UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE ESTUDOS EM SAÚDE COLETIVA

Aprendizado de Máquina Aplicado ao Pós-Processamento do Relacionamento Probabilístico de Bases de Dados de Saúde

Marco Elisio Oliveira Jardim

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Saúde Coletiva do Instituto de Estudos de Saúde Coletiva da Universidade Federal do Rio de Janeiro – UFRJ, como parte dos requisitos necessários à obtenção do título de Doutor em Ciências (Saúde Coletiva).

Orientadora: Rejane Sobrino Pinheiro

Rio de Janeiro

February 8, 2026

[Jardim], [Marco Elisio Oliveira]
CODIGO CUTTER Aprendizado de Máquina Aplicado ao Pós-Processamento de Relacionamento Probabilístico de Bases de Dados de Saúde / Marco Elisio Oliveira Jardim – Rio de Janeiro, 2025.
Número de folhas

Orientadora: Rejane Sobrino Pinheiro
Tese - Universidade Federal do
Rio de Janeiro, Instituto de Estudos em Saúde Coletiva, Programa de Pós-graduação em Saúde Coletiva, 2025.
Referências Bibliográficas: f. [primeira folha das referências biblio]-[última].
1. Relacionamento de registros. 2. Aprendizado de máquina. 3. Linkage probabilístico. 4. Desbalanceamento de classes. 5. Bases de dados de saúde.
6. Tuberculose. I. Pinheiro, Rejane Sobrino,
orient. II. Universidade
Federal do Rio de Janeiro, Instituto de Estudos em Saúde Coletiva, Programa
de Pós-graduação em Saúde Coletiva. III. Título

Resumo

Aprendizado de Máquina Aplicado ao Pós-Processamento do Relacionamento Probabilístico de Bases de Dados de Saúde

Marco Elisio Oliveira Jardim

Orientadora: Rejane Sobrino Pinheiro

Resumo da Tese de Doutorado apresentada ao Programa de Pós-Graduação em Saúde Coletiva do Instituto de Estudos em Saúde Coletiva da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Doutor em Ciências (Saúde Coletiva).

TEXTO DO RESUMO

Palavras-chave: Relacionamento de registros, Aprendizado de máquina, Linkage probabilístico, Desbalanceamento de classes, Bases de dados de saúde, Tuberculose, SIM, Sinan.

Abstract

Machine Learning Applied to Post-Processing of Probabilistic Record Linkage of Health Databases

Marco Elisio Oliveira Jardim

Advisor: Rejane Sobrino Pinheiro

Abstract da Dissertação de Doutorado apresentada ao Programa de Pós-Graduação em Saúde Coletiva do Instituto de Estudos em Saúde Coletiva da Universidade Federal do Rio de Janeiro - UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Ciências (Saúde Coletiva).

TEXTO DO RESUMO EM INGLÊS

Keywords: Record linkage, Machine learning, Probabilistic linkage, Class imbalance, Health databases, Tuberculosis, SIM, Sinan.

Agradecimentos

Agradecimentos. Não esquecer de agradecer à agência pela bolsa concedida.

Sumário

Lista de Figuras

Capítulo 1

Introdução

A avaliação do desempenho de sistemas de saúde depende da capacidade de acompanhar o percurso de pacientes através de múltiplos sistemas de informação, desde a notificação de agravos até o desfecho clínico [?, ?]. No Brasil, os Sistemas de Informação em Saúde (SIS) foram concebidos para finalidades específicas e não dispõem de um identificador unívoco comum, o que impõe a necessidade de técnicas de vinculação de registros para integrar dados de diferentes fontes [?]. O *linkage* de bases de dados constitui, portanto, etapa indispensável para a produção de indicadores de qualidade do cuidado, a identificação de subnotificação e a construção de trajetórias longitudinais de pacientes, subsidiando a vigilância epidemiológica e a gestão em saúde. O presente capítulo apresenta os fundamentos conceituais e metodológicos do *linkage*, as estratégias disponíveis e os desafios específicos que motivam a abordagem proposta nesta tese.

1.1 *Linkage* de bases de dados: conceitos e fundamentos

O *linkage* de bases de dados, também denominado relacionamento de registros (*record linkage*), consiste no processo de identificar, em duas ou mais bases de dados distintas, registros que se referem a uma mesma entidade (tipicamente um mesmo indivíduo), possibilitando a integração de informações provenientes de diferentes fontes e a qualificação de dados para a produção de conhecimento em saúde [?]. Essa técnica tem sido amplamente empregada em estudos epidemiológicos e de avaliação de sistemas e serviços de saúde, na medida em que permite o seguimento longitudinal de pacientes

através de múltiplos sistemas de informação, a identificação de subnotificação de agravos e a construção de indicadores de desempenho que não seriam passíveis de obtenção a partir de bases isoladas [?, ?].

A necessidade de relacionar registros de diferentes fontes decorre, em grande parte, da fragmentação dos sistemas de informação em saúde, nos quais as bases de dados são desenhadas para finalidades específicas (vigilância epidemiológica, faturamento hospitalar, registro de óbitos, entre outras) e não contemplam, em sua concepção, mecanismos padronizados de interligação [?]. Nesse contexto, o *linkage* constitui ferramenta fundamental para a produção de indicadores a partir de dados administrativos, superando as limitações inerentes ao uso de bases isoladas e viabilizando análises longitudinais e intersetoriais, com vistas à melhoria da integralidade do cuidado em saúde.

Originalmente proposto por Newcombe e colaboradores em 1959 [?], o método foi formalizado por Fellegi e Sunter em 1969 [?], que estabeleceram a teoria matemática para o *linkage* probabilístico. Desde então, a técnica evoluiu consideravelmente, incorporando avanços em métricas de comparação de campos textuais [?, ?], estratégias de blocagem para viabilizar o processamento de grandes volumes de dados [?] e, mais recentemente, abordagens baseadas em aprendizado de máquina (*machine learning*) para a classificação automatizada de pares candidatos [?, ?], apontando para um campo em contínua expansão metodológica.

No campo da saúde pública, o *linkage* tem se mostrado particularmente promissor para a avaliação do desempenho de sistemas e serviços de saúde [?], podendo permitir, por exemplo, a identificação de óbitos por tuberculose não notificados ao sistema de vigilância [?, ?], a análise de causas múltiplas de morte em coortes de pacientes [?] e a melhoria da qualidade dos dados registrados em sistemas nacionais [?]. Soma-se a isso o potencial de produção de indicadores inovadores que possibilitem o acompanhamento do itinerário terapêutico do paciente na rede de serviços, identificando nós críticos e desfechos desfavoráveis ao longo da linha de cuidado, subsidiando a tomada de decisão em saúde.

1.2 Bases de dados de saúde no Brasil

O Sistema Único de Saúde (SUS), instituído pela Constituição Federal de 1988, organiza-se como um sistema universal que integra ações e serviços de assistência, vigilância e gestão em saúde em todo o território nacional [?]. Para subsidiar essas atividades, o Brasil dispõe de um amplo conjunto de Sistemas de Informação em Saúde (SIS), cada qual concebido para registrar eventos específicos e atender a finalidades distintas no âmbito do cuidado em saúde [?]. Entre os principais sistemas nacionais, destacam-se:

- **SIM** (Sistema de Informações sobre Mortalidade): registra todos os óbitos ocorridos no território nacional, a partir das Declarações de Óbito, com informações sobre causas básica e associadas de morte;
- **Sinan** (Sistema de Informação de Agravos de Notificação): registra casos de doenças e agravos de notificação compulsória, constituindo instrumento central para a vigilância epidemiológica;
- **SIH-SUS** (Sistema de Informações Hospitalares do SUS): registra internações hospitalares realizadas no âmbito do SUS, incluindo diagnósticos, procedimentos realizados e desfecho da internação;
- **SIA-SUS** (Sistema de Informações Ambulatoriais do SUS): registra procedimentos ambulatoriais de média e alta complexidade;
- **Sinasc** (Sistema de Informações sobre Nascidos Vivos): registra nascimentos a partir das Declarações de Nascido Vivo;
- **GAL** (Gerenciador de Ambiente Laboratorial): registra exames laboratoriais realizados pela rede pública;
- **SITETB** (Sistema de Informação de Tratamentos Especiais da Tuberculose): registra casos de tuberculose drogarresistente e tratamentos especiais.

Uma característica fundamental dessas bases de dados é a **ausência de um identificador unívoco** que permita a vinculação inequívoca de registros referentes a um mesmo

indivíduo entre diferentes sistemas [?]. Embora o Cartão Nacional de Saúde (CNS) tenha sido concebido com essa finalidade, sua cobertura permanece incompleta e sua qualidade de preenchimento é heterogênea entre regiões e sistemas, limitando sua utilidade como chave primária para o *linkage* direto de bases [?]. Essa lacuna impõe a necessidade de métodos indiretos de relacionamento, baseados na comparação de variáveis de identificação comuns (como nome, nome da mãe, data de nascimento, sexo e município de residência), que, por sua natureza, estão sujeitos a erros de digitação, abreviações, homônimos e incompletude, comprometendo potencialmente a qualificação dos dados vinculados [?].

Persistem, portanto, desafios consideráveis para a integração dessas bases: a fragmentação dos sistemas, a heterogeneidade na qualidade dos dados de identificação, a ausência de padronização nos campos de preenchimento e as restrições legais e éticas associadas ao uso de dados pessoais em saúde [?]. Não obstante, o potencial analítico da vinculação dessas bases é expressivo, tendo sido demonstrado em estudos que identificaram subnotificação de tuberculose [?, ?], avaliaram a qualidade da informação por meio de cruzamentos entre SIM e Sinan [?] e construíram coortes populacionais de grande abrangência [?]. Nesse sentido, faz-se necessário o desenvolvimento de abordagens que ampliem a acurácia e a escalabilidade do processo de vinculação, com vistas à qualificação dos dados vinculados e à produção de informação oportuna para a vigilância em saúde.

1.3 Estratégias de *linkage*

Diferentes estratégias têm sido empregadas para o *linkage* em saúde, podendo ser agrupadas em três abordagens principais: o *linkage* determinístico, o *linkage* probabilístico e as abordagens baseadas em aprendizado de máquina. Cada uma dessas estratégias apresenta particularidades distintas em termos de sensibilidade, especificidade e custo computacional, cuja adequação depende das características das bases de dados envolvidas, da qualidade das variáveis de identificação disponíveis e dos objetivos do estudo [?]. O Quadro 1.1 sintetiza as principais vantagens e limitações de cada abordagem.

Tabela 1.1: Comparação entre estratégias de *linkage* de bases de dados.

Estratégia	**Vantagens**	**Limitações**
Determinístico	Simplicidade conceitual e computacional; elevada especificidade dos pares identificados; resultados facilmente auditáveis	Baixa sensibilidade na presença de erros de grafia, campos incompletos ou variações ortográficas; incapacidade de acomodar imperfeições nos dados
Probabilístico	Flexibilidade para acomodar imperfeições nos dados; possibilidade de atribuir pesos diferenciados por variável; ampla utilização no contexto brasileiro (OpenRecLink)	Dependência da calibração de limiares; geração de área cinza que demanda revisão manual; sensibilidade à qualidade dos parâmetros m e u
Aprendizado de máquina	Captura de padrões não lineares; incorporação de múltiplas variáveis e interações; potencial de automatização da classificação	Necessidade de conjunto de treinamento rotulado; sensibilidade ao desbalanceamento de classes; menor interpretabilidade de alguns modelos

1.4 A área cinza e a classificação de pares

No modelo de decisão proposto por Fellegi e Sunter [**?**], dois limiares de classificação dividem o espaço de escores em três regiões: acima do limiar superior, os pares são aceitos como verdadeiros; abaixo do limiar inferior, são descartados como não-pares; entre ambos os limiares, configura-se a denominada "área cinza" (*gray area*), composta por pares candidatos cujos escores de similaridade não são suficientemente elevados para serem aceitos automaticamente, nem suficientemente baixos para serem descartados [**?**, **?**].

A área cinza constitui um dos nós críticos do processo de *linkage*. Sua resolução adequada é determinante para a qualidade final dos vínculos. A extensão dessa região depende da qualidade das variáveis de identificação, do poder discriminatório das métricas de comparação empregadas e da definição dos limiares adotados. Em bases de dados com elevada proporção de campos incompletos, erros de grafia ou homônimos, a área cinza tende a ser expressiva, concentrando tanto pares verdadeiros que não puderam ser identificados com certeza quanto falsos positivos que apresentam semelhança fortuita [**?**].

Tradicionalmente, a resolução da área cinza é realizada por meio de revisão manual

(*clerical review*), na qual revisores humanos examinam individualmente cada par candidato e decidem sobre sua classificação. Esse processo, embora potencialmente acurado, apresenta limitações importantes: é demorado, custoso, não escalável para grandes volumes de dados e sujeito à variabilidade intra e interavaliador [?, ?]. Faz-se necessário, portanto, o desenvolvimento de abordagens automatizadas que possam auxiliar ou substituir a revisão manual, recuperando pares verdadeiros da área cinza e reduzindo a proporção de classificações incertas.

Nesse contexto, a aplicação de técnicas de aprendizado de máquina como etapa de pós-processamento do *linkage* probabilístico apresenta-se como estratégia promissora. Em vez de substituir integralmente o processo probabilístico, a abordagem proposta neste trabalho utiliza os escores de similaridade produzidos pelo comparador (bem como variáveis derivadas desses escores) como atributos (*features*) de entrada para classificadores supervisionados, que aprendem, a partir de exemplos rotulados, a distinguir pares verdadeiros de falsos positivos na região de incerteza. Essa estratégia possibilita a priorização de pares ou não-pares, conforme o objetivo do estudo: em investigações voltadas à identificação de subnotificação, pode-se optar pela maximização da sensibilidade (*recall*), aceitando-se uma taxa maior de falsos positivos; em estudos analíticos que requerem elevada confiabilidade dos vínculos, pode-se priorizar a especificidade e a precisão [?]. A combinação de abordagens automatizadas com regras que incorporem o conhecimento acumulado sobre as particularidades das bases de dados de saúde no Brasil constitui, assim, caminho promissor.

1.5 Desbalanceamento de classes no *linkage*

Uma característica intrínseca ao *linkage* é o severo desbalanceamento entre as classes de pares verdadeiros e não-pares. O fenômeno é estrutural, não contingente. Nas etapas iniciais do processo, a combinação de registros entre duas bases gera um número de pares candidatos que cresce de forma quadrática com o tamanho das bases, enquanto o número de pares verdadeiros cresce linearmente. Mesmo após a aplicação de estratégias de blocagem que restringem as comparações a subconjuntos de registros [?], a proporção

de pares verdadeiros em relação ao total de candidatos permanece tipicamente muito pequena [?].

Esse desbalanceamento impõe desafios específicos para a aplicação de algoritmos de classificação, que tendem a apresentar viés em favor da classe majoritária (os não-pares) quando treinados em conjuntos desbalanceados, resultando em modelos que, embora apresentem elevada acurácia global, podem falhar na identificação de pares verdadeiros [?, ?]. Essa limitação é particularmente relevante no contexto do *linkage* em saúde, no qual a perda de pares verdadeiros pode comprometer a validade de estudos epidemiológicos [?].

Diferentes estratégias têm sido propostas para mitigar o efeito do desbalanceamento de classes em problemas de classificação, podendo ser agrupadas em abordagens de reamostragem, algoritmos sensíveis ao custo e técnicas de *ensemble*:

- **Sobreamostragem da classe minoritária**: consiste na geração de exemplos sintéticos da classe sub-representada, sendo o algoritmo SMOTE (*Synthetic Minority Over-sampling Technique*) [?] a técnica mais amplamente empregada. Variações do SMOTE incluem o Borderline-SMOTE, que concentra a geração de exemplos sintéticos nas regiões de fronteira entre as classes, e o ADASYN (*Adaptive Synthetic Sampling*), que adapta a densidade de geração de acordo com o grau de dificuldade de classificação de cada exemplo;

- **Subamostragem da classe majoritária**: consiste na remoção aleatória de exemplos da classe majoritária para equilibrar a distribuição. Técnicas combinadas, como o SMOTE-Tomek, integram sobreamostragem com remoção de exemplos ruidosos nas fronteiras de decisão;

- **Ponderação de classes (*class weights*)**: consiste na atribuição de pesos diferenciados às classes na função de custo do classificador, penalizando mais severamente os erros cometidos sobre a classe minoritária, sem alterar a composição do conjunto de treinamento;

- **Métodos de *ensemble***: combinam múltiplos classificadores para obter predições mais robustas, podendo incorporar estratégias de balanceamento em cada iteração,

como o *Balanced Random Forest* e o EasyEnsemble [?].

A escolha da estratégia de balanceamento adequada depende das características do problema e dos objetivos do estudo, não havendo consenso na literatura sobre uma abordagem universalmente superior [?]. No que se refere ao *linkage* em bases de saúde, a comparação sistemática de diferentes estratégias de balanceamento constitui, portanto, questão relevante e ainda insuficientemente explorada, cujos resultados podem contribuir para a padronização de protocolos e a melhoria do desempenho de sistemas de informação em saúde.

1.6 Ajustes nos escores e pontos de corte do comparador

Os escores de similaridade produzidos por ferramentas de *linkage* probabilístico, como o OpenRecLink [?], representam uma medida agregada do grau de concordância entre os campos de identificação de cada par candidato. A classificação final dos pares depende da definição de pontos de corte (*thresholds*) sobre esses escores, que delimitam as fronteiras entre pares aceitos, pares rejeitados e a área cinza [?].

A definição de pontos de corte adequados não é uma tarefa trivial. Não há solução analítica geral para esse problema. Pontos de corte excessivamente elevados resultam em alta especificidade, porém com perda de pares verdadeiros que apresentam escores intermediários, frequentemente aqueles com campos de identificação incompletos ou com erros de grafia. Pontos de corte excessivamente baixos, por outro lado, incorporam falsos positivos ao conjunto de pares aceitos, comprometendo a confiabilidade dos vínculos [?]. O desafio reside, portanto, em encontrar um equilíbrio entre sensibilidade e especificidade que seja adequado aos objetivos do estudo.

Neste trabalho, propõe-se a utilização de técnicas de aprendizado de máquina para a análise e otimização dos pontos de corte do comparador, empregando os escores de similaridade individuais (e não apenas o escore composto) como variáveis preditoras. Essa abordagem possibilita a identificação de padrões nos escores que podem indicar pares verdadeiros mesmo em regiões de escore composto intermediário, como situações em que

determinados campos apresentam concordância elevada enquanto outros apresentam discordância explicável por erros de preenchimento. A otimização dos pontos de corte pode ser orientada por diferentes métricas de avaliação, como a medida-F (*F-measure*) [?], a área sob a curva ROC (*AUC-ROC*) ou a área sob a curva precisão-sensibilidade (*AUC-PR*), sendo esta última particularmente adequada para cenários com desbalanceamento severo de classes. Nesse sentido, a integração de técnicas de aprendizado de máquina ao processo de definição de limiares tem potencial para aprimorar a qualificação dos dados vinculados, contribuindo para a produção de indicadores mais acurados e para a melhoria da vigilância epidemiológica.

1.7 Técnicas de aprendizado de máquina utilizadas

Para a classificação de pares candidatos e a recuperação de pares verdadeiros da área cinza, foram empregadas neste trabalho diferentes técnicas de aprendizado de máquina supervisionado, selecionadas por suas propriedades complementares no tratamento de dados desbalanceados e na modelagem de padrões não lineares de similaridade entre campos de identificação.

1.7.1 Regressão Logística

A regressão logística constitui um modelo linear generalizado que estima a probabilidade de um evento binário (neste caso, a pertinência de um par candidato à classe de pares verdadeiros) por meio de uma função logística aplicada a uma combinação linear das variáveis preditoras [?]. Apesar de sua simplicidade, a regressão logística apresenta vantagens importantes como modelo de referência (*baseline*): é computacionalmente eficiente, facilmente interpretável e produz probabilidades calibradas, possibilitando a análise direta da contribuição de cada variável para a classificação.

1.7.2 Floresta Aleatória (*Random Forest*)

O algoritmo de Floresta Aleatória, proposto por Breiman [?], baseia-se na construção de um conjunto (*ensemble*) de árvores de decisão treinadas em amostras aleatórias dos dados de treinamento, com seleção aleatória de subconjuntos de variáveis em cada nó

de decisão. A predição final é obtida por votação majoritária entre as árvores individuais. Essa técnica apresenta elevada robustez ao sobreajuste (*overfitting*), capacidade de captura de relações não-lineares e mecanismos intrínsecos para avaliação da importância relativa das variáveis preditoras, sendo amplamente utilizada em problemas de classificação com desbalanceamento de classes.

1.7.3 Gradient Boosting e suas variantes

Os métodos de *Gradient Boosting* constituem uma família de algoritmos de *ensemble* que constroem modelos de forma sequencial, em que cada novo modelo é treinado para corrigir os erros residuais dos modelos anteriores [**?**]. Neste trabalho, foram empregadas duas implementações de *Gradient Boosting*:

- **XGBoost** (*eXtreme Gradient Boosting*) [**?**]: implementação otimizada que incorpora regularização L1 e L2 na função objetivo, tratamento nativo de valores ausentes e paralelização do treinamento, tendo demonstrado desempenho superior em diversas competições de ciência de dados e aplicações em saúde;

- **LightGBM** (*Light Gradient Boosting Machine*): implementação que utiliza estratégias de amostragem baseadas em gradiente e agrupamento de variáveis para reduzir o custo computacional, mantendo acurácia comparável ao XGBoost, sendo particularmente eficiente para conjuntos de dados de grande porte.

1.7.4 Máquina de Vetores de Suporte (*Support Vector Machine, SVM*)

A Máquina de Vetores de Suporte busca encontrar o hiperplano de separação ótimo entre classes, maximizando a margem de separação entre os exemplos mais próximos de cada classe [**?**]. A utilização de funções de núcleo (*kernel*), como o núcleo de base radial (*Radial Basis Function, RBF*), permite a projeção dos dados em espaços de dimensionalidade superior, possibilitando a separação de classes que não são linearmente separáveis no espaço original.

1.7.5 Rede Neural Artificial (*Multilayer Perceptron, MLP*)

O Perceptron Multicamadas constitui uma rede neural artificial composta por camadas de neurônios artificiais interconectados, que aprendem representações hierárquicas dos dados por meio do algoritmo de retropropagação do erro [?]. A flexibilidade arquitetural dessa técnica permite a modelagem de relações altamente não-lineares entre as variáveis de entrada, embora apresente maior sensibilidade à configuração de hiperparâmetros e ao desbalanceamento de classes.

1.7.6 Métodos de *Ensemble* combinados

Além dos métodos de *ensemble* baseados em *boosting*, foram empregadas técnicas de combinação de classificadores, cuja eficácia para o *linkage* tem sido demonstrada por meio de estudos comparativos entre *bagging*, *bumping*, co-treinamento multiview e aprendizado ativo [?]:

- ***Stacking* (empilhamento)**: consiste no treinamento de um metaclassificador que aprende a combinar as predições de múltiplos classificadores de base, utilizando como entrada as probabilidades preditas por cada modelo individual. Neste trabalho, foram utilizados como classificadores de base XGBoost, LightGBM, Floresta Aleatória e *Gradient Boosting*, com Regressão Logística como metaclassificador;
- **Votação por consenso (*Consensus Voting*)**: classificação baseada na unanimidade ou maioria das predições de múltiplos classificadores independentes, com a vantagem de aumentar a confiabilidade das predições positivas ao exigir concordância entre modelos distintos.

1.7.7 Regras de negócio baseadas em conhecimento de domínio

Complementarmente às técnicas de aprendizado de máquina, foram desenvolvidas regras de classificação baseadas em conhecimento de domínio específico do *linkage* em saúde. Essas regras incorporam critérios de pontuação que consideram a qualidade e a concordância de campos específicos (como a correspondência exata da data de nascimento, a similaridade de nomes acima de limiares pré-estabelecidos e a concordância no município

de residência), podendo ser empregadas isoladamente ou em combinação com modelos de aprendizado de máquina (abordagem híbrida), potencializando a capacidade de classificação ao integrar evidências estatísticas e conhecimento especializado. Essa integração entre classificadores treinados e critérios derivados da prática do *linkage* de bases de saúde constitui aspecto central da abordagem proposta neste trabalho, conforme detalhado no Capítulo 4.

1.8 Estratégia adotada e protocolos de *linkage*

A estratégia adotada neste trabalho consiste no emprego de técnicas de aprendizado de máquina como camada de pós-processamento do *linkage* probabilístico. Os escores de similaridade produzidos pelo comparador OpenRecLink [?], bem como variáveis derivadas desses escores, são utilizados como atributos de entrada para classificadores supervisionados. A abordagem não substitui o processo probabilístico, mas o complementa, focalizando a resolução automatizada da área cinza e a reclassificação de pares situados nas regiões de incerteza do espaço de escores. Abordagens semelhantes têm sido reportadas na literatura internacional com resultados promissores [?, ?, ?].

A escolha dessa estratégia fundamenta-se em três considerações. Primeira, o *linkage* probabilístico já se encontra consolidado no contexto brasileiro e constitui a etapa inicial do processo de vinculação entre as bases do SIM e do Sinan-TB utilizadas neste estudo [?, ?]. Segunda, a área cinza produzida pelo comparador concentra pares cuja classificação não é trivial, demandando a captura de padrões não lineares de concordância que excedem a capacidade de um modelo baseado em combinação linear de escores [?, ?]. Terceira, a disponibilidade de um padrão-ouro obtido por revisão manual viabiliza o treinamento supervisionado, condição necessária para a aplicação de classificadores [?].

Operacionalmente, o protocolo proposto estrutura-se nas seguintes etapas: (i) execução do *linkage* probabilístico mediante OpenRecLink, com múltiplos passos de blocagem [?]; (ii) exportação dos pares candidatos com seus escores de similaridade individuais e escore composto; (iii) engenharia de atributos, incluindo indicadores binários de concordância, escores ponderados, termos de interação e variáveis derivadas do conhecimento

de domínio; (iv) treinamento de classificadores supervisionados sobre o conjunto rotulado, com avaliação por validação cruzada estratificada; (v) aplicação dos classificadores treinados aos pares da área cinza, com possibilidade de priorização de pares (*recall* máximo) ou não-pares (*precision* máxima), conforme o objetivo do estudo; e (vi) integração de regras de negócio baseadas no conhecimento acumulado sobre as particularidades das bases de saúde brasileiras.

Esse protocolo foi desenhado para ser reprodutível e automatizável, executável em ambiente computacional padronizado (Python, *scikit-learn*, *Jupyter/Papermill*) e versionável em repositório Git, atendendo à necessidade de padronização identificada na literatura sobre *linkage* em saúde [**?**, **?**]. Estudos de simulação em larga escala corroboram a importância dessa padronização, demonstrando que diferentes escolhas metodológicas podem produzir viés sistemático na qualidade da vinculação [**?**]. A descrição detalhada de cada etapa encontra-se no Capítulo 4.

Capítulo 2

Justificativa

O *linkage* (relacionamento de registros), conforme apresentado no Capítulo 1, constitui etapa indispensável para a integração de dados entre os múltiplos Sistemas de Informação em Saúde do Brasil [?, ?]. A ausência de um identificador unívoco que perpasse bases como o SIM, o Sinan e o SIH-SUS torna necessário o emprego de métodos probabilísticos baseados na comparação de variáveis de identificação pessoal [?, ?], cujos resultados dependem diretamente da qualidade dos dados disponíveis e da adequação dos limiares de classificação adotados.

O *linkage* probabilístico, fundamentado no modelo de Fellegi e Sunter [?], é amplamente empregado em estudos epidemiológicos brasileiros por meio de ferramentas como o OpenRecLink [?] e estratégias de blocagem [?]. Persistem, contudo, desafios na classificação dos pares situados na "área cinza" do comparador (cf. Seção 1.4): essa faixa intermediária de escores demanda revisão manual (*clerical review*), procedimento dispendioso, pouco escalável e sujeito à variabilidade inter-avaliadores [?]. A resolução automatizada da área cinza constitui, portanto, o nó crítico que motiva a presente investigação.

2.1 Lacuna do conhecimento

Embora a aplicação de técnicas de aprendizado de máquina (*machine learning*) ao *linkage* venha sendo investigada em contextos internacionais [?, ?, ?], a literatura brasileira sobre o tema é incipiente, restringindo-se predominantemente a abordagens determinísticas e probabilísticas tradicionais. No cenário internacional, estudos de simulação em larga escala demonstraram que a escolha do método de vinculação pode introduzir viés

sistemático nas estimativas populacionais [?]. Estudos nacionais que empreguem classificadores supervisionados como pós-processamento do *linkage* probabilístico, com vistas a automatizar a recuperação de pares na área cinza e a identificação de falsos positivos, são escassos. Soma-se a isso a carência de investigações que avaliem sistematicamente o impacto de diferentes estratégias de balanceamento de classes e de ajustes nos pontos de corte do comparador sobre a acurácia do processo de vinculação, apontando para uma lacuna relevante no campo da produção de dados vinculados em saúde no Brasil.

A maioria dos estudos brasileiros emprega protocolos padronizados de *linkage* probabilístico cujos limiares são definidos empiricamente, sem análise sistemática da sensibilidade dos resultados a variações nesses parâmetros [?, ?]. A revisão manual da área cinza, quando realizada, configura etapa artesanal e não reprodutível, comprometendo a comparabilidade entre estudos [?]. Essa limitação agrava-se em contextos de crises sanitárias, nos quais a produção de informação oportuna é requisito para a tomada de decisão.

Faz-se necessário, portanto, investigar abordagens que reduzam a dependência da revisão manual e ampliem a recuperação de pares verdadeiros na área cinza. Estudos recentes em outros contextos demonstraram que abordagens baseadas em *ensemble* de classificadores [?] e métodos híbridos que combinam técnicas probabilísticas com aprendizado supervisionado [?, ?] obtêm ganhos expressivos na acurácia da vinculação. A transposição e a adaptação dessas abordagens ao contexto brasileiro, com suas particularidades em termos de qualidade de dados e volume de registros, constitui contribuição relevante e ainda não explorada.

2.2 Justificativas específicas

Algumas justificativas específicas fundamentam a relevância deste estudo:

1. **Subnotificação da tuberculose e desfechos desfavoráveis.** A tuberculose (TB), reconhecida como condição marcadora da qualidade do cuidado em saúde [?], permanece como problema de saúde pública de grande magnitude no Brasil, com taxas de cura abaixo do preconizado pela Organização Mundial da Saúde e proporção não negligenciável de desfechos desfavoráveis, incluindo óbito, abandono

e resistência medicamentosa [?, ?]. Estudos anteriores demonstraram que o *linkage* entre as bases do SIM e do Sinan-TB permite a identificação de óbitos por tuberculose não notificados ao sistema de vigilância, evidenciando subnotificação significativa [?, ?, ?]. A melhoria na acurácia desse *linkage* tem potencial para ampliar a capacidade de detecção de casos e a qualificação da informação epidemiológica, com impacto direto sobre a análise de causas múltiplas de óbito em coortes de pacientes com TB [?].

2. **Intenso desbalanceamento de classes no *linkage* SIM–Sinan.** O *linkage* entre bases de mortalidade e de agravos de notificação gera um volume de pares candidatos no qual os pares verdadeiros constituem fração extremamente reduzida, frequentemente inferior a 1% do total de comparações [?]. Esse desbalanceamento representa nó crítico para classificadores supervisionados e requer estratégias específicas de tratamento, cuja efetividade comparativa no contexto do *record linkage* em saúde não se encontra adequadamente documentada na literatura brasileira, demandando investigação aprofundada. Hassani e colaboradores [?] propuseram recentemente uma estratégia combinada de sobreamostragem e subamostragem especificamente desenhada para *linkage* de grande escala, evidenciando que o tratamento adequado do desbalanceamento pode elevar substancialmente o desempenho dos classificadores.

3. **Necessidade de protocolos reprodutíveis e automatizados.** A produção de dados vinculados para fins de vigilância epidemiológica e de pesquisa em serviços de saúde demanda agilidade e reprodutibilidade, especialmente em contextos de crises sanitárias nas quais a informação oportuna é requisito para a tomada de decisão no cuidado em saúde [?]. A automatização de etapas do processo de classificação, mediante algoritmos treinados e validados, pode contribuir para a construção de protocolos padronizados de *linkage* que reduzam a variabilidade e ampliem a escalabilidade do método. Nessa direção, diretrizes metodológicas internacionais já recomendam a integração de técnicas de aprendizado de máquina a dados vinculados para a estimação de indicadores populacionais de saúde [?].

4. **Potencial de generalização para outros pares de bases de dados.** Embora o presente estudo tome como caso aplicado o *linkage* SIM–Sinan-TB, as abordagens metodológicas desenvolvidas, incluindo as estratégias de balanceamento, os ajustes nos pontos de corte e os modelos de classificação, possuem potencial de aplicação a outros cenários de vinculação de bases de saúde no Brasil, como SIH-SUS–Sinan, SIM–Sinasc, entre outros, ampliando o alcance das contribuições para a produção de indicadores de desempenho do sistema de saúde.

5. **Experiência institucional acumulada.** O Laboratório de Linkage e Análise de Dados Populacionais do Instituto de Estudo em Saúde Coletiva (IESC) da Universidade Federal do Rio de Janeiro (UFRJ) possui experiência de mais de 20 anos no *linkage* de bases de dados de saúde no Brasil [**?**, **?**, **?**]. Essa trajetória institucional fornece base sólida para o desenvolvimento e a validação de novas abordagens metodológicas, na medida em que dispõe de bases de dados previamente relacionadas, protocolos consolidados e equipe multidisciplinar com conhecimento tanto da área de saúde quanto de ciência de dados, potencializando a produção de conhecimento novo e útil para o campo da saúde coletiva.

2.3 A tuberculose como condição marcadora

A escolha da tuberculose (TB) como condição de estudo neste trabalho fundamenta-se no conceito de condições traçadoras (*tracer conditions*), proposto por Kessner, Kalk e Singer [**?**], segundo o qual determinadas condições de saúde podem funcionar como reveladoras do desempenho do sistema assistencial, desde que sejam inequivocamente identificáveis, possuam prevalência suficiente, tenham história natural modificável pela intervenção e disponham de técnicas de manejo bem estabelecidas. A TB atende a todos esses requisitos: é doença de notificação compulsória, registrada em múltiplos SIS (Sinan, SIM, SIH-SUS, GAL, SITETB), cujo tratamento é padronizado e disponibilizado integralmente pelo SUS [**?**]. Essas propriedades permitem que o percurso do paciente com TB na rede de serviços seja rastreável por meio do *linkage*, revelando atrasos no diagnóstico, irregularidade no tratamento, abandono, internações evitáveis e óbitos que poderiam ter

sido prevenidos [?, ?].

Estudos conduzidos pelo grupo de pesquisa do IESC/UFRJ demonstraram que o *linkage* entre o SIM e o Sinan-TB identificou óbitos por TB não notificados ao sistema de vigilância, evidenciando subnotificação expressiva [?, ?, ?]. Investigações subsequentes qualificaram variáveis do Sinan-TB por meio de regras de *scripting* aplicadas sobre dados vinculados [?] e analisaram as causas múltiplas de morte em coortes de pacientes notificados [?]. A taxa de cura no Brasil permanece abaixo do preconizado pela Organização Mundial da Saúde, e os índices de abandono persistem elevados [?, ?], indicando que a TB continua a revelar fragilidades na organização do cuidado. A melhoria da acurácia do *linkage* entre essas bases tem, portanto, implicações diretas para a avaliação da efetividade do programa de controle da tuberculose.

2.4 Urgência em contextos de crises sanitárias

A necessidade de protocolos automatizados e reprodutíveis de *linkage* é acentuada em contextos de crises sanitárias. A pandemia de COVID-19 provocou sobrecarga nos serviços de saúde, com redução documentada no número de notificações de tuberculose, interrupção de tratamentos e aumento de desfechos desfavoráveis [?, ?]. A queda na detecção de casos durante a pandemia não refletiu redução na incidência da doença, mas a retração do acesso a diagnóstico e a desarticulação de rotinas de vigilância [?]. Cenários semelhantes podem ocorrer em crises climáticas e epidêmicas futuras, reforçando a importância de dispor de métodos de *linkage* que possam ser executados de forma ágil e padronizada, sem depender exclusivamente de revisão manual.

2.5 Vinculação institucional

O presente trabalho insere-se no programa de pós-graduação do Instituto de Estudos em Saúde Coletiva (IESC) da Universidade Federal do Rio de Janeiro (UFRJ), no âmbito da linha de pesquisa em Ciência de Dados aplicada à Saúde. O IESC abriga o Laboratório de Linkage e Análise de Dados Populacionais, que desenvolve, há mais de duas décadas, metodologias de vinculação de bases de dados para a vigilância epidemiológica e a avali-

ação de serviços de saúde [?, ?]. O estudo conta ainda com a colaboração da Secretaria Acadêmica de Saúde, que articula atividades de ensino, pesquisa e extensão voltadas à qualificação dos dados em saúde e ao fortalecimento da capacidade analítica dos sistemas de informação do SUS. Essa vinculação institucional assegura o acesso a bases de dados previamente vinculadas, protocolos consolidados e expertise multidisciplinar necessários para o desenvolvimento e a validação das abordagens propostas.

Capítulo 3

Objetivos

3.1 Objetivo geral

O presente estudo tem como objetivo geral desenvolver e avaliar algoritmos baseados em aprendizado de máquina (*machine learning*) para o pós-processamento do *linkage* probabilístico entre bases de dados de saúde, com vistas a aumentar a acurácia do processo de classificação de pares e recuperar registros da área cinza que permaneceriam não classificados ou incorretamente descartados pelo método probabilístico convencional, contribuindo para a qualificação dos dados vinculados e para a produção de indicadores de desempenho de sistemas e serviços de saúde.

3.2 Objetivos específicos

1. Comparar o desempenho de diferentes técnicas de aprendizado de máquina, a saber: regressão logística, Floresta Aleatória (*Random Forest*), *Gradient Boosting* (XGBoost e LightGBM), Máquina de Vetores de Suporte (*Support Vector Machine, SVM*), redes neurais artificiais (*Multilayer Perceptron, MLP*) e métodos de combinação de modelos (*ensemble*: *Stacking* e votação por consenso), na tarefa de classificação de pares candidatos produzidos pelo *linkage* probabilístico entre o Sistema de Informação sobre Mortalidade (SIM) e o Sistema de Informação de Agravos de Notificação para tuberculose (Sinan-TB).

2. Avaliar e comparar estratégias de balanceamento de classes, incluindo *SMOTE* [**?**], *Borderline-SMOTE*, *ADASYN*, *SMOTE-Tomek* e ponderação de classes (*class*

weights), quanto ao seu efeito sobre a sensibilidade e a especificidade dos classificadores, considerando o severo desbalanceamento inerente ao *linkage*, no que tange à identificação de combinações que possibilitem a melhoria da acurácia do processo de vinculação.

3. Desenvolver e avaliar protocolos de ajuste nos pontos de corte dos escores do comparador, empregando otimização de limiares (*threshold optimization*) e regras de negócio baseadas no conhecimento do domínio, de modo a maximizar a recuperação de pares verdadeiros na área cinza sem comprometer a proporção de falsos positivos, contribuindo para a qualificação dos dados vinculados e para a melhoria do desempenho do comparador probabilístico.

4. Comparar duas estratégias complementares de pós-processamento: uma orientada à maximização da sensibilidade (*recall*), voltada à identificação de subnotificação e à recuperação exaustiva de pares, e outra orientada à maximização da precisão (*precision*), voltada à construção de conjuntos analíticos de alta confiabilidade; avaliando as implicações de cada abordagem para diferentes finalidades de uso dos dados vinculados no âmbito da vigilância, da assistência e da gestão em saúde.

5. Sistematizar os resultados das comparações em quadros e tabelas que possibilitem a reprodução dos experimentos e a identificação das combinações de técnicas, parâmetros e estratégias de balanceamento mais adequadas a cada cenário de aplicação do *linkage* em saúde, com vistas à produção de protocolos reprodutíveis e à padronização de abordagens de pós-processamento.

6. Avaliar o potencial de generalização das abordagens desenvolvidas para outros cenários de *linkage* em saúde, discutindo as condições sob as quais os classificadores treinados e os protocolos propostos podem ser adaptados a diferentes pares de bases de dados e a distintos contextos epidemiológicos.

Capítulo 4

Método

4.1 Desenho do estudo

Trata-se de um estudo metodológico de desenvolvimento e avaliação de algoritmos de aprendizado de máquina (*machine learning*) aplicados ao pós-processamento do *linkage* probabilístico entre bases de dados de saúde. O estudo utiliza dados secundários provenientes de sistemas nacionais de informação em saúde, vinculados por meio de técnicas probabilísticas, e propõe protocolos computacionais para a melhoria da acurácia na classificação de pares candidatos, com vistas à qualificação dos dados vinculados e à produção de indicadores para a vigilância epidemiológica.

4.2 Fontes de dados

Foram utilizados registros provenientes de duas bases de dados nacionais de saúde:

- **Sistema de Informação sobre Mortalidade (SIM):** base de dados que registra todos os óbitos ocorridos no território nacional, a partir das Declarações de Óbito (DO), contendo variáveis demográficas (nome, data de nascimento, nome da mãe, sexo), geográficas (município de residência, endereço) e relativas à causa do óbito codificada pela Classificação Internacional de Doenças (CID-10) [**?**].

- **Sistema de Informação de Agravos de Notificação, Tuberculose (Sinan-TB):** base que registra os casos de tuberculose notificados compulsoriamente no Brasil, contendo variáveis de identificação do paciente, dados clínicos, laboratoriais

e de acompanhamento do tratamento, incluindo a situação de encerramento do caso [?, ?].

Os registros correspondem ao município do Rio de Janeiro e foram previamente submetidos a *linkage* probabilístico por meio do programa OpenRecLink [?], gerando uma base de pares candidatos que constitui o objeto de análise do presente estudo. A escolha dessas bases justifica-se pelas razões detalhadas na Seção 2.2, em particular a relevância da tuberculose como condição marcadora da qualidade do cuidado em saúde e a experiência acumulada do grupo de pesquisa no *linkage* dessas fontes de dados [?, ?, ?].

4.3 Base de pares candidatos

A base de pares candidatos utilizada contém registros classificados pelo OpenRecLink a partir de múltiplos passos de blocagem (*blocking steps*), conforme recomendado na literatura para maximizar a sensibilidade do processo [?]. Cada par candidato é representado por um conjunto de escores de similaridade calculados para as variáveis de identificação disponíveis em ambas as bases:

- Escores de similaridade para o **nome** do indivíduo (fragmentos e variações)
- Escores de similaridade para o **nome da mãe**
- Escore de concordância para a **data de nascimento**
- Escore de concordância para o **município de residência**
- Escores de similaridade para o **endereço**
- **Escore final composto** (*nota final*) calculado pelo OpenRecLink
- **Passo de blocagem** em que o par foi identificado

Cada par possui uma classificação de referência (padrão-ouro) atribuída por revisão manual, categorizada em: par verdadeiro confirmado, par verdadeiro provável e não-par. Para fins de modelagem, os pares verdadeiros confirmados e prováveis foram agrupados

em uma única classe positiva, resultando em uma variável-alvo binária. A disponibilidade dessa classificação de referência possibilita o treinamento e a avaliação dos classificadores supervisionados, na medida em que fornece as observações rotuladas necessárias para o aprendizado.

A base apresenta severo desbalanceamento de classes, com proporção aproximada de 1 par verdadeiro para cada 250 não-pares (cerca de 0,4% de registros positivos), característica inerente ao *linkage* em que o número de combinações candidatas cresce de forma quadrática enquanto os pares verdadeiros crescem linearmente [**?**, **?**]. Esse desbalanceamento constitui nó crítico para a aplicação de classificadores supervisionados, demandando estratégias específicas de tratamento.

4.4 Engenharia de atributos

A partir dos escores brutos de similaridade fornecidos pelo comparador, procedeu-se à derivação de atributos adicionais (*features*) com o objetivo de enriquecer a representação de cada par candidato e possibilitar a captura de padrões não lineares de concordância entre registros. As estratégias de engenharia de atributos incluíram:

- **Indicadores binários de concordância:** variáveis dicotômicas indicando se o escore de similaridade de cada campo ultrapassa limiares predefinidos (concordância "perfeita" e concordância "alta"), com limiares ajustados de acordo com a estratégia de análise empregada, mais permissivos para a estratégia de máximo *recall* e mais restritivos para a estratégia de máxima precisão.

- **Escores agregados e ponderados:** combinações lineares dos escores individuais, atribuindo pesos diferenciados conforme o poder discriminatório de cada variável: maior peso para nome e data de nascimento, peso intermediário para nome da mãe e município, menor peso para endereço.

- **Termos de interação:** produtos entre escores de campos distintos, possibilitando a captura da concordância simultânea de múltiplas variáveis (por exemplo, nome ×

data de nascimento × nome da mãe), cujo valor conjunto pode ser mais informativo do que os escores individuais isoladamente.

- **Indicador de óbito por tuberculose:** variável derivada da situação de encerramento do caso no Sinan-TB, sinalizando registros cuja causa de encerramento indica óbito por tuberculose ou óbito por outras causas, incorporando conhecimento de domínio relevante para a priorização de pares.

4.5 Estratégias de análise

O estudo foi estruturado em três etapas analíticas complementares, cada uma orientada por um objetivo distinto, a saber: comparação ampla de classificadores, maximização da sensibilidade e maximização da precisão, configurando protocolos experimentais que exploram diferentes compromissos entre falsos positivos e falsos negativos na classificação de pares, conforme a estratégia delineada na Seção 1.8.

A seleção dos classificadores empregados neste estudo foi orientada por dois critérios: (i) a representatividade de diferentes famílias de modelos, de modo a cobrir abordagens lineares (regressão logística), baseadas em árvores (Floresta Aleatória, *Gradient Boosting*), baseadas em margens (SVM) e redes neurais (MLP), possibilitando a comparação entre paradigmas de aprendizado distintos; e (ii) a evidência prévia de bom desempenho em problemas com desbalanceamento severo de classes e em aplicações de *linkage* documentadas na literatura [?, ?, ?]. A regressão logística foi incluída como modelo de referência (*baseline*), por sua interpretabilidade e ampla utilização na área de saúde.

Cabe distinguir dois componentes do *pipeline* experimental: os *métodos padrão*, que compreendem o *linkage* probabilístico realizado pelo OpenRecLink [?] com parâmetros e limiares convencionais, representando a prática corrente no contexto brasileiro; e os *métodos propostos*, que compreendem a camada de pós-processamento por aprendizado de máquina, incluindo as estratégias de balanceamento, a engenharia de atributos e as regras de negócio desenvolvidas neste trabalho. Essa distinção permite avaliar o ganho incremental proporcionado pelos métodos propostos em relação ao processo probabilístico convencional.

4.5.1 Análise comparativa de técnicas

Na primeira etapa, procedeu-se à comparação ampla de diferentes classificadores de aprendizado de máquina aplicados à tarefa de classificação de pares. Foram avaliados: regressão logística, Floresta Aleatória (*Random Forest*), *Gradient Boosting*, Máquina de Vetores de Suporte (*SVM*) com núcleo de base radial (*Radial Basis Function, RBF*), rede neural *Multilayer Perceptron* (MLP), Floresta Aleatória com *SMOTE* e combinação por empilhamento (*Stacking Ensemble*). Para cada classificador, foram calculadas métricas de desempenho mediante validação cruzada estratificada, incluindo precisão, sensibilidade, medida-F_1, AUC-ROC e AUC-PR [?, ?]. Adicionalmente, foram geradas curvas de otimização de limiares e análises de importância de atributos, possibilitando a identificação das variáveis de maior poder discriminatório para a classificação de pares.

4.5.2 Estratégia de maximização da sensibilidade

Na segunda etapa, o foco recaiu sobre a maximização da sensibilidade (*recall*), buscando recuperar o maior número possível de pares verdadeiros, particularmente aqueles situados na área cinza do comparador. Para tanto, foram empregadas técnicas de balanceamento de classes (*SMOTE* [?], *Borderline-SMOTE*, *ADASYN* e *SMOTE-Tomek*), seguindo a lógica de combinação de sobreamostragem e subamostragem proposta por Hassani e colaboradores [?], combinadas com classificadores configurados para minimizar falsos negativos: Floresta Aleatória com pesos de classe, *AdaBoost* com sobreamostragem, MLP com *SMOTE* em proporção 1:1, votação suave (*Soft Voting Ensemble*) e classificador em cascata (*Cascade Classifier*) de dois estágios. Os limiares de classificação foram ajustados para valores baixos (entre 0,10 e 0,30), priorizando sensibilidade sobre precisão. Esta abordagem é particularmente relevante para estudos de subnotificação, nos quais a não identificação de um par verdadeiro pode resultar em subestimação da magnitude de desfechos desfavoráveis, comprometendo potencialmente a avaliação do desempenho do sistema de saúde e a identificação de nós críticos no itinerário terapêutico do paciente [?, ?].

4.5.3 Estratégia de maximização da precisão

Na terceira etapa, o foco direcionou-se à maximização da precisão (*precision*), buscando identificar apenas os pares de alta confiabilidade e minimizar falsos positivos. Foram empregados classificadores com forte regularização (XGBoost [?] e LightGBM), Floresta Aleatória calibrada por regressão isotônica, combinação por empilhamento (*Stacking*) com meta-aprendiz de regressão logística, e votação por consenso (unanimidade). Complementarmente, foram desenvolvidas regras de negócio baseadas no conhecimento do domínio, atribuindo pontuação a critérios como qualidade do nome, concordância exata de data de nascimento, similaridade do nome da mãe, concordância de município e endereço, e escore do comparador. Uma abordagem híbrida combinando classificadores de aprendizado de máquina com regras de negócio foi também avaliada. Os limiares foram ajustados para valores elevados (entre 0,60 e 0,90), priorizando a precisão. Esta estratégia é adequada para a construção de conjuntos analíticos de alta confiabilidade, nos quais a inclusão de falsos positivos poderia introduzir viés nas estimativas de associação, sendo, portanto, promissora para estudos que requeiram elevada qualificação dos dados vinculados. Cabe notar que, conforme argumentado por Hand e colaboradores [?], a medida-F pode não refletir adequadamente o desempenho do classificador nesse cenário, recomendando-se a utilização complementar de métricas como AUC-PR.

4.6 Métricas de avaliação

O desempenho dos classificadores foi avaliado por meio das seguintes métricas, reconhecidas na literatura de *linkage* e de aprendizado de máquina [?, ?]:

- **Precisão** (*Precision*): proporção de pares classificados como verdadeiros que são efetivamente pares verdadeiros.
- **Sensibilidade** (*Recall*): proporção de pares verdadeiros corretamente identificados dentre todos os pares verdadeiros existentes.
- **Medida-F$_1$** (F_1-*Score*): média harmônica entre precisão e sensibilidade, sintetizando o equilíbrio entre ambas.

- **AUC-ROC**: área sob a curva *Receiver Operating Characteristic*, que avalia a capacidade discriminatória do classificador em diferentes limiares de classificação.
- **AUC-PR**: área sob a curva Precisão-Sensibilidade (*Precision-Recall*), métrica particularmente informativa em cenários de desbalanceamento severo de classes, na medida em que é menos influenciada pela grande quantidade de verdadeiros negativos [?].

A avaliação foi realizada mediante validação cruzada estratificada, preservando a proporção de classes em cada partição, e os resultados foram apresentados em quadros comparativos que possibilitam a identificação das combinações de técnica, estratégia de balanceamento e limiar mais adequadas a cada cenário de uso, com vistas à padronização de protocolos de pós-processamento para o *linkage* em saúde.

4.7 Ambiente computacional

Todos os experimentos foram implementados na linguagem Python (versão 3.11), utilizando as bibliotecas *scikit-learn* para os classificadores de aprendizado de máquina e métricas de avaliação, *imbalanced-learn* para as técnicas de balanceamento de classes, *XGBoost* e *LightGBM* para os algoritmos de *Gradient Boosting*, e *pandas* e *NumPy* para manipulação e transformação de dados. Os experimentos foram estruturados em cadernos Jupyter (*Jupyter Notebooks*), executados de forma reprodutível por meio do arcabouço (*framework*) *Papermill*, e versionados em repositório Git. A execução automatizada dos cadernos foi configurada em ambiente de integração contínua (*GitHub Actions*), assegurando a reprodutibilidade dos resultados e a rastreabilidade de todas as etapas do processo analítico, em consonância com as diretrizes metodológicas internacionais para a utilização de dados vinculados e técnicas de aprendizado de máquina na estimação de indicadores de saúde [?].

Capítulo 5

Resultados

Este capítulo apresenta os resultados obtidos na etapa de pós-processamento de pares candidatos gerados por *linkage* probabilístico no OpenRecLink [?]. Os achados são organizados a partir de três cenários experimentais (comparação de técnicas, estratégia de maximização da revocação e estratégia de maximização da precisão) e de análises complementares (faixas de escore, estudo de ablação, fronteira de Pareto, validação cruzada e interpretabilidade). Os detalhes de base de dados, engenharia de atributos, métricas e ambiente computacional são descritos no Capítulo 4.

5.1 Base de dados e desbalanceamento

A base analisada foi composta por 61.696 pares candidatos (linhas do arquivo `COMPARADORSEMIDENT.csv`), dos quais 247 foram classificados como pares verdadeiros após rotulagem manual, resultando em um desbalanceamento aproximado de 1:248. Esse cenário é típico em problemas de *record linkage* e impõe restrições importantes ao treinamento e, principalmente, ao uso operacional de classificadores, uma vez que pequenas variações na taxa de falsos positivos podem produzir grandes volumes para revisão [?, ?].

Para possibilitar comparação direta entre estratégias, os experimentos baseados em partição fixa (*hold-out*) utilizaram divisão estratificada 70/30, mantendo a proporção de pares verdadeiros no conjunto de teste. As métricas reportadas seguem a definição apresentada em 4.6, com ênfase em precisão, revocação e F1.

5.2 Faixas de escore do OpenRecLink e zona cinzenta

O OpenRecLink produz, para cada par candidato, um conjunto de escores de similaridade por atributo e um escore agregado final (`nota final`) que sintetiza a evidência de pareamento [?]. Na prática, uma abordagem ingênua consiste em selecionar pares apenas pelo escore agregado acima de um limiar. Para avaliar os limites dessa estratégia, analisou-se a distribuição de pares verdadeiros por faixas de escore.

A Tabela 5.1 e a Figura 5.1 evidenciam dois comportamentos: (i) faixas altas (por exemplo, 9–10 e 10+) apresentam elevadíssima proporção de pares, mas concentram volume pequeno; (ii) uma parcela substancial dos pares verdadeiros está concentrada em uma faixa intermediária, definida aqui como *zona cinzenta* (5–8). Este resultado justifica a aplicação de modelos de aprendizagem de máquina para discriminar pares na região em que o escore agregado do *linkage* probabilístico é insuficiente para decisão automatizada.

Tabela 5.1: Distribuição dos pares candidatos por faixa de escore do OpenRecLink

Faixa de escore	Total	Pares	% Pares
0-3	5,336	0	0.00%
3-5	34,645	3	0.01%
5-6	14,263	17	0.12%
6-7	6,175	57	0.92%
7-8	1,080	43	3.98%
8-9	102	34	33.33%
9-10	44	43	97.73%
10+	51	50	98.04%

5.3 Engenharia de atributos e modelos avaliados

Conforme descrito em 4.4, foram utilizados atributos numéricos derivados dos escores de similaridade por campo (nome, nome da mãe, data de nascimento, município de residência e endereço), além do escore agregado (`nota final`) e do *step* de blocagem. A partir desses escores, foram construídas variáveis derivadas (somas ponderadas, interações e marcadores binários), totalizando conjuntos de atributos com 41 variáveis (cenário comparativo), 58 variáveis (estratégia de revocação) e 50 variáveis (estratégia de precisão).

Em todos os cenários, foram testados classificadores tradicionais e métodos de *en-*

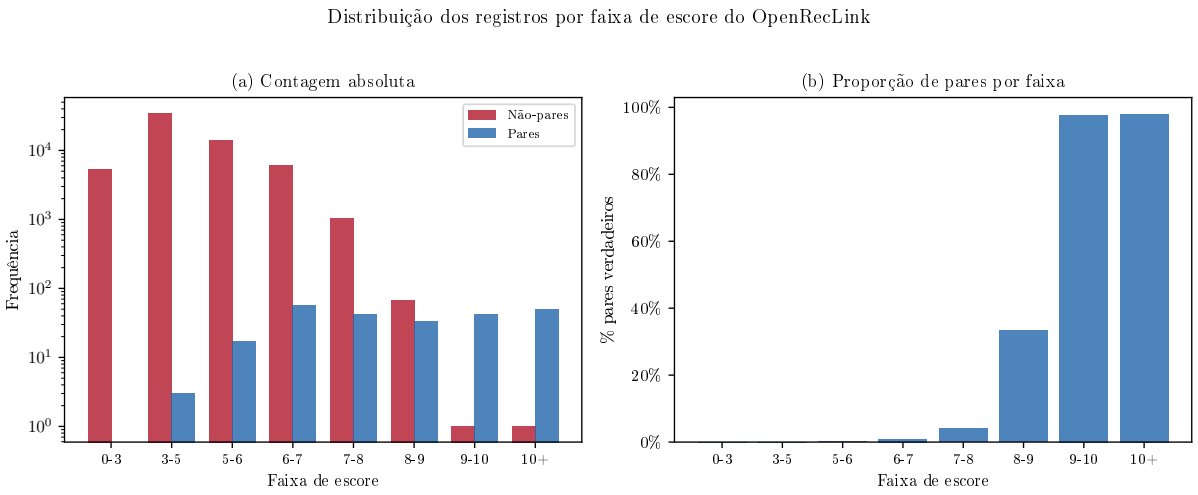

Figura 5.1: Distribuição de pares verdadeiros e não-pares por faixa do escore agregado (`nota final`) do OpenRecLink.

semble. Para lidar com o desbalanceamento, foram avaliadas estratégias de reamostragem do tipo SMOTE e variantes, bem como ponderação de classes, conforme recomendado para bases altamente desbalanceadas [**?**, **?**]. Os modelos e configurações específicas de cada cenário seguem o delineamento apresentado em 4.5.

5.4 Análise comparativa de técnicas (NB01)

No cenário de análise comparativa, foram avaliados sete modelos com limiar padrão de decisão (0,5), de forma a obter um panorama inicial do desempenho em presença de desbalanceamento severo. A Figura 5.2 resume os resultados de precisão, revocação e F1.

Observou-se que a combinação Random Forest com reamostragem (RF+SMOTE) e o Gradient Boosting apresentaram os melhores equilíbrios entre precisão e revocação, enquanto modelos lineares tendem a maximizar revocação ao custo de precisão em bases desbalanceadas. Esse resultado orientou a seleção de modelos base para as estratégias específicas de revocação e de precisão.

5.5 Estratégia de maximização da revocação (NB02)

No cenário de maximização da revocação, buscou-se reduzir ao máximo a chance de perda de pares verdadeiros em contextos de vigilância epidemiológica, aceitando maior carga de revisão manual. Foram avaliadas combinações de modelos, reamostragem e pon-

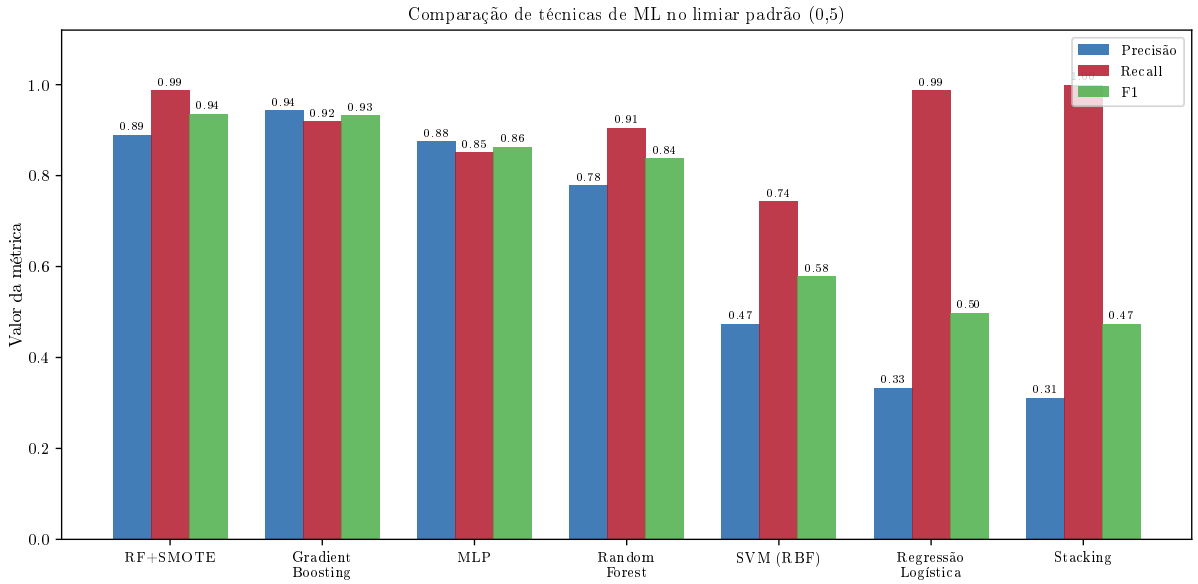

Figura 5.2: Comparação de desempenho (precisão, revocação e F1) entre modelos no cenário NB01, com limiar de decisão padrão.

deração de classes, além de estratégias em cascata. Entre as configurações testadas, o modelo Gradient Boosting com SMOTE e pesos de classe apresentou revocação elevada (0,959) com manutenção de precisão alta (0,934), evidenciando que é possível obter sensibilidade elevada sem degradação extrema da precisão quando há engenharia de atributos apropriada e estratégia de balanceamento adequada.

Por outro lado, configurações extremamente permissivas (por exemplo, *ensemble* com limiar baixo) alcançam revocação próxima de 1,0, mas impõem custos operacionais desproporcionais, o que reforça a necessidade de selecionar pontos operacionais com base em restrições de capacidade de revisão.

5.6 Estratégia de maximização da precisão (NB03)

No cenário de maximização da precisão, o objetivo foi produzir listas de candidatos de alta confiança para investigação, minimizando falsos positivos. Foram comparadas três abordagens principais: (i) regras determinísticas baseadas em evidências fortes (atributos-chave com concordância elevada), (ii) consenso por unanimidade entre modelos e (iii) um classificador híbrido, combinando probabilidade estimada por aprendizagem de máquina e regras determinísticas.

O resultado mais restritivo (regras com limiar alto) produz precisão próxima de

1,0, mas com revocação limitada. O consenso entre modelos aumenta revocação mantendo precisão elevada, enquanto o híbrido permite ajustar o equilíbrio entre precisão e revocação por meio de dois limiares (probabilidade do classificador e escore das regras). Esse desenho motivou a formalização do *framework* híbrido e o estudo de ablação apresentado a seguir.

5.7 Estudo de ablação e fronteira de Pareto

Para tornar explícita a contribuição metodológica do *framework* proposto, foi conduzido um estudo de ablação abrangente, comparando categorias de decisão (limiar ingênuo por escore, regras apenas, aprendizagem de máquina apenas, híbridos do tipo AND e OR, cascatas e consenso). A Tabela 5.2 apresenta a melhor configuração por categoria, enquanto a Tabela 5.3 lista as dez melhores configurações segundo F1.

Tabela 5.2: Melhor configuração por categoria de classificação: estudo de ablação.

Categoria	Configuração	Precisão	Revocação	F1
Limiar ingênuo (escore $\geq t$)	Naive score ≥ 8	0.642	0.581	0.610
Somente regras determinísticas	Rules-only (≥ 6)	0.821	0.865	0.842
Somente ML	ML-only RF+SMOTE (≥ 0.5)	0.923	0.973	**0.947**
Híbrido ML $\cap$ Regras (AND)	Hybrid-AND RF+SMOTE ≥ 0.5 + Rules ≥ 5	0.956	0.878	0.915
Híbrido ML $\cup$ Regras (OR)	Hybrid-OR RF+SMOTE ≥ 0.7 + Rules ≥ 8	0.957	0.905	**0.931**
Cascata ML $\rightarrow$ Regras	Cascade ML$\rightarrow$Rules RF+SMOTE $\geq 0.5 \rightarrow$ Rules ≥ 7	0.965	0.743	0.840
Cascata Regras $\rightarrow$ ML	Cascade Rules$\rightarrow$ML Rules $\geq 5 \rightarrow$ RF+SMOTE ≥ 0.5	0.956	0.878	0.915
Consenso entre modelos ML	Consensus ML-majority (th=0.7, 5 models)	0.958	0.919	**0.938**
Consenso + Regras	Consensus+Rules majority(th=0.5) AND Rules ≥ 6	0.969	0.851	0.906

Os resultados mostram que o limiar ingênuo por escore agregado apresenta desempenho limitado (por exemplo, escore ≥ 8 com F1 inferior ao obtido por modelos), enquanto a abordagem por aprendizagem de máquina, especialmente RF+SMOTE (th=0,5), atinge o melhor ponto de equilíbrio no *hold-out*. As combinações híbridas, por sua vez, deslocam o ponto operacional em direção a maior precisão, sacrificando revocação, o que pode ser

Tabela 5.3: Dez melhores configurações por F1: estudo de ablação.

#	Configuração	Precisão	Revocação	F1
1	ML-only RF+SMOTE (≥ 0.5)	0.923	0.973	**0.947**
2	Consensus ML-majority (th=0.7, 5 models)	0.958	0.919	0.938
3	Consensus ML-majority (th=0.5, 5 models)	0.910	0.959	0.934
4	Hybrid-OR RF+SMOTE ≥ 0.7 + Rules ≥ 8	0.957	0.905	0.931
5	ML-only RF+SMOTE (≥ 0.7)	0.957	0.892	0.923
6	ML-only XGB+SMOTE (≥ 0.9)	0.878	0.973	0.923
7	Hybrid-OR RF+SMOTE ≥ 0.7 + Rules ≥ 9	0.957	0.892	0.923
8	Cascade Rules→ML Rules ≥ 5 → RF+SMOTE ≥ 0.5	0.956	0.878	0.915
9	Hybrid-AND RF+SMOTE ≥ 0.5 + Rules ≥ 5	0.956	0.878	0.915
10	Hybrid-AND RF+SMOTE ≥ 0.5 + Rules ≥ 6	0.970	0.865	0.914

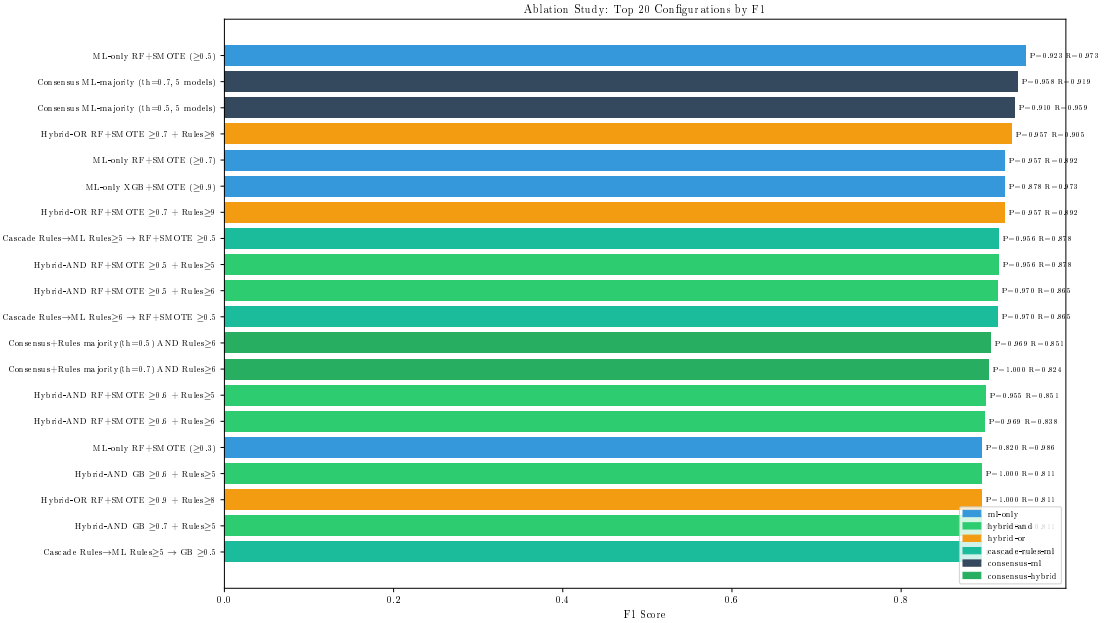

Figura 5.3: Melhor F1 por categoria de decisão no estudo de ablação.

desejável em fluxos de investigação com baixa tolerância a falsos positivos.

A possibilidade de parametrizar a decisão por dois limiares (probabilidade do classificador e escore de regras) permite visualizar e selecionar pontos operacionais na forma de uma fronteira de Pareto (precisão versus revocação), conforme apresentado na Tabela 5.4 e na Figura 5.4.

5.8 Robustez, sensibilidade ao desbalanceamento e interpretabilidade

A robustez das principais configurações foi avaliada por validação cruzada estratificada com cinco partições (5-*fold*), sintetizada na Tabela **??**. Observou-se que a configu-

Tabela 5.4: Fronteira de Pareto: pontos operacionais do classificador híbrido (RF+SMOTE).

θ_{ML}	θ_{Regras}	Precisão	Revocação	F1	Perfil
0.20	2.0	0.667	1.000	0.800	Máx. revocação
0.40	0.0	0.880	0.986	0.930	Máx. revocação
0.50	0.0	0.923	0.973	0.947	**Equilíbrio ótimo**
0.65	4.5	0.932	0.932	0.932	Intermediário
0.70	4.0	0.957	0.892	0.923	Intermediário
0.50	6.0	0.970	0.865	0.914	Intermediário
0.70	5.0	1.000	0.811	0.896	Máx. precisão

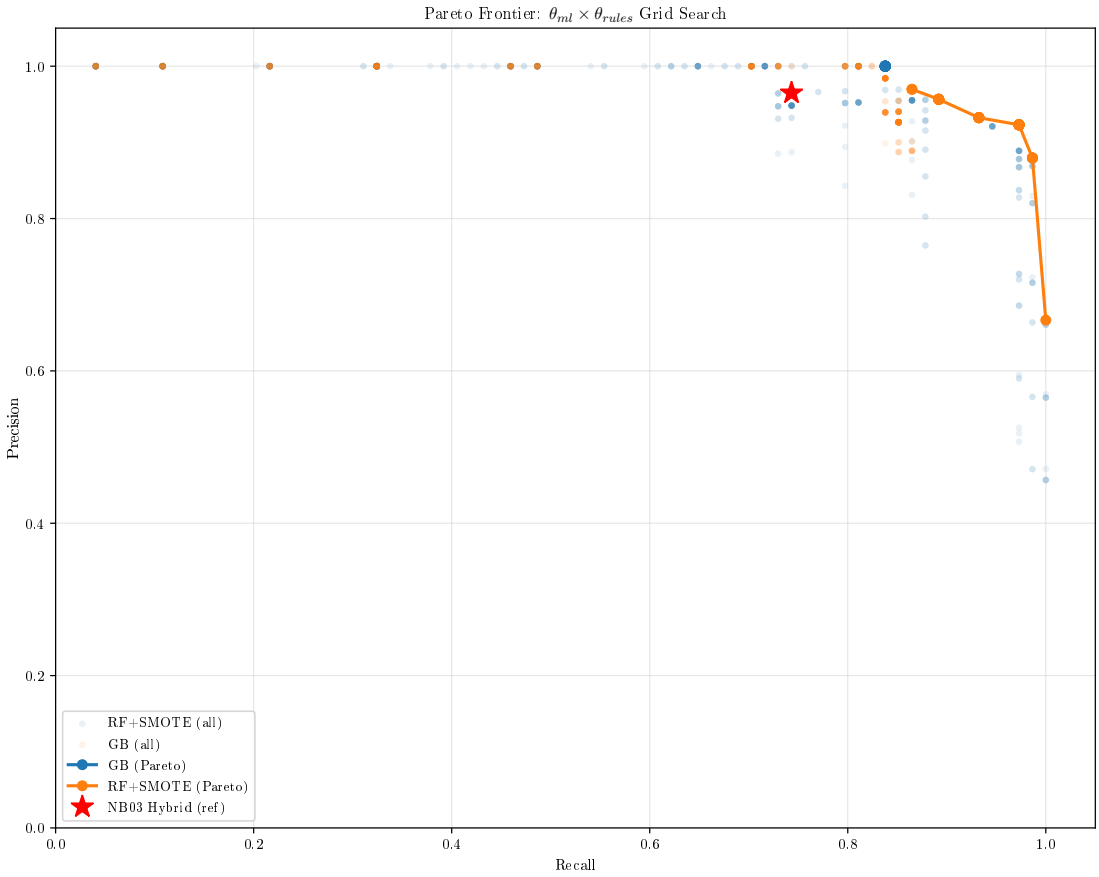

Figura 5.4: Fronteira de Pareto para seleção de pontos operacionais (precisão e revocação) em combinações de limiares do classificador e das regras.

ração híbrida do tipo OR (RF+SMOTE com limiar mais restritivo OU regras com limiar alto) apresentou o maior F1 médio, sugerindo complementaridade mais estável entre os componentes quando a variabilidade amostral é incorporada.

Para avaliar a dependência dos resultados em relação ao tratamento do desbalanceamento, foi conduzida análise de sensibilidade considerando nove estratégias de reamostragem. A Tabela **??** e a Figura **??** mostram que o desempenho permanece em faixa estreita (F1 variando aproximadamente entre 0,823 e 0,903), o que indica compor-

Tabela 5.5: Validação cruzada estratificada (5-*fold*) das principais configurações

Configuração	**Precisão**	**Recall**	**F1**
RF+SMOTE ≥ 0.5	0.846±0.036	0.907±0.031	0.875±0.027
GB ≥ 0.5	0.932±0.022	0.830±0.023	0.878±0.017
Rules ≥ 6	0.814±0.035	0.822±0.029	0.817±0.014
Hybrid-AND RF+SMOTE ≥ 0.5 + Rules ≥ 6	0.948±0.019	0.802±0.038	0.868±0.025
Hybrid-AND GB ≥ 0.6 + Rules ≥ 5	0.979±0.022	0.757±0.027	0.854±0.021
Hybrid-OR RF+SMOTE ≥ 0.7 + Rules ≥ 8	0.905±0.020	0.879±0.032	0.891±0.016
Naive score ≥ 8	0.649±0.074	0.515±0.076	0.571±0.057

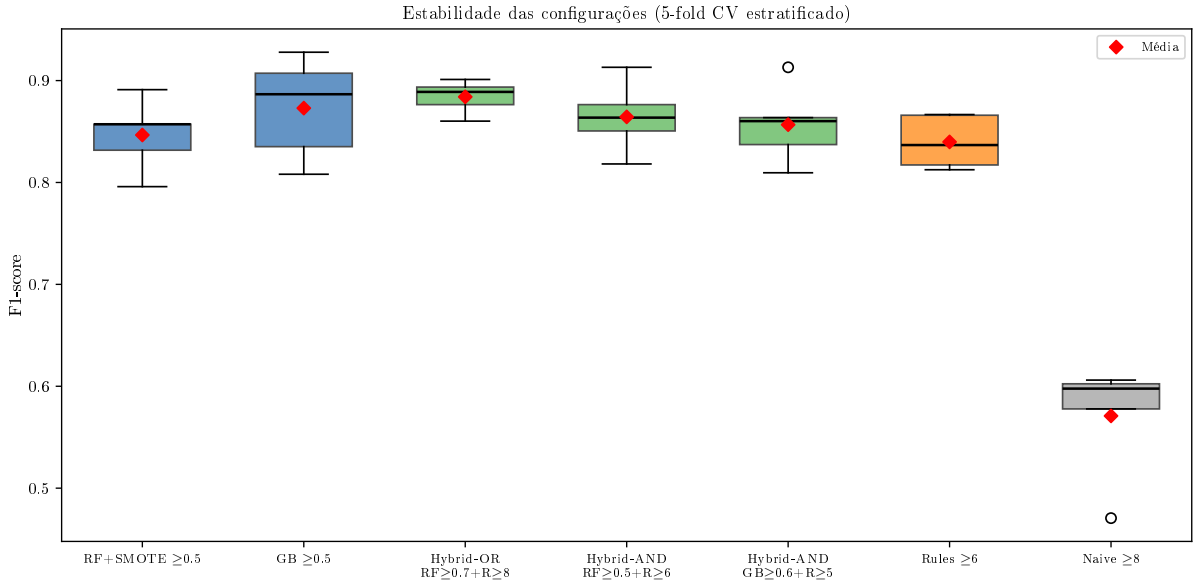

Figura 5.5: Distribuição do F1 por configuração na validação cruzada (5-*fold*).

tamento consistente do classificador sob diferentes estratégias.

Por fim, buscou-se interpretar os fatores que mais contribuem para a classificação correta, com base em valores SHAP (*SHapley Additive exPlanations*) [**?**]. A Tabela **??** e a Figura **??** indicam que o escore agregado (`nota final`) e variáveis relacionadas à concordância de nome tendem a concentrar maior contribuição global. Entretanto, ao restringir a análise à zona cinzenta, atributos relacionados ao nome da mãe ganham relevância, refletindo a necessidade de evidência adicional quando o escore agregado é ambíguo.

Tabela 5.6: Sensibilidade do RF à estratégia de balanceamento (5-*fold* CV)

Estratégia	**Precisão**	**Recall**	**F1**
Sem balanceamento	0.962±0.023	0.834±0.072	0.893±0.047
Class weight only	0.759±0.040	0.899±0.038	0.823±0.032
SMOTE 0.3	0.887±0.030	0.911±0.037	0.898±0.024
SMOTE 0.5	0.879±0.029	0.911±0.037	0.894±0.027
SMOTE 1.0	0.852±0.030	0.903±0.030	0.876±0.021
BorderlineSMOTE 0.3	0.904±0.033	0.903±0.039	0.903±0.023
ADASYN 0.3	0.877±0.044	0.911±0.023	0.893±0.024
RandomUnderSampler 0.01	0.910±0.048	0.875±0.046	0.891±0.024
SMOTETomek 0.3	0.890±0.033	0.911±0.037	0.900±0.028

Tabela 5.7: Importância dos atributos por valores SHAP (média do valor absoluto, classe positiva).

#	Atributo	|SHAP| médio
1	nota final	0.0766
2	NOME qtd frag iguais	0.0671
3	nome_perfeito	0.0611
4	NOME qtd frag muito parec	0.0573
5	NOME qtd frag comuns	0.0362
6	nome_score_total	0.0348
7	dtnasc_score_total	0.0328
8	nome_x_dtnasc	0.0309
9	NOMEMAE qtd frag comuns	0.0267
10	NOMEMAE qtd frag muito parec	0.0144
11	NOMEMAE qtd frag iguais	0.0144
12	DTNASC dt iguais	0.0114
13	dtnasc_perfeito	0.0112
14	mae_score_total	0.0110
15	nome_x_mae	0.0081

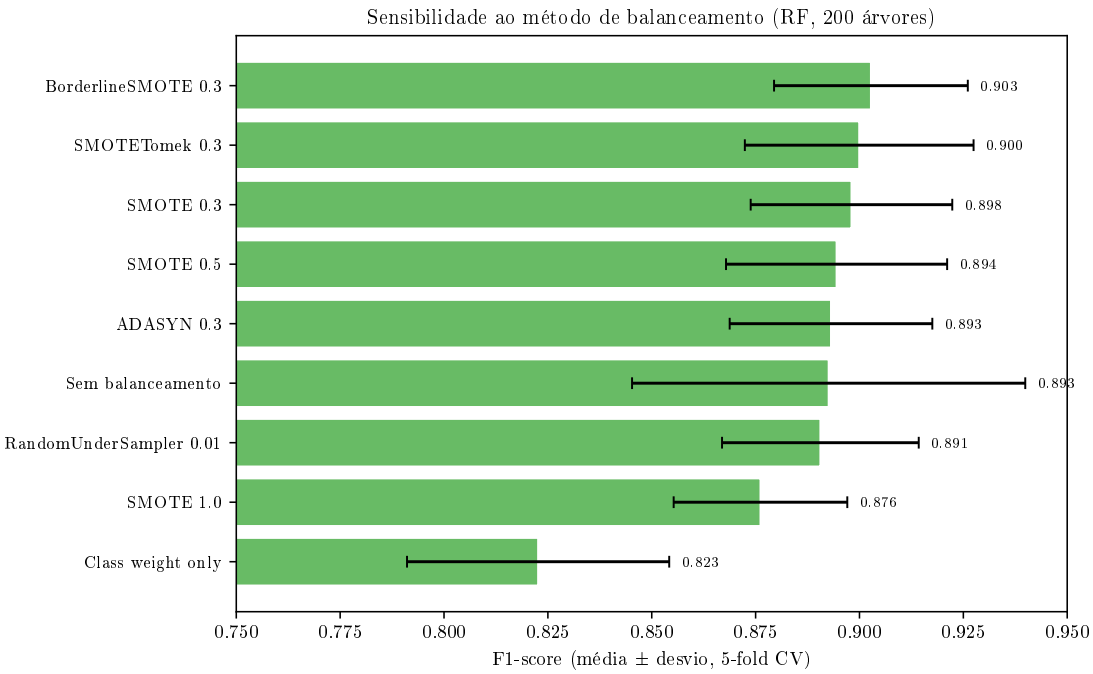

Figura 5.6: Sensibilidade do desempenho (F1) do classificador a diferentes estratégias de reamostragem para desbalanceamento.

5.9 Síntese e recomendações operacionais

Em conjunto, os resultados indicam que decisões baseadas apenas em limiar do escore agregado são insuficientes para recuperar uma parcela relevante dos pares verdadeiros na zona cinzenta. Modelos de aprendizagem de máquina elevam substancialmente a eficiência e a qualidade do *linkage* pós-processado, com possibilidade de ajuste fino do ponto operacional. Além disso, as regras determinísticas funcionam como mecanismo de validação adicional em cenários de alta exigência de precisão, compondo um *framework* configurável.

Uma implicação prática é a recomendação de dois fluxos complementares: (i) um fluxo voltado a vigilância, priorizando revocação, com limiar de ML mais permissivo e estratégia de balanceamento adequada; e (ii) um fluxo voltado a investigação de alta confiança, priorizando precisão, com combinações híbridas do tipo AND e limiares mais restritivos. A discussão das implicações epidemiológicas e operacionais desse desenho é apresentada no Capítulo **??**.

Figura 5.7: Resumo SHAP para a classe positiva, indicando a contribuição média (valor absoluto) de cada atributo para a predição.

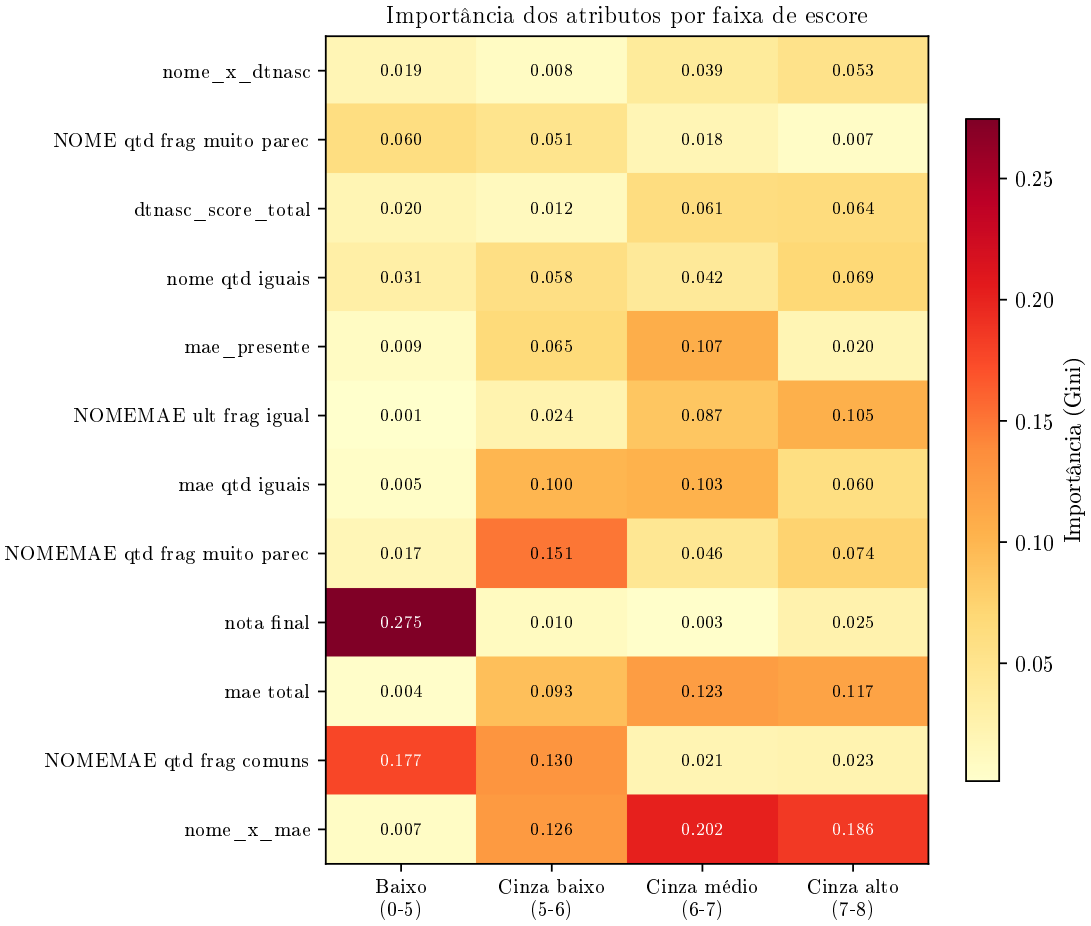

Figura 5.8: Importância dos atributos por faixa de escore, com destaque para mudanças de relevância na zona cinzenta.

Capítulo 6

Discussão

A discussão dos resultados, suas implicações para a prática do *linkage* de registros em saúde pública e as limitações do estudo serão apresentadas neste capítulo.

Referências Bibliográficas

[1] Avedis Donabedian. The quality of care: How can it be assessed? *JAMA*, 260(12):1743–1748, 1988. 1

[2] Francisco Viacava, Maria Alicia Dominguez Ugá, Silvia Marta Porto, Josue Laguardia, and Rodrigo da Silva Moreira. Avaliação de desempenho de sistemas de saúde: um modelo de análise. *Ciência & Saúde Coletiva*, 17(4):921–934, 2012. 1, 2, 16

[3] Kenneth Rochel de Camargo Jr and Cláudia Medina Coeli. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilístico. *Cadernos de Saúde Pública*, 16(2):439–447, 2000. 1, 2, 4, 8, 12, 14, 15, 17, 19, 23, 25

[4] Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012. 1, 2, 4, 5, 6, 7, 8, 13, 14, 15, 24, 25

[5] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. 2, 5, 8, 12, 14

[6] Howard B. Newcombe, James M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130(3381):954–959, 1959. 2

[7] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989. 2

[8] William E. Winkler. String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. Technical Report RR90/05, U.S. Bureau of the Census, Statistical Research Division, 1990. 2

[9] Cláudia Medina Coeli and Kenneth Rochel de Camargo Jr. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. *Revista de Saúde Pública*, 36(4):439–445, 2002. 2, 6, 12, 14, 15, 17, 19, 23

[10] Ted Enamorado, Benjamin Fifield, and Kosuke Imai. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2):353–371, 2019. 2

[11] Luíza Maria Oliveira de Sousa and Rejane Sobrino Pinheiro. Óbitos e internações por tuberculose não notificados no município do Rio de Janeiro. *Revista de Saúde Pública*, 45(1):31–39, 2011. 2, 16, 18, 26

[12] Gisele Pinto de Oliveira, Rejane Sobrino Pinheiro, Cláudia Medina Coeli, Draurio Barreira, and Stefano Barbosa Codenotti. Uso do sistema de informação sobre mortalidade para identificar subnotificação de casos de tuberculose no Brasil. *Revista Brasileira de Epidemiologia*, 15(3):468–477, 2012. 2, 4, 7, 15, 18, 23, 26

[13] Marli Souza Rocha, Gisele Pinto de Oliveira, Flávia Pinheiro Aguiar, Valéria Saraceni, and Rejane Sobrino Pinheiro. Do que morrem os pacientes com tuberculose: causas múltiplas de óbito de uma coorte de casos notificados. *Cadernos de Saúde Pública*, 31(4):709–721, 2015. 2

[14] Patricia Bartholomay, Gisele Pinto de Oliveira, Rejane Sobrino Pinheiro, and Ana Maria Nogales Vasconcelos. Melhoria da qualidade das informações sobre tuberculose a partir do relacionamento entre bases de dados. *Cadernos de Saúde Pública*, 30(11):2459–2470, 2014. 2, 4, 23

[15] Jairnilson Paim, Claudia Travassos, Celia Almeida, Ligia Bahia, and James Macinko. The Brazilian health system: History, advances, and challenges. *The Lancet*, 377(9779):1778–1797, 2011. 3, 22

[16] James Macinko and Matthew J. Harris. Brazil's Family Health Strategy — delivering community-based primary care in a universal health system. *New England Journal of Medicine*, 372(23):2177–2181, 2015. 3

[17] M. Barreto, Myt Ichihara, B. Almeida, M. Barreto, L. Cabral, Rl Fiaccone, RP Carreiro, C. Teles, R. Pitta, G. Penna, M. Barral-Netto, MS Ali, George C. G. Barbosa, S. Denaxas, LC Rodrigues, and L. Smeeth. The centre for data and knowledge integration for health (cidacs): Linking health and social data in brazil. *International Journal of Population Data Science*, 4, 2019. 4

[18] Márcia Elizabeth Marinho da Silva. *Linkage de Bases de Dados Identificadas em Saúde: Consentimento, Privacidade e Segurança da Informação*. Tese de doutorado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2012. Orientadora: Profª Cláudia Medina Coeli. 4

[19] Marcela Lima Santos, Cláudia Medina Coeli, et al. Fatores associados à subnotificação de tuberculose a partir do linkage SINAN-AIDS e SINAN-TB. *Revista Brasileira de Epidemiologia*, 21:e180019, 2018. 4, 23

[20] Daniel Nasseh and Jürgen Stausberg. Evaluation of a binary semi-supervised classification technique for probabilistic record linkage. *Methods of Information in Medicine*, 55(2):136–143, 2016. 6

[21] David J. Hand and Peter Christen. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28:539–547, 2018. 6, 9, 26, 27

[22] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 7, 8, 16, 24, 28

[23] Hossein Hassani, Mohammad Reza Entezarian, Soroosh Zaeimzadeh, Leila Marvian, and Nadejda Komendantova. An oversampling-undersampling strategy for large-scale data linkage. *Frontiers in Big Data*, 8:1542483, 2025. 7, 16, 26

[24] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. 7, 20, 26

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009. 9, 10, 11, 25, 26, 27

[26] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 9

[27] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. 10, 27

[28] Murat Sariyar and Andreas Borg. Bagging, bumping, multiview, and active learning for record linkage with empirical results on patient identity data. *Computer Methods and Programs in Biomedicine*, 108(3):1160–1169, 2012. 11, 25

[29] Thanh Trung Vo and Jongwoo Lee. Statistical supervised meta-ensemble algorithm for medical record linkage. *Journal of Biomedical Informatics*, 95:103220, 2019. 12, 14, 15

[30] Zhenhe Jiao et al. A new hybrid record linkage process to make epidemiological databases interoperable. *BMC Medical Research Methodology*, 21(1):155, 2021. 12, 15

[31] Ali Almadani et al. Linking electronic health records for multiple sclerosis research: Comparison of deterministic, probabilistic, and machine learning linkage methods. *JMIR Medical Informatics*, 14:e79869, 2026. 12, 15

[32] Olivier Binette and Rebecca C. Steorts. (almost) all of entity resolution. *Science Advances*, 8(12):eabi8021, 2022. 12, 14

[33] Murat Sariyar, Andreas Borg, and Klaus Pommerening. Active learning strategies for the deduplication of electronic patient data using classification trees. *Journal of Biomedical Informatics*, 45(5):893–900, 2012. 12

[34] Jana Asher et al. An introduction to probabilistic record linkage with a focus on linkage processing for WIC administrative data. *International Journal of Environmental Research and Public Health*, 17(18):6937, 2020. 13

[35] Rainer Schnell and Stefanie Vivien Weiand. Microsimulation of an educational attainment register to predict future record linkage quality. *International Journal of Population Data Science*, 8(1), 2023. 13, 15

[36] World Health Organization. Global tuberculosis report 2024. Technical report, WHO, 2024. 16, 18

[37] Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente. Boletim epidemiológico de tuberculose 2024. Technical report, Ministério da Saúde, Brasília, 2024. 16, 18

[38] Rejane Sobrino Pinheiro, Vanessa Luiza Andrade, and Gisele Pinto de Oliveira. Subnotificação da tuberculose no SINAN: abandono primário de bacilíferos e captação de casos em Rio de Janeiro. *Cadernos de Saúde Pública*, 28(8):1559–1568, 2012. 16, 18

[39] Marli Souza Rocha, Gisele Pinto de Oliveira, Flávia Pinheiro Aguiar, Valéria Saraceni, and Rejane Sobrino Pinheiro. Do que morrem os pacientes com tuberculose: causas múltiplas de óbito de uma coorte de casos notificados. *Cadernos de Saúde Pública*, 31(4):709–721, 2015. 16, 18

[40] Rana Haneef, Mariken Tijhuis, Rodolphe Thiébaut, Ondřej Májek, Ivan Pristaš, Hanna Tolonen, and Anne Gallay. Methodological guidelines to estimate population-based health indicators using linked data and/or machine learning techniques. *Archives of Public Health*, 80(1):9, 2022. 16, 28

[41] Gisele Pinto de Oliveira, Ana Luiza Bierrenbach, Kenneth Rochel de Camargo Jr, Cláudia Medina Coeli, and Rejane Sobrino Pinheiro. Acurácia do relacionamento probabilístico e determinístico de registros: o caso da tuberculose. *Revista de Saúde Pública*, 50:49, 2016. 17, 23

[42] David M. Kessner, Carolyn E. Kalk, and James Singer. Assessing health quality — the case for tracers. *New England Journal of Medicine*, 288(4):189–194, 1973. 17

[43] Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Manual de recomendações para o controle da tuberculose no Brasil. Technical report, Ministério da Saúde, Brasília, 2019. 17

[44] Marli Souza Rocha, Gisele Pinto de Oliveira, Lucas Calais Tavares Guillen, Cláudia Medina Coeli, Valéria Saraceni, and Rejane Sobrino Pinheiro. Uso de *linkage* entre bases de dados e de regras de *scripting* para qualificação de variáveis do Sinan-TB. *Cadernos de Saúde Pública*, 35(12):e00074318, 2019. 18

[45] Otavio T. Ranzani, Leonardo S. L. Bastos, João Gabriel Machado Gelli, Janaina Figueira Marchesi, Fernando Baiao, Silvio Hamacher, and Fernando A. Bozza. Characterisation of the first 250,000 hospital admissions for COVID-19 in Brazil: A retrospective analysis of nationwide data. *The Lancet Respiratory Medicine*, 9(4):407–418, 2021. 18

[46] Cintya Moura de Maia, Daniela Rezende Martelli, Dayane Machado Silveira, Enaldo Arlindo de Oliveira, Tatiane Cristina Lopes, Rejane Sobrino Pinheiro, and Hercílio Martelli-Júnior. Impact of COVID-19 on tuberculosis indicators in Brazil: A time-series analysis. *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, 29:100325, 2022. DOI not available — original DOI 10.1016/j.jctube.2022.100325 resolves to a different article. 18

[47] Pedro C. Hallal, Fernando P. Hartwig, Bernardo L. Horta, Mariangela F. Silveira, Claudio J. Struchiner, Luiz Paulo Vidaletti, Nelson A. Neumann, Lucia C. Pellanda, Odir A. Dellagostin, Marcelo N. Burattini, Guilherme D. Victora, Ana M. B.

Menezes, Fernando C. Barros, Aluísio J. D. Barros, and Cesar G. Victora. SARS-CoV-2 antibody prevalence in Brazil: Results from two successive nationwide serological household surveys. *The Lancet Global Health*, 8(10):e1390–e1398, 2020. 18

Apêndice A

Glossário de Termos Técnicos

Este glossário apresenta definições de termos técnicos utilizados ao longo desta tese, visando facilitar a compreensão de conceitos que podem ser menos familiares aos profissionais da área de saúde coletiva.

Acurácia Proporção de classificações corretas (verdadeiros positivos e verdadeiros negativos) em relação ao total de pares avaliados.

Área cinza Região de escores intermediários no relacionamento probabilístico, onde os pares candidatos não podem ser classificados automaticamente como verdadeiros ou falsos, demandando revisão adicional.

AUC-PR Área sob a curva precisão-sensibilidade (*Area Under the Precision-Recall Curve*), métrica de desempenho particularmente informativa em cenários de desbalanceamento de classes.

AUC-ROC Área sob a curva característica de operação do receptor (*Area Under the Receiver Operating Characteristic Curve*), métrica que avalia o poder discriminatório de classificadores.

Blocagem Estratégia de redução do espaço de comparação no relacionamento de registros, que agrupa candidatos por chaves comuns (ex.: *Soundex* do nome, ano de nascimento) para evitar a comparação exaustiva de todos os pares possíveis.

Classificação Tarefa de aprendizado supervisionado que atribui rótulos discretos (par verdadeiro ou não-par) a instâncias com base em atributos preditores.

Deduplicação Processo de identificação e remoção de registros duplicados referentes a um mesmo indivíduo dentro de uma única base de dados.

Desbalanceamento de classes Situação em que uma classe (tipicamente os pares verdadeiros) é muito menos frequente que a outra (não-pares), podendo comprometer o desempenho de classificadores.

Ensemble Abordagem que combina múltiplos classificadores para produzir uma decisão agregada, frequentemente superior ao desempenho individual de cada modelo.

Escore de similaridade Valor numérico que quantifica o grau de concordância entre campos de dois registros comparados (ex.: distância de Jaro-Winkler para nomes).

F1-*score* Média harmônica entre precisão e sensibilidade, utilizada como métrica-síntese do desempenho de classificadores.

GAL Gerenciador de Ambiente Laboratorial, sistema do Ministério da Saúde para gestão de exames laboratoriais.

Gradient Boosting Família de algoritmos de aprendizado de máquina que constrói modelos sequenciais, cada um corrigindo erros do anterior, incluindo implementações como XGBoost e LightGBM.

Linkage Ver *Relacionamento de registros*.

OpenRecLink *Software* brasileiro de código aberto para relacionamento probabilístico de registros, desenvolvido por Camargo Jr. e Coeli.

Par candidato Combinação de dois registros, provenientes de bases distintas, que foram selecionados pela etapa de blocagem para comparação detalhada.

Par verdadeiro Par de registros que se refere ao mesmo indivíduo, confirmado por revisão manual ou padrão-ouro.

Precisão Proporção de pares classificados como verdadeiros que são efetivamente verdadeiros (*Positive Predictive Value*).

Random Forest Algoritmo de aprendizado de máquina baseado em múltiplas árvores de decisão treinadas em subamostragens aleatórias dos dados.

Relacionamento de registros Processo de identificação de registros referentes ao mesmo indivíduo em duas ou mais bases de dados distintas (*record linkage*).

Relacionamento determinístico Estratégia de *linkage* baseada em regras exatas de concordância entre campos identificadores.

Relacionamento probabilístico Estratégia de *linkage* fundamentada na teoria de Fellegi e Sunter, que atribui pesos aos campos comparados e calcula um escore composto para classificar pares.

Sensibilidade Proporção de pares verdadeiros corretamente identificados pelo classificador (*Recall*).

SIA-SUS Sistema de Informações Ambulatoriais do SUS, que registra procedimentos ambulatoriais realizados na rede pública.

SIH-SUS Sistema de Informações Hospitalares do SUS, que registra internações hospitalares na rede pública.

SIM Sistema de Informações sobre Mortalidade, que consolida dados das Declarações de Óbito no Brasil.

Sinan Sistema de Informação de Agravos de Notificação, que registra a notificação compulsória de doenças e agravos, incluindo tuberculose.

SITETB Sistema de Informação de Tratamentos Especiais da Tuberculose, que registra casos de tuberculose drogarresistente.

SMOTE *Synthetic Minority Over-sampling Technique*, técnica de sobreamostragem que gera instâncias sintéticas da classe minoritária para atenuar o desbalanceamento.

Stacking Técnica de *ensemble* que utiliza as saídas de múltiplos classificadores de base como atributos de entrada para um meta-classificador.

SUS Sistema Único de Saúde, sistema público de saúde brasileiro de caráter universal.

Tuberculose (TB) Doença infecciosa de notificação compulsória utilizada neste trabalho como condição marcadora para avaliação do relacionamento de bases de dados.