UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
DEPARTAMENTO DE INFORMÁTICA APLICADA
CURSO DE ESPECIALIZAÇÃO EM ENGENHARIA DE SOFTWARE PARA APLICAÇÕES DE CIÊNCIA DE DADOS

GRACE DOS SANTOS FEIJÓ

Explorando o aprendizado de máquina explicável para predizer fatores genéticos relacionados à sobrevivência ao câncer a partir de dados transcriptômicos

Monografia de Conclusão de Curso apresentada como requisito parcial para a obtenção do grau de Especialista em Ciência de Dados

Orientador: Prof. Dra. Mariana Recamonde Mendoza

Porto Alegre
2026

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Feijó, Grace dos Santos

Explorando o aprendizado de máquina explicável para predizer fatores genéticos relacionados à sobrevivência ao câncer a partir de dados transcriptômicos / Grace dos Santos Feijó. – Porto Alegre: PPGC da UFRGS, 2026.

66 f.: il.

Monografia (especialização) – Universidade Federal do Rio Grande do Sul. Curso de Especialização em Engenharia de Software para Aplicações de Ciência de Dados, Porto Alegre, BR–RS, 2026. Orientador: Mariana Recamonde Mendoza.

1. Carcinoma de células claras. 2. Carcinoma papilífero. 3. Biomarcadores prognósticos. 4. RNA-seq. I. Mendoza, Mariana Recamonde. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitora: Prof[a]. Marcia Barbosa
Vice-Reitor: Prof. Pedro de Almeida Costa
Pró-Reitora de Pós-Graduação: Prof[a]. Claudia Wasserman
Diretor do Instituto de Informática: Prof. Luciano Paschoal Gaspary
Coordenadores da Especialização: Prof[a]. Dra. Karin Becker e Prof. Dr. Marcelo Pimenta
Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro

“Nada na vida deve ser temido, apenas compreendido. Agora é hora de compreender mais, para temer menos.”

— MARIE CURIE

AGRADECIMENTOS

A oportunidade de fazer uma pós-graduação em uma área tão diferente da minha formação anterior foi uma experiência incrível. No entanto, nada disso seria possível sem o incansável suporte do meu marido, Erick Puls. Nesses dois anos de estudo nossa vida virou de cabeça para baixo e até demos as boas-vindas ao nosso amado filhote, Dante. Agradeço de coração cada abraço, cada palavra de motivação e toda a parceria nesse período. Agradeço também às nossas famílias que, mesmo de longe, sempre se fizeram presentes, e aos nossos amigos que acabam sendo nossa segunda família longe de casa.

Não menos importante, agradeço à minha orientadora Prof. Dra Mariana Recamonde Mendoza que aceitou me orientar apesar da minha falta de experiência acadêmica sobre o assunto, agradeço de verdade pela paciência e confiança no meu trabalho. Muito muito obrigada!

RESUMO

O uso de técnicas de aprendizado de máquina na pesquisa biomédica tem possibilitado grandes avanços em saúde. Ao permitir a análise de grande quantidade de dados, viabiliza diagnósticos e prognósticos mais precisos e escolhas terapêuticas adaptadas às características individuais de cada paciente. Nesse sentido, o presente estudo teve como objetivo identificar assinaturas gênicas associadas à sobrevivência de pacientes com carcinoma renal dos subtipos KIRC e KIRP por meio da integração de dados transcriptômicos do TCGA e de modelos de aprendizado de máquina (ML) aliados a técnicas de inteligência artificial explicável (XAI). Foram utilizados cinco algoritmos para a predição: *Random Forest*, *Gradient Boosting*, *Logistic Regression*, *MLP* e *CatBoost* Os modelos alcançaram desempenho satisfatório, especialmente para o subtipo KIRC, com destaque para o CatBoost (AUC-ROC = 0.82; C-index = 0.82; acurácia = 0.78). Em KIRP, observou-se maior variabilidade, atribuída ao baixo número de amostras positivas e ao desbalanceamento entre classes, embora o *Random Forest* e o *CatBoost* tenham apresentado resultados promissores. As técnicas de XAI utilizadas, SHAP, PFI e Árvores de Decisão Surrogadas, permitiram identificar como genes de maior importância preditiva em KIRC o *IL20RB*, *SLFN11* e *LINC00839*, e em KIRP o *TOP2A*, *PDIA4* e *ACAA1*. Os resultados reforçam as diferenças moleculares entre os subtipos e apontam potenciais biomarcadores relacionados à resposta inflamatória, replicação celular e resistência terapêutica.

Palavras-chave: Carcinoma de células claras. Carcinoma papilífero. Biomarcadores prognósticos. RNA-seq.

Exploring explainable machine learning to predict genetic factors related to cancer survival from transcriptomic data

ABSTRACT

The application of machine learning techniques in biomedical research has driven significant advancements in healthcare. By enabling the analysis of large datasets, these methods facilitate more precise diagnoses and prognoses, as well as therapeutic choices tailored to individual patient characteristics. In this context, the present study aimed to identify gene signatures associated with the survival of patients with renal cell carcinoma — specifically the clear cell (KIRC) and papillary (KIRP) subtypes — through the integration of transcriptomic data from the TCGA with machine learning (ML) models and explainable artificial intelligence (XAI) techniques. Five predictive algorithms were employed: Random Forest, Gradient Boosting, Logistic Regression, MLP, and CatBoost. The models achieved satisfactory performance, particularly for the KIRC subtype, with CatBoost as the standout performer (AUC-ROC = 0.82; C-index = 0.82; accuracy = 0.78). For KIRP, greater variability was observed, attributed to the low number of positive samples and class imbalance, although Random Forest and CatBoost showed promising results. XAI techniques — including SHAP, PFI, and Surrogate Decision Trees — enabled the identification of genes with the highest predictive importance: *IL20RB*, *SLFN11*, and *LINC00839* for KIRC, and *TOP2A*, *PDIA4*, and *ACAA1* for KIRP. The findings reinforce the molecular distinctions between the subtypes and highlight potential biomarkers related to inflammatory response, cellular replication, and therapeutic resistance.

Keywords: Clear cell renal cell carcinoma, Papillary renal cell carcinoma, Prognostic biomarkers, RNA-seq.

LISTA DE ABREVIATURAS E SIGLAS

ADS	Árvore de Decisão Surrogada
CatBoost	Categorical Boosting
DNNs	Deep Neural Networks
DR	Redução de Dimensionalidade
GIRP	Global Interpretation via Recursive Partitioning
IQR	Intervalo Interquartil
KIRC	Carcinoma de Células Claras
KIRP	Carcinoma Capilífero
LR	Regressão Logística
ML	Aprendizado de Máquina
MLPs	Multi-Layer Perceptrons
mRNA	RNA Mensageiro
ncRNA	RNA Não Codificante
PCA	Análise de Componentes Principais
PFI	Permutation Feature Importance
RF	Random Forest
RNA-seq	Sequenciamento de RNA
SHAP	SHapley Additive exPlanations
t-SNE	t-Distributed Stochastic Neighbor Embedding
TCGA	The Cancer Genome Atlas
UMAP	Uniform Manifold Approximation and Projection
XAI	Inteligência Artificial Explicável

LISTA DE FIGURAS

Figura 2.1 Ilustração da distribuição dos dados na validação cruzada estratificada para os conjuntos de treino e validação. 18
Figura 2.2 Área sob a curva, AUC-ROC. 19
Figura 2.3 Área sob a curva, AUC-PR. AUC = 1 indica um resultado perfeito, enquanto AUC = 0.5 indica uma performance igual ao acaso....20

Figura 4.1 Fluxograma das etapas metodológicas do estudo. 27

Figura 5.1 Processo de ajuste de dimensionalidade dos conjuntos de dados KIRC e KIRP. 31
Figura 5.2 Gráficos da performance dos modelos na validação cruzada para KIRC. 32
Figura 5.3 Gráficos da performance dos modelos na validação cruzada para KIRC. 33
Figura 5.4 Gráficos da performance dos modelos na validação cruzada para KIRP. 35
Figura 5.5 Gráficos da performance dos modelos na validação cruzada para KIRP. 36
Figura 5.6 Análise do desempenho dos modelos no conjunto de dados de teste - KIRC38
Figura 5.7 Gráficos das curvas ROC e PR para todos os modelos avaliados - KIRC 39
Figura 5.8 Gráficos da performance dos modelos no conjunto de dados de teste - KIRP 40
Figura 5.9 Gráficos das curvas ROC e PR para todos os modelos avaliados - KIRP. 41
Figura 5.10 Descrição dos genes de maior importância para KIRC identificados pelas análises de XAI. Apresenta os principais genes associados para cada subtipo de carcinoma renal, indicando se codificam proteínas, e suas funções biológicas conhecidas. 42
Figura 5.11 Descrição dos genes de maior importância para KIRP identificados pelas análises de XAI. Apresenta os principais genes associados para cada subtipo de carcinoma renal, indicando se codificam proteínas, e suas funções biológicas conhecidas 43
Figura 5.12 Distribuição dos valores SHAP dos genes de maior importância para KIRC e KIRP. Cada ponto representa uma amostra, com a posição no eixo X indicando a contribuição do gene para o modelo preditivo. Pontos à direita estão associados a pior prognóstico. As cores refletem os níveis de expressão gênica 44
Figura 5.13 Genes presentes nos nodos principais e secundários das árvores de decisão para cada modelo de ML 46

Figura A.1 Distribuição das amostras após redução de dimensionalidade com PCA, UMAP e t-SNE - KIRC 54
Figura A.2 Distribuição das amostras após redução de dimensionalidade com PCA, UMAP e t-SNE - KIRP....55

Figura B.1 Resultados SHAP nos modelos Randon Forest e Gradient Boosting - KIRC56
Figura B.2 Resultados SHAP no modelo Multilayer Perceptron - KIRC....57
Figura B.3 Resultados SHAP nos modelos Regressão Logística e Categorical Boosting - KIRC....58
Figura B.4 Resultados SHAP nos modelos Randon Forest e Gradient Boosting - KIRP59
Figura B.5 Resultados SHAP no modelo Multilayer Perceptron - KIRP 60
Figura B.6 Resultados SHAP nos modelos Regressão Logística e Categorical Boosting - KIRP 61
Figura B.7 Resultados ADS para o conjunto de dados KIRC 64

Figura B.8 Resultados ADS para o conjunto de dados KIRP.....65

LISTA DE TABELAS

Tabela 4.1 Hiperparâmetros utilizados no teste 29

Tabela 5.1 Distribuição das amostras e número de variáveis após redução de dimensionalidade.....31
Tabela 5.2 Resultados do teste dos modelos no conjunto de dados KIRC. Os valores mais altos em cada métrica estão marcados em verde e os mais baixos em rosa. 37
Tabela 5.3 Resultados do teste dos modelos no conjunto de dados KIRP. Os valores mais altos em cada métrica estão marcados em verde e os mais baixos em rosa. ...39
Tabela 5.4 Resultados da análise com o modelo CatBoost utilizado apenas os genes de maior relevância. 48

Tabela B.1 Resultados das análises de PFI para o conjunto de dados KIRC.....62
Tabela B.2 Resultados das análises de PFI para o conjunto de dados KIRP 62

Tabela C.1 Frequência, por método de XAI, dos 30 genes mais encontrados. 66

SUMÁRIO

1 INTRODUÇÃO **12**
1.1 Objetivo principal **13**
1.1.1 Objetivos específicos 13
2 REFERENCIAL TEÓRICO **14**
2.1 O transcriptoma e biomarcadores na oncologia **14**
2.2 Algoritmos de aprendizado de máquina **15**
2.3 Técnicas de explicabilidade **16**
2.4 Redução de dimensionalidade **17**
2.5 Validação cruzada estratificada e ajuste de hiperparâmetros **18**
2.6 Métricas de avaliação de desempenho **19**
3 REVISÃO DE LITERATURA **22**
3.1 Aprendizado de máquina na saúde **22**
3.2 Aprendizado de máquina aplicado na oncologia e o uso de transcriptômica **23**
3.3 Carcinoma de células claras renais e o carcinoma papilífero renal **23**
3.4 Métodos de ML aplicados na transcriptômica **25**
3.5 Métodos de explicabilidade de modelos **26**
4 METODOLOGÍA **27**
4.1 Coleta e pré-processamento dos dados **27**
4.2 Construção dos modelos de aprendizado de máquina **28**
4.3 Avaliação de desempenho **28**
4.4 Aplicação de técnicas de explicabilidade **30**
4.5 Análise final dos resultados **30**
5 RESULTADOS E DISCUSSÃO **31**
5.1 Redução de dimensionalidade e divisão dos dados de treino e teste **31**
5.2 Desempenho preditivo na validação cruzada **32**
5.2.1 Análise do desempenho na validação cruzada para KIRC 32
5.2.2 Análise da validação cruzada estratificada para KIRP 34
5.3 Desempenho dos modelos de ML no conjunto de teste **37**
5.3.1 Análise do desempenho dos modelos – KIRC 37
5.3.2 Análise do desempenho dos modelos - KIRP 39
5.4 Métodos de explicabilidade **41**
5.4.1 Resultados explicabilidade - KIRC 45
5.4.2 Resultados explicabilidade - KIRP 46
5.5 CatBoost aplicado aos genes de maior relevância **48**
6 CONCLUSÃO **49**
REFERÊNCIAS **50**
APÊNDICE A — RESULTADOS DAS TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE - PCA, UMAP E T-SNE **54**
APÊNDICE B — RESULTADOS COMPLETOS SHAP, PFI, ADS **56**
B.1 SHAP **56**
B.2 PFI **62**
B.3 ADS **63**
APÊNDICE C — FREQUÊNCIA DE GENES POR MÉTODO XAI - TABELA **66**

1 INTRODUÇÃO

O avanço das tecnologias de sequenciamento e o crescimento exponencial de dados biomédicos disponíveis têm impulsionado a aplicação de técnicas de aprendizado de máquina, do inglês, *Machine Learning* (ML) na área da saúde, especialmente na medicina personalizada (GHASSEMI et al., 2020). Ao permitir a análise integrada de informações genômicas, clínicas e ambientais, o ML viabiliza diagnósticos mais precisos, prognósticos mais confiáveis e escolhas terapêuticas adaptadas às características individuais de cada paciente (SHAH; LODHI, 2025).

Entre as fontes de dados mais promissoras para esse tipo de abordagem está a transcriptômica, que possibilita explorar alterações na expressão gênica associadas a processos patológicos complexos, como o câncer (RECHARLA et al., 2025; TRAN et al., 2025). Na oncologia, a análise transcriptômica combinada a modelos de ML tem se mostrado particularmente relevante para compreender a heterogeneidade tumoral, identificar biomarcadores prognósticos e apoiar a estratificação de risco (TRAN et al., 2025).

No carcinoma renal, subtipos como o carcinoma de células claras (KIRC) e o carcinoma papilífero (KIRP) apresentam diferenças clínicas e moleculares significativas, com impacto direto no manejo terapêutico e na sobrevida (TABIBU; VINOD; JAWAHAR, 2019). No entanto, apesar dos avanços, ainda há lacunas na identificação de assinaturas gênicas robustas que diferenciem esses subtipos e que possibilitem uma interpretação biológica clara das decisões algorítmicas (ALBERTSON et al., 2025; SU et al., 2024b).

Embora os métodos de ML apresentem alto desempenho preditivo, a falta de explicabilidade limita sua adoção em contextos clínicos. Nesse sentido, técnicas de Inteligência Artificial Explicável (XAI), emergem como ferramentas fundamentais para compreender os mecanismos internos desses modelos, identificando genes e padrões de expressão mais relevantes para os desfechos clínicos.

Diante desse cenário, este estudo propõe a construção de modelos de ML aplicados a dados transcriptômicos de KIRC e KIRP, aliados a técnicas de inteligência artificial explicável (XAI), como SHAP (*SHapley Additive exPlanations*), PFI (*Permutation Feature Importance*) e GIRP (*Global Interpretation via Recursive Partitioning*). Essa combinação visa não apenas alcançar um bom desempenho preditivo, mas também oferecer interpretações biologicamente significativas que contribuam para o avanço da medicina de precisão no câncer renal.

O trabalho está organizado da seguinte forma: o Capítulo 2 expõe o referencial

teórico, aprofundando os conceitos fundamentais que sustentam a construção e a interpretação dos experimentos realizados. O Capítulo 3 apresenta a revisão de literatura, que reúne e discute estudos anteriores. O Capítulo 4 descreve a metodologia adotada, detalhando as etapas de coleta, processamento e análise dos dados. No Capítulo 5, são apresentados os principais resultados obtidos e a discussão desses achados com base na literatura científica. Por fim, o Capítulo 6 traz as conclusões e limitações do trabalho, e aponta perspectivas futuras para o avanço das pesquisas na área.

1.1 Objetivo principal

Dessa forma, o objetivo principal do estudo se dá em identificar assinaturas gênicas associadas à não sobrevivência de pacientes com carcinoma renal (KIRC e KIRP), por meio de modelos de ML e técnicas de explicabilidade de modelos, visando à descoberta de biomarcadores prognósticos e de auxílio à personalização terapêutica.

1.1.1 Objetivos específicos

- Implementar modelos preditivos baseados em ML para a estimativa de desfechos clínicos (sobrevivência) via dados transcriptômicos.
- Identificar genes candidatos a biomarcadores prognósticos por meio da aplicação de técnicas de XAI.
- Avaliar a significância biológica dos genes identificados, correlacionando-os com vias metabólicas ou processos celulares críticos na progressão tumoral.

2 REFERENCIAL TEÓRICO

No presente capítulo, discutimos o valor do estudo de biomarcadores e assinaturas genéticas usando o transcriptoma, os algoritmos de ML na classificação da sobrevivência, o uso dos métodos de explicabilidade como estratégia para o avanço da área e, finalmente, as métricas e técnicas utilizadas no treino e avaliação dos modelos.

2.1 O transcriptoma e biomarcadores na oncologia

O transcriptoma, representando o conjunto completo de transcritos de RNA em uma célula, tecido ou organismo em um momento específico, é fundamental para entender a regulação gênica e a função biológica em sistemas complexos como o câncer (DIVATE et al., 2022). A análise transcriptômica, frequentemente realizada por sequenciamento de RNA (RNA-seq), permite a quantificação da expressão de RNA mensageiro (mRNA) e RNA não codificante (ncRNA), fornecendo uma visão do estado funcional de um sistema biológico (DIVATE et al., 2022; LIU et al., 2025). Em oncologia, a identificação de assinaturas de expressão gênica distintas entre tumores e tecidos normais é vital, pois essas assinaturas podem delinear a origem tecidual do tumor, diferenciar subtipos moleculares com prognósticos variados, e apontar alvos terapêuticos, como demonstrado em estudos pan-câncer que alcançaram precisão de classificação superior a 97% (DIVATE et al., 2022).

Biomarcadores são características mensuráveis capazes de indicar processos biológicos normais, processos patogênicos ou de resposta farmacológica a uma intervenção terapêutica (TRAN et al., 2025). São ferramentas cruciais na medicina de precisão, pois permitem estratificar pacientes e guiar decisões terapêuticas com base em características moleculares intrínsecas. Um exemplo de gene biomarcador, onde a expressão diferencial prediz a sensibilidade a um tratamento específico, é o *SLFN11*, que prevê a sensibilidade a fármacos que visam o DNA (ex: inibidores de PARP ou agentes de platina) em vários tipos de câncer (ZHOU et al., 2025). Assinaturas genéticas, por sua vez, consistem em painéis de múltiplos genes ou perfis de expressão que, combinados, oferecem maior poder preditivo e prognóstico do que um único marcador, como visto na classificação de subtipos de câncer de mama ou na estratificação de risco no Câncer de Cabeça e Pescoço (NAGY; MUNKÁCSY; GYŐRFFY, 2021; MIAO et al., 2024). A integração dessas informações moleculares com dados clínicos e o uso de métodos avançados de ML, fre-

quentemente auxiliados por XAI, tornam possível construir modelos robustos para prever sobrevida e resposta à imunoterapia, como demonstrado na análise de assinaturas relacionadas a *hallmarks* do câncer ou a vias específicas como a autofagia (TRAN et al., 2025; SU et al., 2024a; MIAO et al., 2024).

2.2 Algoritmos de aprendizado de máquina

A literatura demonstra que a comparação entre diferentes algoritmos de ML é fundamental para identificar a abordagem mais adequada a uma base de dados específica, dado que cada modelo opera sob princípios distintos (PEDUZZI et al., 2025; LI et al., 2024; ALABI et al., 2023; HAMEY; GÖTTGENS, 2019; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Este estudo utiliza cinco classificadores, cujas características são descritas a seguir.

Random Forest: Método de aprendizagem de conjunto formado por múltiplos classificadores (ou regressões) do tipo árvore de decisão. Cada árvore é construída a partir de uma amostra bootstrap dos dados de treinamento e, a cada divisão, apenas um subconjunto aleatório de variáveis é considerado. A predição final é obtida por votação majoritária (classificação) ou pela média das previsões (regressão) entre todas as árvores. O método lida bem com alta dimensionalidade sem exigir seleção explícita de variáveis, é robusto a ruídos de medição e fornece estimativas internas de importância das variáveis, facilitando a identificação de potenciais biomarcadores (PEDUZZI et al., 2025).

Gradient Boosting: Constrói árvores de forma sequencial, em que cada nova árvore é treinada para corrigir os erros residuais do conjunto acumulado até o momento. Captura interações complexas e não lineares entre genes, funciona bem com matrizes esparsas e de alta dimensionalidade e frequentemente atinge maior acurácia preditiva do que métodos baseados em uma única árvore, especialmente em cenários com desbalanceamento entre classes (ALABI et al., 2023).

Multilayer Perceptron (MLP): Rede neural artificial do tipo *feed-forward* composta por uma camada de entrada, uma ou mais camadas ocultas totalmente conectadas e uma camada de saída. Cada neurônio aplica uma combinação linear ponderada de suas entradas seguida de uma função de ativação não linear (como *ReLU* ou *sigmoide*). O treinamento ocorre por meio do algoritmo de retropropagação, que ajusta os pesos para minimizar uma função de perda. As camadas ocultas aprendem representações hierárquicas dos padrões de expressão gênica, permitindo ao modelo capturar relações sutis e não

lineares que podem ser ignoradas por métodos baseados em árvores. Com dados suficientes, MLPs podem alcançar desempenho muito elevado tanto em classificação quanto em regressão(HAMEY; GÖTTGENS, 2019).

Regressão Logística (LR): Modelo amplamente utilizado para prever desfechos binários, como sobrevivência positiva vs. negativa. Ela estima a probabilidade de um evento ocorrer a partir de uma combinação linear das variáveis de entrada. Diferentemente da regressão linear, aplica-se uma função logística (*sigmoide*) para transformar essa combinação em uma probabilidade entre 0 e 1. É um método simples, rápido e interpretável, servindo como uma boa linha de base para problemas de classificação em dados de alta dimensionalidade (LI et al., 2024).

Categorical Boosting (CatBoost): Biblioteca de *gradient boosting* que lida nativamente com variáveis categóricas e reduz o *shift* preditivo por meio do chamado ordered boosting. O método constrói árvores simétricas (oblivious trees), o que torna o modelo mais regularizado e com avaliação mais rápida. Para dados exclusivamente numéricos, como RNA-seq, seus benefícios incluem o tratamento eficiente de matrizes esparsas e de alta dimensionalidade, suporte nativo a valores ausentes e hiperparâmetros padrão robustos, que frequentemente exigem menos ajustes do que outros métodos de boosting (PEDUZZI et al., 2025).

2.3 Técnicas de explicabilidade

Métodos de XAI como SHAP, PFI e GIRP tornam modelos de ML mais transparentes e interpretáveis, especialmente em contextos biomédicos. O PFI é uma técnica de XAI que quantifica a importância de uma variável ao medir a redução média no desempenho do modelo quando seus valores são aleatoriamente permutados no conjunto de validação. Caso essa troca aleatória dos valores leve a uma queda substancial em métricas como acurácia ou AUC, essa variável é considerada essencial para o poder preditivo do modelo (KHAN et al., 2025).

Complementarmente, o SHAP é baseado na teoria dos valores de *Shapley* da teoria dos jogos cooperativos, o modelo atribui a cada variável uma contribuição quantitativa específica para determinada predição, levando em conta todas as combinações possíveis de entrada. Essa abordagem permite identificar quais genes tiveram maior peso na previsão de risco ou desfecho clínico (YAP et al., 2021; PEDUZZI et al., 2025).

Em relação ao GIRP, ele atua em um nível mais global, buscando interpretar mo-

delos complexos por meio da construção de árvores de decisão derivadas das regras internas aprendidas pelos algoritmos originais. Com isso, é possível extrair representações mais simples e compreensíveis, como divisões hierárquicas baseadas em expressões gênicas, que preservam boa parte da acurácia do modelo original, ao mesmo tempo em que tornam suas decisões mais acessíveis à compreensão humana (PEDUZZI et al., 2025; YANG; RANGARAJAN; RANKA, 2018). Em última análise, a aplicação de ferramentas de XAI assegura que os padrões identificados pelo ML não sejam apenas classificatórios, mas também ajudem na compreensão dos mecanismos biológicos subjacentes envolvidos no desfecho clínico.

2.4 Redução de dimensionalidade

Dados transcriptômicos, ao envolverem dezenas de milhares de genes, apresentam alta dimensionalidade, o que inviabiliza análises diretas e consome recursos computacionais significativos (YAP et al., 2021; ZHAO; SHAO; ASMANN, 2022). A redução de dimensionalidade (DR) é, portanto, uma etapa crucial no pré-processamento, visando mitigar o ruído inerente à medição gênica e aumentar a eficiência computacional (KHAN et al., 2025).

Além disso, a DR é fundamental para permitir a visualização de dados de alta dimensão em espaços 2D ou 3D, facilitando a identificação de agrupamentos (*clusters*) que refletem similaridades biológicas entre as amostras (YAP et al., 2021). Técnicas focadas em seleção de recursos (*feature selection*) são usadas para reter apenas as variáveis mais informativas, como aquelas que apresentam maior diferença (avaliadas pelo valor de p) entre os grupos de interesse (por exemplo, tumor vs. normal) (PAN et al., 2024; MAMATJAN, 2024).

As técnicas de DR buscam equilibrar a retenção de informação com a simplificação do conjunto de dados, lidando com desafios específicos como a colinearidade entre a expressão de genes relacionados funcionalmente (YANG; RANGARAJAN; RANKA, 2018). Métodos como a Análise de Componentes Principais (PCA) transformam linearmente os dados para maximizar a variância retida, sendo utilizados para inferir o número ótimo de características a serem selecionadas (SEO et al., 2025). Alternativamente, técnicas não lineares de redução, como t-SNE, são primariamente empregadas para a visualização de dados, projetando as amostras em um espaço de baixa dimensão para revelar a separação de classes, como observado na análise de agrupamento de diferentes tipos de

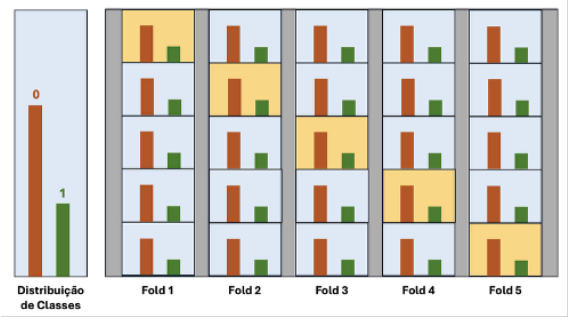
câncer (DIVATE et al., 2022). O uso de *feature selection* por filtros ou *wrappers* também é comum para eliminar variáveis redundantes ou não informativas, focando naquelas com maior impacto preditivo (SEO et al., 2025; KHAN et al., 2025).

Essas abordagens permitem uma representação mais compacta e informativa dos dados transcriptômicos, favorecendo tanto a interpretação biológica quanto o desempenho de métodos analíticos e preditivos. Assim, a escolha adequada da técnica de redução de dimensionalidade ou seleção de atributos é um fator determinante para o sucesso de análises transcriptômicas.

2.5 Validação cruzada estratificada e ajuste de hiperparâmetros

A validação cruzada permite estimar de forma mais confiável o desempenho preditivo, reduzindo o risco de sobreajuste (onde o modelo se adapta excessivamente aos dados de treinamento e perde capacidade preditiva em novos conjuntos). Em cada iteração, quatro folds são usados para treinamento e o quinto para validação; as métricas finais correspondem à média das cinco execuções. A estratificação assegura que a proporção das classes seja mantida em todos os *folds*, aspecto fundamental em análises genômicas marcadas por forte desbalanceamento entre grupos (DIVATE et al., 2022) 2.1

Figura 2.1: Ilustração da distribuição dos dados na validação cruzada estratificada para os conjuntos de treino e validação.



Fonte: Os Autores

Em relação ao ajuste de hiperparâmetros, a otimização *Bayesiana* é amplamente utilizada como uma estratégia eficiente em modelos complexos de ML, com o objetivo de maximizar o desempenho preditivo Diferentemente de abordagens como *Grid Search* ou *Random Search*, a otimização *Bayesiana* utiliza um modelo probabilístico para selecionar os parâmetros mais promissores, tornando o processo mais eficiente e maximizando métricas de desempenho (TRAN et al., 2025; PAN et al., 2024).

Além do ajuste de hiperparâmetros, a otimização de limiares para estratificação de risco é uma etapa fundamental na validação de modelos preditivos, estando frequentemente associada a métricas como o índice de Youden (J) (LI et al., 2024; RAINIO; TEUHO; KLÉN, 2024). O J é particularmente relevante em contextos clínicos, nos quais tanto a identificação correta de casos positivos quanto a minimização de falsos negativos são importantes, e é definido como:

$$J = Sensibilidade + Especificidade - 1$$

2.6 Métricas de avaliação de desempenho

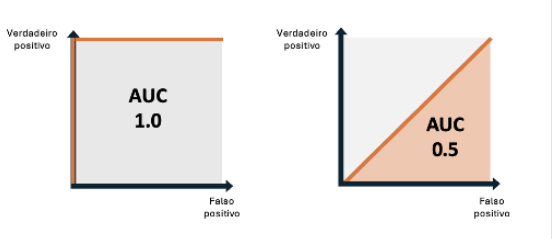
Para avaliar o desempenho dos modelos preditivos, algumas métricas tradicionais de ML podem ser aplicadas (RAINIO; TEUHO; KLÉN, 2024). Para descrever a performance, são considerados os quatro tipos possíveis de classificação:

- Verdadeiros positivos (VP)
- Verdadeiros negativos (VN)
- Falsos positivos (FP)
- Falsos negativos (FN)

Todas as métricas utilizadas na avaliação desse estudo se baseiam nessas quatro classificações. As métricas são explicadas em maiores detalhes abaixo.

Área sob a curva ROC (AUC-ROC): mede a área sob a curva que relaciona a taxa de verdadeiros positivos (sensibilidade) com a taxa de falsos positivos em diferentes limiares de decisão. Valores próximos de 1 indicam excelente capacidade discriminatória, enquanto valores em torno de 0.5 refletem desempenho equivalente ao acaso 2.2.

Figura 2.2: Área sob a curva, AUC-ROC.

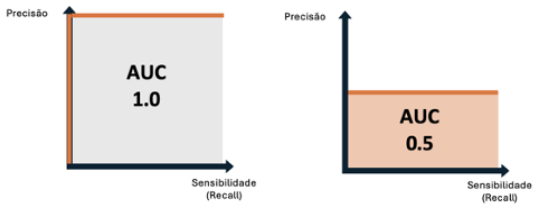


adaptado de: Google Developers, 2025

Área sob a curva precisão-recall (AUC-PR): particularmente útil em conjuntos de

dados desbalanceados, avalia a relação entre precisão e sensibilidade ao longo de diferentes limiares. Assim como a AUC-ROC, valores mais altos indicam melhor desempenho, sendo especialmente informativa quando há poucos exemplos positivos 2.3.

Figura 2.3: Área sob a curva, AUC-PR. AUC = 1 indica um resultado perfeito, enquanto AUC = 0.5 indica uma performance igual ao acaso



adaptado de: Google Developers, 2025

Acurácia: Representa a proporção total de acertos do modelo. Embora amplamente utilizada, sua interpretação deve ser cautelosa em conjuntos de dados desbalanceados. Nesses casos, métricas como a precisão e a sensibilidade também precisam ser exploradas.

$$Acuracia = \frac{VP + VN}{VP + VN + FP + FN}$$

Precisão: indica a proporção de predições positivas que são de fato verdadeiras, refletindo a capacidade do modelo de evitar falsos positivos. É relevante quando o custo de falsos positivos é elevado.

$$Precisao = \frac{VP}{VP + FP}$$

Sensibilidade: ao contrário da precisão, a sensibilidade quantifica a capacidade do modelo de identificar corretamente os casos positivos, reduzindo falsos negativos. É especialmente importante em análises biomédicas, nas quais a omissão de casos relevantes pode comprometer a interpretação clínica.

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

Especificidade: avalia a capacidade do método de detectar resultados negativos. A especificidade complementa a sensibilidade, concentrando-se em verdadeiros negativos.

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

Índice de concordância (C-index): especialmente relevante em modelos de análise de sobrevivência. Mede a concordância entre o risco previsto e a ordem real dos tempos de evento. Um C-index igual a 1 indica uma ordenação perfeita das previsões, enquanto 0.5 indica desempenho aleatório.

Vale ressaltar que a escolha dos modelos mais adequados considerou não apenas a capacidade preditiva, mas também o equilíbrio entre desempenho e interpretabilidade.

3 REVISÃO DE LITERATURA

Este capítulo revisa as principais aplicações do ML na saúde, especificamente na análise de dados transcriptômicos para a identificação de biomarcadores e seleção de terapias personalizadas em oncologia. Além disso, discutimos o uso de abordagens de XAI para análise de modelos preditivos, buscando interpretações biologicamente significativas com foco no carcinoma renal.

3.1 Aprendizado de máquina na saúde

O uso de técnicas de ML na saúde representa uma oportunidade valiosa para o avanço da pesquisa biomédica (GHASSEMI et al., 2020). A capacidade desses métodos de processar grandes volumes de dados tem impulsionado inovações importantes, sobretudo no campo da medicina personalizada, ao promover melhorias tanto no cuidado ao paciente quanto na eficiência dos processos clínicos (SHAH; LODHI, 2025; RECHARLA et al., 2025). Essa abordagem, também chamada de medicina de precisão ou personalizada, propõe um modelo que se afasta do tradicional "*one-size-fits-all*", ao individualizar estratégias de diagnóstico, prevenção e tratamento conforme as particularidades de cada indivíduo. Fatores como perfil genético, sexo, ambiente e estilo de vida são considerados para orientar decisões clínicas mais adequadas (SHAH; LODHI, 2025). Nesse cenário, a genômica assume papel central, pois a análise do genoma permite identificar variações específicas que afetam desde a predisposição a doenças até a resposta a medicamentos, favorecendo intervenções mais precisas e com menor risco de efeitos adversos (RECHARLA et al., 2025).

Os algoritmos de ML são capazes de analisar dados genômicos complexos e revelar padrões e relações sutis entre variáveis genéticas e clínicas que frequentemente escapam aos métodos estatísticos convencionais (RECHARLA et al., 2025; TRAN et al., 2025). Com isso, esses métodos têm se mostrado úteis não apenas no diagnóstico e prognóstico de doenças, mas também na seleção de terapias personalizadas e na identificação de biomarcadores (RECHARLA et al., 2025; TRAN et al., 2025).

3.2 Aprendizado de máquina aplicado na oncologia e o uso de transcriptômica

Na oncologia, o uso de técnicas de ML tem ganhado destaque pela sua capacidade de lidar com as complexidades inerentes aos diferentes tipos de tumores, cujas características moleculares podem variar de forma sutil entre tumores morfologicamente semelhantes (RECHARLA et al., 2025; SEO et al., 2025). Ao integrar diferentes fontes de dados, como informações clínicas, perfis genômicos, imagens histopatológicas e dados transcriptômicos, os algoritmos de ML permitem a construção de modelos preditivos mais precisos e sensíveis. Tais modelos têm sido aplicados com sucesso em diversas etapas do cuidado oncológico, desde o diagnóstico precoce e a classificação de subtipos tumorais até a estratificação de risco e a previsão da resposta terapêutica (SEO et al., 2025).

Nesse sentido, a transcriptômica (ou seja, as análises da expressão gênica) surge como uma das fontes de dados muito promissora para aplicações em ML (SEO et al., 2025). Através de tecnologias como o RNA-seq, é possível capturar as alterações dinâmicas na expressão gênica associadas ao desenvolvimento e à progressão tumoral (SEO et al., 2025; FONTANARI; RECAMONDE-MENDOZA, 2024). A análise desses dados por meio de modelos de aprendizado profundo tem viabilizado a identificação de assinaturas gênicas com valor prognóstico, o mapeamento de vias moleculares associadas à malignidade e a distinção entre tipos de câncer com perfis genéticos semelhantes (SEO et al., 2025; FONTANARI; RECAMONDE-MENDOZA, 2024; DIVATE et al., 2022). Portanto, a combinação entre ML e transcriptômica oferece uma ferramenta poderosa ao permitir uma compreensão mais aprofundada dos mecanismos biológicos que sustentam o câncer.

3.3 Carcinoma de células claras renais e o carcinoma papilífero renal

Entre os diferentes tipos de câncer investigados por meio da transcriptômica e de ML, os tumores renais representam um campo particularmente relevante, tanto pela sua heterogeneidade biológica quanto pelas diferenças nos desfechos clínicos (TABIBU; VINOD; JAWAHAR, 2019; ALBERTSON et al., 2025). Nesse contexto, destacamos dois subtipos: o carcinoma de células claras renais (*kidney renal clear cell carcinoma* – KIRC) e o carcinoma papilífero renal (*kidney renal papillary cell carcinoma* – KIRP). Ambos pertencem ao grupo dos carcinomas de células renais, mas apresentam perfis moleculares e clínicos distintos, com implicações importantes para o diagnóstico, o prognóstico e a

escolha terapêutica (TABIBU; VINOD; JAWAHAR, 2019; ALBERTSON et al., 2025; SU et al., 2024b).

O KIRC é o subtipo mais prevalente, representando aproximadamente 70 a 75% dos casos de carcinoma renal. Origina-se, em geral, dos túbulos contorcidos proximais e apresenta alterações genéticas características. Essas alterações estão frequentemente associadas a uma maior agressividade tumoral, maior taxa de recorrência e propensão a metástases, o que contribui para um prognóstico mais desfavorável em comparação com outros subtipos (ALBERTSON et al., 2025; SU et al., 2024b).

Já o KIRP corresponde a cerca de 15 a 20% dos carcinomas renais e também se origina do túbulo proximal, embora apresente uma morfologia histológica e um perfil molecular diferente. Esse subtipo geralmente apresenta um curso clínico menos agressivo do que o KIRC, no entanto, o KIRP não é isento de desafios terapêuticos, especialmente em casos avançados ou refratários (ALBERTSON et al., 2025; SU et al., 2024b; HWANG et al., 2024).

Estudar os tumores KIRC e KIRP de forma aprofundada é essencial para compreender os mecanismos moleculares que os distinguem, assim como para identificar biomarcadores específicos que possam orientar estratégias terapêuticas.

Nesse sentido, o projeto *The Cancer Genome Atlas* (TCGA) surge como um esforço colaborativo destinado a caracterizar exaustivamente os eventos moleculares em mais de 11.000 tumores através de 33 tipos de câncer, entre eles o KIRC e o KIRP. Uma parte crucial deste projeto é a disponibilização de perfis transcriptômicos (dados de RNA-seq) para esses tumores. Os dados transcriptômicos coletados e curados em larga escala possibilitam estudos comparativos que identificam assinaturas de expressão gênica específicas para o tecido de origem ou para subtipos moleculares, auxiliando na descoberta de biomarcadores para diagnóstico e prognóstico, além de oferecer material para o treinamento e validação de modelos de ML (DIVATE et al., 2022; LIU et al., 2018).

O uso da sobrevivência global como único *endpoint* para os tumores KIRC e KIRP nos dados do TCGA foi anteriormente validado, permitindo estimar a sobrevivência global com alta confiança e garantindo poder estatístico adequado para estudos de sobrevivência (LIU et al., 2018). Essas duas doenças apresentaram acompanhamento suficientemente longo, com tempo médio de seguimento para KIRC de $\approx$ 39 meses e KIRP de $\approx$ 25 meses, superando os períodos mínimos necessários para capturar a maioria dos óbitos, tornando a definição de sobrevivência global (que depende apenas da data de diagnóstico e da data de morte por qualquer causa), a medida menos ambígua e mais robusta nesses

subtipos renais (LIU et al., 2018).

3.4 Métodos de ML aplicados na transcriptômica

Diversos modelos de ML vêm sendo empregados na análise de dados transcriptômicos, com destaque para métodos ensemble, como *Random Forest* (RF), *Categorical Boosting* (CatBoost) e algoritmos de *gradient boosting (XGBoost/LightGBM)*. Esses modelos apresentam boa capacidade de lidar com alta dimensionalidade e variáveis ruidosas, características comuns em dados de RNA-seq, e têm sido aplicados em diferentes contextos oncológicos (DIVATE et al., 2022; PAN et al., 2024; PEDUZZI et al., 2025; LI et al., 2024) O XGBoost, por exemplo, demonstrou desempenho robusto ao prever mortalidade em pacientes com pancreatite aguda grave, atingindo AUC de 0.881 (LI et al., 2024). O RF permanece como uma das abordagens mais consolidadas, sendo utilizado na identificação de genes prognósticos no adenocarcinoma ductal pancreático (SU et al., 2024a) e em classificações pan-câncer baseadas em assinaturas genômicas reduzidas (MAMATJAN, 2024). Já o CatBoost tem sido empregado na modelagem de risco em tumores, beneficiando-se de sua capacidade de lidar eficientemente com variáveis categóricas e relações não lineares (PEDUZZI et al., 2025).

Modelos lineares, como a Regressão Logística (LR), continuam sendo amplamente usados como referência comparativa ou como componentes em abordagens híbridas, incluindo a estratificação de sobrevida em câncer de nasofaringe (ALABI et al., 2023). Por outro lado, modelos baseados em redes neurais, como *Multi-Layer Perceptrons* (MLPs) e *Deep Neural Networks* (DNNs), têm ganhado espaço por sua capacidade de capturar padrões complexos e interações não lineares em dados multi-ômicos integrados (RECHARLA et al., 2025; TRAN et al., 2025; FONTANARI; RECAMONDE-MENDOZA, 2024). As DNNs também têm se mostrado particularmente eficazes em análises pan-câncer, alcançando precisão acima de 97% na identificação de assinaturas transcriptômicas específicas de tecido (DIVATE et al., 2022), reforçando seu potencial para aplicações preditivas em larga escala.

3.5 Métodos de explicabilidade de modelos

A utilização de técnicas de inteligência artificial explicável (XAI) em análises transcriptômicas tem se mostrado essencial para tornar modelos de ML mais transparentes, interpretáveis e aplicáveis à prática clínica. Ao trabalhar com dados de RNA-seq, que são caracterizados por alta dimensionalidade e grande variabilidade biológica, modelos como redes neurais profundas frequentemente operam como "caixas-pretas", dificultando a compreensão dos fatores que geram suas decisões (SEO et al., 2025).

Nesse sentido, métodos como SHAP (*SHapley Additive exPlanations*), PFI (*Permutation Feature Importance*) e GIRP (*Global Interpretation via Recursive Partitioning*) oferecem caminhos distintos e complementares para esclarecer os mecanismos internos dos modelos. O SHAP fornece a contribuição de cada gene para a predição individual; o PFI avalia a importância de variáveis globalmente; e o GIRP constrói árvores derivadas de modelos complexos para visualização interpretável (SEO et al., 2025; YAP et al., 2021; PEDUZZI et al., 2025; YANG; RANGARAJAN; RANKA, 2018; KHAN et al., 2025). Essas abordagens permitem não apenas identificar genes com maior impacto prognóstico em diferentes subtipos tumorais, como KIRC e KIRP, mas também traduzir tais achados em regras interpretáveis que podem apoiar decisões terapêuticas individualizadas.

4 METODOLOGÍA

Trata-se de um estudo retrospectivo, observacional e exploratório com abordagem quantitativa. Este estudo foi conduzido em cinco etapas principais, apresentadas na 4.1. O código está disponível através do link: <https://github.com/GraceFeijo/TCC_especializacao>

Figura 4.1: Fluxograma das etapas metodológicas do estudo.



4.1 Coleta e pré-processamento dos dados

A primeira etapa consistiu na obtenção dos dados transcriptômicos (RNA-seq) e clínicos de pacientes diagnosticados com carcinoma renal dos subtipos KIRC e KIRP, disponibilizados pelo repositório TCGA [1].Os dados transcriptômicos utilizados já haviam passado pelo processamento padronizado do TCGA, incluindo a transformação $\log_2(x + 1)$. Os dados clínicos também foram previamente tratados, com a variável sobrevivência global convertida em binária, classificando os participantes como sobreviventes (0) e não sobreviventes (1), sendo esta última considerada a classe positiva nas análises conduzidas neste estudo. Em seguida, realizou-se a integração dos dados clínicos de tipo de tecido, gênero e status de sobrevivência com os dados transcriptômicos correspondentes.

Após, os dados foram filtrados para incluir apenas amostras de tecido tumoral, e a variável gênero foi convertida em uma variável binária numérica (feminino “1” e masculino “0”). Como os conjuntos de dados apresentavam desequilíbrio entre o número de casos de sobreviventes e não sobreviventes, foram aplicadas técnicas de balanceamento nos modelos da análise, que serão discutidas em mais detalhes no próximo sub-capítulo.

Para reduzir a dimensionalidade, inicialmente foram selecionados apenas os 10% dos genes com maior variabilidade, conforme medido pelo intervalo interquartil (IQR), uma métrica menos sensível à presença de outliers. Em seguida, o IQR dos genes foi comparado entre os grupos de sobreviventes (0) e não sobreviventes (1), mantendo-se apenas aqueles que apresentaram diferença estatisticamente significativa ($p < 0.05$). Posteriormente, foi avaliada a colinearidade entre os genes remanescentes por meio da correlação

[1] https://xenabrowser.net/datapages/

de Spearman, adequada para dados não lineares. Foram removidos um dos genes das duplas com coeficiente de correlação superior a 0.98, uma vez que ambos os genes forneceriam informações essencialmente redundantes.

Dada a alta dimensionalidade dos dados, foram aplicadas técnicas adicionais de redução de dimensionalidade para fins de visualização, como Análise de Componentes Principais (PCA), *Uniform Manifold Approximation and Projection* (UMAP) e *t-Distributed Stochastic Neighbor Embedding* (t- SNE). Contudo, essas abordagens não revelaram uma separação clara entre os grupos de sobrevivência positiva e negativa, não tendo sido utilizadas nas análises subsequentes (apêndice A).

4.2 Construção dos modelos de aprendizado de máquina

Com os dados pré-processados, foram treinados cinco algoritmos de ML, com o objetivo de prever desfechos de sobrevivência com base na expressão gênica. Os algoritmos foram escolhidos por sua capacidade de lidar com dados de alta dimensionalidade e bom desempenho preditivo em contextos biomédicos. Os dados foram divididos em conjuntos de treinamento (80%) e teste (20%), e adotou-se, sobre os dados de treinamento, a validação cruzada estratificada em 5 folds.

Para ajuste de hiperparâmetros, a otimização *Bayesiana* foi aplicada. Neste estudo foram realizadas 30 tentativas (*trials*) de otimização para cada modelo. Devido ao desbalanceamento, todos os métodos ajustam a importância das classes durante o treino. Por fim, ao invés de adotar um threshold fixo de 0.5 para classificação, foi utilizado o Índice de Youden (J) como critério para determinar o ponto de corte ideal. O J foi calculado no conjunto de treino/validação e aplicado no teste. O melhor valor de *threshold* foi selecionado como limiar ótimo de decisão do modelo, garantindo um melhor equilíbrio entre sensibilidade e especificidade. Os hiperparâmetros pós-otimização, e que foram utilizados nos dados de teste, podem ser encontrados na 4.1.

4.3 Avaliação de desempenho

A performance dos modelos foi avaliada por meio de sete métricas complementares, abrangendo diferentes aspectos do desempenho preditivo e permitindo uma análise criteriosa, especialmente em cenários com desbalanceamento entre classes. As métricas

Tabela 4.1: Hiperparâmetros utilizados no teste

	Modelo	**Hiperparâmetros pós otimização**	**Estratégia para dados desbalanceados**	***Threshold otimizado***
KIRC	RF	n_estimators: 500, max_depth: 14, min_samples_leaf: 8, max_features: sqrt	class_weight='balanced'	0.4385
	GB	n_estimators: 200, max_depth: 3, learning_rate: 0.0214033785619142, min_samples_split: 8	sample_weight w0 = n_total / (2 * n_class0) w1 = n_total / (2 * n_class1)	0.4620
	LR	C: 0.0011819136746031138, penalty: l2	class_weight='balanced'	0.4368
	MLP	lr: 0.00834298104257756, h1: 320, h2: 96, d1: 0.5, d2: 0.4, l2w: 2.992914739575636e-05	class_weight cw = n_samples / (2 * n_classes * np.bincount(y))	0.4596
	CatBoost	depth: 5, learning_rate: 0.10184551638502147, iterations: 600, l2_leaf_reg: 5.0	class_weights w0 = n_total / (2 * n_class0) w1 = n_total / (2 * n_class1)	0.4206
KIRP	RF	n_estimators: 550, max_depth: 12, min_samples_leaf: 2, max_features: sqrt	class_weight='balanced'	0.2851
	GB	n_estimators: 400, max_depth: 5, learning_rate: 0.02677945900092808, min_samples_split: 2	sample_weight w0 = n_total / (2 * n_class0) w1 = n_total / (2 * n_class1)	0.4667
	LR	C: 0.0019517224641449498, penalty: l2	class_weight='balanced'	0.5153
	MLP	lr: 0.008461387541700633, h1: 64, h2: 128, d1: 0.5, d2: 0.4, l2w: 2.474476673571584e-06	class_weight cw = n_samples / (2 * n_classes * np.bincount(y))	0.5010
	CatBoost	depth: 6, learning_rate: 0.13125830316209655, iterations: 600, l2_leaf_reg: 6.0	class_weights w0 = n_total / (2 * n_class0) w1 = n_total / (2 * n_class1)	0.4141

utilizadas foram descritas no seção 3.6. Os resultados da performance dos modelos e os valores obtidos em cada métrica estão descritos e discutidos no capítulo 5.

4.4 Aplicação de técnicas de explicabilidade

Na etapa seguinte, foram aplicadas técnicas de XAI para permitir a identificação dos fatores que mais influenciam as predições. Foram utilizados dois métodos (SHAP e PFI) para avaliar tanto a importância global dos genes quanto o impacto individual sobre cada predição. O método GIRP, por outro lado, apresenta algumas limitações que impossibilitaram o seu uso no presente estudo. Devido à complexidade de implementação (o GIRP não está disponível em bibliotecas padrão) e especificidade, visto que o método é projetado para contextos genômicos específicos, a árvore de decisão surrogada (ADS) foi utilizada para oferecer interpretabilidade com implementação mais simples. O uso da ADS facilita a visualização e interpretação das regras que regem o comportamento dos modelos de ML, promovendo melhor compreensão.

4.5 Análise final dos resultados

Por fim, os resultados obtidos com os modelos e com as técnicas de interpretabilidade foram analisados de forma integrada, com o objetivo de identificar genes ou assinaturas gênicas relevantes para a sobrevivência dos pacientes. Esses genes são então investigados em relação a vias moleculares já descritas na literatura, possibilitando uma compreensão mais profunda dos mecanismos associados ao prognóstico. Essa análise foi feita para cada subgrupo tumoral, após, comparamos os resultados entre KIRP e KIRC. Espera-se, com isso, contribuir para a identificação de novos biomarcadores e para a construção de estratégias terapêuticas mais personalizadas no contexto do carcinoma renal.

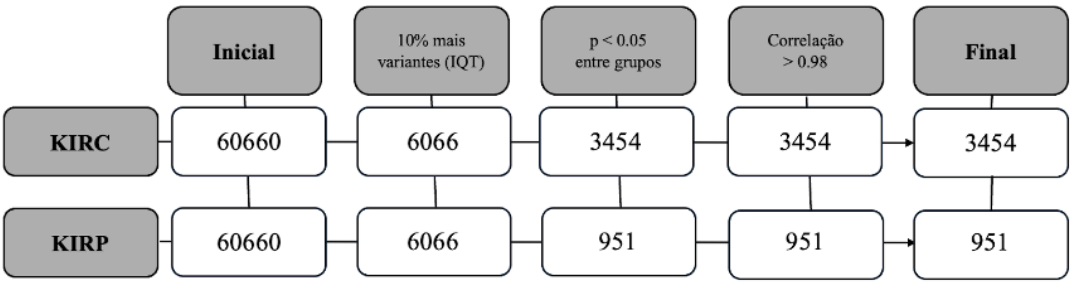
5 RESULTADOS E DISCUSSÃO

Os resultados do estudo são apresentados a seguir. Os modelos de ML foram utilizados para predizer a sobrevivência (de maneira binária) dos pacientes com tumores KIRC ou KIRP. Após, os métodos de explicabilidade foram empregados para auxiliar no entendimento dos mecanismos de decisão internos dos modelos.

5.1 Redução de dimensionalidade e divisão dos dados de treino e teste

Após a fusão dos dados transcriptômicos e clínicos, o primeiro passo foi o de reduzir o número de variáveis a serem utilizadas nos modelos. A etapa de feature selection é de extrema importância para evitar o sobreajuste. Para isso, utilizamos as três abordagens descritas na metodologia, resultando na redução significativa do conjunto inicial de 60660 variáveis transcriptômicas tanto para KIRC quanto para KIRP 5.1.

Figura 5.1: Processo de ajuste de dimensionalidade dos conjuntos de dados KIRC e KIRP.



Após a seleção das variáveis transcriptômicas, as variáveis binárias gênero e sobrevivência foram incluídas, e a divisão dos dados em conjuntos de treino (80%) e teste (20%) foi realizada. Sendo assim, as distribuições resultantes seguiram os números sumarizados na tabela 5.1.

Tabela 5.1: Distribuição das amostras e número de variáveis após redução de dimensionalidade

-	**KIRC**			**KIRP**		
	n positivo	***n***	**Variáveis**	***n positivo***	***n***	**Variáveis**
Conjunto de treino (80%)	138	427	3455	35	230	952
Conjunto de teste (20%)	35	107		9	58	
-	**Total**	534	-	**Total**	288	-

5.2 Desempenho preditivo na validação cruzada

Com o objetivo de maximizar o desempenho preditivo, reduzir o risco de sobre-ajuste e assegurar a manutenção da proporção das classes em todos os folds, a validação cruzada estratificada foi aplicada. Desta forma, os resultados a seguir mostram a variação de desempenho dos modelos ao longo de 5 execuções (i.e. *folds*).

5.2.1 Análise do desempenho na validação cruzada para KIRC

Observa-se que, para o grupo KIRC, os modelos apresentaram desempenho geral equilibrado, demonstrando consistência na capacidade de classificação entre as classes avaliadas. As métricas de curva ROC, acurácia e C-index exibiram os melhores resultados médios, com desvios padrão pequenos, indicando boa capacidade discriminativa e estabilidade dos modelos ao longo dos *folds* da validação cruzada figura 5.2 e figura 5.3.

Figura 5.2: Gráficos da performance dos modelos na validação cruzada para KIRC.

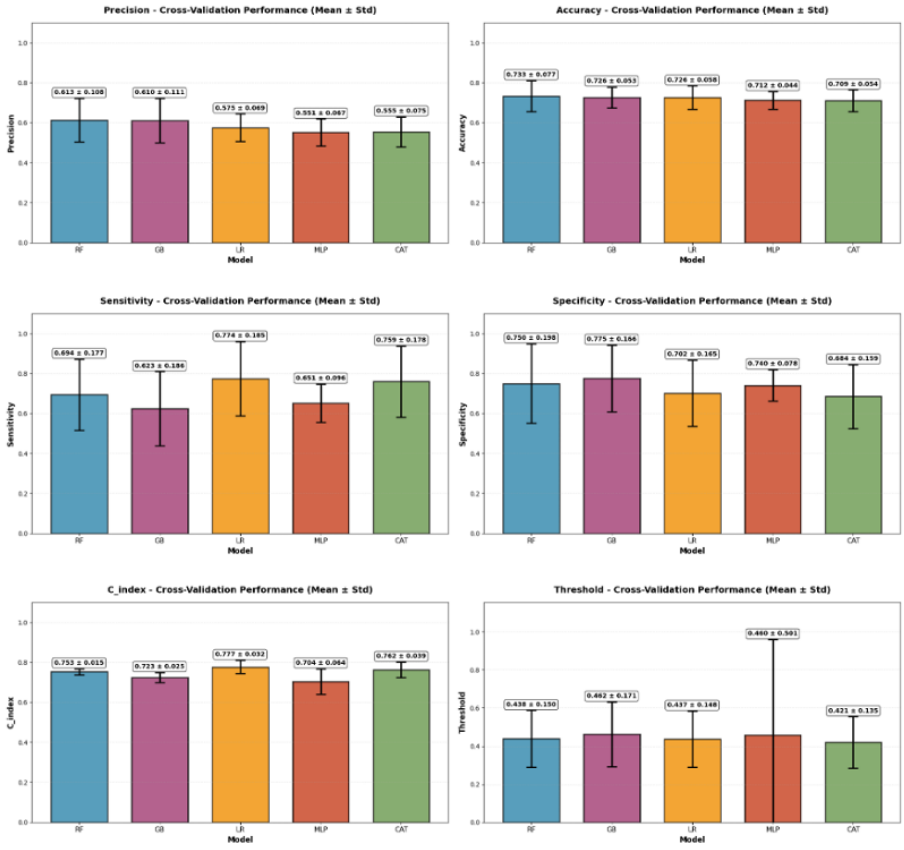
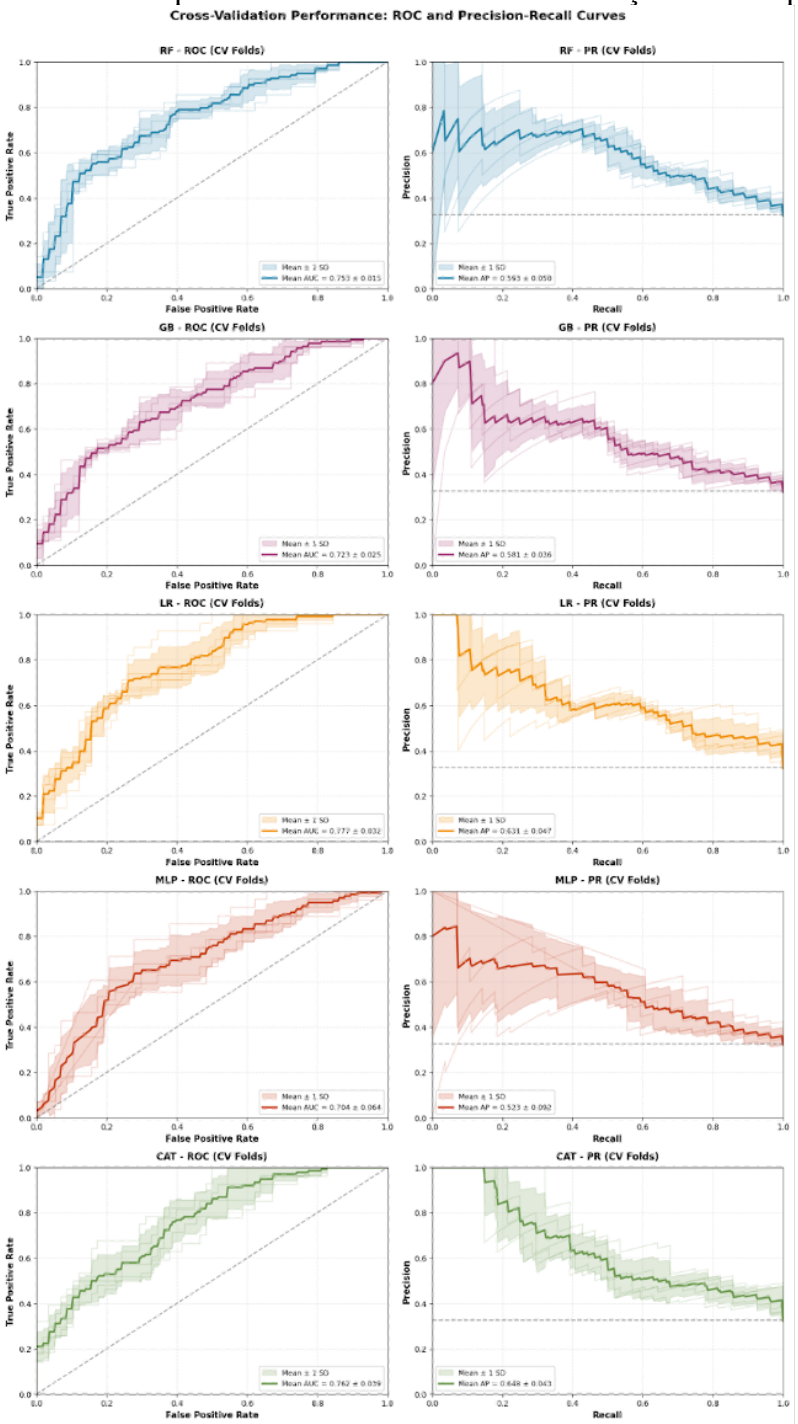


Figura 5.3: Gráficos da performance dos modelos na validação cruzada para KIRC.



Em contraste, métricas como sensibilidade, especificidade e curva PR apresentaram maior variabilidade, com desvios padrão mais amplos. A Regressão Logística (0.77) e o CatBoost (0.75) obtiveram os melhores resultados em sensibilidade, evidenciando maior capacidade de identificar corretamente indivíduos do grupo não sobrevivente, ainda que com uma certa redução na precisão. Por outro lado, o *Gradient Boosting* (0.77) e o *Random Forest* (0.75) apresentaram bons resultados em especificidade, indicando maior acurácia na classificação de pacientes sobreviventes e, portanto, menor taxa de falsos positivos. Na curva PR, os modelos de Regressão Logística (0.77) e CatBoost (0.76) novamente se destacaram, reforçando sua eficiência na distinção entre as classes.

A estabilidade dos limiares de decisão (*thresholds*), concentrados entre 0.42 e 0.46, demonstra coerência na calibração probabilística dos modelos. Além disso, os desvios padrão relativamente estreitos, exceto para o modelo MLP, sugerem baixa variabilidade entre os *folds* e boa reprodutibilidade dos resultados, confirmando a robustez geral das predições.

5.2.2 Análise da validação cruzada estratificada para KIRP

Diferentemente do conjunto KIRC, os resultados obtidos para KIRP apresentaram maiores desvios padrão entre os folds, indicando menor estabilidade dos modelos ao longo da validação cruzada 5.4 e 5.5. Os melhores desempenhos foram observados nas métricas de AUC-ROC, acurácia, especificidade e C-index. Destacamos o modelo MLP, com AUC-ROC e C-index de 0.76, e o *Random Forest*, com acurácia de 0.84 e especificidade de 0.88. Esses resultados sugerem que, apesar da instabilidade geral, alguns modelos conseguem manter boa capacidade de separação entre as classes.

Por outro lado, temos resultados menos favoráveis nas métricas de precisão e, consequentemente, na curva PR, em que o modelo *Gradient Boosting* apresentou valores de 0.37 e 0.33, respectivamente. No conjunto KIRP, o número reduzido de amostras, especialmente do grupo positivo (apenas 35 amostras disponíveis para o treino), contribui para explicar a alta variabilidade entre as métricas e entre os *folds*. Ainda assim, o modelo de Regressão Logística apresentou um comportamento mais equilibrado entre as métricas avaliadas, embora mantendo desvios padrão elevados, o que reforça a baixa estabilidade preditiva em comparação ao conjunto KIRC.Em relação aos *thresholds*, também se observa ampla variabilidade dos desvios padrão entre os modelos. Contudo, o modelo CatBoost se destaca positivamente, com um desvio padrão de apenas 0.09, inferior

Figura 5.4: Gráficos da performance dos modelos na validação cruzada para KIRP.

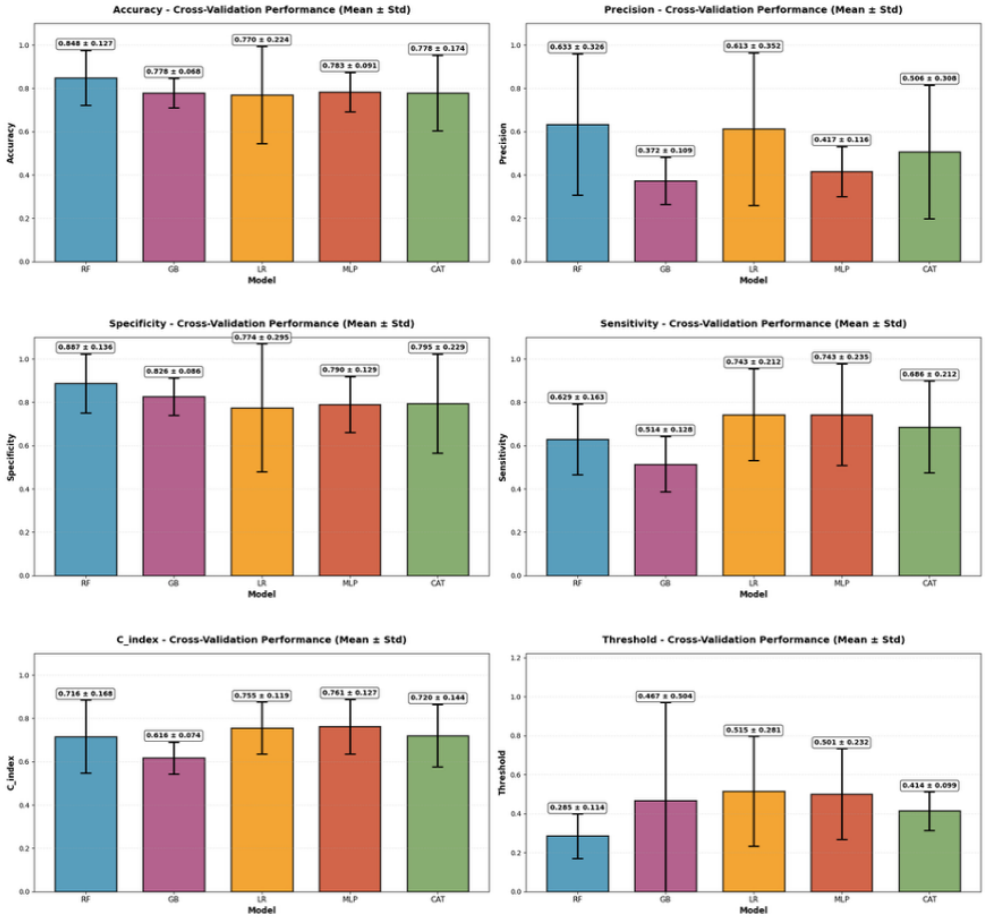
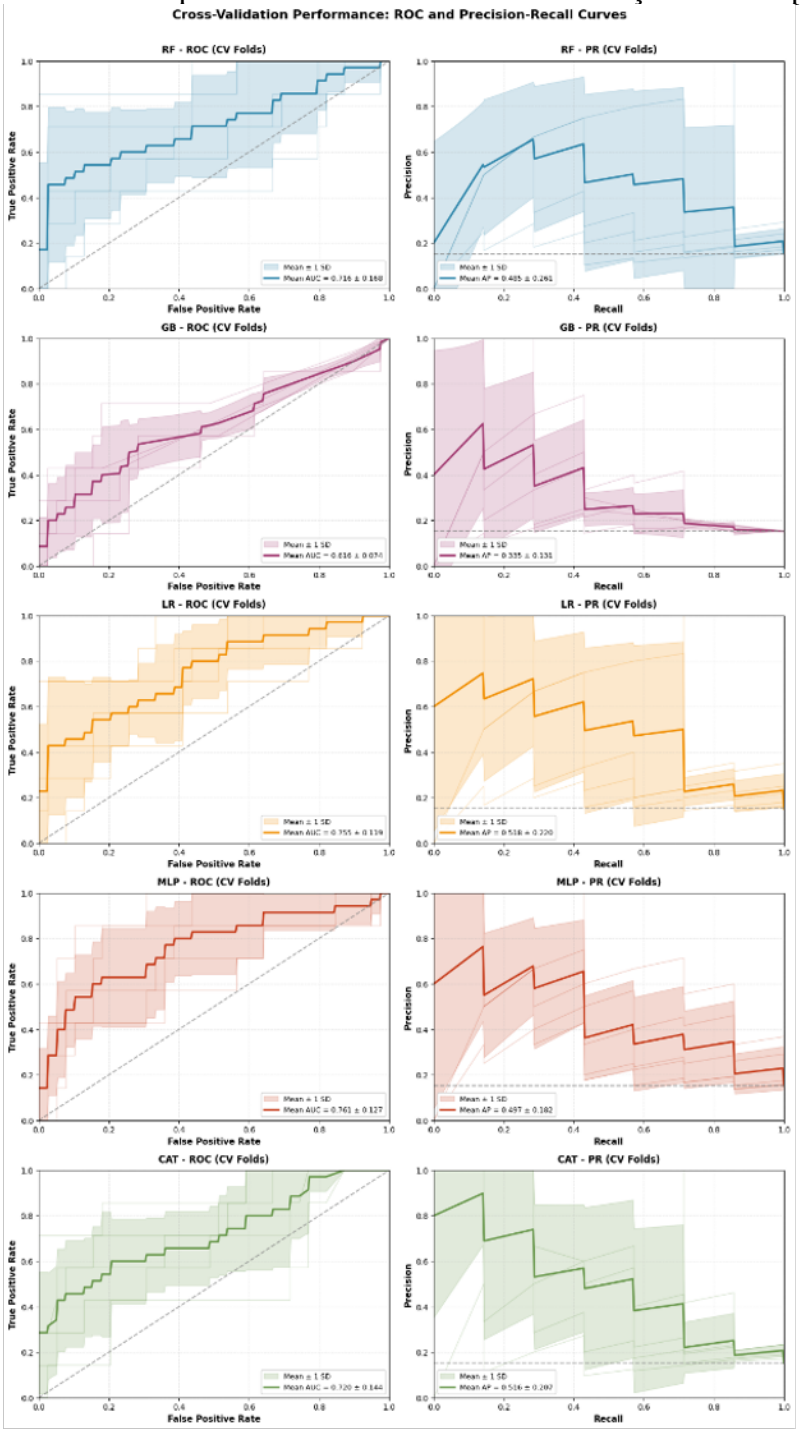


Figura 5.5: Gráficos da performance dos modelos na validação cruzada para KIRP.



inclusive aos valores observados no treino do conjunto KIRC.

5.3 Desempenho dos modelos de ML no conjunto de teste

Após o treinamento, os modelos com as melhores configurações de hiperparâmetros foram aplicados ao conjunto de teste. A tabela 5.2 apresenta os resultados obtidos para KIRC e a tabela 5.3 para KIRP.

5.3.1 Análise do desempenho dos modelos – KIRC

Para o conjunto de dados KIRC, no geral, os modelos obtiveram resultados satisfatórios (Tabela 5.2). Podemos destacar, entre os cinco modelos avaliados, o com melhor desempenho, sendo o CatBoost que alcançou valores de 0.86 em especificidade, 0.82 na AUC-ROC e C-index, além de uma acurácia de 0.78. Já o LR, teve o melhor resultado em sensibilidade (0.77) e o mesmo resultado que o CatBoost em AUC-PR (0.72), apesar de apresentar os piores resultados em acurácia (0.73), precisão (0.57) e especificidade (0.72). O modelo com os piores resultados foi o MLP, com valores como 0.65 em AUC-PR e 0.60 em sensibilidade 5.6.

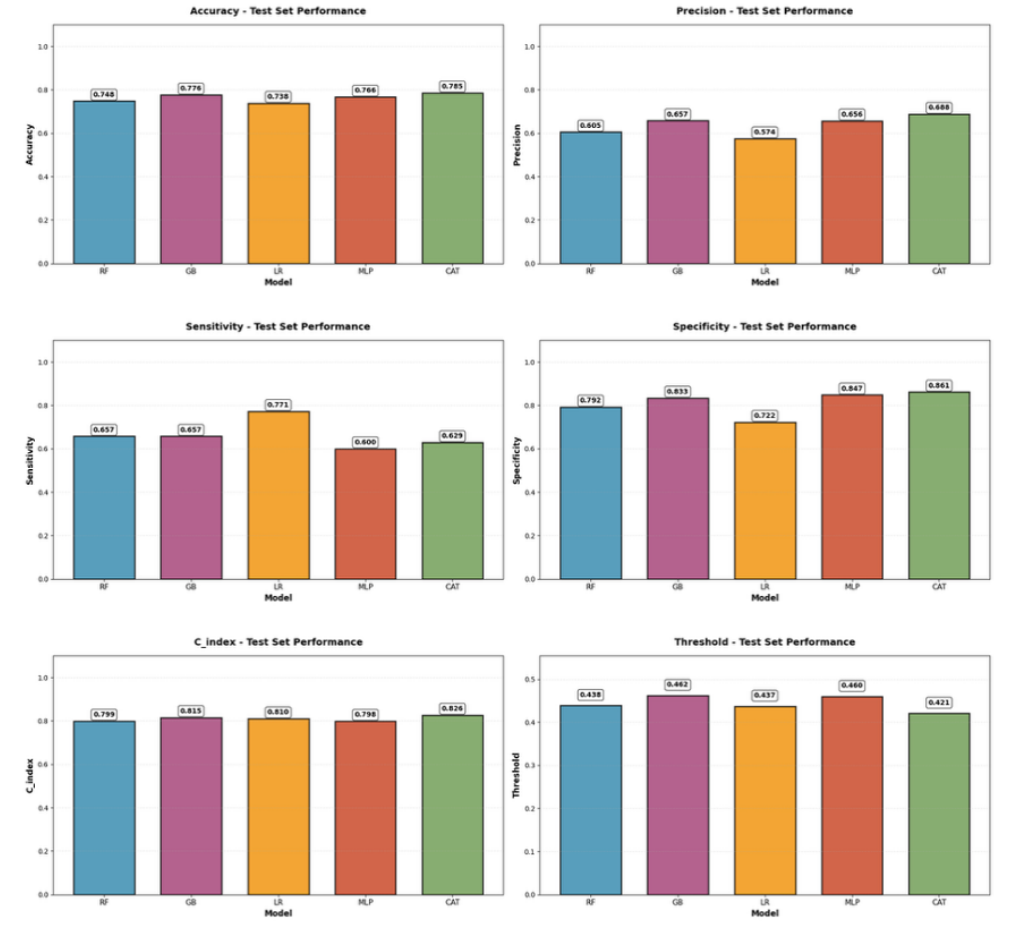
Tabela 5.2: Resultados do teste dos modelos no conjunto de dados KIRC. Os valores mais altos em cada métrica estão marcados em verde e os mais baixos em rosa.

-	KIRC							
	AUC-ROC	**AUC-PR**	**Acurácia**	**Precisão**	**Sensibilidade**	**Especificidade**	**C-index**	**Threshold**
Random Forest	0.798	0.702	0.747	0.605	0.657	0.791	0.798	0.438
Gradient Boosting	0.815	0.679	0.775	0.657	0.657	0.833	0.815	0.462
MLP	0.798	0.651	0.766	0.656	0,600	0.847	0.798	0.459
LR	0.809	0.729	0.738	0.574	0.771	0.722	0.809	0.436
CatBoost	0.826	0.729	0.785	0.687	0.628	0.861	0.826	0.420

Na literatura, o modelo CatBoost também demonstrou desempenho competitivo na previsão de risco para Adenocarcinoma Ductal Pancreático, com a melhor sensibilidade entre os modelos testados, atingindo 0.77 (PEDUZZI et al., 2025). Em um estudo de sobrevivência para Carcinoma Hepatocelular, ao ser aplicado no modelo SurvMLSHAP, o CatBoost obteve acurácia de 0.98 e sensibilidade de 1 no conjunto de teste (PAN et al., 2024). Similarmente, a LR também teve destaque no diagnóstico de câncer de pâncreas usando biomarcadores urinários, com acurácia de 0.86 e AUC-ROC de 0.89 (LI et al., 2024; ALMISNED et al., 2025). Para a previsão de mortalidade em 30 dias em Pancreatite Aguda Grave, a LR obteve um AUC de 0.84 (LI et al., 2024). Além disso, na

estratificação de sobrevivência para Câncer Nasofaríngeo, a LR fez parte de um modelo de voting classifier que alcançou 0.85 de acurácia (ALABI et al., 2023).

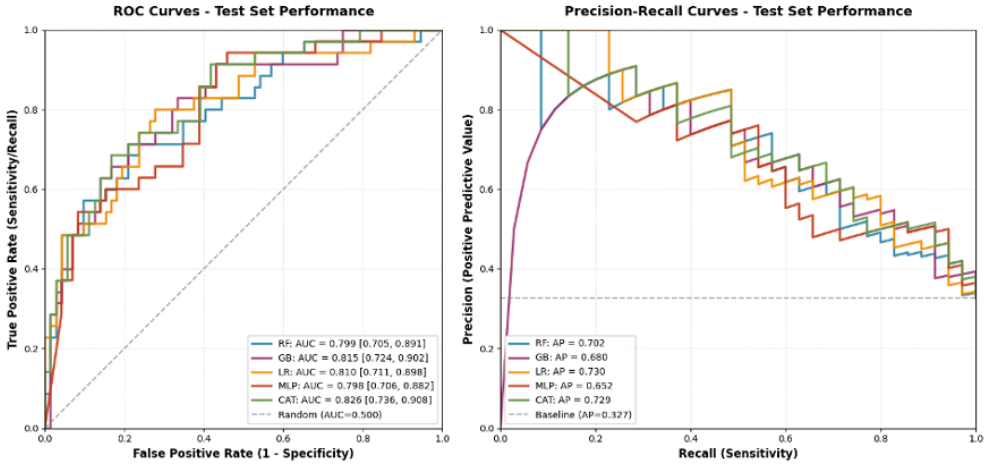
Figura 5.6: Análise do desempenho dos modelos no conjunto de dados de teste - KIRC



Nenhum modelo alcançou junto precisão e sensibilidade ≥ 0.70, indicando alta taxa de falsos positivos e perda relevante dos casos positivos verdadeiros. Mesmo assim, as métricas discriminativas (como AUC-ROC e C-index) apresentaram bons resultados, sugerindo boa capacidade de separação entre classes. Percebemos também que o *threshold* utilizado variou entre 0.42 e 0.46, demonstrando a necessidade de ajuste desse parâmetro em conjuntos de dados desbalanceados.

A análise das curvas ROC e PR (figura 5.7), reforça o padrão observado. Apesar do desbalanço de classes, as curvas ROC mostram que todos os modelos apresentam boa capacidade discriminativa, com AUC variando entre 0.79 e 0.82 aproximadamente. No entanto, a análise da curva PR mostra com mais clareza a dificuldade em identificar corretamente a classe positiva. Embora os modelos atinjam precisão elevada em níveis baixos de sensibilidade, todas as curvas caem rapidamente, indicando o impacto do forte desbalanceamento da variável de sobrevivência.

Figura 5.7: Gráficos das curvas ROC e PR para todos os modelos avaliados - KIRC



5.3.2 Análise do desempenho dos modelos - KIRP

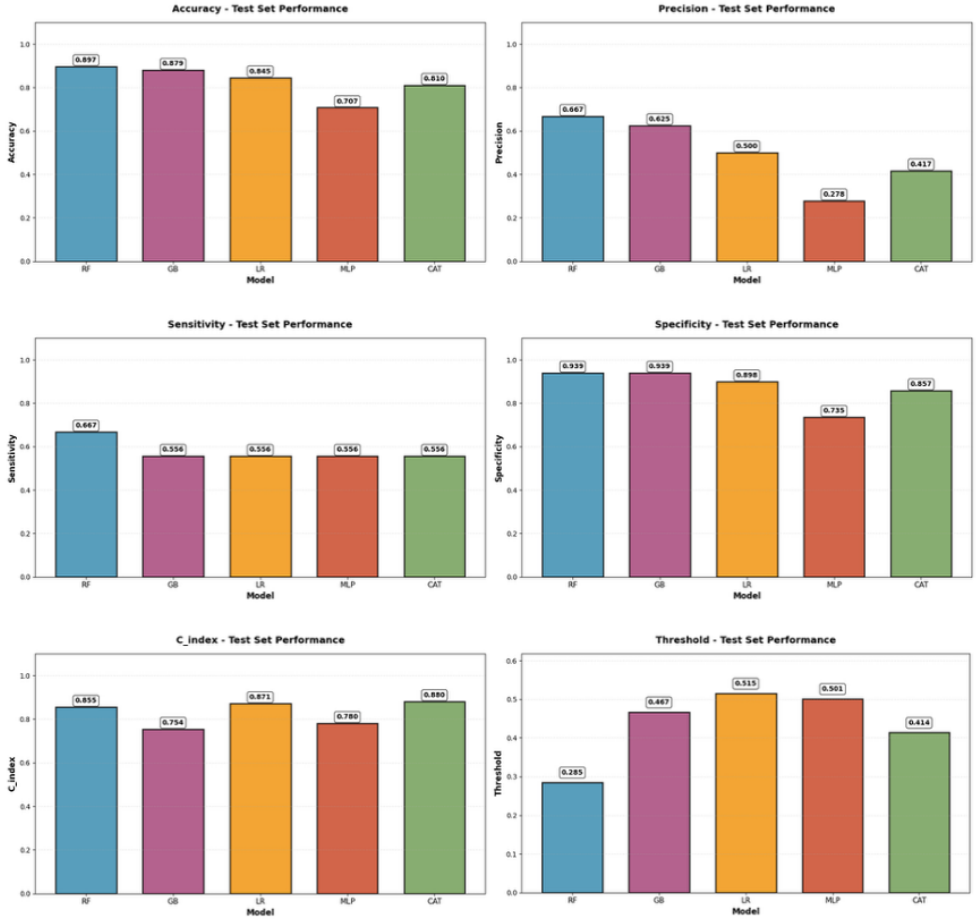
Assim como observado no treinamento, o conjunto de teste do KIRP apresentou desempenho inferior quando comparado ao KIRC (tabela 5.3 e figura 5.8). Podemos observar bons resultados em métricas discriminativas como AUC-ROC (0.75 - 0.87), acurácia (0.70 - 0.89), especificidade (0.73 - 0.93) e C-index (0.75 - 0.87). No entanto, precisão e sensibilidade ficaram muito abaixo do ideal, chegando a valores inferiores ao acaso (< 0.5) em alguns modelos. O *Random Forest* apresentou o melhor desempenho, seguido do CatBoost. Aqui, o MLP também apresentou o pior desempenho, com precisão de apenas 0.27 e AUC-PR de 0.49. Em relação ao *threshold*, percebemos valores muito mais variantes do que KIRC, sendo o mais baixo 0.28 para o *Random Forest* e o mais alto 0.51 para a regressão logística. Mais uma vez vemos a importância da flexibilização do threshold para otimização dos resultados.

Tabela 5.3: Resultados do teste dos modelos no conjunto de dados KIRP. Os valores mais altos em cada métrica estão marcados em verde e os mais baixos em rosa.

-	KIRP							
-	**AUC-ROC**	**AUC-PR**	**Acurácia**	**Precisão**	**Sensibilidade**	**Especificidade**	**C-index**	**Threshold**
Random Forest	0.854	0.529	0.896	0.666	0.666	0.938	0.854	0.285
Gradient Boosting	0.753	0.503	0.879	0.625	0.555	0.938	0.753	0.466
MLP	0.780	0.494	0.706	0.277	0.555	0.734	0.780	0.501
LR	0.870	0.533	0.844	0.500	0.555	0.897	0.870	0.515
CatBoost	0.879	0.702	0.810	0.416	0.555	0.857	0.879	0.414

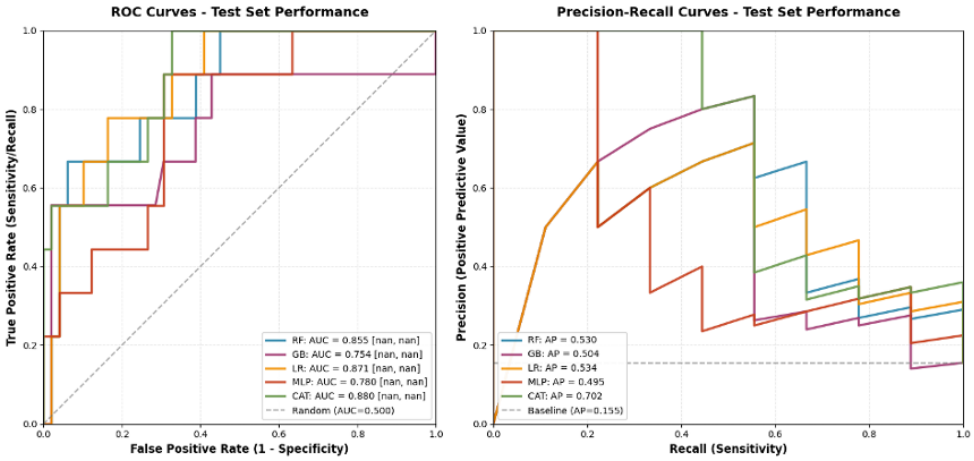
Em relação às curvas ROC e PR do conjunto KIRP (figura 5.9), elas refletem os desafios já observados causados pelo número extremamente reduzido de casos positivos e ao grande desbalanceamento entre as classes. As curvas ROC demonstram que os modelos conseguem manter uma capacidade discriminativa boa a moderada, com AUC

Figura 5.8: Gráficos da performance dos modelos no conjunto de dados de teste - KIRP



variando entre aproximadamente 0.75 e 0.88. No entanto, a análise da curva PR demonstra com clareza a limitação preditiva dos modelos. Embora o CatBoost apresente a maior área (0.70), as demais curvas apresentam grande variabilidade e precisão baixa ao longo de quase todo o eixo X (sensibilidade).

Figura 5.9: Gráficos das curvas ROC e PR para todos os modelos avaliados - KIRP.



5.4 Métodos de explicabilidade

Na pesquisa em saúde é especialmente importante a compreensão e interpretação dos fatores de decisão. Este estudo utilizou três abordagens distintas de XAI que fornecem informações complementares entre si, ao priorizar a consistência dos achados entre os modelos de explicabilidade, buscou-se identificar biomarcadores com sólida relevância preditiva. Os resultados detalhados, discriminados por técnica XAI, estão disponíveis no apêndice B. Como o foco do trabalho está em analisar genes recorrentes, a seguir serão apresentados os genes identificados com maior frequência pelas técnicas de XAI nas coortes de KIRC e KIRP. Os elementos presentes na figura5.10, figura 5.11 e figura 5.12 foram mapeados ao menos 3 instâncias pelos métodos de explicabilidade aplicados nos algoritmos de ML. Com isso, foram identificados 12 genes de interesse para o KIRC e 10 para o KIRP. A tabela com a frequência dos 30 genes mais comuns em cada modelo XAI pode ser encontrada no apêndice C.

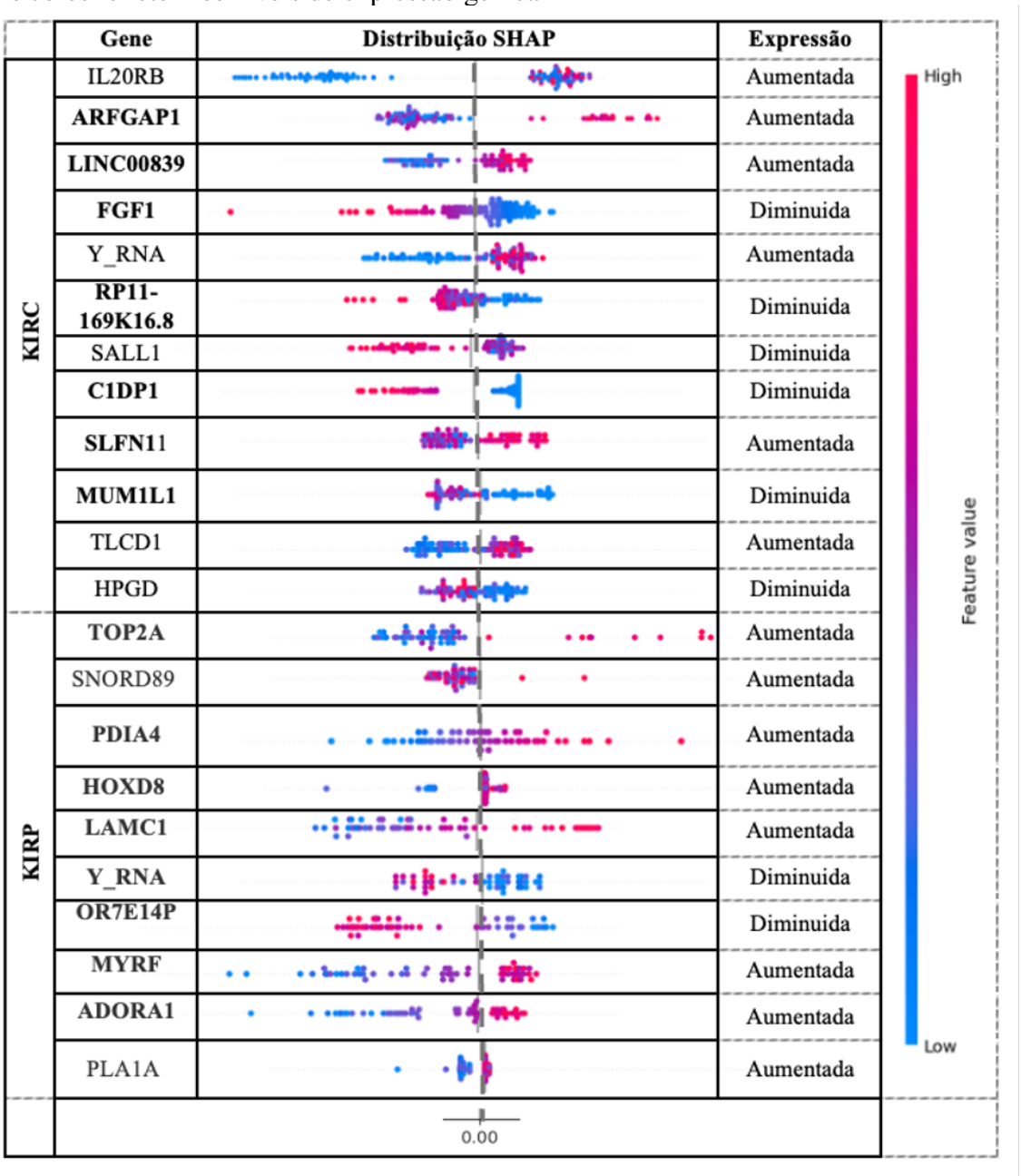
Figura 5.10: Descrição dos genes de maior importância para KIRC identificados pelas análises de XAI. Apresenta os principais genes associados para cada subtipo de carcinoma renal, indicando se codificam proteínas, e suas funções biológicas conhecidas.

	Gene	Proteina	Função
KIRC	IL20RB	Sim	Codifica uma proteína dos receptores de citocinas do tipo II, envolvido em respostas pró-inflamatórias e anti-inflamatórias
	ARFGAP1	Sim	Biomarcador em diversos tumores, mas com impactos opostos, podendo ser considerado supressor ou promotor.
	LINC00839	Não	RNA longo não codificante intergénico. Molécula reguladora intracelular
	FGF1	Sim	Codifica o fator de crescimento de fibroblastos 1. Sua expressão elevada está associada ao crescimento tumoral e à metástases.
	Y_RNA	Não	Lócus para um RNA não codificante, o RNA Y. Não codificam proteínas, mas desempenham um papel em diversos processos celulares essenciais.
	RP11-169K16.8	Não	Pseudogene processado, o que significa que ele não possui íntrons e geralmente não codifica proteínas. Pouco estudado.
	SALL1	Sim	Codifica uma proteína que atua como um fator de transcrição repressor, em alguns tecidos atua como um supressor tumoral
	C1DP1	Não	Pseudogene processado derivado do gene *C1D, que* codifica um correpressor de receptor nuclear envolvido na reparação de DNA e na regulação da transcrição.
	SLFN11	Sim	Codifica uma proteína que atua na resposta celular ao stress do DNA, é determinante crítico da eficácia de muitas quimioterapias comuns
	MUM1L1	Sim	Codifica uma proteína envolvida na regulação epigenética e na manutenção da estabilidade do genoma.
	TLCD1	Sim	Importante biomarcador pan-câncer envolvido no metabolismo tumoral e na resposta imunológica.
	HPGD	Sim	Codifica a principal enzima responsável pela inativação metabólica das prostaglandinas. Atua como um supressor tumoral e tem efeitos anti-inflamatórios

Figura 5.11: Descrição dos genes de maior importância para KIRP identificados pelas análises de XAI. Apresenta os principais genes associados para cada subtipo de carcinoma renal, indicando se codificam proteínas, e suas funções biológicas conhecidas

KIRP	TOP2A	Sim	Codifica uma enzima nuclear que gerencia a forma e enrolamento do DNA - processo essencial para a sobrevivência e proliferação celular. Biomarcador de prognóstico e de resposta à quimioterapia.
	SNORD89	Não	Regulador ativo da progressão tumoral, sendo um dos snoRNAs mais promissores para o desenvolvimento de novas terapias genéticas em oncologia ginecológica.
	PDIA4	Sim	Codifica uma enzima chave que promove a sobrevivência celular tumoral e aumenta a resistência a tratamentos em diversos tipos de câncer
	HOXD8	Sim	Codifica um fator de transcrição que desempenha um papel dual no câncer, agindo como supressor de tumoral ou oncogene dependendo do tipo de tecido e do contexto biológico
	LAMC1	Sim	Codifica uma proteína crucial das membranas basais. Sua sobre-expressão, que promove a proliferação, migração e invasão de células tumorais, está associada a um pior prognóstico.
	Y_RNA	Não	Sua detecção elevada em biópsias é um forte indicativo de alta taxa de replicação celular e instabilidade genômica, características centrais da progressão maligna
	OR7E14P	Não	Foi identificado em estudos de transcriptoma como parte de assinaturas gênicas que distinguem tecidos tumorais de normais
	MYRF	Sim	Codifica um fator de transcrição associado à membrana que desempenha um papel fundamental no desenvolvimento e manutenção de múltiplos órgãos.
	ADORA1	Sim	Codifica o receptor de adenosina A1. Regula a função renal, inibindo a secreção de renina e modulando o tónus vascular renal e a taxa de filtração glomerular.
	PLA1A	Sim	Atua na sinalização lipídica e na modulação da imunidade tumoral. Alta expressão é utilizadaindicativo de pior sobrevida global em pacientes com câncer renal (KIRC) e tumores de trato digestivo superior.

Figura 5.12: Distribuição dos valores SHAP dos genes de maior importância para KIRC e KIRP. Cada ponto representa uma amostra, com a posição no eixo X indicando a contribuição do gene para o modelo preditivo. Pontos à direita estão associados a pior prognóstico. As cores refletem os níveis de expressão gênica



5.4.1 Resultados explicabilidade - KIRC

No cenário do KIRC, observou-se a superexpressão de genes codificadores de proteínas envolvidos na resposta inflamatória, no metabolismo tumoral e na resposta ao estresse genético, como *IL20RB*, *SLFN11* e *TLCD1*. O gene *IL20RB*, por exemplo, foi previamente associado a piores desfechos clínicos e maior infiltração de células imunes no microambiente tumoral (ZHANG et al., 2022). A presença de *SLFN11*, biomarcador pan-câncer de sensibilidade a terapias genotóxicas (ZHOU et al., 2025), reforça a importância de mecanismos de resposta ao dano no DNA nesse subtipo.

Além destes, destacam-se os RNAs não codificantes (*LINC00839* e *Y_RNA*), que atuam na modulação de vias pós-transcricionais e na replicação celular, o que está alinhado com estudos recentes que evidenciam o papel regulatório desses transcritos na progressão tumoral (HU et al., 2024; CAMPO et al., 2025).

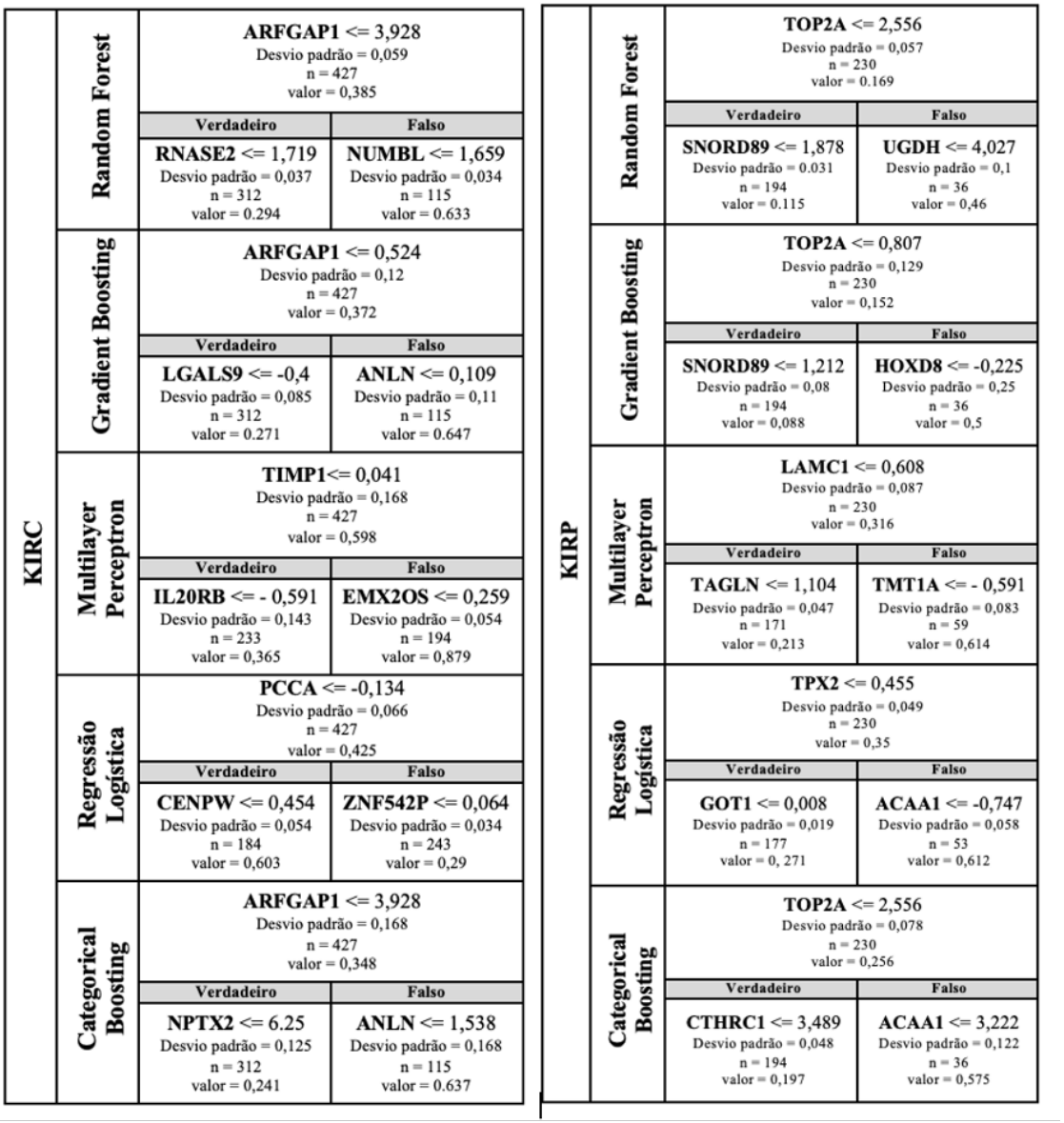
Inversamente, notou-se a sub-expressão de genes relacionados à reparação de DNA e à manutenção da estabilidade genômica, como *C1DP1* e *MUM1L1*, indicando possível perda de controle sobre mecanismos supressores e diferenciação celular. *SALL1*, por exemplo, também sub expresso em KIRC, tem sido apontado como um supressor tumoral cuja baixa expressão se associa a pior prognóstico e menor infiltração linfocitária em tumores renais (QU et al., 2025). De forma semelhante, *HPGD* atua na inativação de prostaglandinas, e sua redução pode contribuir para um ambiente pró-inflamatório favorável à progressão do carcinoma (ZHAI et al., 2024). Assim, os resultados para KIRC sugerem um equilíbrio dinâmico entre genes pró-tumorais e supressores, modulando tanto a imunidade quanto o metabolismo celular.

Ao analisar a distribuição das amostras e o impacto preditivo via valores SHAP (Figura 5.12), observa-se que a distinção entre os grupos de prognóstico nem sempre é linear. Genes como o *C1DP1* exibem uma polarização nítida de expressão entre os grupos, reforçando seu potencial como biomarcador. Por outro lado, o gene *FGF1* apresenta diferenciação clara apenas nos valores extremos, enquanto níveis intermediários de expressão sobrepõem-se em amostras de sobreviventes e não sobreviventes. Mesmo assim, recentemente o gene *FGFI* tem sido considerado promissor como indicador prognóstico para KIRC (ZHANG et al., 2021).

Quanto aos nodos decisórios das árvores de decisão para KIRC (Figura 5.13), a recorrência de genes sugere que sua expressão exerce um papel central na diferenciação dos fenótipos analisados. No caso do KIRC, o *ARFGAP1*, que já vem sendo estudado

como biomarcador em alguns tipos tumorais (MENG et al., 2021) e o *ANLN*, que é um regulador da divisão celular e da organização do citoesqueleto, já foi associado à proliferação tumoral, invasividade e pior prognóstico clínico (LIU et al., 2022). Esses genes podem representar marcadores genéticos potenciais.

Figura 5.13: Genes presentes nos nodos principais e secundários das árvores de decisão para cada modelo de ML



5.4.2 Resultados explicabilidade - KIRP

Em relação ao KIRP, evidenciaram-se genes codificadores de proteínas associados à sobrevivência, proliferação e sinalização extracelular, como *TOP2A*, *PDIA4* e *LAMC1*. A superexpressão de *TOP2A* e *PDIA4* reflete um fenótipo altamente proliferativo e resis-

tente ao estresse celular, com impacto direto na eficácia terapêutica (WANG et al., 2024; KANG et al., 2023), além disso, o aumento de *LAMC1* e *PLA1A*, ambos associados a maior invasividade e menor sobrevida global, reforça a importância das vias de adesão e remodelamento da matriz extracelular no KIRP (BAI et al., 2022; ZHOU et al., 2022).

Também foram identificados genes não codificantes, como o *SNORD89* e variantes de *Y_RNA*, correlacionados à progressão tumoral. De modo geral, o perfil de KIRP é caracterizado pela superexpressão gênica, com exceções notáveis de subexpressão em *Y_RNA* e *OR7E14P*, a diminuição de *Y_RNA* e *OR7E14P* neste subtipo tumoral reflete alterações na dinâmica de replicação e na regulação transcricional. No caso de genes como *TOP2A*, *PDIA4*, *MYRF* e *ADORA1*, a ampla distribuição dos valores SHAP indica uma variabilidade funcional, onde o efeito pode transitar entre protetor (SHAP negativo) e promotor tumoral (SHAP positivo), dependendo do contexto da amostra. A alta expressão de *ADORA1*, gene envolvido na sinalização purinérgica, sugere uma interface entre regulação metabólica e resposta imune renal, de acordo com estudos prévios que destacam sua relevância em processos de inflamação e perfusão renal (XIE et al., 2024).

Por fim, os nodos primários e secundários das árvores de decisão destacaram consistentemente os genes *ACAA1*, *TOP2A* e *SNORD89*, sugerindo sua relevância preditiva no contexto do subtipo KIRP. O gene *ACAA1*, tem sido associado à modulação da infiltração de linfócitos T e à sobrevida em diferentes tipos de câncer, indicando possível papel imunomodulador e prognóstico (FENG; SHEN, 2020). Já *TOP2A*, uma enzima essencial para a topologia e replicação do DNA, foi previamente correlacionado à progressão tumoral e pior prognóstico em carcinomas renais claros, reforçando sua importância como biomarcador de proliferação celular e resposta terapêutica (WANG et al., 2024). Por fim, *SNORD89*, integrante da classe de pequenos RNAs nucleolares, integra assinaturas gênicas de valor prognóstico em tumores ovarianos, com potencial envolvimento em mecanismos epigenéticos e resistência tumoral (ZHU et al., 2022). Esses achados sustentam o potencial desses três genes como marcadores de desfecho e alvos para estudos funcionais em KIRP.

A integração de múltiplos métodos de explicabilidade permite interpretar de forma mais transparente as relações entre os padrões de expressão gênica e os desfechos clínicos em diferentes subtipos tumorais. Curiosamente, apesar de ambos os tumores apresentarem genes diferentes da mesma família não codificante de *Y_RNA*, eles não pertencem ao mesmo locus e todos os outros genes de interesse são diferentes entre KIRC e KIRP, indicando mais uma vez que apesar das similaridades, os dois apresentam características

inerentes ímpares. As assinaturas gênicas obtidas por XAI reforçam que o KIRC tende a apresentar um perfil imunomodulado e epigeneticamente regulado, enquanto o KIRP demonstra características mais proliferativas e dependentes de sinalização metabólica, informações que podem auxiliar na estratificação prognóstica e no delineamento de terapias de precisão.

5.5 CatBoost aplicado aos genes de maior relevância

Dentre os métodos preditivos avaliados, o algoritmo CatBoost demonstrou o desempenho mais robusto, alcançando as métricas superiores nos testes com o conjunto de dados de KIRC e o segundo melhor desempenho para KIRP. Por sua consistência, este modelo foi selecionado para validar um conjunto de dados otimizado, composto exclusivamente pelas variáveis identificadas como de maior relevância pelas técnicas de XAI.

Para a composição deste dataset reduzido, foram consolidadas todas as variáveis apontadas pelos modelos de explicabilidade. Após a remoção de redundâncias, o conjunto final totalizou 117 variáveis para o grupo KIRC e 153 para o grupo KIRP. O treinamento e a validação do modelo CatBoost seguiram a metodologia descrita anteriormente, e os resultados obtidos podem ser observados na tabela 5.4.

Tabela 5.4: Resultados da análise com o modelo CatBoost utilizado apenas os genes de maior relevância.

-	**CatBoost aplicado aos genes de maior relevância**							
	AUC-ROC	**AUC-PR**	**Acurácia**	**Precisão**	**Sensibilidade**	**Especificidade**	**C-index**	**Threshold**
KIRC	0.809	0.720	0.757	0.609	0.714	0.777	0.809	0.397
KIRP	0.918	0.592	0.706	0.346	1	0.653	0.918	0.422

Apesar da redução considerável na dimensionalidade dos dados, o CatBoost manteve um desempenho próximo ao obtido com o conjunto genômico completo. Este achado é significativo, pois indica que a eficácia preditiva pode ser preservada mesmo utilizando menos de 200 genes alvo. Contudo, persistiram as limitações observadas nas análises anteriores, principalmente em relação à baixa precisão do modelo, o que pode ser explicado em parte pelo baixo número de amostras e o desbalanceamento de classes.

6 CONCLUSÃO

Os achados deste estudo demonstram que a integração entre ML e inteligência artificial explicável (XAI) representa uma abordagem promissora para a identificação de biomarcadores prognósticos no carcinoma renal. A aplicação combinada de algoritmos supervisionados e métodos interpretáveis, como SHAP e PFI, permitiu não apenas alcançar bom desempenho preditivo, mas também revelar padrões de expressão gênica biologicamente coerentes. Observou-se que o subtipo KIRC é caracterizado por um perfil imunomodulado e influenciado por RNAs não codificantes, enquanto o KIRP apresenta assinaturas mais associadas à proliferação celular e à sinalização metabólica. Tais diferenças reforçam a heterogeneidade molecular entre os subtipos e evidenciam o potencial das assinaturas gênicas em apoiar a estratificação de risco e a personalização terapêutica.

Entre as limitações do estudo, destacam-se o número reduzido de amostras no grupo KIRP, o desbalanceamento entre classes e o não uso de outros dados clínicos, como a presença de metástases. Além disso, os resultados são baseados exclusivamente em dados públicos do TCGA, o que reforça a necessidade de validação em coortes independentes e contextos clínicos reais. Como perspectivas futuras, a metodologia do estudo pode ser aplicada para outros tumores. Além disso, a ampliação das amostras com o uso de bases de dados distintas no treinamento/validação pode auxiliar na avaliação da generalização dos resultados. Da mesma forma, a integração de mais dados genômicos, epigenômicos e proteômicos, possibilitam uma visão mais abrangente dos mecanismos moleculares associados à sobrevivência. Finalmente, a validação experimental in vitro e in vivo dos genes identificados, aliada à aplicação de novos métodos de XAI em modelos mais complexos, poderá fortalecer a aplicabilidade clínica das descobertas e contribuir para o avanço da medicina de precisão na oncologia.

REFERÊNCIAS

ALABI, R. O. et al. Machine learning explainability in nasopharyngeal cancer survival using lime and shap. **Scientific Reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 8984, 2023.

ALBERTSON, D. et al. Genotype phenotype correlation of renal tumors in the cancer genome atlas database. **International Journal of Surgical Pathology**, SAGE Publications Sage CA: Los Angeles, CA, v. 33, n. 2, p. 289–301, 2025.

ALMISNED, F. A. et al. Incorporation of explainable artificial intelligence in ensemble machine learning-driven pancreatic cancer diagnosis. **Scientific Reports**, Nature Publishing Group UK London, v. 15, n. 1, p. 14038, 2025.

BAI, J. et al. Comprehensive analysis of lamc1 expression and prognostic value in kidney renal papillary cell carcinoma and clear cell carcinoma. **Frontiers in Molecular Biosciences**, Frontiers Media SA, v. 9, p. 988777, 2022.

CAMPO, A. et al. Yrnas: biosynthesis, structure, functions and involvment in cancer development. **Discover Oncology**, Springer, v. 16, n. 1, p. 176, 2025.

DIVATE, M. et al. Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures. **Cancers**, MDPI, v. 14, n. 5, p. 1185, 2022.

FENG, H.; SHEN, W. Acaa1 is a predictive factor of survival and is correlated with t cell infiltration in non-small cell lung cancer. **Frontiers in Oncology**, Frontiers Media SA, v. 10, p. 564796, 2020.

FONTANARI, T.; RECAMONDE-MENDOZA, M. Evaluating the generalization of neural network-based pan-cancer classification models for cohort-specific predictions. In: SBC. **Simpósio Brasileiro de Bioinformática (BSB)**. [S.l.], 2024. p. 12–23.

GHASSEMI, M. et al. A review of challenges and opportunities in machine learning for health. **AMIA Summits on Translational Science Proceedings**, v. 2020, p. 191, 2020.

HAMEY, F. K.; GÖTTGENS, B. Machine learning predicts putative hematopoietic stem cells within large single-cell transcriptomics data sets. **Experimental hematology**, Elsevier, v. 78, p. 11–20, 2019.

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

HU, Y. et al. Linc00839 in human disorders: Insights into its regulatory roles and clinical impact, with a special focus on cancer. **Journal of Cancer**, v. 15, n. 8, p. 2179, 2024.

HWANG, J. et al. Discovery and validation of survival-specific genes in papillary renal cell carcinoma using a customized next-generation sequencing gene panel. **Cancers**, MDPI, v. 16, n. 11, p. 2006, 2024.

KANG, L. et al. Pdia4 confers resistance to ferroptosis via induction of atf4/slc7a11 in renal cell carcinoma. **Cell death & disease**, Nature Publishing Group UK London, v. 14, n. 3, p. 193, 2023.

KHAN, A. et al. Using permutation-based feature importance for improved machine learning model performance at reduced costs. **IEEE Access**, IEEE, 2025.

LI, X. et al. Interpretable prediction of 30-day mortality in patients with acute pancreatitis based on machine learning and shap. **BMC Medical Informatics and Decision Making**, Springer, v. 24, n. 1, p. 328, 2024.

LIU, J. et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. **Cell**, Elsevier, v. 173, n. 2, p. 400–416, 2018.

LIU, K. et al. Pan-cancer analysis of the prognostic and immunological role of anln: an onco-immunological biomarker. **Frontiers in genetics**, Frontiers Media SA, v. 13, p. 922472, 2022.

LIU, Q. et al. Comprehensive analysis of the prognostic, immunological, and diagnostic roles of sirt1 in pan-cancer and its validation in kirc. **Frontiers in Immunology**, Frontiers Media SA, v. 15, p. 1501867, 2025.

MAMATJAN, Y. Pan-cancer classification system with explainable ai interpretation: A feasibility study. In: IEEE. **2024 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.], 2024. p. 1–6.

MENG, D. et al. Arfgap1 inhibits mtorc1 lysosomal localization and activation. **The EMBO journal**, v. 40, n. 12, p. e106412, 2021.

MIAO, K. et al. Exploring explainable machine learning and shapley additive explanations (shap) technique to uncover key factors of hnsc cancer: an analysis of the best practices. **Biomedical Signal Processing and Control**, Elsevier, v. 89, p. 105752, 2024.

NAGY, Á.; MUNKÁCSY, G.; GYŐRFFY, B. Pancancer survival analysis of cancer hallmark genes. **Scientific reports**, Nature Publishing Group UK London, v. 11, n. 1, p. 6047, 2021.

PAN, Q. et al. Explainable machine learning model identified potential biomarkers in liver cancer survival prediction. **Biomedical Signal Processing and Control**, Elsevier, v. 96, p. 106504, 2024.

PEDUZZI, G. et al. Explainable machine learning identifies a polygenic risk score as a key predictor of pancreatic cancer risk in the uk biobank. **Digestive and Liver Disease**, Elsevier, v. 57, n. 4, p. 915–922, 2025.

QU, G.-Y. et al. Downregulation of sall1 and its association with poor prognosis and the immune cell landscape in clear cell renal cell carcinoma. **Frontiers in Oncology**, v. 15, p. 1624651, 2025.

RAINIO, O.; TEUHO, J.; KLÉN, R. Evaluation metrics and statistical tests for machine learning. **Scientific Reports**, Nature Publishing Group UK London, v. 14, n. 1, p. 6086, 2024.

RECHARLA, M. et al. Harnessing ai and machine learning for precision medicine: Advancements in genomic research, disease detection, and personalized healthcare. **American Journal of Psychiatric Rehabilitation**, v. 28, n. 1, p. 112–123, 2025.

SEO, H. et al. Explainable ai based feature selection in cancer rna-seq. **ICT Express**, Elsevier, 2025.

SHAH, H. H.; LODHI, S. K. Ai in personalized medicine: Tailoring treatment plans based on individual patient data. **Global Trends in Science and Technology**, v. 1, n. 1, p. 15–29, 2025.

SU, L. et al. Explainable cancer factors discovery: Shapley additive explanation for machine learning models demonstrates the best practices in the case of pancreatic cancer. **Pancreatology**, Elsevier, v. 24, n. 3, p. 404–423, 2024.

SU, X. et al. Molecular signature of immune-related new survival predictions for subtype of renal cell carcinomas. **Translational Andrology and Urology**, v. 13, n. 10, p. 2180, 2024.

TABIBU, S.; VINOD, P.; JAWAHAR, C. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. **Scientific reports**, Nature Publishing Group UK London, v. 9, n. 1, p. 10509, 2019.

TRAN, D. et al. A comprehensive review of cancer survival prediction using multi-omics integration and clinical variables. **Briefings in bioinformatics**, Oxford University Press, v. 26, n. 2, p. bbaf150, 2025.

WANG, Q. et al. Assessing the role of programmed cell death signatures and related gene top2a in progression and prognostic prediction of clear cell renal cell carcinoma. **Cancer Cell International**, Springer, v. 24, n. 1, p. 164, 2024.

XIE, D. et al. The potential value of the purinergic pathway in the prognostic assessment and clinical application of kidney renal clear cell carcinoma. **Aging (Albany NY)**, v. 16, n. 1, p. 246, 2024.

YANG, C.; RANGARAJAN, A.; RANKA, S. Global model interpretation via recursive partitioning. In: IEEE. **2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)**. [S.l.], 2018. p. 1563–1570.

YAP, M. et al. Verifying explainability of a deep learning tissue classifier trained on rna-seq data. **Scientific reports**, Nature Publishing Group UK London, v. 11, n. 1, p. 2641, 2021.

ZHAI, M. et al. Predictive role of hpgd gene in carcinogenesis and immune environment monitoring in human cervical cancer. **Cancer Biomarkers**, SAGE Publications Sage UK: London, England, v. 41, n. 3-4, p. 18758592241296277, 2024.

ZHANG, H. et al. Interleukin 20 receptor subunit beta (il20rb) predicts poor prognosis and regulates immune cell infiltration in clear cell renal cell carcinoma. **BMC Genomic Data**, Springer, v. 23, n. 1, p. 58, 2022.

ZHANG, X. et al. Bioinformatic analysis identifying fgf1 gene as a new prognostic indicator in clear cell renal cell carcinoma. **Cancer Cell International**, Springer, v. 21, n. 1, p. 222, 2021.

ZHAO, Y.; SHAO, J.; ASMANN, Y. W. Assessment and optimization of explainable machine learning models applied to transcriptomic data. **Genomics, Proteomics & Bioinformatics**, Oxford University Press, v. 20, n. 5, p. 899–911, 2022.

ZHOU, K. et al. Slfn11: a pan-cancer biomarker for dna-targeted drugs sensitivity and therapeutic strategy guidance. **Frontiers in Oncology**, v. 15, p. 1582738, 2025.

ZHOU, Y. et al. Phosphatidylserine-specific phospholipase a1 limits aggressiveness of lung adenocarcinoma by lysophosphatidylserine and protein kinase a–dependent pathway. **The American Journal of Pathology**, Elsevier, v. 192, n. 6, p. 970–983, 2022.

ZHU, W. et al. Identification of a novel nine-snorna signature with potential prognostic and therapeutic value in ovarian cancer. **Cancer medicine**, Wiley Online Library, v. 11, n. 10, p. 2159–2170, 2022.

APÊNDICE A — RESULTADOS DAS TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE - PCA, UMAP E T-SNE

Resultado da visualização após redução de dimensionalidade dos conjuntos de dados KIRC A.1 e KIRP A.2. Foram utilizados três métodos (PCA, UMAP e t-SNE), porém, não foi encontrada uma diferenciação clara entre os grupos nas imagens.

Figura A.1: Distribuição das amostras após redução de dimensionalidade com PCA, UMAP e t-SNE - KIRC

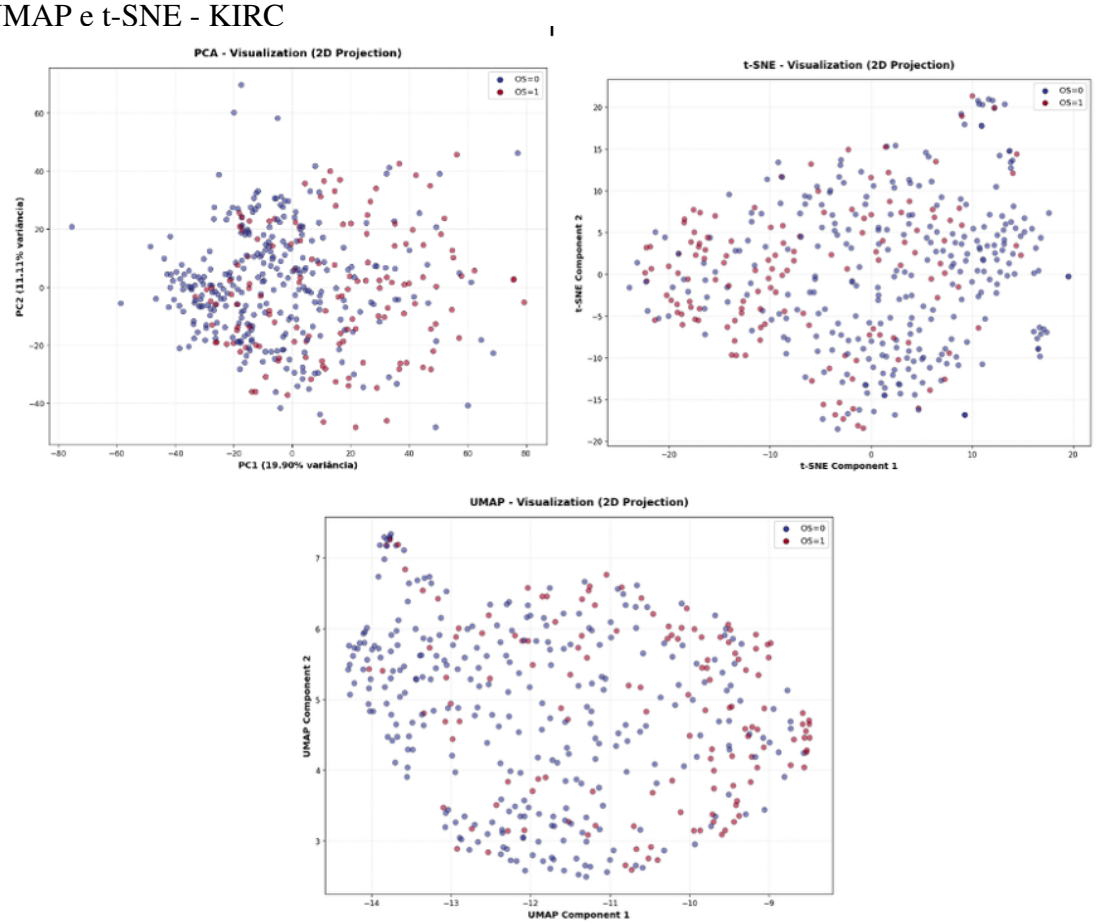
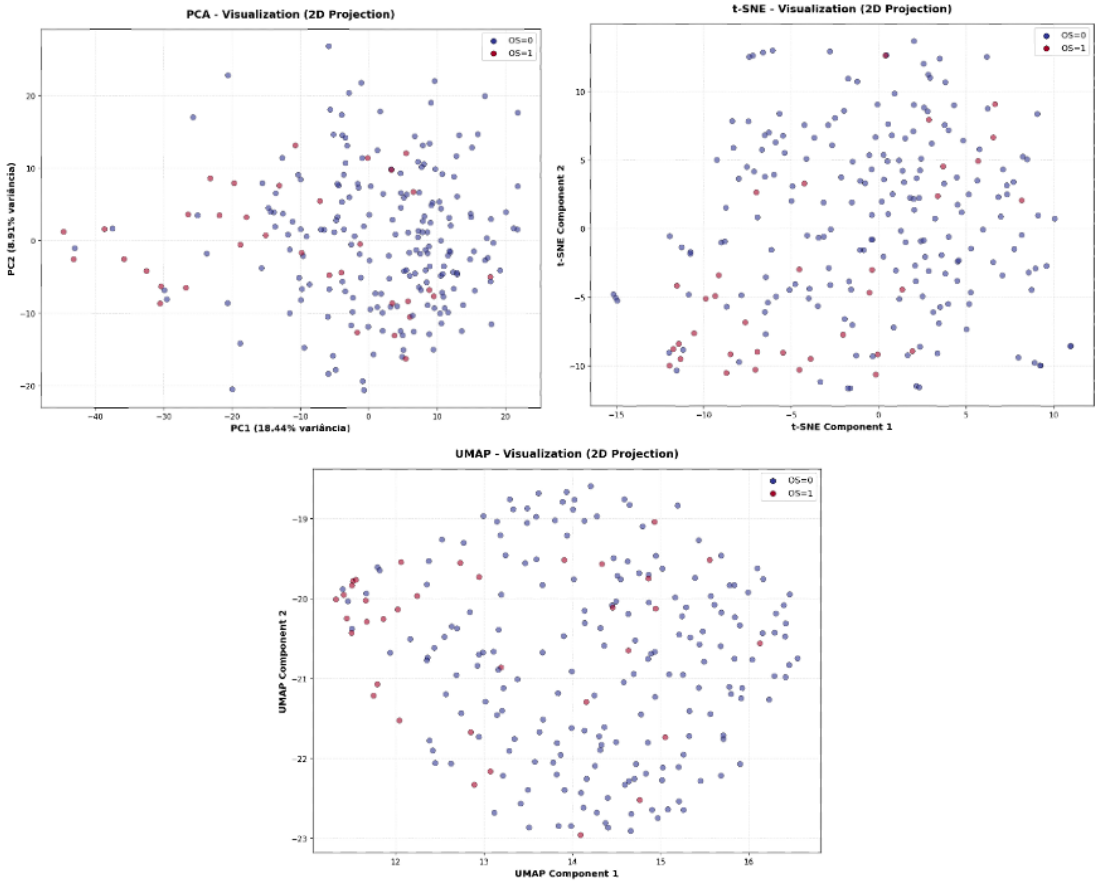


Figura A.2: Distribuição das amostras após redução de dimensionalidade com PCA, UMAP e t-SNE - KIRP



APÊNDICE B — RESULTADOS COMPLETOS SHAP, PFI, ADS

Resultados completos das análises XAI para cada modelo de ML.

B.1 SHAP

Resultados SHAP por modelo de ML.

Figura B.1: Resultados SHAP nos modelos Randon Forest e Gradient Boosting - KIRC

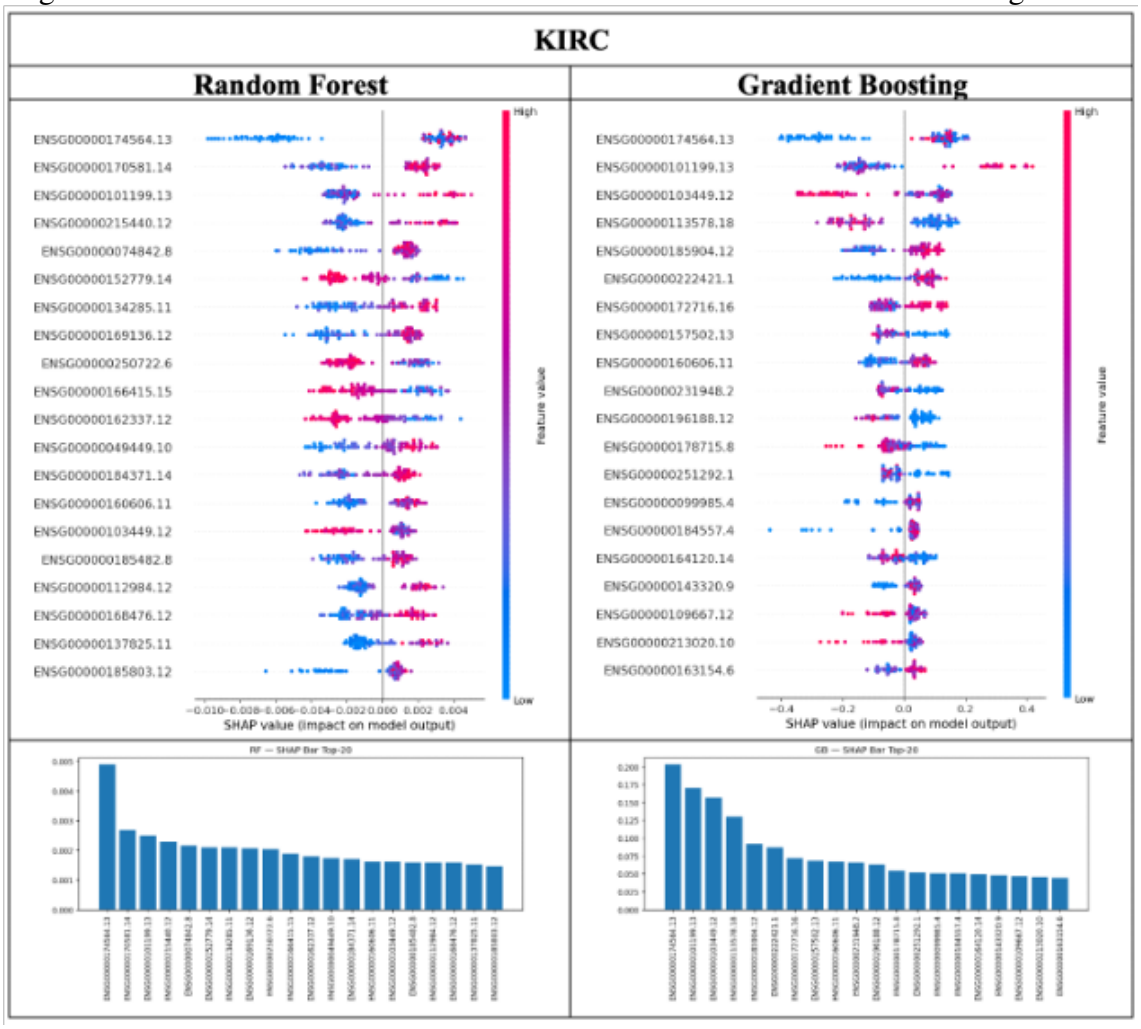


Figura B.2: Resultados SHAP no modelo Multilayer Perceptron - KIRC

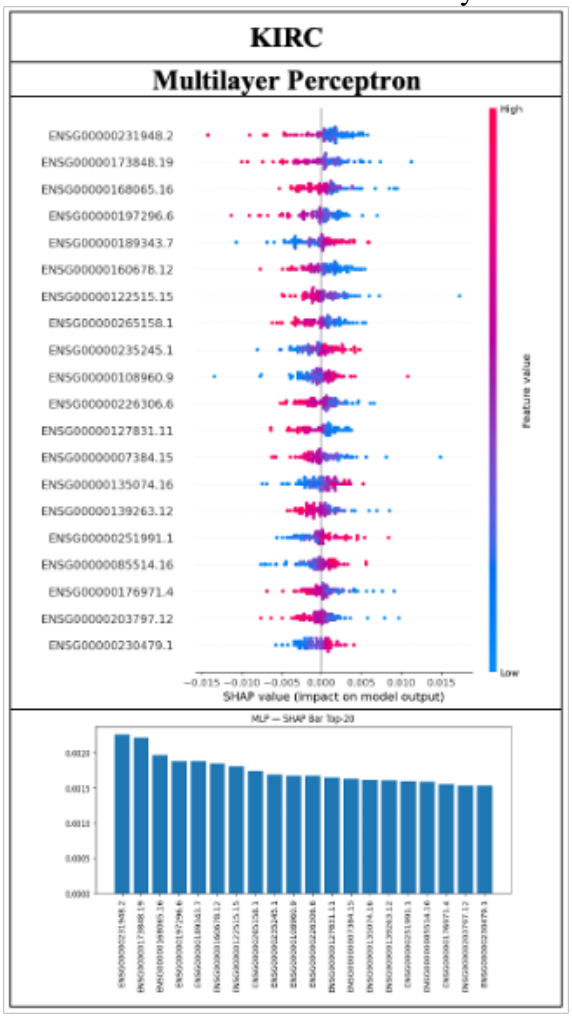


Figura B.3: Resultados SHAP nos modelos Regressão Logística e Categorical Boosting - KIRC

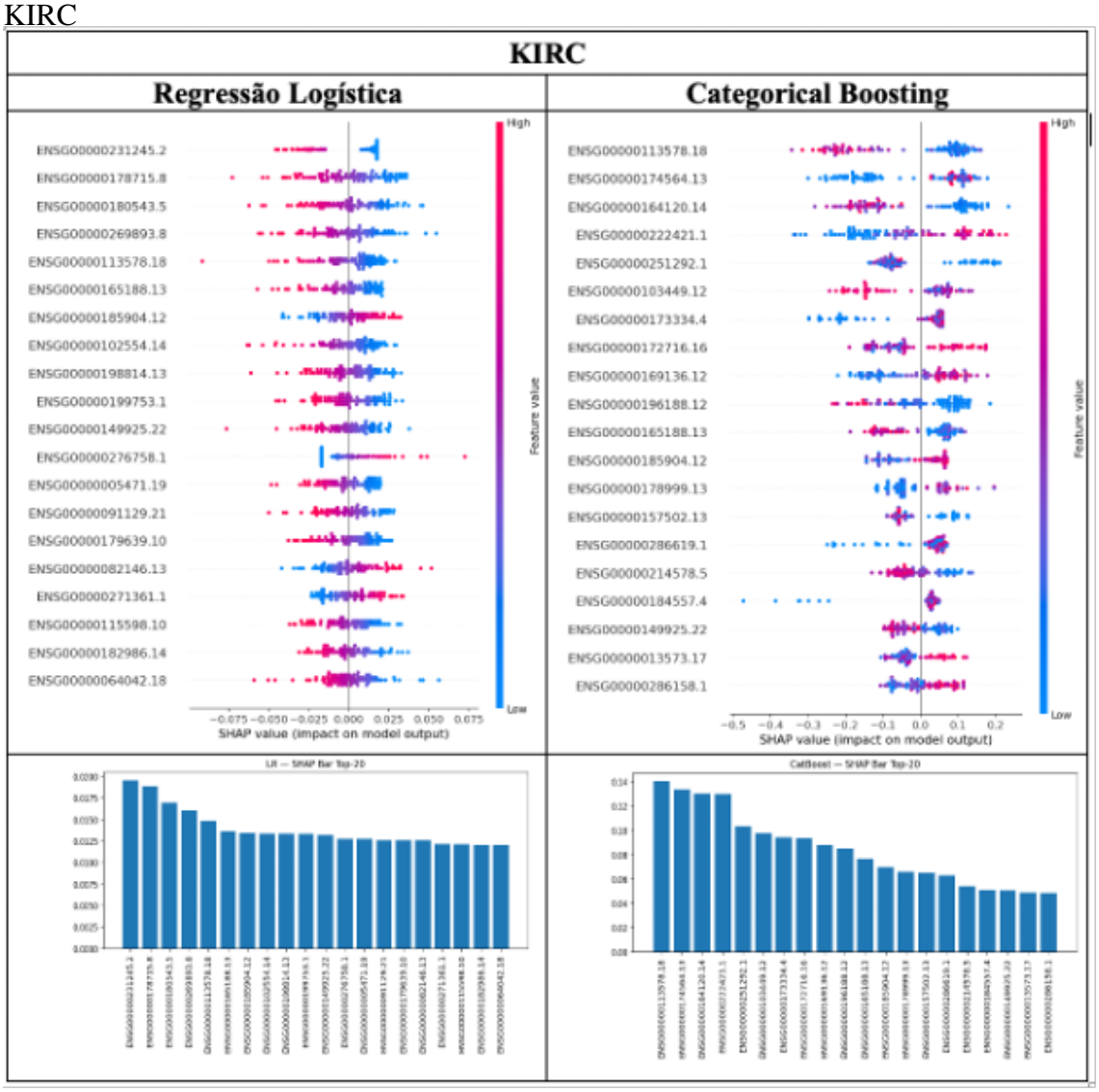


Figura B.4: Resultados SHAP nos modelos Randon Forest e Gradient Boosting - KIRP

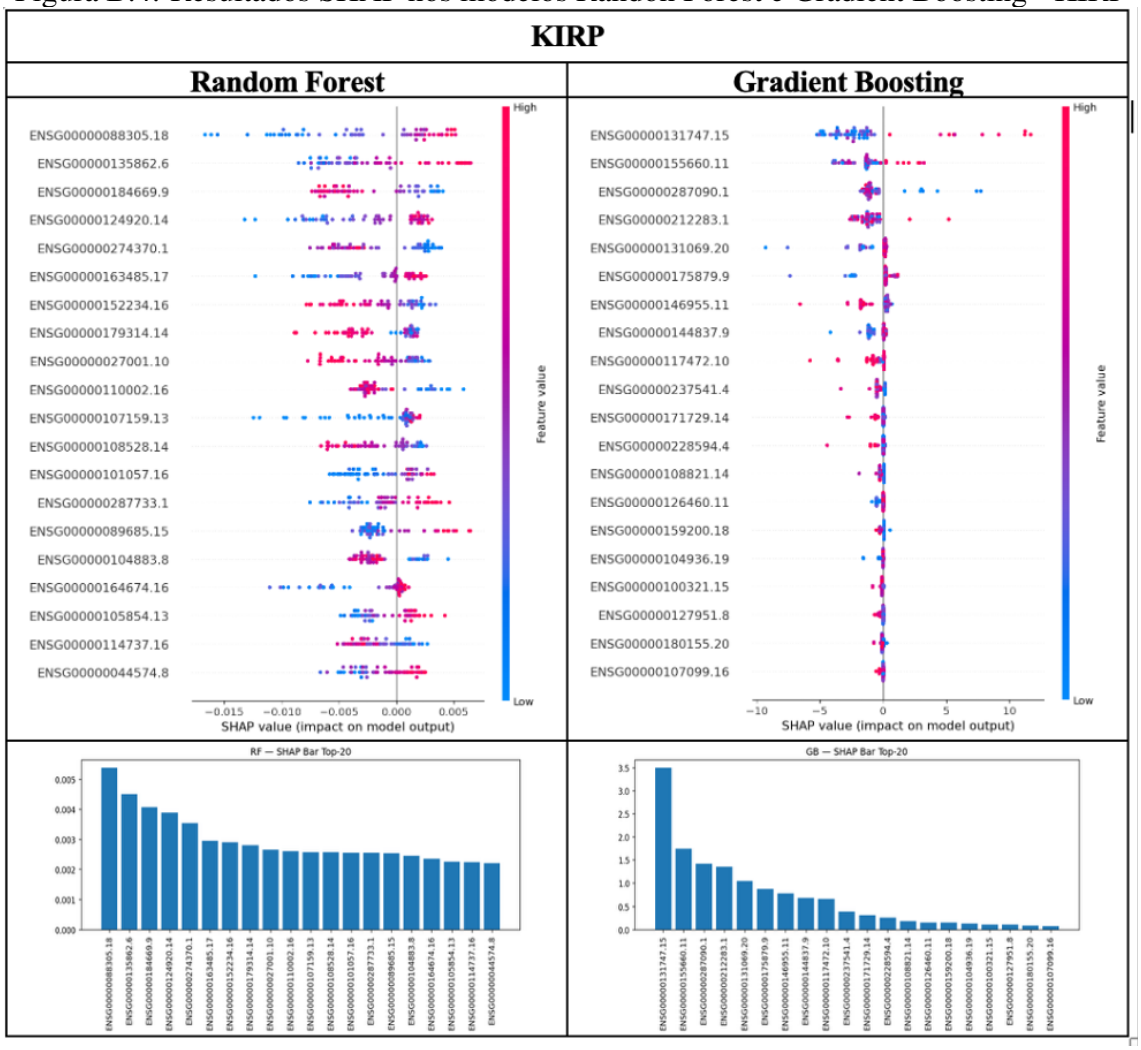


Figura B.5: Resultados SHAP no modelo Multilayer Perceptron - KIRP

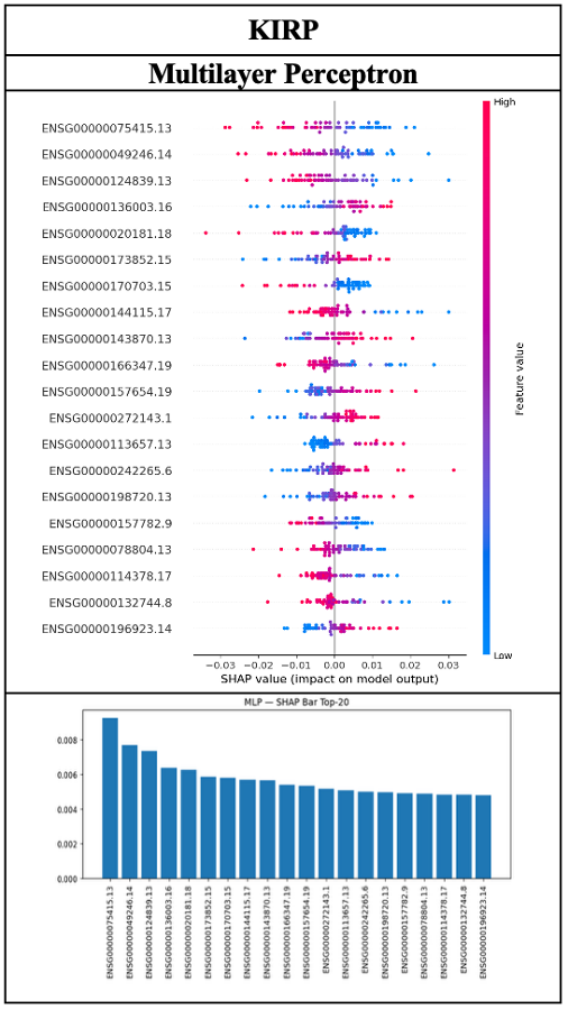
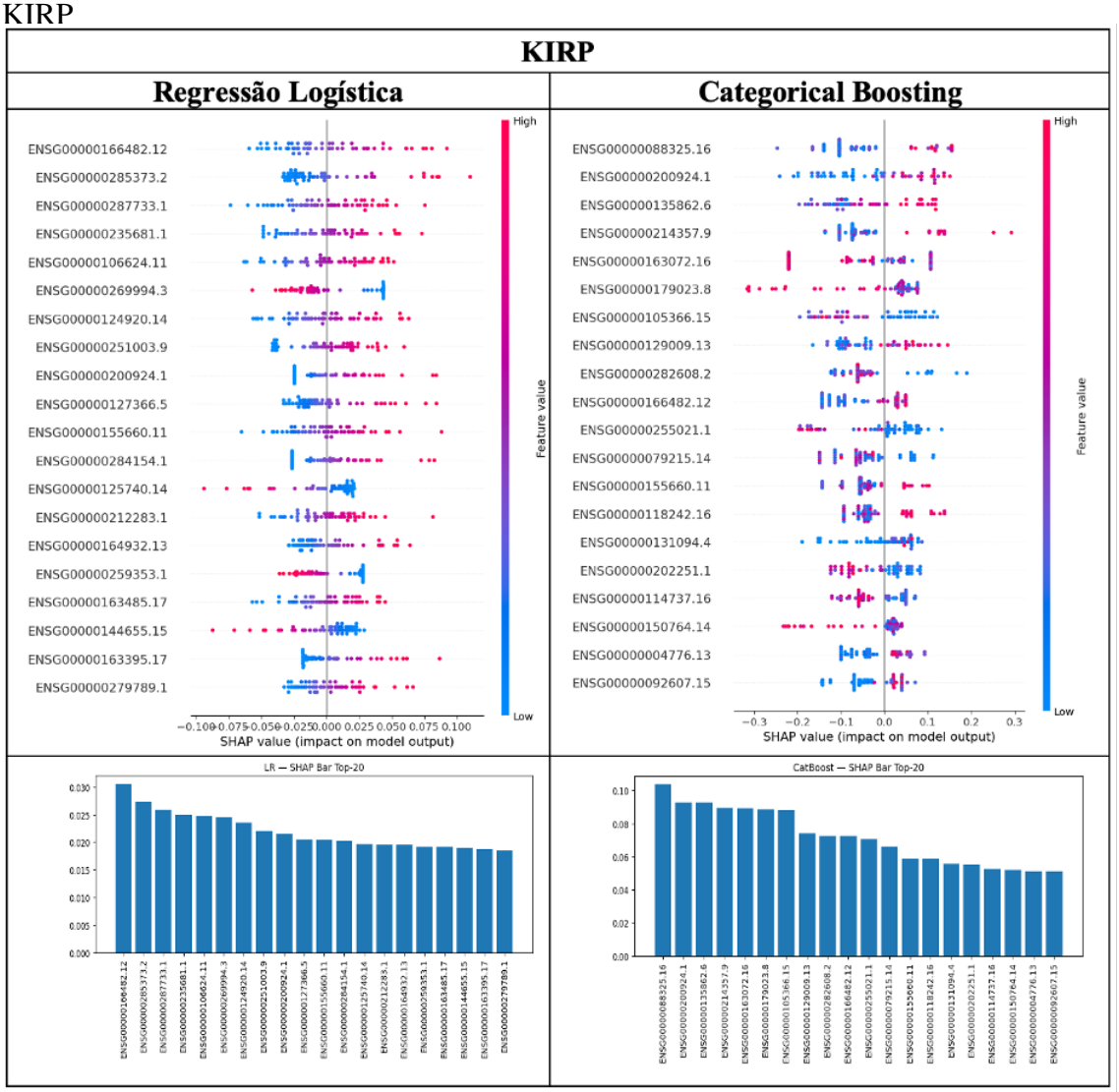


Figura B.6: Resultados SHAP nos modelos Regressão Logística e Categorical Boosting - KIRP



B.2 PFI

Resultado dos genes selecionados pelo *Permutation Feature Importance* por modelo de ML.

Tabela B.1: Resultados das análises de PFI para o conjunto de dados KIRC

KIRC - PFI				
RF - PFI	**GB - PFI**	**MLP - PFI**	**LR - PFI**	**Cat - PFI**
ENSG00000159403	ENSG00000174564	ENSG00000111859	ENSG00000143320	ENSG00000138669
ENSG00000104808	ENSG00000178715	ENSG00000230258	ENSG00000204161	ENSG00000170743
ENSG00000173432	ENSG00000160606	ENSG00000159217	ENSG00000174564	ENSG00000222421
ENSG00000166415	ENSG00000119147	ENSG00000231245	ENSG00000132849	ENSG00000072657
ENSG00000049449	ENSG00000276758	ENSG00000177459	ENSG00000231245	ENSG00000172183
ENSG00000162337	ENSG00000117595	ENSG00000138772	ENSG00000141469	ENSG00000233005
ENSG00000197081	ENSG00000163154	ENSG00000135678	ENSG00000271943	ENSG00000136383
ENSG00000207523	ENSG00000164120	ENSG00000197083	ENSG00000199753	ENSG00000179954
ENSG00000037749	ENSG00000157502	ENSG00000181218	ENSG00000260457	ENSG00000144857
ENSG00000213465	ENSG00000172716	ENSG00000225298	ENSG00000257335	ENSG00000072818
gender.demographic	ENSG00000204894	ENSG00000160179	ENSG00000064886	ENSG00000217801
ENSG00000144891	ENSG00000286619	ENSG00000127533	ENSG00000256269	ENSG00000189171
ENSG00000286064	ENSG00000168961	ENSG00000124749	ENSG00000236969	ENSG00000064989
ENSG00000164687	ENSG00000222421	ENSG00000198932	ENSG00000163328	ENSG00000105122
ENSG00000235245	ENSG00000101199	ENSG00000227121	ENSG00000235173	ENSG00000143878
ENSG00000153993	ENSG00000126016	ENSG00000271335	ENSG00000183397	ENSG00000011028
ENSG00000223552	ENSG00000245694	ENSG00000138669	ENSG00000257093	ENSG00000266104
ENSG00000173482	ENSG00000185904	ENSG00000183742	ENSG00000113578	ENSG00000126264
ENSG00000078114	ENSG00000103742	ENSG00000167371	ENSG00000082146	ENSG00000265982
ENSG00000198837	ENSG00000213903	ENSG00000136158	ENSG00000131080	ENSG00000286064

Tabela B.2: Resultados das análises de PFI para o conjunto de dados KIRP

KIRP - PFI				
RF - PFI	**GB - PFI**	**MLP - PFI**	**LR - PFI**	**Cat - PFI**
ENSG00000104883	ENSG00000131747	ENSG00000115009	ENSG00000213793	ENSG00000229939
ENSG00000089685	ENSG00000155660	ENSG00000162998	ENSG00000123146	ENSG00000100218
ENSG00000131747	ENSG00000117472	ENSG00000213199	ENSG00000270179	ENSG00000135245
ENSG00000184669	ENSG00000131069	ENSG00000132744	ENSG00000135862	ENSG00000155660
ENSG00000152234	ENSG00000144837	ENSG00000204388	ENSG00000085831	ENSG00000183775
ENSG00000144837	ENSG00000142661	ENSG00000276980	ENSG00000163395	ENSG00000161888
ENSG00000106624	ENSG00000171729	ENSG00000267390	ENSG00000184669	ENSG00000079215
ENSG00000175879	ENSG00000271869	ENSG00000150764	ENSG00000121039	ENSG00000266714
ENSG00000163485	ENSG00000175879	ENSG00000132561	ENSG00000100196	ENSG00000278921
ENSG00000154065	ENSG00000228594	ENSG00000242265	ENSG00000102007	ENSG00000272405
ENSG00000027001	ENSG00000212283	ENSG00000165030	ENSG00000101298	ENSG00000137033
ENSG00000149596	ENSG00000100321	ENSG00000127084	ENSG00000112499	ENSG00000144381
ENSG00000088305	ENSG00000123975	ENSG00000198535	ENSG00000179820	ENSG00000171552
ENSG00000101057	ENSG00000237541	ENSG00000215915	ENSG00000092607	ENSG00000139354
ENSG00000179314	ENSG00000287090	ENSG00000163359	ENSG00000110455	ENSG00000202251
ENSG00000110002	ENSG00000104936	ENSG00000151715	ENSG00000131203	ENSG00000162981
ENSG00000108528	ENSG00000137875	ENSG00000163071	ENSG00000241119	ENSG00000160799
ENSG00000124920	gender.demographic	ENSG00000202251	ENSG00000127946	ENSG00000123179
ENSG00000112149	ENSG00000110446	ENSG00000162745	ENSG00000116031	ENSG00000125877
ENSG00000133116	ENSG00000165140	ENSG00000185187	ENSG00000274370	ENSG00000144369

B.3 ADS

Resultados do método de árvore de decisão surrogada por modelo de ML.

Figura B.7: Resultados ADS para o conjunto de dados KIRC

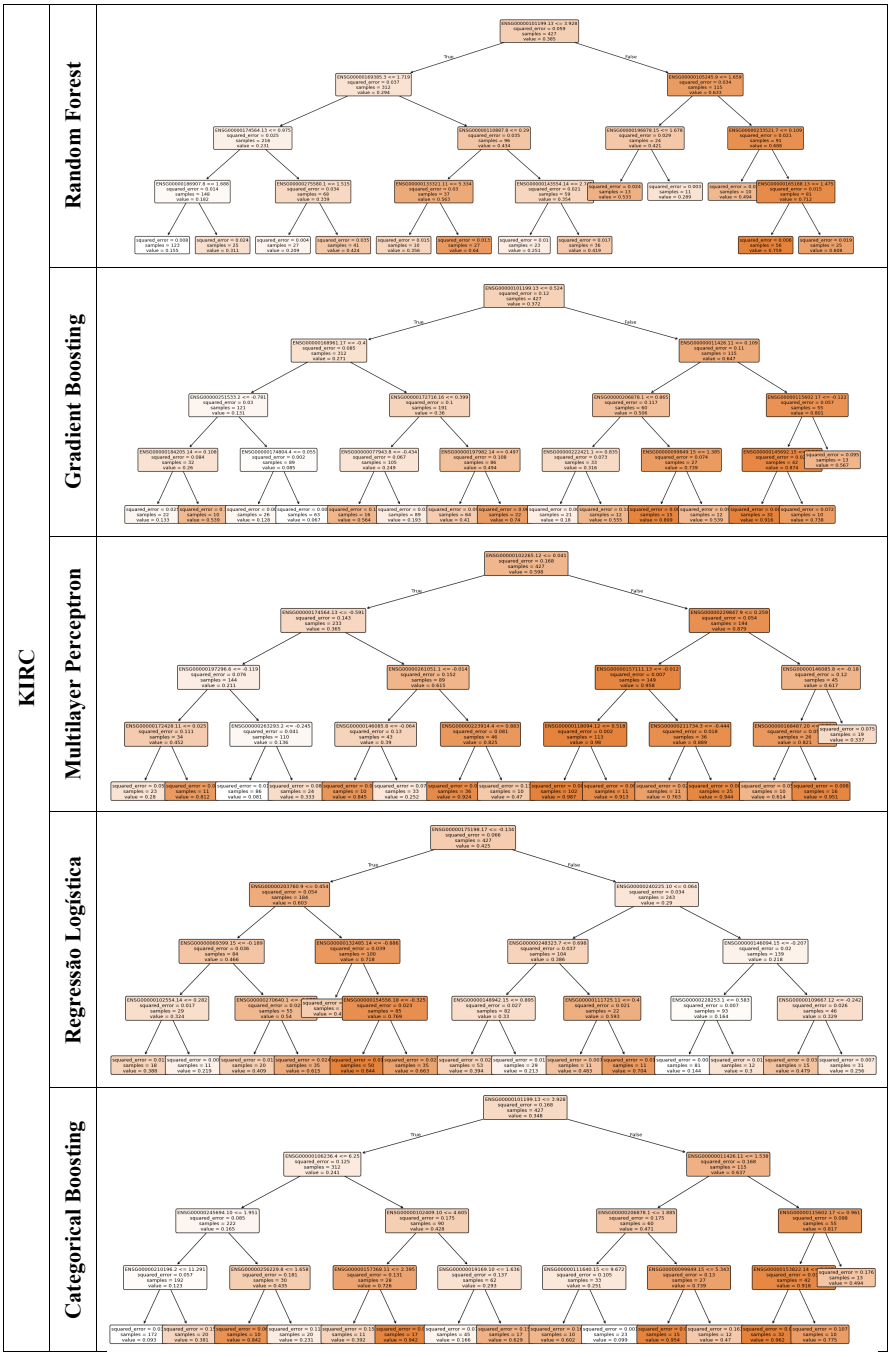
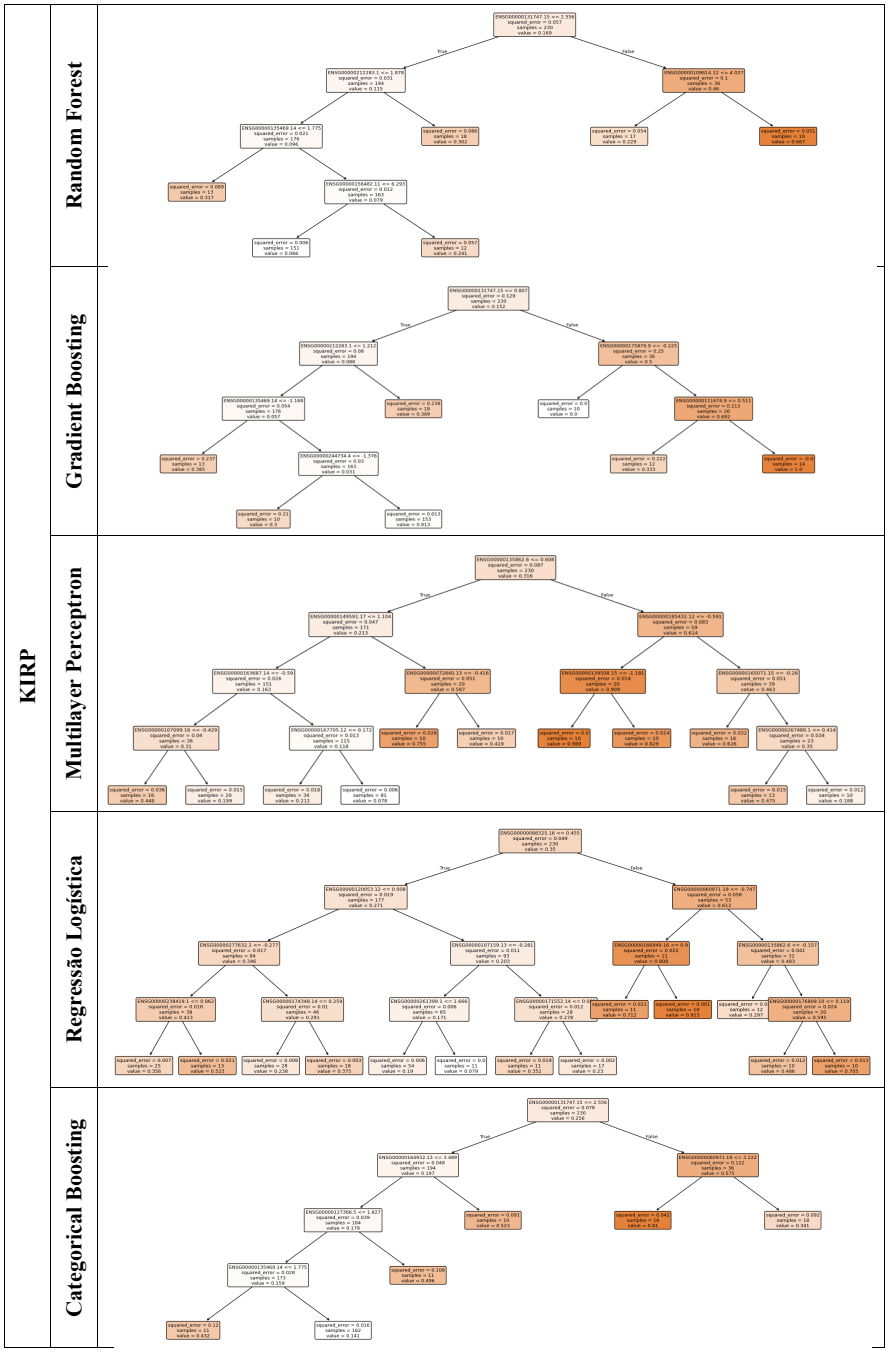


Figura B.8: Resultados ADS para o conjunto de dados KIRP



APÊNDICE C — FREQUÊNCIA DE GENES POR MÉTODO XAI - TABELA

Tabela com a distribuição dos 30 genes principais segundo os métodos de XAI aplicados, detalhando frequências individuais e totais.

Tabela C.1: Frequência, por método de XAI, dos 30 genes mais encontrados.

Gene	SHAP	PFI	ADS	TOTAL
ENSG00000131747.15	1	2	3	6
ENSG00000155660.11	3	2	0	5
ENSG00000135862.6	2	1	2	5
ENSG00000212283.1	2	1	2	5
ENSG00000175879.9	1	2	1	4
ENSG00000124920.14	2	1	0	3
ENSG00000163485.17	2	1	0	3
ENSG00000144837.9	1	2	0	3
ENSG00000184669.9	1	2	0	3
ENSG00000202251.1	1	2	0	3
ENSG00000135469.14	0	0	3	3
ENSG00000114737.16	2	0	0	2
ENSG00000166482.12	2	0	0	2
ENSG00000200924.1	2	0	0	2
ENSG00000287733.1	2	0	0	2
ENSG00000027001.10	1	1	0	2
ENSG00000079215.14	1	1	0	2
ENSG00000088305.18	1	1	0	2
ENSG00000089685.15	1	1	0	2
ENSG00000092607.15	1	1	0	2
ENSG00000100321.15	1	1	0	2
ENSG00000101057.16	1	1	0	2
ENSG00000104883.8	1	1	0	2
ENSG00000104936.19	1	1	0	2
ENSG00000106624.11	1	1	0	2
ENSG00000108528.14	1	1	0	2
ENSG00000110002.16	1	1	0	2
ENSG00000117472.10	1	1	0	2
ENSG00000131069.20	1	1	0	2
ENSG00000132744.8	1	1	0	2