INSTITUTO LATINO-AMERICANO DE CIÊNCIAS DA VIDA E DA NATUREZA

FERNANDO JOSÉ ZARDINELLO BATISTTI

PROJETO I: PREVISÃO PREDITIVA DA SEVERIDADE DO CÂNCER

FOZ DO IGUAÇU

2025

FERNANDO JOSÉ ZARDINELLO BATISTTI

PROJETO I: PREVISÃO PREDITIVA DA SEVERIDADE DO CÂNCER

Project I: Predictive Forecasting of Cancer Severity

Relatório Técnico apresentado à disciplina de Tópicos Especiais em Física Experimental (FIS0006), do curso de Física Aplicada do Instituto Latino-Americano de Ciências da Vida e da Natureza (ILACVN).

Orientador: Prof. Dr. Joylan Nunes Maciel

FOZ DO IGUAÇU
2025

RESUMO

Este trabalho apresenta o desenvolvimento de um modelo de *Machine Learning* para prever a pontuação de severidade do câncer com base em dados multifatoriais de pacientes. Utilizando o algoritmo *RandomForestRegressor*, um *pipeline* de pré-processamento foi construído, incluindo a normalização de dados numéricos e a codificação de variáveis categóricas. O modelo foi treinado, avaliado e otimizado através da técnica de *RandomizedSearchCV*. O modelo final otimizado alcançou um coeficiente de determinação (R^2) de 0.979, demonstrando alta capacidade preditiva e validando a eficácia da abordagem para apoiar a tomada de decisão clínica.

Palavras-chave: Aprendizado de Máquina; Regressão; Random Forest; Previsão de Câncer; Ciência de Dados.

ABSTRACT

This work presents the development of a *Machine Learning* model to predict cancer severity scores based on multifactorial patient data. Using the *RandomForestRegressor* algorithm, a preprocessing *pipeline* was built, including the normalization of numerical data and the encoding of categorical variables. The model was trained, evaluated, and optimized using the *RandomizedSearchCV* technique. The final optimized model achieved a coefficient of determination (R^2) of 0.979, demonstrating high predictive capability and validating the effectiveness of the approach in supporting clinical decision-making.

Keywords: Machine Learning; Regression; Random Forest; Cancer Prediction; Data Science.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTINGS

SUMÁRIO

1 INTRODUÇÃO

O câncer representa um dos desafios mais proeminentes para a saúde pública global, mantendo-se como uma das principais causas de morbidade e mortalidade em todo o mundo, com uma estimativa de 20 milhões de novos casos e 9,7 milhões de óbitos apenas em 2022 (SUNG et al., 2024). O avanço no entendimento da biologia do câncer tem revelado sua natureza heterogênea, onde a progressão e o prognóstico da doença são influenciados por uma complexa interação de fatores demográficos, genéticos, ambientais e de estilo de vida. Essa multifatoriedade torna a avaliação da severidade de cada caso uma tarefa extremamente desafiadora para a prática clínica, limitando a capacidade de personalização dos tratamentos e de alocação eficiente dos recursos de saúde.

Diante dessa complexidade, as abordagens tradicionais de análise podem ser insuficientes para capturar os padrões não lineares e as interdependências sutis entre as variáveis preditoras. É neste cenário que a Inteligência Artificial, especificamente o Aprendizado de Máquina (*Machine Learning*), surge como uma ferramenta poderosa. Modelos de *Machine Learning* são capazes de aprender a partir de grandes volumes de dados, identificando padrões complexos e gerando previsões que podem apoiar de forma significativa a tomada de decisão clínica.

Este trabalho se propõe a explorar essa oportunidade, utilizando um algoritmo clássico de *Machine Learning* para desenvolver um modelo preditivo robusto. A questão central que norteia esta pesquisa é: é possível, utilizando um conjunto de dados multifatoriais de pacientes, construir um modelo de *Machine Learning* com alta acurácia para prever a pontuação de severidade do câncer?

O objetivo principal deste projeto é, portanto, desenvolver, treinar e avaliar um modelo de regressão para prever a pontuação de severidade do câncer (`Target_Severity_Score`) com base em características de pacientes. Um modelo bem-sucedido tem o potencial de oferecer um impacto prático substancial, auxiliando na estratificação de risco dos pacientes, na otimização de planos terapêuticos e na gestão estratégica de recursos hospitalares, contribuindo assim para o avanço da medicina de precisão.

Para atingir tal objetivo, este relatório está estruturado da seguinte forma: a Fundamentação Teórica detalhará os conceitos de *Machine Learning* e os algoritmos utilizados; a seção de Desenvolvimento descreverá o *pipeline* completo, desde a coleta de dados até a implementação do modelo; os Resultados apresentarão as métricas de performance; e, por fim, a Conclusão discutirá os achados, desafios e direções futuras do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os conceitos fundamentais de Aprendizado de Máquina que foram aplicados no desenvolvimento deste projeto, incluindo a definição da tarefa de regressão, o funcionamento do algoritmo *Random Forest*, as métricas de avaliação e a metodologia de otimização de hiperparâmetros.

2.1 Aprendizado de Máquina Supervisionado e Regressão

O *Machine Learning* é um subcampo da Inteligência Artificial que se dedica ao desenvolvimento e estudo de algoritmos capazes de aprender a partir de dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Dentre suas vertentes, o Aprendizado Supervisionado é utilizado quando se dispõe de um conjunto de dados rotulado, ou seja, um conjunto de exemplos de entrada (*features*) acompanhados de suas respectivas saídas corretas (alvo ou *target*).

O objetivo de um modelo supervisionado é aprender uma função de mapeamento f que possa, a partir de um novo vetor de *features* X, prever sua saída $\hat{y}$ da forma mais acurada possível, tal que $\hat{y} \approx f(X)$. O presente projeto se enquadra como uma tarefa de regressão, pois o objetivo é prever a `Target_Severity_Score`, uma variável contínua que quantifica a severidade do câncer.

2.2 Random Forest Regressor

O algoritmo escolhido para este trabalho foi o *Random Forest* (Floresta Aleatória), um dos modelos de *ensemble* mais robustos e populares, conforme descrito por Breiman Breiman (2001). Suas unidades fundamentais são as Árvores de Decisão, que são combinadas para mitigar o problema de sobreajuste (*overfitting*) através de técnicas como *Bagging* e *Random Subspace Method*. A previsão final de um `RandomForestRegressor` é a média das previsões de todas as árvores individuais que compõem a floresta.

2.3 Métricas de Avaliação para Regressão

Para avaliar quantitativamente o desempenho do modelo de regressão, foram utilizadas as seguintes métricas, onde y_i é o valor real, $\hat{y}_i$ é o valor previsto e $\bar{y}$ é a média dos valores reais:

- *Mean Absolute Error* (MAE): Mede a média dos erros absolutos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- *Root Mean Squared Error* (RMSE): É a raiz quadrada do erro quadrático médio, penalizando erros maiores de forma mais significativa.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- *R-squared* (R^2): O Coeficiente de Determinação mede a proporção da variância no alvo que é previsível a partir das *features*.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2.4 Otimização de Hiperparâmetros

Os hiperparâmetros são configurações externas de um modelo definidas antes do treinamento. Neste projeto, foi utilizada a técnica *RandomizedSearchCV*, disponível na biblioteca *Scikit-learn* (PEDREGOSA et al., 2011), para buscar de forma eficiente a combinação ótima de hiperparâmetros.

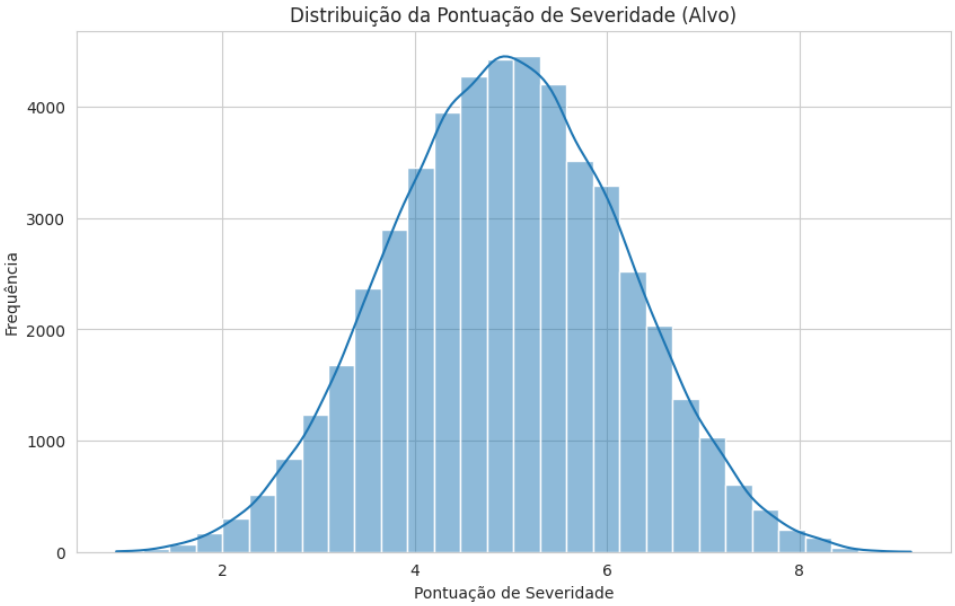
3 DESENVOLVIMENTO

Esta seção detalha o processo prático de construção do modelo, desde a configuração do ambiente até o treinamento e a otimização do algoritmo, implementados em Python.

3.1 Coleta e Análise Exploratória dos Dados

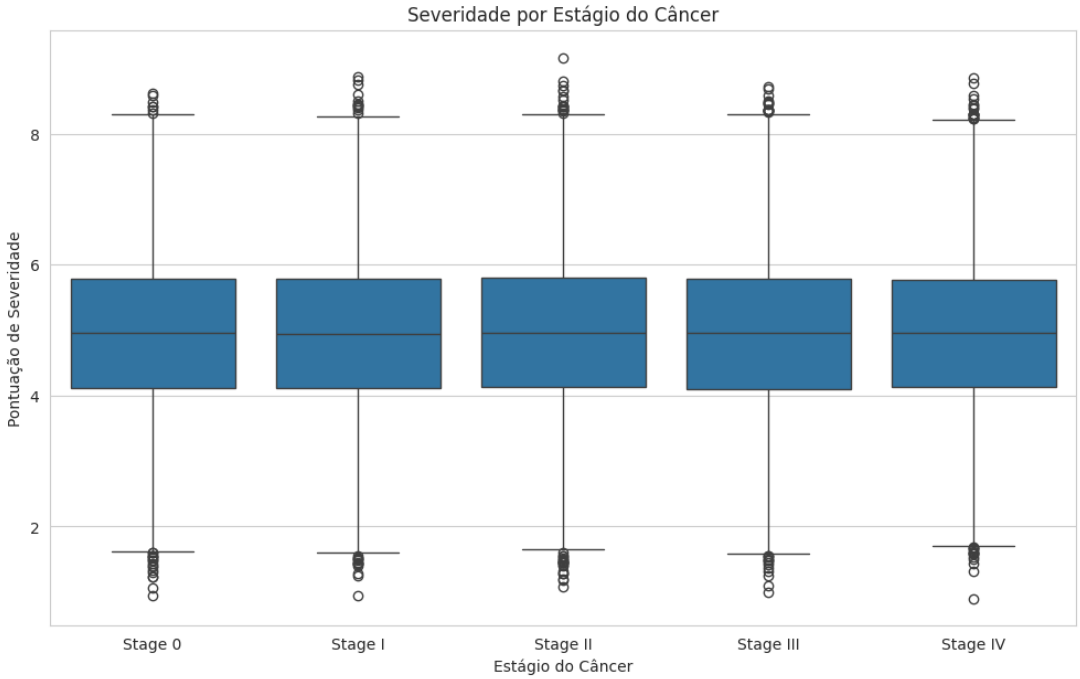
O ponto de partida foi a utilização de um conjunto de dados público (MUGHAL, 2025). É fundamental ressaltar que se trata de um *dataset* sintético, gerado para simular tendências globais de câncer com base em relatórios de saúde pública, não contendo informações de pacientes reais. Após a coleta a partir do arquivo `global_cancer_patients_2015_2024.csv` e a remoção do identificador `Patient_ID`, a Análise Exploratória de Dados (EDA) revelou uma forte correlação positiva entre o estágio do câncer e a severidade, conforme ilustrado na Figura 2. Um desafio técnico identificado foi a presença de numerais romanos na coluna `Cancer_Stage`, o que exigiu uma função de conversão para permitir a correta ordenação visual. A Figura 3 apresenta a matriz de correlação, que apontou outras relações lineares de interesse.

Figura 1 – Distribuição da variável alvo (Target Severity Score).



Fonte: Autor (2025).

Figura 2 – Relação entre o Estágio do Câncer e a Pontuação de Severidade.



Fonte: Autor (2025).

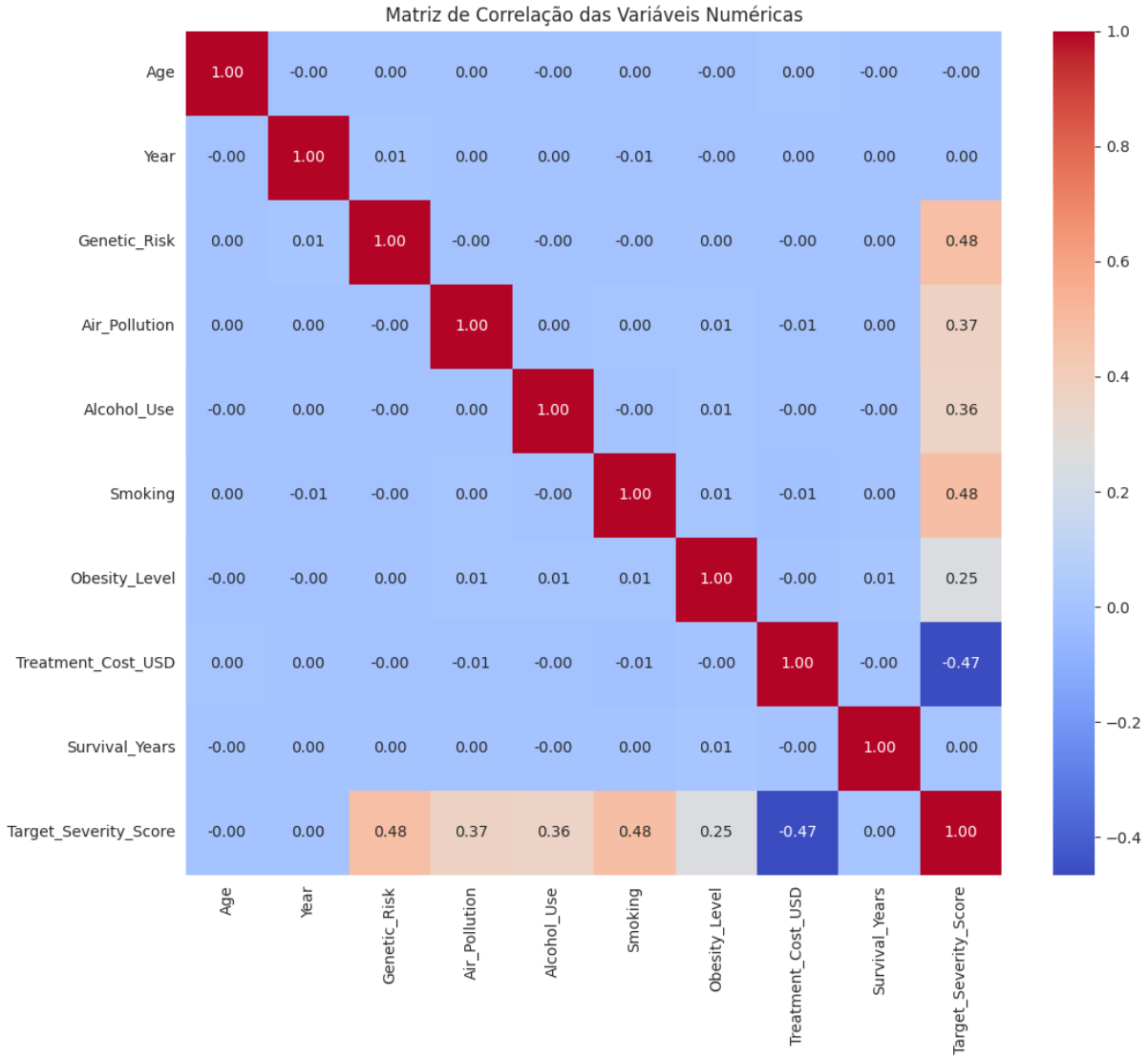
3.2 Pré-processamento e Construção do *Pipeline*

Para garantir a robustez e a reprodutibilidade, foi construído um *Pipeline* com um *ColumnTransformer*, aplicando `StandardScaler` às *features* numéricas e `OneHotEncoder` às categóricas. Esta abordagem previne o vazamento de dados e encapsula todo o fluxo de transformação.

3.3 Treinamento e Otimização do Modelo

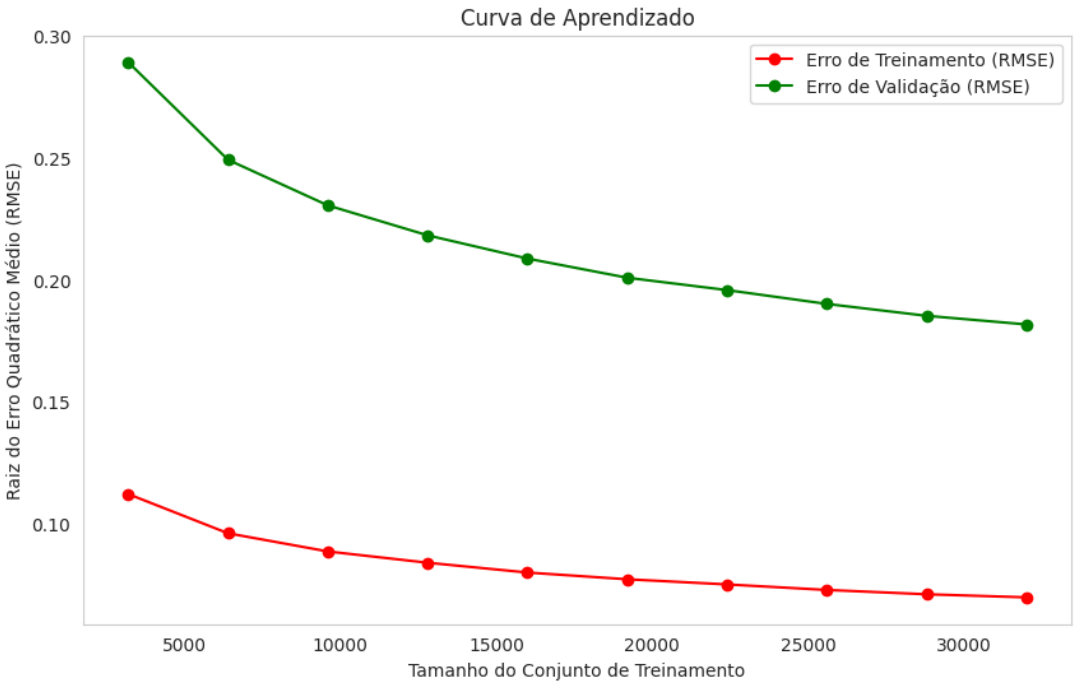
O conjunto de dados foi dividido em 80% para treinamento e 20% para teste. O `RandomForestRegressor` foi treinado primeiramente com seus hiperparâmetros padrão (`n_estimators: 100`; `max_depth: None`; `min_samples_split: 2`; `min_samples_leaf: 1`). A Figura 4 ilustra a curva de aprendizado do modelo base. Subsequentemente, a técnica `RandomizedSearchCV` foi utilizada para otimizar os hiperparâmetros, resultando em um modelo final ajustado.

Figura 3 – Matriz de correlação entre as variáveis numéricas do dataset.



Fonte: Autor (2025).

Figura 4 – Curva de aprendizado do modelo base.



Fonte: Autor (2025).

4 RESULTADOS

Neste capítulo, são apresentados os resultados quantitativos da performance dos modelos.

4.1 Desempenho dos Modelos

O modelo inicial (base) foi avaliado no conjunto de teste para estabelecer uma linha de base de performance. Após o processo de otimização, o modelo com os melhores hiperparâmetros (Listagem 4.1) foi igualmente avaliado.

```
{
    'regressor__n_estimators': 300,
    'regressor__min_samples_split': 2,
    'regressor__min_samples_leaf': 1,
    'regressor__max_depth': None
}
```

Listing 4.1 – Melhores hiperparâmetros encontrados pela busca aleatória.

4.2 Análise Comparativa

A Tabela 1 apresenta uma comparação direta das métricas de performance entre o modelo base e o modelo otimizado. A Figura 5 ilustra a precisão do modelo otimizado.

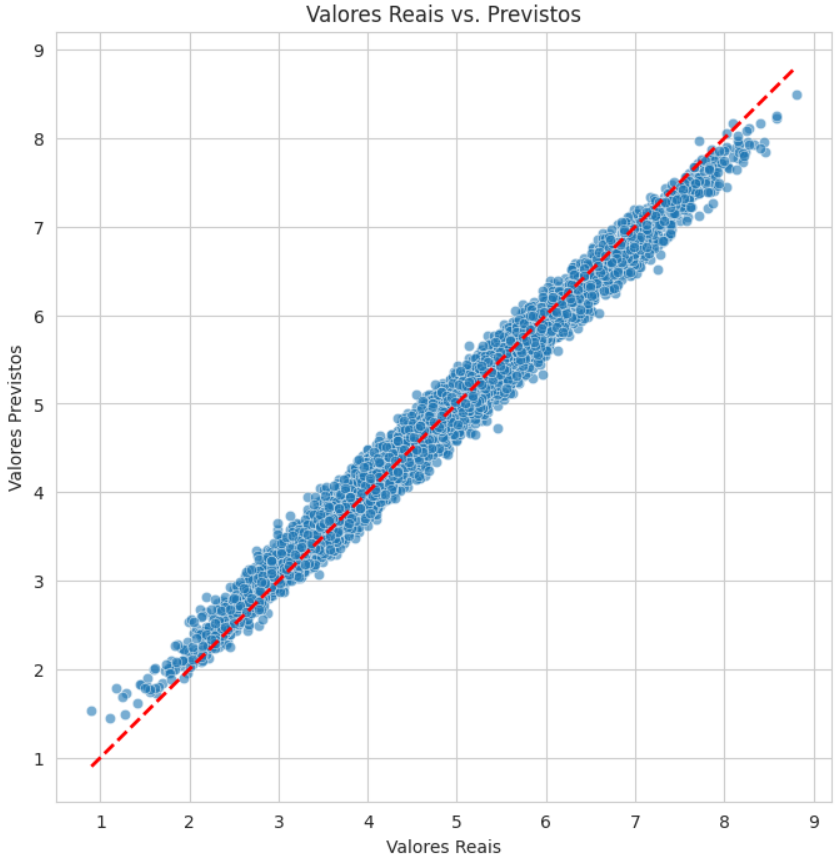
Tabela 1 – Comparação de desempenho entre o modelo base e o otimizado.

Métrica	Modelo Base	Modelo Otimizado
Mean Absolute Error (MAE)	0.136	0.136
Root Mean Squared Error (RMSE)	0.172	0.171
R-squared (R^2)	0.979	0.979

Fonte: Autor (2025).

A análise demonstra que o processo de otimização resultou em uma melhoria marginal, quase insignificante. O coeficiente de determinação (R^2) permaneceu em 0.979. Isso indica que o modelo base, com seus hiperparâmetros padrão, já era extremamente eficaz para este conjunto de dados, atingindo um desempenho próximo do ideal e deixando pouco espaço para melhorias incrementais através do ajuste fino.

Figura 5 – Gráfico de dispersão dos valores reais vs. previstos para o modelo otimizado.



Fonte: Autor (2025).

5 CONCLUSÃO

Este trabalho desenvolveu e avaliou com sucesso uma solução de Inteligência Artificial para prever a severidade do câncer. O modelo `RandomForestRegressor` otimizado atingiu um coeficiente de determinação (R^2) de 0.979, demonstrando uma capacidade preditiva excepcional e respondendo afirmativamente à pergunta de pesquisa.

O sucesso do modelo reforça o potencial do *Machine Learning* como uma ferramenta de apoio à decisão na oncologia, contribuindo para a medicina de precisão. Um desafio técnico notável foi a manipulação de dados não padronizados (numerais romanos), superado com a criação de uma função de conversão, o que ressalta a importância da etapa de pré-processamento.

A principal limitação do trabalho reside no uso de dados sintéticos, o que impede a generalização direta do modelo para populações de pacientes reais sem uma validação prévia com dados clínicos autênticos. Como trabalhos futuros, sugere-se a exploração de outros algoritmos de *ensemble*, como o *Gradient Boosting* (FRIEDMAN, 2001), e uma análise aprofundada da importância das *features* para gerar *insights* clínicos. A implantação do modelo como uma *API* seria o passo subsequente para sua validação em um ambiente real.

Em suma, o projeto entregou uma ferramenta preditiva de alto desempenho e demonstrou o potencial transformador do *Machine Learning* no cuidado quantitativo e personalizado ao paciente com câncer.

REFERÊNCIAS

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

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

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

MUGHAL, Z. **Global Cancer Patients (2015-2024)**. Kaggle, 2025. `https://www.kaggle.com/datasets/zahidmughal2343/global-cancer-patients-2015-2024`. Acesso em: 13 jun. 2025.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

SUNG, H. et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, v. 74, n. 3, p. 209–249, 2024.