



Bryan Nicholas Fontinele Miranda

O Uso de Algoritmos de Classificação de Aprendizado de Máquina Para Predição da Doença de Cálculo Biliar (Pedra Na Vesícula) a Partir de Dados Laboratoriais e de Bioimpedância.

**Sobral-CE
2025**

Bryan Nicholas Fontinele Miranda

O Uso de Algoritmos de Classificação de Machine Learning Para Predição da Doença de Cálculo Biliar (Pedra Na Vesícula) a Partir de Dados Laboratoriais e de Bioimpedância.

Trabalho de Conclusão de Curso apresentado ao corpo docente do Departamento de Engenharia Elétrica do Departamento da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Engenheiro Eletricista.

Orientador: Prof. Dr. Carlos Alexandre.

Sobral-CE

2025

Página reservada para ficha catalográfica.

Utilize a ferramenta online Catalog! para elaborar a ficha catalográfica de seu trabalho
acadêmico, gerando-a em arquivo PDF, disponível para download e/ou impressão.
(<http://www.fichacatalografica.ufc.br/>)

Bryan Nicholas Fontinele Miranda

O Uso de Algoritmos de Classificação de Machine Learning Para Predição da
Doença de Cálculo Biliar (Pedra Na Vesícula) a Partir de Dados Laboratoriais e de
Bioimpedância.

Trabalho de Conclusão de Curso apresentado
ao corpo docente do Departamento de
Engenharia Elétrica do Departamento da
Universidade Federal do Ceará, como requisito
parcial à obtenção do título de Engenheiro
Eletricista.

Aprovada em: 04/12/2025.

BANCA EXAMINADORA

Prof. Dr. Carlos Alexandre (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Jermama Lopes
Universidade Federal do Ceará (UFC)

Prof. Dr. Wendley Souza
Universidade Estadual do Ceará (UECE)

A toda a classe trabalhadora que mantém a
sociedade brasileira em pé.

“OBEY!”. (John Carpenter).

RESUMO

O diagnóstico precoce da doença do cálculo biliar (colelitíase) é um desafio relevante, visto que o método padrão (ultrassonografia) não é universalmente acessível, o que impulsiona a busca por métodos preditivos de baixo custo. Este Trabalho de Conclusão de Curso aborda este desafio através da aplicação de técnicas de Aprendizado de Máquina. Utilizando o *dataset* público "Gallstone-1", o objetivo central foi duplo: desenvolver um modelo preditivo de alta acurácia e identificar os principais fatores de risco que contribuem para o diagnóstico, com base em dados clínicos, laboratoriais e de bioimpedânciia. A metodologia envolveu uma rigorosa análise exploratória e pré-processamento, incluindo a normalização de dados (MinMaxScaler), seguida pela otimização de hiperparâmetros via GridSearchCV com validação cruzada e uma avaliação comparativa de um portfólio de 16 algoritmos de classificação. Os resultados indicaram o Gradient Boosting Classifier como o modelo de melhor performance, alcançando uma acurácia de 91%, superando o benchmark de 85,42% estabelecido na literatura para este mesmo conjunto de dados. A análise de interpretabilidade com SHAP cumpriu o segundo objetivo, revelando que as características mais preditivas estão associadas à inflamação (Proteína C-Reativa), ao metabolismo (Vitamina D) e à composição corporal (Obesidade). O estudo conclui que o modelo desenvolvido se prova uma ferramenta de triagem não invasiva e de eficácia superior, com potencial para otimizar a alocação de recursos diagnósticos e auxiliar na decisão clínica.

Palavras-chave: Aprendizado de Máquina; Cálculo Biliar; Predição de Doenças; Dados Clínicos; Gradient Boosting.

ABSTRACT

The early diagnosis of gallstone disease (cholelithiasis) is a relevant challenge, given that the standard method (ultrasound) is not universally accessible, which drives the search for low-cost predictive methods. This thesis addresses this challenge by applying Machine Learning techniques. Using the public "Gallstone-1" dataset, the central objective was twofold: to develop a high-accuracy predictive model and to identify the main risk factors that contribute to the diagnosis, based on clinical, laboratory, and bioimpedance data. The methodology involved rigorous exploratory analysis and preprocessing, including data normalization (MinMaxScaler), followed by hyperparameter optimization via GridSearchCV with cross-validation and a comparative evaluation of a portfolio of 16 classification algorithms. The results indicated the Gradient Boosting Classifier as the best-performing model, achieving an accuracy of 91%, surpassing the 85.42% benchmark established in the literature for this same dataset. The interpretability analysis with SHAP fulfilled the second objective, revealing that the most predictive features are associated with inflammation (C-Reactive Protein), metabolism (Vitamin D), and body composition (Obesity). The study concludes that the developed model proves to be a non-invasive and highly effective screening tool, with the potential to optimize the allocation of diagnostic resources and assist in clinical decision-making.

Keywords: Machine Learning; Gallstone Disease; Disease Prediction; Clinical Data; Gradient Boosting.

SUMÁRIO

1 INTRODUÇÃO.....	12
1.1 Contextualização.....	12
1.2 Justificativa.....	13
1.3 Objetivos.....	14
2 Fundamentação Teórica.....	15
2.1 Inteligência Artificial e Aprendizado de Máquina.....	15
2.2 Tipos de Aprendizado de Máquina.....	15
2.3 Algoritmos de Classificação.....	16
2.3.1 Modelo de Baseline.....	16
2.3.2 Modelos Lineares.....	16
2.3.3 Modelos Baseados em Proximidade.....	17
2.3.4 Máquinas de Vetores de Suporte.....	17
2.3.5 Modelos Baseados em Árvores e Ensemble Learning.....	17
2.3.7 Modelos Probabilísticos e Discriminantes.....	18
2.3.8 Redes Neurais Artificiais.....	18
2.4 Interpretabilidade com SHAP.....	19
2.5 A Doença do Cálculo Biliar (Colelitíase).....	20
3 REVISÃO DA LITERATURA.....	22
3.1 Estudos Fundamentais e o Desempenho Preditivo com Dados Tabulares.....	23
3.2 Identificação de Fatores de Risco e Interpretabilidade dos Modelos.....	23
3.3 Abordagens e Aplicações Diversificadas.....	24
3.5 Síntese da Literatura e Justificativa da Pesquisa.....	24
4. METODOLOGIA.....	26
4.1 Desenho do Estudo.....	26
4.2 Descrição do Conjunto de Dados.....	26
4.3 Ambiente Computacional e Ferramentas.....	28
4.4 Pré-processamento dos Dados.....	30
4.5 Estratégia Experimental de Seleção de Atributos.....	31
4.6 Modelos de Machine Learning Testados.....	32
4.7 Estratégia de Validação e Métricas de Avaliação.....	33
4.8 Métricas de Avaliação de Desempenho.....	34
5. RESULTADOS E DISCUSSÃO.....	37
5.1. Insights da Análise Exploratória de Dados.....	37
5.2 Desempenho do Modelo de Baseline.....	39
5.3 Desempenho Comparativo dos Modelos de Classificação.....	39
5.4 Análise Detalhada dos Três Modelos de Melhor Performance.....	42
5.5 Análise de Importância dos Atributos dos Três Modelos de Melhor Performance.....	45
6 CONCLUSÃO.....	55
6.1 Síntese do Problema de Pesquisa.....	55
6.2 Síntese dos Principais Achados.....	55
6.3 Conclusão Geral.....	55

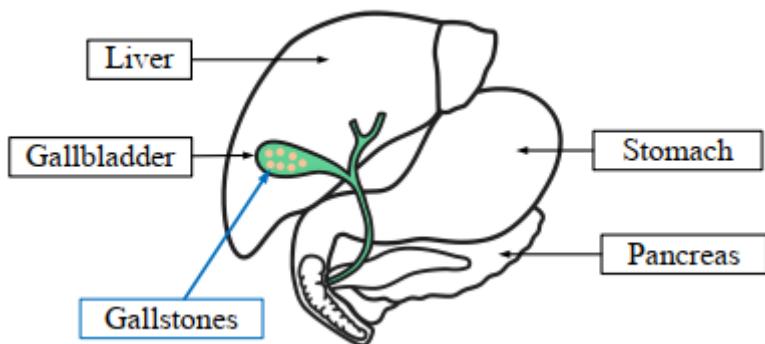
6.4 Sugestões para Trabalhos Futuros.....	56
REFERÊNCIAS.....	57
APÊNDICE A.....	60

1 INTRODUÇÃO

1.1 Contextualização

A doença do cálculo biliar, mais conhecida popularmente como “pedra na vesícula”, figura entre as condições gastrointestinais mais prevalentes em escala global (AHMED et al., 2024). Trata-se de um problema de saúde caracterizado pela formação de depósitos sólidos, os chamados cálculos, no interior da vesícula biliar, um órgão em forma de saco localizado sob o fígado, conforme ilustrado na Figura 1. A origem desse processo é multifatorial, resultante de uma complexa interação de elementos. A bile, um líquido essencial para a digestão de gorduras, pode ter sua composição alterada, sendo o colesterol o principal responsável pela vasta maioria da formação de cálculos. Quando esses depósitos obstruem o fluxo biliar, podem desencadear uma inflamação aguda denominada colecistite (MINISTÉRIO DA SAÚDE, 2022), sendo que Liver significa fígado, gallbladder significa vesícula biliar, gallstone significa pedra na vesícula, stomach significa estômago e pancreas significa pâncreas (Sarker et al. (2025)).

Figura 1 - Ilustração anatômica da vesícula biliar e a localização dos cálculos biliares.



Fonte: Adaptado de Sarker et al. (2025).

A lista de fatores de risco que podem acionar o gatilho para a formação dos cálculos é extensa e abrange desde predisposições genéticas até influências diretas do estilo de vida, como dietas ricas em gorduras e carboidratos, sedentarismo, obesidade, diabetes e hipertensão. Fatores adicionais como o tabagismo, o uso prolongado de anticoncepcionais e a elevação do nível de estrogênio, o que explica a maior incidência da doença em mulheres,

também contribuem significativamente para o seu desenvolvimento (MINISTÉRIO DA SAÚDE, 2022; TAN; JIA; LIU, 2025).

Ainda que muitos indivíduos portadores desses cálculos permaneçam assintomáticos ao longo da vida, a progressão para um quadro sintomático é uma ameaça real e grave. Estudos apontam que até 30% dos casos podem evoluir para manifestações clínicas, incluindo dor intensa, náuseas e febre. O maior perigo, contudo, reside nas complicações severas, como a colecistite aguda, a colangite (infecção dos ductos biliares) e a pancreatite biliar (AHMED et al., 2024). De acordo com o Ministério da Saúde (2022), pacientes que não realizam o tratamento cirúrgico enfrentam um risco de 30% a 50% de sofrerem complicações graves, que exigem intervenções de emergência. A pancreatite, em particular, destaca-se pela sua letalidade, com uma taxa de mortalidade que pode variar de 7% a 15% nesses casos.

Esses desdobramentos não se limitam a afetar diretamente a qualidade de vida dos pacientes; eles também acarretam uma sobrecarga significativa aos sistemas de saúde, tanto públicos quanto privados, gerando custos expressivos relacionados a internações e procedimentos cirúrgicos de alta complexidade. Assim, o cálculo biliar, à primeira vista uma condição silenciosa, revela um importante desafio de saúde pública quando considerado em larga escala. Sua prevalência, associada ao grave impacto clínico e econômico de suas complicações, reforça a necessidade de aprofundar estratégias que viabilizem diagnósticos mais acessíveis e precoces, favorecendo intervenções mais oportunas.

1.2 Justificativa

Diante desse panorama, o diagnóstico precoce se configura como uma das principais ferramentas no enfrentamento da doença do cálculo biliar. A identificação rápida e precisa dos pacientes em risco permite não apenas prevenir complicações graves, mas também otimizar recursos de saúde. Atualmente, a ultrassonografia abdominal é considerada o padrão-ouro para a detecção dos cálculos (SONG et al., 2019). Trata-se de um exame de imagem relativamente seguro, não invasivo e bastante eficaz para a visualização dos depósitos biliares. Contudo, essa abordagem apresenta limitações que não podem ser ignoradas: o custo elevado dos equipamentos, a necessidade de profissionais treinados para a execução e interpretação, além da menor sensibilidade em casos de cálculos muito pequenos ou da presença da chamada “lama biliar” (ESEN et al., 2024).

Nesse contexto de desafios, surgem as contribuições cada vez mais robustas das áreas de ciência de dados e inteligência artificial. A aplicação de técnicas de *machine learning*

(aprendizado de máquina) tem se consolidado como um campo fértil para soluções inovadoras em saúde. Em especial, a possibilidade de desenvolver modelos preditivos com base em dados simples e rotineiros, como exames laboratoriais básicos ou medições de bioimpedância, abre espaço para a criação de ferramentas de triagem acessíveis, não invasivas e escaláveis. Ao identificar padrões invisíveis à análise humana, esses algoritmos podem oferecer classificações de risco mais precisas, direcionando o uso de exames de imagem apenas para os casos realmente mais prováveis.

Portanto, unir a medicina tradicional com o potencial analítico da inteligência artificial não é apenas uma alternativa tecnológica, mas uma estratégia que pode transformar o modo como lidamos com a prevenção e o diagnóstico dessa enfermidade.

1.3 Objetivos

O presente trabalho tem como objetivo geral implementar, treinar e avaliar o desempenho de diferentes modelos de classificação de aprendizado de máquina aplicados à predição da doença do cálculo biliar, utilizando a linguagem Python e um banco de dados composto por informações laboratoriais e de bioimpedância.

Para alcançar esse objetivo mais amplo, foram definidos os seguintes objetivos específicos:

- Realizar o pré-processamento e a análise exploratória do conjunto de dados, de forma a compreender as características e distribuições das variáveis.
- Aplicar e avaliar diferentes técnicas de seleção de atributos (como Regressão Lasso e Random Forest Importance) para determinar o subconjunto de variáveis mais preditivas.
- Implementar e treinar um conjunto de algoritmos de classificação.
- Avaliar e comparar a performance dos modelos utilizando métricas quantitativas como acurácia, precisão, recall e F1-score.
- Identificar as variáveis de maior poder preditivo através de técnicas de interpretabilidade (como a análise SHAP), a fim de compreender quais fatores clínicos e de bioimpedância são mais determinantes para as previsões do modelo de melhor desempenho.

Todos os algoritmos foram implementados na linguagem Python. O código-fonte completo, juntamente com os scripts de pré-processamento e os notebooks de análise, está disponível publicamente em um repositório no GitHub para fins de reproduzibilidade,

disponível em:
<https://github.com/Nicholas-UFC/TCC-sobre-analise-de-dados-de-uma-biblioteca-sobre-gallstone>.

2 Fundamentação Teórica

2.1 Inteligência Artificial e Aprendizado de Máquina

A Inteligência Artificial (IA) é um vasto campo da ciência da computação dedicado a criar sistemas capazes de realizar tarefas que normalmente exigiriam inteligência humana. Dentro da IA, o Aprendizado de Máquina se destaca como uma subárea fundamental, na qual os algoritmos não são explicitamente programados com regras, mas "aprendem" padrões e relações diretamente a partir de dados. O objetivo é construir um modelo matemático que, após ser treinado com um conjunto de dados, consiga fazer previsões ou tomar decisões sobre novos dados nunca antes vistos.

2.2 Tipos de Aprendizado de Máquina

O *Machine Learning* é comumente dividido em paradigmas de aprendizado, sendo o principal para este trabalho o supervisionado. Aprendizado Supervisionado: Nesta abordagem, o algoritmo é treinado com um conjunto de dados "rotulado", onde cada amostra de entrada (vetor de características X) está associada a uma saída ou rótulo correto (y). O objetivo do modelo é aprender a mapear as entradas para as saídas. As tarefas de aprendizado supervisionado são geralmente divididas em:

- Classificação: O objetivo é prever um rótulo de classe discreta. No contexto deste trabalho, a tarefa é classificar um paciente em uma de duas categorias: "com cálculo biliar" (classe 1) ou "sem cálculo biliar" (classe 0).
- Regressão: O objetivo é prever um valor numérico contínuo, como o preço de um imóvel ou a temperatura.

Aprendizado Não Supervisionado: Diferente da abordagem supervisionada, no aprendizado não supervisionado o algoritmo trabalha com dados não rotulados, ou seja, não há uma variável de saída correta para guiar o treinamento. O objetivo aqui é explorar a estrutura intrínseca dos dados para encontrar padrões ou agrupamentos ocultos.

- Agrupamento (*Clustering*): É a tarefa mais comum do aprendizado não supervisionado. O objetivo do agrupamento é particionar o conjunto de dados em grupos (ou clusters), de modo que as amostras dentro do mesmo grupo sejam muito semelhantes entre si e distintas das amostras de outros grupos. É útil para segmentação

de clientes, detecção de anomalias, entre outras aplicações.

2.3 Algoritmos de Classificação

Para a tarefa de classificação, a literatura de aprendizado de máquina oferece um vasto leque de algoritmos, cada um com diferentes abordagens teóricas, vantagens e desvantagens. A escolha do algoritmo ideal frequentemente depende da natureza dos dados, da complexidade do problema e dos objetivos do estudo, como a busca por maior acurácia ou interpretabilidade. A fim de realizar uma análise comparativa robusta, este trabalho implementou modelos de diferentes paradigmas, cujas bases teóricas serão detalhados logo abaixo:

2.3.1 *Modelo de Baseline*

Um modelo de *baseline* (ou modelo de referência) é um modelo muito simples e trivial que serve como um ponto de comparação para todos os seus outros modelos mais complexos. O classificador *Dummy* não aprende padrões a partir dos atributos de entrada (X), servindo como uma linha de base (*baseline*) simples para comparação com outros modelos mais complexos. Ele gera previsões usando regras triviais baseadas apenas na distribuição da variável alvo (y) do conjunto de treino, como, por exemplo, prever sempre a classe mais frequente (majoritária) ou prever aleatoriamente mantendo a proporção das classes. O seu propósito metodológico é estabelecer um limiar de desempenho mínimo; um classificador mais sofisticado só é considerado útil se sua performance for significativamente superior à deste modelo trivial.

2.3.2 *Modelos Lineares*

Modelos lineares buscam aprender uma função linear a partir das características de entrada para realizar a predição. São amplamente utilizados devido à sua simplicidade, rapidez no treinamento e elevada interpretabilidade.

- Regressão Logística: apesar do nome, trata-se de um algoritmo de classificação fundamental, que modela a probabilidade de ocorrência de uma determinada classe (no caso, presença de cálculo biliar) por meio da função logística (sigmóide).
- *Ridge Classifier*: classificador linear que utiliza regularização L2 (*Ridge*). Essa

regularização penaliza coeficientes de grande magnitude, reduzindo o risco de sobreajuste e auxiliando no tratamento da multicolinearidade entre variáveis.

- *SGDClassifier*: Um classificador linear muito eficiente, treinado através do algoritmo de otimização *Stochastic Gradient Descent*. É particularmente poderoso para trabalhar com grandes volumes de dados (*big data*), pois atualiza o modelo a cada amostra (ou pequeno lote), tornando o treinamento extremamente rápido.

2.3.3 Modelos Baseados em Proximidade

Esses modelos realizam a classificação com base na similaridade entre a nova instância e os exemplos do conjunto de treinamento. O *K-Nearest Neighbors (K-NN)* é o algoritmo mais clássico desta categoria. Trata-se de uma técnica não paramétrica baseada em instância. Para classificar um novo ponto, identifica os K vizinhos mais próximos e atribui a classe predominante entre eles.

2.3.4 Máquinas de Vetores de Suporte

Esta classe de algoritmos, conhecida como classificadores de margem máxima, buscam encontrar o hiperplano de separação que melhor divide os dados em diferentes classes, maximizando a distância (ou margem) entre os pontos mais próximos de cada classe. *Support Vector Machine (SVM)* é um classificador robusto que busca o hiperplano ótimo que maximize a margem entre as classes. Por meio da técnica conhecida como *kernel trick*, o SVM é capaz de construir fronteiras de decisão não lineares complexas, sendo eficaz em problemas de alta dimensionalidade.

2.3.5 Modelos Baseados em Árvores e Ensemble Learning

Métodos de *ensemble* combinam as previsões de múltiplos estimadores para alcançar resultados mais robustos e reduzir a variância dos modelos individuais.

- *Random Forest*: algoritmo de *ensemble learning* que gera diversas árvores de decisão a partir de amostras aleatórias dos dados e das variáveis. A predição final é obtida pelo voto majoritário, o que confere ao modelo elevada robustez e menor propensão ao sobreajuste.
- *Extra Trees Classifier*: O classificador *Extremely Randomized Trees* é um método de

ensemble learning que funciona de maneira semelhante ao *Random Forest*, mas introduz mais aleatoriedade na forma como os pontos de corte são selecionados para dividir os nós de uma árvore, o que pode ajudar a reduzir a variância do modelo.

- *Gradient Boosting Classifier*: Uma técnica de *ensemble learning* que constrói modelos de forma sequencial e aditiva. Cada novo modelo é treinado para corrigir os erros (resíduos) do modelo anterior, otimizando gradualmente uma função de perda diferenciável.
- *XGBoost*: implementação otimizada do algoritmo *Gradient Boosting*, caracterizada por alto desempenho computacional. Os modelos são construídos de forma sequencial, corrigindo erros anteriores. Destaca-se pela eficiência e pelo frequente alcance do estado da arte em tarefas com dados tabulares.
- *LightGBM*: Um *framework* de *gradient boosting* que utiliza técnicas baseadas em histogramas, o que o torna notavelmente rápido e eficiente em termos de uso de memória, especialmente com grandes conjuntos de dados.
- *CatBoost Classifier*: Um algoritmo de *gradient boosting* de código aberto que se destaca por sua capacidade de lidar nativamente com variáveis categóricas e pela implementação de árvores de decisão simétricas (*oblivious decision trees*), o que ajuda a prevenir o sobreajuste.

2.3.7 Modelos Probabilísticos e Discriminantes

Os modelos desta categoria utilizam uma abordagem estatística para a classificação, baseando-se no cálculo da probabilidade de uma nova instância pertencer a cada uma das classes a partir das distribuições aprendidas com os dados de treinamento.

- *Quadratic Discriminant Analysis*: classificador probabilístico que assume que os dados de cada classe seguem uma distribuição gaussiana. Aprende fronteiras de decisão quadráticas, o que o torna mais flexível que o *Linear Discriminant Analysis* (LDA).
- *Linear Discriminant Analysis*: Um método usado para encontrar uma combinação linear de características que melhor caracteriza ou separa duas ou mais classes de objetos ou eventos. O modelo resultante pode ser usado como um classificador linear
- *Naive Bayes*: Um classificador probabilístico que aplica o Teorema de *Bayes* com uma forte (ou "ingênua", do inglês *naive*) suposição de independência condicional entre as características, dados o valor da classe.

2.3.8 Redes Neurais Artificiais

Esta classe de algoritmos é inspirada no funcionamento do cérebro humano, utilizando camadas de neurônios interconectados para aprender e modelar relações não-lineares e complexas presentes nos dados.

- *Multilayer Perceptron*: rede neural artificial do tipo *feedforward*, composta por uma camada de entrada, uma ou mais camadas ocultas com funções de ativação não lineares e uma camada de saída. Trata-se de um aproximador universal de funções, capaz de identificar padrões complexos e não lineares nos dados.
- *Deep Learning Multilayer Perceptron*: Um subcampo do machine learning baseado em redes neurais artificiais com múltiplas camadas (arquiteturas profundas) entre a entrada e a saída. A arquitetura específica utilizada neste trabalho é o *Multilayer Perceptron*, também conhecido como Rede Neural Densa. Neste modelo, cada neurônio de uma camada é totalmente conectado aos neurônios da camada seguinte, o que permite a extração progressiva de características de maior nível a partir dos dados brutos, tornando-o capaz de modelar padrões de alta complexidade em dados tabulares.

2.4 Interpretabilidade com SHAP

À medida que os modelos de *machine learning*, como os de *ensemble learning* e as redes neurais, se tornam mais complexos, eles frequentemente operam como "caixas-preta" (*black boxes*), onde a lógica interna que leva a uma determinada predição é opaca para o usuário final. No contexto médico, essa falta de transparência é um obstáculo crítico para a adoção e a confiança em sistemas de IA. Para resolver este problema, surge o campo da Inteligência Artificial Explicável (XAI), cujo objetivo é desenvolver métodos para tornar as decisões dos modelos comprehensíveis para os seres humanos.

Dentro deste campo, o SHAP (*SHapley Additive exPlanations*) consolidou-se como uma das técnicas mais robustas e teoricamente bem-fundamentadas. O SHAP é uma abordagem agnóstica de modelo, baseada nos Valores de *Shapley*, um conceito da teoria dos jogos cooperativos (SARKER et al., 2025). Na teoria dos jogos, os valores de *Shapley* calculam a contribuição justa de cada jogador para o resultado final de um jogo. No *machine learning*, o SHAP adapta essa ideia para explicar uma predição: cada "jogador" é uma

característica (ou atributo) do modelo, e o "resultado do jogo" é a predição para uma única instância.

O método calcula a contribuição de cada característica para "empurrar" a predição desde um valor base (a média de todas as previsões) até o seu valor final. Essa abordagem tem duas vantagens principais:

1. Explicações Locais: O SHAP fornece uma explicação para cada predição individual. Ele mostra, para um paciente específico, quais fatores (ex: um nível alto de PCR, uma idade avançada) contribuíram para aumentar o risco de ter cálculo biliar e quais fatores contribuíram para diminuí-lo.
2. Explicações Globais: Ao agregar as explicações locais de todas as amostras, o SHAP permite entender o comportamento geral do modelo. Visualizações como o summary plot apresentam um design adequado para transmitir múltiplas informações simultaneamente (como a importância global e a direção do impacto das variáveis). Ao expor esses padrões de maneira visual e didática, a ferramenta facilita a análise e interpretação dos dados complexos por parte dos pesquisadores.

Desta forma, o SHAP permite ir além da simples medição de acurácia, respondendo à pergunta crucial: "Por que o modelo tomou esta decisão?". Isso não apenas aumenta a confiança no modelo, mas também permite que especialistas validem se o raciocínio do algoritmo está alinhado com o conhecimento clínico, identificando potenciais vieses e fortalecendo a ponte entre a ciência de dados e a prática médica.

2.5 A Doença do Cálculo Biliar (Colelitíase)

A vesícula biliar é um órgão pequeno, em formato de saco, localizado abaixo do fígado, cuja função principal é armazenar e concentrar a bile. A bile é um fluido produzido pelo fígado, essencial para a digestão de gorduras no intestino. A doença do cálculo biliar, ou colelitíase, é uma condição caracterizada pela formação de depósitos sólidos e cristalinos, conhecidos como cálculos ou pedras, no interior da vesícula biliar ou nos ductos biliares. Trata-se de uma das patologias gastrointestinais mais comuns, com uma prevalência que afeta aproximadamente 10 a 20% da população adulta mundial.

Os cálculos biliares são classificados principalmente por sua composição. A grande maioria, entre 75% a 80% dos casos em países ocidentais, são cálculos de colesterol, que se formam quando a bile contém excesso de colesterol, falta de sais biliares ou outros

fatores que prejudicam o esvaziamento da vesícula, (AHMED et al., 2024). Os demais são cálculos de pigmento (pretos ou marrons), que correspondem a menos de 10% dos casos e estão geralmente associados a outras condições médicas.

A formação da colelitíase é um processo multifatorial, com diversos fatores de risco bem documentados, que incluem:

- Fatores Demográficos e Genéticos: Sexo feminino, envelhecimento e predisposição genética aumentam significativamente o risco.
- Fatores Metabólicos: A obesidade e a síndrome metabólica estão fortemente correlacionadas a um aumento na formação de cálculos. Condições como diabetes, hiperlipidemia (níveis elevados de gordura no sangue) e resistência à insulina também são fatores de risco importantes.
- Fatores de Estilo de Vida: Dietas ricas em gorduras e carboidratos e pobres em fibras, sedentarismo e perda de peso rápida são conhecidas por contribuírem para a doença.
- Outros Fatores: A gravidez e o uso de contraceptivos orais, que elevam o nível de estrogênio, também aumentam a incidência em mulheres.

Muitos indivíduos com cálculos biliares permanecem assintomáticos por longos períodos. No entanto, quando os sintomas ocorrem, o mais comum é uma dor intensa e súbita no lado direito superior do abdômen, que pode irradiar para o tórax e as costas. Esta dor, conhecida como cólica biliar, frequentemente surge após refeições gordurosas e pode ser acompanhada de náuseas, vômitos, febre e, em alguns casos, icterícia (pele e olhos amarelados), Ahmed et al. (2024).

A ausência de tratamento pode levar a complicações graves em até 30% dos pacientes. As principais complicações incluem:

- Colecistite Aguda: Uma inflamação da vesícula biliar causada pela obstrução de um cálculo, que pode levar a infecções severas.
- Coledocolitíase: A migração de um cálculo para o duto biliar principal, podendo causar obstrução e icterícia.
- Colangite: Uma infecção grave dos ductos biliares.
- Pancreatite Biliar: A complicação mais temida, ocorre quando um cálculo obstrui o duto pancreático, causando uma inflamação no pâncreas. Esta condição é grave e apresenta uma taxa de mortalidade que pode variar de 7% a 15%.

O diagnóstico da colelitíase é tipicamente realizado por meio da ultrassonografia

abdominal, um método não invasivo e de alta sensibilidade e especificidade. Uma vez confirmado o diagnóstico, o tratamento definitivo, tanto para pacientes sintomáticos quanto para alguns assintomáticos com alto risco de complicações, é a colecistectomia, a remoção cirúrgica da vesícula biliar, (LI et al., 2024).

3 REVISÃO DA LITERATURA

A presente revisão da literatura adota um critério de seleção com a *string* de busca sendo "*machine learning*" OR "*artificial intelligence*" AND "*gallstone*" OR "*cholelithiasis*", tendo o ano de publicação entre 2020 até 2025, foi utilizado os sites: PubMed, IEEE Xplore e Scopus. Foi priorizando estudos com a máxima relevância para a análise aqui desenvolvida.

PubMed foram obtidos 21 resultados e IEEE Xplore 15 resultados, 98 foram obtidos no Scopus. Após a remoção dos artigos duplicados, foi consolidado um conjunto final de 108 artigos únicos para a triagem.

Ao final da triagem, 91% dos artigos foram excluídos, enquanto apenas 9% (um total de 10 artigos) foram aceitos para a análise aprofundada. Os principais motivos para a exclusão foram:

- Tópico Não Relacionado (36% das exclusões): Muitos estudos aplicavam aprendizado de máquina a outras condições médicas, como doenças renais, diabetes ou cardiovasculares.
- Foco em Análise de Imagens (34% das exclusões): A maior parte dos artigos descartados, apesar de usar IA para cálculo biliar, focava exclusivamente em dados de imagem (ultrassom, tomografia, etc.), o que foge do escopo deste trabalho com dados tabulares.
- Foco em Complicações ou Outros Objetivos (21% das exclusões): Outro grupo de artigos foi excluído por prever complicações da doença (como pancreatite ou câncer) ou a dificuldade cirúrgica, em vez de prever a presença da doença em si.

Dos 9 artigos aceitos, a maioria (6 artigos) foi proveniente da base de dados PubMed, enquanto 3 artigos foram identificados no IEEE Xplore, entretanto foi adicionado mais 1 artigo sobre imagem, pois ele cita o trabalho de Esen et al (2024), que é a publicação original associada à criação e disponibilização deste conjunto de dados. Nenhum artigo da busca na Scopus atendeu aos critérios de inclusão. Essa abordagem de triagem rigorosa resultou em um conjunto focado em publicações de alta relevância, permitindo uma análise aprofundada do estado da arte e a contextualização precisa deste TCC.

O ponto de partida da revisão são os estudos que estabelecem a base de comparação direta para este TCC. Dentre os 10 artigos aceitos, foi dado destaque especial aos trabalhos de Esen et al. (2024) e Sarker et al. (2025), pois ambos utilizam o mesmo conjunto de dados desta pesquisa, permitindo uma análise metodológica rigorosa.

Os demais artigos selecionados serão utilizados para contextualizar estes achados,

seja validando os fatores de risco identificados (como os estudos de Deng et al., 2025), explorando aplicações diversificadas (como Himi et al., 2023 e Thomas et al., 2018) ou estabelecendo um contraponto metodológico com a análise de imagens (como Ahmed et al., 2024). Essa abordagem permite uma análise aprofundada do estado da arte, avaliando a contribuição deste TCC de forma precisa.

Essa abordagem metodológica, que resulta em uma análise aprofundada de cinco artigos centrais, permite uma comparação direta e precisa entre as diferentes abordagens de modelagem, as métricas de desempenho alcançadas e os atributos identificados como mais importantes. Ao focar em trabalhos que partem da mesma base de dados, é possível contextualizar os resultados deste TCC de forma mais rigorosa e avaliar sua contribuição para o estado da arte específica deste problema.

3.1 Estudos Fundamentais e o Desempenho Preditivo com Dados Tabulares

A aplicação de aprendizado de máquina para a predição da colelitíase com dados não-imagéticos foi consolidada por estudos que utilizaram um mesmo banco de dados público, que também serve de base para este TCC. O trabalho de Esen et al. (2024) foi o primeiro a apresentar este *dataset*, composto por 319 amostras e 38 atributos, incluindo dados de bioimpedância e exames laboratoriais. Ao testar múltiplos algoritmos, os autores estabeleceram um *benchmark* de performance, alcançando 85,42% de acurácia com o modelo Gradient Boosting, demonstrando a viabilidade de se prever a doença com alta precisão.

Posteriormente, o estudo de Sarker et al. (2025) utilizou o mesmo *dataset* para aprofundar a análise, focando na otimização do classificador Random Forest e na interpretabilidade dos resultados. Por meio de um algoritmo de otimização, os autores conseguiram reduzir o número de atributos de 38 para 13, mantendo uma acurácia robusta de 79,17%. A consistência dos resultados entre estes dois trabalhos, mesmo com algoritmos e abordagens diferentes, reforça a validade do uso de dados clínicos para esta tarefa de predição.

3.2 Identificação de Fatores de Risco e Interpretabilidade dos Modelos

Um ponto de convergência crucial na literatura é a identificação dos fatores de risco mais importantes através dos próprios modelos de aprendizado de máquina. Tanto Esen et al. (2024) quanto Sarker et al. (2025) destacaram a Proteína C-Reativa (PCR) e a Vitamina D

como os atributos de maior poder preditivo. Este achado é de extrema relevância, pois conecta a predição algorítmica a marcadores biológicos de inflamação e metabolismo, conferindo validade clínica aos modelos.

A importância dos fatores de risco ligados à composição corporal também é um tema recorrente. O estudo de Deng et al. (2025), por exemplo, utilizou aprendizado de máquina em um grande conjunto de dados (da pesquisa NHANES e de hospitais) para investigar especificamente a relação entre índices físicos e a formação de cálculos biliares. A pesquisa concluiu, através de uma análise SHAP, que a massa de gordura relativa (RFM), o índice peso-cintura (WWI) e a circunferência da cintura (WC) são fortes preditores da doença, validando o que os modelos de Esen et al. e Sarker et al. já haviam sinalizado.

3.3 Abordagens e Aplicações Diversificadas

A literatura demonstra a versatilidade do aprendizado de máquina para este problema, com aplicações que vão além da simples classificação. O trabalho de Thomas et al. (2018), por exemplo, usou a colelitíase como estudo de caso para desenvolver um Sistema de Apoio à Decisão Clínica (CDSS), com o objetivo de prever a progressão da doença e a necessidade de intervenção. Eles propuseram uma Rede Neural em Cascata Modificada (ModCNN) que alcançou 96,42% de acurácia para esta tarefa.

Outro exemplo inovador é a pesquisa de Himi, S. T., et al. (2023), que apresentou um sistema para prever 11 doenças comuns, incluindo cálculos biliares, a partir de dados coletados por um *smartwatch*. Utilizando algoritmos como Random Forest, o estudo demonstrou a viabilidade de se usar dados de sensores para a predição de risco, ampliando as fontes de dados para além dos exames laboratoriais tradicionais. Adicionalmente, o estudo de Mena-Camilo et al. (2024) explorou o uso de Redes Neurais Convolucionais 1D com dados clínicos para prever a coledocolitíase, uma complicação da doença, mostrando a aplicação de arquiteturas de *deep learning* para este tipo de dado.

3.5 Síntese da Literatura e Justificativa da Pesquisa

A análise da literatura confirma a viabilidade de se prever a doença do cálculo biliar com alta acurácia usando aprendizado de máquina sobre dados tabulares. Os estudos convergem na identificação de fatores de risco clinicamente relevantes, como marcadores

inflamatórios e de composição corporal. No entanto, as pesquisas existentes se concentraram em modelos específicos ou em técnicas de otimização avançadas. Uma análise comparativa ampla e sistemática, avaliando um portfólio diversificado de algoritmos de classificação sobre um mesmo conjunto de dados, ainda representa uma lacuna. Este trabalho se justifica, portanto, por buscar preencher essa lacuna, oferecendo uma avaliação comparativa que pode guiar o desenvolvimento de futuras ferramentas de apoio ao diagnóstico.

4. METODOLOGIA

4.1 Desenho do Estudo

O presente trabalho enquadra-se como uma pesquisa quantitativa e aplicada, uma vez que busca não apenas compreender, mas também propor soluções práticas para o diagnóstico assistido por computador. O estudo fundamenta-se no uso de técnicas de *machine learning* voltadas ao desenvolvimento e à avaliação de modelos de classificação preditiva, tendo como foco a detecção da colelitíase.

O conjunto de dados empregado é composto por informações clínicas, laboratoriais e de bioimpedância, o que possibilita uma abordagem multidimensional do problema. A partir desses dados, o estudo estabelece um fluxo metodológico que contempla as seguintes etapas principais: análise exploratória dos dados, pré-processamento, seleção dos melhores atributos, treinamento de algoritmos de classificação, avaliação de desempenho e comparação dos resultados obtidos.

Todas as etapas foram implementadas utilizando a linguagem de programação Python, escolhida por sua ampla aceitação na comunidade científica e pela disponibilidade de bibliotecas especializadas em ciência de dados e aprendizado de máquina. Entre as principais bibliotecas empregadas destacam-se Pandas (manipulação e análise de dados), Scikit-learn (modelagem preditiva e métricas de avaliação), Shap (Análise de importância de atributo) Matplotlib e Seaborn (visualização gráfica).

Esse desenho metodológico foi estruturado de forma a garantir a reprodutibilidade do estudo, permitindo que outros pesquisadores possam replicar ou expandir a investigação a partir das mesmas bases técnicas.

4.2 Descrição do Conjunto de Dados

A presente pesquisa utilizou como base o conjunto de dados público *gallstone.csv*, originalmente compilado e disponibilizado nos estudos de Esen et al. (2024) e Sarker et al. (2025). Esse *dataset* constitui uma fonte confiável para o desenvolvimento de modelos de *machine learning* voltados à predição da colelitíase, uma vez que reúne informações clínicas, laboratoriais e de bioimpedância em uma amostra representativa de pacientes.

A base de dados que fundamenta esta pesquisa é constituída por 319 instâncias (registros), sendo cada uma caracterizada por um vetor de 38 variáveis preditoras (*features*) e

uma variável-alvo, intitulada *Gallstone Status*. A variável-alvo, que define o problema de classificação, é binária e assume o valor 1 para indicar um diagnóstico positivo para a doença do cálculo biliar e 0 para um diagnóstico negativo.

Para permitir uma compreensão aprofundada da estrutura e da qualidade dos dados, a Tabela 1, que pode ser encontrada no apêndice A, foi elaborada para detalhar o perfil de cada uma das 38 variáveis preditoras. A Tabela apresenta uma análise descritiva que abrange quatro aspectos cruciais para a etapa de pré-processamento e modelagem:

- Tipo de Dado: Classifica cada variável como Numérica (quantitativa) ou Categórica (qualitativa).
- Inconsistências de Formato: Identifica as variáveis que, em seu formato original, continham valores não numéricos e que necessitam de um tratamento específico para a sua conversão e limpeza.
- Valores Ausentes: Aponta a presença de dados faltantes (NaN) resultantes do processo de limpeza, os quais foram subsequentemente tratados por meio de técnicas de imputação.
- Presença de *Outliers*: Indica se a variável, após a análise exploratória, demonstrou a existência de valores atípicos (*outliers*). Para uma identificação sistemática desses valores, foi empregado o método do intervalo interquartil (KHAN ACADEMY, 2025). Este critério estatístico define que um valor é considerado um *outlier* se estiver significativamente distante da maioria dos outros dados.

A análise da estrutura do conjunto de dados revela que, dos 38 atributos preditores, 31 são de natureza numérica e 7 são categóricos. Uma inspeção inicial da qualidade dos dados identificou inconsistências de formato em três colunas: Vitamina D, Taxa de Filtração Glomerular (TFG) e Área de Músculo Visceral (AMV). Nestas colunas, valores numéricos foram registrados com pontos como separadores de milhar em vez de vírgula decimal (ex: 2.728.571.429), o que exigiu um tratamento específico durante o pré-processamento.

No que se refere à distribuição das classes, o *dataset* apresenta um equilíbrio satisfatório, contendo 161 registros positivos (50,47%) e 158 registros negativos (49,53%). Essa característica é relevante, pois reduz o risco de enviesamento do modelo para uma das classes, favorecendo um aprendizado mais consistente e generalizável.

As variáveis preditoras utilizadas neste estudo são classificadas em duas categorias fundamentais. A primeira são os atributos numéricos, que expressam quantidades e podem ser subdivididos em contínuos (valores que podem assumir qualquer número dentro de

um intervalo, como o peso) ou discretos (valores inteiros, como a idade). A segunda categoria são os atributos categóricos, que descrevem uma qualidade ou característica e cujos valores pertencem a um conjunto predefinido de rótulos, como o gênero.

Esse conjunto de dados, por sua riqueza e diversidade, possibilita a construção de modelos capazes de explorar múltiplos fatores associados à formação de cálculos biliares, constituindo uma base sólida para a investigação proposta.

4.3 Ambiente Computacional e Ferramentas

A execução deste projeto foi realizada integralmente em um ambiente de desenvolvimento baseado na linguagem de programação Python (versão 3.12), amplamente reconhecida pela comunidade científica e de ciência de dados devido à sua robustez, versatilidade e ao extenso ecossistema de bibliotecas especializadas. O trabalho foi conduzido de forma interativa em *Jupyter Notebooks*, ambiente que possibilita a prototipagem rápida, a análise exploratória e a documentação integrada do código.

Para operacionalizar o fluxo de trabalho em *machine learning*, desde a manipulação inicial dos dados até a avaliação e interpretação dos modelos, foram empregadas diferentes bibliotecas de código aberto, cada uma desempenhando um papel específico:

- Pandas: serviu como a principal ferramenta para manipulação e organização dos dados. A estrutura de *DataFrame* foi utilizada para carregar o arquivo gallstone.csv e realizar as etapas iniciais de inspeção (*.info()*, *.describe()*), limpeza (conversão de colunas categóricas em valores numéricos) e organização, fornecendo a base para todo o processo analítico.
- NumPy: atuou como suporte fundamental para operações matemáticas e manipulação eficiente de *arrays*. Como tanto o Pandas quanto o *Scikit-learn* utilizam *arrays* NumPy como estrutura subjacente, essa biblioteca foi essencial para garantir a performance e eficiência computacional, especialmente durante o treinamento dos modelos.
- Matplotlib e Seaborn: foram as principais ferramentas para a Análise Exploratória de Dados (AED). O Matplotlib foi empregado na construção de gráficos básicos, como histogramas de distribuição, enquanto o Seaborn, com foco em visualizações estatísticas mais elaboradas, possibilitou análises gráficas mais claras e interpretativas

dos padrões e relações existentes entre variáveis.

- Scikit-learn: constituiu a biblioteca central do processo de modelagem, reunindo os módulos responsáveis pelas etapas de pré-processamento, divisão do *dataset*, treinamento dos modelos e avaliação de desempenho. Entre os principais recursos utilizados, destacam-se:
 - `preprocessing.MinMaxScaler`: empregado para a normalização dos atributos, redimensionando-os para o intervalo [0, 1]. Essa etapa é fundamental para algoritmos sensíveis à escala, como o *Support Vector Machine* (SVM) e o *Multilayer Perceptron* (MLP).
 - `model_selection.train_test_split`: responsável pela separação entre treino e teste, etapa essencial para avaliar a capacidade de generalização dos modelos em dados não vistos.
 - Modelos de Classificação: foram implementadas as classes de modelos de classificação de `sklearn`.
 - Métricas de Avaliação: utilizaram-se as funções `accuracy_score`(Para mostrar a acurácia do meu modelo), `confusion_matrix`(Para mostrar quantos ele classificou correto e quantos errou), `classification_report`(para saber de dados importantes como precisão, sensibilidade e *f1-score*) para mensurar, de forma quantitativa e qualitativa, o desempenho dos classificadores e um algoritmo próprio para mostrar a sensibilidade do modelo.
- Keras com TensorFlow: Para a implementação do modelo de *Deep Learning*, foi utilizado o *framework* Keras, uma API de alto nível que opera sobre o *backend* do TensorFlow. O Keras foi escolhido por sua simplicidade e flexibilidade, permitindo a construção rápida e intuitiva de arquiteturas de redes neurais. Através dele, foi possível definir a estrutura do modelo, incluindo o número de camadas ocultas, a quantidade de neurônios, as funções de ativação e o otimizador, possibilitando a criação de um classificador mais complexo para capturar padrões não-lineares nos dados.

- SHAP (SHapley Additive exPlanations): por fim, com o intuito de ir além da simples avaliação de métricas e incorporar os princípios da Inteligência Artificial Explicável (XAI), empregou-se a biblioteca SHAP. Em linha com práticas recentes na literatura (Sarker et al., 2025), essa ferramenta foi utilizada para interpretar as previsões dos modelos, quantificando a contribuição de cada variável para cada decisão individual. Essa abordagem confere maior transparência ao modelo e contribui para sua aceitação em contextos médicos, nos quais a interpretabilidade é um fator crítico.

Assim, a combinação dessas ferramentas e bibliotecas proporcionou a infraestrutura necessária para o desenvolvimento, avaliação e explicação dos modelos preditivos, assegurando rigor metodológico e confiabilidade nos resultados obtidos.

4.4 Pré-processamento dos Dados

Com o objetivo de assegurar a qualidade do conjunto de dados e otimizar o desempenho dos modelos preditivos, foi implementada uma sequência de etapas de pré-processamento. Esse processo visou tanto a padronização das informações quanto a prevenção de problemas que poderiam comprometer a acurácia e a generalização dos classificadores. As principais etapas foram:

- Carregamento e limpeza dos dados: após a importação do arquivo *gallstone.csv*, verificou-se que algumas colunas estavam formatadas como texto (*object*), em decorrência da presença de caracteres não numéricos. Essas variáveis foram devidamente convertidas para o tipo numérico (*int*), garantindo compatibilidade com os algoritmos de *machine learning*.
- Separação das variáveis: o conjunto de dados foi dividido em duas partes: a matriz de características X, contendo os 38 atributos preditores, e o vetor alvo Y, correspondente à variável *Gallstone Status*. Essa separação é fundamental para estruturar o fluxo de modelagem preditiva.
- Tratando os dados ausentes: As três colunas que apresentavam inconsistências de formato (Vitamina D, TFG e AMV) foram corrigidas. Primeiramente, os valores inválidos foram convertidos em dados ausentes (*Nan*). Em seguida, esses campos vazios foram preenchidos utilizando a técnica de imputação pela média, onde cada valor ausente foi substituído pela média de sua respectiva coluna. Este método foi escolhido para preservar a tendência central da distribuição original de cada atributo.

- Detecção de valores atípicos(*outliers*) Para isso, foi definido um critério heurístico onde um valor era considerado outlier se fosse superior a 1,5 vezes ou inferior a 0,5 vezes a média de sua respectiva coluna. Embora tenham sido detectados outliers em diversas variáveis, a sua proporção em nenhuma delas excedeu o limiar de 20%. Diante disso, optou-se por manter todos os atributos, preservando a variabilidade original dos dados.
- Normalização das variáveis: Com o objetivo de padronizar a escala das variáveis e garantir que todos os atributos tivessem a mesma ordem de magnitude, foi realizado um processo de normalização. As características numéricas do conjunto de dados foram transformadas para o intervalo [0, 1] utilizando o MinMaxScaler da biblioteca Scikit-learn.

Esse fluxo de pré-processamento assegurou que os dados estivessem em condições ideais para a fase seguinte de modelagem, reduzindo vieses e preservando a integridade metodológica do estudo.

4.5 Estratégia Experimental de Seleção de Atributos

Para otimizar os modelos de classificação e aumentar a interpretabilidade dos resultados, foi conduzido um processo experimental de seleção de atributos. O objetivo desta etapa foi identificar se a remoção de variáveis preditoras de menor relevância poderia reduzir a complexidade do modelo e o tempo de execução, sem comprometer a performance preditiva.

Para determinar o conjunto ótimo de variáveis para a modelagem final, compararam-se três abordagens distintas:

- Seleção via Regularização L1 (Lasso): Utilizando os coeficientes de um modelo Lasso, foi aplicado um limiar de importância de 0,0000001 para selecionar apenas as variáveis mais influentes. Esta abordagem resultou em uma redução drástica para apenas 13 atributos (TIBSHIRANI, 1996). No entanto, em testes preliminares, os modelos treinados com este subconjunto apresentaram uma queda significativa de desempenho, atingindo uma acurácia máxima de apenas 73%, motivo pelo qual esta estratégia foi descartada.
- Seleção via Importância de Atributos (*Random Forest*): Empregando os scores de importância gerados pelo *Random Forest*, foram testados diferentes limites de corte

(ZHANG et al., 2018). Uma remoção conservadora dos 2 atributos menos importantes resultou em uma acurácia de 88% com o modelo *Gradient Boosting*. Contudo, observou-se que a performance diminuía progressivamente à medida que mais atributos eram removidos.

- Filtragem de Atributos Categóricos por Frequência: Foi aplicado um critério de remoção para variáveis categóricas que possuíam categorias raras, definidas como aquelas com menos de 0,5% de representatividade no conjunto de dados. Esta abordagem resultou na eliminação de 4 atributos, que correspondiam a categorias específicas de baixa frequência dentro de variáveis como Comorbidade e Acúmulo de Gordura Hepática (HFA).

Após a comparação experimental, constatou-se que a terceira estratégia (Filtragem por Frequência) foi a que apresentou o melhor equilíbrio entre as técnicas de redução. Entretanto, ao comparar os resultados com o uso integral dos dados, observou-se que a manutenção de todos os atributos proporciona a melhor robustez: sem remover nenhum atributo, o melhor resultado preliminar foi obtido usando o *Gradient Boosting Classifier* com 0,91 de acurácia.

Portanto, a decisão metodológica final foi utilizar o conjunto de dados completo com os 38 atributos. A decisão foi justificada pelo fato de que as estratégias de redução de dimensionalidade testadas não ofereceram um benefício claro de performance em comparação ao modelo completo, indicando que a integridade dos dados originais era essencial para maximizar a acurácia.

4.6 Modelos de *Machine Learning* Testados

Para conduzir uma investigação abrangente e identificar o classificador de melhor desempenho para a predição da doença do cálculo biliar, foi implementado e avaliado um conjunto diversificado de 17 algoritmos de aprendizado de máquina. O portfólio de modelos selecionados abrangeu: Regressão Logística, *Linear Discriminant Analysis* (LDA), *Quadratic Discriminant Analysis* (QDA), *Gaussian Naive Bayes*, *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *SGD Classifier*, *Decision Tree*, *Random Forest*, *Extra Trees*, *AdaBoost*, *Gradient Boosting*, *Bagging Classifier*, *HistGradientBoosting*, *XGBoost*, Rede Neural Artificial (MLP) e o *Dummy Classifier* (utilizado como *baseline*). A escolha por uma gama tão ampla de modelos foi uma decisão metodológica para garantir uma análise comparativa robusta e imparcial, explorando diferentes paradigmas do aprendizado de

máquina. Ao avaliar algoritmos de distintas naturezas, como modelos lineares, probabilísticos, baseados em proximidade, *ensembles* e redes neurais, buscou-se identificar qual abordagem se adapta melhor à complexidade dos dados clínicos e de bioimpedância. Os modelos foram aprofundados no Capítulo 2.

A estratégia de treinamento e validação foi estruturada da seguinte forma:

- Divisão Primária dos Dados: Inicialmente, o *dataset* foi particionado em dois conjuntos: 70% para treinamento e 30% para teste, utilizando a função `train_test_split` da biblioteca Scikit-learn. Para assegurar a reproduzibilidade dos experimentos, o parâmetro `random_state` foi fixado em 42.
- Otimização e Validação Cruzada: O conjunto de teste foi mantido isolado (*held-out*) é utilizado unicamente na avaliação final dos modelos. Todo o processo de otimização de hiperparâmetros foi realizado exclusivamente no conjunto de treinamento, empregando-se o método Grid Search com Validação Cruzada (GridSearchCV). Este processo realiza uma busca exaustiva por uma "grade" (`param_grid`) de hiperparâmetros pré-definidos.
- Execução do GridSearchCV: Para avaliar robustamente cada combinação de parâmetros, o GridSearchCV executa internamente uma validação cruzada de 15 *folds* (`cv=15`). A métrica utilizada para selecionar a melhor combinação foi o F1-score (`scoring='f1'`), escolhida por oferecer um balanço eficaz entre precisão e recall. Ao final da busca, o modelo foi automaticamente re-treinado com os melhores hiperparâmetros encontrados em todo o conjunto de treinamento.

4.7 Estratégia de Validação e Métricas de Avaliação

Para medir a eficácia dos modelos de forma objetiva, adotou-se uma abordagem robusta de validação e análise de erros. Uma prática fundamental para a avaliação robusta de modelos de *machine learning* é a divisão do conjunto de dados em subconjuntos independentes. A abordagem mais completa envolve a criação de três partições, cada uma com um propósito distinto:

- Conjunto de Treino: Corresponde à maior parte dos dados e é utilizado exclusivamente para treinar o modelo, ou seja, para que o algoritmo aprenda os padrões e as relações presentes nos dados.
- Conjunto de Validação: É um subconjunto usado para ajustar os hiperparâmetros do modelo e tomar decisões durante a fase de desenvolvimento. Por exemplo, ao testar

diferentes configurações de um algoritmo, o desempenho em cada uma delas é medido no conjunto de validação. Isso evita que o conjunto de teste seja usado para "espiar" e otimizar o modelo, o que levaria a uma avaliação de desempenho excessivamente otimista.

- Conjunto de Teste: Este subconjunto é mantido completamente isolado durante todo o processo de treinamento e otimização. Ele é utilizado apenas uma vez, no final, para fornecer uma estimativa imparcial da capacidade de generalização do modelo final em dados do mundo real, que ele nunca viu antes.

Neste estudo, a função do Conjunto de Validação foi desempenhada pela técnica de Validação Cruzada (*Cross-Validation*) durante a otimização (GridSearchCV). Nela, o conjunto de treino foi subdividido em k partes, e o modelo foi treinado e validado k vezes, permitindo uma otimização de hiperparâmetros estável sem reduzir drasticamente a quantidade de dados disponíveis para o treino.

Métricas e matriz de confusão, a avaliação final no conjunto de teste baseou-se em métricas consolidadas e na análise da Matriz de Confusão, uma tabela que resume os acertos e erros do classificador em quatro quadrantes:

- Verdadeiros Positivos (VP): Pacientes com cálculo biliar corretamente classificados.
 - Verdadeiros Negativos (VN): Pacientes saudáveis corretamente classificados.
 - Falsos Positivos (FP): Erro Tipo I (paciente saudável classificado como doente).
 - Falsos Negativos (FN): Erro Tipo II (paciente doente classificado como saudável).
- Este é o erro mais crítico no contexto médico.

A partir da matriz, foram calculadas as seguintes métricas:

- Acurácia: Proporção global de acertos.
- Precisão: Proporção de casos positivos reais entre as previsões positivas.
- Especificidade: Capacidade de identificar corretamente os casos negativos.
- Sensibilidade (Recall): Capacidade de identificar corretamente os casos positivos (minimizar Falsos Negativos).
- F1-Score: Média harmônica entre precisão e recall, ideal para avaliar o equilíbrio do modelo.

Por fim, a análise foi complementada pela interpretabilidade visual através dos gráficos SHAP (Summary Plot e Dependence Plots), permitindo validar clinicamente as decisões do modelo. Logo abaixo podemos observar a Figura 2 que demonstra um exemplo de matriz de confusão.

Figura 2 - Exemplo de Matriz de Confusão.

		Detectada	
		Sim	Não
Real	Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: <https://medium.com/@mateuspdua/machine-learning-matriz-de-confus%C3%A3o-524618e0402f>.

4.8 Métricas de Avaliação de Desempenho

A avaliação do desempenho dos modelos foi realizada no conjunto de teste (dados não vistos), utilizando métricas amplamente consolidadas em problemas de classificação binária. Essas métricas possibilitam analisar tanto a performance global quanto os aspectos mais específicos da capacidade preditiva de cada algoritmo:

- Acurácia (*Accuracy*): representa a proporção de previsões corretas em relação ao total de instâncias. É uma métrica global, especialmente informativa em conjuntos de dados平衡ados, como o presente estudo.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

em que VP são os verdadeiros positivos, VN os verdadeiros negativos, FP os falsos positivos e FN os falsos negativos.

- Matriz de Confusão: estrutura em forma de tabela que detalha o desempenho do classificador, evidenciando as quantidades de acertos e erros para cada classe. Ela permite verificar diretamente a distribuição de VP, VN, FP e FN, oferecendo uma visão mais granular do comportamento do modelo.
- Precisão (*Precision*): mede, dentre todas as previsões positivas feitas pelo modelo, qual proporção corresponde efetivamente a casos positivos. Essa métrica é particularmente relevante em cenários nos quais o custo de falsos positivos é elevado.

$$\text{Precisão} = \frac{VP}{VP + FP}$$

- Especificidade (*Specificity*): mede, dentre todos os casos que são efetivamente

negativos (saudáveis), qual proporção o modelo conseguiu classificar corretamente como negativos. Essa métrica é particularmente relevante em cenários nos quais o custo de um falso positivo (um "alarme falso") é elevado.

$$Especificidade = \frac{VN}{VN + FP}$$

- Revocação ou Sensibilidade (*Recall*): quantifica, dentre todas as instâncias que realmente pertencem à classe positiva, quantas foram corretamente identificadas pelo modelo. É especialmente útil quando se busca reduzir o número de falsos negativos.

$$Recall = \frac{VP}{VP + FN}$$

- *F1-Score*: consiste na média harmônica entre precisão e revocação, fornecendo uma métrica única que equilibra ambas. É indicada em cenários nos quais existe assimetria entre o custo de falsos positivos e falsos negativos.

$$F1 - Score = 2 * \frac{Precisão * Recall}{Precisão + Recall}$$

Além das métricas de desempenho quantitativas, a metodologia deste trabalho incluiu uma análise de interpretabilidade para avaliar a validade clínica do modelo de melhor desempenho. Para isso, foi empregada a técnica SHAP, já detalhada na fundamentação teórica. A análise SHAP foi aplicada ao classificador *Gradient Boosting* sobre o conjunto de teste para gerar duas visualizações principais:

- Um Summary Plot (Gráfico Resumo), utilizado para identificar a importância global dos atributos, mostrando quais características (ex: PCR, Vitamina D) tiveram o maior impacto nas previsões e a direção desse impacto (positivo ou negativo).
- Gráficos de Dependência, utilizados para inspecionar o efeito de atributos individuais específicos no resultado do modelo.

Esta abordagem metodológica foi essencial para cumprir o segundo objetivo do trabalho: não apenas criar um modelo preciso, mas também compreender por que ele toma suas decisões.

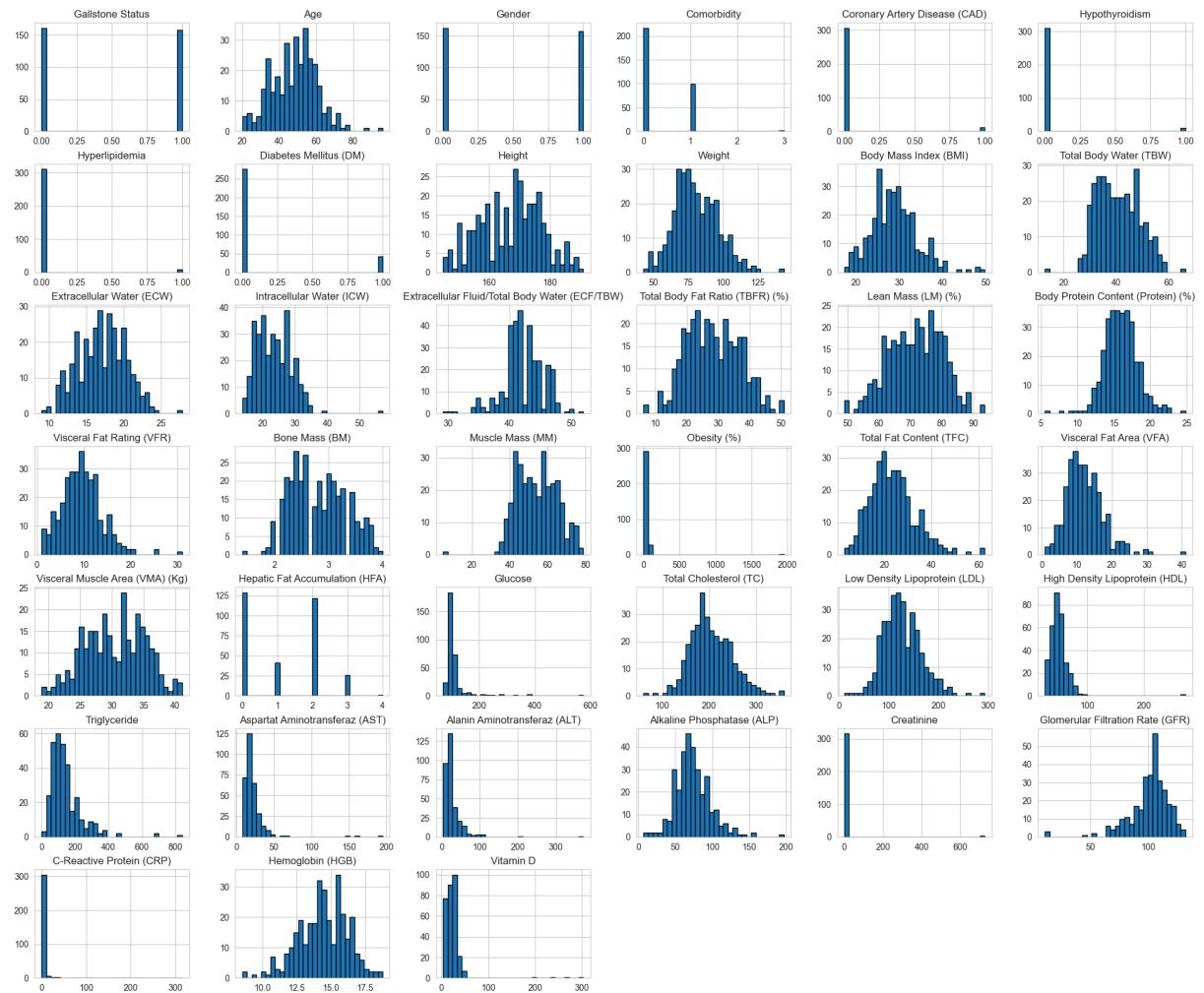
5. RESULTADOS E DISCUSSÃO

Este Capítulo dedica-se à apresentação dos resultados empíricos obtidos através da aplicação da metodologia descrita no Capítulo 4. A exposição dos achados segue uma estrutura lógica: inicia-se com os *insights* obtidos da análise exploratória de dados, depois é falado sobre a eliminação dos atributos de menor importância, depois performance do modelo de *baseline*, que serve como ponto de referência; em seguida, apresenta-se o desempenho comparativo dos um algoritmos de *machine learning* avaliados; realiza-se uma análise mais aprofundada do modelo de melhor performance; e, por fim, detalha-se a análise de importância dos atributos, que revela as variáveis mais influentes no processo de predição.

5.1. *Insights* da Análise Exploratória de Dados

A Figura 3 apresenta um conjunto de histogramas que ilustra a distribuição de frequência de cada variável presente no conjunto de dados, incluindo os 38 atributos preditores e a variável-alvo (*Gallstone Status*). Esta análise visual é um passo fundamental da Análise Exploratória de Dados, permitindo a identificação de padrões, tendências centrais e a dispersão dos dados antes da etapa de modelagem.

Figura 3 - Histograma da distribuição das variáveis.



Fonte: Elaborado pelo autor.

A análise dos histogramas apresentados na Figura 3 permite extrair *insights* valiosos sobre a natureza e a distribuição dos dados, que são cruciais para as etapas subsequentes de modelagem:

- Análise da Variável-Alvo: O histograma da variável *Gallstone Status* evidencia um notável balanceamento entre as duas classes (0 para ausência e 1 para presença da doença). A distribuição quase equitativa entre os grupos constitui um cenário ideal para o treinamento de modelos de classificação, reduzindo o risco de viés e a necessidade de técnicas de reamostragem.
- Análise das Variáveis Categóricas: Os gráficos de contagem para os atributos categóricos revelam padrões distintos. Variáveis como *Gender* (Gênero), *Comorbidity* (Comorbidade) e *Hepatic Fat Accumulation* (HFA) (Acúmulo de Gordura no Fígado)

apresentam uma distribuição relativamente equilibrada entre suas possíveis categorias. Em contrapartida, outras variáveis, como *Hyperlipidemia* (Hiperlipidemia), *Diabetes Mellitus* (DM) e *Coronary Artery Disease* (CAD), exibem um forte desbalanceamento, com uma predominância expressiva de uma única categoria (geralmente a ausência da condição).

- Análise das Variáveis Numéricas: A maioria dos atributos numéricos exibe uma distribuição que se aproxima da normalidade (distribuição gaussiana), com uma clara concentração de valores em torno de uma média central. Este padrão é observado em variáveis como *Age* (Idade), *Body Mass Index* (BMI) e *Weight* (Peso). No entanto, um subconjunto de variáveis, incluindo *Vitamin D*, *Glomerular Filtration Rate* (GFR), *Triglyceride* (Triglicerídeos) e *Aspartat Aminotransferaz* (AST), apresenta uma distribuição assimétrica (*skewed*), com uma cauda mais longa em uma das direções, indicando a presença de valores mais extremos.

Essas observações sobre a forma, a centralidade e a dispersão dos dados são fundamentais para compreender as características intrínsecas do problema e guiar a interpretação dos resultados dos modelos.

5.2 Desempenho do Modelo de Baseline

Conforme a metodologia, o *DummyClassifier* foi utilizado como baseline, configurado para prever sempre a classe majoritária. Dado o balanceamento do conjunto de dados (50,47% de casos positivos), o modelo alcançou uma acurácia de aproximadamente 50,47%. Este valor representa o limiar mínimo de desempenho; qualquer modelo mais complexo deve apresentar uma acurácia superior a este patamar para ser considerado eficaz e agregar valor preditivo.

5.3 Desempenho Comparativo dos Modelos de Classificação

O cerne deste estudo consistiu na avaliação de um portfólio diversificado de 17 algoritmos de classificação. Conforme detalhado na metodologia, cada modelo foi submetido a um rigoroso processo de otimização de hiperparâmetros (via GridSearchCV) utilizando o conjunto de treinamento. O melhor modelo resultante de cada algoritmo foi, então, avaliado no conjunto de testes. A Tabela 2 resume os resultados de performance, apresentando as métricas de acurácia, precisão, recall e F1-Score para a classe positiva (presença de cálculo

biliar).

Tabela 2 - Desempenho Comparativo dos Modelos de Machine Learning no Conjunto de Teste.

Família	Modelo de Classificação	Acurácia	Especificidade	Precisão	<i>Recall</i>	<i>F1-Score</i>
Modelos Baseline	<i>Dummy</i>	0,47	0,51	0,47	0,47	0,47
Modelos Lineares	<i>Regressão Linear</i>	0,75	0,88	0,77	0,75	0,74
Modelos Lineares	<i>Ridge Classifier</i>	0,69	0,81	0,70	0,68	0,68
Modelos Lineares	<i>Stochastic Gradient Descent Classifier</i>	0,67	0,64	0,67	0,67	0,67
Redes Neurais Artificiais	<i>Deep Learning</i>	0,72	0,81	0,73	0,71	0,71
Redes Neurais Artificiais	<i>MLP</i>	0,75	0,90	0,7	0,74	0,74
Máquina de Vetores de Suporte	<i>SVM</i>	0,72	0,81	0,73	0,72	0,71
Baseado em Proximidade	<i>K-NN</i>	0,62	0,60	0,63	0,63	0,62
Probabilístico e Discriminantes	<i>Naive Bayes</i>	0,67	0,75	0,68	0,67	0,67
Probabilístico e Discriminantes	<i>Quadratic Discriminant Analysis</i>	0,69	0,90	0,73	0,68	0,67
Probabilístico e Discriminante	<i>Linear Discriminant Analysis</i>	0,72	0,81	0,73	0,72	0,71

S						
Baseado Em Árvores e Ensembles	<i>Random Forest</i>	0,81	0,88	0,82	0,81	0,81
Baseado Em Árvores e Ensembles	<i>GradientBoostingClassifier</i>	0,91	0,90	0,91	0,91	0,91
Baseado Em Árvores e Ensembles	<i>XGBoost</i>	0,84	0,88	0,85	0,84	0,84
Baseado Em Árvores e Ensembles	<i>CatBoostClassifier</i>	0,84	0,88	0,85	0,84	0,84
Baseado Em Árvores e Ensembles	<i>ExtraTreesClassifier</i>	0,73	0,81	0,74	0,73	0,73
Baseado Em Árvores e Ensembles	<i>Light Gradient Boosting</i>	0,86	0,84	0,86	0,86	0,86

Fonte: Elaborada pelo autor.

A análise da Tabela 2 revela que os modelos da família de *ensembles* baseados em árvores demonstraram um desempenho superior. Em particular, o *Gradient Boosting Classifier* emergiu como o modelo mais robusto, alcançando uma acurácia de 91% no conjunto de teste, com os hiperparâmetros otimizados (`learning_rate = 0,05`; `max_depth = 3`; `n_estimators = 300`; `subsample = 0,8`).

Este resultado do *Gradient Boosting Classifier* é particularmente significativo. O modelo alcançou um desempenho robusto e equilibrado em todas as métricas, obtendo Acurácia de 91%, Especificidade de 90%, Precisão de 91%, *Recall* de 91% e F1-Score de 91%. Este desempenho geral, quando contextualizado com a literatura, supera os *benchmarks* estabelecidos para este mesmo conjunto de dados. A acurácia de 91% é superior aos 85,42% reportados por Esen et al. (2024), que também identificaram o Gradient Boosting como o melhor classificador, e também supera a performance de 79,17% do classificador Random

Forest otimizado de Sarker et al. (2025).

É relevante notar que todos os modelos complexos avaliados apresentaram um desempenho superior ao do modelo de *baseline*, confirmando que todos foram capazes de aprender padrões preditivos relevantes nos dados. Dessa forma, os resultados validam a escolha do *Gradient Boosting* como o algoritmo mais adequado para este problema e demonstram que a otimização de hiperparâmetros permitiu alcançar um novo estado da arte para este conjunto de dados específico.

Dentre os modelos de melhor desempenho, o *Light Gradient Boosting* (LightGBM) destacou-se, alcançando a segunda posição na análise comparativa. O classificador demonstrou uma performance robusta, com uma acurácia de 86% no conjunto de testes.

É relevante notar que, mesmo como o segundo melhor modelo, o desempenho do *LightGBM* supera os *benchmarks* encontrados na literatura para este mesmo conjunto de dados. Sua acurácia de 86% é superior aos 85,42% alcançados pelo *Gradient Boosting* no estudo de Esen et al. (2024), e também excede significativamente os 79,17% obtidos pelo *Random Forest* otimizado de Sarker et al. (2025).

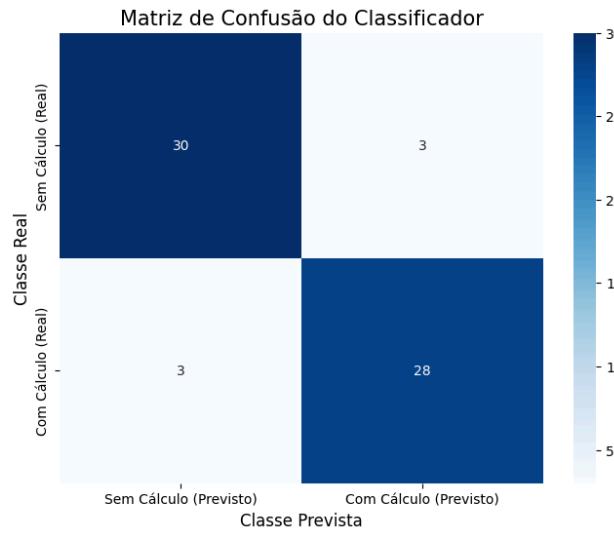
Este resultado reforça a eficácia das implementações modernas de *gradient boosting* para este problema, posicionando o *LightGBM* como uma alternativa de altíssima performance, superada apenas pelo classificador de melhor desempenho deste estudo.

Completando o pódio dos melhores desempenhos, a terceira posição foi compartilhada por dois outros algoritmos da família *gradient boosting*: o *XGBoost* e o *CatBoostClassifier*. Ambos os modelos apresentaram resultados idênticos e robustos, alcançando 84% de acurácia. O desempenho consistente entre as diferentes implementações de *gradient boosting* reforça a superioridade desta abordagem para o presente conjunto de dados.

5.4 Análise Detalhada dos Três Modelos de Melhor Performance

Para uma análise mais aprofundada dos tipos de acertos e erros do modelo *Gradiente Boosting Classifier*, foi gerada sua matriz de confusão, apresentada na Figura 4.

Figura 4 - Matriz de Confusão do Classificador *Gradiente Boosting Classifier*.

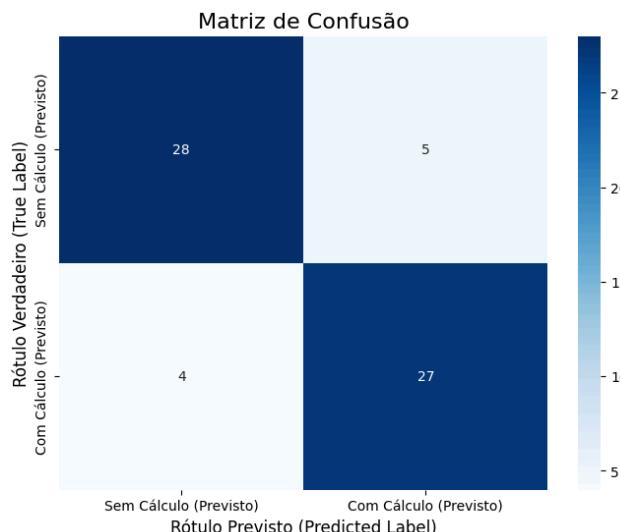


Fonte: Elaborado pelo autor.

A matriz revela que, para o conjunto de teste, o modelo classificou corretamente 28 pacientes como portadores de cálculo biliar e 30 indivíduos saudáveis. Contudo, ocorreram 3 erros do tipo I (controles saudáveis classificados como doentes) e, de maior criticidade clínica, 3 erros do tipo II (pacientes com a doença que não foram identificados pelo modelo).

Para uma análise mais aprofundada dos tipos de acertos e erros do modelo *Light Gradient Boosting Classifier*, foi gerada sua matriz de confusão, apresentada na Figura 5.

Figura 5 - Matriz de Confusão do Classificador *Light Gradient Boosting Classifier*.

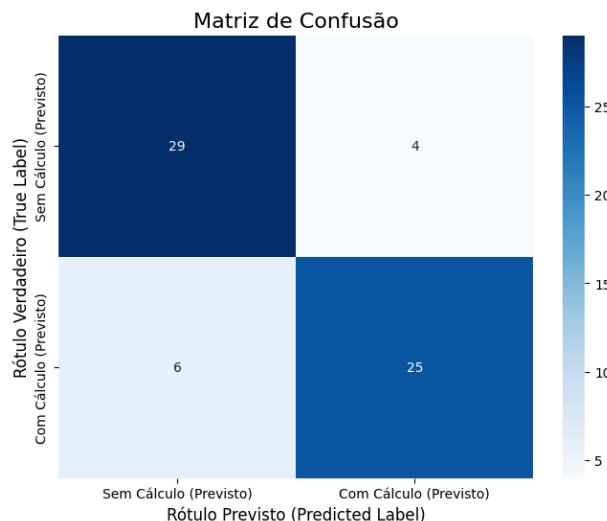


Fonte: Elaborado pelo autor.

A matriz revela que, para o conjunto de teste, o modelo classificou corretamente 27 pacientes como portadores de cálculo biliar e 28 indivíduos saudáveis. Contudo, ocorreram 5 erros do tipo I (controles saudáveis classificados como doentes) e, de maior criticidade clínica, 4 erros do tipo II (pacientes com a doença que não foram identificados pelo modelo).

Para uma análise mais aprofundada dos tipos de acertos e erros do modelo *CatboostClassifier*, foi gerada sua matriz de confusão, apresentada na Figura 6.

Figura 6 - Matriz de Confusão do Classificador *CatBoostClassifier*.

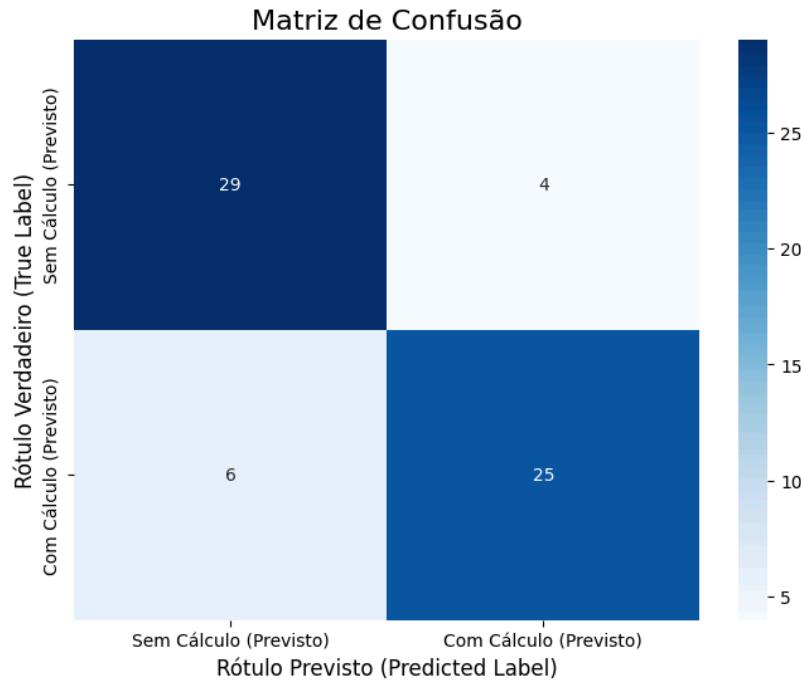


Fonte: Elaborado pelo autor.

A matriz revela que, para o conjunto de teste, o modelo classificou corretamente 25 pacientes como portadores de cálculo biliar e 29 indivíduos saudáveis. Contudo, ocorreram 4 erros do tipo I (controles saudáveis classificados como doentes) e, de maior criticidade clínica, 6 erros do tipo II (pacientes com a doença que não foram identificados pelo modelo).

Para uma análise mais aprofundada dos tipos de acertos e erros do modelo *XGBoost*, foi gerada sua matriz de confusão, apresentada na Figura 7.

Figura 7 - Matriz de Confusão do Classificador *XGBoost*.



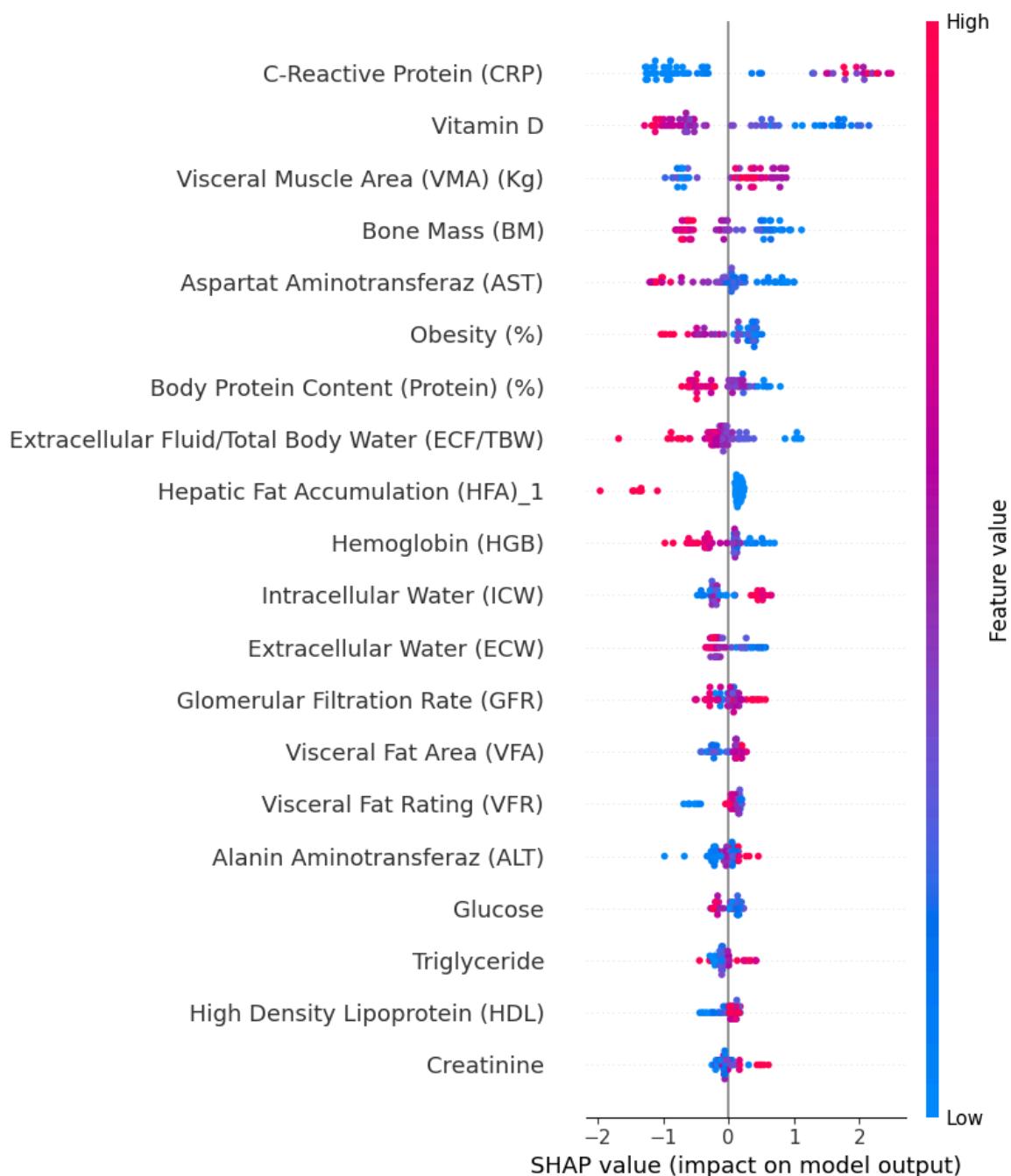
Fonte: Elaborado pelo autor.

A matriz revela que, para o conjunto de teste, o modelo classificou corretamente 25 pacientes como portadores de cálculo biliar e 29 indivíduos saudáveis. Contudo, ocorreram 4 erros do tipo I (controles saudáveis classificados como doentes) e, de maior criticidade clínica, 6 erros do tipo II (pacientes com a doença que não foram identificados pelo modelo).

5.5 Análise de Importância dos Atributos dos Três Modelos de Melhor Performance

Para identificar os fatores mais determinantes para as predições do modelo *Gradiente Boosting Classifier*, foi realizada uma análise de interpretabilidade com a técnica SHAP. A Figura 8 apresenta o gráfico de resumo (*summary plot*), que ordena os atributos por seu impacto médio no modelo.

Figura 8 – Gráfico de Resumo SHAP para os Atributos Mais Importantes do Classificador *Gradiente Boosting Classifier*.



Fonte: Elaborada pelo autor.

A Figura 8 é um Gráfico de Resumo SHAP (*SHAP Summary Plot*), uma visualização que consolida a importância e o impacto de cada atributo nas previsões do modelo. Para uma interpretação correta, o gráfico deve ser analisado a partir de seus três

componentes principais:

- Eixo Vertical (Y) - Importância dos Atributos: As variáveis são ordenadas de cima para baixo, da mais importante para a menos importante. A importância é definida pelo impacto médio que cada atributo tem nas decisões do modelo. Neste gráfico, apenas os atributos mais relevantes são exibidos.
- Eixo Horizontal (X) - Valor SHAP (Impacto na Predição): Este eixo quantifica a influência de um atributo em uma única predição.
 - Valores SHAP Positivos (> 0): Indicam que o valor do atributo "empurrou" a previsão em direção à classe positiva (ex: "Com Cálculo Biliar").
 - Valores SHAP Negativos (< 0): Indicam que o valor do atributo "empurrou" a previsão em direção à classe negativa (ex: "Sem Cálculo Biliar").
- Escala de Cores - Valor do Atributo: A cor de cada ponto representa o valor original do atributo para aquela amostra, normalizado entre 0 e 1.
 - Pontos Vermelhos: Representam valores altos do atributo.
 - Pontos Azuis: Representam valores baixos do atributo.

A combinação desses três componentes permite identificar padrões e relações. A distribuição dos pontos para cada atributo revela se a relação com a variável-alvo é direta (proporcional) ou inversa. Por exemplo, se para um atributo os pontos vermelhos (valores altos) estão concentrados no lado positivo do eixo SHAP, isso indica uma relação direta: quanto maior o valor do atributo, maior a chance de uma predição positiva. O oposto caracteriza uma relação inversa.

A análise SHAP revela que a foi a variável Proteína C-Reativa (PCR) de maior impacto nas predições do modelo. Em seguida, destacam-se vitamina D e Área de Músculo Visceral. O gráfico demonstra, por exemplo, que valores mais altos de obesidade (pontos em vermelho) tendem a aumentar a probabilidade de uma predição positiva para cálculo biliar. Estes achados estão alinhados com a literatura.

A Figura 4 demonstra uma clara correlação positiva: valores elevados de PCR (representados pelos pontos vermelhos) estão consistentemente associados a valores SHAP positivos, indicando que a presença de níveis mais altos deste marcador inflamatório aumenta significativamente a probabilidade de o modelo prever um diagnóstico positivo para cálculo biliar. Inversamente, níveis baixos de PCR (pontos azuis) correspondem a valores SHAP negativos, influenciando o modelo a prever a ausência da doença.

Este comportamento aprendido pelo modelo não é um mero artefato estatístico,

mas reflete uma associação clínica bem documentada na literatura. Um estudo de coorte prospectivo conduzido por Liu et al. (2020) investigou a relação entre a Proteína C-Reativa de alta sensibilidade (hs-CRP) e o risco de desenvolver cálculos biliares em uma população de mais de 96.000 participantes. A pesquisa concluiu que concentrações elevadas de PCR estão independentemente e positivamente associadas a um maior risco para o desenvolvimento da doença do cálculo biliar. Desta forma, os autores estabeleceram a PCR como um fator de risco independente para a formação de cálculos biliares.

A Vitamina D figura como o segundo atributo de maior impacto preditivo, de acordo com a análise do modelo. A interpretação do gráfico SHAP revela uma clara relação inversa: níveis mais baixos de Vitamina D (representados pelos pontos azuis) estão consistentemente associados a valores SHAP positivos, indicando que o modelo aprendeu a relacionar a deficiência desta vitamina com uma maior probabilidade de diagnóstico positivo para cálculo biliar. Inversamente, níveis mais elevados da vitamina influenciam o modelo a prever a ausência da doença. Sendo que esse resultado combina com a publicação de Bin e Zhang(2025), que analisou a associação entre o consumo dietético de Vitamina D e a prevalência de cálculos biliares em adultos norte-americanos. Após ajustar para múltiplas variáveis de confusão, os autores observaram que um maior consumo de Vitamina D estava positivamente associado a uma maior incidência de cálculos biliares.

A Massa Óssea (*Bone Mass*) figura como o quarto atributo de maior importância preditiva, demonstrando uma influência significativa nas decisões do modelo. A análise de interpretabilidade revela uma clara relação inversa: valores mais baixos de massa óssea estão associados a valores SHAP positivos, indicando que o modelo aprendeu a relacionar uma menor densidade óssea a um maior risco de diagnóstico positivo para cálculo biliar. Em contrapartida, valores mais elevados de massa óssea tendem a influenciar a predição para um resultado negativo. Este padrão, identificado pelo algoritmo, é fortemente corroborado por evidências da literatura científica. Um estudo transversal conduzido por Wang et al. (2025) investigou especificamente a associação entre a densidade mineral óssea (DMO) e a prevalência de cálculos biliares em mulheres na pós-menopausa.

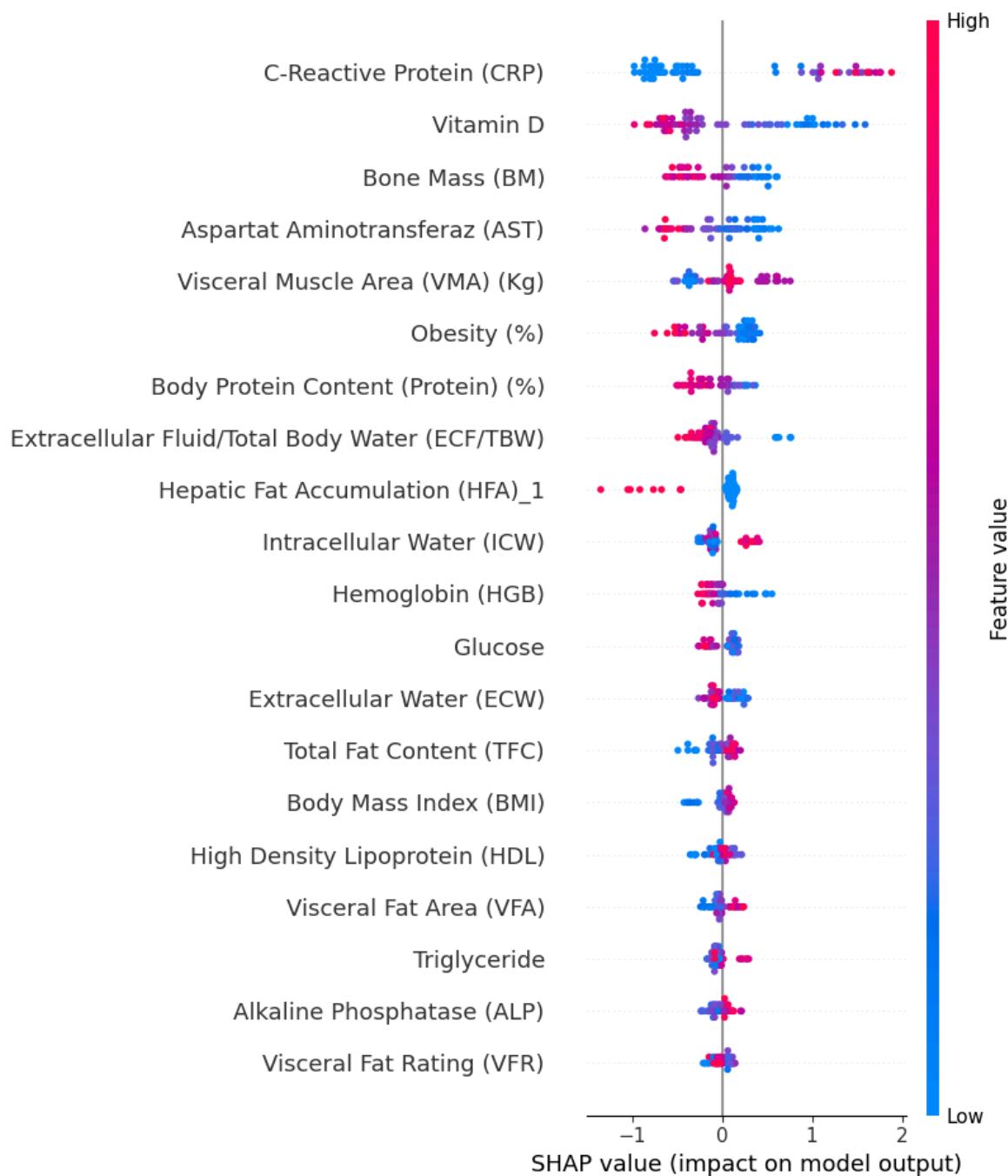
A Área de Músculo Visceral (*Visceral Muscle Area - VMA*) emergiu como o terceiro atributo de maior importância preditiva no modelo desenvolvido. A análise de interpretabilidade aponta para uma relação direta, na qual valores mais elevados de VMA estão associados a uma maior probabilidade de o modelo prever um diagnóstico positivo para cálculo biliar. Uma pesquisa na literatura científica atual não revelou publicações que aprofundem ou expliquem diretamente essa correlação específica.

Curiosamente, este resultado diverge de outras análises sobre o mesmo conjunto de dados. No estudo de Esen et al. (2024), por exemplo, a VMA foi classificada como uma variável de baixa importância preditiva. Essa discrepância sugere que o algoritmo ou a configuração utilizada neste trabalho pode ter sido particularmente sensível a padrões relacionados à VMA que não foram priorizados por outros modelos.

O Aspartato Aminotransferase (AST), uma enzima hepática chave, foi identificado pelo modelo como o quinto atributo mais influente na predição. A análise de interpretabilidade demonstra uma consistente relação inversa: níveis mais elevados de AST estão associados a valores SHAP negativos, indicando que o modelo aprendeu a relacionar uma maior atividade desta enzima com uma menor probabilidade de o paciente ter cálculo biliar. Entretanto não foi achado publicações que abordem o tema sobre a relação entre maiores chances de ter pedra na vesícula e valores baixos de Aspartato aminotransferase.

No estudo de Sarker et al. (2025), a análise SHAP também classificou a AST como a terceira característica mais importante, e sua análise corrobora exatamente a mesma relação inversa encontrada neste trabalho. A Figura 9 o Gráfico de Resumo SHAP (*SHAP Summary Plot*) do classificador *Light Gradient Classifier Boosting*.

Figura 9 – Gráfico de Resumo SHAP para os Atributos Mais Importantes do *Light Classificador Gradiente Boosting Classifier*.



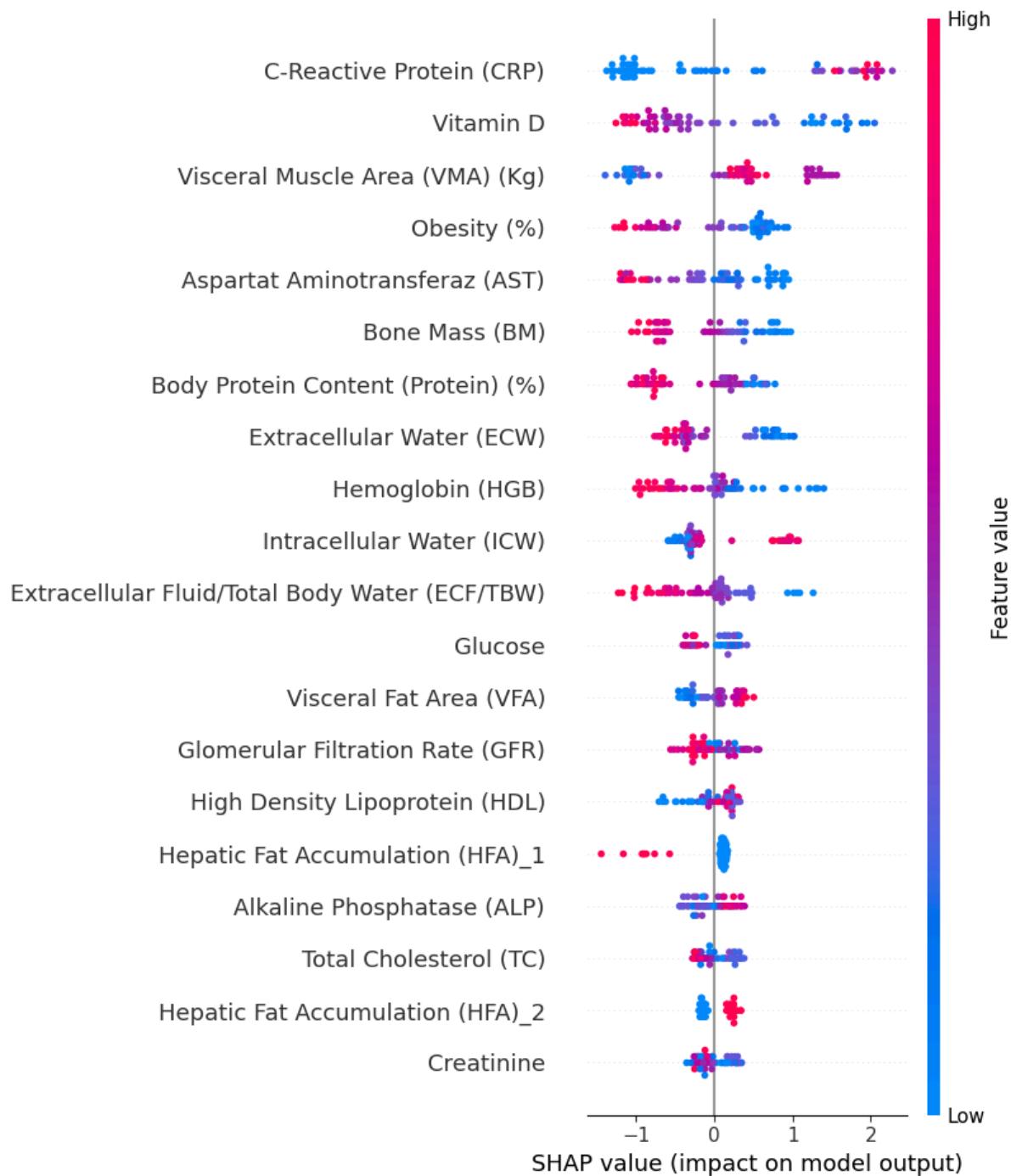
Fonte: Elaborada pelo autor.

A análise de importância de atributos revela uma notável estabilidade nos resultados quando comparada à análise anterior (Figura 8). O conjunto dos cinco preditores mais influentes permaneceu inalterado, reforçando a robustez desses indicadores. Especificamente, os dois atributos de maior impacto, Proteína C-Reativa e Vitamina D, mantiveram suas respectivas primeira e segunda posições, confirmando sua dominância no

modelo preditivo.

Observou-se uma pequena permutação na ordem de classificação entre o terceiro, quarto e quinto atributos mais importantes, uma variação esperada que reflete suas diferenças na ponderação do modelo. Para as demais variáveis de menor *ranking*, ocorreram pequenas alterações em suas posições, mas todas continuam a apresentar valores SHAP médios baixos, confirmando sua contribuição secundária para as decisões do modelo. Essa consistência geral, especialmente no topo da hierarquia de importância, fortalece a confiança na identificação dos principais fatores de risco capturados pelo algoritmo. A Figura 10 é o Gráfico de Resumo SHAP (*SHAP Summary Plot*) do classificador *CatBoostClassifier*.

Figura 10 – Gráfico de Resumo SHAP para os Atributos Mais Importantes do *CatBoostClassifier*.



Fonte: Elaborada pelo autor.

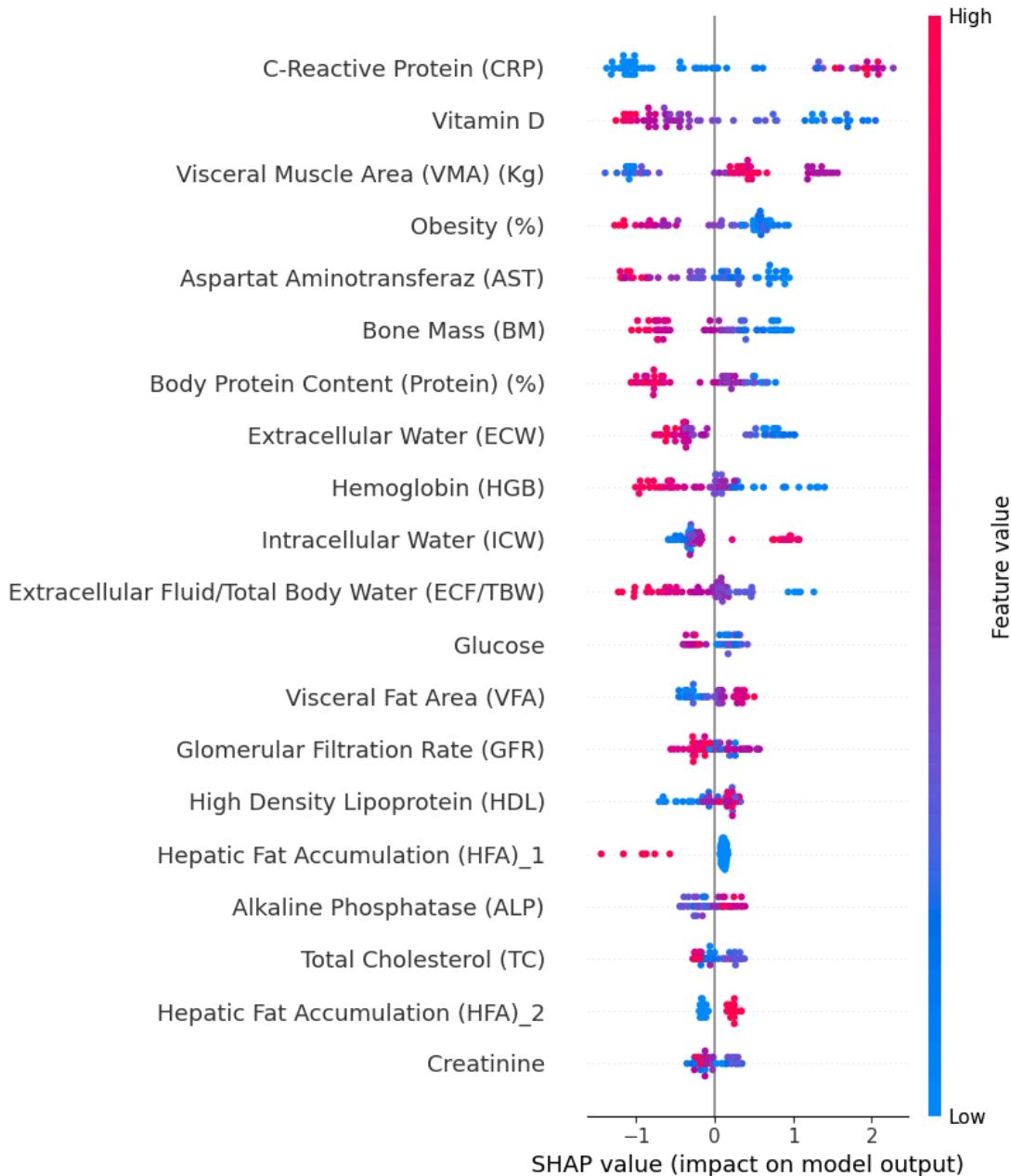
A análise da Figura 10, que detalha a importância dos atributos para o segundo melhor classificador, revela uma notável consistência com os resultados do modelo principal (Figura 8). O conjunto dos atributos mais influentes permanece largamente o mesmo, indicando a estabilidade dos principais preditores da doença. Uma alteração de destaque foi a

ascensão da variável Obesidade (%), que passou da sexta para a quarta posição no *ranking* de importância. A maior relevância atribuída a este atributo está alinhada com o conhecimento clínico consolidado, que aponta a obesidade como um dos principais fatores de risco para a formação de cálculos biliares.

Contudo, a análise aprofundada do gráfico SHAP (Figura 10) revela um padrão contraintuitivo. Embora o modelo tenha identificado a Obesidade (%) como um atributo de alta importância, ele aprendeu uma relação inversa: valores mais altos de obesidade (pontos vermelhos) estão associados a valores SHAP negativos, influenciando o modelo a prever a ausência da doença. Este resultado inesperado pode ser explicado por interações complexas entre as variáveis. É provável que o modelo tenha atribuído o principal poder preditivo a outras características altamente correlacionadas, como o Percentual de Gordura Corporal Total (TBFR) ou a Área de Gordura Visceral (VFA). Assim, a variável Obesidade (%) pode estar atuando como um fator de ajuste secundário dentro do modelo.

Fora esta particularidade, a estrutura geral de importância dos atributos na Figura 9 se assemelha fortemente à da Figura 8, reforçando a consistência dos achados. No entanto, a relação inversamente aprendida para a obesidade destaca a importância da interpretabilidade (XAI) para identificar nuances e complexidades no comportamento do modelo que não seriam visíveis apenas através das métricas de acurácia. A Figura 11 é o Gráfico de Resumo SHAP (*SHAP Summary Plot*) do classificador XGBoost.

Figura 11 – Gráfico de Resumo SHAP para os Atributos Mais Importantes do *XGBoost*.



Fonte: Elaborada pelo autor.

A análise de interpretabilidade da Figura 11, correspondente ao modelo XGBoost, revela uma notável convergência com a análise do CatBoost (Figura 10). O ranking e o impacto das características mais importantes são largamente consistentes entre os dois gráficos, com apenas variações mínimas na magnitude dos valores SHAP. Este achado é coerente com o desempenho preditivo idêntico que ambos os classificadores alcançaram, conforme detalhado na Tabela 2, e reforça a conclusão de que diferentes implementações de *gradient boosting* aprenderam padrões muito semelhantes a partir do conjunto de dados.

6 CONCLUSÃO

6.1 Síntese do Problema de Pesquisa

A doença do cálculo biliar representa um relevante desafio à saúde pública, dada sua alta prevalência e o risco de complicações graves. O diagnóstico, embora eficaz através da ultrassonografia, enfrenta barreiras de custo e acessibilidade. Diante deste cenário, o presente trabalho se propôs a investigar a viabilidade e a eficácia de algoritmos de aprendizado de máquina como uma ferramenta de triagem não invasiva e de baixo custo, utilizando um conjunto de dados clínicos, laboratoriais e de bioimpedância para a predição da doença.

6.2 Síntese dos Principais Achados

Ao longo deste estudo, um portfólio diversificado de 17 algoritmos de classificação foi implementado e avaliado. A análise comparativa demonstrou que os modelos de *ensemble*, em particular o *Gradiente Boosting Classifier*, alcançaram o desempenho mais robusto. O modelo obteve um resultado equilibrado em todas as frentes, com Acurácia de 91%, F1-Score de 91% e *Recall* de 91% no conjunto de teste. Este resultado superou significativamente o modelo de baseline, confirmando a capacidade do *machine learning* de extrair padrões preditivos dos dados.

A análise de interpretabilidade com SHAP foi crucial para validar clinicamente o modelo, revelando que as características de maior impacto estavam alinhadas com os fatores de risco conhecidos da doença. Variáveis ligadas à composição corporal (como Percentual de Obesidade e Massa Magra), a marcadores inflamatórios (Proteína C-Reativa - PCR) e a indicadores metabólicos (Vitamina D, Massa Óssea e AST) emergiram como os preditores mais influentes. A capacidade do modelo de identificar e priorizar estes fatores, que são corroborados pela literatura científica, confere robustez e confiabilidade aos seus resultados.

6.3 Conclusão Geral

Conclui-se que a aplicação de técnicas de *machine learning* sobre dados clínicos e de bioimpedância é uma abordagem viável e de alto potencial para a predição da doença do cálculo biliar. O desempenho alcançado pelo modelo de melhor performance sugere que esta tecnologia pode servir como uma valiosa ferramenta de apoio à decisão clínica. Um sistema

baseado neste modelo poderia ser utilizado na triagem de pacientes, estratificando o risco e ajudando a priorizar a alocação de recursos diagnósticos mais caros, como a ultrassonografia, para os casos de maior suspeita. Desta forma, o estudo demonstra com sucesso que a inteligência artificial pode contribuir para um manejo mais eficiente, proativo e acessível desta prevalente condição.

6.4 Sugestões para Trabalhos Futuros

Com base nos resultados, as seguintes direções para futuras pesquisas são propostas:

- Validação Externa: O passo mais crucial é validar o modelo de melhor desempenho em um conjunto de dados externo, maior e multicêntrico. Isso testaria a sua capacidade de generalização e robustez em diferentes contextos.
- Otimização de Hiperparâmetros: Explorar técnicas avançadas de otimização de hiperparâmetros, para potencialmente aprimorar ainda mais a performance dos classificadores.
- Estudos Longitudinais: Desenvolver pesquisas com acompanhamento de pacientes ao longo do tempo (estudos longitudinais) para criar modelos que possam prever o risco de desenvolvimento futuro da doença, e não apenas detectar a sua presença.
- Desenvolvimento de uma Ferramenta Clínica: Como um passo final, os resultados deste trabalho poderiam ser a base para o desenvolvimento de uma aplicação ou *software* simples, onde um profissional de saúde pudesse inserir os dados do paciente e receber um score de risco.

REFERÊNCIAS

- AHMED, A. S. et al. Advancements in Cholelithiasis Diagnosis: A Systematic Review of Machine Learning Applications in Imaging Analysis. **Cureus**, v. 16, n. 8, p. e66453, ago. 2024.
- BIN, C.; ZHANG, C. The association between vitamin D consumption and gallstones in US adults: A cross-sectional study from the national health and nutrition examination survey. **Journal of the Formosan Medical Association**, 2025.
- BOZDAG, Ahmet et al. Detection of Gallbladder Disease Types Using a Feature Engineering-Based Developed CBIR System. **Diagnostics**, 2025.
- BRASIL. Ministério da Saúde. **Pedra na vesícula (cálculo biliar)**. Biblioteca Virtual em Saúde. Disponível em: <https://bvsms.saude.gov.br/pedra-na-vesicula-calculo-biliar/>. Acesso em: 21 set. 2025.
- CHOLLET, François et al. **Keras**. 2015. Disponível em: <https://keras.io>. Acesso em: 21 set. 2025.
- DENG, L. et al. Relative Fat Mass and Physical Indices as Predictors of Gallstone Formation: Insights From Machine Learning and Logistic Regression. **International Journal of General Medicine**. Disponível em: <https://doi.org/10.2147/IJGM.S507013>. Acesso em: 30 set. 2025.
- ESEN, İrfan et al. Early prediction of gallstone disease with a machine learning-based method from bioimpedance and laboratory data. **Medicine**, 2024.
- GALLSTONE-1 dataset. Irvine, CA: UCI Machine Learning Repository, 2024. Disponível em: <https://archive.ics.uci.edu/dataset/1150/gallstone-1>. Acesso em: 21 set. 2025.
- GRUS, Joel. **Data Science do Zero**: Noções Fundamentais com Python. 2. ed. Rio de Janeiro: Alta Books, 2020.
- HARRIS, Charles R. et al. Array programming with NumPy. **Nature**, 2020.
- HIMI, Shinthi Tasnim et al. MedAi: A Smartwatch-Based Application Framework for Common Diseases Prediction using Machine Learning. **IEEE Access**, [S. l.], 2023. Disponível em: <https://doi.org/10.1109/ACCESS.2023.3236002>. Acesso em: 30 set. 2025.

HUNTER, John D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, 2007.

KHAN ACADEMY. **Identificando outliers com a regra IQR**. Disponível em: <https://pt.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule>. Acesso em: 21 set. 2025.

LIU, T. et al. Relationship between high-sensitivity C reactive protein and the risk of gallstone disease: results from the Kailuan cohort study. **Diabetology & Metabolic Syndrome**, 2020.

LUNDBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. In: **ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS)**, 2017, Long Beach.

MCKINNEY, Wes. Data Structures for Statistical Computing in Python. In: **PROCEEDINGS OF THE 9TH PYTHON IN SCIENCE CONFERENCE**, 2010.

PÁDUA, Mateus. **Machine Learning – Matriz de confusão**. Medium, 10 ago. 2020. Disponível em:<https://medium.com/@mateuspdua/machine-learning-matriz-de-confus%C3%A3o-524618e0402f>. Acesso em: 20 nov. 2025.

PEDREGOSA, Fabian et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, 2011.

PYTHON SOFTWARE FOUNDATION. **Python Language Reference**, version 3.12. 2025. Disponível em: <https://www.python.org>. Acesso em: 21 set. 2025.

SARKER, Proshenjit et al. Gallstone Classification Using Random Forest Optimized by Sand Cat Swarm Optimization Algorithm with SHAP and DiCE-Based Interpretability. **Sensors**, 2025.

SONG, T. et al. U-Next: A Novel Convolution Neural Network With an Aggregation U-Net Architecture for Gallstone Segmentation in CT Images. **IEEE Access**, v. 7, p. 166823–166832, 2019. Disponível em: <https://doi.org/10.1109/ACCESS.2019.2953934>.

TAN, Ludong; JIA, Feng; LIU, Yahui. Advances in research on the role of gut microbiota in the pathogenesis and precision management of gallstone disease. **Frontiers in Medicine**,

2025.

TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267-288, 1996.

WANG, J. et al. Association between bone mineral density and gallstone disease in postmenopausal women: a cross-sectional study. **Frontiers in Public Health**, 2025.

WASKOM, Michael L. seaborn: statistical data visualization. **Journal of Open Source Software**, 2021.

APÊNDICE A

Tabela 1 - Tabela Sobre Os Atributos.

Nome	Tradução	Tipo de Dado	Inconsistência de Formato	Valores Ausentes	Presença de Outliers
<i>Age</i>	Idade	Numérico	0	Não	2
<i>Gender</i>	Gênero	Categórico	0	Não	Não se aplica
<i>Comorbidity</i>	Comorbidade	Categórico	0	Não	Não se aplica
<i>Coronary Artery Disease (CAD)</i>	Doença Arterial Coronariana (DAC)	Categórico	0	Não	Não se aplica
<i>Hypothyroidism</i>	Hipotireoidismo	Categórico	0	Não	Não se aplica
<i>Hyperlipidemia</i>	Hiperlipidemia	Categórico	0	Não	Não se aplica
<i>Diabetes Mellitus (DM)</i>	Diabetes Mellitus (DM)	Categórico	0	Não	Não se aplica
<i>Height</i>	Altura	Numérico	0	Não	0
<i>Weight</i>	Peso	Numérico	0	Não	2
<i>Body Mass Index (BMI)</i>	Índice de Massa Corporal (IMC)	Numérico	0	Não	5
<i>Total Body Water (TBW)</i>	Água Corporal Total (ACT)	Numérico	0	Não	2
<i>Extracellular Water (ECW)</i>	Água Extracelular	Numérico	0	Não	1

	(AEC)				
<i>Intracellular Water (ICW)</i>	Água Intracelular (AIC)	Numérico	0	Não	1
<i>Extracellular Fluid/Total Body Water (ECF/TBW)</i>	Relação Fluido Extracelular/Água Corporal Total	Numérico	0	Não	7
<i>Total Body Fat Ratio (TBFR)</i>	Percentual de Gordura Corporal Total	Numérico	0	Não	0
<i>Lean Mass (LM)</i>	Massa Magra	Numérico	0	Não	0
<i>Body Protein Content (Protein)</i>	Conteúdo de Proteína Corporal	Numérico	0	Não	8
<i>Visceral Fat Rating (VFR)</i>	Classificação da Gordura Visceral	Numérico	0	Não	3
<i>Bone Mass(BM)</i>	Massa Óssea	Numérico	0	Não	0
<i>Muscle Mass (MM)</i>	Massa Muscular	Numérico	0	Não	1
<i>Obesity (%)</i>	Percentual de Obesidade	Numérico	0	Não	7
<i>Total Fat Content (TFC)</i>	Conteúdo Total de Gordura	Numérico	0	Não	7
<i>Visceral Fat Area (VFA)</i>	Área de Gordura Visceral	Numérico	0	Não	5
<i>Visceral Muscle Area</i>	Área de Músculo	Numérico	36	Não	0

(VMA)	Visceral				
<i>Hepatic Fat Accumuation (HFA)</i>	Acúmulo de Gordura no Fígado (Esteatosis Hepática)	Categórico	0	Não	0
<i>Glucose</i>	Glicose	Numérico	0	Não	25
<i>Total Cholesterol (TC)</i>	Colesterol TOTal (CT)	Numérico	0	Não	5
<i>Low Density Lipoprotein (LDL)</i>	Lipoproteína de Baixa Densidade (LDL)	Numérico	0	Não	4
<i>High Density Lipoprotein (HDL)</i>	Lipoproteína de Alta Densidade (HDL)	Numérico	0	Não	5
<i>Triglyceride</i>	Triglicerídeos	Numérico	0	Não	20
<i>Aspartat Aminotransferaz (AST)</i>	Aspartato Aminotransferase (AST)	Numérico	0	Não	23
<i>Alanin Aminotransferaz (ALT)</i>	Alanina Aminotransferase (ALT)	Numérico	0	Não	27
<i>Alkaline Phosphatase (ALP)</i>	Fosfatase Alcalina (FA)	Numérico	0	Não	12
<i>Creatinine</i>	Creatinina	Numérico	0	Não	4
<i>Glomerular Filtration Rate (GFR)</i>	Taxa de Filtração Glomerular (TFG)	Numérico	10	Não	13

<i>C-Reactive Protein (CRP)</i>	Proteína C-Reativa (PCR)	Numérico	0	Não	34
<i>Hemoglobin (HGB)</i>	Hemoglobina (HGB)	Numérico	0	Não	3
<i>Vitamin D</i>	Vitamina D	Numérico	20	Não	5

Fonte: Elaborado pelo autor.