das principais categorias de interesse clínico, excluídas as variáveis município de nascimento e unidade da federação, por sua elevada cardinalidade.

Tabela 7.4: Coeficientes e razões de chance das principais categorias das variáveis nominais.

Categoria	**Coeficiente**	**Razão de chances**
Anomalia congênita – sim	$+1,5191$	4,5681
Anomalia congênita – não	$-1,2378$	0,2900
Anomalia congênita – ignorado	$-0,7395$	0,4774
Nascimento fora do hospital	$+0,1245$	1,1326
Nascimento em hospital	$-0,5732$	0,5637
Mês de início informado	$-0,3401$	0,7117
Mês de início não informado	$-0,1181$	0,8886
Sexo feminino	$-0,3213$	0,7252
Sexo masculino	$-0,1101$	0,8957
Com companheiro	$-0,2594$	0,7715
Sem companheiro	$-0,1766$	0,8381
Raça/cor indígena	$+0,0885$	1,0925
Raça/cor branca	$-0,1482$	0,8623

A variável mais expressiva foi a presença de anomalia congênita: a categoria "sim" apresentou o maior coeficiente positivo de todo o modelo ($+1,5191$), em forte contraste com a categoria "não" ($-1,2378$). Essa diferença acentuada confirma que a presença de anomalia congênita é o fator mais fortemente associado ao aumento do risco de óbito infantil, em consonância com os resultados obtidos pelos modelos baseados em árvores. As demais categorias seguem padrões clinicamente plausíveis quando comparadas dentro de cada variável: o nascimento fora do hospital está associado a maior risco do que o nascimento hospitalar; o sexo feminino apresenta menor risco que o masculino; e a

presença de companheiro, assim como o registro do início do pré-natal, estão associados a menor risco. Esses contrastes reforçam a coerência entre o modelo e o conhecimento epidemiológico sobre mortalidade infantil.

7.4 Discussão e Confronto com a Literatura

Os três modelos, apesar de fundamentos distintos, convergiram quanto aos fatores mais associados ao óbito infantil: o peso ao nascer, a duração da gestação, os índices de Apgar e a presença de anomalia congênita destacaram-se de forma consistente. Esta seção confronta esses achados com análises descritivas da própria base e com a literatura da área, avaliando sua plausibilidade.

O peso ao nascer foi a variável mais influente em praticamente todas as análises. Esse resultado é coerente com a literatura, que aponta o baixo peso ao nascer (inferior a 2.500 g) como um dos fatores isolados de maior influência na sobrevivência nos primeiros anos de vida, associado a maior morbimortalidade neonatal e infantil (MINAGAWA et al., 2012). Estima-se que a mortalidade neonatal seja cerca de vinte vezes maior entre recém-nascidos de baixo peso e até duzentas vezes maior naqueles de muito baixo peso, quando comparados aos de peso adequado (OHLSSON; SHAH, 2008). A distribuição dos valores SHAP obtida neste trabalho reproduz exatamente esse padrão, com os menores pesos deslocando a previsão no sentido do óbito.

A duração da gestação figurou entre as variáveis mais relevantes, o que também encontra respaldo na literatura: a prematuridade é um determinante central do crescimento intrauterino e da sobrevivência, de modo que, quanto mais curta a gestação, maior o risco de mortalidade e morbidade (PRICE et al., 2020). Prematuridade e restrição de crescimento intrauterino são, inclusive, os principais determinantes do próprio baixo peso ao nascer (BARROS et al., 2023), o que ajuda a explicar a atuação conjunta dessas variáveis observada nos modelos.

Os índices de Apgar, sobretudo o do quinto minuto, destacaram-se de forma consistente. Embora o Apgar não seja um preditor absoluto de prognóstico a longo prazo, é um indicador importante de eventos hipóxicos e de risco de morte no período neonatal (Sanarmed, 2025). Estudos que analisam a mortalidade infantil confirmam essa associa-

ção, com Apgar inadequado no quinto minuto apresentando forte razão de chances para o óbito (Research, Society and Development, 2021). O gradiente descrito na base — em que valores baixos deslocam a previsão para o óbito — é plenamente compatível com esse conhecimento.

A presença de anomalia congênita revelou o contraste mais expressivo entre todas as variáveis analisadas: conforme a Tabela 5.24, a taxa de óbito entre recém-nascidos com anomalia foi de 11,91%, mais de vinte vezes superior à observada entre aqueles sem anomalia (0,56%). Esse achado é coerente com a literatura, que aponta as malformações congênitas como causa crescente de mortalidade infantil no Brasil — tendo passado, entre 2000 e 2015, da quarta para a terceira principal causa de óbito em menores de vinte anos (CARDOSO et al., 2021). Nos modelos, a anomalia congênita emergiu como um dos fatores de maior peso, tanto na importância por ganho quanto no maior coeficiente positivo da Regressão Logística.

Além dos fatores clínicos, as análises descritivas realizadas durante o tratamento das variáveis evidenciaram gradientes coerentes com determinantes socioeconômicos e assistenciais amplamente descritos. O número de perdas gestacionais anteriores (Tabela 5.23) apresentou gradiente crescente de mortalidade, e o local de nascimento (Tabela 5.20) mostrou mortalidade duas a três vezes maior fora do ambiente hospitalar, padrões compatíveis com o histórico obstétrico como fator de risco e com o menor acesso a cuidados de emergência. Também as ausências de informação — idade do pai (Tabela 5.12) e mês de início do pré-natal (Tabela 5.18) — associaram-se a maior mortalidade, sugerindo que atuam como marcadores de vulnerabilidade social, frequentemente ligada a piores desfechos perinatais (CABRAL et al., 2015).

Em conjunto, a convergência entre modelos de fundamentos distintos, o suporte das análises descritivas da própria base e a consonância com a literatura reforçam a robustez dos achados e sua plausibilidade clínica, indicando que as relações aprendidas pelos modelos refletem determinantes efetivamente reconhecidos da mortalidade infantil.

8 Síntese dos Resultados

Os três modelos apresentaram capacidade relevante de discriminação entre registros com e sem óbito infantil, com valores de AUC-ROC próximos de 0,91. A diferenciação mais clara ocorreu na AUC-PR, métrica mais adequada para bases desbalanceadas: o XGBoost obteve 0,3561, o *Random Forest* obteve 0,3211 e a Regressão Logística obteve 0,3167. Dessa forma, o XGBoost foi o modelo com melhor desempenho global na identificação da classe de óbito infantil, enquanto o *Random Forest* e a Regressão Logística apresentaram resultados muito próximos entre si.

A escolha do limiar de decisão permite ajustar o comportamento dos modelos de acordo com o objetivo da aplicação. Caso a prioridade seja identificar a maior quantidade possível de casos de risco, a Regressão Logística com limiar de 0,3 alcançou a maior sensibilidade (0,8527), identificando 5.154 óbitos. Esse cenário, porém, gerou 173.426 falsos positivos, o maior número entre os modelos. O XGBoost com limiar de 0,3 ofereceu uma alternativa mais eficiente para esse mesmo objetivo, identificando 4.805 óbitos com cerca de um terço dos falsos positivos da Regressão Logística (63.442).

Caso seja necessário equilibrar a identificação dos casos positivos com uma quantidade menor de falsos positivos, o XGBoost com limiar de 0,9 apresentou o resultado mais favorável, com F1-score de 0,4255 e F2-score de 0,5039, superando os demais modelos. Nesse limiar, o XGBoost identificou 3.473 óbitos com a maior precisão (0,3378) e o menor número de falsos positivos (6.809) entre os três modelos.

Além do desempenho preditivo, a análise da importância das variáveis e dos coeficientes permitiu identificar os principais fatores associados ao óbito infantil. Os três modelos convergiram quanto aos atributos mais relevantes: o peso ao nascer, a duração da gestação, os índices de Apgar e a presença de anomalia congênita destacaram-se de forma consistente. Essa convergência entre modelos de fundamentos distintos reforça a robustez dos achados e sua coerência com o conhecimento clínico sobre mortalidade infantil.

Os resultados mostram que, embora a escolha final do limiar dependa das consequências práticas dos falsos positivos e falsos negativos em uma aplicação relacionada à

saúde pública, o XGBoost apresentou o melhor desempenho entre os modelos avaliados, tanto na métrica global mais adequada ao problema (AUC-PR) quanto na identificação dos casos de óbito infantil nos diferentes limiares analisados. Assim, considerando o objetivo de identificar casos de risco em uma base fortemente desbalanceada, o XGBoost se mostrou o modelo mais adequado entre os três avaliados, aliando bom desempenho preditivo à identificação de fatores clinicamente coerentes.

9 Conclusão e Trabalhos Futuros

Este trabalho teve como objetivo integrar as bases SIM e SINASC da região Sudeste por meio de *linkage* probabilístico e, a partir da base resultante, analisar os fatores associados à mortalidade infantil com o auxílio de técnicas de aprendizado de máquina. O procedimento de vinculação, estruturado em três rodadas com flexibilização progressiva dos critérios de similaridade, permitiu recuperar aproximadamente 82,6% dos óbitos infantis registrados no SIM. Sobre a base integrada, três modelos de classificação — *Random Forest*, Regressão Logística e XGBoost — foram aplicados e comparados em um cenário de forte desbalanceamento de classes, tendo o XGBoost apresentado o melhor desempenho na métrica mais adequada ao problema (AUC-PR). A análise da importância das variáveis, por sua vez, revelou convergência entre os modelos quanto aos principais fatores associados ao óbito infantil — peso ao nascer, duração da gestação, índices de Apgar e presença de anomalia congênita —, em consonância com o conhecimento clínico da área.

A partir dos resultados obtidos e das limitações identificadas ao longo do trabalho, apontam-se algumas direções para pesquisas futuras.

Uma primeira possibilidade consiste em substituir o município de nascimento por um nível de agregação geográfica mais amplo, como as regiões intermediárias ou imediatas dos estados. Essa alternativa reduziria substancialmente a cardinalidade da variável — atenuando a inflação de sua importância observada nos modelos baseados em árvores, sobretudo no XGBoost — e, ao mesmo tempo, permitiria avaliar se alguma região específica se destaca quanto ao risco de óbito infantil, possibilitando a identificação de padrões geográficos que a granularidade municipal, dispersa em centenas de categorias, tende a mascarar.

Outra direção promissora consiste em avaliar e comparar diferentes estratégias de tratamento do desbalanceamento de classes. Além da ponderação de classes adotada neste trabalho, poderiam ser investigadas técnicas de reamostragem, como a geração de exemplos sintéticos da classe minoritária por meio do SMOTE ou a subamostragem da classe majoritária, bem como abordagens híbridas que combinam ambas. A comparação

sistemática dessas estratégias, sob as mesmas condições experimentais, permitiria identificar qual delas oferece o melhor equilíbrio entre sensibilidade e precisão na predição do óbito infantil, aprofundando a análise iniciada aqui.

Uma segunda direção diz respeito à robustez estatística dos resultados. Os modelos finais poderiam ser reexecutados diversas vezes com diferentes sementes aleatórias, reportando-se a média e o desvio-padrão das métricas de desempenho. Esse procedimento permitiria avaliar a estabilidade dos resultados e a significância das diferenças observadas entre os modelos, conferindo maior solidez às comparações realizadas.

Por fim, o próprio processo de *linkage* poderia ser aprimorado com a inclusão das demais regiões do país, o que reduziria a perda de pares associada à mobilidade geográfica das famílias entre o nascimento e o óbito, discutida como limitação deste trabalho.

Bibliografia

BARROS, F. C. et al. Restrição do crescimento intrauterino, prematuridade e baixo peso ao nascer: fenótipos de risco de morte neonatal, estado do rio de janeiro, brasil. *Cadernos de Saúde Pública*, 2023.

Brasil. Ministério da Saúde. *DATASUS — Sistema de Informações sobre Mortalidade (SIM) e Sistema de Informações sobre Nascidos Vivos (SINASC)*. 2021. Disponível em: <http://tabnet.datasus.gov.br>. Acesso em: 23 jun. 2026.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

CABRAL, S. A. et al. Fatores associados à mortalidade infantil e vulnerabilidade social. *Revista Brasileira de Saúde Materno Infantil*, 2015.

CARDOSO, A. C. d. A. et al. Desigualdades nas taxas de mortalidade por malformações do sistema circulatório em crianças menores de 20 anos entre macrorregiões brasileiras. *Arquivos Brasileiros de Cardiologia*, 2021.

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.

CHRISTEN, P. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin: Springer, 2012.

COX, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 20, n. 2, p. 215–242, 1958.

FELLEGI, I. P.; SUNTER, A. B. A theory for record linkage. *Journal of the American Statistical Association*, v. 64, n. 328, p. 1183–1210, 1969.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. 3. ed. Hoboken: John Wiley & Sons, 2013.

Instituto Brasileiro de Geografia e Estatística. *Censo Demográfico 2022: Fecundidade e migração — Resultados preliminares da amostra*. 2023. <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/43815-censo-2022-19-2-milhoes-de-pessoas-vivem-fora-de-sua-regiao-de-nascimento>. Acesso em: [coloque a data de acesso].

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, v. 30, 2017.

MAIA, L. T. d. S.; SOUZA, W. V. d.; MENDES, A. C. G. A contribuição do linkage entre o sim e sinasc para a melhoria das informações da mortalidade infantil em cinco cidades brasileiras. *Revista Brasileira de Saúde Materno Infantil*, v. 15, n. 1, p. 57–66, 2015.

MINAGAWA, Á. T. et al. Peso ao nascer: uma abordagem nutricional. *Comunicação em Ciências Saúde*, v. 23, n. 1, p. 19–30, 2012.

NIAZ, N. U.; SHAHARIAR, K. M. N.; PATWARY, M. J. A. Class imbalance problems in machine learning: A review of methods and future challenges. In: *Proceedings of the 2nd International Conference on Computing Advancements*. New York: Association for Computing Machinery, 2022. p. 485–490.

OHLSSON, A.; SHAH, P. Determinants and prevention of low birth weight: a synopsis of the evidence. *Institute of Health Economics*, 2008.

PRICE, G. C. T. et al. Fatores de risco de baixo peso ao nascer: estudo analítico ecológico de múltiplos grupos. *Jornal Memorial da Medicina*, v. 1, n. 2, p. 46–56, 2020.

RASELLA, D. et al. Effect of a conditional cash transfer programme on childhood mortality: a nationwide analysis of Brazilian municipalities. *The Lancet*, v. 382, n. 9886, p. 57–64, 2013.

Research, Society and Development. Fatores relacionados à mortalidade infantil por anomalias congênitas. *Research, Society and Development*, 2021.

Sanarmed. *Escore de Apgar: como avaliar, interpretar e aplicar na prática clínica neonatal*. 2025. <https://sanarmed.com/escore-de-apgar/>. Acesso em: 30 jun. 2026.

SANDERS, L. S. C. et al. Mortalidade infantil: análise de fatores associados em uma capital do nordeste brasileiro. *Cadernos Saúde Coletiva*, v. 25, n. 1, p. 83–89, 2017.

SANTOS, B. S. D. et al. Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers & Industrial Engineering*, v. 138, p. 106120, 2019.

SCHRAAGEN, M. P. *Aspects of Record Linkage*. Tese (Doutorado) — Leiden University, Leiden, 2014. Disponível em: <http://hdl.handle.net/1887/29716>.

A Dicionário das Variáveis do Modelo

Este apêndice apresenta o dicionário das variáveis explicativas utilizadas na etapa de modelagem, já em sua forma tratada, conforme descrito na Seção 5.2. Para cada variável, indicam-se o nome, o tipo, uma breve descrição e a respectiva faixa de valores válidos ou o conjunto de categorias. As variáveis empregadas exclusivamente no processo de *linkage* não são aqui contempladas. A última linha refere-se à variável resposta.

Tabela A.1: Dicionário das variáveis utilizadas na modelagem.

Variável	**Tipo**	**Descrição**	**Faixa / Categorias**
PESO	Numérica contínua	Peso do recém-nascido ao nascer, em gramas	100 a 7.000 g
IDADEMAE	Numérica discreta	Idade da mãe, em anos completos, no nascimento	10 a 64 anos
SEMAGESTAC	Numérica discreta	Duração da gestação, em semanas completas	19 a 45 semanas
APGAR1	Numérica discreta	Índice de Apgar no 1º minuto de vida	0 a 10
APGAR5	Numérica discreta	Índice de Apgar no 5º minuto de vida	0 a 10
CONSPRENAT	Numérica discreta	Número de consultas pré-natais na gestação	0 a 80
MESPRENAT	Numérica discreta	Mês de gestação de início do pré-natal	1 a 9 (mês)
QTDFILMORT	Numérica discreta	Número de perdas fetais e abortos anteriores	0 a 20
ESCMAE	Ordinal	Escolaridade da mãe, em anos de estudo	1 a 5 (faixas)

Variável	**Tipo**	**Descrição**	**Faixa / Categorias**
SEXO	Categórica nominal	Sexo do recém-nascido	Masculino; Feminino; Ignorado
PARTO	Categórica nominal	Tipo de parto	Vaginal; Cesáreo; Ignorado
RACACOR	Categórica nominal	Raça/cor do recém-nascido	Branca; Preta; Amarela; Parda; Indígena; Ignorado
UF	Categórica nominal	Unidade da federação de nascimento	SP; MG; RJ; ES
MUNNASC	Categórica nominal	Município de nascimento (agrupado)	293 categorias (292 + "Outros")
LOCNASC	Nominal binária	Local de nascimento	Hospital; Fora do hospital
ESTCIVMAE	Nominal binária	Situação conjugal da mãe	Com companheiro; Sem companheiro
IDANOMAL	Categórica nominal	Presença de anomalia congênita	Sim; Não; Ignorado
PAI_ INFORMADO	Nominal binária	Indica se a idade do pai foi informada	Sim; Não
MESPRENAT_ INFORMADO	Nominal binária	Indica se o mês de início do pré-natal foi informado	Sim; Não
OBITO_ INFANTIL	Binária (resposta)	Ocorrência de óbito infantil (desfecho)	0 (não); 1 (sim)