

UNIVERSIDADE FEDERAL DE OURO PRETO  
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS  
DEPARTAMENTO DE COMPUTAÇÃO

PATRICK ARAUJO DE ALMEIDA

**ANÁLISE PREDITIVA E INTERPRETÁVEL DE DADOS CLÍNICOS EM  
TRAUMATISMO CRANIOENCEFÁLICO LEVE COM XGBOOST E  
SHAP**

Ouro Preto, MG  
2025

PATRICK ARAUJO DE ALMEIDA

**ANÁLISE PREDITIVA E INTERPRETÁVEL DE DADOS CLÍNICOS EM  
TRAUMATISMO CRANIOENCEFÁLICO LEVE COM XGBOOST E SHAP**

Monografia apresentada ao Curso de Ciência da Computação da como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

**Orientador:** Prof. Dr. Rafael Alves Bonfim de Queiroz

Ouro Preto, MG  
2025



## FOLHA DE APROVAÇÃO

**Patrick Araujo de Almeida**

### **Análise Preditiva e Interpretável de Dados Clínicos em Traumatismo Cranioencefálico Leve com XGBoost e SHAP**

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 28 de Agosto de 2025.

#### Membros da banca

Rafael Alves Bonfim de Queiroz (Orientador) - Doutor - Universidade Federal de Ouro Preto  
Bárbara Letícia Rodrigues Milagres (Examinadora) - Bacharel - Universidade Federal de Ouro Preto  
Danielle Emely de Souza Almeida (Examinadora) - Mestre

Rafael Alves Bonfim de Queiroz, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 28/08/2025.



Documento assinado eletronicamente por **Rafael Alves Bonfim de Queiroz, PROFESSOR DE MAGISTERIO SUPERIOR**, em 02/09/2025, às 13:18, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site [http://sei.ufop.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **0965548** e o código CRC **4E78B523**.

# Resumo

O avanço das técnicas de aprendizado de máquina, como o *eXtreme Gradient Boosting* (XGBoost), tem revolucionado a análise de dados complexos, especialmente em áreas críticas como a saúde. Contudo, a complexidade desses modelos pode dificultar sua interpretação, tornando métodos como o *SHapley Additive exPlanations* (SHAP) fundamentais para garantir transparência e confiança. Este trabalho realiza uma análise exploratória de dados de pacientes com traumatismo cranioencefálico (TCE), utilizando a sinergia entre o XGBoost e a interpretabilidade SHAP para prever desfechos emocionais — ansiedade e depressão. Foram treinados modelos XGBoost independentes para cada condição, alcançando desempenho satisfatório. A análise SHAP revelou assinaturas biológicas distintas para cada desfecho: para a ansiedade, o biomarcador LIGHT e a idade emergiram como os fatores de maior impacto, enquanto para a depressão, *Lipocalin-2* e MCP-1 foram as variáveis mais influentes. O estudo demonstra que a combinação do poder preditivo do XGBoost com a clareza explicativa do SHAP é uma ferramenta robusta para identificar preditores clinicamente relevantes, reforçando a complexidade fisiopatológica dos transtornos pós-TCE e a necessidade de abordagens personalizadas.

**Palavras-chave:** XGBoost. SHAP. Interpretabilidade. Traumatismo Cranioencefálico.

# Abstract

The advancement of machine learning techniques, such as eXtreme Gradient Boosting (XGBoost), has revolutionized the analysis of complex data, especially in critical areas like healthcare. However, the complexity of these models can hinder their interpretation, making methods like SHapley Additive exPlanations (SHAP) essential to ensure transparency and trust. This work conducts an exploratory analysis of data from patients with traumatic brain injury (TBI), leveraging the synergy between XGBoost and SHAP interpretability to predict emotional outcomes—anxiety and depression. Independent XGBoost models were trained for each condition, achieving satisfactory performance. The SHAP analysis revealed distinct biological signatures for each outcome: for anxiety, the biomarker LIGHT and age emerged as the most impactful factors, whereas for depression, Lipocalin-2 and MCP-1 were the most influential variables. The study demonstrates that combining the predictive power of XGBoost with the explanatory clarity of SHAP provides a robust tool for identifying clinically relevant predictors, reinforcing the pathophysiological complexity of post-TBI disorders and the need for personalized approaches.

**Keywords:** XGBoost. SHAP. Interpretability. Traumatic Brain Injury.

# Lista de Ilustrações

Figura 4.1 – Resultados das métricas de avaliação $MSE$ , $MAE$ e $R^2$ em função do número de árvores nos modelos de predição. <b>Fonte:</b> Elaboração própria a partir dos resultados dos modelos criados. . . . .	15
Figura 4.2 – Importância dos biomarcadores nas predições via <code>plot_importance</code> . <b>Fonte:</b> Elaboração própria a partir dos resultados dos modelos criados. . . . .	16
Figura 4.3 – Impacto médio dos biomarcadores (valores SHAP) nas predições dos modelos XGBoost. <b>Fonte:</b> Elaboração própria a partir dos resultados dos modelos criados. . . . .	18
Figura 4.4 – Distribuição dos valores SHAP dos biomarcadores nas predições dos modelos XGBoost ( <i>Summary Plot</i> ). <b>Fonte:</b> Elaboração própria a partir dos resultados dos modelos criados. . . . .	19

# Lista de Acrônimos

HADS	<i>Hospital Anxiety and Depression Scale.</i>
IA	Inteligência Artificial.
LSTM	<i>Long Short-Term Memory.</i>
ML	<i>Machine Learning.</i>
RF	<i>Random Forest.</i>
SHAP	<i>SHapley Additive exPlanations.</i>
TCE	Traumatismo Cranioencefálico.
UFMG	Universidade Federal de Minas Gerais.
UFOP	Universidade Federal de Ouro Preto.
XGBoost	<i>eXtreme Gradient Boosting.</i>

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	Justificativa	1
1.2	Objetivos	2
1.3	Organização do Trabalho	2
<b>2</b>	<b>Revisão Bibliográfica</b>	<b>3</b>
2.1	Fundamentação Teórica	3
2.1.1	<i>Machine Learning</i>	3
2.1.2	XGBoost ( <i>eXtreme Gradient Boosting</i> )	4
2.1.3	SHAP ( <i>SHapley Additive exPlanations</i> )	6
2.1.4	XGBoost e SHAP	8
2.1.5	Traumatismo Cranioencefálico	8
2.2	Trabalhos Relacionados	8
<b>3</b>	<b>Metodologia</b>	<b>10</b>
3.1	Caracterização do Banco de Dados e do Problema	10
3.2	Abordagem do Problema de Regressão com XGBoost	10
3.3	Otimização e Prevenção de Sobreajuste do Modelo	11
3.4	Interpretação do Modelo e Identificação de Variáveis com XGBoost	11
3.4.1	Análise Preliminar com XGBoost	11
3.4.2	Análise Aprofundada com a Interpretabilidade SHAP	12
3.5	Ferramentas Computacionais	12
<b>4</b>	<b>Resultados Obtidos</b>	<b>14</b>
4.1	Impacto do Número de Árvores na Predição dos Modelos XGBoost	14
4.2	Análise via <code>plot_importance</code> dos Biomarcadores	14
4.3	Análise via SHAP dos Biomarcadores	17
4.4	Comparação entre os Modelos <i>Random Forest</i> e XGBoost	17
<b>5</b>	<b>Considerações Finais</b>	<b>21</b>
5.1	Trabalhos Futuros	21
	<b>Referências</b>	<b>23</b>

# 1 Introdução

O avanço das técnicas de aprendizado de máquina tem revolucionado diversas áreas do conhecimento, oferecendo ferramentas poderosas para a análise de dados complexos e a predição de desfechos em contextos críticos (LUDEMIR, 2021). Dentre essas técnicas, o *eXtreme Gradient Boosting* (XGBoost) (CHEN; GUESTRIN, 2016) tem-se destacado por sua eficácia e eficiência na construção de modelos preditivos robustos, capazes de lidar com grandes volumes de dados e características esparsas. No entanto, a complexidade inerente a esses modelos muitas vezes dificulta a interpretação de suas decisões, especialmente em domínios onde a transparência e a explicabilidade são essenciais, como na área da saúde. É nesse contexto que métodos de interpretabilidade, como o *SHapley Additive exPlanations* (SHAP) (LUNDBERG; LEE, 2017), ganham relevância, permitindo a análise detalhada do impacto de cada variável nas previsões do modelo.

O traumatismo cranioencefálico (TCE) (MATSUNO et al., 2018; GHAJAR, 2000; FONSECA, 2022) representa um dos desafios mais significativos na medicina contemporânea, sendo uma das principais causas de morte e incapacidade, especialmente entre a população jovem. Caracterizado por alterações na função cerebral decorrentes de forças externas, o TCE frequentemente resulta em lesões secundárias que podem levar a complicações graves, como inchaço cerebral, aumento da pressão intracraniana e isquemia. A complexidade e a variabilidade dos fatores envolvidos no TCE demandam abordagens terapêuticas e preditivas baseadas em evidências, capazes de integrar dados clínicos heterogêneos e identificar padrões subjacentes que possam guiar decisões médicas mais precisas e personalizadas.

Nesse cenário, o aprendizado de máquina emerge como uma ferramenta promissora, com estudos recentes demonstrando a superioridade de algoritmos como o XGBoost na previsão de desfechos clínicos em pacientes com TCE. No entanto, a aplicação desses modelos em contextos clínicos exige não apenas alta precisão preditiva, mas também a capacidade de explicar como as variáveis de entrada influenciam as previsões. É aqui que a combinação entre XGBoost e SHAP se mostra particularmente valiosa. Enquanto o XGBoost oferece um modelo robusto e eficiente para a predição, o SHAP fornece uma estrutura matemática sólida para interpretar as contribuições de cada variável, tanto em nível global quanto em nível local, para previsões globais e individuais (SHAPLEY, 1953).

## 1.1 Justificativa

A relevância da combinação entre XGBoost e SHAP é ainda mais evidente em estudos como o de Fonseca (2022), que avaliou diferentes modelos preditivos em conjuntos de dados pediátricos e adultos com TCE. Os resultados demonstraram que modelos baseados em árvores,

como o XGBoost, superaram métodos tradicionais, como a regressão logística, em termos de precisão e identificação de variáveis-chave, como a pontuação na Escala de Coma de Glasgow, níveis de glicose e tempo de protrombina. Além disso, a integração com o SHAP permitiu uma interpretação detalhada do impacto dessas variáveis, reforçando a utilidade clínica desses modelos.

A aplicação dessas técnicas no tratamento de TCE é particularmente importante devido à complexidade e gravidade dessa condição. O TCE é uma das principais causas de morbimortalidade em jovens, e a capacidade de prever desfechos com precisão e interpretar os fatores que influenciam esses desfechos pode transformar a prática clínica (MATSUNO et al., 2018). A interpretabilidade dos modelos de aprendizado de máquina não apenas aumenta a confiança dos profissionais de saúde nas previsões, mas também permite a identificação de variáveis críticas que podem ser alvo de intervenções terapêuticas.

Além disso, a transparência oferecida pelo SHAP é essencial em um contexto clínico, onde decisões erradas podem ter consequências graves. A capacidade de explicar como e por que um modelo chegou a uma determinada previsão é fundamental para a adoção dessas tecnologias na prática médica. Portanto, a investigação da aplicação do XGBoost e do SHAP no tratamento de TCE não apenas contribui para o avanço da neurotraumatologia, mas também promove a integração de técnicas de aprendizado de máquina em um domínio onde a precisão e a explicabilidade são igualmente importantes.

## 1.2 Objetivos

Diante desse contexto, este trabalho tem como objetivo realizar uma análise exploratória de dados reais de pacientes com TCE, utilizando as técnicas XGBoost e SHAP, com ênfase na importância da interpretabilidade para decisões clínicas mais precisas. Para alcançar esse objetivo, foram definidos os seguintes objetivos específicos:

- Identificar as variáveis mais relevantes na predição dos modelos XGBoost;
- Interpretar os modelos XGBoost por meio da técnica SHAP.

## 1.3 Organização do Trabalho

O Capítulo 1 apresenta uma introdução ao tema, a motivação do trabalho e seus objetivos. O Capítulo 2 trata da fundamentação teórica, com a definição dos conceitos essenciais e uma revisão da literatura, destacando estudos relacionados à área de pesquisa. No Capítulo 3, são descritos as ferramentas e os procedimentos utilizados no desenvolvimento do trabalho. Os Capítulos 4 e 5 correspondem ao desfecho da pesquisa: o Capítulo 4 apresenta os resultados obtidos, enquanto o Capítulo 5 expõe as considerações finais.

## 2 Revisão Bibliográfica

Este capítulo de fundamentação teórica e trabalhos relacionados aborda o uso de Machine Learning, destacando o algoritmo XGBoost por sua eficiência e a técnica SHAP por sua capacidade de interpretar modelos de "caixa-preta". O texto explica como a sinergia entre o poder preditivo do XGBoost e a explicabilidade do SHAP é crucial, especialmente em contextos médicos complexos como o do Traumatismo Cranioencefálico, onde a transparência é essencial. A seção também revisa estudos que aplicaram com sucesso essa metodologia para analisar dados clínicos e identificar biomarcadores influentes, validando a abordagem para a tomada de decisão clínica precisa e responsável.

### 2.1 Fundamentação Teórica

Nesta seção, são apresentadas e discutidas as teorias, os conceitos, os modelos e as ideias já consolidados que fundamentam e sustentam a pesquisa. O objetivo é estabelecer uma conexão entre o tema de estudo e o conhecimento já existente na área.

#### 2.1.1 *Machine Learning*

O *Machine Learning* (ML) é um subcampo da Inteligência Artificial (IA) que se concentra no desenvolvimento de algoritmos capazes de aprender padrões a partir de dados, sem serem explicitamente programados para uma tarefa específica (MITCHELL, 1997). Esses algoritmos permitem que sistemas computacionais melhorem seu desempenho por meio da experiência, adaptando-se a novos dados e generalizando conhecimentos para tomar decisões ou fazer previsões (BISHOP, 2006).

Dentre as principais abordagens de ML, destacam-se:

- **Aprendizado Supervisionado:** Utiliza dados rotulados para treinar modelos de classificação ou regressão, como *Random Forest* (BREIMAN, 2001) e XGBoost (CHEN; GUESTRIN, 2016);
- **Aprendizado Não Supervisionado:** Identifica padrões em dados não rotulados, como em algoritmos de agrupamento (*clustering*) e redução de dimensionalidade;
- **Aprendizado por Reforço:** Baseia-se em *feedbacks* de recompensa para otimizar decisões sequenciais, aplicado em robótica e jogos (MATSUNO et al., 2018).

A eficácia dessas técnicas tem revolucionado áreas como saúde, finanças e automação industrial. No entanto, a complexidade de modelos avançados, como *ensembles* de árvores e

redes neurais, muitas vezes os torna “caixas-pretas”, dificultando a interpretação de suas decisões (LUNDBERG; LEE, 2017). Essa limitação é crítica em domínios como a medicina, onde a transparência é essencial para a adoção clínica.

É nesse contexto que métodos de interpretabilidade, como o SHAP, ganham relevância. Baseado na teoria dos jogos cooperativos (SHAPLEY, 1953), o SHAP quantifica a contribuição de cada variável para as previsões do modelo, tanto globalmente (importância geral das *features*) quanto localmente (explicações para instâncias individuais). Sua integração com algoritmos como o XGBoost permite combinar alta precisão preditiva com explicabilidade, tornando-o ideal para aplicações em que a confiança e a auditabilidade são prioritárias, como no prognóstico de TCE (FONSECA, 2022).

### 2.1.2 XGBoost (*eXtreme Gradient Boosting*)

A descrição a seguir do sistema XGBoost é baseada no trabalho seminal de Chen e Guestrin (2016) (CHEN; GUESTRIN, 2016). O XGBoost é um sistema de aprendizado de máquina baseado em uma técnica de *boosting* de árvores, amplamente utilizado por sua eficácia e eficiência. Projetado para ser escalável e rápido, ele permite que cientistas de dados alcancem resultados de ponta em diversos desafios de aprendizado de máquina.

Uma das principais vantagens do XGBoost é sua habilidade de lidar com dados reais, que frequentemente apresentam características como esparsidade ou valores ausentes. Para isso, ele emprega técnicas otimizadas, como o Esboço de Quantis Ponderados (*Weighted Quantile Sketch*). Este algoritmo permite calcular pontos de divisão aproximados em árvores de decisão, levando em conta os pesos das instâncias, o que é crucial em cenários onde a distribuição dos dados varia significativamente. Matematicamente, o cálculo dos pontos de divisão é baseado na minimização da função de perda regularizada  $\mathcal{L}(\phi)$  dada por:

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k),$$

onde:

- $l(y_i, \hat{y}_i)$  é a função de perda que mede a diferença entre o valor real  $y_i$  e a previsão  $\hat{y}_i$ ,
- $\Omega(f_k)$  é o termo de regularização que penaliza a complexidade do modelo, definido como:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2,$$

onde  $T$  é o número de folhas na árvore,  $w$  os pesos das folhas,  $\gamma$  e  $\lambda$  são parâmetros de regularização que controlam a complexidade do modelo,

- $K$  é o número total de árvores no modelo.

Além disso, o XGBoost utiliza uma abordagem de *boosting* sequencial, onde cada nova árvore é treinada para corrigir os erros residuais das árvores anteriores. A previsão final  $\hat{y}_i$  para a instância  $x_i$  é dada pela soma das previsões de todas as árvores:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i),$$

onde  $f_k(x_i)$  é a contribuição da  $k$ -ésima árvore para a instância  $x_i$ .

O XGBoost também se destaca por sua eficiência em termos de memória e tempo de execução. Com suporte para paralelismo, processamento distribuído e algoritmos de compressão, ele é capaz de lidar com conjuntos de dados massivos sem comprometer o desempenho. Essas características fazem do XGBoost uma solução robusta e amplamente adotada em cenários de aprendizado de máquina de grande escala.

O algoritmo XGBoost pode ser descrito pelo pseudocódigo definido pelo *Algorithm 1*.

---

**Algorithm 1** XGBoost.

---

**Input:** Conjunto de dados  $D = \{(x_i, y_i)\}$ , número de árvores  $K$ , taxa de aprendizado  $\eta$ , profundidade máxima da árvore  $d$ .

**Output:** Modelo XGBoost treinado  $F_K(x)$ .

1 Inicialize o modelo com um valor constante:

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma)$$

2 **for**  $k = 1$  **to**  $K$  **do**

3     Calcule os gradientes e hessianos para cada instância:

4      $g_i = \frac{\partial L(y_i, F_{k-1}(x_i))}{\partial F_{k-1}(x_i)}$

5      $h_i = \frac{\partial^2 L(y_i, F_{k-1}(x_i))}{\partial F_{k-1}(x_i)^2}$

6     Ajuste uma árvore de decisão  $f_k(x)$  para prever os valores negativos dos gradientes:

7      $f_k(x) = \arg \min_f \sum_{i=1}^n \left[ g_i f(x_i) + \frac{1}{2} h_i f(x_i)^2 \right] + \Omega(f_k)$

8     Atualize o modelo:

9      $F_k(x) = F_{k-1}(x) + \eta f_k(x)$

10 **end**

11 **return**  $F_K(x)$  ;

// Retorna o modelo treinado

---

O algoritmo *XGBoost* é um método de aprendizado de máquina baseado em *ensemble* que combina múltiplas árvores de decisão de forma sequencial para melhorar a precisão do modelo. O processo inicia com uma predição constante (como a média dos rótulos) e, em cada iteração, calcula gradientes e hessianos da função de perda em relação às predições atuais. Uma nova árvore é então ajustada para aproximar os resíduos negativos (gradientes), utilizando uma função objetivo que incorpora termos de primeira e segunda ordem para otimização mais eficiente, além de penalizações de regularização que controlam a complexidade do modelo. Cada árvore é

adicionada ao *ensemble* com uma taxa de aprendizado ( $\eta$ ), que reduz a contribuição individual de cada árvore, evitando *overfitting* e permitindo um ajuste mais suave. Após um número pré-definido de iterações ( $K$ ), o modelo final é obtido pela combinação ponderada de todas as árvores, resultando em um preditor robusto e altamente ajustável para tarefas de regressão, classificação ou *ranking*.

### 2.1.3 SHAP (*SHapley Additive exPlanations*)

O SHAP (LUNDBERG; LEE, 2017) é uma técnica de interpretabilidade de modelos de aprendizado de máquina. Ele atribui a cada variável de entrada um valor de importância específico em relação a uma previsão, ajudando os usuários a entender como cada característica influencia os resultados do modelo.

Baseado na teoria dos jogos, o SHAP utiliza valores de Shapley (SHAPLEY, 1953) para criar explicações coerentes e matematicamente sólidas. Esses valores garantem uma abordagem justa e intuitiva, alinhando-se às expectativas humanas de como as variáveis devem impactar os resultados. Estudos mostram que os métodos do SHAP são particularmente eficazes em modelos de classificação, proporcionando maior clareza na distinção entre diferentes classes.

O valor de Shapley para uma variável  $i$  é calculado como:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} (v(S \cup \{i\}) - v(S)),$$

onde:

- $F$  é o conjunto de todas as variáveis do modelo,
- $S$  é um subconjunto de variáveis que não inclui  $i$ ,
- $|S|$  é o tamanho do subconjunto  $S$ ,
- $|F|$  é o número total de variáveis,
- $v(S)$  é o valor da função de predição para o subconjunto  $S$ ,
- $v(S \cup \{i\})$  é o valor da função de predição para o subconjunto  $S$  incluindo a variável  $i$ .

Essa fórmula garante que a contribuição de cada variável seja calculada de forma justa, considerando todas as possíveis combinações de variáveis e suas interações. O valor de Shapley  $\phi_i$  representa a contribuição média marginal da variável  $i$  para a previsão do modelo.

Além disso, o SHAP utiliza uma abordagem aditiva para explicar as previsões do modelo. A previsão final  $\hat{y}$  pode ser decomposta como a soma dos valores de Shapley de todas as variáveis

mais um valor base  $\phi_0$ :

$$\hat{y} = \phi_0 + \sum_{i=1}^M \phi_i,$$

onde:

- $\phi_0$  é o valor base, que representa a previsão média do modelo,
- $\phi_i$  é o valor de Shapley da variável  $i$ ,
- $M$  é o número total de variáveis.

O algoritmo SHAP pode ser descrito pelo pseudocódigo definido pelo *Algorithm 2*.

---

**Algorithm 2** SHAP

---

**Result:** Valores SHAP para a instância  $x$

12 **Input:** Modelo  $f$ , instância  $x$ , número de características  $M$

13 **Output:** Valores SHAP  $\phi_1, \phi_2, \dots, \phi_M$

14 Inicialize os valores SHAP:  $\phi_j = 0$  para  $j = 1, 2, \dots, M$

15 **for** cada característica  $j = 1, 2, \dots, M$  **do**

16     **for** cada subconjunto  $S \subseteq \{1, 2, \dots, M\} \setminus \{j\}$  **do**

17         Calcule o valor da coalizão com a característica  $j$ :  $v(S \cup \{j\}) = f(x_{S \cup \{j\}})$

18         Calcule o valor da coalizão sem a característica  $j$ :  $v(S) = f(x_S)$

19         Calcule a contribuição marginal da característica  $j$ :  $\Delta_j(S) = v(S \cup \{j\}) - v(S)$

20         Pese a contribuição marginal pelo coeficiente SHAP:  $w(S) = \frac{|S|!(M - |S| - 1)!}{M!}$

21         Atualize o valor SHAP da característica  $j$ :  $\phi_j = \phi_j + w(S) \cdot \Delta_j(S)$

22     **end**

23 **end**

24 **return**  $\phi_1, \phi_2, \dots, \phi_M$

---

O algoritmo SHAP calcula valores de contribuição para cada característica (*feature*) de um modelo de aprendizado de máquina  $f$  com base na teoria dos jogos cooperativos. O processo inicia com a inicialização dos valores SHAP ( $\phi_j$ ) como zero para todas as  $M$  características. Para cada característica  $j$ , o algoritmo avalia todas as possíveis coalizões (subconjuntos)  $S$  das demais características. Em cada iteração, são calculados: (i) o valor da predição do modelo quando a característica  $j$  é incluída na coalizão ( $v(S \cup \{j\})$ ), (ii) o valor quando  $j$  é excluída ( $v(S)$ ), e (iii) a contribuição marginal  $\Delta_j(S)$  como a diferença entre esses valores. Cada contribuição é ponderada por um coeficiente  $w(S)$  que considera o tamanho da coalizão  $S$ , garantindo que subconjuntos de diferentes tamanhos tenham pesos adequados (equivalente ao peso de Shapley na teoria dos jogos). O valor SHAP final para cada característica  $j$  ( $\phi_j$ ) é a soma ponderada de todas as suas contribuições marginais em todas as coalizões possíveis, proporcionando uma medida justa e axiomática da importância local da característica para a predição do modelo na instância  $x$ .

## 2.1.4 XGBoost e SHAP

A combinação de XGBoost e SHAP é adotada neste trabalho para criar modelos poderosos e explicáveis. Enquanto o XGBoost constrói modelos robustos baseados em árvores para minimizar o erro de previsão, o SHAP fornece uma interpretação detalhada do impacto de cada variável, tanto em nível global quanto local. Essa sinergia permite:

- **Interpretação Global:** Identificar as variáveis mais influentes no desempenho geral do modelo;
- **Interpretação Local:** Explicar como características específicas influenciam previsões individuais.

Essa combinação é especialmente valiosa em setores onde a explicabilidade é essencial, como saúde, finanças e regulamentação, oferecendo previsões precisas e transparência no processo decisório.

## 2.1.5 Traumatismo Cranioencefálico

O TCE (GHAJAR, 2000) é caracterizado por alterações na função cerebral causadas por forças externas, como impactos, desacelerações rápidas ou explosões. É uma das principais causas de morte e incapacidade em jovens, destacando a importância de abordagens terapêuticas baseadas em evidências para melhorar os resultados.

A lesão cerebral secundária é a principal causa de mortalidade hospitalar após um TCE, frequentemente associada a inchaço cerebral, aumento da pressão intracraniana e isquemia devido à redução na perfusão cerebral.

## 2.2 Trabalhos Relacionados

O aprendizado de máquina tem-se mostrado uma ferramenta promissora para prever desfechos clínicos em pacientes com TCE. Estudos recentes, como o de Matsuo *et al.* (2018), destacam a superioridade dos algoritmos de aprendizado de máquina na identificação de padrões complexos em grandes volumes de dados clínicos.

No estudo de Fonseca (FONSECA, 2022), foram avaliados diferentes modelos preditivos em dois conjuntos de dados:

- **IHPTBI (pediátrico):** Modelos baseados em árvores, como *Random Forest* (2001) e XGBoost, demonstraram desempenho superior ao utilizar todas as variáveis;

- **MIMIC-III (adulto):** Algoritmos mais avançados, como *Long Short-Term Memory* (LSTM) (1997) e *Transformers* (2017), apresentaram melhores resultados em dados temporais.

Esses modelos foram avaliados por sua precisão preditiva e relevância clínica, reforçando a eficácia de combinar abordagens tradicionais e avançadas para diferentes populações de pacientes.

O XGBoost (WANG et al., 2022) tem se destacado na previsão de mortalidade em TCE moderado a grave, superando métodos tradicionais como a regressão logística em termos de precisão e identificação de variáveis-chave, como: pontuação na escala de coma de Glasgow, níveis de glicose e tempo de protrombina. Essa abordagem destaca o potencial do aprendizado de máquina na personalização e melhoria dos cuidados clínicos, contribuindo para avanços significativos na área de neurotraumatologia.

Ampliando a aplicação dessas técnicas ao TCE leve, a dissertação de Almeida (2025) propôs uma abordagem inovadora ao combinar modelos de aprendizado de máquina, como *Random Forest* e XGBoost, com inteligência artificial explicável, por meio da técnica SHAP. O diferencial do trabalho reside não apenas na predição de desfechos clínicos, mas também na identificação e interpretação da influência de biomarcadores séricos específicos. A pesquisa destacou biomarcadores como LIGHT, sCD40L, MPO e MCP-1 como os mais relevantes para diferenciar pacientes com TCE leve de outros grupos e para prever sintomas de ansiedade e depressão. Essa abordagem reforça o potencial dos algoritmos baseados em árvores na análise de dados clínicos de TCE e avança na direção da interpretabilidade dos modelos, oferecendo *insights* sobre mecanismos biológicos subjacentes e contribuindo para uma tomada de decisão clínica mais transparente e informada.

A robustez dessas abordagens é corroborada por sua aplicação bem-sucedida em outras áreas complexas da medicina que também dependem da análise de biomarcadores para a predição de desfechos críticos. No contexto da nefrologia, o trabalho de Souza (SOUZA, 2024) utilizou modelos como *Random Forest* e XGBoost para prever o risco de óbito em pacientes em hemodiálise, a partir de dados clínicos e laboratoriais de rotina. De forma análoga ao estudo de Almeida (2025), Souza empregou a metodologia SHAP para conferir interpretabilidade aos modelos, identificando os biomarcadores com maior impacto na mortalidade. Esse estudo reforça a capacidade da combinação de algoritmos baseados em árvores com XAI em extrair conhecimento clinicamente relevante de dados complexos, validando seu uso em tarefas de prognóstico em diferentes especialidades médicas.

# 3 Metodologia

Esta seção detalha os procedimentos metodológicos adotados para a análise do banco de dados, desde a caracterização inicial até a otimização e interpretação do modelo de aprendizado de máquina. A abordagem foi desenhada para explorar as relações internas dos dados, com foco na interpretabilidade dos resultados.

## 3.1 Caracterização do Banco de Dados e do Problema

O presente estudo utiliza um banco de dados clínico específico para uma tarefa de regressão, cujo objetivo é prever desfechos emocionais em pacientes com TCE leve. Os dados foram originalmente coletados em um estudo longitudinal conduzido no Hospital de Emergência João XXIII, em Belo Horizonte/MG, por meio de uma colaboração com pesquisadores da UFMG, e sua coleta foi devidamente aprovada pelo Comitê de Ética em Pesquisa da UFMG-COEP (CAAE: 49623015.0.0000.5149) (ALMEIDA, 2025).

O conjunto de dados para esta análise é o mesmo empregado na abordagem de regressão da dissertação de Almeida (2025), que passou por um tratamento de remoção de *outliers*. O banco de dados final é composto por 21 amostras de pacientes. Para cada amostra, há 31 variáveis preditoras (*features*), que incluem biomarcadores séricos e não séricos (idade e sexo). As duas variáveis-alvo de interesse são os escores da *Hospital Anxiety and Depression Scale* (HADS): HADS *anxiety* e HADS *depression* (ALMEIDA, 2025).

Dada a natureza exploratória da análise e o número reduzido de amostras, o particionamento dos dados em conjuntos de treinamento e teste não foi realizado. A análise é conduzida utilizando o conjunto de dados completo, com foco na identificação das variáveis mais influentes e na compreensão da lógica do modelo para prever os valores contínuos de ansiedade e depressão, em vez de avaliar sua capacidade de generalização para dados não vistos (ALMEIDA, 2025).

## 3.2 Abordagem do Problema de Regressão com XGBoost

Para a tarefa de regressão, é utilizado o algoritmo XGBoost. Esta escolha se deve ao seu alto desempenho e à sua capacidade de modelar relações complexas e não lineares entre as variáveis. São treinados dois modelos XGBoost distintos e independentes:

1. Um modelo para prever a variável HADS *anxiety*;
2. Um modelo para prever a variável HADS *depression*.

Essa abordagem permite uma análise focada e a identificação de padrões específicos para cada uma das condições de saúde mental avaliadas.

### 3.3 Otimização e Prevenção de Sobreajuste do Modelo

Uma etapa fundamental da metodologia consiste em encontrar uma configuração ótima para cada modelo XGBoost. Devido ao número limitado de amostras, há um risco significativo de sobreajuste (*overfitting*), onde o modelo memoriza os dados de treinamento em vez de aprender padrões generalizáveis (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Para mitigar esse risco de forma robusta, adota-se uma abordagem focada e parcimoniosa, investigando-se o número de árvores (`n_estimators`), que é o principal hiperparâmetro que controla a complexidade do modelo. Manter os demais hiperparâmetros fixos em valores conservadores e padrão evita o “viés de otimização” no processo de validação, garantindo que o desempenho estimado do modelo não seja excessivamente otimista e resultado de um ajuste fortuito a um subconjunto de dados (CAWLEY; TALBOT, 2010).

### 3.4 Interpretação do Modelo e Identificação de Variáveis com XGBoost

Uma etapa fundamental da metodologia consiste na interpretação dos modelos XGBoost treinados para ansiedade e depressão. O objetivo é transcender a simples predição e compreender os mecanismos subjacentes que levam o modelo a tomar suas decisões, garantindo transparência e confiabilidade nos resultados. Para isso, é adotada uma abordagem em duas fases, utilizando tanto a funcionalidade de importância nativa do XGBoost quanto a biblioteca SHAP para uma análise mais aprofundada.

#### 3.4.1 Análise Preliminar com XGBoost

Como passo inicial na interpretação do modelo, utiliza-se a função `plot_importance` da biblioteca XGBoost para estabelecer um ranking da importância global das variáveis (The XGBoost developers, 2024). Esta função permite quantificar a contribuição de cada característica para o desempenho do modelo por meio de diferentes métricas, selecionáveis através do parâmetro `importance_type`. Para este estudo, foram consideradas as seguintes métricas:

- **Ganho (*Gain*):** A métrica principal utilizada, que representa a melhoria média na redução da função de perda (erro do modelo) proporcionada por uma variável sempre que ela é utilizada para uma divisão nas árvores de decisão. É a medida mais relevante para avaliar o impacto preditivo de uma variável.

- **Frequência (*F-score / Weight*):** Quantifica o número de vezes que uma variável foi utilizada para criar uma divisão em todas as árvores. Embora útil, esta métrica pode favorecer variáveis que são frequentemente usadas em divisões de baixo impacto.

A análise destas métricas fornece uma avaliação inicial robusta, permitindo identificar o conjunto de biomarcadores com maior influência global na predição, que será posteriormente investigado em detalhe com a análise SHAP.

### 3.4.2 Análise Aprofundada com a Interpretabilidade SHAP

Para garantir que o modelo final seja não apenas preciso, mas também transparente e explicável, a sua interpretabilidade é aprofundada através da biblioteca SHAP. Esta abordagem permite identificar as variáveis preditoras de maior impacto para HADS *anxiety* e HADS *depression*, fornecendo informações sobre a contribuição de cada característica para as predições do modelo.

Para visualizar essas informações, duas ferramentas principais são geradas:

1. **Gráfico de Barras (Bar Plot):** Este gráfico mostra o impacto médio absoluto dos valores SHAP, ordenando os biomarcadores por sua importância global. Ele é fundamental para destacar as variáveis com maior influência geral nas predições do modelo.
2. **Gráfico de Resumo (Summary Plot):** Esta visualização oferece uma análise mais aprofundada, exibindo a distribuição dos valores SHAP para cada biomarcador e variável clínica. O gráfico detalha a direção e a intensidade do impacto de cada variável, utilizando cores para representar seus valores: o vermelho indica valores altos do biomarcador, enquanto o azul representa valores baixos. Desta forma, é possível analisar crucialmente como os diferentes níveis de cada variável (altos ou baixos) influenciam os resultados da previsão.

## 3.5 Ferramentas Computacionais

Para a execução da metodologia proposta e a obtenção dos resultados desejados, são utilizadas ferramentas computacionais e bibliotecas de alto desempenho.

O ambiente de desenvolvimento é a linguagem Python, em sua versão mais recente, assegurando compatibilidade com bibliotecas modernas e acesso às funcionalidades atualizadas. Todo o processamento é realizado na plataforma *Google Colab*.

O *Google Colab* é um ambiente de notebook Jupyter gratuito que roda na nuvem. Ele permite a escrita e execução de código, uso de textos, gráficos e outros recursos em um único documento, sem a necessidade de nenhuma configuração no computador pessoal. Sua principal

característica é o acesso a recursos de hardware gratuitos, como GPUs e TPUs, o que o torna ideal para projetos de aprendizado de máquina e ciência de dados (Google, 2025).

A biblioteca XGBoost é utilizada para a construção do modelo de aprendizado de máquina, sendo escolhida por sua eficiência no tratamento de grandes volumes de dados e pela comprovada eficácia em tarefas complexas de classificação e regressão.

A biblioteca SHAP é aplicada para a interpretação e explicação das previsões geradas pelo modelo, permitindo a análise da contribuição de cada variável e promovendo transparência e confiabilidade nos resultados obtidos.

## 4 Resultados Obtidos

Este capítulo apresenta os resultados obtidos neste trabalho. Na Seção 4.1, analisa-se o impacto do número de árvores na predição dos modelos XGBoost. A Seção 4.2 apresenta os resultados da análise via `plot_importance`, destacando os biomarcadores mais relevantes nas predições. A Seção 4.3 detalha a análise dos biomarcadores por meio da técnica SHAP. Por fim, a Seção 4.4 compara os resultados dos modelos *Random Forest* (ALMEIDA, 2025) e XGBoost.

### 4.1 Impacto do Número de Árvores na Predição dos Modelos XGBoost

A Figura 4.1 apresenta os gráficos das métricas de avaliação ( $MSE$ ,  $MAE$  e  $R^2$ ) em função do número de árvores para os modelos de predição de ansiedade (HADS *Anxiety*) e depressão (HADS *Depression*).

Para ambos os modelos, observa-se uma melhoria significativa das métricas nos estágios iniciais (até aproximadamente 20 árvores), seguida por uma estabilização do desempenho, indicando que o modelo atinge sua capacidade máxima de generalização.

O coeficiente de determinação ( $R^2$ ) apresenta valores satisfatórios, sugerindo que o XGBoost consegue explicar uma parcela relevante da variabilidade dos dados. As curvas de  $MSE$  e  $MAE$  apresentam comportamentos similares: redução acentuada no início (1–20 árvores), com convergência progressiva após o ponto de saturação.

Os resultados da Figura 4.1 evidenciam a eficácia dos modelos XGBoost, com desempenho estável a partir de uma complexidade intermediária (40–60 árvores). A estabilização das métricas sugere que a adição de mais árvores não melhora significativamente a capacidade preditiva, conforme indicado por (CHEN; GUESTRIN, 2016).

### 4.2 Análise via `plot_importance` dos Biomarcadores

Os gráficos de importância apresentados na Figura 4.2 revelam os biomarcadores mais relevantes nas predições. Esses gráficos foram gerados com a função `xgb.plot_importance` (20 árvores).

Para a predição da variável-alvo HADS *Anxiety*, os cinco biomarcadores mais importantes foram: LIGHT, PAI-1, *age* (idade), *Angiotensin 2* e *Angiotensin 1-7*.

Já para HADS *Depression*, os cinco principais foram: *Lipocalin-2*, NCAM, BDNF, *Cathepsin D* e *Neuropilin-1*.

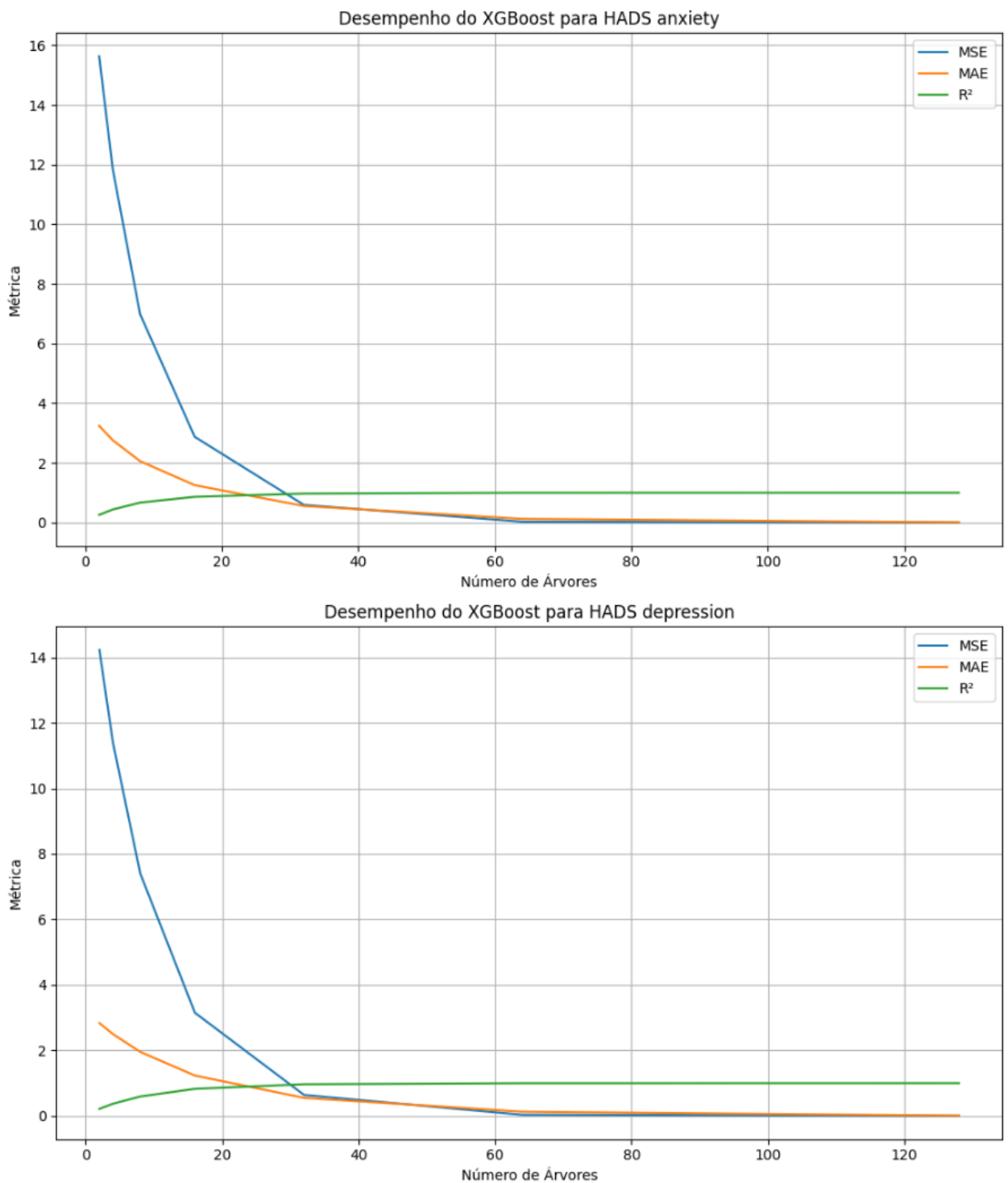


Figura 4.1 – Resultados das métricas de avaliação  $MSE$ ,  $MAE$  e  $R^2$  em função do número de árvores nos modelos de predição. **Fonte:** Elaboração própria a partir dos resultados dos modelos criados.

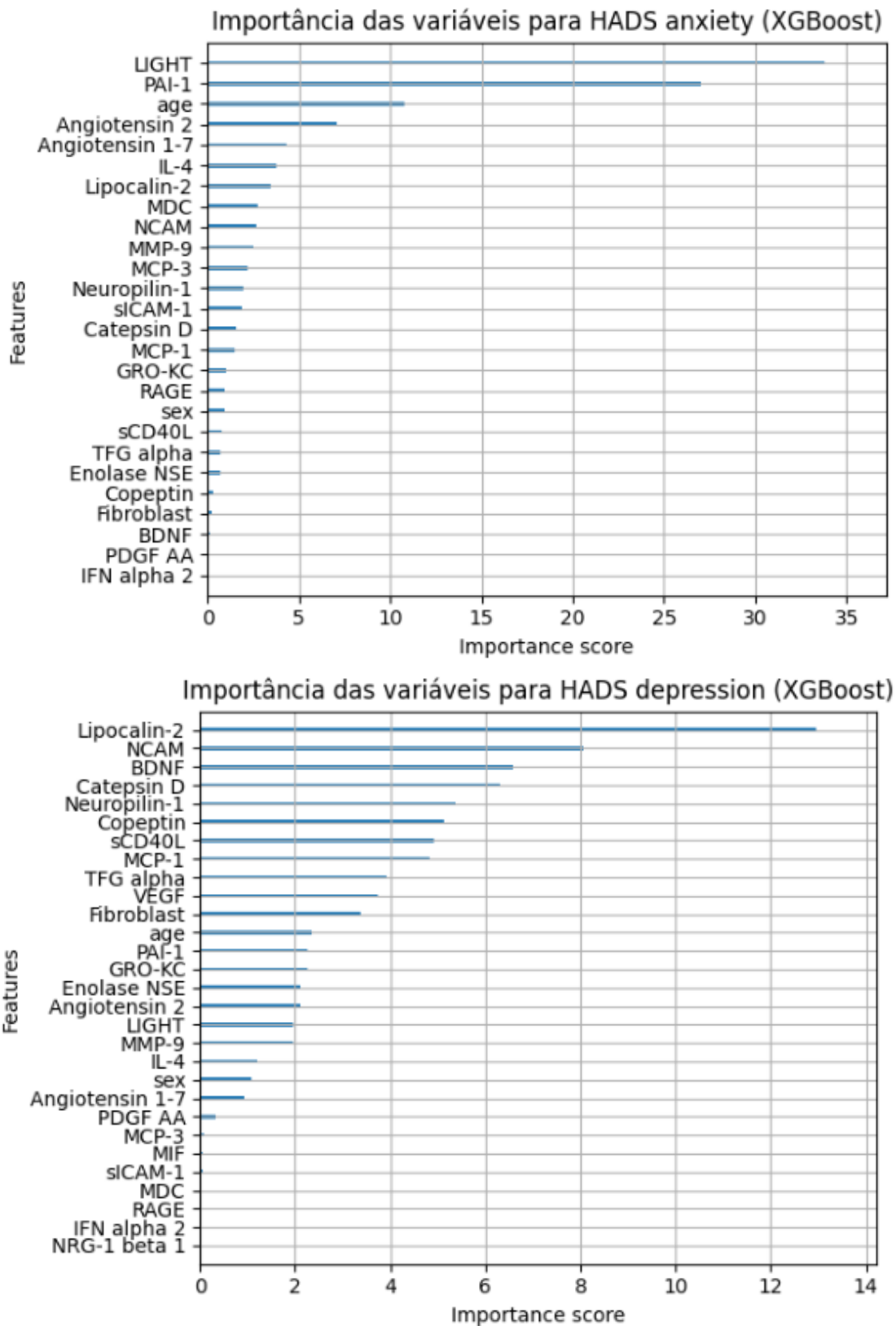


Figura 4.2 – Importância dos biomarcadores nas predições via plot\_importance. **Fonte:** Elaboração própria a partir dos resultados dos modelos criados.

Destacam-se os seguintes pontos:

- **Consistência biológica:** os biomarcadores mais relevantes estão associados a vias reconhecidas na fisiopatologia dos transtornos como MCP-1, sCD40L e *Lipocalin-2*;
- **Especificidade:** os perfis distintos entre ansiedade e depressão sugerem que o modelo capta padrões específicos a cada condição.

### 4.3 Análise via SHAP dos Biomarcadores

Esta seção apresenta os resultados da aplicação da técnica SHAP em conjunto com o XGBoost para analisar a importância e o impacto dos biomarcadores nas previsões de ansiedade e depressão.

Para HADS *Anxiety*, o gráfico de barras (Figura 4.3) mostra como principais biomarcadores: LIGHT, *age*, PAI-1, *Lipocalin-2* e *sex*.

Para HADS *Depression*, os mais importantes foram: *Lipocalin-2*, MCP-1, *age*, *Cathepsin D* e sCD40L.

Considerando a Figura 4.4, tem-se:

#### Para HADS *Anxiety*:

- LIGHT e *age*: valores baixos favorecem a previsão da ansiedade;
- *Lipocalin-2*: valores altos favorecem a previsão;
- *sex*: o valor “1” (masculino) favorece a previsão da ansiedade.

#### Para HADS *Depression*:

- *Lipocalin-2* e *Cathepsin D*: valores altos favorecem a previsão;
- *age* e SCD40L: valores baixos favorecem a previsão.

A análise evidencia que, embora alguns biomarcadores (como *age* e *Lipocalin-2*) sejam relevantes para ambos os desfechos, a magnitude do impacto e os biomarcadores que os acompanham variam entre ansiedade e depressão. Isso valida a modelagem separada e interpretável para cada condição.

### 4.4 Comparação entre os Modelos *Random Forest* e XGBoost

A análise comparativa entre os resultados de Almeida (2025), que utilizou o modelo *Random Forest* (RF), e os do presente trabalho, que utilizou o XGBoost, mostra que ambos os modelos foram eficazes na identificação de biomarcadores associados a sintomas de ansiedade

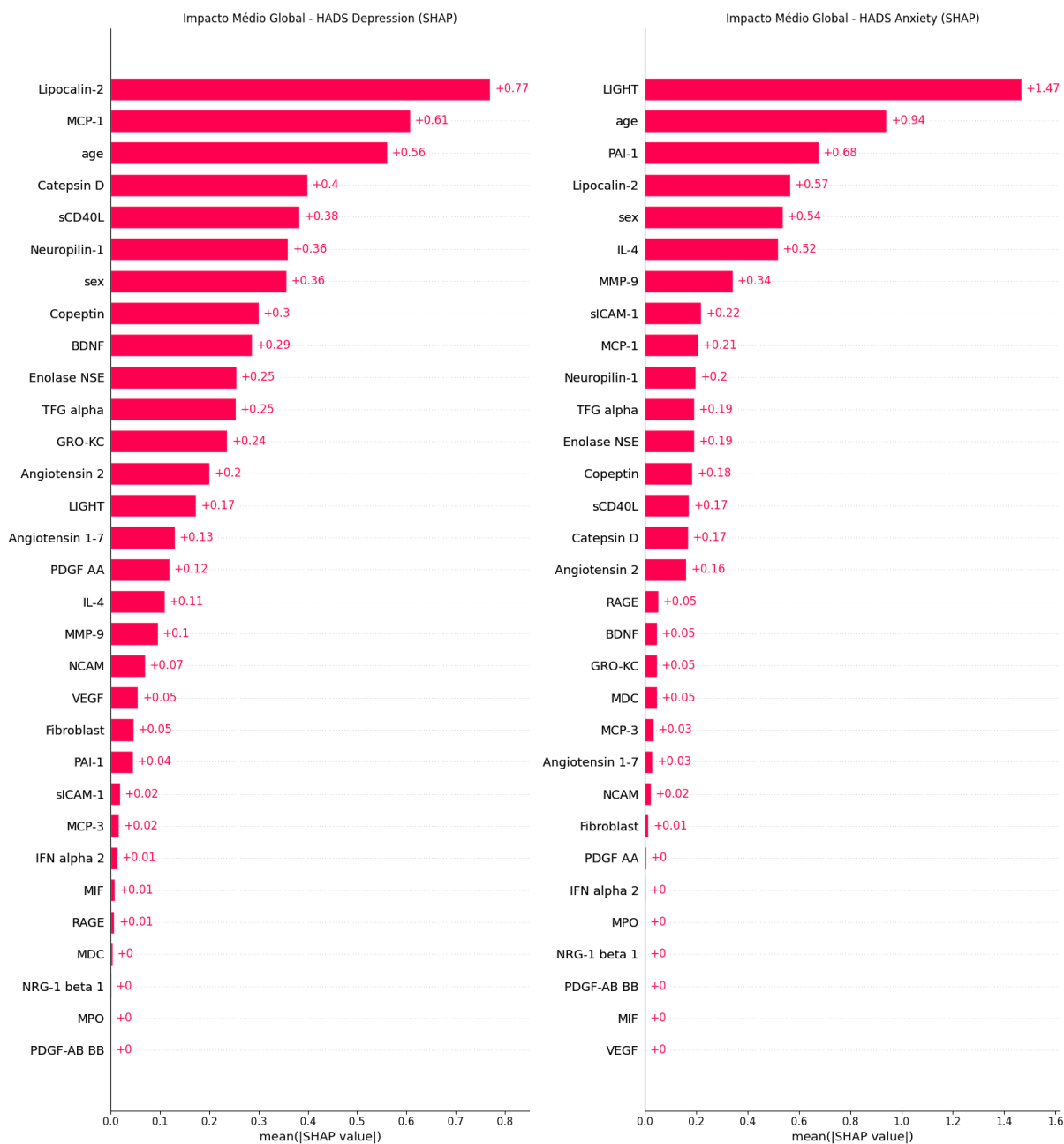


Figura 4.3 – Impacto médio dos biomarcadores (valores SHAP) nas predições dos modelos XGBoost. **Fonte:** Elaboração própria a partir dos resultados dos modelos criados.

e depressão em pacientes com TCE. No entanto, observaram-se convergências e divergências relevantes entre as abordagens.

**Para HADS *Anxiety*:**

- A idade (variável clínica) e o biomarcador *Lipocalin-2* foram apontados como preditores relevantes por ambos os modelos;
- O XGBoost atribuiu maior importância a LIGHT e PAI-1; já o modelo RF destacou MCP-1

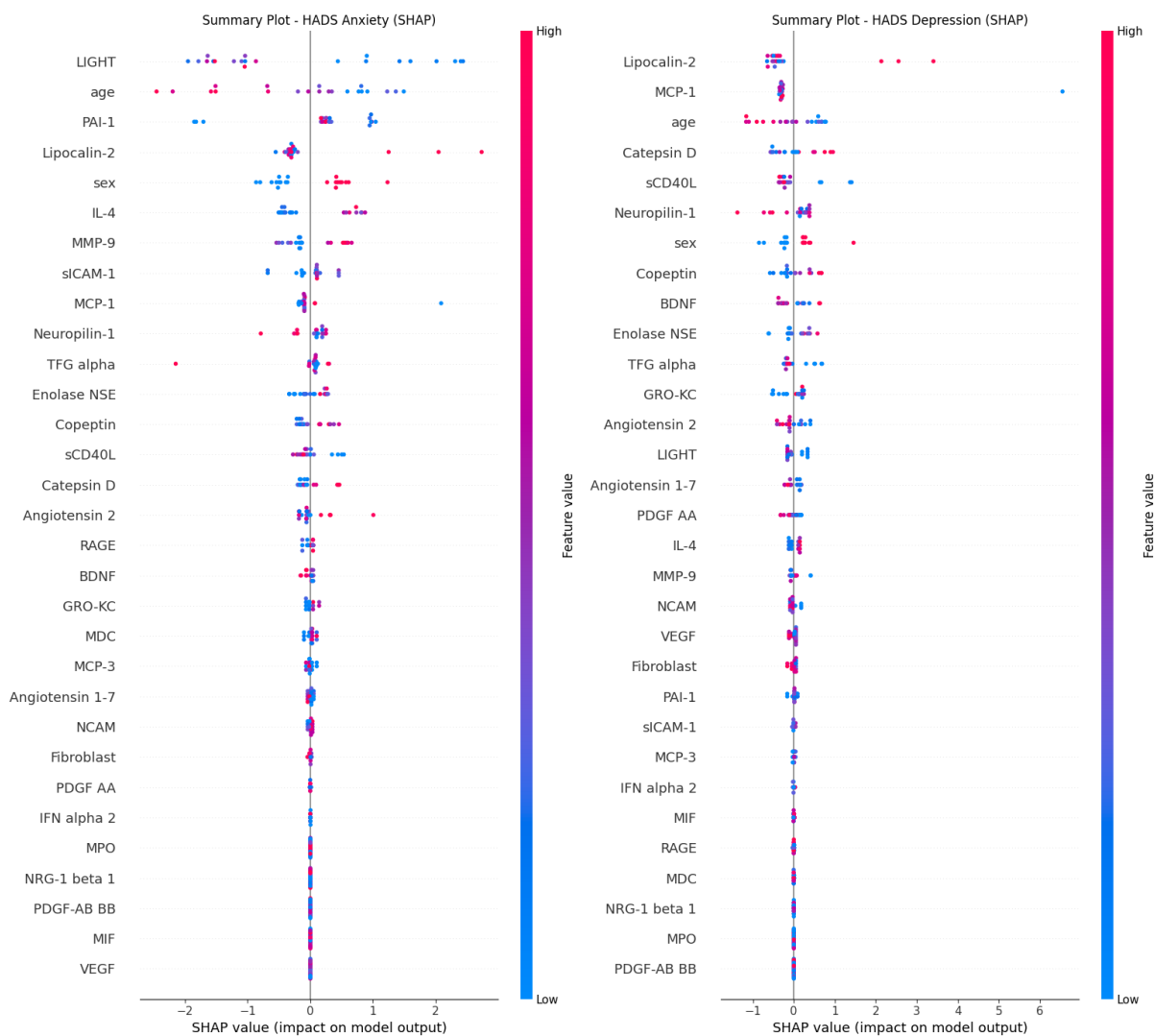


Figura 4.4 – Distribuição dos valores SHAP dos biomarcadores nas predições dos modelos XGBoost (*Summary Plot*). **Fonte:** Elaboração própria a partir dos resultados dos modelos criados.

e RAGE.

**Para HADS *Depression*:**

- Ambos reconheceram *Lipocalin-2* como preditor de destaque;
- O XGBoost deu ênfase a MCP-1 e *Cathepsin D*, enquanto o RF destacou MCP-1, *Fibroblast* e VEGF;
- *age* teve menor impacto no RF do que no XGBoost.

Esses resultados sugerem que os modelos capturam aspectos distintos das relações não lineares entre biomarcadores e desfechos. O XGBoost apresenta uma hierarquia mais concentrada,

enquanto o RF mostra uma distribuição mais homogênea. Essa complementaridade reforça a importância de utilizar múltiplos algoritmos para obter uma compreensão mais abrangente dos fenômenos biológicos.

Com base em uma análise aprofundada dos modelos, os resultados sugerem que tanto o RF quanto o XGBoost capturam aspectos distintos das relações não lineares entre biomarcadores e os desfechos clínicos, embora com um desempenho preditivo comparável. Na busca pelo melhor desempenho do XGBoost, observou-se que as métricas de avaliação melhoraram até aproximadamente 20 árvores e depois estabilizaram, atingindo um desempenho ótimo com 40 a 60 árvores. Em comparação, o modelo RF de Almeida (2025) apresentou um desempenho ligeiramente superior.

Apesar da performance similar indicada por essas métricas, a análise da importância das variáveis revela abordagens de aprendizado distintas. O XGBoost apresenta uma hierarquia de importância mais concentrada, atribuindo maior relevância a biomarcadores como LIGHT e PAI-1 para a predição de ansiedade, e a *Lipocalin-2* e MCP-1 para a depressão. Por outro lado, o RF mostra uma distribuição de importância mais homogênea entre os preditores, destacando MCP-1 e RAGE para ansiedade, e MCP-1, *Fibroblast* e VEGF para depressão. Essa complementaridade, onde ambos os algoritmos alcançam alta precisão, mas por meio de diferentes ponderações das características biológicas, reforça a importância de utilizar múltiplos modelos para obter uma compreensão mais abrangente dos fenômenos biológicos complexos.

# 5 Considerações Finais

Este trabalho demonstrou com sucesso a aplicação sinérgica do algoritmo XGBoost e da técnica de interpretabilidade SHAP para a análise de dados tabulares complexos no contexto de desfechos emocionais pós-TCE leve. A combinação provou ser uma ferramenta poderosa, não apenas para construir modelos preditivos de alta acurácia para ansiedade e depressão, mas também para desvendar a “caixa-preta” desses modelos, oferecendo *insights* clinicamente relevantes.

Os resultados evidenciaram que o XGBoost é capaz de modelar as relações não-lineares presentes nos dados, alcançando um desempenho satisfatório na predição dos escores de HADS *Anxiety* e HADS *Depression*. A análise de interpretabilidade com SHAP foi fundamental para validar a robustez do modelo, revelando que as predições são guiadas por variáveis biologicamente plausíveis.

De forma notável, a análise SHAP identificou assinaturas biológicas distintas para a ansiedade e a depressão. Para a ansiedade, o biomarcador LIGHT e a idade emergiram como os fatores de maior impacto. Já para a depressão, a *Lipocalin-2* e o MCP-1 foram as variáveis mais influentes. A identificação desses biomarcadores específicos e a diferenciação de seus papéis em cada condição reforçam a complexidade fisiopatológica desses transtornos e a necessidade de abordagens diagnósticas e terapêuticas personalizadas.

O estudo valida a arquitetura de aprendizado de máquina proposta como uma abordagem eficaz para identificar preditores em dados de saúde, justificando a decisão de utilizar modelos separados para cada desfecho e abrindo caminho para investigações mais aprofundadas sobre os mecanismos biológicos subjacentes.

## 5.1 Trabalhos Futuros

As conclusões e os *insights* gerados por este trabalho abrem diversas avenidas para pesquisas futuras, que podem aprofundar e expandir os resultados aqui apresentados. Sugere-se as seguintes direções:

- **Ampliação da Base de Dados:** Utilizar conjuntos de dados mais amplos e diversificados para validar os achados em populações maiores e com diferentes perfis clínicos. Essa abordagem é fundamental para reduzir possíveis vieses amostrais e aumentar a capacidade de generalização dos modelos preditivos, garantindo sua robustez e aplicabilidade em contextos variados;
- **Investigação de Relações Não-Lineares e Interações:** Explorar mais a fundo as relações não-lineares e as interações complexas entre as variáveis, que a técnica SHAP é capaz de

revelar. Uma análise detalhada dos *summary plots* e dos gráficos de dependência pode fornecer *insights* mais profundos sobre a fisiopatologia do TCE e seus impactos em desfechos emocionais, identificando sinergias e antagonismos entre biomarcadores;

- **Desenvolvimento de uma Ferramenta de Apoio à Decisão Clínica:** A longo prazo, os modelos validados e refinados poderiam ser integrados a uma ferramenta de *software* para apoio à decisão clínica. Tal ferramenta poderia auxiliar profissionais de saúde a estratificar o risco de pacientes desenvolverem transtornos de humor pós-TCE, com base em seu perfil biológico individual, permitindo a implementação de intervenções terapêuticas precoces e personalizadas.

# Referências

- ALMEIDA, D. E. de S. *Aplicação de inteligência artificial explicável na identificação de biomarcadores associados ao traumatismo cranioencefálico leve*. Dissertação (Dissertação de Mestrado) — Universidade Federal de Ouro Preto, Ouro Preto, MG, 2025. Programa de Pós-Graduação em Biotecnologia.
- BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006.
- BREIMAN, L. Random forests. *Machine Learning*, Springer, v. 45, n. 1, p. 5–32, 2001.
- CAWLEY, G. C.; TALBOT, N. L. C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, v. 11, n. 70, p. 2079–2107, 2010.
- CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.]: ACM, 2016. (KDD '16), p. 785–794.
- FONSECA, J. P. B. *AI-Based Models to Predict The Traumatic Brain Injury Outcome*. 2022. Acesso em: 9 dez. 2024. Disponível em: <<https://hdl.handle.net/10216/142988>>.
- GHAJAR, J. Traumatic brain injury. *Lancet*, v. 356, n. 9233, p. 923–929, 2000.
- Google. *Perguntas Frequentes do Google Colab*. 2025. Acessado em: 3 set. 2025. Disponível em: <<https://research.google.com/colaboratory/intl/pt-BR/faq.html>>.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. ed. New York: Springer, 2009.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.
- LUDEMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, v. 35, n. 101, p. 85–94, 2021.
- LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA, USA: [s.n.], 2017. p. 4765–4774.
- MATSUNO, K.; AIHARA, H.; NAKAI, T.; MORISHITA, A.; TOHMA, Y.; KOHMURA, E. Machine learning to predict in-hospital morbidity and mortality after traumatic brain injury. *Journal of Neurotrauma*, v. 35, n. 1, p. 1–10, 2018.
- MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997.
- SHAPLEY, L. S. A value for n-person games. In: KUHN, H. W.; TUCKER, A. W. (Ed.). *Contributions to the Theory of Games, Volume II*. [S.l.]: Princeton University Press, 1953. p. 307–318.

SOUZA, D. C. de. *Modelos de Aprendizado de Máquina para Predição de Óbito de Pacientes em Hemodiálise*. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2024.

The XGBoost developers. *XGBoost Python API Reference*. 2024. <[https://xgboost.readthedocs.io/en/stable/python/python\\_api.html](https://xgboost.readthedocs.io/en/stable/python/python_api.html)>. Acessado em: 3 de setembro de 2025.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*. [S.l.]: Curran Associates, Inc., 2017. p. 5998–6008.

WANG, R.; WANG, L.; ZHANG, J.; HE, M.; XU, J. XGBoost machine learning algorithm performed better than regression models in predicting mortality of moderate-to-severe traumatic brain injury. *World Neurosurgery*, v. 163, p. e617–e622, 2022.