CENTRO DE CIÊNCIAS SOCIAIS APLICADAS E SAÚDE

PROGRAMA DE PÓS-GRADUAÇÃO STRICTO SENSU EM SAÚDE COLETIVA

DOUTORADO EM SAÚDE COLETIVA

MARCIO LUIZ PIVA

ANÁLISE PREDITIVA DAS COBERTURAS VACINAIS DA BCG E TRÍPLICE VIRAL DOS MUNÍCIPIOS DO ESTADO DE SÃO PAULO ENTRE 2016 E 2023 UTILIZANDO MACHINE LEARNING

SANTOS

2026

[Dados Internacionais de Catalogação]
Departamento de Bibliotecas da Universidade Católica de Santos
Viviane Santos da Silva – CRB 8/6746

P693a Piva, Marcio Luiz
Análise preditiva das coberturas vacinais da BCG e Tríplice viral dos municípios do estado de São Paulo entre 2016 e 2023 utilizando Machine Learning / Márcio Luiz Piva; orientadora Lourdes Conceição Martins – 2026.
188 f.

Tese (doutorado)- Universidade Católica de Santos, Programa de Pós-Graduação stricto sensu em Saúde Coletiva, 2026.
Inclui bibliografia

1. Saúde Pública. 2. Cobertura Vacinal. 3. Machine Learning. 4. Modelos Preditivos. I. Martins, Lourdes Conceição. II. Título.

CDU: Ed. 1997 -- 614(043.2)

ANÁLISE PREDITIVA DAS COBERTURAS VACINAIS DA BCG E TRÍPLICE VIRAL DOS MUNÍCIPIOS DO ESTADO DE SÃO PAULO ENTRE 2016 E 2023 UTILIZANDO MACHINE LEARNING

Texto apresentado ao Programa de Pós-Graduação em Saúde Coletiva da Universidade Católica de Santos como requisito para obtenção do título de Doutor em Saúde Coletiva.

Área de Concentração: Saúde, Ambiente e Mudanças Sociais.

Linha de Pesquisa: Epidemiologia Ambiental.

Orientadora: Profa. Dra. Lourdes Conceição Martins

SANTOS

2026

MARCIO LUIZ PIVA

Análise Preditiva das Coberturas Vacinais da BCG e Tríplice Viral dos Munícipios do estado de São Paulo entre 2016 e 2023 utilizando Machine Learning

Tese apresentada ao Programa de Doutorado em Saúde Coletiva da Universidade Católica de Santos como exigência ao título de Doutor em Saúde Coletiva. Área de Concentração: Saúde, Ambiente e Mudanças Sociais.

Aprovado em 08/10/2026.

Banca examinadora

Prof. Dra. Lourdes Conceição Martins
Universidade Católica de Santos – UNISANTOS

Prof. Dr. Alfésio Luis Ferreira Braga
Universidade Católica de Santos – UNISANTOS

Prof. Dr. Daniel Alberto Pamplona
Instituto Tecnológico de Aeronáutica – ITA

Prof. Dr. José João Sumbana
Universidade Eduardo Mondlane (Moçambique) – UEM

Prof. Dr. Ricardo Alves de Olinda
Universidade Estadual da Paraíba – UEPB

DEDICATÓRIA

Dedico esse trabalho de pesquisa à família, que é e sempre será a base para tudo. À minha esposa Cristina, pelo apoio e paciência, aos meus filhos, Arthur e Anna, pela compreensão pelas intermináveis horas de ausência.

AGRADECIMENTOS

Diante do desafio, agradeço primeiramente a Deus pelo amparo, iluminação e fortalecimento necessários para transpor as dificuldades dessa caminhada.

Aos meus pais (*in memoriam*), por proporcionarem as condições e as oportunidades necessárias para que eu chegasse até aqui; sem eles, nada disso seria possível.

À minha orientadora, Profa. Dra. Lourdes Conceição Martins, pelo apoio e, sobretudo, pela confiança depositada para a realização desta pesquisa.

Expresso, finalmente, meus sinceros agradecimentos à Universidade Católica de Santos, pelo suporte institucional e concessão da bolsa de estudos, recurso fundamental para o pleno desenvolvimento desta pesquisa.

RESUMO

Introdução: A cobertura vacinal é considerada um importante indicador de saúde pública, sendo amplamente utilizada em âmbito global. As baixas coberturas vacinais têm provocado sérios problemas à saúde da população brasileira, como o retorno de doenças imunopreveníveis. O estado de São Paulo é composto por 645 municípios que apresentam características socioeconômicas, demográficas, de saúde, educacionais e de cobertura vacinal bastante heterogêneas.

Objetivo: Análise preditiva das coberturas vacinais das vacinas BCG e Tríplice Viral (SCR), utilizando dados secundários dos municípios do estado de São Paulo, referentes ao período de 2016 a 2023, através de métodos de Machine Learning.

Metodologia: Esta pesquisa tratou-se de um Estudo Epidemiológico Ecológico do Tipo Misto focando na análise dados no período de 2016 a 2023 provenientes dos diversos repositórios públicos utilizando Machine Learning, uma subárea da Inteligência Artificial. Os dados foram preparados, agrupados e analisados estatisticamente para a realização de análise preditiva comparativa entre seis modelos de Machine Learning: Support Vector Machine, Logistic Regression, MultiLayer Perceptron (MLP), Random Forest, LightGBM e XGBoost.

Resultados: O modelo XGBoost apresentou o melhor desempenho nas análises comparativas, utilizando-se, sobretudo, a métrica curva ROC-AUC, apresentando valores aproximados de 0,84 para a vacina BCG e 0,70 para a vacina Tríplice Viral (SCR). Por fim, como forma de explicação dos resultados do modelo, foram utilizadas as técnicas SHAP e LIME, permitindo a identificação e listagem dos 20 principais atributos preditores.

Conclusão: A vacinação e suas coberturas, como constatado, não são apenas atos clínicos, mas um fenômeno regido por aspectos multidimensionais, como mostraram os preditores apontados pelo modelo, tais como as dimensões econômicas, educacionais, de saúde e de infraestrutura, que tiveram variáveis destacadas nos achados. Dessa forma, modelos preditivos como os aqui

empregados podem servir como um sistema de alerta antecipado para a gestão de saúde, permitindo identificar municípios em risco do não atingimento das metas, possibilitando propostas de mitigação.

Palavras-chave: Saúde Pública; Cobertura Vacinal; Machine Learning; Modelos Preditivos.

ABSTRACT

Introduction: Vaccination coverage is an important public health indicator and a key measure for monitoring the performance of immunization programs. Persistently low vaccination coverage poses a significant public health challenge in Brazil and has contributed to the resurgence of vaccine-preventable diseases. The state of São Paulo comprises 645 municipalities characterized by substantial socioeconomic, demographic, health, educational, infrastructural, and vaccination coverage heterogeneity.

Objective: To develop predictive models of vaccination coverage for Bacillus Calmette–Guérin (BCG) and Measles, Mumps, and Rubella (MMR) vaccines using secondary data from municipalities in the state of São Paulo, Brazil, for the period from 2016 to 2023, through machine learning methods.

Methods: This study was designed as a mixed ecological study, focusing on the analysis of data in the period from 2016 to 2023 obtained from multiple public repositories using machine learning, a subfield of artificial intelligence. The data were prepared, integrated, and statistically analyzed to perform a comparative predictive analysis using six machine learning models: Support Vector Machine, Logistic Regression, Multilayer Perceptron (MLP), Random Forest, LightGBM, and XGBoost.

Results: XGBoost achieved the highest ROC-AUC among the evaluated models, with values of approximately 0.84 for BCG vaccination coverage and 0.70 for MMR vaccination coverage. The application of SHAP and LIME enabled the interpretation of model predictions and the identification of the 20 most relevant predictor features. These features encompassed variables related to economic, educational, health, demographic, and infrastructure dimensions, highlighting the multidimensional nature of vaccination coverage across municipalities.

Conclusion: As demonstrated by the findings, vaccination and vaccination coverage are not merely clinical phenomena but are influenced by multidimensional factors. The predictors identified by the models included

variables related to economic, educational, health, and infrastructure dimensions. Therefore, predictive models such as those employed in this study may serve as early warning systems for public health management, enabling the identification of municipalities at risk of failing to achieve vaccination coverage targets and supporting the development of timely mitigation strategies.

Keywords: Public Health; Vaccination Coverage; Machine Learning; Predictive Models.

LISTA DE ABREVIATURAS

ANN	-	Artificial Neural Networks ou Redes Neurais Artificiais
ARIMA	-	Autoregressive integrated 11orna11 average ou Média móvel integrada autorregressiva
AUC	-	Area Under the Curve ou Área Sob a Curva
BCG	-	Bacilo de Calmette e Guérin, usada como vacina contra Tuberculose
CRIE	-	Centros de referência para Imunobiológicos especiais
CSV	-	*Comma-Separated Values* ou Valores Separados por Vírgula
CV	-	Cobertura Vacinal
FN	-	Falso negativo
FP	-	Falso positivo
IA	-	Inteligência Artificial
IDHM	-	Índice de Desenvolvimento Humano Municipal
IEPS	-	Instituto de estudos para políticas de saúde
IQR	-	*Inter-Quartile Range* ou Intervalo interquartil
K-NN		K-Nearest Neighbors ou K-vizinhos mais próximos
LIME	-	Local Interpretable Model-Agnostic Explanation ou Explicações locais interpretáveis e agnósticas ao modelo
MAE	-	Mean Absolute error ou Erro médio absoluto
ML	-	*Machine Learning* ou Aprendizado de Máquina
MSE	-	Mean squared error ou Erro quadrático médio
OMS	-	Organização Mundial de Saúde
PCA	-	Principal Component Analysis ou Análise de componentes principais
PIB	-	Produto Interno Bruto
PNI	-	Programa Nacional de Imunizações
R2	-	Coeficiente de determinação
RMSE	-	Root squared MSE ou Raiz quadrado do MSE
RNA	-	Ácido Ribonucleico

ROC	-	Receiver Operating Characteristic ou Característica de Operação do Receptor
SARIMA	-	Seasonal ARIMA
SCR	-	Vacina tríplice viral, contra Sarampo, Caxumba e Rubéola
SHAP	-	Shapley Additive Explanations ou Explicações Aditivas de Shapley
SI	-	Sistema de informação
SINASC	-	Sistema de informação sobre nascidos vivos
SIPNI	-	Sistema de informação do programa nacional de Imunização
SVM	-	Support Vector Machine ou Máquina de vetores de suporte
VN	-	Verdadeiro negativo
VP	-	Verdadeiro positivo
XGBoost	-	Extreme Gradient Boosting

LISTA DE FIGURAS

Figura 1 – História das diversas áreas da Inteligência Artificial 42
Figura 2 – Visão macro das divisões da Inteligência Artificial.....45
Figura 3 – Fluxo geral para análise de dados e aplicação de modelos de Machine Learning.....50
Figura 4 – Divisão conceitual do conjunto de dados 53
Figura 5 – Divisão do conjunto de treinamento em *K-folds* 54
Figura 6 – Exemplo de curva ROC-AUC.....58
Figura 7 – Mapa do estado de São Paulo 68
Figura 8 – Estado de São Paulo: Concentração populacional.....76
Figura 9 – Visão dos municípios do estado de São Paulo segundo porte populacional.....78
Figura 10 – Cobertura Vacinal da BCG: Municípios 102
Figura 11 – Cobertura Vacinal da BCG: Regiões Administrativas 106
Figura 12 – Cobertura vacinal da BCG: Departamentos Regionais de Saúde 109
Figura 13 – Cobertura vacinal da Tríplice Viral (SCR): Municípios 113
Figura 16 – Cobertura vacinal da Tríplice Viral (SCR): Regiões Administrativas 118
Figura 17 – Cobertura vacinal da Tríplice Viral (SCR): Departamentos Regionais de Saúde 121
Figura 18 – Matriz de correlação das 20 variáveis independentes principais para predição da cobertura vacinal da BCG.....136
Figura 19 – Matriz de correlação das 20 variáveis independentes principais para predição da cobertura vacinal da Tríplice Viral (SCR).....144

LISTA DE GRÁFICOS

Gráfico 1 – Distribuição dos municípios por Regiões Administrativas 68
Gráfico 2 – Distribuição dos municípios por Departamentos Regionais de Saúde 72
Gráfico 3 – Distribuição dos municípios por porte 78
Gráfico 4 – Variável IPPI: Índice Paulista da Primeira Infância.....95
Gráfico 5 – Variável IPDM: Índice Paulista de Desenvolvimento Municipal 95
Gráfico 6 – Distribuição dos munícipios por classe de cobertura vacinal da BCG 103
Gráfico 7 – Comparação da distribuição da cobertura vacinal da BCG 105
Gráfico 8 – Evolução da cobertura vacinal da BCG por Regiões Administrativas – Maiores e menores, em relação a quantidade de habitantes 108
Gráfico 9 – Evolução da cobertura vacinal da BCG por Departamento Regional de Saúde – Maiores e menores, em relação a quantidade de habitantes 111
Gráfico 10 – Evolução da cobertura vacinal da BCG por porte municipal 112
Gráfico 11 – Distribuição dos munícipios por classe de cobertura vacinal da Tríplice Viral (SCR) 114
Gráfico 12 – Distribuição da cobertura vacinal da Tríplice Viral (SCR).....117
Gráfico 13 – Evolução da cobertura vacinal da Tríplice Viral (SCR) por Região Administrativa – Maiores e menores, em relação a quantidade de habitantes 120
Gráfico 14 – Evolução da cobertura vacinal da Tríplice Viral (SCR) por Departamento Regional de Saúde – Maiores e menores, em relação a quantidade de habitantes.....123
Gráfico 15 – Evolução da cobertura vacinal da Tríplice Viral (SCR) por porte municipal.....124
Gráfico 16 – Evolução da cobertura vacinal da BCG por ano.....125
Gráfico 17 – Distribuição da cobertura vacinal da BCG por ano.....125
Gráfico 18 – Comparativo geral da cobertura vacinal da BCG com teste de significância estatística 126
Gráfico 19 – Comparação das coberturas vacinais da BCG por ano.....127
Gráfico 20 – Curva ROC-AUC da BCG com o modelo XGBoost.....131

Gráfico 21 – Curva ROC-AUC da BCG: Comparativo de todos os modelos ..132
Gráfico 22 – Explicação dos resultados do modelo *XGBoost* com SHAP *values* resumo.....133
Gráfico 23 – Explicação dos resultados do modelo *XGBoost* com SHAP, LIME e Ganho de Informação 134
Gráfico 24 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo XGBoost 139
Gráfico 25 – Curva ROC-AUC da Tríplice Viral (SCR): Comparativo de todos os modelos 140
Gráfico 26 – Explicação dos resultados do modelo *XGBoost* com SHAP *values* resumo.....141
Gráfico 27 – Explicação dos resultados do modelo *XGBoost* com SHAP, LIME e Ganho de Informação 142
Gráfico 28 – Curva ROC-AUC da BCG com o modelo Support Vector Machine (SVM).....173
Gráfico 29 – Curva ROC-AUC da BCG com o modelo Logistic Regression ...174
Gráfico 30 – Curva ROC-AUC da BCG com o modelo Mutilayer Perceptron (MLP) 174
Gráfico 31 – Curva ROC-AUC da BCG com o modelo Random Forest 175
Gráfico 32 – Curva ROC-AUC da BCG com o modelo LightGBM 175
Gráfico 33 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo Support Vector Machine (SVM) 177
Gráfico 34 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo Logistic Regression.....178
Gráfico 35 – Curva ROC-AUC da Tríplice Viral (SCR) com o Mutilayer Perceptron (MLP).....178
Gráfico 36 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo Random Forest.....179
Gráfico 37 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo LightGBM 179

LISTA DE TABELAS

Tabela 1 – Distribuição populacional dos municípios de São Paulo 77
Tabela 2 – Grupos das variáveis independentes (totais) 93
Tabela 3 – Grupos das variáveis independentes (seleção para predição).....96
Tabela 4 – Análise de cardinalidade das variáveis independentes 96
Tabela 5 – Análise descritiva da cobertura vacinal da BCG por ano 104
Tabela 6 – Análise descritiva da cobertura vacinal da BCG por Regiões Administrativas com média geral do estado de São Paulo 107
Tabela 7 – Análise descritiva da cobertura vacinal da BCG por Departamentos Regionais de Saúde com média do estado de São Paulo 110
Tabela 8 – Cobertura vacinal da Tríplice Viral (SCR) por ano 115
Tabela 9 – Análise descritiva da cobertura vacinal da Tríplice Viral (SCR) por Regiões Administrativas com média geral do estado de São Paulo 119
Tabela 10 – Análise descritiva da cobertura vacinal da Tríplice Viral (SCR) por Departamentos Regionais de Saúde com média do estado de São Paulo 122
Tabela 11 – Análise descritiva comparada da BCG e Tríplice Viral (SCR) por ano 128
Tabela 12 – Comparativo dos métodos de Machine Learning para cobertura vacinal da BCG 130
Tabela 13 – Matriz de Confusão da cobertura vacinal da BCG com o modelo *XGBoost* 130
Tabela 14 – Correlação das métricas de explicação do modelo XGBoost para a cobertura vacinal da BCG 135
Tabela 15 – Comparativo dos métodos de Machine Learning para cobertura vacinal da Tríplice Viral (SCR) 138
Tabela 16 – Matriz de Confusão da cobertura vacinal da Tríplice Viral (SCR) com modelo XGBoost 139
Tabela 17 – Correlação das métricas de explicação do modelo *XGBoost* para a cobertura vacinal da Tríplice Viral (SCR) 142
Tabela 18 – Matriz de Confusão da BCG com o modelo Support Vector Machine (SVM).....172

Tabela 19 – Matriz de Confusão da BCG com o modelo Logistic Regression 172
Tabela 20 – Matriz de Confusão da BCG com o modelo Mutilayer Perceptron (MLP) 172
Tabela 21 – Matriz de Confusão da BCG com o modelo Random Forest 173
Tabela 22 – Matriz de Confusão da BCG com o modelo LightGBM 173
Tabela 23 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo Support Vector Machine (SVM) – 176
Tabela 24 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo Logistic Regression 176
Tabela 25 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo Mutilayer Perceptron (MLP) 176
Tabela 26 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo Random Forest 176
Tabela 27 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo LightGBM 177

LISTA DE QUADROS

Quadro 1 – Tipos de estudos epidemiológicos 26
Quadro 2 – Histórico das vacinas humanas 29
Quadro 3 – Calendário Nacional de Vacinação – 0 a 4 meses 31
Quadro 4 – Calendário Nacional de Vacinação – 5 a 15 meses 32
Quadro 5 – Calendário Nacional de Vacinação – 4 a 14 anos 33
Quadro 6 – Cronologia resumida da Inteligência Artificial 43
Quadro 7 – Matriz de avaliação de dados 51
Quadro 8 – Exemplos dos métodos e algoritmos de Machine Learning 52
Quadro 9 – Métricas utilizadas nos modelos de regressão 55
Quadro 10 – Representação da Matriz de Confusão 56
Quadro 11 – Principais métricas obtidas a partir da Matriz de Confusão 57
Quadro 12 – Revisão de trabalhos da área da saúde e cobertura vacinal com Machine Learning 60
Quadro 13 – Regiões Administrativas e seus municípios 69
Quadro 14 – Departamentos Regionais de Saúde e seus municípios 72
Quadro 15 – Fontes utilizadas para obtenção dos dados 81
Quadro 16 – Esboço de um conjunto de dados 82
Quadro 17 – Relação de ferramentas e bibliotecas de software empregadas na pesquisa 84
Quadro 18 – Dados completos sem informações ou faltantes 87
Quadro 19 – Quatro classes da cobertura vacinal da BCG 91
Quadro 20 – Quatro classes da cobertura vacinal da Tríplice Viral (SCR) 92
Quadro 21 – Modelo de duas classes da cobertura vacinal da BCG 92
Quadro 22 – Modelo de duas classes da cobertura vacinal da Tríplice Viral (SCR) 92
Quadro 23 – Modelos de Machine Learning utilizados 98
Quadro 24 – Coberturas vacinais por classe 99
Quadro 25 – 20 principais variáveis independentes preditoras do modelo XGBoost para a cobertura vacinal da BCG 137

Quadro 26 – 20 principais variáveis independentes preditoras do modelo XGBoost para a cobertura vacinal da Tríplice Viral (SCR) 145

SUMÁRIO

1. INTRODUÇÃO 23
1.1. Os Fundamentos da Epidemiologia 23
1.2. Conceitos Básicos de Imunização 26
1.2.1. Origens das Vacinas 27
1.2.2. Calendário Vacinal no Brasil 30
1.2.2.1. A Vacina BCG 33
1.2.2.2. A Vacina Tríplice Viral (SCR) 34
1.2.3. Cobertura vacinal e controle de doenças 35
1.2.4. O Programa Nacional de Imunizações (PNI) 37
1.2.5. Determinantes da Cobertura Vacinal 38
1.2.6. Consequências da baixa cobertura 40
1.3. Fundamentos da Inteligência Artificial 41
1.3.1. Evolução histórica da Inteligência Artificial 42
1.3.2. Divisões da Inteligência Artificial 44
1.3.3. *Machine Learning*: Conceitos e aplicações 45
1.3.3.1. Tipos de aprendizado e a predição 46
1.3.3.1.1. Aprendizado supervisionado 47
1.3.3.1.2. Aprendizado não supervisionado 49
1.3.3.2. Fluxo de análise de dados com Machine Learning 50
1.3.3.3. Métodos e algoritmos para análise preditiva 51
1.3.3.4. Métodos de treinamento e validação 52
1.3.3.5. Métricas de desempenho 54
1.3.3.6. Interpretação e explicação dos modelos e seus resultados 58
1.3.4. Machine Learning na saúde e cobertura vacinal 59
2. JUSTIFICATIVAS 65

3. OBJETIVOS 66
3.1. Objetivos Gerais 66
3.2. Objetivos Específicos 66
4. MÉTODO 67
4.1. A delimitação geográfica da pesquisa 67
4.2. Obtenção dos dados 79
4.3. Preparação dos dados 82
4.3.1. Ferramentas e bibliotecas de software 82
4.3.2. Processo de carga de dados 86
4.3.2.1. Ajustes e carga de dados faltantes 87
4.4. Análise estatística preliminar 93
4.5. Aplicação de Machine Learning 97
5. RESULTADOS 101
5.1. Análise exploratória dos dados 101
5.1.1. Análise exploratória da BCG 101
5.1.2. Análise exploratória da Tríplice Viral (SCR) 112
5.1.3. Análise exploratória comparada 124
5.2. Análise preditiva das coberturas vacinais 129
5.2.1. Análise preditiva da BCG 129
5.2.2. Análise preditiva da Tríplice Viral (SCR) 137
6. DISCUSSÃO 146
7. CONCLUSÃO 156
8. BIBLIOGRAFIAS 158
9. APÊNDICES 172
9.1. Matriz de Confusão dos demais modelos utilizados para a BCG 172
9.2. Curva ROC-AUC dos demais modelos utilizados para a BCG 173

9.3. Matriz de Confusão dos demais modelos utilizados para a Tríplice Viral (SCR) 176
9.4. Curva ROC-AUC dos modelos utilizados para a Tríplice Viral (SCR) ..177
9.5. Estrutura de ajustes dos hiper parâmetros dos modelos (Hyper Parameter *Tuning*) 180
9.5.1. Parâmetros para Treinamento – BCG e Tríplice Viral (SCR).....180
9.5.2. Parâmetros Ajustados para a BCG 184
9.5.3. Parâmetros Ajustados para a Tríplice Viral (SCR).....186

1. INTRODUÇÃO

Este capítulo concentra-se na contextualização que permeia o objeto desta pesquisa, passando pelos conceitos de Epidemiologia, Imunização e Cobertura Vacinal e, finalmente, adentrando à Inteligência Artificial e uma das suas grandes áreas, Machine Learning (ML) ou Aprendizado de Máquina. Descreve, portanto, todo o referencial teórico necessário para o seu devido embasamento.

Essa pesquisa, além da exploração profunda dos métodos e técnicas da Inteligência Artificial e sua subárea Machine Learning, é, antes de tudo, um estudo epidemiológico observacional descritivo, ainda, com a característica ecológica do tipo misto (híbrido), abordando tanto aspectos geográficos (áreas), quanto a temporalidade dos eventos analisados.

1.1. Os Fundamentos da Epidemiologia

Definir epidemiologia é crucial porque ela é a ciência que estuda a distribuição e os determinantes de problemas de saúde nas populações, fornecendo a base para planejar, implementar e avaliar ações de saúde pública, desde o controle de epidemias até a promoção do bem-estar, identificando quem, quando e onde as doenças ocorrem para melhor direcionar recursos e criar políticas eficazes. Sem essa definição, não seria possível entender padrões de doenças, grupos de risco, nem desenvolver estratégias de prevenção e controle que melhoram a saúde coletiva (Palmeira, 2000).

Como ciência, estuda, portanto, a distribuição das doenças e suas causas em populações humanas. É possível defini-la como a abordagem dos fenômenos da saúde-doença-cuidado com utilização de métodos quantitativos, usando bastante o cálculo matemático e as técnicas estatísticas de amostragem e de análise (Almeida, 2012).

Os estudos epidemiológicos são, ainda, instrumentos para a observação sistemática de fenômenos de interesse na área da saúde, para permitir a análise

e a interpretação dos dados de maneira sistemática, possibilitando a comparação entre grupos. Tais estudos podem ser classificados de acordo com o controle do pesquisador sobre a exposição de interesse, em relação ao eixo temporal de desenvolvimento do estudo e em função da unidade de análise (Braga, 2015).

A epidemiologia se firmou enquanto ciência, baseada em pesquisas e evidências científicas que visam a determinação das condições de saúde da população e a busca sistemática dos agentes etiológicos das doenças ou dos fatores de risco envolvidos no seu aparecimento, através de diferentes tipos de estudos (Gomes, 2015).

Os estudos epidemiológicos, em geral, se dividem entre observacionais (descritivos e analíticos) e experimentais (Zetzsche, 2015). Os estudos observacionais descritivos permitem que a natureza determine o seu curso: o investigador mede, mas não intervém. Um estudo descritivo limita-se a descrever a ocorrência de uma doença em uma população, sendo, frequentemente, o primeiro passo de uma investigação epidemiológica. Já um estudo analítico aborda, com mais profundidade, as relações entre o estado de saúde e as outras variáveis (Bonita, 2010).

Dentre os observacionais descritivos, encontram-se os estudos ecológicos que se limitam a descrever a ocorrência de uma doença em uma população, sendo, frequentemente, o primeiro passo de uma investigação epidemiológica (Braga, 2015).

Duas classes de seres são, portanto, estudadas pela Epidemiologia: agregados humanos (coletivos de homens e mulheres) e indivíduos membros desses agregados. Os agregados de que trata a pesquisa epidemiológica são quase sempre referidos à uma base geográfica e temporal, caracterizando, assim, um estudo ecológico misto. Os desenhos observacionais podem ser transversais ou longitudinais, de acordo com a temporalidade do processo de produção de dados (Almeida, 2012).

Os estudos ecológicos abordam áreas geográficas ou blocos de população bem delimitados, analisando comparativamente variáveis globais, quase sempre por meio da correlação entre indicadores de condições de vida e indicadores de

situação de saúde. Os indicadores de cada área ou bloco constituem-se em médias referentes à sua população total, tomada como um agregado integral (Bonita, 2010).

Os estudos ecológicos podem ser classificados em dois subtipos, a depender da natureza do agregado base de referência para a produção dos dados: as investigações de base territorial e os estudos de agregados institucionais (Palmeira, 2000).

Investigações de base territorial utilizam uma referência geográfica para a definição das suas unidades de informação, em qualquer nível de abrangência (por exemplo, bairros, distritos, municípios, estados, nações e continentes). Os estudos de agregados institucionais tomam organizações coletivas de qualquer natureza como referência para a definição da sua unidade de informação (Almeida, 2012).

Os estudos classificados como agregados-observacionais podem ser também longitudinais (estudos tipo agregado-observacional-longitudinal) a partir de uma analogia com os estudos longitudinais de base individualizada. Nos estudos de séries temporais uma mesma área ou população é investigada em momentos distintos no tempo, classificados pelos manuais de Epidemiologia como um subtipo de estudo ecológico (arquitetura híbrida ou mista, com desenho simultaneamente ecológico e de tendência temporal) (Gomes, 2015).

O Quadro 1 apresenta um detalhamento mais completo e abrangente de todos os tipos de estudos compreendidos pela Epidemiologia.

Quadro 1 – Tipos de estudos epidemiológicos

Tipo de estudo	Nome alternativo	Unidade de estudo
Estudos observacionais		
Estudos descritivos		
Estudos analíticos		
Ecológico	Correlação	População (conjunto)
Transversal	Prevalência	Indivíduo
Casos e controles	Caso-referência	Indivíduo
Coorte	Longitudinal (Follow-up)	Indivíduo
Estudos experimentais	**Estudos de intervenção**	
Ensaio clínico randomizado controlado	Ensaios clínicos	Pacientes
Ensaio clínico randomizado controlado com grupos (clusters)		Grupos
Ensaios de campo		
Ensaios comunitários	Estudos de intervenção na comunidade	Indivíduos saudáveis na comunidade

Fonte: Adaptado de Bonita, 2010

1.2. Conceitos Básicos de Imunização

Ao invadir um organismo, bactérias e vírus atacam o organismo, se multiplicam e atingem diferentes estruturas e órgãos. Esta invasão é chamada de infecção e é isso que causa a doença. Para proteger nossa saúde, as vacinas precisam estimular o sistema imunológico – também chamado de sistema imunitário ou imune – a produzir anticorpos, um tipo de proteína, agentes de defesa que atuam contra os micróbios que provocam doenças infecciosas (Ballalai, 2016).

Segundo a World Health Organization (WHO, 2024), a “Imunização é uma história de sucesso global em saúde, salvando milhões de vidas a cada ano. As vacinas reduzem o risco de se contrair uma doença, trabalhando com o próprio sistema de defesa natural do organismo na criação de proteção. Quando se toma uma vacina, o sistema imunológico responde.”

A imunização é a intervenção em saúde pública com a melhor relação custo-benefício na promoção da saúde e na proteção de indivíduos imunizados, sendo um importante fator de mudança do perfil de morbimortalidade, por doenças

transmissíveis, e capaz de evitar de dois a três milhões de mortes por ano. O impacto da imunização vai além da área da saúde, pois o benefício resultante da vacinação reflete diretamente no potencial econômico e social do país (Lima, 2020).

1.2.1. Origens das Vacinas

Vacinar diz respeito a imunizar um ser vivo, submetendo-o ao contato direto com antígenos de agentes etiológicos com o objetivo de proteger de uma determinada doença (Brasil, 2019c).

Após a aplicação da vacina, contendo vírus ou bactérias vivas atenuadas, mortas, inativadas, fragmentos ou por engenharia genética, esta irá sensibilizar o Sistema Imunológico (SI), induzindo-o a produzir anticorpos específicos e protegendo a população contra doenças ou as respectivas formas graves dela (Brasil, 2019b).

Ao longo da história, surgiram doenças com o potencial de exterminar populações através da sua rápida propagação, da forma que seus agravos logo se disseminaram pelo mundo transformando-se em pandemias, dizimando a vida de grandes populações. Uma forma de reduzir o impacto dessas doenças é buscando estratégias preventivas, o que impulsionou a criação das vacinas, como uma estratégia eficaz contra os agentes causadores de doenças (Oliveira, 2021).

A primeira vacina foi criada pelo médico inglês Edward Jenner em 1796 e protegia contra o vírus da varíola, tornando-se um dos mais importantes avanços da medicina (Oliveira, 2021). Essa virose, possivelmente originária da Ásia com as grandes navegações, espalhou-se para os diversos continentes e disseminou-se pelos países ocidentais e no continente africano (Homma, 2020).

A vacina humanizada foi introduzida no Brasil em 1804 pelo marechal Caldeira Brand Pontes (marquês de Barbacena), que enviou a Lisboa alguns escravos e um especialista a fim de que a vacina fosse transmitida de braço a braço até a

Bahia, de onde foi da mesma forma trazida para o Rio de Janeiro, ainda em 1804 (Fernandes, 1999).

No período entre 1900 e 1901, foram criados no país o Instituto Soroterápico do Rio de Janeiro, com direção do médico Oswaldo Cruz, e o Instituto Serumtherápico, com Vital Brazil, em São Paulo, que se dedicavam à produção da vacina para varíola e dos soros contra a peste bubônica e difteria. Essas instituições dariam origem às atuais, respectivamente, Fundação Oswaldo Cruz (Fiocruz) e Instituto Butantan (Santos, 2021).

Em 1904, no esforço de controlar a varíola na cidade do Rio de Janeiro, onde era frequente a ocorrência de epidemias por essa doença, Oswaldo Cruz conseguiu a aprovação da lei da obrigatoriedade da vacinação (Homma, 2020).

O Quadro 2 sintetiza a cronologia inicial e o histórico da evolução das vacinas humanas.

Quadro 2 – Histórico das vacinas humanas

Período	Vacinas vivas atenuadas	Microrganismos inteiros inativados	Proteínas purificadas de microrganismos ou polissacarídeos	Reassortantes	Engenharia genética (vacinas recombinantes)
Século XVIII	Varíola				
Século XIX	Raiva	Febre tifoide			
		Cólera			
		Peste			
Início do século XX	BCG (Tuberculose)	Coqueluche (célula inteira)	Difteria		
	Febre amarela	Influenza (gripe)	Tétano		
		Riquétsias			
Final do século XX	Poliomielite (OPV – vacina oral)	Influenza (gripe)	Pneumococo	Influenza (vírus morto + vivo)	Hepatite B recombinante
					Toxina da cólera
					Toxina da coqueluche
	Sarampo	Poliomielite (IPV – vacina inativada)	Meningococo		
	Caxumba	Raiva (nova)	Hepatite B (derivada do plasma)		
	Rubéola (adaptada ao frio)	Antraz			
	Adenovírus	Encefalite japonesa	*Haemophilus influenzae*		
	Febre tifoide (Ty21a)	Hepatite A	Febre tifoide (antígeno Vi)		
	Varicela	Encefalite transmitida por carrapatos	Coqueluche acelular		
	Cólera (CVD103)	*Escherichia coli* (+ CTB)	*Haemophilus influenzae* e meningococo (vacinas conjugadas com proteína)		
Século XXI	Herpes-zóster		Vacina pneumocócica conjugada	Rotavírus	Papilomavírus humano (HPV)

Fonte: Adaptado de Plotkin, 2011 (em tradução livre)

O desenvolvimento de uma vacina segue altos padrões de exigência e qualidade em todas as suas fases, o que inclui a pesquisa inicial, os testes em animais e

humanos, sob rigoroso protocolo de procedimentos éticos, até o processo de avaliação de resultados pelas agências reguladoras governamentais (Ballalai, 2016).

1.2.2. Calendário Vacinal no Brasil

Vacinar é, portanto, a melhor forma de proteger o indivíduo, sobretudo a criança, contra doenças graves. O Calendário Nacional de Vacinação orienta quais vacinas são necessárias em cada idade. Nos primeiros 1000 dias de vida, a vacinação é essencial para o crescimento e desenvolvimento saudável. A vacinação é reconhecida como uma das mais eficazes estratégias para preservar a saúde da população e fortalecer uma sociedade saudável e resistente. Além de prevenir doenças graves, a imunização contribui para reduzir a disseminação desses agentes infecciosos na comunidade, protegendo aqueles que não podem ser vacinados, por exemplo por motivos de saúde (BRASIL, 2025b).

No Quadro 3 encontra-se o calendário de vacinação direcionado a crianças de 0 a 4 meses, com publicação atualizada em 2025 pelo Ministério da Saúde. A BCG, um dos objetos desse estudo, é a primeira vacina do calendário.

Quadro 3 – Calendário Nacional de Vacinação – 0 a 4 meses

Idade	Vacina	Dose	Doenças evitadas
Ao nascer	BCG	1 dose	Formas graves e disseminadas da tuberculose e, também, efeito protetor contra a hanseníase.
	Hepatite B	1 dose	Hepatite B, hepatite D.
2 meses	Penta (DTP+Hib+HB)	1ª dose	Difteria, tétano, coqueluche, infecções pelo *H. influenzae* tipo b, hepatite B.
	Poliomielite inativada (VIP)	1ª dose	Poliomielite ou paralisia infantil.
	Pneumocócica 10-valente	1ª dose	Doenças pneumocócicas invasivas (pelos sorogrupos contidos na vacina).
	Rotavírus humano	1ª dose	Gastroenterite viral (diarreia, vômito).
3 meses	Meningocócica C	1ª dose	Doenças meningocócicas (meningite, encefalite, meningoencefalite) pelo meningococo tipo C.
4 meses	Penta (DTP+Hib+HB)	2ª dose	Difteria, tétano, coqueluche, infecções pelo *H. influenzae* tipo b, hepatite B.
	Poliomielite inativada (VIP)	2ª dose	Poliomielite ou paralisia infantil.
	Pneumocócica 10-valente	2ª dose	Doenças pneumocócicas invasivas (pelos sorogrupos contidos na vacina).
	Rotavírus humano	2ª dose	Gastroenterite viral (diarreia, vômito).

Fonte: Adaptado de BRASIL, 2025b

O Quadro 4 é a sequência do calendário do Quadro 1, agora voltado a crianças de 5 a 15 meses. Nele encontra-se a segunda vacina que é parte do objeto desta pesquisa, a Tríplice Viral (SCR) em sua 2ª. dose (no calendário administrada como Tetraviral (SRCV) 1ª. dose, incluindo o imunizante para a Varicela).

Quadro 4 – Calendário Nacional de Vacinação – 5 a 15 meses

Idade	Vacina	Dose	Doenças evitadas
5 meses	Meningocócica C	2ª dose	Doenças meningocócicas (meningite, encefalite, meningoencefalite) pelo meningococo tipo C.
6 meses	Penta (DTP+Hib+HB)	3ª dose	Difteria, tétano, coqueluche, infecções pelo *H. influenzae* tipo b, hepatite B.
	Poliomielite inativada (VIP)	3ª dose	Poliomielite ou paralisia infantil.
	Influenza trivalente	1ª dose	Influenza (gripe).
	Covid-19	1ª dose	Formas graves da covid-19 e óbitos causados pelo vírus SARS-CoV-2.
6 a 8 meses	Febre amarela	1 dose, em casos excepcionais[4]	Febre amarela.
7 meses	Covid-19	2ª dose	Formas graves da covid-19 e óbitos causados pelo vírus SARS-CoV-2.
9 meses	Covid-19	3ª dose	Formas graves da covid-19 e óbitos causados pelo vírus SARS-CoV-2.
	Febre amarela	1 dose	Febre amarela.
12 meses	Pneumocócica 10-valente	1 dose (reforço)	Doenças pneumocócicas invasivas (pelos sorogrupos contidos na vacina).
	Meningocócica ACWY	1 dose	Doenças meningocócicas (meningite, encefalite, meningoencefalite) por meningococos dos tipos A, C, W e Y.
	Tríplice viral (SCR)	1 dose	Sarampo, caxumba, rubéola e síndrome da rubéola congênita (futuramente, na gravidez).
15 meses	DTP	1ª dose (reforço)	Difteria, tétano, coqueluche.
	Poliomielite inativada (VIP)	1ª dose (reforço)	Poliomielite ou paralisia infantil.
	Tetraviral (SCRV)	1 dose	Sarampo, caxumba, rubéola, síndrome da rubéola congênita (futuramente, na gravidez) e varicela.
	Hepatite A	1 dose	Hepatite A.

Fonte: Adaptado de BRASIL, 2025b

Finalizando as vacinas previstas no calendário, o Quadro 5 elenca as vacinas direcionadas ao público da faixa etária de 4 a 14 anos.

Quadro 5 – Calendário Nacional de Vacinação – 4 a 14 anos

Idade	Vacina	Dose	Doenças evitadas
4 anos	DTP	2ª dose (reforço)	Difteria, tétano, coqueluche.
	Febre amarela	1 dose (reforço)	Febre amarela.
	Varicela	1 dose	Varicela ou catapora.
5 anos (somente indígena, sem histórico vacinal com pneumo conjugada)	Pneumocócica 23-valente	1 dose	Doenças pneumocócicas invasivas (pelos sorogrupos contidos na vacina).
A partir de 7 anos (todas as idades)	dT	3 doses, conforme histórico vacinal	Difteria, tétano.
9 a 14 anos	HPV4	1 dose	Infecções causadas pelo papilomavírus humano.

Fonte: Adaptado de BRASIL, 2025b

O calendário nacional de vacinação contempla, na rotina dos serviços, 19 vacinas que protegem o indivíduo em todos os ciclos de vida, desde o nascimento. Entre as doenças imunopreveníveis por essas vacinas estão a poliomielite, sarampo, rubéola, tétano, coqueluche e outras doenças graves e muitas vezes fatais. (BRASIL, 2025b).

Como já destacado, para essa pesquisa, duas vacinas e suas respectivas coberturas foram selecionadas como objeto de estudo, a BCG, presente no calendário para aplicação na criança ao nascer, e a Tríplice Viral (SCR) 2ª. dose, presente no calendário para aplicação na criança aos 15 meses de vida (BRASIL, 2025b).

1.2.2.1. A Vacina BCG

A BCG é composta pelo Bacilo de Calmette-Guérin – origem do nome BCG – obtido pela atenuação (enfraquecimento) de uma das bactérias que causam a tuberculose. Completam sua composição, o glutamato de sódio e a solução fisiológica (soro a 0,9%) (Ballalai, 2016).

Visa a prevenção, por meio da ativação do sistema imunológico, contra a bactéria que causa a meningite tuberculosa, formas pulmonares e disseminadas da doença. Sua administração é também recomendada para contatos domiciliares de hanseníase, como forma de prevenir a doença em indivíduos expostos ao *Mycobacterium leprae* (Lima, 2020).

A tuberculose é uma doença bacteriana infecciosa causada pela *Mycobacterium tuberculosis*, que é transmitida entre humanos pela via respiratória e afeta mais comumente os pulmões, mas pode danificar qualquer tecido; o contágio se dá de pessoa para pessoa por meio de aerossóis de gotículas da garganta e dos pulmões de pessoas com doença respiratória ativa. A importância da tuberculose entre as doenças infecciosas é atribuível à alta taxa de letalidade entre os casos não tratados ou tratados inadequadamente (Holmes, 2017).

A vacinação é a medida central para proteger as pessoas não infectadas de adoecerem por tuberculose, caso venham a se infectar, protegendo principalmente contra as formas primárias da doença (Procópio, 2014).

A hanseníase é uma doença infecciosa crônica causada pelo *Mycobacterium leprae*. A predileção pela pele e nervos periféricos confere características peculiares a esta moléstia, tornando o seu diagnóstico simples, na maioria dos casos. Em contrapartida, o dano neurológico responsabiliza-se pelas sequelas que podem surgir. Já constituiu importante problema de saúde pública no Brasil e em vários países do mundo (Araújo, 2003).

1.2.2.2. A Vacina Tríplice Viral (SCR)

A Tríplice Viral (SCR) trata-se de uma vacina atenuada, contendo vírus vivos enfraquecidos do sarampo, da rubéola e da caxumba (Ballalai, 2016).

É indicada, portanto, para a imunização ativa contra caxumba, sarampo e rubéola, composta por uma preparação das cepas de vírus atenuados, de sarampo, de caxumba, de rubéola, obtidas separadamente por propagação em cultura de ovos embrionados de galinha (sarampo e caxumba) ou em células diploides humanas (rubéola) (Lima, 2020).

O sarampo é uma doença altamente contagiosa causada pela infecção por um vírus pertencente ao gênero *Morbillivirus* da família *Paramyxoviridae*. A infecção se manifesta clinicamente como febre e exantema, seguidos de tosse, conjuntivite ou coriza. Mortes podem ocorrer como resultado de complicações como pneumonia, diarreia, desnutrição, otite média e encefalite. Sequelas

neurológicas permanentes também podem ocorrer, incluindo surdez e cegueira (Muloliwa, 2013).

A caxumba ou parotidite infecciosa é uma doença altamente contagiosa, causada pelo vírus RNA Paramyxovirus, transmitido por contato direto com perdigotos de pessoas infectadas. Uma das principais características da doença é o aumento das glândulas salivares próximas aos ouvidos que fazem o rosto inchar, e, em casos graves, pode causar surdez, meningite e, mais raramente, levar à morte. Após a puberdade, pode causar inflamação e inchaço doloroso dos testículos ou dos ovários e levar à esterilidade (Oliveira, 2021).

A rubéola, também chamada de sarampo alemão, é uma doença viral contagiosa que normalmente começa com febre baixa e linfadenopatia, seguidas pelo aparecimento breve e característico de uma erupção cutânea maculopapular eritematosa generalizada. Na ausência de gravidez, a rubéola geralmente é uma infecção leve, autolimitada e relativamente benigna. No entanto, a infecção materna por rubéola, durante o primeiro trimestre da gravidez, pode ser catastrófica e pode resultar em morte fetal ou no nascimento de um bebê com uma constelação de anomalias congênitas, denominada síndrome da rubéola congênita (Leung, 2019).

1.2.3. Cobertura vacinal e controle de doenças

A cobertura vacinal é um importante indicador de saúde, utilizado em âmbito global. Segundo a Organização Mundial de Saúde (OMS), a cobertura vacinal é definida como a proporção de uma determinada população que foi vacinada em um determinado período. É estimado para cada vacina e, para vacinas multidose, para cada dose recebida (Barbieri, 2021).

As coberturas de vacinação são decorrentes das ações e das realidades técnicas, de gestão, políticas e sociais de cada país em particular (Barbieri, 2021).

As baixas coberturas vacinais têm provocado sérios problemas à saúde dos brasileiros, principalmente pelo aumento das infecções por pneumococo, responsável por elevadas cargas de morbimortalidade (Brasil, 2018).

A OMS recomenda pelo menos 95% de cobertura vacinal para erradicação, eliminação ou controle de doenças imunopreveníveis, além de outros indicadores, como as proporções de municípios com coberturas adequadas (Dietz, 2004).

A taxa de cobertura vacinal é o indicador fundamental da vacinação. De maneira geral, espera-se que a imunidade de rebanho seja alcançada a partir de uma cobertura vacinal de 95% da população, como já preconizado pela OMS (Brasil, 2015).

É direta e proporcional a relação entre números da queda da mortalidade infantil e o aumento da cobertura vacinal no Brasil e no mundo, principalmente na faixa etária de crianças menores de 1 ano de vida (Machado, 2018).

A cobertura vacinal incorpora, portanto, aspectos individuais – capacidade de reação imunológica protetora – e coletivos – redução ou eliminação dos suscetíveis da população (Moraes, 2003).

Segundo Alves (Alves, 2026), em pesquisa conduzida nos estados brasileiros entre 2010 e 2021, em todo o país, houve queda na cobertura vacinal da BCG, seguido da poliomielite e da Tríplice Viral (SCR) 2ª. dose, além de todas as vacinas ficarem abaixo da cobertura vacinal mínima. Algumas vacinas, embora tenham apresentado aumento da cobertura vacinal no período, como a meningococo C, ainda apresentaram grandes diferenças percentuais abaixo da meta.

Na cidade de Cubatão no estado de São Paulo, uma investigação em crianças nascidas em 2019 e 2020, para o esquema vacinal completo, constatou que apenas 25,7% das crianças foram vacinadas sem atraso. Entre os fatores associados ao atraso vacinal aos 24 meses, estavam a escolaridade da mãe e o recebimento do benefício do Bolsa Família. Embora ainda mantendo uma

cobertura vacinal relevante, atraso para aplicação das doses representa um problema de saúde pública (Rozman, 2025).

Além das campanhas de vacinação suplementares, na hipótese constatada de baixa cobertura vacinal, é necessário um conjunto mais amplo de intervenções. Métodos aprimorados para monitorar a cobertura vacinal em nível individual e comunitário são imprescindíveis, juntamente com uma vigilância reforçada. A tomada de decisões precisa ser descentralizada para desenvolver estratégias que sejam, de fato, sustentáveis (Cutts, 2021).

Estudo transversal conduzido por Holipah (Holipah, 2020) objetivou determinar os fatores que promovem maiores taxas de imunização de crianças com a vacina DTP3 e constatou que, entre os fatores individuais, a presença de um profissional de saúde durante o parto foi o fator mais influente, porém, a nível distrital, os fatores variaram, concluindo que o baixo progresso na cobertura vacinal com a terceira dose da vacina DTP na Indonésia deveu-se às enormes disparidades entre as ilhas do país, à densidade dos serviços de saúde e à situação socioeconômica das famílias.

1.2.4. O Programa Nacional de Imunizações (PNI)

O Programa Nacional de Imunizações (PNI) foi criado em 1973 por determinação do Ministério da Saúde, com o objetivo de coordenar as ações de imunizações que se caracterizavam, até então, pela descontinuidade, pelo caráter episódico e pela reduzida área de cobertura (Brasil, 2025a).

O PNI adquire, distribui e normatiza também o uso dos imunobiológicos especiais, indicados para situações e grupos populacionais específicos que serão atendidos nos Centros de Referência para Imunobiológicos Especiais (CRIE). É também de sua responsabilidade a implantação do Sistema de Informação e a consolidação dos dados de cobertura vacinal em todo o país (Brasil, 2025a).

O PNI adota estratégias diferenciadas para definição das metas de coberturas para cada vacina, considerando eficácia, segurança e custo-efetividade do produto. Nesse sentido, foi estabelecida a meta de 90% para as vacinas BCG e VRH e de 95% para as demais vacinas da criança: Penta (DTP + HB + Hib), Poliomielite (VIP/VOP), Pneumo 10, Meningo C, Tríplice Viral (SCR), Hepatite A e Febre amarela. Para cálculo da cobertura vacinal, utiliza-se como numerador o total de doses que completam o esquema primário de cada vacina, e como denominador, o número de nascidos vivos no município, registrado no Sistema de Informações sobre Nascidos Vivos (SINASC), última base de dados disponíveis no âmbito nacional (Barbieri, 2021).

Algumas modificações têm sido propostas para os Sistemas de Informação (SI) que registram as doses aplicadas de vacinação. O Decreto 12.650, de 23 de julho de 2025, dispõe sobre a Rede Nacional de Dados em Saúde (RNDS) e sobre as Plataformas SUS Digital, melhor definindo o uso compartilhado de dados, bem como as atividades de tratamento desses dados (Brasil, 2025c), assim centralizando o registro de dados de vacinação no país, como dispõe a Portaria GM/MS de 5.663, de 31 de outubro de 2024. Delibera, inclusive, que os estados e municípios podem escolher os próprios sistemas de informação para registro dos dados de vacinação, desde que compatíveis com as regras de interoperabilidade com a RNDS do Ministério da Saúde (Brasil, 2024).

1.2.5. Determinantes da Cobertura Vacinal

O não comparecimento da população à unidade de saúde, seja para iniciar ou dar continuidade aos esquemas vacinais, tem múltiplas causalidades que merecem ser identificadas para adoção de soluções adequadas (horário de funcionamento, moradias de difícil acesso, ausência de meios de transporte nas áreas de abrangência da unidade, desconhecimento dos benefícios e indicações de vacinas etc.) (Barbieri, 2021).

Análise inédita do Instituto de Estudos para Políticas de Saúde (IEPS), com base em dados do Ministério da Saúde, mostra que menos da metade dos municípios brasileiros atingiu a meta estabelecida pelo Plano Nacional de Imunizações (PNI)

para nove vacinas, entre elas as que protegem contra hepatites, poliomielite, tuberculose e sarampo (Colucci, 2022).

Os motivos da queda das coberturas vacinais são múltiplos. Desde a percepção enganosa da população de que não é mais preciso vacinar porque as doenças desapareceram, o sucesso das ações de vacinação causou a falsa sensação de que não há mais necessidade de se vacinar, horário de funcionamento das unidades de saúde incompatíveis com as novas rotinas da população, problemas com o sistema informatizado de registro de vacinação, o desconhecimento de quais vacinas integram o calendário nacional de vacinação, o medo de que as vacinas causem reações prejudiciais ao organismo, o receio de que o número elevado de vacinas causem reações prejudiciais ao organismo e sobrecarregue o sistema imunológico (Domingues, 2019) (Zorzeto, 2018).

Ainda, a relação entre uma condição econômica desfavorável e uma menor cobertura vacinal pode ser compreendida, sob a perspectiva da intervenção, como um aspecto da condição de vida que dificulta o acesso a ela. Isto é, piores condições econômicas podem significar uma menor oferta de serviços e/ou uma maior dificuldade de acesso à intervenção (Moraes, 2008).

A ampliação da cobertura vacinal está relacionada à múltiplas condições, passando pela quebra da hesitação vacinal, que é influenciada por fatores como a complacência, conveniência e a confiança (MacDonalds, 2015).

O estudo das coberturas vacinais segundo condições de vida fornece uma oportunidade para a investigação acerca da concretização de políticas públicas compensatórias, isto é, intervenções programadas e executadas preferencialmente pelo poder público, que buscam diminuir as diferenças entre os grupos sociais (Moraes, 2000).

A redução destas desigualdades entre grupos sociais coloca-se como um dos principais objetivos do PNI, podendo ser pensada particularmente pela relação à posição socioeconômica, à raça, ao gênero e à localização geográfica (Moraes, 2008).

Em grandes centros urbanos, como é o caso do município de São Paulo, a diversidade de condições de vida poderia estar se refletindo em coberturas vacinais diferenciadas, não reveladas pela média municipal, pois é possível que nas áreas com piores condições de vida o acesso aos programas de imunização seja diferenciado, em relação à outras áreas da cidade (Moraes, 2000).

Quanto melhor estruturada a rede de atenção básica, maiores serão o acesso e a adesão ao Programa Nacional de Imunizações. Outras características que interferem nesse indicador são a proporção de população vivendo em zona rural, a dispersão da população no território e o nível de desenvolvimento socioeconômico (Moraes, 2003).

1.2.6. Consequências da baixa cobertura

A epidemia de sarampo, já caracterizado anteriormente, ocorrida na Grande São Paulo em 1997, surpreendeu o programa de vigilância epidemiológica e serviu de alerta para a importância de se conhecer, correta e oportunamente, a cobertura vacinal em menores de 1 ano (Moraes, 2000).

Ainda, atendo-se à saúde, as "*fake news*" sobre vacinação podem aumentar a hesitação vacinal e reduzir a busca pelos imunizantes disponíveis. Um exemplo brasileiro atual é a queda na cobertura vacinal de adultos e crianças. De acordo com dados do Ministério da Saúde, de 2019 a 2021, a porcentagem de vacinação dos brasileiros caiu de 73% para menos de 59% e, como resultado, registra-se o recrudescimento de doenças evitáveis pela vacinação, como o sarampo (Veneu, 2023).

As vacinas são vantajosas porque seu custo é menor do que o custo cumulativo de consequências como hospitalizações, tratamentos e perda de dias de trabalho, entre outras, resultantes da ocorrência de doenças imunopreveníveis. Por exemplo, nos Estados Unidos, a vacina Tríplice Viral permite economizar 16,34 dólares, apenas em custos médicos diretos, para cada dólar gasto com a

vacina, enquanto a Tríplice Bacteriana permite economizar 6,21 dólares para cada dólar gasto (Barata, 2005).

A vacinação com uma vacina segura e eficaz disponível desde o início da década de 1960 é a intervenção de saúde pública mais custo-efetiva para a prevenção do sarampo (Muloliwa, 2013).

Braz descreve (Braz, 2016) a classificação de risco de doenças imunopreveníveis nos municípios brasileiros ressurgirem. Dos 5.570 municípios brasileiros, 12,0% foram classificados como de risco muito baixo, 29,6% de risco baixo, 2,2% de risco médio, 54,3% de risco alto e 1,8% de risco muito alto, tornando imprescindível a vigilância permanente das coberturas vacinais.

Uma vez que, tradicionalmente, destinada a indivíduos saudáveis como forma de prevenir o adoecimento, a vacinação possui uma relação custo-benefício melhor que alternativas indicadas apenas para tratar, remediar efeitos ou mesmo curar enfermidades (Homma, 2020).

1.3. Fundamentos da Inteligência Artificial

A Inteligência Artificial (IA) pode ser definida como o ramo da ciência da computação que se ocupa da automação do comportamento inteligente (...). Porém, o problema de definir o campo inteiro da IA é semelhante ao de definir a própria inteligência: ela é uma faculdade ou apenas um nome para uma coleção de capacidades distintas e não relacionadas? (Luger, 2013).

IA pode ser vista como a capacidade de computadores digitais, ou robôs controlados por computador, de resolver problemas que normalmente são associados às maiores capacidades de processamento intelectual dos humanos (Russel, 2013).

Ainda, pode ser encarada como o estudo de como fazer os computadores fazerem coisas nas quais, no momento, as pessoas são melhores (Russel, 2013).

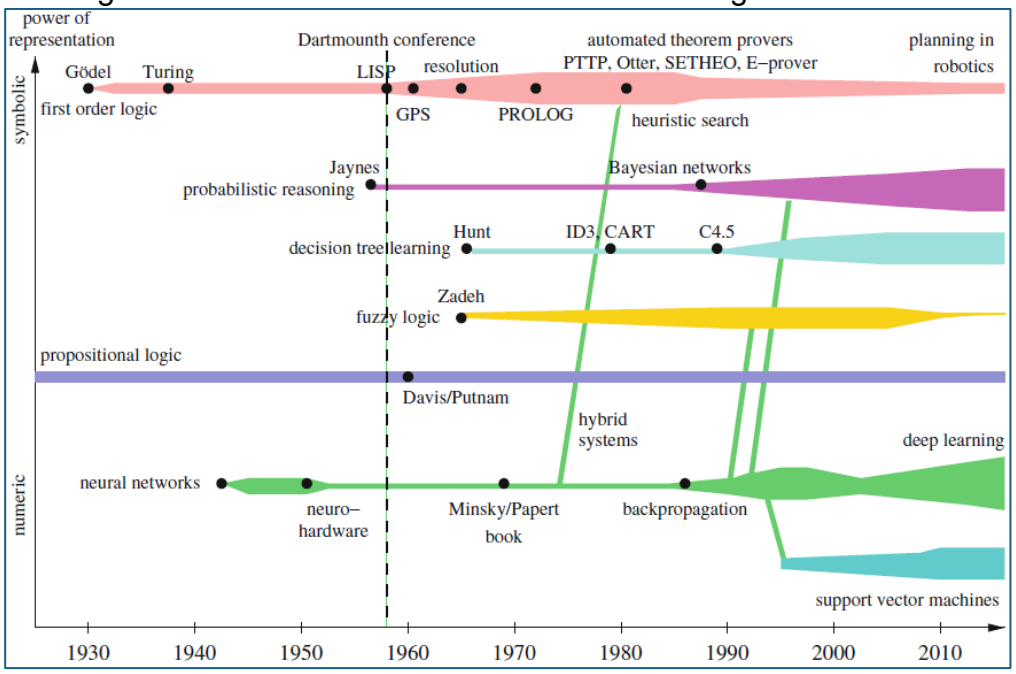
Como a maioria das ciências, a IA é composta de uma série de disciplinas que, ao mesmo tempo em que compartilha uma abordagem essencial para a solução de problemas, está preocupada com aplicações diferentes (Luger, 2013).

1.3.1. Evolução histórica da Inteligência Artificial

A IA baseia-se em muitas conquistas científicas do passado que não são mencionadas aqui, pois, a Inteligência Artificial como ciência em si, só existe desde meados do século XX (Ertel, 2017).

A Figura 1 mostra, de forma breve, alguns marcos importantes, de forma cronológica, na evolução da Inteligência Artificial.

Figura 1 – História das diversas áreas da Inteligência Artificial



Fonte: Ertel, 2017

O Quadro 6 é um resumo dos principais marcos da história da Inteligência Artificial, identificando as principais personalidades que contribuíram ao longo da evolução cronológica.

Quadro 6 – Cronologia resumida da Inteligência Artificial

1931 – 1955	O austríaco Kurt Gödel demonstra que, na lógica de predicados de primeira ordem, todas as afirmações verdadeiras são deriváveis.
	Alan Turing aponta os limites das máquinas inteligentes. McCulloch e Pitts modelam redes neurais e estabelecem a conexão com a lógica proposicional.
	Alan Turing define a inteligência artificial com o teste de Turing.
	Marvin Minsky desenvolve uma máquina de rede neural, com 3000 válvulas eletrônicas, simulando 40 neurônios.
	Arthur Samuel (IBM) cria um programa de Damas que aprende a jogar e que joga melhor do que seu desenvolvedor.
1956 – 1969	McCarthy organiza uma conferência no Dartmouth College e o nome Inteligência Artificial é introduzido pela primeira vez.
	McCarthy inventa no MIT (Instituto de Tecnologia de Massachusetts) a linguagem de alto nível LISP.
	McCarthy funda o Laboratório de Inteligência Artificial na Universidade de Stanford.
1970 – 1995	O cientista francês Alain Colmerauer inventa a linguagem de programação lógica PROLOG.
	Shortliffe e Buchanan desenvolvem o MYCIN, um sistema especialista para diagnóstico de doenças infecciosas, capaz de lidar com incertezas.
	O Japão inicia, com grande investimento, o “Projeto Quinta Geração” com o objetivo de construir uma poderosa máquina PROLOG.
	Renascimento das redes neurais por meio de, entre outros, Rumelhart, Hinton e Sejnowski.
	A partir da teoria da aprendizagem estatística, Vapnik desenvolve as máquinas de vetores de suporte, que são muito importantes ainda hoje em dia.
1996 – 2016	O computador de xadrez Deep Blue da IBM derrota o campeão mundial de xadrez Gary Kasparov.
	A robótica de serviço se torna uma importante área de pesquisa em Inteligência Artificial.
	O primeiro carro autônomo do Google trafega em uma rodovia da Califórnia.
	Watson entende linguagem natural e consegue responder a perguntas difíceis muito rapidamente.
	O aprendizado profundo possibilita uma excelente classificação de imagens.
2017	Mecanismos de atenção se tornaram parte integrante de modelos de modelagem e transdução de sequência convincentes em diversas tarefas (Vaswani, 2017).
2018	Apresentou-se uma estrutura para alcançar uma forte compreensão da linguagem natural com um único modelo agnóstico à tarefa, por meio de pré-treinamento generativo e ajuste fino discriminativo (Radford, 2018).

Fonte: Adaptado de Russel (Russel, 2017) e Ertel (Ertel, 2017)

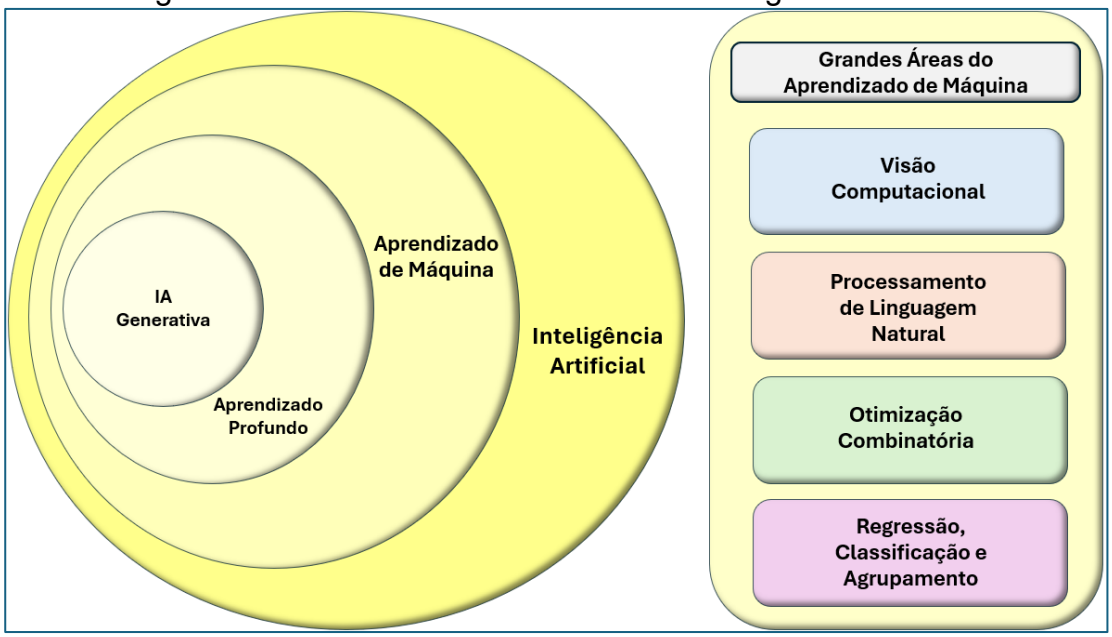
1.3.2. Divisões da Inteligência Artificial

É possível, de forma simplista e direta, dividir a IA em dois campos. A IA fraca (ou *narrow* IA ou IA restrita), que se limita à uma única tarefa, definida de forma restrita; aqui situam-se a maioria das soluções de IA da atualidade. Esses sistemas são desenvolvidos para lidar com um único problema e, geralmente, não são capazes de resolver outros problemas, mesmo os relacionados. Em contraste, existe a concepção da IA forte, onde um computador programado adequadamente possuiria uma “mente” no mesmo sentido dos seres humanos. Nesta, as máquinas poderiam resolver múltiplos problemas que não sejam definidos de forma estrita, alcançando algum nível de inteligência geral. Nenhum sistema de IA ainda alcançou a inteligência geral (Bartneck, 2021).

Muitas, senão a maioria, das publicações que mencionam IA são posicionadas em sete grandes segmentos: busca e otimização, sistemas *fuzzy*, planejamento e tomada de decisões, processamento de linguagem natural e representação do conhecimento, visão computacional, aprendizado de máquina, raciocínio probabilístico e redes neurais. Isso ressalta que a IA não é uma tecnologia única, mas pode ser mais bem compreendida como um conjunto de técnicas e subdisciplinas (Stahl, 2021).

A Figura 2 sintetiza as divisões da IA e às relaciona aos tipos de aplicações e foco de uso.

Figura 2 – Visão macro das divisões da Inteligência Artificial



Fonte: Elaborado pelo autor (2026)

1.3.3. *Machine Learning*: Conceitos e aplicações

Machine Learning (ML), ou Aprendizado de Máquina, é a intersecção entre ciência da computação teoricamente sólida e dados praticamente ruidosos. *Machine Learning* é, portanto, um tipo de inteligência artificial em que um algoritmo ou método extrairá padrões contidos nos dados (Kirk, 2015).

Ainda, segundo Jason Bell (Bell, 2015), ML é o ramo da IA que, usando computação, torna-se possível projetar sistemas que podem aprender com dados da maneira que são treinados. Os sistemas, assim, podem aprender e melhorar com a experiência e, com o tempo, refinar um modelo que pode ser usado para prever resultados de perguntas com base no aprendizado anterior.

ML tem sido uma parte estabelecida da pesquisa em IA desde a sua concepção, mas os recentes avanços no poder computacional e a disponibilidade de dados levaram a um aumento na sua aplicação numa vasta gama de domínios. ML, apesar de seus impressionantes sucessos, é ainda caracterizada como um exemplo de IA restrita (Stahl, 2021).

ML tem inúmeras aplicações em diversos setores. Na área da saúde, por exemplo, pode ser usada para analisar imagens médicas e dados de pacientes para melhorar o diagnóstico e o tratamento de doenças. No setor financeiro, pode ser aplicada para analisar tendências de mercado e tomar decisões de investimento. No transporte, existem aplicações para otimizar o fluxo de tráfego e melhorar a segurança viária. Na educação, é possível personalizar o aprendizado e melhorar o desempenho dos alunos (Banafa, 2024).

ML trata-se de um campo interdisciplinar que engloba ciência da computação, matemática e estatística. Diferentemente da IA simbólica baseada em regras, a ideia de utilizar ML para IA é aprender com dados (exemplos e experiências), não programando explicitamente regras elaboradas manualmente, mas, construindo um modelo para previsão, alimentando um algoritmo de ML com dados para que o algoritmo aprenda uma função otimizada com base nestes dados e na tarefa específica (Weng, 2020).

1.3.3.1. Tipos de aprendizado e a predição

ML é focada, como já destacado, na criação de algoritmos que utilizam a experiência, com relação à uma classe de tarefas, e o feedback, na forma de uma métrica de avalição para aprimorar seu desempenho nessa tarefa. A esse processo dá-se o nome de aprendizado, subdividindo-o, normalmente, em três tipos (Bartneck, 2021).

Mais adiante, ainda neste capítulo, serão apresentadas as definições dos tipos de aprendizado, atendo-se ao supervisionado e ao não supervisionado, relacionando-os, respectivamente, com seu principal objetivo de predição, seja uma classificação, uma regressão ou um agrupamento de dados. O terceiro tipo de aprendizado, o por reforço, uma vez que esta pesquisa é centrada na análise de dados estruturados tabulares, por ter aplicação diversa, não será revisado.

A modelagem preditiva, que trata o tema central da pesquisa, é o processo de desenvolvimento de uma ferramenta matemática ou um modelo de ML que

possa gerar uma previsão precisa e específica para um problema a ser abordado (Kuhn, 2013).

A análise preditiva é, portanto, o processo de utilização de formulações matemáticas avançadas, algoritmos de computação e estatística poderosos, ferramentas de software eficientes e serviços para representar, analisar e interpretar dados complexos. Como o próprio nome sugere, o objetivo central da análise preditiva é prever tendências, prever padrões nos dados ou prognosticar o comportamento do processo (Dinov, 2018).

1.3.3.1.1. Aprendizado supervisionado

Aprendizado supervisionado se refere a trabalhar com um conjunto de dados para treinamento previamente rotulado, ou seja, onde os dados já foram historicamente observados e, portanto, existem respostas conhecidas para cada instância. Para cada exemplo presente nestes dados de treinamento, existe, pelo menos, um objeto de entrada (variáveis independentes) e um objeto de saída (variável dependente) (Bell, 2015).

O aprendizado supervisionado pede que a máquina aprenda com os dados quando se especifica uma variável dependente como alvo. Isso reduz a tarefa da máquina a apenas “adivinhar” algum padrão dos dados de entrada para predizer a variável alvo (Harrington, 2012).

O treinamento supervisionado ocorre quando a saída que se deseja do algoritmo é conhecida. Assim, o treinamento supervisionado sempre vai lidar com dados conhecidos previamente (Heaton, 2013).

O treinamento supervisionado pode ser aplicado tanto à problemas de classificação, quanto de regressão. Por classificação deve ser entendido como a tentativa de determinar uma classe (discreta) na qual os dados de entrada se enquadram. Na classificação de dados, o resultado esperado é a identificação de uma classe que representa os dados (Heaton, 2013).

Em problemas de classificação utilizando um classificador baseado em ML, as únicas possibilidades de resultado são: (1) o exemplo é positivo e o classificador o reconhece como tal (verdadeiro positivo); (2) o exemplo é negativo e o classificador o reconhece corretamente (verdadeiro negativo); (3) o exemplo é positivo, mas o classificador o rotula como negativo (falso negativo); e (4) o exemplo é negativo, mas o classificador o rotula como positivo (falso positivo) (Kubat, 2017).

Já nos problemas de regressão, o foco é prever um valor numérico absoluto, normalmente não discreto, para uma série de entradas. Sob certos aspectos, isso pode ser considerado uma possível melhoria em relação à classificação, porque existe a possibilidade de prever um valor exato (contínuo) em vez de uma categoria (classe) discreta (Harrington, 2012).

Alguma medida de precisão é sempre usada para avaliar a eficácia do modelo. Há diferentes maneiras de se medir a precisão, cada uma com sua própria nuance. Para entender os pontos fortes e fracos de um modelo específico, confiar apenas em uma única métrica pode ser muito restrito. Visualizações do ajuste do modelo, particularmente gráficos residuais, são essenciais para entender se o modelo é adequado para o propósito sendo pesquisado (Kuhn, 2013).

Quando a variável dependente (rótulo) é ordinal ou contínua, a classificação por meio de dicotomização forçada inicial, numa tentativa de simplificar o problema, pode resultar em arbitrariedade e perda significativa de informação, mesmo quando se utiliza o ponto de corte ideal, que pode ser, por exemplo, a mediana (Harrell, 2015).

Ainda, dentro do aprendizado supervisionado, é possível lidar com análise de séries temporais, que podem ser definidas como um conjunto de observações, de certa condição ou variável, cada uma sendo registrada em um tempo específico, podendo levar em conta marcos de tempo discretos, onde o conjunto de tempos em que as observações são feitas é um conjunto discreto, como é o caso, por exemplo, quando as observações são feitas em intervalos de tempo fixos, e os marcos de tempos contínuos, onde as séries de tempo são obtidas

com observações registradas continuamente ao longo de algum intervalo de tempo (Brockwell, 2016).

1.3.3.1.2. Aprendizado não supervisionado

Na aprendizagem não supervisionada, o algoritmo tem o objetivo de procurar por um padrão oculto nos dados utilizados para o treinamento, ou seja, não existe identificação prévia dos rótulos das instâncias (Bell, 2015).

Neste tipo de aprendizagem, pode nem se quer existir uma variável definida e marcada como alvo (Harrington, 2012). Assim, esse tipo de treinamento é utilizado e adequado para quando não são fornecidas as saídas esperadas para o algoritmo (Heaton, 2013).

A aprendizagem não supervisionada envolve descobrir o que torna um determinado subconjunto dos dados especial. Por exemplo, quando existem muitos pontos comuns nos dados é possível agrupá-los por similaridade ou, talvez, determinar quais variáveis são melhores do que outras (Kirk, 2015).

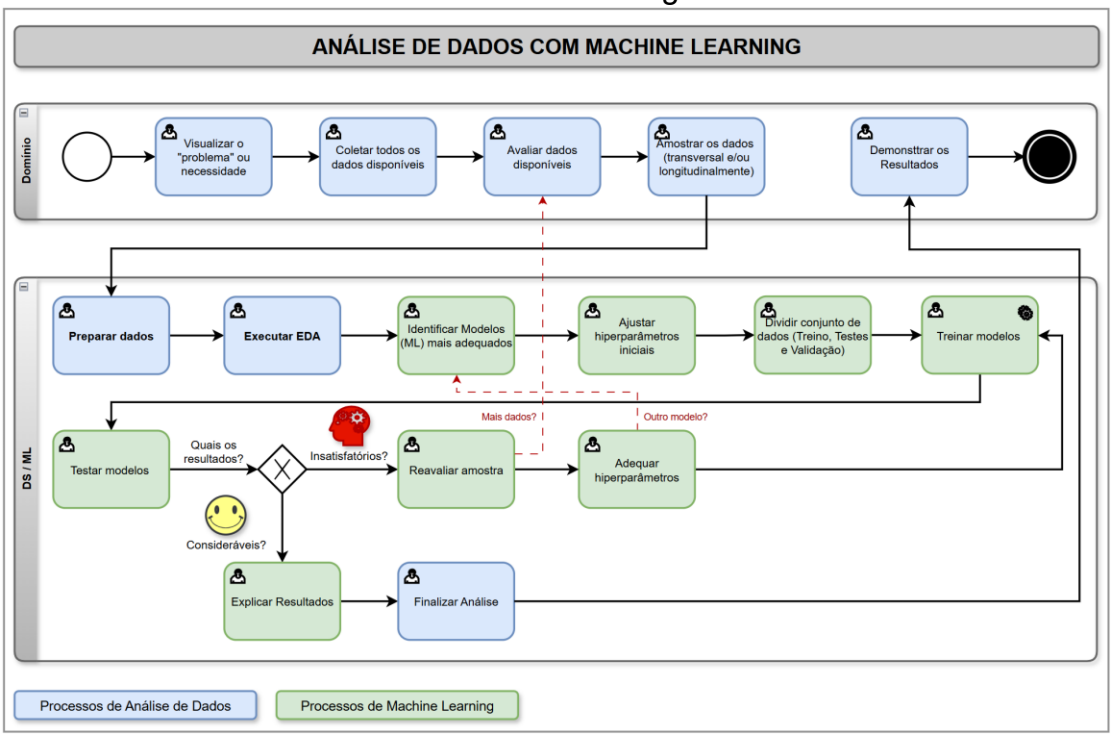
Com o aprendizado não supervisionado, não há resposta certa ou errada; trata-se apenas de executar o algoritmo de ML e observar quais padrões e resultados ocorrem, através da análise dos grupos formados (Bell, 2015).

Como objetivo principal do aprendizado não supervisionado, está a segmentação (ou agrupamento), que é um conjunto de métodos usados para dividir itens, geralmente pessoas ou produtos, em grupos lógicos mutuamente exclusivos, chamados segmentos. Esses métodos de segmentação incluem diversas funções para identificar as melhores variáveis a serem usadas na definição dos grupos, as variáveis a serem usadas na análise desses grupos, a atribuição de observações aos grupos, a análise/validação desses grupos e a criação de perfis para identificar as características comuns a eles (Dean, 2014).

1.3.3.2. Fluxo de análise de dados com Machine Learning

A Figura 3 tem como objetivo detalhar, no formato de fluxo, o que pode ser compreendido como passos fundamentais (ou essenciais) para adoção de ML para análise preditiva.

Figura 3 – Fluxo geral para análise de dados e aplicação de modelos de Machine Learning



Fonte: Elaborado pelo autor (2026).

Antes de qualquer etapa que envolva ML, é preciso que existam dados em quantidade e qualidade suficientes para análise. Ainda, é imprescindível compreender o problema a ser resolvido. Isso pode parecer óbvio, mas os projetos raramente vêm pré-definidos com dados claros e inequívocos. É importante entender os pontos fortes e as limitações dos dados, pois, raramente há uma correspondência exata com o problema (Provost, 2013).

“Lixo entra, lixo sai” é o princípio fundamental da análise de dados. O caminho dos dados brutos até um conjunto de dados limpo e analisável pode ser longo. Muitos problemas potenciais podem advir da limpeza de dados para análise (Skiena, 2017).

A matriz de avaliação dos dados de Mouga (Mouga, 2023) exibida no Quadro 7 relaciona quatro dimensões de completude em conjunto de dados (Insuficiente, Em desenvolvimento, Funcional e Ideal) e faz um cruzamento com o que se pode esperar em cada uma delas, nos quesitos Adequação, Prontidão, Confiabilidade, Sensibilidade e Suficiência.

Quadro 7 – Matriz de avaliação de dados

	Insuficiente	**Em desenvolvimento**	**Funcional**	**Ideal**
Adequação	Dados não relacionados à(s) questão(ões) de pesquisa	Variáveis-alvo inadequadas para a questão de pesquisa	Características não totalmente adequadas para a questão de pesquisa	Relação clara entre as características e as variáveis-alvo
Prontidão	Nenhum dado coletado ou metodologia de coleta definida	Alguns dados coletados ou metodologia de coleta clara em vigor	Dados coletados, mas que requerem alguma fusão e unificação	Dados coletados e limpos
Confiabilidade	Dados fortemente enviesados	Fonte de viés indeterminada	Fonte de viés determinada que pode ser explicada	Viés fraco ou isento de viés
Sensibilidade	Dados sensíveis de alto risco	Requerem medidas de segurança extensivas	Os dados não apresentam risco significativo, mas podem ser confidenciais	Riscos limitados associados aos dados, geralmente dados disponíveis publicamente
Suficiência	A experiência anterior demonstra baixa taxa de sucesso	Não atinge o nível de eventos anteriores bem-sucedidos	Eventos anteriores operaram com sucesso neste nível	Dados extensos em comparação com eventos anteriores bem-sucedidos

Fonte: Adaptado de Mougan (Mougan, 2023)

1.3.3.3. Métodos e algoritmos para análise preditiva

São diversos os métodos e algoritmos para análise preditiva dentro da ML. O detalhamento completo de todos os métodos e técnicas, cobrindo todos os princípios matemáticos e detalhes estruturais e de funcionamento de cada um, está além do escopo desta pesquisa.

No Quadro 8 encontra-se a menção sintética destes métodos e algoritmos agrupados por tipo de aprendizado e seus respectivos focos de utilização.

Quadro 8 – Exemplos dos métodos e algoritmos de Machine Learning

MÉTODO	ALGORITMO	APRENDIZADO	FOCO
Linear	Linear Regression	Supervisionado	Regressão
	Logistic Regression	Supervisionado	Classificação
	Support Vector Machines (SVM)	Supervisionado	Classificação
	Lasso Regression	Supervisionado	Classificação
	Ridge Regression	Supervisionado	Classificação
Árvores	Decision Trees	Supervisionado	Classificação Regressão
	Random Forest (Ensemble – Bagging)	Supervisionado	Classificação Regressão
	Gradient Boosting (Boosting)	Supervisionado	Classificação Regressão
	CatBoost (Ensemble – Boosting)	Supervisionado	Classificação Regressão
	LightGBM (Ensemble – Boosting)	Supervisionado	Classificação Regressão
	XGBoost (Ensemble – Boosting)	Supervisionado	Classificação Regressão
Probabilístico	Naive Bayes	Supervisionado	Classificação Regressão
	Gaussian Mixtures	Não Supervisionado	Agrupamento
	ARIMA (Autoregressive Integrated Moving Average)	Supervisionado	Séries Temporais
	SARIMA (Seasonal Autoregressive Integrated Moving Average)	Supervisionado	Séries Temporais
Similaridade	K-Nearest Neighbors	Supervisionado	Classificação
	K-Means	Não Supervisionado	Agrupamento
	DBScan	Não Supervisionado	Agrupamento
	HDBScan	Não Supervisionado	Agrupamento
Cognitivo	Artificial Neural Networks	Supervisionado	Classificação Regressão
	Self Organizing Maps	Não Supervisionado	Agrupamento

Fonte: Elaborado pelo autor (2026)

1.3.3.4. Métodos de treinamento e validação

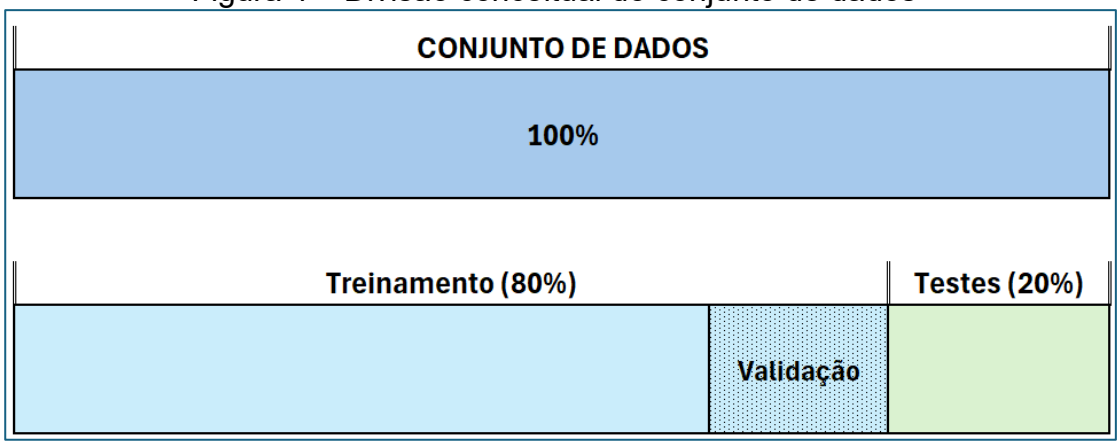
A tarefa primária para iniciar o que é chamado de “Processo de Treinamento” do modelo e algoritmo consiste em dividir o conjunto de dados em conjunto de

dados de treinamento (*training set*) e conjunto de dados de testes (*testing set ou holdout set*). Habitualmente, essa divisão é de 80% para treino e 20% para testes, mas, não há impedimentos formais e técnicos para que essa proporção seja ajustada, mantendo-se sempre a premissa de que o conjunto de treino seja o maior deles (Ellis, 2024).

Com dados suficientes, ou seja, quando a quantidade de instâncias (ou padrões) é razoável, pode-se, ainda, segregar uma parte do conjunto de treino, o que é chamado de conjunto de validação cruzada (*cross-validation set*), que será utilizado logo após o processo de treinamento, ajudando na avaliação do desempenho do modelo, ajustando os parâmetros nele utilizados, assim contribuindo para a prevenção da memorização dos dados, o que também é conhecido como *overfitting* (Hastie, 2017).

Como mencionado, o método bastante utilizado é a divisão do conjunto de treinamento para possibilitar o já descrito como validação cruzada. A Figura 4 ilustra essa sistemática de divisão do conjunto de dados.

Figura 4 – Divisão conceitual do conjunto de dados



Fonte: Elaborado pelo autor (2026)

A aplicação da técnica de validação cruzada pode ser modificada para considerar a divisão em vários (K) conjuntos menores, ou partições (*folds*). Na validação cruzada utilizando K-*fold*s, o conjunto de treinamento é dividido aleatoriamente em K partes sem reposição, onde K-1 partes são usadas para o treinamento do modelo e uma parte é usada para avaliação de desempenho.

Esse procedimento é repetido K vezes para que sejam obtidos K modelos e as suas respectivas estimativas de desempenho (Raschka, 2019).

A Figura 5 mostra o processo de divisão utilizando o princípio *k-folds.*

Figura 5 – Divisão do conjunto de treinamento em *K-folds*



Fonte: Raschka, 2019

1.3.3.5. Métricas de desempenho

As métricas de desempenho aqui elencadas e analisadas considerarão o tipo de aprendizado supervisionado, uma vez que o aprendizado não supervisionado ampliaria desnecessariamente o escopo desta revisão de literatura, por não ter sido aplicado nos métodos de pesquisa.

Avaliar o desempenho de um modelo de ML é um requisito fundamental e está basicamente relacionado em como medir o desempenho de determinado modelo e algoritmo e, mais, quando entender que o objetivo foi alcançado, ou seja, quando não é mais necessário modificação de parâmetros para evoluir o algoritmo (Mahdavinejad, 2018).

Em um modelo supervisionado, como já descrito, o valor esperado é conhecido previamente. O erro (E) do modelo é, portanto, o resultado da comparação da

predição (saída do modelo) com o valor realmente esperado, constante nos padrões presentes no conjunto de dados de treinamento.

O Quadro 9 se concentra nas métricas principais relacionadas a modelos de regressão, ou seja, aqueles modelos que têm por objetivo fazer a predição de um valor absoluto, não uma classe discreta.

Quadro 9 – Métricas utilizadas nos modelos de regressão

MÉTRICA	**DESCRIÇÃO**
MSE	*Mean Squared Error* ou Erro quadrático médio, que se trata simplesmente do valor médio dos erros elevados ao quadrado, sendo bastante útil para comparar diferentes modelos de regressão ou para ajustar seus parâmetros, seja por força bruta (tentativa e erro) ou por meio de busca utilizando validação cruzada, visto na sequência (Raschka, 2019).
RMSE	*Root Squared* MSE ou Raiz Quadrado do MSE, proporcionando uma análise mais interpretativa, uma vez que retorna à mesma unidade base do valor da predição (Ellis, 2024).
MAE	*Mean Absolute Error* ou Média do Erro Absoluto, que trata o erro de forma igualitária, mantendo na mesma unidade da predição (Ellis, 2024).
R2	R *Squared* ou R Quadrado ou, ainda, Coeficiente de Determinação, que corresponde a fração da variância da resposta que é capturada pelo modelo. É o resultado do MSE dividido pela variância da resposta do modelo (Hastie, 2017).

Fonte: Elaborado pelo autor (2026)

Para iniciar a revisão das métricas para análise dos resultados dos modelos de classificação, a seguir, o Quadro 10 apresenta a chamada “Matriz de Confusão” em versão simplificada para problemas binários, ou seja, aqueles modelos voltados a somente uma classe de predição, positiva ou negativa.

Uma “Matriz de Confusão” é uma tabela que resume o sucesso do modelo de classificação em prever a qual classe os exemplos (padrões) pertencem. Um eixo da matriz de confusão representa a classe prevista pelo modelo; o outro eixo representa o rótulo real (Raschka, 2019).

Quadro 10 – Representação da Matriz de Confusão

MATRIZ DE CONFUSÃO		PREDIÇÃO	
		SIM	**NÃO**
REAL	**SIM**	**VP** (Verdadeiro Positivo) Predição: SIM Real: SIM	**FN** (Falso Negativo) Predição: NÃO Real: SIM
	NÃO	**FP** (Falso Positivo) Predição: SIM Real: NÃO	**VN** (Verdadeiro Negativo) Predição: NÃO Real: NÃO

Fonte: Elaborado pelo autor (2026)

A partir da Matriz de Confusão, algumas métricas importantes podem ser calculadas e analisadas. No Quadro 11 pode ser encontrado um detalhamento dessas métricas e suas aplicabilidades para análise da resposta do modelo e algoritmo utilizado.

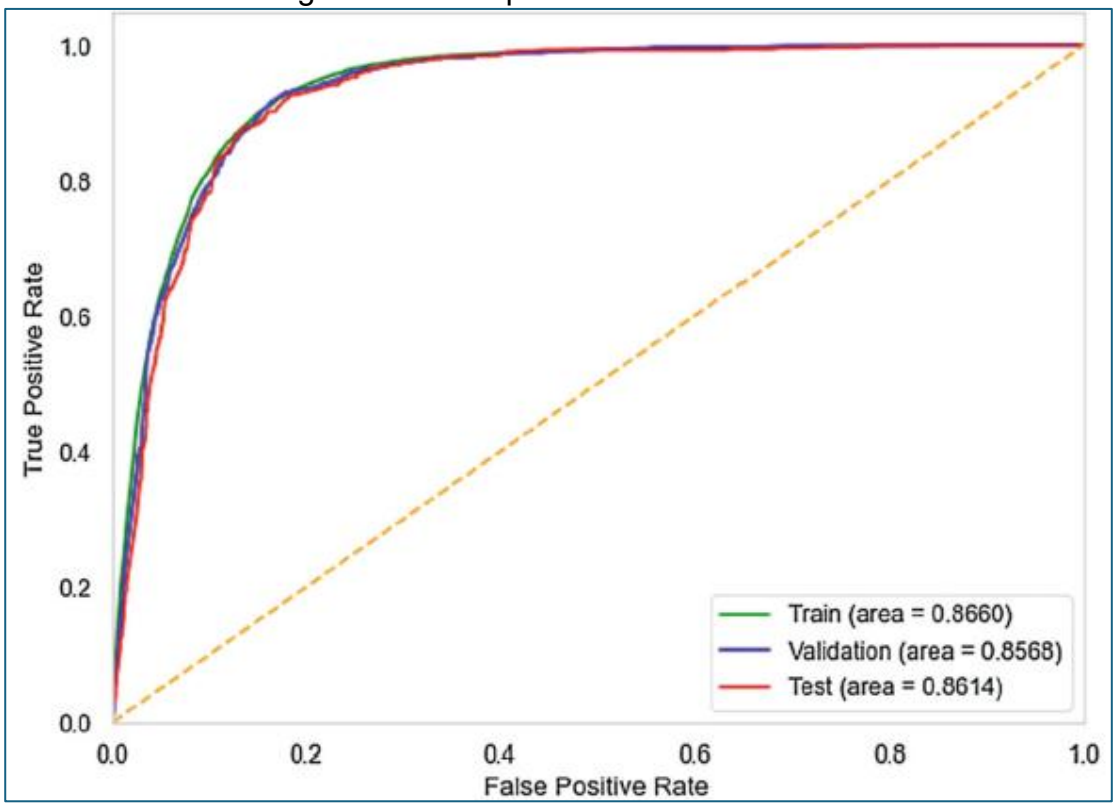
Quadro 11 – Principais métricas obtidas a partir da Matriz de Confusão

MÉTRICA	FORMULAÇÃO	DESCRIÇÃO
Acurácia	$\frac{(VP + VN)}{(VP + FN + FP + VN)}$	Em interpretação literal, é a proporção de instâncias classificadas corretamente em relação ao total de instâncias no conjunto de dados. Possui certa restrição de uso em conjuntos de dados com classes não proporcionais (Wulff, 2025).
Precisão	$\frac{VP}{(VP + FP)}$	A precisão é a proporção das instâncias positivas efetivamente encontrados na lista de todas as instâncias analisadas.
Sensibilidade (Recall)	$\frac{VP}{(VP + FN)}$	Proporção de instâncias verdadeiramente positivas entre todas as instâncias positivas reais, possibilitando a análise da capacidade do modelo de identificar corretamente as instâncias positivas.
Pontuação Fβ (Fβ Score)	$(1 + \beta^2) \frac{Precisão * Sensibilidade}{\beta^2 * Precisão + Sensibilidade}$	A pontuação Fβ é a média da precisão e da sensibilidade, fornecendo uma medida equilibrada de ambas as métricas. O valor mais comum de β é 1 (F1-Score), que, matematicamente, balanceia proporcionalmente as duas métricas.

Fonte: Elaborado pelo autor (2026)

Completando as métricas que são base para análise dos modelos de classificação, a Figura 6 mostra o gráfico da Curva ROC-AUC (*Receiver Operating Characterisc – Area Under the Curve* ou Área sob a Curva de Característica de Operação do Receptor, em tradução livre) que mede a capacidade de um modelo de distinguir entre instâncias positivas e negativas, mostrando a taxa de Verdadeiros Positivos (Sensibilidade) contra a taxa de Falsos Positivos, para diferentes limiares. Um valor de 0,5 significa que o modelo é tão bom quanto um chute aleatório, enquanto 1 indica um desempenho perfeito (Wulff, 2025).

Figura 6 – Exemplo de curva ROC-AUC



Fonte: Dumpert, 2025.

1.3.3.6. Interpretação e explicação dos modelos e seus resultados

A interpretabilidade (ou explicabilidade) de um modelo ou algoritmo de ML depende da facilidade com que se consegue compreender os procedimentos e técnicas que foram utilizados para chegar aos seus resultados. Enxergar os modelos como “caixas-pretas”, sem nenhuma maneira de saber o que acontece internamente, torna impossível explicar os resultados a quem quer que esteja interessado (Ganatra, 2024).

Duas formas usualmente implementadas, com o objetivo de explicar os modelos e seus resultados, são SHAP e LIME, que podem ser definidos como:

- SHAP (*Shapley Additive exPlanations*): atribui a cada característica um valor de importância derivado dos valores de *Shapley*[1] na teoria dos jogos cooperativos, garantindo que as explicações satisfaçam propriedades

[1] Valor Shapley: Estudo sobre teoria dos jogos cooperativos de Lloyd Shapley. Shapley, L.S. A value for n-person games. Princeton University Press. Princeton, 1953.

desejáveis, incluindo precisão local, ausência de dados e consistência (Lamane, 2024).

- LIME (*Local Interpretable Model-Agnostic Explanations* ou explicações locais interpretáveis e independentes de modelo): gera explicações perturbando instâncias de entrada (de forma estocástica), obtendo novas previsões do modelo para essas perturbações e ajustando um modelo linear ponderado para aproximar o comportamento local. Pode ser aplicado em diversas arquiteturas, desde modelos lineares até redes neurais profundas (May, 2025).

Ambos, como pode ser interpretado pelas descrições, estão focados e preocupados em demonstrar quais são os atributos (variáveis independentes) mais importantes e quais suas respectivas contribuições para os resultados obtidos, objetivando dar mais segurança e confiança na utilização dos modelos e algoritmos de ML adotados para abordar os possíveis e mais diversos problemas.

1.3.4. Machine Learning na saúde e cobertura vacinal

ML tem sido amplamente utilizada para problemas relacionados à área da saúde com algumas aplicações, também, para análise da Cobertura Vacinal. O Quadro 12 sintetiza uma revisão de literatura com diversos exemplos de aplicações, tanto em sentido amplo, como, em alguns casos, estritamente voltados à Cobertura Vacinal, o foco desta pesquisa.

Foram revisados 22 artigos, os mais antigos do ano de 2020 e os mais recentes do ano de 2024. Todos os artigos revisados aplicam ML, porém, com abordagens e métodos diferenciados, fazendo da revisão uma ferramenta importante e essencial, uma vez que vai de encontro ao objeto da pesquisa e contribui sobremaneira na fundamentação das escolhas dos métodos, trazendo ainda mais argumentos para sustentar as comparações e como elas colaboram para os métodos que serão aqui demonstrados.

Quadro 12 – Revisão de trabalhos da área da saúde e cobertura vacinal com Machine Learning

AUTOR	TÍTULO	OBJETIVOS	MÉTODOS ML
Alegado, 2020	***Statistical and Machine Learning Methods for Vaccine Demand Forecasting: A Comparative Analysis***	Encontrar um modelo adequado para prever o estoque apropriado de vacinas para evitar escassez e excesso de oferta.	Aprendizado supervisionado utilizando ARIMA e *Artificial Neural Networks*.
Avirappattu, 2022	***Optimized machine learning model for identifying socio-economic, demographic and health-related variables associated with low vaccination levels that vary across ZIP codes in California***	Utilizar ML para identificar variáveis socioeconômicas, demográficas e relacionadas à saúde associadas a baixos níveis de vacinação que variam entre os códigos postais da Califórnia.	Aprendizado supervisionado utilizando *Gradient Boost Regressor*.
Bar-Lev, 2021	***Prediction of vaccine hesitancy based on social media traffic among Israeli parents using machine learning strategies***	Utilizar estratégias de ML para avaliar como o conteúdo online sobre vacinação afeta a hesitação em relação à vacina.	Aprendizado supervisionado utilizando *Logistic Regression*, *Random Forest* e *Artificial Neural Networks*.
Biswas, 2023	***Performance of predictive algorithms in estimating the risk of being a zero-dose child in India, Mali and Nigeria***	Examinar a capacidade dos algoritmos preditivos de caracterizar o risco de uma criança ser “dose zero (ZD)” com base em variáveis preditoras disponíveis em dados administrativos de rotina.	Aprendizado Supervisionado utilizando *Ridge Classifier*, *Decision Tree*, K-NN e *Artificial Neural Networks*.
Carrieri, 2021	***Predicting vaccine hesitancy from area-level indicators: A machine learning approach***	Prever comunidades com alto risco de hesitação vacinal com base em indicadores de nível de área facilmente acessíveis aos formuladores de políticas.	Aprendizado supervisionando utilizando *Lasso Regression*, *Gradient Boosting Machine, Artificial Neural Networks, Random Forest*.
Chang, 2024	***Measuring vaccination coverage and concerns of vaccine holdouts from web search logs***	Usar os registros de mecanismos de busca e o aprendizado de máquina para ajudar no entendimento da hesitação vacinal, ou seja, quem está relutante em se vacinar e por que,	Aprendizado supervisionando utilizando *Graph Neural Networks*.

		contribuindo para a elaboração de políticas de vacinação eficazes.	
Couto, 2022	***Covid-19 vaccination priorities defined on machine learning***	Definir grupos prioritários para vacinação como fator crítico para reduzir as taxas de mortalidade.	Aprendizado supervisionado utilizando *XGBoost*.
Demsah, 2023	***Machine learning algorithms' application to predict childhood vaccination among children aged 12–23 months in Ethiopia: Evidence 2016 Ethiopian Demographic and Health Survey dataset***	Prever a vacinação infantil em crianças de 12 a 23 meses de idade utilizando o melhor (ou mais adequado) algoritmo de aprendizado de máquina.	Aprendizado supervisionado utilizando *Decision Tree*, *Random Forest* e *Artificial Neural Networks*.
Doddo, 2024	***Using machine learning algorithms to predict COVID-19 vaccine uptake: A year after the introduction of COVID-19 vaccines in Ghana***	Realizar um estudo transversal utilizando questionários autoaplicáveis para determinar os fatores, incluindo as barreiras, que predizem a adesão à vacina contra a COVID-19 entre pacientes que frequentam um hospital terciário e quaternário, utilizando alguns algoritmos de ML.	Aprendizado supervisionando utilizando K-NN e *Random Forest*.
Ebulue, 2024	***Leveraging machine learning for vaccine distribution in resource limited settings: A synthesis of approaches***	Sintetizar várias abordagens de ML para solucionar os desafios da distribuição de vacinas em ambientes com recursos limitados, incluindo uma revisão da literatura sobre aplicações de ML em saúde e distribuição de vacinas, destacando as principais descobertas e metodologias.	Aprendizado supervisionado utilizando *Random Forests* e *Support Vector Machines*; otimização Combinatória com *Genetic Algorithms* e *Simulated Annealing*, entre outros
Kumar, 2023	***Machine Learning and Deep Learning Based Time Series Prediction and Forecasting of Nations' COVID-19 Pandemic***	Investigar e prever as futuras circunstâncias ambientais da COVID-19 para minimizar seus efeitos usando técnicas de inteligência artificial.	Aprendizado supervisionado utilizando K-NN, *XGBoost*, *Linear Regression*, entre outros.

Lu, 2023	***Reinforcement learning relieves the vaccination dilemma***	Estudar como uma regra de tomada de decisão para vacinação pode afetar a disseminação de epidemias, explorando o modelo de Bush-Mosteller (BM), uma das metodologias de aprendizado por reforço em inteligência artificial (IA), que pode realizar o processo sistemático de aprendizado em humanos, em redes complexas.	Aprendizado por Reforço utilizando *Artificial Neural Networks*.
Lui, 2020	***Robust Predictive Models in Clinical Data – Random Forest and Support Vector Machines***	Detalhar os princípios que tornam Random forest e Support Vector Machine ferramentas de modelagem e previsão bem-sucedidas para diversas aplicações, incluindo um exemplo ilustrativo utilizando o banco de dados MIMIC III.	Aprendizado supervisionado utilizando *Random Forest* e *Suport Vector Machine*.
Magazzino, 2022	***A machine learning algorithm to analyse the effects of vaccination on COVID-19 mortality***	Desenvolver um experimento de ML para avaliar os efeitos da vacinação na taxa de mortalidade da pandemia de COVID-19.	Aprendizado supervisionado utilizando *Artificial Neural Networks*.
Revathi, 2024	***Correlation between Vaccination and Child Mortality Rate Using Multivariate Linear Regression Model***	Utilizando a base de dados da OMS sobre a taxa de consumo de vacinas em diversos países, analisar as razões do baixo acesso às vacinas em relação aos níveis de mortalidade e pobreza.	Aprendizado supervisionado utilizando *Multivariate Linear Regression*.
Salgado, 2020a	***Machine Learning for Patient Stratification and Classification Part 2: Unsupervised Learning with Clustering***	Estratificação dos pacientes admitidos na ICU (Intensive Care Unit – Unidade de Tratamento Intensivo, em tradução livre) entre 2001 e 2012.	Aprendizado não supervisionado para agrupamento utilizando K-*means*.
Salgado, 2020b	***Machine Learning for Patient Stratification and Classification Part 3: Supervised Learning***	Aplicação de modelos de classificação para prever a mortalidade dos pacientes admitidos na ICU (Intensive Care Unit – Unidade de Tratamento Intensivo, em tradução livre) entre 2001 e 2012.	Aprendizado supervisionado utilizando *Logistic Regression*, *Decision Tree* e *Random Forest*.

Sameen, 2022	**Improving Routine Immunization Coverage Through Optimally Designed Predictive Models**	Propor um modelo preditivo baseado de ML para prever quais crianças faltarão ou não às próximas consultas de vacinação e para examinar o efeito dos fatores subjacentes que contribuem para essa falta.	Aprendizado supervisionado utilizando *Random Forest*.
Sanderse, 2021	**Application of Machine Learning Models to Predict the Covid-19 Vaccine Coverage of Sub-National Areas in Flanders Using Socio-Demographic Variables**	Explicar a relação entre as variáveis sociodemográficas de áreas subnacionais e a cobertura vacinal contra a Covid-19 em Flandres, testando a relação entre variáveis sociodemográficas em um conjunto mais amplo de variáveis sociodemográficas.	Aprendizado supervisionado utilizando *Multiple Linear Regression*, Random *Forest* e K-NN.
Tadese, 2024	**Prediction of incomplete immunization among under-five children in East Africa from recente demographic and health surveys: a machine learning approach**	Identificar preditores de vacinação incompleta em crianças menores de 5 anos na África Oriental através da análise de dados secundários de seis países da África Oriental, utilizando o conjunto de dados da Pesquisa Demográfica e de Saúde (DHS) de 2016 a 2021.	Aprendizado supervisionado utilizando *XGBoost*.
Weng, 2020	**Machine Learning for Clinical Predictive Analytics**	Aplicação de técnicas de aprendizado de máquina para tarefas de previsão clínica, incluindo estudos de caso utilizando dados reais de unidades de terapia intensiva (UTI) provenientes de um conjunto de dados publicamente disponível, o PhysioNet Challenge 2012, e dados sobre tumores de mama do banco de dados de diagnóstico do câncer de mama de Wisconsin.	Aprendizado não supervisionado para agrupamento com K-*means* e aprendizado supervisionado com *Logistic Regression*, *Decision Tree* e *Suport Vector Machine*.

Xie, 2020	**Biomedical Signal Processing: ECG Application**	Demonstrar a extração e aplicação da transformada *wavelet* contínua em sinais de ECG multicanal de pacientes com arritmias para desenvolver um classificador de batimentos cardíacos de alto desempenho, capaz de distinguir entre vários tipos de batimentos regulares e irregulares.	Aprendizado supervisionado utilizando *Logistic Regression*, K-NN, *Support Vector Machine* e *Artificial Neural Networks*.

Fonte: Elaborado pelo autor (2026).

2. JUSTIFICATIVAS

A cobertura vacinal, no Brasil e no mundo, tem sofrido uma queda expressiva, fazendo com que especialistas de todos os países fiquem preocupados, uma vez que a vacinação foi responsável por grandes conquistas na saúde pública mundial, como já destacado nas revisões aqui presentes. A queda da cobertura vacinal pode levar a um retrocesso na área da saúde brasileira, com retorno de doenças que estavam erradicadas, eliminadas ou controladas, ocorrendo, entre outros fatores, pela não completude do esquema vacinal. O estado de São Paulo é composto por 645 municípios que possuem características diferentes, mas, que guardam entre si, taxas de mortalidade infantil semelhantes.

Atualmente, existe uma grande variedade de modelos que se utilizam de técnicas de aprendizado de máquina nas mais diversas áreas, incluindo a área da saúde, como fora revisado. A investigação destes modelos aplicados aos elementos constantes nos diversos repositórios de dados, os socioeconômicos, demográficos, de saúde, educacionais e de cobertura vacinal, visto os importantes e promissores exemplos apresentados, mostra-se com grande potencial para obtenção de bons resultados.

Ademais, a proposta de uma metodologia orientada a dados é considerada abordagem de vanguarda em diversas outras áreas de pesquisa, tais como, visão computacional, processamento de linguagem natural e tradução de fala para texto, entre outras aplicações igualmente complexas, sendo adequadas não apenas a contextos teóricos, mas, sobretudo, aos desafios encontrados no mundo real.

3. OBJETIVOS

3.1. Objetivos Gerais

Como objetivos gerais, a presente pesquisa consolidou uma análise preditiva das coberturas vacinais da BGC e da Tríplice Viral (SCR) através de modelos de Machine Learning, utilizando os dados dos munícipios do estado de São Paulo entre os anos 2016 e 2023.

3.2. Objetivos Específicos

O presente trabalho de pesquisa teve como objetivos específicos, ainda:

- Promover análise exploratória dos dados utilizando métodos de estatística descritiva e inferencial;
- Utilizar estratégias de extrapolação de dados para tratar lacunas temporais;
- Testar e avaliar os modelos e algoritmos padrões da ML apontando o melhor método diante dos cenários de pesquisa;
- Destacar o conjunto de atributos que melhor explicam os resultados do modelo de ML com o melhor resultado.

4. MÉTODO

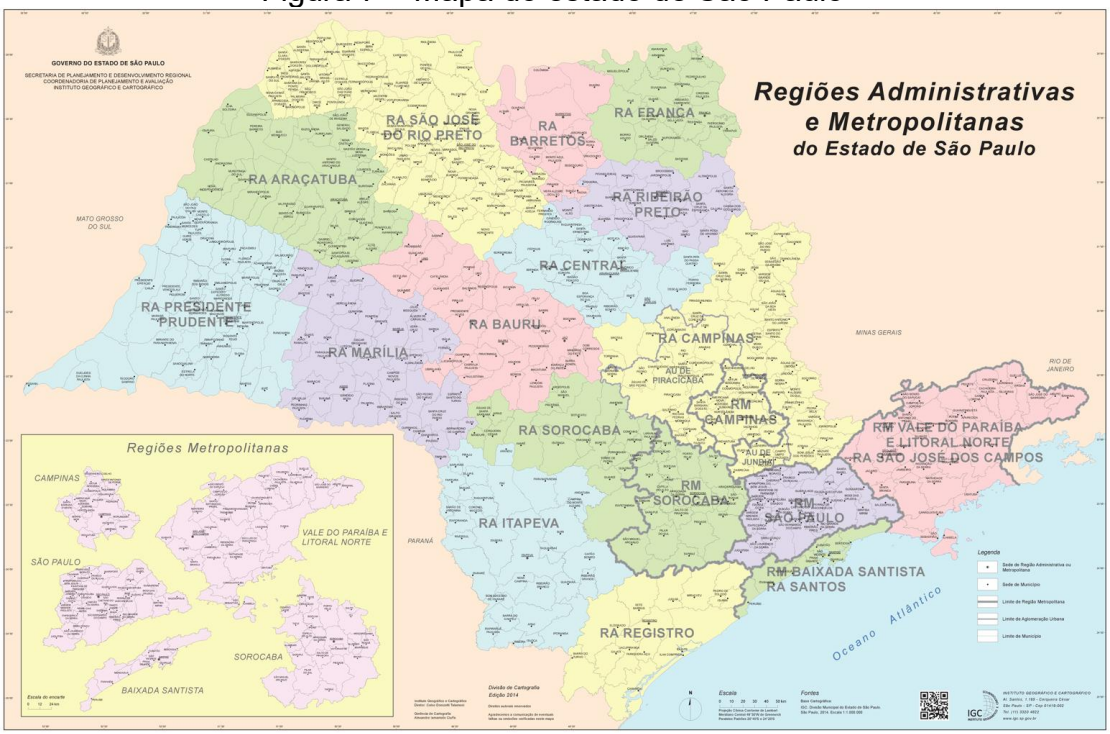
A presente pesquisa tratou-se de um Estudo Epidemiológico Ecológico do Tipo Misto, ou seja, concentrou-se em analisar dados tanto de uma determinada delimitação geográfica, neste caso, os municípios do estado de São Paulo, quanto de uma temporalidade específica e definida, entre os anos de 2016 e 2023, inclusive.

A análise conduzida nesta pesquisa se baseou no fluxo da Figura 3 (tópico “1.3.3.2. Fluxo de Análise de Dados com Machine Learning”). Para cada processo constante do fluxo geral proposto, teve-se, portanto, a aplicação de um método considerado o mais apropriado, subdividindo-o ou não, tendo como pano de fundo os objetivos específicos ora elencados.

4.1. A delimitação geográfica da pesquisa

Como já destacado, a pesquisa concentrou-se nos municípios do estado de São Paulo. A Figura 7 mostra a divisão do estado de São Paulo em Regiões Administrativas e Regiões Metropolitanas, como o estado atualmente está organizado para melhor governança (Seade, 2024).

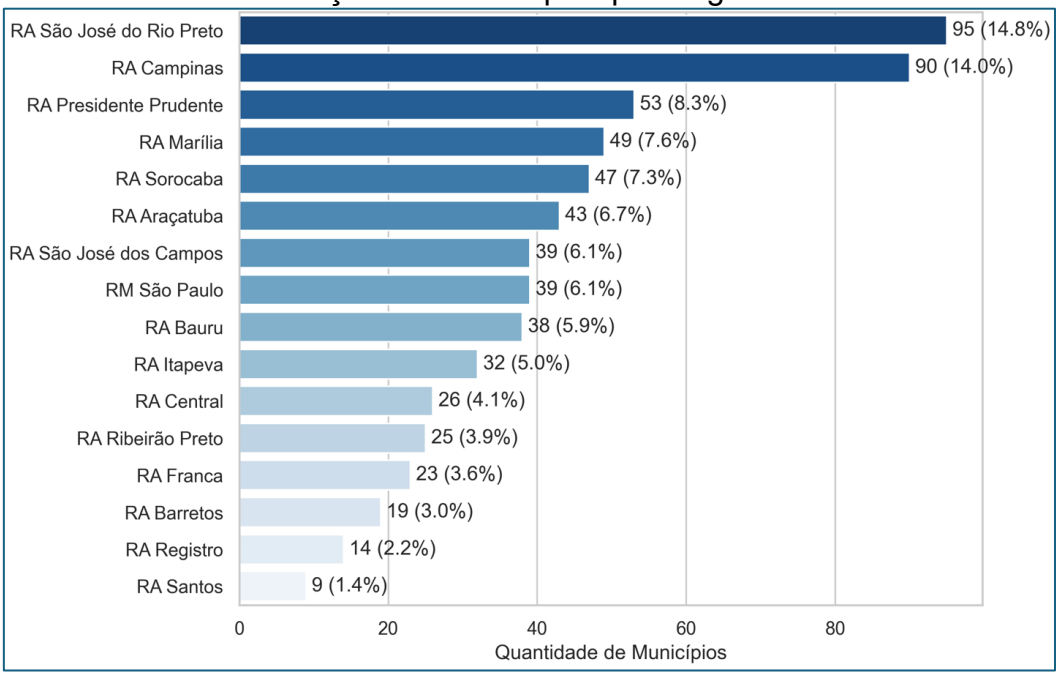
Figura 7 – Mapa do estado de São Paulo



Fonte: IGC, 2014.

Com base nos dados constantes do repositório explorado, foi produzido o Gráfico 1 que mostra as quantidades de municípios em cada uma das Regiões Administrativas.

Gráfico 1 – Distribuição dos municípios por Regiões Administrativas



Fonte: Elaborado pelo autor (2026)

Ainda, com auxílio do mesmo conjunto de dados, o Quadro 13 elenca quais são os municípios pertencentes a quais Regiões Administrativas, destaque para a "RA Santos", área que contém o município de Santos e que leva seu nome, inclusive.

Quadro 13 – Regiões Administrativas e seus municípios

REGIÃO	MUNICÍPIOS
RA Araçatuba	Alto Alegre, Andradina, Araçatuba, Auriflama, Avanhandava, Barbosa, Bento de Abreu, Bilac, Birigui, Braúna, Brejo Alegre, Buritama, Castilho, Clementina, Coroados, Gabriel Monteiro, Gastão Vidigal, General Salgado, Glicério, Guaraçaí, Guararapes, Guzolândia, Ilha Solteira, Itapura, Lavínia, Lourdes, Luiziânia, Mirandópolis, Murutinga do Sul, Nova Castilho, Nova Independência, Nova Luzitânia, Penápolis, Pereira Barreto, Piacatu, Rubiácea, Santo Antônio do Aracanguá, Santópolis do Aguapeí, São João de Iracema, Sud Mennucci, Suzanápolis, Turiúba, Valparaíso
RA Barretos	Altair, Barretos, Bebedouro, Cajobi, Colina, Colômbia, Embaúba, Guaíra, Guaraci, Jaborandi, Monte Azul Paulista, Olímpia, Pirangi, Severínia, Taiaçu, Taiúva, Terra Roxa, Viradouro, Vista Alegre do Alto
RA Bauru	Agudos, Arealva, Avaí, Balbinos, Bariri, Barra Bonita, Bauru, Bocaina, Boracéia, Borebi, Cabrália Paulista, Cafelândia, Dois Córregos, Duartina, Getulina, Guaiçara, Guaimbê, Guarantã, Iacanga, Igaraçu do Tietê, Itaju, Itapuí, Jaú, Lençóis Paulista, Lins, Lucianópolis, Macatuba, Mineiros do Tietê, Paulistânia, Pederneiras, Pirajuí, Piratininga, Pongaí, Presidente Alves, Promissão, Reginópolis, Sabino, Ubirajara, Uru
RA Campinas	Aguaí, Águas da Prata, Águas de Lindóia, Águas de São Pedro, Americana, Amparo, Analândia, Araras, Artur Nogueira, Atibaia, Bom Jesus dos Perdões, Bragança Paulista, Brotas, Cabreúva, Caconde, Campinas, Campo Limpo Paulista, Capivari, Casa Branca, Charqueada, Conchal, Cordeirópolis, Corumbataí, Cosmópolis, Divinolândia, Elias Fausto, Engenheiro Coelho, Espírito Santo do Pinhal, Holambra, Hortolândia, Indaiatuba, Ipeúna, Iracemápolis, Itapira, Itatiba, Itirapina, Itobi, Itupeva, Jaguariúna, Jarinu, Joanópolis, Jundiaí, Leme, Limeira, Lindóia, Louveira, Mococa, Mogi Guaçu, Mogi Mirim, Mombuca, Monte Alegre do Sul, Monte Mor, Morungaba, Nazaré Paulista, Nova Odessa, Paulínia, Pedra Bela, Pedreira, Pinhalzinho, Piracaia, Piracicaba, Pirassununga, Rafard, Rio Claro, Rio das Pedras, Saltinho, Santa Bárbara D'Oeste, Santa Cruz da Conceição, Santa Cruz das Palmeiras, Santa Gertrudes, Santa Maria da Serra, Santo Antônio de Posse, Santo Antônio do Jardim, São João da Boa Vista, São José do Rio Pardo, São Pedro, São Sebastião da Grama, Serra Negra, Socorro, Sumaré, Tambaú, Tapiratiba, Torrinha, Tuiuti, Valinhos, Vargem, Vargem Grande do Sul, Várzea Paulista, Vinhedo, Estiva Gerbi

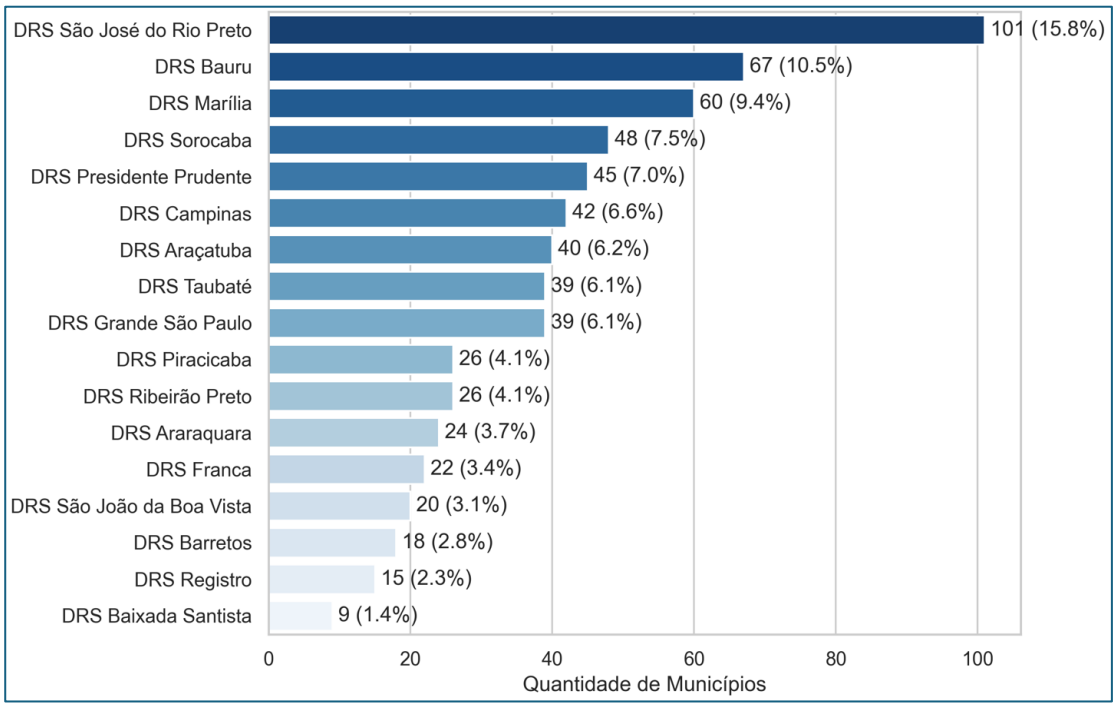
RA Central	Américo Brasiliense, Araraquara, Boa Esperança do Sul, Borborema, Cândido Rodrigues, Descalvado, Dobrada, Dourado, Fernando Prestes, Gavião Peixoto, Ibaté, Ibitinga, Itápolis, Matão, Motuca, Nova Europa, Porto Ferreira, Ribeirão Bonito, Rincão, Santa Ernestina, Santa Lúcia, Santa Rita do Passa Quatro, São Carlos, Tabatinga, Taquaritinga, Trabiju
RA Franca	Aramina, Batatais, Buritizal, Cristais Paulista, Franca, Guará, Igarapava, Ipuã, Itirapuã, Ituverava, Jeriquara, Miguelópolis, Morro Agudo, Nuporanga, Orlândia, Patrocínio Paulista, Pedregulho, Restinga, Ribeirão Corrente, Rifaina, Sales Oliveira, São Joaquim da Barra, São José da Bela Vista
RA Itapeva	Angatuba, Apiaí, Arandu, Barão de Antonina, Barra do Chapéu, Bom Sucesso de Itararé, Buri, Campina do Monte Alegre, Capão Bonito, Coronel Macedo, Fartura, Guapiara, Iporanga, Itaberá, Itaí, Itaoca, Itapeva, Itapirapuã Paulista, Itaporanga, Itararé, Nova Campina, Paranapanema, Piraju, Ribeira, Ribeirão Branco, Ribeirão Grande, Riversul, Sarutaiá, Taguaí, Taquarituba, Taquarivaí, Tejupá
RA Marília	Álvaro de Carvalho, Alvinlândia, Arco-Íris, Assis, Bastos, Bernardino de Campos, Borá, Campos Novos Paulista, Cândido Mota, Canitar, Cruzália, Echaporã, Espírito Santo do Turvo, Fernão, Florínea, Gália, Garça, Herculândia, Iacri, Ibirarema, Ipaussu, João Ramalho, Júlio Mesquita, Lupércio, Lutécia, Maracaí, Marília, Ocauçu, Óleo, Oriente, Oscar Bressane, Ourinhos, Palmital, Paraguaçu Paulista, Parapuã, Pedrinhas Paulista, Platina, Pompéia, Quatá, Queiroz, Quintana, Ribeirão do Sul, Rinópolis, Salto Grande, Santa Cruz do Rio Pardo, São Pedro do Turvo, Tarumã, Timburi, Tupã, Vera Cruz, Chavantes
RA Presidente Prudente	Adamantina, Alfredo Marcondes, Álvares Machado, Anhumas, Caiabu, Caiuá, Dracena, Emilianópolis, Estrela do Norte, Euclides da Cunha Paulista, Flora Rica, Flórida Paulista, Iepê, Indiana, Inúbia Paulista, Irapuru, Junqueirópolis, Lucélia, Marabá Paulista, Mariápolis, Martinópolis, Mirante do Paranapanema, Monte Castelo, Nantes, Narandiba, Nova Guataporanga, Osvaldo Cruz, Ouro Verde, Pacaembu, Panorama, Paulicéia, Piquerobi, Pirapozinho, Pracinha, Presidente Bernardes, Presidente Epitácio, Presidente Prudente, Presidente Venceslau, Rancharia, Regente Feijó, Ribeirão dos Índios, Rosana, Sagres, Salmourão, Sandovalina, Santa Mercedes, Santo Anastácio, Santo Expedito, São João do Pau dAlho, Taciba, Tarabai, Teodoro Sampaio, Tupi Paulista
RA Registro	Barra do Turvo, Cajati, Cananéia, Eldorado, Iguape, Ilha Comprida, Itariri, Jacupiranga, Juquiá, Miracatu, Pariquera-Açu, Pedro de Toledo, Registro, Sete Barras
RA Ribeirão Preto	Altinópolis, Barrinha, Brodowski, Cajuru, Cássia dos Coqueiros, Cravinhos, Dumont, Guariba, Guatapará, Jaboticabal, Jardinópolis, Luís Antônio, Monte Alto, Pitangueiras, Pontal, Pradópolis, Ribeirão Preto, Santa Cruz da Esperança, Santa Rosa de Viterbo, Santo Antônio da Alegria, São Simão, Serra Azul, Serrana, Sertãozinho, Taquaral

RA Santos	Bertioga, Cubatão, Guarujá, Itanhaém, Mongaguá, Peruíbe, Praia Grande, Santos, São Vicente
RA Sorocaba	Águas de Santa Bárbara, Alambari, Alumínio, Anhembi, Araçariguama, Araçoiaba da Serra, Areiópolis, Avaré, Bofete, Boituva, Botucatu, Capela do Alto, Cerqueira César, Cerquilho, Cesário Lange, Conchas, Guareí, Iaras, Ibiúna, Iperó, Itapetininga, Itatinga, Itu, Jumirim, Laranjal Paulista, Mairinque, Manduri, Pardinho, Pereiras, Piedade, Pilar do Sul, Porangaba, Porto Feliz, Pratânia, Quadra, Salto, Salto de Pirapora, São Manuel, São Miguel Arcanjo, São Roque, Sarapuí, Sorocaba, Tapiraí, Tatuí, Tietê, Torre de Pedra, Votorantim
RA São José do Rio Preto	Adolfo, Álvares Florence, Américo de Campos, Aparecida D'Oeste, Ariranha, Aspásia, Bady Bassitt, Bálsamo, Cardoso, Catanduva, Catiguá, Cedral, Cosmorama, Dirce Reis, Dolcinópolis, Elisiário, Estrela D'Oeste, Fernandópolis, Floreal, Guapiaçu, Guarani D'Oeste, Ibirá, Icém, Indiaporã, Ipiguá, Irapuã, Itajobi, Jaci, Jales, José Bonifácio, Macaubal, Macedônia, Magda, Marapoama, Marinópolis, Mendonça, Meridiano, Mesópolis, Mira Estrela, Mirassol, Mirassolândia, Monções, Monte Aprazível, Neves Paulista, Nhandeara, Nipoã, Nova Aliança, Nova Canaã Paulista, Nova Granada, Novais, Novo Horizonte, Onda Verde, Orindiúva, Ouroeste, Palestina, Palmares Paulista, Palmeira D'Oeste, Paraíso, Paranapuã, Parisi, Paulo de Faria, Pedranópolis, Pindorama, Planalto, Poloni, Pontalinda, Pontes Gestal, Populina, Potirendaba, Riolândia, Rubinéia, Sales, Santa Adélia, Santa Albertina, Santa Clara D'Oeste, Santa Fé do Sul, Santana da Ponte Pensa, Santa Rita D'Oeste, Santa Salete, São Francisco, São João das Duas Pontes, São José do Rio Preto, Sebastianópolis do Sul, Tabapuã, Tanabi, Três Fronteiras, Turmalina, Ubarana, Uchoa, União Paulista, Urânia, Urupês, Valentim Gentil, Vitória Brasil, Votuporanga, Zacarias
RA São José dos Campos	Aparecida, Arapeí, Areias, Bananal, Caçapava, Cachoeira Paulista, Campos do Jordão, Canas, Caraguatatuba, Cruzeiro, Cunha, Guaratinguetá, Igaratá, Ilhabela, Jacareí, Jambeiro, Lagoinha, Lavrinhas, Lorena, Monteiro Lobato, Natividade da Serra, Paraibuna, Pindamonhangaba, Piquete, Potim, Queluz, Redenção da Serra, Roseira, Santa Branca, Santo Antônio do Pinhal, São Bento do Sapucaí, São José do Barreiro, São José dos Campos, São Luiz do Paraitinga, São Sebastião, Silveiras, Taubaté, Tremembé, Ubatuba
RM São Paulo	Arujá, Barueri, Biritiba Mirim, Caieiras, Cajamar, Carapicuíba, Cotia, Diadema, Embu das Artes, Embu-Guaçu, Ferraz de Vasconcelos, Francisco Morato, Franco da Rocha, Guararema, Guarulhos, Itapecerica da Serra, Itapevi, Itaquaquecetuba, Jandira, Juquitiba, Mairiporã, Mauá, Mogi das Cruzes, Osasco, Pirapora do Bom Jesus, Poá, Ribeirão Pires, Rio Grande da Serra, Salesópolis, Santa Isabel, Santana de Parnaíba, Santo André, São Bernardo do Campo, São Caetano do Sul, São Lourenço da Serra, São Paulo, Suzano, Taboão da Serra, Vargem Grande Paulista

Fonte: Elaborado pelo autor (2026)

O Gráfico 2 e o Quadro 14 também foram produzidos com base no conjunto de dados explorado e mostram outra divisão importante do estado de São Paulo, agora por Departamentos Regionais de Saúde.

Gráfico 2 – Distribuição dos municípios por Departamentos Regionais de Saúde



Fonte: Elaborado pelo autor (2026)

Quadro 14 – Departamentos Regionais de Saúde e seus municípios

REGIÃO	MUNICÍPIOS
DRS Araraquara	Américo Brasiliense, Araraquara, Boa Esperança do Sul, Borborema, Cândido Rodrigues, Descalvado, Dobrada, Dourado, Gavião Peixoto, Ibaté, Ibitinga, Itápolis, Matão, Motuca, Nova Europa, Porto Ferreira, Ribeirão Bonito, Rincão, Santa Ernestina, Santa Lúcia, São Carlos, Tabatinga, Taquaritinga, Trabiju
DRS Araçatuba	Alto Alegre, Andradina, Araçatuba, Auriflama, Avanhandava, Barbosa, Bento de Abreu, Bilac, Birigui, Braúna, Brejo Alegre, Buritama, Castilho, Clementina, Coroados, Gabriel Monteiro, Glicério, Guaraçaí, Guararapes, Guzolândia, Ilha Solteira, Itapura, Lavínia, Lourdes, Luiziânia, Mirandópolis, Murutinga do Sul, Nova Castilho, Nova Independência, Nova Luzitânia, Penápolis, Pereira Barreto, Piacatu, Rubiácea, Santo Antônio do Aracanguá, Santópolis do Aguapeí, Sud Mennucci, Suzanápolis, Turiúba, Valparaíso
DRS Baixada Santista	Bertioga, Cubatão, Guarujá, Itanhaém, Mongaguá, Peruíbe, Praia Grande, Santos, São Vicente
DRS Barretos	Altair, Barretos, Bebedouro, Cajobi, Colina, Colômbia, Guaíra, Guaraci, Jaborandi, Monte Azul Paulista, Olímpia, Severínia,

	Taiaçu, Taiúva, Taquaral, Terra Roxa, Viradouro, Vista Alegre do Alto
DRS Bauru	Águas de Santa Bárbara, Agudos, Anhembi, Arandu, Arealva, Areiópolis, Avaí, Avaré, Balbinos, Barão de Antonina, Bariri, Barra Bonita, Bauru, Bocaina, Bofete, Boracéia, Borebi, Botucatu, Brotas, Cabrália Paulista, Cafelândia, Cerqueira César, Conchas, Coronel Macedo, Dois Córregos, Duartina, Fartura, Getulina, Guaiçara, Iacanga, Iaras, Igaraçu do Tietê, Itaí, Itaju, Itaporanga, Itapuí, Itatinga, Jaú, Laranjal Paulista, Lençóis Paulista, Lins, Lucianópolis, Macatuba, Manduri, Mineiros do Tietê, Paranapanema, Pardinho, Paulistânia, Pederneiras, Pereiras, Piraju, Pirajuí, Piratininga, Pongaí, Porangaba, Pratânia, Presidente Alves, Promissão, Reginópolis, Sabino, São Manuel, Sarutaiá, Taguaí, Taquarituba, Tejupá, Torre de Pedra, Torrinha, Uru
DRS Campinas	Águas de Lindóia, Americana, Amparo, Artur Nogueira, Atibaia, Bom Jesus dos Perdões, Bragança Paulista, Cabreúva, Campinas, Campo Limpo Paulista, Cosmópolis, Holambra, Hortolândia, Indaiatuba, Itatiba, Itupeva, Jaguariúna, Jarinu, Joanópolis, Jundiaí, Lindóia, Louveira, Monte Alegre do Sul, Monte Mor, Morungaba, Nazaré Paulista, Nova Odessa, Paulínia, Pedra Bela, Pedreira, Pinhalzinho, Piracaia, Santa Bárbara D'Oeste, Santo Antônio de Posse, Serra Negra, Socorro, Sumaré, Tuiuti, Valinhos, Vargem, Várzea Paulista, Vinhedo
DRS Franca	Aramina, Buritizal, Cristais Paulista, Franca, Guará, Igarapava, Ipuã, Itirapuã, Ituverava, Jeriquara, Miguelópolis, Morro Agudo, Nuporanga, Orlândia, Patrocínio Paulista, Pedregulho, Restinga, Ribeirão Corrente, Rifaina, Sales Oliveira, São Joaquim da Barra, São José da Bela Vista
DRS Grande São Paulo	Arujá, Barueri, Biritiba Mirim, Caieiras, Cajamar, Carapicuíba, Cotia, Diadema, Embu das Artes, Embu-Guaçu, Ferraz de Vasconcelos, Francisco Morato, Franco da Rocha, Guararema, Guarulhos, Itapecerica da Serra, Itapevi, Itaquaquecetuba, Jandira, Juquitiba, Mairiporã, Mauá, Mogi das Cruzes, Osasco, Pirapora do Bom Jesus, Poá, Ribeirão Pires, Rio Grande da Serra, Salesópolis, Santa Isabel, Santana de Parnaíba, Santo André, São Bernardo do Campo, São Caetano do Sul, São Lourenço da Serra, São Paulo, Suzano, Taboão da Serra, Vargem Grande Paulista
DRS Marília	Adamantina, Álvaro de Carvalho, Alvinlândia, Arco-Íris, Assis, Bastos, Bernardino de Campos, Borá, Campos Novos Paulista, Cândido Mota, Canitar, Cruzália, Echaporã, Espírito Santo do Turvo, Fernão, Flórida Paulista, Florínea, Gália, Garça, Guaimbê, Guarantã, Herculândia, Iacri, Ibirarema, Inúbia Paulista, Ipaussu, Júlio Mesquita, Lucélia, Lupércio, Lutécia, Maracaí, Mariápolis, Marília, Ocauçu, Óleo, Oriente, Oscar Bressane, Osvaldo Cruz, Ourinhos, Pacaembu, Palmital, Paraguaçu Paulista, Parapuã, Pedrinhas Paulista, Platina, Pompéia, Pracinha, Queiroz, Quintana, Ribeirão do Sul, Rinópolis, Sagres, Salmourão, Salto Grande, Santa Cruz do Rio Pardo, São Pedro do Turvo, Tarumã, Timburi, Tupã, Ubirajara, Vera Cruz, Chavantes

DRS Piracicaba	Águas de São Pedro, Analândia, Araras, Capivari, Charqueada, Conchal, Cordeirópolis, Corumbataí, Elias Fausto, Engenheiro Coelho, Ipeúna, Iracemápolis, Itirapina, Leme, Limeira, Mombuca, Piracicaba, Pirassununga, Rafard, Rio Claro, Rio das Pedras, Saltinho, Santa Cruz da Conceição, Santa Gertrudes, Santa Maria da Serra, São Pedro
DRS Presidente Prudente	Alfredo Marcondes, Álvares Machado, Anhumas, Caiabu, Caiuá, Dracena, Emilianópolis, Estrela do Norte, Euclides da Cunha Paulista, Flora Rica, Iepê, Indiana, Irapuru, João Ramalho, Junqueirópolis, Marabá Paulista, Martinópolis, Mirante do Paranapanema, Monte Castelo, Nantes, Narandiba, Nova Guataporanga, Ouro Verde, Panorama, Paulicéia, Piquerobi, Pirapozinho, Presidente Bernardes, Presidente Epitácio, Presidente Prudente, Presidente Venceslau, Quatá, Rancharia, Regente Feijó, Ribeirão dos Índios, Rosana, Sandovalina, Santa Mercedes, Santo Anastácio, Santo Expedito, São João do Pau dAlho, Taciba, Tarabai, Teodoro Sampaio, Tupi Paulista
DRS Registro	Barra do Turvo, Cajati, Cananéia, Eldorado, Iguape, Ilha Comprida, Iporanga, Itariri, Jacupiranga, Juquiá, Miracatu, Pariquera-Açu, Pedro de Toledo, Registro, Sete Barras
DRS Ribeirão Preto	Altinópolis, Barrinha, Batatais, Brodowski, Cajuru, Cássia dos Coqueiros, Cravinhos, Dumont, Guariba, Guatapará, Jaboticabal, Jardinópolis, Luís Antônio, Monte Alto, Pitangueiras, Pontal, Pradópolis, Ribeirão Preto, Santa Cruz da Esperança, Santa Rita do Passa Quatro, Santa Rosa de Viterbo, Santo Antônio da Alegria, São Simão, Serra Azul, Serrana, Sertãozinho
DRS Sorocaba	Alambari, Alumínio, Angatuba, Apiaí, Araçariguama, Araçoiaba da Serra, Barra do Chapéu, Boituva, Bom Sucesso de Itararé, Buri, Campina do Monte Alegre, Capão Bonito, Capela do Alto, Cerquilho, Cesário Lange, Guapiara, Guareí, Ibiúna, Iperó, Itaberá, Itaoca, Itapetininga, Itapeva, Itapirapuã Paulista, Itararé, Itu, Jumirim, Mairinque, Nova Campina, Piedade, Pilar do Sul, Porto Feliz, Quadra, Ribeira, Ribeirão Branco, Ribeirão Grande, Riversul, Salto, Salto de Pirapora, São Miguel Arcanjo, São Roque, Sarapuí, Sorocaba, Tapiraí, Taquarivaí, Tatuí, Tietê, Votorantim
DRS São José do Rio Preto	Adolfo, Álvares Florence, Américo de Campos, Aparecida D'Oeste, Ariranha, Aspásia, Bady Bassitt, Bálsamo, Cardoso, Catanduva, Catiguá, Cedral, Cosmorama, Dirce Reis, Dolcinópolis, Elisiário, Embaúba, Estrela D'Oeste, Fernandópolis, Fernando Prestes, Floreal, Gastão Vidigal, General Salgado, Guapiaçu, Guarani D'Oeste, Ibirá, Icém, Indiaporã, Ipiguá, Irapuã, Itajobi, Jaci, Jales, José Bonifácio, Macaubal, Macedônia, Magda, Marapoama, Marinópolis, Mendonça, Meridiano, Mesópolis, Mira Estrela, Mirassol, Mirassolândia, Monções, Monte Aprazível, Neves Paulista, Nhandeara, Nipoã, Nova Aliança, Nova Canaã Paulista, Nova Granada, Novais, Novo Horizonte, Onda Verde, Orindiúva, Ouroeste, Palestina, Palmares Paulista, Palmeira D'Oeste, Paraíso, Paranapuã, Parisi, Paulo de Faria, Pedranópolis, Pindorama, Pirangi, Planalto, Poloni, Pontalinda, Pontes Gestal, Populina, Potirendaba, Riolândia, Rubinéia, Sales, Santa Adélia, Santa Albertina, Santa Clara D'Oeste, Santa Fé do

	Sul, Santana da Ponte Pensa, Santa Rita D'Oeste, Santa Salete, São Francisco, São João das Duas Pontes, São João de Iracema, São José do Rio Preto, Sebastianópolis do Sul, Tabapuã, Tanabi, Três Fronteiras, Turmalina, Ubarana, Uchoa, União Paulista, Urânia, Urupês, Valentim Gentil, Vitória Brasil, Votuporanga, Zacarias
DRS São João da Boa Vista	Aguaí, Águas da Prata, Caconde, Casa Branca, Divinolândia, Espírito Santo do Pinhal, Itapira, Itobi, Mococa, Mogi Guaçu, Mogi Mirim, Santa Cruz das Palmeiras, Santo Antônio do Jardim, São João da Boa Vista, São José do Rio Pardo, São Sebastião da Grama, Tambaú, Tapiratiba, Vargem Grande do Sul, Estiva Gerbi

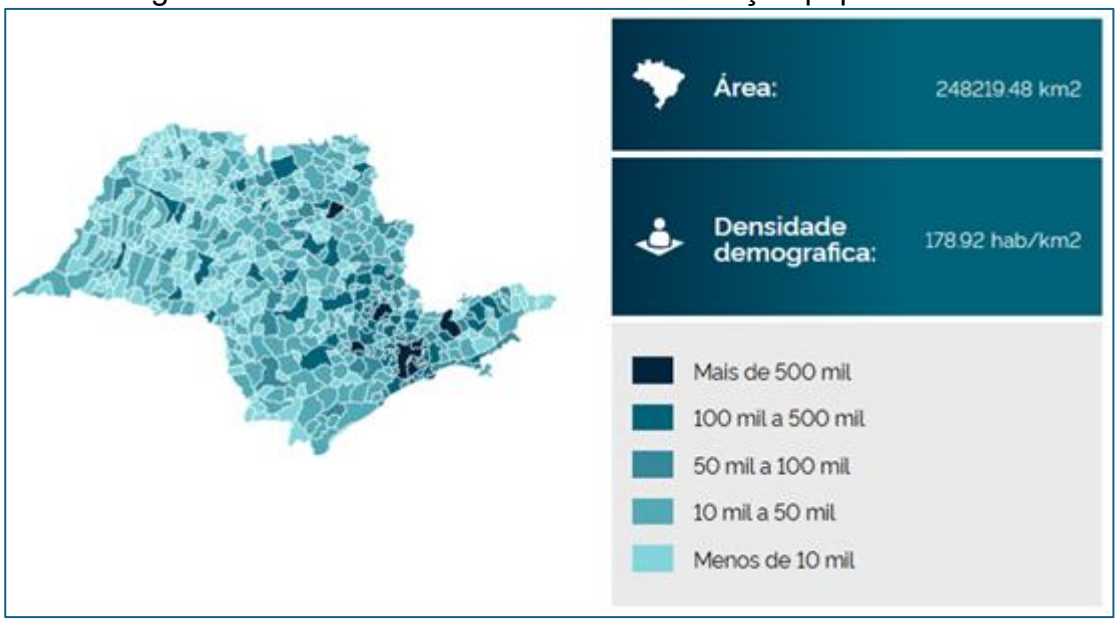
Fonte: Elaborado pelo autor (2026)

O estado de São Paulo, composto pelos seus 645 municípios, segundo o IBGE (Instituto Brasileiro de Geografia e Estatística) (IBGE, 2023b), está distribuído em uma área territorial de 248,2 milhões de Km², segundo dados do Censo 2022. É uma área que corresponde à aproximadamente 3% da extensão territorial do Brasil (IBGE, 2023a), ocupando a 12ª posição entre os estados do país (IBGE, 2023a).

Segundo dados da Fundação SEADE (Fundação Sistema Estadual de Análise de Dados), possuía em 2024, uma população total de 44,6 milhões de habitantes (SEADE, 2024), correspondendo à aproximadamente 22% da população total brasileira, sendo o 1º estado do ranking de concentração populacional (IBGE, 2023c). Segundo o IBGE, estima-se que, em 2025, haveria um crescimento populacional no estado de São Paulo de aproximadamente 3%, saltando para 46,1 milhões de habitantes (IBGE, 2023b).

Na Figura 8 é possível observar a distribuição populacional dos municípios do estado, dentro de faixas para melhor análise, assim como a informação da densidade demográfica, que mede o número médio de habitantes por km^2. O estado de São Paulo possui ainda uma densidade demográfica de 178,9 habitantes por km^2, aproximadamente 7,5 vezes superior à média do Brasil, que é de 23,8 habitantes por km^2 (IBGE, 2023c).

Figura 8 – Estado de São Paulo: Concentração populacional



Fonte: IBGE, 2023c

Em 2023, os municípios com até 10 mil habitantes no estado de São Paulo constituíam a maior parcela, 42,3%, apesar de concentrarem percentual bem reduzido da população estadual, 3,1%. Já aqueles com mais de 400 mil habitantes representavam somente 2,3% do total de municípios, mas abrigavam quase metade dos residentes no Estado, 47% (SEADE, 2024). A Tabela 1 mostra uma comparação entre os anos de 2000 e 2023 com suas respectivas distribuições.

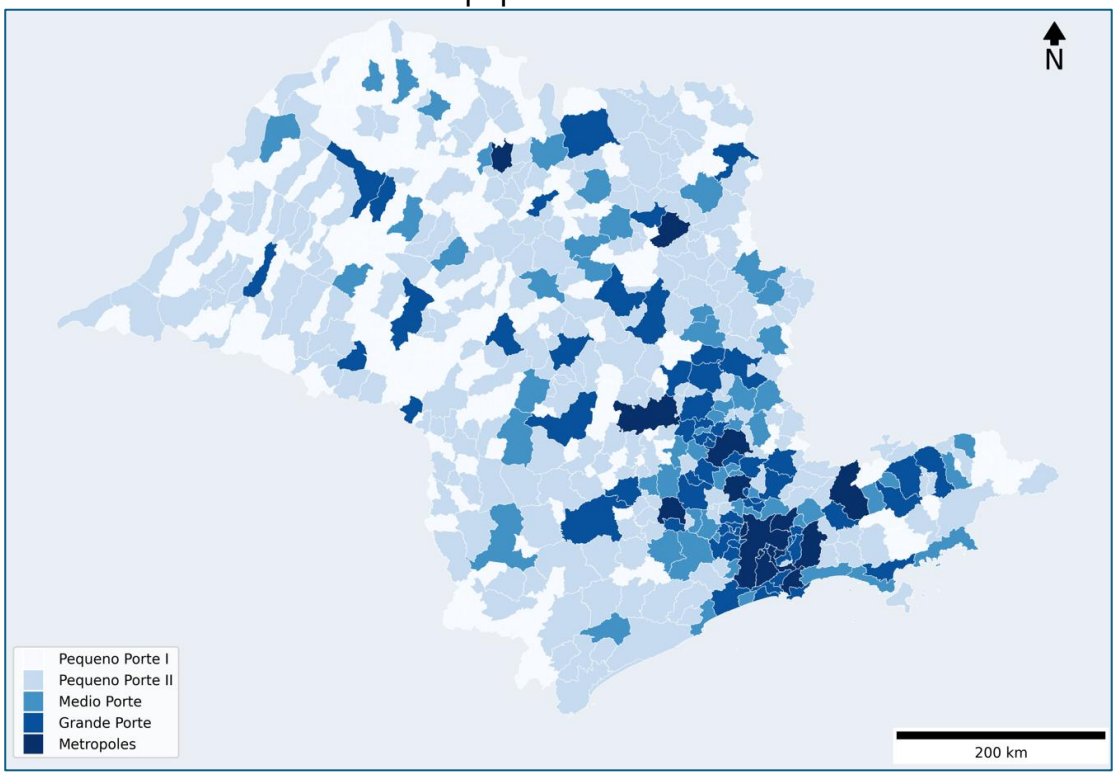
Tabela 1 – Distribuição populacional dos municípios de São Paulo

MUNICÍPIOS	População		Proporção da população (%)		Nº de municípios		Proporção de municípios (%)	
	2000	2023	2000	2023	2000	2023	2000	2023
Estado de São Paulo	36.974.378	44.539.225	100	100	645	645	100	100
Até 10 mil	1.397.860	1.373.853	3,78	3,08	295	273	45,74	42,33
De 10 mil a 50 mil	5.275.979	5.357.258	14,27	12,03	234	235	36,28	36,43
De 50 mil a 100 mil	3.818.498	4.160.522	10,33	9,34	54	59	8,37	9,15
De 100 mil a 400 mil	10.062.716	12.700.183	27,22	28,51	52	63	8,06	9,77
Mais de 400 mil	16.419.325	20.947.409	44,41	47,03	10	15	1,55	2,33

Fonte: Adaptado de SEADE, 2024.

Também com base no conjunto de dados explorado, a Figura 9 foi produzida para ilustrar a distribuição dos municípios, segundo seus portes, classificados conforme legenda demonstrada.

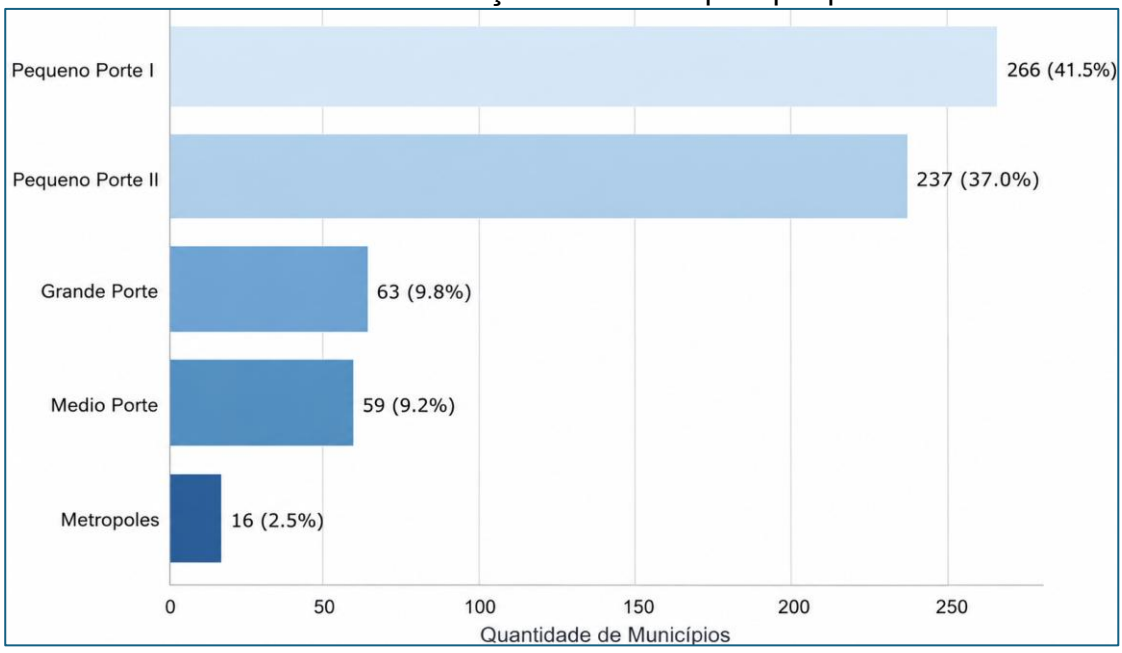
Figura 9 – Visão dos municípios do estado de São Paulo segundo porte populacional



Fonte: Elaborado pelo autor (2026)

O Gráfico 3 mostra a distribuição dos municípios do estado de São Paulo, segundo seus portes, que leva em consideração a população.

Gráfico 3 – Distribuição dos municípios por porte



Fonte: Elaborado pelo autor (2026)

A idade média da população do estado de São Paulo vem se elevando gradativamente e passou de 26,1 para 37,1 anos, entre 1980 e 2022, com ampliação de 11 anos nesse período. (SEADE, 2024)

Outras características importantes do estado de São Paulo dizem respeito aos aspectos econômicos e de renda da populacional. O estado apresentou, em 2023, um PIB (Produto Interno Bruto) de 3,4 trilhões, o que responde à aproximadamente 31% do PIB brasileiro no mesmo ano, PIB *per capita*, número que se refere a cada habitante, de 60,5 mil, superando a média brasileira em 17%, e um salário médio por habitante de 4,5 mil, 32% acima da média nacional (IBGE, 2025; SEADE, 2024).

4.2. Obtenção dos dados

Reestabelecendo a temática central da pesquisa, buscou-se conectar a cobertura vacinal da BCG e Tríplice Viral (SCR) nos Municípios do estado de São Paulo aos possíveis fatores externos, ou seja, que poderiam afetar seu declínio ao longo do período estudado, os anos de 2016 a 2023.

Além dos dados sobre as doses aplicadas dos imunobiológicos (vacinas) para os cálculos das coberturas vacinais, os fatores externos analisados foram compostos de variáveis separadas por tipo (ou grupos), as socioeconômicas, as demográficas, de saúde, educacionais e sobre a infraestrutura de saúde dos municípios, coletadas de fontes de dados públicos, conforme relacionado no Quadro 5.

É preciso destacar que todos os dados utilizados nesta pesquisa são secundários e de domínio público, em sua grande maioria, disponibilizados no formato de arquivos texto do tipo “CSV” (*Comma-Separated Values* ou Valores Separados por Vírgula, em tradução livre), um padrão bastante utilizado para publicação de dados tabulares, definido primariamente por Yakov Shafranovich (Shafranovich, 2005).

O processo de carregamento, agrupamento e, na maior parte dos casos, conversão de tipos e transformação dos dados, está detalhado no processo tópico.

Quadro 15 – Fontes utilizadas para obtenção dos dados

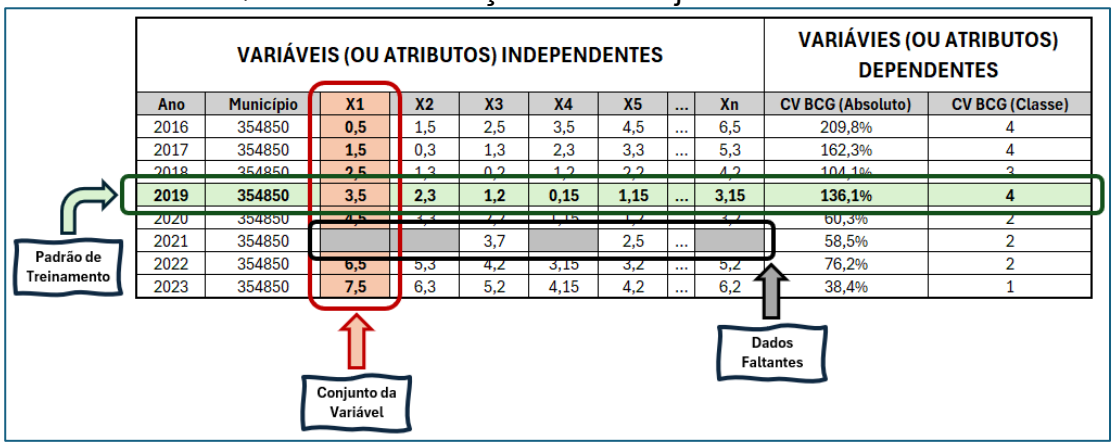
GRUPO DO DADO	DESCRIÇÃO	FONTE	PERÍODO
DEMOGRÁFICO	Projeção Populacional	DATASUS – Tecnologia da Informação a Serviço do SUS	2016 – 2023
ECONÔMICO	Salário-Mínimo Anual	Instituto de Pesquisa Econômica Aplicada – IPEA	2016 – 2023
	PIB – Produto Interno Bruto	Sistema IBGE de Recuperação Automática – SIDRA	2016 – 2023
EDUCACIONAL	IDEB – Índice de Desenvolvimento da Educação Básica	Ministério da Educação	2016 – 2023
INFRAESTRURA	CNES – Recursos Físicos	DATASUS – Tecnologia da Informação a Serviço do SUS	2016 – 2023
	CNES – Estabelecimentos	DATASUS – Tecnologia da Informação a Serviço do SUS	2016 – 2023
SAÚDE	Doses aplicadas dos Imunobiológicos	DATASUS – Tecnologia da Informação a Serviço do SUS	2016 – 2022
	Doses aplicadas dos Imunobiológicos	Ministério da Saúde – Vacinação do Calendário Nacional	2023
	Cobertura da atenção básica	e-Gestor Atenção Primária à Saúde	2016 – 2020
	Cobertura da atenção primária	e-Gestor Atenção Primária à Saúde	2021 – 2023
	Nascidos Vivos	DATASUS – Tecnologia da Informação a Serviço do SUS	2016 – 2023
	Cobertura de Pré-Natal	DATASUS – Tecnologia da Informação a Serviço do SUS	2016 – 2023
	Óbitos Infantis	DATASUS – Tecnologia da Informação a Serviço do SUS	2016 – 2023
	Morbidade Hospitalar do Sus	DATASUS – Tecnologia da Informação a Serviço do SUS	2016 – 2023
	IPPI – Índice Paulista da Primeira Infância	SEADE – Fundação Sistema Estadual de Análise de Dados	2016 – 2023
	IDH – Índice de Desenvolvimento Humano (Estados)	Atlas do Desenvolvimento Humano no Brasil	2010 – 2021
SOCIAL	IDHM – Índice de Desenvolvimento Humano dos Municípios	Atlas do Desenvolvimento Humano no Brasil	2010
	IPDM – Índice Paulista de Desenvolvimento Municipal	SEADE – Fundação Sistema Estadual de Análise de Dados	2014 – 2022
	Dados de saneamento (água e esgotamento sanitário)	SEADE – Fundação Sistema Estadual de Análise de Dados	2017 – 2021
SÓCIOECONÔMICO e DEMOGRÁFICO	Censo demográfico 2022	Sistema IBGE de Recuperação Automática – SIDRA	2022
	Censo demográfico 2010	Sistema IBGE de Recuperação Automática – SIDRA	2010

Fonte: Elaborado pelo autor (2026)

4.3. Preparação dos dados

O resultado esperado e necessário do processo de preparação de dados é um único conjunto de dados onde cada linha deste conjunto (também conhecida como padrão) deve conter todas as variáveis dependentes e independentes alinhadas, ou seja, ocupando uma única dimensão, como exemplificado no Quadro 16.

Quadro 16 – Esboço de um conjunto de dados



Fonte: Elaborado pelo autor (2026).

4.3.1. Ferramentas e bibliotecas de software

Como destacado no tópico anterior, os dados coletados foram obtidos no formato “CSV” (arquivos texto) com tipos, disposições internas, formas de agrupamento e dimensões, muitas vezes, distintas. Para possibilitar o processo de manuseio destes dados, carga e os ajustes necessários, algumas ferramentas e bibliotecas de software foram empregadas no processo.

Não faz parte do escopo desta pesquisa, detalhar os princípios de funcionamento, detalhes técnicos ou origens de cada uma dessas ferramentas ou bibliotecas de software.

O Quadro 17 mostra a relação completa das ferramentas e bibliotecas de software utilizadas para carregamento, concentração, agrupamento, análise de consistência, conversão de tipos, transformação e visualização dos dados

secundários, bem como as utilizadas para aplicação dos modelos de classificação de ML e análise de resultados. A escolha das ferramentas deu-se, principalmente, por afinidade e experiência do autor no uso de cada uma delas.

Quadro 17 – Relação de ferramentas e bibliotecas de software empregadas na pesquisa

FERRAMENTA	DESCRIÇÃO	OBJETIVO	VERSÃO	REFERÊNCIA
SQLite	Trata-se de software de banco de dados que suporte linguagem de consulta SQL (Structured Query Language ou Linguagem Estruturada de Pesquisa) de grande utilização pela facilidade de uso (Machado, 2014).	Concentrar e agrupar todos os dados em tabelas	3.49.0	https://sqlite.org/about.html
Dbeaver Community	Ferramenta de gerenciamento de banco de dados gratuita e de código aberto para projetos pessoais.	Criar e manipular os bancos de dados SQLite	25.0.0	https://dbeaver.io/about/
Visual Studio Code	Editor de programas de computador gratuito	Escrever os programas em linguagem Python	1.97	https://code.visualstudio.com/
Python	Linguagem de programação	Criar os programas utilizados para analisar de dados	3.12.3	https://www.python.org/about/
Pandas	Biblioteca Python para análise e manipulação de dados	Carregar e manipular os dados	2.2.3	https://pandas.pydata.org/about/
GeoPandas	Biblioteca Python para gráficos coropléticos baseados em *shapefiles*	Plotar gráficos para análise geoespacial	1.1.2	https://geopandas.org/©/stable/
Matplotlib	Biblioteca Python para criar visualizações de dados	Plotar gráficos para análise dos dados	3.10.1	https://matplotlib.org/
Seaborn	Biblioteca Python para visualização de dados e criação de gráficos estatísticos	Plotar gráficos para análise dos dados	0.13.2	https://seaborn.pydata.org/
SciPy	Biblioteca Python com algoritmos para otimização e estatística	Calcular desvios, outliers e outros	1.15.2	https://scipy.org/

Imbalaced Learn	Biblioteca Python que fornece ferramentas para lidar com classificação em classes desbalanceadas.	Balancear classes em conjuntos de dados	0.13.0	https://imbalanced-learn.org/stable/
SHAP	Biblioteca Python que utiliza a abordagem da teoria dos jogos para explicar a saída dos modelos de aprendizado de máquina	Analisar a importâncias das variáveis para os modelos de aprendizado	0.47.1	https://shap.readthedocs.io/©/latest/
LIME	Biblioteca Python usada para explicar as previsões de um classificador ou regressor.	Analisar a importâncias das variáveis para os modelos de aprendizado	0.2.0	https://lime-ml.readthedocs.io/
Ydata Profiling	Biblioteca Python para a análise de qualidade de dados e a análise exploratória de dados	Trabalhar a etapa de análise descritiva e exploratória dos dados	4.16.1	https://docs.profiling.ydata.ai/latest/
Scikit-learn	Biblioteca Python base aprendizado de máquina, com aprendizado supervisionado e não supervisionado, classificação e regressão, entre outros.	Aplicar e analisar o resultado de diversos modelos de aprendizado	1.6.1	https://scikit-learn.org/stable/
XGBoost	Biblioteca Python que implementa *Gradient Boosting* como estrutura para os modelos de aprendizado	Treinar e analisar resultado com o modelo de aprendizado	3.0.0	https://xgboost.readthedocs.io/
LightGBM	Biblioteca Python que implementa *Gradient Boosting* como estrutura para os modelos de aprendizado	Treinar e analisar resultado com o modelo de aprendizado	4.6.0	https://lightgbm.readthedocs.io/

Fonte: Elaborado pelo autor (2026)

4.3.2. Processo de carga de dados

Para o agrupamento de todos os dados secundários coletados, foi criado um modelo de banco de dados com a ferramenta *Dbeaver* utilizando o sistema gerenciador de banco de dados *SQLite*, objetivando concentrá-los em uma única base para facilitar o manuseio, a manutenção e a análise dos dados, tanto com linguagens de acesso aos dados diretamente em bandos de dados, como favorecendo as consultas com as bibliotecas da linguagem Python já elencadas no Quadro 17.

Como exibido no Quadro 15, as fontes dos dados analisados foram diversas, sendo que algumas já forneciam todo o período que foi foco da análise, ou seja, os dados das variáveis independentes já contemplavam de 2016 a 2023, mas, em determinados casos, dados faltantes existiam e precisaram ser tratados ou produzidos.

A primeira parte do processo de carregamento dos dados coletados concentrou-se no cenário onde não houve detecção de dados faltantes, onde o processo de carregamento deu-se de maneira direta e mais simplificada, apenas com atenção aos tipos de dados, evitando problemas de conversão e consistência, as checagem de eventuais necessidades de transformações e os ajustes nas disposições das variáveis (linhas versus colunas).

Conteúdos obtidos de maneira direta, sem necessidade de atenção às possíveis informações ausentes, estão mostrados no Quadro 18.

Quadro 18 – Dados completos sem informações ou faltantes

GRUPO	DESCRIÇÃO	TABELA SQLite
DEMOGRÁFICO	Projeção Populacional	POPULACAO
ECONÔMICO	PIB – Produto Interno Bruto	PIB
	Salário-Mínimo Anual	SALARIO
EDUCACIONAL	IDEB – Índice de Desenvolvimento da Educação Básica	IDEB
INFRAESTRURA	CNES – Recursos Físicos	CNES_RECURSO
	CNES – Estabelecimentos	CNES_ESTABELECIMENTO
SAÚDE	Doses Aplicadas dos Imunológicos	VACINACAO
	Cobertura da Atenção Básica e Primária	AB
	Nascidos Vivos	NASCIDOSVIVOS
	Cobertura de Pré-Natal	NASCIDOSVIVOS
	Óbitos Infantis	OBITOINFANTIL
	Morbidade Hospitalar do SUS	MORBIDADE
	IPPI – Índice Paulista da Primeira Infância	IPPI_SP

Fonte: Elaborado pelo autor (2026)

4.3.2.1. Ajustes e carga de dados faltantes

A segunda parte do processo de carga de dados foi direcionada aos conjuntos de variáveis com dados faltantes (ou ausentes). Lidar com dados faltantes é de extrema importância na ciência de dados, pois representa um desafio significativo com potencial para afetar negativamente os processos de tomada de decisão e os resultados de pesquisa (Zhou, 2024).

É comum separar os dados faltantes em categorias, sendo: dados faltantes completamente ao acaso (*Missing completely at Random* – MCAR), onde os dados estão ausentes por razões que não estão relacionadas a quaisquer características; dados faltantes ao acaso (*Missing at Random* – MAR), onde os dados não estão faltando aleatoriamente, mas a probabilidade de um valor estar faltando depende dos valores das variáveis que foram efetivamente medidas; dados faltantes não ao acaso, mas informativos (Not missing at Random – NMAR ou *Informative missing* – IM): a tendência de uma variável estar ausente é uma função dos dados que podem não estar disponíveis, de fato, no momento da coleta (Harrell, 2015).

Esta pesquisa se deparou com a terceira categoria, a dos dados faltantes informativos, ou seja, a ausência factual para alguns anos específicos e que estavam dentro do período de análise.

As técnicas comumente usadas para lidar com o problema dos dados faltantes são diversas, como, por exemplo, preenchimento com zeros, exclusão completa dos padrões de treinamento, e métodos de imputação, incluindo média, mediana, moda, regressão, k-vizinhos mais próximos (KNN) e sinalização (Atoum, 2025).

Analisou-se cada um dos conjuntos de variáveis que necessitavam complementação, aplicando-se técnicas de ajustes e geração de dados específicas para cada caso, como detalhado a seguir.

- **IDHM – Índice de Desenvolvimento Humano Municipal –** Fonte de dados: Atlas do Desenvolvimento Humano no Brasil

Os dados coletados continham, para os municípios brasileiros, somente o ano de 2010. Porém, para os estados, a fonte dispunha do período de 2010 a 2021.

O IDHM é composto por três dimensões: renda, educação e longevidade. O valor do IDHM é a média dessas três dimensões.

Para a geração das quatro dimensões do conjunto IDHM citadas, no período de análise, 2016 a 2023, adotou-se a seguinte estratégia:

1. Utilizando o período de 2010 a 2021 dos estados, fez-se a projeção dos valores para os anos de 2022 e 2023, comparando duas estratégias de Regressão[2], a linear e a polinomial (grau 3, testado e já ajustado), optando pela que melhor acomodou a projeção;
2. Com o conjunto completo dos dados dos estados contendo de 2016 a 2023, foi feito o cálculo das diferenças (positivas ou negativas) entre os anos;

[2] Um problema de regressão é aquele em que os valores alvo pertencem ao conjunto dos números reais. A abordagem linear segue a premissa de que os valores têm relação direta e linear, enquanto a polinomial admite sinuosidade e até mesmo inversões (Ellis, 2024).

3. Adotou-se os valores do ano de 2010 como base para os municípios no ano de 2016 e calculou-se os demais anos com os valores apurados das diferenças dos estados[3].

- **IPDM – Índice Paulista de Desenvolvimento Municipal –** Fonte de dados: SEADE – Fundação Sistema Estadual de Análise de Dados

Esse índice é disponibilizado a cada dois anos (bienal), portanto, apresentando ausência de informação para os anos de 2017, 2019, 2021 e 2023. Duas estratégias distintas foram utilizadas:

1. Anos de 2017, 2019 e 2021: por encontrar-se em intervalos com dois valores disponíveis nos anos anteriores e posteriores, optou-se por utilizar a média simples;
2. Ano de 2023: com base na série completa e gerada no item anterior, optou-se por adotar as mesmas estratégias de regressão (linear e polinomial), já utilizadas para a projeção do IDHM.

- **Dados de saneamento (água e esgotamento sanitário) –** Fonte de dados: SEADE – Fundação Sistema Estadual de Análise de Dados

Os dados de saneamento estavam disponíveis para o período de 2017 a 2021, ou seja, ausentes para os anos de 2016, 2022 e 2023, necessários para o processo de predição.

Optou-se, também, por adotar as mesmas estratégias de regressão (linear e polinomial), já utilizadas para a projeção do IDHM e do IPDM, discutidas anteriormente.

- **Censo Demográfico 2010 e 2022 –** Fonte de dados: Sistema IBGE de Recuperação Automática – SIDRA

Os dados censitários contribuíram com grande parte das variáveis dependentes analisadas. Porém, dados prontos para carga direta só estavam disponíveis nos

[3] Apesar do estudo ter foco no estado de São Paulo, foram utilizados todos os estados brasileiros para os ajustes dos modelos.

anos do Censo Demográfico, 2010 e 2022, deixando muitas lacunas para os demais períodos. As estratégias para utilização dos dados censitários foram:

1. Cópia literal dos valores das variáveis do ano de 2010 para os anos de 2016, 2017, 2018 e 2019 e do ano de 2022 para 2020, 2021 e 2023;
2. Ajuste das variáveis com base na população de cada município, oriunda de outra fonte de dados, de acordo com sua representatividade relativa ao dado de origem (2010 ou 2022);
3. Ajustes em variáveis específicas, tais como, “Índice de Envelhecimento”, com base na melhor variável semântica correspondente, como a população com mais de 60 anos de idade, neste exemplo;
4. Ajustes nas variáveis representativas de faixas de rendimento com base em variáveis que representam crescimento econômico (PIB *per capita*) e melhora das condições sociais (IDHM), além das variações salariais.

Finalizando a construção do conjunto de dados, a última atenção foi à variável dependente, a Cobertura Vacinal (CV).

Como já observado, a CV é demonstrada como um valor percentual para cada município e ano, e, como não está disponível como um dado pronto nos repositórios de dados secundários utilizados, foi necessário seu cálculo com base nas doses de cada imunológico.

O cálculo da cobertura vacinal é feito pelo número de doses aplicadas para determinada vacina no período e local estabelecidos, sobre o denominador de nascidos vivos naquele ano e local (Brasil, 2019).

As doses aplicadas são obtidas pelo Sistema de Informação do PNI (SI-PNI), sendo usadas como numerador (doses aplicadas de cada antígeno ou vacina num determinado local e ano) podendo ser utilizado filtro de idade, vacina, local e ano e, no denominador, a população-alvo por meio da população estimada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) ou os nascidos vivos pelo Sistema de Informações sobre Nascidos Vivos (SINASC) (Domingues, 2013) (Pamplona, 2020).

O formato de cálculo da cobertura utiliza os dados coletados das doses aplicadas de cada imunobiológico e dados demográficos (Quadro 15), separados por município e ano de análise, como detalhado no tópico **1.2.3** deste trabalho. A Cobertura Vacinal (CV) para a vacina BCG considera as doses aplicadas da vacina no ano (Do) e os nascidos vivos no mesmo ano (Na), como segue:

$$CV(BCG) = \frac{Do}{Na}$$

A Cobertura Vacinal da Tríplice Viral (SCR) considera, além das variáveis já utilizadas para o cálculo da BCG, o número de óbitos infantis no ano anterior ao cálculo (Ob), como segue:

$$CV(SCR) = \frac{Do}{Na - Ob}$$

Ao final do processo de cálculo, os valores das coberturas vacinais foram armazenados em tabelas específicas e as devidas classificações também foram calculadas e armazenadas.

As classificações das coberturas obedeceram ao demonstrado nos Quadro 19 e 20, seguindo os padrões já adotados pela literatura, separados cada vicina analisada (BCG e SCR), uma vez que possuem metas diferentes para Cobertura Vacinal “Adequada”, 90% para BCG e 95% para a Tríplice Viral (SCR), segundo o PNI.

Quadro 19 – Quatro classes da cobertura vacinal da BCG

CLASSE	DESCRIÇÃO	DE	ATÉ
1	Muito baixa	< 50%	
2	Baixa	50%	< 90%
3	Adequada	90%	< 120%
4	Elevada	> 120%	

Fonte: Elaborado pelo autor (2026)

Quadro 20 – Quatro classes da cobertura vacinal da Tríplice Viral (SCR)

CLASSE	DESCRIÇÃO	DE	ATÉ
1	Muito baixa	< 50%	
2	Baixa	50%	< 95%
3	Adequada	95%	< 120%
4	Elevada	> 120%	

Fonte: Elaborado pelo autor (2026)

Para as classes a serem utilizadas para predição nos modelos de ML (que serão demonstrados nos próximos tópicos), foram atribuídas novas faixas de classificação, como mostram os Quadro 21 e 22.

Quadro 21 – Modelo de duas classes da cobertura vacinal da BCG

CLASSE	DESCRIÇÃO	DE	ATÉ
0	Não adequada (Baixa)	< 90%	
1	Adequada	>= 90%	

Fonte: Elaborado pelo autor (2026)

Quadro 22 – Modelo de duas classes da cobertura vacinal da Tríplice Viral (SCR)

CLASSE	DESCRIÇÃO	DE	ATÉ
0	Não adequada (Baixa)	< 95%	
1	Adequada	>= 95%	

Fonte: Elaborado pelo autor (2026)

A composição final do banco de dados, com todas as variáveis, dependentes e independentes, dispostas como padrões de treinamento, ficaram com as seguintes características:

- 11 atributos apenas informativos, utilizados para base, controle e visualizações;
- 4 variáveis dependentes, correspondentes às coberturas vacinais absolutas e suas respectivas classes, utilizadas nos modelos de forma individual;

- 293 variáveis independentes, candidatas à utilização direta nos modelos de ML, ou para posterior agrupamento e engenharia de atributos, caso necessário;
- 5160 padrões de treinamento (registros ou linhas), contemplando os 645 municípios do estado de São Paulo no período de 8 anos, 2016 a 2023.

A Tabela 2 mostra as divisões das variáveis independentes em grupos, facilitando a compreensão das suas principais características e origens.

Tabela 2 – Grupos das variáveis independentes (totais)

GRUPOS DAS VARIÁVEIS INDEPENDENTES		
GRUPO	**QUANTIDADE**	**(%)**
DEMOGRAFICO	73	24,9
ECONOMICO	37	12,6
EDUCACIONAL	21	7,2
SAUDE	74	25,3
SOCIAL	88	30,0
TOTAL	**293**	**100,0**

Fonte: Elaborado pelo autor (2026)

4.4. Análise estatística preliminar

Com base o conjunto de dados já construído, passou-se para a análise estatística preliminar, com foco na checagem de consistência e coerência dos dados, utilizando as ferramentas e bibliotecas de software relacionadas no Quadro 15. Para o desenvolvimento das análises utilizou-se a linguagem *Python*, com ambiente de desenvolvimento do *Visual Studio Code*, incluindo o uso das bibliotecas *Pandas, Seaborn, Matplotlib, SciPy, Ydata Profiling e Scikit-learn*.

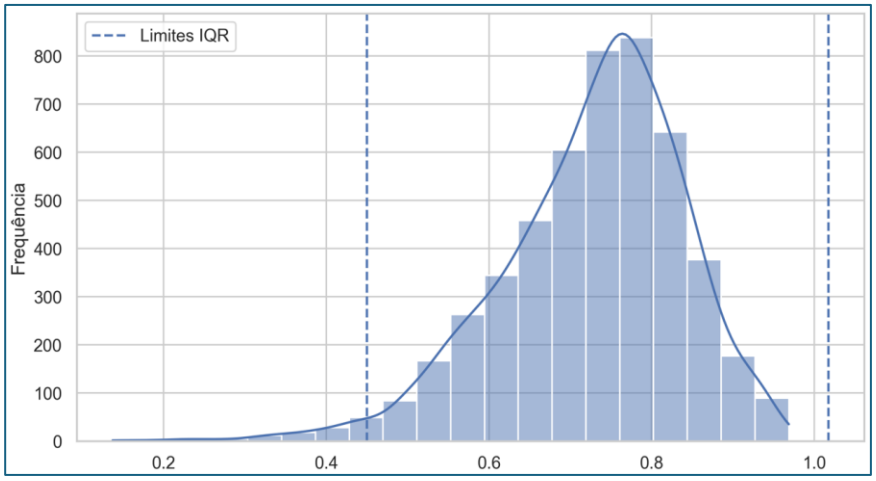
Essa etapa contemplou as fases abaixo, nesta ordem:

- Análise de valores discrepantes (*outliers*), considerando o padrão de distribuição de cada variável, sendo:

  - Variáveis com distribuição normal[4] (exemplo no Gráfico 4): dados ajustados utilizando percentil inferior e superior: [0,10; 0,90];
  - Variáveis com distribuição não-normal (*skewed* ou enviesado, em tradução livre) (exemplo no Gráfico 5): ajustes utilizando Intervalo Interquartil (IQR): [0,25; 0,75];
- Padronização e/ou normalização dos dados: cada variável foi padronizada dentro do intervalo [0, 1] seguindo a estratégia de Normalização Mínimos e Máximos (*MinMaxScaler* ou Normalizador Min-Max);
- Análise de Variância: variáveis com variância menor ou igual a 0,01 foram considerados sem Ganho de Informação[5] e, portanto, candidatas a remoção, após análise semântica, ou seja, se apresentavam ou não relação com a pesquisa ou domínio;
- Análise de Correlação: foi utilizado valores de correlação acima de 0,95 como indicador de alta correlação[6], sugerindo variáveis também candidatas a remoção, também após análise semântica;
- Análise Semântica: após aplicadas técnicas estatísticas já mencionadas, as variáveis candidatas à remoção foram analisadas quanto a seu valor semântico, ou seja, sua representatividade no domínio de pesquisa, a cobertura vacinal e a área da saúde, em si.

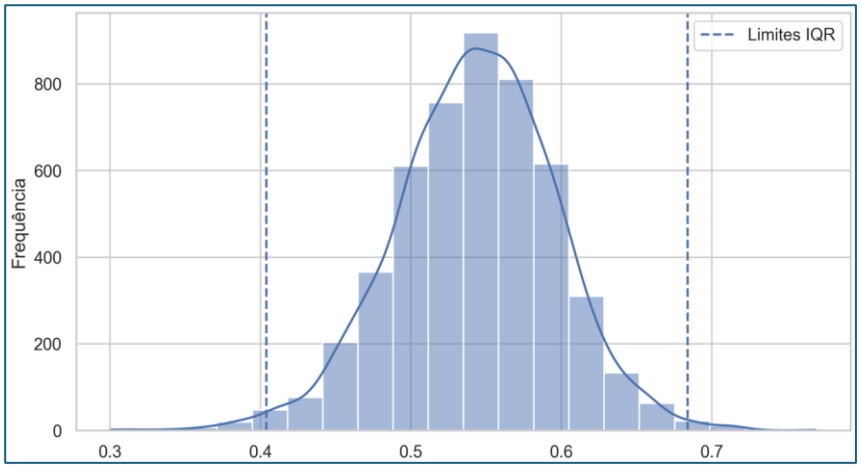
[4] Uma distribuição normal é aquela onde existe uma simetria na distribuição dos dados ao longo da análise da tendência central. Nesse caso, dado às características do conjunto de dados, utilizou-se o teste de *Kolmogorov-Smirnov* para detecção (Heumann, 2016).
[5] Ganho de informação é parte da Teoria da Informação e ajuda a quantificar a pureza de uma variável presente em um conjunto de dados (Forsyth, 2018).
[6] Correlação é uma medida estatística que mede a relação de tendências entre duas variáveis. Utilizou-se o coeficiente de correlação de Pearson, para variáveis com distribuição normal, e o coeficiente de Spearman, para as não normalmente distribuídas (Granda, 2017).

Gráfico 4 – Variável IPPI: Índice Paulista da Primeira Infância



Fonte: Elaborado pelo autor (2026)

Gráfico 5 – Variável IPDM: Índice Paulista de Desenvolvimento Municipal



Fonte: Elaborado pelo autor (2026)

O resultado desse processo deu origem a um novo conjunto de dados, agora reduzido para 94 variáveis independentes, com as características mostradas na Tabela 3. Esse novo conjunto de dados foi utilizado como base para aplicação e comparação dos modelos, detalhado no próximo tópico.

Tabela 3 – Grupos das variáveis independentes (seleção para predição)

GRUPOS DAS VARIÁVEIS INDEPENDENTES		
GRUPO	**QUANTIDADE**	**(%)**
DEMOGRAFICO	16	17,0
ECONOMICO	15	16,0
EDUCACIONAL	7	7,4
SAUDE	18	19,2
SOCIAL	38	40,4
TOTAL	**94**	**100,0**

Fonte: Elaborado pelo autor (2026)

Ainda, outra informação adicional importante, finalizando essa análise preliminar, está demonstrada na Tabela 4. Em primeira instância, a análise da cardinalidade, que mostra os valores distintos dentro do conjunto de dados na dimensão específica da variável, é um achado que levanta a possibilidade de discretização ou estratificação por quartis, pois, pode influenciar modelos de predição sensíveis a essas condições.

Tabela 4 – Análise de cardinalidade das variáveis independentes

ANÁLISE DE CARDINALIDADE		
VARIÁVEL	**REGISTROS**	**(%)**
pib_pib_per_capita	5127	100,0
pib_pib	5116	99,8
sp_saneamento_perda_ligacao	4622	90,1
pop_adulto	4532	88,4
pop_jovem	4114	80,2
pop_idoso	3695	72,1
ab_pop_total	3653	71,2
ab_pop_atendida	3483	67,9
p_saneamento_perda_distribuicao	3151	61,5
sp_ippi_ippi	2915	56,8
ab_pop_esf	2843	55,4
sp_saneamento_trata_esgoto	2634	51,4

Fonte: Elaborado pelo autor (2026).

4.5. Aplicação de Machine Learning

Com o conjunto de dados preparado, como detalhado no tópico anterior, passou-se para a aplicação dos modelos de ML, como descrito no Quadro 23, utilizando o mesmo ambiente de desenvolvimento utilizado nas análises anteriores, acrescentando as bibliotecas de software *Imbaleced Learn*, *LIME*, *SHAP*, *XGBoost* e *LightGBM*.

A escolha dos modelos específicos para comparação deu-se com base na nas revisões dos trabalhos de pesquisa na área de saúde e cobertura vacinal, já detalhados no Quadro 12.

Quadro 23 – Modelos de Machine Learning utilizados

Support Vector Machine (SVM)	Uma máquina de vetores de suporte (SVM) é essencialmente uma técnica para classificar objetos. É um método de aprendizado supervisionado. Assumindo que existe uma forma linear para separar duas classes, as máquinas de vetores de suporte definem a margem máxima que separa esse plano linear (Bell, 2015).	Biblioteca Python Scikit-Learn (sklearn)
Logistic Regression	Na regressão logística, com um conjunto de dados é proposta uma equação objetivando a classificação dos objetos, portanto, um método de aprendizado supervisionado. Regressão significa tentar encontrar um conjunto de parâmetros que melhor se ajuste aos seus objetivos (Harrington, 2012).	Biblioteca Python Scikit-Learn (sklearn)
Redes Neurais Artificiais – *MultiLayer Perceptron* (MLP)	Redes Neurais são eficazes no mapeamento de dados observados para uma função. É efetiva em aproximação de funções e problemas de aprendizado supervisionado, com poucas restrições quanto aos tipos de aplicações, mas tem desafios em termos de complexidade e velocidade (Kirk, 2014).	Biblioteca Python Scikit-Learn (sklearn)
Random Forest	*Random Forest* é um método baseado em árvores, que fornecem um mecanismo simples tanto para regressão quanto para classificação, onde a ideia principal é dividir um espaço de variáveis independentes em regiões menores e ajustar uma função de predição simples a cada região (Kroese, 2024).	Biblioteca Python Scikit-Learn (sklearn)
LightGBM	*Gradient Boosting* baseia-se no princípio de aprimorar iterativamente os classificadores fracos, por padrão, árvores de decisão, direcionando o foco para observações problemáticas que foram difíceis de prever em iterações anteriores, sendo construído e melhorado, portanto, em etapas (Hastie, 2017).	Bibloteca Python LightGBM
XGBoost		Biblioteca Python XGBoost

Fonte: Elaborado pelo autor (2026).

Os 6 modelos de ML apresentados no Quadro 23 foram utilizados para classificação das coberturas vacinais estudadas (BCG e Tríplice Viral (SCR)) baseando-se em duas classes, como mostrado no Quadro 24.

Quadro 24 – Coberturas vacinais por classe

COBERTURA VACINAL								
CLASSE	DESCRIÇÃO	BCG				TRÍPLICE VIRAL (SCR)			
		DE	ATÉ	PADRÕES		DE	ATÉ	PADRÕES	
1	Baixa	< 90%		3377	65,4%	< 95%		2616	50,7%
2	Adequada	> 90%		1783	34,6%	> 95%		2544	49,3%
TOTAL DE PADRÕES		5160							

Fonte: Elaborado pelo autor (2026)

Essa etapa contemplou as fases abaixo, nesta ordem:

- Segregação do conjunto de dados: conjunto de treinamento contemplando 80% do total, com 20% separado para os testes, observando as boas práticas para não vazamento de dados entre conjuntos;
- Balanceamento das classes: como pôde-se observar no Quadro 21 acima, as classes de cobertura vacinal apresentaram diferenças na quantidade de padrões dentro conjunto de dados. Comparou-se técnicas para balanceamento dos dados, iniciando pelo controle do balanceamento entre classes positivas e negativas e também aplicando técnicas direcionadas à geração de dados sintéticos para ajustes na classe minoritária, utilizando para isso o SMOTE (*Synthetic Minority Oversampling Technique*) e o ADASYN (*Adaptive Synthetic Sampling*);
- Redução de dimensionalidade: dois dos modelos utilizados nas comparações, SVM e *Logistic Regression*, por conceito, enfrentam dificuldades com conjuntos de dados com alta dimensionalidade, como é o caso do presente estudo, que possui 94 variáveis independentes. Assim, optou-se pelo uso da técnica PCA (*Principal Component Analysis*) para redução de dimensionalidade para testes destes dois modelos, reduzindo as variáveis independentes de 94 para 20 e 40, respectivamente;
- Testes dos modelos: o processo de otimização dos hiper parâmetros utilizados para testes de cada modelo foi automatizado com uso da

biblioteca de software Python *GridsearchCV*, que percorre de forma exaustiva todas as combinações possíveis, dentro de limites pré-estabelecidos;

- Validação dos modelos: para validação e apontamento do melhor modelo para classificação de cada cobertura vacinal, foram utilizadas as métricas: Acurácia, Precisão, Sensibilidade (Recall), Pontuação F1 (F1-Score) e Curva ROC–AUC;
- Explicabilidade dos resultados: finalmente, para o modelo que apresentou melhor performance na classificação, utilizou-se SHAP (*Shapley Additive Explanations*) e LIME (*Local Interpretable Model-agnostic Explanations*) para elencar as variáveis independentes que mais influenciaram no processo decisório (classificação).

5. RESULTADOS

Este tópico detalha as análises do conjunto de dados, bem como os resultados encontrados com a aplicação dos modelos de ML.

5.1. Análise exploratória dos dados

Para facilitar a análise exploratória do conjunto de dados pesquisado e também para contextualização, foram utilizados mapas coropléticos do estado de São Paulo das coberturas vacinais (BCG e Tríplice Viral (SCR)) utilizando a biblioteca *Geopandas*.

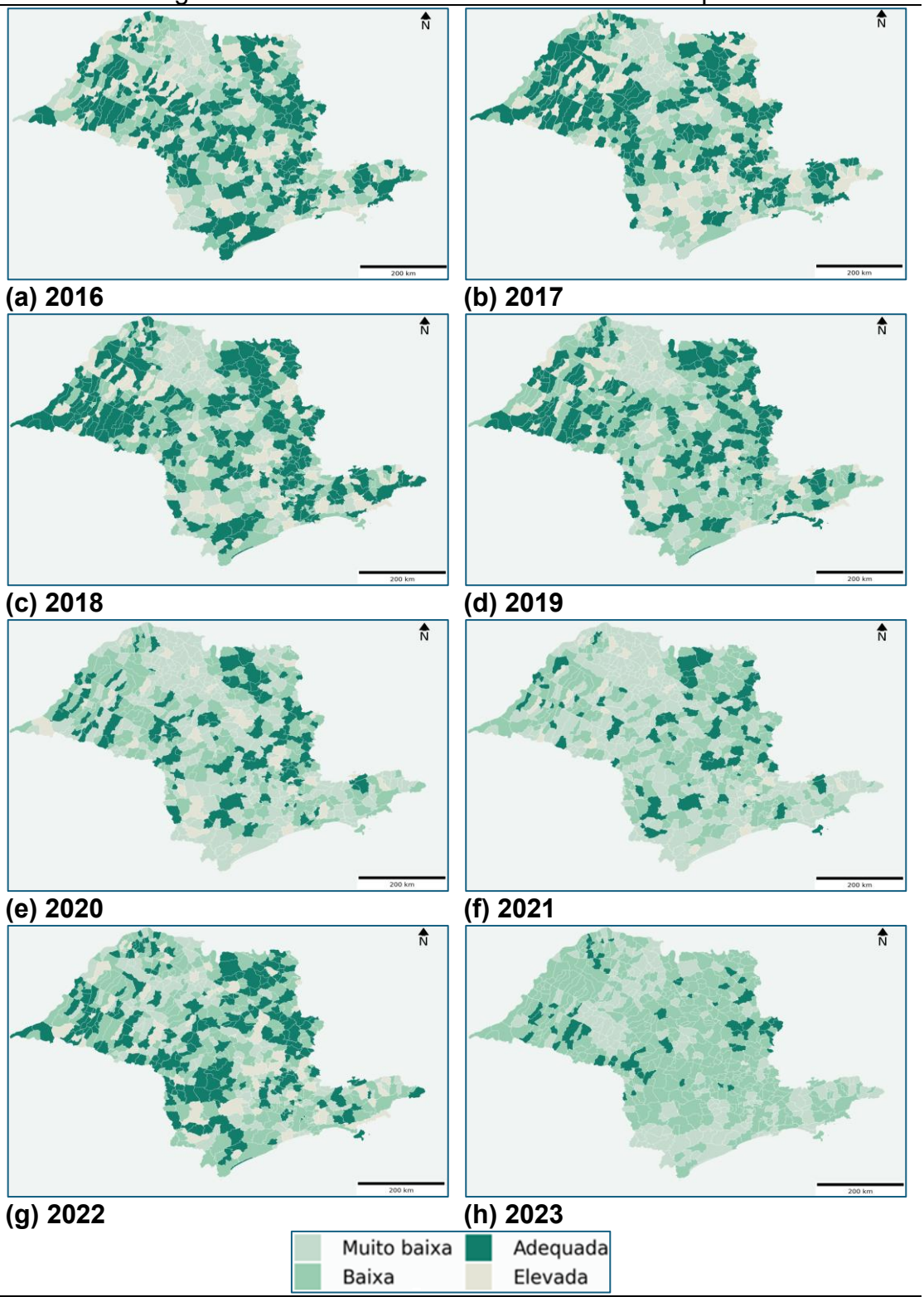
Foi demonstrado, primeiramente, uma visão geral por municípios (Figura 10 para BCG e Figura 13 para Tríplice Viral (SCR)), na sequência, a divisão por Regiões Administrativos (Figura 11 para BCG e Figura 14 para Tríplice Viral (SCR)), passando pelos Departamentos Regionais de Saúde (Figura 12 para BCG e Figura 15 para Tríplice Viral (SCR)).

As visões das coberturas vacinais estão apresentadas e segregadas por ano, para proporcionar uma visão abrangente da evolução (ou involução) ao longo do período da pesquisa, de 2016 a 2023.

5.1.1. Análise exploratória da BCG

Os mapas exibidos na Figura 10 possibilitam a visão da evolução da cobertura vacinal da BCG ao longo dos anos de 2016 a 2023.

Figura 10 – Cobertura Vacinal da BCG: Municípios



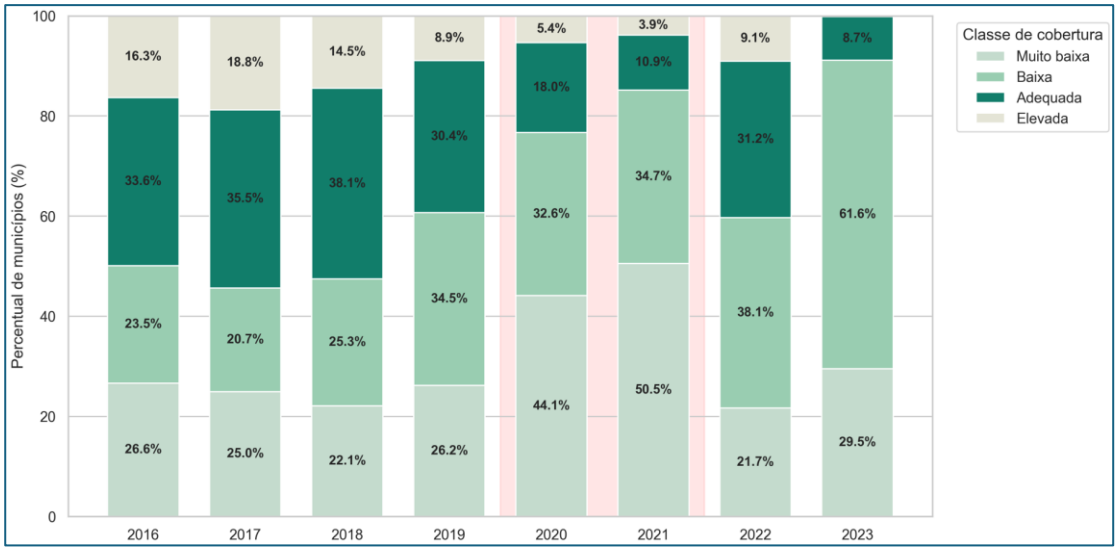
Fonte: Elaborado pelo autor (2026)

A área destacada no Gráfico 6 corresponde ao período da Pandemia da Covid-19, com aumento considerável do percentual de municípios com baixa cobertura,

o que também pode ser visualizado nos mapas dos anos de 2020 e 2021 (Figura 10 (e) e (f)), com diminuição nas áreas de cobertura “Adequada” e aumento nas áreas de cobertura “Muito baixa”. Em 2022, há queda no percentual de municípios com cobertura “Muito baixa”, e elevação da média percentual de cobertura, conforme Tabela 5.

Ainda, é possível notar que em 2023 há um aumento da cobertura, porém, ainda com um percentual grande de municípios com cobertura “Baixa”, na ordem de 61,6%, ou seja, mais que a metade dos municípios nessa condição.

Gráfico 6 – Distribuição dos munícipios por classe de cobertura vacinal da BCG



Fonte: Elaborado pelo autor (2026)

Na análise descritiva completa da cobertura detalhada na Tabela 5 é possível destacar:

a. Desvio padrão relevante, evidenciando distribuição de cobertura não uniforme, reforçado pelo Coeficiente de Variação (%) bastante alto;
b. Valores mínimos e máximos muito distantes das médias, chamando atenção para possíveis erros de entrada de dados ou municípios com cobertura vacinal muito discrepante (*outliers*);
c. Ano de 2022, imediatamente pós-pandemia, apresenta o maior valor Máximo, a maior Assimetria de Distribuição e o maior Curtose, o que reforça possíveis problemas com entrada de dados.

Tabela 5 – Análise descritiva da cobertura vacinal da BCG por ano

INDICADOR	2016	2017	2018	2019	2020	2021	2022	2023
Média (%)	82,4	87,3	85,2	76,2	58,1	50,2	79,6	62,1
Mediana (%)	89,9	93,0	91,5	83,8	57,0	47,6	81,2	62,7
Desvio Padrão (%)	48,1	50,5	48,5	45,4	42,1	40,4	43,3	21,4
Coeficiente Variação (%)	58,4	57,8	56,9	59,6	72,4	80,5	54,4	34,4
Mínimo (%)	0,4	0,7	0,3	0,7	0,3	0,3	0,5	2,9
Máximo (%)	294,1	475,1	593,9	572,7	220,5	285,7	647,5	133,8
Percentil 25 (%)	46,9	50,8	55,3	48,4	17,5	13,9	54,7	45,7
Percentil 75 (%)	108,6	112,9	106,2	98,8	88,7	77,9	100,0	80,0
Assimetria	0,4	0,9	2,0	2,2	0,6	1,0	3,8	-0,2
Curtose	0,6	5,3	19,1	22,3	0,2	2,4	46,3	-0,4

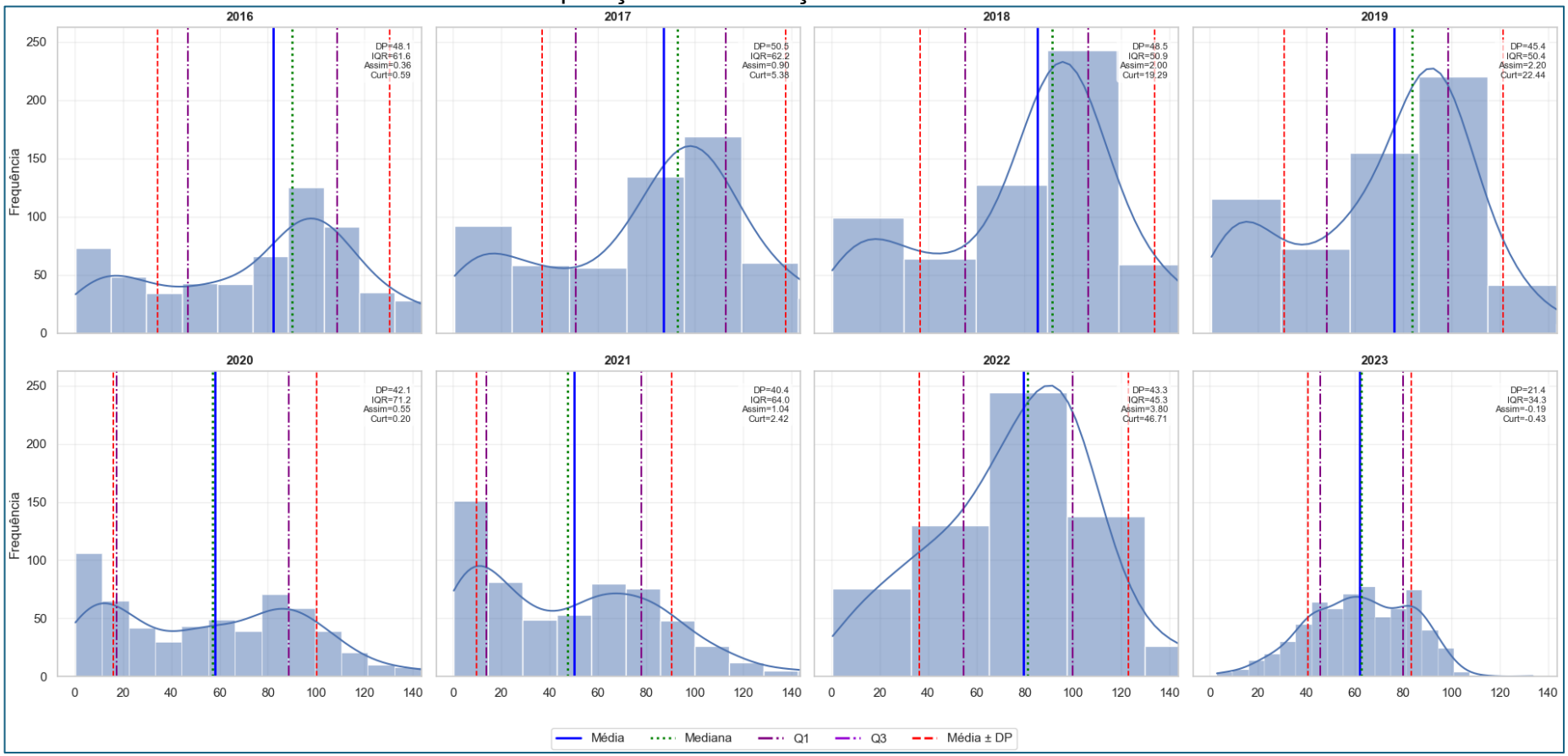
Fonte: Elaborado pelo autor (2026).

Para melhor visualizar a distribuição da cobertura da cobertura vacinal ao longo dos anos, o Gráfico 7 apresenta outra forma de análise.

É possível notar que não há distribuição igualitária entre os anos analisados e há deslocamento da média (linha azul dos gráficos) a partir de 2020, retomando em 2022, involuindo novamente em 2023.

Com a diminuição considerável dos intervalos interquartis (IQR) (linhas roxas) e do intervalo da Média e Desvio Padrão (positivo e negativo) (linhas vermelhas) é possível notar que o ano de 2023 tem menor dispersão e variabilidade (média mais confiável) e melhor consistência nos dados, porém, é marcado pela baixa cobertura vacinal geral entre todos os municípios do estado.

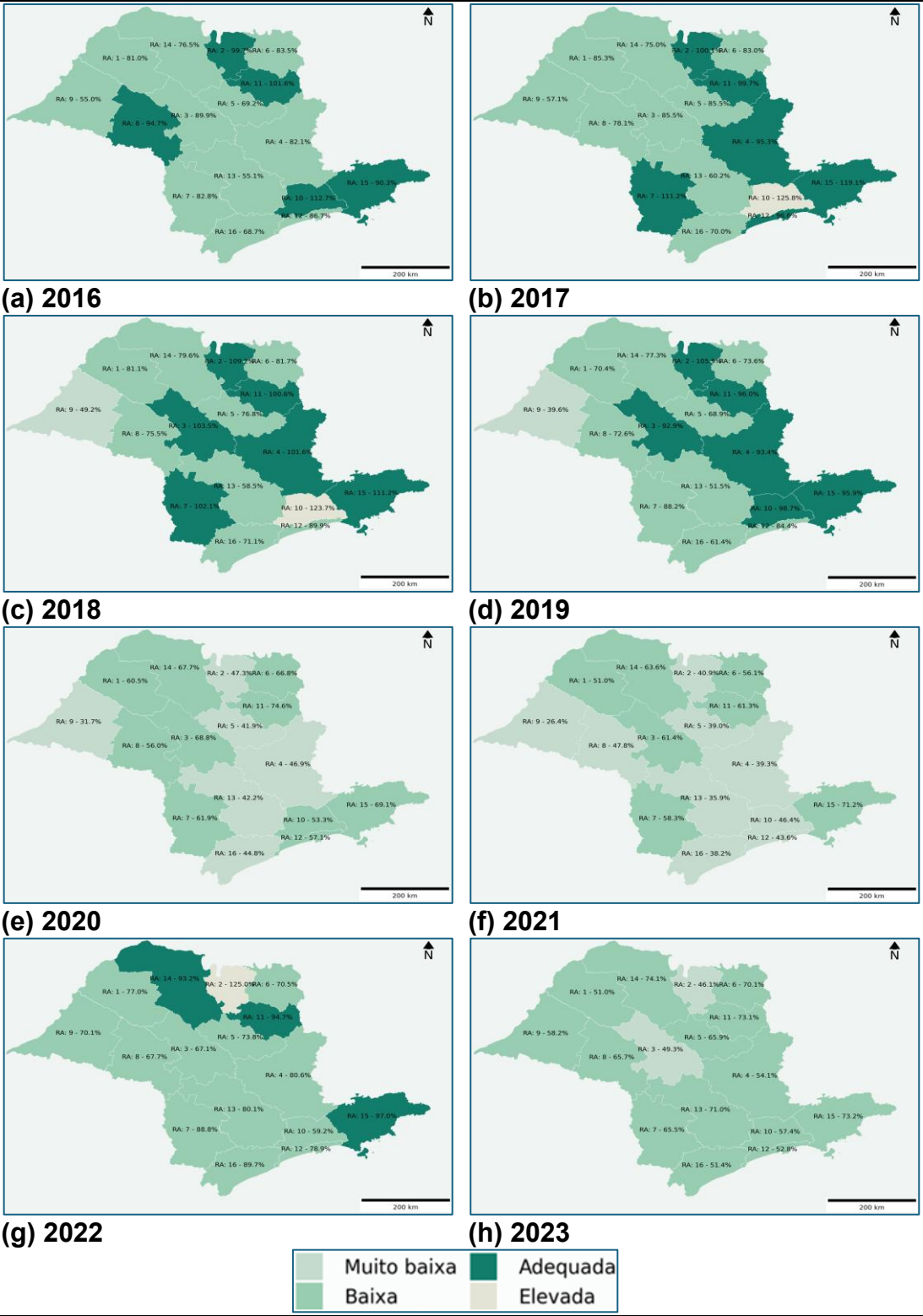
Gráfico 7 – Comparação da distribuição da cobertura vacinal da BCG



Fonte: Elaborado pelo autor (2026).

A Figura 11 apresenta os mapas da cobertura vacinal da BCG, com a visão comparada das áreas administrativas do estado de São Paulo.

Figura 11 – Cobertura Vacinal da BCG: Regiões Administrativas



Fonte: Elaborado pelo autor (2026)

Tabela 6 – Análise descritiva da cobertura vacinal da BCG por Regiões Administrativas com média geral do estado de São Paulo

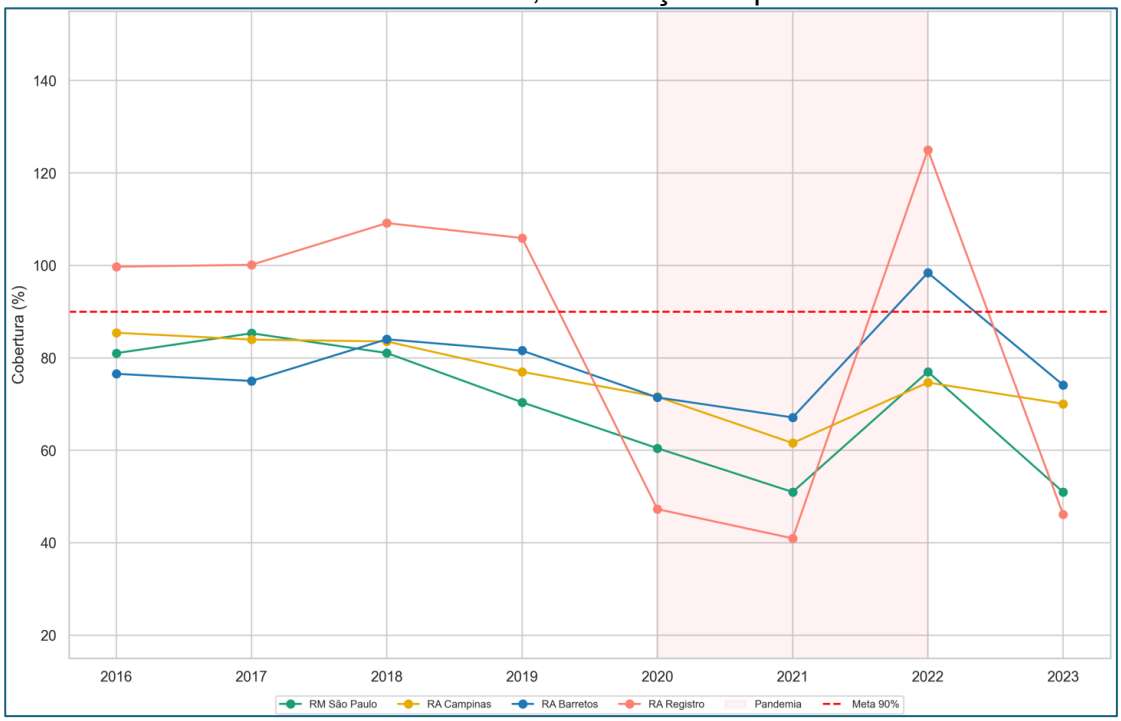
REGIÃO ADMINISTRATIVA	Habitantes (2023)	Municípios (2023)	Média (%)	Mediana (%)	Desvio Padrão (%)	Coeficiente Variação (%)	Mínimo (%)	Máximo (%)	Percentil 25 (%)	Percentil 75 (%)
RM São Paulo	21.515.186	39	69,7	67,7	42,7	61,3	0,7	252,1	38,0	92,3
RA Campinas	7.227.439	90	76,1	82,9	37,2	48,9	0,3	207,3	53,2	99,1
RA Sorocaba	2.709.819	47	66,5	72,2	42,2	63,4	0,3	290,0	30,5	91,0
RA São José dos Campos	2.580.764	39	76,1	77,0	46,3	60,8	0,4	247,3	41,9	101,8
RA Santos	1.856.233	9	77,3	70,5	55,7	72,0	4,7	247,5	39,5	93,4
RA São José do Rio Preto	1.644.015	96	51,7	43,1	48,8	94,3	0,5	294,1	9,6	76,9
RA Ribeirão Preto	1.425.082	25	83,2	88,1	33,2	39,9	1,2	177,9	58,6	103,5
RA Bauru	1.139.288	39	70,7	77,4	36,1	51,0	1,4	193,4	42,1	97,1
RA Central	1.051.890	26	62,2	63,6	38,1	61,2	1,6	160,9	28,4	92,1
RA Marília	998.444	51	73,9	81,7	36,7	49,7	1,7	193,6	42,4	98,1
RA Presidente Prudente	880.624	53	87,9	92,3	35,7	40,6	4,7	235,0	70,0	105,3
RA Araçatuba	794.486	43	85,1	82,1	51,7	60,7	1,2	300,0	47,3	108,8
RA Franca	761.249	23	90,9	92,1	32,3	35,5	7,6	169,4	73,8	109,4
RA Itapeva	536.826	32	68,6	76,9	44,1	64,2	0,7	206,3	28,5	96,5
RA Barretos	449.289	19	78,5	83,9	26,9	34,3	5,6	148,0	62,0	95,6
RA Registro	279.936	14	84,3	63,2	103,7	123,0	1,9	647,5	38,4	91,3
ESTADO DE SÃO PAULO			**72,9**	**76,5**	**45,2**	**62,0**	**0,3**	**647,5**	**40,3**	**98,2**

Fonte: Elaborado pelo autor (2026)

É possível notar, pela análise descritiva comparativa da Tabela 6, que as Regiões Administrativas do estado de São Paulo possuem realidades e cenários distintos, quando se analisa as coberturas vacinais da BCG, o que pode indicar políticas públicas distintas e/ou ações de governança diferentes.

O Gráfico 8 ilustra as médias ao longo dos anos. Para facilitar a leitura, foram exibidas somente quatro áreas (maiores e menores em quantidades de habitantes), como indicado na legenda. As linhas verde e laranja representam as áreas administrativas de maior quantidade de habitantes e as outras duas, linhas azul e vermelha, representam as áreas com menor concentração de habitantes.

Gráfico 8 – Evolução da cobertura vacinal da BCG por Regiões Administrativas – Maiores e menores, em relação a quantidade de habitantes

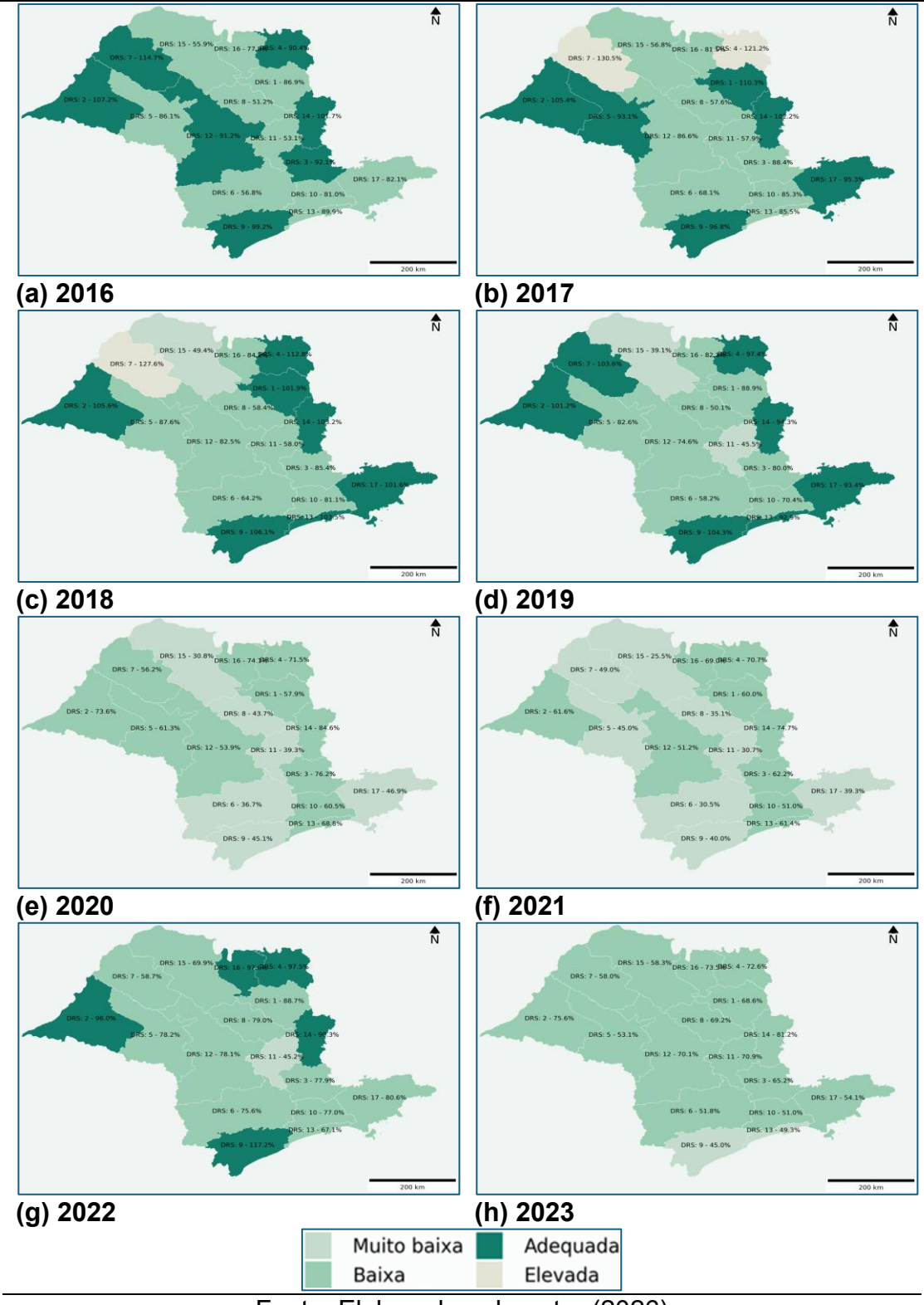


Fonte: Elaborado pelo autor (2026)

Cabe comentar, considerando o exibido no gráfico, melhor cobertura vacinal entre 2016 e 2019, para a região de menor concentração populacional, reduzindo somente no período da Pandemia da Covid-19, faixa destacada do gráfico.

Segue-se com a visão comparada dos Departamentos Regionais de Saúde, como exibido nos mapas da Figura 12 e na Tabela 7.

Figura 12 – Cobertura vacinal da BCG: Departamentos Regionais de Saúde



Fonte: Elaborado pelo autor (2026)

Tabela 7 – Análise descritiva da cobertura vacinal da BCG por Departamentos Regionais de Saúde com média do estado de São Paulo

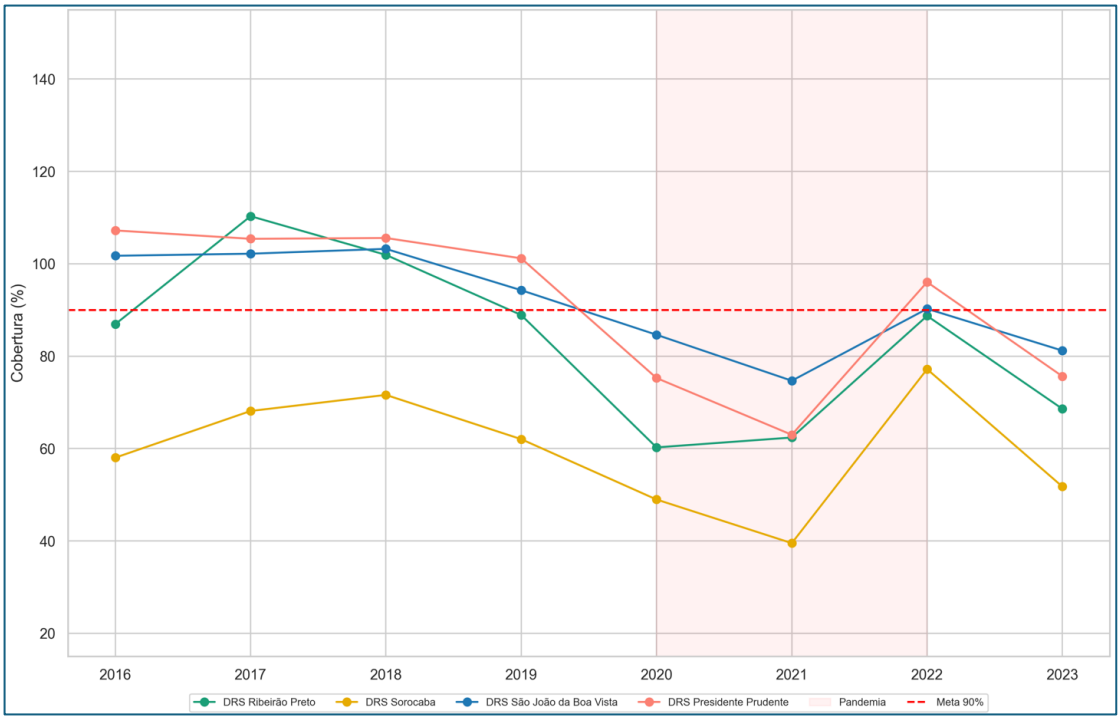
DEPARTAMENTO REGIONAL DE SAÚDE	Habitantes (2023)	Municípios (2023)	Média (%)	Mediana (%)	Desvio Padrão (%)	Coeficiente Variação (%)	Mínimo (%)	Máximo (%)	Percentil 25 (%)	Percentil 75 (%)
DRS Grande São Paulo	21.515.186	39	69,7	67,7	42,7	61,3	0,7	252,1	38,0	92,3
DRS Campinas	4.770.262	42	78,4	83,1	34,2	43,6	5,9	207,3	54,9	97,8
DRS Sorocaba	2.621.900	48	60,4	58,5	46,0	76,2	0,3	206,3	18,7	88,5
DRS Taubaté	2.580.764	39	76,1	77,0	46,3	60,8	0,4	247,3	41,9	101,8
DRS Baixada Santista	1.856.233	9	77,3	70,5	55,7	72,0	4,7	247,5	39,5	93,4
DRS Bauru	1.776.445	68	74,2	81,3	36,0	48,5	1,4	290,0	50,4	97,3
DRS São José do Rio Preto	1.679.058	102	51,7	43,9	48,0	92,9	0,5	294,1	10,0	76,7
DRS Piracicaba	1.597.101	26	57,9	55,4	44,8	77,5	0,3	161,7	14,0	92,5
DRS Ribeirão Preto	1.507.619	26	83,7	88,2	32,6	38,9	1,2	177,9	61,0	103,5
DRS Marília	1.131.679	62	73,4	80,5	37,4	51,0	1,7	198,7	41,4	97,6
DRS Araraquara	1.020.567	24	61,3	63,1	38,3	62,5	1,6	160,9	27,2	90,2
DRS São João da Boa Vista	826.248	20	91,5	94,4	23,9	26,2	20,6	202,6	78,8	103,8
DRS Araçatuba	778.885	40	87,8	84,9	51,9	59,1	1,2	300,0	50,0	110,4
DRS Presidente Prudente	764.711	45	91,3	94,1	33,4	36,6	4,7	235,0	76,4	107,5
DRS Franca	701.363	22	91,8	93,9	32,4	35,3	7,6	169,4	73,9	109,7
DRS Barretos	438.519	18	80,0	84,4	25,6	32,1	10,0	148,0	63,0	95,7
DRS Registro	284.030	15	81,7	62,9	100,9	123,4	1,9	647,5	37,3	91,0
ESTADO DE SÃO PAULO			**72,9**	**76,5**	**45,2**	**62,0**	**0,3**	**647,5**	**40,3**	**98,2**

Fonte: Elaborado pelo autor (2026)

É possível notar, pela análise descritiva na Tabela 7, realidades de cobertura vacinal da BCG também diversificadas, como já comparado entre as áreas administrativas. Como já discutido anteriormente, a divisão em Departamentos Regionais de Saúde tem foco na descentralização de políticas voltadas à saúde e, diferenças nas coberturas vacinais, como demonstrado, também chamam a atenção para conduções distintas entre os departamentos regionais.

Como forma de ampliar o entendimento dessas variações, assim como nas regiões administrativas, o Gráfico 9 a seguir foi também construído tomando quatro cenários diferentes escolhidos de acordo com a quantidade de habitantes em cada, sendo dois cenários com as maiores quantidades (linhas verde e laranja) e dois cenários com as menores quantidades (linhas azul e vermelha).

Gráfico 9 – Evolução da cobertura vacinal da BCG por Departamento Regional de Saúde – Maiores e menores, em relação a quantidade de habitantes



Fonte: Elaborado pelo autor (2026)

Apesar de tratar-se de Departamento Regional de Saúde geograficamente distante, a característica de cobertura vacinal mais consistente também é apresentada para concentração menor de habitantes, como pode ser visto no gráfico, abrindo margem para investigações mais detalhadas nesse sentido.

O Gráfico 10 a seguir fecha a análise exploratória para a BCG atendo-se as relações com a quantidade de habitantes, assim como foi feito para as Regiões Administrativas e os Departamentos Regionais de Saúde.

Essa análise concentrou-se nos portes dos municípios e mostrou que, para a BCG, municípios de menor porte (Pequeno Porte I) tiveram as piores médias anuais, somente ficando à frente dos municípios de Grande Porte e Metrópolis no ano de 2023.

Gráfico 10 – Evolução da cobertura vacinal da BCG por porte municipal

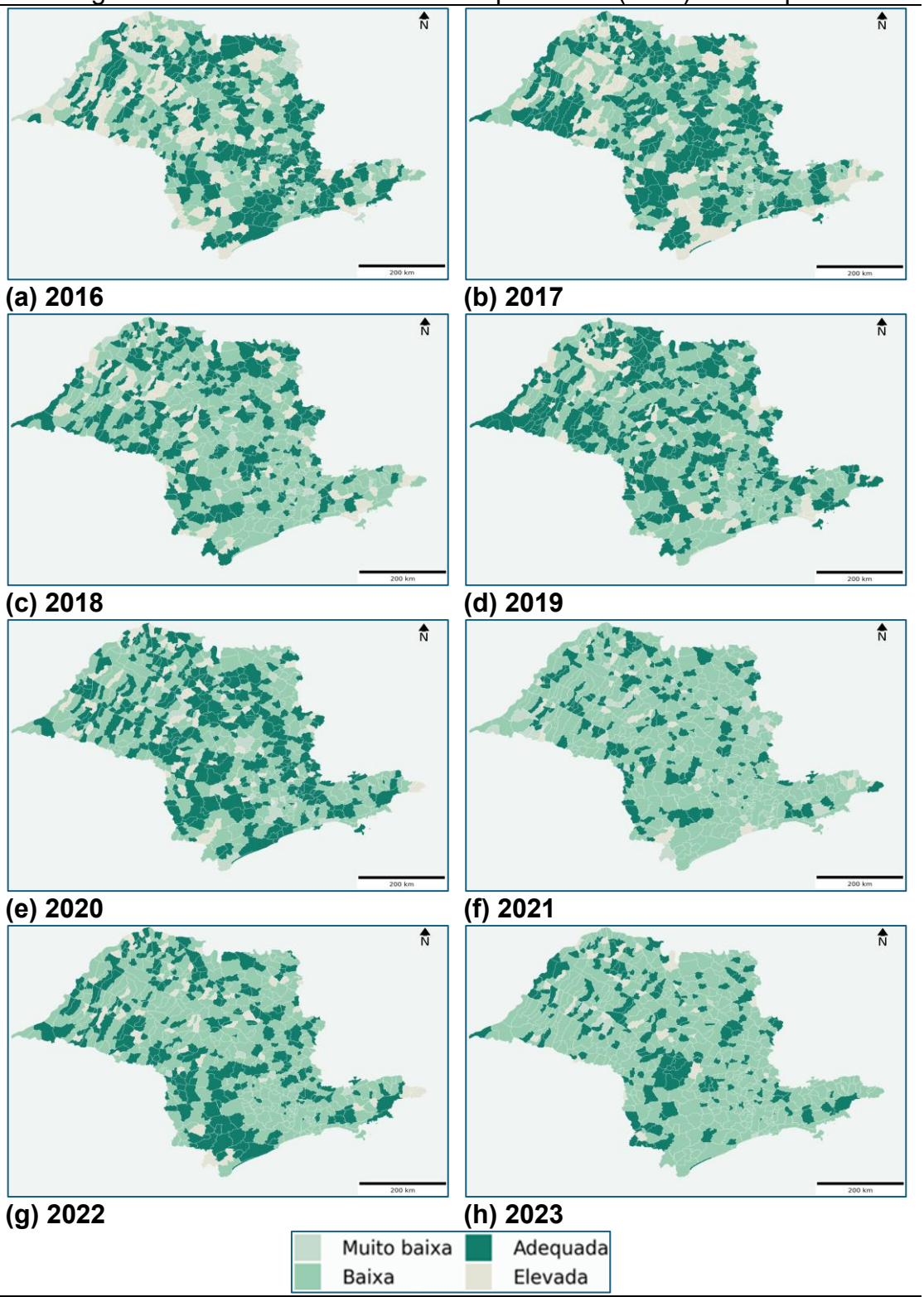


Fonte: Elaborado pelo autor (2026)

5.1.2. Análise exploratória da Tríplice Viral (SCR)

A análise do cenário da cobertura vacinal da Tríplice Viral (SCR) foi iniciado com os mapas exibidos na Figura 13, focando na visão da evolução ao longo dos anos de 2016 a 2023.

Figura 13 – Cobertura vacinal da Tríplice Viral (SCR): Municípios

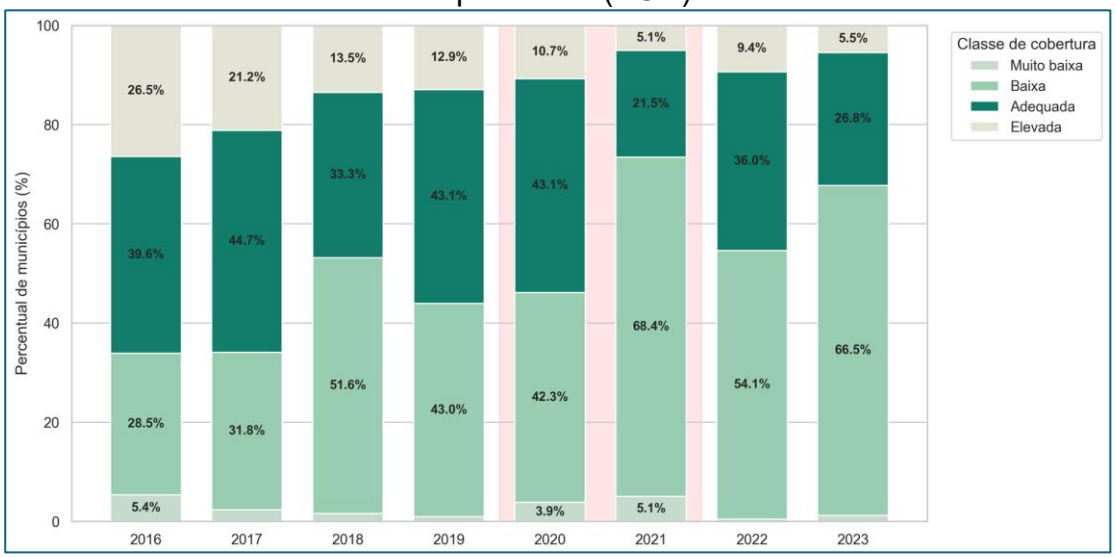


Fonte: Elaborado pelo autor (2026)

A cobertura vacinal da Tríplice Viral (SCR), como é possível ser observado no Gráfico 11, também é penalizada nos anos 2020 e 2021 (anos marcados pela

Pandemia da Covid-19), dando sinais de retomada em 2022, porém, com queda, novamente, no ano de 2023. Diferentemente da BCG, a queda da cobertura dessa vacina tem características diferentes, visto que, ao longo de todo o período é possível notar percentual de municípios com cobertura além da “Adequada”, pertencentes a Classe de cobertura “Elevada”, mas, também é possível notar presença de cobertura “Muito baixa”, o que não ocorre na análise da BCG.

Gráfico 11 – Distribuição dos munícipios por classe de cobertura vacinal da Tríplice Viral (SCR)



Fonte: Elaborado pelo autor (2026)

Na análise descritiva completa da cobertura detalhada na Tabela 8, é possível destacar:

a. Maior desvio padrão no ano de 2016 (35,1%) e o menor em 2023 (16,7%), porém, apesar de menor, ainda evidenciando distribuição de cobertura pouco uniforme, reforçado pelo Coeficiente de Variação (%) bastante alto;
b. Valores mínimos e máximos distantes das médias, chamando atenção para municípios com cobertura vacinal muito discrepantes (*outliers*);
c. A Tríplice Viral (SCR) apresenta valores médios dentro (ou muito próximos) da classe de cobertura “Adequada”, como pode ser comprovado pelas áreas verdes escuras nos mapas da Figura 13;

d. A Assimetria (mais bem notada no Gráfico 12) apresenta valores baixos, reforçando a ideia de uma distribuição de cobertura mais normalizada ao longo dos anos.

Tabela 8 – Cobertura vacinal da Tríplice Viral (SCR) por ano

INDICADOR	2016	2017	2018	2019	2020	2021	2022	2023
Média (%)	105,6	106,8	98,8	101,1	95,3	85,0	95,2	89,6
Mediana (%)	102,5	100,1	93,9	96,8	96,8	85,5	93,7	89,2
Desvio Padrão (%)	35,1	34,5	27,9	26,8	23,4	22,2	19,3	16,7
Mínimo (%)	5,2	8,6	24,3	19,1	4,8	18,9	23,8	15,8
Máximo (%)	287,5	336,4	275,8	311,1	214,3	228,6	231,3	153,3
Percentil 25 (%)	89,2	92,2	83,7	88,3	84,1	73,4	85,0	80,4
Percentil 75 (%)	120,8	116,4	107,9	107,2	107,7	96,3	102,7	97,7
Coeficiente Variação (%)	33,3	32,3	28,3	26,5	24,5	26,2	20,3	18,6
Assimetria	0,6	1,9	2,1	2,5	0,1	0,7	1,1	0,1
Curtose	2,7	10,4	8,6	12,7	2,7	5,2	5,5	2,2

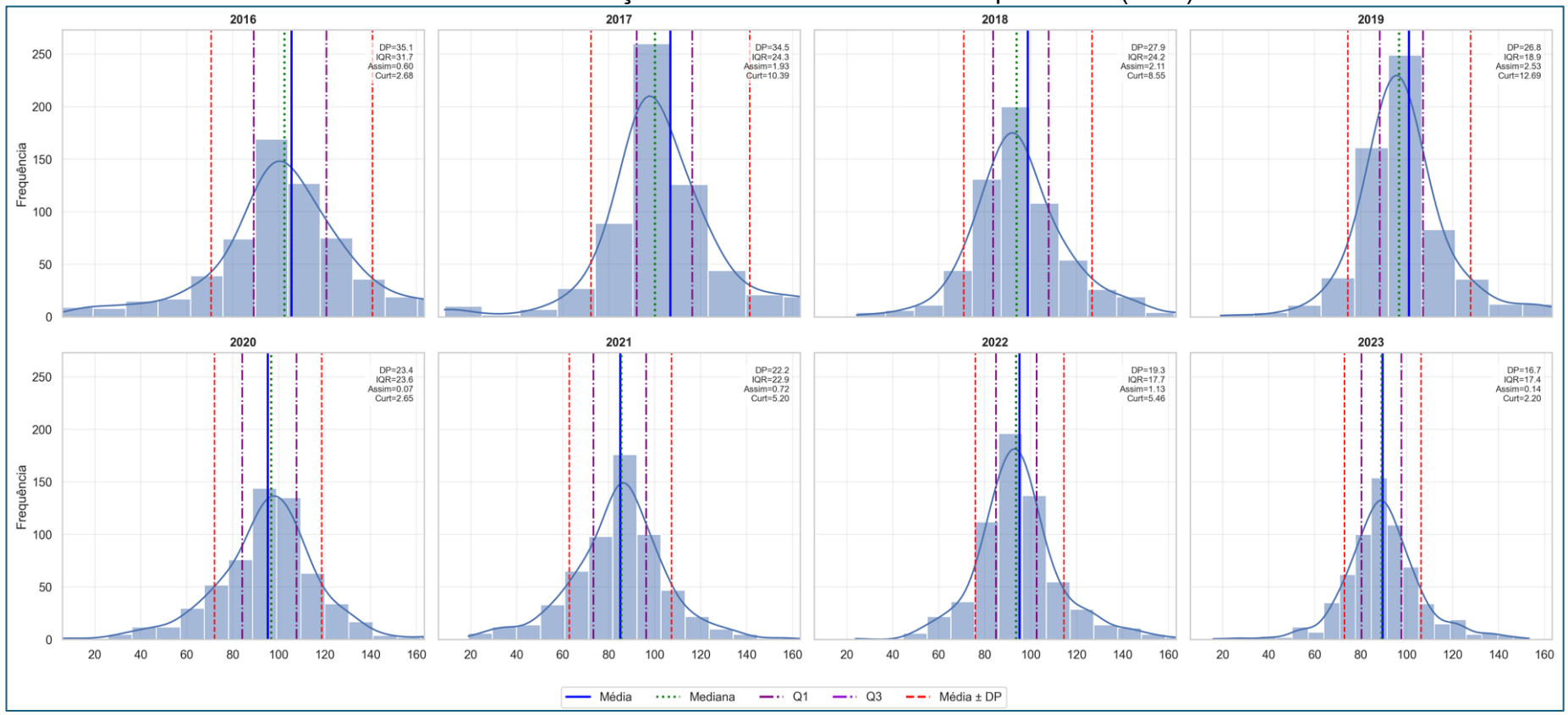
Fonte: Elaborado pelo autor (2026)

No Gráfico 12, é possível analisar a distribuição aproximadamente uniforme entre os anos analisados, assimetria maior entre 2016 e 2019 (média um pouco distante da mediana), porém, já praticamente se equivalem a partir de 2020, melhorando a distribuição e indicando menos dados discrepantes (*outliers*).

Os intervalos interquartis (IQR) (linhas roxas) e o intervalo da Média e Desvio Padrão (positivo e negativo) (linhas vermelhas) representa uma maior consistência nos dados quando comparado com a BCG, menor variabilidade e

as médias com mais valor representativo, com destaque ainda mais positivo para o ano de 2023.

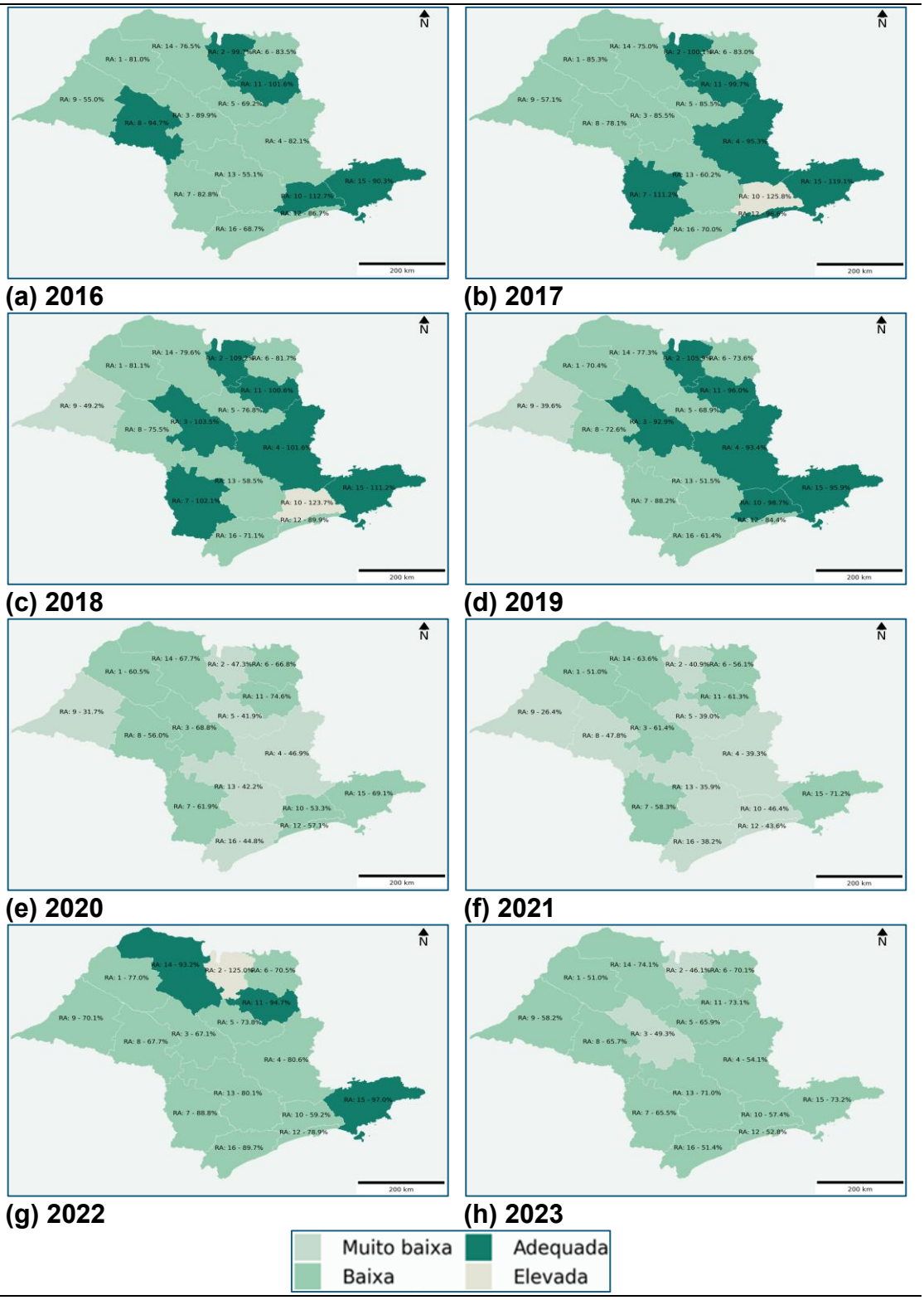
Gráfico 12 – Distribuição da cobertura vacinal da Tríplice Viral (SCR)



Fonte: Elaborado pelo autor (2026)

Os mapas da Figura 14 e a Tabela 9 são os instrumentos de análise nas divisões em Regiões Administrativas do estado de São Paulo.

Figura 16 – Cobertura vacinal da Tríplice Viral (SCR): Regiões Administrativas



Fonte: Elaborado pelo autor (2026)

Tabela 9 – Análise descritiva da cobertura vacinal da Tríplice Viral (SCR) por Regiões Administrativas com média geral do estado de São Paulo

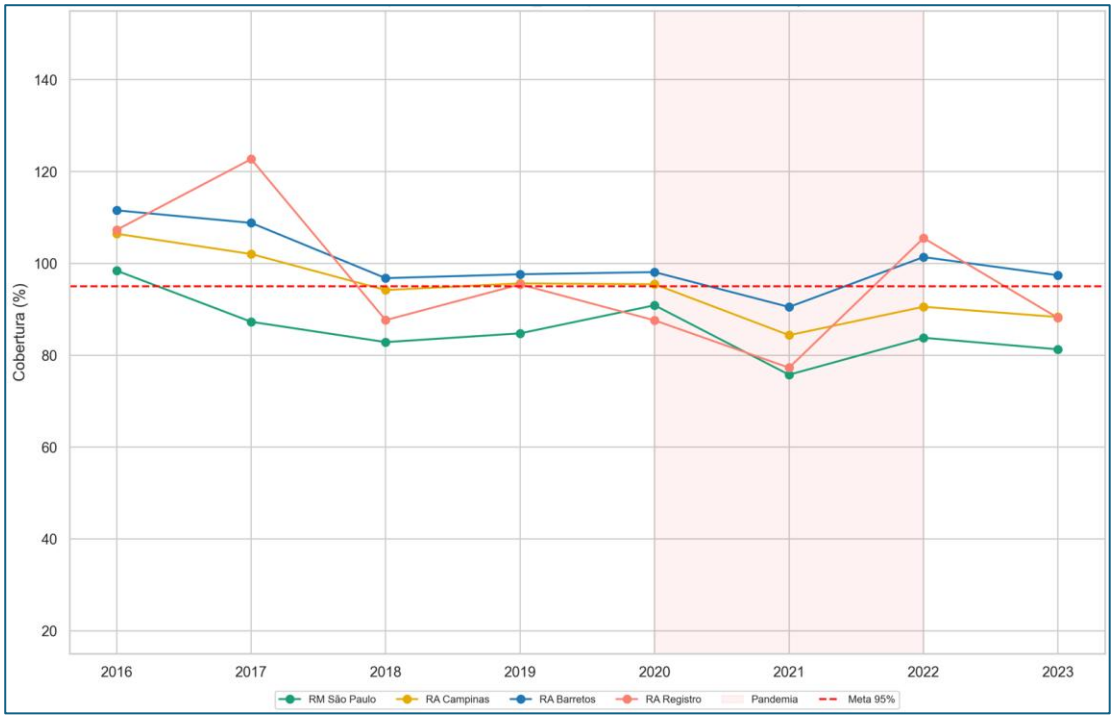
REGIÃO ADMINISTRATIVA	Habitantes (2023)	Municípios (2023)	Média (%)	Mediana (%)	Desvio Padrão (%)	Coeficiente Variação (%)	Mínimo (%)	Máximo (%)	Percentil 25 (%)	Percentil 75 (%)
RM São Paulo	21.515.186	39	85,6	87,0	23,2	27,1	9,2	189,2	75,4	96,5
RA Campinas	7.227.439	90	96,5	93,8	22,9	23,8	28,9	178,8	83,4	106,3
RA Sorocaba	2.709.819	47	80,1	80,4	21,9	27,3	8,9	130,0	70,1	92,5
RA São José dos Campos	2.580.764	39	94,4	90,6	31,1	33,0	19,1	224,1	77,8	104,1
RA Santos	1.856.233	9	95,1	94,0	26,6	28,0	8,6	290,0	83,0	105,5
RA São José do Rio Preto	1.644.015	96	94,7	94,0	19,4	20,5	4,8	231,6	85,9	103,1
RA Ribeirão Preto	1.425.082	25	93,4	90,6	28,0	30,0	19,2	235,7	78,4	101,7
RA Bauru	1.139.288	39	97,6	94,6	27,8	28,5	9,7	223,7	82,8	108,1
RA Central	1.051.890	26	103,5	98,7	29,2	28,2	13,3	336,4	88,9	113,5
RA Marília	998.444	51	106,4	96,0	41,9	39,4	10,9	287,5	84,3	115,2
RA Presidente Prudente	880.624	53	98,8	97,5	26,4	26,7	16,7	235,3	86,2	109,0
RA Araçatuba	794.486	43	97,2	95,0	25,7	26,4	5,2	217,8	84,9	106,6
RA Franca	761.249	23	96,0	94,2	18,4	19,2	48,2	153,2	85,9	104,3
RA Itapeva	536.826	32	100,4	97,0	21,9	21,8	41,5	215,7	87,9	109,3
RA Barretos	449.289	19	96,2	94,0	28,2	29,3	11,8	177,8	81,3	110,3
RA Registro	279.936	14	101,0	98,2	21,9	21,7	36,4	201,2	89,2	109,2
ESTADO DE SÃO PAULO			**97,26**	**94,66**	**27,42**	**28,19**	**4,75**	**336,36**	**84,05**	**106,71**

Fonte: Elaborado pelo autor (2026)

Nota-se, pela análise descritiva comparativa da Tabela 9, que as Áreas Administrativas do estado de São Paulo apresentam realidades e condutas semelhantes para a cobertura da Tríplice Viral (SCR). Exceto a “RA Sorocaba” e a “RM São Paulo”, todas as demais possuem cobertura próximas ou superior à adequada (acima de 95%). As distâncias dentre quartis (IQR) anotados como “Percentil 25%” e “Percentil 75%” mantem-se, praticamente, como média constante entre todas as áreas, favorecendo a análise das médias e reforçando a menor variabilidade e boa consistência nos dados.

O Gráfico 13 ilustra as médias ao longo dos anos. Assim como a BCG, são mostradas somente quatro áreas, como indicado na legenda. As linhas verde e laranja representam as áreas administrativas de maior quantidade de habitantes e as outras duas, linhas azul e vermelha, representam as áreas com menor concentração de habitantes. Destaques para a “RA Barretos” em 2018 e a “RA Registro” em 2022.

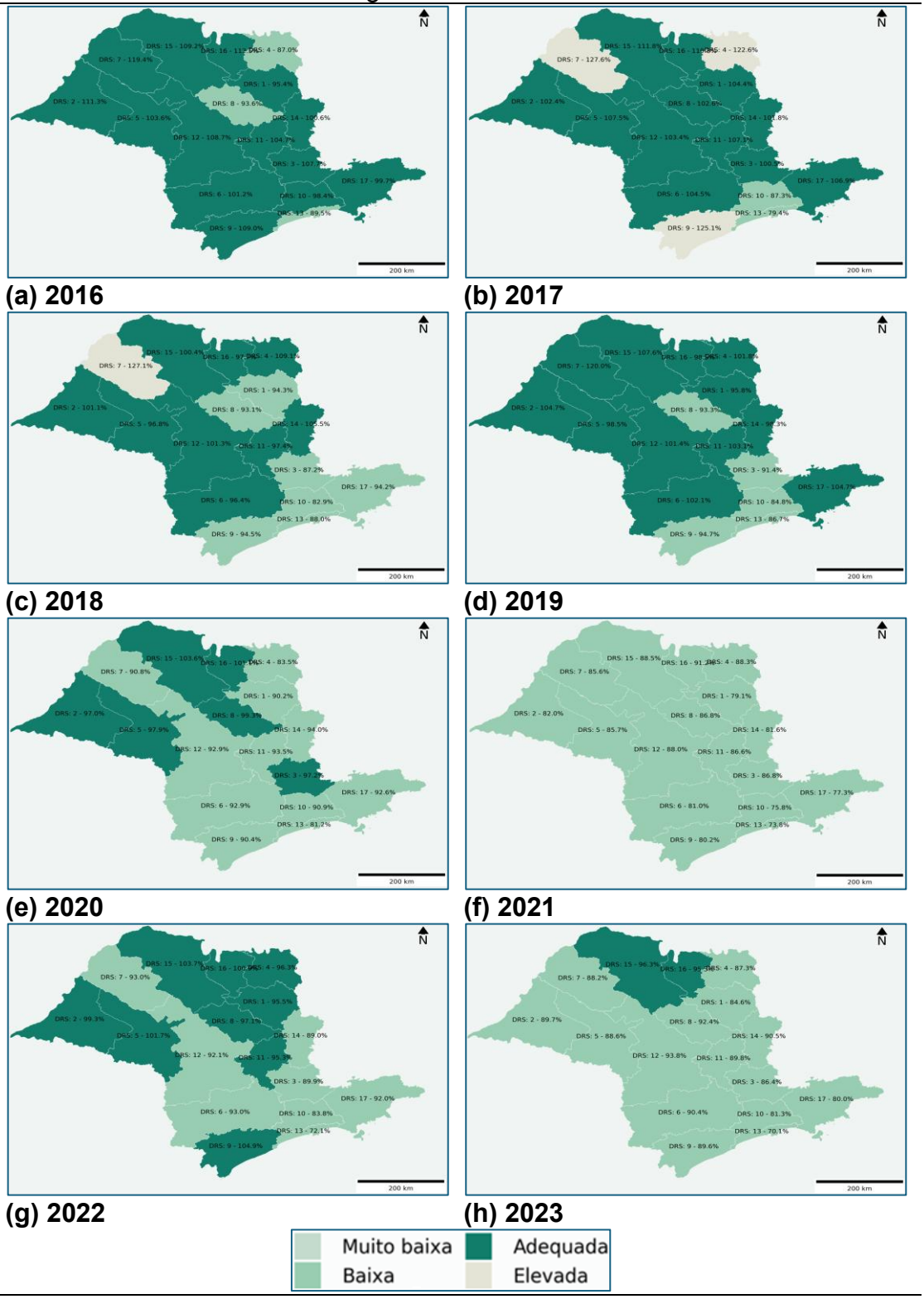
Gráfico 13 – Evolução da cobertura vacinal da Tríplice Viral (SCR) por Região Administrativa – Maiores e menores, em relação a quantidade de habitantes



Fonte: Elaborado pelo autor (2026)

Na Figura 15 e Tabela 10 são considerados os mapas e análise descritiva na visão dos Departamentos Regionais de Saúde do estado de São Paulo.

Figura 17 – Cobertura vacinal da Tríplice Viral (SCR): Departamentos Regionais de Saúde



Fonte: Elaborado pelo autor (2026)

Tabela 10 – Análise descritiva da cobertura vacinal da Tríplice Viral (SCR) por Departamentos Regionais de Saúde com média do estado de São Paulo

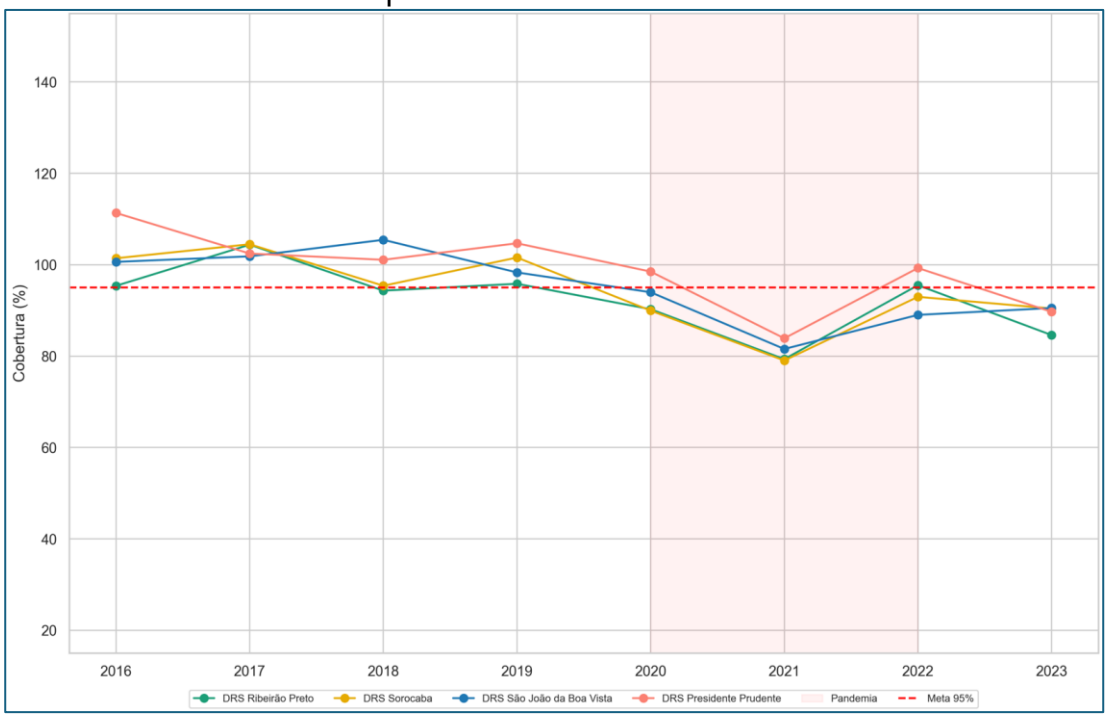
DEPARTAMENTO REGIONAL DE SAÚDE	Habitantes (2023)	Municípios (2023)	Média (%)	Mediana (%)	Desvio Padrão (%)	Coeficiente Variação (%)	Mínimo (%)	Máximo (%)	Percentil 25 (%)	Percentil 75 (%)
DRS Grande São Paulo	21.515.186	39	92,5	90,4	27,0	29,2	19,2	235,7	78,4	101,1
DRS Campinas	4.770.262	42	98,9	98,0	27,3	27,6	16,7	235,3	86,3	108,7
DRS Sorocaba	2.621.900	48	93,4	93,5	17,5	18,8	8,8	188,5	85,1	100,5
DRS Taubaté	2.580.764	39	97,0	95,2	28,5	29,4	11,8	177,8	81,9	112,9
DRS Baixada Santista	1.856.233	9	97,6	95,0	24,9	25,5	5,2	217,8	85,0	106,8
DRS Bauru	1.776.445	68	95,0	93,3	24,6	25,9	8,6	201,2	82,9	104,5
DRS São José do Rio Preto	1.679.058	102	106,7	95,5	43,0	40,3	10,9	287,5	84,3	115,2
DRS Piracicaba	1.597.101	26	95,3	93,7	18,7	19,7	48,2	153,2	85,6	104,2
DRS Ribeirão Preto	1.507.619	26	98,5	95,3	24,9	25,3	28,9	189,8	84,7	109,3
DRS Marília	1.131.679	62	85,6	87,0	23,2	27,1	9,2	189,2	75,4	96,5
DRS Araraquara	1.020.567	24	97,4	95,3	18,1	18,6	55,7	170,7	87,0	104,6
DRS São João da Boa Vista	826.248	20	98,1	96,3	27,1	27,6	4,8	290,0	86,2	107,0
DRS Araçatuba	778.885	40	80,1	80,4	21,9	27,3	8,9	130,0	70,1	92,5
DRS Presidente Prudente	764.711	45	95,2	94,3	21,7	22,8	26,6	231,6	86,9	105,5
DRS Franca	701.363	22	103,4	98,6	28,8	27,8	13,3	336,4	88,9	113,9
DRS Barretos	438.519	18	101,0	97,6	22,4	22,2	41,5	215,7	87,8	110,3
DRS Registro	284.030	15	94,4	90,6	31,1	33,0	19,1	224,1	77,8	104,1
ESTADO DE SÃO PAULO			**97,3**	**94,7**	**27,4**	**28,2**	**4,8**	**336,4**	**84,1**	**106,7**

Fonte: Elaborado pelo autor (2026)

Pelos indicadores descritivos da Tabela 10 é possível notar a confirmação da uniformidade na cobertura vacinal da Tríplice Viral (SCR) também entre os Departamentos Regionais de Saúde.

Abaixo, o Gráfico 14, assim como nos de mesmo tipo anteriores, foi construído tomando quatro cenários diferentes escolhidos de acordo com a quantidade de habitantes em cada, sendo dois cenários com as maiores quantidades (linhas verde e laranja) e dois cenários com as menores quantidade (linhas azul e vermelha).

Gráfico 14 – Evolução da cobertura vacinal da Tríplice Viral (SCR) por Departamento Regional de Saúde – Maiores e menores, em relação a quantidade de habitantes



Fonte: Elaborado pelo autor (2026)

Assim como na análise da BCG, novamente o Departamento Regional com menor concentração de habitantes apresentou médias de cobertura mais consistentes, ao longo do período analisado.

Em uma última análise da cobertura vacinal da Tríplice Viral (SCR), o Gráfico 15 mostra os comparativos de cobertura por porte dos municípios. Os municípios

de menor quantidade absoluta de habitantes apresentam um melhor cenário, ao longo de todo o período analisado.

Gráfico 15 – Evolução da cobertura vacinal da Tríplice Viral (SCR) por porte municipal



Fonte: Elaborado pelo autor (2026)

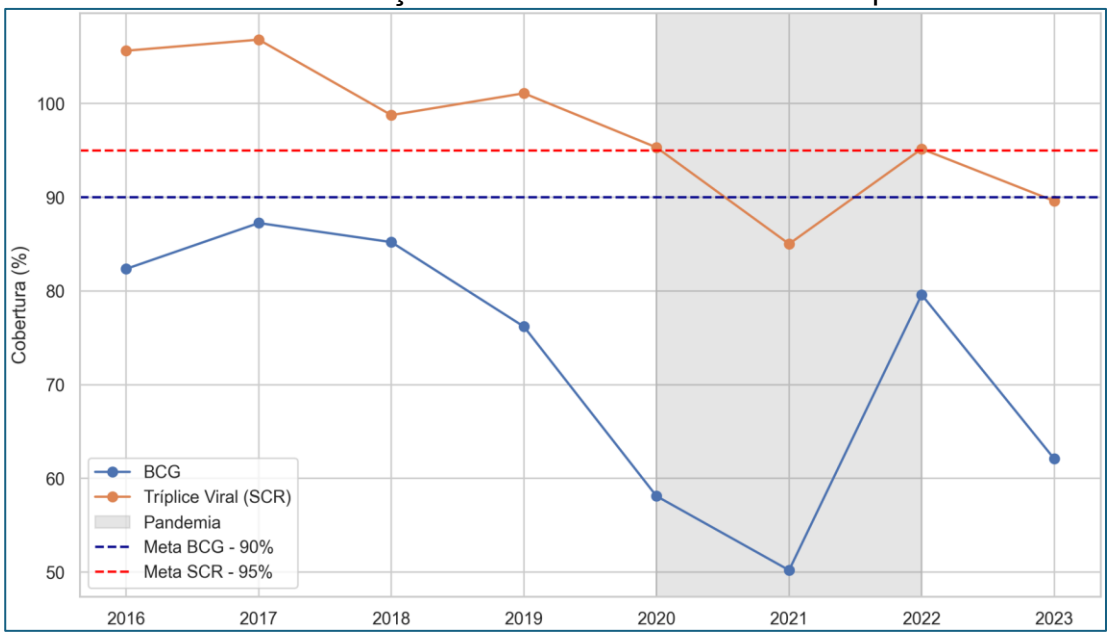
5.1.3. Análise exploratória comparada

Para iniciar análise estatística descritiva comparada, inicialmente, o Gráfico 16 ilustra a evolução das médias de cobertura das duas vacinas ao longo do período estudado.

O comportamento da cobertura vacinal da vacina Tríplice Viral (SCR) é bastante superior em todos os anos, mantendo médias superiores a 95% de cobertura (adequado), exceto em 2021, dentro do período já citado da Pandemia da Covid-19.

A cobertura da BCG consegue ficar próxima da cobertura ideal em apenas dois anos, 2017 e 2018 e, nos demais, tem médias muito aquém do ideal.

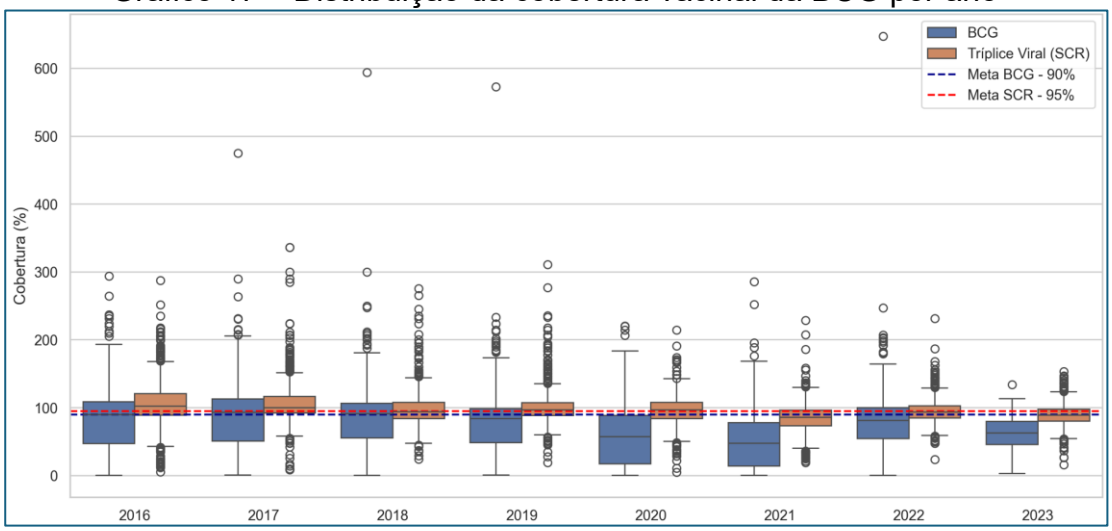
Gráfico 16 – Evolução da cobertura vacinal da BCG por ano



Fonte: Elaborado pelo autor (2026)

Na exibição das coberturas vacinais com outras visões (gráficos 17 e 19), é possível comparar as características das distribuições. A BCG tem uma distribuição bastante desuniforme, com dados que se distanciam bastante das análises de tendências centrais, enquanto a Tríplice Viral (SCR) mantém distribuição próximas das medidas de tendências centrais, facilitando as análises, evidenciando melhor consistência dos dados e aumentado a confiabilidade em suas leituras.

Gráfico 17 – Distribuição da cobertura vacinal da BCG por ano



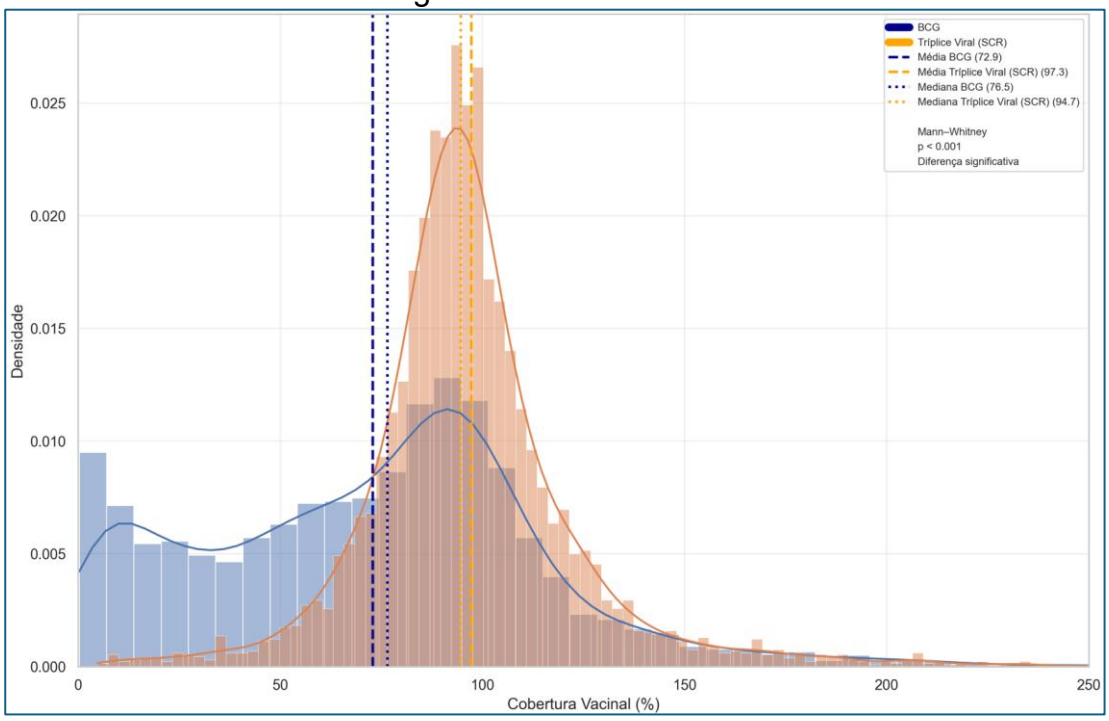
Fonte: Elaborado pelo autor (2026)

Ainda, na visão do Gráfico 18, onde as coberturas são comparadas segundo as distribuições de probabilidade com uma análise de densidade, é possível consolidar a visão do espalhamento dos dados da vacina BCG e das concentrações mais uniformes da vacina Tríplice Viral (SCR).

As maiores densidades da Tríplice Viral (SCR) marcam a melhor qualidade da cobertura vacinal ao longo de todos os anos do período pesquisado.

O Gráfico 18 compara as distribuições em única visão e inclui teste de significância entre as medianas. Foi utilizado *Mann–Whitney* para a comparação, uma vez que nenhuma das vacinas possuem distribuição normal.

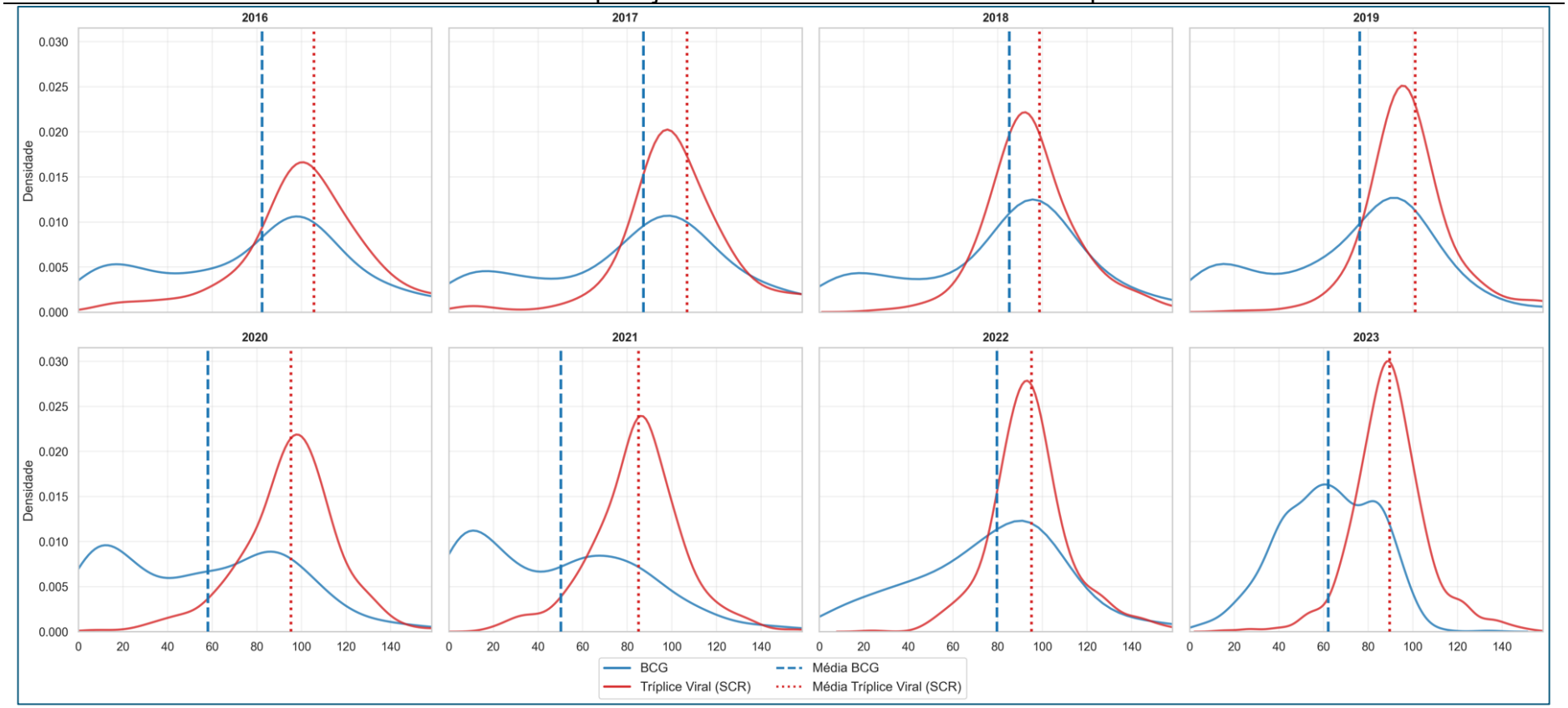
Gráfico 18 – Comparativo geral da cobertura vacinal da BCG com teste de significância estatística



Fonte: Elaborado pelo autor (2026)

A análise comparada finaliza com os indicadores de tendência central, mínimos, máximos, coeficiente de variação (%), assimetria e curtose da Tabela 11, de forma a consolidar os achados demonstrados até o presente momento.

Gráfico 19 – Comparação das coberturas vacinais da BCG por ano



Fonte: Elaborado pelo autor (2026)

Tabela 11 – Análise descritiva comparada da BCG e Tríplice Viral (SCR) por ano

INDICADOR	2016		2017		2018		2019		2020		2021		2022		2023		GERAL	
	BCG	SCR	BCG	SCR	BCG	SCR	BCG	SCR	BCG	SCR	BCG	SCR	BCG	SCR	BCG	SCR	BCG	SCR
Média (%)	82,4	105,6	87,3	106,8	85,2	98,8	76,2	101,1	58,1	95,3	50,2	85,0	79,6	95,2	62,1	89,6	**72,9**	**97,3**
Mediana (%)	89,9	102,5	93,0	100,1	91,5	93,9	83,8	96,8	57,0	96,8	47,6	85,5	81,2	93,7	62,7	89,2	**76,5**	**94,7**
Desvio Padrão (%)	48,1	35,1	50,5	34,5	48,5	27,9	45,4	26,8	42,1	23,4	40,4	22,2	43,3	19,3	21,4	16,7	**45,2**	**27,4**
Coeficiente Variação (%)	58,4	33,3	57,8	32,3	56,9	28,3	59,6	26,5	72,4	24,5	80,5	26,2	54,4	20,3	34,4	18,6	**62,0**	**28,2**
Mínimo (%)	0,4	5,2	0,7	8,6	0,3	24,3	0,7	19,1	0,3	4,8	0,3	18,9	0,5	23,8	2,9	15,8	**0,3**	**4,8**
Máximo (%)	294,1	287,5	475,1	336,4	593,9	275,8	572,7	311,1	220,5	214,3	285,7	228,6	647,5	231,3	133,8	153,3	**647,5**	**336,4**
Percentil 25 (%)	46,9	89,2	50,8	92,2	55,3	83,7	48,4	88,3	17,5	84,1	13,9	73,4	54,7	85,0	45,7	80,4	**40,3**	**84,1**
Percentil 75 (%)	108,6	120,8	112,9	116,4	106,2	107,9	98,8	107,2	88,7	107,7	77,9	96,3	100,0	102,7	80,0	97,7	**98,2**	**106,7**
Assimetria	0,4	0,6	0,9	1,9	2,0	2,1	2,2	2,5	0,6	0,1	1,0	0,7	3,8	1,1	-0,2	0,1	**1,5**	**1,6**
Curtose	0,6	2,7	5,3	10,4	19,1	8,6	22,3	12,7	0,2	2,7	2,4	5,2	46,3	5,5	-0,4	2,2	**13,3**	**9,0**

Fonte: Elaborado pelo autor (2026).

5.2. Análise preditiva das coberturas vacinais

O processo de análise preditiva foi conduzido com a aplicação dos modelos de ML escolhidos e já mencionados nos métodos no capítulo 4.

Antes do detalhamento dos resultados encontrados, cabe voltar novamente a atenção às divisões de classes para predição já consideradas no Quadro 24, sobretudo no que diz respeito às fronteiras do que é considerado CV “Adequada” (maior ou igual à 90% para BCG e maior ou igual a 95% para a Tríplice Viral (SCR)).

Os modelos foram treinados e testados utilizando o conjunto de padrões para predição da cobertura vacinal “Adequada” (classe 1 ou positiva) ou “Não adequada / Baixa” (classe 0 ou negativa), para as duas vacinas, a BCG e a Tríplice Viral (SCR).

O conjunto de dados, para uso durante todo o processo de treinamento e testes, foi dividido utilizando o critério 80-20, ou seja, 80% dos padrões para treinamento (ou 4132 padrões) e 20% dos padrões para os testes de consistência dos modelos (ou 1026 padrões).

5.2.1. Análise preditiva da BCG

A Tabela 12 traz a comparação dos resultados dos métodos de ML utilizados para a predição das classes da BCG.

Tabela 12 – Comparativo dos métodos de Machine Learning para cobertura vacinal da BCG

MODELO	Acuracidade	Precisão	Sensibilidade	Pontuação F1	Curva ROC-AUC
XGBoost	0,800	0,713	0,713	0,713	0,840
Random Forest	0,805	0,734	0,685	0,708	0,840
LightGBM	0,798	0,692	0,752	0,721	0,835
Support Vector Machine (SVM)	0,766	0,700	0,566	0,626	0,805
Multilayer Perceptron (MLP)	0,722	0,598	0,603	0,600	0,761
Logistic Regression	0,656	0,502	0,628	0,558	0,713

Fonte: Elaborado pelo autor (2026)

A capacidade discriminatória foi o item mais focado na análise do desempenho, medido pela métrica “Curva ROC-AUC”. Para essa métrica, tanto o *XGBoost* quanto *Random Forest* tiveram o mesmo desempenho. Como segundo critério de avaliação, optou-se pela métrica “Pontuação F1”, que calcula uma média harmônica entre os falsos positivos e falsos negativos. Nesta métrica, o *XGBoost* levou pequena vantagem sendo considerado, portanto, o método de melhor performance.

A Tabela 13 mostra a Matriz de Confusão do modelo *XGBoost* com os 1026 padrões utilizados para testes. O modelo classificou adequadamente 822 padrões, sendo 569 padrões na classe de cobertura baixa ou não adequada e 253 padrões na classe de cobertura adequada. Os falsos positivos e falsos negativos somaram 204 padrões do total de padrões analisados.

Tabela 13 – Matriz de Confusão da cobertura vacinal da BCG com o modelo *XGBoost*

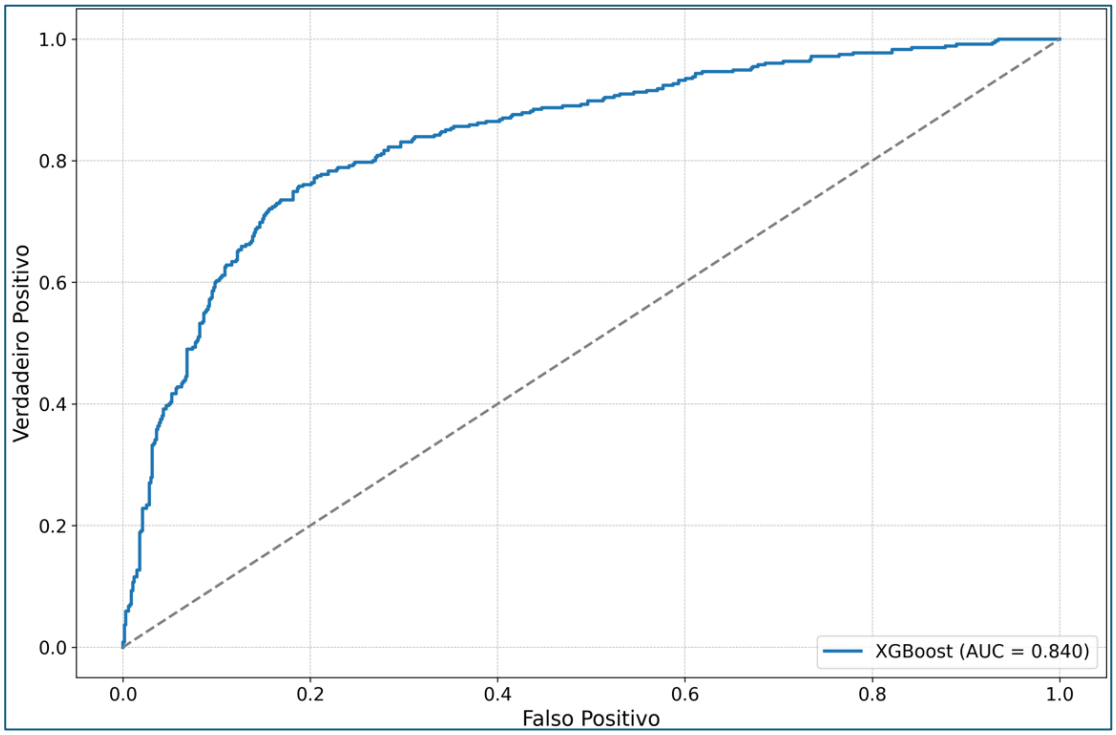
Classe Real	Predita: Baixa	Predita: Adequada
Baixa	569 (84,8%)	102 (15,2%)
Adequada	102 (28,7%)	253 (71,3%)

Fonte: Elaborado pelo autor (2026)

Para referência e comparativos, as Matrizes de Confusão de todos os demais modelos testados são apresentadas no apêndice desta pesquisa.

O Gráfico 20 apresenta o gráfico da “Curva ROC-AUC do modelo *XGBoost* onde é possível observar a evolução da fronteira discriminatória das classes durante o processamento dos padrões de testes.

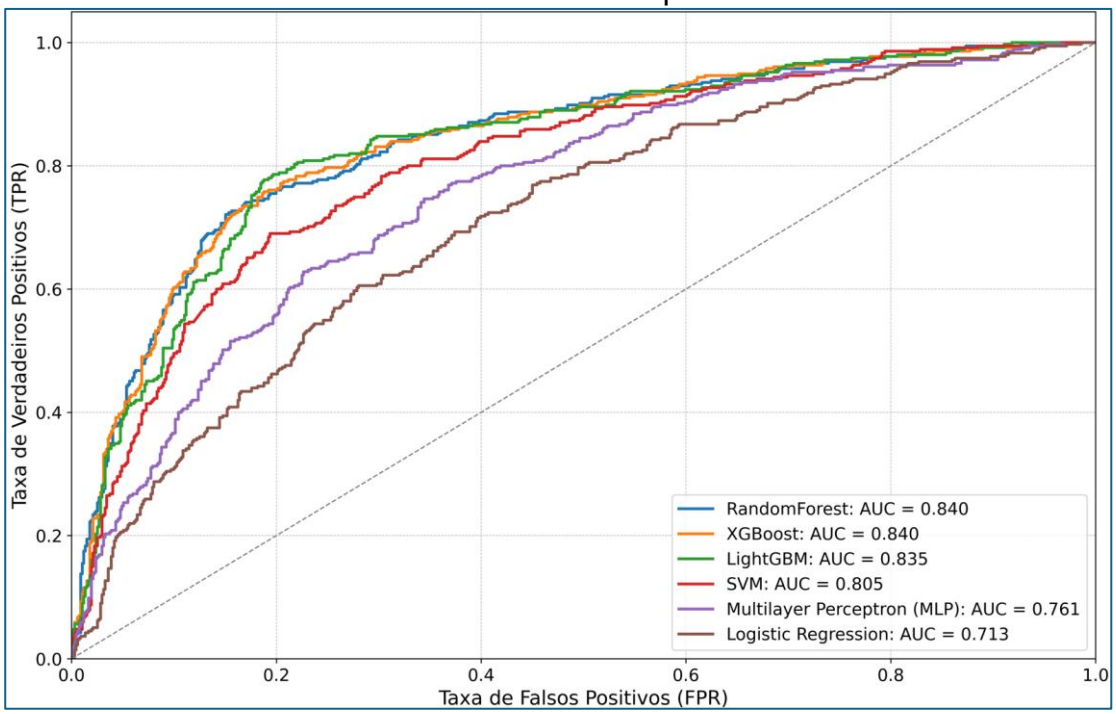
Gráfico 20 – Curva ROC-AUC da BCG com o modelo XGBoost



Fonte: Elaborado pelo autor (2026)

Para referência e comparação, o Gráfico 21 apresenta a performance para a cobertura da BCG de todos os modelos analisados, segundo a métrica “Curva ROC-AUC”.

Gráfico 21 – Curva ROC-AUC da BCG: Comparativo de todos os modelos



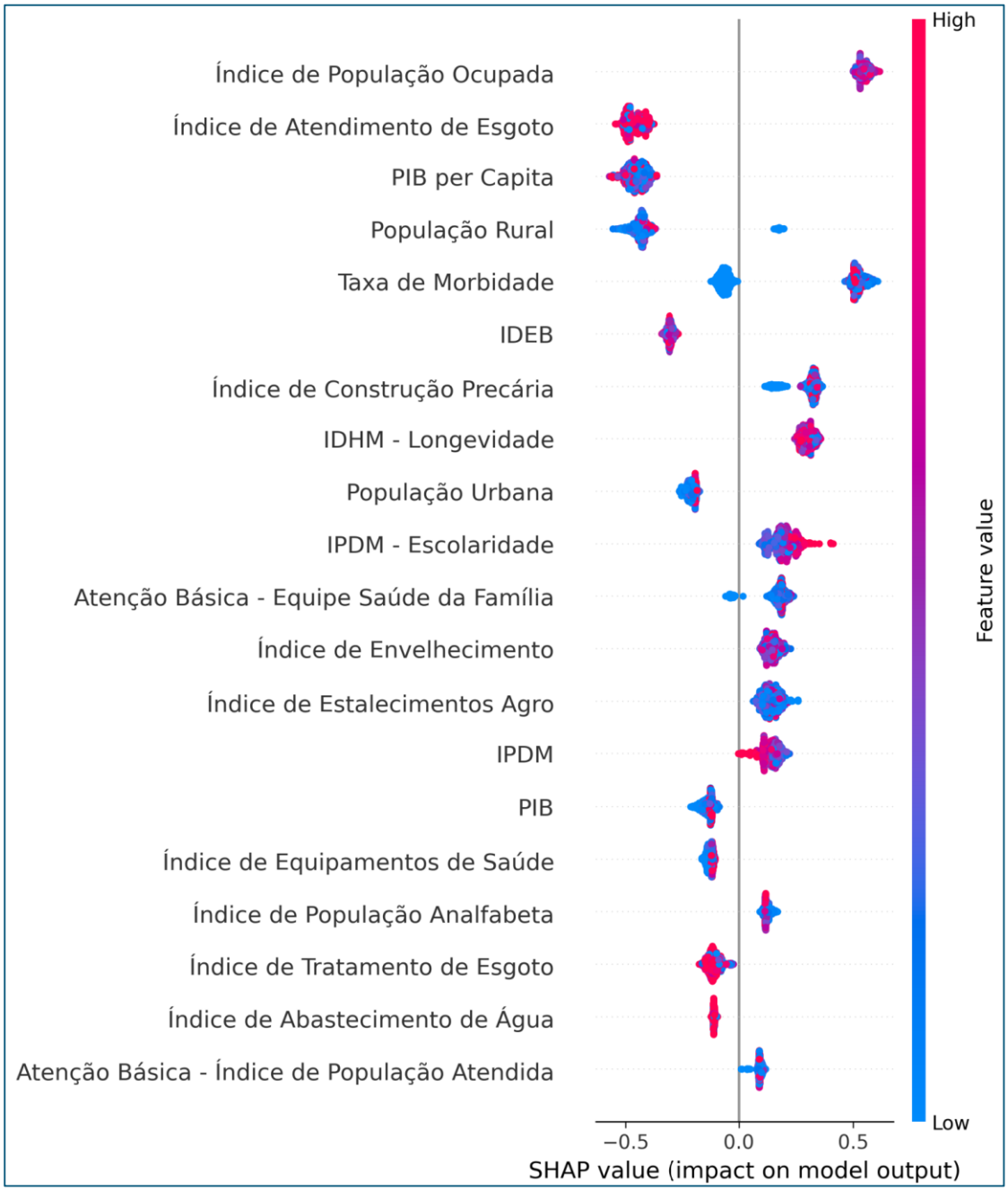
Fonte: Elaborado pelo autor (2026)

Os gráficos com as Curvas ROC-AUC de todos os modelos também são apresentados no apêndice desta pesquisa.

Foram utilizados *SHAP* e *LIME* para composição da explicabilidade do modelo *XGBoost*, o modelo com melhor resultado na comparação e, posteriormente, adicionando uma relação com o ganho de informação de cada variável que participaram da construção da base do modelo.

Iniciando a explicação dos resultados do modelo de melhor performance, o *XGBoost*, na perspectiva da relação dos atributos preditores, são apresentados os Gráficos 22 e 23.

Gráfico 22 – Explicação dos resultados do modelo *XGBoost* com SHAP *values* resumo

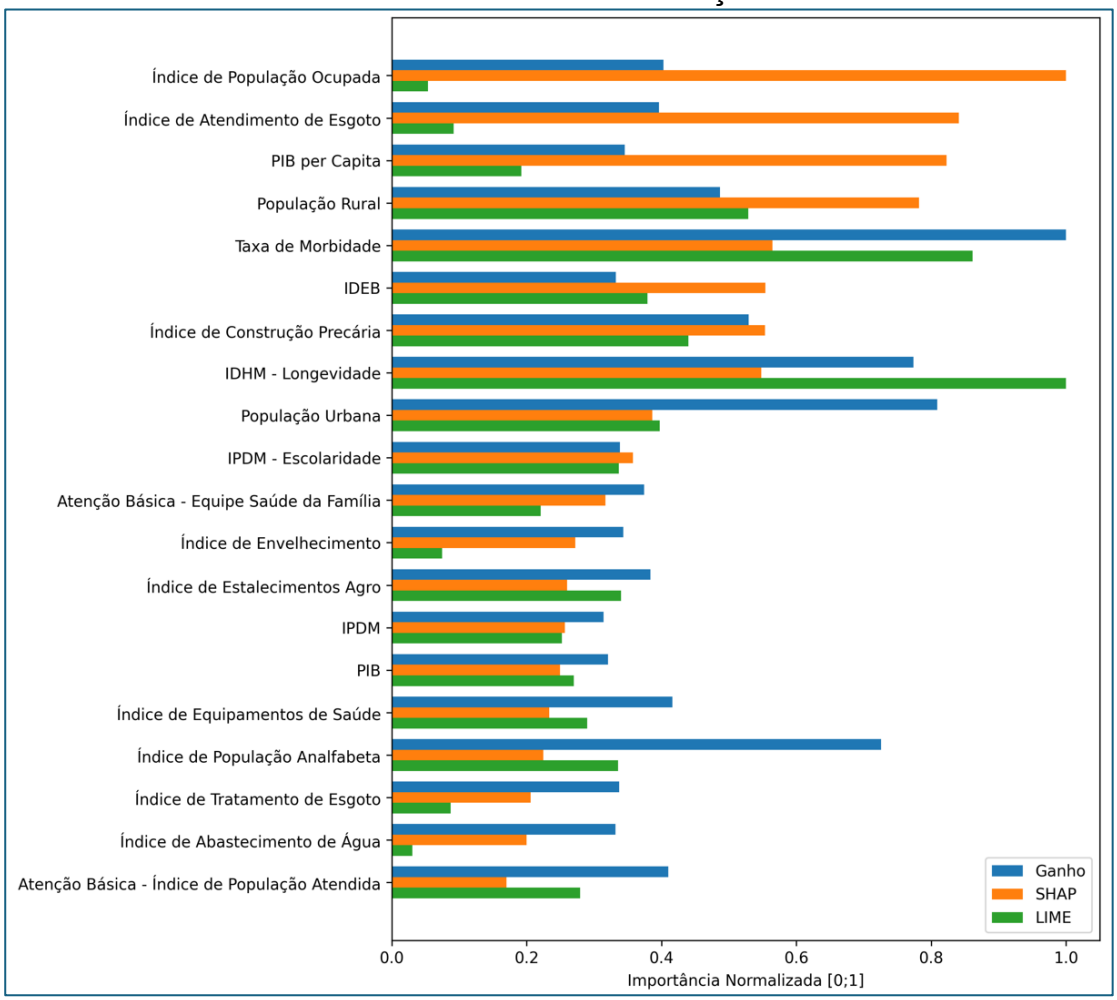


Fonte: Elaborado pelo autor (2026)

No eixo das ordenadas, de cima para baixo, encontram-se a escala das variáveis com maior influência nas previsões. Os valores *SHAP* no eixo das abscissas indicam a direção e a magnitude de sua contribuição, enquanto a cor representa o valor da variável, sendo o rosa para valor alto e o azul para valor baixo (Salih, 2024). Valores à esquerda representam maior influência para a decisão na

classe de baixa cobertura, e à direita na classe de cobertura adequada. As cores complementam a análise indicando seus respectivos valores (ou pesos).

Gráfico 23 – Explicação dos resultados do modelo *XGBoost* com SHAP, LIME e Ganho de Informação



Fonte: Elaborado pelo autor (2026)

XGBoost é um modelo que utiliza, por padrão, árvores de decisão como base, seguindo o conceito de preditores fracos combinados (*ensemble)* para construir a resposta final. Assim, ampliando as formas de explicabilidade, também foram analisados os ganhos de informação de cada variável independente utilizada no conjunto de dados.

A Tabela 14 descreve a correlação entre formas utilizadas para explicação dos resultados do modelo, o Ganho de Informação, o *SHAP* e o *LIME*.

Tabela 14 – Correlação das métricas de explicação do modelo XGBoost para a cobertura vacinal da BCG

	CORRELAÇÃO
Ganho vs SHAP	0,27
SHAP vs LIME	0,70
Ganho vs LIME	0,36

Fonte: Elaborado pelo autor (2026)

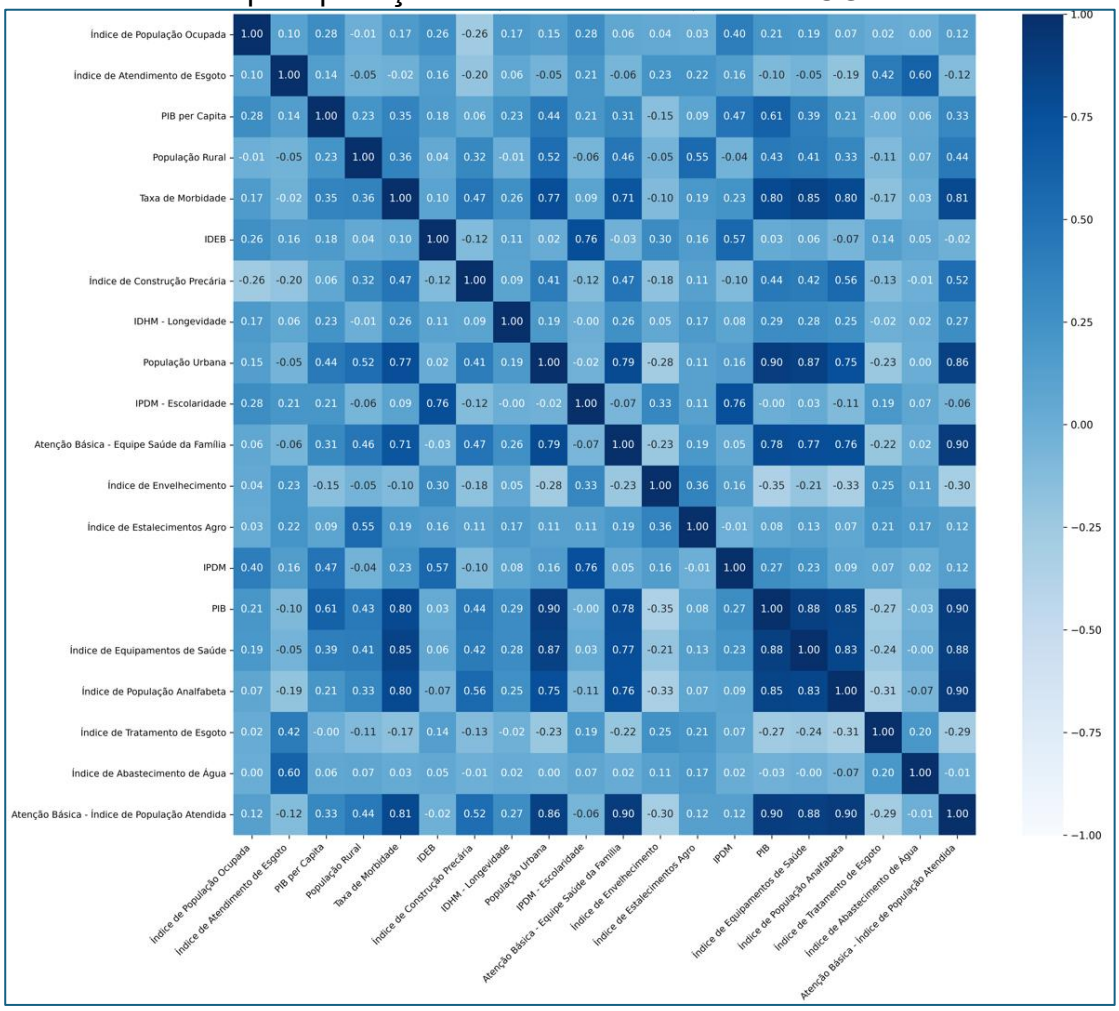
Ganho de Informação, por participar somente da estrutura de base do modelo (arvores de decisão rasas, ou seja, os preditores fracos), tem as menores correlações, porém, como *SHAP* e *LIME* mantém correlação acima de 0,50, as variáveis podem, de fato, serem consideradas como preditores consistentes para o modelo.

Encerrando a análise do modelo, uma vez que a explicabilidade está fundamentada, adicionalmente, na Figura 16, encontra-se a matriz de correlação entre os 20 atributos independentes destacados como os mais importantes para a predição.

É possível notar que muitos preditores mantém alta correlação entre si, ou seja, trata-se de um conjunto de dados que apresenta a característica da multicolinearidade, característica difícil de ser manejada pelos modelos de ML.

A multicolinerialidade já fora analisada anteriormente e, em casos extremos, eliminadas, mantendo-se nos níveis demonstrados na Figura 16, uma vez que modelos que utilizam combinação de métodos (*ensemble*), como é o caso do *XGBoost*, pela sua essência arquitetural, oferecem o suporte mínimo para lidar com esse aspecto.

Figura 18 – Matriz de correlação das 20 variáveis independentes principais para predição da cobertura vacinal da BCG



Fonte: Elaborado pelo autor (2026)

Em última análise, o Quadro 25 apresenta a classificação dos 20 principais preditores, segundo seus tipos ou grupos semânticos, visando retorno melhor demonstrável e direcionado aos profissionais de área de saúde.

Quadro 25 – 20 principais variáveis independentes preditoras do modelo XGBoost para a cobertura vacinal da BCG

VARIÁVEL	DESCRIÇÃO	GRUPO
pop_ocup_10mais_perc	Índice de População Ocupada	DEMOGRAFICO
sp_saneamento_atende_esgoto	Índice de Atendimento de Esgoto	SOCIAL
pib_pib_per_capita	PIB per Capita	ECONOMICO
pop_rural	Índice de População Rural	DEMOGRAFICO
morb_internacao	Taxa de Morbidade	SAUDE
ideb_ideb	Índice de Desenvolvimento da Educação Básica	EDUCACIONAL
dom_construcao_precaria	Índice de Construção Precária	DEMOGRAFICO
idh_idhm_longevidade	IDH – Longevidade	SAUDE
pop_urbana	Índice de População Urbana	DEMOGRAFICO
sp_ipdm_ipdm_escolaridade	IPDM – Escolaridade	EDUCACIONAL
ab_pop_esf	Atenção Básica – Equipe de Saúde da Família	SAUDE
pop_enve_indice	Índice de Envelhecimento Populacional	DEMOGRAFICO
dom_comercial_agro	Índice de Estabelecimentos Agro	DEMOGRAFICO
sp_ipdm_ipdm	Índice Paulista de Desenvolvimento Municipal	SOCIAL
pib_pib	PIB	ECONOMICO
cnes_rec_eqpto_total	Índice de Equipamentos de Saúde	SAUDE
pop_15mais_naoalfa	Índice de População Analfabeta	EDUCACIONAL
sp_saneamento_trata_esgoto	Índice de Tratamento de Esgoto	SOCIAL
sp_saneamento_atende_agua	Índice de Abastecimento de Água	SOCIAL
ab_pop_atendida	Atenção Básica – Índice de População Atendida	SAUDE

RESUMO	DEMOGRAFICO	6
	SAUDE	5
	SOCIAL	4
	EDUCACIONAL	3
	ECONOMICO	2

Fonte: Elaborado pelo autor (2026)

5.2.2. Análise preditiva da Tríplice Viral (SCR)

A Tabela 15 traz a comparação dos resultados dos métodos de ML utilizados para a predição das classes da Tríplice Viral (SCR).

Tabela 15 – Comparativo dos métodos de Machine Learning para cobertura vacinal da Tríplice Viral (SCR)

MODELO	Acuracidade	Precisão	Sensibilidade	Pontuação F1	Curva ROC-AUC
XGBoost	0,647	0,617	0,755	0,679	0,700
LightGBM	0,651	0,648	0,643	0,646	0,698
Random Forest	0,642	0,636	0,647	0,641	0,691
Logistic Regression	0,638	0,613	0,726	0,665	0,672
Multilayer Perceptron (MLP)	0,631	0,630	0,611	0,621	0,665
Support Vector Machine (SVM)	0,626	0,591	0,787	0,675	0,661

Fonte: Elaborado pelo autor (2026)

Assim como no estudo da BCG, o modelo de melhor resultado foi o *XGBoost*, porém, com performance inferior. A capacidade de discriminação das classes medida pela métrica “Curva ROC-AUC” ficou em 70%, indicando ainda poder de separar adequadamente as classes positivas das não positivas. A segunda métrica comparada foi a Pontuação F1, como já destacado, sendo a média harmônica em a Sensibilidade (capacidade de predição correta para os verdadeiros negativos, nesse caso, a cobertura Baixa) e a Precisão (capacidade de predição para os verdadeiros positivos, nesse caso, a cobertura Adequada).

A Tabela 16 mostra a Matriz de Confusão do modelo XGBoost com os 1026 padrões utilizados para testes. O modelo classificou adequadamente 664 padrões, sendo 281 padrões na classe de cobertura baixa ou não adequada e 383 padrões na classe de cobertura adequada. Os falsos positivos e falsos negativos somaram 362 padrões do total de padrões analisados.

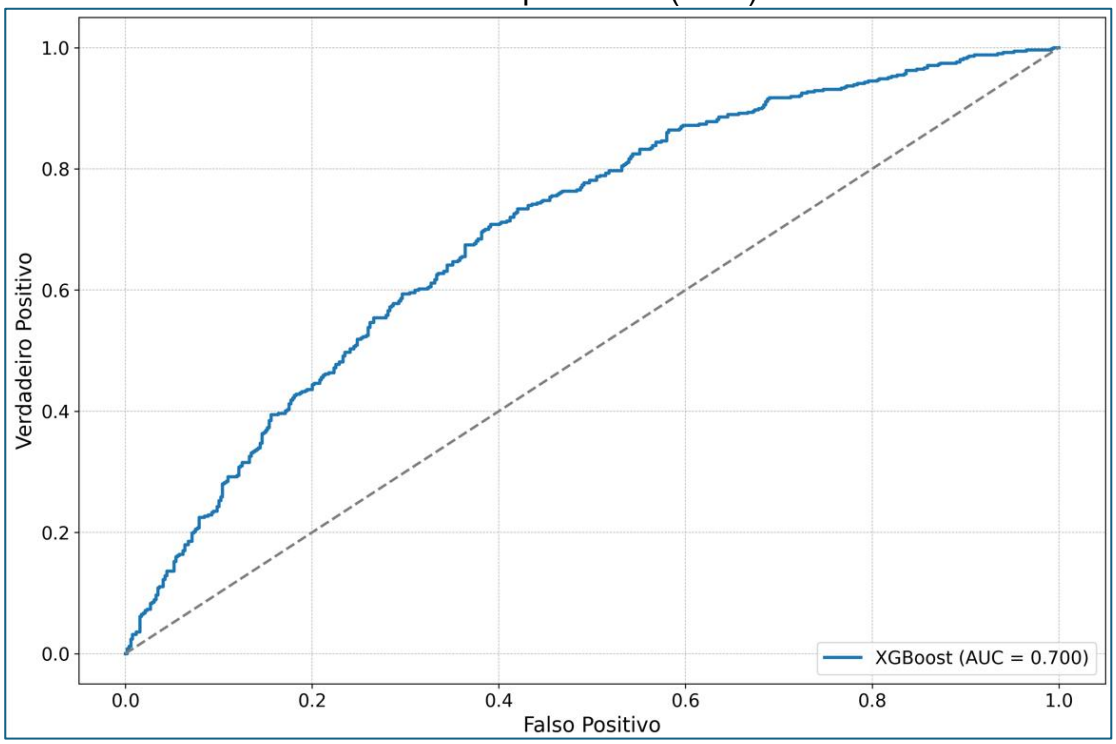
Tabela 16 – Matriz de Confusão da cobertura vacinal da Tríplice Viral (SCR) com modelo XGBoost

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	281 (54,0%)	238 (45,9%)
Adequada	124 (24,5%)	383 (75,5%)

Fonte: Elaborado pelo autor (2026)

O Gráfico 24 apresenta o gráfico da “Curva ROC-AUC do modelo *XGBoost* onde é possível observar a evolução da fronteira discriminatória das classes durante o processamento dos padrões de testes.

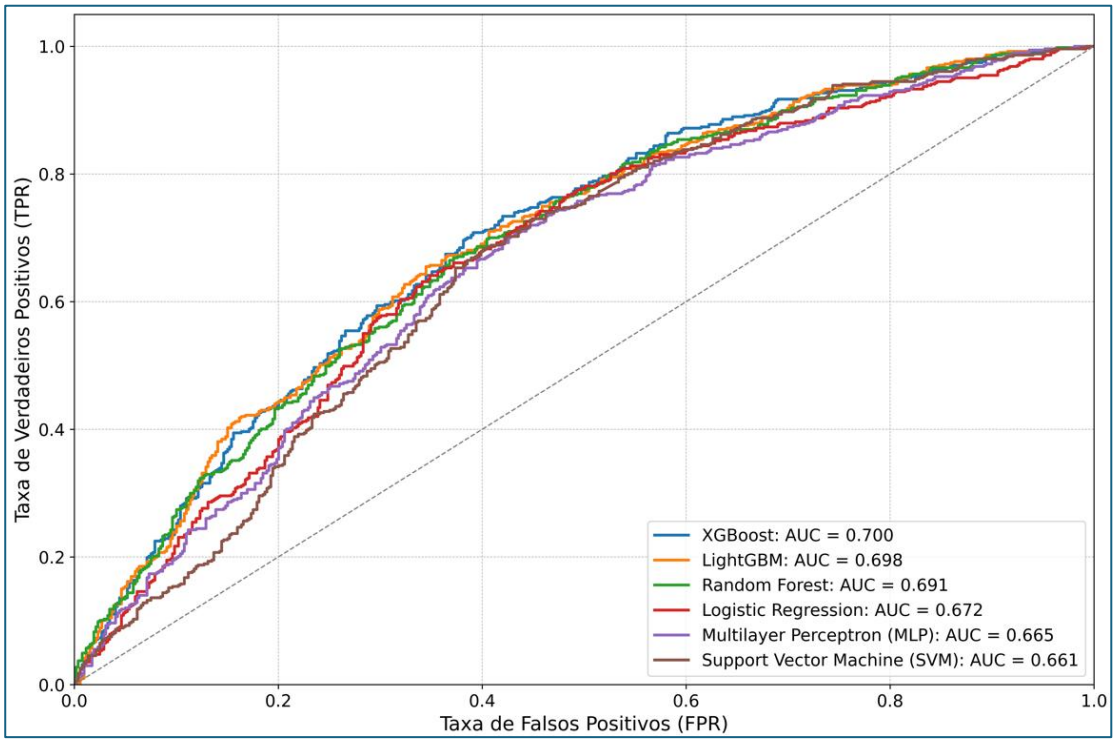
Gráfico 24 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo XGBoost



Fonte: Elaborado pelo autor (2026)

O Gráfico 25 apresentada a comparação de resultados dos modelos, segundo a métrica “Curva ROC-AUC”.

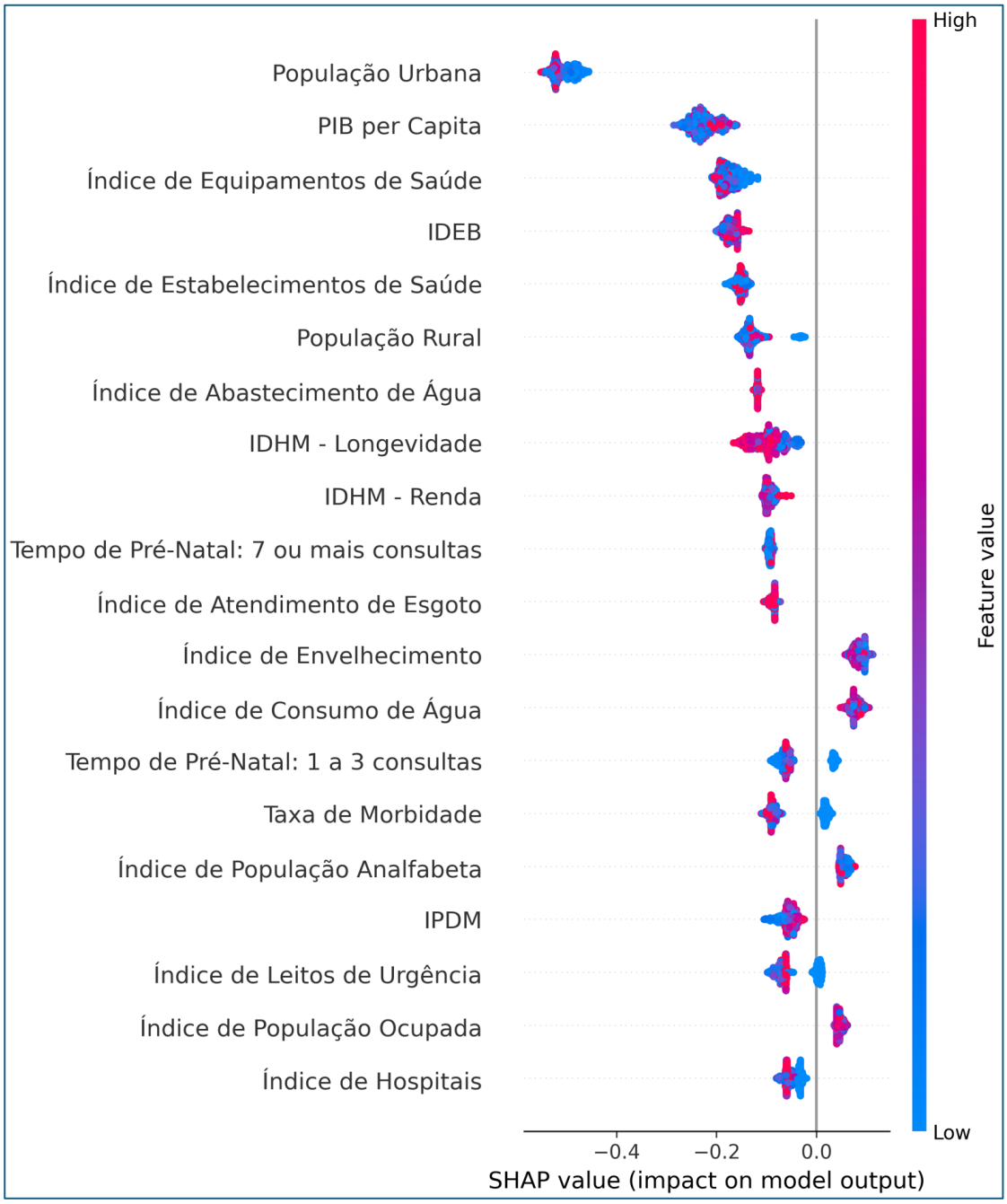
Gráfico 25 – Curva ROC-AUC da Tríplice Viral (SCR): Comparativo de todos os modelos



Fonte: Elaborado pelo autor (2026)

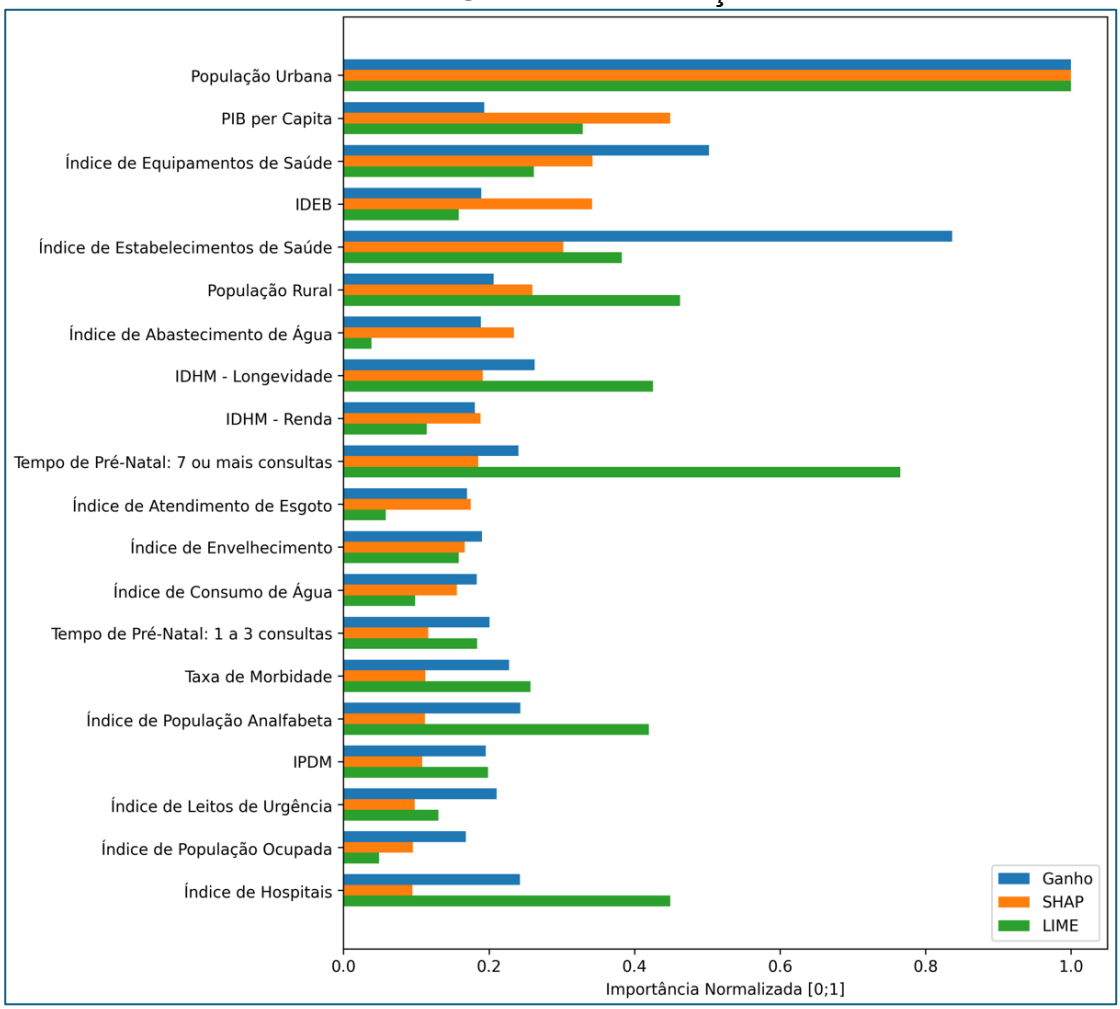
Assim como na avalição da BCG, para explicação dos resultados do modelo de melhor resultado, o *XGBoost*, são apresentados os Gráficos 26 e 27.

Gráfico 26 – Explicação dos resultados do modelo *XGBoost* com SHAP *values* resumo



Fonte: Elaborado pelo autor (2026)

Gráfico 27 – Explicação dos resultados do modelo *XGBoost* com SHAP, LIME e Ganho de Informação



Fonte: Elaborado pelo autor (2026)

SHAP, *LIME* e Ganho de informação, como já destacado na avaliação da BCG, foram as bases para composição da explicabilidade do modelo.

A Tabela 17 descreve a correlação entre formas de explicação do modelo.

Tabela 17 – Correlação das métricas de explicação do modelo *XGBoost* para a cobertura vacinal da Tríplice Viral (SCR)

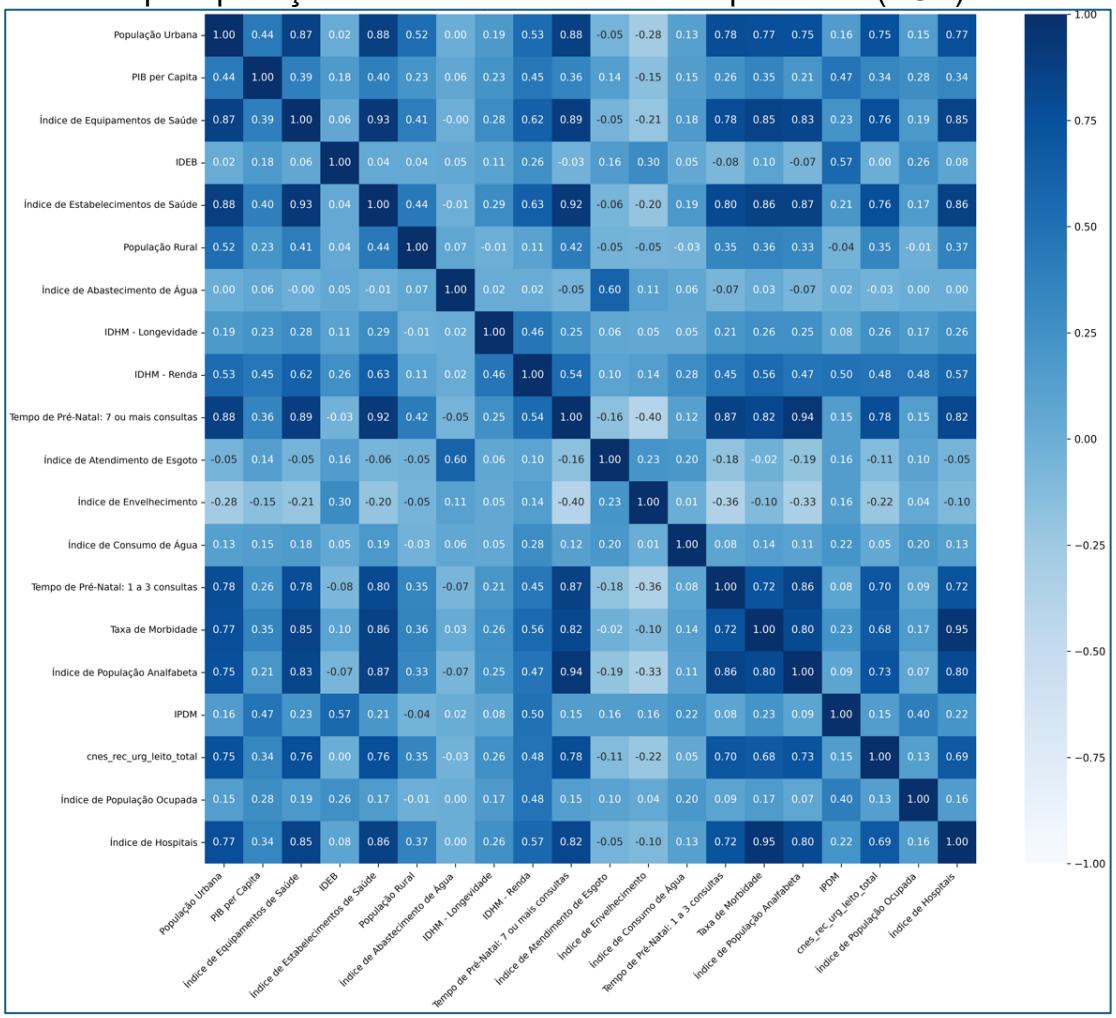
	CORRELAÇÃO
Ganho vs SHAP	0,39
SHAP vs LIME	0,56
Ganho vs LIME	0,63

Fonte: Elaborado pelo autor (2026)

Além da consistência “SHAP vs LIME” já citada anteriormente, na aplicação para avaliação da Tríplice Viral (SCR, Ganho de Informação, elemento estrutural do modelo, também se destacou em consistência quando comparado com LIME, que é um “explicador” de resultados que tem fundamentos de abordagem local, porém, dado ao tamanho reduzido do conjunto de dados, pôde ser aplicado abrangendo quase 98% dos padrões de testes (aproximadamente 1000 padrões).

A Figura 17 apresenta a matriz de correlação entre os 20 principais preditores do modelo. Observa-se a presença de correlações elevadas entre diversas variáveis, caracterizando um cenário de multicolinearidade presente no conjunto de dados. Tal característica fica ainda mais evidente na avaliação das classes pertencentes à primeira dose da vacina Tríplice Viral (SCR). Embora a multicoliearidade tenha sido previamente analisada e eliminada em casos extremos, optou-se por manter esse nível de correlação, considerando que modelos baseados em conjunto de preditores (*ensemble*), como o XGBoost, apresentam certa robustez preditiva diante dessa característica.

Figura 19 – Matriz de correlação das 20 variáveis independentes principais para predição da cobertura vacinal da Tríplice Viral (SCR)



Fonte: Elaborado pelo autor (2026)

Em última análise, assim como para a BCG, o Quadro 26 apresenta a classificação dos principais preditores (20 principais variáveis independentes já mostradas acima), segundo seus tipos ou grupos semânticos, fornecendo ainda mais subsídios para as considerações dos profissionais do campo da saúde coletiva.

Quadro 26 – 20 principais variáveis independentes preditoras do modelo XGBoost para a cobertura vacinal da Tríplice Viral (SCR)

VARIÁVEL	**DESCRIÇÃO**	**GRUPO**
pop_urbana	Índice de População Urbana	DEMOGRAFICO
pib_pib_per_capita	PIB per Capita	ECONOMICO
cnes_rec_eqpto_total	Índice de Equipamentos de Saúde	SAÚDE
ideb_ideb	Índice de Desenvolvimento da Educação Básica	EDUCACIONAL
cnes_estab_total	Índice de Estabelecimentos de Saúde	SAÚDE
pop_rural	Índice de População Rural	DEMOGRAFICO
sp_saneamento_atende_agua	Índice de Abastecimento de Água	SOCIAL
idh_idhm_longevidade	IDH – Longevidade	SOCIAL
idh_idhm_renda	IDH – Renda	ECONOMICO
nv_tempoprenatal_7mais	Tempo de Pré-Natal: 7 ou mais consultas	SAÚDE
sp_saneamento_atende_esgoto	Índice de Atendimento de Esgoto	SOCIAL
pop_enve_indice	Índice de Envelhecimento da População	DEMOGRAFICO
sp_saneamento_consumo_agua	Índice de Consumo de Água	SOCIAL
nv_tempoprenatal_1ª3	Tempo de Pré-Natal: 1 a 3 consultas	SAÚDE
morb_internacao	Índice de Morbidade	SAÚDE
pop_15mais_naoalfa	Índice de População Analfabeta	EDUCACIONAL
sp_ipdm_ipdm	Índice Paulista de Desenvolvimento Municipal	SOCIAL
cnes_rec_urg_leito_total	Índice de Leitos Hospitalares (UTI)	SAÚDE
pop_ocup_10mais_perc	Índice de População com Ocupação	DEMOGRAFICO
cnes_rec_hosp_leito_int_total	Índice de Leitos Hospitalares	SAÚDE

RESUMO	SAUDE	7
	SOCIAL	5
	DEMOGRAFICO	4
	EDUCACIONAL	2
	ECONOMICO	2

Fonte: Elaborado pelo autor (2026)

6. DISCUSSÃO

Esta pesquisa teve por objetivo, a partir da análise de dados secundários obtidos das diversas fontes descritas, fazer uma predição para as coberturas vacinais da BCG e da Tríplice Viral (SCR), utilizando métodos e técnicas de *Machine Learning* (Aprendizado de Máquina). Tratou-se, portando, de um estudo epidemiológico do tipo misto, realizado com os dados dos municípios do estado de São Paulo entre 2016 e 2023.

Ainda, na direção do objetivo amplo, o trabalho foi estruturado seguindo as etapas de tratamento de dados, com uso de técnicas de extrapolação de dados, principalmente, para provisão de dados faltantes, análise exploratória, para aprofundar o conhecimento do conjunto de dados utilizando estatística descritiva e inferencial, testes dos modelos de ML para obtenção do de melhor desempenho e, fechando, para garantir a transparência dos resultados do modelo, a explicabilidade para destacar o conjunto de variáveis independentes preditoras principais.

A premissa da análise de dados e posterior aprofundamento para predição é de que os dados utilizados devem apresentar a consistência necessária, sob o risco de não produzir resultados satisfatórios, ou seja, todos os modelos de ML são dependentes da qualidade dos dados de entrada, o que foi descrito por Skiena (Skiena, 2017) como o princípio “entra lixo, sai lixo”. Essa pesquisa apresentou o desafio de lidar com lacunas temporais nos conjuntos de dados utilizados, não por falta de qualidade no processo de geração (ou entrada) desses dados, mas por eles não existirem em determinadas faixas de tempo dentro do período analisado (por características inerentes às fontes produtoras), os chamados dados faltantes não ao acaso ou informativos (NMAR – IM) destacado por Zhou (Zhou, 2024), cabendo aqui discutir as técnicas empregadas para provisão destes dados, de acordo com os dados de suporte disponíveis.

Sue (Sue, 2022) e Atoum (Atoum, 2025) revisam diversas técnicas para imputação de dados faltantes e, por tratar principalmente, de dados médicos, tornou-se relevante como referência. Nessa pesquisa optou-se por técnicas diferentes dos trabalhos revisados, uma vez que eles partem da premissa da

provisão generalizada dos dados, sem ater-se às possíveis presenças de outras variáveis que pudessem dar apoio à geração dos dados, concentrando-se na utilização somente da variável em questão, sem qualquer aspecto semântico. Outras variáveis pertencentes ao conjunto de dados ofereceram a possibilidade de uma geração de dados menos sintética, porém, demandando maior tempo para aprofundamento no domínio com estudo mais abrangente do cenário, mas, criando possibilidades de provisões mais realistas.

O IDHM (Índice de Desenvolvimento Humano – Municipal) foi complementado utilizando Regressão Polinomial grau 3 utilizando, primeiramente, os dados dos estados, com foco na manutenção de uma coerência macro. Os dados censitários ausentes foram providos envolvendo ajustes dinâmicos, baseados na população real de cada município e indicadores de crescimento econômico (PIB *per capita*) e o próprio IDHM, indicador social. O IPDM (Índice Paulista de Desenvolvimento Municipal) tem periodicidade bienal. Utilizou-se interpolação com médias simples para anos intercalados e estratégias de regressão para os anos finas (extrapolação). Ao final do processo, o conjunto de dados tornou-se mais robusto e menos propenso a causar interferência na capacidade discriminativa dos modelos de ML para predição das classes.

A reflexão que permanece após esse trabalho de ajustes de dados faltantes é a ausência de políticas consistentes de geração e manutenção desses dados secundários abertos que sejam, de fato, voltadas para Ciência de Dados e *Machine Learning*. Os dados secundários públicos disponíveis são valiosos e em grande número, mas ainda carecem de grande esforço do pesquisador para torná-los adequados e menos ruidosos para análise estatística e aplicação dos modelos de predição.

Atendo-se ao estado de São Paulo sob a ótica territorial, a pesquisa revelou um estado de certa complexidade, composto por 645 municípios com características socioeconômicas e demográficas altamente heterogêneas. O estado lidera o ranking dos estados brasileiros em concentração populacional e é somente o décimo segundo em extensão territorial, o que demonstra uma elevada densidade demográfica (7,5 vezes superior à média do pais, segundo o IBGE (IBGE, 2023c), aspecto frequentemente retomado na literatura por facilitar a disseminação de doenças transmissíveis, principalmente pelas vias

respiratórias, como destacado por Kadi (Kadi; Khelfaoui, 2020). Ou seja, o estado abriga 22% da população brasileira em apenas 3% do território nacional e ainda tem uma concentração populacional equivalente a 47% dos residentes em apenas 2,3% dos municípios, segundo dados da Fundação SEADE (SEADE, 2024), o que certamente cria desafios logísticos e de gestão de saúde.

Voltando-se à análise exploratória, o comportamento estatístico da cobertura da vacina BCG que, conforme calendário vacinal brasileiro, é a primeira vacina a ser administrada (logo ao nascer) (BRASIL, 2025b) e tem meta de cobertura de 90% segundo o PNI (Brasil, 2025a), mostra que a média estadual só se aproximou do ideal em 2017 (87,3%) e 2018 (85,2%), ficando muito aquém nos anos subsequentes. Nestes mesmos anos, a mediana supera um pouco a meta, o que indica heterogeneidade nas coberturas dos municípios, com assimetria à esquerda (ou negativa), fazendo com que a média aritmética seja inferior.

O período da Pandemia da COVID-19, como esperado, traz a cobertura da BCG para o patamar histórico mais baixo, dentro do período analisado, marcando média de 50,2% em 2021, o que também ficou marcante na visualização dos mapas e gráficos, com aumento expressivo da classe “Muito Baixa” nos anos de 2020 e 2021. Pós-pandemia, há uma pequena retomada na média, passando para 79,6%, porém segue-se uma nova queda em 2023, marcando 62,1% de cobertura, onde 61,6% dos municípios terminaram o período analisado com cobertura classificada como “Baixa”.

O comportamento da cobertura da BCG destaca os aspectos da variabilidade e da dispersão, o que chama a atenção para a qualidade dos dados. A BCG apresenta um alto coeficiente de variação, com pico de 80,5% em 2021, indicando que a cobertura é muito heterogênea entre os municípios do estado. A presença de valores discrepantes, os chamados *outliers*, pode indicar possíveis problemas de registro de doses ou variações no número de nascidos vivos, o que traz distorções para o cálculo. Exemplo claro é a marca de 647,5% de cobertura em 2022. Anos com dispersões menores, por exemplo 2023, apesar da média baixa, possui um menor desvio padrão o que sugere mais consistência nos dados, com um cenário de cobertura mais generalizado, possivelmente, mais próximo da realidade.

Sob a ótica da demografia e da regionalização, seja administrativa ou de saúde, a BCG apresenta também desigualdades de cobertura. Regiões administrativas com menor concentração populacional, como a de Franca (90,9%) e Presidente Prudente (87,9%), mantiveram desempenho superiores quando comparada com regiões mais populosas, como a de São Paulo (69,7%) e a de São José do Rio Preto (51,7%), o que contrasta com o impacto do porte municipal, onde os municípios de considerados de pequeno porte (até 10 mil habitantes, segundo a Fundação SEADE (SEADE, 2024)), apresentaram as piores médias anuais, o que abre uma frente para discussão dos motivos pelos quais as cidades pequenas, com redes de saúde, em teoria, mais próximas da população, tem mais falhas na cobertura da BCG, quando comparadas às cidades de médio e grande porte.

A análise estatística para a cobertura da Tríplice Viral (SCR), como comportamento central, demonstrou uma realidade diferente e superior à BCG. Apesar da meta de cobertura preconizada pelo PNI ter patamar superior à BCG, meta de 95% (Brasil, 2025a), durante o período analisado, ela apresentou cobertura “Adequada” com estabilidade maior (menor dispersão). A distribuição de probabilidade mostrou uma concentração muito maior em torno das medidas de tendência central, evidenciando um sistema de imunização mais uniforme para esta vacina.

Assim como na BCG, a cobertura da Tríplice Viral (SCR) também foi bastante impactada pela Pandemia da COVID-19, apresentando em 2021 uma queda drástica chegando a 85%. O que se seguiu, pós-pandemia, foi uma recuperação modesta em 2022 (95,2%) e, novamente, queda abaixo da meta em 2023, marcando 89,6% de cobertura média. Sob o prisma da consistência e qualidade dos dados, os achados para 2023, que apresenta o menor desvio padrão (16,7%) e o menor coeficiente de variação (18,6%) de todo o período, sugere que a baixa cobertura não é um problema de municípios isolados, mas um cenário de queda generalizada e real em todo o estado, com 66,5% dos municípios apresentando cobertura classificada como “Baixa” (abaixo dos 95% da meta) o que enfatiza a necessidade de olhares para o risco de recrudescimento do Sarampo, por exemplo.

Com vistas para os contrastes regionais, assim como para a BCG, regiões de menor concentração de habitantes também apresentaram médias de cobertura mais consistentes, a exemplo das regiões de Marília (106,4%) e Central (103,5%), que sustentam médias elevadas durante todo o intervalo analisado. Porém, diferente da BCG, os municípios classificados como de pequeno porte mostraram um melhor cenário de cobertura.

Em última análise sob a perspectiva estatística, a comparação direta das duas vacinas ressalta uma cobertura substancialmente superior da Tríplice Viral (SCR) sobre a BCG em todos os anos da série. Enquanto a Tríplice Viral (SCR) se concentra em torno da média, com uma curva mais estreita e alongada verticalmente a favor de melhor cobertura, a BCG apresenta um espalhamento e uma curva achatada, evidenciando que o sistema de saúde consegue manter níveis mais uniformes para a Tríplice Viral (SCR), seja através de campanhas ou melhor constância na logística de distribuição e aplicação.

Voltando os olhares para a análise preditiva, os resultados dos testes com os modelos de *Machine Learning* para a vacina BCG apontaram o *XGBoost* com melhor performance dentre os demais, com a métrica “Curva ROC-AUC” de 0,84, a mesma obtida pelo modelo *Random Forest*, porém, como segundo critério de comparação, a métrica “Pontuação F1” de 0,713 se destacou. O fato dos modelos baseados em árvores de decisão combinadas, os chamados e*nsembles*, apresentarem uma melhor eficácia nos testes com a BCG reforça os resultados já apresentados na revisão da literatura. O valor obtido de 0,84 na AUC indica forte capacidade discriminatória do modelo em separar municípios com coberta “Adequada” daqueles com cobertura “Baixa”.

Em contraste, modelos lineares como a Regressão Logística (ROC-AUC 0,71) tiveram desempenho inferior, sugerindo que a relação entre os determinantes sociais, demográficos, econômicos, educacionais e de saúde com a cobertura da BCG é complexa e não-linear, sendo mais bem capturada por modelos que utilizam conjuntos de preditores fracos combinados em sua estrutura. A multicolinearidade entre os principais preditores também penaliza modelos baseados em Redes Neurais, como o Multilayer Perceptron (MLP) utilizado nos testes.

O modelo *XGBoost* para a vacina BCG classificou corretamente 822 exemplos de um total de 1016 utilizados para teste. As métricas “Precisão” e “Sensibilidade” obtiveram resultados idênticos, o que significa que o modelo é igualmente equilibrado para identificar corretamente as classes e não gerar excesso de “Falsos Positivos”. Ainda, sobre os casos de “Falsos Negativos”, que correspondem a cobertura “Baixa”, em saúde pública ou qualquer área de saúde, pecar pelo excesso representa melhor conduta diante do risco da não atuação prévia ou antecipada.

Sob o quesito da explicabilidade do modelo, as técnicas SHAP e LIME apontaram as variáveis que mais contribuíram para o modelo. É preciso salientar que o fato de aparecerem como preditores não significa, necessariamente, relação causal, demandando sempre análises semânticas mais profundas e relacionamento com a literatura para essa atribuição. Os destaques sociais e econômicos foram o “Índice de Atendimento de Esgoto” e o “PIB *per capita*”, chamando a atenção para aspectos ligados à infraestrutura municipal e o nível de desenvolvimento econômico municipal. Para as variáveis demográficas, chamou a atenção o “Índice de População Ocupada”, o que remete a dinâmica de trabalho nos municípios paulistas. IDEB e o “Índice de População Analfabeta” também aparecem nos preditores, o que reforça a necessidade de atenção à educação, sobretudo no quesito de possíveis causas de hesitação vacinal.

A despeito da necessidade de análise criteriosa sobre causalidade, com o índice de 84% de confiança na predição para risco ou não de baixa cobertura, baseado nas variáveis empregadas, algumas possíveis ações intervencionistas nas políticas públicas podem ser discutidas e planejadas com maior antecedência.

Na análise preditiva da Tríplice Viral (SCR) os modelos testados não tiveram a mesma performance encontrada na BCG. Os motivos apontam para a maior estabilidade e uniformidade dos dados da Tríplice Viral (SCR), como já descrito na discussão da análise descritiva, o que acaba tornando as fronteiras de decisão menos destacadas dificultando para os modelos encontrarem padrões que melhor distinguam municípios de melhor ou pior cobertura. Na BCG, como mostrado, os dados possuem maior espalhamento.

Como na BCG, o modelo que melhor capturou as nuances nos dados e apresentou, assim, melhor performance foi novamente o *XGBoost*, com a métrica “Curva ROC-AUC” de 0,70, refletindo capacidade discriminatória moderada, porém, ainda destacando-se diante dos modelos lineares ou baseados em redes neurais. Os outros modelos baseados em árvores de decisão agregadas como preditores fracos (*ensembles*) que foram testados (*Random Forest* e *LightGBM*) também tiveram resultados próximos, reforçando a direção no sentido deste tipo de modelo ser preferível para manusear dados tabulares, com as características dos encontrados nesta pesquisa.

Superior à BCG, a métrica “Sensibilidade” de 0,755 obtida nos testes da Tríplice Viral (SCR) que indica “Verdadeiros Positivos”, ou seja, cobertura “Adequada” pode ter uso direcionado para a validação de estratégias que apontam para o sucesso na cobertura vacinal.

Diferentemente da BCG, no item explicabilidade do modelo, destacou-se preditores ligados à dimensão saúde, o que pode refletir que o sucesso vacinal depende diretamente da infraestrutura física e do acompanhamento prévio das famílias no sistema de saúde, como “Tempo de Pré-Natal”, elencado pelo modelo. Outros elementos, como o “PIB *per capita”* e o IDEB voltaram a se repetir também na predição da cobertura da Tríplice Viral (SCR).

Cabe destacar e reforçar essa discussão salientando as literaturas que corroboram os achados desta pesquisa e reforçam as escolhas dos modelos utilizados.

Couto (Couto et al, 2022), em seu estudo para definir prioridades vacinais, destaca que o modelo *XGBoost* obteve 0,803 na métrica “Curva ROC-AUC”, superando o modelo *Random Forest* e demonstrando maior robustez para lidar com os dados de gestão de saúde.

O trabalho de Tadese (Tadese et al, 2024) focou na predição da imunização incompleta entre crianças abaixo de cinco anos de idade no leste da África utilizando, também, modelos baseados em árvores decisão combinadas, obtendo resultados similares à essa pesquisa, novamente destacando o *XGBoost* como modelo mais performático, superando até mesmo o *Random Forest*. Ainda compartilha que as possíveis correlações não-lineares (hipótese

levantada por esse trabalho) são muito presentes nos determinantes sociais de saúde e que trazem dificuldades e falhas para modelos como a Regressão Logística.

Sobre a superação de modelos lineares ou baseados em redes neurais, o trabalho de Avirappattu (Avirappattu et al, 2022) acorda com os achados nesta pesquisa, porém, utilizando modelo diferente do XGBoost, mas que compartilha da mesma ideia e lógica de *boosting*. Ainda destaca, da mesma forma, as restrições rígidas de normalidade e presença de multicolinearidade, problema comum em dados municipais, tais como os enfrentados nessa pesquisa.

No quesito explicabilidade e as variáveis preditoras, educação e renda, assim como mostrado nos achados sobre IDEB, “Índice de População Analfabeta” e “PIB *per capita*”, em estudos realizados na Itália por Carrieri (Carrieri, Lagravinese e Resce, 2021) e Etiópia por Demsash (Demsah et al, 2023) e Tadese (Tadese et al, 2024) confirmam que o nível educacional da população e a riqueza dos domicílios são preditores universais do sucesso vacinal.

Acesso à saúde, como “Tempo de Pré-Natal”, citado nessa pesquisa, é reforçado por Demsash (Demsah et al, 2023) e Biswas (Biswas, Trucker e Bauhoff, 2023) que identificaram as visitas de pré-natal como atributo de ganho considerável para garantir a vacinação infantil.

Infraestrutura e urbanização, como “População Urbana” e variáveis que remetem a questões de saneamento aqui levantadas são ratificadas por estudos que mostram como a densidade populacional e a infraestrutura urbana influenciam na logística e a adesão vacinal, como apresentado por Carrieri (Carrieri, Lagravinese e Resce, 2021), Sanderse (Sanderse, 2021) e Ebulue (Ebulue, 2024).

Assim, após essa breve revisão, todos os trabalhos reforçam a superioridade dos modelos baseados em árvores e fundamentam que os achados não podem ser considerados casos isolados, mas uma tendência da Ciência de Dados e *Machine Learning* aplicados à Saúde Coletiva, especialmente quando se trata de heterogeneidade municipal e determinantes demográficos, sociais, econômicos, educacionais e de saúde complexos, com variações não lineares e, muitas vezes, altamente correlacionados.

Para finalizar e enriquecer os questionamentos e achados que respondem às hipóteses levantadas pela pesquisa, cabe destacar alguns pontos de limitações aqui presentes, além de sugerir ampliações para o debate e aprofundamentos futuros.

A dependência exclusiva de bases de dados públicas (DATASUS, IBGE, SEADE etc.) sujeita o modelo a erros de registro na origem, o que levanta observações sobre a qualidade e fidedignidade dos dados secundários. É exemplo a ocorrência de valores discrepantes (*outliers*), como coberturas vacinais superiores a 600% que podem indicar subestimativas populacionais ou erros na inserção de doses aplicadas nos sistemas SI-PNI e e-SUS AB.

O desafio dos dados faltantes da categoria NMAR exigiu estratégias de imputação, interpolação e extrapolação, como as lacunas entre Censos, indicadores bienais (IPDM) ou, simplesmente, ausência de dados para municípios (IDHM). Embora métodos como a regressão polinomial sejam efetivos e tenham sido aplicados para preenchimento das lacunas, essas estimativas podem não capturar variações locais importantes ocorridas nos anos não presentes nos repositórios públicos.

A simplificação e dicotomização para uma classificação binária (classe “Baixa” e classe “Adequada”), apesar de embasado pelas metas do PNI, pode resultar em perdas de informações estatísticas que porventura pudessem ser capturadas por modelos de regressão contínua ou, ainda, pela classificação multiclasse, com variações mais granulares entre as classes “Baixa” e “Adequada”, ou inferiores ou superiores a elas.

Finalizando, trabalhos futuros poderiam ampliar o escopo geográfico e/ou sugerir o emprego da análise de outras vacinais. Contextos socioeconômicos distintos podem contribuir para validar os modelos e técnicas aqui empregados. A Poliomielite e a Pentavalente também enfrentam quedas de cobertura e, igualmente, participam do calendário vacinal.

A queda observada pós-pandemia em ambas as coberturas analisadas abre, também, frente clara para investigação futura. Fatores como a hesitação vacinal e o advento das *Fake News*, marcante do período pandêmico, podem ter se

tornado determinantes ainda mais latentes do que os indicadores socioeconômicos tradicionais.

A critério de refinamento da modelagem de *Machine Learning* proposto, explorar o uso de modelos de aprendizado não-supervisionado para agrupamentos de municípios por perfis de risco vacinal antes da predição, pode oferecer novos preditores de igual ou superior consistência ante aos achados apresentados.

7. CONCLUSÃO

A necessidade premente de entender o comportamento das coberturas vacinais e seus riscos inerentes, que podem levar, em caso de baixas coberturas, ao recrudescimento de doenças imuno evitáveis, como o sarampo, foi o grande motivador para a realização desta pesquisa.

Foi possível observar que a cobertura vacinal da Tríplice Viral (SCR) é superior e mais estável do que a BCG, embora ambas tenham sofrido quedas expressivas, especialmente durante o período da Pandemia da COVID-19. Porém, é enfático destacar que, mesmo no pós-pandemia, por exemplo em 2023, ainda o cenário apresentou baixa generalizada e real, como pode ser constatado através de uma menor dispersão dos dados, indicando que o problema passa a não ser mais de municípios isolados.

Análise comparada de municípios de menor porte populacional mostrou que há maior dificuldades na manutenção da cobertura vacinal da BCG, aplicada na criança ao nascer, mas são mais eficazes em manter as metas para a Tríplice Viral (SCR), vacinas aplicadas aos 12 e 15 meses de vida.

Do ponto de vista da Inteligência Artificial, mais especificamente do *Machine Learning*, os algoritmos baseados em arranjos (ou conjuntos) árvores de decisão (*ensembles*) se mostram mais robustos para lidar com a heterogeneidade dos dados secundários utilizados, atingindo capacidade discriminatória de 0,84 para a BCG e 0,70 para a Tríplice Viral (SCR).

Foi possível concluir que a vacinação e as coberturas não são apenas atos clínicos, mas um fenômeno regido por aspectos multidimensionais, como mostrou os preditores apontados pelo *XGBoost*, modelo de melhor resultado nos testes. As dimensões econômicas, educacionais, de saúde e de infraestrutura tiveram variáveis selecionadas e destacadas nos achados.

Modelos preditivos, como os utilizados por este trabalho, apesar de distar do apontamento de causalidade direta pelas suas próprias complexidades intrínsecas, podem servir como um sistema de alerta antecipado para a gestão de saúde, permitindo identificar municípios em risco do não atingimento das metas, possibilitando propostas de mitigação.

Finalizando, mesmo considerando que os objetivos propostos foram alcançados em sua plenitude, cabe retomar que a qualidade da predição é indissociável da fidedignidade dos dados secundários a serem utilizados, sugerindo ações enfáticas para modernização e ampliação de seus registros. Dessa forma, a Ciência de Dados e o *Machine Learning* podem continuar representando, ainda com mais sucesso, ferramentas importantes para auxílio à saúde no enfrentamento do retrocesso das conquistas vacinais.

8. BIBLIOGRAFIAS

ABEP. Associação Brasileira de Empresas de Pesquisa. Critério Brasil. Disponível em https://abep.org/criterio-brasil/. Acesso em 20 out 2025.

Alegado, R.T., Tumibay, G.M. Statistical and Machine Learning Methods for Vaccine Demand Forecasting: A Comparative Analysis. Journal of Computer and Communications, 2020.

Almeida F., N. Epidemiologia & Saúde: Fundamentos, Métodos, Aplicações. Rio de Janeiro, Guanabara Koogan, 2012.

Alves, M. H. S.; et al. Heterogeneidades temporal e espacial da cobertura vacinal infantil nas capitais brasileiras: proposta metodológica utilizando razão de doses aplicadas e esperadas. Cadernos de Saúde Pública, 2026.

Araujo, M. G. Hanseníase no Brasil. Revista da Sociedade Brasileira de Medicina Tropical. Minas Gerais, 2003.

Atoum, I. The critical role of evaluation metrics in handling missing data in machine learning. International Journal of Advanced and Applied Sciences, 2025.

Avirappattu, G., et al. An optimized machine learning model for identifying socio-economic, demographic and health-related variables associated with low vaccination levels that vary across ZIP codes in California. Preventive Medicine Reports. Elsevier, 2022.

Bagrichevsky, M.; Estevão, A. Saúde Coletiva: Dialogando Sobre Interfaces Temáticas. Editora da UESC, 2015.

Ballalai, I.; Bravo, F. Imunização: tudo o que você sempre quis saber. RMCOM. Rio de Janeiro, 2016.

Banafa, A. Introduction to Artificial Intelligence (AI). River Publishers. Denmark, 2024.

Barata, R. B.; et. al. Inquérito de cobertura vacinal: avaliação empírica da técnica de amostragem por conglomerados proposta pela Organização Mundial da Saúde. Revista Panamericana Salud Publica, 2005.

Barbieri, C. L.; et al. Imunização e cobertura vacinal: passado, presente e futuro. Santos, SP. Editora Universitária Leopoldianum, 2021.

Bar-Lev, S., Reichman, S., Barnett-Itzhaki, Z. Prediction of vaccine hesitancy based on social media traffic among Israeli parentes using machine learning strategies. sr J Health Policy Res, 2021.

Bartneck, C.; et. al. An Introduction to Ethics in Robotics and AI. SpringerBriefs in Ethics. Switzerland, 2021.

Bell, J. Machine Learning: Hands-On for Developers and Technical Professionals. Indianapolis, Indiana. John Wiley & Sons, 2015.

Biswas A., Tucker J., Bauhoff S. Performance of predictive algorithms in estimating the risk of being a zero-dose child in India, Mali and Nigeria. BMJ Glob Health, 2023.

Bonita, R. Beaglehole, R., Kjellström, T. Epidemiologia básica. São Paulo, Santos, 2010.

Braga, A. L. F.; et. al. Manual de Epidemiologia Ambiental. Ed. Universitária Leopoldianum. Santos, 2015.

Brasil. Ministério da Saúde. Portaria nº 2.499, de 23 de setembro de 2019. Altera a Portaria de Consolidação nº 1, de 28 de setembro de 2017. Diário Oficial da União, 24 set., 2019a.

Brasil. Ministério da Saúde. Programa Nacional de Imunizações. Brasília, DF. 2019b.

Brasil. Ministério da Saúde. Programa Nacional de Imunizações. Brasília, 2025a. Disponível em: https://www.gov.br/saude/pt-br/acesso-a-informacao/acoes-e-programas/pni. Acesso em: 10 set. 2025.

Brasil. Ministério da Saúde. Saúde de A a Z. Brasília, 2025a. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z. Acesso em: 10 out. 2025.

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Instrução normativa referente ao calendário nacional de vacinação. Brasília, 2019c.

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Programa Nacional de Imunizações: Coberturas vacinais no Brasil período 2010-2014. Brasília, DF: Ministério da Saúde; Secretaria de Vigilância em Saúde, 2015a.

Brasil. Ministério da Saúde. Vacinação. Brasília, 2025b. Disponível em: https://www.gov.br/saude/pt-br/vacinacao/. Acesso em: 10 out. 2025.

Brasil. Presidência da República – Casa Civil. Decreto no. 12.560, de 23 de julho de 2025. Dispõe sobre a Rede Nacional de Dados em Saúde e sobre as Plataformas SUS Digital e regulamenta o art. 47 e o art. 47-A, caput, § 1º e § 2º, da Lei nº 8.080, de 19 de setembro de 1990. Brasília, 2025c. Disponível em https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/decreto/d12560.htm. Acesso em: 12 de nov. 2025.

Brasil. Ministério da Saúde. Portaria GM/MS no. 5.663, de 31 de outubro de 2024. Altera a Portaria de Consolidação GM/MS nº 01, de 27 de setembro de 2017. Diário Oficial da República Federativa do Brasil. Brasília, 2024. Disponível em:

https://www.in.gov.br/en/web/dou/-/portaria-gm/ms-n-5.663-de-31-de-outubro-de-2024-593693777. Acesso em: 12 nov. 2025.

Brasil. Ministério da Saúde. Portaria nº 1.520, de 30 de maio de 2018. Altera os Anexos IMUNIZAÇÃO E COBERTURA VACINAL 58 XCVIII e XCIX à Portaria de Consolidação nº 5/GM/MS, de 28 de setembro de 2017, com a inclusão de metas e indicadores do Programa de Qualificação das Ações de Vigilância em Saúde - PQA-VS, a partir de 2018. Diário Oficial da República Federativa do Brasil. Brasília, (DF), Seção 1:47-53, 6 junho de 2018.

Braz, R. M.; et. al. Classificação de risco de transmissão de doenças imunopreveníveis a partir de indicadores de coberturas vacinais nos municípios brasileiros. Epidemiologia Serviço Saúde. Brasília, 2016.

Brockwell, P. J.; Davis, R. A. Introduction to Time Series and Forecasting. Springer International Publishing, Switzerland, 2016.

Carrieri, V., Lagravinese, R., Resce, G. Predicting vaccine hesitancy from area-level indicators: A machine learning approach. *Magna Graecia University.* Italy, 2021.

Celi, L. A.; et. al. Leveraging Data Science for Global Health. Springer International Publishing, Switzerland, 2020.

Chang, S.; Fourney, A.; Horvitz, E. Measuring vaccination coverage and concerns of vaccine holdouts from web search logs. Nature Communications, 2024.

Collucci, C. Folha de São Paulo. Saúde. Com pandemia, taxa de cobertura vacinal no país despenca e abre brecha para novos surtos. São Paulo, 2022. Disponível em: https://www1.folha.uol.com.br/equilibrioesaude/2021/06/ com-pandemia-taxa-decobertura-vacinal-no-pais-despenca-e-abre-brecha-para-novos-surtos.shtml.

Couto R. C., et al. Covid-19 vaccination priorities defined on machine learning. Revista Saúde Pública, 2022.

Cutts, F. T.; et al. Vaccination strategies for measles control and elimination: time to strengthen local initiatives. BMC Medicine, 2021.

Dean, J. Big Data, Data Mining, and Machine Learning. Value creation for business Leaders and practicioers. Wiley & SAS business series. United States of America, 2014.

Demsash A. W., et al. Machine learning algorithms' application to predict childhood vaccination among children aged 12–23 months in Ethiopia: Evidence 2016 Ethiopian Demographic and Health Survey dataset. PLoS ONE, 2023.

Dietz, V.; et al. Assessing and monitoring vaccination coverage levels: lessons from the Americas. Rev Panam Salud Publica, v. 16, n. 6, p. 432-42, 2004.

Dinov, I. D. Data Science and Predictive Analytics. Springer International Publishing, Switzerland, 2018.

Dodoo, C. C., et al. Using machine learning algorithms to predict COVID-19 vaccine uptake: A year after the introduction of COVID-19 vaccines in Ghana. Vaccine: X. Elsevier, 2024.

Domingues, C. M. A. S.; Teixeira, A. M. S. Coberturas vacinais e doenças imunopreveníveis no Brasil no período 1982-2012: avanços e desafios do Programa Nacional de Imunizações. Epidemiol. Serv. Saúde, Brasília, v. 22, n. 1, p. 9-27, 2013.

Domingues, C. M.; et al. Coberturas vacinais: como explicar a queda? Controvérsias em Imunizações 2019. 160p. 2019.

Dumpert, F. Foundations and Advances of Machine Learning in Official Statistics. Springer Nature. Switzerland, 2025.

Ebulue, C. C., et al. Leveraging machine learning for vaccine distribution in resourcelimited settings: A synthesis of approaches. International Medical Science Research Journal, 2024.

Ellis, C. M. The Orange Book of Machine Learning – Orange Edition. Carl McBride Ellis, 2024.

Ertel, W. Introduction to Artificial Intelligence. Weingarten, Germany. Springer International Publishing, 2017.

Fernandes, T.: Vacina antivariólica: seu primeiro século no Brasil (da vacina jenneriana à animal). História, Ciências, Saúde. Manguinhos, 1999.

Forsyth, D. Probability and Statistics for Computer Science. Springer International Publishing, 2018.

Ganatra, A.; et. al. Introduction to Explainable AI. Springer Nature Singapore. Singapore, 2024.

Gomes, E. C. de S. Conceitos e Ferramentas da Epidemiologia. Recife. Ed. Universitária da UFPE, 2015.

Granda, C. F. Probability and Statistics for Data Science. Center for Data Science. New York, 2017.

Harrell, F. E. Regression Modeling Strategies. With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer Series in Statistics. Switzerland, 2015.

Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. Springer International Publishing, Switzerland, 2017.

Heaton, J. Artificial Intelligence for Humans, Volume 1: Fundamental Algorithms. Heaton Research, Inc. St. Louis, 2013.

Heumann, C.; Shalabh, M. S. Introduction to Statistics and Data Analysis. Springer International Publishing Switzerland, 2016.

Holipah, H.; et al. Trends, Spatial Disparities, and Social Determinants of DTP3 Immunization Status in Indonesia 2004–2016. Vaccines, 2020.

Holmes, K. K.; et. al. Major Infectious Diseases. International Bank for Reconstruction and Development / The World Bank. Washington, 2017.

Homma, A.; et. al. Vacinas e vacinação no Brasil: horizontes para os próximos 20 anos. Edições Livres. Rio de Janeiro, 2020.

Hurrington, P. Machine Learnig in Action. Manning Publications Co. New York, 2012.

IBGE. Instituto Brasileiro de Geografia e Estatística. Censo 2022: Panorama. Rio de Janeiro, 2023c. Disponível em https://censo2022.ibge.gov.br/panorama. Acesso em 10 nov 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. Cidades e Estados. Rio de Janeiro, 2023a. Disponível em https://www.ibge.gov.br/cidades-e-estados/sp/. Acesso em 10 nov 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. Panorama: São Paulo. Rio de Janeiro, 2023b. Disponível em https://cidades.ibge.gov.br/brasil/sp/panorama. Acesso em 10 nov 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. Produto Interno Bruto dos Municípios. Rio de Janeiro, 2025. Disponível em https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html. Acesso em 10 jan 2026.

IGC SP. Mapa do Estado - Regiões Administrativas. São Paulo, 2014. Disponível em http://geoportal.igc.sp.gov.br/produtos/arquivos/IGC_Regioes_Administrativas_e_Metropolitanas_2014.jpg. Acesso em 20 out. 2025.

Kadi, N.; Khelfaoui, M. Population density, a factor in the spread of COVID-19 in Algeria: statistic study. Bull Natl Res Cent, 2020.

Kirk, M. Thoughtful Machine Learning. Sebastopol, CA. O'Reilly Media, 2015.

Kubat, M. An Introduction to Machine Learning. Springer International Publishing, Switzerland, 2017.

Kuhn, M. Johnson, K. Applied Predictive Modeling. Springer Science and Business Media. New York, 2013.

Kumar, Y., et al. Machine Learning and Deep Learning Based Time Series Prediction and Forecasting of Ten Nations' COVID-19 Pandemic. SN Computer Science, 2023.

Lamane, H.; et. al. Interpreting machine learning models based on SHAP values in predicting suspended sediment concentration. International Journal of Sediment Research, 2024.

Leung, A. K. C.; Hon, K. L.; Leong, K. F. Rubella (German measles) revisited. Hong Kong Medical Journal. Hong Kong, 2019.

Lima, E. J. F. Rotinas de Imunização. Vaccine. Recife, 2020.

Liu, S., Du, H., Feng, M. Robust Predictive Models in Clinical Data—Random Forest and Support Vector Machines. In: Celi, L., Majumder, M., Ordóñez, P., Osorio, J., Paik, K., Somai, M. (eds) Leveraging Data Science for Global Health. Springer, 2020.

Lu, Y. et al. Reinforcement learning relieves the vaccination dilema. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2023.

Luger, G. F. Inteligência Artificial. São Paulo, SP. Pearson Education do Brasil, 2013.

Luiz, R. R.; Struchiner, C. J. Inferência Causal em Epidemiologia. O Modelo de Respostas Potenciais. Rio de Janeiro. Ed. FIOCRUZ, 2002.

Macdonald, N. E. Vaccine hesitancy: Definition, scope and determinants. Elsevier Ltd, 2015.

Machado, F. N. R. Projeto e implementação de banco de dados. São Paulo Erica, 2014.

Machado, R.; Cardoso, B. SBIM 20 anos: conectando conhecimentos e promovendo a prevenção. Rio de Janeiro: Sociedade Brasileira de Imunizações, 2018.

Magazzino C., Mele M., Coccia M. A machine learning algorithm to analyse the effects of vaccination on COVID-19 mortality. Epidemiology and Infection, 2022.

Mahdavinejad, M. S.; et. al. Machine Learning for Internet of Things Data Analysis: A Survey. Journal of Digital Communications and Networks, 2018.

May, S.; Stewart, A. A. Interpreting Machine Learning Predictions with SHAP and LIME for Transparent Decision Making. International Journal of Computer Science and Mathematical Theory, 2025.

Moraes, J. C.; et. al. Cobertura vacinal no primeiro ano de vida em quatro cidades do estado de São Paulo, Brasil. Revista Panamericana Salud Publica, 2000.

Moraes, J. C.; et. al. Qual é a cobertura vacinal real? Epidemiologia e Serviços de Saúde 2003, 2003.

Moraes, J. C.; Ribeiro, M. C. S. A. Desigualdades sociais e cobertura vacinal: uso de inquéritos domiciliares. Revista Brasileira de Epidemiologia, 2008.

Mougan, C.; et. al. How to Data in Datathons. 37th Conference on Neural Information Processing Systems. New Orleans, 2023.

Muloliwa, A. M.; et. al. Impact of vaccination on the incidence of measles in Mozambique in the period 2000 to 2011. Cad. Saúde Pública, Rio de Janeiro, 2013.

Observatório das Vacinas. Mapa .de Cobertura. Santos, 2024. Disponível em https://www.observatoriodasvacinas.com.br. Acesso em 15 set 2025.

Oliveira, T. R. S. Vacinas: Cenário histórico e importância como medida essencial no enfrentamento das pandemias. Synapse Editora. Belo Horizonte, 2021.

Palmeira, G. Epidemiologia. In: ROZENFELD, S., org. Fundamentos da Vigilância Sanitária [online]. Rio de Janeiro: Editora FIOCRUZ, 2000, pp. 135-194.

Pamplona, Y. A. P.; Martins, L. C.; Barbieri, C. L. A. Tutorial de coleta de dados no TABNET. Santos: Leopoldianum, 2020.

Plotkin, S. A. History of Vaccine Developlment. Springer Science. New York, 2011.

Procópio, M. J. Controle da Tuberculose: Uma proposta de integração ensino-serviço. Editora FIOCRUZ. Rio de Janeiro, 2014.

Provost, F.; Fawcett, T. Data Science for Business. O'Reilly Media. Sebastopol, 2013.

Radford, A.; et. al. Improving Language Understanding by Generative Pre-Training. Computer Science, Linguistics. Chicago, 2018.

Raschka, S.; Mirjalili, V. Python Machine Learning. Packt Publishing. Birmingham, 2019.

Revathi, A.; et. al. Correlation between Vaccination and Child Mortality Rate Using Multivariate Linear Regression Model. Journal of Information & Knowledge Management. World Scientific Publishing Co, 2024.

Rozman, L. M.; et al. Cobertura e atraso vacinal nas coortes de nascidos em 2019 e 2020: inquérito domiciliar em Cubatão, estado de São Paulo, Brasil. Cadernos de Saúde Pública, 2025.

Russel, S.; Norvig, P. Inteligência Artificial. Rio de Janeiro, RJ. Elsevier, 2013.

Salih, A.; et al. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. Advanced Intelligent System, 2024.

Salgado, C.M., Vieira, S.M. Machine Learning for Patient Stratification and Classification Part 1: Data Preparation and Analysis. In: Celi, L., Majumder, M., Ordóñez, P., Osorio, J., Paik, K., Somai, M. (eds) Leveraging Data Science for Global Health. Springer, 2020a.

Salgado, C.M., Vieira, S.M. Machine Learning for Patient Stratification and Classification Part 3: Supervised Learning. In: Celi, L., Majumder, M., Ordóñez,

P., Osorio, J., Paik, K., Somai, M. (eds) Leveraging Data Science for Global Health. Springer, 2020b.

Sameen, F., et al. Improving Routine Immunization Coverage Through Optimally Designed Predictive Models. Tech Science Press, 2022.

Sanderse, T. An Application of Machine Learning Models to Predict the Covid-19 Vaccine Coverage of Sub-National Areas in Flanders Using Socio-Demographic Variables. Data Science & Society, 2021.

Santos, F. B. O.; et. al. Imunização e sua história sob as lentes da Enfermagem. Editora ABen. Brasilia, 2021.

SEADE. Fundação Sistema Estadual de Análise de Dados. SEADE Municípios. São Paulo, 2024. Disponível em https://municipios.seade.gov.br/. Acesso em 20 out 2025.

Shafranovich, Y. Common format and MIME type for Comma-Separated Values (CSV) files. Internet Engineering Task Force − IETF, 2005. Disponível em: https://datatracker.ietf.org/doc/html/rfc4180.

Skiena, S. S. The Data Science Design Manual. Springer International Publishing, Switzerland, 2017.

Stahl, B. C. Artificial Intelligence for a Better Future An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies. SpringerBriefs in Research and Innovation Governance. Switzerland, 2021.

Seu, K.; Kang, M.; Lee, H. An Intelligent Missing Data Imputation Techniques: A Review. International Journal on Informatics Visualization, 2022.

Tadese, Z. B.; et al. Prediction of incomplete immunization among under-five children in East Africa from recent demographic and health surveys: A machine learning approach. In: Scientific Reports, 2024.

Thomaz, B. A. F.; et al. Cadernos de Saúde da Família: Conceitos e Ferramentas da Epidemiologia. São Luis. EDUFMA, 2015.

Vaswani, A.; et. al. Attention Is All You Need. 31st Conference on Neural Information Processing Systems. Long Beach, 2017.

Veneu, F.; et. al. Que ideias nos transmitem as fake news sobre as vacinas contra a Covid-19? Desafios para o Ensino de Ciências e a divulgação científica. Ciência & Educação. Bauru, 2023.

Weng, W.H. Machine Learning for Clinical Predictive Analytics. In: Celi, L., Majumder, M., Ordóñez, P., Osorio, J., Paik, K., Somai, M. (eds) Leveraging Data Science for Global Health. Springer, 2020.

WORLD HEALTH ORGANIZATION (WHO). Vaccines and immunization. Genebra WHO, 2024. Disponível em: https://www.who.int/health-topics/vaccines-and-immunization#tab=tab_1. Acesso em: 10 set. 2024.

Wulff, P.; Kubsch, M.; Krist, C. Applying Machine Learning in Science Education Research. Springer Texts in Education. Switzerland, 2025.

Xie, C. Biomedical Signal Processing: An ECG Application. In: Celi, L., Majumder, M., Ordóñez, P., Osorio, J., Paik, K., Somai, M. (eds) Leveraging Data Science for Global Health. Springer, 2020.

Zetzsche, M. F. Epidemiologia. Indaial. UNIASSELVI, 2015.

Zhou, Y.; Aryal, S.; Bornadjenek, M. R. A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms. Elsevier, 2024.

Zorzetto, R. As razões da queda na vacinação. Pesquisa FAPESP, São Paulo, v. 270, p. 19-24, 2018.

9. APÊNDICES

9.1. Matriz de Confusão dos demais modelos utilizados para a BCG

Tabela 18 – Matriz de Confusão da BCG com o modelo Support Vector Machine (SVM)

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	585 (87,2%)	86 (12,8%)
Adequada	154 (38,8%)	243 (61,2%)

Fonte: Elaborado pelo autor (2026)

Tabela 19 – Matriz de Confusão da BCG com o modelo Logistic Regression

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	450 (67,1%)	221 (32,9%)
Adequada	132 (37,2%)	223 (62,8%)

Fonte: Elaborado pelo autor (2026)

Tabela 20 – Matriz de Confusão da BCG com o modelo Mutilayer Perceptron (MLP)

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	527 (78,5%)	144 (21,5%)
Adequada	141 (39,7%)	214 (60,3%)

Fonte: Elaborado pelo autor (2026)

Tabela 21 – Matriz de Confusão da BCG com o modelo Random Forest

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	583 (86,9%)	88 (13,1%)
Adequada	112 (31,5%)	243 (68,5%)

Fonte: Elaborado pelo autor (2026)

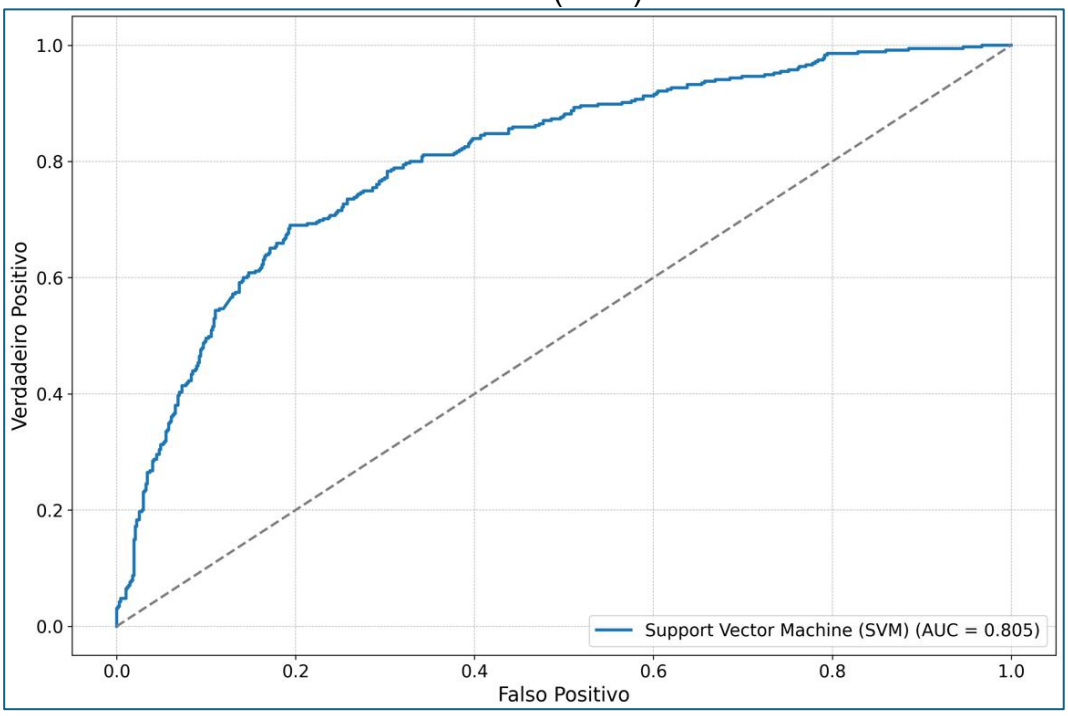
Tabela 22 – Matriz de Confusão da BCG com o modelo LightGBM

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	552 (82,2%)	119 (17,7%)
Adequada	88 (24,8%)	267 (75,2%)

Fonte: Elaborado pelo autor (2026)

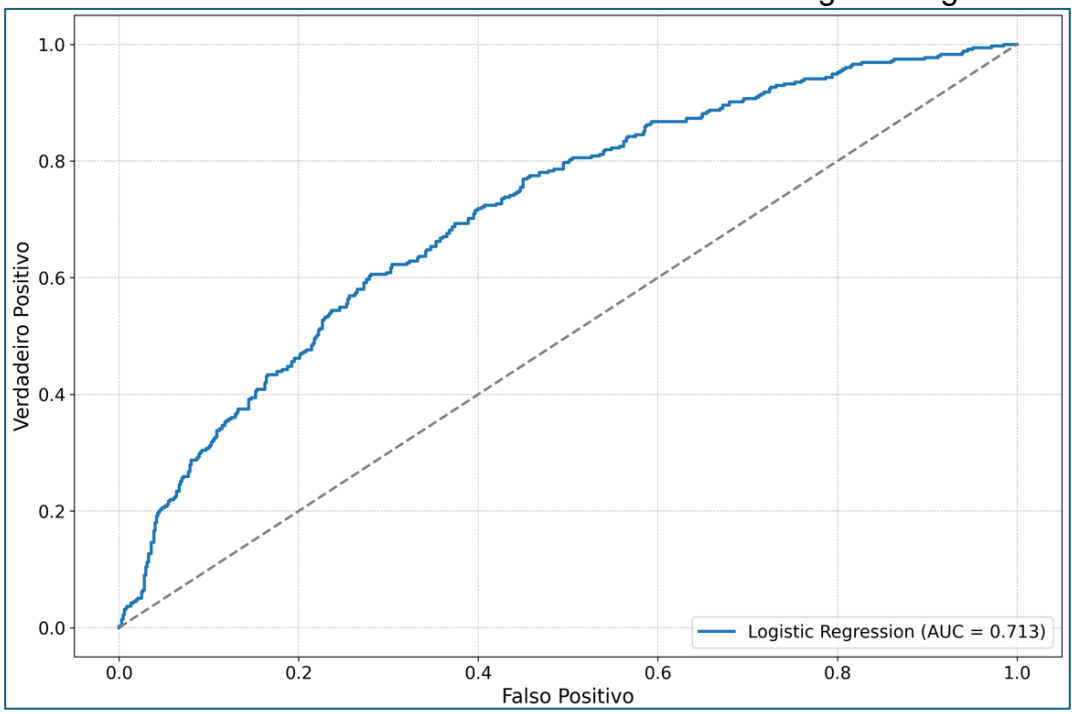
9.2. Curva ROC-AUC dos demais modelos utilizados para a BCG

Gráfico 28 – Curva ROC-AUC da BCG com o modelo Support Vector Machine (SVM)



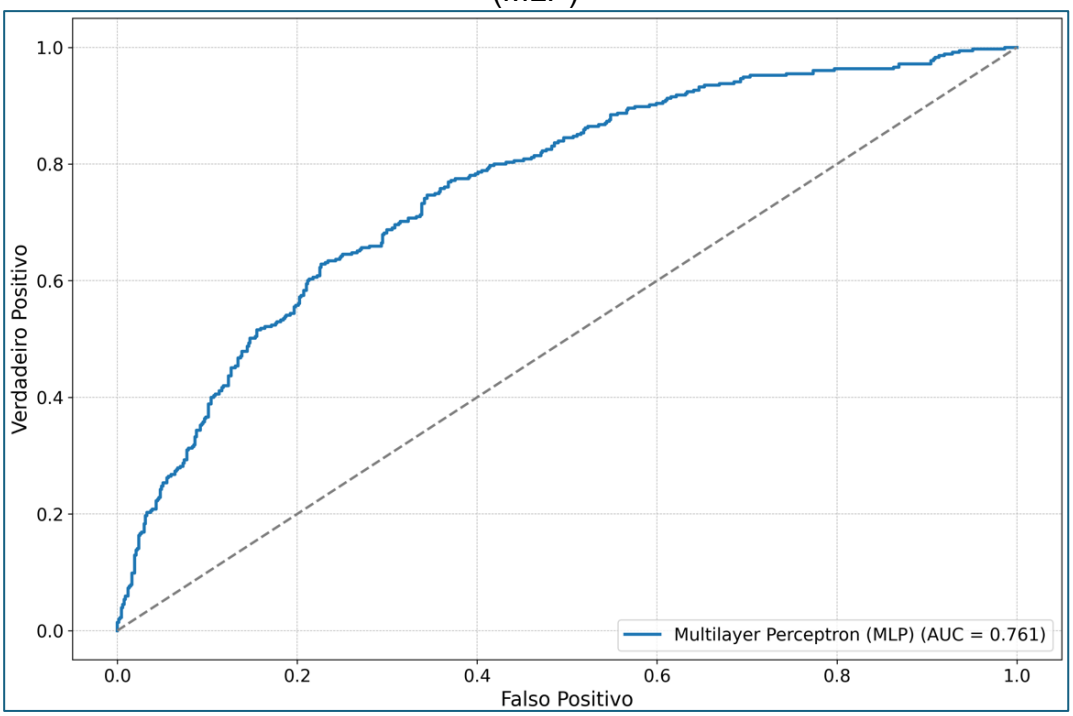
Fonte: Elaborado pelo autor (2026)

Gráfico 29 – Curva ROC-AUC da BCG com o modelo Logistic Regression



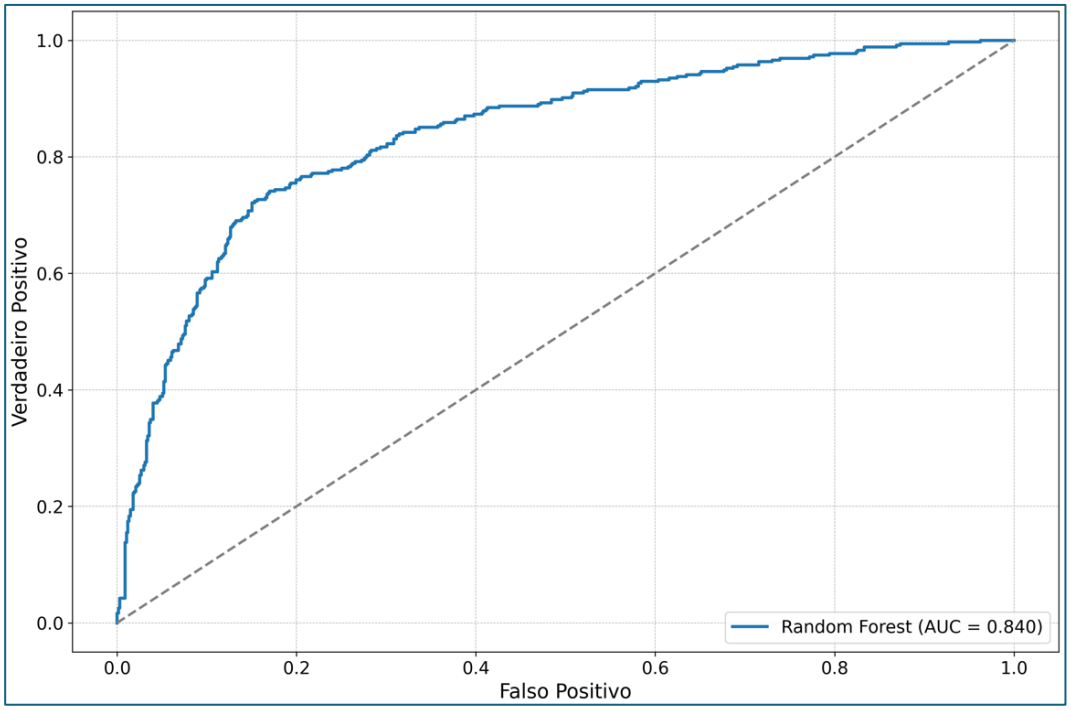
Fonte: Elaborado pelo autor (2026)

Gráfico 30 – Curva ROC-AUC da BCG com o modelo Mutilayer Perceptron (MLP)



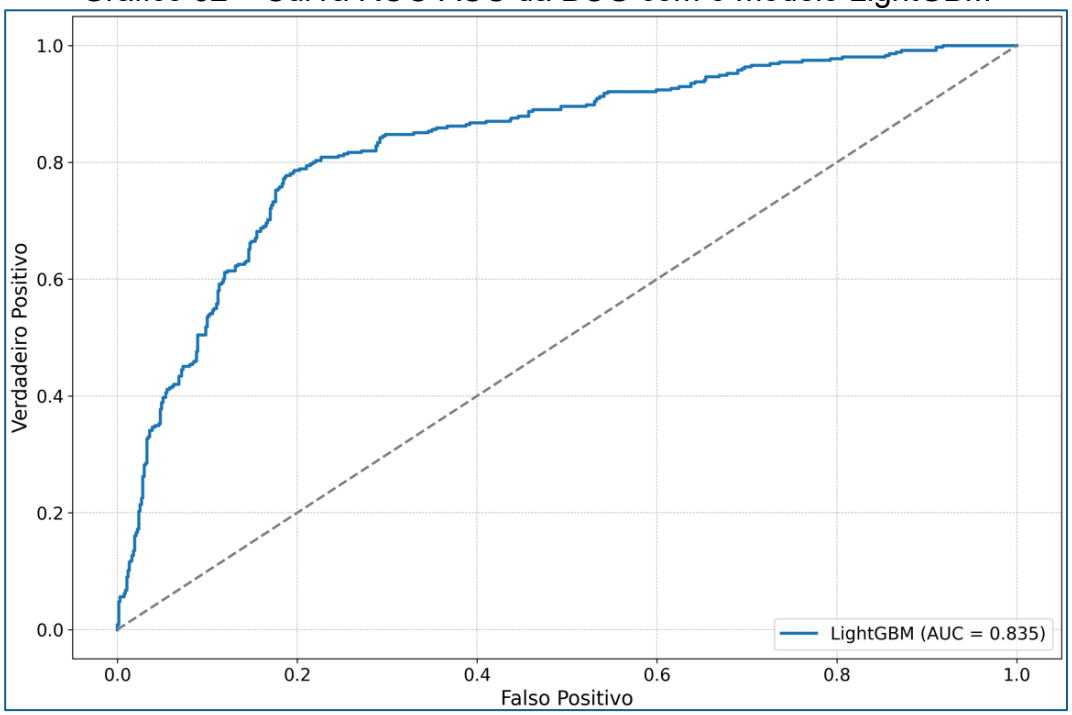
Fonte: Elaborado pelo autor (2026)

Gráfico 31 – Curva ROC-AUC da BCG com o modelo Random Forest



Fonte: Elaborado pelo autor (2026)

Gráfico 32 – Curva ROC-AUC da BCG com o modelo LightGBM



Fonte: Elaborado pelo autor (2026)

9.3. Matriz de Confusão dos demais modelos utilizados para a Tríplice Viral (SCR)

Tabela 23 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo Support Vector Machine (SVM) –

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	243 (46,8%)	276 (53,2%)
Adequada	108 (21,3%)	399 (78,7%)

Fonte: Elaborado pelo autor (2026)

Tabela 24 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo Logistic Regression

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	287 (55,3%)	232 (44,7%)
Adequada	139 (27,4%)	368 (72,6%)

Fonte: Elaborado pelo autor (2026)

Tabela 25 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo Mutilayer Perceptron (MLP)

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	337 (64,9%)	182 (35,1%)
Adequada	197 (38,9%)	310 (61,1%)

Fonte: Elaborado pelo autor (2026)

Tabela 26 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo Random Forest

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	331 (63,8%)	188 (36,2%)
Adequada	179 (35,3%)	328 (64,7%)

Fonte: Elaborado pelo autor (2026)

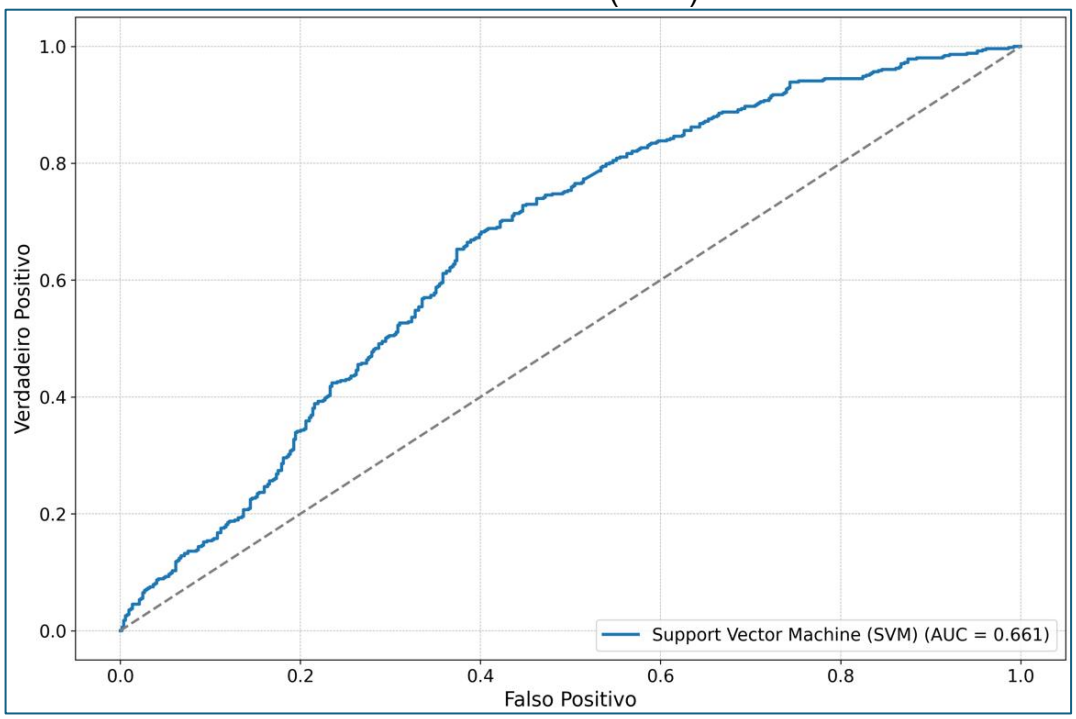
Tabela 27 – Matriz de Confusão da Tríplice Viral (SCR) com o modelo LightGBM

Classe Real	Predita: Baixa	Predita: Adequada
Baixa	342 (65,9%)	177 (34,1%)
Adequada	181 (35,7%)	326 (64,3%)

Fonte: Elaborado pelo autor (2026)

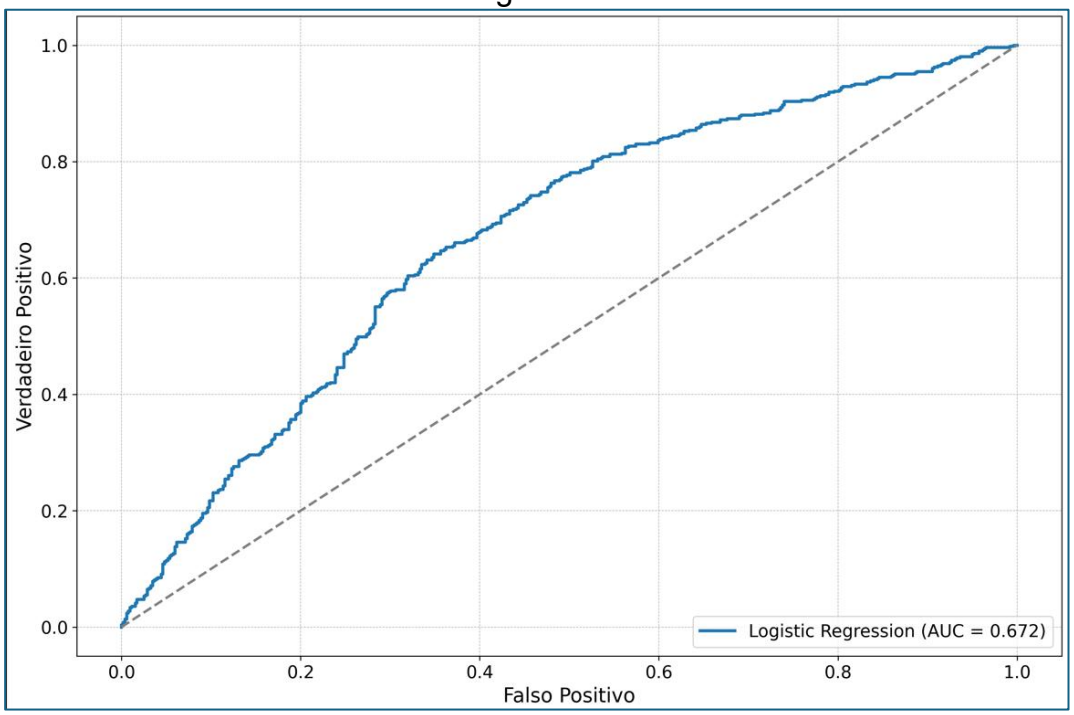
9.4. Curva ROC-AUC dos modelos utilizados para a Tríplice Viral (SCR)

Gráfico 33 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo Support Vector Machine (SVM)



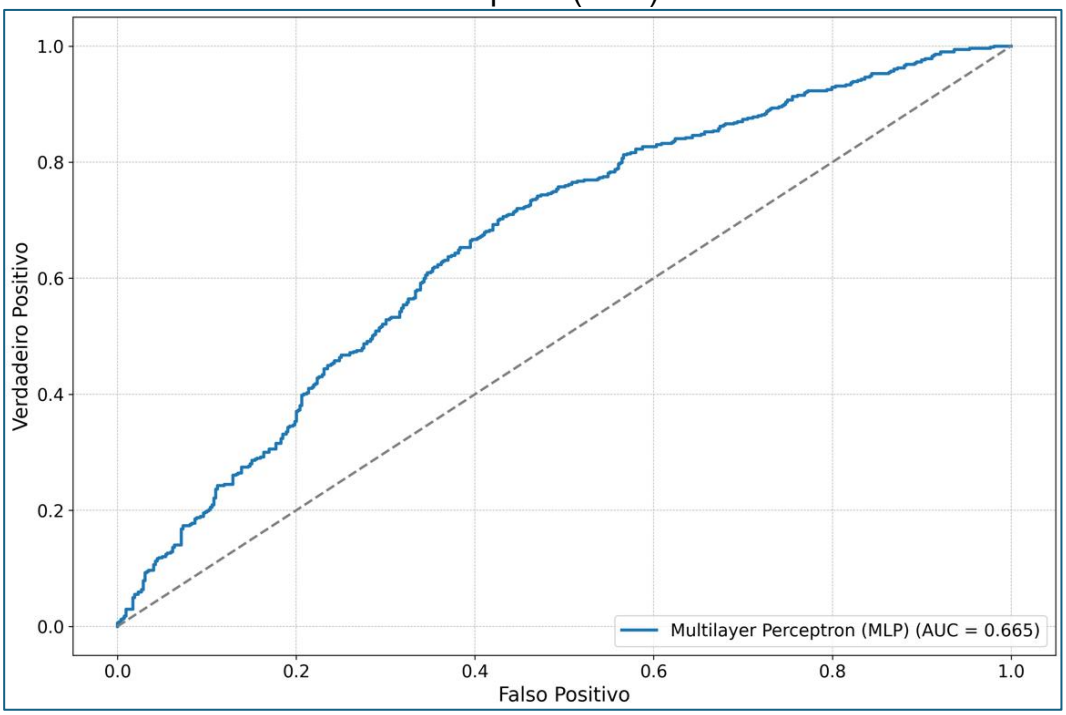
Fonte: Elaborado pelo autor (2026)

Gráfico 34 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo Logistic Regression



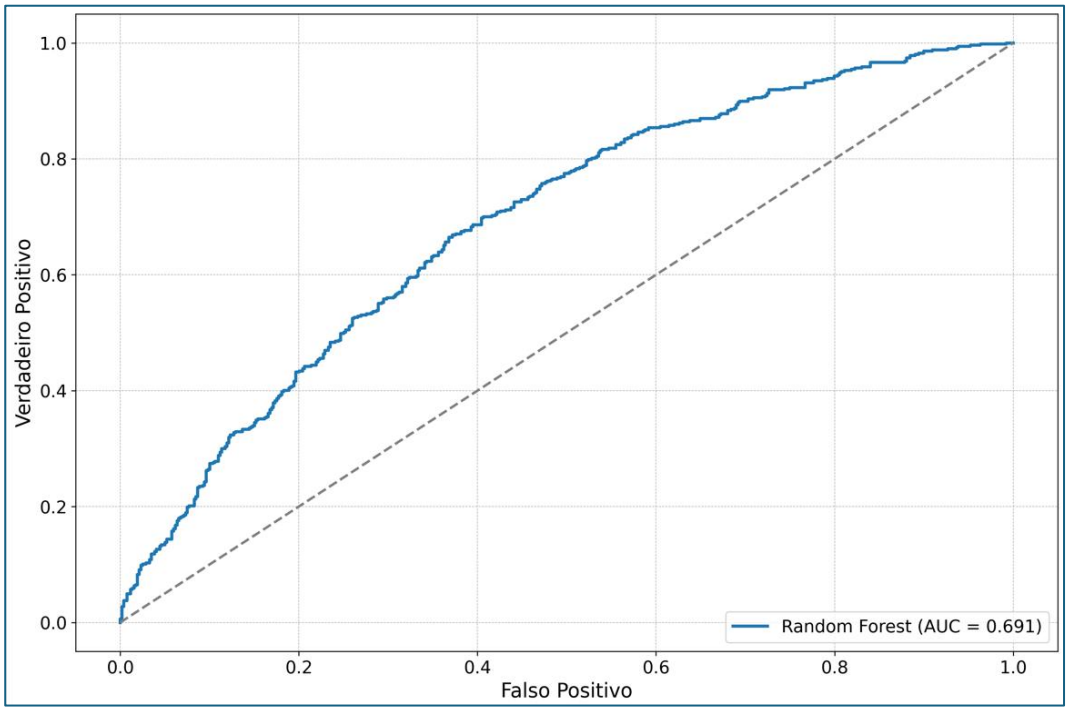
Fonte: Elaborado pelo autor (2026)

Gráfico 35 – Curva ROC-AUC da Tríplice Viral (SCR) com o Mutilayer Perceptron (MLP)



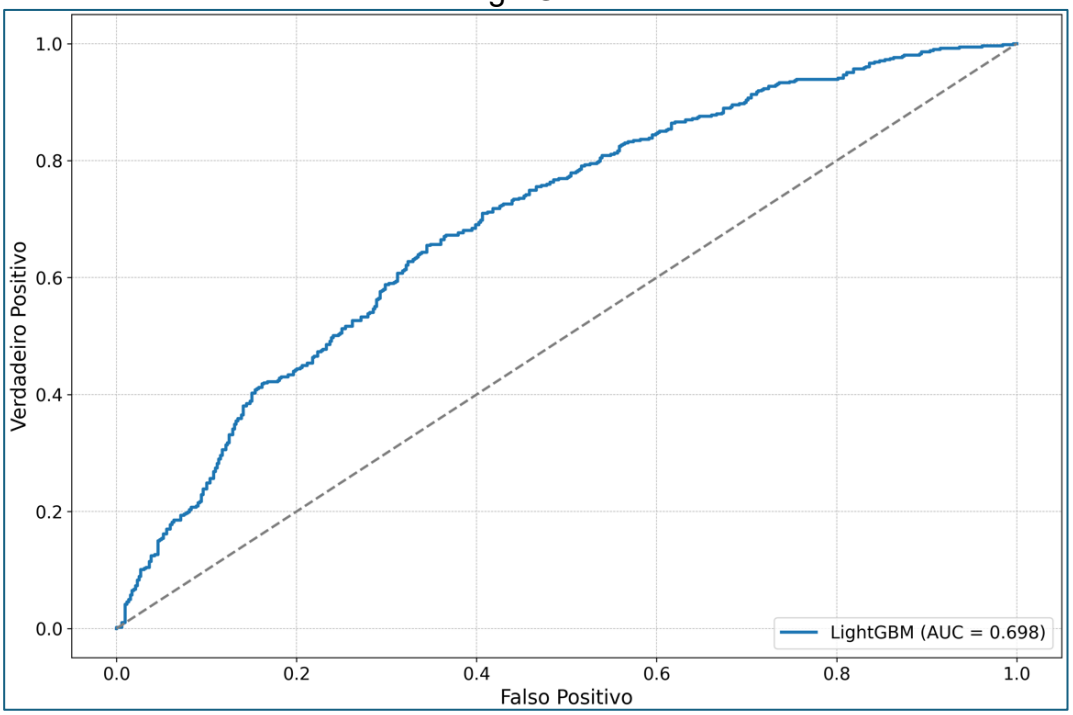
Fonte: Elaborado pelo autor (2026)

Gráfico 36 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo Random Forest



Fonte: Elaborado pelo autor (2026)

Gráfico 37 – Curva ROC-AUC da Tríplice Viral (SCR) com o modelo LightGBM



Fonte: Elaborado pelo autor (2026)

9.5. Estrutura de ajustes dos hiper parâmetros dos modelos (Hyper Parameter *Tuning*)

9.5.1. Parâmetros para Treinamento – BCG e Tríplice Viral (SCR)

Código 2 – Pipeline e Parâmetros – Support Vector Machine (SVM)

```
pipelines["Support Vector Machine (SVM)"] = ImbPipeline([
    ("smote", SMOTE(random_state=42)),
    ("scaler", StandardScaler()),
    ("pca", PCA(n_components=20, random_state=42)),
    ("clf", SVC(probability=True,
                class_weight="balanced",
                random_state=42))
])

param_grids["Support Vector Machine (SVM)"] = [
    # RBF (principal)
    {
        "clf__kernel": ["rbf"],
        "clf__C": [0.1, 1, 10],
        "clf__gamma": ["scale", 0.01, 0.001]
    },

    # Polynomial (secundário)
    {
        "clf__kernel": ["poly"],
        "clf__C": [0.1, 1, 10],
        "clf__gamma": ["scale", 0.01],
        "clf__degree": [2, 3]
    }
]
```

Fonte: Elaborado pelo autor (2026)

Código 3 – Pipeline e Parâmetros – Logistic Regression

```
pipelines["Logistic Regression"] = ImbPipeline([
    ("smote", SMOTE(random_state=42)),
    ("scaler", StandardScaler()),
    ("pca", PCA(n_components=20, random_state=42)),
    ("clf", LogisticRegression(max_iter=5000,
                               class_weight="balanced",
                               random_state=42,
                               n_jobs=-1))
])

param_grids["Logistic Regression"] = [
    # L2 (baseline principal)
    {
        "clf__penalty": ["l2"],
        "clf__C": [0.01, 0.1, 1, 10],
        "clf__solver": ["lbfgs"]
    },

    # L1 (seleção de variáveis)
    {
        "clf__penalty": ["l1"],
        "clf__C": [0.01, 0.1, 1, 10],
        "clf__solver": ["saga"]
    },

    # Elastic Net (melhor para features correlacionadas)
    {
        "clf__penalty": ["elasticnet"],
        "clf__C": [0.01, 0.1, 1, 10],
        "clf__l1_ratio": [0.2, 0.5, 0.8],
        "clf__solver": ["saga"]
    }
]
```

Fonte: Elaborado pelo autor (2026)

Código 4 – Pipeline e Parâmetros – Multilayer Perceptron (MLP)

```
pipelines["Multilayer Perceptron (MLP)"] = ImbPipeline([
    ("smote", SMOTE(random_state=42)),
    ("scaler", StandardScaler()),
    ("pca", PCA(n_components=20, random_state=42)),
    ("clf", MLPClassifier(activation="relu",
                          solver="adam",
                          early_stopping=True,
                          validation_fraction=0.15,
                          n_iter_no_change=20,
                          max_iter=1000,
                          random_state=42))
])

param_grids["Multilayer Perceptron (MLP)"] = {
    "clf__hidden_layer_sizes": [
        (64,),
        (64, 32),
        (128, 64)
    ],
    "clf__alpha": [0.001, 0.01],
    "clf__learning_rate_init": [0.001, 0.003],
    "clf__batch_size": [32]
}
```

Fonte: Elaborado pelo autor (2026)

Código 1 – Pipeline e Parâmetros – Random Forest

```
pipelines["Random Forest"] = ImbPipeline([
    ("smote", SMOTE(random_state=42)),
    ("clf", RandomForestClassifier(random_state=42,
                                   n_jobs=-1,
                                   class_weight="balanced_subsample",
                                   oob_score=True))
])

param_grids["Random Forest"] = {
    "clf__n_estimators": [200, 400],
    "clf__max_depth": [None, 8, 16],
    "clf__min_samples_split": [2, 10],
    "clf__min_samples_leaf": [1, 4],
    "clf__max_features": ["sqrt", 0.3, 0.5],
    "clf__bootstrap": [True]
}
```

Fonte: Elaborado pelo autor (2026)

Código 5 – Pipeline e Parâmetros – LightGBM

```
pipelines["LightGBM"] = ImbPipeline([
    ("smote", SMOTE(random_state=42)),
    ("preprocess", StandardScaler()),
    ("clf", LGBMClassifier(objective="binary",
                           metric="auc",
                           random_state=42,
                           device="gpu",
                           n_jobs=-1,
                           class_weight="balanced",
                           verbosity=-1))
])

param_grids["LightGBM"] = {
    "clf__n_estimators": [300, 500],
    "clf__learning_rate": [0.03, 0.05],
    "clf__num_leaves": [31, 63],
    "clf__max_depth": [4, 6],
    "clf__min_child_samples": [20, 40],
    "clf__subsample": [0.8],
    "clf__colsample_bytree": [0.6, 0.8],
    "clf__reg_alpha": [0.1],
    "clf__reg_lambda": [1, 3]
}
```

Fonte: Elaborado pelo autor (2026)

Código 6 – Pipeline e Parâmetros – XGBoost

```
pipelines["XGBoost"] = ImbPipeline([
    ("smote", SMOTE(random_state=42)),
    ("preprocess", StandardScaler()),
    ("clf", XGBClassifier(objective="binary:logistic",
                          eval_metric="auc",
                          tree_method="hist",
                          device="cuda",
                          random_state=42,
                          class_weight="balanced",
                          n_jobs=-1,
                          verbosity=0))
])

param_grids["XGBoost"] = {
    "clf__n_estimators": [300, 500],
    "clf__learning_rate": [0.03, 0.05],
    "clf__max_depth": [4, 6],
    "clf__min_child_weight": [3, 5],
    "clf__subsample": [0.8],
    "clf__colsample_bytree": [0.6, 0.8],
    "clf__gamma": [0, 0.2],
    "clf__reg_alpha": [0.1],
    "clf__reg_lambda": [1, 3]
}
```

Fonte: Elaborado pelo autor (2026)

9.5.2. Parâmetros Ajustados para a BCG

Código 1 – Parâmetros Ajustados – Random Forest

```
param_grids_best["Random Forest"] = {
    "clf__n_estimators": [200],
    "clf__max_depth": [None],
    "clf__min_samples_split": [2],
    "clf__min_samples_leaf": [4],
    "clf__max_features": ["sqrt"],
    "clf__bootstrap": [True]
}
```

Fonte: Elaborado pelo autor (2026)

Código 2 – Parâmetros Ajustados – LightGBM

```
param_grids_best["LightGBM"] = {
    "clf__n_estimators": [300],
    "clf__learning_rate": [0.05],
    "clf__num_leaves": [63],
    "clf__max_depth": [6],
    "clf__min_child_samples": [40],
    "clf__subsample": [0.8],
    "clf__colsample_bytree": [0.8],
    "clf__reg_alpha": [0.1],
    "clf__reg_lambda": [3]
}
```

Fonte: Elaborado pelo autor (2026)

Código 3 – Parâmetros Ajustados – XGBoost

```
param_grids_best["XGBoost"] = {
    "clf__n_estimators": [500],
    "clf__learning_rate": [0.05],
    "clf__max_depth": [6],
    "clf__min_child_weight": [5],
    "clf__subsample": [0.8],
    "clf__colsample_bytree": [0.8],
    "clf__gamma": [0],
    "clf__reg_alpha": [0.1],
    "clf__reg_lambda": [3]
}
```

Fonte: Elaborado pelo autor (2026)

Código 4 – Parâmetros Ajustados – Support Vector Machine (SVM)

```
param_grids_best["Support Vector Machine (SVM)"] = [
    {
        "clf__kernel": ["rbf"],
        "clf__C": [10],
        "clf__gamma": ["scale"]
    }
]
```

Fonte: Elaborado pelo autor (2026)

Código 5 – Parâmetros Ajustados – Logistic Regression

```
param_grids_best["Logistic Regression"] = [
    {
        "clf__penalty": ["l1"],
        "clf__C": [0.1],
        "clf__solver": ["saga"]
    }
]
```

Fonte: Elaborado pelo autor (2026)

Código 6 – Parâmetros Ajustados – Multilayer Perceptron (MLP)

```
param_grids_best["Multilayer Perceptron (MLP)"] = {
    "clf__hidden_layer_sizes": [
        (64, 32)
    ],
    "clf__alpha": [0.01],
    "clf__learning_rate_init": [0.003],
    "clf__batch_size": [32]
}
```

Fonte: Elaborado pelo autor (2026)

9.5.3. Parâmetros Ajustados para a Tríplice Viral (SCR)

Código 1 – Parâmetros Ajustados – Random Forest

```
param_grids_best["Random Forest"] = {
    "clf__n_estimators": [400],
    "clf__max_depth": [8],
    "clf__min_samples_split": [2],
    "clf__min_samples_leaf": [4],
    "clf__max_features": ["sqrt"],
    "clf__bootstrap": [True]
}
```

Fonte: Elaborado pelo autor (2026)

Código 2 – Parâmetros Ajustados – LightGBM

```
param_grids_best["LightGBM"] = {
    "clf__n_estimators": [300],
    "clf__learning_rate": [0.03],
    "clf__num_leaves": [31],
    "clf__max_depth": [4],
    "clf__min_child_samples": [40],
    "clf__subsample": [0.8],
    "clf__colsample_bytree": [0.6],
    "clf__reg_alpha": [0.1],
    "clf__reg_lambda": [3]
}
```

Fonte: Elaborado pelo autor (2026)

Código 3 – Parâmetros Ajustados – XGBoost

```
param_grids_best["XGBoost"] = {
    "clf__n_estimators": [300],
    "clf__learning_rate": [0.03],
    "clf__max_depth": [6],
    "clf__min_child_weight": [5],
    "clf__subsample": [0.8],
    "clf__colsample_bytree": [0.6],
    "clf__gamma": [0.2],
    "clf__reg_alpha": [0.1],
    "clf__reg_lambda": [3]
}
```

Fonte: Elaborado pelo autor (2026)

Código 4 – Parâmetros Ajustados – Support Vector Machine (SVM)

```
param_grids_best["Support Vector Machine (SVM)"] = [
    {
        "clf__kernel": ["poly"],
        "clf__C": [1],
        "clf__degree": [3],
        "clf__gamma": [0.01]
    }
]
```

Fonte: Elaborado pelo autor (2026)

Código 5 – Parâmetros Ajustados – Logistic Regression

```
param_grids_best["Logistic Regression"] = [
    {
        "clf__penalty": ["elasticnet"],
        "clf__C": [0.01],
        "clf__l1_ratio": [0.8],
        "clf__solver": ["saga"]
    }
]
```

Fonte: Elaborado pelo autor (2026)

Código 6 – Parâmetros Ajustados – Multilayer Perceptron (MLP)

```
param_grids_best["Multilayer Perceptron (MLP)"] = {
    "clf__hidden_layer_sizes": [
        (128, 64)
    ],
    "clf__alpha": [0.001],
    "clf__learning_rate_init": [0.003],
    "clf__batch_size": [32]
}
```

Fonte: Elaborado pelo autor (2026)