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

GUSTAVO PRESOTI SALES BRITO
Orientador: Prof. Dr. Rafael Alves Bonfim de Queiroz
Coorientadora: Profa. Dra. Aline Silva de Miranda

ANÁLISE EXPLORATÓRIA DE DADOS DE TRAUMATISMO CRANIOENCEFÁLICO USANDO FLORESTA ALEATÓRIA

Ouro Preto, MG
2023

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

GUSTAVO PRESOTI SALES BRITO

ANÁLISE EXPLORATÓRIA DE DADOS DE TRAUMATISMO CRANIOENCEFÁLICO USANDO FLORESTA ALEATÓRIA

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

Orientador: Prof. Dr. Rafael Alves Bonfim de Queiroz
Coorientadora: Profa. Dra. Aline Silva de Miranda

Ouro Preto, MG
2023

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Gustavo Presoti Sales Brito

Análise exploratória de dados de traumatismo cranioencefálico usando floresta aleatória

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 23 de Março de 2023.

Membros da banca

Rafael Alves Bonfim de Queiroz (Orientador) - Doutor - Universidade Federal de Ouro Preto
Aline Silva de Miranda (Coorientadora) - Doutora - Universidade Federal de Minas Gerais
Valéria de Carvalho Santos (Examinadora) - Doutora - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto

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 23/03/2023.

Documento assinado eletronicamente por **Rafael Alves Bonfim de Queiroz**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 29/03/2023, às 20:34, 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 , informando o código verificador **0493045** e o código CRC **17CF947D**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Agradecimentos

O desenvolvimento deste trabalho só se tornou possível por ter a participação e ajuda de pessoas incríveis e que são fundamentais em minha vida. Gostaria de agradecer:

Aos meus pais, Weliton Aloir de Souza Brito e Dircineia de Sousa Sales por todo o amor, educação, compreensão e por acreditarem nesse sonho junto comigo e me darem todo o suporte para realizá-lo.

Ao meu irmão Eduardo Henrique Sales Brito pelo apoio e por servir de inspiração.

A Caroline Alves, por não ter me deixado desistir e ter me dado suporte, amor e compreensão durante esse tempo e por me levantar nos momentos mais difíceis. Aos amigos Alan Erse, Jonata Souza e Luis Roberto por acreditarem em mim e não me deixarem desistir, assim como por todos os momentos que compartilhamos juntos. O apoio de vocês foi fundamental para que eu me tornasse uma pessoa melhor e pudesse alcançar essa conquista.

A minha amada república Kamikaze, que me proporcionou todo o suporte emocional durante todo esse tempo e pelo apoio incondicional.

Ao Departamento de Computação (DECOM) e a Universidade Federal de Ouro Preto (UFOP), pelo ensino de qualidade e por todo o suporte dado e também o projeto CuidaIdoso e todos os que estiveram envolvidos, ajudando no meu desenvolvimento pessoal e profissional.

Ao meu orientador Rafael por ter confiado este trabalho a mim e por toda a paciência, apoio e confiança. A minha coorientadora Aline por ter fornecido os dados clínicos para análise exploratória.

Resumo

Esse trabalho traz uma análise exploratória em dados de traumatismo cranioencefálico por meio do uso de aprendizado de máquina para a identificação dos atributos mais importantes na predição do desfecho de variáveis-alvo. As lesões causadas pelo traumatismo cranioencefálico são uma das maiores causas de mortes no mundo e também estão diretamente ligadas com diversos outros problemas neurológicos pós-trauma. Diversos algoritmos de aprendizado de máquina vêm sendo estudados e aplicados em análise de dados clínicos. Dessa forma, esse trabalho surgiu da necessidade identificada em se realizar uma análise exploratória de dados de traumatismo cranioencefálico. Modelos de floresta aleatória são utilizados para identificação dos atributos mais relevantes para a predição de desfechos. Para tal, foi utilizada a classe *RandomForestRegressor* da biblioteca *scikit-learn* do Python para a regressão dos dados, definindo a pontuação média de cada um dos atributos na predição do desfecho e, diante disso, a identificação dos atributos importantes. Os resultados produzidos consideraram a predição de duas variáveis-alvo, ambas relacionadas com a escala HAD. Através da aplicação da metodologia proposta, foram identificados os atributos que apresentam maior relevância na predição do desfecho de cada variável-alvo, sendo de extrema importância para o entendimento e análise do comportamento dos dados e do algoritmo para melhorias futuras.

Palavras-chave: Traumatismo Cranioencefálico. Aprendizado de Máquina. Floresta Aleatória. Importância de Atributo. Seleção de Atributo.

Abstract

This work brings an exploratory analysis of traumatic brain injury data through machine learning to identify the most important attributes in predicting the outcome of target variables. Traumatic brain injury is one of the largest causes of death in the world and is also directly linked to several other neurological problems after the trauma. Several machine learning algorithms have been studied and applied in cases of this kind. Thus, this work arose from the identified need to perform an exploratory analysis of traumatic brain injury data. Random forest models are employed to identify the most relevant attributes for predicting outcomes. The RandomForestRegressor class from Python's scikit-learn library was used to regress the data, defining the average score for each attribute in predicting the outcome and, based on this, identifying the important attributes. The results considered the prediction of two target variables related to HAD scale. Through the application of the proposed methodology, the attributes that present the most significant relevance in predicting the outcome of each target variable were identified, being of extreme importance for understanding and analyzing the behavior of the data and the algorithm for future improvements.

Keywords: Traumatic Brain Injury, Machine Learning, Random Forest. Feature Importance. Feature Selection.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

TBI	*Traumatic Brain Injury*
ML	*Machine Learning*
RF	*Random Forest*
CT	*Computed Tomography*
MRI	*Magnetic Resonance Imaging*
GCS	*Glasgow Coma Scale*
FDP	*Fibrinogen Degradation Products*
CART	*Classification and Regression Trees*
PSO	*Particle Swarm Optimization*

Sumário

1 Introdução

1.1 Problema abordado

As lesões decorrentes do traumatismo cranioencefálico (TBI - *Traumatic Brain Injury*) são um dos maiores causadores de mortes no mundo e, além disso, também estão diretamente relacionadas com diversas outras deficiências neurológicas pós-trauma (KINNUNEN et al., 2011; TRIFAN et al., 2017). Em geral, a avaliação dos procedimentos a serem adotados e a predição do resultado clínico são feitas pelo próprio médico, com o auxílio dos dados já coletados em exames e o estado clínico observado. Atualmente, diversos algoritmos de aprendizado de máquina (ML - *Machine Learning*) estão sendo estudados e aplicados como ferramentas de predição para casos clínicos (KAVOSI et al., 2015; IJ, 2018), a fim de auxiliar na predição do desfecho dos casos a partir dos dados conhecidos do paciente.

Em geral, os problemas da predição e de ML podem ser definidos quando, a partir de um conjunto de valores de atributos observados, o objetivo é prever o valor desconhecido de outro atributo. Com o avanço de ML e seus diferentes algoritmos, a aplicação em diversas áreas é cada vez mais explorada para a resolução de problemas de predição. Por depender comumente de tomadas de decisões e diagnóstico baseado em dados históricos, a área da saúde de emergência, principalmente em traumas, é uma das áreas que trazem bons desafios para a aplicação de ML. Um dos algoritmos de ML mais utilizados e que apresenta bons resultados para predição de traumatismo cranioencefálico é o de floresta aleatória (RF - *Random Forest*) (THARA; THAKUL, 2021).

Para se realizar a regressão dos dados utilizando um modelo de RF, é necessário definir a importância dos atributos presentes nas amostras de dados para obtenção de quais características são relevantes ou não para o modelo preditivo, com o intuito de simplificar e melhorar o modelo. Para tal, é definida uma pontuação para cada característica baseada na relevância para predizer uma determinada variável.

Dessa forma, este trabalho tem como finalidade a identificação dos principais atributos para definição do modelo de predição envolvendo dados clínicos de traumatismo cranioencefálico. Para tanto, o modelo RF será investigado para tal finalidade.

1.2 Justificativa

De acordo com o exposto por Hsia et al. (2018) e Carteri e Silva (2021), o número de mortes decorrentes do TBI tem diminuído com o tempo, mas a quantidade de pacientes com consequências pós traumáticas cresce em ritmo acelerado. Além disso, também geram um alto custo com despesas hospitalares que estão em sua grande maioria relacionadas à necessidade de realização de exames custosos, como a tomografia computadorizada (CT - *Computed Tomography*) e a ressonância magnética (MRI - *Magnetic Resonance Imaging*). Por conta disso e da complexidade, os atendimentos de emergência para esses casos dependem ainda mais de avaliação rápida e tomada de decisões (KORLEY et al., 2016; HUANG et al., 2019) para tentar identificar o possível desfecho do trauma.

Diante disso, a motivação do presente trabalho vem da necessidade da construção de modelos de predição para realização de uma análise exploratória dos dados em casos de TBI. Porém, devido a pouca amostragem dos dados a serem utilizados, é preciso a aplicação de modelos de RF para realizar a quantificação da importância dos atributos ao invés de uma análise com o uso de métodos estatísticos.

1.3 Objetivos

O objetivo principal deste trabalho é avaliar a importância de atributos para predição de desfecho em dados clínicos de TBI.

Como objetivos específicos desta monografia, este trabalho visa:

- Construir modelos de RF usando a linguagem de programação Python (ROSSUM; DRAKE, 1995);

- Investigar a influência de parâmetros na construção dos modelos de RF.
- Identificar os atributos mais importantes para predição usando o modelo de RF.

1.4 Organização do Trabalho

A estrutura de organização do presente trabalho é apresentada abaixo.

No Capítulo 1 será feita uma breve introdução ao tema do trabalho e, posteriormente, uma apresentação da motivação que gerou à produção do trabalho e também os seus objetivos. Em seguida, no Capítulo 2, é apresentada toda a fundamentação teórica do trabalho, destacando as definições dos conceitos mais importantes para o melhor entendimento dos assuntos e das tecnologias a serem utilizadas e simplificação do desenvolvimento. Ainda no mesmo capítulo, também é feita uma revisão da literatura, trazendo alguns trabalhos relacionados que apresentam similaridades ou que fazem parte da mesma área de pesquisa. O Capítulo 3 traz a especificação de como o desenvolvimento do trabalho foi feito, apresentando as ferramentas e procedimentos utilizados. Os dois últimos capítulos trazem o desfecho do trabalho, sendo que no Capítulo 4 são apresentados os resultados obtidos e a discussão dos mesmos, enquanto no Capítulo 5 é feita a conclusão de todo o exposto, fazendo a análise de tudo o que foi feito e quais são os próximos passos já identificados para a evolução deste trabalho.

2 Revisão Bibliográfica

Este capítulo contém a discussão acerca dos trabalhos relacionados, bem como dos temas associados ao trabalho proposto.

2.1 Traumatismo Cranioencefálico

O traumatismo cranioencefálico (PARIKH; KOCH; NARAYAN, 2007) é um dos tipos mais comuns de traumas, principalmente em atendimentos de emergência, sendo considerado o trauma que mais causa mortes e deficiências no mundo todo. O traumatismo é caracterizado por lesão física do tecido cerebral, ocasionando danos temporários ou até mesmo permanentes na função cerebral.

O diagnóstico para esse tipo de traumatismo geralmente é realizado por meio da avaliação geral rápida do trauma, definição da escala de coma modificada de Glasgow, exame neurológico, entre outros julgados necessários pelo(a) médico(a). No banco de dados utilizado neste trabalho, estão presentes diversos biomarcadores utilizados para identificar a gravidade do trauma e a consequente predição de desfecho. Esse banco de dados foi utilizado para a identificação dos atributos (biomarcadores) mais importantes para determinados desfechos através do uso do algoritmo de floresta aleatória.

2.2 Árvores de Decisão

As árvores de decisão (MORGAN; SONQUIST, 1963) são uma forma de representação de dados simples e baseadas na técnica de divisão e conquista (QUINLAN, 2014). Dessa forma, uma árvore de decisão realiza diversas divisões seguidas nos dados até que os dados façam parte da mesma classe. Devido a essa capacidade, as árvores de decisão são utilizadas para se calcular o valor esperado de determinadas decisões de acordo com um conjunto de opções e as probabilidades associadas com cada uma dessas decisões. Dessa forma,

a árvore de decisão apresenta as condições e alternativas para classificar o resultado ou prever um desfecho.

A Figura 2.1 mostra a representação de como uma árvore de decisão é construída. Neste exemplo, é possível observar que a árvore é constituída por um conjunto de regras que partem da raiz da árvore (decisão) e finalizam em um nó folha (resultado).

Figura 2.1 – Árvore de decisão
Fonte: Elaborado pelo autor.

O principal objetivo ao utilizar árvores de decisão é separar os dados em grupos menores e cada vez mais homogêneos em relação ao resultado desejado. Os preditores utilizados em cada ponto da árvore são decididos pelo algoritmo, assim como o tamanho da divisão da amostra. A profundidade da árvore também deve ser definida e controlada para evitar o *overfit* [1], que é uma das maiores desvantagens das árvores de decisão, pois dificulta a generalização.

[1] overfit: quando o modelo se torna sobre-ajustado, ou seja, aprende demais com os dados e não consegue generalizar com dados novos

2.3 Métodos baseados em comitê de classificadores

Os métodos de aprendizado de comitê (*ensemble*) fazem uso de classificadores, como as árvores de decisão, para agregar os resultados das predições feitas e identificar aquele resultado de maior frequência (em cenários de classificação).

Um dos métodos mais conhecidos é o de ensacamento (*bagging*) (BREIMAN, 1996). Nesse método uma parte aleatória do conjunto de dados de treino é escolhida para ser substituída, gerando diversas amostras de dados. Em seguida, os modelos são treinados de forma independente e são obtidos resultados preditivos que, no geral, são mais precisos. É importante ressaltar que os resultados dependem se o tipo de tarefa a ser executada é de classificação ou regressão.

2.4 Floresta Aleatória

A técnica de aprendizado de máquina conhecida como *Random Forest* (RF) teve inspiração no trabalho de Amit e Geman (1997) e foi introduzida primeiro por Breiman et al. (1984), Breiman (2001) como Árvores de Classificação e Regressão (CART - *Classification and Regression Trees*). Assim como o nome já diz, consiste em diversas árvores de decisão que podem ser utilizadas para solucionar problemas de classificação e regressão. A construção dessas árvores é feita através de métodos que buscam criar "florestas" onde essas árvores não estejam correlacionadas ou tenham a menor correlação possível. É uma técnica que realiza as predições de acordo com o resultado das árvores de decisão, considerando a maioria dos votos ou a média dessas saídas. Dessa forma, quanto maior o número de árvores, maior tende a ser a precisão do resultado.

Na definição dada por Breiman (2001), o método de floresta aleatória é representado como uma coleção de árvores de decisão $h(x, \theta_k, k = 1, ...)$, onde os θ_k são vetores aleatórios independentes e igualmente distribuídos, onde cada árvore apresenta um voto único da classe mais popular da entrada x.

Na Figura 2.2 é possível visualizar a principal diferença entre as árvores de decisão e as florestas aleatórias: a geração e consideração dos subconjunto aleatórios de atributos onde não são mais consideradas todas as divisões de atributos, somente os subconjuntos (EDUCATION, 2022).

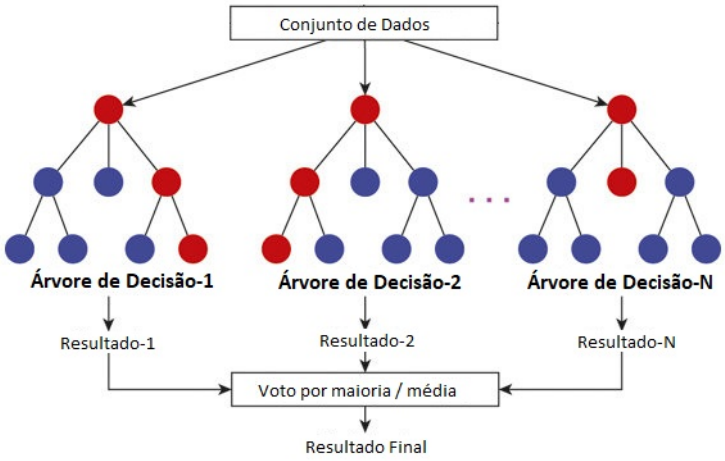


Figura 2.2 – Representação de uma floresta aleatória e sua tomada de decisão. Fonte: Adaptado de Education (2022).

O método RF pode ser considerado uma extensão do método de ensacamento, acrescentando ao método flexibilidade e aleatoriedade nos dados, se tornando uma alternativa ao *overfit*.

Para construção de uma RF, é necessário ajustar o tamanho dos nós, o número de árvores e o número de atributos das amostras antes do treinamento do modelo. O ajuste desses parâmetros estão diretamente relacionados com o processo de construção de uma RF, que pode ser dividido em quatro passos principais, descritos a seguir.

- O primeiro passo antes de se iniciar o processo de escolha dos atributos que a RF utilizará no modelo, é a obtenção do *bootstrap dataset* (conjunto de dados de bootstrap). Esse conjunto de dados é criado a partir do conjunto de dados original, onde são selecionadas algumas amostras aleatórias dentro dos dados, afim de prover a variância dos dados.

- O segundo passo é a criação das árvores de decisão que irão fazer parte da floresta. Para tal, é necessário realizar a seleção de N atributos, também

de maneira aleatória, de forma que esses atributos possam ser utilizados na criação da primeira árvore de decisão.

- No terceiro passo, a partir do subconjunto selecionado, é feita a verificação do atributo que melhor separa os dados e esse atributo é escolhido para ser a raiz da árvore sendo construída. Em seguida, são escolhidos outros dois atributos a partir dos outros que não foram selecionados anteriormente para separar os dados e compor a árvore. Esse mesmo processo é repetido para a formação de todas as árvores que compõem a floresta. É importante ressaltar que o tamanho das árvores geradas não são necessariamente iguais, podendo variar.

- O quarto passo é percorrer cada árvore da floresta gerada e definir qual o resultado da predição obtido daquela árvore. No caso da classificação, o desfecho considera o resultado com maior número de acontecimentos e, no caso da regressão, a média desses resultados.

A separação necessária para se gerar novos ramos e construir as árvores devem sempre ser melhores que a anterior. Para saber a qualidade dessa separação, são utilizados critérios diferentes de acordo com o tipo de problema sendo tratado.

Em problemas de classificação, são utilizados frequentemente os critérios de Entropia (Equação 2.1) e Impureza de Gini (Equação 2.2).

A entropia usa a probabilidade de uma determinado desfecho para decidir se a separação deve ser feita ou não. Na fórmula, p_i e C representa o número de classes.

$$E = \sum_{i=1}^{C} -p_i * \log_2(p_i) \tag{2.1}$$

O índice de impureza de Gini define qual dos ramos tem mais probabilidade de acontecer a partir da classe e da probabilidade de cada ramo ocorrer. Na fórmula, p_i representa a frequência relativa da classe sendo observada no conjunto de dados e C representa o número de classes.

$$G = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{2.2}$$

Já em problemas de regressão, o critério mais frequentemente utilizado é o Erro Quadrado Médio (MSE - *Mean Squared Error*) (2.3).

O MSE faz o cálculo da distância de cada nó partindo do valor real com o objetivo de decidir qual ramo apresenta melhor valor para a construção da floresta. Na fórmula, N é o número de dados, fi é o valor retornado pelo modelo e yi é o valor atual do dado sendo testado em um determinado nó.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2 \tag{2.3}$$

Para medir o nível em que o número médio de votos da classificação para a classe certa de **X**, sendo **Y, X** um conjunto de treinamento, excede a classificação para outra classe, temos que:

Dado um conjunto de classificadores $h_1(\boldsymbol{X}), h_2(\boldsymbol{X}), ..., h_k(\boldsymbol{X})$, é possível demonstrar que:

$$mg(\boldsymbol{X}, Y) = P_\Theta(h(\boldsymbol{X}, \Theta) = Y) - max_{j \neq Y}(P_\Theta(h(\boldsymbol{X}, \Theta) = j), \tag{2.4}$$

onde Θ representa vetores aleatórios distribuídos de forma independente. Além disso, a equação para o erro de generalização é dada por:

$$PE = P_{\boldsymbol{X},Y}(mg(\boldsymbol{X}, Y)) < 0. \tag{2.5}$$

Através da Lei Forte dos Grandes Números, é possível demonstrar também que o erro de generalização converge para:

$$P_{\boldsymbol{X},Y}(P_\Theta(h(\boldsymbol{X}, \Theta) = Y) - max_{j \neq Y}(P_\Theta(h(\boldsymbol{X}, \Theta) = j)) < 0). \tag{2.6}$$

Com isso, é possível perceber que mesmo com o aumento da quantidade de árvores, há uma restrição para o erro de generalização. Dessa forma, também acontece uma restrição do *overfit*.

Diante de todo o exposto, é perceptível a importância de um bom ajuste dos parâmetros para uma melhor acurácia do modelo. Como cada caso tem uma especificidade e um conjunto de dados diferente, o ajuste adequado pode fazer com que o resultado da predição do modelo sendo utilizado seja muito melhor e mais preciso para o caso sendo trabalhado.

Como cada caso traz uma possibilidade enorme de variações, não existem valores definidos para os parâmetros que garantam um bom resultado. Diante disso, o ajuste destes parâmetros está muito mais relacionado com a experimentação dos resultados do que com a definição dos mesmo, ou seja, o melhor ajuste para cada caso deve ser descoberto através de inúmeras tentativas, fazendo a validação e análise dos resultados dos experimentos.

A seguir, são apresentados alguns parâmetros que podem ser variados dentro do algoritmo de RF e que serão utilizados na experimentação do presente trabalho.

2.4.1 *n_estimators*

O parâmetro *n_estimators* define o número de árvores que serão construídas e, consequentemente, formarão a floresta que irá produzir os resultados da classificação ou regressão. Por produzir uma variedade maior, esse parâmetro pode trazer uma melhoria significativa para o resultado do modelo, mas também irá reduzir o tempo de execução. Dessa forma, o ideal é o ajuste dentro do poder de processamento, para que os resultados sejam mais precisos e estáveis.

2.4.2 *min_samples_split*

O *min_samples_split* é o parâmetro que controla o número de amostras necessárias em um nó para que seja considerada uma divisão pela árvore. O ajuste desse parâmetro com um valor muito baixo contribui para uma chance maior de *overfitting* e o ajuste com um valor muito alto pode levar a um efeito contrário, ou seja, maior chance de *underfitting* (modelo não consegue se ajustar aos dados de treinamento).

2.4.3 *min_samples_leaf*

O parâmetro *min_samples_leaf* é outro parâmetro importante para ser ajustado. Com ele, é possível ajustar o número mínimo de atributos que devem compor os nós folhas de cada árvore. Por impactar diretamente no resultado de cada árvore, um número baixo de nós folhas fará com que o modelo seja menos preciso e mais propenso a falha.

2.4.4 *max_features*

O parâmetro *max_features* é o que define o número máximo de atributos que o modelo irá considerar para a construção de uma árvore e, consequentemente, para realizar a divisão de uma árvore nova.

2.4.5 *criterion*

O parâmetro *criterion* serve como medida de qualidade para a divisão das árvores. O critério definido é utilizado para definir como as variáveis e os limites para divisão das amostras serão feitas. O erro será minimizado de acordo com o critério escolhido.

2.5 Importância e Seleção de Atributos

De acordo com Brownlee (2020), importância de atributo se refere a um conjunto de técnicas para definir as pontuações dos atributos de entrada para um modelo preditivo que indica a importância relativa de cada atributo ao fazer uma predição. Esse conjunto de técnicas podem ser aplicadas em problemas de regressão e de classificação, colaborando para o melhor entendimento dos dados e do modelo.

O uso de importância de atributo está diretamente associado à obtenção de quais características são relevantes ou não para o modelo preditivo. Dessa forma, a utilização desse conjunto de técnicas permite a simplificação e melhoria do modelo. Por ser um conjunto de técnicas, cada técnica pode gerar uma nova perspectiva dos dados e, consequentemente, proporcionar melhorias de diferentes formas. Porém, a definição de quais técnicas deve levar em conta o

caso e o modelo a ser utilizado, sendo recomendado a realização de testes com diferentes técnicas para obtenção do melhor resultado.

Segundo Brownlee (2020), seleção de atributos se refere às técnicas para selecionar o subconjunto de atributos de entrada que é mais relevante para a variável que está sendo prevista. O objetivo do uso de seleção de atributos é a redução das variáveis de entrada do modelo de predição através da eliminação do uso de dados irrelevantes.

As técnicas também são subdivididas em dois grupos: supervisionada ou não supervisionada. Os modelos supervisionados são aqueles que utilizam as variáveis alvo para identificar quais são as variáveis que podem gerar melhoria no modelo. Por outro lado, os modelos não supervisionados são aqueles que não dependem da variável alvo para selecionar os atributos.

A Figura 2.3 ilustra uma visão geral das técnicas de seleção de atributos, onde é possível observar que as árvores utilizam essa técnica intrinsecamente na sua construção, permitindo que atributos sem relevância sejam filtrados e, dessa forma, árvores com melhores divisões sejam formadas.

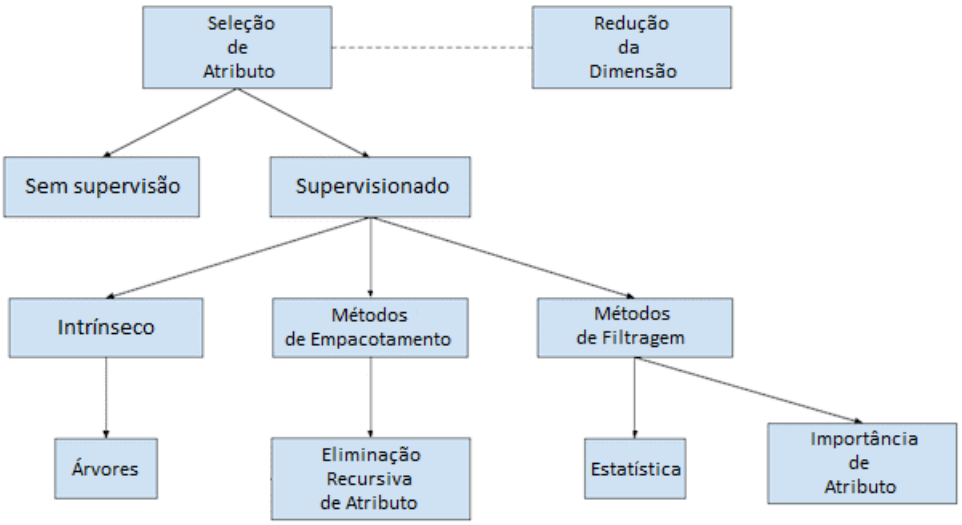
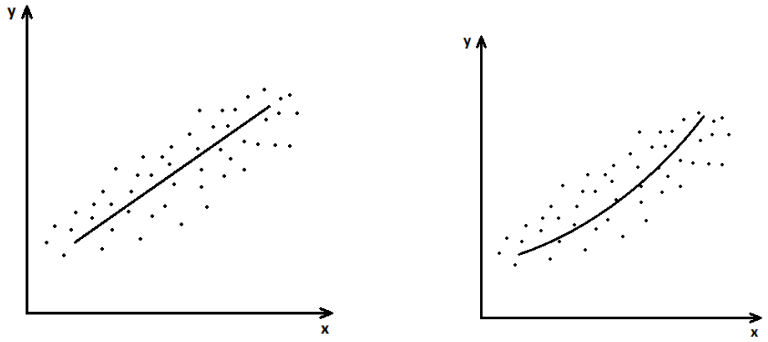


Figura 2.3 – Visão geral das técnicas de seleção de atributos.
Fonte: Adaptado de Brownlee (2020).

2.6 Modelo de Regressão

O principal objetivo de um modelo de regressão é obter uma equação que seja capaz de demonstrar a relação entre uma variável de resposta e outra(s) explicativa(s) de forma satisfatória, facilitando a tarefa de predizer a variável de interesse. A relação pode ser tanto linear quanto não linear conforme ilustra a Figura 2.4.



(a) Linear (b) Não linear

Figura 2.4 – Gráfico de dispersão linear e não linear de relação entre pares de variáveis

Fonte: Elaborado pelo autor.

De acordo com o apresentado em Montgomery, Peck e Vining (2021), o modelo de regressão linear é definido pela equação:

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{2.7}$$

onde β_0 e β_1 são constantes determinadas a partir dos dados. O valor ε é a diferença entre o valor observado (ou seja, o dado conhecido) de y e o seu valor estimado pelo modelo. De acordo com a definição de um modelo de regressão apresentado, temos que x é a variável de regressão e y é a variável resposta. A equação pode apresentar uma ou mais variáveis de regressão, sendo definidas como modelo de regressão linear simples (Equação 2.7) (apenas uma variável) e modelo de regressão linear múltiplo (Equação 2.8) (duas ou mais variáveis).

Ainda de acordo com o apresentado em Montgomery, Peck e Vining (2021), o modelo de regressão linear múltiplo é definido pela equação:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon. \tag{2.8}$$

Um objetivo importante da análise de regressão é estimar os parâmetros desconhecidos ($\beta's$) no modelo de regressão (MONTGOMERY; PECK; VINING, 2021). Para realizar essa adaptação ao modelo de dados, existem diversas técnicas conhecidas para se estimar os parâmetros.

2.7 Trabalhos Relacionados

As predições e classificações utilizando técnicas e algoritmos de ML vêm sendo utilizadas em diversas áreas nos últimos tempos, tais como redes sociais, engenharia, medicina, entre outras. Diversos trabalhos Zikic et al. (2012), Rao et al. (2014), Molaei et al. (2016), Matsuo et al. (2020), Thara e Thakul (2021) apresentaram resultados muito satisfatórios alcançados por modelos de ML e, diante disso, existem diversos estudos que propõem o uso de algoritmos de ML em casos de TBI.

Em Zikic et al. (2012), é apresentado um método para segmentação automática de gliomas de alto grau e suas sub regiões a partir de imagens de MRI multicanal. Neste trabalho, os autores propuseram uma abordagem diferente, fazendo uso de florestas de decisão com características espaciais sensíveis ao contexto. Os resultados do método mostram que a classificação simultânea dos tipos individuais de tecidos utilizada na abordagem proposta o tornam eficiente e menos complexo, permitindo realizar segmentações precisas e com menor custo.

No trabalho de Rao et al. (2014) é proposto um segmentador de contusões automático para MRI multi-modalidade utilizando a abordagem de RF. Para tal, foi utilizado um conjunto de dados de MRI de 23 pessoas e, para uma melhor identificação de cada tipo de biomarcador, as imagens são corrigidas de diferentes formas em uma fase de pré-processamento. Para o treinamento, cada *voxel*[1] da imagem de teste passa por uma floresta gerada de 500 árvores, cada uma com a sua profundidade máxima, onde é definida a probabilidade daquele *voxel* pertencer ou não a uma contusão. A avaliação foi feita utilizando uma validação cruzada de 6 vezes e a quantificação da qualidade da segmentação foi feita utilizando o coeficiente de similaridade dice (DICE, 1945; SORENSEN,

[1] voxel: menor unidade em espessura na imagem tomográfica

1948). A abordagem proposta obteve uma boa segmentação para casos mais conhecidos, mas ainda apresenta resultados com baixa segmentação, muito provavelmente por esses casos mais particulares não possuírem uma quantidade de exemplos suficientes para o treinamento.

Em Molaei et al. (2016), foi definido um conjunto de características para avaliar a necessidade de se realizar uma tomografia computadorizada em pacientes com TBI.

A classificação realizada neste trabalho utiliza a abordagem de RF, onde foi identificado que o número de árvores de decisão a serem executadas é um dos parâmetros que mais impacta na obtenção de um resultado positivo, além do número de predições feitas em cada árvore.

No trabalho de Matsuo et al. (2020) é feita uma análise comparativa do uso de nove algoritmos de ML aplicados na predição de casos de TBI para um conjunto de dados de 232 casos com a utilização de diversos parâmetros possíveis. O resultado obtido indica que a utilização de RF apresenta melhor desempenho na previsão de resultados ruins e o uso de *ridge regression* para previsão de mortalidade. Além disso, os parâmetros mais relevantes para ambos os casos foram a idade, escala de coma de Glasgow (GCS - *Glasgow Coma Scale*), Produto da Degradação do Fibrinogênio (FDP - *fibrin/fibrinogen degradation products*) e glicose. O trabalho traz uma visão mais ampla de como a escolha dos parâmetros tem impacto importante no resultado dos modelos propostos.

Por fim, o trabalho de Thara e Thakul (2021), traz uma aplicação de ML para predizer o resultado de TBI pediátricos. Casos pediátricos são desconsiderados em diversos outros estudos por conta da diferença na definição das características a serem utilizadas como parâmetros. O trabalho realiza uma análise através do uso de diferentes algoritmos supervisionados, incluindo RF, para os modelos de treinamento. Os resultados obtidos seguem a tendência de que esses algoritmos apresentam um grande potencial para auxiliar nas decisões, por terem uma alta sensibilidade. Porém, ainda é preciso evoluir para que a acurácia destes modelos apresentem resultados mais satisfatórios.

Os trabalhos anteriormente descritos auxiliaram na decisão da utilização do algoritmo de floresta aleatória para a construção de modelos para

identificação dos atributos importantes. Essa decisão se baseou no adequado e promissor desempenho obtido com esse algoritmo na maioria dos estudos, além da versatilidade dele, dado que são possíveis diversos ajustes para um melhor desempenho caso a caso.

Destaca-se também que os trabalhos mencionados buscam apresentar soluções para a utilização de algoritmos de aprendizado de máquina como ferramenta auxiliar para predizer desfechos em casos de TBI. O presente trabalho segue essa mesma ideia, buscando identificar atributos que apresentem maior importância para predição de desfecho o que permite reduzir o número de atributos (biomarcadores) a serem considerados para construção de modelos de predição lineares. Por fim, identificar os biomarcadores mais importantes no contexto de TBI implicará em redução de custos financeiros na prática clínica, pois não necessitaria utilizar um grande número de biomarcadores para auxiliar no diagnóstico médico.

3 Desenvolvimento

Neste capítulo são apresentadas as ferramentas e métodos utilizados para o desenvolvimento do trabalho. Nas seções a seguir serão apresentadas as definições e detalhes das principais estratégias e metodologias utilizadas na abordagem proposta.

A análise exploratória dos dados clínicos envolvendo traumatismo cranioencefálico é explicada neste capítulo. Esta análise visa investigar uma relação exploratória entre biomarcadores, idade e sexo com função executiva em pacientes com traumatismo cranioencefálico. Esses biomarcadores são indicadores do estado médico de um paciente, que mostram a incidência de alguma função de um organismo ou uma resposta a um determinado agente farmacológico (STRIMBU; TAVEL, 2010). Dentro deste contexto, biomarcadores, idade e sexo são atributos e as funções executivas são variáveis-alvo.

A Tabela 3.1 apresenta os atributos e variáveis-alvo do conjuntos de dados que contém 21 amostras, isto é, dados de 21 pacientes. Salienta-se que estes dados utilizados foram anteriormente pré-processados para remover atributos que apresentavam valores discrepantes. Dessa forma, o conjunto de dados não possui dados espúrios (*outliers*). Devido ao pequeno conjunto de pacientes (dados), esse pré-processamento foi necessário para que não fosse necessário excluir amostras (dados de pacientes), ou seja, removeu-se todos biomarcadores discrepantes das amostras.

É importante informar que os dados clínicos aqui utilizados já foram aprovados pelo comitê de ética e são disponíveis através de parcerias que o orientador desta monografia possui dentro do Grupo Mineiro de Estudo em Traumatismo Cranioencefálico.

As variáveis alvo a serem analisadas estão associadas a *Hospital Anxiety and Depression Scale* (HADS) (ZIGMOND; SNAITH, 1983). Essa avaliação envolve uma série de perguntas feitas ao paciente, que são pontuadas com o intuito de definir se o paciente apresenta algum nível de ansiedade e/ou depressão.

Tabela 3.1 – Conjunto de dados: atributos e variáveis alvos.

Atributos	**Variáveis-Alvos**
sex, age, Copeptin, Angiotensin 1-7, Angiotensin 2, TFG alpha, IFN alpha 2, GRO-KC, MCP-3, MDC, sCD40L, IL-4, MCP-1, BDNF, Catepsin D, sICAM-1, MPO, PDGF AA, NCAM, PDGF-AB BB, PAI-1, MMP-9, Fibroblast, Neuropilin-1, Lipocalin-2, VEGF, MIF, LIGHT, RAGE, Enolase NSE, NRG-1 beta 1, HADS anxiety, HADS depression	HADS anxiety, HADS depression

Fonte: Elaborado pelo autor.

A regressão a ser feita nos dados com as variáveis alvo HADS *anxiety* e HADS *depression* tem como objetivo inicial a busca da relação dos demais atributos com o desfecho positivo para esses casos. Dessa forma, seria possível reduzir os atributos observados para a predição do desfecho para novos casos.

Para a determinação da pontuação de importância para cada atributo, são considerados os atributos sem *outliers* já obtidos com o propósito de assinalar um valor significativo para atributos de entrada baseados em quão útil eles são na predição de uma variável-alvo. Nesse passo, o algoritmo RF para regressão é investigado utilizando o *scikit-learn* (PEDREGOSA et al., 2011), uma ferramenta simples e eficiente para análise preditiva de dados em Python.

Para o melhor entendimento do conjunto de dados inicial e da associação entre os atributos (descrição da existência de relação entre uma variável e outra ou não), foi gerada uma matriz de correlação entre os atributos e as duas variáveis alvo (HADS *anxiety* e HADS *depression*).

A medida do grau de relação que uma variável tem com outras é realizada com a definição de coeficientes. Para a determinação do grau de correlação deste trabalho, foi utilizado o coeficiente padrão da função *corr*, que é o coeficiente de Pearson. O teste feito para definir o grau de correlação do coeficiente de Pearson é feito por meio do cálculo de direção (positiva ou negativa), onde assume valores no intervalo de -1 e 1. Essa correlação pode ser positiva, negativa

ou inexistente.

No contexto desta monografia, o algoritmo RF é aplicado com a variação dos parâmetros existentes definidos pela classe *RandomForestRegressor* do *scikit-learn*. Essa variação foi dividida em diversos casos, que são apresentados na Tabela 3.2. Para os parâmetros que assumem valores numéricos, foi feito o aumento progressivo do valor para um novo teste, afim de cobrir um intervalo maior de valores. Devido ao tamanho do conjunto de dados, preferiu-se trabalhar com valores mais baixos. Para a execução do parâmetro *n_estimators*, foram utilizados todos os outros parâmetros fixados com os respectivos valores padrão da classe. Já para os demais parâmetros, o valor do parâmetro *n_estimators* foi fixado com o valor 800 e os demais parâmetros com os valores padrão da classe.

Tabela 3.2 – Atributos com pontuação média maior ou igual ao limiar

Parâmetro	**Valores assumidos**
n_estimators	25, 50, 100, 200, 400 e 800
min_samples_split	2, 3, 4, 7, 10 e 15
max_features	sqrt, log2, 1, 2, 4 e 8
min_samples_leaf	1, 2, 3, 4, 7 e 10
criterion	*squared_error*, *absolute_error*, *friedman_mse* e *poisson*

Fonte: Elaborado pelo autor.

Para se realizar a avaliação da importância desses atributos é necessário realizar o cálculo do limiar (*mean score* de cada variável-alvo, que é a pontuação média geral de todos os atributos na predição daquela variável. Com o limiar definido, é feita a comparação entre a pontuação média obtida para o atributo em análise e o limiar. Caso o valor da pontuação desse atributo seja maior ou igual ao limiar, esse atributo é considerado importante para a predição.

Salienta-se que com os resultados obtidos e a avaliação realizada, é necessária uma validação desses resultados. Esse processo é realizado por um profissional da saúde (e.g. médico) capaz de definir se os atributos definidos como importantes apresentam sentido para a predição das variáveis.

Além dessa avaliação feita por um profissional, também é utilizada a medida R^2 (CHICCO; WARRENS; JURMAN, 2021 apud WRIGHT, 1921), também conhecida como R-*squared*, que assume valores entre 0 e 1 e, por definição, "[...] realiza a quantificação do quanto uma variável dependente é determinada pelas variáveis independentes, em termos de proporção de variância" (CHICCO; WARRENS; JURMAN, 2021). A fórmula geral para a determinação dessa medida é dada pela Equação 3.2, sendo utilizado o resultado da Equação 3.1, que faz o cálculo do valor médio dos valores verdadeiros.

$$\overline{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i^2, \tag{3.1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (X_i - Y_i)^2}{\sum_{i=1}^{m} (\overline{Y}_i - Y_i)^2}, \tag{3.2}$$

nos quais o X representa os valores preditos e o Y representa os valores verdadeiros.

Para a determinação da medida R^2, foi utilizada a própria classe *RandomForestRegressor*, que possui um método chamado *score* que retorna a pontuação R^2 calculada para o modelo.

4 Resultados

Neste capítulo são mostrados e analisados os resultados obtidos com o presente trabalho. Na Figura 4.1 é possível visualizar a matriz de correlação gerada com base no conjunto de dados inicial, com uma escala de cores (laranja, vermelho e roxo) e uma escala de -1 a 1.

A interpretação do grau de correlação irá seguir a definição do coeficiente de Pearson. Dessa forma, as variáveis dentro da escala de 0,7 a 1 (positivas e negativas) e na tonalidade laranja (positivas), e roxo escuro (negativas), possuem correlação forte. Já as variáveis entre 0,5 a 0,7 (positivas e negativas) e na tonalidade vermelha (positivas) e roxo médio (negativas), possuem correlação moderada. No caso das variáveis na escala de 0,3 a 0,5 (positivas e negativas) possuem correlação baixa. Por fim, as variáveis de 0 a 0,3 (positivas e negativas) e com a tonalidade vermelha, não possuem nenhuma correlação.

Para a análise dos resultados da execução do modelo para cada um dos parâmetros com os valores pré determinados, foram geradas 10 figuras contendo os gráficos de pontuação de importância de cada atributo, sendo 5 para a variável alvo HADS *anxiety* e 5 para a variável alvo HADS *depression*. O limiar (*mean score*) é sempre o mesmo e teve o valor calculado como sendo 0.032. Nas figuras com os gráficos das pontuações dos atributos, a linha tracejada vermelha representa o limiar. Além disso, foram geradas 10 tabelas, sendo que 5 contam com os atributos mais importantes e outras 5 contam com o resultado da pontuação R^2, ambas para cada variável alvo, com base nos parâmetros analisados. Dessa forma, as seções seguintes apresentam os resultados para cada um dos parâmetros.

Vale destacar que, para efeito de análise, serão tomados apenas os atributos e biomarcadores que estejam presente em 80% ou mais dos casos.

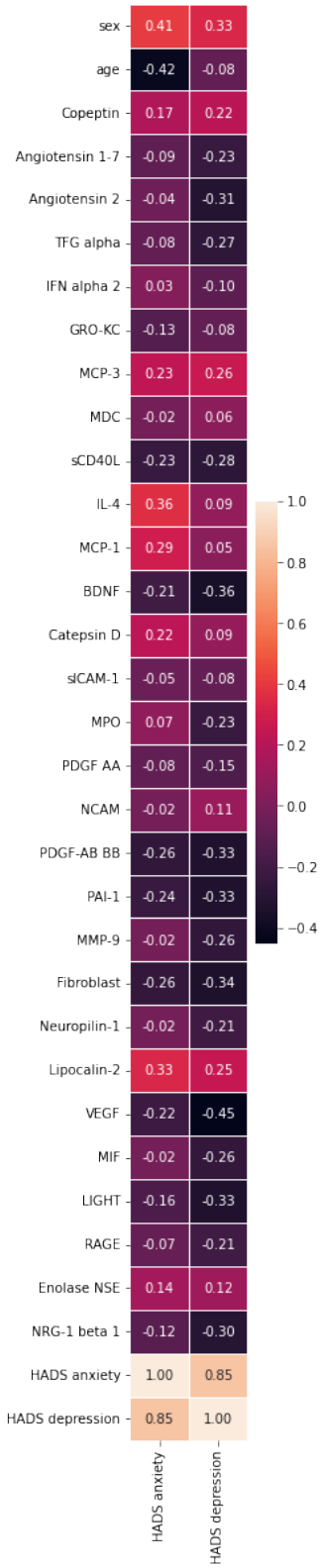


Figura 4.1 – Matriz de correlação entre as variáveis da HADS*anxiety* e HADS *depression* e as demais variáveis

Fonte: Elaborado pelo autor.

4.1 *n_estimators*

Nessa seção serão apresentados os resultados e análises para o parâmetro *n_estimators*, que define a quantidade de árvores de decisão a serem geradas na formação da floresta.

As Figuras 4.2 e 4.3 mostram os gráficos das pontuações dos atributos para as variáveis HADS *anxiety* e HADS *depression* e, com o auxílio da Tabela 4.1, é possível identificar que os atributos que obtiveram maior pontuação e, portanto, mais relevância.

Para o caso da HADS *anxiety*, os atributos e biomarcadores mais relevantes foram os de sexo e idade, MCP-1, LIGHT e RAGE, esses aparecendo em todos os casos de teste, e também o IL-4.

Para o caso da HADS *depression*, os atributos e biomarcadores mais relevantes foram os de idade, GRO-KC, sCD40L, MCP-1, Fibroblast, VEGF e LIGHT, aparecendo em todos os casos de teste, mas também os biomarcadores BDNF, MMP-9, PDGF-AB BB.

Diante dos resultados, temos que a idade, o MCP-1 e o LIGHT foram os atributos que mais apresentaram relevância de forma geral, tendo pontuações acima do limiar em todos os casos de teste para ambas as variáveis alvo.

Os valores apresentados para o R^2 foram bem próximos em ambos os casos, mas com valores melhores para um número maior de árvores, tendo os melhores valores para *n_estimator* de 200 ou acima. O melhor desempenho, nessa medida, foi com o valor de *n_estimator*=800, chegando a 80% para HADS *anxiety* e 82% para HADS *depression*. Essas medidas indicam que 80% da variação dos dados é explicada pela relação analisada para o caso da HADS *anxiety* e, para o caso da HADS *depression*, que 82% da variação dos dados é explicada.

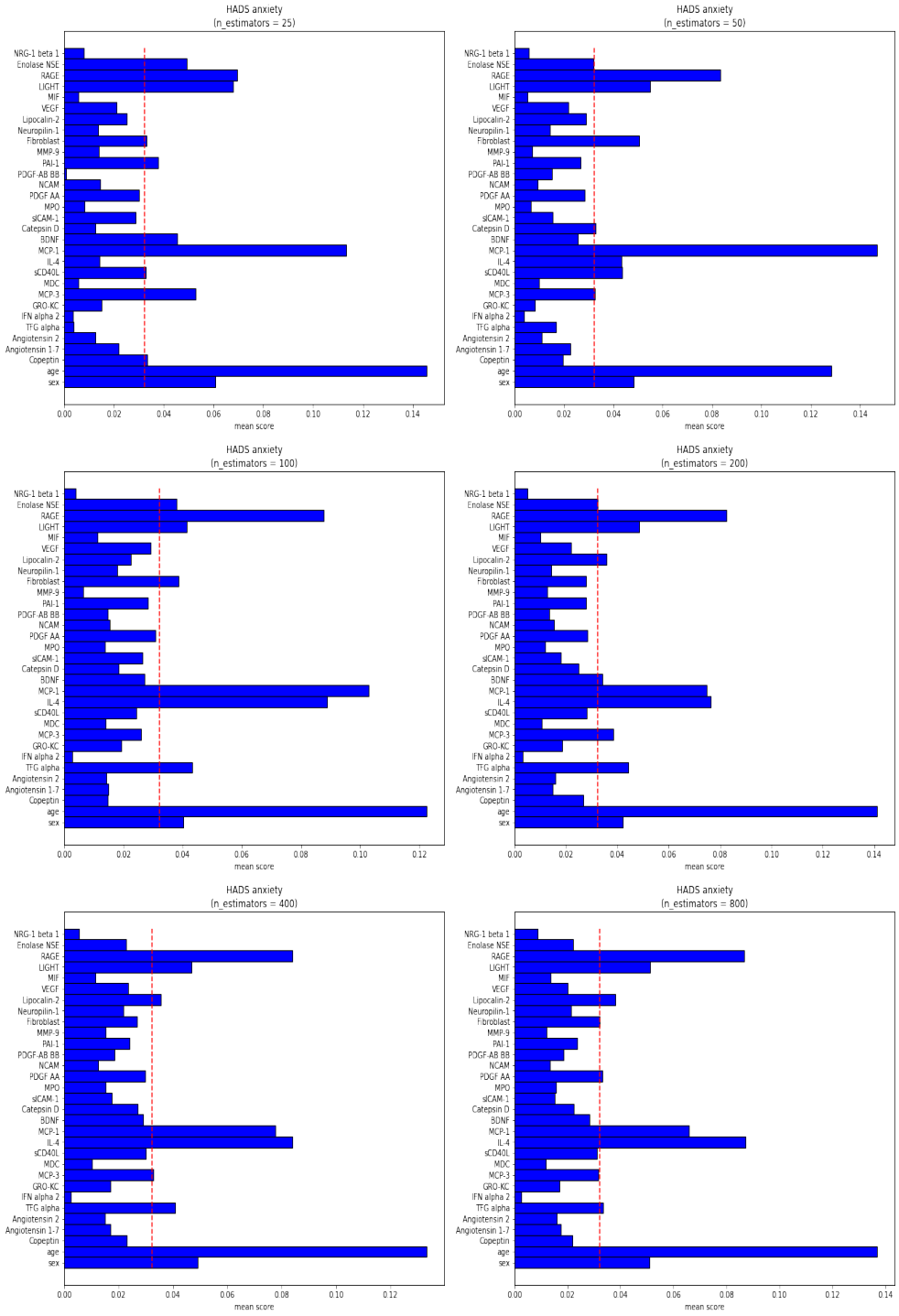


Figura 4.2 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. n_estimators= 25, 50, 100, 200, 400 e 800
Fonte: Elaborado pelo autor.

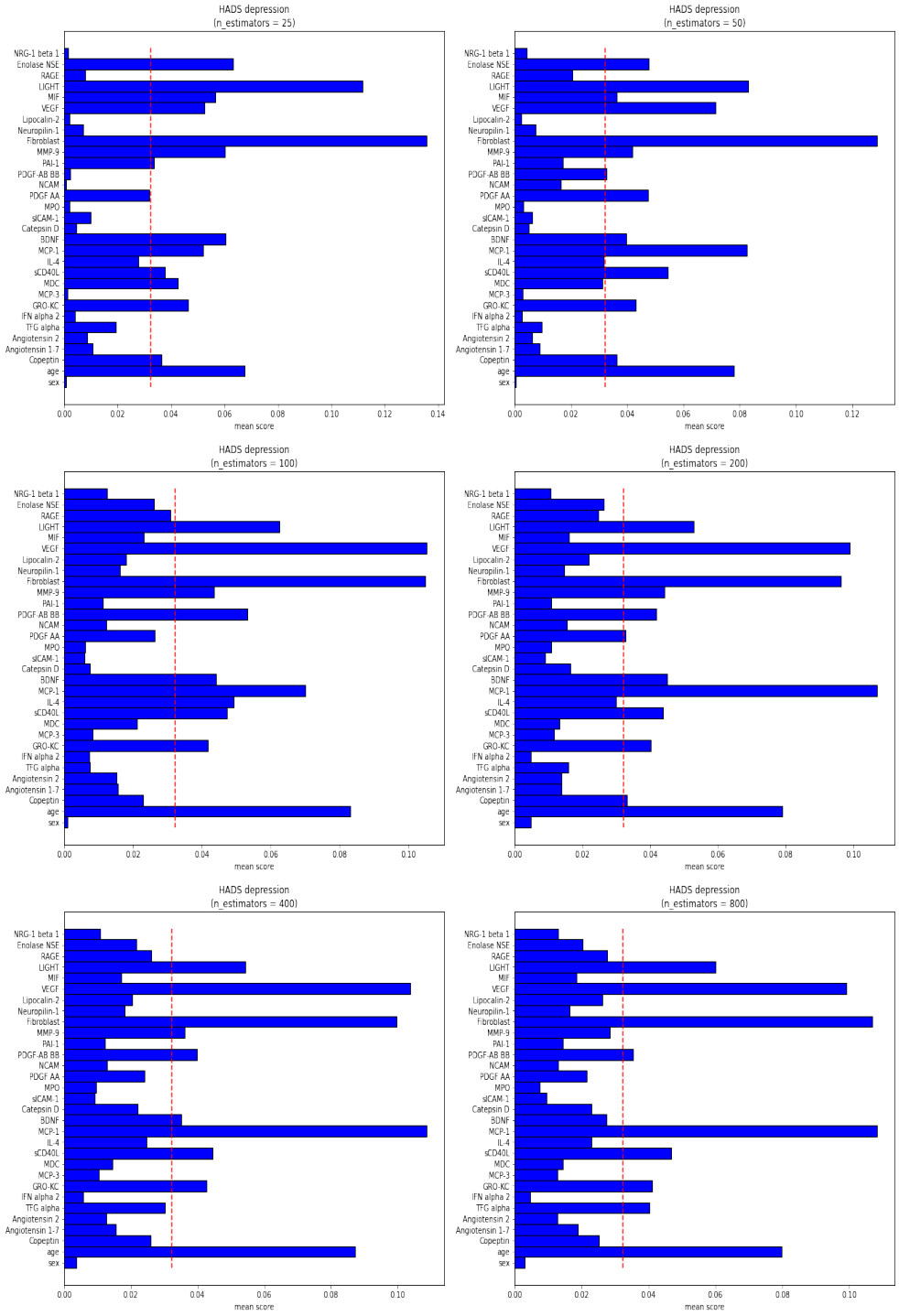


Figura 4.3 – Pontuação média das variáveis da HADS *depression* em casos de TBI. n_estimators= 25, 50, 100, 200, 400 e 800

Fonte: Elaborado pelo autor.

Tabela 4.1 – Atributos com pontuação média maior ou igual ao limiar para o parâmetro *n_estimators*

n_estimators	**HADS *anxiety***	**HADS *depression***
25	sex, age, Copeptin, MCP-3, sCD40L, MCP-1, BDNF, PAI-1, Fibroblast, LIGHT, RAGE, Enolase NSE	age, Copeptin, GRO-KC, MDC, sCD40L, MCP-1, BDNF, PAI-1, MMP-9, Fibroblast, VEGF, MIF, LIGHT, Enolase NSE
50	sex, age, MCP-3, sCD40L, IL-4, MCP-1, Catepsin D, Fibroblast, LIGHT, RAGE	age, Copeptin, GRO-KC, sCD40L, MCP-1, BDNF, PDGF AA, PDGF-AB BB, MMP-9, Fibroblast, VEGF, MIF, LIGHT, Enolase NSE
100	sex, age, TFG alpha, IL-4, MCP-1, Fibroblast, LIGHT, RAGE, Enolase NSE	age, GRO-KC, sCD40L, IL-4, MCP-1, BDNF, PDGF-AB BB, MMP-9, Fibroblast, VEGF, LIGHT
200	sex, age, TFG alpha, MCP-3, IL-4, MCP-1, BDNF, Lipocalin-2, LIGHT, RAGE, Enolase NSE	age, Copeptin, GRO-KC, sCD40L, MCP-1, BDNF, PDGF AA, PDGF-AB BB, MMP-9, Fibroblast, VEGF, LIGHT
400	sex, age, TFG alpha, MCP-3, IL-4, MCP-1, Lipocalin-2, LIGHT, RAGE	age, GRO-KC, sCD40L, MCP-1, BDNF, PDGF-AB BB, MMP-9, Fibroblast, VEGF, LIGHT
800	sex, age, TFG alpha, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT

Fonte: Elaborado pelo autor.

Tabela 4.2 – Pontuação R^2 para o parâmetro *n_estimators*

n_estimators	**HADS *anxiety***	**HADS *depression***
25	77%	77%
50	76%	76%
100	77%	76%
200	79%	81%
400	80%	81%
800	80%	82%

Fonte: Elaborado pelo autor.

4.2 *min_samples_split*

Nessa seção serão apresentados os resultados e análises para o parâmetro do *min_samples_split*, que define a quantidade de amostras necessárias para ser considerada a divisão de um nó interno dentro da árvore de decisão.

As Figuras 4.4 e 4.5 mostram os gráficos das pontuações dos atributos para as variáveis HADS *anxiety* e HADS *depression* e, com o auxílio da Tabela 4.3, é possível identificar que os atributos que obtiveram maior pontuação e, portanto, mais relevância.

Para o caso da HADS *anxiety*, os atributos e biomarcadores mais relevantes foram os de sexo, idade, IL-4, MCP-1, Lipocalin-2, LIGHT e RAGE, esses aparecendo em todos os casos de teste, e também o MCP-3.

Para o caso da HADS *depression*, os atributos e biomarcadores mais relevantes foram os de idade, TFG alpha, MCP-1, LIGHT, GRO-KC, sCD40L, Fibroblast e VEGF, aparecendo em todos os casos de teste, e também o PDGF-AB BB.

Diante dos resultados, temos que a idade, MCP-1 e LIGHT foram os atributos que mais apresentaram relevância de forma geral, tendo pontuações acima do limiar em todos os casos de teste para ambas as variáveis alvo.

Os valores apresentados para o R^2 tiveram bastante variação, apresentando queda considerável a cada aumento do valor. Os melhores valores para *min_samples_split* foram de 4 ou abaixo, sendo que o melhor desempe-

nho, nessa medida, foi com o valor de *min_samples_split*=2 (valor padrão), chegando a 80% para HADS *anxiety* e 82% para HADS *depression*.

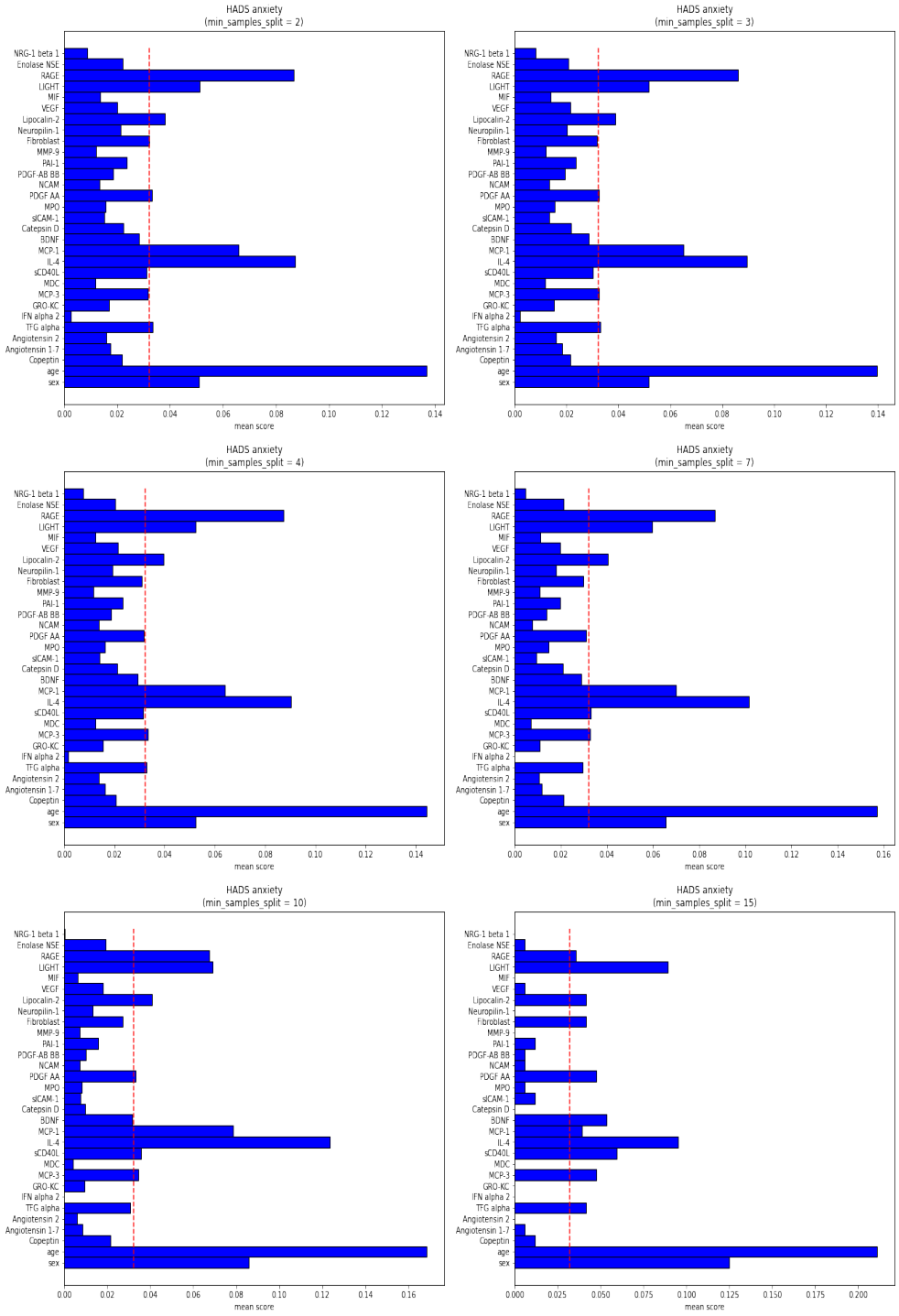


Figura 4.4 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. min_samples_split= 2, 3, 4, 7, 10, 15
Fonte: Elaborado pelo autor.

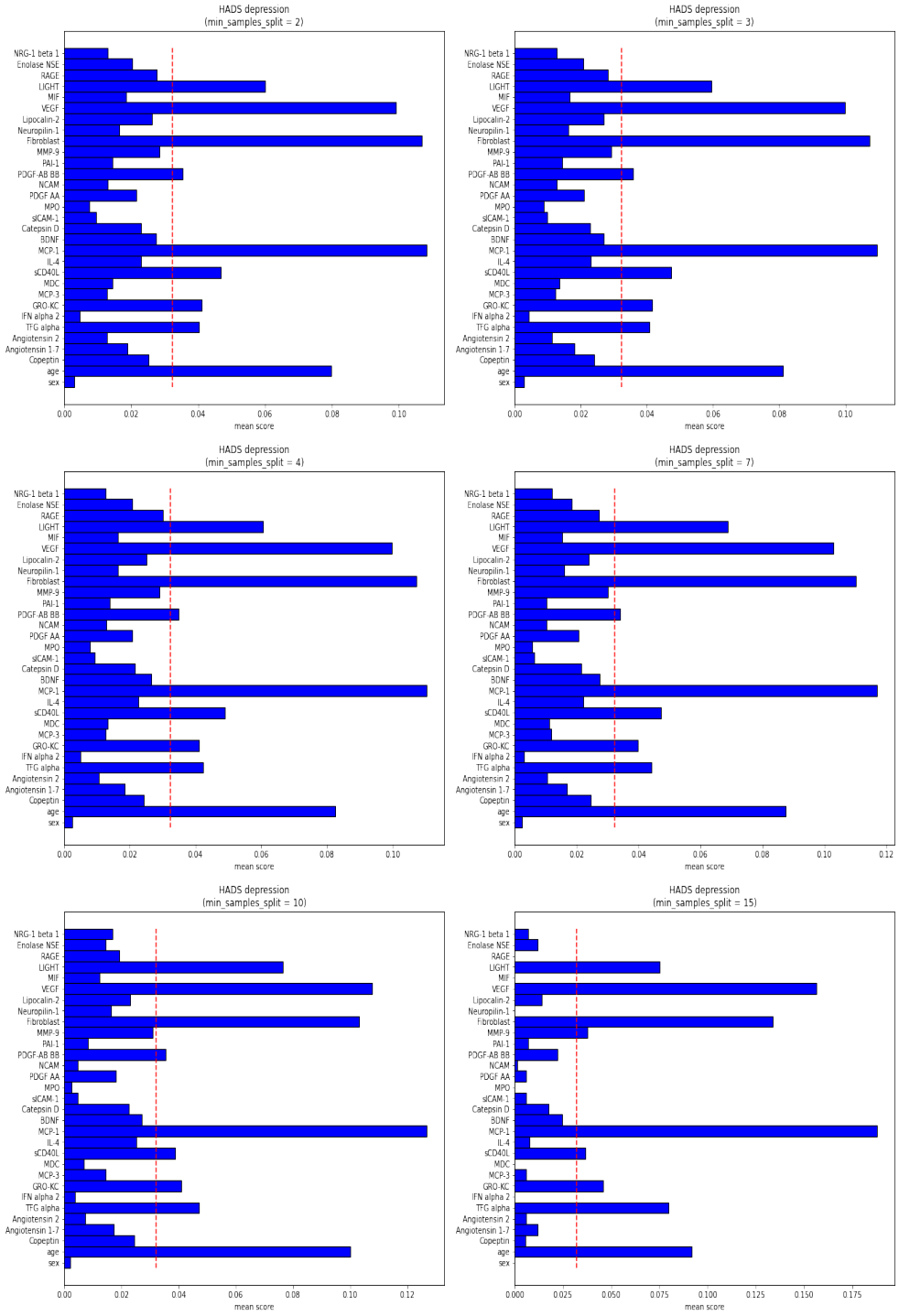


Figura 4.5 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. min_samples_split= 2, 3, 4, 7, 10, 15

Fonte: Elaborado pelo autor.

Tabela 4.3 – Atributos com pontuação média maior ou igual ao limiar para o parâmetro *min_samples_split*

min_samples_split	**HADS *anxiety***	**HADS *depression***
2	sex, age, TFG alpha, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT
3	sex, age, TFG alpha, MCP-3, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT
4	sex, age, TFG alpha, MCP-3, IL-4, MCP-1, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT
7	sex, age, MCP-3, sCD40L, IL-4, MCP-1, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT
10	sex, age, MCP-3, sCD40L, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT
15	sex, age, TFG alpha, MCP-3, sCD40L, IL-4, MCP-1, BDNF, PDGF AA, Fibroblast, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, MMP-9, Fibroblast, VEGF, LIGHT

Fonte: Elaborado pelo autor.

Tabela 4.4 – Pontuação R^2 para o parâmetro *min_samples_split*

min_samples_split	**HADS *anxiety***	**HADS *depression***
2	80%	82%
3	79%	80%
4	77%	79%
7	65%	72%
10	53%	64%
15	11%	16%

Fonte: Elaborado pelo autor.

4.3 *max_features*

Nessa seção serão apresentados os resultados e análises para o parâmetro do *max_features*, que define a quantidade máxima de atributos que o modelo considera ao fazer uma divisão na árvore de decisão.

As Figuras 4.6 e 4.7 mostram os gráficos das pontuações dos atributos para as variáveis HADS *anxiety* e HADS *depression* e, com o auxílio da Tabela 4.5, é possível identificar que os atributos que obtiveram maior pontuação e, portanto, mais relevância.

Para o caso da HADS *anxiety*, os atributos e biomarcadores mais relevantes foram os de idade, TFG alpha, MCP-1, PDGF-AB BB, Lipocalin-2, VEGF, LIGHT e RAGE, esses aparecendo em todos os casos de teste, e também os biomarcadores IL-4 e PDGF AA.

Para o caso da HADS *depression*, os atributos e biomarcadores mais relevantes foram os de idade, TFG alpha, MCP-1, PDGF-AB BB, VEGF, LIGHT, GRO-KC, sCD40L, BDNF e Fibroblast, aparecendo em todos os casos de teste, mas também os biomarcadores Lipocalin-2 e MMP-9.

Diante dos resultados, temos que a idade, TFG alpha, MCP-1, PDGF-AB BB, VEGF e LIGHT foram os atributos que mais apresentaram relevância de forma geral, tendo pontuações acima do limiar em todos os casos de teste para ambas as variáveis alvo.

Os valores apresentados para o R^2 tiveram valores muito semelhantes para todos os casos analisados, apresentando quase o mesmo valor em todos os

casos. Dessa forma, todos os valores testados para *max_features* apresentaram boas medidas, sendo que o melhor desempenho, nessa medida, foi com o valor de *max_features*=1 (valor padrão), chegando a 84% para HADS *anxiety* e 85% para HADS *depression*.

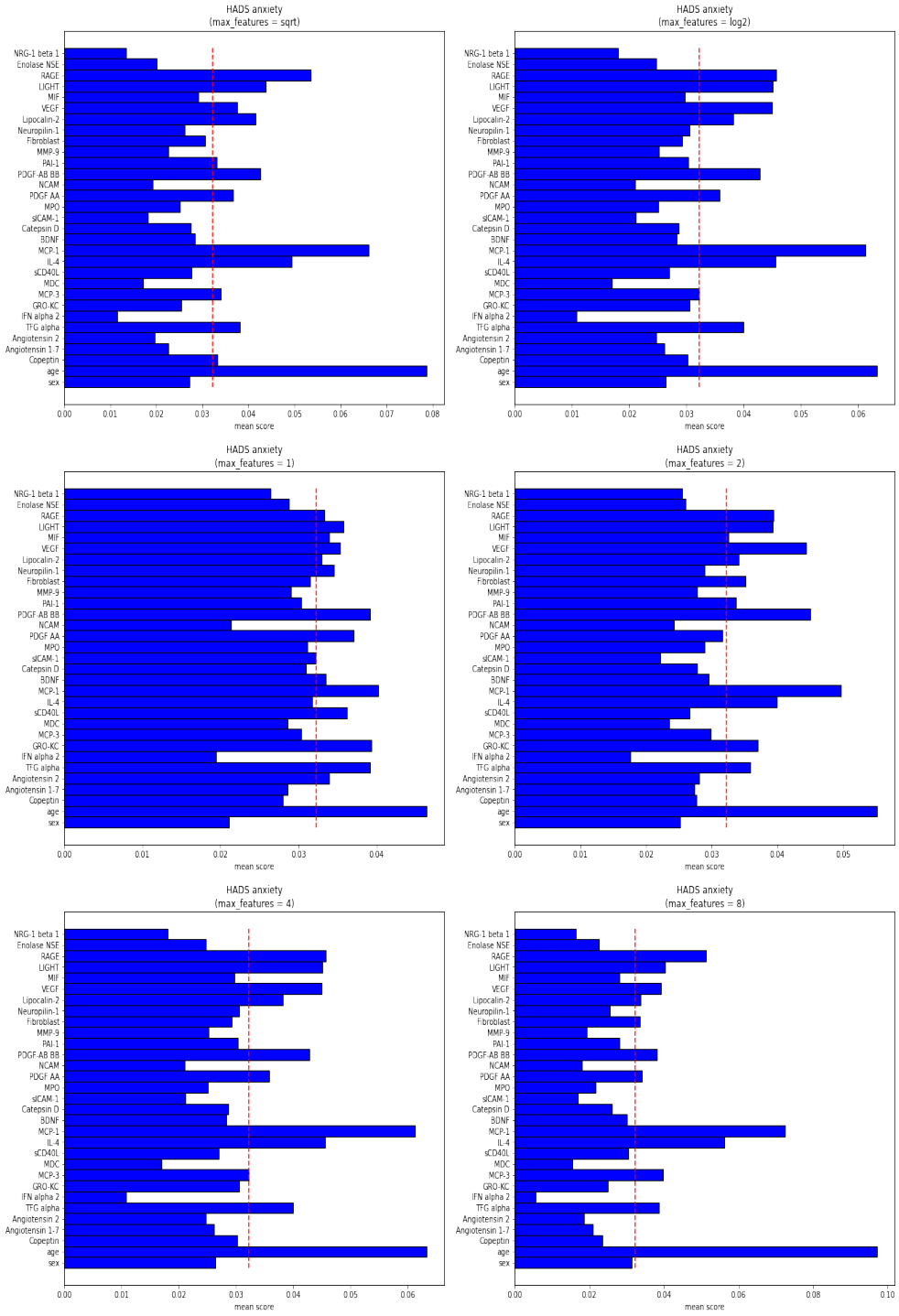


Figura 4.6 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. max_features= sqrt, log2, 1, 2, 4, 8

Fonte: Elaborado pelo autor.

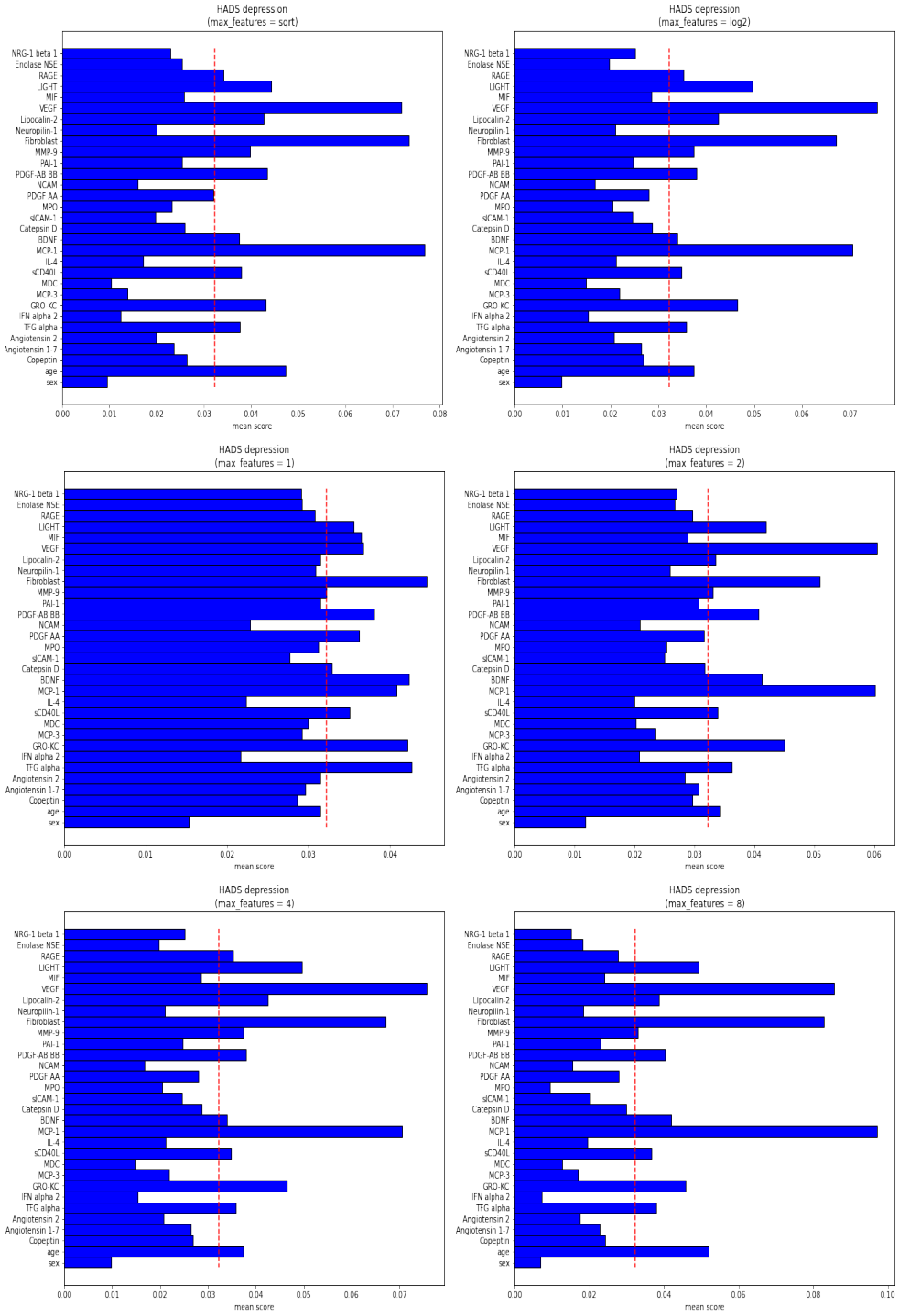


Figura 4.7 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. max_features= sqrt, log2, 1, 2, 4, 8

Fonte: Elaborado pelo autor.

Tabela 4.5 – Atributos com pontuação média maior ou igual ao limiar para o parâmetro *max_features*

max_features	**HADS *anxiety***	**HADS *depression***
sqrt	age, Copeptin, TFG alpha, MCP-3, IL-4, MCP-1, PDGF AA, PDGF-AB BB, PAI-1, Lipocalin-2, VEGF, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, BDNF, PDGF-AB BB, MMP-9, Fibroblast, Lipocalin-2, VEGF, LIGHT, RAGE
log2	age, TFG alpha, IL-4, MCP-1, PDGF AA, PDGF-AB BB, Lipocalin-2, VEGF, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, BDNF, PDGF-AB BB, MMP-9, Fibroblast, Lipocalin-2, VEGF, LIGHT, RAGE
1	age, Angiotensin 2, TFG alpha, GRO-KC, sCD40L, MCP-1, BDNF, PDGF AA, PDGF-AB BB, Neuropilin-1, Lipocalin-2, VEGF, MIF, LIGHT, RAGE	TFG alpha, GRO-KC, sCD40L, MCP-1, BDNF, Catepsin D, PDGF AA, PDGF-AB BB, Fibroblast, VEGF, MIF, LIGHT
2	age, TFG alpha, GRO-KC, IL-4, MCP-1, PDGF-AB BB, PAI-1, Fibroblast, Lipocalin-2, VEGF, MIF, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, BDNF, PDGF-AB BB, MMP-9, Fibroblast, Lipocalin-2, VEGF, LIGHT
4	age, TFG alpha, IL-4, MCP-1, PDGF AA, PDGF-AB BB, Lipocalin-2, VEGF, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, BDNF, PDGF-AB BB, MMP-9, Fibroblast, Lipocalin-2, VEGF, LIGHT, RAGE
8	age, TFG alpha, MCP-3, IL-4, MCP-1, PDGF AA, PDGF-AB BB, Fibroblast, Lipocalin-2, VEGF, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, BDNF, PDGF-AB BB, MMP-9, Fibroblast, Lipocalin-2, VEGF, LIGHT

Fonte: Elaborado pelo autor.

Tabela 4.6 – Pontuação R^2 para o parâmetro *max_features*

max_features	**HADS *anxiety***	**HADS *depression***
sqrt	83%	83%
log2	83%	83%
1	84%	85%
2	83%	84%
4	83%	83%
8	82%	83%

Fonte: Elaborado pelo autor.

4.4 *min_samples_leaf*

Nessa seção serão apresentados os resultados e análises para o parâmetro do *min_samples_leaf*, que define o número mínimo de amostras que são necessárias em um nó folha dentro da árvore de decisão.

As Figuras 4.8 e 4.9 mostram os gráficos das pontuações dos atributos para as variáveis HADS *anxiety* e HADS *depression* e, com o auxílio da Tabela 4.7, é possível identificar que os atributos que obtiveram maior pontuação e, portanto, mais relevância.

Para o caso da HADS *anxiety*, os atributos e biomarcadores mais relevantes foram os de sexo, idade, IL-4, MCP-1, PDGF AA, Lipocalin-2 e LIGHT, esses aparecendo em todos os casos de teste, e também o MCP-3.

Para o caso da HADS *depression*, os atributos e biomarcadores mais relevantes foram os de idade, LIGHT e Fibroblast, aparecendo em todos os casos de teste, além dos biomarcadores TFG alpha, sCD40L, Copeptin e BDNF.

Diante dos resultados, temos que a idade e o biomarcador LIGHT foram os atributos que mais apresentaram relevância de forma geral, tendo pontuações acima do limiar em todos os casos de teste para ambas as variáveis alvo.

Os valores apresentados para o R^2 tiveram muita variação, apresentando queda drástica em cada aumento do valor, chegando a não produzir os resultados para o modelo quando testado com o valor 10. Os melhores valores para *min_samples_leaf* foram de 2 ou abaixo, sendo que o melhor desempenho,

nessa medida, foi com o valor de *min_samples_leaf* =1 (valor padrão), chegando a 80% para HADS *anxiety* e 82% para HADS *depression*.

Figura 4.8 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. *min_samples_leaf* = 1, 2, 3, 4, 7, 10

Fonte: Elaborado pelo autor.

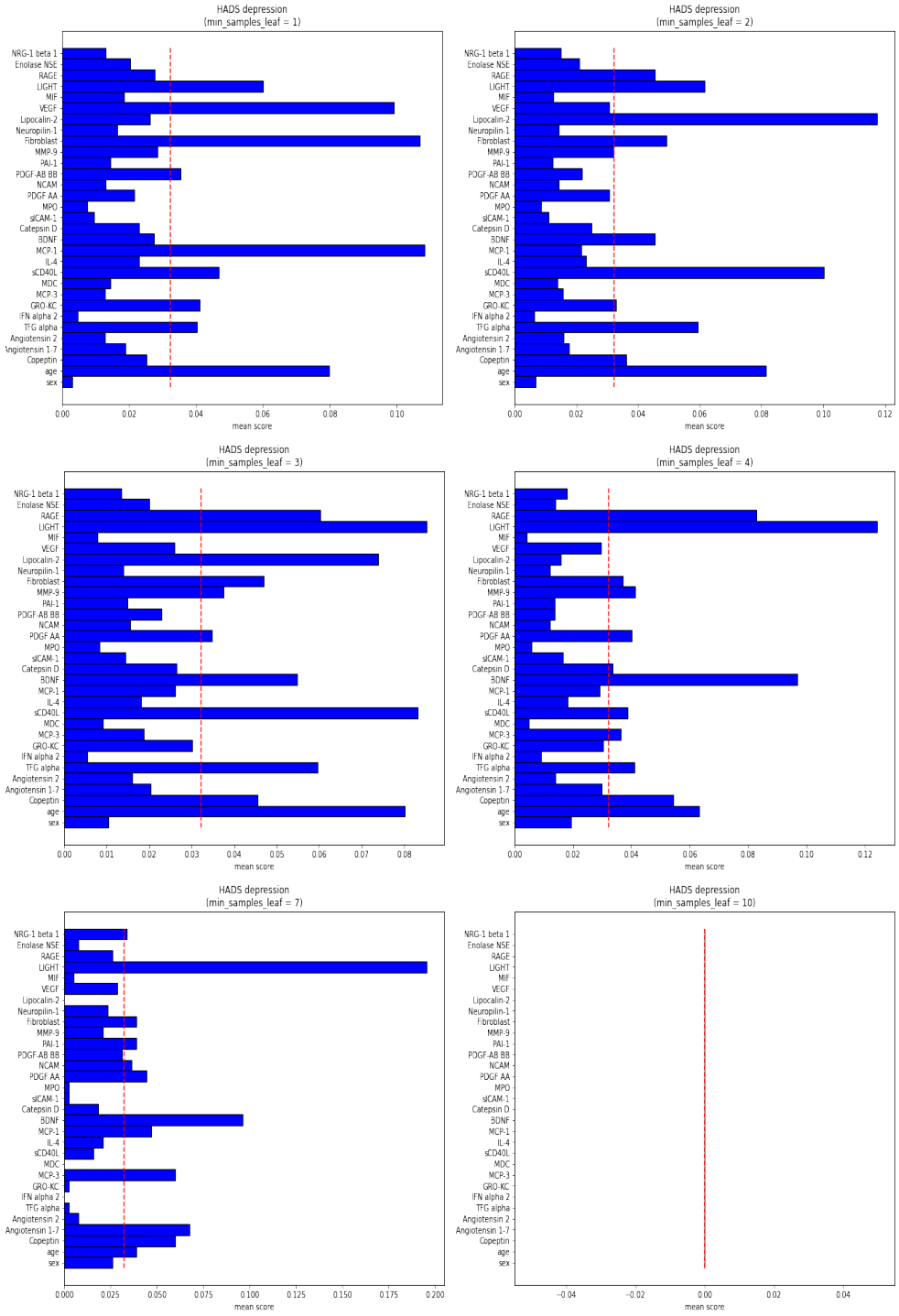


Figura 4.9 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. *min_samples_leaf* = 1, 2, 3, 4, 7, 10

Fonte: Elaborado pelo autor.

Tabela 4.7 – Atributos com pontuação média maior ou igual ao limiar para o parâmetro *min_samples_leaf*

min_samples_leaf	**HADS *anxiety***	**HADS *depression***
1	sex, age, TFG alpha, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT
2	sex, age, MCP-3, sCD40L, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, Copeptin, TFG alpha, GRO-KC, sCD40L, BDNF, Fibroblast, Lipocalin-2, LIGHT, RAGE
3	sex, age, MCP-3, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, Copeptin, TFG alpha, sCD40L, BDNF, PDGF AA, MMP-9, Fibroblast, Lipocalin-2, LIGHT, RAGE
4	sex, age, MCP-3, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, Copeptin, TFG alpha, MCP-3, sCD40L, BDNF, Catepsin D, PDGF AA, MMP-9, Fibroblast, LIGHT, RAGE
7	sex, age, Copeptin, MCP-3, sCD40L, IL-4, MCP-1, PDGF AA, PAI-1, Lipocalin-2, LIGHT	age, Copeptin, Angiotensin 1-7, MCP-3, MCP-1, BDNF, PDGF AA, NCAM, PAI-1, Fibroblast, LIGHT, NRG-1 beta 1
10	-	-

Fonte: Elaborado pelo autor.

Tabela 4.8 – Pontuação R^2 para o parâmetro *min_samples_leaf*

min_samples_leaf	**HADS *anxiety***	**HADS *depression***
1	80%	82%
2	73%	68%
3	64%	59%
4	53%	49%
7	19%	19%
10	-	-

Fonte: Elaborado pelo autor.

4.5 *criterion*

Nessa seção serão apresentados os resultados e análises para o parâmetro do *criterion*, que define a função utilizada para realizar a medida da qualidade de cada divisão feita pelo modelo.

As Figuras 4.10 e 4.11 mostram os gráficos das pontuações dos atributos para as variáveis HADS *anxiety* e HADS *depression* e, com o auxílio da Tabela 4.9, é possível identificar que os atributos que obtiveram maior pontuação e, portanto, mais relevância. Como o conjunto de parâmetros a serem analisados é menor, foram considerados os atributos e biomarcadores presentes em 70% ou mais dos casos.

Para o caso da HADS *anxiety*, os atributos e biomarcadores mais relevantes foram os de sexo, idade, IL-4, MCP-1, LIGHT e RAGE, esses aparecendo em todos os casos de teste, e também o TFG alpha.

Para o caso da HADS *depression*, os atributos e biomarcadores mais relevantes foram os de idade, MCP-1, GRO-KC, sCD40L, Fibroblast, VEGF e LIGHT, aparecendo em todos os casos de teste, mas também os biomarcadores TFG alpha e PDGF-AB BB.

Diante dos resultados, temos que a idade e os biomarcadores MCP-1 e LIGHT foram os atributos que mais apresentaram relevância de forma geral, tendo pontuações acima do limiar em todos os casos de teste para ambas as variáveis alvo.

Os valores apresentados para o R^2 não sofreu muita variação, apresentando valores bem próximo. Diante disso, todos os valores testados para *criterion* apresentaram boas medidas, sendo que o melhor desempenho, nessa medida, foi com o valor de *criterion*=absolute_error, chegando a 84% para HADS *anxiety* e o valores de *criterion*=squared_error e *criterion*=friedman_mse, chegando a 82% para HADS *depression*, em ambos os casos.

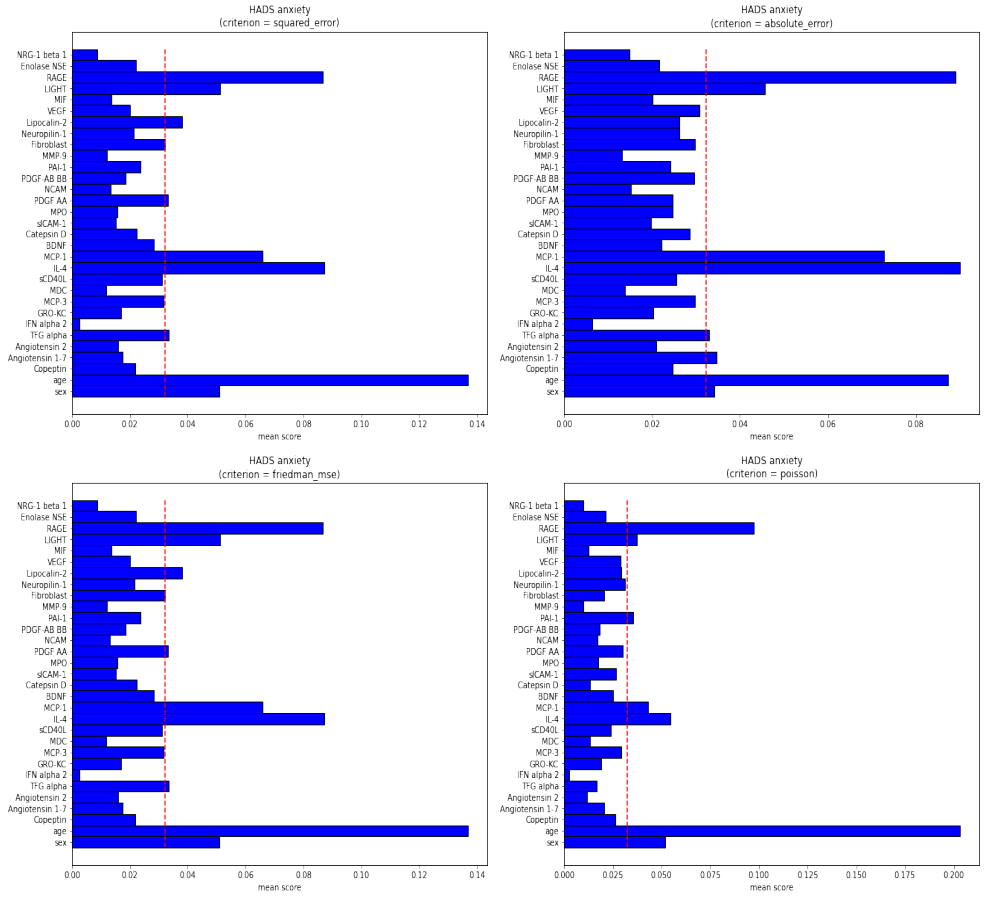


Figura 4.10 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. criterion= *squared_error*, *absolute_error*, *friedman_mse*, *poisson*

Fonte: Elaborado pelo autor.

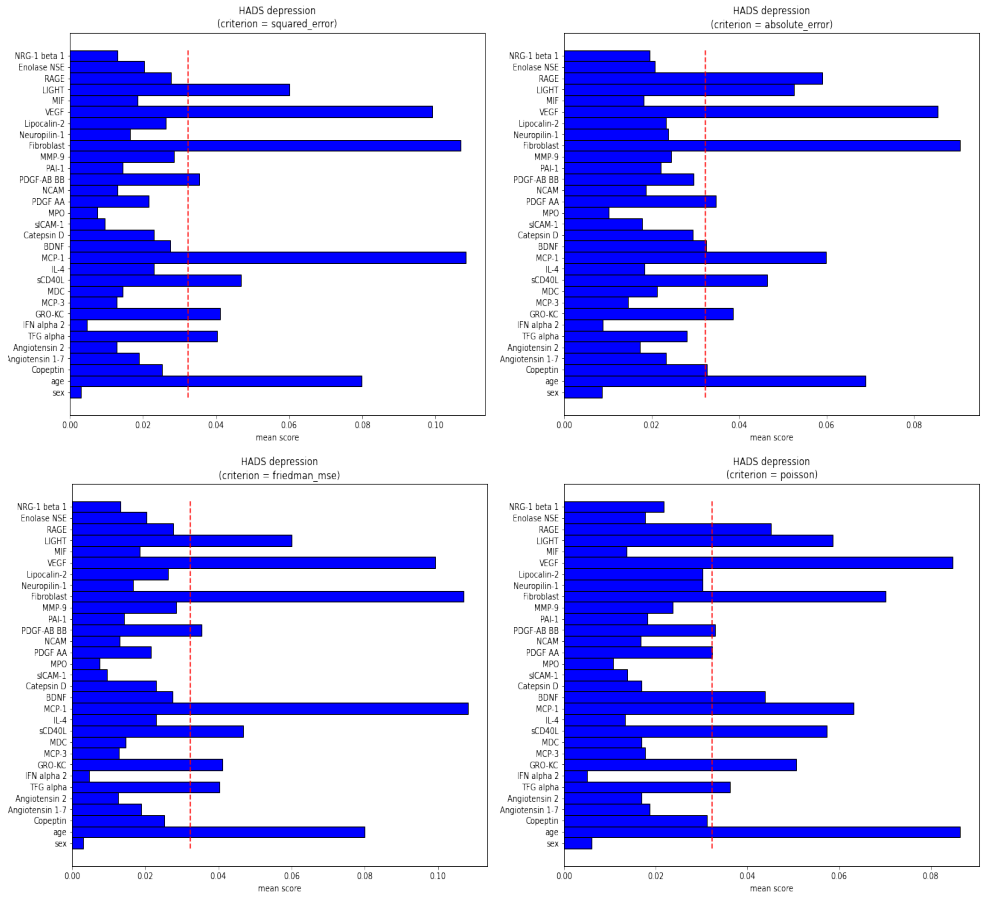


Figura 4.11 – Pontuação média das variáveis da HADS *anxiety* em casos de TBI. criterion= *squared_error*, *absolute_error*, *friedman_mse*, *poisson*

Fonte: Elaborado pelo autor.

Tabela 4.9 – Atributos com pontuação média maior ou igual ao limiar para o parâmetro *criterion*

criterion	**HADS *anxiety***	**HADS *depression***
squared_error	sex, age, TFG alpha, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT
absolute_error	sex, age, Angiotensin 1-7, TFG alpha, IL-4, MCP-1, LIGHT, RAGE	age, Copeptin, GRO-KC, sCD40L, MCP-1, BDNF, PDGF AA, Fibroblast, VEGF, LIGHT, RAGE
friedman_mse	sex, age, TFG alpha, IL-4, MCP-1, PDGF AA, Lipocalin-2, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, PDGF-AB BB, Fibroblast, VEGF, LIGHT
poisson	sex, age, IL-4, MCP-1, PAI-1, LIGHT, RAGE	age, TFG alpha, GRO-KC, sCD40L, MCP-1, BDNF, PDGF-AB BB, Fibroblast, VEGF, LIGHT, RAGE

Fonte: Elaborado pelo autor.

Tabela 4.10 – Pontuação R^2 para o parâmetro *criterion*

criterion	**HADS *anxiety***	**HADS *depression***
squared_error	80%	82%
absolute_error	81%	81%
friedman_mse	80%	82%
poisson	78%	81%

Fonte: Elaborado pelo autor.

Com base em todos os resultados obtidos para a variação de cada parâmetro, existem atributos e biomarcadores que estão presentes em todos ou quase todos os casos, independentemente do parâmetro e do valor assumido.

Portanto, esses atributos e biomarcadores representam maior relevância dentro do modelo. Em contrapartida, existem outras características que possuem baixíssima importância no modelo, uma vez que aparecem poucas vezes com pontuação acima do limiar ou sequer chegam a alcançar esse valor em algum dos casos analisados.

É importante ressaltar também que foi possível perceber que uma correlação de Pearson alta entre cada atributo e as variáveis-alvo (ver Figura 4.1) não correspondeu, necessariamente, na importância dos atributos determinadas pelos modelos de floresta aleatória. Isto se explica pelo fato que estes modelos de aprendizado de máquina conseguem levar em conta também as correlações entre os atributos para implicar na variável-alvo.

De forma geral, as características coletadas que representam maior importância para o modelo foram as de idade e LIGHT, MCP-1, TFG-alpha, IL-4, sCD40L, Fibroblast, VEGF e PDGF-AB BB.

5 Considerações Finais

Este capítulo apresenta algumas considerações finais sobre o trabalho desenvolvido e os objetivos atingidos, além de apresentar quais são os próximos passos identificados para o prosseguimento do mesmo.

5.1 Conclusão

Neste trabalho foi desenvolvido um modelo de RF para regressão do *scikit-learn* a fim de investigar os parâmetros e também os diversos atributos existentes no conjunto de dados e avaliar a importância de cada um na predição do desfecho em casos de TBI.

A partir do modelo de RF desenvolvido, foram investigados alguns parâmetros para analisar o comportamento da classe e o impacto do ajuste dos mesmos nas medidas de pontuação geradas para se ter uma definição melhor de quais valores os parâmetros podem assumir para um melhor desempenho e maior acurácia. Para tal, os parâmetros tiveram diversos valores testados e o resultado da pontação dos atributos, do limiar (pontuação média geral) e do R^2 foram utilizados para medir a qualidade do ajuste. Essas medidas permitiram identificar quais os valores de ajuste que cada um dos parâmetros pode assumir para se ter um desempenho satisfatório.

Em seguida, foram identificados os atributos mais importantes na predição do desfecho das variáveis alvo HADS *anxiety* e HADS *depression* a partir dos resultados obtidos com o modelo de RF para cada um dos ajustes de parâmetros. A identificação foi feita analisando os atributos que apresentaram pontuação acima do limiar para a maioria dos casos. Dessa forma, foi possível fazer a avaliação das características que apresentaram maior relevância.

Como contribuição deste trabalho, é possível destacar a identificação dos valores de ajuste para os parâmetros da classe *RandomForestRegressor* que produzem resultados com boa medida de desempenho. Além disso, este trabalho também traz, como principal objetivo, a verificação dos atributos

mais importantes para se realizar predição em casos clínicos de TBI, tendo alcançado estes resultados e conseguido identificar diversos atributos relevantes para diferentes casos.

5.2 Trabalhos Futuros

Como prosseguimento deste trabalho, o recomendado seria fazer ajustes para os parâmetros de forma conjunta, para analisar o impacto dentro do modelo ao se utilizar diferentes valores para vários parâmetros, uma vez que o os valores ajustados individualmente podem levar a um desempenho não tão bom, ao se agregar com outros.

Para calibração dos parâmetros dos modelos de floresta aleatória, pretende-se explorar os seguintes métodos: *Grid Search*, *Particle Swarm Optimization* (PSO), algoritmos genéticos.

Além disso, é importante que sejam feitos testes considerando outras variáveis-alvo, de forma que se tenha melhor avaliação do impacto que eles causam nos modelos de predição ou até mesmo uma predição multirrótulo, com o objetivo de predizer dois atributos ao mesmo tempo.

Por fim, é possível agregar mais técnicas para a melhor identificação dos atributos importantes para a predição do modelo, como a técnica de *permutation importance* (ALTMANN et al., 2010).

Referências

ALTMANN, A.; TOLOŞI, L.; SANDER, O.; LENGAUER, T. Permutation importance: a corrected feature importance measure. *Bioinformatics*, Oxford University Press, v. 26, n. 10, p. 1340–1347, 2010.

AMIT, Y.; GEMAN, D. Shape quantization and recognition with randomized trees. *Neural computation*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 9, n. 7, p. 1545–1588, 1997.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996.

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

BREIMAN, L.; FRIEDMAN, J.; STONE, C. J.; OLSHEN, R. *Classification and regression tree analysis*. [S.l.]: CRC Press: Boca Raton, FL, USA, 1984.

BROWNLEE, J. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. [S.l.]: Machine Learning Mastery, 2020.

CARTERI, R. B. K.; SILVA, R. A. d. Incidência hospitalar de traumatismo craniencefálico no brasil: uma análise dos últimos 10 anos. *Revista Brasileira de Terapia Intensiva*, SciELO Brasil, v. 33, p. 282–289, 2021.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, PeerJ Inc., v. 7, p. e623, 2021.

DICE, L. R. Measures of the amount of ecologic association between species. *Ecology*, JSTOR, v. 26, n. 3, p. 297–302, 1945.

EDUCATION, I. C. Random forest. *IBM. Retrieved*, v. 2020, 2022.

HSIA, R. Y.; MARKOWITZ, A. J.; LIN, F.; GUO, J.; MADHOK, D. Y.; MANLEY, G. T. Ten-year trends in traumatic brain injury: a retrospective cohort study of california emergency department and hospital revisits and readmissions. *BMJ open*, British Medical Journal Publishing Group, v. 8, n. 12, p. e022297, 2018.

HUANG, Y.-T.; HUANG, Y.-H.; HSIEH, C.-H.; LI, C.-J.; CHIU, I.-M. Comparison of injury severity score, glasgow coma scale, and revised trauma score in predicting the mortality and prolonged icu stay of traumatic young children: a cross-sectional retrospective study. *Emergency Medicine International*, Hindawi, v. 2019, 2019.

IJ, H. Statistics versus machine learning. *Nat Methods*, v. 15, n. 4, p. 233, 2018.

KAVOSI, Z.; JAFARI, A.; HATAM, N.; ENAAMI, M. The economic burden of traumatic brain injury due to fatal traffic accidents in shahid rajaei trauma hospital, shiraz, iran. *Archives of trauma research*, Kowsar Medical Institute, v. 4, n. 1, 2015.

KINNUNEN, K. M.; GREENWOOD, R.; POWELL, J. H.; LEECH, R.; HAWKINS, P. C.; BONNELLE, V.; PATEL, M. C.; COUNSELL, S. J.; SHARP, D. J. White matter damage and cognitive impairment after traumatic brain injury. *Brain*, Oxford University Press, v. 134, n. 2, p. 449–463, 2011.

KORLEY, F. K.; KELEN, G. D.; JONES, C. M.; DIAZ-ARRASTIA, R. Emergency department evaluation of traumatic brain injury in the united states, 2009–2010. *The Journal of head trauma rehabilitation*, NIH Public Access, v. 31, n. 6, p. 379, 2016.

MATSUO, 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*, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New ..., v. 37, n. 1, p. 202–210, 2020.

MOLAEI, S.; KORLEY, F. K.; SOROUSHMEHR, S. R.; FALK, H.; SAIR, H.; WARD, K.; NAJARIAN, K. A machine learning based approach for identifying traumatic brain injury patients for whom a head ct scan can be avoided. In: IEEE. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. [S.l.], 2016. p. 2258–2261.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2021.

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, Taylor & Francis, v. 58, n. 302, p. 415–434, 1963.

PARIKH, S.; KOCH, M.; NARAYAN, R. K. Traumatic brain injury. *International anesthesiology clinics*, LWW, v. 45, n. 3, p. 119–135, 2007.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011.

QUINLAN, J. R. *C4. 5: programs for machine learning*. [S.l.]: Elsevier, 2014.

RAO, A.; LEDIG, C.; NEWCOMBE, V.; MENON, D.; RUECKERT, D. Contusion segmentation from subjects with traumatic brain injury: a random forest framework. In: IEEE. *2014 IEEE 11th International Symposium on biomedical imaging (ISBI)*. [S.l.], 2014. p. 333–336.

ROSSUM, G. V.; DRAKE, F. L. *Python library reference*. [S.l.]: Centrum voor Wiskunde en Informatica, 1995.

SORENSEN, T. A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, v. 5, p. 1–34, 1948.

STRIMBU, K.; TAVEL, J. A. What are biomarkers? *Current Opinion in HIV and AIDS*, NIH Public Access, v. 5, n. 6, p. 463, 2010.

THARA, T.; THAKUL, O. Application of machine learning to predict the outcome of pediatric traumatic brain injury. *Chinese Journal of Traumatology*, Chinese Medical Journals Publishing House Co., Ltd. 42 Dongsi Xidajie ..., v. 24, n. 06, p. 350–355, 2021.

TRIFAN, G.; GATTU, R.; HAACKE, E. M.; KOU, Z.; BENSON, R. R. Mr imaging findings in mild traumatic brain injury with persistent neurological impairment. *Magnetic resonance imaging*, Elsevier, v. 37, p. 243–251, 2017.

WRIGHT, S. Correlation and causation. 1921.

ZIGMOND, A. S.; SNAITH, R. P. The hospital anxiety and depression scale. *Acta psychiatrica scandinavica*, Wiley Online Library, v. 67, n. 6, p. 361–370, 1983.

ZIKIC, D.; GLOCKER, B.; KONUKOGLU, E.; CRIMINISI, A.; DEMIRALP, C.; SHOTTON, J.; THOMAS, O. M.; DAS, T.; JENA, R.; PRICE, S. J. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr. In: SPRINGER. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. [S.l.], 2012. p. 369–376.