Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Desempenho e consistência entre grupos raciais em modelos de aprendizado de máquina para o desfecho pós-transplante de células-tronco hematopoiéticas

Rafael de Oliveira Vargas

JUIZ DE FORA
JULHO, 2026

Desempenho e consistência entre grupos raciais em modelos de aprendizado de máquina para o desfecho pós-transplante de células-tronco hematopoiéticas

Rafael de Oliveira Vargas

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Profa. D.Sc. Luciana Conceição Dias Campos

Coorientador: Profa. D.Sc. Priscila Vanessa Zabala Capriles Goliatt

JUIZ DE FORA

JULHO, 2026

Desempenho e consistência entre grupos raciais em modelos de aprendizado de máquina para o desfecho pós-transplante de células-tronco hematopoiéticas

Rafael de Oliveira Vargas

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Profa. D.Sc. Luciana Conceição Dias Campos
Universidade Federal de Juiz de Fora

Profa. D.Sc. Priscila Vanessa Zabala Capriles Goliatt
Universidade Federal de Juiz de Fora

Prof. D.Sc. Alex Borges Vieira
Universidade Federal de Juiz de Fora

Prof. D.Sc. Igor de Oliveira Knop
Universidade Federal de Juiz de Fora

JUIZ DE FORA
17 DE JULHO, 2026

Resumo

Este trabalho compara modelos de aprendizado de máquina para a classificação do indicador de sobrevivência livre de eventos (`efs`) em dados sintéticos de pacientes após o Transplante de Células-Tronco Hematopoiéticas (TCTH), oriundos da competição Equity in Post-HCT Survival Predictions, promovida pelo CIBMTR, avaliando tanto o desempenho global quanto a consistência da capacidade discriminativa entre grupos raciais. O objetivo central foi comparar o desempenho e a equidade de diferentes modelos de aprendizado de máquina, investigando se a superioridade preditiva de determinados métodos se mantém quando analisados possíveis vieses, especialmente relacionados a grupos raciais.

Para isso, foram comparados modelos de diferentes naturezas, incluindo algoritmos baseados em árvores, como XGBoost e Random Forest, técnicas de ensemble, como o Stacking Ensemble, e uma rede neural recorrente do tipo LSTM. A avaliação considerou métricas de desempenho global, como AUC-ROC, acurácia e precisão, além de métricas específicas de equidade, com destaque para o índice de concordância estratificado.

Os resultados indicam que no conjunto de teste analisado, XGBoost e Stacking Ensemble apresentaram os maiores valores numéricos nas métricas globais e de concordância. A configuração LSTM avaliada apresentou resultados inferiores e maior amplitude de C-index entre os grupos raciais. Como não foram estimados intervalos de confiança nem realizadas comparações inferenciais, as diferenças entre os modelos devem ser interpretadas de forma descritiva. Entretanto, esses resultados já indicam que determinadas abordagens podem contribuir para a construção de modelos mais equitativos.

Conclui-se que os resultados mostram que a ordenação dos modelos pode variar conforme a métrica utilizada e reforçam a importância de reportar o desempenho por subgrupos. Entretanto, por se tratar de uma análise de uma base sintética, sem validação externa e com avaliação de equidade restrita à variação do C-index, os achados não devem ser interpretados como evidência de equidade clínica ou de aplicabilidade direta à tomada de decisão, mas sim como uma contribuição para futuros estudos para a avaliação de suporte clínico.

Palavras-chave: Predição, Transplante de Células-Tronco Hematopoiéticas, *Machine Learning*, Equidade.

Abstract

This study compares machine learning models for the classification of the event-free survival indicator (`efs`) in synthetic data from patients after Hematopoietic Stem Cell Transplantation (HSCT), derived from the Equity in Post-HCT Survival Predictions competition promoted by the CIBMTR. The analysis evaluates both overall predictive performance and the consistency of discriminative ability across racial groups. The central objective was to compare the performance and fairness of different machine learning models, investigating whether the predictive superiority of certain methods is maintained when potential biases are analyzed, especially those related to racial groups.

To this end, models of different natures were compared, including tree-based algorithms, such as XGBoost and Random Forest; ensemble techniques, such as the Stacking Ensemble; and a recurrent neural network of the LSTM type. The evaluation considered global performance metrics, such as AUC-ROC, accuracy, and precision, as well as fairness-specific metrics, with emphasis on the stratified concordance index.

The results indicate that, in the analyzed test set, XGBoost and the Stacking Ensemble achieved the highest numerical values in the global and concordance metrics. The evaluated LSTM configuration showed lower results and a greater range of C-index values across ethnic groups. Since confidence intervals were not estimated and no inferential comparisons were performed, the differences between the models should be interpreted descriptively. Nevertheless, these results already suggest that certain approaches may contribute to the development of more equitable models.

It is concluded that the results show that the ranking of the models may vary according to the metric used and reinforce the importance of reporting performance by subgroups. However, because this analysis is based on a synthetic dataset, without external validation and with fairness evaluation restricted to C-index variation, the findings should not be interpreted as evidence of clinical fairness or direct applicability to decision-making, but rather as a contribution to future studies on the evaluation of clinical support

systems.

Keywords: Survival prediction, Hematopoietic Cell Transplantation, Machine Learning, Equity.

Agradecimentos

A todos os meus parentes, pelo encorajamento e apoio.

Às professoras Luciana Conceição Dias Campos e Priscila Vanessa Zabala Capriles Goliatt pela orientação, amizade e principalmente, pela paciência, sem a qual este trabalho não se realizaria.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC	Departamento de Ciência da Computação
UFJF	Universidade Federal de Juiz de Fora
TCTH	Transplante de Células-Tronco Hematopoéticas
CIBMTR	*Center for International Blood and Marrow Transplant Research*
LSTM	*Long Short-Term Memory*
RNA	Redes Neurais Artificiais
RF	*Random Forest*

1 Introdução

A competição *Equity in Post-HCT Survival Predictions*, promovida pelo CIBMTR (*Center for International Blood and Marrow Transplant Research*) e disponibilizada na plataforma *Kaggle* (CIBMTR, 2024), propõe o desenvolvimento de modelos preditivos para sobrevivência pós-transplante de células-tronco hematopoéticas. O diferencial da competição está no fato de que a avaliação não se restringe à precisão das previsões, mas também considera sua consistência entre grupos raciais.

As células-tronco hematopoéticas (CTH) são células precursoras localizadas principalmente na medula óssea e responsáveis pela formação das células do sangue. O transplante de CTH, popularmente conhecido como transplante de medula óssea, consiste na substituição da medula óssea doente ou danificada por células-tronco saudáveis, obtidas do próprio paciente ou de um doador compatível. Esse procedimento é utilizado no tratamento de doenças hematológicas graves, como leucemias, linfomas e anemias (COPELAN, 2006). Ainda assim, a mortalidade associada ao transplante permanece relevante, especialmente em razão de complicações como doença do enxerto contra o hospedeiro (*graft-versus-host disease* – GVHD), infecções, falência do enxerto e recaída da doença de base (APPELBAUM, 2001).

Nesse contexto, classificar o indicador de sobrevivência livre de eventos (`efs`), que assume valor 1 quando o evento clínico de interesse é observado e 0 quando o registro é censurado, e ordenar o risco pós-TCTH é uma tarefa complexa. Do ponto de vista computacional, os dados clínicos disponíveis costumam reunir variáveis de diferentes naturezas, valores ausentes, alta dimensionalidade e relações não lineares entre fatores prognósticos e desfechos clínicos (RAJKOMAR; DEAN; KOHANE, 2019; SORROR et al., 2005). Essas características dificultam a construção de modelos preditivos simples e motivam o uso de técnicas de aprendizado de máquina capazes de explorar interações mais complexas entre os atributos (GOLDSTEIN et al., 2017).

Entretanto, em aplicações clínicas, bom desempenho médio não é suficiente. Modelos treinados sobre bases de dados marcadas por desigualdades estruturais podem apre-

sentar previsões menos confiáveis para determinados grupos populacionais, mesmo quando suas métricas globais indicam desempenho satisfatório (OBERMEYER et al., 2019a). Por isso, a avaliação por subgrupos torna-se importante para verificar se o modelo mantém comportamento semelhante entre diferentes grupos raciais.

Este trabalho compara diferentes abordagens de aprendizado de máquina para a classificação do indicador `efs` e ordenação de risco pós-TCTH. São avaliados modelos baseados em árvores, uma arquitetura neural recorrente adaptada a dados tabulares e uma estratégia de empilhamento. A comparação considera tanto métricas globais de desempenho quanto métricas estratificadas por grupo racial, buscando verificar se os modelos com melhor desempenho geral também apresentam maior estabilidade entre os grupos analisados.

As hipóteses que orientam essa comparação são apresentadas na Seção 1.1. Os conceitos relacionados aos modelos e às métricas de avaliação são discutidos nos capítulos seguintes.

1.1 Hipóteses de Pesquisa

As hipóteses de pesquisa foram formuladas para orientar a comparação entre os modelos avaliados, considerando duas questões centrais: desempenho preditivo e equidade entre grupos raciais. Portanto, não se busca apenas identificar o modelo com melhor resultado global, mas verificar se esse desempenho se mantém de forma consistente entre os subgrupos analisados.

A primeira hipótese parte da natureza do conjunto de dados. Como a base contém variáveis clínicas, demográficas e procedimentais, investigou-se se modelos baseados em árvores, como Random Forest, XGBoost e LightGBM, seriam mais adequados do que uma LSTM adaptada a esse contexto. (BREIMAN, 2001; CHEN; GUESTRIN, 2016; KE et al., 2017; HOCHREITER, 1991).

A segunda hipótese considera que técnicas de *ensemble* podem produzir previsões mais robustas ao combinar diferentes modelos (DIETTERICH, 2000). Por isso, avaliou-se se o empilhamento dos modelos baseados em árvores também contribuiria para maior estabilidade entre grupos raciais, refletida na métrica de avaliação.

A terceira hipótese decorre da possibilidade de que métricas globais ocultem diferenças relevantes entre subgrupos. Em aplicações clínicas, modelos com bom desempenho médio podem apresentar previsões menos confiáveis para determinados grupos populacionais (OBERMEYER et al., 2019b; RAJKOMAR et al., 2020). Assim, investigou-se se o modelo com melhor desempenho global também seria o mais estável entre os grupos raciais avaliados.

Com base nesses pressupostos, foram definidas as seguintes hipóteses de pesquisa:

- **Hipótese 1:** Modelos baseados em árvores (*Random Forest*, *XGBoost* e *LightGBM*) apresentam desempenho preditivo superior ao modelo *LSTM* na classificação do indicador `efs` em dados tabulares.
- **Hipótese 2:** O uso de *empilhamento* de modelos (*Stacking*) reduz as diferenças observadas de desempenho entre grupos raciais, resultando em um *C-index* estratificado superior aos modelos individuais.
- **Hipótese 3:** Modelos com maior desempenho preditivo global não necessariamente apresentam maior estabilidade de desempenho entre diferentes grupos raciais.

1.2 Objetivos

1.2.1 Objetivo Geral

Avaliar modelos de aprendizado de máquina para a classificação do indicador de sobrevivência livre de eventos (`efs`) e a ordenação de risco pós-TCTH, considerando simultaneamente o desempenho preditivo global e a consistência do desempenho entre grupos raciais por meio de métricas gerais e estratificadas.

1.2.2 Objetivos Específicos

- **Pré-processamento dos dados:** Preparar o conjunto de dados da competição *CIBMTR* disponível na plataforma *Kaggle* (CIBMTR, 2024)
- **Implementar Modelos:** Treinar modelos de aprendizado de máquina, ajustando hiperparâmetros e avaliando seu desempenho.

- **Avaliar o desempenho por subgrupos:** analisar a consistência do desempenho dos modelos entre grupos raciais por meio do C-index por grupo, do C-index estratificado e da amplitude racial.

1.3 Organização do Trabalho

Este trabalho está estruturado da seguinte forma. No Capítulo 1, é apresentada a introdução do trabalho, contextualizando o problema da classificação do indicador `efs` e a relevância da avaliação por subgrupos em modelos de aprendizado de máquina na área da saúde, as hipóteses de pesquisa, os objetivos e a organização geral da monografia.

No Capítulo 2, são discutidos os fundamentos teóricos necessários para o desenvolvimento do trabalho, incluindo os modelos de aprendizado de máquina utilizados e as métricas de avaliação, com destaque para o *C-index* e o *C-index* estratificado.

No Capítulo 3, é apresentada a revisão sistemática e bibliográfica, abordando os principais estudos relacionados à predição de sobrevivência e equidade em modelos de aprendizado de máquina.

No Capítulo 4, são descritos os materiais e métodos adotados, incluindo o conjunto de dados, as etapas de pré-processamento, o treinamento dos modelos e os procedimentos de avaliação.

No Capítulo 5, são apresentados e analisados os resultados obtidos, com foco na comparação entre os modelos e na avaliação da variação de desempenho entre grupos raciais.

Por fim, no Capítulo 6, são apresentadas as conclusões do trabalho e são discutidas possíveis direções para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os fundamentos conceituais e metodológicos que embasam o presente estudo, organizado em duas seções. Primeiramente, descrevem-se os principais modelos de aprendizado de máquina aplicados à classificação do indicador `efs` e à ordenação de risco, com ênfase em algoritmos baseados em árvores de decisão e em redes neurais recorrentes. Na segunda seção, detalham-se as métricas de avaliação de desempenho, destacando o Índice de Concordância (*C-index*) e sua variante estratificada, utilizada neste trabalho para mensurar a variação da capacidade discriminativa entre grupos.

2.1 Análise de Sobrevivência

Neste trabalho, utilizam-se variáveis e uma métrica originárias da Análise de Sobrevivência, abordagem estatística voltada ao estudo do tempo até a ocorrência de um evento específico. A tarefa, porém, foi formulada como classificação binária do indicador `efs`, e não como modelagem direta do tempo até o evento. No conjunto de dados utilizado, a variável `efs` indica a ocorrência do evento clínico de interesse ou a censura do registro, enquanto a variável `efs_time` representa o tempo de acompanhamento até o evento ou até a censura. Como a documentação da base não especifica a natureza clínica exata do evento, não é possível determinar se ele corresponde, por exemplo, ao óbito, à falha do transplante ou a outro desfecho clínico. Por esse motivo, ele é tratado neste trabalho, de forma geral, como o evento associado à sobrevivência livre de eventos após o Transplante de Células-Tronco Hematopoiéticas (TCTH).

Para essa tarefa, os modelos baseiam-se em um conjunto de variáveis clínicas, demográficas e procedimentais coletadas sobre cada paciente. Entre elas, incluem-se dados demográficos, como idade, sexo e grupo racial; informações sobre a doença, como o diagnóstico e seu estágio; e características do transplante, como o tipo de doador e o regime de condicionamento utilizado.

O principal desafio computacional decorre da quantidade e da heterogeneidade

das variáveis disponíveis, além da possibilidade de relações não lineares entre essas informações e o desfecho observado. Por exemplo, o efeito da idade sobre o desfecho pode variar de acordo com o tipo de doença e o tratamento recebido.

Nesse contexto, modelos de aprendizado de máquina são empregados por sua capacidade de representar relações não lineares e interações entre múltiplas variáveis, sem exigir que todas essas relações sejam previamente especificadas (GOLDSTEIN et al., 2017).

2.2 Modelos de Aprendizado de Máquina

2.2.1 Algoritmos Baseados em Árvores de Decisão

Modelos de árvore de decisão dividem iterativamente o espaço de entradas por meio de regras lógicas até chegar em uma resposta, oferecendo boa interpretabilidade e robustez (BREIMAN et al., 1984). Em termos práticos, esse tipo de modelo realiza uma sequência de decisões baseadas em condições simples, como por exemplo, "se a idade for maior que 30", criando uma estrutura em forma de árvore. Cada divisão (ou *split*) separa os dados em grupos, permitindo que o modelo capture relações complexas de maneira intuitiva.

Random Forest

O *Random Forest* (BREIMAN, 2001) é um método de aprendizado de máquina do tipo *ensemble*[1] que combina múltiplas árvores de decisão para aumentar a robustez e a precisão preditiva do modelo. Cada árvore é treinada com uma amostra aleatória dos dados, utilizando a técnica de *bootstrap*[2], e avalia apenas um subconjunto aleatório de variáveis a cada divisão. Esse processo introduz diversidade entre as árvores, reduzindo a correlação entre elas. Durante a predição, os resultados das árvores são agregados por votação, no caso de classificação, ou por média, no caso da regressão, como ilustrado na Figura 2.1, o que contribui para reduzir a variância do modelo e mitigar o risco de sobreajuste. Por

[1] O termo *ensemble* refere-se a uma técnica que combina as previsões de múltiplos modelos base, geralmente fracos, para formar um modelo mais robusto e preciso. Essa abordagem busca reduzir a variância, o viés ou ambos, melhorando o desempenho preditivo geral (DIETTERICH, 2000).

[2] O *bootstrap* é um método de amostragem com reposição, no qual diferentes subconjuntos de dados são gerados a partir do conjunto original. Essa estratégia permite estimar a variabilidade do modelo e introduz diversidade entre as árvores do *ensemble* (EFRON, 1979).

sua capacidade de lidar bem com dados tabulares e seu bom desempenho em diferentes contextos clínicos, o *Random Forest* é amplamente utilizado em tarefas de análise de sobrevivência.

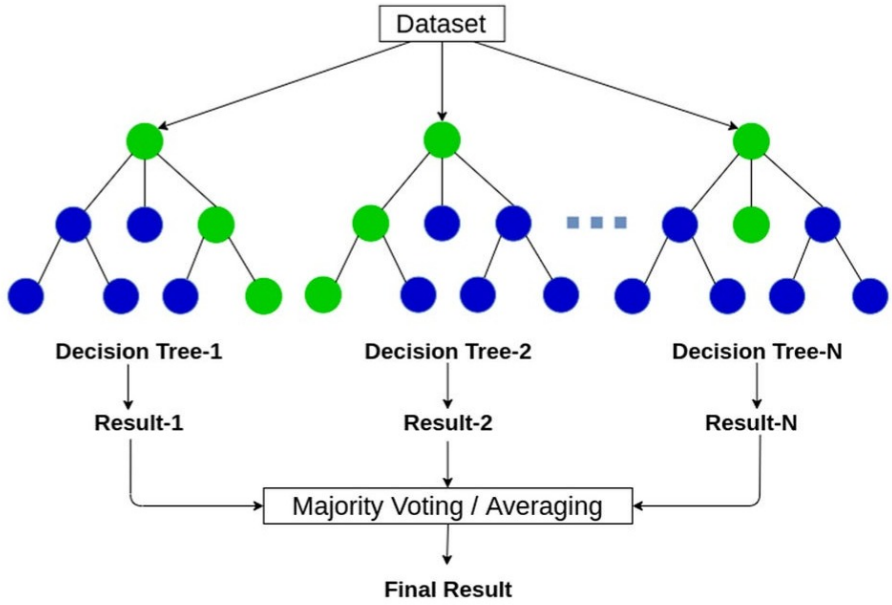

Figura 2.1: Representação gráfica de uma *Random Forest*, composta por múltiplas árvores de decisão cujos resultados são combinados por votação (classificação) ou média (regressão) para gerar a predição final. Adaptado de (JAIN, 2024).

XGBoost

O *XGBoost* (*Extreme Gradient Boosting*) (CHEN; GUESTRIN, 2016) é um algoritmo de aprendizado de máquina baseado em árvores que utiliza a técnica de *boosting*[3] para construir modelos sequenciais, onde cada nova árvore é treinada para corrigir os erros cometidos pelas anteriores. O processo é ilustrado na Figura 2.2, mostrando como o modelo concentra-se em exemplos mal preditos e ajusta-se iterativamente. O *XGBoost* se destaca por sua eficiência computacional, escalabilidade e robustez, incorporando técnicas para controlar o sobreajuste e otimizar o uso de memória durante o treinamento. Além disso, sua implementação permite o tratamento nativo de dados ausentes, tornando-o particularmente eficaz em cenários clínicos com bases incompletas ou heterogêneas.

[3] O *boosting* é uma técnica de aprendizado de máquina que combina vários modelos fracos, normalmente árvores de decisão rasas, de forma sequencial. A cada iteração, o algoritmo dá maior peso aos exemplos que foram mal classificados nas etapas anteriores, de modo que o novo modelo aprenda com os erros acumulados, resultando em um preditor final mais robusto e preciso. (SCHAPIRE, 1999)

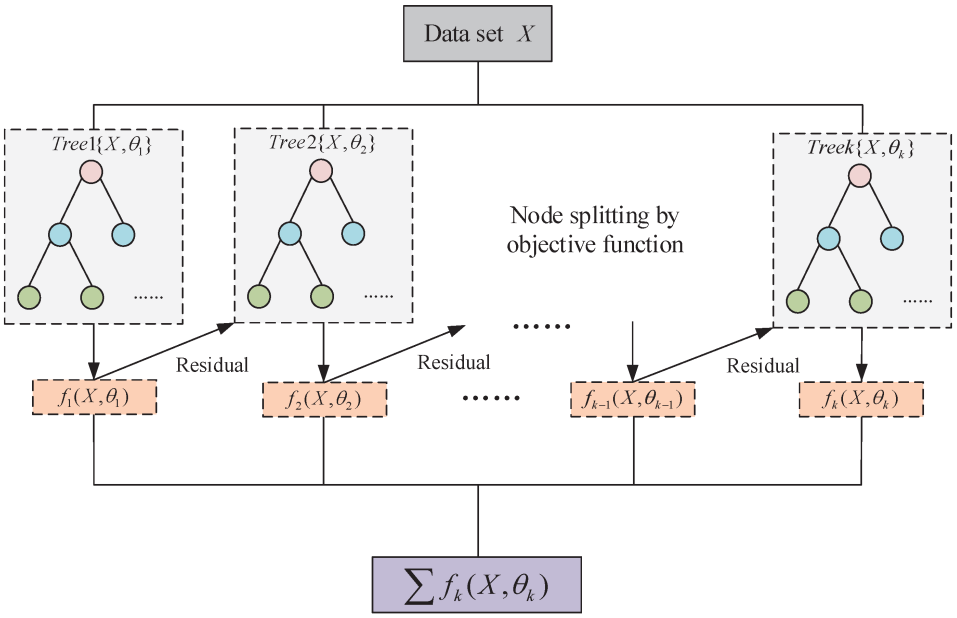

Figura 2.2: Representação esquemática do funcionamento do XGBoost. As árvores são construídas sequencialmente, de modo que cada nova árvore é ajustada a partir dos resíduos do modelo acumulado. A predição final resulta da soma das contribuições das árvores. Fonte: adaptado de (ZHOU et al., 2020).

LightGBM

O *LightGBM* (*Light Gradient Boosting Machine*) (KE et al., 2017) é uma biblioteca de *gradient boosting* concebida para alta eficiência computacional e escalabilidade. Analogamente ao *XGBoost*, o algoritmo constrói um empilhamento sequencial de árvores de decisão, em que cada árvore subsequente visa corrigir os resíduos das anteriores, atribuindo maior ponderação às instâncias com erro de predição elevado.

A distinção primária em relação a métodos convencionais reside na estratégia de expansão da estrutura de árvore. Enquanto abordagens como o *XGBoost* adotam um crescimento *level-wise* (por níveis), o *LightGBM* emprega uma expansão *leaf-wise* (por folhas), priorizando a divisão dos nós com ganho mais significativo. Essa metodologia promove uma convergência mais rápida e frequentemente resulta em maior acurácia, particularmente em conjuntos de dados de grande volume. (KE et al., 2017)

A eficiência do algoritmo é substancialmente amplificada pela integração de duas técnicas fundamentais: *Gradient-based One-Side Sampling* (GOSS) e *Exclusive Feature*

Bundling (EFB) (KE et al., 2017). O GOSS opera por meio de uma otimização amostral seletiva, preservando todas as instâncias com gradientes elevados e subamostrando estocasticamente aquelas com gradientes reduzidos, reduzindo custo computacional sem comprometer a precisão. O EFB, por sua vez, aborda a dimensionalidade por meio da agregação de características mutuamente exclusivas, variáveis que raramente assumem valores não nulos simultaneamente, em feixes (*bundles*), uma estratégia particularmente eficaz para dados esparsos.

A combinação da expansão *leaf-wise* com as técnicas GOSS e EFB confere ao *LightGBM* ganhos expressivos em velocidade e otimização de memória, posicionando-o como uma ferramenta distintamente adequada para aplicações de larga escala e alto volume de dados (KE et al., 2017).

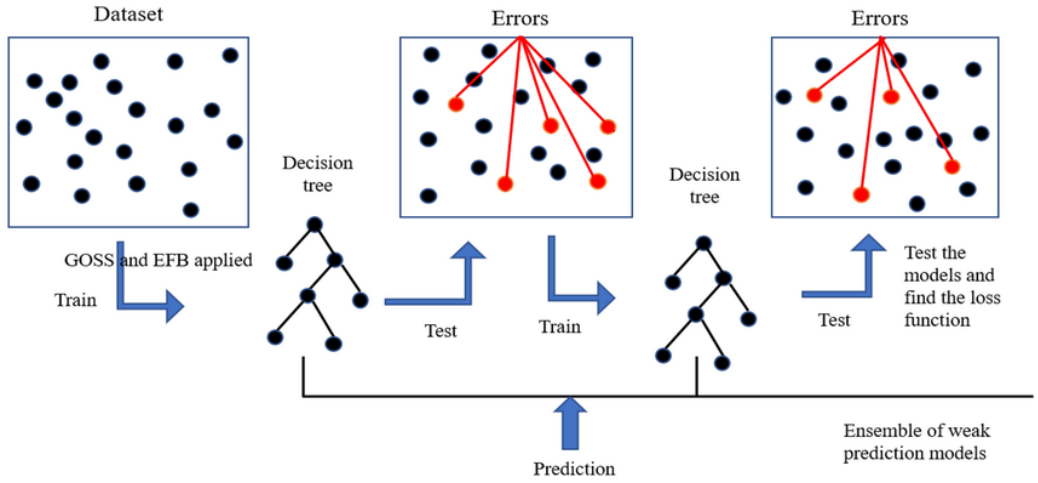

Figura 2.3: Representação simplificada do processo de treinamento do *LightGBM*, construção sequencial de árvores com correção de erros e aplicação das técnicas *GOSS* e *EFB*. Fonte: adaptado de (LIU et al., 2024).

2.2.2 Redes Neuronais Artificiais (RNAs)

As redes neuronais artificiais (RNAs) são modelos computacionais inspirados no cérebro humano, compostos por camadas de neurônios que processam informações, como podemos observar na Figura 2.4 (GOODFELLOW; BENGIO; COURVILLE, 2016). De modo geral, seu funcionamento baseia-se na interconexão entre unidades computacionais simples, capazes de representar relações complexas entre variáveis e extrair padrões a partir dos dados (BISHOP, 2006).

No contexto do aprendizado de máquina, é importante distinguir dois paradigmas fundamentais. No aprendizado supervisionado, o modelo é treinado com dados rotulados, isto é, exemplos em que cada entrada está associada a uma saída conhecida, permitindo o aprendizado de uma função de mapeamento entre variáveis de entrada e saída (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Já no aprendizado não supervisionado, os dados não possuem rótulos, e o objetivo do modelo passa a ser a identificação de estruturas, agrupamentos ou representações latentes presentes nos dados (BISHOP, 2006).

No caso das RNAs, sua aplicação é mais frequente em tarefas supervisionadas, como classificação e regressão, embora também existam arquiteturas empregadas em contextos não supervisionados. O conceito fundamental de neurônio artificial foi introduzido por McCulloch e Pitts (MCCULLOCH; PITTS, 1943), que é uma unidade computacional que recebe um conjunto de entradas, normalmente representadas por valores numéricos, e calcula uma saída a partir da combinação dessas entradas. Cada entrada é associada a um peso, que indica a importância daquela informação no processo de decisão do modelo. O neurônio realiza uma combinação linear dessas entradas ponderadas e, em seguida, aplica uma função de ativação para produzir a saída. Esse mecanismo permite que o modelo represente relações entre variáveis e aprenda padrões presentes nos dados. Esse funcionamento é ilustrado na Figura 2.4.

A formulação básica de um neurônio é descrita como na equação 2.1

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{2.1}$$

onde:

- y é a saída do neurônio,
- f é a função de ativação,
- w_i são os pesos associados às entradas x_i,
- b é o termo de viés (*bias*).

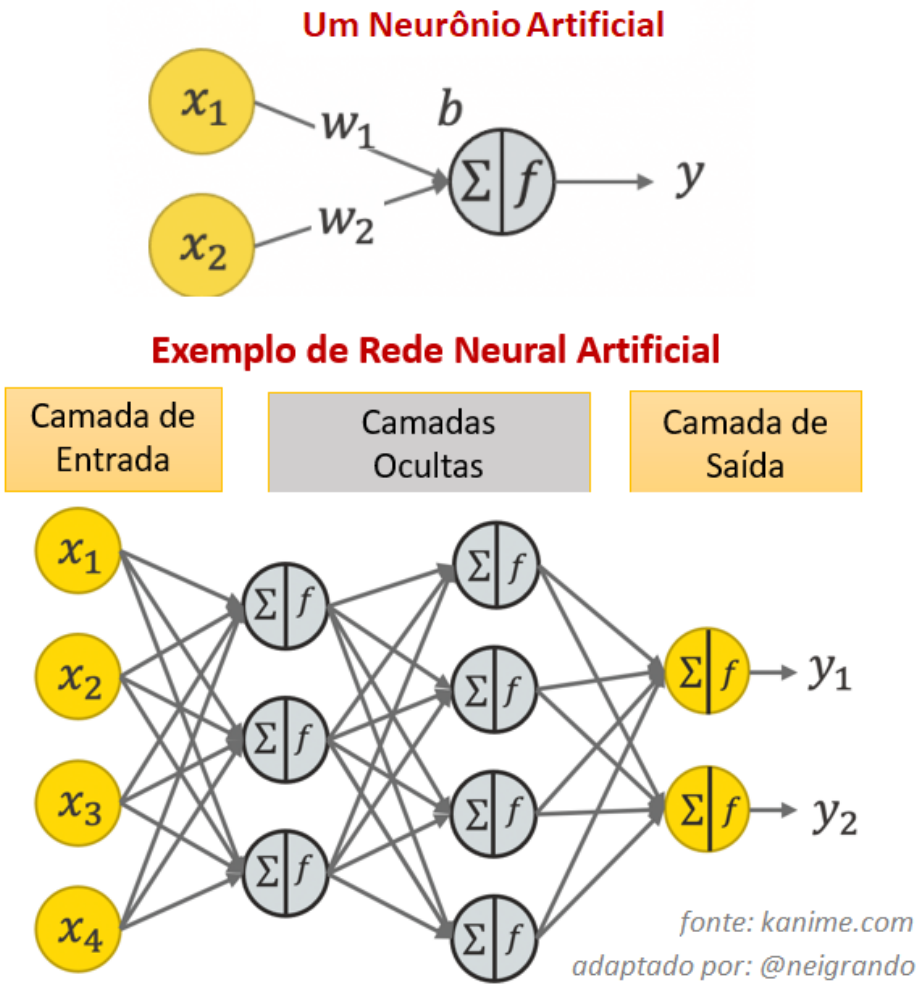

Figura 2.4: Ilustração de um neurônio artificial e uma rede neuronal (Neigrando, 2022)

O aprendizado das redes neuronais artificiais ocorre por meio do ajuste iterativo dos pesos e vieses dos neurônios, que constituem os parâmetros do modelo (GOODFELLOW; BENGIO; COURVILLE, 2016). Durante o treinamento, o objetivo é encontrar valores para esses parâmetros que permitam ao modelo produzir previsões cada vez mais próximas dos valores observados nos dados. Para isso, define-se uma função de custo (ou função de perda), responsável por quantificar o erro entre as previsões do modelo e os valores reais presentes no conjunto de treinamento (GOODFELLOW; BENGIO; COURVILLE, 2016). Dependendo da tarefa considerada, essa função pode assumir diferentes formas, como o erro quadrático médio em problemas de regressão ou a entropia cruzada em tarefas de classificação.

A otimização dessa função de custo é realizada com base no conceito de gradiente, que corresponde ao vetor de derivadas parciais da função em relação aos parâmetros da

rede (GOODFELLOW; BENGIO; COURVILLE, 2016). O gradiente indica a direção de maior variação da função de custo e fornece informação sobre como os parâmetros devem ser ajustados para reduzir o erro. Com base nesse princípio, utiliza-se frequentemente o algoritmo de descida do gradiente, que atualiza iterativamente os parâmetros do modelo.

O ajuste dos pesos é realizado, de forma geral, pela regra de atualização presente na equação (2.2).

$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \nabla_\theta J(\theta^{(t)}) \tag{2.2}$$

onde $\theta^{(t)}$ representa os parâmetros da rede na iteração t, η é a taxa de aprendizado e $\nabla_\theta J(\theta^{(t)})$ corresponde ao gradiente da função de custo J em relação aos parâmetros da rede naquele instante. Dessa forma, a cada iteração os parâmetros são ajustados na direção que reduz o valor da função de custo, conduzindo o modelo gradualmente a uma solução de menor erro. (GOODFELLOW; BENGIO; COURVILLE, 2016)

Durante o treinamento de redes profundas podem ocorrer problemas relacionados ao gradiente, como o desaparecimento do gradiente e a explosão do gradiente (HOCHREITER, 1991). O desaparecimento do gradiente ocorre quando os valores das derivadas se tornam muito pequenos ao serem propagados ao longo de múltiplas camadas ou passos temporais, dificultando a atualização dos parâmetros e levando à estagnação do aprendizado. Por outro lado, a explosão do gradiente ocorre quando esses valores se tornam excessivamente grandes, causando instabilidade no treinamento e dificultando a convergência do modelo. (GOODFELLOW; BENGIO; COURVILLE, 2016)

As redes neuronais são amplamente utilizadas em tarefas de classificação, regressão e reconhecimento de padrões, sendo capazes de modelar relações não lineares complexas entre as variáveis de entrada e saída (Neigrando, 2022).

2.2.3 Redes Neuronais Recorrentes (RNNs)

Redes neuronais tradicionais (*feedforward*) são geralmente eficazes para processar dados estáticos; entretanto, elas não conseguem capturar informações temporais entre instantes consecutivos, uma vez que assumem independência entre as entradas (GOODFELLOW; BENGIO; COURVILLE, 2016). Para lidar com dados sequenciais, como séries temporais,

textos ou sinais fisiológicos [4], é necessário um modelo capaz de considerar o contexto das observações anteriores.

As Redes Neuronais Recorrentes (RNNs) foram projetadas para lidar com esse tipo de dependência temporal (GOODFELLOW; BENGIO; COURVILLE, 2016). Elas introduzem conexões recorrentes que permitem que a saída de um neurônio em um instante de tempo influencie o processamento nos instantes seguintes, criando uma espécie de "memória" interna do sistema. Como ilustrado na Figura 2.5(a), as RNNs possuem conexões que realimentam a saída para o próprio modelo ao longo do tempo, permitindo a modelagem de dependências temporais. Em contraste, redes *feedforward*, representadas na Figura 2.5(b), processam cada entrada de forma independente, sem considerar informações de estados anteriores. Essa diferença estrutural torna as RNNs especialmente adequadas para modelar sequências temporais e dados contextuais.

No entanto, RNNs tradicionais enfrentam dificuldades para aprender dependências de longo prazo devido aos problemas de desaparecimento e explosão do gradiente durante o treinamento. Ressalta-se que, embora a arquitetura recorrente seja concebida para dados com estrutura temporal, os dados utilizados neste trabalho não possuem sequência temporal por paciente. A adaptação adotada, com comprimento de sequência igual a um, é detalhada na Figura 2.7 e retomada nas limitações.

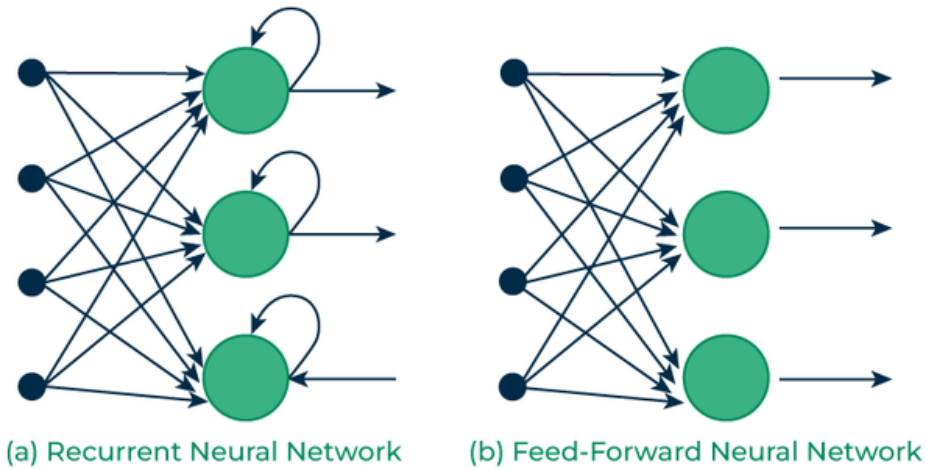

Figura 2.5: Comparação de redes neurais recorrentes (a) versus redes neurais padrão (b). Fonte: adaptado de (GEEKSFORGEEKS, 2025).

[4]Sinais fisiológicos são medidas geradas pelo organismo e registradas de forma contínua ao longo do tempo, como o eletrocardiograma (ECG) e o eletroencefalograma (EEG).

2.2.4 Long Short-Term Memory (LSTM)

A *Long Short-Term Memory (LSTM)* é uma arquitetura avançada de rede neural recorrente proposta por Hochreiter e Schmidhuber (HOCHREITER; SCHMIDHUBER, 1997), especificamente projetada para mitigar o problema do desaparecimento do gradiente em redes recorrentes tradicionais.

A principal inovação da LSTM está na introdução de uma célula de memória, representada por C_t, que permite a propagação da informação ao longo do tempo de forma mais estável. Essa célula é controlada por mecanismos conhecidos como portões (*gates*) de entrada, esquecimento e saída que regulam quais informações devem ser armazenadas, atualizadas ou descartadas ao longo das iterações. (HOCHREITER; SCHMIDHUBER, 1997).

Diferentemente das RNNs tradicionais, nas quais o gradiente é repetidamente multiplicado ao longo dos passos temporais, a LSTM realiza a atualização do estado da célula por meio de combinações aditivas. Em particular, o portão de esquecimento controla a proporção da informação anterior que deve ser mantida, permitindo que o gradiente seja preservado ao longo do tempo quando necessário. Esse mecanismo reduz o efeito de atenuação progressiva do gradiente, possibilitando o aprendizado de dependências de longo prazo de forma mais estável (HOCHREITER; SCHMIDHUBER, 1997). A estrutura de uma célula LSTM é apresentada na Figura 2.6, evidenciando os mecanismos responsáveis pelo controle do fluxo de informações ao longo do processamento temporal.

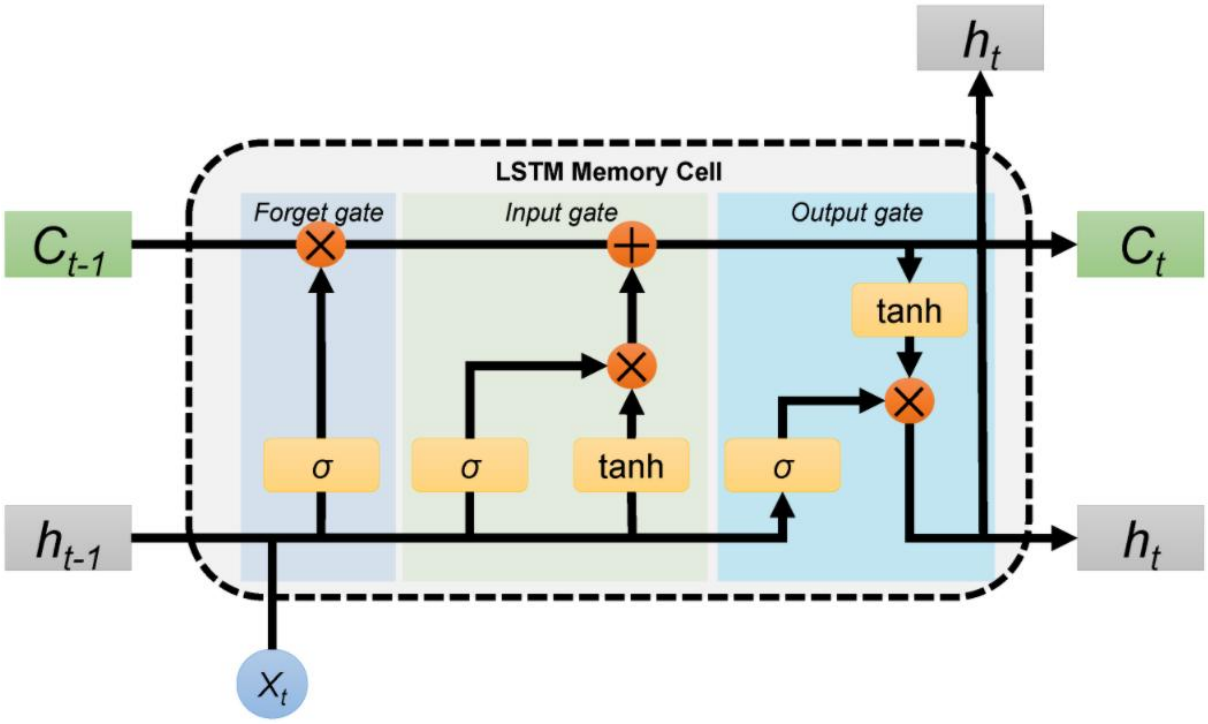

Figura 2.6: Estrutura de uma célula LSTM, destacando os portões de entrada, esquecimento e saída. Fonte: adaptado de (Analytics Vidhya, 2020).

Na Figura 2.6, o fluxo horizontal superior representa a propagação do estado da memória da célula ao longo do tempo, permitindo a preservação de dependências temporais relevantes. O termo C_{t-1} corresponde ao estado da memória proveniente do instante anterior, enquanto C_t representa o estado atualizado da célula no instante atual t.

A entrada da rede no instante atual é representada por x_t, enquanto h_{t-1} e h_t correspondem, respectivamente, aos estados ocultos anterior e atual da célula LSTM. Os portões de esquecimento, entrada e saída regulam o fluxo de informações durante o processamento, utilizando funções de ativação sigmoide (σ) para controlar quais informações devem ser preservadas, incorporadas ou propagadas.

Os blocos identificados por tanh são responsáveis pela transformação não linear das informações processadas, enquanto os círculos com os operadores de soma e multiplicação representam, respectivamente, operações de combinação e filtragem dos dados ao longo da atualização do estado interno da célula.

Nas Equações 2.3–2.8, h_{t-1} corresponde ao estado oculto do instante anterior, W e b representam, respectivamente, os pesos e vieses aprendidos durante o treinamento, enquanto σ denota a função de ativação sigmoide. A função σ produz valores no intervalo entre 0 e 1, permitindo controlar o fluxo de informações nos portões da arquitetura LSTM. Já a função tanh produz valores entre -1 e 1, sendo utilizada na atualização do estado interno da célula.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.4}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.5}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{2.6}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.7}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{2.8}$$

Na Equação 2.3, o portão de esquecimento f_t controla quais informações do estado anterior devem ser mantidas. Na Equação 2.4, o portão de entrada i_t define quais novas informações serão incorporadas ao estado da célula. A Equação 2.5 calcula o conteúdo candidato $\tilde{C}_t$ a ser adicionado à memória. Na Equação 2.6, o estado da célula C_t é atualizado combinando o estado anterior e o novo conteúdo. A Equação 2.7 define o portão de saída o_t, responsável por controlar a saída da célula. Por fim, a Equação 2.8 calcula a saída h_t da célula, baseada no estado atualizado.

Diferentemente das RNNs tradicionais, a LSTM permite que o gradiente seja propagado ao longo do estado da célula C_t de forma mais estável, reduzindo o problema do desaparecimento do gradiente e permitindo o aprendizado de dependências de longo

prazo (HOCHREITER; SCHMIDHUBER, 1997).

A estrutura de uma célula LSTM, ilustrada na Figura 2.6, evidencia como os diferentes portões controlam o fluxo de informação ao longo do tempo, permitindo que o modelo retenha ou descarte informações conforme necessário.

Devido a essa capacidade, as *LSTMs* são amplamente aplicadas em tarefas que envolvem séries temporais e linguagem natural, sendo eficazes na modelagem de dados sequenciais e dependências de longo prazo. No contexto deste trabalho, a arquitetura LSTM foi considerada devido à sua capacidade de modelar relações não lineares complexas entre variáveis clínicas associadas ao desfecho pós-transplante observado na base. Embora originalmente desenvolvida para dados sequenciais, a LSTM tem sido aplicada em diferentes problemas biomédicos e de predição clínica, especialmente em cenários com múltiplos atributos correlacionados e padrões complexos de interação entre variáveis (GOODFELLOW; BENGIO; COURVILLE, 2016).

Camadas LSTM recebem, em geral, entradas tridimensionais no formato (n, t, p), em que n representa o número de amostras, t representa o número de passos temporais e p representa o número de atributos observados em cada passo. Em dados sequenciais, t é maior que um e permite que a rede modele dependências ao longo do tempo.

É importante distinguir dois conceitos próximos. O comprimento da sequência, denotado por T, corresponde à segunda dimensão do tensor de entrada e determina quantos passos compõem cada amostra. Já o índice temporal t, empregado na representação dos estados da célula, identifica um passo específico da recorrência, como ocorre no estado C_{t-1} da Equação 2.6; esse índice equivale ao t utilizado nas Equações 2.3–2.8.

A dependência recorrente entre diferentes observações de uma mesma amostra só pode ser explorada quando $T > 1$, pois apenas nessa situação o estado C_{t-1} produzido em um passo participa do processamento do passo seguinte. No caso adotado neste trabalho, em que $T = 1$, não há transição temporal interna à amostra: a entrada da LSTM tem a forma $(n, 1, p)$, mas cada amostra contém apenas um vetor com p atributos. Dessa forma, ainda que a camada preserve sua estrutura de portas e estados internos, ela não chega a explorar dependências entre passos temporais. A Figura 2.7 ilustra a diferença entre essas duas configurações.

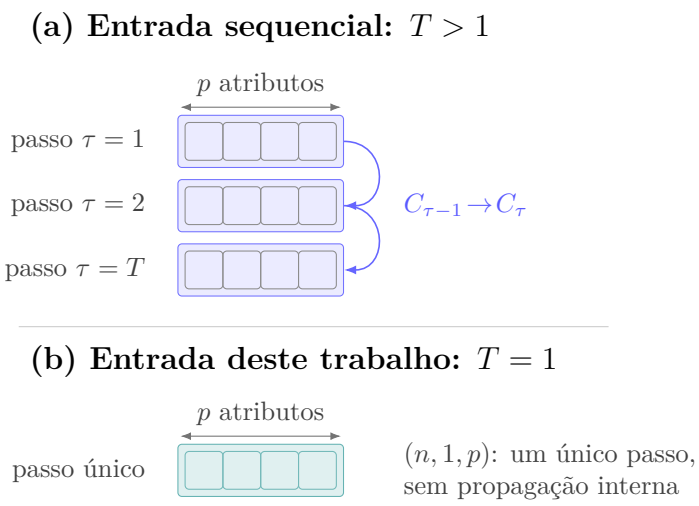

Figura 2.7: Comparação entre a entrada sequencial de uma LSTM, com $T > 1$ e propagação do estado C_{t-1} entre passos, e a entrada adotada neste trabalho, com $T = 1$, equivalente a uma observação tabular. Fonte: elaborado pelo autor.

Essa configuração, com (T=1), é retomada na Seção 4.2, ao se descrever a organização dos dados de entrada da LSTM.

2.3 Preparação de Dados para Modelos Preditivos

Em problemas de aprendizado de máquina aplicados a dados clínicos tabulares, a etapa de preparação dos dados é parte essencial do processo de modelagem. Bases dessa natureza frequentemente combinam variáveis numéricas, variáveis categóricas, valores ausentes e escalas de medida distintas. Essas características exigem transformações antes do treinamento dos modelos, de modo que os algoritmos possam processar os dados de forma adequada e sem introduzir inconsistências entre treinamento e avaliação (KUHN; JOHNSON, 2013; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Valores ausentes podem ocorrer por diferentes motivos, como indisponibilidade de exames, ausência de registro clínico ou diferenças nos protocolos de coleta. Uma estratégia simples e reprodutível para lidar com esse problema é a imputação, na qual valores ausentes são substituídos por estimativas calculadas a partir dos dados disponíveis. Em variáveis numéricas, a mediana é frequentemente utilizada por ser menos sensível a valores extremos do que a média. Em variáveis categóricas, uma alternativa simples consiste em substituir valores ausentes pela categoria mais frequente observada no conjunto de

treinamento (KUHN; JOHNSON, 2013).

Outra etapa comum é a padronização de variáveis numéricas. Essa transformação ajusta as variáveis para uma escala comum, geralmente subtraindo a média e dividindo pelo desvio padrão. Embora modelos baseados em árvores sejam menos sensíveis à escala das variáveis, algoritmos baseados em otimização numérica, como redes neurais, tendem a se beneficiar de entradas com escalas comparáveis (GOODFELLOW; BENGIO; COURVILLE, 2016).

2.3.1 Codificação de Variáveis Categóricas

Variáveis categóricas precisam ser convertidas para representações numéricas antes de serem utilizadas por algoritmos de aprendizado de máquina. Uma abordagem amplamente utilizada é o *one-hot encoding*, que transforma cada categoria de uma variável em uma coluna binária. Essa técnica evita impor uma ordem artificial entre categorias nominais, mas pode aumentar a dimensionalidade dos dados quando há muitas categorias.

Outra possibilidade é a codificação ordinal, na qual cada categoria é substituída por um valor inteiro. Por exemplo, uma variável com três categorias distintas pode ser representada por valores 0, 1 e 2. Essa abordagem preserva uma única coluna por variável categórica e produz uma matriz numérica mais compacta. Entretanto, quando a variável não possui ordem natural, essa codificação pode introduzir uma ordenação numérica artificial. Por isso, sua utilização deve ser interpretada como uma decisão prática de representação dos dados, e não como evidência de hierarquia real entre as categorias.

2.4 Validação Cruzada, Hiperparâmetros e Vazamento de Dados

A avaliação de modelos preditivos requer separação cuidadosa entre os dados utilizados para ajuste e os dados utilizados para avaliação. Uma prática comum é dividir a base em subconjuntos de treinamento e teste. O conjunto de treinamento é usado para ajustar os modelos e selecionar configurações, enquanto o conjunto de teste deve permanecer isolado até a avaliação final, fornecendo uma estimativa mais realista da capacidade de

generalização.

Além da divisão entre os conjuntos de treinamento e teste, a validação cruzada é frequentemente utilizada para estimar o desempenho durante a seleção de modelos e hiperparâmetros. Na validação cruzada em k partições, o conjunto de treinamento é dividido em k subconjuntos. A cada iteração, $k-1$ subconjuntos são utilizados para o ajuste do modelo, enquanto o subconjunto restante é empregado para validação. Esse processo é repetido até que todos os subconjuntos tenham sido utilizados uma vez como conjunto de validação (KOHAVI, 1995).

A Figura 2.8 sintetiza esse procedimento para $k = 5$ e destaca que o conjunto de teste não participa da seleção de modelos ou hiperparâmetros.

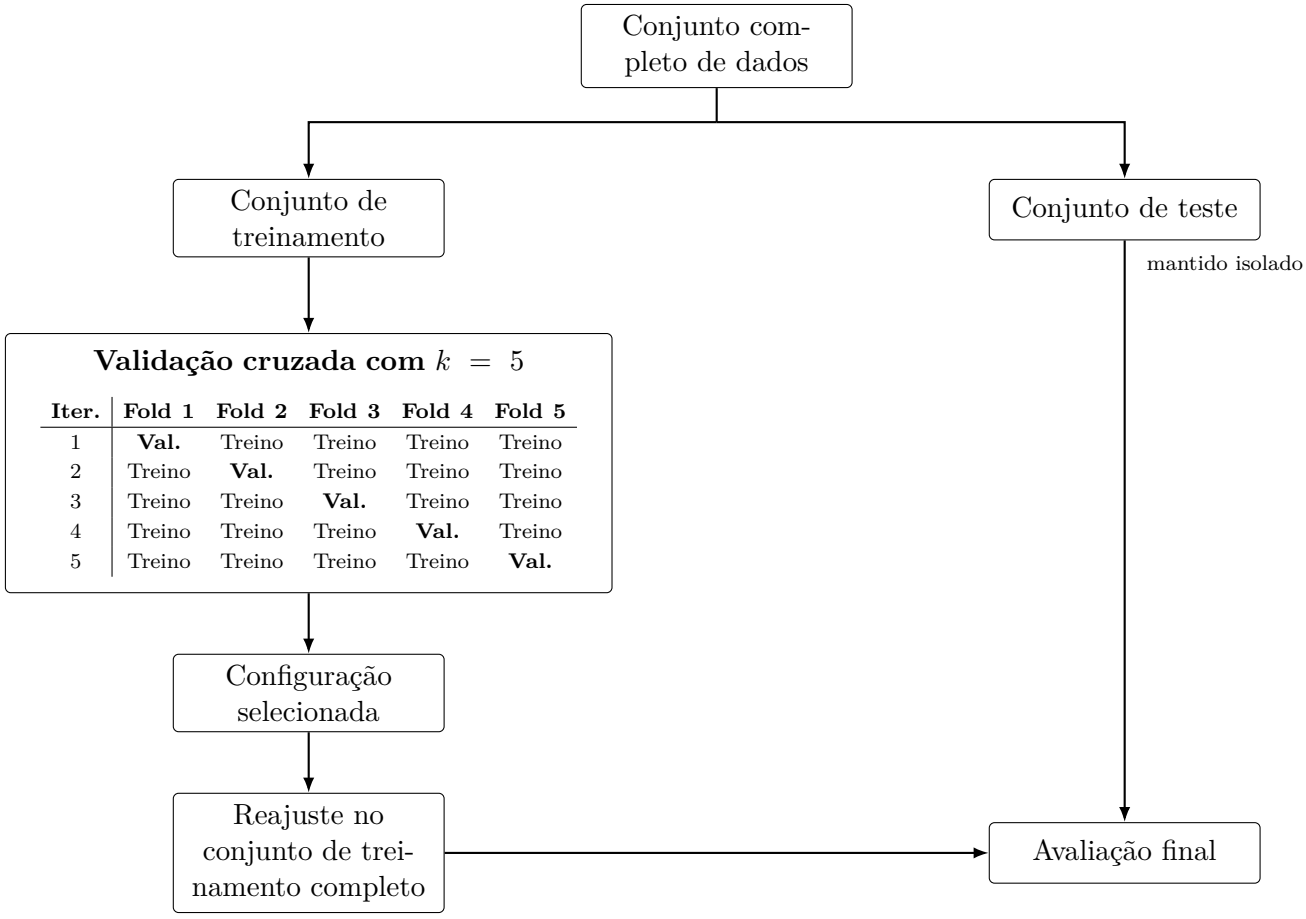

Figura 2.8: Esquema da divisão treino-teste e da validação cruzada em cinco partições. Fonte: elaborado pelo autor com base em (KOHAVI, 1995).

Um cuidado importante nesse processo é evitar o vazamento de dados. O vazamento ocorre quando informações do conjunto de teste influenciam direta ou indiretamente o treinamento, produzindo estimativas otimistas de desempenho. Por esse motivo, trans-

formações como imputação, padronização e codificação devem ser ajustadas apenas nos dados de treinamento e posteriormente aplicadas aos dados de validação ou teste com os parâmetros já aprendidos (KUHN; JOHNSON, 2019).

2.5 Métricas de Avaliação para Modelos de Sobrevivência

Avaliar modelos de sobrevivência exige métricas capazes de lidar com dados censurados. Em análise de sobrevivência, a censura ocorre quando o evento de interesse não é observado para determinados indivíduos durante o período de acompanhamento do estudo. Isso pode ocorrer, por exemplo, quando um paciente ainda está vivo ao final do período de observação, quando abandona o estudo antes da ocorrência do evento ou quando o acompanhamento é interrompido por outros motivos. Nesses casos, sabe-se apenas que o tempo até o evento é maior do que o tempo observado, mas não se conhece o momento exato de sua ocorrência. Diferentemente de tarefas tradicionais de classificação ou regressão, os modelos de sobrevivência lidam explicitamente com o tempo até a ocorrência de um evento (como morte, recaída ou falência de um órgão), o que exige métodos específicos de avaliação (PARK et al., 2021).

Para modelos de classificação binária, uma das métricas mais utilizadas é a Área Sob a Curva ROC (*Area Under the Receiver Operating Characteristic Curve* — AUC-ROC), que mede a capacidade discriminativa do modelo ao longo de todos os limiares de classificação possíveis. A curva ROC representa a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (1-especificidade), e a área sob essa curva fornece uma medida global do desempenho do classificador (FAWCETT, 2006).

Entretanto, métricas como a AUC-ROC não consideram explicitamente o tempo até a ocorrência do evento nem tratam adequadamente a presença de censura, características centrais dos dados de sobrevivência. Por esse motivo, utilizam-se métricas específicas para esse tipo de problema, entre as quais se destaca o Índice de Concordância (*Concordance Index* ou C-index), amplamente empregado na avaliação de modelos de sobrevivência.

2.5.1 Índice de Concordância (C-index)

O Índice de Concordância, também conhecido como *Concordance Index* ou C-index, é uma métrica amplamente utilizada para avaliar a capacidade discriminativa de modelos em problemas de análise de sobrevivência (JR et al., 1982). Diferentemente de métricas tradicionais de classificação, como acurácia ou AUC-ROC, o C-index considera a ordenação relativa dos indivíduos de acordo com o tempo até a ocorrência do evento, sendo, portanto, mais adequado para problemas em que o desfecho envolve tempo de acompanhamento e censura.

De forma intuitiva, o C-index mede a proporção de pares comparáveis de indivíduos para os quais o modelo atribui maior risco ao indivíduo que apresentou o evento primeiro. Assim, um modelo possui boa concordância quando consegue ordenar corretamente os pacientes segundo o risco estimado de ocorrência do evento.

Considere dois indivíduos i e j, com tempos observados T_i e T_j, indicadores de evento δ_i e δ_j, e escores de risco preditos r_i e r_j. Neste trabalho, r_i representa o escore de risco produzido pelo modelo, isto é, valores maiores indicam maior risco estimado de ocorrência do evento e, consequentemente, menor tempo esperado livre de evento.

Um par é considerado comparável quando é possível determinar qual indivíduo apresentou o evento primeiro. De modo simplificado, se $T_i < T_j$ e o evento foi observado para o indivíduo i, isto é, $\delta_i = 1$, então espera-se que o modelo atribua maior risco a i do que a j. Nesse caso, o par é concordante se $r_i > r_j$.

A formulação geral do C-index, considerando escores de risco, é expressa pela equação 2.9

$$C = \frac{\sum_i \sum_j I(T_i < T_j) \cdot I(r_i > r_j) \cdot \delta_i}{\sum_i \sum_j I(T_i < T_j) \cdot \delta_i}, \tag{2.9}$$

em que $I(\cdot)$ é a função indicadora, que assume valor 1 quando a condição é verdadeira e 0 caso contrário. O termo δ_i garante que apenas pares em que o evento foi observado para o indivíduo com menor tempo sejam considerados na comparação.

O valor do C-index varia entre 0 e 1. Valores próximos de 0,5 indicam desempenho semelhante ao acaso; valores próximos de 1 indicam elevada capacidade de ordenação

correta dos riscos; e valores próximos de 0 indicam ordenação inversa. No contexto deste trabalho, o C-index foi utilizado para avaliar se os modelos conseguem atribuir maior risco aos pacientes que apresentam menor tempo livre de evento.

2.5.2 Índice de Concordância Estratificado

Embora o C-index seja adequado para avaliar a capacidade de ordenação dos riscos em problemas de sobrevivência, sua versão global pode mascarar diferenças relevantes de desempenho entre subgrupos da população. Em contextos clínicos, essa limitação é especialmente importante, pois um modelo pode apresentar bom desempenho médio e, ainda assim, produzir previsões menos confiáveis para determinados grupos demográficos ou populacionais (HARTMAN et al., 2023).

Neste trabalho, essa comparação foi operacionalizada pelo cálculo do C-index separadamente para cada grupo racial. Esse procedimento segue a lógica da competição *Equity in Post-HCT Survival Predictions*, cuja métrica busca penalizar modelos com maior variação de desempenho entre os grupos analisados (CIBMTR, 2024).

Seja G o número de grupos raciais avaliados e seja C_g o C-index calculado para o grupo racial g. Inicialmente, calcula-se a média dos C-index dos grupos:

$$\bar{C} = \frac{1}{G} \sum_{g=1}^{G} C_g. \tag{2.10}$$

Em seguida, calcula-se o desvio padrão dos valores de C-index entre os grupos:

$$s_C = \sqrt{\frac{1}{G} \sum_{g=1}^{G} (C_g - \bar{C})^2}. \tag{2.11}$$

Por fim, o C-index estratificado é definido como:

$$C_{\text{estratificado}} = \bar{C} - s_C. \tag{2.12}$$

Essa formulação penaliza modelos que apresentam grande variação de desempenho entre grupos raciais. Portanto, um modelo só obtém valor elevado nessa métrica quando apresenta, simultaneamente, boa capacidade de ordenação dos riscos e desempenho relativamente estável entre os grupos avaliados.

Além do C-index estratificado, este trabalho também utiliza a amplitude racial como medida complementar de disparidade. A amplitude racial é definida como a diferença entre o maior e o menor C-index observado entre os grupos:

$$A_{\text{racial}} = \max(C_g) - \min(C_g). \tag{2.13}$$

Enquanto o C-index estratificado resume desempenho e estabilidade em um único valor, a amplitude racial permite observar diretamente a distância entre o grupo com melhor desempenho e o grupo com pior desempenho. Dessa forma, as duas medidas são complementares: valores maiores de $C_{\text{estratificado}}$ indicam melhor desempenho ajustado pela variação entre grupos, enquanto valores menores de A_{racial} indicam menor disparidade entre grupos raciais.

É importante destacar que neste trabalho, o termo equidade é empregado em sentido operacional e restrito, referindo-se à semelhança da capacidade discriminativa do modelo entre os grupos definidos por `race_group`. É de conhecimento que o C-index estratificado e a amplitude entre grupos não são suficientes para demonstrar equidade clínica, ausência de viés ou distribuição equitativa de benefícios e danos (PFOHL; FORY-CIARZ; SHAH, 2021). Logo, essas conclusões exigiriam métricas adicionais, avaliação de calibração, quantificação de incerteza e análise do contexto de utilização do modelo, não sendo no momento o foco principal deste trabalho.

3 Revisão Sistemática e Bibliográfica

Este capítulo apresenta a revisão sistemática e bibliográfica que fundamenta a pesquisa desenvolvida. O objetivo central é identificar, analisar e discutir os principais estudos relacionados à predição de sobrevida pós-transplante de células-tronco hematopoiéticas (TCTH), com ênfase no uso de algoritmos de aprendizado de máquina e na consideração da equidade nas previsões.

Para isso, o capítulo apresenta a estratégia utilizada para condução da revisão, incluindo a definição da pergunta norteadora, os critérios de busca, inclusão e exclusão e o processo de seleção dos estudos. Em seguida, são sintetizados e discutidos os trabalhos selecionados, com destaque para suas abordagens metodológicas, principais resultados e limitações relacionadas ao escopo desta pesquisa.

3.1 Método PICO

Para garantir rigor metodológico na formulação da questão de pesquisa e na seleção dos artigos, empregou-se a estratégia **PICO**, amplamente utilizada em revisões sistemáticas da área da saúde. O acrônimo **PICO** é formado por quatro componentes principais:

- **P (*Population*)**: define a população ou pacientes de interesse
- **I (*Intervention*)**: descreve a intervenção ou exposição a ser investigada
- **C (*Comparison*)**: indica a intervenção ou condição de comparação
- **O (*Outcome*)**: especifica o desfecho clínico de interesse

Segundo (RICHARDSON et al., 1995), a estruturação de perguntas de pesquisa por meio do modelo PICO é essencial para orientar a busca bibliográfica de forma eficiente, uma vez que possibilita transformar questões clínicas gerais em perguntas bem definidas e passíveis de investigação sistemática. Dessa forma, a utilização do PICO contribui para maior clareza e objetividade na revisão da literatura, assegurando a seleção de estudos que efetivamente respondam ao problema de pesquisa.

A partir dessa definição, foram estabelecidas: a pergunta norteadora, as palavras-chave, critérios de busca e bases de dados utilizadas (*Scopus*, *PubMed* e Google Acadêmico). Posteriormente, o processo de seleção foi organizado por meio do fluxograma *PRISMA*, que permitiu documentar as etapas de inclusão e exclusão dos artigos. Por fim, apresentam-se e discutem-se os estudos selecionados, destacando seus métodos, resultados e lacunas, de modo a justificar a proposta e relevância desta monografia.

3.2 Pergunta Norteadora

"Em pacientes submetidos a transplantes de células-tronco hematopoéticas (TCTH), métodos de aprendizado de máquina conseguem prever desfechos pós-transplante com maior precisão e apresentar desempenho consistente entre grupos populacionais, contribuindo para melhores estratégias de apoio à decisão clínica?"

3.3 Palavras-Chave

- *Machine Learning*
- Transplantes de células-tronco hematopoéticas (TCTH)
- Predição de sobrevivência
- Equidade em algoritmos de saúde

3.4 Regras de Busca

As buscas foram realizadas nas bases *Scopus*, Google Acadêmico e *PubMed*. A regra utilizada no Google Acadêmico e *PubMed* foi:

```
‘‘machine learning’’ OR ‘‘deep learning’’ OR ‘‘predictive modeling’’
AND (transplantation OR ‘‘transplant’’)
AND (hematopoietic)
AND (survival OR risk)
```

```
AND (retrospective OR ‘‘single-center’’ OR pediatric OR leukemia OR
CIBMTR)
AND ANO >= 2022 E ANO <= 2025
```

Enquanto a mesma regra formatada para a pesquisa avançada no *Scopus* foi:

```
(TITLE-ABS-KEY(‘‘machine learning’’ OR ‘‘deep learning’’ OR ‘‘predictive
modeling’’))
AND (TITLE-ABS-KEY(transplantation OR transplant))
AND (TITLE-ABS-KEY(hematopoietic))
AND (TITLE-ABS-KEY(survival OR risk))
AND (TITLE-ABS-KEY(retrospective OR ‘‘single-center’’ OR pediatric
OR leukemia OR CIBMTR))
AND (PUBYEAR > 2021 AND PUBYEAR < 2026)
```

3.5 Processo de Seleção dos Estudos

A seleção dos estudos foi conduzida seguindo as recomendações do protocolo PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), amplamente utilizado na condução e relato de revisões sistemáticas (PAGE et al., 2021).

O processo metodológico compreendeu as etapas de identificação dos estudos nas bases de dados, triagem inicial por título e resumo, avaliação de elegibilidade por leitura completa e inclusão final dos artigos selecionados para análise.

A definição dos estudos elegíveis foi realizada com base na estratégia PICO apresentada anteriormente, buscando assegurar coerência entre a pergunta de pesquisa, os objetivos do estudo e os trabalhos incluídos na revisão.

As subseções seguintes apresentam os critérios de inclusão e exclusão adotados, bem como o fluxograma PRISMA correspondente ao processo de seleção dos artigos.

3.5.1 Critérios de Inclusão e Exclusão

Os critérios de inclusão e exclusão foram definidos com base na estratégia PICO apresentada na Seção 3.1, estabelecendo os elementos centrais considerados relevantes para a

seleção dos estudos.

Os componentes do PICO considerados neste trabalho foram:

- **P (Population):** pacientes submetidos ao transplante de células-tronco hematopoéticas (TCTH);
- **I (Intervention):** aplicação de técnicas de aprendizado de máquina para predição de sobrevivência pós-transplante;
- **C (Comparison):** comparação entre diferentes modelos preditivos e abordagens de análise;
- **O (Outcome):** predição de sobrevivência, risco clínico ou desfechos relacionados ao período pós-transplante.

Com base nessa definição, foram adotados os seguintes critérios de inclusão:

- estudos envolvendo pacientes submetidos ao TCTH;
- trabalhos que aplicassem técnicas de aprendizado de máquina ou aprendizado profundo;
- estudos voltados à predição de sobrevivência, risco clínico ou desfechos pós-transplante;
- artigos publicados entre 2022 e 2025;
- estudos disponíveis integralmente em língua inglesa.

Os critérios de exclusão considerados foram:

- estudos com populações distintas daquelas definidas no PICO, como transplantes não hematopoéticos ou contextos clínicos não relacionados ao TCTH;
- estudos que utilizavam exclusivamente métodos estatísticos tradicionais, sem aplicação de técnicas de aprendizado de máquina;
- publicações que não correspondiam a estudos originais, como editoriais, comentários, revisões narrativas ou relatos;

- trabalhos fora do escopo principal da pesquisa, especialmente estudos sem foco em modelagem preditiva baseada em dados clínicos;
- artigos sem acesso ao texto completo.

3.5.2 Fluxograma PRISMA

A Figura 3.1 apresenta o fluxograma do processo de seleção dos estudos incluídos nesta revisão sistemática, seguindo as recomendações do protocolo PRISMA (PAGE et al., 2021).

Inicialmente, foram identificados 366 registros distribuídos entre as bases Scopus, PubMed e Google Acadêmico. Após a remoção de 147 registros duplicados, permaneceram 219 estudos para a etapa de triagem por título e resumo.

Na etapa de elegibilidade, 38 artigos foram selecionados para leitura completa, dos quais 4 não estavam disponíveis integralmente. Assim, 34 estudos foram avaliados quanto à elegibilidade metodológica.

A aplicação dos critérios de inclusão e exclusão resultou na exclusão de 27 publicações. Dentre essas, 4 apresentavam populações distintas das definidas no PICO, 13 utilizavam exclusivamente métodos estatísticos tradicionais sem aplicação de técnicas de aprendizado de máquina, 3 correspondiam a editoriais, comentários ou relatos de caso e 7 encontravam-se fora do escopo principal da pesquisa.

Ao final do processo, 7 estudos foram incluídos na revisão sistemática e utilizados como base para a discussão teórica e metodológica deste trabalho.

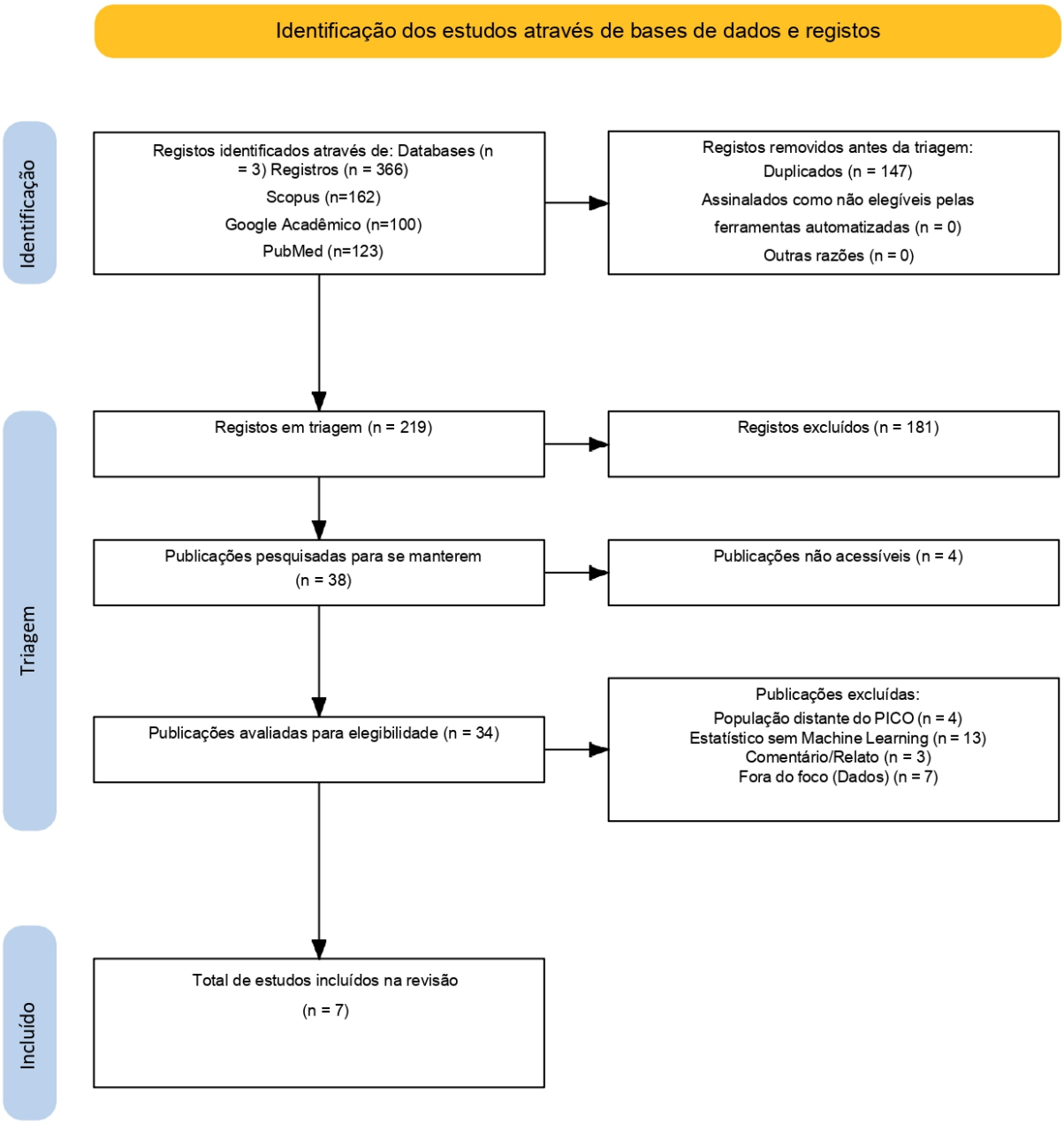

Figura 3.1: Ilustração do diagrama PRISMA da revisão

3.6 Artigos Selecionados

Os artigos finais aceitos na revisão sistemática foram:

- (RAJKOMAR et al., 2020): Discussão sobre princípios para garantir justiça em modelos de aprendizado de máquina na saúde, com ênfase em auditoria de desempenho por subgrupos, definição explícita de equidade e estratégias para reduzir disparida-

des algorítmicas.

- (RATUL et al., 2022): Análise retrospectiva em grupo pediátrico comparando classificadores (*SVM*, *XGBoost*, *Random Forest*) e otimização de hiperparâmetros via teste qui-quadrado.

- (CHADAGA et al., 2023): *Framework* de *Explainable AI* para prever a eficácia de TCTH em pacientes pediátricos, com geração de mapas de saliência e regras de decisão.

- (CHO et al., 2023): Uso da calculadora de sobrevida de um ano do *CIBMTR* como referência para validação retrospectiva de modelos preditivos clínicos.

- (LI; JIANG; ZHANG, 2024): Abordagem baseada em *transformers* para prever fatores de risco pós-transplante hepático, incorporando estratégias de justiça algorítmica.

- (SHOURABIZADEH et al., 2024): Estudo aplicando diversos algoritmos de aprendizado de máquina para prever a sobrevida pós-transplante alogênico de células-tronco hematopoéticas.

- (ROUZBAHANI et al., 2025): Modelagem preditiva baseada em *ensemble* (*Random Forest* e *boosting*) para resultados em pacientes com leucemia aguda submetidos a TCTH.

3.6.1 Discussão dos artigos

A predição de sobrevida pós-TCTH tem sido aprimorada por meio de técnicas de aprendizado de máquina que exploram tanto modelos clássicos quanto arquiteturas de *deep learning*. Nesse sentido, Rajkomar et al. discutem de forma abrangente os princípios necessários para garantir equidade em sistemas de aprendizado de máquina aplicados à saúde, enfatizando práticas como a definição clara de justiça, a auditoria de desempenho em subgrupos sensíveis e a integração de perspectivas interdisciplinares (RAJKOMAR et al., 2020). A análise destaca que, sem uma estrutura ética robusta, mesmo modelos tecnicamente avançados podem perpetuar ou ampliar desigualdades já existentes.

Ratul et al. trabalharam analisando um grupo pediátrico, enfatizando a importância da otimização de hiperparâmetros por meio de testes qui-quadrado e validação cruzada para elevar a robustez preditiva dos modelos tradicionais (*SVM*, *XGBoost*, etc.) (RATUL et al., 2022). A calibração e validação retrospectiva, por sua vez, foram discutidas por Cho et al., que utilizaram a calculadora de sobrevida em um ano do *Center for International Blood and Marrow Transplant Research* (CIBMTR), uma ferramenta baseada em dados de registro em larga escala que fornece estimativas individualizadas da probabilidade de sobrevida após transplante hematopoético. Nesse estudo, os autores compararam as probabilidades preditas pela calculadora com as estimativas observadas por meio do método de Kaplan-Meier, avaliando a concordância entre os resultados em um cenário clínico real (CHO et al., 2023). Outra importante abordagem é a transparência dos modelos e a capacidade de interpretação, que são cruciais para a adoção clínica. Chadaga et al. propuseram um *framework* de *explainable AI* para transplante pediátrico, que fornece mapas de saliência e regras de decisão extraídas de árvores de decisão para explicar individualmente cada previsão, facilitando a confiança do médico na ferramenta (CHADAGA et al., 2023).

Em um estudo observacional, Shourabizadeh et al. utilizaram diversos algoritmos (incluindo regressão logística, árvore de decisão e redes neurais) para avaliar a acurácia na previsão de sobrevida em pacientes submetidos a transplante alogênico, identificando variáveis clínicas de maior peso na decisão prognóstica (SHOURABIZADEH et al., 2024). A adoção de arquiteturas de *deep learning* vem incorporando preocupações com justiça algorítmica e mitigação de viés. Li et al. adaptaram *transformers* para prever fatores de risco em transplantes hepáticos, incluindo estratégias de balanceamento de subgrupos para reduzir disparidades raciais e de gênero nas estimativas de risco (LI; JIANG; ZHANG, 2024). Rouzbahani et al. complementaram esse enfoque ao aplicar métodos de *ensemble*, como *Random Forest* e métodos de *boosting*, em pacientes com leucemia aguda, demonstrando que combinações de modelos podem superar classificadores individuais em métricas de performance (ROUZBAHANI et al., 2025).

Em síntese, a literatura converge para a ideia de que não existe um modelo universal, a integração de múltiplas abordagens, variando desde *ensembles* de árvores e

validações robustas até arquiteturas de *deep learning* justas e explicáveis, representam o futuro da predição de sobrevida pós-TCTH. A reflexão crítica sobre vieses, transparência e aplicabilidade clínica, como evidenciado nos trabalhos citados, fornece um alicerce sólido para futuras pesquisas e para a construção de sistemas de apoio à decisão que sejam, simultaneamente, precisos e éticos.

A partir da análise dos trabalhos relacionados, observa-se que, embora existam avanços significativos na aplicação de técnicas de aprendizado de máquina para predição de sobrevivência no contexto do TCTH, ainda há lacunas relevantes na literatura. Em geral, os estudos concentram-se na otimização do desempenho preditivo global, utilizando métricas agregadas como AUC-ROC ou acurácia, sem considerar de forma sistemática possíveis disparidades entre grupos demográficos.

Além disso, apesar do uso recorrente de técnicas de ensemble e modelos avançados, não foram identificados trabalhos que realizem uma análise comparativa abrangente entre diferentes famílias de modelos, como modelos baseados em árvores, redes neurais e métodos de empilhamento, sob a perspectiva simultânea de desempenho e equidade.

Dessa forma, este trabalho diferencia-se ao propor uma análise integrada que considera não apenas a capacidade preditiva dos modelos, mas também a consistência da capacidade discriminativa entre diferentes subgrupos populacionais, utilizando métricas específicas como o *C-index* estratificado. Ao investigar conjuntamente diferentes abordagens de modelagem e etapas de pré-processamento, busca-se contribuir para o desenvolvimento de soluções mais robustas e justas no contexto da predição de sobrevivência pós-TCTH.

3.6.2 Nuvem de Palavras

A nuvem de palavras apresentada na Figura 3.2 sintetiza os principais conceitos e temas recorrentes nos trabalhos analisados neste estudo. Observa-se a centralidade de termos como *fairness*, *data*, *model* e *survival*, evidenciando o foco dos artigos na construção de modelos preditivos baseados em dados clínicos, com ênfase na predição de sobrevida em contextos de saúde. A presença destacada de palavras como *patients*, *transplant* e *health* reforça o caráter aplicado dessas pesquisas, especialmente no domínio do transplante e do

cuidado médico orientado por dados.

Além disso, termos como *bias*, *learning* e *machine* indicam uma preocupação crescente com os vieses introduzidos por métodos de aprendizado de máquina e pela qualidade dos dados utilizados. A recorrência de conceitos associados à justiça algorítmica (*fairness*) sugere um movimento da literatura no sentido de não apenas maximizar desempenho preditivo, mas também garantir equidade nas decisões automatizadas, particularmente em cenários clínicos sensíveis. Dessa forma, a nuvem de palavras evidencia a convergência entre modelagem preditiva, ética, e justiça algorítmica, destacando os desafios e direções atuais da pesquisa na área.

Figura 3.2: Ilustração da nuvem de palavras dos artigos selecionados

4 Material e Métodos

Este capítulo descreve o procedimento experimental adotado para preparação dos dados, treinamento dos modelos e avaliação das predições. A descrição foi organizada com foco em reprodutibilidade: cada etapa corresponde ao fluxo implementado no pipeline computacional final, executado sobre os dados da competição *Equity in Post-HCT Survival Predictions*, do CIBMTR (CIBMTR, 2024).

O experimento foi formulado como uma tarefa de classificação binária, ou seja, os modelos foram treinados para predizer um escore de risco associado à ocorrência do evento. Esse escore foi posteriormente comparado com os tempos de fato observados.

A formulação adotada nesse trabalho constitui uma classificação do indicador observado `efs`, e não uma modelagem direta do tempo até o evento. A variável `efs_time` e a informação de censura não participaram da função de treinamento dos classificadores, sendo utilizadas apenas na avaliação posterior do C-index. Assim, é importante observar que as saídas dos modelos devem ser interpretadas como escores associados à classe `efs=1`, e não como estimativas calibradas da função de sobrevivência ou da probabilidade de evento em um horizonte temporal específico.

4.1 Conjunto de Dados

O conjunto de dados utilizado neste trabalho foi disponibilizado na plataforma *Kaggle* para a competição *Equity in Post-HCT Survival Predictions*, promovida pelo CIBMTR (CIBMTR, 2024). A base reúne informações demográficas, clínicas, imunogenéticas e procedimentais relacionadas ao transplante de células-tronco hematopoéticas (TCTH). Os dados disponibilizados pela competição são sintéticos, gerados a partir de padrões presentes em dados reais do CIBMTR, com o objetivo de preservar características relevantes do problema sem expor informações de pacientes.

A base de dados utilizada nos experimentos reúne 28 800 registros e 60 características. A característica `efs` indica se o evento foi observado durante o acompanha-

mento ou se o registro foi censurado, enquanto `efs_time` registra o tempo até o evento ou até o encerramento do acompanhamento.

Na base de dados, `efs=1` indica que o evento foi observado, enquanto `efs=0` representa uma observação censurada. Já `efs_time` informa o tempo, em meses, até a ocorrência do evento ou até o encerramento do acompanhamento. A documentação da competição caracteriza `efs` como um indicador de sobrevivência livre de evento, mas não detalha quais ocorrências clínicas compõem esse desfecho. Por esse motivo, o evento é interpretado conforme a codificação fornecida pela base, sem ser atribuído exclusivamente ao óbito, à recaída ou a outro resultado clínico específico.

O identificador `ID` foi utilizado exclusivamente para a identificação dos registros e, por esse motivo, não foi incluído entre os preditores do modelo. As características `efs` e `efs_time` também foram removidas das entradas do modelo por constituírem o desfecho descrito anteriormente, sendo `efs` utilizada como *target*. Após essas remoções, permaneceram 57 características originais candidatas a preditores.

Durante a preparação dos dados, foi acrescentada uma característica derivada, descrita na Seção 4.2, elevando o total para 58 características antes da codificação das categóricas.

A distribuição de `efs` é apresentada na Tabela 4.1. Foram identificados 15 532 registros com evento observado e 13 268 registros censurados. As duas situações apresentam proporções próximas, embora os eventos observados correspondam a pouco mais da metade da base, indicando um desbalanceamento aceitável.

Situação	**Registros**	**Percentual**
Censura (`efs=0`)	13 268	46,07%
Evento observado (`efs=1`)	15 532	53,93%

Tabela 4.1: Distribuição do indicador `efs`.

Exclusivamente para facilitar a explicação da base, as características preditoras foram organizadas em grupos, de acordo com o tipo de informação que representam:

Compatibilidade imunogenética: reúne as características relacionadas ao sistema HLA

(*Human Leukocyte Antigen*) [5], responsável pela compatibilidade imunológica entre doador e receptor. Estão incluídas medidas de correspondência nos loci A, B, C, DQB1 e DRB1,[6] calculadas em diferentes níveis de resolução,[7] além de contagens agregadas para painéis de 6, 8 e 10 marcadores.[8]

Características clínicas do receptor: inclui informações sobre a condição de saúde do paciente no período do transplante, como os escores de comorbidade e de desempenho funcional, representados por `comorbidity_score`, que sintetiza a carga de comorbidades clínicas do paciente, e `karnofsky_score`, que expressa seu grau de capacidade funcional, com valores mais elevados indicando maior independência (ver Apêndice A). Esse grupo também contém indicadores de diabetes, arritmia, obesidade e comprometimentos pulmonares, hepáticos e renais, além de informações sobre a doença que motivou o transplante.

Características do procedimento: compreende informações sobre a realização do transplante e os tratamentos associados. Entre elas estão o tipo de enxerto, a intensidade do regime de condicionamento, o uso de irradiação corporal total, a profilaxia da doença do enxerto contra o hospedeiro e o uso de medicamentos durante o procedimento.

Características demográficas e do par doador-receptor: reúne características como a idade do receptor, a idade do doador, o grupo racial, a etnia e a compatibilidade de sexo entre doador e receptor. Também pertencem a esse grupo as informações que descrevem o vínculo entre o doador e o paciente.

[5]Sistema formado por genes e proteínas presentes na superfície das células, responsável por auxiliar o sistema imunológico no reconhecimento do que pertence ou não ao organismo. No TCTH, a compatibilidade HLA entre doador e receptor é um dos principais critérios para a seleção do doador.

[6]Os loci correspondem a posições específicas do material genético nas quais se encontram genes do sistema HLA. As designações A, B, C, DQB1 e DRB1 identificam diferentes genes considerados na avaliação da compatibilidade entre doador e receptor. Ver Apêndice A (GRAGERT et al., 2014).

[7]O nível de resolução indica o grau de detalhamento da tipagem HLA. A baixa resolução identifica grupos mais amplos de variantes, enquanto a alta resolução permite distinguir alelos específicos com maior precisão.

[8]Essas contagens representam o número de correspondências entre os alelos do doador e do receptor. Como cada indivíduo possui dois alelos em cada locus, os painéis de 6, 8 e 10 marcadores correspondem, respectivamente, à avaliação de três, quatro e cinco loci HLA. Em geral, são considerados HLA-A, HLA-B e HLA-DRB1 no painel de 6; acrescenta-se HLA-C no painel de 8 e HLA-DQB1 no painel de 10. Ver Apêndice A.

Essa organização possui finalidade descritiva e não altera a forma como as características são fornecidas aos modelos. O dicionário completo das características, com seus nomes, tipos e descrições individuais, é apresentado no Apêndice A.

Como parte da caracterização inicial da base, calcularam-se as correlações entre os preditores numéricos, além da correlação entre cada característica numérica e o indicador binário `efs`. A Figura 4.1 apresenta o mapa de calor obtido. O aspecto mais evidente é a alta correlação entre diversas características relacionadas à compatibilidade HLA, o que ocorre porque algumas delas medem os mesmos loci em diferentes níveis de resolução. As associações lineares entre os preditores numéricos individuais e `efs` apresentam, em geral, baixa magnitude, indicando que nenhuma das características numéricas avaliadas possui, isoladamente, associação linear forte com o indicador de evento. Esse resultado, contudo, não exclui a existência de associações não lineares ou de interações entre diferentes características.

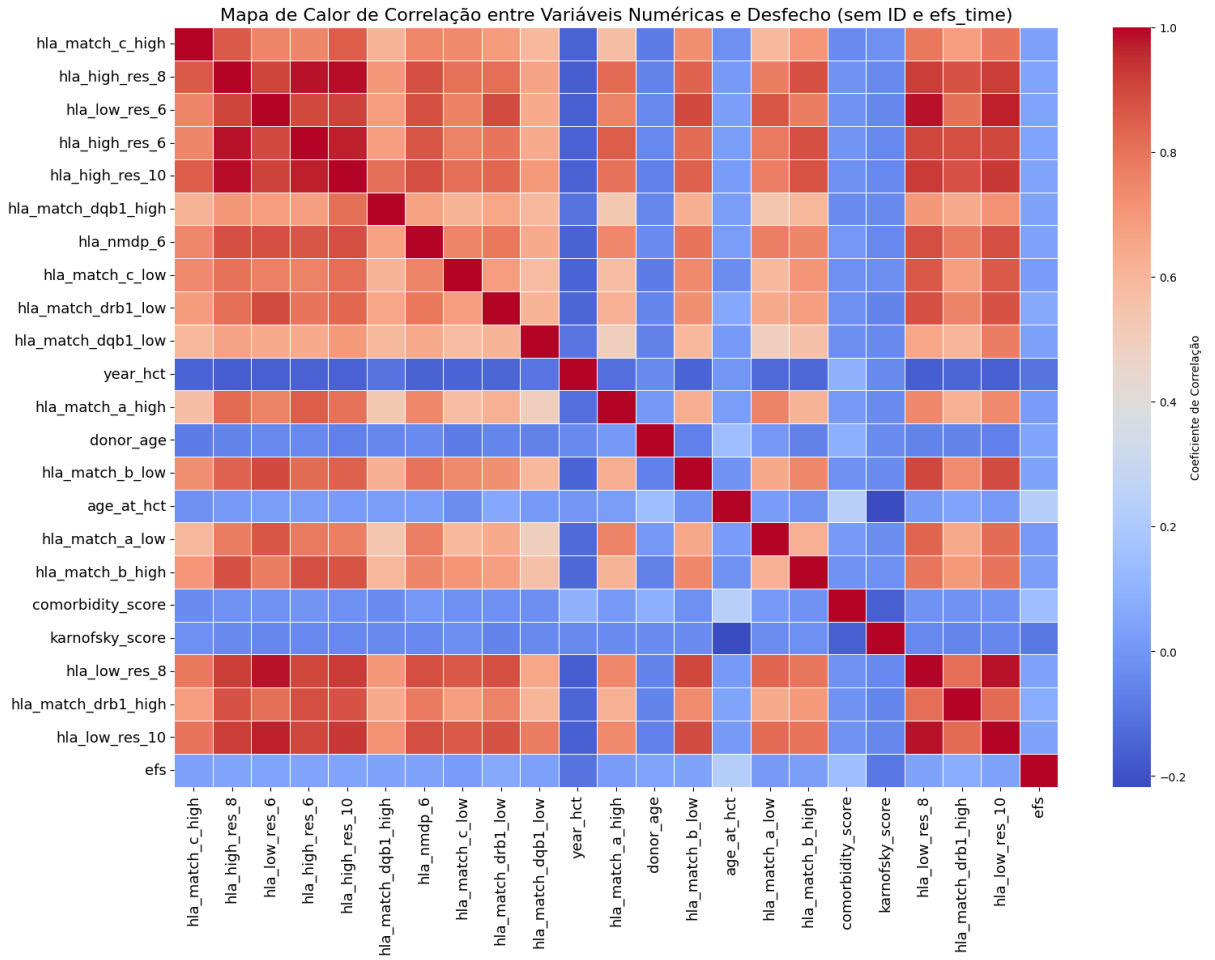

Figura 4.1: Mapa de calor das correlações entre os preditores numéricos e o indicador `efs`. Valores próximos de 1 indicam associação linear positiva, enquanto valores próximos de -1 indicam associação linear negativa.

A característica `race_group`, que registra o grupo racial atribuído ao paciente na base de dados (ver Apêndice A), foi mantida como preditor e também utilizada na avaliação estratificada dos modelos, na qual o desempenho foi calculado separadamente para cada grupo racial. Essa escolha acompanha o protocolo da competição, cuja métrica oficial considera o desempenho obtido entre esses grupos (CIBMTR, 2024).

Neste trabalho, `race_group` é interpretada como uma característica demográfica de categorização racial, não devendo ser considerada uma medida direta de ancestralidade genética. Embora categorias raciais e medidas de ancestralidade possam apresentar associações estatísticas com a distribuição de haplótipos HLA, elas não são equivalentes nem intercambiáveis (HOLLENBACH et al., 2015; DAMOTTE et al., 2024). No contexto do

transplante, diferenças observadas entre grupos raciais podem refletir tanto a distribuição populacional dos haplótipos HLA e a disponibilidade desigual de doadores compatíveis (GRAGERT et al., 2014) quanto barreiras sociais e institucionais relacionadas ao encaminhamento, ao acesso e à utilização do transplante (LANDRY, 2021).

Por isso, sua utilização como preditor e a interpretação dos resultados por grupo devem ser feitas com cautela. A distribuição dos registros segundo `race_group` é apresentada na Tabela 4.2. Os seis grupos possuem quantidades semelhantes de observações, aspecto relacionado ao processo de construção do conjunto sintético da competição.

Tabela 4.2: Distribuição dos registros por grupo racial.

Grupo racial	**Registros**
More than one race	4 845
Asian	4 832
White	4 831
Black or African-American	4 795
American Indian or Alaska Native	4 790
Native Hawaiian or other Pacific Islander	4 707

Foram identificados valores ausentes em 50 das 60 características, num total de 189 575 registros sem preenchimento, concentrados sobretudo em características de compatibilidade imunogenética e de caracterização da doença. Dentro desse conjunto, as características de tipagem e compatibilidade HLA[9] formam o maior grupo de características afetadas, ainda que não correspondam aos maiores percentuais individuais. Esse padrão é coerente com a forma como a tipagem HLA é registrada, uma vez que nem todos os pares doador-receptor são tipados em todos os loci nem no mesmo nível de resolução, o que torna as medidas de alta resolução e os painéis mais completos mais suscetíveis à ausência. Esse diagnóstico foi realizado sobre a base completa, antes da divisão em

[9]O sistema HLA (*Human Leukocyte Antigen*) é responsável pela compatibilidade imunológica entre doador e receptor e constitui um dos principais critérios na escolha do doador no TCTH. A compatibilidade é verificada em posições específicas do material genético, chamadas *loci* (A, B, C, DQB1 e DRB1), e pode ser avaliada em diferentes níveis de resolução: a baixa resolução distingue grupos mais amplos de variantes, enquanto a alta resolução identifica variantes mais específicas. Os painéis de 6, 8 e 10 marcadores resumem essas comparações em contagens agregadas, considerando um número crescente de loci (GRAGERT et al., 2014).

treinamento e teste descrita na Seção 4.3. A Figura 4.2 apresenta as trinta características com maior percentual de ausência nesse levantamento inicial, cujos percentuais decrescem de forma acentuada a partir das primeiras posições. Os registros não foram removidos em razão dos valores ausentes. Em vez disso, o tratamento dessas ocorrências foi incorporado ao pré-processamento por meio de técnicas de imputação, conforme descrito na Seção 4.2. Os parâmetros da imputação foram estimados apenas com os dados de treinamento, já isolados do teste, evitando que informação do conjunto de teste influenciasse essa etapa.

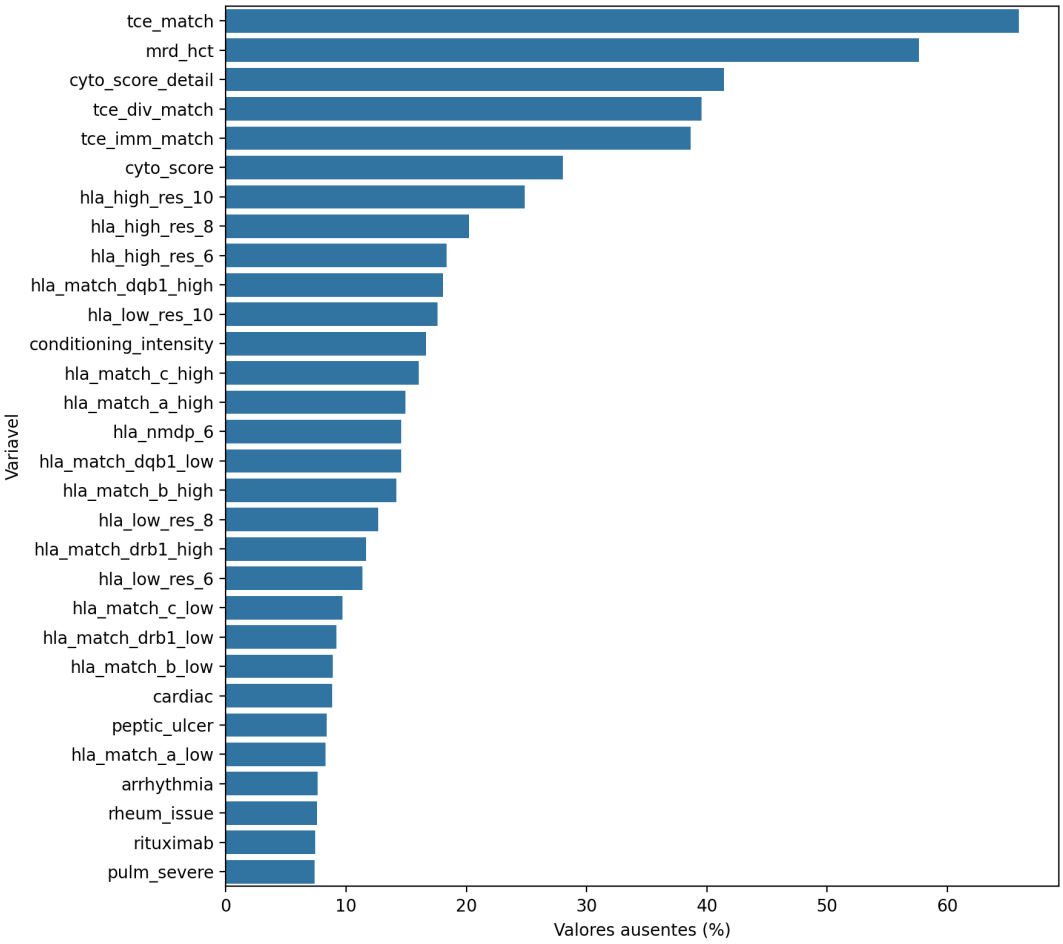

Figura 4.2: Trinta características com maior percentual de valores ausentes na base completa, antes da divisão em treinamento e teste.

Após a caracterização inicial, a base foi separada em conjuntos de treinamento e teste. O ajuste dos modelos foi realizado utilizando o conjunto de treinamento, com validação cruzada para a seleção dos hiperparâmetros. O conjunto de teste permaneceu separado até a avaliação final. O procedimento de divisão e os critérios de estratificação são apresentados na Seção 4.3.

É importante destacar que o balanceamento aproximado dos grupos raciais decorre da construção sintética da base e não deve ser interpretado como representação da frequência desses grupos na população submetida ao TCTH. Da mesma forma, diferenças de desempenho observadas nesse conjunto não constituem evidência de desigualdades clínicas reais.

4.2 Preparação e Transformação dos Dados

Antes do treinamento dos modelos, os dados foram submetidos a etapas de preparação e transformação, com o objetivo de adequar as características ao formato de entrada exigido pelos algoritmos e manter a consistência do fluxo experimental.

Como etapa de construção de características, foi criada a seguinte característica: `donor_recipient_age_diff`, definida como a diferença entre `donor_age`, que representa a idade do doador, e `age_at_hct`, que corresponde à idade do receptor no momento do transplante (ver Apêndice A):

$$\texttt{donor\_recipient\_age\_diff} = \texttt{donor\_age} - \texttt{age\_at\_hct}.$$

A diferença permite distinguir situações em que o doador é mais velho do que o receptor, representadas por valores positivos, daquelas em que o doador é mais jovem, representadas por valores negativos. Embora seja matematicamente derivada de duas características já presentes na base, sua inclusão explícita pode facilitar a captura da relação etária pelos modelos baseados em árvores, que de outra forma dependeriam de divisões sucessivas para aproximar a mesma informação. Por se tratar de uma operação aplicada individualmente a cada registro, sem depender de nenhuma estatística agregada da base, sua criação não introduz vazamento de informação entre treino e teste.

A motivação clínica para considerar relações envolvendo a idade decorre de estudos que identificaram associação entre a idade do doador e os resultados do transplante. Arai et al. observaram pior prognóstico entre receptores com anemia aplástica submetidos a transplante alogênico não aparentado quando utilizados doadores mais velhos (ARAI et al., 2016). No entanto, essa evidência não estabelece que a diferença etária seja, isola-

damente, um fator prognóstico para a população analisada neste trabalho. Sua inclusão possui caráter exploratório e busca verificar se a relação relativa entre as idades contém informação preditiva complementar às idades absolutas do doador e do receptor.

O pré-processamento foi implementado dentro de *pipelines*, de modo que os imputadores, codificadores e escalonadores — responsáveis, respectivamente, por preencher valores ausentes, converter características categóricas em valores numéricos e ajustar características numéricas a uma escala comum — fossem ajustados apenas nos dados de treinamento de cada etapa. Essa estratégia evita vazamento de informação, pois estatísticas como a mediana, a categoria mais frequente e os parâmetros de padronização não são calculadas com base no subconjunto de teste.

Todos os modelos avaliados neste trabalho — *Random Forest*, *XGBoost*, *LightGBM*, *Stacking Ensemble* e LSTM — compartilharam a mesma definição de pré-processamento, composta pelas seguintes etapas:

- características numéricas: imputação pela mediana do conjunto de treinamento e padronização por `StandardScaler`;
- características categóricas: imputação pela categoria mais frequente do conjunto de treinamento e codificação por *one-hot encoding*. Categorias presentes nos conjuntos de validação ou teste, mas não observadas durante o treinamento, foram processadas sem a criação de novas colunas, mantendo-se a mesma estrutura de entrada dos modelos.

A mediana foi utilizada na imputação das características numéricas por ser menos sensível a valores extremos do que a média, aspecto relevante em dados clínicos heterogêneos. A característica derivada `donor_recipient_age_diff`, por ser numérica, segue o mesmo procedimento de imputação e padronização aplicado às demais desse tipo.

Adotar um pipeline de pré-processamento idêntico para todos os modelos teve como objetivo isolar o efeito da arquitetura de cada algoritmo: assim, diferenças de desempenho entre os modelos podem ser atribuídas às suas características de modelagem, e não a etapas distintas de preparação dos dados. Por esse mesmo motivo, optou-se por manter `StandardScaler` também para a LSTM, em vez de uma normalização min-max no intervalo $[0, 1]$ – prática também comum em redes neurais. Como a padronização já é

adequada à arquitetura LSTM e permite comparação direta com os modelos baseados em árvores sob a mesma escala de entrada, sua adoção uniforme reforça a comparabilidade dos resultados reportados no Capítulo 5.

A aplicação do *one-hot encoding* às 35 características categóricas expandiu a base de 58 para 184 características no conjunto de treinamento. As 23 características numéricas foram preservadas em sua forma original, enquanto as 35 categóricas deram origem a 161 colunas binárias. Essa base de 184 características constitui a entrada efetivamente recebida por todos os modelos.

4.3 Divisão dos Dados e Validação

A divisão dos dados foi realizada após a definição das características de entrada e antes do ajuste dos transformadores de pré-processamento. Foram reservados 70% dos registros para treinamento e 30% para teste final. A divisão foi estratificada, ou seja, subagrupada, pela combinação de `efs` e `race_group`: em vez de sortear os registros de forma totalmente aleatória, a base foi primeiro dividida em subgrupos definidos por cada combinação de valores dessas duas informações, por exemplo, registros com `efs=1` e `race_group` igual a *Asian* formam um desses subgrupos, e a proporção 70/30 foi aplicada dentro de cada subgrupo separadamente. Cada um desses subgrupos é chamado, ao longo deste capítulo, de estrato. Esse procedimento preserva simultaneamente a distribuição do desfecho e dos grupos raciais nos conjuntos de treinamento e teste, o que não seria garantido por uma divisão puramente aleatória. A escolha se justifica porque essas duas informações são centrais para o objetivo do estudo: `efs` define o desfecho modelado, enquanto `race_group` define os grupos usados na avaliação de equidade.

Durante a preparação do experimento, foram avaliadas combinações adicionais de estratificação, incluindo `ethnicity` e `sex_match`, para verificar se outras características demográficas também poderiam ser preservadas na divisão sem fragmentar excessivamente a base. A Tabela 4.3 apresenta o diagnóstico dessas combinações. A estratificação por `efs + race_group` resultou em 12 estratos, com o menor deles contendo 1 809 registros, sem ocorrência de estratos pequenos. A inclusão simultânea de `ethnicity` e `sex_match`, por sua vez, elevou o número de estratos para 226 e produziu estratos com apenas um

registro, o que seria incompatível com a validação cruzada em cinco partições, etapa descrita adiante nesta seção. Por esse motivo, características como `ethnicity` e `sex_match` permaneceram disponíveis como preditores, mas não foram usadas como critérios de estratificação. Não se assume, portanto, preservação controlada de suas proporções além daquela decorrente da divisão aleatória estratificada por `efs` e `race_group`.

Estratificação	**Nº de estratos**	**Menor estrato**	**Estratos < 20**	**Estratos < 50**
`efs`	2	13 268	0	0
`efs + race_group`	12	1 809	0	0
`efs + ethnicity`	8	161	0	0
`efs + race_group + ethnicity`	48	16	2	18
`efs + race_group + sex_match`	60	7	6	12
`efs + race_group + ethnicity + sex_match`	226	1	123	138

Tabela 4.3: Diagnóstico das combinações avaliadas para estratificação da divisão treino-teste.

As Figuras 4.3 e 4.4 traduzem em escala visual os mesmos números já reunidos na Tabela 4.3, facilitando a comparação entre as seis combinações avaliadas. A Figura 4.3 evidencia a redução do tamanho do menor estrato à medida que novas características são combinadas à estratificação, tornando visível o ponto a partir do qual a fragmentação compromete a validação cruzada em cinco partições. A Figura 4.4 mostra o efeito complementar: o crescimento do número total de estratos conforme mais características são incluídas. Em conjunto, esses resultados justificam a escolha por `efs + race_group`: essa combinação preserva o desfecho e o grupo racial, que são centrais para a análise, sem fragmentar excessivamente a amostra.

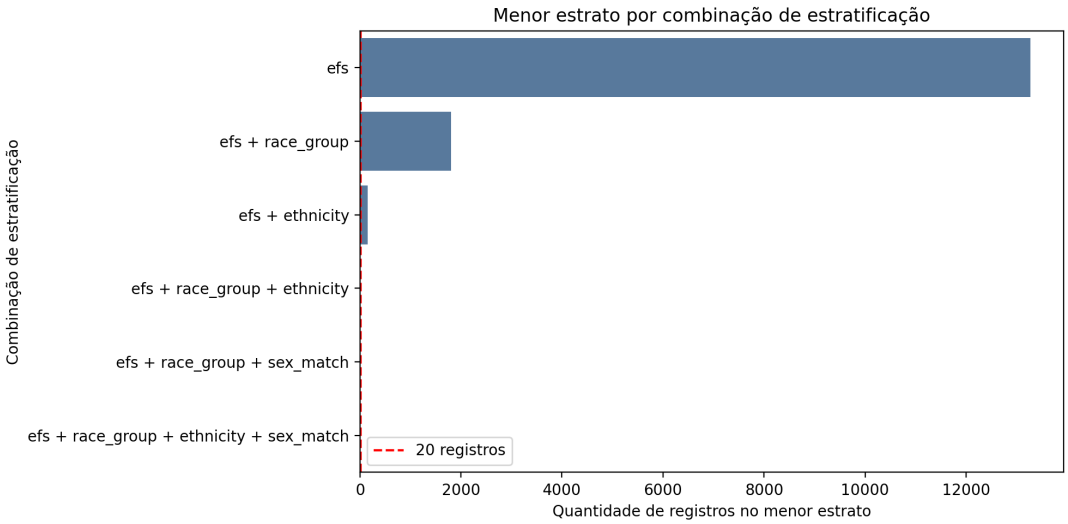

Figura 4.3: Tamanho do menor estrato em cada combinação avaliada para estratificação.

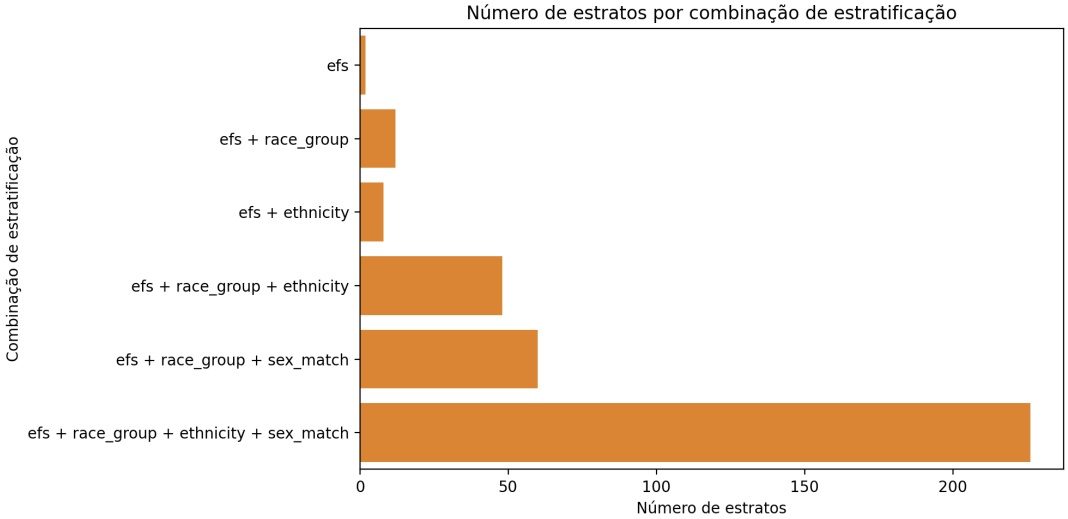

Figura 4.4: Número total de estratos gerados por cada combinação avaliada.

A Tabela 4.4 mostra a distribuição do desfecho após a divisão, evidenciando que a proporção entre as classes foi mantida nos subconjuntos.

Tabela 4.4: Distribuição do desfecho nos subconjuntos de treinamento e teste.

Subconjunto	**efs=0**	**efs=1**
Treinamento	9 287	10 873
Teste	3 981	4 659

Além do desfecho, foi necessário preservar a composição racial dos subconjun-

tos, pois a avaliação de equidade depende da comparação entre grupos. A Tabela 4.5 apresenta essa distribuição no treinamento e no teste, demonstrando que todos os grupos permaneceram representados nos dois subconjuntos.

Grupo racial	**Treinamento**	**Teste**
More than one race	3 392	1 453
Asian	3 383	1 449
White	3 381	1 450
Black or African-American	3 356	1 439
American Indian or Alaska Native	3 353	1 437
Native Hawaiian or other Pacific Islander	3 295	1 412

Tabela 4.5: Distribuição dos grupos raciais nos subconjuntos de treinamento e teste.

O subconjunto de teste foi mantido isolado durante o ajuste dos modelos. A seleção de hiperparâmetros e a análise de estabilidade foram realizadas apenas no subconjunto de treinamento, por validação cruzada estratificada com cinco partições. Em cada iteração da validação cruzada, quatro partições foram usadas para ajuste e uma para validação. O procedimento foi repetido até que cada partição atuasse uma vez como validação. A Figura 4.5 ilustra a rotação dos *folds* aplicada neste trabalho, particularizando, para a divisão adotada, o esquema geral de validação cruzada já apresentado na Figura 2.8.

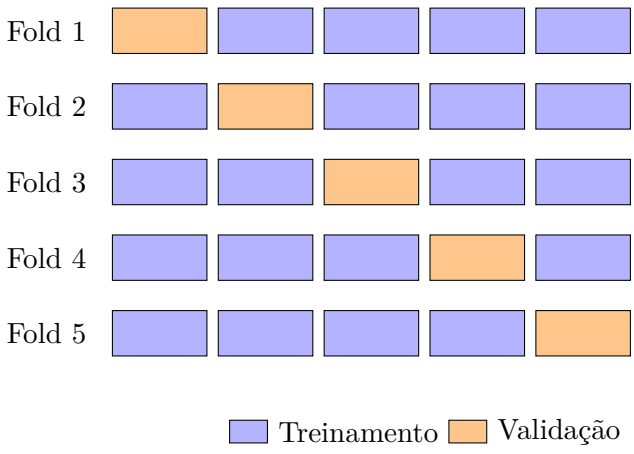

Figura 4.5: Esquema da validação cruzada com cinco partições no conjunto de treinamento.

4.4 Modelos Preditivos

Foram avaliados cinco modelos: *Random Forest*, *XGBoost*, *LightGBM*, *Stacking Ensemble* e LSTM. Os três primeiros foram ajustados por busca em grade de hiperparâmetros; o *Stacking Ensemble* combinou os três modelos baseados em árvores.

Neste trabalho, os hiperparâmetros considerados para avaliação foram: número de árvores, profundidade máxima, taxa de aprendizado e número de unidades de uma camada neural. Para os modelos baseados em árvores, a seleção foi feita com *GridSearchCV*, utilizando AUC-ROC como métrica de escolha no conjunto de treinamento.

A opção por AUC-ROC nessa etapa, em vez do C-index, decorre de uma diferença prática entre as duas métricas: o C-index é calculado sobre pares de indivíduos comparáveis e depende do tempo até o evento e do indicador de censura, o que o torna mais custoso de recalcular a cada uma das combinações de hiperparâmetros testadas na busca em grade. Porém, o AUC-ROC opera diretamente sobre o rótulo binário `efs` e é nativamente suportada como métrica de pontuação pelo *GridSearchCV* do scikit-learn.

Devido a restrições de implementação e de custo computacional, a seleção de hiperparâmetros foi realizada com base na AUC-ROC sobre o indicador binário `efs`. Entende-se que esse critério não incorpora o tempo de acompanhamento nem trata explicitamente a censura e, portanto, não deve ser considerado equivalente ou necessariamente alinhado ao C-index. Consequentemente, as configurações selecionadas foram otimizadas para a tarefa auxiliar de classificação, enquanto o C-index foi utilizado apenas na avaliação final. Portanto, essa diferença entre o objetivo de treinamento, o critério de seleção e a métrica principal de avaliação constitui uma limitação desse estudo.

Modelo	**Espaço de busca**	**Configuração selecionada**
Random Forest	$n\_estimators \in \{100, 150, 200\}$	$n\_estimators = 200$
	$max\_depth \in \{None, 10, 12, 20\}$	$max\_depth = 20$
	$min\_samples\_split \in \{2, 5, 10\}$	$min\_samples\_split = 10$
	$class\_weight \in \{balanced, None\}$	$class\_weight = None$
XGBoost	$n\_estimators \in \{100, 200, 300\}$	$n\_estimators = 300$
	$max\_depth \in \{3, 5, 7\}$	$max\_depth = 3$
	$learning\_rate \in \{0.01, 0.1\}$	$learning\_rate = 0.1$
	$subsample \in \{0.7, 1.0\}$	$subsample = 0.7$
LightGBM	$n\_estimators \in \{100, 200\}$	$n\_estimators = 100$
	$num\_leaves \in \{31, 50, 100\}$	$num\_leaves = 31$
	$learning\_rate \in \{0.01, 0.1, 0.3\}$	$learning\_rate = 0.1$

Tabela 4.6: Espaço de busca e hiperparâmetros selecionados para os modelos baseados em árvores.

No XGBoost, as árvores são adicionadas sequencialmente, e o aumento de *max_depth* eleva a complexidade de cada estimador e o risco de sobreajuste (CHEN; GUESTRIN, 2016). Por esse motivo, foram avaliadas profundidades relativamente conservadoras, $3, 5, 7$. Para o Random Forest, foi considerado um intervalo mais amplo, `None`, $10, 12, 20$, pois o método agrega as predições de múltiplas árvores aleatorizadas, reduzindo a variância associada aos estimadores individuais (BREIMAN, 2001). No LightGBM, a complexidade estrutural foi controlada principalmente por *num_leaves* e *max_depth*, em razão do crescimento *leaf-wise*, que pode produzir árvores assimétricas e profundas (KE et al., 2017). O intervalo de *learning_rate*, incluindo o valor 0.3, foi avaliado conjuntamente com diferentes números de estimadores, considerando a relação entre a magnitude das atualizações e a quantidade de iterações do boosting.

As faixas foram, portanto, definidas a partir das características de cada algoritmo, dos valores sugeridos nas respectivas documentações e do orçamento computacional disponível, buscando explorar configurações plausíveis para cada modelo.

O *Stacking Ensemble* foi composto por Random Forest, XGBoost e LightGBM como estimadores base, configurados com os hiperparâmetros selecionados pelos respectivos procedimentos de busca, apresentados na Tabela 4.6. As predições fora da amostra dos estimadores base, geradas por meio das cinco partições compartilhadas de validação cruzada, foram utilizadas como características de entrada para o meta-modelo de regressão logística (WOLPERT, 1992). O meta-modelo foi mantido em sua configuração padrão, sem busca de hiperparâmetros específica. Optou-se por não ajustar seus hiperparâmetros porque sua função no empilhamento é apenas combinar linearmente as predições dos três estimadores base já ajustados, papel para o qual a regressão logística com regularização padrão é suficiente. Após a geração dessas predições, os estimadores base foram ajustados sobre todo o conjunto de treinamento.

4.4.1 Modelo LSTM

A arquitetura da LSTM foi definida a partir de experimentos preliminares e de valores convencionais para problemas de classificação binária com redes recorrentes. A rede foi composta por duas camadas recorrentes com 50 unidades cada, intercaladas por camadas de *dropout* com taxa de 0,2. Após as camadas recorrentes, foi utilizada uma camada densa com 25 neurônios e ativação ReLU, seguida por uma camada de saída com um neurônio e ativação sigmoide, adequada à tarefa de classificação binária por restringir a saída da rede ao intervalo entre 0 e 1, interpretável como um escore associado à classe `efs=1`. O modelo foi treinado com otimizador Adam, função de perda *binary cross-entropy*, tamanho de lote igual a 32 e limite máximo de 100 épocas. A parada antecipada monitorou a perda de validação, com paciência de 10 épocas e restauração dos melhores pesos. A Figura 4.6 apresenta a arquitetura utilizada, indicando a ordem das camadas e os principais hiperparâmetros da rede.

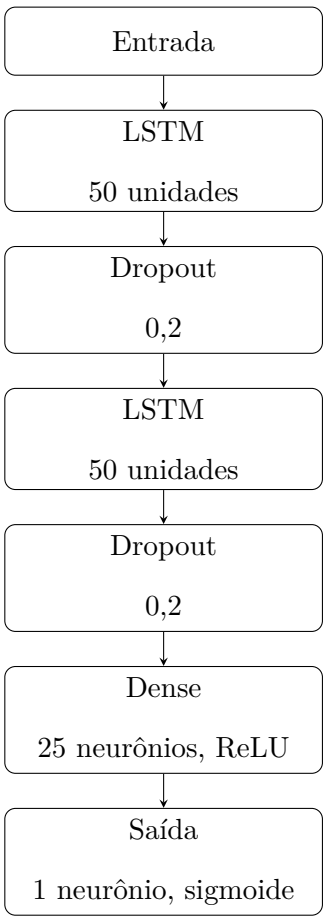

Figura 4.6: Arquitetura da LSTM utilizada no experimento.

4.5 Métricas de Avaliação

A avaliação final foi realizada no subconjunto de teste. Os modelos de classificação produzem, para cada paciente, uma probabilidade contínua entre 0 e 1 de pertencer à classe `efs=1`. Essa probabilidade foi convertida em uma classe prevista por meio de um limiar de decisão fixado em 0,5: probabilidades iguais ou superiores a esse valor foram classificadas como `efs=1`, e as demais como `efs=0`. A partir das classes previstas e das classes reais, foram calculadas acurácia, precisão, revocação e F1-score por classe, métricas derivadas da matriz de confusão entre rótulos previstos e observados e cujas definições formais são apresentadas no Capítulo 2. A AUC-ROC, por sua vez, foi calculada diretamente sobre as probabilidades estimadas para `efs=1`, sem depender do limiar de 0,5.

Também foi calculado o C-index, cuja formulação é apresentada no Capítulo 2, utilizando `efs_time`, `efs` e o escore de risco produzido por cada modelo. Para todos os

classificadores, o escore de risco correspondeu à probabilidade prevista para `efs=1`: nesse contexto, dois pacientes só formam um par comparável quando é possível determinar qual deles apresentou o evento primeiro, e o par é considerado concordante quando o modelo atribui maior probabilidade de evento ao paciente com menor tempo observado. Dessa forma, a avaliação considerou se pacientes com maior escore de risco tendiam a apresentar menor tempo livre de evento.

Para a análise de equidade, foi utilizada a métrica estratificada da competição, cuja formulação geral também é apresentada no Capítulo 2. O C-index foi calculado separadamente para cada grupo racial. Em seguida, a métrica final foi obtida pela média dos C-index dos grupos menos o desvio padrão entre esses valores, conforme a Eq. (4.1):

$$C_{\text{estratificado}} = \bar{C} - s_C, \tag{4.1}$$

em que $\bar{C}$ representa a média dos C-index calculados nos grupos raciais e s_C representa o desvio padrão desses valores.

Foi calculada também a amplitude racial, definida como a diferença entre o maior e o menor C-index entre os grupos. Diferentemente do C-index estratificado, que é a métrica oficial da competição, a amplitude racial não corresponde a uma métrica consagrada na literatura de aprendizado de máquina; ela foi adotada neste trabalho como medida complementar de disparidade, por expressar de forma direta a distância entre o grupo de melhor e o de pior desempenho, informação que a média e o desvio padrão, isoladamente, não tornam imediatamente visível.

Na base em estudo, `efs=0` indica censura, ou seja, o evento não foi observado até o último acompanhamento. Por isso, acurácia, precisão, revocação e F1-score descrevem apenas a classificação do status observado ao final do período disponível, e não a capacidade de prever a ocorrência futura do evento, já que um registro censurado não afirma que o evento jamais ocorrerá. Essas métricas são, portanto, auxiliares e descritivas. O C-index é mais adequado à estrutura temporal do conjunto, embora também apresente limitações, como a dependência da distribuição de censura dos dados (UNO et al., 2011).

4.6 Ferramentas Computacionais e Reprodutibilidade

Os experimentos foram implementados em Python, com semente aleatória definida como `42` nos procedimentos que permitem controle por `random_state`. Foram utilizadas bibliotecas voltadas à manipulação de dados, aprendizado de máquina, redes neurais, visualização e avaliação de sobrevivência, incluindo *pandas*, *NumPy*, *scikit-learn*, *TensorFlow/-Keras*, *XGBoost*, *LightGBM*, *lifelines*, *Matplotlib* e *Seaborn*.

Ao final do experimento, os principais resultados — resumos da base, distribuições do desfecho e dos grupos raciais, levantamento de valores ausentes, hiperparâmetros selecionados, métricas finais e C-index por grupo racial — foram registrados de forma automática, juntamente com as versões das bibliotecas utilizadas e um resumo metodológico. Esse registro garante a rastreabilidade entre o experimento conduzido, as tabelas geradas e os resultados apresentados no trabalho.

5 Resultados

Nesta seção são apresentados e discutidos os resultados obtidos a partir dos experimentos descritos no capítulo anterior. A análise é conduzida em duas etapas complementares. Inicialmente, realiza-se uma avaliação global do desempenho dos modelos, considerando métricas tradicionais de classificação aplicadas ao conjunto de teste. Essa etapa permite comparar a capacidade preditiva média dos diferentes algoritmos. Em seguida, é realizada uma análise estratificada do desempenho, na qual os resultados são examinados separadamente entre diferentes grupos populacionais. Essa segunda etapa permite investigar possíveis variações de desempenho entre grupos demográficos, aspecto central para a avaliação de desempenho por subgrupos nas previsões.

5.1 Desempenho Global dos Modelos

A Tabela 5.1 apresenta as métricas globais de classificação obtidas no conjunto de teste. Foram consideradas acurácia, AUC-ROC, precisão, revocação e F1-score por classe. Essa tabela permite comparar a capacidade discriminativa dos modelos e o equilíbrio entre os acertos nas duas classes.

Para precisão, revocação e F1-score, os valores são apresentados por classe, no formato classe 0 / classe 1, em que a classe 0 corresponde a efs=0 (observação censurada) e a classe 1 a efs=1 (evento observado). Optou-se por reportar as duas classes separadamente porque um dos objetivos da análise é justamente avaliar o equilíbrio dos modelos entre elas, aspecto que um valor agregado ocultaria.

Modelo	Acurácia	AUC-ROC	Precisão (0/1)	Revocação (0/1)	F1-score (0/1)
Stacking Ensemble	**0,6863**	**0,7531**	0,6877 / **0,6854**	0,5848 / 0,7731	**0,6321** / **0,7266**
XGBoost	0,6848	0,7523	0,6843 / 0,6852	**0,5865** / 0,7688	0,6317 / 0,7246
LightGBM	0,6791	0,7448	0,6780 / 0,6798	0,5780 / 0,7654	0,6240 / 0,7200
Random Forest	0,6763	0,7361	**0,6925** / 0,6673	0,5350 / 0,7970	0,6037 / 0,7264
LSTM	0,6617	0,7199	0,6865 / 0,6496	0,4891 / **0,8092**	0,5712 / 0,7206

Tabela 5.1: Desempenho global dos modelos no conjunto de teste.

O *Stacking Ensemble* apresentou a maior acurácia (0,6863) e a maior AUC-ROC (0,7531), seguido de perto pelo XGBoost, que obteve acurácia de 0,6848 e AUC-ROC de 0,7523. A diferença entre esses dois modelos foi pequena, indicando desempenho global muito próximo. O LightGBM e o Random Forest apresentaram desempenho intermediário, enquanto a LSTM obteve os menores valores de acurácia e AUC-ROC entre os modelos avaliados.

Observa-se também que todos os modelos apresentaram revocação maior para a classe `efs=1` do que para a classe `efs=0`. Esse padrão indica maior sensibilidade para identificar registros associados ao evento, e menor capacidade de identificar registros censurados como pertencentes à classe `efs=0`, considerando o limiar fixo adotado. A LSTM foi o caso mais acentuado: obteve revocação de 0,8092 para a classe `efs=1`, mas apenas 0,4891 para a classe `efs=0`. Portanto, embora a LSTM tenha identificado relativamente bem a classe positiva, seu desempenho global foi prejudicado pelo menor equilíbrio entre as classes.

5.2 C-index e Avaliação de Desempenho por Subgrupos

A Tabela 5.2 apresenta o C-index normal, o C-index estratificado e a amplitude racial de cada modelo. O C-index estratificado corresponde à métrica oficial da competição, calculada como a média dos C-index por grupo racial menos o desvio padrão entre esses grupos.

As métricas da Seção 5.1 foram calculadas sobre o rótulo binário `efs`. O C-index funciona de outra forma, pois usa o tempo até o evento (`efs_time`) e a censura. Com ele, avalia-se se o modelo ordena bem os pacientes pelo risco, ou seja, se um bom modelo atribui risco maior a quem apresenta o evento primeiro. A acurácia e a AUC-ROC não medem essa ordenação ao longo do tempo, que é o principal objetivo ao usar o C-index como avaliação temporal.

Modelo	**C-index normal**	**C-index estratificado**	**Amplitude racial**
XGBoost	**0,6604**	**0,6434**	**0,0459**
Stacking Ensemble	0,6603	0,6427	0,0487
LightGBM	0,6543	0,6376	0,0463
Random Forest	0,6471	0,6281	0,0555
LSTM	0,6378	0,6181	0,0596

Tabela 5.2: C-index, C-index estratificado e amplitude racial por modelo.

O XGBoost obteve os maiores valores numéricos de C-index, tanto global quanto estratificado, enquanto o Stacking Ensemble apresentou resultados muito próximos. Como não foram calculados intervalos de confiança ou testes de comparação, não é possível determinar se essas diferenças representam variação sistemática entre os modelos ou flutuação amostral da divisão utilizada.

A Tabela 5.3 detalha o C-index calculado separadamente para cada grupo racial e serve de base para as métricas de dispersão do C-index entre grupos da Tabela 5.2. Para cada modelo, o C-index estratificado corresponde à média dos seis valores de C-index por grupo menos o desvio padrão entre eles, conforme a Eq. (4.1), e a amplitude racial é a diferença entre o maior e o menor desses valores. A Tabela 5.2 resume desempenho e disparidade em três números por modelo, enquanto a Tabela 5.3 mostra o desempenho em cada grupo, de onde esses números vêm.

A título de ilustração, para o XGBoost os seis valores por grupo da Tabela 5.3 têm média $\bar{C} = 0{,}6582$ e desvio padrão $s_C = 0{,}0148$, o que resulta em $C_{\text{estratificado}} = 0{,}6582 - 0{,}0148 = 0{,}6434$, exatamente o valor reportado na Tabela 5.2. A amplitude

racial do mesmo modelo, $0{,}6765 - 0{,}6306 = 0{,}0459$, corresponde à diferença entre seus grupos de melhor e pior desempenho.

A amplitude racial é reportada junto com o C-index estratificado porque as duas medem coisas diferentes. O C-index estratificado junta em um só valor o desempenho médio dos grupos e a variação entre eles. Um valor alto, então, pode vir de um desempenho médio alto ou de uma variação pequena, e o número sozinho não separa os dois casos. A amplitude racial resolve isso ao mostrar apenas a variação, já que expressa a distância entre o grupo de melhor e o de pior desempenho. Com as duas medidas juntas, é possível ver o resultado resumido e também o quanto ele muda entre os grupos.

A comparação entre o C-index normal e o estratificado ajuda a entender o papel da penalização pela variação de desempenho entre grupos. Os dois índices ordenam os cinco modelos da mesma maneira, mas a queda de um para o outro não é igual em todos. Os modelos menos estáveis entre grupos caem mais: a LSTM e o Random Forest perdem cerca de 0,019 do C-index normal para o estratificado, enquanto o XGBoost e o LightGBM perdem cerca de 0,017. A penalização pelo desvio padrão modifica numericamente as pontuações dos modelos, mas, sem estimativa de incerteza, não permite classificá-los de forma conclusiva quanto à estabilidade entre grupos. Ele também muda qual modelo aparece na frente. Na avaliação global da Seção 5.1, o *Stacking Ensemble* liderava em acurácia e AUC-ROC. Aqui, nas métricas de concordância e de variação entre grupos, quem lidera é o XGBoost. O modelo de melhor desempenho global nem sempre é o que apresenta menor dispersão do C-index entre os grupos, e é por isso que uma métrica focada na variação entre grupos faz diferença.

Grupo racial	**Random Forest**	**XGBoost**	**LightGBM**	**LSTM**	**Stacking Ensemble**
American Indian or Alaska Native	0,6667	0,6765	0,6721	0,6642	**0,6775**
Asian	0,6551	**0,6680**	0,6598	0,6321	**0,6680**
Black or African-American	0,6400	0,6505	0,6447	0,6337	**0,6506**
More than one race	0,6503	0,6576	0,6578	0,6453	**0,6592**
Native Hawaiian or other Pacific Islander	0,6485	**0,6657**	0,6517	0,6350	0,6651
White	0,6112	**0,6306**	0,6258	0,6047	0,6288

Tabela 5.3: C-index por grupo racial nos modelos avaliados. Os valores em negrito indicam o melhor resultado por grupo racial.

A leitura por linha da Tabela 5.3 mostra que o melhor C-index de cada grupo racial ficou quase sempre com o *Stacking Ensemble* ou com o XGBoost, que se revezam nas primeiras posições e empatam no grupo *Asian*. A LSTM ficou com o menor C-index na maioria dos grupos, o que acompanha seu desempenho global mais baixo. O grupo *American Indian or Alaska Native* teve os maiores valores de concordância em todos os modelos, e o grupo *White* teve os menores. Como esse comportamento se repete nas cinco arquiteturas, a diferença entre grupos não parece depender do modelo escolhido.

A Figura 5.1 mostra de forma visual o padrão de variação entre grupos. Como a tabela já traz os valores de cada grupo, o que o gráfico acrescenta é a ordem entre eles, que se mantém parecida de um modelo para outro: o grupo *White* fica sempre na base, o *American Indian or Alaska Native* no topo e os demais em posições próximas. Essa regularidade sugere que a disparidade está ligada aos dados e ao problema, e não a uma arquitetura em particular. Por isso, faz sentido avaliar o desempenho por subgrupos, e não apenas pelas métricas gerais. É preciso, porém, ter cuidado com a leitura desse resultado. A diferença entre grupos diz respeito ao desempenho dos modelos, não a uma característica dos pacientes, e sua interpretação deve levar em conta as ressalvas da Seção 4.1 sobre o uso de `race_group` como categoria demográfica.

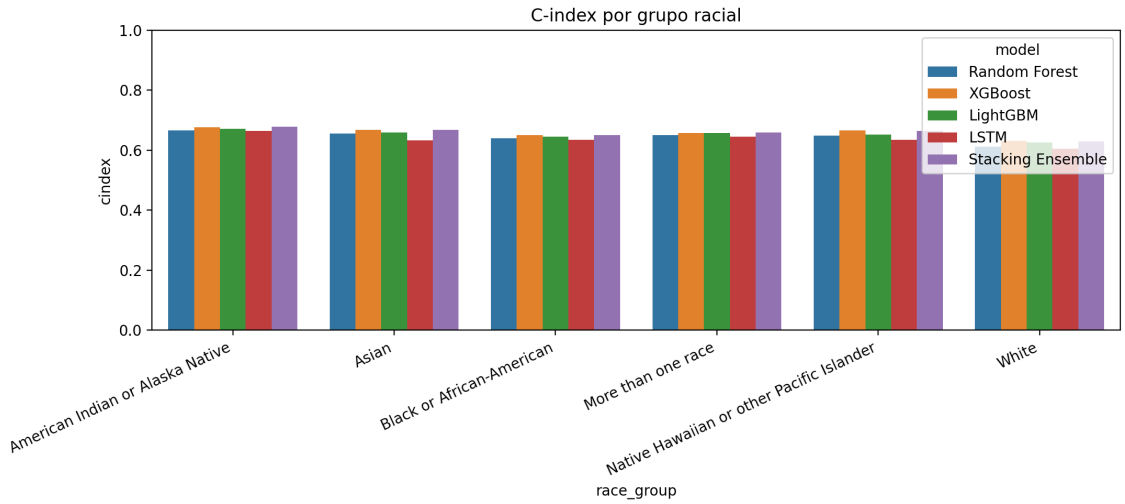

Figura 5.1: C-index por grupo racial para cada modelo avaliado.

5.3 Avaliação das Hipóteses de Pesquisa

Os resultados obtidos permitiram reavaliar as hipóteses formuladas na Seção 1.1. A Tabela 5.4 sintetiza o comportamento das hipóteses investigadas com base nos resultados experimentais obtidos nas Seções 5.1 e 5.2 e logo após a Tabela são analisadas em detalhes cada uma das proposições.

Hipótese	**Descrição**	**Resultado**	**Evidência observada**
H1	Modelos baseados em árvores apresentariam desempenho superior à LSTM na classificação do indicador `efs` e na ordenação de risco pós-TCTH em dados tabulares.	Confirmada	Random Forest, XGBoost e LightGBM superaram a LSTM nas métricas globais de classificação e concordância.
H2	O empilhamento de modelos reduziria as diferenças observadas de desempenho discriminativo entre grupos demográficos.	Parcialmente sustentada	O Stacking Ensemble apresentou melhor desempenho global, porém o XGBoost obteve maior C-index estratificado e menor amplitude racial.
H3	Modelos com maior desempenho global não necessariamente apresentariam maior estabilidade de desempenho entre grupos raciais.	Parcialmente sustentada	A análise estratificada mostrou diferenças entre desempenho global e estabilidade entre grupos raciais, especialmente na comparação entre Stacking Ensemble, XGBoost e LSTM.

Tabela 5.4: Síntese da avaliação das hipóteses de pesquisa.

Hipótese 1

A primeira hipótese afirmava que modelos baseados em árvores, como *Random Forest*, *XGBoost* e *LightGBM*, apresentariam desempenho preditivo superior ao modelo LSTM na classificação do indicador `efs` e na ordenação de risco pós-TCTH em dados tabulares.

Essa hipótese foi confirmada no contexto experimental deste trabalho. Os três modelos baseados em árvores superaram a LSTM em acurácia, AUC-ROC, C-index normal e C-index estratificado. A LSTM apresentou revocação maior para a classe `efs=1`, mas esse ganho ocorreu acompanhado de pior desempenho na classe `efs=0` e menor desempenho global. Portanto, a confirmação da hipótese deve ser entendida em termos de desempenho global e de concordância, não como superioridade dos modelos de árvore em todas as métricas isoladas.

Hipótese 2

A segunda hipótese afirmava que o uso de empilhamento de modelos reduziria diferenças de desempenho discriminativo entre grupos demográficos, resultando em C-index estratificado superior aos modelos individuais.

Essa hipótese foi apenas parcialmente sustentada. O *Stacking Ensemble* apresentou o melhor desempenho global de classificação, com maior acurácia e maior AUC-ROC. No entanto, ele não obteve o maior C-index estratificado nem a menor amplitude racial. Nessas duas métricas, o XGBoost individual apresentou desempenho ligeiramente superior. Assim, os resultados indicam que o empilhamento melhorou a discriminação global, mas não demonstrou vantagem clara na dispersão do C-index entre os grupos quando comparado ao melhor modelo individual.

Hipótese 3

A terceira hipótese afirmava que modelos com maior desempenho preditivo global não necessariamente apresentariam menor dispersão do C-index entre diferentes grupos raciais.

Os resultados obtidos sustentam parcialmente essa hipótese. Observou-se que os modelos com melhor desempenho global não foram exatamente os mesmos que apresen-

taram os melhores indicadores de estabilidade entre grupos raciais. O *Stacking Ensemble*, por exemplo, obteve a maior AUC-ROC global entre os modelos avaliados, enquanto o XGBoost apresentou menor amplitude racial e maior C-index estratificado.

Além disso, a análise estratificada revelou diferenças de desempenho entre grupos raciais que não seriam perceptíveis apenas por métricas globais agregadas. Esse comportamento foi particularmente evidente na comparação com a LSTM, que apresentou desempenho global inferior e maior variação entre os grupos analisados.

Os resultados indicam, portanto, que métricas globais de classificação e concordância não são suficientes, isoladamente, para caracterizar o desempenho por subgrupos dos modelos preditivos. Nesse contexto, a utilização de métricas estratificadas mostrou-se relevante para identificar disparidades de desempenho entre subgrupos populacionais.

6 Conclusão

Este trabalho desenvolveu e avaliou modelos de aprendizado de máquina para a classificação do indicador de sobrevivência livre de eventos (`efs`) após transplante de células-tronco hematopoéticas, utilizando dados da competição *Equity in Post-HCT Survival Predictions*, do CIBMTR. A modelagem foi conduzida como uma tarefa de classificação binária do desfecho `efs`, enquanto a variável `efs_time` foi utilizada na avaliação por C-index.

Os resultados indicaram que modelos baseados em árvores foram mais adequados à tarefa do que a LSTM avaliada. O *Stacking Ensemble* apresentou o melhor desempenho global de classificação, com acurácia de 0,6863 e AUC-ROC de 0,7531. O XGBoost, por sua vez, obteve o melhor desempenho nas métricas de concordância e de variação entre grupos, com C-index normal de 0,6604, C-index estratificado de 0,6434 e a menor amplitude racial entre os modelos avaliados.

A comparação entre XGBoost e *Stacking Ensemble* mostrou diferenças pequenas. Dessa forma, a conclusão mais adequada não é que um modelo tenha superado amplamente o outro, mas que ambos se destacaram sob perspectivas diferentes: o *Stacking Ensemble* nas métricas globais de classificação e o XGBoost nas métricas baseadas em C-index e variação entre grupos raciais.

A análise estratificada mostrou que a avaliação por métricas agregadas é insuficiente para compreender plenamente o comportamento dos modelos. Mesmo os modelos com melhor desempenho global apresentaram variações entre grupos raciais. Nos resultados finais, o grupo *White* apresentou o menor C-index em todos os modelos, enquanto os maiores valores ocorreram, em geral, no grupo *American Indian or Alaska Native*. Esse achado reforça a necessidade de analisar o desempenho por subgrupos, em vez de depender apenas de métricas médias.

Os resultados obtidos demonstram que a comparação entre modelos preditivos em aplicações clínicas não deve considerar apenas métricas globais de classificação, mas também critérios relacionados à estabilidade entre subgrupos populacionais. Embora os

modelos avaliados tenham apresentado desempenho competitivo, observou-se variação consistente entre grupos raciais, evidenciando a relevância da análise estratificada em cenários de predição clínica.

Além da comparação entre arquiteturas, o trabalho reforça a importância da avaliação de desempenho por subgrupos em modelos aplicados à saúde, especialmente em contextos envolvendo decisões clínicas potencialmente sensíveis. Nesse sentido, métricas como o C-index estratificado mostraram-se relevantes para complementar a análise tradicional baseada em acurácia e AUC-ROC.

6.1 Limitações e Trabalhos Futuros

Os resultados apresentados devem ser interpretados à luz de algumas limitações relacionadas à formulação do problema, ao conjunto de dados e às escolhas metodológicas adotadas. Esta seção reúne essas limitações e, a partir delas, aponta direções para trabalhos futuros.

A primeira limitação decorre da forma como o problema foi formulado. Os modelos foram treinados como classificadores binários do indicador `efs`, de modo que o tempo de acompanhamento (`efs_time`) e a informação de censura não participaram do treinamento, sendo usados apenas na avaliação por C-index, conforme descrito na Seção 4.5. As saídas produzidas correspondem a escores associados à classe `efs=1` e não a funções de sobrevivência ou a probabilidades calibradas de ocorrência do evento em um horizonte de tempo definido. Como `efs=0` representa uma observação censurada, ou seja, um registro para o qual o evento não foi observado até o fim do acompanhamento, as métricas de acurácia, precisão, revocação e F1-score descrevem a classificação do status observado na base e não a capacidade de prever a ocorrência futura do evento.

Uma segunda limitação está na diferença entre o critério de seleção de hiperparâmetros e a métrica principal de avaliação. Os modelos baseados em árvores foram ajustados pela AUC-ROC calculada sobre o rótulo binário `efs`, enquanto a comparação principal utilizou o C-index, que considera o tempo observado e a censura. Como esses critérios não são equivalentes, as configurações selecionadas podem ser adequadas para a tarefa de classificação sem serem as que maximizariam a concordância em uma modelagem temporal, conforme já observado na Seção 4.4.

A avaliação também foi conduzida sobre uma única divisão entre treinamento e teste. Além disso, cada modelo foi treinado e avaliado uma única vez, com semente fixa e sem repetição da execução sob diferentes inicializações. Não foram realizados particionamentos repetidos, validação cruzada externa, reamostragem por *bootstrap* ou cálculo de intervalos de confiança para as métricas finais. Por esse motivo, diferenças pequenas, como as observadas entre XGBoost e *Stacking Ensemble*, não permitem afirmar a superioridade de um modelo sobre o outro, e a comparação deve ser lida de forma descritiva.

Outra limitação diz respeito ao escopo da avaliação por subgrupos. Neste trabalho, o desempenho entre grupos foi resumido pela dispersão do C-index entre os grupos de `race_group`, por meio da média dos C-index menos o desvio padrão e da amplitude entre o maior e o menor valor. Essas medidas identificam diferenças de capacidade discriminativa entre grupos, mas não avaliam a calibração por grupo, as diferenças nas distribuições de censura, os falsos positivos e falsos negativos clinicamente relevantes, o benefício ou dano clínico, a interseccionalidade entre características demográficas nem os mecanismos causais das diferenças observadas. Uma menor dispersão descritiva do C-index, portanto, não constitui demonstração de equidade clínica, alerta recorrente na literatura sobre justiça algorítmica em predição clínica de risco.

A característica `race_group` foi utilizada ao mesmo tempo como preditor e como critério para estratificar a avaliação. Não foi realizado um experimento de ablação comparando modelos com e sem essa característica, o que impede determinar se seu uso direto contribuiu para reduzir ou ampliar as diferenças de desempenho entre os grupos. Como observado na Seção 4.1, essas categorias são demográficas e não devem ser interpretadas como medidas de ancestralidade genética.

O caráter sintético da base impõe uma limitação adicional. Os dados foram construídos para uma competição e apresentam grupos raciais aproximadamente balanceados por construção, o que não reproduz a distribuição de pacientes submetidos ao TCTH em populações reais. O processo de geração sintética também pode não capturar integralmente os mecanismos clínicos, sociais e de censura presentes em registros observacionais. Assim, os resultados não devem ser generalizados diretamente para instituições de saúde.

No pré-processamento, foram identificados valores ausentes em 50 das 60 carac-

terísticas. Para preservar os registros disponíveis e manter condições equivalentes entre os modelos, adotou-se um único *pipeline*, com imputação pela mediana nas características numéricas e pela categoria mais frequente nas categóricas. Essa estratégia é simples e uniforme, mas pode não representar bem situações em que a ausência tem significado clínico. Na característica `cyto_score`, por exemplo, a ausência concentra-se em pacientes cuja doença não define o escore citogenético, o que sugere um mecanismo de ausência potencialmente não aleatório. Nesses casos, a imputação pela categoria mais frequente pode suprimir parte da informação representada pela própria ausência (RUBIN, 1976). Não foi realizado um estudo que quantificasse separadamente o efeito de cada transformação sobre o desempenho e a dispersão do C-index entre grupos.

Também foram mantidos preditores que representam informações semelhantes ou deterministicamente relacionadas. As contagens agregadas de compatibilidade HLA para painéis de 6, 8 e 10 marcadores, por exemplo, derivam das medidas por locus, introduzindo redundância entre os preditores (KUHN; JOHNSON, 2013). Essa redundância, combinada às diferentes taxas de preenchimento, não foi tratada por um procedimento sistemático de seleção de características.

A comparação com a LSTM deve ser lida com cautela semelhante. A entrada foi organizada com comprimento de sequência igual a um, de modo que a rede não explorou dependências recorrentes entre instantes. Além disso, os modelos baseados em árvores passaram por busca em grade de hiperparâmetros, enquanto a arquitetura neural foi definida a partir de experimentos preliminares. Os resultados referem-se, portanto, à configuração específica avaliada e não permitem concluir que redes neurais sejam, de modo geral, menos adequadas que modelos baseados em árvores para esse problema.

Essas limitações não comprometem a comparação interna realizada, uma vez que todos os modelos receberam o mesmo conjunto de características e o mesmo pré-processamento, ajustado apenas sobre os dados de treinamento. Elas afetam, porém, os valores absolutos das métricas e restringem a generalização dos resultados para outras estratégias de preparação e para dados reais.

A partir dessas limitações, várias direções de pesquisa se mostram promissoras:

- **Formular o problema com métodos de análise de sobrevivência:** adotar

modelos em que o tempo e a censura participem do treinamento, como modelos de riscos proporcionais, florestas de sobrevivência, *boosting* com funções objetivo de sobrevivência e arquiteturas neurais voltadas a dados de tempo até o evento. A avaliação poderia incluir, além do C-index, a AUC dependente do tempo, o escore de Brier integrado e curvas de calibração em horizontes clinicamente definidos.

- **Quantificar a incerteza das comparações:** empregar validação cruzada repetida ou aninhada, reamostragem por *bootstrap* e intervalos de confiança para as diferenças entre modelos e entre grupos, de modo a distinguir variação sistemática de flutuação amostral.

- **Ampliar a avaliação por subgrupos:** considerar a calibração dentro de cada grupo, as diferenças nos mecanismos de censura, as interseções entre características demográficas e medidas de utilidade clínica, além de um experimento de ablação com e sem `race_group` como preditor.

- **Tratar os valores ausentes conforme o motivo da ausência:** comparar alternativas como indicadores explícitos de ausência, categorias de "não aplicável", imputação múltipla e métodos que tratem valores ausentes de forma nativa, verificando se elas afetam os grupos de `race_group` de maneiras diferentes (RUBIN, 1976).

- **Selecionar características e reduzir redundância:** investigar procedimentos de seleção e comparar configurações com e sem preditores deterministicamente relacionados, como algumas contagens agregadas de compatibilidade HLA e suas medidas por locus, avaliando se modelos mais simples preservam o desempenho (KUHN; JOHNSON, 2013).

- **Explorar e otimizar arquiteturas neurais:** aplicar otimização sistemática de hiperparâmetros, por meio de técnicas como *Bayesian Optimization* ou *Neural Architecture Search*, e avaliar variantes como *Gated Recurrent Units* (GRUs), redes convolucionais e mecanismos de atenção.

- **Realizar validação externa:** avaliar os modelos em dados clínicos reais, de di-

ferentes centros e períodos, com recalibração e análise de benefício clínico, etapa necessária antes de qualquer consideração sobre uso na tomada de decisão.

Até que essas avaliações sejam realizadas, os resultados deste trabalho devem ser entendidos como evidência metodológica exploratória obtida sobre uma base sintética, e não como validação de uma ferramenta destinada à tomada de decisão clínica.

A Dicionário de Variáveis

Este apêndice apresenta o dicionário das variáveis utilizadas neste trabalho. Os tipos indicados correspondem à representação lida pelo *pipeline* antes das etapas de imputação, padronização e codificação.

Variável	**Tipo de dado**	**Descrição**
`ID`	Numérico	Identificador do registro no conjunto de dados. Foi removido dos preditores por não representar uma característica clínica, demográfica ou procedimental do paciente.
`dri_score`	Categórico	Classificação do *Disease Risk Index* (DRI), utilizada para representar o risco associado à doença de base no contexto do transplante.
`psych_disturb`	Categórico	Indicação da presença de distúrbio psiquiátrico registrado antes do transplante.
`cyto_score`	Categórico	Classificação citogenética associada ao prognóstico da doença.
`diabetes`	Categórico	Indicação de diabetes como comorbidade registrada no histórico clínico.
`hla_match_c_high`	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-C em alta resolução.
`hla_high_res_8`	Numérico	Contagem agregada de compatibilidade HLA em alta resolução considerando oito marcadores.

Continua na próxima página

Tabela A.1 – continuação

Variável	Tipo de dado	Descrição
tbi_status	Categórico	Situação relacionada ao uso de irradiação corporal total no regime de condicionamento pré-transplante.
arrhythmia	Categórico	Indicação de histórico de arritmia cardíaca.
hla_low_res_6	Numérico	Contagem agregada de compatibilidade HLA em baixa resolução considerando seis marcadores.
graft_type	Categórico	Tipo de enxerto utilizado no transplante de células-tronco hematopoéticas.
vent_hist	Categórico	Indicação de histórico de ventilação mecânica.
renal_issue	Categórico	Indicação de comprometimento renal ou comorbidade renal registrada.
pulm_severe	Categórico	Indicação de comprometimento pulmonar grave.
prim_disease_hct	Categórico	Doença primária que motivou a realização do transplante.
hla_high_res_6	Numérico	Contagem agregada de compatibilidade HLA em alta resolução considerando seis marcadores.
cmv_status	Categórico	Situação sorológica para citomegalovírus no par doador-receptor.
hla_high_res_10	Numérico	Contagem agregada de compatibilidade HLA em alta resolução considerando dez marcadores.
hla_match_dqb1_high	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-DQB1 em alta resolução.

Continua na próxima página

Tabela A.1 – continuação

Variável	Tipo de dado	Descrição
tce_imm_match	Categórico	Classificação de compatibilidade imunológica relacionada ao *T-cell epitope* (TCE).
hla_nmdp_6	Numérico	Medida de compatibilidade HLA segundo critério NMDP considerando seis marcadores.
hla_match_c_low	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-C em baixa resolução.
rituximab	Categórico	Indicação de uso de rituximabe no contexto terapêutico registrado.
hla_match_drb1_low	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-DRB1 em baixa resolução.
hla_match_dqb1_low	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-DQB1 em baixa resolução.
prod_type	Categórico	Tipo de produto celular utilizado no transplante.
cyto_score_detail	Categórico	Classificação citogenética detalhada associada à doença de base.
conditioning_intensity	Categórico	Intensidade do regime de condicionamento aplicado antes do transplante.
ethnicity	Categórico	Categoria de etnia registrada no conjunto de dados.
year_hct	Numérico	Ano de realização do transplante.
obesity	Categórico	Indicação de obesidade como condição clínica registrada.

Continua na próxima página

Tabela A.1 – continuação

Variável	**Tipo de dado**	**Descrição**
`mrd_hct`	Categórico	Indicação relacionada à presença de doença residual mensurável no momento do transplante.
`in_vivo_tcd`	Categórico	Indicação de depleção de células T *in vivo*.
`tce_match`	Categórico	Classificação de compatibilidade por *T-cell epitope* entre doador e receptor.
`hla_match_a_high`	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-A em alta resolução.
`hepatic_severe`	Categórico	Indicação de comprometimento hepático grave.
`donor_age`	Numérico	Idade do doador no momento relacionado ao transplante.
`prior_tumor`	Categórico	Indicação de tumor prévio registrado no histórico clínico.
`hla_match_b_low`	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-B em baixa resolução.
`peptic_ulcer`	Categórico	Indicação de úlcera péptica como condição clínica registrada.
`age_at_hct`	Numérico	Idade do paciente no momento do transplante.
`hla_match_a_low`	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-A em baixa resolução.
`gvhd_proph`	Categórico	Estratégia de profilaxia para doença do enxerto contra o hospedeiro.
`rheum_issue`	Categórico	Indicação de doença ou condição reumatológica registrada.

Continua na próxima página

Tabela A.1 – continuação

Variável	**Tipo de dado**	**Descrição**
sex_match	Categórico	Compatibilidade ou combinação de sexo entre doador e receptor.
hla_match_b_high	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-B em alta resolução.
race_group	Categórico	Grupo racial registrado no conjunto de dados; foi utilizado como preditor e também como variável de estratificação para avaliação de equidade.
comorbidity_score	Numérico	Escore agregado de comorbidades do paciente.
karnofsky_score	Numérico	Escore funcional de Karnofsky, utilizado para representar o estado geral de desempenho do paciente.
hepatic_mild	Categórico	Indicação de comprometimento hepático leve.
tce_div_match	Categórico	Classificação de compatibilidade ou permissividade relacionada à divergência de *T-cell epitope*.
donor_related	Categórico	Indicação de relação de parentesco ou vínculo entre doador e receptor.
melphalan_dose	Categórico	Categoria relacionada à dose de melfalano utilizada no regime terapêutico.
hla_low_res_8	Numérico	Contagem agregada de compatibilidade HLA em baixa resolução considerando oito marcadores.
cardiac	Categórico	Indicação de condição cardíaca ou comorbidade cardiovascular registrada.

Continua na próxima página

Tabela A.1 – continuação

Variável	Tipo de dado	Descrição
hla_match_drb1_high	Numérico	Grau de compatibilidade entre doador e receptor para o locus HLA-DRB1 em alta resolução.
pulm_moderate	Categórico	Indicação de comprometimento pulmonar moderado.
hla_low_res_10	Numérico	Contagem agregada de compatibilidade HLA em baixa resolução considerando dez marcadores.
efs	Numérico	Indicador binário de sobrevivência livre de eventos. Neste trabalho, foi utilizado como alvo de treinamento dos modelos de classificação.
efs_time	Numérico	Tempo de acompanhamento associado ao desfecho efs. Neste trabalho, foi utilizado na avaliação por C-index.
donor_recipient_age_diff	Numérico	Atributo derivado criado no *pipeline*, definido como a diferença entre donor_age e age_at_hct.

Tabela A.1: Dicionário de dados: variáveis do estudo, tipos e descrições.

Bibliografia

Analytics Vidhya. *Fake News Classifier on US Election News – LSTM*. 2020. Acesso em: 25 abr. 2026. Disponível em: ⟨https://www.analyticsvidhya.com/blog/2020/12/fake-news-classifier-on-us-election-news/⟩.

APPELBAUM, F. R. Hematopoietic cell transplantation as curative therapy for leukemia. *Best Practice & Research Clinical Haematology*, Elsevier, v. 14, n. 4, p. 585–596, 2001.

ARAI, Y. et al. Allogeneic unrelated bone marrow transplantation from older donors results in worse prognosis in recipients with aplastic anemia. *Haematologica*, v. 101, n. 5, p. 644–652, 2016.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006.

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

BREIMAN, L. et al. *Classification and Regression Trees*. Belmont, California: Wadsworth International Group, 1984.

CHADAGA, K. et al. A machine learning and explainable artificial intelligence approach to predict the efficacy of hematopoietic stem cell transplant in pediatric patients. *Artificial Intelligence in Medicine*, v. 144, p. 102456, 2023.

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.: s.n.], 2016.

CHO, C. et al. Application of the cibmtr one year survival outcomes calculator as a tool for retrospective analysis. *Bone Marrow Transplantation*, v. 58, n. 10, p. 1089–1095, 2023.

CIBMTR. *Equity in Post-HCT Survival Predictions*. 2024. Kaggle Competition. Available at: ⟨https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions/⟩.

COPELAN, E. A. Hematopoietic stem-cell transplantation. *The New England Journal of Medicine*, Massachusetts Medical Society, v. 354, n. 17, p. 1813–1826, 2006.

DAMOTTE, V. et al. Multiple measures for self-identification improve matching donors with patients in unrelated hematopoietic stem cell transplant. *Communications Medicine*, v. 4, p. 189, 2024.

DIETTERICH, T. G. Ensemble methods in machine learning. *Multiple Classifier Systems*, Springer, p. 1–15, 2000.

EFRON, B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 7, n. 1, p. 1–26, 1979.

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

GEEKSFORGEEKS. *Introduction to Recurrent Neural Networks*. 2025. ⟨https://www.geeksforgeeks.org/machine-learning/introduction-to-recurrent-neural-network/⟩. Acesso em: 07 out. 2025.

GOLDSTEIN, B. A. et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, Oxford University Press, v. 24, n. 1, p. 198–208, 2017.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. ⟨http://www.deeplearningbook.org⟩.

GRAGERT, L. et al. Hla match likelihoods for hematopoietic stem-cell grafts in the u.s. registry. *New England Journal of Medicine*, v. 371, n. 4, p. 339–348, 2014.

HARTMAN, N. et al. Pitfalls of the concordance index for survival outcomes. *Statistics in Medicine*, v. 42, n. 13, p. 2179–2190, 2023. Disponível em: ⟨https://pmc.ncbi.nlm.nih.gov/articles/PMC10219847/⟩.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. [S.l.]: Springer, 2009.

HOCHREITER, S. *Untersuchungen zu dynamischen neuronalen Netzen*. 1991. Diploma thesis, Institut für Informatik, Technische Universität München.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997.

HOLLENBACH, J. A. et al. Race, ethnicity and ancestry in unrelated transplant matching for the national marrow donor program: A comparison of multiple forms of self-identification with genetics. *PLOS ONE*, v. 10, n. 8, p. e0135960, 2015.

JAIN, A. *Everything about Random Forest*. 2024. Imagem: "Random forest — workflow". Disponível em: ⟨https://medium.com/@abhishekjainindore24/everything-about-random-forest-90c106d63989⟩.

JR, F. E. H. et al. Evaluating the yield of medical tests. *JAMA*, American Medical Association, v. 247, n. 18, p. 2543–2546, 1982.

KE, G. et al. LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems 30*. [S.l.: s.n.], 2017.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. [S.l.]: Morgan Kaufmann, 1995. p. 1137–1143.

KUHN, M.; JOHNSON, K. *Applied Predictive Modeling*. New York: Springer, 2013.

KUHN, M.; JOHNSON, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. [S.l.]: CRC Press, 2019.

LANDRY, I. Racial disparities in hematopoietic stem cell transplant: a systematic review of the literature. *Stem Cell Investigation*, v. 8, p. 24, 2021.

LI, C.; JIANG, X.; ZHANG, K. A transformer-based deep learning approach for fairly predicting post-liver transplant risk factors. *Journal of Biomedical Informatics*, v. 149, p. 104545, 2024. Originally published as arXiv preprint arXiv:2304.02780.

LIU, Y. et al. *Illustration of LightGBM model creation [Figura].* 2024. Figura 4 do artigo *Prediction of soil organic carbon content using machine learning models based on visible and near-infrared spectroscopy*. ⟨https://www.researchgate.net/figure/Illustration-of-Light-GBM-model-creation_fig4_377930134⟩, acessado em 7 de outubro de 2025.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.

Neigrando. *Neurônios e Redes Neurais Artificiais*. 2022. Acesso em: 25 jul. 2025. Disponível em: ⟨https://neigrando.com/2022/03/03/neuronios-e-redes-neurais-artificiais/⟩.

OBERMEYER, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, American Association for the Advancement of Science, v. 366, n. 6464, p. 447–453, 2019.

OBERMEYER, Z. et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, American Association for the Advancement of Science, v. 366, n. 6464, p. 447–453, 2019.

PAGE, M. J. et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, BMJ Publishing Group, v. 372, 2021.

PARK, S. Y. et al. Review of statistical methods for evaluating the performance of survival models. *Journal of Clinical Oncology*, American Society of Clinical Oncology, v. 39, n. 11, p. 1225–1235, 2021.

PFOHL, S. R.; FORYCIARZ, A.; SHAH, N. H. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, v. 113, p. 103621, 2021.

RAJKOMAR, A.; DEAN, J.; KOHANE, I. Machine learning in medicine. *New England Journal of Medicine*, Massachusetts Medical Society, v. 380, n. 14, p. 1347–1358, 2019.

RAJKOMAR, A. et al. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, v. 172, n. 9, p. 589–594, 2020.

RATUL, I. et al. Survival prediction of children undergoing hematopoietic stem cell transplantation using different machine learning classifiers by performing chi-square test and hyperparameter optimization: A retrospective analysis. *Computational and Mathematical Methods in Medicine*, v. 2022, p. 9391136, 2022.

RICHARDSON, W. S. et al. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, v. 123, n. 3, p. A12–A13, 1995.

ROUZBAHANI, M. et al. Predictive modeling of outcomes in acute leukemia patients undergoing allogeneic hematopoietic stem cell transplantation using machine learning techniques. *Leukemia Research*, v. 148, p. 107619, 2025.

RUBIN, D. B. Inference and missing data. *Biometrika*, Oxford University Press, v. 63, n. 3, p. 581–592, 1976.

SCHAPIRE, R. E. A brief introduction to boosting. *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1999.

SHOURABIZADEH, H. et al. Machine learning for the prediction of survival post-allogeneic hematopoietic cell transplantation: A single-center experience. *Acta Haematologica*, S.KargerAG, v. 147, n. 3, p. 280–291, September 2024. Disponível em: ⟨https://doi.org/10.1159/000533665⟩.

SORROR, M. L. et al. Comorbidity-age index: a clinical measure of biologic age before allogeneic hematopoietic cell transplantation. *Journal of Clinical Oncology*, American Society of Clinical Oncology, v. 23, n. 16, p. 3833–3841, 2005.

UNO, H. et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, v. 30, n. 10, p. 1105–1117, 2011.

WOLPERT, D. H. Stacked generalization. *Neural Networks*, v. 5, n. 2, p. 241–259, 1992.

ZHOU, F. et al. Degradation state recognition of piston pump based on iceemdan and xgboost. *Applied Sciences*, v. 10, n. 18, p. 6593, 2020.