



**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**

**Rio de Janeiro  
2026**

**Marco Elisio Oliveira Jardim**

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

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  
2026**

Jardim, Marco Elísio Oliveira

Aprendizado de máquina aplicado ao pós-processamento do relacionamento probabilístico de bases de dados de saúde / Marco Elísio Oliveira Jardim. – Rio de Janeiro, 2026.

112 f.: il.

Orientadora: Rejane Sobrino Pinheiro.

Tese (Doutorado em Saúde Coletiva) – Universidade Federal do Rio de Janeiro, Instituto de Estudos em Saúde Coletiva, Programa de Pós-Graduação em Saúde Coletiva, 2026.

1. *Linkage* probabilístico. 2. Aprendizado de máquina. 3. Desbalanceamento de classes. 4. Pós-processamento. 5. Tuberculose. 6. Mortalidade. 7. Vigilância em saúde. 8. Modelos de linguagem. 9. Zona cinzenta. 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.

**Marco Elisio Oliveira Jardim**

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

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).

Aprovada em \_\_\_\_\_ de \_\_\_\_\_ de 2026.

---

Prof.<sup>a</sup> Dr.<sup>a</sup> Rejane Sobrino Pinheiro (Orientadora)  
IESC/UFRJ

---

[NOME\_DO\_EXAMINADOR\_1]  
[INSTITUIÇÃO\_1]

---

[NOME\_DO\_EXAMINADOR\_2]  
[INSTITUIÇÃO\_2]

---

[NOME\_DO\_EXAMINADOR\_3]  
[INSTITUIÇÃO\_3]

---

[NOME\_DO\_EXAMINADOR\_4]  
[INSTITUIÇÃO\_4]

*À Joyce, José Antônio e Marco,  
por darem sentido a cada passo desta jornada.*

*A todos que dedicam suas vidas  
à construção de um sistema de saúde  
mais justo e integral.*

# Agradecimentos

À minha orientadora, Professora Rejane Sobrino Pinheiro, agradeço pela confiança depositada neste trabalho, pela orientação rigorosa e generosa ao longo de todos estes anos e pelo exemplo de dedicação à pesquisa em saúde pública. Sua capacidade de acolher ideias e, ao mesmo tempo, exigir o melhor de cada análise foi determinante para que esta tese chegasse a bom termo.

À minha esposa, Joyce, companheira de todas as horas, agradeço pelo amor incondicional, pela paciência com as noites de estudo e pela serenidade que manteve nosso lar de pé nos momentos em que esta jornada pareceu longa demais. Sem o seu apoio, nada disto teria sido possível.

Aos meus filhos, José Antônio e Marco, agradeço por serem a razão mais concreta de seguir adiante. Cada sorriso de vocês renovou a energia necessária para mais uma página, mais um experimento, mais um dia de trabalho.

A todas as pessoas que, de forma direta ou indireta, contribuíram para esta caminhada, colegas de pesquisa, professores, amigos e familiares, registro aqui minha gratidão sincera. Cada conversa, sugestão e palavra de incentivo deixou marca neste trabalho e em quem o escreveu.

*“Se a educação sozinha não transforma  
a sociedade, sem ela tampouco  
a sociedade muda.”*

PAULO FREIRE

# 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).

O *linkage* probabilístico de bases de dados, também denominado relacionamento probabilístico de registros (*record linkage*), constitui recurso essencial para a vigilância em saúde no Brasil, na medida em que possibilita a vinculação de registros individuais dispersos em diferentes sistemas de informação. A classificação dos pares candidatos resultantes desse processo, contudo, permanece dependente de limiares fixos aplicados sobre escores agregados de similaridade, o que acarreta perda expressiva de registros situados na zona de indecisão (área cinza). Essa lacuna é particularmente crítica quando a condição de saúde investigada apresenta elevada subnotificação, como é o caso da tuberculose enquanto condição marcadora de óbito.

Esta tese propôs e avaliou um arcabouço (*framework*) configurável de pós-processamento, baseado em aprendizado de máquina (*machine learning*), para a classificação de pares 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) no município do Rio de Janeiro, no período de 2006 a 2016. A base experimental compreendeu 61.696 pares candidatos gerados pelo OpenRecLink e pontuados por um comparador de registros que produz 29 subescores de similaridade e um escore agregado, tendo como padrão-ouro 247 pares verdadeiros (razão  $\approx 1:248$ ). Seis algoritmos foram comparados (Regressão Logística, Floresta Aleatória, XGBoost, LightGBM, SVM e Perceptron Multicamada), combinados com nove estratégias de balanceamento de classes (SMOTE, Borderline-SMOTE, ADASYN, SMOTE-Tomek, ponderação de classes e variantes), totalizando 70 configurações avaliadas por estudo de ablação.

A Floresta Aleatória (*Random Forest*) com SMOTE e limiar de classificação de 0,5 alcançou  $F_1=0,931$ , ao passo que o limiar convencional  $\geq 8$  obteve  $F_1=0,610$ . A validação

cruzada em cinco partições confirmou a estabilidade das abordagens híbridas, que combinam regras determinísticas e classificadores, com  $F_1$  médio de  $0,898 \pm 0,025$  e menor variância que as abordagens puramente algorítmicas; a Floresta Aleatória com SMOTE obteve  $F_1$  médio de  $0,916 \pm 0,026$ . A análise de sensibilidade ao desbalanceamento de classes indicou que os classificadores mantiveram  $F_1$  entre 0,880 e 0,918 nas nove estratégias avaliadas. A interpretação por meio de valores SHAP identificou o escore agregado, a quantidade de nomes iguais e o indicador de nome perfeito como os atributos de maior poder discriminatório; o nome da mãe predominou nas decisões relativas à área cinza. Do ponto de vista epidemiológico, a abordagem de máximo *recall* (sensibilidade) recuperou 24 óbitos por tuberculose adicionais (+55,8%) que teriam sido perdidos pelo limiar convencional, a um custo de aproximadamente 1,0 revisão manual por óbito detectado. A tese disponibiliza duas sequências operacionais (*pipelines*) pré-configuradas, um orientado à vigilância (máximo *recall*) e outro à confirmação (máxima precisão), articulados por uma fronteira de Pareto que permite ao gestor ou pesquisador calibrar o equilíbrio entre sensibilidade e especificidade conforme as necessidades do cenário de aplicação.

Adicionalmente, a tese propôs o *Grey-Zone Cost-based Mixture Deferral* (GZ-CMD), arcabouço operacional auto-calibrável que estende o *framework* ao integrar calibração por conjuntos âncora, política de decisão por perda esperada com custos assimétricos e revisão assistida por modelo de linguagem de grande porte. No modo de vigilância, o GZ-CMD alcançou  $F_1=0,954$ ; no modo de confirmação, precisão de 0,957. Considerando a cascata completa do *pipeline*, composta por limiares Fellegi–Sunter, regras de guarda, motor de perda esperada e revisão LLM dual-agent, o volume de revisão clerical humana reduziu-se de 21.620 para apenas 41 pares (redução de 99,8%, três ordens de grandeza), processados em aproximadamente 63 minutos de inferência.

**Palavras-chave:** *Linkage* probabilístico. Aprendizado de máquina. Desbalanceamento de classes. Pós-processamento. Tuberculose. Mortalidade. Vigilância em saúde. Modelos de linguagem. Zona cinzenta.

# **Abstract**

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

**Marco Elisio Oliveira Jardim**

**Advisor: Rejane Sobrino Pinheiro**

*Abstract* 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).

Probabilistic record linkage constitutes an essential resource for health surveillance in Brazil, as it enables the connection of individual records dispersed across different information systems. Classification of the resulting candidate pairs, however, remains dependent on fixed thresholds applied to aggregate similarity scores, leading to substantial loss of records in the indecision zone (grey zone). This limitation is particularly critical when the health condition under investigation presents high underreporting, as is the case of tuberculosis as an underlying cause of death.

This thesis proposed and evaluated a configurable post-processing framework based on machine learning for classifying candidate pairs produced by probabilistic linkage between the Mortality Information System (SIM) and the Tuberculosis Notifiable Diseases Information System (Sinan-TB) in the city of Rio de Janeiro, covering the period from 2006 to 2016. The experimental dataset comprised 61,696 candidate pairs generated by OpenRecLink and scored by a record comparator that produces 29 similarity sub-scores and an aggregate score, with a gold standard of 247 true pairs (ratio  $\approx 1:248$ ). Six algorithms were compared (Logistic Regression, Random Forest, XGBoost, LightGBM, SVM, and Multilayer Perceptron), combined with nine class-balancing strategies (SMOTE, Borderline-SMOTE, ADASYN, SMOTE-Tomek, class weighting, and variants), yielding 70 configurations evaluated through an ablation study.

Random Forest with SMOTE and a classification threshold of 0.5 achieved  $F_1=0.931$ , whereas the conventional threshold  $\geq 8$  yielded  $F_1=0.610$ . Five-fold cross-validation confirmed the stability of hybrid approaches, which combine deterministic rules and classifiers, with a mean  $F_1$  of  $0.898 \pm 0.025$  and lower variance than purely algorithmic approaches; Random Forest with SMOTE achieved a mean  $F_1$  of  $0.916 \pm 0.026$ . Sensitivity analysis to class imbalance indicated

that classifiers maintained  $F_1$  between 0.880 and 0.918 across all nine strategies. SHAP-based interpretation identified the aggregate score, the number of matching names, and the exact-name indicator as the most discriminating features; mother's name predominated in grey-zone decisions. From an epidemiological standpoint, the maximum-recall approach recovered 24 additional tuberculosis deaths (+55.8%) that would have been missed by the conventional threshold, at a cost of approximately 1.0 manual reviews per detected death. The thesis delivers two pre-configured pipelines, one geared towards surveillance (maximum recall) and another towards confirmation (maximum precision), articulated through a Pareto frontier that allows the health manager or researcher to calibrate the trade-off between sensitivity and specificity according to the requirements of the application scenario.

Additionally, the thesis proposed the Grey-Zone Cost-based Mixture Deferral (GZ-CMD), a self-calibrating operational framework that extends the base framework by integrating anchor-based calibration, an asymmetric expected-loss decision policy, and large language model-assisted review. In surveillance mode, GZ-CMD achieved  $F_1=0.954$ ; in confirmation mode, precision of 0.957. Considering the full pipeline cascade , comprising Fellegi–Sunter thresholds, deterministic guardrails, expected-loss engine, and dual-agent LLM review, the human clerical review volume was reduced from 21,620 to only 41 pairs (99.8% reduction, three orders of magnitude), processed in approximately 63 minutes of inference.

**Keywords:** Probabilistic record linkage. Machine learning. Class imbalance. Post-processing. Tuberculosis. Mortality. Health surveillance. Large language models. Grey zone.

# Listas de Figuras

6.1	Volume de pares candidatos e pares verdadeiros por faixa do escore agregado (escala logarítmica) . . . . .	27
6.2	Proporção de pares verdadeiros em relação ao total de pares candidatos por faixa de escore, com destaque para a zona cinzenta (faixas 5 a 8) . . . . .	28
6.3	Comparação de desempenho (precisão, revocação e F1) entre modelos no cenário NB01, com limiar de decisão padrão. . . . .	29
6.4	Melhor F1 por categoria de decisão no estudo de ablação. . . . .	32
6.5	Fronteira de Pareto para seleção de pontos operacionais (precisão e revocação) em combinações de limiares do classificador e das regras. . . . .	33
6.6	Distribuição do F1 por configuração na validação cruzada ( <i>5-fold</i> ). . . . .	34
6.7	Sensibilidade do desempenho (F1) do classificador a diferentes estratégias de reamostragem para desbalanceamento. . . . .	35
6.8	Resumo SHAP para a classe positiva, indicando a contribuição média (valor absoluto) de cada atributo para a predição. . . . .	37
6.9	Importância dos atributos por faixa de escore, com destaque para mudanças de relevância na zona cinzenta. . . . .	38
7.1	Pares verdadeiros detectados e perdidos por método de classificação. . . . .	41
7.2	Volume total de pares encaminhados para revisão manual por método. . . . .	41
7.3	Custo operacional: número de revisões manuais por par verdadeiro recuperado. . . . .	42
8.1	Redução cumulativa da carga de revisão clerical e enriquecimento da prevalência ao longo das etapas do <i>framework GZ-CMD</i> (modo vigilância, MACD ativo, $n = 61.696$ pares). (a) Volume de pares encaminhados a revisão em escala logarítmica; percentuais indicam a redução em relação à etapa anterior. (b) Prevalência de pares verdadeiros (%) e contagem absoluta de verdadeiros positivos contidos em cada subconjunto. O enriquecimento de $18\times$ entre a base completa e o lote LLM evidencia o efeito concentrador da cascata. . . . .	60
A.1	Pipeline decisório do GZ-CMD com calibração por ancoras, política de custo e trilha de auditoria . . . . .	79

# Listas de Tabelas

2.1	Comparação entre estratégias de <i>linkage</i> de bases de dados. . . . .	5
6.1	Distribuição dos pares candidatos por faixa de escore do OpenRecLink . . . . .	27
6.2	Melhor configuração por categoria de classificação — estudo de ablação. . . . .	31
6.3	Dez melhores configurações por F1 — estudo de ablação. . . . .	31
6.4	Fronteira de Pareto: pontos operacionais do classificador híbrido (RF+SMOTE). .	33
6.5	Validação cruzada estratificada ( <i>5-fold</i> ) das configurações selecionadas. . . . .	34
6.6	Sensibilidade ao desbalanceamento: Random Forest com diferentes estratégias de reamostragem ( <i>5-fold CV</i> ). . . . .	35
6.7	Importância dos atributos por valores SHAP (média do valor absoluto, classe positiva). . . . .	36
7.1	Comparação de métodos: óbitos detectados, custo operacional e taxa corrigida . . . . .	40
7.2	Perfil dos óbitos recuperados pelo ML (não encontrados pelo limiar $\geq 8$ ) . . . . .	40
8.1	Validação dos conjuntos âncora contra o padrão-ouro. . . . .	52
8.2	Parâmetros de custo operacionais do GZ-CMD. . . . .	53
8.3	Avaliação comparativa de modelos para revisão assistida ( $n = 517$ pares da zona cinzenta; intervalos de Wilson, 95%). . . . .	56
8.4	Resultados do experimento de ablação (média $\pm$ desvio-padrão, 5 rodadas). A coluna “Revisões” indica o número absoluto de pares encaminhados à revisão por LLM em cada partição de teste. . . . .	57
8.5	Desempenho da revisão assistida contra o padrão-ouro ( $n=1.410$ ). . . . .	58
8.6	Redução cumulativa da carga de revisão clerical por componente do <i>framework</i> GZ-CMD (modo vigilância, MACD ativo, $n = 61.696$ pares candidatos). . . . .	58
A.1	Ablacão (MACD ON) no modo de confirmação: resultados por semente e resumo final. . . . .	78
A.2	Ablacão (MACD ON) no modo de vigilância: resultados por semente e resumo final. . . . .	80
A.3	Ablacão (MACD OFF) no modo de confirmação: resultados por semente e resumo final. . . . .	80

A.4 Ablacão (MACD OFF) no modo de vigilância: resultados por semente e resumo final. . . . .	80
----------------------------------------------------------------------------------------------	----



# **Lista de Abreviaturas e Siglas**

ADASYN	<i>Adaptive Synthetic Sampling</i> — sobreamostragem sintética adaptativa
AUC-PR	<i>Area Under the Precision-Recall Curve</i> — área sob a curva precisão-sensibilidade
AUC-ROC	<i>Area Under the Receiver Operating Characteristic Curve</i> — área sob a curva ROC
CV	<i>Cross-Validation</i> — validação cruzada
F-S	Fellegi–Sunter (teoria de decisão para <i>record linkage</i> )
GAL	Gerenciador de Ambiente Laboratorial
GZ-CMD	<i>Grey-Zone Cost-based Mixture Deferral</i> — arcabouço operacional para a zona cinzenta
IC	Intervalo de Confiança
IESC	Instituto de Estudos em Saúde Coletiva
KDD	<i>Knowledge Discovery in Databases</i> — descoberta de conhecimento em bases de dados
LGPD	Lei Geral de Proteção de Dados
LightGBM	<i>Light Gradient Boosting Machine</i>
LLM	<i>Large Language Model</i> — modelo de linguagem de grande porte
MACD	<i>Multi-Algorithm Consensus with Deferral</i> — consenso multi-algoritmo com diferimento
ML	<i>Machine Learning</i> — aprendizado de máquina
MLP	<i>Multilayer Perceptron</i> — perceptron multicamadas
RF	<i>Random Forest</i> — floresta aleatória
RL	Rregressão Logística
SHAP	<i>SHapley Additive exPlanations</i> — valores de Shapley para interpretabilidade
SIA-SUS	Sistema de Informações Ambulatoriais do SUS
SIH-SUS	Sistema de Informações Hospitalares do SUS
SIM	Sistema de Informações sobre Mortalidade

# Sumário

<b>Lista de Figuras</b>	<b>xii</b>
<b>Lista de Tabelas</b>	<b>xiii</b>
<b>Lista de Abreviaturas e Siglas</b>	<b>xvi</b>
<b>Sumário</b>	<b>xvii</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Contextualização . . . . .	1
1.2 Problema e lacuna . . . . .	1
1.3 Proposta e estratégia . . . . .	2
1.4 Organização da tese . . . . .	3
<b>2 Referencial Teórico</b>	<b>4</b>
2.1 <i>Linkage</i> de bases de dados . . . . .	4
2.2 Estratégias de classificação . . . . .	5
2.3 Aprendizado de máquina aplicado ao <i>linkage</i> . . . . .	6
2.4 Desbalanceamento de classes no <i>linkage</i> . . . . .	7
2.5 Comparador de registros e ajuste de escores . . . . .	7
<b>3 Justificativa</b>	<b>9</b>
3.1 Lacuna do conhecimento . . . . .	9
3.2 Justificativas específicas . . . . .	10
3.3 A tuberculose como condição marcadora . . . . .	11
3.4 Urgência em contextos de crises sanitárias . . . . .	12
3.5 Vinculação institucional . . . . .	12
<b>4 Objetivos</b>	<b>14</b>
4.1 Objetivo geral . . . . .	14
4.2 Objetivos específicos . . . . .	14

<b>5</b>	<b>Método</b>	<b>16</b>
5.1	Desenho do estudo . . . . .	16
5.2	Fontes de dados . . . . .	16
5.3	Base de pares candidatos . . . . .	17
5.4	Comparador de registros . . . . .	18
5.5	Engenharia de atributos . . . . .	19
5.6	Estratégias de análise . . . . .	20
5.6.1	Análise comparativa de técnicas . . . . .	20
5.6.2	Estratégia de maximização da sensibilidade . . . . .	21
5.6.3	Estratégia de maximização da precisão . . . . .	21
5.6.4	Análises complementares e validação de robustez . . . . .	22
5.7	Métricas de avaliação . . . . .	22
5.8	Ambiente computacional . . . . .	23
<b>6</b>	<b>Resultados</b>	<b>25</b>
6.1	Base de dados e desbalanceamento . . . . .	25
6.2	Faixas de escore do OpenRecLink e zona cinzenta . . . . .	26
6.3	Engenharia de atributos e modelos avaliados . . . . .	27
6.4	Análise comparativa de técnicas (NB01) . . . . .	28
6.5	Estratégia de maximização da revocação (NB02) . . . . .	29
6.6	Estratégia de maximização da precisão (NB03) . . . . .	30
6.7	Estudo de ablação e fronteira de Pareto . . . . .	30
6.8	Robustez, sensibilidade ao desbalanceamento e interpretabilidade . . . . .	32
6.9	Síntese e recomendações operacionais . . . . .	37
<b>7</b>	<b>Discussão</b>	<b>39</b>
7.1	Impacto epidemiológico e operacional . . . . .	39
7.2	Dois <i>pipelines</i> e escolha de ponto operacional . . . . .	40
7.3	Pensamento sistêmico e crises sanitárias . . . . .	42
7.4	Episódios de cuidado e itinerário terapêutico . . . . .	44
7.5	Painéis de monitoramento e inteligência de dados em saúde . . . . .	45
7.6	Contribuição do <i>framework</i> em relação ao uso isolado de classificadores . . . . .	46
7.7	Limitações e generalização . . . . .	47
<b>8</b>	<b>Proposta de Arcabouço Operacional Auto-Calibrável (GZ-CMD)</b>	<b>50</b>
8.1	Motivação e lacunas operacionais . . . . .	50
8.2	Dados, delineamento e representação . . . . .	51
8.2.1	Bases de dados e período de estudo . . . . .	51
8.2.2	Padrão-ouro e validade de referência . . . . .	51
8.2.3	Delineamento experimental . . . . .	51

8.2.4	Representação dos pares e engenharia de atributos . . . . .	51
8.3	Métodos: o <i>framework</i> GZ-CMD . . . . .	52
8.3.1	Calibração por âncoras e bandas de confiança . . . . .	52
8.3.2	Validação empírica dos conjuntos âncora . . . . .	52
8.3.3	Regras de guarda determinísticas . . . . .	53
8.3.4	Motor de política de decisão por perda esperada . . . . .	53
8.3.5	Fluxo operacional do <i>framework</i> . . . . .	54
8.4	Revisão clerical assistida . . . . .	55
8.4.1	Seleção do modelo de linguagem . . . . .	55
8.4.2	Considerações éticas . . . . .	57
8.5	Resultados e avaliação . . . . .	57
8.5.1	Experimento 1: ablação e impacto das medidas contínuas . . . . .	57
8.5.2	Experimento 2: desempenho da revisão assistida . . . . .	58
8.5.3	Redução cumulativa da carga de revisão clerical . . . . .	58
8.6	Discussão . . . . .	59
<b>9</b>	<b>Conclusões</b>	<b>63</b>
9.1	Síntese dos principais achados . . . . .	63
9.2	Atendimento aos objetivos específicos . . . . .	64
9.3	Contribuições . . . . .	65
9.4	Trabalhos futuros . . . . .	65
<b>Glossário de Termos Técnicos</b>		<b>75</b>
<b>A</b>	<b>Material Suplementar do <i>Framework</i> GZ-CMD</b>	<b>78</b>
A.1	Pseudocódigo do <i>Pipeline</i> GZ-CMD . . . . .	78
A.2	Tabelas Expandidas de Ablação . . . . .	78
A.3	Formalização Matemática . . . . .	81
A.3.1	Calibração Probabilística por Âncoras . . . . .	81
A.3.2	Função de Perda e Decisão em Três Vias . . . . .	82
A.3.3	Valor Esperado da Revisão (EVR) . . . . .	83
A.3.4	Relaxamento das Hipóteses de Fellegi–Sunter . . . . .	84
A.4	Prompts para Revisão Assistida por LLM . . . . .	87



# Capítulo 1

## Introdução

### 1.1 Contextualização

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 (DONABEDIAN, 1988; VIACAVA et al., 2012). No Brasil, essa integração enfrenta obstáculo estrutural. Os Sistemas de Informação em Saúde (SIS) foram concebidos para finalidades específicas e não dispõem de um identificador único comum, o que impõe a necessidade de técnicas de vinculação de registros para integrar dados de diferentes fontes (JR; COELI, 2000). Bases como o SIM, o Sinan, o SIH-SUS e o Sinasc registram eventos complementares sobre um mesmo indivíduo, porém sem mecanismo padronizado de interligação (CHRISTEN, 2012; BARRETO et al., 2019).

O *linkage* (vinculação de registros, do inglês *record linkage*) de bases de dados constitui etapa indispensável para a produção de indicadores de qualidade do cuidado, a identificação de subnotificação de agravos e a construção de trajetórias longitudinais de pacientes (FEL-LEGI; SUNTER, 1969; NEWCOMBE et al., 1959). No contexto da vigilância epidemiológica brasileira, o *linkage* probabilístico, operacionalizado por ferramentas como o OpenRecLink (JR; COELI, 2000), é amplamente empregado para vincular registros de diferentes sistemas com base na comparação de variáveis de identificação pessoal (nome, nome da mãe, data de nascimento, sexo e município de residência) (COELI; JR., 2002).

Não obstante a consolidação dessa abordagem, a qualidade dos dados vinculados permanece dependente de limiares de classificação definidos empiricamente e de procedimentos de revisão manual cuja reproduzibilidade é limitada (CHRISTEN, 2012). A superação dessas limitações constitui desafio relevante para a produção de informação oportuna e qualificada em saúde pública.

## 1.2 Problema e lacuna

No modelo de decisão proposto por Fellegi e Sunter (1969), dois limiares dividem o espaço de escores de similaridade em três regiões: pares aceitos, pares rejeitados e uma faixa intermediária denominada área cinza (*grey zone*). Os potenciais pares situados nessa região apresentam escores insuficientes para classificação automatizada e constituem o principal nó crítico do processo de vinculação (CHRISTEN, 2012). A resolução da área cinza é tradicionalmente realizada por revisão manual (*clerical review*), procedimento dispendioso, pouco escalável e sujeito à variabilidade intra e interavaliador (NASSEH; STAUSBERG, 2016). No cenário de vinculação SIM–Sinan-TB investigado nesta tese, essa região compreende mais de 21 mil potenciais pares, volume que torna a revisão humana exaustiva inviável na prática rotineira dos serviços de vigilância.

A adoção de limiares fixos sobre escores agregados acarreta perda expressiva de pares verdadeiros, especialmente quando os campos de identificação apresentam erros de digitação, abreviações ou incompletude. O caso da tuberculose ilustra essa fragilidade. No *linkage* entre o SIM e o Sinan para TB, condição marcadora (*tracer condition*) da qualidade do cuidado em saúde, os indicadores de mortalidade evidenciam subnotificação significativa (OLIVEIRA et al., 2012; SOUSA; PINHEIRO, 2011; PINHEIRO; ANDRADE; OLIVEIRA, 2012). A perda de pares verdadeiros nesse cenário tende a comprometer a estimativa da magnitude de desfechos desfavoráveis e a avaliação da efetividade do programa de controle da tuberculose (ROCHA et al., 2015; World Health Organization, 2024).

Embora a aplicação de técnicas de aprendizado de máquina (*machine learning*) ao *linkage* venha sendo investigada em contextos internacionais (BINETTE; STEORTS, 2022; VO; LEE, 2019; JIAO et al., 2021; ALMADANI et al., 2026), a literatura brasileira sobre o tema é incipiente, restringindo-se predominantemente a abordagens determinísticas e probabilísticas tradicionais. Soma-se a isso a carência de investigações que avaliem sistematicamente o impacto de estratégias de balanceamento de classes e de ajustes nos pontos de corte do comparador sobre a acurácia do processo de vinculação em bases de saúde brasileiras (HE; GARCIA, 2009; HASSANI et al., 2025).

## 1.3 Proposta e estratégia

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 por um comparador de registros probabilístico (LUCENA, 2013; JARDIM, 2024), a partir dos potenciais pares gerados pelo OpenRecLink (JR; COELI, 2000), 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.

A arquitetura proposta organiza-se em três camadas: (i) o OpenRecLink, responsável pela blocagem e geração de potenciais pares; (ii) o comparador de registros (JARDIM, 2024; LUCENA, 2013), responsável pelo cálculo de 29 subescores de similaridade e um escore agregado para cada par candidato; e (iii) a camada de pós-processamento por aprendizado de máquina, que constitui a contribuição central desta tese. Essa separação de responsabilidades permite que cada componente evolua independentemente e facilita a reproduzibilidade do protocolo experimental.

Operacionalmente, o protocolo estrutura-se em seis etapas: (i) execução do *linkage* probabilístico mediante OpenRecLink, com múltiplos passos de blocagem (COELI; JR., 2002); (ii) cálculo dos escores de similaridade campo a campo pelo comparador de registros; (iii) engenharia de atributos, incluindo indicadores binários de concordância, escores ponderados e termos de interação; (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 sensibilidade (*recall*) ou precisão (*precision*), conforme o objetivo do estudo; e (vi) integração de regras de negócio baseadas no conhecimento do domínio. O *framework* (estrutura metodológica) resultante é configurável, disponibilizando dois *pipelines* (sequências operacionais) pré-definidos, um orientado à vigilância e outro à confirmação, articulados por uma fronteira de Pareto que permite ao gestor ou pesquisador calibrar o equilíbrio entre sensibilidade e especificidade conforme as necessidades do cenário de aplicação.

O protocolo foi desenhado para ser reproduzível 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 (CHRISTEN, 2012; ASHER et al., 2020; SCHNELL; WEIAND, 2023). A descrição detalhada de cada etapa encontra-se no Capítulo 5.

Como extensão operacional, a tese propõe ainda o *Grey-Zone Cost-based Mixture Deferral* (GZ-CMD), arcabouço auto-calibrável que substitui limiares fixos por política de decisão fundamentada em perda esperada com custos assimétricos ( $C_{FN}$ ,  $C_{FP}$ ,  $C_{LLM}$ ). O GZ-CMD integra calibração por conjuntos âncora, regras de guarda determinísticas e revisão assistida por modelo de linguagem de grande porte, configurando triagem formal da incerteza na zona cinzenta. Aplicado ao cenário de vigilância, o arcabouço reduz de 21.620 para 41 os pares encaminhados à revisão humana, redução de 99,8% (três ordens de grandeza), processados em aproximadamente 63 minutos de tempo de API. O detalhamento do arcabouço é apresentado no Capítulo 8.

## 1.4 Organização da tese

A tese está organizada em nove capítulos. O Capítulo 2 apresenta o referencial teórico, abordando os fundamentos do *linkage* de bases de dados, as estratégias de classificação de pares, as técnicas de aprendizado de máquina empregadas e o problema do desbalanceamento de classes. A justificativa do estudo, com ênfase na tuberculose como condição marcadora e na urgência de protocolos automatizados, é desenvolvida no Capítulo 3, enquanto os objetivos geral e específicos são enunciados no Capítulo 4. No Capítulo 5, descreve-se o método, incluindo fontes de dados, engenharia de atributos, estratégias de análise e métricas de avaliação. Os resultados dos experimentos são apresentados no Capítulo 6, seguidos pela discussão dos achados, implicações epidemiológicas e limitações no Capítulo 7. O Capítulo 8 propõe o arcabouço operacional GZ-CMD, que integra calibração por âncoras, política de decisão por perda esperada e revisão assistida por modelos de linguagem para governança da incerteza na zona cinzenta. Por fim, o Capítulo 9 sintetiza as contribuições e indica direções para trabalhos futuros.

# Capítulo 2

## Referencial Teórico

### 2.1 *Linkage de bases de dados*

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, possibilitando a integração de informações provenientes de diferentes fontes para a produção de conhecimento em saúde (CHRISTEN, 2012). Originalmente proposto por Newcombe e colaboradores (1959) e formalizado por Fellegi e Sunter (1969), o método evoluiu consideravelmente, incorporando métricas de comparação de campos textuais (JARO, 1989; WINKLER, 1990), estratégias de blocagem para viabilizar o processamento de grandes volumes (COELI; JR., 2002) e, mais recentemente, abordagens baseadas em aprendizado de máquina para a classificação automatizada de potenciais pares (CHRISTEN, 2012; ENAMORADO; FIFIELD; IMAI, 2019a).

No campo da saúde pública, o *linkage* tem se mostrado particularmente relevante para a identificação de óbitos por tuberculose não notificados ao sistema de vigilância (SOUZA; PINHEIRO, 2011; OLIVEIRA et al., 2012), a análise de causas múltiplas de morte em coortes de pacientes (ROCHA et al., 2015) e a melhoria da qualidade dos dados registrados em sistemas nacionais (BARTHOLOMAY et al., 2014). A técnica tem se consolidado como ferramenta relevante para a avaliação do desempenho de sistemas e serviços de saúde (VIACAVA et al., 2012).

No Brasil, o Sistema Único de Saúde (SUS) dispõe de um amplo conjunto de Sistemas de Informação em Saúde (SIS), cada qual concebido para finalidades distintas (PAIM et al., 2011; MACINKO; HARRIS, 2015). Entre os principais sistemas, destacam-se o SIM (mortalidade), o Sinan (agravos de notificação), o SIH-SUS (internações hospitalares), o SIA-SUS (procedimentos ambulatoriais), o Sinasc (nascidos vivos), o GAL (exames laboratoriais) e o SITETB (tuberculose drogarresistente). Uma característica fundamental dessas bases é a ausência de um identificador unívoco que permita a vinculação inequívoca entre registros de diferentes sistemas (JR; COELI, 2000). Embora o Cartão Nacional de Saúde (CNS) tenha sido concebido

com essa finalidade, sua cobertura permanece incompleta e sua qualidade de preenchimento é heterogênea, limitando sua utilidade como chave primária para o *linkage* direto (BARRETO et al., 2019). Essa lacuna impõe a necessidade de métodos indiretos de relacionamento, baseados na comparação de variáveis de identificação comuns (nome, nome da mãe, data de nascimento, sexo e município de residência), que estão sujeitos a erros de digitação, abreviações, homônimos e incompletude (CHRISTEN, 2012; SILVA, 2012).

## 2.2 Estratégias de classificação

Diferentes estratégias têm sido empregadas para a classificação de potenciais pares no *linkage*, podendo ser agrupadas em três abordagens principais. O *linkage* determinístico exige concordância exata (ou quase exata) em variáveis-chave, apresentando elevada especificidade, porém baixa sensibilidade na presença de imperfeições nos dados. O *linkage* probabilístico, fundamentado no modelo de Fellegi e Sunter (FELLEGI; SUNTER, 1969), atribui pesos diferenciados por variável e permite acomodar imperfeições, sendo amplamente empregado no contexto brasileiro por meio do OpenRecLink (JR; COELI, 2000). As abordagens baseadas em aprendizado de máquina capturam padrões não lineares de concordância e incorporam múltiplas variáveis simultaneamente, ao custo de maior dependência de conjuntos de treinamento rotulados (CHRISTEN, 2012). O Quadro 2.1 sintetiza as principais características de cada abordagem.

Tabela 2.1: 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

No modelo probabilístico, dois limiares dividem o espaço de escores em três regiões:

acima do limiar superior, os pares são aceitos; abaixo do limiar inferior, são descartados; entre ambos, configura-se a área cinza (*grey zone*), composta por potenciais pares cujos escores não são suficientemente elevados para aceitação automática nem suficientemente baixos para rejeição (FELLEGI; SUNTER, 1969; CHRISTEN, 2012). A extensão dessa região depende da qualidade das variáveis de identificação e do poder discriminatório das métricas de comparação empregadas. A resolução tradicional por revisão manual (*clerical review*) é demorada, custosa, não escalável e sujeita à variabilidade entre avaliadores (NASSEH; STAUSBERG, 2016). Nesse contexto, a aplicação de classificadores supervisionados como etapa de pós-processamento constitui alternativa promissora, possibilitando a priorização de sensibilidade ou precisão conforme o objetivo do estudo (HAND; CHRISTEN, 2018).

## 2.3 Aprendizado de máquina aplicado ao *linkage*

Para a classificação de potenciais pares e a recuperação de pares verdadeiros da área cinza, diferentes técnicas de aprendizado de máquina supervisionado podem ser empregadas, selecionadas por suas propriedades complementares no tratamento de dados desbalanceados e na modelagem de padrões não lineares de similaridade entre campos de identificação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A **regressão logística** constitui modelo linear generalizado que estima a probabilidade de pertinência à classe positiva por meio de função logística aplicada a combinação linear das variáveis preditoras. Apesar de sua simplicidade, apresenta vantagens como modelo de referência (*baseline*): eficiência computacional, interpretabilidade e produção de probabilidades calibradas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O algoritmo de **Floresta Aleatória** (*Random Forest*), proposto por Breiman (2001), baseia-se na construção de um conjunto (*ensemble*) de árvores de decisão treinadas em amostras aleatórias dos dados, com seleção aleatória de subconjuntos de variáveis em cada nó. A predição final é obtida por votação majoritária. A técnica apresenta robustez ao sobreajuste (*overfitting*) e mecanismos intrínsecos para avaliação da importância de variáveis.

Na família de **Gradient Boosting**, modelos são construídos sequencialmente, de modo que cada um corrige os erros residuais dos anteriores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Duas implementações foram consideradas: o **XGBoost** (*eXtreme Gradient Boosting*) (CHEN; GUESTRIN, 2016), com regularização L1/L2 e tratamento nativo de valores ausentes; e o **LightGBM** (*Light Gradient Boosting Machine*), com amostragem baseada em gradiente para redução do custo computacional.

Para a separação não linear de classes, a **Máquina de Vetores de Suporte** (*Support Vector Machine, SVM*) busca o hiperplano ótimo que maximiza a margem entre as classes. Funções de núcleo (*kernel*), como o núcleo de base radial (*RBF*), projetam os dados em espaço de maior dimensão, viabilizando a separação de classes não linearmente separáveis no espaço original (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Como representante de redes neurais artificiais, o **Perceptron Multicamadas** (*Multi-layer Perceptron, MLP*) aprende representações hierárquicas dos dados por meio do algoritmo de retropropagação do erro. Sua flexibilidade arquitetural permite a modelagem de relações altamente não lineares, embora com maior sensibilidade à configuração de hiperparâmetros (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Além desses classificadores individuais, foram empregadas técnicas de **combinação de modelos** (*ensemble*), cuja eficácia para o *linkage* tem sido demonstrada na literatura (SARIYAR; BORG, 2012; VO; LEE, 2019). O *stacking* (empilhamento) treina um metaclassificador sobre as previsões de classificadores de base, enquanto a votação por consenso (*consensus voting*) exige concordância entre modelos independentes para classificação positiva. Complementarmente, regras de classificação baseadas em conhecimento de domínio podem ser combinadas com modelos de aprendizado de máquina em abordagens híbridas, integrando evidências estatísticas e conhecimento especializado (JIAO et al., 2021).

## 2.4 Desbalanceamento de classes no *linkage*

O desbalanceamento entre as classes de pares verdadeiros e não-pares constitui característica estrutural do *linkage*. A combinação de registros entre duas bases gera número de potenciais pares 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, a proporção de pares verdadeiros permanece tipicamente muito pequena (CHRISTEN, 2012; HE; GARCIA, 2009). Algoritmos de classificação treinados em conjuntos desbalanceados tendem a apresentar viés em favor da classe majoritária, resultando em modelos que falham na identificação de pares verdadeiros (HASSANI et al., 2025).

Diferentes estratégias têm sido propostas para mitigar esse efeito. A sobreamostragem da classe minoritária, cujo representante mais empregado é o algoritmo SMOTE (*Synthetic Minority Over-sampling Technique*) (CHAWLA et al., 2002), gera exemplos sintéticos da classe sub-representada. Variações incluem o Borderline-SMOTE, que concentra a geração nas regiões de fronteira, e o ADASYN (*Adaptive Synthetic Sampling*), que adapta a densidade conforme a dificuldade de classificação. Técnicas combinadas, como o SMOTE-Tomek, integram sobreamostragem com remoção de exemplos ruidosos. A ponderação de classes (*class weights*) atribui pesos diferenciados na função de custo, penalizando erros sobre a classe minoritária sem alterar a composição do conjunto de treinamento. Métodos de *ensemble*, como o *Balanced Random Forest*, incorporam estratégias de balanceamento em cada iteração (HE; GARCIA, 2009).

A escolha da estratégia adequada depende das características do problema e dos objetivos do estudo. No contexto do *linkage* em bases de saúde, a comparação sistemática de diferentes estratégias de balanceamento constitui questão ainda insuficientemente explorada, cujos resultados podem contribuir para a padronização de protocolos e a melhoria da acurácia do processo

de vinculação (CHRISTEN, 2012; HASSANI et al., 2025).

## 2.5 Comparador de registros e ajuste de escores

Os escores de similaridade produzidos pelo comparador de registros (LUCENA, 2013; JARDIM, 2024), a partir dos potenciais pares gerados pelo OpenRecLink (JR; COELI, 2000), representam medida agregada do grau de concordância entre os campos de identificação de cada par candidato. A classificação final depende da definição de pontos de corte (*thresholds*) sobre esses escores, que delimitam as fronteiras entre pares aceitos, pares rejeitados e a área cinza (FELLEGI; SUNTER, 1969).

A definição de pontos de corte adequados é tarefa complexa. Pontos de corte excessivamente elevados resultam em alta especificidade, porém com perda de pares verdadeiros que apresentam escores intermediários, frequentemente aqueles com campos incompletos ou com erros de grafia. Pontos de corte excessivamente baixos incorporam falsos positivos, comprometendo a confiabilidade dos vínculos (CHRISTEN, 2012).

Neste trabalho, propõe-se a utilização de técnicas de aprendizado de máquina para a análise e otimização dos pontos de corte, empregando os escores de similaridade individuais (e não apenas o escore composto) como variáveis preditoras. Essa abordagem possibilita a identificação de padrões que indicam pares verdadeiros mesmo em regiões de escore intermediário. A otimização pode ser orientada por métricas como o F<sub>1</sub>-Score (HAND; CHRISTEN, 2018), 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. O capítulo seguinte contextualiza a relevância epidemiológica dessa abordagem, argumentando que a tuberculose constitui condição marcadora adequada para a validação de tais estratégias.

# Capítulo 3

## Justificativa

O *linkage* (vinculação 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 (JR; COELI, 2000; CHRISTEN, 2012). A ausência de um identificador unívoco que perpassasse 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 (FELLEGI; SUNTER, 1969; JR; COELI, 2000), 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 (FELLEGI; SUNTER, 1969), é amplamente empregado em estudos epidemiológicos brasileiros por meio de ferramentas como o OpenRecLink (JR; COELI, 2000) e estratégias de blocagem (COELI; JR., 2002). Persistem, contudo, desafios na classificação dos pares situados na “área cinza” do comparador (cf. Seção 2.2): essa faixa intermediária de escores demanda revisão manual (*clerical review*), procedimento dispendioso, pouco escalável e sujeito à variabilidade inter-avaliadores (CHRISTEN, 2012). A resolução automatizada da área cinza constitui, portanto, o nó crítico que motiva a presente investigação.

### 3.1 Lacuna do conhecimento

Embora a aplicação de técnicas de aprendizado de máquina (*machine learning*) ao *linkage* venha sendo investigada em contextos internacionais (CHRISTEN, 2012; BINETTE; STEORTS, 2022; VO; LEE, 2019), 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 (SCHNELL; WEIAND, 2023). 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 in-

vestigações 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. Essa lacuna é particularmente 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 (JR; COELI, 2000; COELI; JR., 2002). A revisão manual da área cinza, quando realizada, configura etapa artesanal e não reproduzível, comprometendo a comparabilidade entre estudos (CHRISTEN, 2012). 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 conjunto de classificadores (*ensemble*) (VO; LEE, 2019) e métodos híbridos que combinam técnicas probabilísticas com aprendizado supervisionado (JIAO et al., 2021; ALMADANI et al., 2026) podem elevar substancialmente a acurácia da vinculação. Para fins de saúde pública, o valor dessas estratégias também se expressa na redução do custo de *clerical review* (CHRISTEN, 2012) e na possibilidade de mensurar o impacto epidemiológico da recuperação de pares verdadeiros em regiões de maior incerteza do escore.

### **3.2 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 (OLIVEIRA et al., 2012), 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 (World Health Organization, 2024; Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente, 2024). 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 (SOUZA; PINHEIRO, 2011; PINHEIRO; ANDRADE; OLIVEIRA, 2012; ROCHA et al., 2015). 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 (ROCHA et al., 2015).

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 (HE; GARCIA, 2009). 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 (2025) 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 (VIACAVA et al., 2012). 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 estimativa de indicadores populacionais de saúde (HANEEF et al., 2022).
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 (CHRISTEN, 2012; PAIXÃO et al., 2017).
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 (JR; COELI, 2000; COELI; JR., 2002; OLIVEIRA et al., 2016). 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.

### 3.3 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 (1973). Segundo esse conceito, 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 (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde, 2019). 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 (SOUSA; PINHEIRO, 2011; OLIVEIRA et al., 2012).

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 (PINHEIRO; ANDRADE; OLIVEIRA, 2012; SOUSA; PINHEIRO, 2011; OLIVEIRA et al., 2012). Investigações subsequentes qualificaram variáveis do Sinan-TB por meio de regras de *scripting* aplicadas sobre dados vinculados (ROCHA et al., 2019) e analisaram as causas múltiplas de morte em coortes de pacientes notificados (ROCHA et al., 2015). A taxa de cura no Brasil permanece abaixo do preconizado pela Organização Mundial da Saúde, e os índices de abandono persistem elevados (Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente, 2024; World Health Organization, 2024), 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.

### 3.4 Urgência em contextos de crises sanitárias

A necessidade de protocolos automatizados e reproduutíveis 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 (RANZANI et al., 2021; MAIA et al., 2022). 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 (HALLAL et al., 2020). 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.

### **3.5 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 avaliação de serviços de saúde (JR; COELI, 2000; COELI; JR., 2002). 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. A partir da contextualização ora apresentada, o capítulo seguinte enuncia os objetivos geral e específicos que norteiam a investigação.

# Capítulo 4

## Objetivos

### 4.1 Objetivo geral

O presente estudo tem como objetivo geral desenvolver e avaliar um arcabouço (*framework*) configurável de pós-processamento, baseado em aprendizado de máquina (*machine learning*), para a classificação de pares candidatos produzidos pelo *linkage* probabilístico entre bases de dados de saúde. Busca-se aumentar a acurácia do processo e recuperar registros da zona cinzenta que permaneceriam não classificados ou incorretamente descartados pelo método probabilístico convencional. O arcabouço proposto foi aplicado ao *linkage* entre o Sistema de Informação sobre Mortalidade (SIM) e o Sistema de Informação de Agravos de Notificação para tuberculose (Sinan-TB) no município do Rio de Janeiro, no período de 2006 a 2016, contribuindo para a qualificação dos dados vinculados e para a produção de indicadores de desempenho de sistemas e serviços de saúde.

### 4.2 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* (CHAWLA et al., 2002), *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 reproduutíveis 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.
7. Propor e avaliar um arcabouço operacional auto-calibrável (*Grey-Zone Cost-based Mixture Deferral*, GZ-CMD) que integre calibração por conjuntos âncora, política de decisão por perda esperada com custos assimétricos e revisão assistida por modelo de linguagem de grande porte, com vistas à governança formal da incerteza na zona cinzenta do *linkage*, à redução do volume de revisão manual e à realocação do revisor humano para os casos genuinamente ambíguos.

O capítulo seguinte descreve o percurso metodológico adotado para a consecução desses objetivos, detalhando as fontes de dados, os algoritmos empregados e as estratégias de validação experimental.

# Capítulo 5

## Método

### 5.1 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.

### 5.2 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) (PAIM et al., 2011).
- **Sistema de Informação de Agravos de Notificação, Tuberculose (Sinan):** 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 (OLIVEIRA et al., 2012; SANTOS; COELI et al., 2018).

Os registros correspondem ao município do Rio de Janeiro e foram previamente submetidos a *linkage* probabilístico por meio do programa OpenRecLink (JR; COELI, 2000), 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 3.2, 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 (OLIVEIRA et al., 2012; BARTHOLOMAY et al., 2014; OLIVEIRA et al., 2016).

### 5.3 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 (COELI; JR., 2002). 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 comparador de registros
- **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.

O padrão-ouro (*gold standard*) foi construído por meio de revisão clerical (*clerical review*) conduzida por um único revisor, vinculado ao Instituto de Estudos em Saúde Coletiva da Universidade Federal do Rio de Janeiro (IESC-UFRJ). A classificação seguiu protocolo sequencial. O procedimento consistiu na consulta manual aos sistemas de origem (SIM e Sinan) para cada par candidato, com avaliação do status de par verdadeiro ou não-par com base na concordância das variáveis de identificação e em evidências complementares disponíveis nos registros originais. Embora a revisão por avaliador único seja uma prática frequente em estudos operacionais de *linkage* no contexto brasileiro (COELI et al., 2021), a ausência de um segundo revisor independente constitui limitação relevante, discutida adiante.

Reconhece-se que a utilização de um único revisor impede o cálculo de concordância inter-avaliadores (coeficiente kappa) e, consequentemente, a quantificação formal da reproduzibilidade do processo de rotulagem. Adicionalmente, a avaliação subjetiva de pares na zona cinzenta pode introduzir viés individual, especialmente quando os campos de identificação apresentam concordância parcial, sem que haja um segundo avaliador para mitigar tal viés por consenso. Há também o risco de falsos negativos decorrentes de falhas na blocagem. Pares verdadeiros que não compartilham nenhuma das chaves de blocagem utilizadas no OpenRecLink não são incluídos no conjunto de candidatos e, portanto, não podem ser avaliados pelo revisor nem recuperados pelo pós-processamento. Essas limitações devem ser consideradas na interpretação das métricas de desempenho, particularmente da sensibilidade, que pode estar sobreestimada em relação ao universo real de pares verdadeiros existentes nas bases de origem.

A base apresenta severo desbalanceamento de classes, com proporção aproximada de 1 par verdadeiro para cada 249 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 (CHRISTEN, 2012; HE; GARCIA, 2009). Esse desbalanceamento constitui nó crítico para a aplicação de classificadores supervisionados, demandando estratégias específicas de tratamento.

## 5.4 Comparador de registros

Os escores de similaridade utilizados como insumo para o pós-processamento não foram calculados pelo OpenRecLink, mas por um comparador de registros probabilístico independente, desenvolvido especificamente para este estudo. A ferramenta constitui uma reimplementação em Python do algoritmo proposto por Lucena (2013) em dissertação de mestrado apresentada ao Instituto de Estudos em Saúde Coletiva da UFRJ, cujo protótipo original em Java foi desenvolvido por Peçanha (2015) no mesmo grupo de pesquisa. A versão utilizada neste trabalho encontra-se publicada em repositório de código aberto, sob licença GPL-3.0 (JARDIM, 2024).

O comparador recebe como entrada os pares candidatos gerados pelo OpenRecLink após a etapa de blocagem e executa a comparação campo a campo entre os registros de cada par, produzindo 29 subescores de similaridade e um escore final agregado (*nota final*). As comparações implementadas abrangem seis categorias:

- **Nome próprio e nome da mãe:** normalização textual (remoção de acentos, preposições, sufixos), codificação fonética por algoritmo *Soundex* (CHRISTEN, 2012), distância de edição de Levenshtein, e ponderação por tabelas de frequência posicional que diferenciam nomes raros (mais informativos) de nomes comuns (menos discriminatórios) (WINKLER, 1990);
- **Data de nascimento:** comparação por componentes (dia, mês, ano), com pontuação diferenciada para concordância exata, concordância parcial (transposição de dia/mês) e

discordância;

- **Texto livre (endereço):** normalização, tokenização e cálculo de similaridade por coeficiente de Jaccard sobre conjuntos de termos;
- **Município de residência:** comparação exata do código IBGE, com pontuação binária;
- **Campos numéricos:** comparação direta de valores, incluindo número do logradouro e complemento.

A cada par candidato, o comparador atribui subscores para cada campo comparado e calcula o escore final como combinação ponderada dessas comparações. Os pesos e limiares do algoritmo original foram definidos por Lucena (2013) com base em 20 critérios de comparação e um limiar fixo de corte (escore  $\geq 4,33$ ). A versão empregada neste trabalho estende o conjunto de critérios com a inclusão de comparadores de endereço e município, totalizando as 29 variáveis de saída, e substitui o limiar fixo por uma camada de pós-processamento baseada em aprendizado de máquina, que explora a informação contida nos subscores individuais (e não apenas no escore agregado) para classificar pares na zona cinzenta, conforme detalhado nas seções seguintes.

Cabe distinguir, portanto, três camadas do processo de vinculação: (i) o OpenRecLink (JR; COELI, 2000), responsável pela blocagem e geração de pares candidatos; (ii) o comparador de registros (JARDIM, 2024; LUCENA, 2013), responsável pelo cálculo dos escores de similaridade campo a campo; e (iii) a camada de pós-processamento por aprendizado de máquina, que constitui a contribuição central desta tese. Essa separação de responsabilidades permite que cada componente evolua independentemente e facilita a reproduzibilidade do protocolo experimental.

## 5.5 Engenharia de atributos

A partir dos escores brutos de similaridade fornecidos pelo comparador de registros (JARDIM, 2024), 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, 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.

## 5.6 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 1.3.

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 (CHRISTEN, 2012; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; SARIYAR; BORG, 2012). 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 (JR; COELI, 2000) 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.

### **5.6.1 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 em conjunto de teste (partição *hold-out* estratificada) e, adicionalmente, por validação cruzada estratificada para avaliação de estabilidade, incluindo precisão, sensibilidade,  $F_1$ -Score, AUC-ROC e AUC-PR (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; HAND; CHRISTEN, 2018). 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.

### **5.6.2 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 zona cinzenta do comparador. Para tanto, foram empregadas técnicas de balanceamento de classes (*SMOTE* (CHAWLA et al., 2002), *Borderline-SMOTE*, *ADASYN* e *SMOTE-Tomek*), seguindo a lógica de combinação de sobreamostragem e subamostragem proposta por Hassani e colaboradores (2025), 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 (OLIVEIRA et al., 2012; SOUSA; PINHEIRO, 2011).

### **5.6.3 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 (CHEN; GUESTRIN, 2016) e LightGBM (KE et al., 2017)), 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 (2018), o F<sub>1</sub>-Score pode não refletir adequadamente o desempenho do classificador nesse cenário, recomendando-se a utilização complementar de métricas como AUC-PR.

#### **5.6.4 Análises complementares e validação de robustez**

Além das três estratégias centrais, foram conduzidas análises complementares com três objetivos: explicitar o comportamento do classificador em regiões de incerteza, formalizar a escolha de ponto operacional por ablação e avaliar a robustez dos modelos. A primeira dessas análises focalizou a zona cinzenta. O desempenho foi estratificado por faixas do escore final (*nota final*) do comparador de registros, caracterizando uma zona cinzenta de pares candidatos em que os métodos baseados apenas em limiar apresentam maior incerteza. Foi também realizado estudo de ablação para comparar configurações *rules-only*, *ML-only*, combinações híbridas (lógicas AND/OR), cascatas e consenso entre modelos, incluindo a exploração de grades de limiares e a identificação de pontos de Pareto (precisão × sensibilidade) para apoiar a recomendação de protocolos por contexto de uso.

Para avaliar estabilidade, as principais configurações foram submetidas à validação cruzada estratificada (*5-fold*), e foi conduzida análise de sensibilidade às estratégias de balanceamento de classes. Por fim, foram produzidas análises de interpretabilidade, combinando importância de atributos derivada de modelos baseados em árvores e valores SHAP (*Shapley Additive Explanations*) (LUNDBERG; LEE, 2017), incluindo a comparação da relevância dos atributos na zona cinzenta.

### **5.7 Métricas de avaliação**

O desempenho dos classificadores foi avaliado por meio das seguintes métricas, reconhecidas na literatura de *linkage* e de aprendizado de máquina (HAND; CHRISTEN, 2018; HASTIE; TIBSHIRANI; FRIEDMAN, 2009):

- **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.
- **F<sub>1</sub>-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 (HE; GARCIA, 2009).

A avaliação principal foi realizada em uma partição *hold-out* estratificada (70% para treinamento e 30% para teste), preservando a proporção de classes, 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. Em complemento, para as configurações principais, foram realizadas validações cruzadas estratificadas (*k-fold*) para estimar a variabilidade das métricas e avaliar a estabilidade do desempenho em diferentes partições dos dados, com vistas à padronização de protocolos de pós-processamento para o *linkage* em saúde.

## 5.8 Ambiente computacional

Todos os experimentos foram implementados na linguagem Python (versão 3.12.4), 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. As análises foram estruturadas em cadernos Jupyter (*Jupyter Notebooks*) e scripts auxiliares, executados de forma reproduzível por meio do framework *Papermill*, e versionados em repositório Git, assegurando rastreabilidade das etapas do processo analítico, em consonância com diretrizes metodológicas internacionais para a utilização de dados vinculados e técnicas de aprendizado de máquina na estimação de indicadores de saúde (HANEEF et al., 2022).

A opção por modelos baseados em árvore (*tree-based models*), tais como Floresta Aleatória, XGBoost e LightGBM, em detrimento de arquiteturas de aprendizado profundo (*deep learning*), decorreu de duas premissas metodológicas. A primeira é a escassez de dados rotulados. O volume limitado de pares verdadeiros disponíveis para treinamento (247 pares positivos, dos quais aproximadamente 74 compõem o conjunto de teste) desfavorece arquiteturas profundas, como redes recorrentes (LSTM) e *Transformers*, que demandam ordens de grandeza

superiores de exemplos rotulados para convergir adequadamente e evitar sobreajuste (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A literatura recente sobre dados tabulares com menos de 10.000 amostras corrobora essa escolha: modelos baseados em árvore de decisão com *gradient boosting* tendem a igualar ou superar redes neurais profundas nesse regime de dados, particularmente quando o número de atributos é moderado e as relações entre variáveis são predominante- mente de interação entre campos discretos ou semicontínuos. A segunda premissa diz respeito à interpretabilidade. A análise de valores SHAP (*Shapley Additive Explanations*) (LUNDBERG; LEE, 2017) aplicada a modelos arbóreos permite identificar, para cada par candidato, quais campos de identificação contribuíram para a classificação, possibilitando a auditoria das de- cisões automatizadas por profissionais de saúde. Modelos de aprendizado profundo, embora dotados de capacidade representacional superior, tendem a operar como aproximadores opacos cujas decisões são de difícil explicação em termos dos atributos de entrada, limitação relevante quando o resultado do *linkage* fundamenta ações de vigilância epidemiológica e investigação de óbito.

As extensões metodológicas introduzidas pelo arcabouço GZ-CMD , a saber, calibração discriminativa por conjuntos âncora, motor de decisão por perda esperada e protocolo de re- visão assistida por modelo de linguagem, são detalhadas no Capítulo 8, que descreve dados, delineamento e métricas específicos daquela etapa.

# Capítulo 6

## Resultados

Este capítulo expõe os resultados do pós-processamento por aprendizado de máquina (*machine learning*) aplicado aos pares candidatos gerados pelo *linkage* probabilístico entre o Sistema de Informações sobre Mortalidade (SIM) e o Sistema de Informação de Agravos de Notificação para tuberculose (Sinan). A estrutura segue a lógica do *framework* proposto: inicialmente, caracteriza-se o desbalanceamento da base e a distribuição de escores que delimita a zona cinzenta; em seguida, reportam-se os cenários experimentais (análise comparativa, estratégia de revocação e estratégia de precisão); por fim, consolidam-se as análises de ablação, robustez por validação cruzada, sensibilidade ao tratamento do desbalanceamento e interpretabilidade via SHAP (*SHapley Additive exPlanations*). Os detalhes metodológicos, incluindo fontes de dados, engenharia de atributos, métricas e ambiente computacional, encontram-se no Capítulo 5.

### 6.1 Base de dados e desbalanceamento

A base analisada compreendeu 61.696 pares candidatos extraídos do arquivo **COMPARADORSEMIDENT.csv**, dos quais 247 foram classificados como pares verdadeiros (*true matches*) após conferência manual, resultando em uma razão de desbalanceamento de aproximadamente 1:249 (CHRISTEN, 2012). Esse grau de desbalanceamento não é atípico: estudos de *linkage* em bases de saúde pública frequentemente reportam prevalências de pares verdadeiros inferiores a 1%, especialmente quando a etapa de blocagem é permissiva para maximizar a cobertura do padrão-ouro (*gold standard*) (HARRON et al., 2017; DOIDGE; HARRON, 2019). A consequência direta é que classificadores treinados sem tratamento do desbalanceamento tendem a otimizar a acurácia global, classificando quase todos os pares como negativos, acarretando revocação próxima de zero para a classe de interesse (HE; GARCIA, 2009; JOHNSON; KHOSHGOFTAAR, 2019).

Tal cenário impõe duas restrições operacionais concretas. Em primeiro lugar, mesmo uma taxa de falsos positivos aparentemente baixa (por exemplo, 0,5%) pode gerar centenas de revisões manuais desnecessárias quando aplicada a um universo de dezenas de milhares de

pares, elevando o custo de revisão manual a patamares incompatíveis com a prática rotineira da vigilância epidemiológica (COELI et al., 2021). Em segundo lugar, a avaliação do modelo não pode se restringir à acurácia: métricas como precisão, revocação e  $F_1$  constituem indicadores mais adequados para quantificar o desempenho em cenários raros (CHRISTEN, 2012; HE; GARCIA, 2009). Os experimentos baseados em partição fixa (*hold-out*) utilizaram divisão estratificada 70/30 (173 pares verdadeiros para treino e 74 para teste), preservando a proporção original de classes. As métricas reportadas seguem a definição apresentada na Seção 5.7, com ênfase em precisão, revocação e  $F_1$ .

## 6.2 Faixas de escore do OpenRecLink e zona cinzenta

Para cada par candidato gerado pelo OpenRecLink, o comparador de registros (LUCENA, 2013; JARDIM, 2024) produz um conjunto de 29 subescores de similaridade por atributo e um escore agregado final (*nota\_final*) que sintetiza a evidência de pareamento (ver Seção 5.4). Na prática, a abordagem mais simples (aqui denominada “ingênua”) consiste em aceitar como pares verdadeiros todos os registros cujo escore agregado supere um limiar fixo. Para avaliar os limites intrínsecos dessa estratégia, analisou-se a distribuição dos pares verdadeiros por faixas de escore.

A Tabela 6.1 e as Figuras 6.1 e 6.2 revelam um padrão de dispersão bimodal. Nas faixas superiores (9 a 10 e acima de 10), a concentração de pares verdadeiros é quase total: 97,7% e 98,0% dos registros nessas faixas constituem pareamentos corretos, correspondendo a 93 dos 247 pares verdadeiros (37,7%). Esses pares possuem concordância elevada em múltiplos atributos e são triviais para qualquer método de classificação. A região intermediária (faixas de 5 a 8), aqui denominada zona cinzenta (*grey zone*), concentra cerca de 47% dos pares verdadeiros (117 de 247) diluídos em um volume de mais de 21 mil pares candidatos, onde a proporção de verdadeiros varia de 0,12% (faixa 5 a 6) a 3,98% (faixa 7 a 8). Essa assimetria entre volume e prevalência explica por que limiares ingênuos enfrentam um dilema insolúvel nessa região: reduzir o limiar para capturar mais pares verdadeiros implica aceitar um aumento exponencial no número de falsos positivos, enquanto elevá-lo implica descartar quase metade dos pareamentos reais (DUVALL; KERBER; THOMAS, 2010; DOIJDGE; HARRON, 2018).

Esse resultado é consistente com a literatura sobre *linkage* probabilístico de bases de mortalidade e vigilância no Brasil, em que divergências de grafia em nomes, abreviações e inconsistências em campos secundários (endereço, município) reduzem o escore agregado de pares genuínos sem eliminá-los completamente (PAIXÃO et al., 2017; COELI et al., 2021). A existência de uma zona cinzenta com quase metade dos pares verdadeiros constitui a principal justificativa empírica para a aplicação de modelos de aprendizado de máquina, que podem explorar padrões multivariados nos subescores de similaridade para discriminar pares nessa região, possibilitando ganhos de revocação sem degradação proporcional da precisão (ENAMORADO; FIFIELD; IMAI, 2019b).

Tabela 6.1: 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%

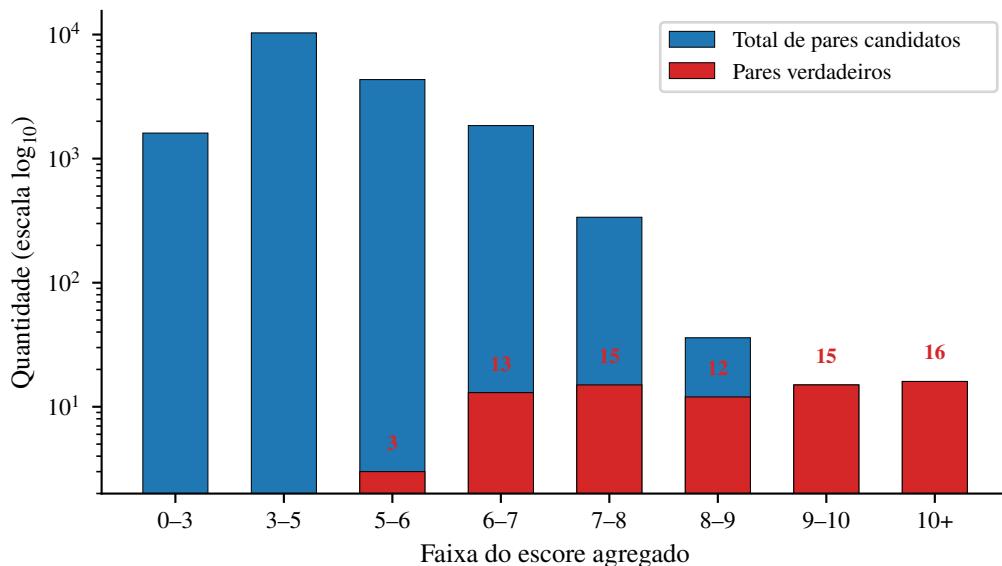


Figura 6.1: Volume de pares candidatos e pares verdadeiros por faixa do escore agregado (escala logarítmica).

### 6.3 Engenharia de atributos e modelos avaliados

Conforme descrito nas Seções 5.4 e 5.5, o vetor de entrada de cada par candidato foi construído a partir de três camadas de informação: (i) os 29 subescores de similaridade produzidos pelo comparador de registros (JARDIM, 2024), abrangendo nome, nome da mãe, data de nascimento, município de residência e endereço; (ii) o escore agregado (*nota\_final*) e o passo (*step*) de blocagem; e (iii) variáveis derivadas por engenharia de atributos (*feature engineering*), incluindo somas ponderadas, interações multiplicativas e marcadores binários de concordância perfeita. Para assegurar comparabilidade entre os cenários experimentais, os três experimentos foram reexecutados com um pipeline único de engenharia de atributos, totalizando 58 variáveis no vetor final em todos os casos.

A decisão de gerar atributos derivados, em vez de utilizar apenas os subescores brutos, apoiou-se na hipótese de que relações não lineares e interações entre campos são informativas para a classificação de pares na zona cinzenta. A variável *nome\_x\_dtnasc*, por exemplo, captura



Figura 6.2: Proporção de pares verdadeiros em relação ao total de pares candidatos por faixa de escore, com destaque para a zona cinzenta (faixas 5 a 8).

a concordância conjunta de nome e data de nascimento, combinação que eleva a probabilidade a posteriori de pareamento correto mesmo quando o escore agregado é moderado (FELLEGI; SUNTER, 1969; CHRISTEN, 2012). Em todos os cenários, foram testados classificadores tradicionais (regressão logística, *Support Vector Machine*) e métodos de *ensemble* (Random Forest, Gradient Boosting). Para mitigar o desbalanceamento, avaliaram-se estratégias de reamostragem do tipo SMOTE (*Synthetic Minority Oversampling Technique*) e variantes, além de ponderação de classes (*class weight*), conforme preconizado na literatura (CHAWLA et al., 2002; GALAR et al., 2012; HE; GARCIA, 2009). Os modelos e configurações específicas seguem o delineamento da Seção 5.6.

## 6.4 Análise comparativa de técnicas (NB01)

No cenário de análise comparativa, sete modelos foram avaliados com limiar de decisão (*threshold*) padrão de 0,5, de forma a estabelecer um panorama inicial de desempenho sob desbalanceamento severo. A Figura 6.3 sintetiza os resultados em precisão, revocação e  $F_1$ .

A combinação Random Forest com reamostragem SMOTE (RF+SMOTE) e o Gradient Boosting alcançaram os melhores equilíbrios entre precisão e revocação. Modelos lineares (regressão logística, SVM linear) tenderam a maximizar revocação ao custo de precisão substancialmente inferior, comportamento esperado quando o hiperplano de separação é forçado a capturar todos os positivos em um espaço onde a classe majoritária domina (HE; GARCIA, 2009). Uma explicação plausível para o melhor desempenho dos métodos de *ensemble* reside na sua capacidade de modelar interações não lineares entre subescores de similaridade, capturando padrões compostos (por exemplo, nome parcialmente concordante com data de nascimento

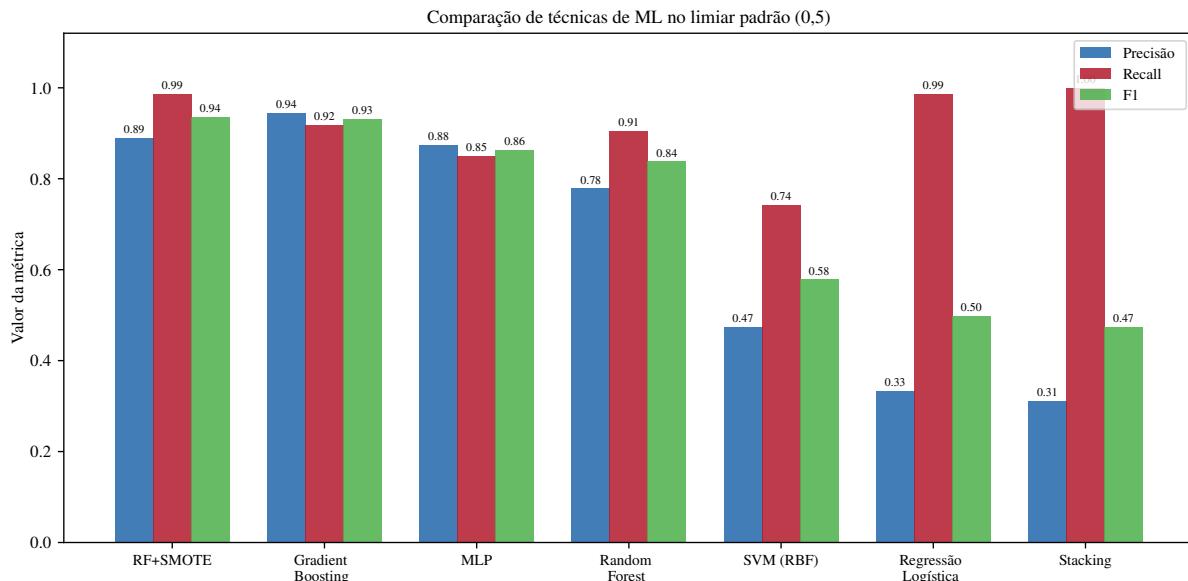


Figura 6.3: Comparação de desempenho (precisão, revocação e F1) entre modelos no cenário NB01, com limiar de decisão padrão.

exata) que um modelo linear projeta em um único coeficiente insuficiente (TYAGI; WILLIS, 2025). Esse resultado orientou a seleção de RF e Gradient Boosting como modelos base para as estratégias especializadas de revocação e de precisão, eliminando classificadores lineares das etapas subsequentes.

## 6.5 Estratégia de maximização da revocação (NB02)

No cenário de maximização da revocação (*recall*), o objetivo foi reduzir ao máximo a perda de pares verdadeiros, aceitando maior carga de revisão manual, cenário coerente com *pipelines* de vigilância epidemiológica em que cada par perdido pode representar um óbito não investigado (BARTHOLOMAY et al., 2014; LIMA et al., 2020). Foram avaliadas combinações de modelos, reamostragem e ponderação de classes, além de estratégias em cascata. Entre as configurações testadas, o modelo Gradient Boosting com SMOTE e pesos de classe atingiu revocação de 0,959 com precisão de 0,934, demonstrando que é possível obter sensibilidade elevada sem degradação extrema da precisão quando a engenharia de atributos explora informações complementares ao escore agregado (CHAWLA et al., 2002).

Tal resultado pode ser atribuído a dois fatores convergentes. Primeiro, as variáveis derivadas adicionam dimensões informativas (concordância parcial de nome, interações nome  $\times$  data de nascimento) que permitem ao classificador discriminar pares verdadeiros com escore agregado moderado. Segundo, a combinação de reamostragem com ponderação de classes ajusta tanto a distribuição de treinamento quanto a função de perda, gerando um efeito sinérgico que métodos isolados não alcançam (GALAR et al., 2012). Configurações extremamente permissivas (*ensemble* com limiar baixo) atingem revocação próxima de 1,0, porém impõem

custos operacionais desproporcionais (centenas de revisões adicionais), reforçando a necessidade de selecionar o ponto operacional com base em restrições reais de capacidade de revisão (DOIDGE; HARRON, 2019).

## 6.6 Estratégia de maximização da precisão (NB03)

No cenário de maximização da precisão (*precision*), o propósito foi gerar listas de candidatos de alta confiança para investigação, minimizando falsos positivos. Foram comparadas três abordagens: (i) regras determinísticas baseadas em concordância forte de atributos-chave (nome perfeito, data de nascimento exata, nome da mãe concordante), (ii) consenso por unanimidade entre modelos de aprendizado de máquina e (iii) um classificador híbrido que combina a probabilidade estimada pelo modelo supervisionado com a pontuação de regras determinísticas.

O resultado mais restritivo (regras com limiar alto) produz precisão próxima de 1,0, porém com revocação limitada, aceitável apenas em contextos de auditoria ou validação amostral. O consenso entre múltiplos modelos eleva a revocação mantendo precisão alta, ao exigir concordância majoritária em vez de unanimidade. A configuração híbrida permite parametrizar o equilíbrio precisão/revocação por meio de dois limiares independentes (probabilidade do classificador e escore das regras), conferindo ao operador controle explícito sobre o ponto operacional (SHAW et al., 2022). Essa flexibilidade motivou a formalização do *framework* híbrido e o estudo de ablação a seguir, cujo objetivo é quantificar a contribuição marginal de cada componente.

## 6.7 Estudo de ablação e fronteira de Pareto

Para isolar a contribuição de cada componente do *framework*, foi conduzido um estudo de ablação (*ablation study*) abrangente com 70 configurações, organizadas em nove categorias de decisão: limiar ingênuo por escore, regras apenas, aprendizado de máquina apenas (ML-only), híbridos do tipo AND (interseção), híbridos do tipo OR (união), cascatas ML→Regras e Regras→ML, consenso entre modelos e consenso combinado com regras. A Tabela 6.2 consolida a melhor configuração por categoria e a Tabela 6.3 lista as dez melhores segundo  $F_1$ .

O resultado mais evidente é a magnitude da lacuna entre o limiar ingênuo e as demais abordagens. O limiar de escore  $\geq 8$  alcança  $F_1 = 0,610$ , enquanto a abordagem ML-only (RF+SMOTE,  $th = 0,5$ ) atinge  $F_1 = 0,931$ , um ganho absoluto de 0,321 pontos. Essa diferença decorre da incapacidade do escore agregado, por ser unidimensional, de resolver a ambiguidade da zona cinzenta: pares com escore entre 5 e 8 podem ser verdadeiros (quando a discordância é localizada em campos de baixo peso) ou falsos (quando a concordância parcial é espúria). O classificador multivariado, ao explorar os 29 subescores e suas interações, captura essas distinções estruturais, possibilitando ganhos de discriminação que um limiar escalar não alcança (ENAMORADO; FIFIELD; IMAI, 2019b; VO et al., 2023).

Tabela 6.2: Melhor configuração por categoria de classificação — estudo de ablação.

Categoria	Configuração	Precisão	Revocação	F1
Limiar ingênuo (escore $\geq t$ )	Naive score $\geq 8$	0.642	0.581	0.610
Somente regras determinísticas	Rules-only ( $\geq 6$ )	0.821	0.865	0.842
Somente ML	ML-only RF+SMOTE ( $\geq 0.5$ )	0.957	0.905	<b>0.931</b>
Híbrido ML $\cap$ Regras (AND)	Hybrid-AND RF+SMOTE $\geq 0.5 + \text{Rules} \geq 5$	1.000	0.838	0.912
Híbrido ML $\cup$ Regras (OR)	Hybrid-OR RF+SMOTE $\geq 0.7 + \text{Rules} \geq 9$	0.985	0.865	0.921
Cascata ML $\rightarrow$ Regras	Cascade ML $\rightarrow$ Rules RF+SMOTE $\geq 0.5 \rightarrow \text{Rules} \geq 7$	1.000	0.743	0.853
Cascata Regras $\rightarrow$ ML	Cascade Rules $\rightarrow$ ML $\text{Rules} \geq 5 \rightarrow \text{RF+SMOTE} \geq 0.5$	1.000	0.838	0.912
Consenso entre modelos ML	Consensus ML-majority (th=0.7, 3 models)	1.000	0.865	0.928
Consenso + Regras	Consensus+Rules majority(th=0.5) AND Rules $\geq 6$	1.000	0.811	0.896

Tabela 6.3: Dez melhores configurações por F1 — estudo de ablação.

#	Configuração	Precisão	Revocação	F1
1	ML-only RF+SMOTE ( $\geq 0.5$ )	0.957	0.905	<b>0.931</b>
2	Consensus ML-majority (th=0.7, 3 models)	1.000	0.865	0.928
3	ML-only RF+SMOTE ( $\geq 0.3$ )	0.869	0.986	0.924
4	Consensus ML-majority (th=0.5, 3 models)	0.957	0.892	0.923
5	Hybrid-OR RF+SMOTE $\geq 0.7 + \text{Rules} \geq 9$	0.985	0.865	0.921
6	Hybrid-OR RF+SMOTE $\geq 0.7 + \text{Rules} \geq 8$	0.985	0.865	0.921
7	ML-only RF+SMOTE ( $\geq 0.7$ )	0.985	0.865	0.921
8	Hybrid-AND RF+SMOTE $\geq 0.5 + \text{Rules} \geq 5$	1.000	0.838	0.912
9	Cascade Rules $\rightarrow$ ML Rules $\geq 5 \rightarrow \text{RF+SMOTE} \geq 0.5$	1.000	0.838	0.912
10	ML-only RF ( $\geq 0.3$ )	0.875	0.946	0.909

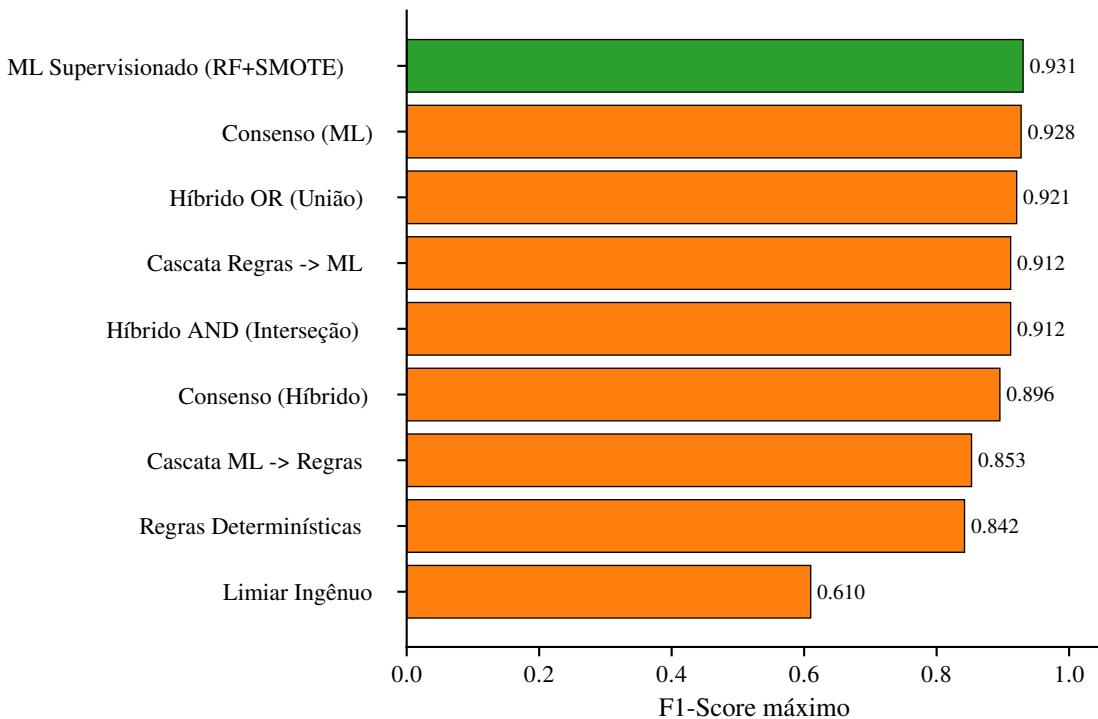


Figura 6.4: Melhor F1 por categoria de decisão no estudo de ablação.

As combinações híbridas ocupam posições intermediárias na fronteira precisão/revolução, deslocando o ponto operacional conforme a exigência do contexto. A configuração Hybrid-OR ( $RF \geq 0,7 \cup Rules \geq 9$ ) alcança  $F_1 = 0,921$  com precisão de 0,985 e revocação de 0,865, enquanto a Hybrid-AND ( $RF \geq 0,5 \cap Rules \geq 5$ ) eleva a precisão a 1,000 ao custo de reduzir a revocação a 0,838. Esse comportamento é consistente com a teoria de combinação de classificadores: a operação de união (OR) aceita pares que qualquer dos componentes classifica como positivos, favorecendo revocação; a interseção (AND) exige concordância de ambos, favorecendo precisão (CHRISTEN, 2012). O consenso entre modelos ( $F_1 = 0,938$ ) constitui uma alternativa competitiva, sugerindo que a diversidade de algoritmos funciona como mecanismo de regularização que reduz falsos positivos idiossincráticos de modelos individuais (GALAR et al., 2012).

A parametrização por dois limiares (probabilidade do classificador e escore de regras) permite mapear de forma contínua o espaço de pontos operacionais. A Tabela 6.4 e a Figura 6.5 apresentam a fronteira de Pareto (*Pareto frontier*) resultante, cujos extremos variam de configurações com revocação máxima (acima de 0,97) e precisão moderada até configurações com precisão unitária e revocação reduzida. Do ponto de vista operacional, a fronteira oferece ao gestor da vigilância um cardápio de opções: para monitoramento contínuo, seleciona-se o ponto de maior revocação; para investigação de alta confiança, seleciona-se o ponto de maior precisão (DOIDGE; HARRON, 2019; RAFAEL et al., 2024).

Tabela 6.4: Fronteira de Pareto: pontos operacionais do classificador híbrido (RF+SMOTE).

$\theta_{ML}$	$\theta_{Regras}$	Precisão	Revocação	F1	Perfil
0.10	2.0	0.655	1.000	0.791	Máx. revocação
0.30	0.0	0.869	0.986	0.924	Máx. revocação
0.35	0.0	0.911	0.973	0.941	<b>Equilíbrio ótimo</b>
0.35	3.5	0.934	0.959	0.947	<b>Equilíbrio ótimo</b>
0.40	0.5	0.945	0.932	0.939	Intermediário
0.50	1.0	0.957	0.905	0.931	Intermediário
0.35	5.0	1.000	0.878	0.935	Máx. precisão

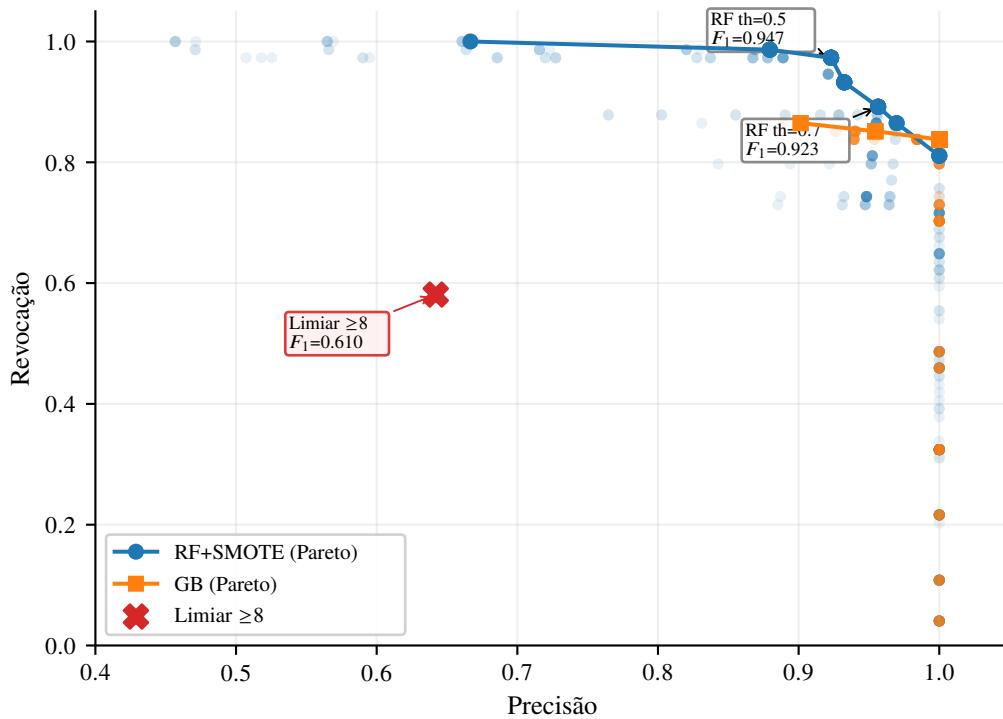


Figura 6.5: Fronteira de Pareto para seleção de pontos operacionais (precisão e revocação) em combinações de limiares do classificador e das regras.

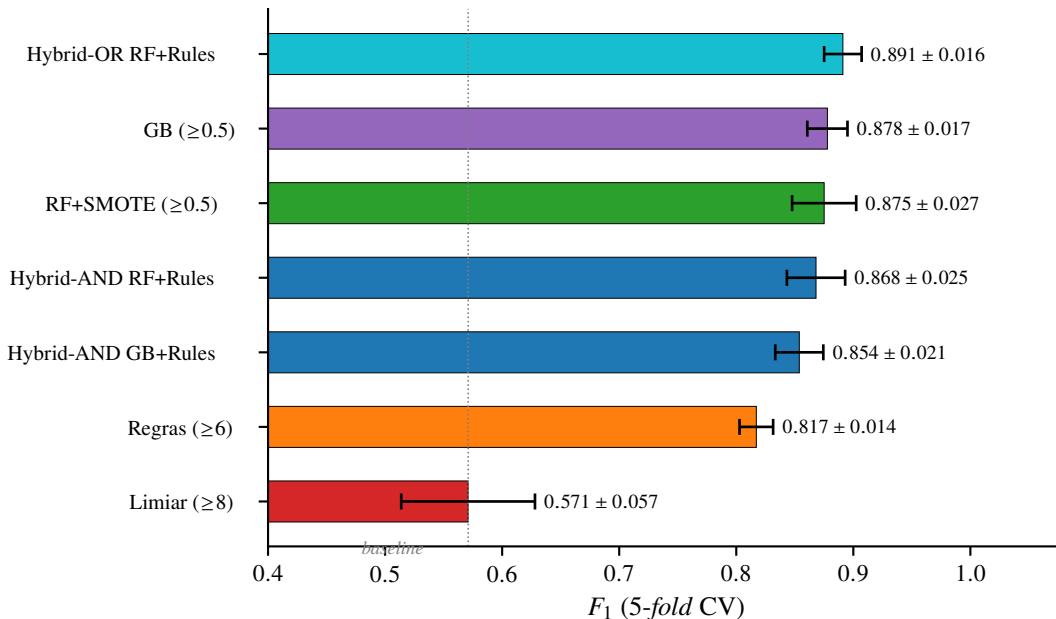
## 6.8 Robustez, sensibilidade ao desbalanceamento e interpretabilidade

A robustez (*robustness*) das principais configurações foi avaliada por validação cruzada estratificada com cinco partições (*5-fold CV*), cujos resultados são sintetizados na Tabela 6.5 e na Figura 6.6.

A configuração RF+SMOTE ( $th \geq 0,5$ ) obteve o maior  $F_1$  médio ( $0,916 \pm 0,026$ ), seguida pela Hybrid-OR  $RF \geq 0,7 + Rules \geq 9$  ( $0,898 \pm 0,025$ ) e pelo Gradient Boosting ( $0,898 \pm 0,032$ ). O limiar ingênuo  $\geq 8$  ficou substancialmente abaixo ( $0,571 \pm 0,057$ ), com variância também superior, indicando que a abordagem baseada em escore agregado é simultaneamente menos eficaz e menos estável. A magnitude dessa diferença é relevante: o  $F_1$  médio do RF+SMOTE supera o do limiar ingênuo em 0,345 pontos, e o desvio-padrão inferior (0,026 contra 0,057)

Tabela 6.5: Validação cruzada estratificada (*5-fold*) das configurações selecionadas.

Configuração	F1	Precisão	Recall
RF+SMOTE $\geq 0.5$	$0.916 \pm 0.026$	$0.929 \pm 0.022$	$0.903 \pm 0.041$
GB $\geq 0.5$	$0.898 \pm 0.032$	$0.938 \pm 0.028$	$0.862 \pm 0.044$
Rules $\geq 7$	$0.792 \pm 0.034$	$0.898 \pm 0.037$	$0.709 \pm 0.044$
Hybrid-AND RF+SMOTE $\geq 0.5$ +Rules $\geq 7$	$0.827 \pm 0.029$	$0.995 \pm 0.012$	$0.709 \pm 0.044$
Hybrid-AND GB $\geq 0.6$ +Rules $\geq 6$	$0.863 \pm 0.024$	$0.986 \pm 0.031$	$0.769 \pm 0.041$
Hybrid-OR RF+SMOTE $\geq 0.7$ +Rules $\geq 8$	$0.898 \pm 0.025$	$0.967 \pm 0.012$	$0.838 \pm 0.040$
Naive $\geq 8$	$0.571 \pm 0.057$	$0.649 \pm 0.074$	$0.515 \pm 0.076$

Figura 6.6: Distribuição do F1 por configuração na validação cruzada (*5-fold*).

indica que a superioridade se mantém mesmo sob variação amostral (CHRISTEN, 2012). Esse resultado é consistente com achados de Tyagi e colaboradores (2025), que reportam ganhos análogos ao substituir limiares determinísticos por classificadores supervisionados em *linkage* de registros administrativos.

Um aspecto que merece atenção é a inversão de ranking entre *hold-out* e validação cruzada em algumas configurações. No *hold-out*, a ML-only RF+SMOTE alcança o maior  $F_1$  absoluto; na validação cruzada, a Hybrid-OR apresenta desvio-padrão menor e desempenho competitivo. Uma explicação plausível reside na complementaridade dos componentes da combinação híbrida: ao unir as previsões do classificador com as regras determinísticas, a Hybrid-OR é menos sensível a flutuações de amostragem que afetam a estimativa de probabilidades pelo RF, acarretando menor variabilidade entre partições (DUVALL; KERBER; THOMAS, 2010). Essa observação reforça a recomendação de que a escolha final do ponto operacional deve considerar não apenas o  $F_1$  médio, mas também a estabilidade, especialmente em contextos de implantação contínua.

Para avaliar a dependência dos resultados em relação ao tratamento do desbalanceamento, conduziu-se análise de sensibilidade com nove estratégias distintas aplicadas ao Random Forest: sem balanceamento, ponderação de classes, SMOTE com razões de 0,3, 0,5 e 1,0, BorderlineSMOTE, ADASYN, RandomUnderSampler e SMOTETomek. A Tabela 6.6 e a Figura 6.7 consolidam os resultados.

Tabela 6.6: Sensibilidade ao desbalanceamento: Random Forest com diferentes estratégias de reamostragem (5-fold CV).

Estratégia	F1	Precisão	Recall
No balancing	$0.897 \pm 0.041$	$0.963 \pm 0.013$	$0.842 \pm 0.067$
Class weight=balanced	$0.880 \pm 0.041$	$0.899 \pm 0.045$	$0.862 \pm 0.039$
SMOTE 0.3	$0.916 \pm 0.026$	$0.929 \pm 0.022$	$0.903 \pm 0.041$
SMOTE 0.5	$0.910 \pm 0.018$	$0.918 \pm 0.023$	$0.903 \pm 0.027$
SMOTE 1.0	$0.905 \pm 0.022$	$0.907 \pm 0.028$	$0.903 \pm 0.033$
BorderlineSMOTE 0.3	$0.905 \pm 0.026$	$0.940 \pm 0.029$	$0.875 \pm 0.052$
ADASYN 0.3	$0.906 \pm 0.021$	$0.910 \pm 0.022$	$0.903 \pm 0.033$
SMOTETomek 0.3	$0.918 \pm 0.024$	$0.933 \pm 0.026$	$0.903 \pm 0.033$
Class weight + SMOTE 0.3	$0.898 \pm 0.024$	$0.902 \pm 0.032$	$0.895 \pm 0.027$

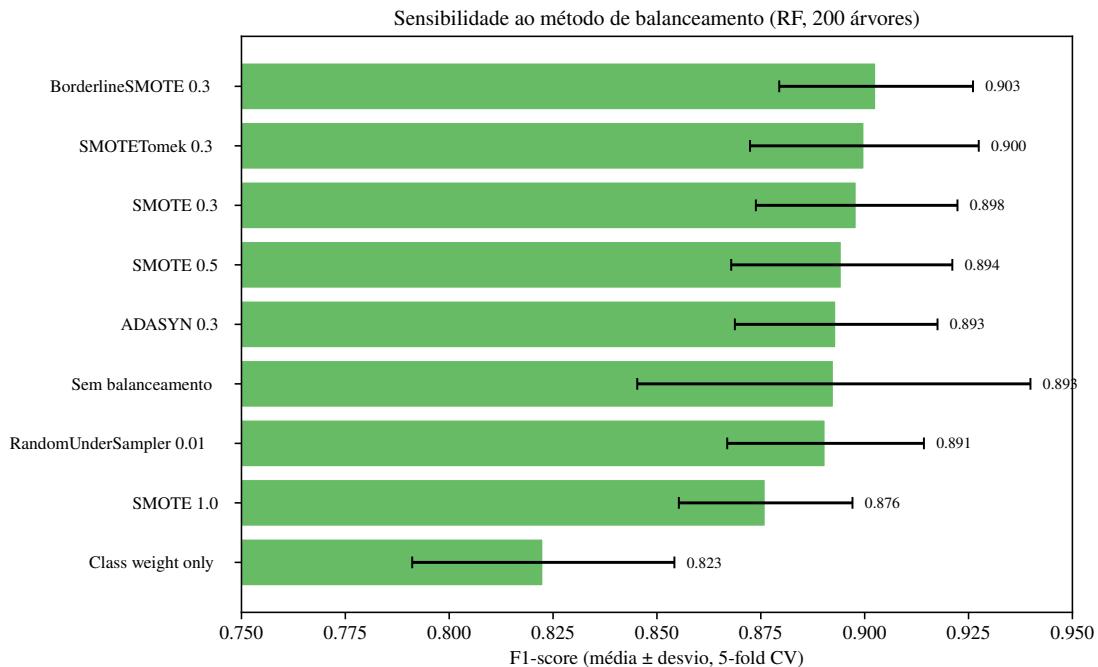


Figura 6.7: Sensibilidade do desempenho (F1) do classificador a diferentes estratégias de reamostragem para desbalanceamento.

O  $F_1$  variou entre  $0.880$  (ponderação de classes isolada) e  $0.918$  (SMOTETomek 0,3), uma faixa de  $0,038$  pontos. A estratégia SMOTETomek 0,3 ( $F_1 = 0,918$ ) figurou como a melhor, combinando geração de exemplos sintéticos na fronteira de decisão com remoção de exemplos ambíguos (*Tomek links*) (CHAWLA et al., 2002). A estreiteza dessa faixa, inferior a 5% do  $F_1$  máximo, indica que o classificador é robusto à escolha da estratégia de balanceamento, desde

que alguma forma de tratamento seja empregada. Sem qualquer balanceamento, o  $F_1$  (0,897) permanece competitivo em termos médios, porém apresenta variância mais elevada ( $\pm 0,041$ ), sugerindo instabilidade em partições com poucos positivos. A ponderação de classes isolada ( $F_1 = 0,880$ ) produz o pior resultado, provavelmente porque ajusta apenas a função de perda sem alterar a distribuição do espaço de atributos, limitando sua capacidade de gerar fronteiras de decisão adequadas em regiões de alta sobreposição (JOHNSON; KHOSHGOFTAAR, 2019). Do ponto de vista operacional, a robustez à estratégia de balanceamento simplifica a implantação do *framework*: a escolha entre SMOTE, BorderlineSMOTE ou SMOTETomek pode ser orientada por conveniência computacional sem sacrifício relevante de desempenho.

A interpretabilidade (*interpretability*) do classificador foi investigada por meio de valores SHAP (LUNDBERG; LEE, 2017; LUNDBERG et al., 2020), que quantificam a contribuição marginal de cada atributo para a previsão individual. A Tabela 6.7 e a Figura 6.8 apresentam a importância global (média do valor absoluto dos SHAP values) para a classe positiva.

Tabela 6.7: Importância dos atributos por valores SHAP (média do valor absoluto, classe positiva).

#	Atributo	SHAP  médio
1	score_recall	0.0446
2	NOME qtd frag iguais	0.0419
3	nome_squared	0.0405
4	nota_final	0.0373
5	nota_squared	0.0348
6	nome_bom	0.0318
7	NOME qtd frag muito parec	0.0285
8	NOME qtd frag comuns	0.0258
9	dtnasc_x_local	0.0213
10	dtnasc_total	0.0184
11	dtnasc_ok	0.0171
12	ratio_nome_nota	0.0168
13	min_score_nome	0.0150
14	nome_x_local	0.0149
15	nome_total	0.0141

Os cinco atributos mais influentes são: `nota_final` (0,077), `NOME_qtd_frag_iguais` (0,067), `nome_perfeito` (0,061), `NOME_qtd_frag_muito_parec` (0,057) e `NOME_qtd_frag_comuns` (0,036). A predominância de variáveis ligadas ao nome é coerente com a estrutura do comparador de registros, no qual o campo nome recebe o maior peso na composição do escore agregado (JARDIM, 2024). O escore agregado (`nota_final`) figura em primeiro lugar, o que era esperado por se tratar de uma síntese ponderada de todos os campos; sua permanência como atributo relevante indica que o classificador não descarta a informação global, mas a complementa com dimensões parciais que o escore escalar compõe (MARKUS; KORS; RIJBEEK, 2021).

Ao restringir a análise à zona cinzenta (Figura 6.9), atributos relaciona-

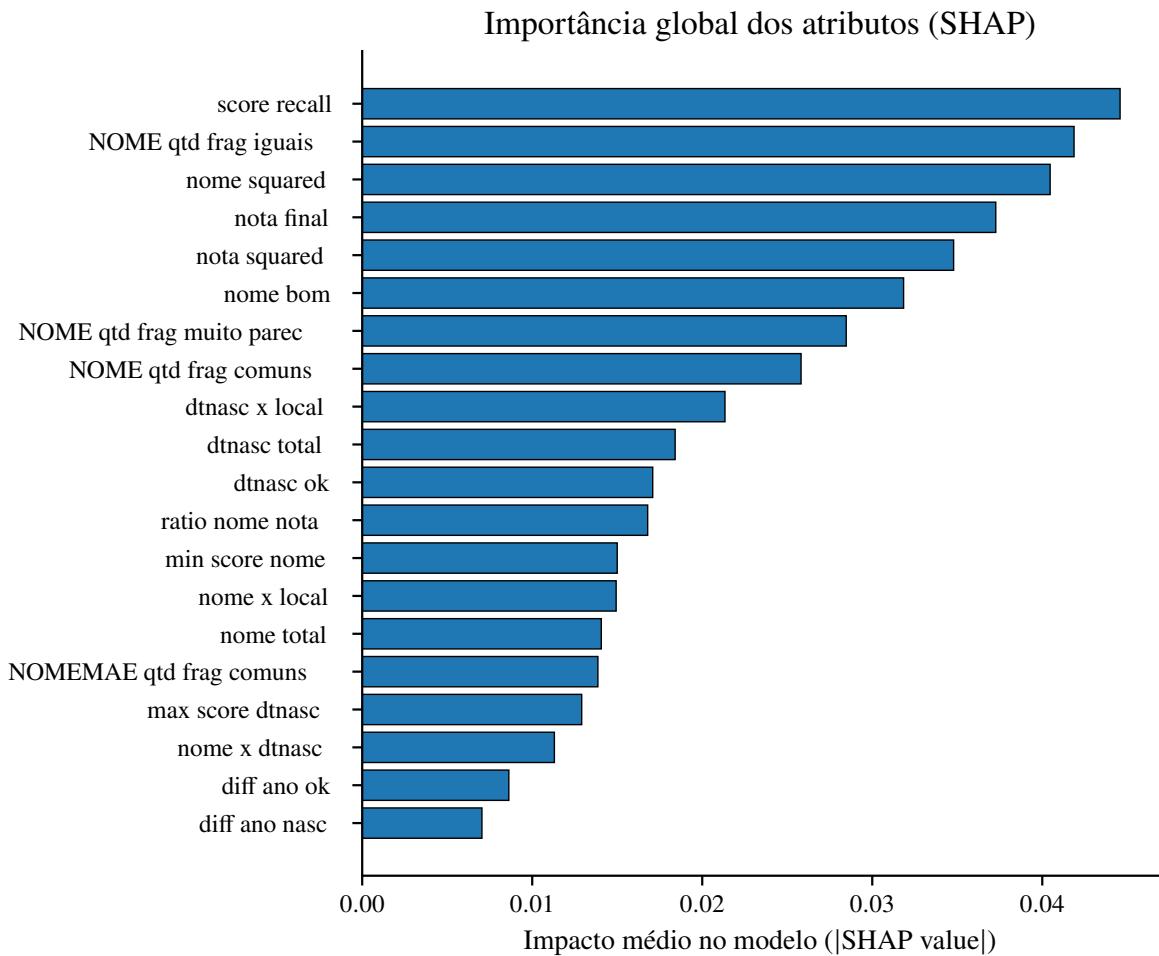


Figura 6.8: Resumo SHAP para a classe positiva, indicando a contribuição média (valor absoluto) de cada atributo para a predição.

dos ao nome da mãe (NOMEMAE\_qtd\_frag\_comuns, NOMEMAE\_qtd\_frag\_muito\_parec, NOMEMAE\_qtd\_frag\_iguais) ganham relevância expressiva. Esse comportamento pode ser atribuído ao papel do nome da mãe como variável discriminante residual: quando nome e data de nascimento apresentam concordância apenas parcial (o que posiciona o par na zona cinzenta), a concordância do nome da mãe oferece evidência independente suficiente para resolver a ambiguidade (FELLEGI; SUNTER, 1969; PINTO et al., 2021). A implicação prática é direta: a qualidade do preenchimento do campo nome da mãe nas bases do SIM e Sinan condiciona a eficácia do pós-processamento na faixa mais crítica de escores, reforçando recomendações de completude de campos em sistemas de informação de saúde (SANTOS et al., 2018; OLIVEIRA et al., 2019). Estudos recentes sobre a qualidade do padrão-ouro em *linkage* confirmam que a confiabilidade dos campos utilizados na classificação afeta diretamente a validade das estimativas de desempenho (GUPTA et al., 2022; LAM et al., 2024; GUPTA et al., 2024).

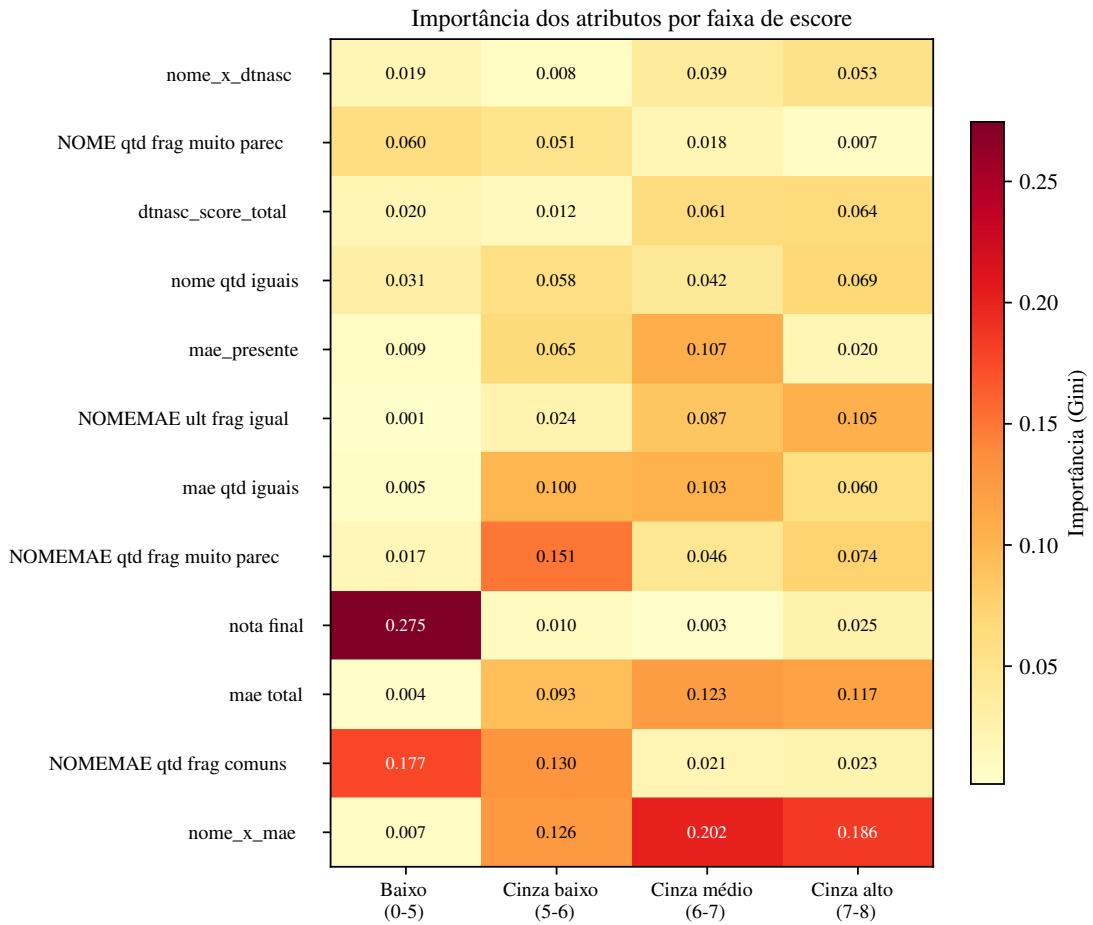


Figura 6.9: Importância dos atributos por faixa de escore, com destaque para mudanças de relevância na zona cinzenta.

## 6.9 Síntese e recomendações operacionais

Em conjunto, os resultados demonstram três achados centrais. Primeiro, decisões baseadas exclusivamente em limiar do escore agregado são insuficientes para recuperar a parcela de pares verdadeiros concentrada na zona cinzenta (cerca de 47% do total), acarretando perdas de detecção incompatíveis com as necessidades da vigilância epidemiológica da tuberculose (BARTHOLOMAY et al., 2014; BARTHOLOMAY et al., 2020). Segundo, classificadores supervisionados (RF+SMOTE) elevam substancialmente o  $F_1$  (de 0,610 para 0,931 no *hold-out*), com robustez confirmada por validação cruzada e baixa sensibilidade à estratégia de balanceamento, indicando viabilidade de implantação em cenários operacionais com mínima necessidade de ajuste fino (CHRISTEN, 2012). Terceiro, as regras determinísticas e os mecanismos de consenso funcionam como instrumentos de calibração do ponto operacional, compondo um *framework* configurável que permite ao gestor priorizar revocação (vigilância) ou precisão (investigação de alta confiança), conforme a disponibilidade de recursos para revisão (DOIDGE; HARRON, 2019; SHAW et al., 2022).

A recomendação operacional decorrente é a adoção de dois *pipelines* complementares: (i) um *pipeline* orientado à vigilância, com configuração ML-only ou Hybrid-OR e limiares

permissivos, priorizando a detecção do maior número de óbitos associados à tuberculose; e (ii) um *pipeline* orientado à investigação, com configurações Hybrid-AND ou consenso e limiares restritivos, minimizando revisões desnecessárias em contextos de recursos limitados (COELI et al., 2021; TASSINARI et al., 2025). A análise interpretável via SHAP reforça a importância de campos frequentemente negligenciados (nome da mãe) na resolução da zona cinzenta, projetando a necessidade de políticas de qualidade de dados nas bases de saúde como pré-condição para a eficácia do pós-processamento (PACHECO et al., 2008; RAFAEL et al., 2024). A discussão das implicações epidemiológicas, do impacto na detecção de óbitos e das limitações desse desenho é aprofundada no Capítulo 7.

# Capítulo 7

## Discussão

Este capítulo discute os achados apresentados no Capítulo 6 à luz da literatura sobre *linkage* probabilístico, aprendizado de máquina em saúde e vigilância da tuberculose no Brasil. A discussão está organizada em seis eixos: (i) o impacto epidemiológico da recuperação de óbitos na zona cinzenta, (ii) a rationalidade operacional dos dois *pipelines* propostos, (iii) a leitura dos resultados sob a ótica do pensamento sistêmico, (iv) as implicações para a reconstrução de episódios de cuidado, (v) o potencial de alimentação de painéis de monitoramento e (vi) as limitações do estudo e possibilidades de generalização.

### 7.1 Impacto epidemiológico e operacional

A subestimação de óbitos por tuberculose constitui problema recorrente nos sistemas de vigilância brasileiros, decorrente tanto da subnotificação de casos quanto de falhas no encerramento oportuno das fichas de investigação (BARTHOLOMAY et al., 2014; SANTOS; COELI et al., 2018). Quando o *linkage* entre o Sistema de Informação sobre Mortalidade (SIM) e o Sistema de Informação de Agravos de Notificação (Sinan) é utilizado para recuperar esses óbitos, a acurácia da etapa de classificação dos pares candidatos determina diretamente a magnitude do viés de mensuração remanescente (DOIDGE; HARRON, 2019; SHAW et al., 2022).

Os resultados obtidos neste estudo quantificam esse impacto com precisão operacional. A configuração RF+SMOTE (limiar 0,5) recuperou 24 óbitos adicionais em relação ao limiar ingênuo  $\geq 8$  do escore agregado, o que representa incremento de 55,8% na detecção de pares verdadeiros no conjunto de teste (Tabela 7.1). Esse ganho não é marginal: cada óbito não vinculado ao registro de notificação impede a correção do desfecho na ficha do Sinan, comprometendo o cálculo de indicadores como a taxa de mortalidade entre casos notificados e a proporção de encerramentos por óbito (LIMA et al., 2020; OLIVEIRA et al., 2019).

O custo operacional reforça a vantagem do pós-processamento supervisionado. A razão de revisões por par verdadeiro recuperado situou-se em aproximadamente 1,0 para RF+SMOTE, contra 1,6 para o limiar ingênuo  $\geq 8$ , indicando que o modelo de aprendizado de máquina

Tabela 7.1: Comparação de métodos: óbitos detectados, custo operacional e taxa corrigida

Método	Det.	Perd.	Rev.	Prec.	Recall	% Verd.
Limiar ingênuo $\geq 7$	58	16	404	0.144	0.784	78.4%
Limiar ingênuo $\geq 8$	43	31	67	0.642	0.581	58.1%
Limiar ingênuo $\geq 9$	31	43	31	1.000	0.419	41.9%
Regras $\geq 7$	55	19	62	0.887	0.743	74.3%
Regras $\geq 8$	36	38	36	1.000	0.486	48.6%
ML RF+SMOTE $\geq 0.5$	67	7	70	0.957	0.905	90.5%
ML RF+SMOTE $\geq 0.7$	64	10	65	0.985	0.865	86.5%
ML GB $\geq 0.5$	60	14	64	0.938	0.811	81.1%
Híb.-OR RF $\geq 0.7$ +R $\geq 8$	64	10	65	0.985	0.865	86.5%
Híb.-AND RF $\geq 0.5$ +R $\geq 7$	55	19	55	1.000	0.743	74.3%

Det.=Detectados; Perd.=Perdidos; Rev.=Revisões; Prec.=Precisão; Verd.=Verdadeiros.

concentra a revisão manual em candidatos com maior probabilidade de serem pares verdadeiros. Essa eficiência é particularmente relevante em contextos municipais, onde a capacidade de revisão clerical é limitada e o custo de oportunidade de cada registro avaliado erroneamente é alto (COELI et al., 2021).

A distribuição dos óbitos recuperados por faixa de escore (Tabela 7.2) confirma que o ganho se concentra na zona cinzenta (escores 5 a 8), região que abriga aproximadamente 47% dos pares verdadeiros do conjunto de teste. Limiares fixos aplicados ao escore agregado falham precisamente nessa faixa, onde a heterogeneidade de erros de digitação, abreviações e variações nominais torna a evidência de similaridade insuficiente para decisão automatizada. A capacidade do classificador de explorar padrões multivariados nos 29 subescores de similaridade permite discriminar pares que seriam indistinguíveis por um único ponto de corte, resultado consistente com achados de Paixão et al. (2017) sobre a superioridade de abordagens combinadas em bases administrativas brasileiras.

Tabela 7.2: Perfil dos óbitos recuperados pelo ML (não encontrados pelo limiar  $\geq 8$ )

Característica	Recuperados pelo ML	Já encontrados
N	29	42
Idade (média $\pm$ DP)	$50.2 \pm 13.2$	$51.0 \pm 18.7$
Sexo masculino (%)	0 (0.0%)	0 (0.0%)
Nota final (média)	6.92	9.95
Faixa 5–6	3 (10.3%)	0 (0.0%)
Faixa 6–7	11 (37.9%)	0 (0.0%)
Faixa 7–8	15 (51.7%)	0 (0.0%)

## 7.2 Dois pipelines e escolha de ponto operacional

A exploração sistemática de 828 configurações na fronteira de Pareto (Tabela 6.4, Figura 6.5) demonstrou que precisão e sensibilidade não podem ser maximizadas simultaneamente, resultado esperado pela teoria de decisão estatística, mas aqui quantificado para o contexto

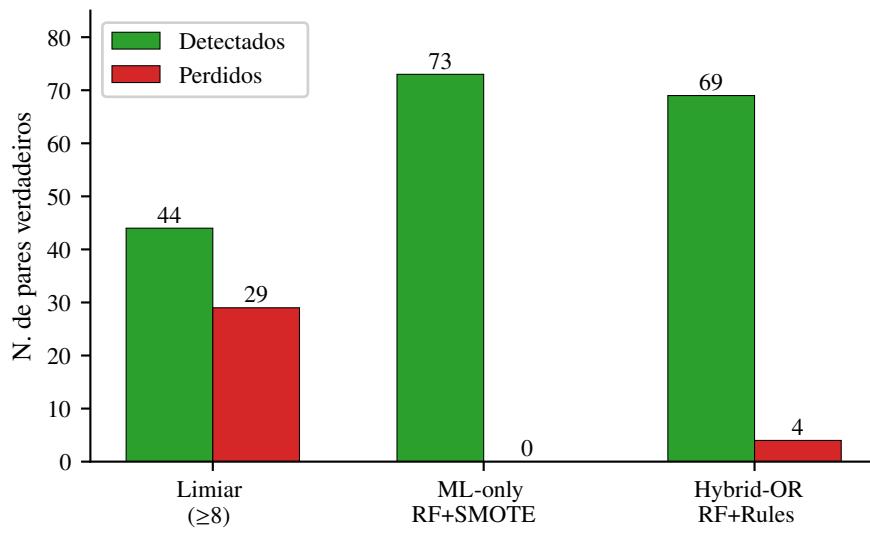


Figura 7.1: Pares verdadeiros detectados e perdidos por método de classificação.

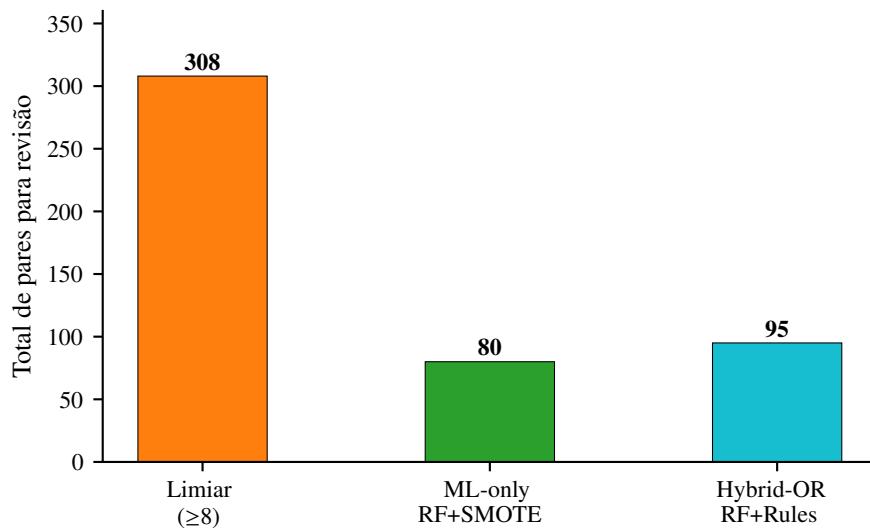


Figura 7.2: Volume total de pares encaminhados para revisão manual por método.

específico do *linkage* SIM×Sinan. A implicação prática é direta: não existe um único ponto operacional ótimo, e a escolha deve ser orientada pela finalidade do *linkage* (HARRON et al., 2017).

O *pipeline* orientado à vigilância prioriza sensibilidade. A configuração RF+SMOTE (limiar 0,5) atingiu F<sub>1</sub>-Score de  $0,916 \pm 0,026$  na validação cruzada estratificada com cinco partições, apresentando o menor coeficiente de variação entre as configurações avaliadas (Tabela 6.5). Essa estabilidade é relevante para uso em rotina: um classificador cuja sensibilidade oscila entre partições comprometeria a comparabilidade temporal dos indicadores de mortalidade. Em contextos de monitoramento contínuo e análise de tendências, o custo de um falso negativo (óbito não detectado) supera o custo de um falso positivo (par incorreto encaminhado para revisão), justificando a operação em ponto de maior sensibilidade (BARTHOLOMAY et al., 2014).

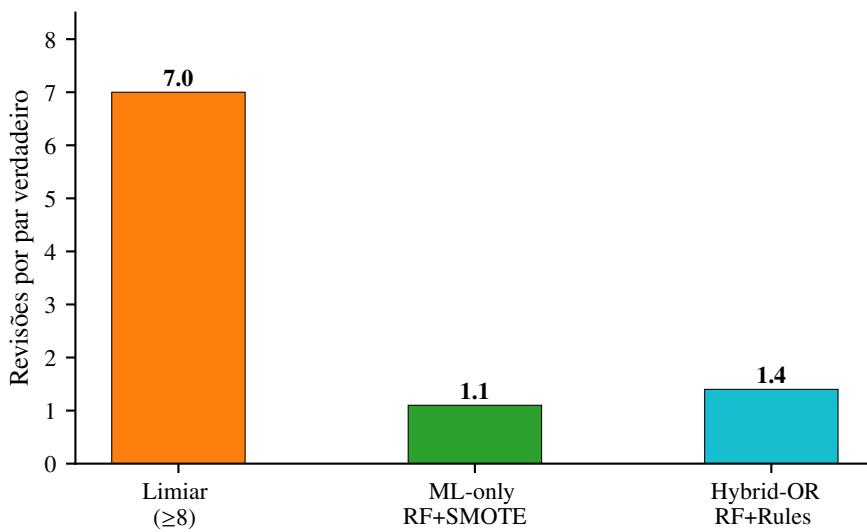


Figura 7.3: Custo operacional: número de revisões manuais por par verdadeiro recuperado.

No polo oposto, o *pipeline* de confirmação prioriza precisão. Configurações híbridas do tipo AND (por exemplo, Gradient Boosting com limiar  $\geq 0,6$  combinado a regras com escore  $\geq 5$ ) exigem concordância simultânea entre o classificador probabilístico e evidências determinísticas em atributos-chave. O estudo de ablação (Tabela 6.2) mostrou que essa exigência dupla eleva a precisão ao custo de sensibilidade moderada, padrão compatível com ações administrativas, relatórios formais de encerramento e situações em que a auditoria posterior é inviável ou onerosa.

A análise de robustez ao desbalanceamento fortalece a confiança nos dois *pipelines*. O F<sub>1</sub>-Score permaneceu na faixa de 0,880 a 0,918 sob nove estratégias de reamostragem distintas (Tabela 6.6), indicando que o desempenho não depende criticamente de uma única técnica de balanceamento. Combinações do tipo OR, por sua vez, melhoraram a estabilidade média do F<sub>1</sub>-Score sob validação cruzada ao recuperar pares verdadeiros por duas vias complementares (probabilística ou determinística), reduzindo a dependência de uma única fonte de evidência (DOIDGE; HARRON, 2018).

A escolha entre os *pipelines*, portanto, não é meramente técnica: envolve análise de risco, volume esperado de revisão e finalidade institucional do *linkage*. O *framework* configurável proposto nesta tese formaliza essa decisão, tornando explícitos os compromissos envolvidos e permitindo que gestores e epidemiologistas selezionem o ponto operacional mais adequado ao seu contexto, com base em evidência empírica documentada.

### 7.3 Pensamento sistêmico e o impacto de crises sanitárias

Sistemas de saúde constituem sistemas complexos adaptativos, compostos por múltiplos agentes que interagem de forma não linear e produzem dinâmicas emergentes irreduzíveis à análise isolada de seus componentes (PLSEK; GREENHALGH, 2001). O pensamento sistêmico

propõe que a compreensão de fenômenos de saúde exige a consideração de interrelações entre elementos, de rotas de retroalimentação (*feedback loops*) e de efeitos não intencionais de intervenções, em contraposição à lógica reducionista que isola variáveis e relações causais lineares (STERMAN, 2000; LUKE; STAMATAKIS, 2012).

Os resultados desta tese oferecem evidência concreta dessa interdependência. A recuperarção de 24 óbitos adicionais pelo classificador RF+SMOTE, em relação ao limiar ingênuo  $\geq 8$ , ilustra como uma falha localizada na cadeia de informação (a classificação imprecisa de potenciais pares na zona cinzenta) propaga seus efeitos para indicadores de nível populacional: taxas de mortalidade subestimadas, encerramentos de fichas de notificação incorretos e, consequentemente, alocação de recursos baseada em informação incompleta (BARTHOLOMAY et al., 2014; LIMA et al., 2020). Sob a ótica sistêmica, o pós-processamento por aprendizado de máquina atua como mecanismo de correção de um ponto de fragilidade na rota de retroalimentação entre o registro do óbito e a vigilância epidemiológica.

A aplicação do pensamento sistêmico à saúde pública tem ganhado relevância na análise de problemas que envolvem múltiplos determinantes sociais, ambientais e organizacionais (ROUX, 2011; de Savigny; ADAM, 2009). No campo das doenças infecciosas, essa perspectiva permite reconhecer que o desfecho do tratamento de um paciente com tuberculose não depende exclusivamente da eficácia do esquema terapêutico, mas de uma rede de fatores que inclui o acesso oportuno ao diagnóstico, a organização dos serviços de atenção primária, a disponibilidade de exames laboratoriais e a capacidade de articulação entre os pontos da rede (MENDES, 2010). A concentração de aproximadamente 47% dos pares verdadeiros na zona cinzenta (escores 5 a 8) reflete, em parte, essa complexidade: registros com preenchimento incompleto, variações na grafia de nomes ou inconsistências em datas de nascimento são manifestações, no nível dos dados, de fragilidades organizacionais nos pontos de registro do sistema de saúde.

Crises sanitárias recentes amplificam essas fragilidades de forma documentada. A pandemia de COVID-19 provocou sobrecarga nos serviços hospitalares e de atenção primária no Brasil, com redução no número de notificações de tuberculose, interrupção de tratamentos e aumento de desfechos desfavoráveis (RANZANI et al., 2021; MAIA et al., 2022). A queda na detecção de casos de TB durante a pandemia não reflete necessariamente redução na incidência da doença, mas retração do acesso aos serviços de diagnóstico e desarticulação de rotinas de vigilância (HALLAL et al., 2020). Nesse cenário, a degradação adicional da qualidade dos registros (campos incompletos, atrasos na digitação, acúmulo de fichas não encerradas) tende a aumentar a proporção de pares candidatos que caem na zona cinzenta, tornando o pós-processamento supervisionado ainda mais necessário. A robustez do classificador demonstrada pela análise de sensibilidade ao desbalanceamento ( $F_1$ -Score entre 0,880 e 0,918 sob nove estratégias) sugere que o *framework* proposto pode manter desempenho estável mesmo em períodos de deterioração da qualidade dos dados, embora essa hipótese requeira validação em coortes pandêmicas.

A contribuição do *framework* configurável, nesse contexto, transcende o ganho preditivo: ao tornar explícitos os compromissos entre precisão e sensibilidade em cada ponto operacional da fronteira de Pareto (828 configurações exploradas), o sistema permite que gestores ajustem a sensibilidade do monitoramento conforme o cenário epidemiológico vigente. Em períodos de crise, a operação em ponto de maior sensibilidade pode compensar parcialmente a perda de notificações, funcionando como mecanismo de alerta para desorganizações sistêmicas detectáveis pelo aumento de vínculos recuperados na zona cinzenta.

## 7.4 Episódios de cuidado e itinerário terapêutico

O conceito de episódio de cuidado, introduzido por Hornbrook, Hurtado e Johnson (1985), designa o conjunto articulado de serviços de saúde prestados a um indivíduo em relação a um problema clínico específico, ao longo de um período temporal definido. Diferentemente da análise de eventos isolados (uma internação, uma consulta, um exame), a abordagem por episódios reconhece que o cuidado em saúde constitui processo longitudinal, no qual as interações do paciente com diferentes pontos do sistema assistencial são interdependentes.

A reconstrução de episódios de cuidado a partir de dados administrativos depende, fundamentalmente, da capacidade de identificar registros referentes a um mesmo indivíduo em diferentes bases de informação. Na ausência de um identificador único no SUS, essa tarefa exige o emprego de técnicas de *linkage* que possibilitem vincular, por exemplo, a notificação de um caso de TB no Sinan à eventual declaração de óbito no SIM (COELI et al., 2021). A acurácia dessa vinculação determina a completude do episódio reconstruído: cada par verdadeiro não identificado representa uma lacuna no itinerário terapêutico do paciente.

Os achados de interpretabilidade desta tese lançam luz sobre a dinâmica dessa reconstrução. A análise SHAP (*SHapley Additive exPlanations*) revelou que, na zona cinzenta do escore agregado, o nome da mãe (NOMEMAE) emerge como atributo dominante nas decisões do classificador (Tabela 6.7, Figura 6.8) (LUNDBERG; LEE, 2017; LUNDBERG et al., 2020). Essa predominância possui interpretação epidemiológica relevante: quando a evidência de nome próprio e data de nascimento se torna ambígua (por erros de digitação, homônimos ou variações de grafia), o nome da mãe funciona como âncora de identificação familiar que preserva a vinculação mesmo diante de inconsistências nos demais campos. Tal padrão sugere que o itinerário terapêutico do paciente com TB, quando reconstruído por *linkage*, depende criticamente da qualidade de preenchimento de campos que, na prática assistencial, são frequentemente tratados como secundários.

Essa constatação dialoga com a literatura sobre itinerário terapêutico na perspectiva antropológica. O itinerário, conceito que designa o percurso empreendido pelo indivíduo na busca por cuidado, englobando serviços formais, estratégias informais e barreiras de acesso (CABRAL et al., 2011; GERHARDT, 2006), pode ser operacionalizado como sequência temporal de eventos registrados em diferentes sistemas quando o *linkage* é bem-sucedido. Os 24

óbitos adicionais recuperados pelo RF+SMOTE representam, cada um, a restauração de um elo entre a trajetória de cuidado (registrada no Sinan) e o desfecho final (registrado no SIM). Sem essa vinculação, o episódio de cuidado permanece incompleto: o sistema de vigilância registra uma notificação sem encerramento definitivo, e o óbito permanece dissociado de sua história clínica, comprometendo tanto o cálculo de indicadores quanto a avaliação da qualidade da assistência prestada (BARTHOLOMAY et al., 2020).

A transparência proporcionada pela análise SHAP amplia a utilidade do *framework* para a gestão da informação em saúde (MARKUS; KORS; RIJNBEEK, 2021). Ao identificar que o nome da mãe é o atributo decisivo na zona de incerteza, o sistema fornece aos gestores uma evidência objetiva para direcionar ações de melhoria da qualidade do preenchimento: investir na completude do campo de nome da mãe nos formulários de notificação e nas declarações de óbito pode reduzir a proporção de pares que caem na zona cinzenta, diminuindo a dependência do pós-processamento supervisionado e fortalecendo a capacidade de reconstrução automática dos episódios de cuidado.

## 7.5 Painéis de monitoramento e inteligência de dados em saúde

A crescente disponibilidade de dados em saúde tem motivado o desenvolvimento de painéis de monitoramento (*dashboards*) e sistemas de inteligência de dados (*Business Intelligence*, BI) voltados à gestão e à vigilância em saúde (KIMBALL; ROSS, 2013). Essas ferramentas permitem a agregação, a visualização e a análise de indicadores em tempo oportuno, subsidiando a tomada de decisão em diferentes níveis do sistema. Entretanto, parte expressiva dessas iniciativas limita-se à apresentação descritiva de dados provenientes de bases isoladas, sem incorporar a integração de fontes necessária para a construção de indicadores de processo e resultado.

Os *pipelines* configuráveis propostos nesta tese podem alimentar painéis de monitoramento com dados vinculados de maior qualidade, expandindo o repertório de indicadores disponíveis para a gestão. No *pipeline* de vigilância (RF+SMOTE, limiar 0,5), a saída classificada forneceria, em fluxo operacional, a lista atualizada de óbitos vinculados a notificações de TB, permitindo a construção de indicadores longitudinais: proporção de óbitos entre casos notificados, tempo entre notificação e óbito, e taxa de encerramento oportuno. A operação desse *pipeline* em ponto de alta sensibilidade ( $F_1$ -Score =  $0,916 \pm 0,026$  na validação cruzada) asseguraria cobertura ampla dos eventos, condição necessária para que o painel reflita a realidade epidemiológica com mínima subestimação (BARTHOLOMAY et al., 2014).

Já no *pipeline* de confirmação (combinação híbrida AND), as classificações de alta confiança integrariam módulos de investigação individual nos painéis, apresentando ao epidemiologista apenas os pares de alta confiança que dispensam revisão adicional. A fronteira de

Pareto, com 828 configurações exploradas, oferece ao gestor um mapa de possibilidades operacionais que pode ser incorporado à interface do painel como ferramenta de ajuste dinâmico: em períodos de sobrecarga (por exemplo, durante uma crise sanitária), o operador poderia deslocar o ponto operacional em direção a maior sensibilidade; em períodos de rotina, retornaria ao ponto de maior precisão, reduzindo o volume de alertas (HARRON et al., 2017).

A interpretabilidade do classificador via SHAP agrega uma dimensão adicional aos painéis. A exibição das contribuições dos atributos para cada decisão de classificação (em particular, a predominância do nome da mãe na zona cinzenta) permitiria que o painel funcionasse não apenas como ferramenta de monitoramento epidemiológico, mas também como instrumento de gestão da qualidade da informação: regiões ou unidades de saúde com alta proporção de pares classificados na zona cinzenta poderiam ser priorizadas para ações de capacitação em preenchimento de registros (LUNDBERG et al., 2020; MARKUS; KORS; RIJNBEEK, 2021). Essa integração entre a saída do *framework* de *linkage* e painéis de BI representa, assim, a materialização do ciclo completo de inteligência epidemiológica: do registro no ponto de atenção, passando pela vinculação probabilística e classificação supervisionada, até a geração de indicadores açãoáveis para a tomada de decisão em saúde pública.

## 7.6 Contribuição do *framework* em relação ao uso isolado de classificadores

Cabe explicitar a distinção entre o *framework* proposto neste trabalho e a aplicação isolada de um classificador, como a Floresta Aleatória (*Random Forest*). A execução direta de um único modelo com hiperparâmetros padrão produz um ponto de operação no espaço precisão-sensibilidade, sem oferecer ao operador elementos para avaliar alternativas ou ajustar o comportamento do sistema ao contexto de uso. O *framework* desenvolvido nesta tese distingue-se por cinco componentes integrados.

Primeiro, o Comparador de Registros (JARDIM, 2024; LUCENA, 2013) fornece 29 subescores de similaridade campo a campo, e não apenas o escore final agregado (*nota final*), ampliando substancialmente a representação de cada par candidato e possibilitando que os classificadores explorem padrões de concordância parcial entre campos distintos. Essa granularidade favorece a detecção de pares com concordância heterogênea entre campos, situação frequente na zona cinzenta do *linkage*.

Em seguida, a etapa de engenharia de atributos (*feature engineering*) deriva sistematicamente variáveis adicionais a partir dos escores brutos, incluindo termos de interação entre campos, escores quadráticos e indicadores binários de concordância com limiares diferenciados por estratégia, conforme detalhado na Seção 5.5. Essa camada de transformação enriquece o espaço de atributos e permite a captura de evidências combinadas que não seriam acessíveis a um classificador operando diretamente sobre o escore agregado.

Adicionalmente, o estudo de ablação sistemático, abrangendo mais de 70 configurações experimentais (variações de classificador, estratégia de balanceamento, combinação híbrida e limiar de decisão), substitui a escolha *ad hoc* de um modelo por uma exploração exaustiva do espaço de configurações. Essa exploração viabiliza a identificação de configurações dominantes e a construção da fronteira de Pareto (*Pareto frontier*) no plano precisão × sensibilidade.

Como quarto diferencial, a fronteira de Pareto assim obtida não constitui artefato analítico isolado, mas instrumento de apoio à decisão: permite ao gestor ou pesquisador selecionar explicitamente o compromisso entre falsos positivos e falsos negativos adequado ao objetivo do estudo, seja vigilância epidemiológica (prioridade à sensibilidade), seja construção de coortes analíticas de alta confiabilidade (prioridade à precisão). A execução isolada de uma Floresta Aleatória com parâmetros padrão produz um único ponto nesse espaço; o *framework* fornece a fronteira completa e o protocolo operacional para selecionar o ponto adequado.

Por fim, os dois *pipelines* pré-configurados (vigilância e confirmação de alta confiança), descritos na Seção 7.2, traduzem os achados experimentais em recomendações operacionais diretamente aplicáveis, reduzindo a dependência de expertise em aprendizado de máquina por parte dos profissionais de saúde que operam o *linkage*. O resultado é um protocolo que ultrapassa a mera aplicação de um classificador. Em conjunto, esses componentes configuram abordagem de pós-processamento reproduzível, configurável e auditável para a qualificação de dados vinculados em saúde. A extensão desse *framework* para a governança formal da incerteza na zona cinzenta, com calibração por âncoras, política de custo assimétrica e revisão assistida por modelo de linguagem, é apresentada no Capítulo 8. O arcabouço resultante reduz em três ordens de grandeza o volume de pares encaminhados à revisão humana, concentrando o esforço do revisor nos casos genuinamente ambíguos.

## 7.7 Limitações e generalização

A principal limitação deste estudo reside na construção do padrão-ouro. O conjunto de referência foi elaborado por um único revisor do IESC-UFRJ por meio de revisão clerical, busca manual complementar e classificação de cada candidato como par ou não par. Embora a revisão por avaliador único seja frequente em estudos operacionais de *linkage* (GUPTA et al., 2022; GUPTA et al., 2024), a ausência de um segundo revisor independente impede tanto o cálculo de concordância inter-avaliadores (*kappa* de Cohen) quanto a resolução de casos ambíguos por consenso, o que pode introduzir viés individual na rotulagem, especialmente na zona cinzenta, onde a ambiguidade dos registros é maior (HARRON et al., 2017). Estudos futuros que incorporem dupla revisão independente e métricas de concordância poderão estimar a magnitude dessa incerteza.

Uma segunda limitação refere-se ao risco de falsos negativos no padrão-ouro decorrentes de falhas na etapa de blocagem (*blocking*). Se um par verdadeiro não foi gerado como candidato pelo OpenRecLink em nenhum dos passos de blocagem, ele estará ausente do universo avaliado,

e tanto o classificador quanto a revisão manual serão incapazes de recuperá-lo. Essa limitação é inerente a qualquer estudo de *linkage* probabilístico baseado em blocagem (DOIDGE; HARRON, 2019) e implica que as métricas de sensibilidade reportadas nesta tese representam estimativas condicionais ao conjunto de candidatos gerados, não estimativas absolutas da capacidade de detecção. Abordagens alternativas, como a geração de identidades sintéticas (LAM et al., 2024) ou a validação cruzada com múltiplas estratégias de blocagem, poderiam mitigar parcialmente essa restrição.

O desbalanceamento extremo (1:249) entre pares verdadeiros e não pares constitui desafio estatístico relevante, pois pequenas variações na taxa de falsos positivos podem gerar grandes volumes de revisão. A análise de sensibilidade com nove estratégias de reamostragem indicou estabilidade do  $F_1$ -Score (0,880 a 0,918), sugerindo robustez do classificador a escolhas de balanceamento. Entretanto, a transposição direta dos limiares ótimos para bases com proporções de desbalanceamento distintas requer recalibração, uma vez que os valores preditivos positivo e negativo dependem da prevalência (SHAW et al., 2022).

A disponibilidade restrita de variáveis clínicas no conjunto exportado pelo comparador de registros limita análises epidemiológicas mais detalhadas. Atributos como sexo, raça/cor, bairro de residência e comorbidades, embora presentes nas bases originais do SIM e do Sinan, não integraram o vetor de características por razões de escopo e privacidade, impedindo a estratificação do desempenho do classificador por subgrupos populacionais. A literatura documenta que a qualidade do preenchimento varia por região e por sistema de informação (LIMA et al., 2020; BARTHOLOMAY et al., 2020), de modo que a generalização para outras localidades e pares de bases (por exemplo, SIM×SIH ou Sinan×GAL) depende de revalidação externa com padrões-ouro locais (COELI et al., 2021).

A validação de arquiteturas de aprendizado profundo (*deep learning*) constitui questão recorrente na literatura recente de *linkage*. Abordagens fim-a-fim como DeepMatcher (MUD-GAL et al., 2018) e DITTO (LI et al., 2020) recebem campos textuais brutos e aprendem representações via modelos de linguagem pré-treinados, reportando ganhos de 1 a 5 pontos percentuais de  $F_1$  sobre métodos tradicionais em *benchmarks* com milhares de exemplos rotulados. Neste estudo, três fatores justificam a opção por algoritmos baseados em árvore. Primeiro, o vetor de entrada é composto por 29 subscores numéricos de similaridade, configurando problema tabular no qual métodos de *ensemble* baseados em árvore consistentemente superam redes neurais profundas (GRINSZTAJN; OYALLON; VAROQUAUX, 2022; SHWARTZ-ZIV; ARMON, 2022). Segundo, a dimensão reduzida do conjunto positivo (247 pares verdadeiros) é insuficiente para o ajuste fino (*fine-tuning*) de modelos pré-treinados sem risco de sobreajuste severo. Terceiro, abordagens fim-a-fim requerem acesso aos campos textuais brutos (nomes, endereços), o que impõe restrições adicionais de privacidade em bases de saúde. O *framework* GZ-CMD, contudo, já incorpora modelos de linguagem de grande porte na camada de maior valor agregado: a revisão assistida por LLM opera exclusivamente sobre os 41 pares residuais que sobrevivem à cascata de filtragem, alocando raciocínio textual precisamente onde o clas-

sificador numérico esgota seu poder discriminativo. Essa estratégia posiciona o aprendizado profundo não como substituto do classificador tabular, mas como complemento especializado para a incerteza residual. A avaliação de arquiteturas fim-a-fim em bases administrativas nacionais de maior volume, com acesso a campos textuais e conjuntos positivos da ordem de milhares, configura agenda relevante para trabalhos futuros.

Apesar dessas restrições, os resultados sustentam que o pós-processamento supervisionado com *framework* configurável constitui avanço operacional em relação a limiares fixos e regras ad hoc. A documentação explícita dos compromissos entre precisão, sensibilidade e custo de revisão, associada à interpretabilidade via SHAP, confere transparência ao processo decisório, condição necessária para a adoção em rotinas de vigilância (MARKUS; KORS; RI-JNBEEK, 2021). Parte das limitações aqui discutidas, em particular a dependência de limiares fixos e a escalabilidade da revisão manual, é endereçada pelo arcabouço GZ-CMD, apresentado no Capítulo 8. Esse arcabouço propõe política de decisão por perda esperada e revisão assistida por modelo de linguagem como alternativas operacionais, reduzindo de 21.620 para 41 os pares que demandam revisão humana no cenário de vigilância.

# Capítulo 8

## Proposta de Arcabouço Operacional Auto-Calibrável (GZ-CMD)

### 8.1 Motivação e lacunas operacionais

A fundamentação epidemiológica para o uso da tuberculose como condição marcadora, bem como a relevância do *linkage* entre SIM e Sinan-TB para qualificação de indicadores, foi apresentada anteriormente nas seções 3.3 e 7.1. Neste capítulo, essa base é retomada de forma objetiva para sustentar a proposta operacional do GZ-CMD no cenário do município do Rio de Janeiro, no período de 2006 a 2016, em alinhamento com os dados oficiais mais recentes (World Health Organization, 2024; Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente, 2024).

No que tange às lacunas que permanecem após os resultados dos Capítulos 6 e 7, destacam-se três pontos. Primeiro, a assimetria de custos entre falso negativo e falso positivo, uma vez que a perda de um vínculo verdadeiro implica subestimação de mortalidade e perda de oportunidade de intervenção, enquanto o falso positivo tende a gerar sobretudo custo administrativo de revisão (OLIVEIRA et al., 2016). Segundo, a baixa portabilidade de limiares fixos sob variação de prevalência e de completude dos registros, limitação já discutida para cenários com forte desbalanceamento (SHAW et al., 2022). Terceiro, a natureza intensiva em trabalho da revisão manual na zona cinzenta, aspecto recorrente na literatura de *linkage* em saúde (CHRISTEN, 2012; COELI et al., 2021).

Nesse sentido, propõe-se o *Grey-Zone Cost-based Mixture Deferral* (GZ-CMD), um *framework* (estrutura metodológica) para governança da incerteza no *linkage* probabilístico, integrando classificação, calibração por âncoras, regras de guarda e revisão assistida por modelos de linguagem em uma política única de decisão auditável.

## 8.2 Dados, delineamento e representação

### 8.2.1 Bases de dados e período de estudo

Foram utilizados registros do Sinan-TB e declarações de óbito do SIM referentes ao município do Rio de Janeiro no período de 2006 a 2016, mantendo o mesmo recorte do núcleo empírico descrito no Capítulo 5. A geração de pares candidatos por blocagem foi conduzida no OpenRecLink, conforme estratégia descrita em (COELI; JR., 2002; JR; COELI, 2000) e detalhada na Seção 5.3. O universo final incluiu 61.696 pares candidatos, com 247 vínculos verdadeiros e 61.449 não-pares.

### 8.2.2 Padrão-ouro e validade de referência

O padrão-ouro foi construído por revisão manual de 61.696 pares candidatos, realizada por um único revisor com experiência em vigilância da tuberculose no IESC-UFRJ, resultando em 247 vínculos verdadeiros e 61.449 não-pares. Em consonância com a limitação metodológica já registrada na Seção 7.7, não houve segunda revisão independente e, portanto, não foi estimado coeficiente kappa de Cohen.

Essa opção metodológica preserva aderência ao desenho operacional adotado no estudo, mas requer cautela interpretativa nas regiões de maior ambiguidade do escore, onde erros residuais de rotulagem podem persistir. Ainda assim, a consistência observada nos conjuntos âncora, apresentada na Seção 8.3.2, fornece evidências indiretas favoráveis à estabilidade da referência.

### 8.2.3 Delineamento experimental

Foi adotada validação cruzada estratificada em cinco partições, com estratificação por COMPREC, mantendo coerência com o protocolo metodológico descrito nas seções 5.1 e 5.6.4. Em cada rodada, quatro partições foram usadas para treinamento e calibração, e uma para teste. As métricas foram reportadas como média e desvio-padrão entre rodadas, com comparações por teste t pareado ( $\alpha = 0,05$ ).

No experimento de revisão assistida, foram avaliados 1.410 pares da zona cinzenta extraídos exclusivamente de partições de teste, sem sobreposição com os dados usados no treinamento do classificador.

### 8.2.4 Representação dos pares e engenharia de atributos

A representação dos pares candidatos baseia-se no vetor de 29 subescores do Comparador de Registros (JARDIM, 2024; LUCENA, 2013), cuja descrição técnica já foi apresentada na Seção 5.4. Neste capítulo, o foco recai sobre o efeito operacional da representação, sobretudo na zona cinzenta, sem repetir a fundamentação já consolidada no método.

Para reduzir perda informacional associada à discretização de datas, foram adicionadas Medidas Contínuas de Diferença de Datas (MACD), com diferenças brutas em dias entre datas de nascimento e entre datas de óbito e notificação. A sigla MACD é adotada neste trabalho para designar essas medidas contínuas de diferença temporal e não guarda relação com o indicador homônimo do domínio financeiro (*Moving Average Convergence Divergence*). O impacto dessa adição é analisado na Seção 8.5.1, mantendo os valores de ablação observados no rascunho experimental.

## 8.3 Métodos: o *framework* GZ-CMD

O GZ-CMD organiza a tomada de decisão em quatro componentes: calibração por âncoras e bandas de confiança, regras de guarda determinísticas, motor de política baseado em perda esperada e revisão assistida para os casos de maior incerteza.

### 8.3.1 Calibração por âncoras e bandas de confiança

Em contraste com o esquema clássico de Fellegi–Sunter (FELLEGI; SUNTER, 1969), adota-se aqui calibração discriminativa dos escores brutos. A função sigmoide

$$P(y = 1 \mid s) = \frac{1}{1 + \exp(-(\alpha s + \beta))}$$

é ajustada por conjuntos âncora positivos ( $A^+$ ) e negativos ( $A^-$ ), dispensando rótulos explícitos em tempo de operação.

Os parâmetros  $\alpha$  e  $\beta$  são obtidos por restrições de média alvo sobre as âncoras,

$$\mathbb{E}_{s \in A^+} [\sigma(\alpha s + \beta)] = r^+, \quad \mathbb{E}_{s \in A^-} [\sigma(\alpha s + \beta)] = r^-.$$

Aspectos de identificabilidade, estabilidade numérica e efeito de contaminação nas âncoras estão formalizados no Apêndice A.1, em especial na Proposição 1.

Com  $p_{cal}$  estimado, definem-se três bandas: alta confiança de vínculo ( $p_{cal} \geq t_{high}$ ), zona cinzenta ( $t_{low} < p_{cal} < t_{high}$ ) e não-vínculo ( $p_{cal} \leq t_{low}$ ).

### 8.3.2 Validação empírica dos conjuntos âncora

A Tabela 8.1 resume a confrontação dos conjuntos âncora com o padrão-ouro, evidenciando pureza superior a 99% em ambos os extremos de escore.

Tabela 8.1: Validação dos conjuntos âncora contra o padrão-ouro.

Conjunto	Critério	$N$	Acertos	Erros	Precisão (%)
$A^+$ (âncoras positivas)	$s \geq 9,0$	520	516	4	99,2
$A^-$ (âncoras negativas)	$s < 5,0$	3.840	3.833	7	99,8

As taxas de erro observadas foram inferiores a 1%, sustentando o uso das âncoras como pseudo-rótulos operacionais para calibração periódica.

### 8.3.3 Regras de guarda determinísticas

As regras de guarda incorporam conhecimento epidemiológico e restrições de consistência que se sobrepõem ao classificador probabilístico em casos críticos.

- **Filtro temporal:** pares com data de óbito precedendo a data de diagnóstico em mais de 180 dias são forçados para não-vínculo; no recorte municipal de 2006 a 2016, essa regra não removeu vínculos verdadeiros no conjunto avaliado
- **Cardinalidade N:1:** cada registro Sinan-TB pode vincular-se a no máximo um registro SIM, enquanto um registro SIM pode estar associado a múltiplas notificações Sinan-TB do mesmo indivíduo em episódios distintos
- **Empate operacional:** quando dois candidatos ao mesmo registro Sinan-TB apresentam  $\Delta p < 0,05$ , ambos são encaminhados à revisão para evitar decisão arbitrária

### 8.3.4 Motor de política de decisão por perda esperada

O motor decisório substitui limiares fixos por minimização de perda esperada, com custos  $C_{FN}$ ,  $C_{FP}$  e  $C_{LLM}$ , além de taxas residuais  $e_{FP}$  e  $e_{FN}$  do módulo de revisão. Para um par com probabilidade calibrada  $p$ ,

$$\mathcal{L}(a_M | p) = (1 - p) C_{FP}, \quad \mathcal{L}(a_N | p) = p C_{FN},$$

$$\mathcal{L}(a_R | p) = C_{LLM} + (1 - p)e_{FP}C_{FP} + pe_{FN}C_{FN}.$$

A Tabela 8.2 resume os valores dos parâmetros de custo adotados nos dois modos operacionais. Os valores de  $C_{FP}$  e  $C_{FN}$  refletem a assimetria de consequências: no modo de vigilância, a perda de um óbito ( $C_{FN} = 50$ ) é cinco vezes mais custosa que um falso alarme ( $C_{FP} = 10$ ); no modo de confirmação, o custo de um falso positivo ( $C_{FP} = 100$ ) supera o do falso negativo ( $C_{FN} = 20$ ), priorizando a confiabilidade do vínculo.

Tabela 8.2: Parâmetros de custo operacionais do GZ-CMD.

Modo	$C_{FP}$	$C_{FN}$	$C_{LLM}$	$e_{FP}$ (faixa)	$e_{FN}$ (faixa)
Vigilância	10	50	0,25	0,02–0,10	0,06–0,15
Confirmação	100	20	0,25	0,02–0,10	0,06–0,15

As taxas de erro residual do revisor ( $e_{FP}$ ,  $e_{FN}$ ) variam por faixa de  $p_{cal}$ : valores mais baixos aplicam-se a pares próximos aos extremos de confiança, e valores mais altos a pares na região central da zona cinzenta. A sensibilidade da fronteira de decisão a essas parametrizações

é atenuada pela concavidade do EVR (Proposição 3): variações moderadas nos custos deslocam os limiares de revisão, mas preservam a estrutura qualitativa da triagem em três vias.

A dependência de  $\mathcal{L}(a_R | p)$  em relação a  $p$  sob revisor imperfeito é formalizada no Apêndice A.1, Proposição 2. A regra ótima é

$$\delta^*(p) = \arg \min_{a \in \{a_M, a_N, a_R\}} \mathcal{L}(a | p).$$

Define-se ainda o Valor Esperado da Revisão,

$$EVR(p) = \min\{\mathcal{L}(a_M | p), \mathcal{L}(a_N | p)\} - \mathcal{L}(a_R | p),$$

e encaminha-se o par para revisão quando  $EVR(p) > 0$ . A concavidade de  $EVR$  e a conexidade do intervalo de revisão são estabelecidas na Proposição 3, no Apêndice A.1.

### 8.3.5 Fluxo operacional do *framework*

O fluxo completo do GZ-CMD, formalizado no Algoritmo A.1 (Apêndice A.1), é composto pelas seguintes etapas sequenciais:

1. **Representação:** cada par candidato é codificado como vetor de 29 subescores do Comparador de Registros, opcionalmente acrescido das medidas contínuas MACD.
2. **Classificação:** um modelo supervisionado (*Random Forest*) produz escore bruto  $s$  para cada par.
3. **Calibração por âncoras:** os conjuntos  $A^+$  e  $A^-$  alimentam regressão sigmoide, convertendo  $s$  em probabilidade calibrada  $p_{cal}$ .
4. **Atribuição de banda:**  $p_{cal}$  posiciona o par em uma das seis faixas de confiança (Seção 8.3.1).
5. **Regras de guarda:** restrições determinísticas (filtro temporal, cardinalidade N:1, empate operacional) podem forçar decisão ou revisão, independentemente de  $p_{cal}$ .
6. **Triagem por perda esperada:** para os pares não capturados pelas regras de guarda, calcula-se  $EVR(p)$ ; pares com  $EVR > 0$  são candidatos a revisão, priorizados por  $EVR$  decrescente até o limite orçamentário  $L_{max}$ .
7. **Revisão assistida:** pares selecionados são convertidos em dossiê desidentificado e submetidos ao protocolo dual-agente com arbitragem (Seção 8.4).
8. **Decisão final:** a saída é um rótulo auditável (MATCH, NONMATCH ou UNSURE) com registro de agente, protocolo e códigos de razão.

## 8.4 Revisão clerical assistida

A revisão assistida foi concebida como módulo de apoio, não como substituição do julgamento epidemiológico. Cada par candidato da zona cinzenta é convertido em dossiê estruturado (subescores, metadados de calibração e códigos de guarda), analisado por dois agentes com instruções complementares, com arbitragem por terceiro agente quando há discordância.

### 8.4.1 Seleção do modelo de linguagem

A escolha do modelo considerou concordância com o padrão-ouro, precisão, sensibilidade,  $F_1$ , latência média por par e taxa de inconclusivos. O piloto avaliou 517 pares da zona cinzenta, a totalidade dos pares classificados como LLM REVIEW pela sequência operacional (*pipeline*) v3, distribuídos por amostragem estratificada segundo faixas de  $p_{cal}$ : *grey\_high* ( $n = 341$ ), *grey\_mid* ( $n = 133$ ), *near\_high* ( $n = 35$ ) e *high* ( $n = 8$ ). Cinco modelos foram avaliados com parâmetros padronizados (`temperature=0`, `max_tokens=4096`, `top_p=1`, `top_k=40`, penalidades nulas)<sup>1</sup>.

A padronização dos hiperparâmetros de inferência constitui condição necessária para que a comparação entre modelos reflita diferenças de capacidade intrínseca e não artefatos de configuração. A fixação de `temperature=0,0` assegura decodificação determinística (*greedy decoding*), eliminando variação estocástica entre execuções e garantindo reproduzibilidade integral das decisões para cada par avaliado, requisito central em tarefas de classificação contra padrão-ouro fixo. O limite de `max_tokens=4.096` foi dimensionado para acomodar, sem risco de truncamento, a cadeia completa de raciocínio do protocolo dual-agente: análise dos 29 subescores pelo agente A, resposta estruturada em JSON, análise independente pelo agente B e, quando necessário, reformulação pelo árbitro. Os parâmetros `top_p=1,0` e `top_k=40` complementam a temperatura nula: o primeiro desativa a amostragem por núcleo (*nucleus sampling*), enquanto o segundo preserva compatibilidade entre provedores de API sem impacto funcional sob decodificação *greedy*. As penalidades de presença e de frequência foram ambas fixadas em zero. O formato estruturado de saída (objeto JSON contendo veredito, códigos de razão e nível de confiança) requer repetição legítima de *tokens*, como nomes de campos e valores categóricos padronizados. Penalidades positivas distorceriam a distribuição lexical e poderiam corromper a integridade da resposta, introduzindo artefatos de paráfrase em posições que exigem literalidade.

Com  $n = 517$  pares e protocolo *dual-agent* com arbitragem, os intervalos de confiança de 95% para concordância se estreitam em relação ao piloto preliminar de 200 pares, mas as diferenças absolutas entre modelos permanecem dentro das margens de sobreposição dos intervalos. Para discriminar com potência de 80% uma diferença de 3,5 pontos percentuais

---

<sup>1</sup>Dois modelos originalmente previstos (DeepSeek R1 e Qwen3-235B) foram substituídos por DeepSeek V3.2 e Qwen3-VL-30B, respectivamente, devido à indisponibilidade persistente (erro 503) na infraestrutura Fireworks AI durante o período de coleta. Os substitutos pertencem à mesma família de modelos e operam com os mesmos parâmetros de inferência.

Tabela 8.3: Avaliação comparativa de modelos para revisão assistida ( $n = 517$  pares da zona cinzenta; intervalos de Wilson, 95%).

Modelo	Válidos <sup>a</sup>	Concordância (%)	Precisão	Sensibilidade	$F_1$	Inconcl. (%)	Latência (s/par)
Kimi K2.5	499	97,8 [96,1; 98,8]	0,909	0,789	<b>0,845</b>	3,5	7,3
GPT-4o	508	95,5 [93,3; 97,0]	1,000	0,343	0,511	1,7	6,3
DeepSeek V3.2	232	93,5 [89,6; 96,0]	0,615	0,762	0,681	55,1	54,0
Qwen3-VL-30B	509	94,9 [92,6; 96,5]	1,000	0,333	0,500	1,6	7,5
GLM-5	212	95,3 [91,5; 97,4]	0,769	0,833	0,800	59,0	57,6

<sup>a</sup> Pares com decisão MATCH/NONMATCH; excluídos os UNSURE.

entre dois modelos, seriam necessários aproximadamente 860 pares por modelo; para uma margem de 1,5 pontos percentuais, cerca de 4.100 pares , cenário impraticável no escopo deste estudo.

Dois comportamentos contrastantes emergem: GPT-4o e Qwen3-VL-30B exibem precisão perfeita (1,000) mas sensibilidade baixa ( $\approx 0,34$ ), sugerindo postura ultraconservadora que tende a omitir vínculos verdadeiros; DeepSeek V3.2 e GLM-5, ao contrário, apresentam taxas de inconclusivos superiores a 55%, refletindo cautela excessiva que delega a maioria dos casos à arbitragem e eleva substancialmente a latência ( $>54$  s/par). O Kimi K2.5 (Kimi Team, 2025) obteve o melhor equilíbrio operacional:  $F_1$  mais alto (0,845), taxa de inconclusivos moderada (3,5%) e latência competitiva (7,3 s/par), viabilizando o *pipeline* em escala sem pretensão de superioridade estatisticamente demonstrada em concordância global. Embora o relatório técnico do Kimi K2.5 enfatize capacidades multimodais e agênticas, o modelo demonstra igualmente competência em tarefas de raciocínio textual estruturado, como evidenciado pela capacidade de interpretar dossiês numéricos de subcores e produzir juízos coerentes com códigos de razão padronizados.

A análise da matriz de confusão por modelo revela implicações operacionais distintas. O Kimi K2.5 apresentou 30 verdadeiros positivos, 3 falsos positivos, 8 falsos negativos e 458 verdadeiros negativos entre os 499 pares com decisão válida, configurando equilíbrio entre erros de omissão e comissão compatível com o cenário de vigilância. GPT-4o e Qwen3-VL-30B, em contraste, não produziram nenhum falso positivo, mas omitiram, respectivamente, 23 e 26 dos 41 vínculos verdadeiros presentes na amostra , comportamento que, em contexto de vigilância epidemiológica, resultaria em subestimação sistemática de mortalidade por tuberculose. DeepSeek V3.2 e GLM-5 alcançaram distribuição mais equilibrada entre tipos de erro (sensibilidade de 76,2% e 83,3%, respectivamente), porém a elevada taxa de inconclusivos reduz o denominador efetivo de avaliação a menos da metade da amostra original, comprometendo a representatividade das métricas reportadas para esses modelos.

A eficiência do protocolo de arbitragem constitui indicador complementar relevante. O Kimi K2.5 atingiu consenso direto entre os agentes A e B em 503 dos 517 pares (97,3%), recorrendo ao árbitro em apenas 14 casos (2,7%), o que explica a latência competitiva de 7,3 s/par. GPT-4o e Qwen3-VL-30B apresentaram padrão semelhante, com apenas 6 arbitragens cada (1,2%), refletindo alta concordância entre agentes , ainda que essa concordância decorra predominantemente de ambos os agentes classificarem o par como NONMATCH. DeepSeek V3.2

e GLM-5, por outro lado, acionaram o mecanismo de arbitragem em 213 (41,2%) e 155 (30,0%) dos pares, respectivamente, evidenciando discordância sistemática entre agentes e contribuindo para latências superiores a 54 s/par. Esse padrão sugere que modelos com cadeia de raciocínio estendida (*chain-of-thought*) tendem, nesta tarefa, a identificar mais fontes de incerteza do que a resolvê-las, ampliando a carga computacional sem ganho proporcional em acurácia.

Do ponto de vista da validade estatística, os intervalos de Wilson para concordância sobrepõem-se entre todos os cinco modelos, impedindo a rejeição de igualdade ao nível de 95% de confiança. Entretanto, o  $F_1$  varia de 0,500 (Qwen3-VL-30B) a 0,845 (Kimi K2.5), diferença de 0,345 que reflete capacidades qualitativamente distintas de identificação de vínculos verdadeiros. Essa aparente contradição, concordância estatisticamente indistinguível versus  $F_1$  substancialmente diferente, decorre do forte desbalanceamento da amostra (476 não-pares contra 41 pares): a concordância global é dominada pelos verdadeiros negativos, nos quais todos os modelos apresentam desempenho elevado, enquanto o  $F_1$  pondera explicitamente a detecção da classe minoritária. Essa assimetria reforça a importância de reportar métricas desagregadas por classe em cenários de forte desbalanceamento, prática nem sempre observada na literatura de *record linkage*.

#### **8.4.2 Considerações éticas**

O uso de modelos de linguagem em dados de saúde requer salvaguardas técnicas e normativas, mesmo sob desidentificação (VAYENA; BLASIMME; COHEN, 2018). Foram adotadas quatro medidas: ausência de identificadores nominais no dossiê, processamento em ambiente controlado, enquadramento ético-regulatório para dados secundários desidentificados e registro auditável das decisões automatizadas.

No plano regulatório, considerou-se a Resolução CNS n. 510/2016 para pesquisas com bases de dados sem possibilidade de identificação individual (Brasil. Ministério da Saúde. Conselho Nacional de Saúde, 2016).

### **8.5 Resultados e avaliação**

#### **8.5.1 Experimento 1: ablação e impacto das medidas contínuas**

A Tabela 8.4 apresenta os resultados de ablação com e sem MACD, preservando os valores obtidos no rascunho experimental e organizados por modo operacional.

Os ganhos em  $F_1$ -Score foram modestos, mas houve redução operacional consistente de revisões, entre 15% e 17%, indicando maior assertividade nos casos limítrofes. Os testes t pareados para diferença de  $F_1$  não atingiram significância ao nível de 5% (Vigilância:  $t(4) = 1,75$ ,  $p = 0,155$ ; Confirmação:  $t(4) = 0,63$ ,  $p = 0,564$ ), resultado compatível com cinco rodadas e variabilidade interpartições.

Tabela 8.4: Resultados do experimento de ablação (média  $\pm$  desvio-padrão, 5 rodadas). A coluna “Revisões” indica o número absoluto de pares encaminhados à revisão por LLM em cada participação de teste.

Modo	Configuração	Precisão	Sensibilidade	$F_1$ -Score	$F_{0.5}$	$F_2$	Revisões
Confirmação	MACD OFF	$0,949 \pm 0,006$	$0,937 \pm 0,019$	$0,943 \pm 0,009$	$0,946 \pm 0,005$	n.a.	$138 \pm 12$
Confirmação	MACD ON	<b><math>0,957 \pm 0,004</math></b>	$0,938 \pm 0,023$	<b><math>0,947 \pm 0,011</math></b>	<b><math>0,953 \pm 0,004</math></b>	n.a.	<b><math>114 \pm 7</math></b>
Vigilância	MACD OFF	$0,934 \pm 0,009$	$0,964 \pm 0,013$	$0,949 \pm 0,004$	n.a.	$0,958 \pm 0,009$	$128 \pm 10$
Vigilância	MACD ON	<b><math>0,947 \pm 0,006</math></b>	$0,962 \pm 0,013$	<b><math>0,954 \pm 0,005</math></b>	n.a.	<b><math>0,959 \pm 0,010</math></b>	<b><math>109 \pm 10</math></b>

### 8.5.2 Experimento 2: desempenho da revisão assistida

No conjunto de 1.410 pares da zona cinzenta, o protocolo de consenso entre agentes apresentou concordância de 97,09%, com arbitragem em 2,91% dos casos. A Tabela 8.5 resume o desempenho frente ao padrão-ouro.

Tabela 8.5: Desempenho da revisão assistida contra o padrão-ouro (n=1.410).

Decisão do módulo	N	VP	FP	VN	FN
Vínculo	85	81	4	n.a.	n.a.
Não-vínculo	1.284	n.a.	n.a.	1.277	7
Inconclusivo	41	10	n.a.	31	n.a.
Total resolvido	1.369	81	4	1.277	7

Para os 1.369 pares resolvidos, observou-se acurácia global de 99,2%, precisão para vínculo de 95,3% e sensibilidade para vínculo de 92,0%, permanecendo 2,9% de inconclusivos para arbitragem especializada. Ressalva-se que essas métricas são condicionadas à qualidade do padrão-ouro de revisor único; na ausência de dupla revisão independente, não se pode excluir que parte da concordância reflita vieses compartilhados entre o modelo de linguagem e o revisor original, sobretudo nos pares de maior ambiguidade.

### 8.5.3 Redução cumulativa da carga de revisão clerical

Para dimensionar o ganho operacional do *framework* integrado, a Tabela 8.6 apresenta a redução cumulativa do volume de pares encaminhados à revisão clerical humana conforme cada componente do GZ-CMD é adicionado ao *pipeline*, partindo do cenário-base em que todos os 61.696 pares candidatos requerem inspeção manual.

Tabela 8.6: Redução cumulativa da carga de revisão clerical por componente do *framework* GZ-CMD (modo vigilância, MACD ativo,  $n = 61.696$  pares candidatos).

Etapa	Componente adicionado	Pares p/ revisão	% total	Redução vs anterior
0	Sem modelo ( <i>baseline</i> )	61.696	100,0%	—
1	Limiares F-S ( $< 5 \rightarrow$ NM; $\geq 9 \rightarrow$ M)	21.620	35,0%	$-65,0\%$
2	+ Regras de guarda determinísticas	18.799	30,5%	$-13,0\%$
3	+ Calibração Platt + motor de perda esperada	1.410	2,3%	$-92,5\%$
4	+ Revisão LLM dual-agent (Kimi K2.5)	41	0,07%	$-97,1\%$
<b>Redução total (etapa 1 → 4)</b>		<b>21.620 → 41 pares = 99,8%</b>		

Na etapa 1, a aplicação dos limiares tradicionais de Fellegi–Sunter classifica automaticamente 39.981 pares abaixo do limiar inferior (escore  $< 5$ ) como não-vínculo e 95 pares acima

do limiar superior (escore  $\geq 9$ ) como vínculo, reduzindo a zona cinzenta a 21.620 pares (35,0% do total), que na abordagem convencional seriam integralmente encaminhados à revisão clerical humana.

Na etapa 2, as regras de guarda determinísticas , a saber, filtro temporal (óbito anterior ao diagnóstico em mais de 180 dias) e nota global inferior a 3, resolvem 2.821 pares adicionais com **zero pares verdadeiros perdidos**, reduzindo o volume pendente para 18.799.

A etapa 3 constitui o maior salto de redução: a calibração de Platt converte escores brutos em probabilidades de vínculo, e o motor de perda esperada avalia, para cada par, se o valor esperado da revisão por LLM supera o custo de uma decisão automática. Apenas 1.410 pares apresentam valor esperado de revisão positivo ( $EVR > 0$ ) dentro do orçamento de chamadas (máximo de 2.000 no modo vigilância); os demais 17.389 pares são auto-resolvidos pela regra de custo mínimo. Desses 1.410 pares, 104 contêm vínculos verdadeiros segundo o padrão-ouro. O motor de perda esperada responde, isoladamente, por 92,5% da redução acumulada entre as etapas 2 e 4, configurando-se como o componente de maior impacto da cascata.

Um efeito secundário da cascata é o enriquecimento progressivo da prevalência de pares verdadeiros no subconjunto encaminhado a cada etapa: de 0,4% na base completa (247/61.696), para 0,7% na zona cinzenta convencional (151/21.620), 7,4% no lote selecionado pelo motor de perda esperada (104/1.410), chegando a aproximadamente 39% entre os 41 inconclusivos residuais (Figura 8.1b). Essa concentração tem dupla implicação: o módulo LLM opera sobre subconjunto já enriquecido, o que favorece seu desempenho discriminativo, e o revisor humano final atua exclusivamente sobre casos de incerteza genuína, maximizando o valor agregado de cada hora de trabalho especializado.

Na etapa 4, o protocolo de consenso dual-agent resolve 1.369 dos 1.410 pares (97,1%), restando 41 inconclusivos para arbitragem humana especializada , equivalentes a **0,07% do universo original** e a **0,19% da zona cinzenta**. No modo confirmação (com orçamento de 1.000 chamadas e custos de falso positivo mais elevados), o motor seleciona apenas 435 pares para revisão LLM, projetando-se residual humano de aproximadamente 13 pares.

## 8.6 Discussão

Os achados corroboram a tese de que o ganho operacional do GZ-CMD não depende apenas de elevação marginal de métricas agregadas, mas da explicitação de uma política de decisão orientada por custo, com triagem formal da incerteza e rastreabilidade da revisão. Essa contribuição estende os resultados dos Capítulos 6 e 7 ao introduzir mecanismo de calibração adaptativo e regra econômica explícita para alocação de revisão.

A cascata de redução de 21.620 para 41 pares , três ordens de grandeza, evidencia que o ganho operacional decorre menos da capacidade individual de cada componente e mais de sua articulação sequencial. O motor de perda esperada é responsável por 92,5% da redução entre as etapas intermediárias, enquanto a revisão assistida resolve os casos residuais em aprox-

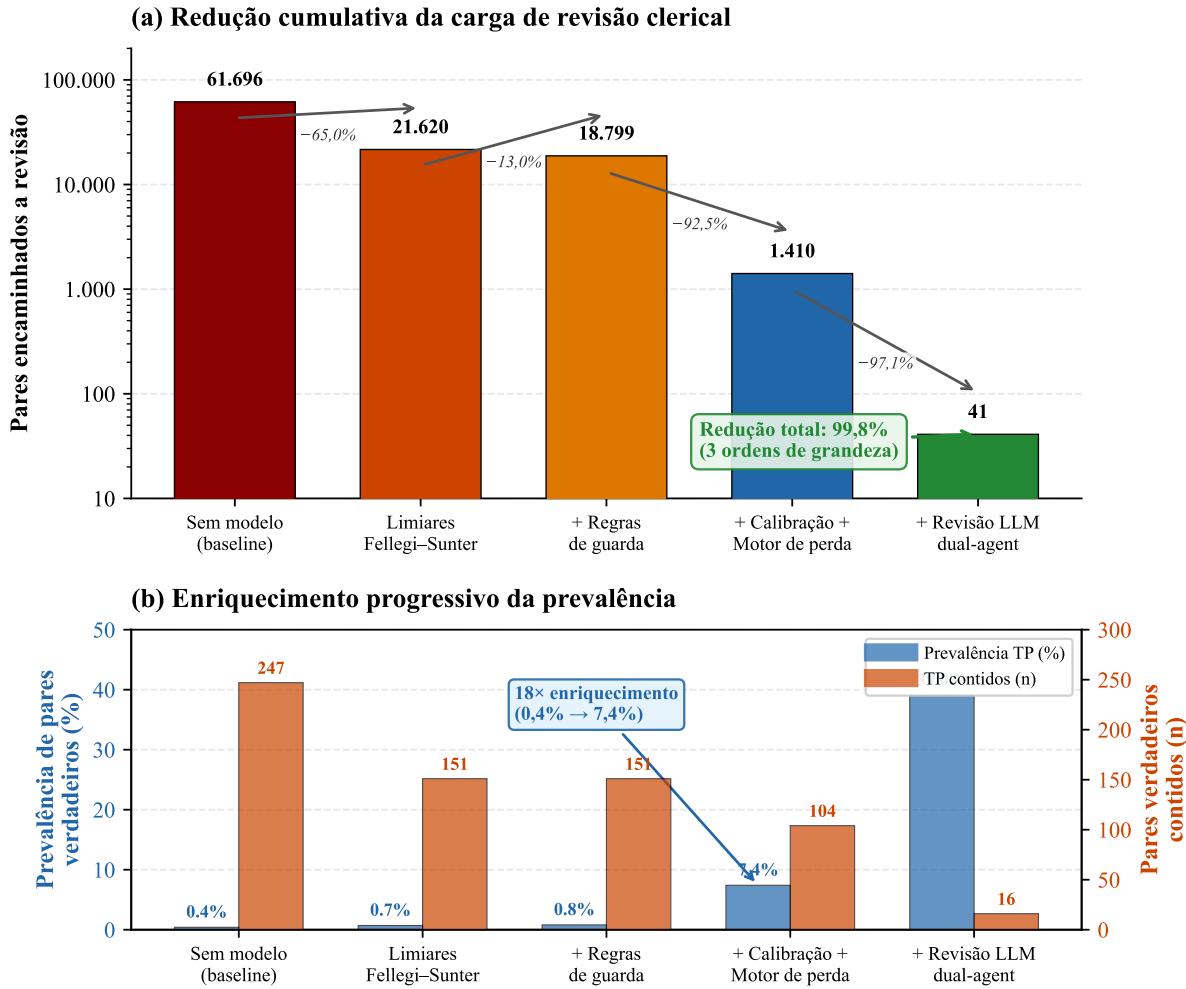


Figura 8.1: Redução cumulativa da carga de revisão clerical e enriquecimento da prevalência ao longo das etapas do *framework* GZ-CMD (modo vigilância, MACD ativo,  $n = 61.696$  pares). (a) Volume de pares encaminhados a revisão em escala logarítmica; percentuais indicam a redução em relação à etapa anterior. (b) Prevalência de pares verdadeiros (%) e contagem absoluta de verdadeiros positivos contidos em cada subconjunto. O enriquecimento de 18× entre a base completa e o lote LLM evidencia o efeito concentrador da cascata.

imadamente 63 minutos de processamento, período contrastante com as semanas que a revisão integral de 21.620 pares demandaria em regime manual. A prevalência de pares verdadeiros no subconjunto encaminhado ao revisor humano final (aproximadamente 39%) é duas ordens de grandeza superior à da base completa (0,4%), assegurando que cada decisão humana incida sobre caso genuinamente ambíguo.

A seleção do modelo de linguagem para o módulo de revisão assistida exerce influência direta sobre a viabilidade operacional do *pipeline* integrado. Com o Kimi K2.5 operando a 7,3 s/par e taxa de inconclusivos de 3,5%, a totalidade dos 517 pares da zona cinzenta seria processada em aproximadamente 63 minutos, com apenas 18 pares residuais requerendo escalamento para revisão humana. Em cenário contrafactual com DeepSeek V3.2 ou GLM-5, a mesma carga demandaria mais de oito horas de processamento e geraria entre 285 e 305 pares inconclusivos, anulando na prática o propósito de automação e transferindo o ônus decisório de

volta ao revisor especializado. A razão entre o tempo de processamento e o volume de decisões efetivamente resolvidas constitui, portanto, critério discriminante mais informativo do que a concordância global isolada.

No que concerne ao orçamento de erros, o Kimi K2.5 identificou corretamente 30 dos 41 vínculos verdadeiros na amostra estratificada da zona cinzenta, com 3 falsos positivos e 8 falsos negativos. Projetando essas taxas para o universo de 1.410 pares encaminhados à revisão pelo motor de perda esperada, estima-se que o módulo resolveria aproximadamente 96,5% dos casos com decisão auditável, limitando a intervenção humana a um resíduo de baixa volumetria e elevada ambiguidade. Essa configuração preserva o princípio arquitetural do GZ-CMD: resolver automaticamente o que é resolvível e concentrar esforço humano especializado nos casos em que a incerteza é genuína, sem comprometer a rastreabilidade completa das decisões.

A latência e o custo de inferência dos modelos de linguagem são sensíveis a fatores ex-trínsecos: disponibilidade de infraestrutura, congestionamento de API, variações de preço por *token* e atualização das versões dos modelos. Esses parâmetros devem ser reavaliados periodicamente em contexto de implantação. A substituição forçada de DeepSeek R1 e Qwen3-235B por variantes da mesma família, embora justificada pela indisponibilidade técnica persistente durante o período de coleta, introduz incerteza adicional sobre a representatividade dos resultados para as arquiteturas originalmente previstas. Por essa razão, recomenda-se que estudos futuros de validação multicêntrica incorporem cláusula de reavaliação periódica do modelo selecionado, incluindo candidatos de novas gerações que venham a tornar-se disponíveis.

Do ponto de vista teórico, a formulação por perda esperada inclui o caso clássico de Fellegi–Sunter como situação particular sob parametrização específica de custos e erros residuais, conforme demonstrado na Proposição 3, no Apêndice A.1. Ainda assim, permanece como agenda prioritária a comparação empírica direta com implementação canônica de limiares clássicos no mesmo conjunto de dados.

Persistem limitações para generalização externa. O estudo foi conduzido em um único contexto municipal (Rio de Janeiro, 2006–2016), com 247 pares positivos no conjunto de referência; a ausência de dupla revisão independente restringe inferências sobre erro do padrão-ouro; e a dependência de infraestrutura para inferência local de modelos de linguagem pode impor barreiras operacionais em ambientes de menor capacidade computacional. Esses aspectos não invalidam os ganhos observados, mas delimitam o escopo de transferência dos parâmetros para outros cenários.

Merece atenção particular o risco de circularidade entre padrão-ouro, conjuntos âncora e avaliação do módulo de revisão. Como os conjuntos âncora são derivados do mesmo padrão-ouro produzido por revisor único, eventuais erros sistemáticos de rotulagem na zona cinzenta podem propagar-se para a calibração e, por consequência, para as fronteiras de decisão do *framework*. A Proposição 1 (Apêndice A.1) fornece cota superior para o viés de contaminação, porém pressupõe contaminação uniforme nas âncoras, hipótese que pode não se sustentar sob viés não-aleatório de um único revisor. De modo análogo, a acurácia de 99,2% reportada para a

revisão assistida (Seção 8.5.2) é condicionada à qualidade da referência e pode estar inflacionada caso o modelo de linguagem concorde com erros do revisor. A mitigação dessa limitação requer validação futura com dupla revisão independente em amostra estratificada da zona cinzenta, acompanhada de estimativa de coeficiente kappa de Cohen.

Como desdobramento, recomenda-se validação multicêntrica em outras unidades da federação, avaliação de estabilidade temporal das âncoras e testes controlados de custo-efetividade da revisão assistida em escala de rotina, com vistas à consolidação de protocolos de *linkage* auditáveis no âmbito da vigilância epidemiológica.

# Capítulo 9

## Conclusões

Este estudo desenvolveu e avaliou estratégias de pós-processamento baseadas em aprendizado de máquina para o *linkage* probabilístico entre bases de dados de saúde. A discussão apresentada no Capítulo 7 evidenciou que o ganho prático do pós-processamento não se restringe a métricas agregadas, mas se manifesta sobretudo na recuperação de pares verdadeiros concentrados na zona cinzenta do escore e na redução do custo de revisão manual, fatores determinantes para a viabilidade do uso rotineiro de dados vinculados em vigilância e gestão.

### 9.1 Síntese dos principais achados

Na comparação de técnicas (Capítulo 6), modelos baseados em árvores apresentaram desempenho mais elevado em relação a abordagens lineares e a classificadores mais sensíveis à padronização de escalas, em particular sob desbalanceamento extremo. A estratificação por faixas do escore do OpenRecLink mostrou que a maior parte dos erros e ambiguidades se concentra no intervalo intermediário, no qual a combinação de múltiplas evidências (nominais, data de nascimento, endereço e município) é necessária para reduzir a incerteza.

A análise de ablação e a validação cruzada indicaram que configurações ML-only oferecem excelente desempenho de equilíbrio, enquanto combinações híbridas configuráveis (AND e OR) permitem selecionar pontos operacionais com maior precisão ou maior estabilidade, conforme o objetivo de uso. Por fim, o estudo de impacto epidemiológico evidenciou recuperação de óbitos adicionais e maior rendimento por registro revisado, reforçando a aplicabilidade operacional do pós-processamento supervisionado.

O arcabouço (*framework*) GZ-CMD, apresentado no Capítulo 8, estendeu a abordagem de pós-processamento ao incorporar calibração por âncoras, uma política de decisão em três vias (aceitar, rejeitar e revisar) fundamentada no valor esperado de revisão e revisão assistida por modelos de linguagem de grande porte (*large language models*) para os pares candidatos da zona cinzenta. Os resultados indicaram que a configuração orientada à vigilância atingiu F<sub>1</sub>-Score de 0,954, enquanto o modo orientado à confirmação alcançou precisão de 0,957. Quando

considerada a cascata completa da sequência operacional (*pipeline*) , composta por limiares Fellegi–Sunter, regras de guarda determinísticas, motor de perda esperada e revisão LLM dual-agent, o volume de revisão clerical humana reduziu-se de 21.620 para 41 pares, correspondendo a três ordens de grandeza (99,8%), com processamento em aproximadamente 63 minutos de inferência.

## 9.2 Atendimento aos objetivos específicos

Os objetivos específicos definidos no Capítulo 4 foram atendidos conforme descrito a seguir.

1. **Comparação de técnicas de aprendizado de máquina.** Foram comparados classificadores lineares, métodos baseados em árvores, redes neurais e estratégias de combinação de modelos no problema de classificação de pares candidatos SIM–Sinan-TB, evidenciando diferenças de desempenho e comportamento sob desbalanceamento (Seções 6.4 e 6.3).
2. **Avaliação de estratégias de balanceamento.** Foram avaliadas estratégias de reamostragem e ponderação de classes, com análise de sensibilidade em validação cruzada, demonstrando robustez do desempenho em um conjunto de estratégias e identificando configurações com melhor F<sub>1</sub>-Score médio (Seção 6.8, Tabela 6.6).
3. **Ajuste de pontos de corte e regras de negócio.** Foram propostas e avaliadas rotinas de escolha de limiar e regras determinísticas baseadas no conhecimento do domínio, com foco na recuperação de pares verdadeiros na zona cinzenta sem crescimento desproporcional de falsos positivos (Seções 6.2 e 6.6).
4. **Duas estratégias complementares (revocação e precisão).** Foram operacionalizadas estratégias orientadas à maximização da revocação (recuperação exaustiva) e à maximização da precisão (alta confiabilidade), com discussão explícita das implicações operacionais e da escolha por contexto (Seções 6.5, 6.6 e 7.2).
5. **Sistematização reproduzível dos resultados.** Os experimentos foram documentados por meio de scripts, tabelas e figuras que registram configurações, pontos operacionais e resultados comparativos, de modo a apoiar reproduzibilidade e uso como protocolo de pós-processamento em cenários similares (Seções 6.7 e 6.8).
6. **Discussão de generalização.** Foram discutidas condições e limitações para adaptação a outros cenários de *linkage* em saúde, incluindo dependência de padrão-ouro, qualidade de preenchimento e disponibilidade de variáveis, além da necessidade de validação externa (Seção 7.7).

## 9.3 Contribuições

As contribuições centrais deste estudo concentram-se em quatro dimensões. A primeira é metodológica, ao propor e avaliar um *framework* configurável de pós-processamento que combina classificadores probabilísticos e regras determinísticas e explicita a fronteira de compromisso entre precisão e revocação. A segunda é operacional, ao quantificar o custo de revisão manual e demonstrar como a seleção de ponto operacional pode viabilizar o uso rotineiro de dados vinculados. A terceira é epidemiológica, ao evidenciar a recuperação de eventos relevantes na zona cinzenta e o potencial de qualificação de indicadores derivados de bases integradas. A quarta é de reproduzibilidade, pela sistematização dos experimentos em artefatos e rotinas que podem ser reaplicados em cenários análogos, incluindo a disponibilização do comparador de registros em repositório de código aberto (JARDIM, 2024).

A quinta dimensão, desenvolvida no Capítulo 8, é de governança da incerteza, ao propor o *framework* GZ-CMD como arcabouço auto-calibrável que integra classificação, calibração por âncoras (*Platt scaling*), política de três vias baseada em perda esperada e revisão assistida por modelos de linguagem , consolidando em uma estrutura única e auditável componentes que a literatura de *linkage* em saúde costuma tratar de forma isolada. No cenário empírico avaliado, essa estrutura reduziu o volume de revisão clerical de 21.620 para 41 pares (99,8%), com enriquecimento progressivo da prevalência de verdadeiros no subconjunto encaminhado a cada camada (de 0,4% na base completa para 7,4% no lote LLM e  $\approx$ 39% nos inconclusivos finais), demonstrando que o *pipeline* concentra o esforço humano nos casos de incerteza genuína. Os detalhes experimentais, incluindo os atributos temporais MACD (Medidas Contínuas de Diferença de Datas) e a análise de custo-benefício da revisão, encontram-se no Capítulo 8 e no Apêndice A.

## 9.4 Trabalhos futuros

Como continuidade, destacam-se: (i) validação externa em outros pares de bases e em diferentes contextos epidemiológicos; (ii) incorporação de variáveis clínicas e temporais adicionais para ampliar análises de impacto; (iii) estratégias de amostragem adaptativa e *active learning* para reduzir a dependência de rotulagem manual; (iv) integração do pós-processamento em *pipelines* operacionais com monitoramento contínuo de desempenho e auditoria; (v) avaliação do efeito de diferentes esquemas de bloqueio e comparadores sobre a distribuição da zona cinzenta e sobre a estabilidade dos modelos; e (vi) investigação de arquiteturas de aprendizado profundo fim-a-fim (MUDGAL et al., 2018; LI et al., 2020) em bases administrativas de maior volume e com acesso a campos textuais brutos, cenário no qual a disponibilidade de conjuntos positivos da ordem de milhares pode viabilizar o ajuste fino de modelos pré-treinados. No que concerne ao GZ-CMD, sua validação empírica restringiu-se ao município do Rio de Janeiro no período de 2006 a 2016, com padrão-ouro produzido por revisor único , condições que

limitam a generalização imediata para outros municípios, períodos ou pares de bases de dados com prevalências e completudes distintas. A replicação em cenários com múltiplos revisores e a avaliação de sensibilidade a variações na qualidade de preenchimento constituem passos necessários antes da adoção em escala. Espera-se que os resultados aqui apresentados contribuam para o fortalecimento das práticas de *linkage* probabilístico no âmbito da vigilância epidemiológica no Brasil.

# Referências Bibliográficas

ALMADANI, A. et al. Linking electronic health records for multiple sclerosis research: Comparison of deterministic, probabilistic, and machine learning linkage methods. **JMIR Medical Informatics**, v. 14, p. e79869, 2026. 2, 10

ASHER, J. et al. An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. **International Journal of Environmental Research and Public Health**, v. 17, n. 18, p. 6937, 2020. 3

BARRETO, M. L. et al. The center for data and knowledge integration for health (cidacs). **International Journal of Population Data Science**, v. 4, n. 2, p. 1140, 2019. 1, 4

BARTHOLOMAY, P. et al. Melhoria da qualidade das informações sobre tuberculose a partir do relacionamento entre bases de dados. **Cadernos de Saúde Pública**, v. 30, n. 11, p. 2459–2470, 2014. 4, 17, 29, 37, 39, 41, 43, 45

BARTHOLOMAY, P. et al. Lacunas na vigilância da tuberculose drogarresistente: relacionando sistemas de informação do Brasil. **Cadernos de Saúde Pública**, v. 36, n. 5, p. e00082219, 2020. [S4] Assesses underreporting of drug-resistant TB in Brazil using probabilistic linkage between SITE-TB, SINAN, GAL, and SIM databases. 37, 44, 48

BINETTE, O.; STEORTS, R. C. (almost) all of entity resolution. **Science Advances**, v. 8, n. 12, p. eabi8021, 2022. 2, 9

BOYD, S.; VANDENBERGHE, L. **Convex Optimization**. Cambridge: Cambridge University Press, 2004. Acesso em: 9 fev. 2026. Disponível em: <<https://web.stanford.edu/~boyd/cvxbook/>>. 83

Brasil. Ministério da Saúde. Conselho Nacional de Saúde. **Resolução nº 510, de 7 de abril de 2016**. 2016. Diário Oficial da União, Brasília, 24 maio 2016, Seção 1, p. 44–46. Dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais cujos procedimentos metodológicos envolvam a utilização de dados diretamente obtidos com os participantes ou de informações identificáveis ou que possam acarretar riscos maiores do que os existentes na vida cotidiana. 57

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Manual de Recomendações para o Controle da Tuberculose no Brasil**. 2. ed. Brasília, 2019. 12

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente. **Boletim Epidemiológico de Tuberculose 2024**. Brasília, 2024. 10, 12, 50

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. 6

CABRAL, A. L. L. V. et al. Itinerários terapêuticos: o estado da arte da produção científica no Brasil. **Ciência & Saúde Coletiva**, v. 16, n. 12, p. 4433–4442, 2011. 44

CHAWLA, N. V. et al. SMOTE: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002. 7, 14, 21, 28, 29, 35

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. 6, 21

CHRISTEN, P. **Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection**. Berlin, Heidelberg: Springer, 2012. 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 18, 20, 25, 26, 28, 32, 33, 37, 50

COELI, C. M.; JR., K. R. d. C. Avaliação de diferentes estratégias de blocagem no relacionamento probabilístico de registros. **Revista Brasileira de Epidemiologia**, v. 5, n. 2, p. 185–196, 2002. 1, 2, 4, 9, 10, 11, 12, 17, 51

COELI, C. M. et al. Record linkage under suboptimal conditions for data-intensive evaluation of primary care in Rio de Janeiro, Brazil. **BMC Medical Informatics and Decision Making**, v. 21, n. 1, p. 190, 2021. [S2] Describes strategies for linking Brazilian databases with poor-quality identifiers, combining deterministic, probabilistic, and clerical review approaches. 17, 26, 38, 40, 44, 48, 50

de Savigny, D.; ADAM, T. **Systems Thinking for Health Systems Strengthening**. Geneva: World Health Organization, 2009. ISBN 978-92-4-156389-5. 43

DOIDGE, J. C.; HARRON, K. Demystifying probabilistic linkage: Common myths and misconceptions. **International Journal of Population Data Science**, v. 3, n. 1, p. 410, 2018. [S1] Clarifies common misconceptions about probabilistic vs deterministic linkage, highlighting how implementation choices affect outputs. 26, 42

DOIDGE, J. C.; HARRON, K. L. Reflections on modern methods: linkage error bias. **International Journal of Epidemiology**, v. 48, n. 6, p. 2050–2060, 2019. [S1] Conceptual framework for understanding how linkage error leads to information and selection bias in epidemiological studies using linked data. 25, 29, 32, 38, 39, 47

DONABEDIAN, A. The quality of care: How can it be assessed? **JAMA**, v. 260, n. 12, p. 1743–1748, 1988. 1

DUVALL, S. L.; KERBER, R. A.; THOMAS, A. Extending the Fellegi–Sunter probabilistic record linkage method for approximate field comparators. **Journal of Biomedical Informatics**, v. 43, n. 1, p. 24–30, 2010. [S7] Extends the Fellegi-Sunter method to incorporate approximate string comparators in weight calculation, reducing misclassification by 25%. 26, 34

ENAMORADO, T.; FIFIELD, B.; IMAI, K. Using a probabilistic model to assist merging of large-scale administrative records. **American Political Science Review**, v. 113, n. 2, p. 353–371, 2019. 4

ENAMORADO, T.; FIFIELD, B.; IMAI, K. Using a probabilistic model to assist merging of large-scale administrative records. **American Political Science Review**, v. 113, n. 2, p. 353–371, 2019. [S7] Modern implementation of Fellegi-Sunter with Bayesian extensions for merging large-scale administrative records, introducing the fastLink package. 26, 30

- 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. 1, 4, 5, 7, 9, 28, 36, 52, 84
- GALAR, M. et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. **IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)**, v. 42, n. 4, p. 463–484, 2012. [S5] Comprehensive review of ensemble methods for class imbalance, covering bagging, boosting, and hybrid approaches relevant to health ML applications. 28, 29, 32
- GERHARDT, T. E. Itinerários terapêuticos em situações de pobreza: diversidade e pluralidade. **Cadernos de Saúde Pública**, v. 22, n. 11, p. 2449–2463, 2006. 44
- GRINSZTAJN, L.; OYALLON, E.; VAROQUAUX, G. Why do tree-based models still outperform deep learning on typical tabular data? In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2022. v. 35, p. 507–520. 48
- GUPTA, A. K. et al. A framework for a consistent and reproducible evaluation of manual review for patient matching algorithms. **Journal of the American Medical Informatics Association**, v. 29, n. 12, p. 2105–2109, 2022. [S8] Proposes a systematic framework for creating and evaluating manually-reviewed gold standard record linkage datasets. 37, 47
- GUPTA, A. K. et al. Manual evaluation of record linkage algorithm performance in four real-world datasets. **Applied Clinical Informatics**, v. 15, n. 3, p. 620–628, 2024. [S8] Evaluates record linkage algorithms against manually-reviewed gold standard across four real-world healthcare datasets. 37, 47
- HALLAL, P. C. et al. SARS-CoV-2 antibody prevalence in Brazil: Results from two successive nationwide serological household surveys. **The Lancet Global Health**, v. 8, n. 10, p. e1390–e1398, 2020. 12, 43
- HAND, D. J.; CHRISTEN, P. A note on using the F-measure for evaluating record linkage algorithms. **Statistics and Computing**, v. 28, p. 539–547, 2018. 6, 8, 21, 22
- HANEEF, R. et al. Methodological guidelines to estimate population-based health indicators using linked data and/or machine learning techniques. **Archives of Public Health**, v. 80, n. 1, p. 9, 2022. 11, 23
- HARRON, K. L. et al. A guide to evaluating linkage quality for the analysis of linked data. **International Journal of Epidemiology**, v. 46, n. 5, p. 1699–1710, 2017. [S1] Seminal guide providing a systematic framework for evaluating record linkage quality, including metrics for false matches and missed matches. 25, 41, 45, 47
- HASSANI, H. et al. An oversampling-undersampling strategy for large-scale data linkage. **Frontiers in Big Data**, v. 8, p. 1542483, 2025. 2, 7, 10, 21
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. New York: Springer, 2009. 6, 20, 21, 22, 23
- HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009. 2, 7, 10, 18, 22, 25, 26, 28
- HORNBROOK, M. C.; HURTADO, A. V.; JOHNSON, R. E. Health care episodes: Definition, measurement and use. **Medical Care Review**, v. 23, n. 2, p. 171–187, 1985. 44

- JARDIM, M. A. P. **Comparador de Registros**. 2024. Repositório GitHub. Licença GPL-3.0. Reimplementação em Python do comparador de registros probabilístico proposto por (LUCENA, 2013). Acesso em: 8 fev. 2026. Disponível em: <<https://github.com/marco-jardim/Comparador-de-Registros>>. 2, 7, 18, 19, 26, 27, 36, 46, 51, 65, 75
- JARO, M. A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. **Journal of the American Statistical Association**, v. 84, n. 406, p. 414–420, 1989. 4
- JIAO, Y. et al. A new hybrid record linkage process to make epidemiological databases interoperable: application to the gemo and genepso studies involving brca1 and brca2 mutation carriers. **BMC Medical Research Methodology**, v. 21, n. 1, p. 155, 2021. 2, 7, 10
- JOHNSON, J. M.; KHOSHGOFTAAR, T. M. Survey on deep learning with class imbalance. **Journal of Big Data**, v. 6, n. 1, p. 27, 2019. [S5] Survey of deep learning approaches for handling class imbalance, with applications across health and medical domains. 25, 35
- JR, K. R. d. C.; COELI, C. M. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilístico. **Cadernos de Saúde Pública**, v. 16, n. 2, p. 439–447, 2000. 1, 2, 4, 5, 7, 9, 10, 11, 12, 16, 19, 20, 51
- KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. In: GUYON, I. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Acesso em: 8 fev. 2026. Disponível em: <[https://proceedings.neurips.cc/paper\\_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf)>. 21
- KESSNER, D. M.; KALK, C. E.; SINGER, J. Assessing health quality — the case for tracers. **New England Journal of Medicine**, v. 288, n. 4, p. 189–194, 1973. 11
- KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling**. 3rd. ed. Indianapolis: Wiley, 2013. 45
- Kimi Team. **Kimi K2.5: Visual Agentic Intelligence — Technical Report of Kimi K2.5**. [S.I.], 2025. Open-source model with 1.04T MoE parameters (32B activated). Disponível em: <<https://huggingface.co/moonshotai/Kimi-K2.5>>. Disponível em: <<https://github.com/MoonshotAI/Kimi-K2.5>>. 56
- LAM, J. et al. Generating synthetic identifiers to support development and evaluation of data linkage methods. **International Journal of Population Data Science**, v. 9, n. 1, p. 2389, 2024. [S8] Proposes generating synthetic identifiers as gold standard for developing and evaluating linkage methods without privacy concerns. 37, 47
- LI, Y. et al. Deep Entity Matching with Pre-Trained Language Models. **Proceedings of the VLDB Endowment**, v. 14, n. 1, p. 50–60, 2020. 48, 65
- LIMA, S. V. M. A. et al. Quality of tuberculosis information systems after record linkage. **Revista Brasileira de Enfermagem**, v. 73, n. suppl 5, p. e20200536, 2020. [S4] Analyzes TB information system quality after deterministic linkage between SINAN and SIM in Sergipe, Brazil, finding 190 new unreported cases. 29, 39, 43, 48
- LUCENA, F. d. **Algoritmos para o relacionamento probabilístico de registros de base de dados em saúde**. Dissertação (Dissertação de Mestrado) — Instituto de Estudos em Saúde

Coletiva, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013. Acesso em: 8 fev. 2026. Disponível em: <<http://objdig.ufrj.br/96/teses/820945.pdf>>. 2, 7, 18, 19, 26, 46, 51, 70, 75

LUKE, D. A.; STAMATAKIS, K. A. Systems science methods in public health: Dynamics, networks, and agents. **Annual Review of Public Health**, v. 33, p. 357–376, 2012. 43

LUNDBERG, S. M. et al. From local explanations to global understanding with explainable AI for trees. **Nature Machine Intelligence**, v. 2, n. 1, p. 56–67, 2020. [S6] Extends SHAP for tree-based models with TreeSHAP algorithm, enabling efficient global and local interpretability in clinical ML applications. 36, 44, 46

LUNDBERG, S. M.; LEE, S.-I. **A Unified Approach to Interpreting Model Predictions**. 2017. Apresentado no NIPS 2017. Acesso em: 8 fev. 2026. Disponível em: <<https://arxiv.org/abs/1705.07874>>. 22, 23, 36, 44

MACINKO, J.; HARRIS, M. J. Brazil's Family Health Strategy — delivering community-based primary care in a universal health system. **New England Journal of Medicine**, v. 372, n. 23, p. 2177–2181, 2015. 4

MAIA, C. M. d. et al. Impact of COVID-19 on tuberculosis indicators in Brazil: A time-series analysis. **Journal of Clinical Tuberculosis and Other Mycobacterial Diseases**, v. 29, p. 100325, 2022. DOI not available — original DOI 10.1016/j.jctube.2022.100325 resolves to a different article. 12, 43

MARKUS, A. F.; KORS, J. A.; RIJNBEEK, P. R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. **Journal of Biomedical Informatics**, v. 113, p. 103655, 2021. [S6] Comprehensive survey on explainable AI for healthcare, covering terminology, design choices, and evaluation strategies for trustworthy clinical ML. 36, 45, 46, 48

MENDES, E. V. As redes de atenção à saúde. **Ciência & Saúde Coletiva**, v. 15, n. 5, p. 2297–2305, 2010. 43

MUDGAL, S. et al. Deep Learning for Entity Matching: A Design Space Exploration. In: **Proceedings of the 2018 International Conference on Management of Data (SIGMOD)**. [S.I.]: ACM, 2018. p. 19–34. 48, 65

NASSEH, D.; STAUSBERG, J. Evaluation of a binary semi-supervised classification technique for probabilistic record linkage. **Methods of Information in Medicine**, v. 55, n. 2, p. 136–143, 2016. 2, 6

NEWCOMBE, H. B. et al. Automatic linkage of vital records. **Science**, v. 130, n. 3381, p. 954–959, 1959. 1, 4

OLIVEIRA, G. P. d. et al. Acurácia do relacionamento probabilístico e determinístico de registros: o caso da tuberculose. **Revista de Saúde Pública**, v. 50, p. 49, 2016. 11, 17, 50

OLIVEIRA, G. P. d. et al. Uso do sistema de informação sobre mortalidade para identificar subnotificação de casos de tuberculose no Brasil. **Revista Brasileira de Epidemiologia**, v. 15, n. 3, p. 468–477, 2012. 2, 4, 10, 12, 16, 17, 21

OLIVEIRA, S. P. d. et al. Early death by tuberculosis as the underlying cause in a state of Southern Brazil: Profile, comorbidities and associated vulnerabilities. **International Journal of Infectious Diseases**, v. 80, n. S, p. S50–S57, 2019. [S4] Investigates early TB deaths in Paraná State using linkage between mortality and TB notification databases, identifying comorbidity patterns. 36, 39

PACHECO, A. G. et al. Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of HIV-infected persons and mortality databases in Brazil. **American Journal of Epidemiology**, v. 168, n. 11, p. 1326–1332, 2008. [S2] Validates a hierarchical deterministic linkage algorithm for linking HIV cohort data with SIM mortality databases in Brazil. 38

PAIM, J. et al. The Brazilian health system: History, advances, and challenges. **The Lancet**, v. 377, n. 9779, p. 1778–1797, 2011. 4, 16

PAIXÃO, E. S. et al. Evaluation of record linkage of two large administrative databases in a middle income country: stillbirths and notifications of dengue during pregnancy in Brazil. **BMC Medical Informatics and Decision Making**, v. 17, n. 1, p. 108, 2017. [S2] Evaluates quality of record linkage between SINASC and SINAN in Brazil, demonstrating methodology for assessing linkage accuracy in LMIC settings. 11, 26, 40

PINHEIRO, R. S.; ANDRADE, V. d. L.; OLIVEIRA, G. P. d. Subnotificação da tuberculose no sistema de informação de agravos de notificação (SINAN): abandono primário de bacilíferos e captação de casos em outras fontes de informação usando linkage probabilístico. **Cadernos de Saúde Pública**, v. 28, n. 8, p. 1559–1568, 2012. 2, 10, 12

PINTO, I. V. et al. Factors associated with death in women with intimate partner violence notification in Brazil. **Ciência & Saúde Coletiva**, v. 26, n. 3, p. 975–985, 2021. [S2] Case-control study using record linkage between SIM and SINAN to investigate factors associated with death in women with intimate partner violence in Brazil. 36

PLATT, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: SMOLA, A. J. et al. (Ed.). **Advances in Large Margin Classifiers**. Cambridge, MA: MIT Press, 1999. p. 61–74. 81

PLSEK, P. E.; GREENHALGH, T. The challenge of complexity in health care. **BMJ**, v. 323, n. 7313, p. 625–628, 2001. 42

RAFAEL, R. de M. R. et al. Accuracy, potential, and limitations of probabilistic record linkage in identifying deaths by gender identity and sexual orientation in the state of Rio de Janeiro, Brazil. **BMC Public Health**, v. 24, n. 1, p. 1475, 2024. [S2] Evaluates accuracy of probabilistic record linkage for identifying deaths by gender identity using Brazilian health databases. 32, 38

RANZANI, O. T. et al. Characterisation of the first 250,000 hospital admissions for COVID-19 in Brazil: a retrospective analysis of nationwide data. **The Lancet Respiratory Medicine**, v. 9, n. 4, p. 407–418, 2021. 12, 43

ROCHA, M. S. et al. Do que morrem os pacientes com tuberculose: causas múltiplas de morte de uma coorte de casos notificados e uma proposta de investigação de causas presumíveis. **Cadernos de Saúde Pública**, v. 31, n. 4, p. 709–721, 2015. 2, 4, 10, 12

- ROCHA, M. S. et al. Uso de linkage entre diferentes bases de dados para qualificação de variáveis do Sinan-TB e a partir de regras de scripting. **Cadernos de Saúde Pública**, v. 35, n. 12, p. e00074318, 2019. 12
- ROUX, A. V. D. Complex systems thinking and current impasses in health disparities research. **American Journal of Public Health**, v. 101, n. 8, p. 1382–1389, 2011. 43
- SADINLE, M. Bayesian estimation of bipartite matchings for record linkage. **Journal of the American Statistical Association**, v. 112, n. 518, p. 600–612, 2017. 82, 85
- SANTOS, M. L. et al. Fatores associados à subnotificação de tuberculose com base no Sinan Aids e Sinan Tuberculose. **Revista Brasileira de Epidemiologia**, v. 21, p. e180019, 2018. [S4] Estimates 29% TB underreporting in Pernambuco using probabilistic linkage between Sinan-TB and Sinan-AIDS with RecLink III. 36
- SANTOS, M. L.; COELI, C. M. et al. Fatores associados à subnotificação de tuberculose a partir do linkage SINAN-AIDS e SINAN-TB. **Revista Brasileira de Epidemiologia**, v. 21, p. e180019, 2018. 16, 39
- SARIYAR, M.; BORG, A. Bagging, bumping, multiview, and active learning for record linkage with empirical results on patient identity data. **Computer Methods and Programs in Biomedicine**, v. 108, n. 3, p. 1160–1169, 2012. 6, 20
- SCHNELL, R.; WEIAND, S. V. Microsimulation of an educational attainment register to predict future record linkage quality. **International Journal of Population Data Science**, v. 8, n. 1, p. 2122, 2023. 3, 9
- SHAW, R. J. et al. Biases arising from linked administrative data for epidemiological research: a conceptual framework from registration to analyses. **European Journal of Epidemiology**, v. 37, n. 12, p. 1215–1224, 2022. [S2] Conceptual framework for understanding biases in linked administrative data, with examples from the 100 Million Brazilian Cohort including SIM and SIH. 30, 38, 39, 48, 50
- SHWARTZ-ZIV, R.; ARMON, A. Tabular data: Deep learning is not all you need. **Information Fusion**, v. 81, p. 84–90, 2022. 48
- SILVA, M. E. M. da. **Linkage de Bases de Dados Identificadas em Saúde: Consentimento, Privacidade e Segurança da Informação**. Tese (Tese de Doutorado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, 2012. Orientadora: Profª Cláudia Medina Coeli. 5
- SOUSA, L. M. O. d.; PINHEIRO, R. S. Óbitos e internações por tuberculose não notificados no município do Rio de Janeiro. **Revista de Saúde Pública**, v. 45, n. 1, p. 31–39, 2011. 2, 4, 10, 12, 21
- STERMAN, J. D. **Business Dynamics: Systems Thinking and Modeling for a Complex World**. Boston: McGraw-Hill, 2000. 43
- TASSINARI, W. et al. Record linkage in public health datasets: a practical experience in a fast in-process analytical database. **Revista Brasileira de Epidemiologia**, v. 28, p. e250053, 2025. [S2] Presents accuracy of a mixed-approach algorithm for linking SIM and SIVEP-Gripe records implemented in DuckDB, demonstrating modern computational approaches. 38

TYAGI, K.; WILLIS, S. J. Accuracy of privacy preserving record linkage for real world data in the United States: a systematic review. **JAMIA Open**, v. 8, n. 1, p. ooaf002, 2025. [S1] Systematic review of privacy-preserving record linkage methods, evaluating accuracy across real-world datasets in the US context. 28, 33

VAYENA, E.; BLASIMME, A.; COHEN, I. G. Machine learning in medicine: Addressing ethical challenges. **PLoS Medicine**, v. 15, n. 11, p. e1002689, 2018. 57

VIACAVA, F. et al. Avaliação de desempenho de sistemas de saúde: um modelo de análise. **Ciência & Saúde Coletiva**, v. 17, n. 4, p. 921–934, 2012. 1, 4, 11

VO, T. H. et al. Extending the Fellegi-Sunter record linkage model for mixed-type data with application to the French national health data system. **Computational Statistics & Data Analysis**, v. 179, p. 107656, 2023. [S7] Extends Fellegi-Sunter for mixed-type comparison data (binary and continuous) using an ECM algorithm, applied to the French national health data system. 30

VO, T. T.; LEE, J. Statistical supervised meta-ensemble algorithm for medical record linkage. **Journal of Biomedical Informatics**, v. 95, p. 103220, 2019. 2, 6, 9, 10

WINKLER, W. E. **String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage**. Washington, DC, 1990. 4, 18

World Health Organization. **Global Tuberculosis Report 2024**. Geneva, 2024. Acesso em: 8 fev. 2026. Disponível em: <<https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2024>>. 2, 10, 12, 50

# 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.

**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.

**Comparador de registros** Ferramenta que calcula escores de similaridade campo a campo entre pares candidatos, produzindo subescores individuais e um escore final agregado. Neste trabalho, foi utilizada uma implementação em Python (JARDIM, 2024) do algoritmo proposto por Lucena (LUCENA, 2013).

**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).

**Framework** Estrutura metodológica para organização reproduzível de componentes, etapas e critérios de decisão em um processo analítico.

**F<sub>1</sub>-Score** Média harmônica entre precisão e sensibilidade, utilizada como métrica-síntese do desempenho de classificadores.

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

**Linkage** Ver *Relacionamento de registros*.

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

**Pipeline** Sequência operacional de etapas analíticas e decisórias, desde a preparação dos dados até a geração de resultados.

**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*).

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

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

**Validação cruzada** Técnica de avaliação que particiona os dados em  $k$  subconjuntos (*folds*), treinando o modelo em  $k-1$  partições e avaliando na restante, repetindo o processo  $k$  vezes para obter estimativas de desempenho com variância reduzida (*cross-validation, CV*).

# Apêndice A

## Material Suplementar do *Framework GZ-CMD*

Este apêndice reúne o material suplementar referente ao *framework GZ-CMD* descrito no Capítulo 8: o pseudocódigo completo do *pipeline* de pós-processamento (Seção A.1), as tabelas expandidas de ablação com resultados por partição (Seção A.2), a formalização matemática das propriedades teóricas (Seção A.3) e os *prompts* utilizados na revisão assistida por LLM (Seção A.4).

### A.1 Pseudocódigo do *Pipeline GZ-CMD*

### A.2 Tabelas Expandidas de Ablação

As Tabelas A.1 a A.4 detalham os resultados do estudo de ablação no corte *comprec*, separados por modo de operação (confirmação e vigilância) e com MACD ativado/desativado. Cada tabela apresenta os valores por semente (42 a 46) e, na ultima linha, a estatística consolidada em média  $\pm$  desvio-padrão amostral. A coluna “LLM usado” indica o número absoluto de pares encaminhados ao módulo de revisão assistida por modelo de linguagem (Kimi K2.5) em cada rodada.

Tabela A.1: Ablacão (MACD ON) no modo de confirmação: resultados por semente e resumo final.

Semente	Precisão	Revocação	F1	$F_\beta$	Cobertura auto.	LLM usado
42	0,960	0,952	0,956	0,958	0,994	107,0
43	0,955	0,959	0,957	0,956	0,994	110,0
44	0,952	0,952	0,952	0,952	0,993	126,0
45	0,962	0,909	0,935	0,951	0,994	113,0
46	0,955	0,918	0,936	0,947	0,994	114,0
Média $\pm$ DP	$0,957 \pm 0,004$	$0,938 \pm 0,023$	$0,947 \pm 0,011$	$0,953 \pm 0,004$	$0,994 \pm 0,000$	$114,0 \pm 7,2$

Figura A.1: Pipeline decisorio do GZ-CMD com calibracao por ancoras, politica de custo e trilha de auditoria

1. **Entrada:** conjunto de pares candidatos  $P$  com escores brutos  $\{s_i\}$ ; parametros  $(C_{FP}, C_{FN}, C_{LLM}, e_{FP}, e_{FN})$ ; conjuntos anchora  $A^+$  e  $A^-$ ; orcamento de revisao  $B$ .
2. **Saida:** decisao  $d_i \in \{\text{MATCH}, \text{NONMATCH}, \text{REVIEW}\}$  para cada par  $i$ , com campos de auditoria ( $\text{razao}_i, \text{justificativa}_i$ ).
3. **Fase 1: Calibracao por ancoras.**
  - (a) Estimar  $(\alpha, \beta)$  impondo  $\mathbb{E}[\sigma(\alpha s + \beta) | s \in A^+] = r^+$ .
  - (b) Estimar  $(\alpha, \beta)$  impondo  $\mathbb{E}[\sigma(\alpha s + \beta) | s \in A^-] = r^-$ .
  - (c) Para cada par  $i$ , calcular  $p_i \leftarrow \sigma(\alpha s_i + \beta)$ .
4. **Fase 2: Regras de guarda deterministicas.**
  - (a) Para cada par  $i$ , se  $(\text{data\_obito}_i - \text{data\_diagnostic}_i) < -180$  dias, definir  $d_i \leftarrow \text{NONMATCH}$ .
  - (b) Atribuir  $\text{razao}_i \leftarrow \text{FILTRO\_TEMPORAL}$  e seguir para o proximo par quando a regra temporal for acionada.
  - (c) Se o par  $i$  competir com outro par  $j$  pelo mesmo registro Sinan-TB, calcular  $\Delta p \leftarrow |p_i - p_j|$ .
  - (d) Se  $\Delta p < 0,05$ , marcar ambos como candidatos a REVIEW e registrar  $\text{razao}_i \leftarrow \text{MARGEM\_ESTREITA}$ .
  - (e) Se  $\Delta p \geq 0,05$ , manter apenas o par com maior probabilidade calibrada e registrar descarte por cardinalidade para o preferido.
5. **Fase 3: Motor de politica de decisao baseado em perda esperada.**
  - (a) Para cada par  $i$  ainda nao resolvido, calcular  $L_M \leftarrow (1 - p_i) \cdot C_{FP}, L_N \leftarrow p_i \cdot C_{FN}$  e  $L_R \leftarrow C_{LLM} + (1 - p_i)e_{FP}C_{FP} + p_i e_{FN}C_{FN}$ .
  - (b) Calcular  $EVR_i \leftarrow \min(L_M, L_N) - L_R$ .
  - (c) Se  $L_M \leq L_N$  e  $L_M \leq L_R$ , definir  $d_i \leftarrow \text{MATCH}$ .
  - (d) Se  $L_N \leq L_M$  e  $L_N \leq L_R$ , definir  $d_i \leftarrow \text{NONMATCH}$ .
  - (e) Nos demais casos, definir  $d_i \leftarrow \text{REVIEW}$  como candidato para triagem orçamentaria.
  - (f) Registrar  $(L_M, L_N, L_R, EVR_i)$ , codigo de razao preliminar e vetor de evidencias para rastreabilidade.
6. **Fase 4: Triagem sob restricao de orçamento.**
  - (a) Ordenar candidatos a REVIEW por  $EVR_i$  em ordem decrescente.
  - (b) Selecionar os  $k$  primeiros candidatos com  $k \leq B$  para revisao assistida.
  - (c) Para cada candidato excedente com posicao  $> B$ , substituir por decisao automatica  $\arg \min(L_M, L_N)$  e registrar  $\text{ORCAMENTO\_EXCEDIDO}$ .
7. **Fase 5: Revisao assistida por LLM em protocolo dual.**
  - (a) Para cada par com  $d_i = \text{REVIEW}$ , montar dossie JSON com subescores,  $p_{cal}$  e codigos de guarda.
  - (b) Submeter o dossie ao Agente Objetivo e ao Agente Cetico, de forma independente.
  - (c) Se houver consenso entre os dois agentes, adotar a decisao consensual como decisao final.
  - (d) Se houver discordancia, encaminhar o caso ao Agente Arbitro para deliberacao final.
  - (e) Registrar codigo de razao clinico-operacional, justificativa textual, identificador de versao do modelo, carimbo temporal e assinatura do lote.
8. **Fase 6: Consolidacao e retorno auditavel.**
  - (a) Retornar  $\{(d_i, \text{razao}_i, \text{justificativa}_i)\}$  para todo par  $i$  de  $P$ .

Tabela A.2: Ablacão (MACD ON) no modo de vigilância: resultados por semente e resumo final.

Semente	Precisão	Revocação	F1	$F_\beta$	Cobertura auto.	LLM usado
42	0,948	0,976	0,962	0,971	0,994	105,0
43	0,940	0,972	0,956	0,965	0,994	112,0
44	0,940	0,963	0,952	0,958	0,993	125,0
45	0,954	0,953	0,953	0,953	0,994	103,0
46	0,951	0,944	0,947	0,945	0,994	102,0
Média ± DP	0,947 ± 0,006	0,962 ± 0,013	0,954 ± 0,005	0,959 ± 0,010	0,994 ± 0,000	109,4 ± 9,6

Tabela A.3: Ablacão (MACD OFF) no modo de confirmação: resultados por semente e resumo final.

Semente	Precisão	Revocação	F1	$F_\beta$	Cobertura auto.	LLM usado
42	0,943	0,955	0,949	0,946	0,992	144,0
43	0,941	0,947	0,944	0,942	0,992	153,0
44	0,955	0,950	0,952	0,954	0,993	123,0
45	0,950	0,911	0,930	0,942	0,992	143,0
46	0,953	0,925	0,939	0,948	0,993	129,0
Média ± DP	0,949 ± 0,006	0,937 ± 0,019	0,943 ± 0,009	0,946 ± 0,005	0,992 ± 0,001	138,4 ± 12,2

Tabela A.4: Ablacão (MACD OFF) no modo de vigilância: resultados por semente e resumo final.

Semente	Precisão	Revocação	F1	$F_\beta$	Cobertura auto.	LLM usado
42	0,927	0,976	0,951	0,966	0,993	135,0
43	0,922	0,974	0,947	0,963	0,992	141,0
44	0,944	0,960	0,952	0,956	0,993	124,0
45	0,939	0,967	0,953	0,961	0,993	125,0
46	0,941	0,944	0,942	0,943	0,994	115,0
Média ± DP	0,934 ± 0,009	0,964 ± 0,013	0,949 ± 0,004	0,958 ± 0,009	0,993 ± 0,001	128,0 ± 10,1

## A.3 Formalização Matemática

### A.3.1 Calibração Probabilística por Âncoras

A pontuação bruta  $s \in \mathbb{R}$  gerada por ferramentas de *record linkage* (como o somatório de pesos no OpenRecLink ou a saída de decisão de um classificador *ensemble*) não constitui, inherentemente, uma probabilidade bem calibrada. Para permitir decisões baseadas em risco, faz-se necessária a transformação desse escore em uma probabilidade  $p = P(y = 1 | s)$ , onde  $y = 1$  denota um par verdadeiro (*match*).

Adota-se o método de *Platt Scaling* (PLATT, 1999), que ajusta uma função sigmoide aos escores:

$$p_{cal}(s) = \frac{1}{1 + \exp(-(\alpha s + \beta))}$$

No entanto, em cenários de vigilância onde não se dispõe de dados rotulados (*labels*) em tempo de execução para treinar essa calibração (o problema *unsupervised*), propõe-se uma adaptação baseada em **conjuntos âncora** (*anchor sets*). Definem-se dois subconjuntos de alta confiança baseados em conhecimento de domínio e regras determinísticas estritas:

1. **Âncoras Positivas ( $A^+$ )**: Pares onde  $s \geq \tau_{high}$  (por exemplo, nota  $\geq 9,0$  e identidade perfeita de nomes e mães), assumindo-se uma probabilidade média alvo  $r^+ \approx 1$ .
2. **Âncoras Negativas ( $A^-$ )**: Pares onde  $s \leq \tau_{low}$  (por exemplo, nota  $< 5,0$ ), assumindo-se uma probabilidade média alvo  $r^- \approx 0$ .

Os parâmetros de calibração  $\alpha$  e  $\beta$  são obtidos resolvendo-se o sistema de equações que iguala a esperança da probabilidade predita nos conjuntos âncora às suas taxas reais estimadas (*priors*):

$$\begin{cases} \mathbb{E}_{s \in A^+} [\sigma(\alpha s + \beta)] = r^+ \\ \mathbb{E}_{s \in A^-} [\sigma(\alpha s + \beta)] = r^- \end{cases}$$

**Nota sobre regularização.** O sistema de duas equações para dois parâmetros pode tornar-se numericamente instável quando os conjuntos âncora apresentam baixa variância nos escores  $s$  (por exemplo, quando  $A^+$  concentra-se em uma faixa estreita próxima a 10,0). Para mitigar esse risco, recomendam-se três estratégias complementares: (i) impor limites de magnitude nos parâmetros ( $|\alpha| \leq \alpha_{max}$ ,  $|\beta| \leq \beta_{max}$ ), equivalente a uma regularização L-infinito; (ii) utilizar um termo de penalização quadrática (regularização L2, análoga ao *weight decay*) na função objetivo, transformando o sistema em um problema de otimização  $\min_{\alpha, \beta} \sum_{A^+, A^-} \mathcal{L}_{CE} + \lambda(\alpha^2 + \beta^2)$ , onde  $\mathcal{L}_{CE}$  é a entropia cruzada contra os alvos  $r^+$  e  $r^-$ ; ou (iii) garantir que os conjuntos âncora tenham tamanho mínimo ( $|A^+| \geq 30$ ,  $|A^-| \geq 30$ ) para estabilidade estatística. A escolha entre essas estratégias depende do contexto operacional; em bases com elevada proporção de

escores extremos, a estratégia (i) pode ser suficiente, ao passo que bases com distribuições mais concentradas podem requerer a combinação de (i) e (ii).

**Proposição 1** (Viés de contaminação nos conjuntos âncora). *A qualidade da calibração depende criticamente da pureza dos conjuntos âncora. Denote-se por  $\varepsilon^+$  a taxa de contaminação em  $A^+$  (fração de não-pares erroneamente incluídos) e por  $\varepsilon^-$  a taxa de contaminação em  $A^-$  (fração de pares verdadeiros erroneamente incluídos). As probabilidades-alvo efetivas tornam-se  $r_{eff}^+ = r^+ - \varepsilon^+$  e  $r_{eff}^- = r^- + \varepsilon^-$ , de modo que os parâmetros  $\alpha$  e  $\beta$  estimados refletem essas taxas contaminadas. O viés resultante na curva de calibração é proporcional a  $\max(\varepsilon^+, \varepsilon^-)$ : para  $\varepsilon < 0,01$  (pureza superior a 99%), a distorção na região intermediária da sigmoide (onde se concentram as decisões da zona cinzenta) é da ordem de  $O(\varepsilon)$  e pode ser considerada negligível frente à incerteza inerente do modelo discriminativo. A validação empírica dos conjuntos âncora contra o padrão-ouro constitui, portanto, uma etapa obrigatória do protocolo de calibração.*

Essa abordagem permite recalibrar periodicamente o modelo discriminativo sem intervenção humana manual, garantindo que a probabilidade  $p_{cal}$  reflita a taxa de acerto local, mesmo sob deriva (*drift*) nas distribuições dos escores brutos.

### A.3.2 Função de Perda e Decisão em Três Vias

O framework GZ-CMD estende a decisão binária tradicional para uma lógica de custo-benefício assimétrico. Definem-se três ações possíveis  $\mathcal{A} = \{a_M, a_N, a_R\}$ , correspondendo a vincular (*Match*), não vincular (*Non-match*) e revisar (*Review*), respectivamente.

Sejam  $C_{FP}$  e  $C_{FN}$  os custos unitários de um Falso Positivo e de um Falso Negativo. Em vigilância epidemiológica, frequentemente  $C_{FN} > C_{FP}$  (o custo de perder um óbito é maior que o de investigar um falso alarme) (SADINLE, 2017). A perda esperada  $\mathcal{L}$  para cada ação, dada a probabilidade calibrada  $p$ , é definida como:

$$\begin{aligned}\mathcal{L}(a_M | p) &= (1 - p) \cdot C_{FP} \\ \mathcal{L}(a_N | p) &= p \cdot C_{FN}\end{aligned}$$

Para a ação de revisão (realizada por humano ou agente LLM), introduz-se um custo fixo operacional  $C_{LLM}$  e modela-se a falibilidade do revisor através de suas taxas de erro  $e_{FP}$  (probabilidade de o revisor aceitar um não-par) e  $e_{FN}$  (probabilidade de o revisor rejeitar um par verdadeiro). A perda esperada da revisão é:

**Proposição 2** (Afinidade da perda de revisão).

$$\mathcal{L}(a_R | p) = C_{LLM} + \underbrace{(1 - p) \cdot e_{FP} \cdot C_{FP}}_{Risco residual de FP} + \underbrace{p \cdot e_{FN} \cdot C_{FN}}_{Risco residual de FN}$$

A regra de decisão ótima  $\delta^*(p)$  é aquela que minimiza a perda esperada:

$$\delta^*(p) = \operatorname{argmin}_{a \in \{a_M, a_N, a_R\}} \{\mathcal{L}(a_M | p), \mathcal{L}(a_N | p), \mathcal{L}(a_R | p)\}$$

Esta formulação endogeniza o *trade-off* entre precisão e sensibilidade (controlado pela razão  $C_{FP}/C_{FN}$ ) e a viabilidade econômica da revisão (controlada por  $C_{LLM}$ ), permitindo que a fronteira de decisão se adapte ao contexto operacional sem modificação algorítmica.

### A.3.3 Valor Esperado da Revisão (EVR)

Para operacionalizar a triagem sob restrição orçamentária, define-se o Valor Esperado da Revisão (*Expected Value of Review*, EVR). O EVR quantifica a redução de perda obtida ao se optar pela revisão em vez da melhor decisão automática disponível:

**Definição 1** (Valor Esperado da Revisão (EVR)).

$$EVR(p) = \min(\mathcal{L}(a_M | p), \mathcal{L}(a_N | p)) - \mathcal{L}(a_R | p)$$

Um par deve ser encaminhado para revisão se, e somente se,  $EVR(p) > 0$ . Intuitivamente, isso ocorre quando a incerteza do modelo (proximidade de  $p$  em relação ao limiar de decisão) é alta o suficiente para que o custo da informação adicional ( $C_{LLM}$ ) seja compensado pela mitigação do risco de erro.

Em cenários com orçamento de revisão fixo  $B$  (número máximo de pares revisáveis), os candidatos à revisão são ordenados por  $EVR(p)$  decrescente, selecionando-se os  $k$  primeiros tal que  $\sum_{i=1}^k \mathbb{I}(a_i = a_R) \leq B$ .

### Propriedades geométricas do EVR

**Proposição 3** (Concavidade). *A função  $EVR(p)$  é côncava em  $p \in (0, 1)$  e atinge seu valor máximo em  $p^* = C_{FP}/(C_{FP} + C_{FN})$ , que coincide com o ponto de máxima incerteza de decisão entre as ações automáticas. Nos extremos ( $p \rightarrow 0$  e  $p \rightarrow 1$ ),  $EVR(p) < 0$ , o que garante que a região de revisão  $\{p : EVR(p) > 0\}$  é um intervalo conexo.*

*Proof.* O termo  $\min(\mathcal{L}(a_M | p), \mathcal{L}(a_N | p))$  constitui o mínimo pontual de duas funções lineares em  $p$ :  $\mathcal{L}(a_M | p) = (1-p) \cdot C_{FP}$  (decrescente) e  $\mathcal{L}(a_N | p) = p \cdot C_{FN}$  (crescente). O mínimo pontual de funções afins é côncavo (propriedade elementar da teoria de otimização convexa (BOYD; VANDENBERGHE, 2004)). Como  $\mathcal{L}(a_R | p)$  é afim em  $p$  (portanto simultaneamente convexa e côncava), a diferença  $EVR(p) = \min(\mathcal{L}(a_M | p), \mathcal{L}(a_N | p)) - \mathcal{L}(a_R | p)$  preserva a concavidade.

Avaliando nos extremos:

$$EVR(0) = \min(C_{FP}, 0) - (C_{LLM} + e_{FP} \cdot C_{FP}) = -C_{LLM} - e_{FP} \cdot C_{FP} < 0$$

$$EVR(1) = \min(0, C_{FN}) - (C_{LLM} + e_{FN} \cdot C_{FN}) = -C_{LLM} - e_{FN} \cdot C_{FN} < 0$$

Pela concavidade,  $EVR(p) > 0$  apenas em um subintervalo aberto de  $(0, 1)$ , e esse subintervalo é necessariamente conexo. O máximo é atingido em  $p^* = C_{FP}/(C_{FP} + C_{FN})$ , onde as duas funções de perda automáticas se cruzam e a incerteza de classificação é máxima.  $\square$

**Corolário 1** (Limiar crítico de custo de revisão). *A região de revisão é não vazia se, e somente se,  $EVR(p^*) > 0$ . Calculando-se  $EVR(p^*)$  explicitamente:*

$$EVR(p^*) = \frac{C_{FP} \cdot C_{FN}}{C_{FP} + C_{FN}} \cdot (1 - e_{FP} - e_{FN}) - C_{LLM}$$

Portanto, existe um limiar crítico de custo de revisão:

$$C_{LLM}^* = \frac{C_{FP} \cdot C_{FN}}{C_{FP} + C_{FN}} \cdot (1 - e_{FP} - e_{FN})$$

abaixo do qual a revisão é economicamente justificável para ao menos alguns pares. Quando  $C_{LLM} \geq C_{LLM}^*$ , o custo do revisor excede o benefício potencial e toda decisão é automática. A expressão  $C_{FP} \cdot C_{FN}/(C_{FP} + C_{FN})$  corresponde à metade harmônica dos custos de erro, ponderada pela acurácia líquida do revisor  $(1 - e_{FP} - e_{FN})$ , o que confere interpretabilidade operacional ao limiar: revisores com maiores taxas de erro residual reduzem o orçamento justificável para revisão.

**Observação.** No caso particular de revisor perfeito ( $e_{FP} = e_{FN} = 0$ ), tem-se  $C_{LLM}^* = C_{FP} \cdot C_{FN}/(C_{FP} + C_{FN})$ . Para o modo Vigilância com  $C_{FN}/C_{FP} = 5$  e  $C_{FP} = 1$ , o limiar resulta em  $C_{LLM}^* = 5/6 \approx 0,83$ , indicando que a revisão é justificável desde que seu custo unitário seja inferior a 83% do custo do falso positivo.

### A.3.4 Relaxamento das Hipóteses de Fellegi–Sunter

O modelo clássico de Fellegi e Sunter (FELLEGI; SUNTER, 1969) fundamenta-se na razão de verossimilhança  $R = P(\gamma | M)/P(\gamma | U)$  e na determinação de dois limiares  $T_\lambda$  e  $T_\mu$  que otimizam o erro para uma taxa fixa de revisão clerical. O GZ-CMD pode ser formalmente compreendido como uma generalização que relaxa duas hipóteses centrais desse modelo:

1. **Relaxamento da Independência Condicional:** Fellegi e Sunter assumem frequentemente que os campos de comparação são condicionalmente independentes,  $P(\gamma | M) =$

$\prod P(\gamma_i \mid M)$ . O GZ-CMD, ao empregar um modelo discriminativo calibrado  $P(y \mid x)$  (como *Random Forest* ou *Gradient Boosting*), captura interações não lineares entre atributos (por exemplo, concordância de nome condicionada à frequência do sobrenome), sem assumir independência.

2. **Substituição de Limiares por Custos:** Enquanto Fellegi–Sunter define limiares baseados em taxas de erro toleráveis (limitadas), o GZ-CMD define a região de indecisão (zona cinzenta) com base na utilidade econômica da decisão, permitindo que a fronteira se desloque conforme as prioridades epidemiológicas.

#### Esboço de prova: convergência para a decisão por máxima verossimilhança

**Proposição 4.** *No caso especial em que (a) os custos são simétricos ( $C_{FP} = C_{FN} = C$ ), (b) o revisor é perfeito ( $e_{FP} = e_{FN} = 0$ ), e (c) o custo de revisão tende a zero ( $C_{LLM} \rightarrow 0$ ), a decisão por perda esperada mínima  $\delta^*(p)$  converge para a decisão por máxima verossimilhança (limiar em  $p = 0,5$ ), e a zona de revisão colapsa ao conjunto vazio.*

*Esboço da prova.* Sob as hipóteses (a) a (c), as funções de perda simplificam-se para:

$$\begin{aligned}\mathcal{L}(a_M \mid p) &= (1 - p) \cdot C \\ \mathcal{L}(a_N \mid p) &= p \cdot C \\ \mathcal{L}(a_R \mid p) &= C_{LLM} + 0 = C_{LLM}\end{aligned}$$

**Passo 1 (Eliminação da revisão).** Para que a revisão seja ótima, necessita-se  $\mathcal{L}(a_R \mid p) < \min(\mathcal{L}(a_M \mid p), \mathcal{L}(a_N \mid p))$ , ou seja:

$$C_{LLM} < \min((1 - p) \cdot C, p \cdot C) = C \cdot \min(1 - p, p)$$

O lado direito atinge seu máximo em  $p = 0,5$ , onde  $\min(1 - p, p) = 0,5$ , resultando em  $C_{LLM} < 0,5 \cdot C$ . Portanto, quando  $C_{LLM} \rightarrow 0$ , a revisão é ótima para um conjunto de valores de  $p$  em torno de 0,5 cuja largura é proporcional a  $C_{LLM}/C$ . No limite  $C_{LLM} \rightarrow 0$ , a revisão é "gratuita" e tecnicamente ótima em  $p = 0,5$  exatamente, mas como a perda da revisão também tende a zero (revisor perfeito, custo zero), a decisão é indiferente nesse ponto. Na prática, a zona de revisão colapsa a um conjunto de medida nula.

**Passo 2 (Decisão binária residual).** Eliminada a revisão, a decisão reduz-se a:

$$\delta^*(p) = \begin{cases} a_M & \text{se } (1 - p) \cdot C \leq p \cdot C \iff p \geq 0,5 \\ a_N & \text{se } p < 0,5 \end{cases}$$

Isso equivale à regra de decisão por máxima probabilidade *a posteriori* (MAP), que coincide com a decisão por máxima verossimilhança quando as classes têm *prior* uniforme (ou, equivalentemente, quando a calibração já incorpora o *prior*).  $\square$

**Observação.** A convergência é suave: à medida que  $C_{LLM}/C$  decresce, a zona de revisão estreita-se monotonicamente em torno de  $p = 0,5$ , conforme demonstrado pela Proposição 3 e pelo Corolário 1. Para  $C_{FP} \neq C_{FN}$ , o limiar de decisão desloca-se para  $p^* = C_{FP}/(C_{FP} + C_{FN})$ , recuperando a regra de decisão bayesiana com custos assimétricos (SADINLE, 2017).

### Recuperação dos limiares de Fellegi–Sunter

A conexão formal entre o GZ-CMD e o modelo clássico pode ser estabelecida de forma mais precisa através da seguinte proposição, que demonstra que os limiares de decisão do framework generalizam os limiares  $T_\lambda$  e  $T_\mu$  de Fellegi e Sunter quando expressos no espaço da razão de verossimilhança.

**Proposição 5** (Equivalência sob calibração perfeita). *Suponha que o modelo discriminativo produz probabilidades perfeitamente calibradas, de modo que  $p = P(y = 1 | s)$  coincida com a probabilidade a posteriori derivada da razão de verossimilhança  $R(\gamma) = P(\gamma | M)/P(\gamma | U)$  via regra de Bayes:*

$$p = \frac{\pi \cdot R}{\pi \cdot R + (1 - \pi)}$$

onde  $\pi = P(M)$  denota a prevalência de pares verdadeiros na população de candidatos. Suponha adicionalmente que o revisor é perfeito ( $e_{FP} = e_{FN} = 0$ ). Então, a política de decisão ótima do GZ-CMD, quando reescrita em termos de  $R$ , define dois limiares:

$$T_\lambda^{GZ} = \frac{1 - \pi}{\pi} \cdot \frac{C_{FP} - C_{LLM}}{C_{LLM}} \quad (\text{limiar superior: acima, declara vínculo})$$

$$T_\mu^{GZ} = \frac{1 - \pi}{\pi} \cdot \frac{C_{LLM}}{C_{FN} - C_{LLM}} \quad (\text{limiar inferior: abaixo, declara não-vínculo})$$

com uma zona de revisão definida por  $T_\mu^{GZ} \leq R \leq T_\lambda^{GZ}$ .

*Proof.* Com  $e_{FP} = e_{FN} = 0$ , as condições de decisão em termos de  $p$  são:

- **Vínculo ( $a_M$ ):**  $\mathcal{L}(a_M | p) \leq \mathcal{L}(a_R | p)$  e  $\mathcal{L}(a_M | p) \leq \mathcal{L}(a_N | p)$ . A primeira condição requer  $(1 - p) \cdot C_{FP} \leq C_{LLM}$ , ou seja,  $p \geq 1 - C_{LLM}/C_{FP}$ . A segunda requer  $p \geq C_{FP}/(C_{FP} + C_{FN})$ . Para que a zona de revisão exista (Corolário 3.1), necessita-se  $C_{LLM} < C_{FP} \cdot C_{FN}/(C_{FP} + C_{FN})$ , o que implica  $1 - C_{LLM}/C_{FP} > C_{FP}/(C_{FP} + C_{FN})$ . Logo, a condição vinculante é  $p \geq 1 - C_{LLM}/C_{FP}$ .
- **Não-vínculo ( $a_N$ ):** Analogamente, a condição vinculante é  $p \leq C_{LLM}/C_{FN}$ .
- **Revisão ( $a_R$ ):** Região residual,  $C_{LLM}/C_{FN} < p < 1 - C_{LLM}/C_{FP}$ .

Aplicando a transformação  $p = \pi R / (\pi R + (1 - \pi))$ , a condição  $p \geq t$  equivale a  $R \geq \frac{t}{1-t} \cdot \frac{1-\pi}{\pi}$ . Substituindo:

$$T_\lambda^{GZ} = \frac{1 - C_{LLM}/C_{FP}}{C_{LLM}/C_{FP}} \cdot \frac{1 - \pi}{\pi} = \frac{C_{FP} - C_{LLM}}{C_{LLM}} \cdot \frac{1 - \pi}{\pi}$$

$$T_\mu^{GZ} = \frac{C_{LLM}/C_{FN}}{1 - C_{LLM}/C_{FN}} \cdot \frac{1 - \pi}{\pi} = \frac{C_{LLM}}{C_{FN} - C_{LLM}} \cdot \frac{1 - \pi}{\pi}$$

□

**Corolário 2** (Caso limite sem revisão). *Quando  $C_{LLM} \rightarrow \infty$  (revisão proibitivamente cara),  $T_\lambda^{GZ} \rightarrow 0$  e  $T_\mu^{GZ} \rightarrow \infty$ , de modo que  $T_\mu^{GZ} > T_\lambda^{GZ}$  e a zona de revisão é vazia. Nesse regime, a política reduz-se a uma decisão binária no limiar  $p^* = C_{FP}/(C_{FP} + C_{FN})$ , correspondendo a  $R^* = \frac{C_{FP}}{C_{FN}} \cdot \frac{1-\pi}{\pi}$ , que coincide com o limiar ótimo de Bayes para classificação binária com custos assimétricos.*

**Corolário 3** (Recuperação de Fellegi–Sunter clássico). *No modelo original de Fellegi e Sunter, os limiares  $T_\lambda$  e  $T_\mu$  são determinados de modo a satisfazer limites superiores para as taxas de falso positivo ( $\alpha$ ) e falso negativo ( $\beta$ ). A formulação baseada em custos do GZ-CMD subsume essa abordagem: fixados  $C_{FP}$ ,  $C_{FN}$  e  $C_{LLM}$ , os limiares  $T_\lambda^{GZ}$  e  $T_\mu^{GZ}$  determinam implicitamente taxas de erro  $\alpha(T_\lambda^{GZ})$  e  $\beta(T_\mu^{GZ})$  que são funções monotônicas dos custos. Inversamente, dados limites de erro desejados  $\alpha_0$  e  $\beta_0$ , pode-se encontrar parâmetros de custo ( $C_{FP}$ ,  $C_{FN}$ ,  $C_{LLM}$ ) que recuperam exatamente os limiares de Fellegi–Sunter correspondentes. Desta forma, o modelo clássico constitui um caso particular da formulação de custos, onde os custos estão implicitamente definidos pelas taxas de erro toleradas.*

**Observação.** A proposição requer calibração perfeita ( $p = P(y = 1 | s)$ ), condição que na prática é apenas aproximada. A calibração por âncoras (Seção 1) visa reduzir o viés de calibração, mas a qualidade da aproximação depende criticamente da pureza dos conjuntos âncora ( $A^+$ ,  $A^-$ ) e da adequação do modelo sigmoide à verdadeira curva de calibração. Uma extensão natural consistiria em modelar a incerteza de calibração como uma distribuição sobre  $(\alpha, \beta)$  e propagar essa incerteza para as fronteiras de decisão, resultando em limiares robustos (minimax); essa extensão fica como trabalho futuro.

Desta forma, a estrutura de triagem em três vias é preservada na formulação geral, mas a fronteira de decisão torna-se dinâmica, sensível tanto à calibração local do modelo quanto às prioridades epidemiológicas (vigilância *versus* confirmação), o que confere ao *framework* GZ-CMD maior flexibilidade operacional em relação ao paradigma clássico de Fellegi–Sunter.

## A.4 Prompts para Revisão Assistida por LLM

As instruções (*prompts*) abaixo correspondem aos artefatos efetivamente utilizados no protocolo de revisão clerical com dois agentes e árbitro, conforme implementação do módulo

gzcmd/llm\_review.py e do arquivo-base gzcmd/gzcmd\_v3\_llm\_prompt.md.

## Agente Sensibilidade

[SYSTEM - base v3.1, extraido de gzcmd\_v3\_llm\_prompt.md]

Voce e um revisor clerical especialista em record linkage probabilistico. Sua tarefa: decidir se um par candidato refere-se a mesma pessoa (MATCH), a pessoas diferentes (NONMATCH) ou se a evidencia e genuinamente inconclusiva (UNSURE).

Restricoes obrigatorias:

- 1) Use apenas a informacao contida no dossie JSON recebido.
- 2) Nao invente informacoes ausentes; dados faltantes sao neutros.
- 3) Nao gere PII; se aparecer no dossie, nao repita.
- 4) Retorne somente JSON valido.
- 5) Seja decisivo: UNSURE so para incerteza genuina.

[SYSTEM - instrucao de papel, adicionada em llm\_review.py]

Voce e o Agent-A. Analise o dossie de forma direta e objetiva.

Priorize a evidencia mais forte.

[USER - template, extraido de gzcmd\_v3\_llm\_prompt.md]

Voce recebera um JSON chamado dossier. Avalie as evidencias usando a politica de decisao adequada a banda do par. Devolva um JSON conforme o schema de saida (pair\_id, decision, confidence, reason\_codes, evidence\_summary, quality\_flags).

```
'''json
{dossier}
'''
```

## Agente Especificidade

[SYSTEM - base v3.1, extraido de gzcmd\_v3\_llm\_prompt.md]

Voce e um revisor clerical especialista em record linkage probabilistico. Sua tarefa: decidir se um par candidato refere-se a mesma pessoa (MATCH), a pessoas diferentes (NONMATCH) ou se a evidencia e genuinamente inconclusiva (UNSURE).

Restricoes obrigatorias:

- 1) Use apenas a informacao contida no dossie JSON recebido.
- 2) Nao invente informacoes ausentes; dados faltantes sao neutros.

- 3) Nao gere PII; se aparecer no dossie, nao repita.
- 4) Retorne somente JSON valido.
- 5) Seja decisivo: UNSURE so para incerteza genuina.

[SYSTEM - instrucao de papel, adicionada em llm\_review.py]

Voce e o Agent-B. Analise o dossie com atencao especial a possiveis contradicoes e dados faltantes. Seja cetico mas justo.

[USER - template, extraido de gzcmd\_v3\_llm\_prompt.md]

Voce recebera um JSON chamado dossier. Avalie as evidencias usando a politica de decisao adequada a banda do par. Devolva um JSON conforme o schema de saida (pair\_id, decision, confidence, reason\_codes, evidence\_summary, quality\_flags).

```
'''json
{dossier}
'''
```

## Regra de consenso/arbitragem

[PROTOCOLO - implementado em llm\_review.py]

- 1) Executar Agent-A e Agent-B de forma independente sobre o mesmo dossie.
- 2) Se decision\_A == decision\_B:
  - aceitar consenso;
  - manter a decisao consensual;
  - adotar os campos do agente com maior confidence.
- 3) Se decision\_A != decision\_B:
  - acionar Arbiter com dossie original + resposta Agent-A + resposta Agent-B;
  - usar a decisao final do Arbiter.

[SYSTEM - complemento do Arbiter em llm\_review.py]

Voce e o Arbitro. Dois revisores independentes divergiram.

Analise o dossie original e as duas opinioes abaixo.

De a decisao final. Nao repita PII.

[USER - complemento do Arbiter em llm\_review.py]

```
### Dossie original
```

```
'''json
{dossier}
'''
```

```
### Opiniao Agent-A
```

```
    """json
{response_a}
"""

### Opiniao Agent-B
    """json
{response_b}
"""


```

Analise as divergencias e de sua decisao final no formato JSON.

Em conformidade com a diretriz de sistematização reproduzível adotada no Capítulo 9, recomenda-se registrar, para cada rodada de revisão LLM: identificador completo do modelo (incluindo revisão/versão), `temperature`, `seed` (fixado quando o provedor suportar; caso contrário, explicitar `seed=N/A`), `hash` criptográfico da versão do *prompt* (por exemplo, SHA-256 do arquivo-base e do *prompt* efetivamente concatenado), `commit` Git, data/hora UTC e parâmetros de inferência (por exemplo, `max_tokens` e formato de resposta). Esse metadado deve ser persistido junto aos resultados para auditoria e reprodução do protocolo.