



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

PREDITORES DE PERDA DENTÁRIA NA POPULAÇÃO BRASILEIRA: UMA
ABORDAGEM DE APRENDIZADO DE MÁQUINA

Thiago Burgos Medeiros Rodrigues

Orientador

Carlos Eduardo Ribeiro de Mello

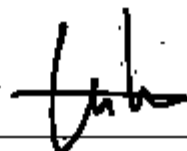
RIO DE JANEIRO, RJ – BRASIL
AGOSTO DE 2023

PREDITORES DE PERDA DENTÁRIA NA POPULAÇÃO BRASILEIRA: UMA
ABORDAGEM DE APRENDIZADO DE MÁQUINA

Thiago Burgos Medeiros Rodrigues

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO
DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM
INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
(UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:



Carlos Eduardo Ribeiro de Mello, D.Sc. – UNIRIO



Pedro Nuno de Souza Moura, D.Sc. – UNIRIO



Leandro Guimarães Marques Alvim, D.Sc. – UFRRJ

RIO DE JANEIRO, RJ – BRASIL
AGOSTO DE 2023

Catálogo informatizada pelo(a) autor(a)

R696	Rodrigues, Thiago Burgos Medeiros Preditores de perda dentária na população brasileira: uma abordagem de aprendizado de máquina / Thiago Burgos Medeiros Rodrigues. -- Rio de Janeiro, 2023. 86 Orientador: Carlos Eduardo Ribeiro de Mello. Dissertação (Mestrado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2023. 1. Aprendizado de máquina. 2. Inteligência artificial. 3. Modelo preditivo. 4. Perda dentária. 5. PNS. I. Mello, Carlos Eduardo Ribeiro de, orient. II. Título.
------	--

Dedico este trabalho à minha esposa
Paloma, que me deu todo o apoio para
este projeto.

AGRADECIMENTOS

A Deus, sempre presente, e que colocou no meu caminho pessoas especiais. Aquele que me concede forças para vencer os obstáculos da vida.

À Paloma e ao Nadal, que estiveram ao meu lado durante todo o caminho percorrido.

Aos meus pais e minha sogra pelo apoio incondicional.

Ao meu orientador Carlos Eduardo pela confiança, amizade e ensinamentos.

Aos professores do Mestrado em Informática, pelo interesse e esforços em prol da pesquisa e do crescimento acadêmico da instituição.

Aos meus amigos Guilherme, Vinícius e Aline, pela inspiração e incentivo.

Aos meus companheiros de trabalho Wagner, Daniel e Tarcísio que me ajudaram com precisas e incisivas pontuações.

Aos meus companheiros de estudo Cesar, Daniel e Rodrigo aqueles com quem criei verdadeira amizade e com os quais compartilhamos êxitos e dificuldades.

*Nossa grande tarefa não é ver o que está
obscuramente à distância, e sim fazer o
que está claramente a mão. – Thomas
Carlyle*

RODRIGUES, Thiago Burgos Medeiros. **Preditores de perda dentária na população brasileira: uma abordagem de aprendizado de máquina.** UNIRIO, 2023. 75 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Os problemas de saúde bucal afetam a qualidade de vida de forma geral, uma vez que podem gerar dores crônicas, afetar a mastigação e a fala, trazer repercussões estéticas e impactar, ainda, em outras partes do corpo, a exemplo de enfermidades cardíacas. A saúde bucal comprometida pode ter consequências importantes ao bem-estar físico, mental e social. Este tema tem sua relevância acentuada à medida que a população envelhece, pois a idade é um dos fatores de risco preponderantes no processo de evolução das doenças bucais até a perda dentária, o último estágio da patologia. Este trabalho visa apresentar o desenvolvimento e a avaliação de modelos de aprendizado de máquina, ainda pouco investigadas no tema de pesquisa acerca da predição da perda dentária, com o objetivo de prever a perda dentária. Consideraram-se dados sobre aspectos socioeconômicos, comportamentos da saúde oral, condições médicas crônicas e estilo de vida. Para o estudo, utilizou-se a base da Pesquisa Nacional de Saúde (PNS), realizada pelo IBGE. O PNS consiste de uma pesquisa conduzida por meio de inquérito, onde são coletados dados de um estrato da população contendo informações socioeconômicas, condições médicas (saúde bucal, inclusive) e estilo de vida. Pretende-se, portanto, propor e construir um modelo de aprendizado de máquina otimizado para predição de perda dentária a partir dos dados do PNS, considerando preditores individuais, não intrusivos (sem necessidade de exames) e de baixo custo operacional. Uma análise crítica dos resultados obtidos com o modelo de aprendizado de máquina permitiu identificar e analisar os principais preditores para perda dentária. Estes achados corroboram com a hipótese, apontada por outros trabalhos, que as características socioeconômicas têm grande importância no risco de desenvolvimento de doença dentária.

Palavras-chave: Aprendizado de máquina; Inteligência artificial; Modelo preditivo; Perda dentária; PNS.

ABSTRACT

Oral health problems affect the quality of life in general, since they can generate chronic pain, affect chewing and speech, bring aesthetic repercussions and also impact other parts of the body, such as heart disease. Compromised oral health can have important consequences for physical, mental and social well-being. This theme becomes increasingly relevant as the population ages, as age is one of the most important risk factors in the process of evolution from oral diseases to tooth loss, the last stage of the pathology. This work aims to present the development and evaluation of machine learning models, still little investigated in the research topic about the prediction of tooth loss, with the objective of predicting tooth loss. Data on socioeconomic aspects, oral health behaviors, chronic medical conditions and lifestyle were considered. For the study, the basis of the National Health Survey (PNS), carried out by IBGE, was used. The PNS is a survey-based research that collects data from a stratum of the population containing socioeconomic information, medical conditions (including oral health), and lifestyle details. It is intended, therefore, to propose and build an optimized machine learning model for predicting tooth loss from PNS data, considering individual, non-intrusive predictors (no need for tests) and with low operational cost. A critical analysis of the results obtained with the machine learning model allowed identifying and analyzing the main predictors of tooth loss. These findings corroborate the hypothesis, pointed out by other studies, that socioeconomic characteristics are of great importance in the risk of developing dental disease.

Keywords: *Machine Learning*; Artificial intelligence; Predictive model; Tooth loss; PNS.

SUMÁRIO

1	INTRODUÇÃO	15
1.1	Contexto e Motivação	15
1.2	Justificativa e relevância	15
1.3	Hipótese.....	16
1.4	Objetivos	16
1.4.1	Objetivo geral	16
1.4.2	Objetivos específicos.....	16
1.5	Metodologia	17
1.6	Contribuições	18
1.6.1	Contribuições científicas.....	18
1.6.2	Contribuições tecnológicas.....	18
1.7	Estrutura da dissertação	18
2	CARACTERIZAÇÃO DO PROBLEMA	19
2.1	Acesso à saúde bucal no Brasil	19
2.2	Principais doenças bucais e desfechos.....	20
2.3	Inquéritos de Saúde no Brasil	21
2.4	A Pesquisa Nacional de Saúde (PNS)	21
2.5	Estudo de referência para o desenvolvimento desta pesquisa	23
2.5.1	Sobre o artigo	23
2.6	Sobre o conjunto de dados National Health and Nutrition Examination Survey (NHANES).....	24
3	FUNDAMENTAÇÃO TEÓRICA.....	26
3.1	Inteligência Artificial e Aprendizado de Máquina	26
3.2	Algoritmos de Aprendizado de Máquina.....	28
3.2.1	Regressão logística	28
3.2.2	Floresta Aleatória (<i>Random Forest</i>).....	31
3.2.3	Extreme Gradient boosting Trees (XGBoost).....	33
3.2.4	Light <i>Gradient boosting</i> Machine (LightGBM)	35
3.2.5	Redes Neurais MultiLayer Perceptron (MLP).....	36

4	METODOLOGIA	38
4.1	Coleta de Dados	38
4.2	Seleção de Recursos	38
4.3	Pré-processamento.....	39
4.3.1	Reamostragem	41
4.3.2	Criação de variáveis derivadas (Feature Engineering).....	42
4.4	Modelagem	43
4.5	Interpretação das Métricas de Avaliação.....	44
5	RESULTADOS E DISCUSSÃO	47
5.1	Análise Exploratória de Dados	47
5.1.1	Características gerais da amostra	49
5.1.2	Análises	50
5.2	Análise dos Resultados	60
5.2.1	Importância das Variáveis (Feature Importance)	66
5.3	DISCUSSÃO	72
5.3.1	Discussão dos Resultados.....	72
5.3.2	Comparação com Estudo referência.....	74
6	CONSIDERAÇÕES FINAIS	79
6.1	Limitações e trabalhos futuros.....	79
	REFERÊNCIAS	80
	APÊNDICE A – TABELA DE PSEUDÔNIMOS X NOMENCLATURAS PNS 2013 E 2019.....	83
	APÊNDICE B – TABELA DE CONFIGURAÇÃO FINAL DE HIPERPARÂMETROS.....	85
	APÊNDICE C – RESULTADOS OBTIDOS A PARTIR DA AVALIAÇÃO DO CONJUNTO DE TESTE.....	86

LISTA DE FIGURAS

Figura 1 – Classificação dos algoritmos de Aprendizado de Máquina utilizados no estudo.	28
Figura 2 – Função logística	29
Figura 3 – Gráfico de função logística	29
Figura 4 – Diagrama ilustrativo do funcionamento do algoritmo Floresta Aleatória (Random Forest)	32
Figura 5 – Diagrama ilustrativo do funcionamento do algoritmo Extreme Gradient boosting Trees (XGBoost)	33
Figura 6 – Diagrama ilustrativo do funcionamento do algoritmo Light Gradient boosting Machine (LightGBM)	35
Figura 7 – Ilustração de uma RNA do tipo MLP	36
Figura 8 – Metodologia aplicada no desenvolvimento do experimento	38
Figura 9 – Código para dividir o conjunto de dados em treino e teste de forma estratificada	40
Figura 10 – Código para aplicar a padronização nas variáveis contínuas do conjunto de dados.	41
Figura 11 – Código de oversample SMOTE aplicado aos três tipos de desfechos. .	41
Figura 12 – Processo de formação da variável dependente: ‘Auto cuidado em Saúde Bucal’.	43
Figura 13 – Exemplo de Matriz de Confusão para Algoritmos de Classificação.	45
Figura 14 – Distribuição Percentual da Condição da Dentição	50
Figura 15 – Gráfico ilustrativo da distribuição completa entre faixas etárias e características dentárias.	51
Figura 16 – Gráfico ilustrativo da distribuição da perda dentária entre homens e mulheres.	51
Figura 17 – Gráfico ilustrativo da porcentagem de perda dentária por grupo étnico. .	52
Figura 18 - Gráfico ilustrativo da Porcentagem de perda dentária por estado civil. ..	53
Figura 19 – Gráfico ilustrativo da associação entre grau de instrução e perda dentária	54
Figura 20 – Gráfico ilustrativo da associação entre IMC perda dentária	55
Figura 21 – Gráfico ilustrativo da relação entre Doenças Crônicas e Incidência de Perda Dentária	56

Figura 22 – Gráfico ilustrativo da relação entre qualidade de auto cuidado em saúde bucal e incidência de perda dentária.	57
Figura 23 – Gráfico ilustrativo da relação entre perda dentária e faixa de renda per capita.	58
Figura 24 – Gráfico ilustrativo da relação entre perda dentária e condição de emprego.	59
Figura 25 – Gráfico ilustrativo da relação entre perda dentária e benefício de plano de saúde.	60
Figura 26 – Gráfico da Curva ROC de desempenho dos cinco algoritmos para o desfecho “Edentulismo”.	61
Figura 27 – Gráfico da Curva ROC de desempenho dos cinco algoritmos para o desfecho “Perda da Dentição Funcional”.	62
Figura 28 – Gráfico da Curva ROC de desempenho dos cinco algoritmos para o desfecho “Ausência de Dentes ou Perda de Um ou Mais Dentes”.	64
Figura 29 – Desempenho dos cinco algoritmos em cada desfecho.	66
Figura 30 – Os cinco principais preditores do desfecho "Edentulismo", com o modelo gerado a partir do algoritmo LGBM.	67
Figura 31 – Os cinco principais preditores do desfecho "Edentulismo", com o modelo gerado a partir do algoritmo Regressão Logística.	67
Figura 32 – Os cinco principais preditores do desfecho "Perda da Dentição Funcional", com o modelo gerado a partir do algoritmo LGBM.	68
Figura 33 – Os cinco principais preditores do desfecho "Perda da Dentição Funcional", com o modelo gerado a partir do algoritmo XGBoost.	68
Figura 34 – Os cinco principais preditores do desfecho "Ausência de dentes ou perda de um ou mais dentes", com o modelo gerado a partir do algoritmo XGBoost.	69
Figura 35 – Os cinco principais preditores do desfecho "Ausência de dentes ou perda de um ou mais dentes", com o modelo gerado a partir do algoritmo LGBM. ..	69
Figura 36 – Os sete principais preditores do desfecho "Perda da Dentição Funcional", com o modelo gerado a partir do algoritmo XGBoost.	71

LISTA DE TABELAS

Tabela 1 – Categorias definidas no PNS para variáveis categóricas selecionadas. .	48
Tabela 2 – Categorias codificadas para variáveis do PNS selecionadas.	49
Tabela 3 – Amostra de instâncias.	49
Tabela 4 – Distribuição completa entre faixas etárias e características dentárias. ...	50
Tabela 5 – Comparação da distribuição da perda dentária entre homens e mulheres.	51
Tabela 6 – Porcentagem de perda dentária por grupo étnico.	52
Tabela 7 – Porcentagem de perda dentária por estado civil.	52
Tabela 8 – Variação do Índice de perda dentária ao longo da idade.	53
Tabela 9 – Associação entre grau de instrução e perda dentária.	54
Tabela 10 - Associação entre IMC e incidência de perda dentária.	55
Tabela 11 – Relação entre Doenças Crônicas e Incidência de Perda Dentária.	56
Tabela 12 – Relação entre qualidade de auto cuidado em saúde bucal e incidência de perda dentária.	57
Tabela 13 – Relação entre perda dentária e faixa de renda per capita.	58
Tabela 14 – Relação entre perda dentária e condição de emprego.	59
Tabela 15 – Relação entre perda dentária e benefício de plano de saúde.	59
Tabela 16 – Resultados obtidos para o desfecho edentulismo.	61
Tabela 17 – Resultados obtidos para o desfecho perda da dentição funcional.	63
Tabela 18 – Resultados obtidos para o desfecho ausência de dentes ou perda de um ou mais dentes.	64
Tabela 19 – Proporção dos desfechos PNS e NHANES.	75
Tabela 20 – Comparativos de desempenho entre os resultados obtidos nesse estudo e os obtidos no estudo de referência para o desfecho edentulismo.	76
Tabela 21 – Comparativos de desempenho entre os resultados obtidos nesse estudo e os obtidos no estudo de referência para o desfecho perda da dentição funcional.	77
Tabela 22 – Comparativos de desempenho entre os resultados obtidos nesse estudo e os obtidos no estudo de referência para o desfecho ausência de dentes ou perda de um ou mais dentes.	78

LISTA DE NOMENCLATURAS

AAAI	Association for the Advancement of Artificial Intelligence
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
LightGBM	Light <i>Gradient boosting</i> Machine
MLP	Redes Neurais MultiLayer Perceptron
NHANES	National Health and Nutrition Examination Survey
PNS	Pesquisa Nacional de Saúde
SBBrazil	Sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas não Transmissíveis
SMOTE	Synthetic Minority Over-sampling Technique
SUS	Sistema Único de Saúde
XGBoost	<i>Extreme Gradient boosting Trees</i>

1 INTRODUÇÃO

1.1 Contexto e Motivação

Desde o nascimento, o ar entra nos pulmões pela boca e marca o primeiro contato com a vida. Entretanto, para sobreviver, é preciso mais do que respirar: nutrir o corpo é uma condição física necessária à sobrevivência humana. Neste sentido, o alimento que fornece energia para o crescimento também precisa ser obtido pela boca, por meio da amamentação e da alimentação.

No entanto, a boca não é apenas um órgão biológico. Ela é, ainda, reflexo da história de vida de cada indivíduo. As cáries e outras doenças bucais podem ser explicadas por diversos fatores, entre eles a genética, a dieta, o estilo de vida e a higiene bucal.

Por isso, faz-se necessário analisar a boca além das funções básicas de mastigação e fala, o que permite concluir sobre a saúde em geral e o bem-estar emocional do indivíduo. É através da observação atenta e cuidadosa desses fatores que se torna possível compreender a trajetória de vida de cada indivíduo.

Os problemas de saúde bucal têm influência na qualidade de vida, uma vez que podem gerar dores e afetar a mastigação e a fala. Além das repercussões estéticas, as patologias bucais estão associadas a enfermidades cardíacas (PAULETTO, Guilherme, et al.).

O indivíduo com a saúde bucal comprometida tem comprometido seu bem-estar físico, mental e social. À medida que a população envelhece, a idade se mostra um dos fatores preponderantes no processo de evolução das doenças bucais até a perda dentária, último estágio da patologia, sendo potencialmente prejudicial ao estado nutricional de uma pessoa idosa (MILAGRES, Clarice Santana, et al.).

1.2 Justificativa e relevância

A pesquisa parte do pressuposto de que “a medição de condições clínicas de saúde bucal e a dependência do diagnóstico profissional têm se mostrado cada vez mais limitadas para atender ao papel orientador das políticas públicas no campo da Odontologia, uma vez que excluem a relação das doenças bucais com as condições sociais, econômicas e culturais e como próprio olhar do indivíduo acerca de seu adoecimento” (PINTO, 2013).

Neste contexto, visa-se a fornecer outra perspectiva à observação dos dados da Pesquisa Nacional de Saúde (PNS) - realizada pelo Ministério da Saúde em parceria com o Instituto Brasileiro de Geografia e Estatística (IBGE) nos anos de 2013 e 2019. Averiguar-se-á as possibilidades de análise da população, por meio de algoritmos de aprendizado de máquina, a partir da apresentação de resultados mais embasados para a elaboração de políticas públicas. Por fim, nortear a busca a prevenção de problemas de saúde bucal.

1.3 Hipótese

Diante das limitações econômicas para as medições clínicas, dados como condição social, econômica e cultural podem fornecer uma alternativa adicional e eficiente para prever doenças bucais de um indivíduo (PINTO, 2013). Deste modo, seria possível projetar modelos de aprendizado de máquina capazes de prever perda dentária em indivíduos com grande eficácia.

1.4 Objetivos

1.4.1 Objetivo geral

Este trabalho visa a apresentar o desenvolvimento e a avaliação de modelos de aprendizado de máquina, metodologias ainda pouco investigadas no desafio de predição da perda dentária.

Para isso, utilizaram-se, como base da modelagem, características relacionadas aos aspectos socioeconômicos e condições médicas crônicas disponíveis nas pesquisas de 2013 e de 2019.

1.4.2 Objetivos específicos

Pretende-se experimentar e determinar um modelo de aprendizado de máquina otimizado, com vistas à predição de perda dentária.

A partir de dados subjetivos, não intrusivos (sem necessidade de exames) e de baixo custo operacional, aspira-se disponibilizar dados suficientes para reprodução dos modelos estudados, como ferramenta para subsidiar a elaboração de políticas públicas de alta impacto na prevenção de problemas de saúde bucal na população brasileira.

Por fim, reforça-se a importância das pesquisas realizadas pelo IBGE para a sociedade e comparar os resultados obtidos no presente estudo, que retrata a

realidade brasileira, com o estudo “Predictors of tooth loss: A *Machine Learning* approach” elaborado por Elani et al. (2021), realizado com dados de residentes nos E.U.A.

1.5 Metodologia

O presente projeto de pesquisa caracteriza-se como estudo transversal, quantitativo, com análise multidimensional e, tem como objetivo aplicar técnicas de *Data Mining* em Saúde (abordagem computacional capaz de extrair conhecimento de grandes conjuntos de dados mediante identificação de padrões) para realizar análises preditivas.

A pesquisa foi conduzida a partir de dados oriundos do inquérito populacional PNS (Pesquisa Nacional de Saúde), de âmbito nacional, realizada pelo Ministério da Saúde em parceria com o Instituto Brasileiro de Geografia e Estatística (IBGE). Foram utilizados os dados coletados no ano de 2019.

Para condução da pesquisa, foram utilizadas técnicas de classificação. Essa abordagem identifica a qual classe cada registro em um conjunto de dados pertence, com base em características e atributos específicos. Dessa forma, é possível prever com determinada precisão quais indivíduos têm maior probabilidade de desenvolver determinada condição.

Os modelos concebidos são potencialmente úteis no auxílio na administração da saúde pública. Fornecem, também, suporte à decisão clínica no planejamento e execução de medidas preventivas, por intermédio de dados não intrusivos, como características socioeconômicas e de saúde médica, a fim de reduzir a incidência de perda dentária, um desfecho comum de doenças bucais.

O experimento definido tem como principais etapas a análise exploratória de dados e a análise de desempenho entre cinco modelos de aprendizado de máquina, treinados a partir da base do PNS. Por fim, os resultados obtidos no artefato foram comparados com os reportados no artigo “Predictors of tooth loss: A *Machine Learning* approach”, elaborado por Elani et al. (2021), em que o mesmo experimento foi realizado com a base National Health and Nutrition Examination Survey (NHANES), que discorre acerca da realidade dos residentes nos E.U.A de forma a validar a hipótese da pesquisa.

1.6 Contribuições

1.6.1 Contribuições científicas

Investigar, a partir dos dados do PNS, com uma metodologia de aprendizado de máquina, preditores de risco para a evolução de doença dentária. Além disso, a análise comparativa entre resultados obtidos neste estudo e aqueles obtidos por Elani et al. (2021), realizado com dados de residentes nos E.U.A.

1.6.2 Contribuições tecnológicas

Buscou-se experimentar e projetar um modelo de aprendizado de máquina otimizado para predição de risco de perda dentária a partir de dados subjetivos, não intrusivos e de baixo custo operacional. Experimentou-se os seguintes métodos de aprendizado para classificação: *Extreme Gradient boosting Trees*, *Random Forests*, *Neural Networks*, *Light Gradient boosting Machine* e *Logistic Regression*.

1.7 Estrutura da dissertação

Esta pesquisa está organizada da seguinte maneira: no Capítulo 1 são abordados pontos-chave que guiam toda a dissertação, começando pela Motivação, seguida pelo Contexto, Justificativa, Hipótese, Objetivos, Metodologia, Principais contribuições e Estrutura. No Capítulo 2, é apresentada a caracterização do problema. No Capítulo 3, é realizada a fundamentação teórica das ferramentas que compõem tanto o experimento desenvolvido quanto o reproduzido. No Capítulo 4, é apresentada uma proposta de abordagem baseada em algoritmos de aprendizado de máquina para modelagem de classificadores robustos aplicados ao contexto de perda dentária e, a partir do experimento implementado no Capítulo 5, os resultados são discutidos e analisados no Capítulo 6. Por fim, no Capítulo 7 são apresentados os comentários, limitações do estudo e considerações finais.

2 CARACTERIZAÇÃO DO PROBLEMA

Este capítulo visa a contextualizar em que posição o objeto de pesquisa se insere na realidade brasileira, indicando uma lacuna ainda não preenchida no estudo de *Machine Learning* aplicado à Saúde. As informações aqui apresentadas fornecerão subsídios para a pesquisa realizada neste projeto.

Na Seção 2.1, aborda-se o acesso à saúde bucal no Brasil. Na Seção 2.2, apresentam-se as principais doenças bucais, com ênfase na cárie dentária e na doença periodontal, as mais comuns relacionadas à perda dentária, e seus principais desfechos. A Seção 2.3 aborda os inquéritos que fornecem informações sobre saúde bucal no Brasil, em especial a Pesquisa Nacional de Saúde (PNS), fonte de dados utilizada para o desenvolvimento deste estudo. Na seção 2.4, é apresentado pequeno resumo sobre o artigo “Predictors of tooth loss: A *Machine Learning* approach”, elaborado por Elani et al. (2021), que foi utilizado como estudo de referência para o desenvolvimento desta pesquisa.

2.1 Acesso à saúde bucal no Brasil

“Um país se conhece pelos dentes”
Arnaldo Luis Miranda

Em termos gerais, a Constituição Federal de 1988 estabeleceu a saúde como direito de todos e dever do Estado (BRASIL, 2023a, art. 196). Para tornar realidade este direito, foi criado o Sistema Único de Saúde (SUS), com o objetivo de oferecer atendimento gratuito e universal, sem discriminação. O SUS é financiado com recursos públicos e sua gestão é compartilhada pelos governos federal, estadual e municipal.

No que tange à saúde bucal, esta corresponde a um conjunto de condições, objetivas (biológicas) e subjetivas (psicológicas), que possibilita ao ser humano exercer funções como mastigação, deglutição e fonação e, também, pela dimensão estética inerente à região anatômica, exercitar a autoestima e relacionar-se socialmente sem inibição ou constrangimento. Essas condições devem corresponder à ausência de doença ativa em níveis tais que permitam ao indivíduo exercer as mencionadas funções do modo que lhe pareça adequado e lhe permita sentir-se bem, contribuindo desta forma para sua saúde geral.

Por outro lado, apesar da garantia constitucional, “as condições dentárias se mostram um dos mais significativos sinais de exclusão social.” (MINISTÉRIO DA

SAÚDE, p. 7). Isso porque dentes e gengivas registram o impacto das precárias condições de vida de milhões de pessoas em todo o país. A escolaridade deficiente, a renda baixa, a falta de trabalho, enfim, a má qualidade de vida produz efeitos devastadores sobre gengivas, dentes e outras estruturas da boca.

A OMS destaca que as doenças bucais implicam restrições de atividades na escola, no trabalho e na vida doméstica.

2.2 Principais doenças bucais e desfechos

Segundo Lindhe, Karring e Lang (1999), as doenças bucais de maior prevalência na população mundial são as cáries e a doença periodontal. A cárie é a segunda enfermidade mais comum do mundo, sendo a primeira o resfriado. Trata-se de uma doença multifatorial, na qual a estrutura do dente passa por um processo de desmineralização: diante da ingestão de carboidratos fermentáveis, bactérias que habitam naturalmente a boca se multiplicam rapidamente e formam a placa bacteriana, situação que é facilitada por ser a boca um ambiente úmido (FEIJÓ; IWASAKI, 2014).

A bactéria sobrevive da degradação desta matéria orgânica contida na boca e, como consequência, há formação de ácidos que são responsáveis pelo enfraquecimento da estrutura superficial do dente (esmalte). Com o passar do tempo, diante da ausência de cuidados, a estrutura do dente se mostra mais fraca e a parte mais interna (dentina) é atingida, momento em que o indivíduo passa a sentir dor (ibid.). Assim, caso persista o processo de corrosão, o esmalte quebra e inicialmente é formada uma cavidade; posteriormente, ocorre a fratura e, por fim, a destruição do dente (BARATIERI et al., 2002).

O aparecimento da cárie, portanto, depende da existência da bactéria *streptococcus mutans* e a acessibilidade deste microrganismo ao açúcar, diante de uma má higienização bucal (REIS; MELO, 2003).

Em seu turno, a doença periodontal é uma doença crônica e progressiva que leva à perda do dente, cuja principal causa é também a falta ou deficiência na higiene dental. Segundo Buffon (2005), inicia-se com a gengivite, que é a inflamação da gengiva, a qual ocorre por conta de um agente agressor, que é a placa bacteriana. A gengivite causa sangramento, vermelhidão e inchaço na gengiva e, caso não tratada, ocorre a migração para os tecidos de sustentação do dente. Inicia-se a periodontite, que consiste na inflamação no ligamento periodontal e, ainda, uma reabsorção óssea:

há a retração gengival e ocorre a exposição da raiz (ibid.). De acordo com Narvai e Frazão (2008), trata-se de uma doença crônica e que, em muitos casos, não causa forte dor; com o passar do tempo, o tecido ósseo continua o processo de reabsorção, até o dente ter mobilidade e, ao fim, soltar-se da raiz.

O estágio final das doenças dentárias, conhecido como perda dentária, pode ter um impacto significativo na saúde geral do indivíduo, afetando negativamente sua qualidade de vida e bem-estar. Estudos como o de Thomson (2012) enfatizam que a perda dentária é uma consequência grave das doenças dentárias. Além disso, pesquisas recentes, como as de Schierz et al. (2021) e Romandini et al. (2021), destacam como tal situação pode comprometer a saúde geral do indivíduo, ressaltando a importância de prevenir e tratar as doenças dentárias precocemente, a fim de preservar a função oral e promover uma melhor qualidade de vida.

Para o presente estudo, tomou-se por definição que, por dentição completa, entende-se ter ao menos vinte e oito (28) dentes, ou seja, não houve perda dentária, excetuados os sisos. Para dentes ausentes, indica-se que o indivíduo possui vinte e sete (27) dentes ou menos, pois perdeu um (1) ou mais dentes, porém mantém a sua dentição funcional. A perda da dentição funcional define pessoas que possuem menos de vinte e um (21) dentes (Figura 14). O edentulismo, estágio final de doenças dentárias, descreve a total ausência de dentes.

2.3 Inquéritos de Saúde no Brasil

Os Inquéritos de Saúde são pesquisas realizadas para coletar informações sobre a saúde e o bem-estar da população brasileira. Essas pesquisas têm como objetivo fornecer dados para o planejamento, monitoramento e avaliação de políticas públicas de saúde, bem como para a produção de indicadores epidemiológicos e demográficos.

Os inquéritos domiciliares representam uma das principais metodologias empregadas em estudos de base populacional. No Brasil, destaca-se o papel histórico desempenhado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) na condução das pesquisas mais importantes na produção de dados sociais.

2.4 A Pesquisa Nacional de Saúde (PNS)

A Pesquisa Nacional de Saúde (PNS) é realizada pelo IBGE em parceria com o Ministério da Saúde, em periodicidade quinquenal, com diversos níveis de

desagregação geográfica e socioeconômica, o que permite maior precisão nas estimativas. Os agentes de coleta do IBGE visitam domicílios no país para investigar informações relacionadas à saúde e ao bem-estar da população brasileira.

O questionário é dividido em três blocos. O primeiro, de entrevista domiciliar, envolve questões generalistas sobre o domicílio, como posse de bens, existência de animais, coleta de lixo e esgotamento sanitário. O segundo consiste em entrevista com um informante do domicílio acerca de perguntas que envolvem todos os moradores, nas quais se questiona sobre características socioeconômicas, cobertura de plano de saúde e pessoas com deficiência.

Por fim, o último bloco consiste em entrevista individual, na qual há questões específicas sobre comorbidade e estilos de vida de cada indivíduo maior de 15 anos, selecionado aleatoriamente. Neste momento, são arguidas perguntas sobre percepção do estado de saúde, estilo de vida, doenças crônicas, atividade sexual e doenças transmissíveis, condições de trabalho, violências sofridas e saúde bucal.

Assim, são encontradas na pesquisa questões relativas a doenças crônicas, deficiências, realização de exames preventivos e atividades físicas, bem como acidentes e agressões sofridas, de forma a possibilitar construir um retrato da saúde no país.

O objetivo geral da PNS é acompanhar e avaliar o estado de saúde da população brasileira, através do acesso aos serviços básicos de saúde, seja ele público ou privado, além de acompanhar a evolução de doenças crônicas. As respostas ajudam a consolidar informações sobre os determinantes, condicionantes e necessidades de saúde, além do uso e a avaliação dos serviços de saúde junto à população; servem, portanto, de base para o planejamento de políticas públicas na área de saúde.

Em 2013, foi lançada a primeira PNS, o maior inquérito domiciliar em saúde já realizado no Brasil, com enfoque em doenças crônicas não transmissíveis, estilos de vida e acesso ao atendimento médico. A mesma pesquisa foi novamente realizada em 2019, com a finalidade de ampliar a produção de informações em saúde, as quais foram inferidas a partir das respostas obtidas em entrevista aplicada a uma amostra probabilística da população analisada; atingiu ao redor de 100 mil domicílios em todo o território nacional.

2.5 Estudo de referência para o desenvolvimento desta pesquisa

O artigo “Predictors of tooth loss: A *Machine Learning* approach”, de autoria de Elani et al. (2021) foi utilizado para balizar este estudo, servindo de fonte primária para direcionar a metodologia aplicada na mineração dos dados.

2.5.1 Sobre o artigo

Publicado em junho de 2021 no periódico PLOS ONE, o objetivo do artigo foi desenvolver e testar algoritmos de *Machine Learning* para prever perda dentária completa ou incremental, a partir de duas análises: dados socioeconômicos e de saúde auto relatados e dados de rotina clínica, obtidos por exame.

A justificativa do seu desenvolvimento baseia-se na ausência de estudos que apontem com clareza o papel das condições socioeconômicas e outras características de saúde na perda dentária e no benefício de se diagnosticar e tratar no estágio inicial as doenças dentárias. Deste modo, o estudo poderia ajudar médicos na implementação da prevenção precoce e servir de subsídio para políticas que visem a garantir acesso ao atendimento odontológico e melhorar a saúde bucal de populações vulneráveis.

O estudo dividiu-se em duas análises: primária, onde foram utilizadas características socioeconômicas, rotina de cuidado dental e condições médicas crônicas; e secundária, com utilização apenas de variáveis clínicas, além da idade, sexo e raça. A presente dissertação manteve seu foco na análise primária, uma vez que a Pesquisa Nacional de Saúde (PNS), *conjunto de dados* brasileiro, não dispõe de variáveis clínicas com mesma natureza semântica.

Os resultados alcançados no artigo foram obtidos a partir dos dados do ciclo de 2011-2012 (treino) e 2013-2014 (teste) do *National Health and Nutrition Examination Survey* (NHANES) - Pesquisa Nacional de Saúde e Nutrição dos E.U.A.

Adotou-se o recorte de adultos com 18 anos ou mais e consideraram-se três possíveis desfechos de interesse: edentulismo (perda completa de todos os dentes), perda da dentição funcional (ter menos de 21 dentes) e ausência de um ou mais dentes. A metodologia proposta tratou estas classes de forma dicotômica, de modo a permitir apenas dois possíveis valores 1 ou 0. Ainda no pré-processamento, foram aplicadas as técnicas de *One-Hot Encoding* nas variáveis categóricas e a padronização nas variáveis contínuas.

Devido à baixa prevalência de valores positivos para a classe edentulismo, foi aplicada a técnica de reamostragem com seleção unilateral quando analisado esse desfecho. Os ajustes dos hiperparâmetros dos algoritmos foram realizados com o método de otimização Bayesiana nos conjuntos de dados de treino para cada desfecho.

Ao final da análise, os valores de *Shapley* foram investigados para cada modelo preditivo, visando avaliar a contribuição de cada atributo e valor para a previsão dos desfechos. Toda a implementação foi realizada com Python (biblioteca *scikit-learn*) e STATA 15.1 (para análise). Cinco algoritmos de aprendizado foram investigados: Regressão Logística, Floresta Aleatória(*Random Forest*), *Light Gradient boosting Machine*, *Extreme Gradient boosting Trees* e Redes Neurais *MultiLayer Perceptron* (MLP).

Os seguintes resultados foram reportados na análise primária:

- Todos os modelos apresentaram desempenho semelhantes com alta AUC (>81,5%);
- *Extreme Gradient boosting Trees* obteve o melhor desempenho;
- Variáveis Predictoras mais significativas: Idade, educação, atendimento odontológico de rotina, emprego, relação entre renda familiar e nível de pobreza, raça e casa própria foram fortes preditores de perda dentária;
- Variáveis Predictoras menos significativas: condições médicas como artrite, diabetes, colesterol alto, hipertensão e doenças cardiovasculares;
- Principais preditores para a maioria dos modelos: idade, educação e raça.

2.6 Sobre o conjunto de dados National Health and Nutrition Examination Survey (NHANES)

Para maior esclarecimento e possibilitar a comparação, descrevemos brevemente o conjunto de dados NHANES, utilizado na pesquisa referência desta dissertação, de modo a compreender importantes nuances e comparativos entre os dois trabalhos.

O *National Health and Nutrition Examination Survey* (Pesquisa Nacional de Saúde e Nutrição / NHANES) é um programa de pesquisa conduzido nos Estados Unidos pelo *National Center for Health Statistics*. Seu objetivo é avaliar a saúde e o estado nutricional dos residentes naquele país.

Desde 1960, a pesquisa coleta informações com base em entrevistas pessoais e exames físicos, obtidas em todos os condados de forma aleatória, com o objetivo de verificar tendências e problemas de saúde emergentes, rastrear estimativas de prevalência e traçar um retrato do estado nutricional das pessoas.

A partir de 1999, o NHANES passou a coletar anualmente tais dados, examinando 5.000 indivíduos a cada período - cada pessoa pode representar até 65.000 outras no país. Os entrevistadores visitam os domicílios amostrados para inquirir informações acerca da saúde e dos hábitos alimentares de cada pessoa selecionada no domicílio.

Posteriormente, estes indivíduos são encaminhados para exames físicos com equipes de profissionais de saúde, que ocorrem dentro de centros móveis. Cada equipe de examinadores inclui um médico, um dentista, um entrevistador especializado em nutrição, técnicos de laboratório e profissionais de coleta de dados. Os indivíduos que se submetem ao exame completo têm sua altura e peso medidos, o que permite estimar a porcentagem de obesidade da população. Outros participantes passam, ainda, por coleta de sangue para medição de colesterol e glicose. Ao fim de cada exame, é fornecido prêmio em dinheiro em agradecimento pelo tempo e esforço despendidos.

Um condado é selecionado aleatoriamente para representar outros semelhantes dentro do território norte-americano, por meio de informações obtidas junto ao US Census Bureau. Todos os condados são divididos em 15 grupos com base em suas características; em seguida, um condado é selecionado de cada grupo e, juntos, são formados os 15 condados na pesquisa NHANES para o ano.

Os dados do NHANES desempenham importante papel no monitoramento da saúde nos Estados Unidos e são essenciais para decisões em saúde pública: agências governamentais e organizações do setor privado os utilizam para estabelecer políticas e planejar programas de pesquisa, educação e prevenção de doenças.

3 FUNDAMENTAÇÃO TEÓRICA

Os conceitos apresentados neste capítulo serão utilizados ao longo de todo o processo de desenvolvimento desta pesquisa e facilitarão a compreensão da lógica aplicada. Na Seção 3.1 serão introduzidos os conceitos de Inteligência Artificial, com ênfase nas particularidades de Aprendizado de Máquina quando aplicada à área de Saúde. Por fim, na Seção 3.2, são apresentados conceitos gerais acerca de Algoritmos de Aprendizado de Máquina, bem como especificidades dos aplicados neste trabalho.

3.1 Inteligência Artificial e Aprendizado de Máquina

Nos últimos anos, tem-se observado um considerável aumento na quantidade de dados disponíveis, oriundas das mais diversas fontes: dispositivos interconectados, entidade física - como telefones celulares, sensores ou dispositivos GPS, ou ainda virtual como redes sociais. No mesmo caminho, segundo Batista e Chiavegatto Filho (2019), a área de saúde tem gerado um grande volume de dados úteis sobre diagnóstico e tratamento de doenças, informações demográficas de pacientes, planos de tratamento, pagamento e cobertura de seguro, atraindo a atenção de médicos e cientistas, que abordam diferentes dimensões da aplicação de mineração de dados.

Segundo Russel e Norvig (2020), a Inteligência Artificial (IA) é um campo da computação que estuda a construção de entidades inteligentes, ou seja, máquinas que parecem ter inteligência humana. Os mesmos autores descrevem quatro categorias de definições de inteligência artificial: sistemas que agem como seres humanos, sistemas que pensam como seres humanos, sistemas que pensam racionalmente e sistemas que agem racionalmente (ibid.).

A IA pode ser dividida em subáreas de acordo com suas aplicações. A *Association for the Advancement of Artificial Intelligence* (AAAI) na sua última chamada de trabalhos dividiu as aplicações em nove subáreas:

- Pesquisa;
- *Machine Learning*, Data Mining e *Big Data*;
- Planejamento Automatizado;
- Representação de Conhecimento;
- Raciocínio (Probabilístico ou não);

- Processamento de Linguagem Natural;
- Robótica;
- Sistema de Agente;
- Multi-Agente e Aplicações.

A Inteligência Artificial (IA) tornou-se uma ferramenta importante no campo de medicina. Algoritmos de aprendizado de máquina hoje são capazes de identificar padrões, termos e tendências de dados que não são tão evidentes ao olho humano. Isso permite que os profissionais médicos façam testes mais precisos e previsões sobre o diagnóstico e prognóstico do paciente, bem como tomar decisões sobre seu tratamento (SILVA et al., 2023).

Aprendizado de máquina é a ciência (e a arte) da programação de computadores de modo que eles possam aprender com os dados, isto é, concentra-se no desenvolvimento de algoritmos e modelos que, permitem aos computadores adquirirem a habilidade de aprender com experiências anteriores. (GÉRON, 2019). No campo da saúde, essa tecnologia promissora possui um potencial enorme para transformar os resultados dos pacientes, elevando-os a um novo patamar de excelência, enquanto aprimora a eficiência e a eficácia do sistema de saúde como um todo (SILVA et al., 2023). Ao capacitar os computadores a analisarem, compreenderem e interpretar informações médicas complexas, o aprendizado de máquina permite a detecção precoce de doenças, diagnósticos mais precisos, tratamentos personalizados e até mesmo a prevenção de problemas de saúde.

Conforme Batista e Chiavegatto Filho (2019, p. 2), *Machine Learning* (ML) é uma das áreas da Inteligência Artificial que vem apresentando maior destaque na área da saúde e em outras áreas do conhecimento". Esta baseia-se na teoria de que os computadores podem aprender com os dados sem a necessidade de programação explícita (manual) para executar determinadas atividades. Desta forma, o aprendizado ocorre a partir de um conjunto de dados que expressa toda a experiência que se deseja passar.

Neste projeto de pesquisa serão utilizados cinco algoritmos de *Machine Learning* (Figura 1):

- Regressão Logística;
- Floresta Aleatória (*Random Forest*);

- *Extreme Gradient boosting Trees* (XGBoost);
- *Light Gradient boosting Machine* (LightGBM);
- *Redes Neurais MultiLayer Perceptron* (MLP).

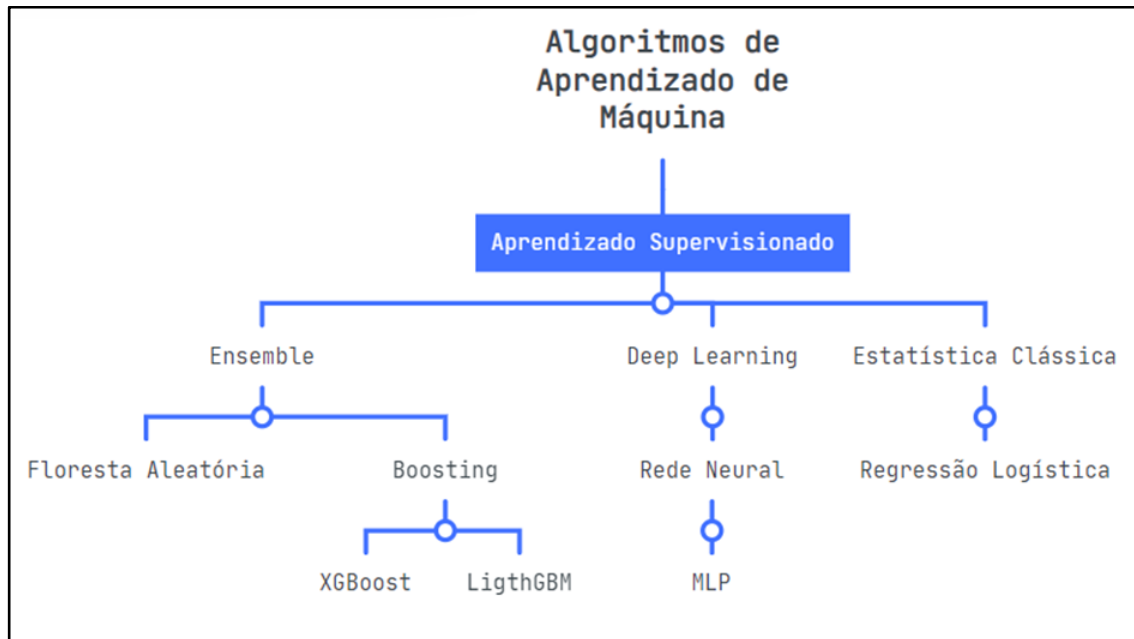


Figura 1 – Classificação dos algoritmos de Aprendizado de Máquina utilizados no estudo.

Fonte: Desenvolvido pelo Autor (2023)

A seguir, será tecida uma breve introdução sobre estes algoritmos. Serão pontuados de forma sucinta a definição de cada algoritmo, suas vantagens e desvantagens, seu funcionamento e seus hiperparâmetros mais importantes.

3.2 Algoritmos de Aprendizado de Máquina

3.2.1 Regressão logística

A regressão logística é um algoritmo de aprendizado de máquina supervisionado – classe de algoritmo que emprega dados rotulados para treinar modelos – utilizado para classificação (Figura 2). Este se comporta como um modelo linear, ou seja, realiza a predição através da soma ponderada dos atributos (*features*), além de uma constante chamada viés (*bias*), também chamado de constante (*intercept*). No entanto, ao invés de retornar um valor contínuo (números reais), retorna a função logística desse valor (Figura 3).

$$\text{Logística}(x) = \frac{1}{1 + e^{(-x)}}$$

Figura 2 – Função logística

Fonte: STOROPOLI; SOUZA (2020). Recuperado de <https://github.com/storopoli/ciencia-de-dados>

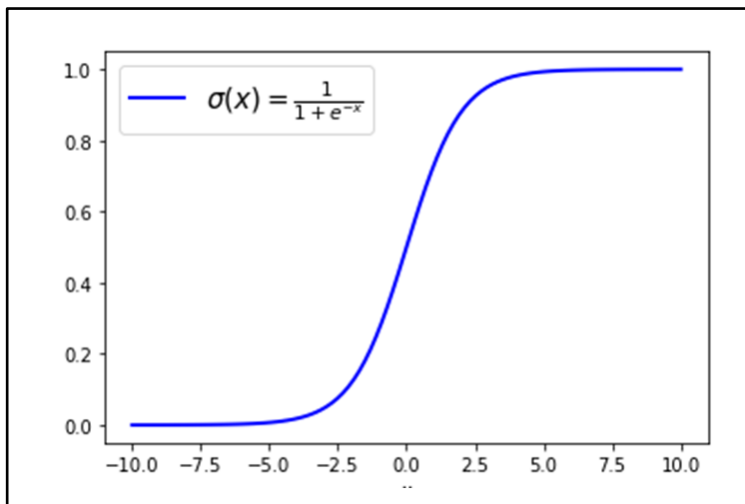


Figura 3 – Gráfico de função logística

Fonte: STOROPOLI; SOUZA (2020). Recuperado de <https://github.com/storopoli/ciencia-de-dados>

Alguns **exemplos de aplicação de regressão logística** podem ser:

- Segmentação e categorização de imagens;
- Reconhecimento de escrita/caligrafia;
- Diagnóstico médico;
- Previsões climáticas.

O algoritmo também apresenta algumas vantagens e desvantagens, a saber:

Vantagens:

- O modelo de regressão logística não atua somente como um classificador, ele também fornece um valor de probabilidade para cada instância classificada. Esta é uma característica poderosa para inferir na certeza sobre a classificação, uma vez que um registro classificado com $p = 0,99$ é muito mais provável de estar correto do que um registro classificado com $p = 0,51$;
- É um modelo relativamente fácil de implementar, interpretar, é rápido e possui um bom desempenho (principalmente quando os dados são linearmente separáveis);

- Os coeficientes da regressão logística dão uma ideia do quão relevante é um atributo para a predição (peso) e também a direção de associação com a variável alvo (podendo ser positivo ou negativo).

Desvantagens

- Embora seja menos suscetível a *overfitting* (sobreajuste), i.e., quando o modelo se ajusta em demasia aos dados de treino e se torna ineficaz para prever/classificar novos resultados) que outros algoritmos de classificação, esse problema ainda pode ocorrer quando o modelo é submetido a um conjunto de dados de alta dimensionalidade. Nesses cenários, técnicas de regularização podem ser utilizadas para contornar o problema;
- Requer uma atenção maior com o pré-processamento dos dados quando comparado a outros modelos de classificação;
- Podem aparecer problemas nos casos em que uma variável tem uma influência muito maior para a classificação do que as demais. Nesses cenários pode haver a não convergência do coeficiente/peso associado a essa variável. Esse problema pode ser solucionado introduzindo uma “penalização” aos pesos.

Alguns dos **hiperparâmetros importantes da regressão logística** são:

- `max_iter` - int - Número máximo de iterações do *Gradient Descent*
 - padrão 100
 - O Gradiente Descendente é um algoritmo padrão para otimizar funções complexas iterativamente dentro de um programa de computador. Seu objetivo é: dada alguma função arbitrária, encontrar um mínimo local.
 - Em se tratando de uma simples regressão linear, o método de gradiente descendente só é recomendado quando temos dados com muitas dimensões.
- `random_state` - int - *seed* do gerador de número randômicos (replicabilidade)
- `n_jobs` - int - Número de núcleos de processamento que serão utilizados em paralelo
 - None - padrão
 - -1 - usa todos os núcleos
- `solver` - {'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'} - Algoritmo a ser utilizado no problema de otimização.
 - padrão lbfgs

- Para escolher um solucionador, considerar os seguintes aspectos:
- Para conjuntos de dados pequenos, 'liblinear' é uma boa escolha, enquanto 'sag' e 'saga' são mais rápidos para grandes;
- Para problemas multiclasse, apenas 'newton-cg', 'sag', 'saga' e 'lbfgs' lidam com perda multinomial;
- 'liblinear' é limitado a esquemas um-contra-resto.
- A escolha do algoritmo depende da penalidade escolhida. Para maiores informações consultar o sklearn.
- penalty - {'l1', 'l2', 'elasticnet', 'none'} - Especifica a norma da pena:
 - 'none': nenhuma penalidade é adicionada;
 - 'l2': adiciona um termo de penalidade L2 e é a escolha padrão;
 - 'l1': adicionar um termo de penalidade L1;
 - 'elasticnet': ambos os termos de penalidade L1 e L2 são adicionados.
- C - float - Inverso da força de regularização; deve ser um float positivo. Valores menores especificam uma regularização mais forte.
 - padrão 1.0

Em suma, construído a partir de uma transformação/função (denominada função logística ou sigmoide) sobre a Regressão Linear, a Regressão Logística é utilizada para classificação. Este algoritmo de aprendizado de máquina supervisionado pode ser aplicado para prever eventos binomiais e multinomiais, com potencial de realizar até classificação de imagens.

3.2.2 Floresta Aleatória (*Random Forest*)

Os algoritmos a seguir: Floresta Aleatória, *Extreme Gradient boosting Machine* (XGBoost) e *Light Gradient boosting Trees* (LightGBM), têm a sua origem no de Árvore de Decisão e são classificados como *ensemble* (comitê). Isso significa que, para obter previsões, eles combinam o resultado de múltiplos modelos para obter um melhor desempenho.

O algoritmo Floresta Aleatória pode ser utilizado tanto para tarefas de classificação quanto para regressão. Ele estende o de árvore de decisão ao combinar diversas árvores individuais (floresta) para gerar um modelo mais robusto e preciso.

Por ser um algoritmo muito popular, sua aplicação inclui diversas áreas como finanças, bioinformática, marketing e publicidade e medicina.

A ideia central do Floresta Aleatória é criar uma "floresta" de árvores de decisão, onde cada árvore é treinada em uma amostra aleatória dos dados sem uma estar vinculada à outra. Deste modo, cada árvore faz sua própria previsão e, no caso de tarefas de classificação - foco deste estudo, o resultado final é determinada pela previsões da maioria das árvores individuais.

Segundo Storopoli e Souza (2020), o funcionamento do *Ensemble* (Comitê) pode ser descrito assim:

Comitê (*Ensemble*)

- Cada Árvore de Decisão Individual faz um “voto” pela previsão e é utilizada a previsão que tem a maioria dos “votos”;
- Essa classe de algoritmos são chamadas de comitê (*ensemble*);
- Essa lógica se dá pela sabedoria da multidão (*wisdom of the crowd*):
 - Um grande número de modelos (árvores) relativamente não correlacionados que operam como um comitê superará qualquer um dos modelos constituintes individuais (Figura 4).

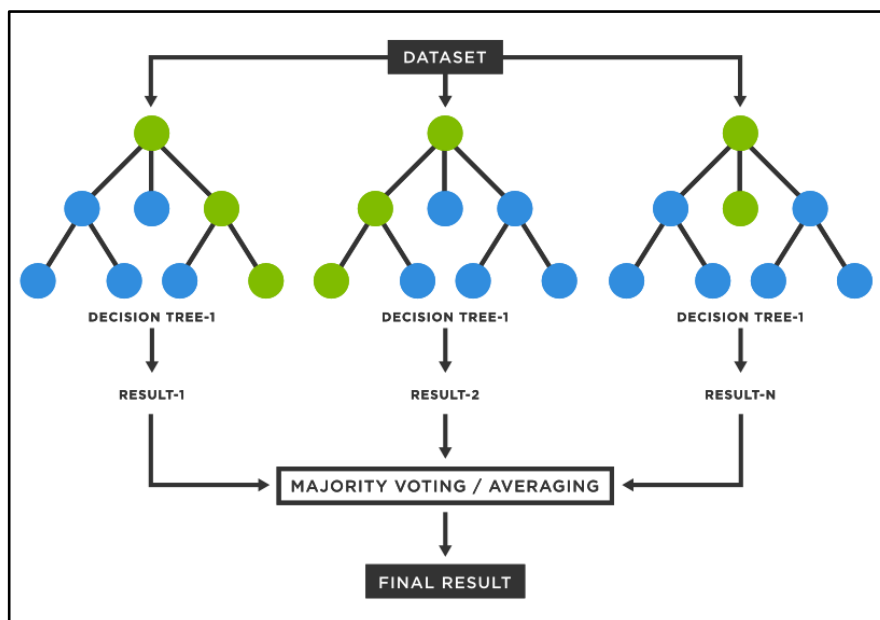


Figura 4 – Diagrama ilustrativo do funcionamento do algoritmo Floresta Aleatória (Random Forest)

Fonte: Recuperado de <https://www.tibco.com/pt-br/reference-center/what-is-a-random-forest>

Alguns **hiperparâmetros importantes das Florestas Aleatórias** incluem:

- Número de Árvores da Floresta;

- Todos das Árvores de Decisão:
 - Profundidade Máxima das Árvores: Altura máxima de nós que devem existir em uma dada árvore;
 - Quantidade Mínima de Observações para uma Decisão: Mínimo de observações que um nó deve ter para ser separado.

3.2.3 Extreme Gradient boosting Trees (XGBoost)

Extreme Gradient boosting (XGBoost) é um algoritmo de aprendizado supervisionado que se baseia na técnica de boosting para criar modelos preditivos (CHEN; GUESTRIN, 2016). Foi desenvolvido por Tianqi Chen e Carlos Guestrin na Universidade de Washington e se tornou uma escolha popular em competições de ciência de dados devido ao seu desempenho e eficácia. De acordo com Brownlee (2020, p. 364), “trata-se de um algoritmo *ensemble* de árvores de decisão, onde novas árvores corrigem erros daquelas árvores que já fazem parte do modelo. Árvores são adicionadas até que nenhuma outra melhoria possa ser feita no modelo”.

Além disso, o XGBoost pertence à família de algoritmos de *gradient boosting*, que combinam vários modelos de aprendizado fraco (geralmente árvores de decisão) para formar um modelo preditivo mais forte (Figura 5). Ele é projetado para lidar com problemas de regressão e classificação.

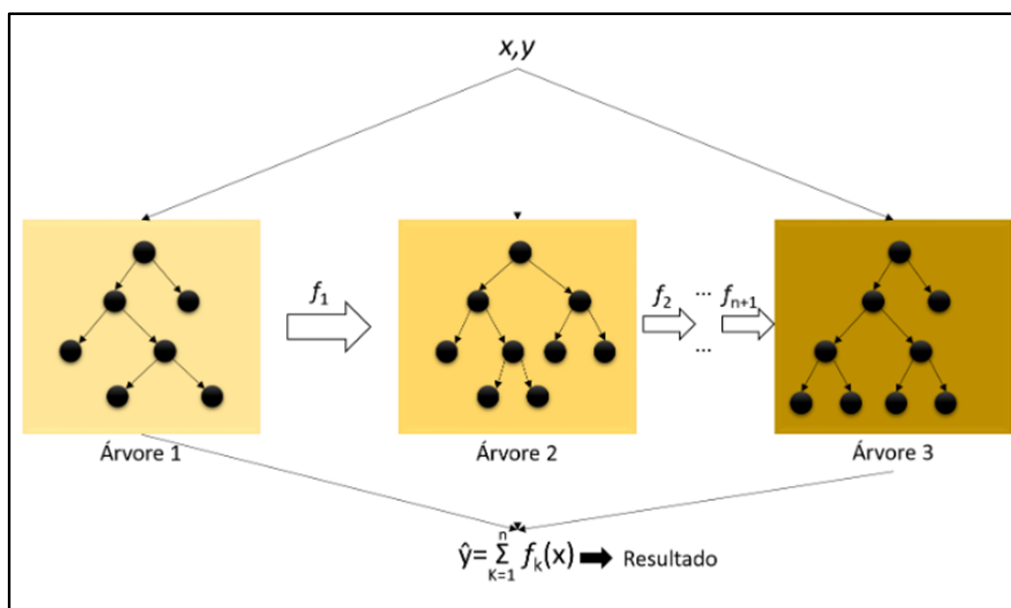


Figura 5 – Diagrama ilustrativo do funcionamento do algoritmo Extreme Gradient boosting Trees (XGBoost)

Fonte: Recuperado de <https://dadosaocubo.com/o-guia-do-xgboost-com-python/>

Este algoritmo é amplamente utilizado nas mais diversas áreas como previsão de demanda, detecção de fraudes, sistemas de recomendação e diagnóstico médico, entre outros. Suas vantagens incluem:

- Bom desempenho: a utilização de otimizações eficientes e técnicas avançadas, como paralelização de tarefas, permite acelerar o treinamento e a previsão;
- Eficácia na manipulação de dados complexos e de alta dimensionalidade: possui capacidade de lidar com uma grande quantidade de recursos (variáveis) mantendo resultados precisos;
- Suporte a diferentes linguagens de programação: Julia, Scala, Java, C++, R e Python;
- Lida com dados desbalanceados: O XGBoost é eficiente em lidar com problemas de desequilíbrio de classes em tarefas de classificação. Ele permite ajustar os pesos das amostras de treinamento para dar maior ênfase às classes minoritárias, ajudando a melhorar a precisão nas previsões.

Embora o XGBoost seja um algoritmo poderoso e de ampla utilização, existem desvantagens que devem ser consideradas, tais como:

- Tempo de treinamento mais longo: Embora seja conhecido por seu desempenho superior em relação a outros algoritmos de *gradient boosting*, o treinamento do modelo pode levar mais tempo em comparação com métodos mais simples, como regressão linear ou árvores de decisão isoladas;
- Sensível a ruídos e *outliers*: O *Gradient boosting* é muito sensível a *outliers* nos dados de treinamento;
- Uso de recursos computacionais: Devido à sua complexidade e capacidade de paralelização, o XGBoost pode exigir recursos computacionais significativos ao lidar com grandes conjuntos de dados.

Alguns dos **hiperparâmetros importantes do XGBoost** incluem:

- Controle de *overfitting*
 - `learning_rate`: valores ideais situam-se entre 0,01-0,2;
 - `max_depth`;
 - `Min_child_weight`: semelhante a `min_childleaf`; padrão é 1;
- Controle de velocidade
 - `colsample_bytree`: taxa de subamostra de colunas;
 - `subsample`: taxa de subamostra da instância de treinamento;

- `n_estimators`: número máximo de árvores de decisão; valor alto pode levar a *overfitting*.

3.2.4 Light Gradient boosting Machine (LightGBM)

Desenvolvido pela Microsoft, o *Light Gradient boosting Machine* (LightGBM) é um algoritmo de aprendizado de máquina baseado em *gradient boosting*, conhecido por seu alto desempenho e eficiência com grandes conjuntos de dados.

Este algoritmo apresenta basicamente as mesmas vantagens e desvantagens que o XGBoost, entretanto, apresenta maior velocidade de treinamento que ele. Outro fator importante que o difere do XGBoost é em relação às especificidades das otimizações, já que o LightGBM contém parâmetros que permite escolher como é realizado o crescimento da árvore (FERREIRA, 2020a), como mostra a Figura 6 a seguir.

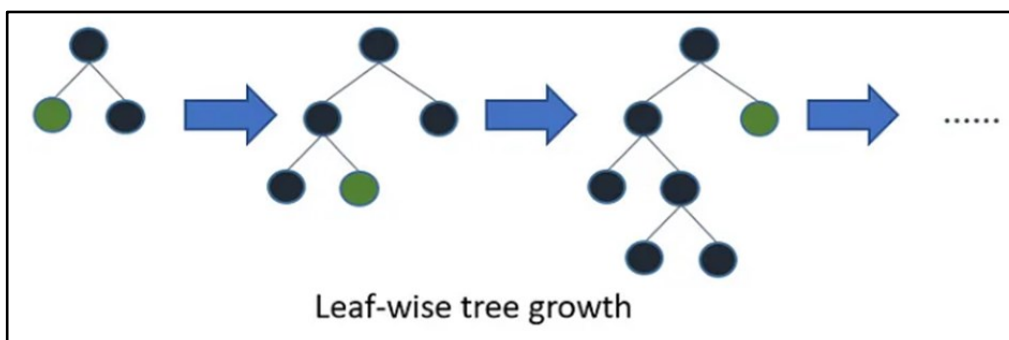


Figura 6 – Diagrama ilustrativo do funcionamento do algoritmo Light Gradient boosting Machine (LightGBM)

Fonte: Recuperado de <https://towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db>

Alguns dos **hiperparâmetros importantes** do (LightGBM) incluem:

- Controle de overfitting
 - `learning_rate`;
 - `max_depth`: o padrão é 20. Importante notar que a árvore ainda cresce *leaf-wise*. Portanto, é importante ajustar `num_leaves` (número de folhas em uma árvore) que deve ser menor que $2^{(\text{max_depth})}$. É um parâmetro muito importante para LGBM;
 - `min_data_in_leaf`: default=20, alias=min_data, min_child_samples;
- parâmetro para valores categóricos:

- `categorical_feature`: especificar os recursos categóricos que queremos usar para treinar nosso modelo;
- Parâmetros para controle de velocidade
 - `feature_fraction`: fração de recursos a serem obtidos para cada iteração;
 - `bagging_fraction`: dados a serem usados para cada iteração e geralmente é usado para acelerar o treinamento e evitar o *overfitting*;
 - `num_iterations`: número de iterações de reforço a serem executadas. padrão=100.

3.2.5 Redes Neurais MultiLayer Perceptron (MLP)

De acordo com Alves (2020),

As Redes Neurais Artificiais (RNA) são métodos de Deep Learning com capacidade de reconhecimento de padrões complexos ou numerosos e com função de aprendizado dentro da própria rede. As RNA se baseiam na arquitetura dos neurônios humanos e se destinam a reproduzir o aprendizado por meio do desenvolvimento de sistemas que aprendem com exemplos de treinamento (ALVES, 2020, online).

Assim como Alves (2020), Stropoli (2021) afirma que as RNA são modelos computacionais inspirados pelo sistema nervoso central (em particular o cérebro) que são capazes de realizar o aprendizado de máquina bem como o reconhecimento de padrões. Redes neurais artificiais geralmente são apresentadas como sistemas de “neurônios interconectados, que podem computar valores de entradas”, simulando o comportamento de redes neurais biológicas. A Figura 7 abaixo ilustra mais claramente um RNA do tipo MLP:

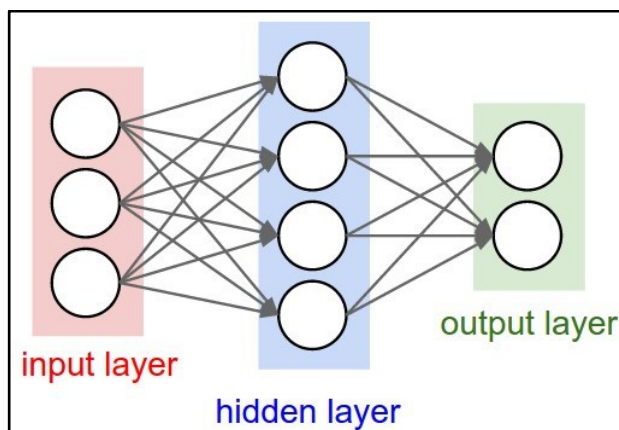


Figura 7 – Ilustração de uma RNA do tipo MLP

Fonte: Recuperado de <https://cs231n.github.io/neural-networks-1/>

Segundo Goodfellow (2016), existem dez principais arquiteturas de RNA, entre elas as *Redes Multilayer Perceptron* (MLP). O MLP é normalmente usado em aprendizado supervisionado como, por exemplo, reconhecimento e classificação de

idiomas em arquivos de áudios. Estudos sobre previsão de série temporal de poluição e evolução de preços de habitação vêm comparando o aprendizado por MLP com modelos de previsões tradicionais de série atemporais.

De forma sucinta, pode-se descrever o funcionamento de Rede Neural da seguinte forma: em cada neurônio há uma função de ativação (*activation function*) que processa uma combinação linear entre *inputs* e pesos sinápticos, gerando assim um sinal de saída. A informação flui da *input layer* para as *hidden layers* e, por fim, para a *output layer*. Nesse fluxo os inputs de dados da input layer são alimentados para os neurônios das *hidden layers* que finalmente alimentam os neurônios finais da *output layer*.

A primeira passada de informação (propagação) pela rede é geralmente feita com parâmetros aleatórios para as funções de ativação dos neurônios. Ao realizar a propagação, chamada de *feed forward*, temos sinais de saídas nos neurônios da *output layer*. No fim da propagação, a função de custo (uma métrica de erro) é calculada e o modelo então ajusta os parâmetros dos neurônios na direção de um menor custo (por meio do gradiente – derivada multivariada). Assim, uma nova propagação é gerada e a numa nova função custo é calculada, da mesma forma como é realizado a atualização dos parâmetros dos neurônios. O nome desse algoritmo é Retro-propagação (*backpropagation*). Cada vez que ele é executado denomina-se como época (*epoch*). E quando as épocas estabelecidas se encerram, a rede neural encerra o seu treinamento/aprendizado.

Alguns dos **hiperparâmetros importantes do MLP** são:

- activation (function);
- alpha;
- batch_size;
- hidden_layers_sizes.

No Capítulo 4 (Metodologia), serão descritas métricas utilizadas para análise e, no Capítulo 5 (Resultados e Discussão), será realizado comparação do desempenho dos algoritmos para determinar qual obteve melhor performance na classificação dos desfechos pretendidos (edentulismo, perda da dentição funcional e perda de um ou mais dentes).

4 METODOLOGIA

Esta seção descreve a metodologia adotada para realizar o estudo proposto na presente dissertação. A metodologia engloba coleta de dados, seleção de recursos, pré-processamento e aplicação de técnicas de modelagem, conforme ilustra o esquema na Figura 8.

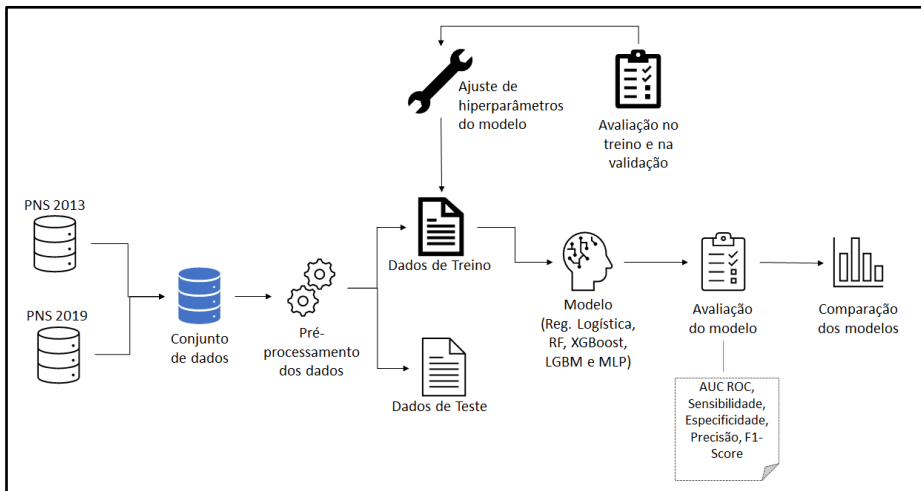


Figura 8 – Metodologia aplicada no desenvolvimento do experimento.

Fonte: Desenvolvido pelo Autor (2023)

4.1 Coleta de Dados

Neste estudo, foram utilizados dados da Pesquisa Nacional de Saúde (PNS), relativos aos inquéritos de 2013 e 2019.

Os dados, em formato *.csv, foram obtidos por meio do portal <https://www.pns.icict.fiocruz.br/bases-de-dados/>, mantido pela Fundação Oswaldo Cruz (Fiocruz). As informações disponibilizadas foram extraídas dos inquéritos PNS 2013 e 2019 fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em seu portal.

4.2 Seleção de Recursos

A seleção de recursos foi baseada na literatura. Realizou-se o mapeamento das variáveis utilizadas no estudo “Predictors of tooth loss: A *Machine Learning* approach” (ELANI et al., 2021), que analisa a influência de aspectos socioeconômicos da população dos Estados Unidos na perda dentária completa e incremental entre adultos; mostraram-se possíveis os seguintes desfechos: edentulismo, perda funcional dentária ou ausência de dentes.

Para o presente estudo, foram selecionadas variáveis que correspondam ou se assemelham às utilizadas no estudo de referência, com intuito de avaliar aquela modelagem no contexto dos dados e da realidade brasileira, permitindo assim adotar uma perspectiva entre essas populações, por meio dos dados do PNS.

Adicionalmente, foram inseridas características para criação da variável derivada **auto_cuidado_saude_bucal**, conforme especificado na tese de Bordin (2017), com vistas a melhorar o desempenho do experimento.

Para facilitar a compreensão das informações apresentadas no estudo, foram atribuídos pseudônimos às nomenclaturas encontradas na PNS, cujas originais e seus correspondentes pseudônimos podem ser encontradas no APÊNDICE A – TABELA DE PSEUDÔNIMOS X NOMENCLATURAS PNS 2013 E 2019.

4.3 Pré-processamento

Inicialmente, a proposta era utilizar o conjunto de dados do PNS 2013 para treino e o PNS 2019 para teste, de forma análoga ao do estudo de Elani et al. (2021) em que foram utilizados dados do NHANES do ciclo 2011 a 2012 para treino e do ciclo 2013 a 2014 para teste.

Por outro lado, observou-se que o PNS 2013 não contém registro de desfecho edentulismo, ou seja, não dispunha de instâncias com este desfecho. Diante desta ausência no PNS 2013, optou-se por unificar as bases PNS 2013 e 2019 em um único conjunto, já que o aumento do volume de dados enseja melhora no treino dos algoritmos para os demais desfechos (perda funcional dentária ou ausência de dentes).

Para realizar a unificação a partir do PNS 2019, conjunto de dados definido como principal, realizou-se os seguintes processamentos no conjunto de dados de 2013:

- Renomear variáveis com informações iguais, mas que constam com nomes diferentes;
- Criação da variável derivada VDF004 (faixa_renda_per_capita) a partir do Rendimento Domiciliar per Capita (VDF003) e do Salário-Mínimo em 2013, classificando os valores de acordo com as categorias declaradas no dicionário de dados do PNS 2019 para a variável VDF004;

- Redefinição das categorias divergentes entre as duas bases, como foram os casos das variáveis C011 (estado_civil), Q030 (diabetes-Q03001), Q002 (hipertensão-Q00201).

Em seguida, foi aplicado o filtro para selecionar apenas as instâncias de adultos com idade igual ou superior a 18 anos. Os dados nulos/vazios ou categóricos com valor ignorado (9 e/ou 99) foram removidos. Empregou-se a técnica de *one-hot encoding* para as variáveis *raça*, *educacao_nivel*, *faixa_renda_per_capita* e *estado_civil*, a fim de aumentar o desempenho do algoritmo. De acordo com Batista e Chiavegatto Filho (2019, p. 7), “[...] se as variáveis qualitativas [...] forem nominais, é interessante transformá-las em um conjunto de variáveis indicadoras (denominadas *dummies*), as quais são representadas por valores binários [...]”.

A técnica de *one-hot encoding* é aplicada para representar essas variáveis de forma numérica, permitindo que os algoritmos de aprendizado de máquina manipulem variáveis categóricas ou discretas de forma matemática, sem impor uma ordem ou relação numérica entre elas, evitando pesos indesejados.

O conjunto de dados consolidado foi então dividido de forma estratificada para os desfechos (classes) com reserva de 70% para treino e os 30% restante para teste. A estratificação é especialmente recomendada para casos em que se realiza a separação em conjuntos de dados desbalanceados (Figura 9). Esta prática visa garantir a representatividade adequada das classes nos conjuntos de treinamento e teste, evitando vieses e permitindo que o modelo aprenda e generalize corretamente para todas as classes do problema em questão.

```
X_treino, X_teste, y_treino, y_teste = train_test_split(X, y, test_size=0.3, random_state=2811, stratify=y)
```

Figura 9 – Código para dividir o conjunto de dados em treino e teste de forma estratificada.

Fonte: Desenvolvido pelo Autor (2023)

Em relação às variáveis quantitativas contínuas (*idade*, *num_quartos_casa*, *imc* e *tamanho_familia*), aplicou-se a padronização nos dados de treino e posterior transformação nos dados de teste.

A padronização é uma etapa fundamental no pré-processamento dos dados, pois é através dele que se normalizam os dados, colocando-os em uma escala comum. Com o uso desta técnica, os valores originais da variável são redimensionados para terem média igual a 0 e desvio-padrão igual a 1. Conforme

Batista e Chiavegatto Filho (2019, p. 5), “[...] algoritmos de predição apresentam melhores resultados quando as variáveis preditoras estão sob uma mesma escala”.

A exemplo do que foi realizado neste estudo, o ajuste da padronização deve ser realizado apenas nos dados de treino, de modo a prevenir vazamento de informações, já que se aplicado a dados de teste, estaria utilizando informações desse conjunto para ajustar os parâmetros e fornecendo um conhecimento prévio sobre os dados de teste ao modelo (Figura 10).

```
# Padronizando
from sklearn.preprocessing import StandardScaler
con_vars = ['idade', 'num_quartos_casa', 'imc', 'tamanho_familia']
scaler = StandardScaler()

X_treino[con_vars] = scaler.fit_transform(X_treino[con_vars])
X_teste[con_vars] = scaler.transform(X_teste[con_vars])
```

Figura 10 – Código para aplicar a padronização nas variáveis contínuas do conjunto de dados.

Fonte: Desenvolvido pelo Autor (2023)

4.3.1 Reamostragem

O desbalanceamento de dados ocorre quando a distribuição das classes em um conjunto de dados é significativamente desigual: algumas classes têm um número muito maior de instâncias do que outras.

Constatou-se esta característica de desbalanceamento no conjunto de dados, onde os desfechos representam, do total do *conjunto de dados*, respectivamente: Edentulismo 1,19%; Perda da dentição funcional 8,6% e Ausência de dentes 20,59% (Figura 11). Foram avaliadas três técnicas: duas de subamostragem (*undersampling*) (OneSidedSelection e NearMiss) e uma de superamostragem (*oversample*) SMOTE.

```
SMOTE

In [ ]: smote = SMOTE()

Desfecho 'edentulismo'

In [ ]: X_treino_edentulismo_smote, y_treino_edentulismo_smote = smote.fit_resample(X_treino, y_treino.edentulismo)
        y_treino_edentulismo_smote.value_counts()

Desfecho 'perda_denticao_funcional'

In [ ]: X_treino_perda_denticao_funcional_smote, y_treino_perda_denticao_funcional_smote = smote.fit_resample(
        X_treino, y_treino.perda_denticao_funcional)
        y_treino_perda_denticao_funcional_smote.value_counts()

Desfecho 'dentes_ausentes'

In [ ]: X_treino_dentes_ausentes_smote, y_treino_dentes_ausentes_smote = smote.fit_resample(X_treino, y_treino.dentes_ausentes)
        y_treino_dentes_ausentes_smote.value_counts()
```

Figura 11 – Código de oversample SMOTE aplicado aos três tipos de desfechos.

Fonte: Desenvolvido pelo Autor (2023)

A técnica SMOTE possibilitou o melhor desempenho, porém sua utilização limitou-se apenas ao treino do algoritmo de Rede Neural - MLP, uma vez que, em relação aos demais algoritmos, o ajuste de hiperparâmetros permitiu melhores resultados.

4.3.2 Criação de variáveis derivadas (Feature Engineering)

Para melhorar a performance dos modelos, foram criadas as variáveis `auto_cuidado_saude_bucal` e `doencaocronica`, ambas a partir do estudo elaborado por Bordin (2017).

A variável `doencaocronica` representa a presença ou ausência de alguma das doenças crônicas selecionadas para este estudo (asma, artrite, derrame, ataque_cardiaco, diabetes, colesterol_alto e hipertensao). A introdução desse recurso se fez interessante, uma vez que se verificou que este aspecto pouco contribui para determinar o desfecho, evitando assim demasiado aumento de dimensionalidade. Desta forma, as variáveis originais foram excluídas, mantendo apenas a derivada.

Por sua vez, a variável `auto_cuidado_saude_bucal` contempla informações que, segundo a mencionada tese, possuem alto valor preditivo. Deste modo, foram compiladas informações acerca de: Tempo desde a última consulta ao dentista, Frequência de escovação, Materiais utilizados para auto limpeza e Frequência de substituição da escova dental do seguinte modo:

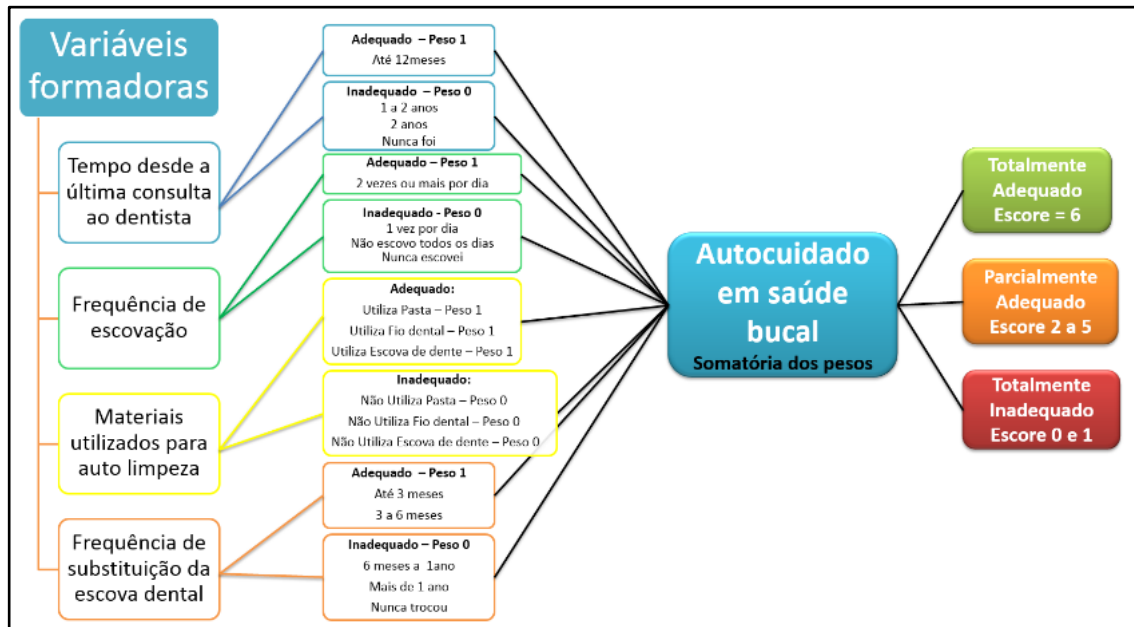


Figura 12 – Processo de formação da variável dependente: ‘Auto cuidado em Saúde Bucal’.

Fonte: Bordin (2017, p. 35).

Para evitar problemas de colinearidade (relação entre variáveis cuja característica é indesejada, pois pode levar o modelo a gerar resultados instáveis numericamente), com exceção de “tempo_ultima_visita_dentista”, as variáveis originais utilizadas para a formação de auto_cuidado_saude_bucal e doencao_cronica foram removidas, mantendo-se apenas a correspondente derivada.

4.4 Modelagem

Após a seleção de recursos, foram aplicadas técnicas de modelagem para construir os modelos preditivos. Neste estudo, explorou-se cinco algoritmos de aprendizado de máquina: Regressão Logística, Floresta Aleatória (*Random Forest*), *Light Gradient boosting Machine* (LGBM), *Extreme Gradient boosting Trees* (XGBoost) e Rede Neural Multi-layer Perceptron classifier (MLP), os mesmos utilizados no estudo de Elani et al. (2021). Os modelos foram implementados utilizando a biblioteca *Scikit-learn* em Python.

Os cinco algoritmos foram parametrizados de modo a trabalhar com classes desbalanceadas, uma vez que dois dos desfechos representam uma pequena parcela dos dados: Edentulismo (1,19%), Perda da dentição funcional (8,6%) e Ausência de dentes (20,59%). Além dos hiperparâmetros específicos para tratamento de dados desbalanceados, decidiu-se por aqueles com maior êxito relatado na literatura. Os resultados para cada algoritmo são obtidos através do esquema de

RandomizedSearchCV, que realiza uma busca aleatória em um subconjunto dos hiperparâmetros candidatos para buscar os melhores candidatos.

A configuração final dos hiperparâmetros definida para cada algoritmo, apresentada na Tabela de Configuração Final de Hiperparâmetros, pode ser encontrada no Apêndice A.

4.5 Interpretação das Métricas de Avaliação

A escolha das métricas de avaliação é fundamental para a compreensão e interpretação dos resultados obtidos neste estudo. Nesta seção, serão especificadas as métricas selecionadas, discutindo-se sua importância no contexto do problema em estudo. Essas métricas foram escolhidas com objetivo de fornecer uma medida precisa do desempenho dos modelos propostos e permitir uma comparação adequada entre os diferentes algoritmos.

Em avaliação de resultados de pesquisa em saúde, deve-se levar em conta as seguintes características:

- A maioria dos problemas são preditivos do tipo classificação;
- O modelo a ser desenvolvido lida com dados desbalanceados;
- O desfecho analisado normalmente é um evento raro, portanto, representando uma classe minoritária.

Por se tratar de um estudo de *Machine Learning* em Saúde, as seguintes métricas foram adotadas: Sensibilidade ou Recall, Especificidade, Precisão, F1-Score e a Curva ROC.

Tudo começa a partir de quatro estatísticas básicas passíveis de cálculo a partir da previsão do modelo:

- Verdadeiros Positivos (VP) - O modelo classifica o exemplo como positivo e o rótulo real também é positivo;
- Verdadeiros Negativos (VN) - O modelo classifica o exemplo como negativo e o rótulo real também é negativo;
- Falsos Positivos (FP) - O modelo classifica o exemplo como positivo, mas o rótulo real é negativo;
- Falsos Negativos (FN) - O modelo classifica o exemplo como negativo, mas o rótulo real é positivo.

Essas métricas podem ser representadas através de uma tabulação cruzada, denominada de Matriz de Confusão, de classes observadas e preditas (Figura 13):

		Resposta Observada		total
		p	n	
Resposta Predita	p'	VP (a)	FP (b)	a+b
	n'	FN (c)	VN (d)	b+d
total		a+c	b+d	

Figura 13 – Exemplo de Matriz de Confusão para Algoritmos de Classificação.

Fonte: Figura extraída do minicurso *Machine Learning aplicado à Saúde*, disponível em <https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/29/94/244-1?inline=1>

A partir de VP, VN, FP e FN então, são calculadas as métricas adotadas (Quadro 1):

MÉTRICA (NOMENCLATURA)	CÁLCULO	DESCRIÇÃO
Sensibilidade ou <i>Recall</i>	$Sensibilidade = \frac{VP}{VP + FN}$	Proporção de verdadeiros positivos corretamente identificados.
Especificidade	$Especificidade = \frac{VN}{VN + FP}$	Proporção de verdadeiros negativos corretamente identificados.
Precisão	$Precisão = \frac{VP}{VP + FP}$	Quantifica o número de predições de classes positivas que pertence, de fato, à classe positiva.
F1-Score	$F1 = \frac{2 * VP}{2 * VP + FP + FN}$	Medida harmônica entre a precisão e a sensibilidade

Quadro 1 – Métricas

Fonte: Elaborado pelo Autor (2023).

A curva de característica de operação (ROC), também adotada neste estudo, é outra ferramenta comum usada com classificadores binários. Representa a taxa de verdadeiros positivos (VP) em relação à taxa de positivos falsos (FP) (GÉRON, 2021, p. 176). A métrica acurácia, muito utilizada na área de *Machine Learning*, não foi

considerada pois indicam a proporção de acertos (observações classificadas corretamente), o que não configura uma boa medida, já que a base de dados se apresenta desbalanceada.

5 RESULTADOS E DISCUSSÃO

Neste capítulo, serão apresentados e discutidos os resultados obtidos no experimento. A análise dos resultados será organizada nas seguintes seções: 5.1 Análise Exploratória de Dados; 5.2 Apresentação dos resultados com análise da performance dos algoritmos e importância das variáveis preditoras; 5.3 Discussão dos resultados em relação aos trabalhos relacionados, com uma análise comparativa com o estudo “Predictors of tooth loss: A *Machine Learning* approach”, que foi estabelecido como ponto de partida e referência para esta dissertação.

5.1 Análise Exploratória de Dados

Nesta seção, será realizada a Análise Exploratória de Dados, etapa importante em uma pesquisa em ciência de dados. Este procedimento possibilita compreender as informações obtidas a partir das variáveis do PNS selecionadas para estudo. Como complemento, também foram incluídas variáveis derivadas a partir da transformação das originais do PNS..

Foram selecionadas do PNS as seguintes variáveis: 'num_quartos_casa', 'sexo', 'idade', 'raca', 'estado_civil', 'plano_de_saude', 'peso', 'altura', 'tamanho_familia', 'educacao_nivel', 'empregado', 'faixa_renda_per_capita', 'utiliza_escova_dente', 'utiliza_pasta', 'utiliza_fio_dental', 'frequencia_escovacao', 'frequencia_substituicao_escova_dental', 'tempo_ultima_visita_dentista', 'hipertensao', 'diabetes', 'colesterol_alto', 'ataque_cardiaco', 'derrame', 'asma', 'artrite'.

Para variáveis categóricas serão utilizadas as categorias definidas no próprio PNS: 'raca', 'plano_de_saude', 'educacao_nivel', 'faixa_renda_per_capita', 'sexo', 'estado_civil'. Estas estão listadas na Tabela 1.

Atributo	Índice	Categoria
<u>raça</u>	1	Branca
	2	Preta
	3	Amarela
	4	Parda
	5	Indígena
	9	Ignorado
<u>plano de saúde</u>	1	Sim
	2	Não
	9	Ignorado
<u>educação nível</u>	1	Sem instrução
	2	Fundamental incompleto ou equivalente
	3	Fundamental completo ou equivalente
	4	Médio incompleto ou equivalente
	5	Médio completo ou equivalente
	6	Superior incompleto ou equivalente
	7	Superior completo
<u>faixa renda per capita</u>	1	Até ¼ salário mínimo
	2	Mais de ¼ até ½ salário mínimo
	3	Mais de ½ até 1 salário mínimo
	4	Mais de 1 até 2 salários mínimos
	5	Mais de 2 até 3 salários mínimos
	6	Mais de 3 até 5 salários mínimos
	7	Mais de 5 salários mínimos
<u>sexo</u>	1	Homem
	2	Mulher
<u>estado civil</u>	1	Casado(a)
	2	Divorciado(a) ou desquitado(a) ou separado(a) judicialmente
	3	Viúvo(a)
	4	Solteiro(a)
	9	Ignorado

Tabela 1 – Categorias definidas no PNS para variáveis categóricas selecionadas.
Fonte: Elaborado pelo Autor (2023).

Outras variáveis que originalmente não eram categóricas ou que apresentavam categorias fora do escopo deste estudo receberam nova codificação, de forma a facilitar a leitura e a compreensão durante essa etapa: 'tempo_ultima_visita_dentista', 'imc', 'doenca_cronica', 'auto_cuidado_saude_bucal'. Encontram-se na Tabela 2.

Atributo	Índice	Categoria
<u>tempo ultima visita dentista</u>	0	Demais categorias
	1	Até 1 ano.
<u>imc</u>	1	Baixo Peso
	2	Peso Adequado
	3	Sobrepeso
	4	Obesidade
<u>doencao cronica</u>	0	Não possui doença crônica
	1	Possui ao menos uma doenças crônica
<u>auto cuidado saude bucal</u>	1	Totalmente Adequado
	2	Parcialmente Adequado
	3	Totalmente Inadequado
	4	Não informado

Tabela 2 – Categorias codificadas para variáveis do PNS selecionadas.
Fonte: Elaborado pelo Autor (2023).

Os dados serão explorados conforme os interesses definidos nas seções a seguir.

5.1.1 Características gerais da amostra

A amostra completa totaliza 114.279 instâncias, distribuídas conforme a Tabela 3:

CARACTERÍSTICA	n	%
Dentição Completa	79549	69,61
Dentes Ausentes	23535	20,59
Perda da Dentição Funcional	9830	8,6
Edentulismo	1365	1,19

Tabela 3 – Amostra de instâncias.
Fonte: Elaborado pelo Autor (2023).

Para este trabalho, que se propõe prever a perda dentária, foram considerados como desfechos de interesse somente aqueles com ausência de dentes, perda de dentição funcional e edentulismo.

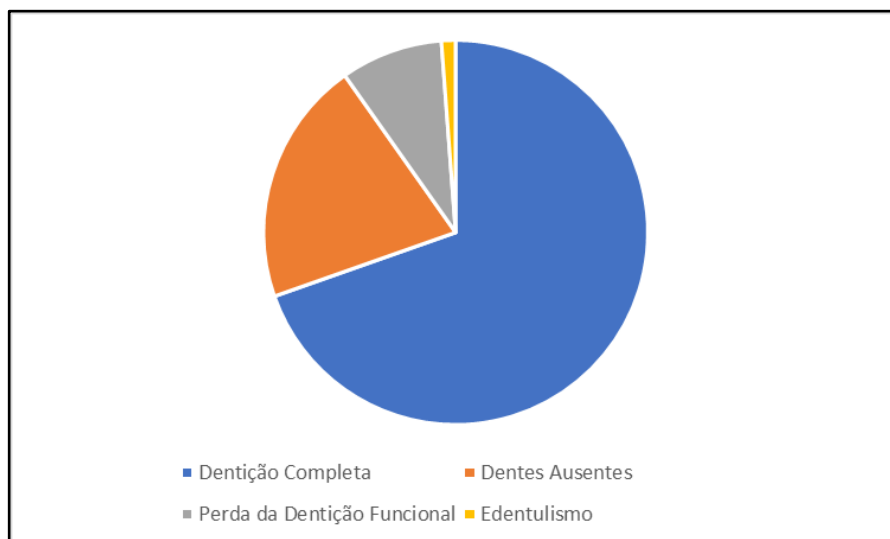


Figura 14 – Distribuição Percentual da Condição da Dentição.
Fonte: Elaborado pelo Autor (2023).

5.1.2 Análises

5.1.2.1 Características socioeconômicas

Os dados relativos à idade foram classificados conforme 3 categorias: jovem, adulto e idoso.

O critério adotado foi o da legislação brasileira. A Lei 12.852/13 (Estatuto da Juventude) descreve como jovem o indivíduo com idade entre 15 e 29 anos, enquanto a Lei 10.741/03 (Estatuto do Idoso) identifica como idoso aquele com idade partir de 60 anos; por sua vez, adotou-se como adulto um critério residual, ou seja, todo indivíduo de idade entre 30 e 59 anos.

A investigação identificou que, com o envelhecimento, a perda dentária se torna um problema naturalmente mais recorrente. Em torno de 47,7% dos indivíduos idosos relataram sofrer de algum tipo de perda dentária. Em relação aos adultos, a ocorrência foi confirmada por 30,99% e, entre os jovens, apenas 3,54%.

A distribuição completa entre as faixas etárias e as características dentárias pode ser observada na Tabela 4 e Figura 15:

FAIXA ETÁRIA	DENTIÇÃO COMPLETA (%)	DENTES AUSENTES (%)	PERDA DA DENTIÇÃO FUNCIONAL (%)	EDENTULISMO (%)
Jovens	96,46	3,38	0,16	0
Adultos	69,01	23,49	7,11	0,39
Idosos	52,31	25,64	18,09	3,97

Tabela 4 – Distribuição completa entre faixas etárias e características dentárias.
Fonte: Elaborado pelo Autor (2023).

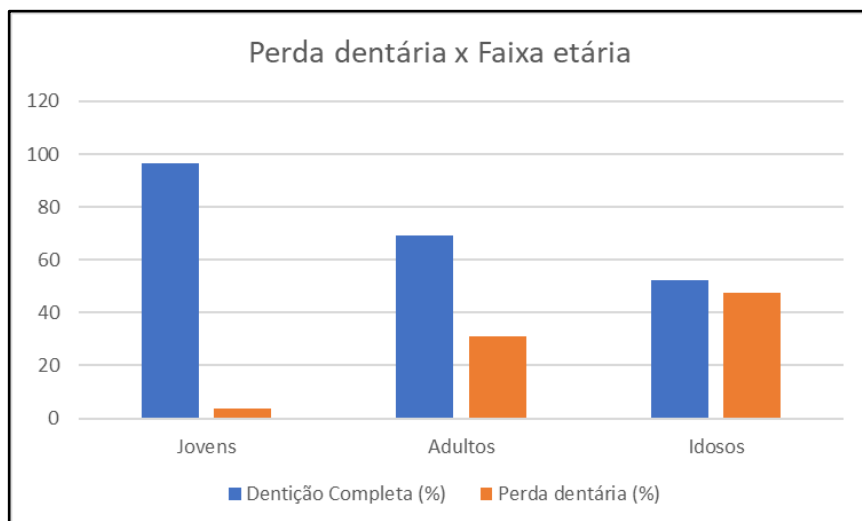


Figura 15 – Gráfico ilustrativo da distribuição completa entre faixas etárias e características dentárias.

Fonte: Elaborado pelo Autor (2023).

Em relação à incidência de desfechos por sexo, não foi possível observar diferença significativa de perda dentária entre homens e mulheres. A distribuição dos desfechos também foi semelhante para ambos os sexos (Tabela 5 e Figura 16).

SEXO	DENTIÇÃO COMPLETA (%)	DENTES AUSENTES (%)	PERDA DA DENTIÇÃO FUNCIONAL (%)	EDENTULISMO (%)
Homem	69,65	20,42	8,94	0,99
Mulher	69,57	20,73	8,33	1,37

Tabela 5 – Comparação da distribuição da perda dentária entre homens e mulheres.

Fonte: Elaborado pelo Autor (2023).

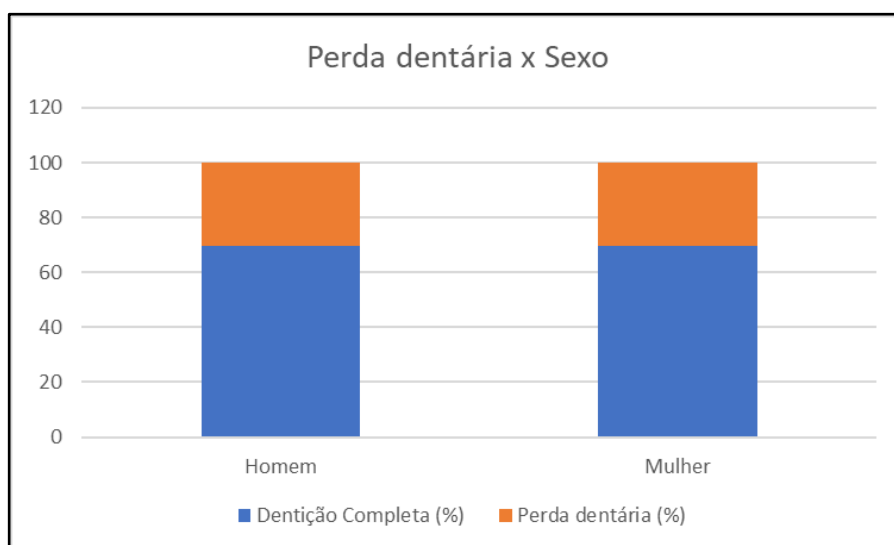


Figura 16 – Gráfico ilustrativo da distribuição da perda dentária entre homens e mulheres.

Fonte: Elaborado pelo Autor (2023).

Os dados analisados sugerem que, em nosso país, a etnia indígena apresenta a maior perda dentária (33,86%), seguida da parda (32,03%), da preta (31,94%) e da

branca (28,04%). A etnia amarela apresentou o melhor desfecho, com 27,53%, como mostram a Tabela 6 e a Figura 17:

ETNIA	DENTIÇÃO COMPLETA (%)	DENTES AUSENTES (%)	PERDA DA DENTIÇÃO FUNCIONAL (%)	EDENTULISMO (%)
Amarela	72,47	17,84	9,07	0,62
Branca	71,96	19,21	7,65	1,18
Indígena	66,15	22,18	10,51	1,17
Parda	67,96	21,64	9,19	1,2
Preta	68,06	21,27	9,42	1,25

Tabela 6 – Porcentagem de perda dentária por grupo étnico.
Fonte: Elaborado pelo Autor (2023).

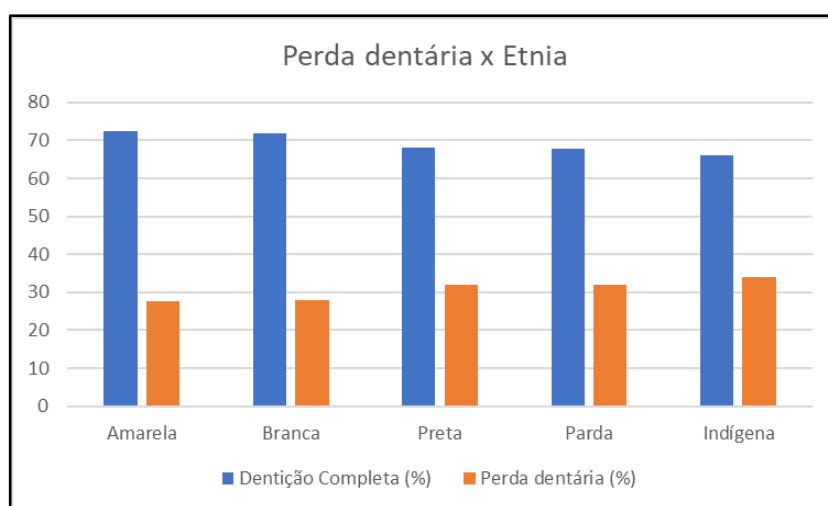


Figura 17 – Gráfico ilustrativo da porcentagem de perda dentária por grupo étnico.
Fonte: Elaborado pelo Autor (2023).

Em relação ao estado civil dos indivíduos, observou-se maior incidência de perda dentária entre os viúvos, 42,72%, e os divorciados, 38,07%. Vide Tabela 7 e Figura 18.

ESTADO CIVIL	DENTIÇÃO COMPLETA (%)	DENTES AUSENTES (%)	PERDA DA DENTIÇÃO FUNCIONAL (%)	EDENTULISMO (%)
Solteiro	76,67	16,74	6,06	0,54
Casado	66,86	22,7	9,27	1,17
Divorciado	61,93	26,08	10,57	1,42
Viúvo	57,28	22,98	15,51	4,23

Tabela 7 – Porcentagem de perda dentária por estado civil.
Fonte: Elaborado pelo Autor (2023).

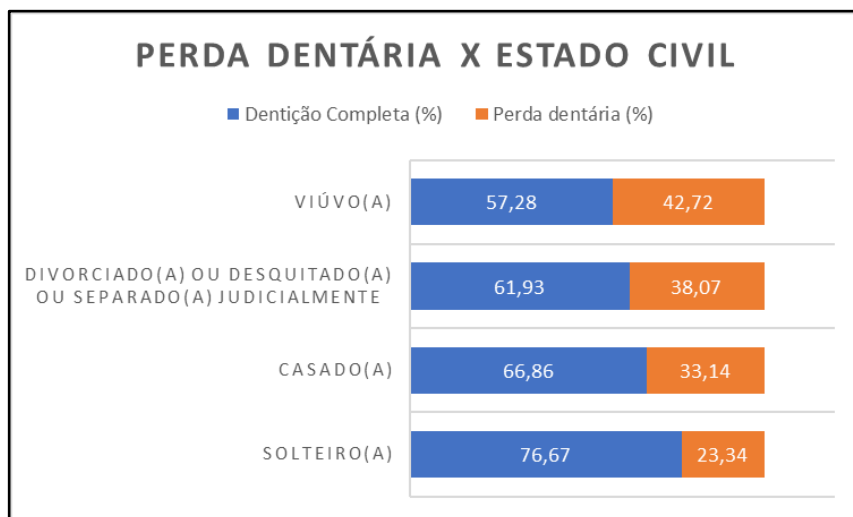


Figura 18 - Gráfico ilustrativo da Porcentagem de perda dentária por estado civil.
Fonte: Elaborado pelo Autor (2023).

Conforme a Tabela 8, a informação acima corrobora com os indicativos de que com a idade, o índice de perda dentária tende a aumentar, uma vez que os estados civis destacados estão relacionados a faixas etárias mais elevadas.

ESTADO CIVIL	JOVENS (%)	ADULTOS (%)	IDOSOS (%)
Solteiro	32,04	57,73	10,23
Casado	8,62	65,29	26,09
Divorciado	3,28	66,5	30,22
Viúvo	0,38	22,01	77,61

Tabela 8 – Variação do Índice de perda dentária ao longo da idade.
Fonte: Elaborado pelo Autor (2023).

A menor perda dental acumulada apresentou-se entre os solteiros (23,34%), seguidos pelos casados (33,14%).

Outra variável que, aparentemente, marca as condições dentárias dos indivíduos no Brasil é o grau de instrução. Este é um importante marcador de desigualdade no país. Os dados analisados sugerem que, de modo geral, um maior grau de instrução está associado a uma menor perda dentária.

Na Tabela 9 e Figura 19, observa-se que indivíduos com grau de instrução “fundamental incompleto ou equivalente” registram o maior percentual de perda dentária (44,37%), enquanto quem se declarou “sem instrução”, apesar de menor perda dentária acumulada (42,69%), apresenta o maior percentual nos desfechos finais, i.e. edentulismo e perda de dentição funcional.

NÍVEL DE INSTRUÇÃO	DENTIÇÃO COMPLETA (%)	DENTES AUSENTES (%)	PERDA DA DENTIÇÃO FUNCIONAL (%)	EDENTULISMO (%)
Superior completo ou equivalente	81,25	15,04	3,45	0,25
Superior incompleto ou equivalente	87,41	10,62	1,9	0,07
Médio completo ou equivalente	76,41	18	5,19	0,4
Médio incompleto ou equivalente	76,56	17,56	5,13	0,75
Fundamental completo ou equivalente	66,26	24,06	8,69	1
Fundamental incompleto ou equivalente	55,63	27,56	14,51	2,3
Sem instrução	57,31	21,52	17,73	3,44

Tabela 9 – Associação entre grau de instrução e perda dentária.
Fonte: Elaborado pelo Autor (2023).

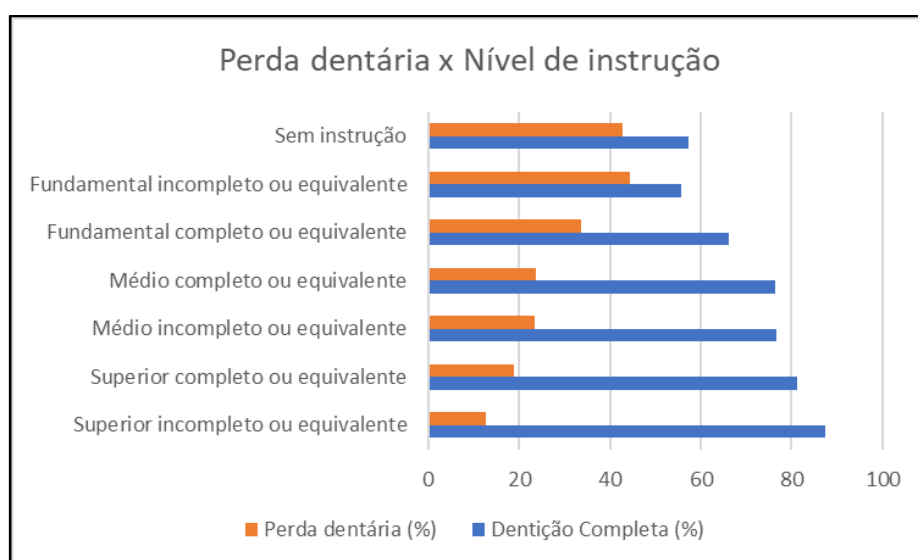


Figura 19 – Gráfico ilustrativo da associação entre grau de instrução e perda dentária.

Fonte: Elaborado pelo Autor (2023).

Outro marcador de risco associado à perda dentária é a obesidade. O Índice de Massa Corporal (IMC) é utilizado para classificar o estado nutricional dos indivíduos. Esta é medida recomendada para avaliar obesidade, tanto em nível populacional quanto na prática clínica. De acordo com a Norma Técnica da Vigilância Alimentar e Nutricional do Sistema de Vigilância Alimentar e Nutricional (SISVAN), o índice é a combinação entre duas medidas antropométricas, peso e altura, e pode ser classificado em:

- Valores de IMC menores que 18,5: adulto com baixo peso;
- Valores de IMC entre 18,5 e 25,0: adulto com peso adequado (eutrófico);
- Valores de IMC entre 25,0 e 30,0: adulto com sobrepeso;
- Valores de IMC maiores ou iguais a 30,0: adulto com obesidade.

O estudo em tela, considerando os dados fornecidos na pesquisa, observou que o aumento da massa corporal, medido por meio de IMC, está associado à maior incidência de perda dentária em geral (Tabela 10 e Figura 20).

CLASSE DE IMC	DENTIÇÃO COMPLETA (%)	DENTES AUSENTES (%)	PERDA DA DENTIÇÃO FUNCIONAL (%)	EDENTULISMO (%)
Baixo Peso	74,43	14,34	9,17	2,06
Peso Adequado	72,4	18,38	8	1,22
Sobrepeso	68,29	21,75	8,8	1,16
Obesidade	66,21	23,32	9,34	1,13

Tabela 10 - Associação entre IMC e incidência de perda dentária.

Fonte: Elaborado pelo Autor (2023).

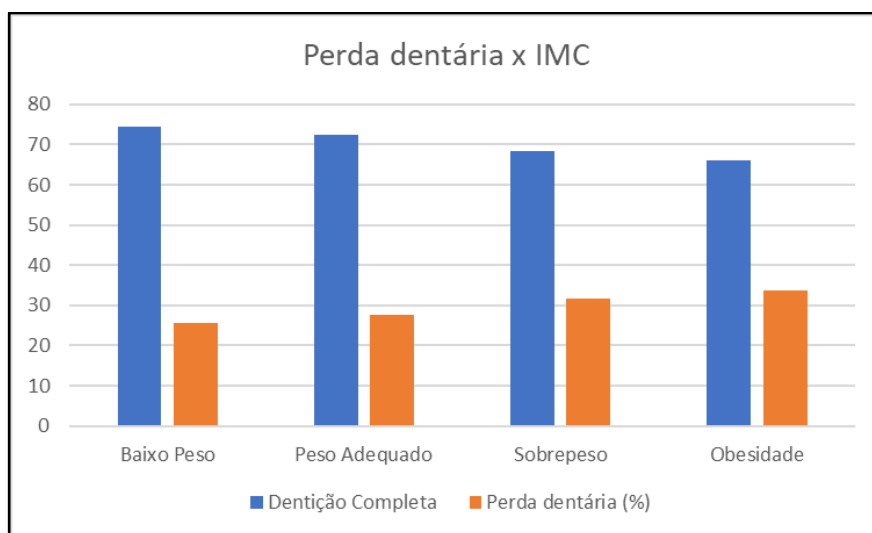


Figura 20 – Gráfico ilustrativo da associação entre IMC e perda dentária.

Fonte: Elaborado pelo Autor (2023).

De acordo com os dados registrados no PNS 2013 e 2019, a presença de pelo menos uma das doenças crônicas analisadas no estudo (hipertensão, diabetes, colesterol alto, ataque cardíaco, derrame, asma e artrite) sugere relação com a perda dentária.

Para os indivíduos que relataram um dos desfechos de perda dentária e presença de pelo menos uma doença crônica, o percentual acumulado de perda foi

de 40,69%, enquanto para os que não possuíam doença crônica, foi de 22,54%, ou seja, 1,80 vezes maior, conforme a Tabela 11 e a Figura 21.

DIAGNÓSTICO DE DOENÇA CRÔNICA	DENTIÇÃO COMPLETA	DENTES AUSENTES	PERDA DA DENTIÇÃO FUNCIONAL	EDENTULISMO
Não	77,46	16,65	5,35	0,54
Sim	59,31	25,77	12,87	2,05

Tabela 11 – Relação entre Doenças Crônicas e Incidência de Perda Dentária.
Fonte: Elaborado pelo Autor (2023).

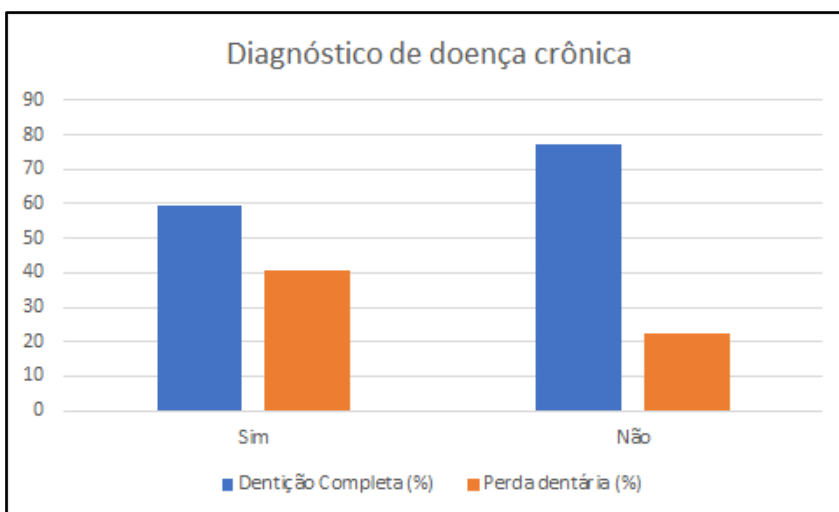


Figura 21 – Gráfico ilustrativo da relação entre Doenças Crônicas e Incidência de Perda Dentária.

Fonte: Elaborado pelo Autor (2023).

5.1.2.2 Comportamento de saúde oral

Conforme especificado na Metodologia (Capítulo 4), a variável `auto_cuidado_saude_bucal` foi criada para pontuar o cumprimento das boas práticas de saúde bucal. A eficácia das boas práticas pode ser observada através da Tabela 12 e Figura 22, nas quais indivíduos que pontuaram com totalmente adequado apresentaram menor índice de perda dentária (23,41%) e os que se avaliaram como totalmente inadequado apresentaram o maior índice de perda (34,56%).

AUTO CUIDADO EM SAÚDE BUCAL	DENTIÇÃO COMPLETA	DENTES AUSENTES	PERDA DA DENTIÇÃO FUNCIONAL	EDENTULISMO
Totalmente Adequado	76,59	18,38	4,81	0,22
Não informado	76,41	4,81	12,09	6,69
Parcialmente Adequado	67,45	21,52	9,64	1,38
Totalmente Inadequado	65,43	17,28	17,28	0

Tabela 12 – Relação entre qualidade de auto cuidado em saúde bucal e incidência de perda dentária.

Fonte: Elaborado pelo Autor (2023).

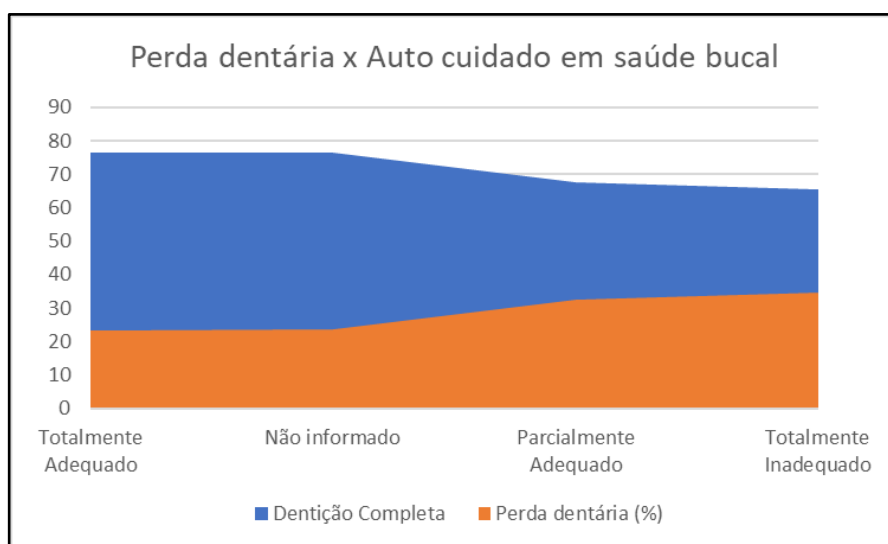


Figura 22 – Gráfico ilustrativo da relação entre qualidade de auto cuidado em saúde bucal e incidência de perda dentária.

Fonte: Elaborado pelo Autor (2023).

5.1.2.3 Condições econômicas

De acordo com os fatores socioeconômicos relatados, identificou-se predominância da perda dentária na classe com menor faixa de renda per capita, até 1/4 de salário-mínimo (34,19%), ao passo que, na classe com maior faixa de renda, mais de 5 salários-mínimos, houve menor registro de perda dentária (23,35%), como indicam a Tabela 13 e a Figura 23 em sequência.

FAIXA DE RENDA PER CAPITA	DENTIÇÃO COMPLETA	DENTES AUSENTES	PERDA DA DENTIÇÃO FUNCIONAL	EDENTULISMO
Mais de 5 salários-mínimos	76,66	17,93	4,81	0,61
Mais de 3 até 5 salários-mínimos	73,74	19,4	6,21	0,65
Mais de 2 até 3 salários-mínimos	71,57	20,1	7,36	0,97
Mais de 1 até 2 salários-mínimos	69,78	20,16	8,72	1,34
Mais de 1/4 até 1/2 salário-mínimo	68,28	21,69	9,15	0,88
Mais de 1/2 até 1 salário-mínimo	67,56	20,6	10,12	1,73
Até 1/4 de salário-mínimo	65,81	24,4	9,15	0,64

Tabela 13 – Relação entre perda dentária e faixa de renda per capita.
Fonte: Elaborado pelo Autor (2023).

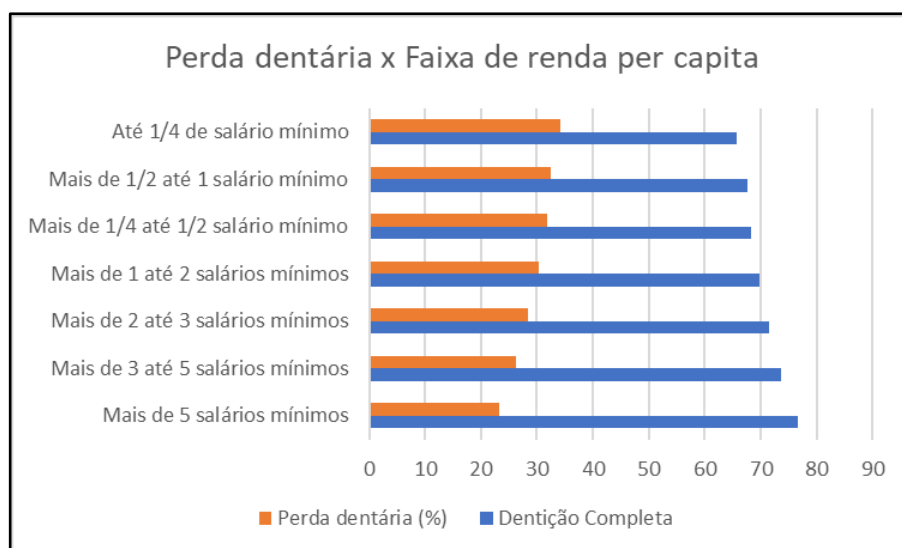


Figura 23 – Gráfico ilustrativo da relação entre perda dentária e faixa de renda per capita.

Fonte: Elaborado pelo Autor (2023).

Outro aspecto importante relativo à condição econômica é a ocupação (empregado/desempregado), esta sugere forte relação com a perda dentária, uma vez que os participantes da pesquisa que relataram estar desempregados acumularam maior perda (38,85%) do que os empregados (25,73%). Os dados mencionados são apresentados de forma mais detalhada na Tabela 14 e Figura 24.

EMPREGADO	DENTIÇÃO COMPLETA (%)	DENTES AUSENTES (%)	PERDA DA DENTIÇÃO FUNCIONAL (%)	EDENTULISMO (%)
Sim	74,27	19,27	5,98	0,48
Não	61,16	23	13,36	2,49

Tabela 14 – Relação entre perda dentária e condição de emprego.

Fonte: Elaborado pelo Autor (2023).

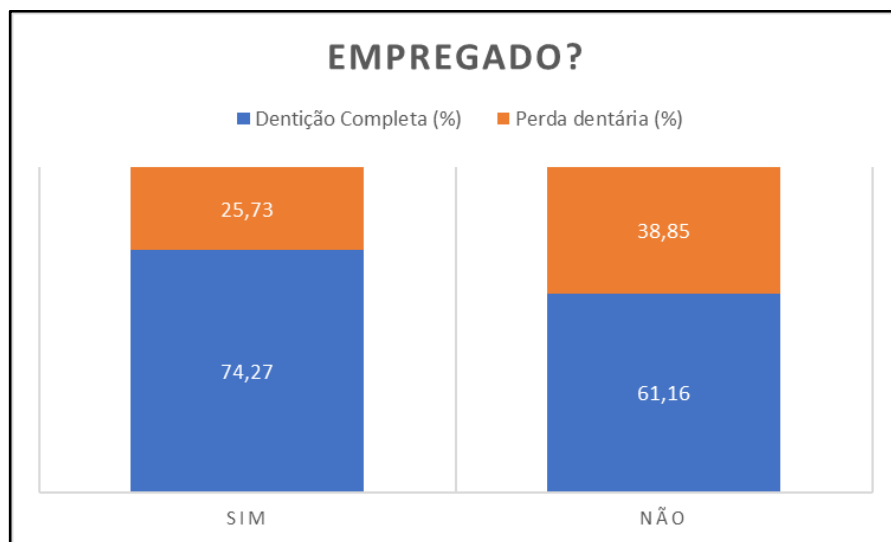


Figura 24 – Gráfico ilustrativo da relação entre perda dentária e condição de emprego.

Fonte: Elaborado pelo Autor (2023).

Por fim, ao analisar se o indivíduo é beneficiário de plano de saúde ou não, mais uma vez observa-se o impacto da desigualdade na dentição da população.

Entre os participantes da pesquisa que relataram não ter plano de saúde, 32,9% registram perda dentária, em contraste com 24,1% dos que indicaram ter plano de saúde (Tabela 15 e Figura 25).

BENEFICIÁRIO DE PLANO DE SAÚDE	DENTIÇÃO COMPLETA (%)	DENTES AUSENTES (%)	PERDA DA DENTIÇÃO FUNCIONAL (%)	EDENTULISMO (%)
Sim	75,91	17,86	5,54	0,7
Não	67,1	21,69	9,82	1,39

Tabela 15 – Relação entre perda dentária e benefício de plano de saúde.

Fonte: Elaborado pelo Autor (2023).

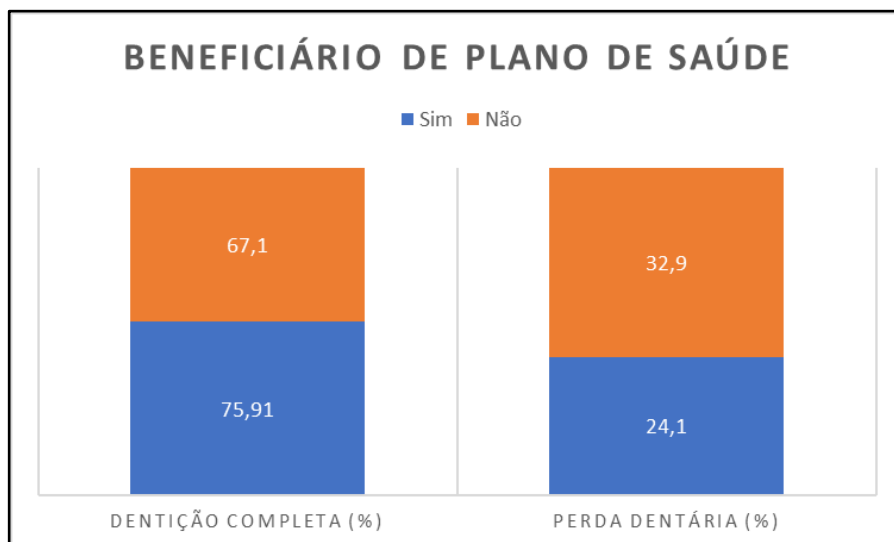


Figura 25 – Gráfico ilustrativo da relação entre perda dentária e benefício de plano de saúde.

Fonte: Elaborado pelo Autor (2023).

5.2 Análise dos Resultados

Para o desfecho edentulismo, o modelo treinado com Light Gradient boosting Machine (LightGBM) obteve o melhor desempenho, apresentando AUC ROC de 89,2%, sensibilidade de 80,4% e especificidade de 81,6%. Esse resultado indica que o modelo de classificação possui alta capacidade de distinguir corretamente entre as classes, ou seja, alta taxa de verdadeiros positivos (sensibilidade) e uma baixa taxa de falsos positivos (especificidade).

Em sequência, o modelo Regressão Logística, Floresta Aleatória (Random Forest), *Extreme Gradient boosting Trees* (XGBoost) e Redes Neurais MultiLayer Perceptron (MLP) também foram avaliados. A Regressão Logística e o XGBoost obtiveram resultados próximos ao do modelo de LightGBM, enquanto *Random Forest* e MLP apresentaram desempenhos inferiores em relação à métrica sensibilidade se comparados aos anteriores.

O modelo do algoritmo Redes Neurais MultiLayer Perceptron (MLP), por exemplo, alcançou uma AUC ROC de 87.1%, uma sensibilidade de 52,6% e uma especificidade de 90,2%. Esses valores indicam uma capacidade limitada de distinguir corretamente entre as classes, com uma sensibilidade relativamente baixa e uma especificidade alta (Tabela 16).

Portanto, com base nos resultados obtidos, o modelo de Light Gradient boosting Machine (LightGBM) se destaca como o mais adequado para lidar com o problema de classificação de edentulismo.

A Figura 26 apresenta o resultado de execução dos algoritmos sob o conjunto de teste.

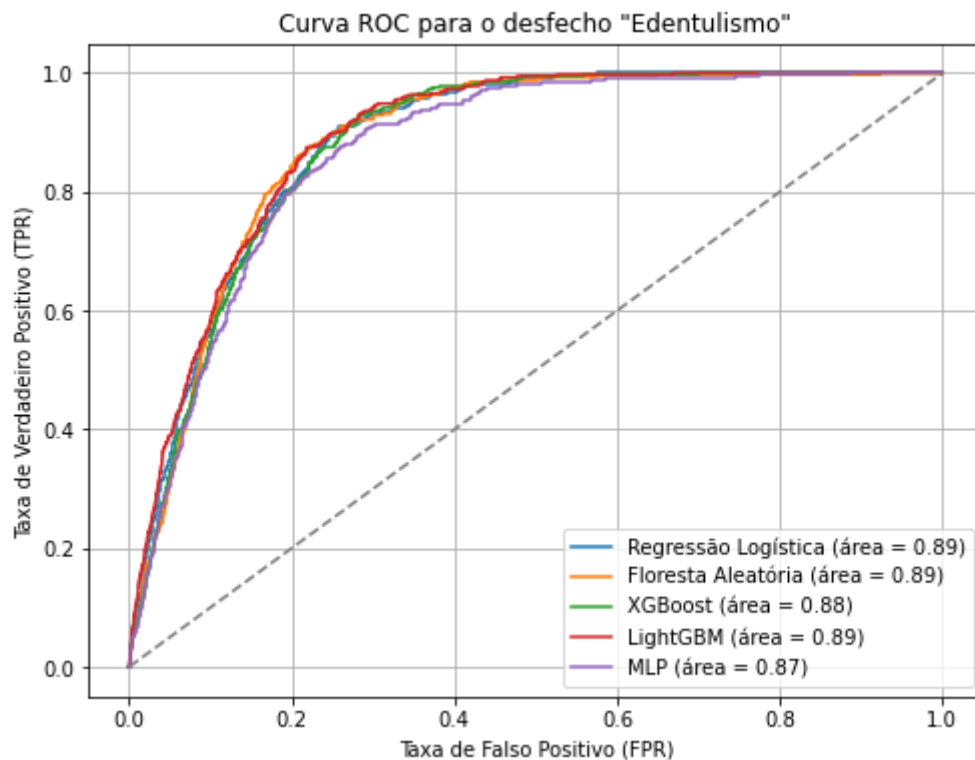


Figura 26 – Gráfico da Curva ROC de desempenho dos cinco algoritmos para o desfecho "Edentulismo".

Fonte: Elaborado pelo Autor (2023).

ALGORITMO	SENSIBILIDADE	ESPECIFICIDADE	AUC ROC
Light Gradient boosting Machine (LightGBM)	80,4	81,6	89,2
Regressão Logística	83,1	78,7	88,6
Floresta Aleatória (Random Forest)	57,7	90,1	88,5
Extreme Gradient boosting Trees (XGBoost)	73,6	83,8	88,2
Redes Neurais MultiLayer Perceptron (MLP)	52,6	90,2	87,1

Tabela 16 – Resultados obtidos para o desfecho edentulismo.

Fonte: Elaborado pelo Autor (2023).

Na avaliação do desfecho perda da dentição funcional, classe com proporção de 8,6% instâncias no *conjunto de dados*, o melhor desempenho foi registrado pelo algoritmo LightGBM, seguido pelo XGBoost. Os algoritmos MLP, *Random Forest* e Regressão Logística apresentaram desempenho inferior aos anteriores.

O algoritmo LightGBM, com registro de melhor desempenho, obteve um bom resultado. Sua AUC ROC foi de 79,1%, indicando que o modelo possui uma capacidade moderada de distinguir corretamente entre as classes. Além disso, a sensibilidade do modelo foi de 85,8%, o que significa que ele é capaz de identificar corretamente 85,8% dos casos positivos. Por outro lado, a especificidade foi de 61,9%, o que indica que o modelo possui uma taxa relativamente alta de falsos positivos (38,1%).

Para este desfecho (perda da dentição funcional), o algoritmo Redes Neurais MultiLayer Perceptron (MLP) foi o que alcançou o pior resultado de desempenho com AUC ROC de 73,8%, Sensibilidade de 27,5% e Especificidade de 89,2%, como mostra a Tabela 17.

A Figura 27 apresenta os resultados dos modelos quando aplicados ao conjunto de teste.

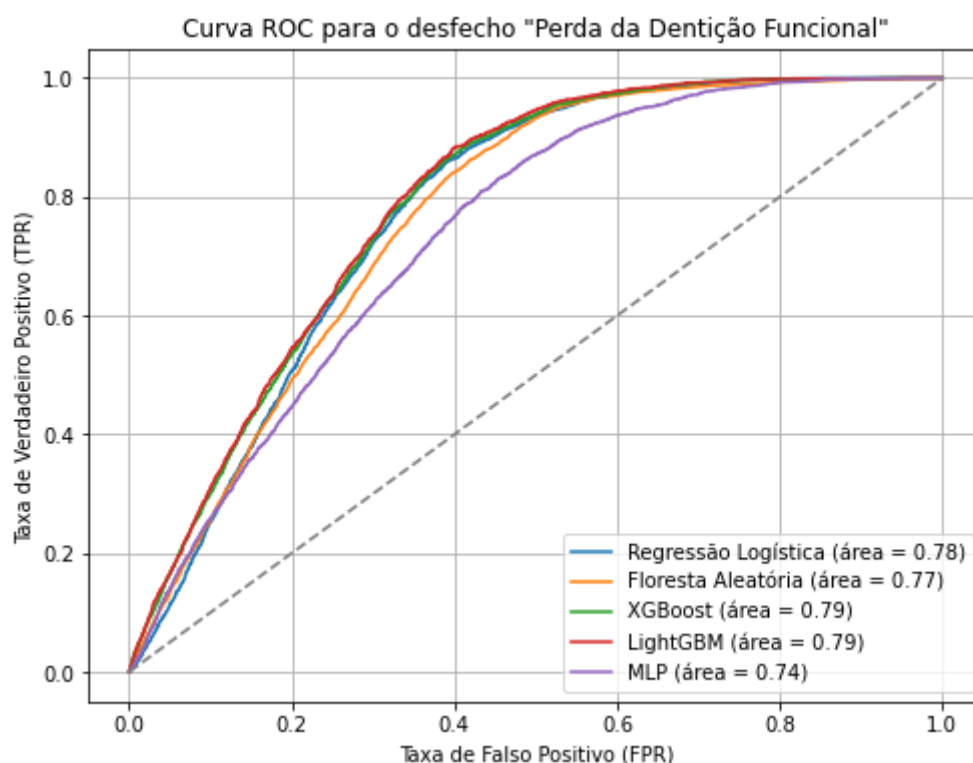


Figura 27 – Gráfico da Curva ROC de desempenho dos cinco algoritmos para o desfecho "Perda da Dentição Funcional".

Fonte: Elaborado pelo Autor (2023).

ALGORITMO	SENSIBILIDADE	ESPECIFICIDADE	AUC ROC
Light <i>Gradient boosting</i> Machine (LightGBM)	85,8	61,9	79,1
<i>Extreme Gradient boosting Trees</i> (XGBoost)	82,5	63,9	78,7
Regressão Logística	73,6	69,2	77,5
Floresta Aleatória (<i>Random Forest</i>)	0,8	99,7	76,7
Redes Neurais MultiLayer Perceptron (MLP)	27,5	89,2	73,8

Tabela 17 – Resultados obtidos para o desfecho perda da dentição funcional.
Fonte: Elaborado pelo Autor (2023).

No terceiro desfecho analisado, **ausência de dentes ou perda de um ou mais dentes**, classe que apresenta maior número de instâncias entre os possíveis desfechos (20.59%), o algoritmo XGBoost foi o que melhor previu os desfechos, seguido por LightGBM e MLP. Floresta Aleatória e Regressão Logística apresentaram resultados inferiores.

Nesse contexto, o modelo XGBoost demonstrou um desempenho interessante. Obteve uma AUC ROC de 72,0%, o que indica que o modelo possui uma capacidade moderada de distinguir entre as classes.

A sensibilidade alcançou um valor elevado de 94,1%. No entanto, a especificidade sugere uma taxa relativamente alta de falsos positivos (33,7%).

O LightGBM registrou resultado semelhante ao algoritmo anterior, com valor próximo de AUC ROC (71,6%) e pequena variação nos resultados de sensibilidade (91,6%) e especificidade (38,3%).

O algoritmo de Regressão Logística apresentou baixo desempenho preditivo para o desfecho ausência de dentes se comparado aos outros. Com AUC ROC de 66,4%, capacidade razoável de distinguir entre as classes; sensibilidade de 64,8%, sendo capaz de identificar corretamente a maior parte dos casos positivos em relação à variável-alvo; e especificidade 59,6%, taxa relativamente baixa de acertos na detecção dos casos negativos (taxa considerável de falsos positivos), como mostra a Tabela 18 a seguir.

A Figura 28 apresenta o resultado de execução dos algoritmos sob o conjunto de teste.

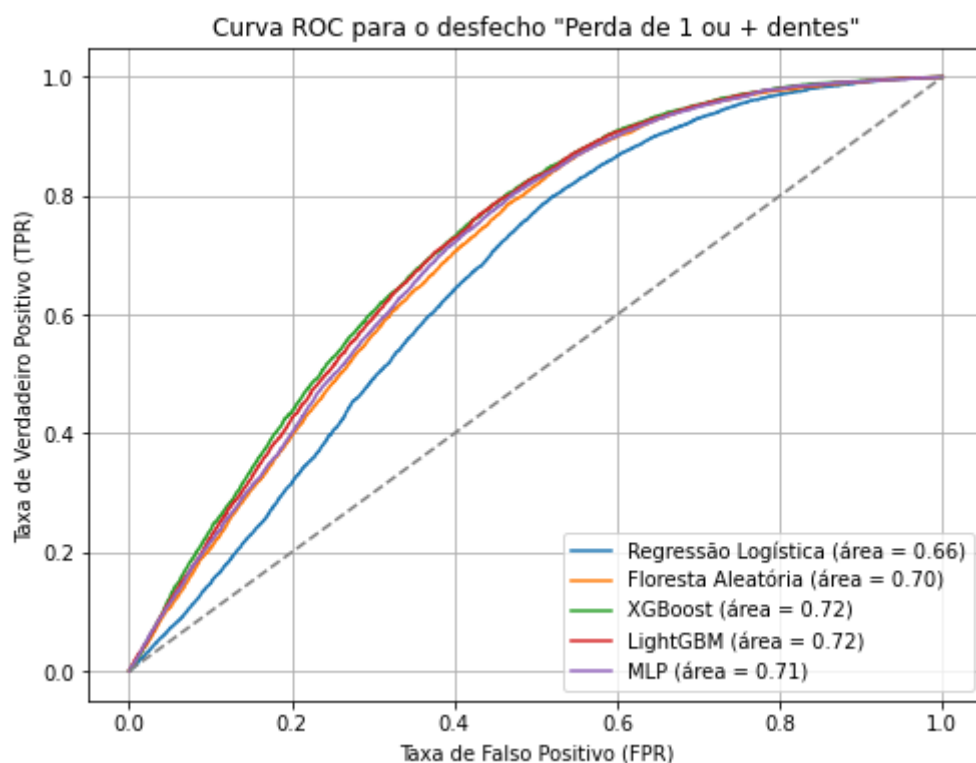


Figura 28 – Gráfico da Curva ROC de desempenho dos cinco algoritmos para o desfecho “Ausência de Dentes ou Perda de Um ou Mais Dentes”.

Fonte: Elaborado pelo Autor (2023).

ALGORITMO	SENSIBILIDADE	ESPECIFICIDADE	AUC ROC
<i>Extreme Gradient boosting Trees</i> (XGBoost)	94,1	33,7	72
<i>Light Gradient boosting Machine</i> (LightGBM)	91,6	38,3	71,6
Redes Neurais MultiLayer Perceptron (MLP)	64,2	65,6	70,9
Floresta Aleatória (<i>Random Forest</i>)	27,4	86,7	70,3
Regressão Logística	64,8	59,6	66,4

Tabela 18 – Resultados obtidos para o desfecho ausência de dentes ou perda de um ou mais dentes.

Fonte: Elaborado pelo Autor (2023).

Por meio do presente resultado, é possível determinar que o algoritmo LightGBM foi o que proporcionou o melhor desempenho entre os cinco. Melhor resultado para classificação para os desfechos Edentulismo e Perda da Dentição Funcional e segundo melhor para o desfecho Ausência de Dentes. O XGBoost também se apresentou uma boa alternativa, já que alcançou resultados semelhantes ao anterior, sendo o primeiro para classificação de Ausência de Dentes, segundo para

Perda de Dentição Funcional e, apesar de quarto para Edentulismo, apresentou boa taxa de sensibilidade.

A Regressão Logística registrou o segundo melhor desempenho para classificação de edentulismo, no entanto, obteve o pior resultado entre os cinco para o desfecho Ausência de Dentes. De acordo com Borisov et al. (2022), rede neural, já conhecida por sua “dificuldade” em lidar com dados tabulares, obteve baixo desempenho na classificação de perda de Edentulismo e Perda da Dentição Funcional e razoável para Ausência de Dentes.

Por fim, o algoritmo de *Random Forest* apesar de registrar razoável AUC ROC em geral, apresentou baixo desempenho para métrica de sensibilidade em todos os casos, indicando dificuldade para identificar corretamente casos positivos em relação à variável-alvo.

A partir das considerações acima, a título de ilustração pode-se ranquear os algoritmos da seguinte forma:

- 1º *Light Gradient boosting Machine* (LightGBM);
- 2º *Extreme Gradient boosting Trees* (XGBoost);
- 3º Regressão Logística;
- 4º Floresta Aleatória (*Random Forest*);
- 5º Redes Neurais *MultiLayer Perceptron* (MLP);

Apresenta-se na Figura 29 uma tabela consolidada com o desempenho dos cinco algoritmos em cada desfecho. Vale ressaltar que a tabela completa, contendo uma análise mais abrangente, pode ser encontrada no Apêndice B.

	Sensibilidade (Recall)	Especificidade	Precisão	F1-Score	AUC ROC
Edentulismo					
<u>Classificador</u>					
Regressão Logística	83,1	78,7	4,5	8,5	88,6
Floresta Aleatória (Random Forest)	57,7	90,1	6,6	11,8	88,5
Extreme Gradient Boosting Trees (XGBoost)	73,6	83,8	5,2	9,7	88,2
Light Gradient Boosting Machine (LightGBM)	80,4	81,6	5	9,4	89,2
Redes Neurais MultiLayer Perceptron (MLP)	52,6	90,2	6,1	10,9	87,1
Perda funcional dentária					
<u>Classificador</u>					
Regressão Logística	73,6	69,2	18,3	29,4	77,5
Floresta Aleatória (Random Forest)	0,8	99,7	18,9	1,6	76,7
Extreme Gradient Boosting Trees (XGBoost)	82,5	63,9	17,7	29,2	78,7
Light Gradient Boosting Machine (LightGBM)	85,8	61,9	17,5	29,1	79,1
Redes Neurais MultiLayer Perceptron (MLP)	27,5	89,2	19,3	22,7	73,8
Ausência de dentes (perda de um ou mais dentes)					
<u>Classificador</u>					
Regressão Logística	64,8	59,6	29,4	40,4	66,4
Floresta Aleatória (Random Forest)	27,4	86,7	34,8	30,7	70,3
Extreme Gradient Boosting Trees (XGBoost)	94,1	33,7	26,9	41,8	72
Light Gradient Boosting Machine (LightGBM)	91,6	38,3	27,8	42,6	71,6
Redes Neurais MultiLayer Perceptron (MLP)	64,2	65,6	32,6	43,3	70,9

Figura 29 – Desempenho dos cinco algoritmos em cada desfecho.

Fonte: Elaborado pelo Autor (2023).

5.2.1 Importância das Variáveis (Feature Importance)

Em uma aplicação da área de saúde, a transparência é de grande importância para os profissionais e, também, para o paciente. Para que um diagnóstico seja confiável e as decisões de tratamento sejam melhor embasadas, é necessário possibilitar a compreensão dos motivos que levaram o software a realizar determinado diagnóstico.

A avaliação de importância das variáveis preditoras revela a contribuição relativa de cada uma no processo de tomada de decisão do modelo. Para interpretar o processo de tomada de decisão dos algoritmos, calculou-se os valores de Shapley para cada desfecho, selecionando os dois algoritmos que obtiveram melhor desempenho na tarefa de classificação a partir das métricas AUC ROC, Sensibilidade e Especificidade.

As figuras a seguir apresentam a importância de cada variável, para as cinco mais discriminantes para o desfecho de saída. Estas foram extraídas a partir dos dois algoritmos de melhor desempenho em cada desfecho.

Edentulismo

LightGBM

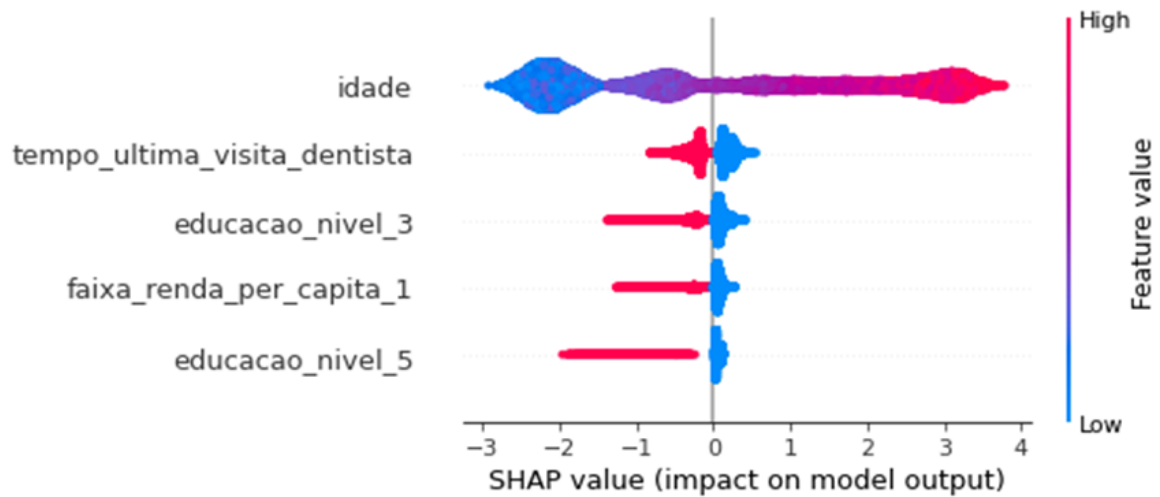


Figura 30 – Os cinco principais preditores do desfecho "Edentulismo", com o modelo gerado a partir do algoritmo LightGBM.

Fonte: Elaborado pelo Autor (2023).

Regressão Logística

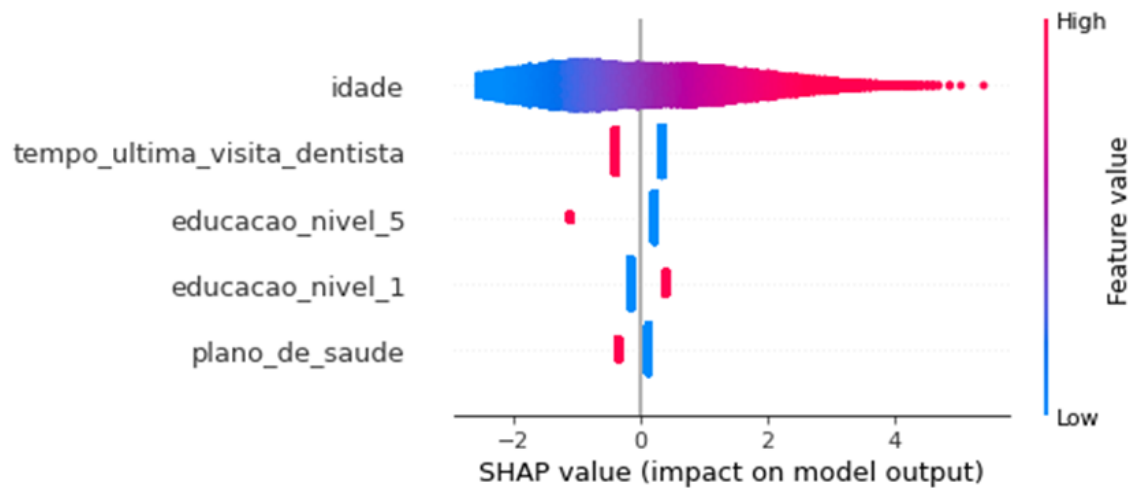


Figura 31 – Os cinco principais preditores do desfecho "Edentulismo", com o modelo gerado a partir do algoritmo Regressão Logística.

Fonte: Elaborado pelo Autor (2023).

Perda da Dentição Funcional

LightGBM

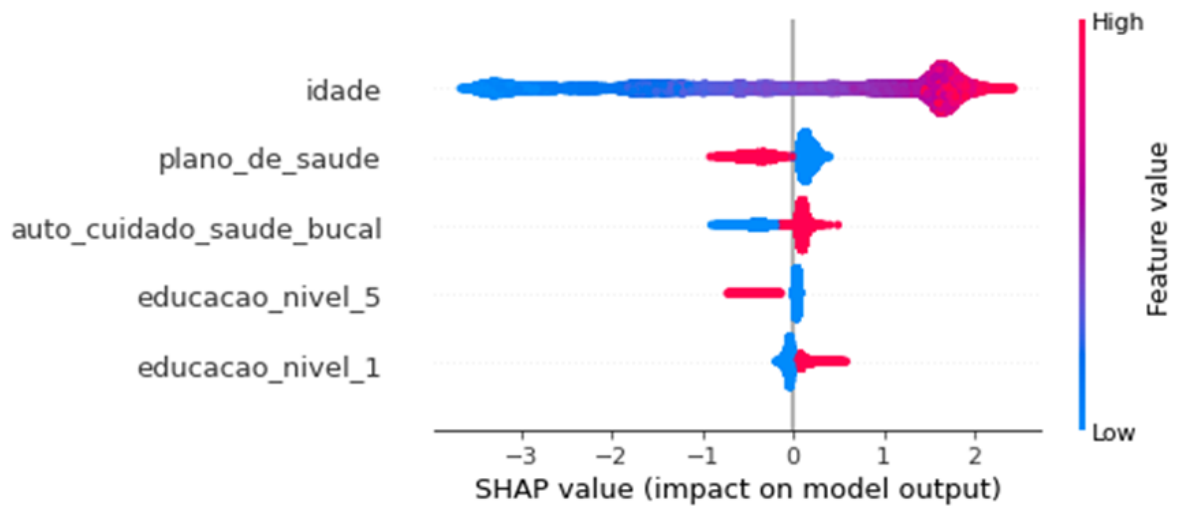


Figura 32 – Os cinco principais preditores do desfecho "Perda da Dentição Funcional", com o modelo gerado a partir do algoritmo LightGBM.

Fonte: Elaborado pelo Autor (2023).

XGBoost

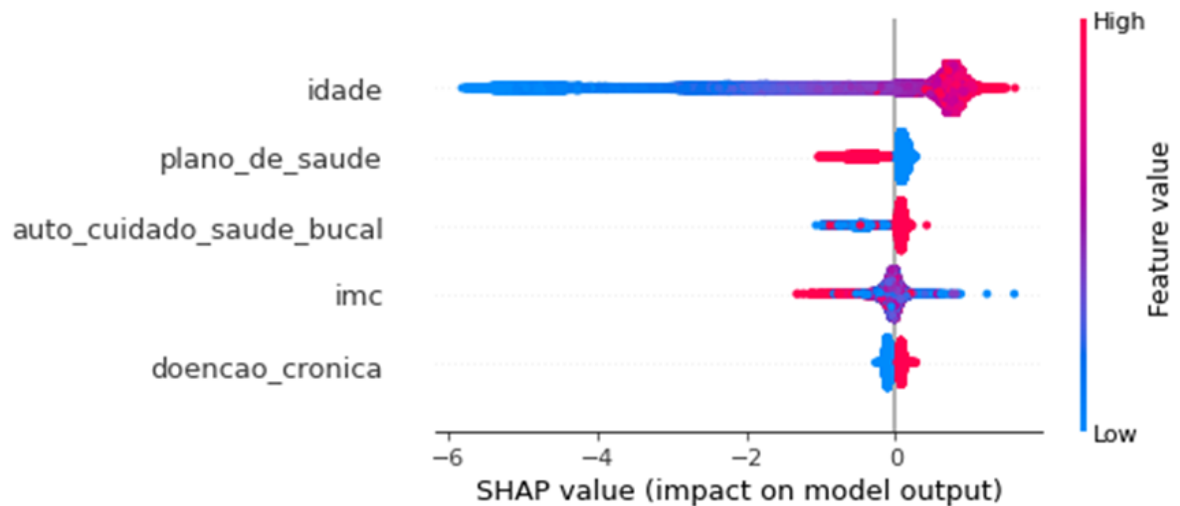


Figura 33 – Os cinco principais preditores do desfecho "Perda da Dentição Funcional", com o modelo gerado a partir do algoritmo XGBoost.

Fonte: Elaborado pelo Autor (2023).

Ausência de Dentes (perda de um ou mais dentes)

XGBoost

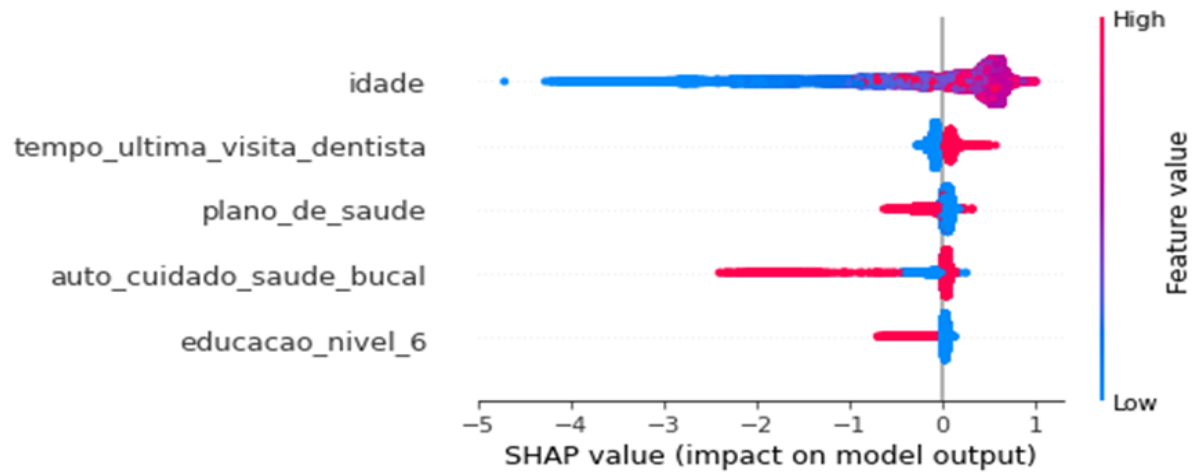


Figura 34 – Os cinco principais preditores do desfecho "Ausência de dentes ou perda de um ou mais dentes", com o modelo gerado a partir do algoritmo XGBoost.

Fonte: Elaborado pelo Autor (2023).

LightGBM

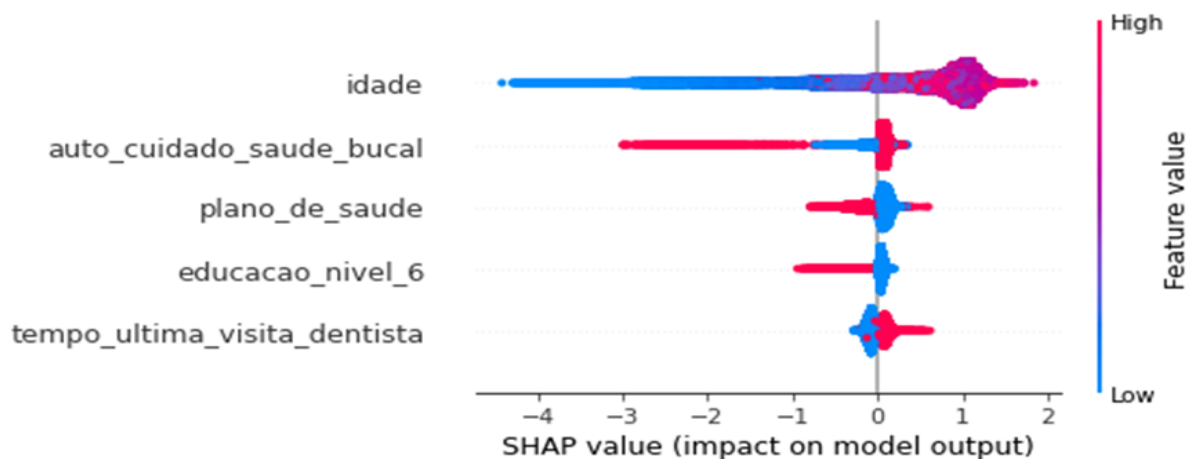


Figura 35 – Os cinco principais preditores do desfecho "Ausência de dentes ou perda de um ou mais dentes", com o modelo gerado a partir do algoritmo LightGBM.

Fonte: Elaborado pelo Autor (2023).

Conforme observado, a “idade” é a variável mais importante no processo de decisão de todos os modelos para todos os desfechos. De maneira geral, observou-se que os indivíduos com maior idade tiveram mais probabilidade de sofrer de perda dentária.

A variável “tempo_ultima_visita_dentista” foi a segunda mais apontada a nível de importância para predição e reforça que pessoas que tiveram acesso a consulta dentária em um período inferior ou igual a um ano tiveram probabilidade reduzida de sofrer de perda dentária em comparação a quem não teve.

Como será citado em momento posterior, as variáveis preditoras “idade” e “tempo_ultima_visita_dentista” são conhecidas na área como indicadores mais significativos para determinar um desfecho de saúde bucal.

Nível de educação (variáveis com prefixo “educacao_nivel_”) e Plano de saúde (“plano_de_saude”) também foram listadas entre os mais importantes preditores.

Na avaliação do desfecho edentulismo, a regressão logística avaliou o nível de educação “Sem Instrução” como um fator de aumento de probabilidade para a total ausência de dentes, ao passo que possuir o ensino médio completo ou equivalente afasta essa possibilidade.

Avaliação semelhante foi identificada pelo modelo gerado a partir do algoritmo LightGBM, que, para o mesmo desfecho, elencou os níveis de educação Ensino Médio Completo (novamente) e Fundamental Completo como atenuantes, destacando o grau de importância desta variável, uma vez que, como foi observado, à medida que aumenta o nível de instrução, diminui a probabilidade de edentulismo.

Na avaliação do desfecho “Perda da dentição funcional”, o algoritmo LGBM (melhor performance para esse resultado) sugeriu a relação do nível de educação “Sem Instrução” com o aumento do desfecho, ao passo que possuir nível de instrução “Médio Completo” reduz a incidência. O segundo algoritmo, XGBoost, não apresentou o nível de educação entre os cinco maiores preditores; no entanto, registrou a mesma relação que o algoritmo anterior, posicionando tais atributos (“Sem Instrução” e “Ensino Médio Completo”) nas posições sexta e sétima de importância.

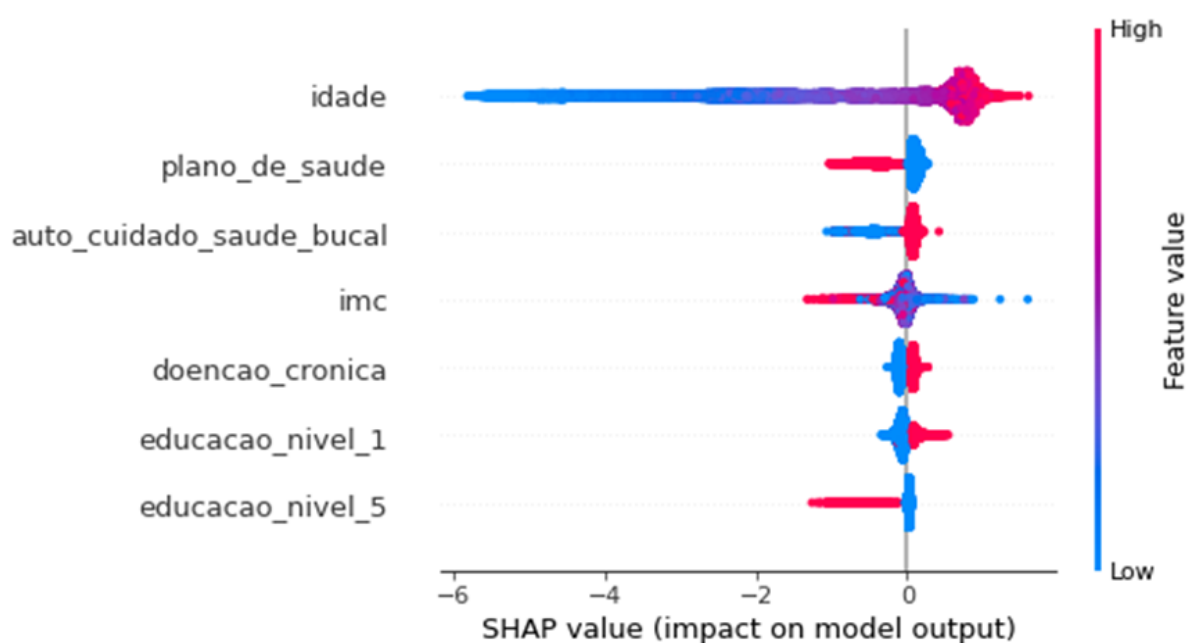


Figura 36 – Os sete principais preditores do desfecho "Perda da Dentição Funcional", com o modelo gerado a partir do algoritmo XGBoost.

Fonte: Elaborado pelo Autor (2023).

Para o desfecho “Dentes Ausentes” (Perda de um ou mais dentes), o nível de instrução “Superior Incompleto” esteve entre os cinco atributos mais relevantes para apontar menor propensão à perda dentária.

Assim como os atributos anteriores, possuir ou não plano de saúde também teve forte impacto na saída dos modelos e esteve na lista dos cinco preditores mais importantes.

Em todos os casos, o indivíduo que possui um plano de saúde tende à menor perda dentária.

As variáveis faixa de renda (prefixo “faixa_renda_per_capita_”) “até ¼ de salário mínimo”, “auto_cuidado_saude_bucal”, “IMC” e doença crônica (“doenca_cronica”) também apareceram em, pelo menos, um dos algoritmos na avaliação de um dos desfechos.

No modelo gerado pelo algoritmo LightGBM, a faixa de renda “até ¼ de salário mínimo” indicou menor chance para o desfecho edentulismo, em que pese esta faixa tenha apresentado maior perda dentária acumulada.

A variável “auto_cuidado_saude_bucal”, nos algoritmos LGBM e XGBoost, afastou a possibilidade do desfecho “Ausência de Dentes” (perda de um ou mais dentes).

Observou-se, com o algoritmo XGBoost, que a presença de uma ou mais doenças crônicas aumentam a incidência do desfecho “Perda da dentição funcional”.

Por fim, o algoritmo XGBoost também registrou um maior impacto da variável IMC, à medida que seu valor aumenta para o desfecho “Perda da dentição funcional”.

5.3 DISCUSSÃO

5.3.1 Discussão dos Resultados

Conforme o conhecimento disponível, este estudo é um dos primeiros a explorar a aplicação de técnicas de Inteligência Artificial (Aprendizado de Máquina) para avaliar a perda dentária completa e incremental em adultos utilizando dados de uma base de dados de contexto brasileiro. Os principais pontos desse estudo incluem: seleção de atributos (baseado na literatura), técnicas de pré-processamento, definição de cinco algoritmos e respectivos ajustes para geração e avaliação de modelos preditores.

Foram utilizadas para avaliação de desempenho dos modelos as métricas: AUC ROC, sensibilidade (recall), especificidade, precisão e F1-Score. Conforme explicitado na seção “Análise dos Resultados”, o algoritmo LightGBM foi o que proporcionou o melhor desempenho entre os cinco: melhor resultado para classificação dos desfechos Edentulismo e Perda da Dentição Funcional e segundo melhor para o desfecho Ausência de Dentes.

Por outro lado, Cooray et al. (2021) buscaram compreender o comportamento das variáveis preditoras de perda dentária em idosos japoneses maiores de 65 anos e obtiveram bom rendimento com o algoritmo XGBoost, que apresentou acurácia de 90.5%. O mesmo ocorreu na pesquisa de Elani et al. (2021) que objetivou prever a perda dentária completa e incremental entre adultos estadunidenses e comparar o desempenho preditivo desses modelos. Foram testados cinco algoritmos: XGBoost, LightGBM, floresta aleatória, regressão logística, rede neural MLP. Nesse caso, a principal métrica adotada foi a mesma do presente estudo, AUC ROC, que obteve o seguinte resultado separado por desfecho: edentulismo (AUC = 88,7%), ausência de dentição funcional (AUC = 88,3%) e falta de qualquer dente (AUC = 83,2%).

Sob outra perspectiva, a pesquisa de Silva et al. (2023), que objetivou prever o risco de morte em pacientes com câncer, avaliou seis algoritmos diferentes de aprendizado de máquina: CatBoost, XGBoost, LightGBM, classificador de aumento de gradiente (*gradient boosting classifier*), floresta aleatória e regressão logística; igualmente adotando como métrica principal o AUC ROC, identificou o algoritmo CatBoost como o de melhor desempenho, com resultado 94.6%.

Em síntese, a presente pesquisa buscou identificar a capacidade de algoritmos de IA para determinar a possibilidade de perda dentária a partir de três desfechos. Foram utilizadas informações fornecidas por meio dos inquéritos PNS (2013 e 2019) no que tange a condições socioeconômicas, acesso a atendimento médico-odontológico, rotina de cuidados bucais e coexistência de doenças crônicas. Tal circunstância sugere a relevância de conhecer sexo, nível de escolaridade e renda.

Com isso, uma vez que medição de condições clínicas de saúde bucal e a dependência do diagnóstico profissional se mostram limitadas diante do desafio do planejamento de políticas públicas no campo de Odontologia (PINTO, 2013), o estudo visou ao desenvolvimento de modelos preditivos utilizando características que independem de exame odontológico.

A Hipótese de pesquisa de que seria possível implementar modelos de aprendizado de máquina com grande taxa de acerto a partir de dados socioeconômicos é reforçada pelos estudos (ANDRADE et al., 2019) que sugerem associação entre a dentição funcional e escolaridade, sexo e cor/gênero (ELANI et al., 2021); e que realizou experimento semelhante, avaliando aspectos socioeconômicos com aprendizado de máquina a partir de dados do inquérito de origem americana (U.S) NHANES. Este último serviu de referência para o desenvolvimento do experimento desta dissertação.

O estudo tende a confirmar a literatura (ROBERTO et al., 2019, FERREIRA et al., 2020b, MATSUYAMA; JÜRGES; LISTL, 2019, ELANI et al., 2021) que identifica idade, escolaridade e condições socioeconômicas para os desfechos já mencionados. Por outro lado, não se obteve informações sobre a relevância das condições médicas preexistentes como determinante de perda dentária.

Mostra-se evidente a necessidade de estudos futuros com coortes longitudinais com maior separação de tempo entre os conjuntos de dados de treinamento e testes para avaliar a validade externa dos modelos aqui propostos e confirmar sua

capacidade de inferência. Ademais, foram excluídas instâncias de dados com valores indefinidos para variáveis selecionadas para o estudo.

Por fim, mesmo com o desempenho AUC de 89,2%, cabe informar que o desfecho edentulismo, presente em apenas 1,19% das amostras, configura-se como “evento raro” e, por esse motivo, pode ter afetado o desempenho dos modelos.

As conclusões não permitem estabelecer relação de causalidade (inferência); por outro lado, indicam a possibilidade de verificar efeito previsível (predição). Ainda sim, podem fornecer subsídios para que estudos futuros utilizem aprendizado de máquina no intuito de identificar quais nichos populacionais estão expostos a outros desfechos odontológicos.

O presente estudo, que utiliza modelos baseados em aprendizado de máquina, demonstra sua relevância na medida em que existem limitações estatísticas nos modelos multivariáveis convencionais, pois as técnicas paramétricas não utilizam dados multidimensionais com associações complexas, limitando assim a capacidade preditiva dos modelos.

Em contraposição, os modelos baseados em aprendizado de máquina podem fazer previsões com maior precisão, considerando uma gama mais ampla de fatores sem suposições estritas sobre os preditores e o resultado.

5.3.2 Comparação com Estudo referência

A partir dos dados obtidos neste estudo, torna-se possível tecer comparações, realizando uma breve análise entre os resultados encontrados neste e na análise primária realizada no estudo de referência (artigo “Predictors of tooth loss: A *Machine Learning* approach”, de ELANI et al., 2021).

• Ponto 1: conjunto de dados

Para o desenvolvimento desta dissertação, foram utilizados *conjuntos de dados* oriundos dos inquéritos do PNS de 2013 e 2019, enquanto no estudo de referência foram utilizados os *conjuntos de dados* NHANES dos ciclos de 2011-2012 e 2013-2014.

Os dados de treino e teste neste estudo foram obtidos com a junção dos *conjuntos de dados* de 2013 e 2019 e distribuídos a partir do recurso Split, de forma estratificada, reservando 70% da massa para treino e outros 30% para teste. No artigo

de referência, utilizou-se o conjunto de dados do ciclo 2011-2012 para treino e do ciclo 2013-2014 para teste.

A base do PNS apresenta a seguinte proporção dos desfechos estudados em relação ao total de amostras: Edentulismo 1,19%; Perda da dentição funcional 8,6% e Ausência de dentes 20,59%. Juntos os dois ciclos do NHANES apresentam os seguintes números: Edentulismo 6%; Perda da dentição funcional 22% e Ausência de dentes 58%, como mostra a Tabela 19 a seguir.

	TOTAL DE AMOSTRAS AVALIADAS	EDENTULISMO	PERDA DA DENTIÇÃO FUNCIONAL	AUSÊNCIA DE DENTES
PNS	114.279	1,19%	8,6%	20,59%
NHANES	11.977	5,3%	18,5%	58,3%

Tabela 19 – Proporção dos desfechos PNS e NHANES.

Fonte: Elaborado pelo Autor (2023).

• Ponto 2: Pré-processamento

Nesta seção serão pontuadas as diferenças entre os experimentos da presente dissertação e do estudo de referência.

Durante o processo de definição das variáveis de estudo, para o experimento desta dissertação, foram criadas duas variáveis derivadas a fim de melhorar o modelo: `auto_cuidado_saude_bucal` e `doenca_cronica`. Deste modo, foram substituídas do modelo as variáveis de condições médicas crônicas (asma, artrite, derrame, ataque_cardiaco, diabetes, colesterol_alto e hipertensao) pela última supracitada.

No experimento desta dissertação foi aplicada a técnica de reamostragem do tipo *oversample* com o algoritmo *Synthetic Minority Over-sampling Technique* (SMOTE) para o treino do algoritmo MLP em todos os desfechos, enquanto para o experimento do artigo referência, o autor informou ter utilizado a técnica do tipo *undersample*, algoritmo One-Sided Selection apenas para o desfecho edentulismo para o mesmo fim.

• Ponto 3: Algoritmos

Para implementação dos modelos, foram utilizados os cinco algoritmos em ambos os estudos: Regressão Logística, Floresta Aleatória (*Random Forest*), *Extreme Gradient boosting Trees* (XGBoost), *Light Gradient boosting Machine* (LightGBM) e Redes Neurais *MultiLayer Perceptron* (MLP).

• Ponto 4: Resultados

A seguir será realizada uma análise comparativa entre os resultados obtidos nos dois estudos a partir das métricas AUC ROC, Sensibilidade e Especificidade.

Em cada tabela, foram designados os títulos PNS para indicar os resultados obtidos no experimento desta dissertação e NHANES para os valores colhidos no artigo referência.

Além disso, especificou-se a coluna “Ranking”, que indica em uma escala de 1 a 5, qual algoritmo obteve o melhor desempenho dentro do estudo anunciado na subcoluna (PNS ou NHANES).

Para o desfecho “Edentulismo”, neste estudo, o algoritmo de Light Gradient boosting Machine (LightGBM) foi o que obteve o melhor desempenho (AUC ROC: 89,2% , Sensibilidade: 80,4% e Especificidade: 81,6%). Na pesquisa de referência, segundo os autores, o algoritmo XGBoost foi quem atingiu o melhor resultado (AUC ROC: 88,7%, Sensibilidade: 74,3% e Especificidade: 84,5%). A seguir encontra-se a Tabela 20 comparativa de desempenho entre os estudos para o desfecho Edentulismo:

Algoritmo	EDENTULISMO							
	AUC ROC		Sensibilidade		Especificidade		Ranking	
	PNS	NHANES	PNS	NHANES	PNS	NHANES	PNS	NHANES
Light Gradient boosting Machine (LightGBM)	89,2	88,4	80,4	76,4	81,6	83,1	1º	3º
Regressão Logística	88,6	86,5	83,1	71,9	78,7	84,6	2º	5º
Floresta Aleatória (Random Forest)	88,5	88,5	57,7	73,7	90,1	85,1	3º	2º
Extreme Gradient boosting Trees (XGBoost)	88,2	88,7	73,6	74,3	83,8	84,5	4º	1º
Redes Neurais MultiLayer Perceptron (MLP)	87,1	87,7	52,6	78,5	90,2	82,5	5º	4º

Tabela 20 – Comparativos de desempenho entre os resultados obtidos nesse estudo e os obtidos no estudo de referência para o desfecho edentulismo.

Fonte: Elaborado pelo Autor (2023).

Com o desfecho “Perda de dentição funcional”, o algoritmo LightGBM alcançou o melhor resultado (AUC ROC: 79,1%, Sensibilidade: 85,8% e Especificidade: 61,9%). Para a pesquisa de referência, o algoritmo MLP teve a melhor avaliação (AUC ROC: 88,3%, Sensibilidade: 74,1% e Especificidade: 84,2%). Abaixo a Tabela 21 mostra o comparativo de desempenho entre os estudos para o desfecho perda de dentição funcional.

PERDA DA DENTIÇÃO FUNCIONAL								
Algoritmo	AUC ROC		Sensibilidade		Especificidade		Ranking	
	PNS	NHANES	PNS	NHANES	PNS	NHANES	PNS	NHANES
<i>Light Gradient boosting Machine (LightGBM)</i>	79,1	87,2	85,8	53,9	61,9	92,1	1º	5º
<i>Extreme Gradient boosting Trees (XGBoost)</i>	78,7	87,7	82,5	58,0	63,9	91,3	2º	3º
Regressão Logística	77,5	87,6	73,6	48,4	69,2	93,7	3º	4º
Floresta Aleatória (<i>Random Forest</i>)	76,7	88,1	0,8	56,4	99,7	92,0	4º	2º
Redes Neurais <i>MultiLayer Perceptron (MLP)</i>	73,8	88,3	27,5	74,1	89,2	84,2	5º	1º

Tabela 21 – Comparativos de desempenho entre os resultados obtidos nesse estudo e os obtidos no estudo de referência para o desfecho perda da dentição funcional.

Fonte: Elaborado pelo Autor (2023).

No desfecho “Ausência de dentes” (perda de um ou mais dentes), foi observado no presente estudo que o algoritmo Extreme Gradient boosting Trees (XGBoost) registrou o melhor desempenho (AUC ROC: 72,0%, Sensibilidade: 94,1% e Especificidade: 33,7%). Na pesquisa de referência, o mesmo algoritmo obteve melhor resultado (AUC ROC: 83,2%, Sensibilidade: 95,9% e Especificidade: 29,6%). Abaixo pode-se observar a Tabela 22 comparativa de desempenho entre os estudos para o desfecho “Ausência de dentes”:

AUSÊNCIA DE DENTES								
Algoritmo	AUC ROC		Sensibilidade		Especificidade		Ranking	
	PNS	NHANES	PNS	NHANES	PNS	NHANES	PNS	NHANES
<i>Extreme Gradient boosting Trees</i> (XGBoost)	72,0	83,2	94,1	95,9	33,7	29,6	1º	1º
<i>Light Gradient boosting Machine</i> (LightGBM)	71,6	81,9	91,6	93,7	38,3	33,8	2º	5º
Redes Neurais MultiLayer Perceptron (MLP)	70,9	83,1	64,2	85,8	65,6	59,7	3º	2º
Floresta Aleatória (<i>Random Forest</i>)	70,3	82,7	27,4	89,5	86,7	55,6	4º	4º
Regressão Logística	66,4	83,1	64,8	85,6	59,6	59,4	5º	3º

Tabela 22 – Comparativos de desempenho entre os resultados obtidos nesse estudo e os obtidos no estudo de referência para o desfecho ausência de dentes ou perda de um ou mais dentes.

Fonte: Elaborado pelo Autor (2023).

6 CONSIDERAÇÕES FINAIS

Este capítulo encerra o estudo, ao realizar a identificação e discussão das restrições, limitações ou desafios encontrados durante a realização da pesquisa, juntamente com possíveis oportunidades de extensão e melhoria do trabalho.

6.1 Limitações e trabalhos futuros

Por se tratar a PNS de um questionário no qual as informações são obtidas por autorrelato, e não mediante análise clínica, não se pode atestar a veracidade dos dados fornecidos, já que os entrevistados podem responder o que entendem ser ideal e não a realidade.

Mostrou-se, ainda, evidente a necessidade de estudos futuros com coortes longitudinais com maior separação de tempo entre os conjuntos de dados de treinamento e testes para avaliar a validade externa dos modelos aqui propostos e confirmar sua capacidade de inferência.

Houve redução do conjunto de dados, uma vez que foram excluídas instâncias de dados com valores indefinidos para variáveis selecionadas para o estudo.

Por fim, cabe informar outra limitação: mesmo com o desempenho AUC de 89.0%, o desfecho edentulismo, presente em apenas 1,19% das amostras, configura-se como “evento raro” e, por esse motivo, pode ter afetado o desempenho dos modelos.

Como trabalhos futuros, sugere-se o desenvolvimento de modelo com o algoritmo Catboost para avaliar seu desempenho frente aos algoritmos analisados neste estudo. Conjuntamente, explorar mais recursos de *Deep Learning* visando a otimização dos resultados obtidos neste trabalho com a Rede Neural MLP.

Em conclusão, uma vez que o estudo delimitou seu foco na análise dos dados da PNS, propor-se-ia estender a análise ao conjunto de dados oriundo da pesquisa do SBBrazil (Sistema de Vigilância de Fatores de Risco e Proteção para Doenças Crônicas não Transmissíveis), que é um programa de vigilância em saúde bucal realizado no Brasil. Deste modo, seria possível verificar se resultados confirmam os achados desta pesquisa e explorar outras nuances que possam se mostrar relevantes ao tema de saúde bucal no país.

REFERÊNCIAS

- ALVES, Priscila Mello. “Inteligência Artificial e Redes Neurais”. *IPEA – Centro de Pesquisa em Ciência, Tecnologia e Sociedade*, [s. l.], 23 de dezembro de 2020. Artigos. Disponível em <https://www.ipea.gov.br/cts/pt/central-de-conteudo/artigos/artigos/106-inteligencia-artificial-e-redes-neurais>. Acesso em 04 jan. 2023.
- ANDRADE, Fabíola Bof; TEIXEIRA, Doralice Severo da Cruz; FRAZÃO, Paulo; DUARTE, Yeda Aparecida Oliveira; LEBRÃO, Maria Lúcia; ANTUNES, José Leopoldo Ferreira. Perfil de saúde bucal de idosos não institucionalizados e sua associação com autoavaliação da saúde bucal. *Revista brasileira de epidemiologia*, v. 21, 2019.
- BARATIERI, Luiz Narciso et al. “Odontologia restauradora: fundamentos e possibilidades”. In: *Odontologia restauradora: fundamentos e possibilidades*. São Paulo – SP - Brasil: Santos, 2002.
- BATISTA, André Filipe de Moraes; CHIAVEGATTO FILHO, Alexandre Dias Porto. “Machine Learning aplicado à Saúde”. In: *Anais do 19º Simpósio Brasileiro de Computação Aplicada à Saúde*, pp. 1-42, Niterói, RJ, 11 a 14 de junho de 2019. Disponível em <https://sol.sbc.org.br/livros/index.php/sbc/catalog/download/29/94/244-1?inline=1>. Acesso em 12 fev. 2023.
- BORDIN, Danielle. “Determinantes da condição percebida de saúde bucal e da adesão ao autocuidado em adultos brasileiros”. Tese de D.Sc., Pós-Graduação em Odontologia Preventiva e Social/UNESP, Araçatuba, SP, Brasil. 2017.
- BORISOV, Vadim; LEEMANN, Tobias; SEßLER, Kathrin; HAUG, Johannes; PAWELCZYK, Martin; KASNECI, Gjergji. *Deep Neural Networks and tabular data: A survey*. IEEE Transactions on Neural Networks and Learning Systems, 2022.
- BROADBENT, J. M.; THOMSON, W. M.; POULTON, R. “Progression of dental caries and tooth loss between the third and fourth decades of life: a birth cohort study”. *Caries Res.*, v. 40, n. 6, pp. 459–65, 2006.
- BROWNLEE, Jason. “Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning”. Edição do Kindle. [s. n.]. 2021.
- BUFFON, Marilene da Cruz Magalhães. “Aplicação do extrato de agrião d’água (*nasturtium officinale* r. br.) no controle da placa bacteriana: uma proposta para a saúde pública”. Tese de D.Sc., Pós-Graduação em Produção Vegetal/UFPR, Curitiba, PR, Brasil. 2005.
- CHEN, Tianqi; GUESTRIN, Carlos. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, San Francisco, CA, Estados Unidos, 13 de Agosto de 2016.
- COORAY, Upul; WATT Richard G.; TSAKOS, Georgios; HEILMANN Aja; HARIYAMA, Masanori; YAMAMOTO, Takafumi; KURUPPUARACHCHIGE, Isuruni; KONDO, Katsunori; OSAKA Ken; AIDA, Jun. Importance of socioeconomic factors in predicting

tooth loss among older adults in Japan: Evidence from a Machine Learning analysis. *Social Science & Medicine*, v. 291, p. 114486, 2021

ELANI, Hawazin W.; BATISTA, André F. M.; THOMSON, W. Murray; KAWACHI, Ichiro; CHIAVEGATTO FILHO, Alexandre D. P. "Predictors of tooth loss: A *Machine Learning* approach". *PLOS ONE*, v. 16, n. 6, (Jun.), pp. 1-14. 2021. Disponível em <https://doi.org/10.1371/journal.pone.0252873>. Acesso em 25 mar. 2023.

FEIJÓ, Isadora da Silva; IWASAKI, Kesley Merry Katherine. "Cárie e dieta alimentar". *Uningá Review*, v. 19, n. 3, pp. 44-50, 2014.

FERREIRA, Pedro Miguel Santos. "Modelos de Aprendizagem Computacional para previsão da eficácia de medicamentos avaliados em linhas celulares". Dissertação de M.Sc., Pós-Graduação em Engenharia Informática e Computação/Universidade do Porto, Porto Portugal. 2020a

FERREIRA, Raquel Conceição; SENNA, Maria Inês Barreiros; RODRIGUES, Lorrany Gabriela; CAMPOS, Fernanda Lamounier; MARTINS, Andrea Eleuterio Barros Lima; KAWACHI, Ichiro. Education and income-based inequality in tooth loss among Brazilian adults: does the place you live make a difference?. *BMC Oral Health*, v. 20, p. 1-13, 2020b.

GÉRON, Aurélien. *Mãos A Obra: Aprendizado De Máquina Com Scikit-Learn, Keras & TensorFlow*. 2 ed. Rio de Janeiro, Alta Books. Edição do Kindle, 2021.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep learning*. 1 ed. Vol. 1. Cambridge, MIT press, 2016.

LINDHE, Jan; KARRING, Thorkild; LANG, Niklaus. "Tratado de periodontia clínica e implantodologia oral". In: *Tratado de periodontia clínica e implantodologia oral*. 3 ed. Rio de Janeiro, Guanabara Koogan, 1999

MATSUYAMA, Yusuke; JÜRGES, Hendrik; LISTL, Stefan. The causal effect of education on tooth loss: evidence from United Kingdom schooling reforms. *American journal of epidemiology*, v. 188, n. 1, p. 87-95, 2019.

MINISTÉRIO DA SAÚDE (BRASIL). Conselho Nacional de Saúde. *3ª Conferência Nacional de Saúde Bucal. Saúde Bucal: acesso e qualidade, superando a exclusão social*. Relatório Final. Brasília: Ministério da Saúde; 2004. Disponível em https://bvsms.saude.gov.br/bvs/publicacoes/3_conferencia_nacional_saude_bucal_relatorio_final.pdf

NARVAI, Paulo Capel; FRAZÃO, Paulo. *Saúde bucal no Brasil: muito além do céu da boca*. Rio de Janeiro, Editora FIOCRUZ. Edição do Kindle, 2008.

ONG, Grace. "Periodontal disease and tooth loss". *Int Dent J.*, v. 48, n. 3, (Jun), pp. 233–238. 1998.

REIS, João.; MELO, Paulo. "A cárie dentária, uma doença infecciosa". *Revista Portuguesa de Saúde Pública*, v. 21, n. 1, pp. 35-40, (Jan./Jun). 2003.

ROBERTO, Luana L; CRESPO, Thaisa S.; MONBTEIRO-JUNIOR, Renato S.;

MARTINS, Andréa M. E. B. I.; DE PAULA, Alfredo M. B.; FERREIRA, Efigênia F.; HAIKAL, Desirée S. Sociodemographic determinants of edentulism in the elderly population: A systematic review and meta-analysis. *Gerodontology*, v. 36, n. 4, p. 325-337, 2019.

ROMANDINI, M.; BAIMA, G.; ANTONOGLIOU, G.; BUENO, J.; FIGUERO, E.; SANZ, M. "Periodontitis, Edentulism, and Risk of Mortality: A Systematic Review with Meta-analyses". *Journal of Dental Research*, v. 100, n. 1, pp. 37-49, 2021.

RUSSELL, Stuart; NORVIG, Peter. *Artificial intelligence: a modern approach*. 4 ed. Londres, Pearson Education Limited, 2020.

SCHIERZ, O.; BABA, K.; FUEKI, K. "Functional oral health-related quality of life impact: a systematic review in populations with tooth loss: . *J Oral Rehabil*, v. 48, n. 3, pp. 256-270, 2021.

SILVA, G. F. S.; DUARTE, L. S.; SHIRASSU, M. M.; PERES, S. V.; DEMORAES, M. A.; CHIAVEGATTO FILHO, A. "mortality risk prediction in patients with malignant neoplasm in São Paulo, Brazil". *Artificial Intelligence in the Life Sciences*, v. 3 (Dez.), pp. 1-8, 2023. Disponível em <https://www.sciencedirect.com/science/article/pii/S2667318523000053>

STOROPOLI, Jose. *Ciência de Dados e Computação Científica com Julia*. [online]. Disponível em <https://storopoli.github.io/Computacao-Cientifica>.

STOROPOLI, José Eduardo; SOUZA, Edson Melo de. *Ciência de Dados com Python: pandas, matplotlib, Scikit-Learn, TensorFlow e PyTorch*. [online]. 2020. Disponível em <https://github.com/storopoli/ciencia-de-dados>

THOMSON, William Murray. "Monitoring edentulism in older New Zealand adults over two decades: a review and commentary". *International Journal of Dentistry*, . 2012;2012:375407.

APÊNDICE A – TABELA DE PSEUDÔNIMOS X NOMENCLATURAS PNS 2013 E 2019

PSEUDÔNIMOS	NOMENCLATURA PNS	
	2013	2019
CARACTERÍSTICAS SOCIOECONÔMICAS		
idade	C008	C008
sexo	C006	C006
raca	C009	C009
nacionalidade		-
educacao_nivel	VDD004A	VDD004A
estado_civil	C011	C011
tamanho_familia	VDC001	VDC001
faixa_renda_per_capita	VDF003	VDF004
empregado	VDE001	VDE001
casa_propria	-	-
num_quartos_casa	A010	A01001
plano_de_saude	I001	I00102
imc	Peso/altura2 = 'P00101'/'P00401'x' P00401'	Peso/altura2 = 'P00104'/'P00404'x'P00404'
COMPORTAMENTO DE SAÚDE ORAL		
tempo_ultima_visita_dentista	J013	J01301
utiliza_escova_dente	U00201	U00204
utiliza_pasta	U00202	U00205
utiliza_fio_dental	U00203	U00206
frequencia_escovacao	U001	U00101

frequencia_substituicao_escova_dental	U004	U00401
---------------------------------------	------	--------

CONDIÇÕES MÉDICAS CRÔNICAS

asma	Q074	Q074
artrite	Q079	Q079
derrame	Q068	Q068
ataque_cardiaco	Q063	Q06306
diabetes	Q030	Q03001
colesterol_alto	Q060	Q060
hipertensao	Q002	Q00201

VARIÁVEIS DE DESFECHO

num_dentes_naturais	32 – ('U02301' + 'U02401')	32 – ('U02303' + 'U02403')
---------------------	----------------------------	----------------------------

edentulismo	[Sim] num_dentes_naturais == 0, [Não] qualquer outra condição
-------------	---

denticao_funcional (perda da dentição funcional)	[Sim] num_dentes_naturais > 0 e < 21, [Não] qualquer outra condição
---	---

dentes_ausentes	[Sim] num_dentes_naturais >=21 e <28, [Não] qualquer outra condição
-----------------	---

APÊNDICE B – TABELA DE CONFIGURAÇÃO FINAL DE HIPERPARÂMETROS

ALGORITMO	EDENTULISMO	PERDA DA DENTIÇÃO FUNCIONAL (TER MENOS QUE 21 DENTES)	AUSÊNCIA DE DENTES (PERDA DE ALGUM DENTE)
LogisticRegression			
class_weight	balanced	balanced	balanced
C	0.01	0.01	0.01
penalty	l2	l2	l2
solver	liblinear	liblinear	liblinear
RandomForestClassifier			
bootstrap	False	-	False
class_weight	balanced	{0: 1, 1: 10.625}	balanced
criterion	entropy	gini	entropy
max_depth	15	None	142
max_features	-	auto	sqrt
min_samples_leaf	-	1	2
min_samples_split	-	2	2
n_estimators	1000	1000	538
XGBClassifier			
binary	binary:logistic	binary:logistic	binary:logistic
eval_metric	auc	auc	auc
eta	0.1	0.1	0.1
max_depth	6	6	6
subsample	0.8	0.8	0.8
colsample_bytree	0.8	0.8	0.8
scale_pos_weight	82.67677824267783	10.625490481034733	3.856
LGBMClassifier			
objective	binary	binary	-
metric	binary_logloss	binary_logloss	-
num_leaves	31	31	147
learning_rate	0.05	0.05	0.1000
feature_fraction	0.9	0.9	-
bagging_fraction	0.8	0.8	-
bagging_freq	5	5	-
class_weight	{0: 1, 1: 83.33}	{0: 1, 1: 10.62}	-
colsample_bytree	-	-	0.6046
min_child_samples	-	-	20
n_estimators	-	-	100
reg_alpha	-	-	0.0080
reg_lambda	-	-	0.3946
scale_pos_weight	-	-	3.856
MLPClassifier			
activation	tanh	relu	relu
alpha	0.0100	0.0001	0.0100
batch_size	15	auto	5
hidden_layer_sizes	(7,)	100,	(11,)
solver	-	adam	-
learning_rate	-	constant	-
learning_rate_init	-	0.001	-
max_iter	-	200	-

APÊNDICE C – RESULTADOS OBTIDOS A PARTIR DA AVALIAÇÃO DO CONJUNTO DE TESTE

DESFECHO/ALGORITMO	SENSIBILIDADE (RECALL)	ESPECIFICIDADE	PRECISÃO	F1-SCORE	AUC ROC
EDENTULISMO					
<u>Classificador</u>					
Regressão Logística	83,1	78,7	4,5	8,5	88,6
Floresta Aleatória (<i>Random Forest</i>)	57,7	90,1	6,6	11,8	88,5
<i>Extreme Gradient boosting Trees</i> (XGBoost)	73,6	83,8	5,2	9,7	88,2
Light <i>Gradient boosting</i> Machine (LightGBM)	80,4	81,6	5	9,4	89,2
Redes Neurais MultiLayer Perceptron (MLP)	52,6	90,2	6,1	10,9	87,1
PERDA FUNCIONAL DENTÁRIA					
<u>Classificador</u>					
Regressão Logística	73,6	69,2	18,3	29,4	77,5
Floresta Aleatória (<i>Random Forest</i>)	0,8	99,7	18,9	1,6	76,7
<i>Extreme Gradient boosting Trees</i> (XGBoost)	82,5	63,9	17,7	29,2	78,7
Light <i>Gradient boosting</i> Machine (LightGBM)	85,8	61,9	17,5	29,1	79,1
Redes Neurais MultiLayer Perceptron (MLP)	27,5	89,2	19,3	22,7	73,8
AUSÊNCIA DE DENTES					
(perda de um ou mais dentes)					
<u>Classificador</u>					
Regressão Logística	64,8	59,6	29,4	40,4	66,4
Floresta Aleatória (<i>Random Forest</i>)	27,4	86,7	34,8	30,7	70,3
<i>Extreme Gradient boosting Trees</i> (XGBoost)	94,1	33,7	26,9	41,8	72
Light <i>Gradient boosting</i> Machine (LightGBM)	91,6	38,3	27,8	42,6	71,6
Redes Neurais MultiLayer Perceptron (MLP)	64,2	65,6	32,6	43,3	70,9