Giovanna Aguiar de Castro

Sistema de suporte à decisão para a escolha do protocolo terapêutico para pacientes com leucemia mieloide aguda

Sorocaba, SP

17 de fevereiro de 2023

Giovanna Aguiar de Castro

Sistema de suporte à decisão para a escolha do protocolo terapêutico para pacientes com leucemia mieloide aguda

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCC-So) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Linha de pesquisa: Computação Científica e Inteligência Computacional.

Universidade Federal de São Carlos – UFSCar

Centro de Ciências em Gestão e Tecnologia – CCGT

Programa de Pós-Graduação em Ciência da Computação – PPGCC-So

Orientador: Prof. Dr. Tiago A. Almeida

Sorocaba, SP

17 de fevereiro de 2023

Castro, Giovanna A.

Sistema de suporte à decisão para a escolha do protocolo terapêutico para pacientes com leucemia mieloide aguda / Giovanna A. Castro -- 2023.
122f.

Dissertação (Mestrado) - Universidade Federal de São Carlos, campus Sorocaba, Sorocaba
Orientador (a): Tiago Agostinho de Almeida
Banca Examinadora: Tiago Agostinho de Almeida, Ana Carolina Lorena, Ricardo Cerri
Bibliografia

1. Leucemia mieloide aguda. 2. Aprendizado de máquina. 3. Sistema de suporte à decisão. I. Castro, Giovanna A.. II. Título.

Ficha catalográfica desenvolvida pela Secretaria Geral de Informática (SIn)

DADOS FORNECIDOS PELO AUTOR

Bibliotecário responsável: Maria Aparecida de Lourdes Mariano - CRB/8 6979

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências em Gestão e Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Giovanna Aguiar de Castro, realizada em 17/02/2023.

Comissão Julgadora:

Prof. Dr. Tiago Agostinho de Almeida (UFSCar)

Profa. Dra. Ana Carolina Lorena (ITA)

Prof. Dr. Ricardo Cerri (UFSCar)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

Dedico este trabalho a minha mãe, Maria Ivelte, a primeira pessoa que acreditou no meu futuro por meio da pesquisa.

Agradecimentos

Agradeço,

a Deus, que me permitiu ter um verdadeiro encontro junto a ele e, principalmente, por me mostrar à todo momento o quão valiosas são as ações e práticas da vida, se feitas com carinho e dedicação.

a minha querida e amada mãe, Maria Ivelte, que incansáveis vezes colocou os sonhos e objetivos de seus filhos em primeiro lugar na sua vida. Em inúmeros momentos se tornou meu ponto de refúgio quando sempre o que pensava era desistir e não acreditava em mim. Em relação à família, não posso deixar de citar minha cachorrinha Ada, sempre presente e minha companheira de vida.

ao meu Professor e Orientador, Dr. Tiago A. Almeida, que me incentivou e arduamente acreditou no meu potencial para desenvolver e completar mais esta etapa de minha vida com êxito e dedicação. Sendo ele totalmente compreensivo e aplicado na pesquisa e nas suas orientações.

aos meus companheiros de pesquisa, Breno e, em especial, a Jade Almeida que sempre esteve presente na condução desta pesquisa. De certa forma, aprendi muito com o seu comprometimento e o seu jeito de realizar a pesquisa.

aos meus amigos da faculdade, Gislainy, Karine, Mariana e Marcus, por sempre me motivarem e apoiarem minha pesquisa. Responsáveis pelos pontos de descontração, me mostrando que a vida é baseada em um equilíbrio entre as atividades que propomos fazer.

a família Fávero (Ariane, Emílio e Giovanni), que serviu como suporte em uma nova cidade e que hoje são parte integrante da minha vida. A Ariane (Ari) de modo específico se tornou uma grande amiga presente e virou uma ajuda nos piores momentos da minha vida.

ao meu amigo, Vinícius Armele, que com o seu jeito descontraído de levar a vida me mostrou que as obrigações não precisam ser mais pesadas do que já são. Não posso esquecer de sua esposa Fabiana, que sempre se mostrou como uma pessoa sábia e objetiva.

ao meu gestor, Ivo Medeiros, que no final do mestrado serviu de grande ajuda e compreensão para que pudesse conciliar trabalho e pesquisa acadêmica. Além de ser um ponto de referência acadêmica e profissional.

a Universidade Federal de São Carlos, por oferecer um curso de pós-graduação de excelência com destaque (inter)nacional.

ao Prof. Dr. João A. Machado Neto, por disponibilizar os dados para a realização desta pesquisa e ter atuado como coorientador informal para a condução da mesma.

a todos aqui não citados, que de certa forma fizeram parte da minha trajetória de vida para a construção deste momento.

“ Minha cabeça e meu coração parecem me pregar uma peça cruel, me deixando com a falsa sensação de juventude ao encarar o mundo todos os dias com os olhos idealistas e travessos de uma criança rebelde que enxerga valor e felicidade nas coisas mais básicas e simples.”

– Dave Grohl

Resumo

A Leucemia Mieloide Aguda é uma doença de caráter crônico e incapacitante. Ela é heterogênea e se apresenta de diferentes formas. Após o diagnóstico, o paciente recebe um prognóstico de risco costumeiramente dividido em três grupos: favorável, intermediário e adverso. Esta classificação é frequentemente utilizada por especialistas para auxiliá-los na personalização de decisões terapêuticas. Nas últimas décadas, o tratamento padrão tem sido terapia intensiva com a combinação dos medicamentos citarabina e antraciclina. A classificação de risco atual é conservadora e frequentemente requer que os especialistas recorram a mais informações, como resultados de outros exames e análises, para selecionar o tratamento adequado, ainda que com pouca ou nenhuma evidência de eficácia. Esse processo pode ocasionar no atraso no início do tratamento e no agravamento no quadro clínico do paciente. Neste trabalho, foram realizados estudos na literatura a fim de entender o comportamento da doença e suas implicações. Especificadamente, foi realizado um mapeamento sistemático da literatura que buscou categorizar e analisar os guias de tratamentos existentes e, principalmente, descobrir, se existe algum que emprega técnicas de Aprendizado de Máquina. Os protocolos terapêuticos foram agrupados conforme a intensidade de cada tratamento existente. A doença se manifesta de formas diferentes nos pacientes e, por isso, é difícil decidir um curso terapêutico genérico. Dessa forma, as estratégias terapêuticas têm se tornado cada vez mais personalizadas e isoladas para realidades clínicas individualizadas. Existem diversas variáveis que podem influenciar a escolha de guias de tratamento, como a idade do paciente, recaídas e medicamentos inibidores de ações proteicas. Contudo, a combinação destes e outros critérios é dificultada em uma análise manual, principalmente quando dados genéticos são empregados. Por isso, é importante recorrer a ferramentas que possam realizar essas análises de forma automatizada, a fim de auxiliar os especialistas na escolha dos melhores protocolos de tratamento para seus pacientes. Neste contexto, este trabalho propõe um sistema de suporte à decisão que visa recomendar automaticamente conjuntos de tratamentos adequados para pacientes portadores de leucemia mieloide aguda em função da predição automática da mortalidade/sobrevivência. Para isso, foram utilizados dados de natureza clínica e genética (mutação e expressão gênica) provenientes de duas bases dados de domínio público. O sistema proposto é formado por dois comitês de classificação compostos pelos melhores modelos de predição obtidos. Os resultados indicam que o sistema proposto é promissor e pode ser utilizado como ferramenta de apoio à decisão dos especialistas, com o potencial de reduzir a subjetividade e o tempo nos processos de escolha de tratamentos, resultando em recomendações mais assertivas, com menos efeitos adversos, que poderá contribuir para aumentar o tempo de sobrevida e a qualidade de vida dos pacientes.

Palavras-chave: leucemia mieloide aguda; aprendizado de máquina, sistema de suporte à

decisão.

Abstract

Acute Myeloid Leukemia is a chronic and disabling disease. It is heterogeneous and presents in different ways. After the diagnosis, the patient receives a risk prognosis of outcomes usually divided into three groups: favorable, intermediate, and adverse. Specialists frequently use this classification to customize therapeutic decisions. In recent decades, the standard treatment has been intensive therapy with a combination of cytarabine and anthracycline. Current risk classification is conservative and often requires specialists to resort to more information, such as the results of other exams and analyses, to select the appropriate treatment, even with little or no evidence of efficacy. This process can delay the start of treatment and worsen the patient's clinical status. In this study, we have investigated the behavior of the disease and its implications. Specifically, we carried out a systematic mapping of the literature to categorize and analyze the existing treatment guides and, mainly, to find out if any employ Machine Learning techniques. The therapeutic protocols were grouped according to the intensity of each current treatment. The disease manifests in different ways in patients; therefore, deciding on a generic therapeutic course is challenging. Thus, therapeutic strategies have become increasingly personalized and isolated for individualized clinical realities. Several variables can influence the choice of treatment guidelines, such as the patient's age, relapses, and drugs that inhibit protein actions. However, combining these and other criteria is difficult in a manual analysis, especially when genetic data are used. Therefore, it is important to resort to tools that can perform these analyzes automatically to assist specialists in choosing the best treatment protocols for their patients. In this context, this study proposes a decision support system that aims to automatically recommend sets of appropriate treatments for patients with acute myeloid leukemia by automatically predicting their mortality/survival. We used clinical and genetic data (mutation and gene expression) from two public-domain databases. The proposed system is formed by two classification committees composed of the best prediction models obtained. The results indicate that the proposed system is promising and can be used as a decision support tool for specialists, with the potential to reduce subjectivity and time in the processes of choosing treatments, resulting in more strong recommendations with fewer adverse effects, which may contribute to increasing the survival and quality of life of patients.

Keywords: acute myeloid leukemia; machine learning; decision support system.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

3-NN	*3-Nearest Neighboors*
ALL	*Acute Lymphoblastic Leukemia*
AML	*Acute Myeloid Leukemia*
ANOVA	*Analysis of Variance*
AUC	*Area under the ROC curve*
BM	*Bone Marrow*
CE	Critério de exclusão
CI	Critério de inclusão
CHI-2	*Chi-square test*
CLIN	Dados clínicos
CV	*Cross-Validation*
EI	Endocardite Infecciosa
ELN	*European Leukemia Net*
EXP	Dados de expressão gênica
FAB	*French, American and British research group*
FN	Falso negativo
FP	Falso positivo
GS	*Grid Search*
H0	Hipótese nula
H1	Hipótese alternativa
HIT	*High-intensity therapy*
KDD	*Knowledge Discovery in Databases*
LIT	*Low-intensity therapy*

LOO	*Leave-one out*
LR	*Logistic Regression*
MDS	*Myelodysplastic syndromes*
ML	*Machine Learning*
MPN	*Myeloproliferative neoplasms*
MSL	Mapeamento Sistemático da Literatura
MUT	Dados de mutação gênica
OHSU	*Oregon Health and Science University*
PB	*Peripheral Blood*
PICOC	*Population, Intervention, Comparison, Outcomes and Context*
QP	Questão de Pesquisa
RF	*Random Forests*
ROC	*Recieve operation characteristic*
RSL	Revisão Sistemática da Literatura
RT	*Regular therapy*
SVM	*Support Vector Machine*
TCGA	*The Cancer Genome Atlas Program*
TMB	*The total number of mutations (changes) found in the DNA of cancer cells*
TT	*Target therapy*
VN	Verdadeiro negativo
VP	Verdadeiro positivo
WBC	*White Blood Count*

Sumário

1 Introdução

A Leucemia Mieloide Aguda (do inglês, *Acute Myeloid Leukemia* – AML) é um tipo de câncer que atua no sangue, caracterizado pela infiltração de células cancerígenas na medula óssea. A AML tem caráter crônico e possui taxas de remissão decrescentes em relação à idade do paciente. A taxa mediana de sobrevida global é apenas de 12 a 18 meses após o diagnóstico (PELCOVITS; NIROULA, 2020; ROSE-INMAN; KUEHL, 2014).

Em geral, o desenvolvimento da AML é incapacitante para o paciente e os tratamentos envolvem quimioterapia intensiva, com uma alta taxa de mortalidade em resposta a este tipo de terapia. A combinação dos medicamentos citarabina e antraciclina tem sido nas últimas décadas utilizada como padrão de tratamento (KIM et al., 2015). O uso destes medicamentos é feito de forma intravenosa e envolve internações de 3 a 4 semanas, normalmente causando desconfortos aos pacientes (BRYANT et al., 2020).

Em 2010, o grupo internacional de pesquisa *European LeukemiaNet* (ELN) publicou recomendações para o diagnóstico e tratamento da AML (DÖHNER et al., 2010). Estas recomendações foram amplamente adotadas por médicos e especialistas da área. Já, em 2017, houve uma atualização dessas recomendações, incorporando novas descobertas sobre o comportamento da doença (DÖHNER et al., 2017). A sua última atualização foi realizada em 2022, considerando novas descobertas genéticas a respeito do curso da doença (DÖHNER et al., 2022).

Para um diagnóstico de AML, é preciso que haja ao menos 10% de mieloblastos (tipo de célula sanguínea) na medula óssea ou sangue periférico (ARBER et al., 2022). Esta análise é realizada conforme a Classificação de Tumores de Hematopoiéticos e Linfoides (do inglês, *Classification of Tumours of Haematopoietic and Lymphoid Tissues*), divulgada e atualizada pela Organização Mundial da Saúde. A sua última atualização foi realizada no ano de 2022 referente a subtipos da doença, como AML com mielodisplasias relacionadas à mutação gênica ou alterações citogenéticas.

Além do diagnóstico, o paciente com AML recebe um prognóstico de risco da doença, dividido em três classes: favorável, intermediário e adverso. Tal divisão é dada por características citogenéticas (relacionadas a estrutura das células) e moleculares (Cancer Genome Atlas Research Network et al., 2013). As características citogenéticas são oriundas de certas alterações cromossômicas. As moleculares, por sua vez, são determinadas pelo valor da expressão do gene *FLT3-ITD* e análise do número de mutações nos genes *NPM1*, *RUNX1*, *ASXL1*, *TP53*, *BCOR*, *EZH2*, *SF3B1*, *SRSF2*, *STAG2* e *ZRSR2*. A classificação de risco da ELN é vastamente usada por especialistas para tomar decisões importantes sobre o curso de cada tratamento e, consequentemente, pode impactar diretamente na

qualidade e expectativa de vida dos pacientes.

Os grupos de risco da ELN são frequentemente utilizados como guias de tratamentos. No entanto, em sua definição não há a consideração da idade do paciente. Além disso, pacientes classificados em um mesmo grupo de risco podem apresentar respostas terapêuticas diferentes em relação a um mesmo medicamento. Nesse sentido, os especialistas acabam recorrendo a outros tipos de análises e considerações para definir o tratamento adequado aos seus pacientes.

Pacientes com prognóstico de risco favorável costumam apresentar uma boa resposta ao tratamento quimioterápico (KADIA et al., 2015). Por outro lado, os que se enquadram na classe de risco adversa tendem a não apresentar um bom resultado em relação a este tipo de terapia, e, consequentemente, recorrem a outros tratamentos, como transplante da medula óssea (Cancer Genome Atlas Research Network et al., 2013). O problema da classificação de risco atual é que há grande heterogeneidade de pacientes em um mesmo grupo de risco. Além disso, não há uma definição clara quanto à classe de risco intermediário, uma vez que esses pacientes não apresentam um padrão de resposta aos tratamentos (KADIA et al., 2015).

A maioria dos pacientes com AML recebe o prognóstico de risco intermediário (DÖHNER et al., 2010), o que faz com que os especialistas precisem de mais informações, como resultados de outros exames e análises, para selecionar o tratamento adequado, ainda que com pouca ou nenhuma evidência de eficácia. Esse processo pode ocasionar no atraso no início do tratamento e no possível agravamento no quadro clínico do paciente.

Com o avanço recente no estudo sobre AML, diversos dados de expressão gênica dos pacientes passaram a ser disponibilizados (Cancer Genome Atlas Research Network et al., 2013). Contudo, a quantidade massiva de informações dificulta a análise e avaliação dos especialistas, o que demanda soluções automatizadas para explorar, confirmar e/ou descobrir padrões nesses dados. Neste cenário, técnicas de Aprendizado de Máquina (do inglês, *Machine Learning* – ML) têm se mostrado cada vez mais eficazes na descoberta de padrões e relações em grandes bases de dados, sendo que muitas delas vêm se popularizando na área médica por conta do seu alto potencial em suportar decisões dos especialistas acerca de diagnósticos, prognósticos e tratamentos (KOUROU et al., 2015). Por meio de métodos de ML, é possível realizar uma análise profunda de conjuntos gigantescos de dados, uma vez que o processo pode ser automatizado e otimizado. Com a automatização, ainda é possível realizar a personalização de tratamentos a partir da descoberta de novos padrões. Adicionalmente, o uso de ML tem se mostrado eficiente na descoberta de padrões, predições de risco e predições de sobrevivência, em doenças como: câncer de nasofaringe (JING et al., 2020) e de pulmão (YU et al., 2021), doenças cardiovasculares (ROSS et al., 2019), esquizofrenia (LI et al., 2021) e COVID-19 (TANG et al., 2021b).

Neste contexto, esse trabalho assume a hipótese de que o emprego de métodos de ML

tem potencial para contribuir na descoberta de novos padrões gênicos e clínicos que possam auxiliar os especialistas em suas decisões terapêuticas. Para esse fim, com a colaboração de especialistas em AML[1], o principal objetivo deste trabalho é propor um sistema de apoio à decisão que visa recomendar automaticamente conjuntos de tratamentos adequados para pacientes com AML em função da predição automática da mortalidade/sobrevivência. Com isso, é esperado reduzir significativamente a subjetividade e o tempo nos processos de escolha de tratamentos, resultando em recomendações mais assertivas, com menos efeitos adversos, que poderá contribuir para aumentar o tempo de sobrevida e a qualidade de vida dos pacientes.

Essa dissertação está estruturada da seguinte forma:

- No Capítulo 2, são contextualizados os trabalhos relacionados a este por meio de revisões da literatura em conjunto com um mapeamento sistemático. Os trabalhos relacionados cobrem desde as primeiras definições de AML, seus prognósticos de risco, os guias de tratamento existentes para a doença e estudos que utilizaram técnicas de ML para gerar sistemas de suporte à decisão para outras doenças relacionadas;
- No Capítulo 3, é detalhado o processo de obtenção, pré-processamento e análise dos dados utilizados para criar o sistema de apoio a decisão proposto;
- No Capítulo 4, é apresentado uma descrição detalhada da proposta do sistema de suporte à decisão. O processo envolve a combinação de conjuntos de atributos para o treinamento de modelos individuais especializados, a composição de um comitê de classificação e a recomendação terapêutica;
- No Capítulo 5, são apresentados os resultados experimentais obtidos pelos modelos individuais e pelo comitê de classificação; e
- Por fim, no Capítulo 6, são apresentadas as conclusões, bem como os direcionamentos para estudos futuros.

[1] Este trabalho foi desenvolvido em colaboração com o grupo de pesquisa do Dr. João A. Machado Neto <jamachadoneto@usp.br>, professor do Departamento de Farmacologia, Instituto de Ciências Biomédicas da Universidade de São Paulo.

2 Trabalhos relacionados

A revisão bibliográfica inicia com a análise dos principais trabalhos introdutórios a respeito de AML e do seu prognóstico de risco (Seção 2.1). Em seguida, é apresentado um mapeamento sistemático da literatura para contextualizar os guias de tratamento existentes para AML (Seção 2.2). Por fim, a Seção 2.3 apresenta estudos que empregaram com sucesso técnicas de ML para auxiliar o prognóstico e diagnóstico de AML e de outras doenças relacionadas.

2.1 Classificação de risco no contexto da AML

A AML foi inicialmente subclassificada no ano de 1976 em um estudo publicado por um grupo de estudos cooperativo de franceses, americanos e britânicos (do inglês, *French, American and British* – FAB). A classificação estabelecida pela FAB é exclusivamente baseada em características morfológicas da medula óssea e sangue periférico. Nela, há seis grupos principais (M1, M2, M3, M4, M5 e M6) definidos de acordo com (*i*) a diferenciação de um grupo de células, e (*ii*) o grau de maturação (desenvolvimento celular) das mesmas (BENNETT et al., 1976).

Os grupos M1, M2 e M3 se diferem na predominância dos granulócitos (células do tipo glóbulos brancos), suas extensões e maturação. No caso do grupo M4, há a diferenciação granulocítica e monocítica (células produzidas na medula óssea). O M5 tem predominância na diferenciação monocítica e, por fim, o M6 se difere na predominância eritroblástica (células vermelhas do sangue). Estes grupos podem ser descritos da seguinte forma: M1 é a Leucemia mieloblástica aguda sem maturação; M2 é a Leucemia mieloblástica aguda com maturação; M3 é a Leucemia promielocítica aguda; M4 é a Leucemia mielomonocítica aguda; M5 é a Leucemia monocítica; por fim, M6 é a Leucemia eritroide aguda (BENNETT et al., 1976).

Em 2010, houve uma nova categorização considerando as condições de gravidade da doença, baseada em características citogenéticas e moleculares (DÖHNER et al., 2010). As características citogenéticas estão relacionadas com alterações cromossômicas (translocação, inversão, cópia ou deleção) encontradas nas células cancerígenas. Uma translocação pode ser descrita como a divisão de um cromossomo que se liga a outro, enquanto uma inversão se trata de mudanças em um gene que serão posteriormente ligadas ao mesmo, mas de maneira invertida. A análise molecular está relacionada ao número de mutações gênicas e desajustes em suas expressões.

O grupo de risco **citogenético** favorável é dado por translocações entre os cro-

mossomos 8 e 21, ou 15 e 17, ou, inversão do 16. De maneira semelhante, o grupo de risco adverso é dado por translocações nos cromossomos 5 e 7, ou 6 e 9, ou, 9 e 22; além disto, ainda há a possibilidade de perdas cromossômicas. O grupo intermediário é dado para pacientes que não se encaixam no risco favorável e adverso (DÖHNER et al., 2010). Consequentemente, indivíduos deste grupo se mostram muito diferentes uns dos outros, e é esta heterogeneidade um dos agravantes para a incerteza quanto ao seu curso do tratamento, podendo resultar na indefinição e consequente atraso no início do tratamento.

Para o paciente ser classificado no grupo de risco **molecular** favorável, é necessário que hajam mutações no gene *NPM1* e uma expressão menor que 0,5 do *FLT3-ITD*. Já para grupo intermediário, é preciso que esta expressão seja maior que 0,5 e que também hajam mutações no *NPM1*. Por fim, o grupo de risco molecular adverso também possui as características descritas para o grupo intermediário e se diferencia nas mutações dos seguintes genes: *RUNX1*, *ASXL1* e *TP53* (DÖHNER et al., 2010). Esta classificação é mais efetiva que a citogenética, contudo é mais difícil realizá-la em uma primeira visita clínica devido ao alto custo envolvido na coleta e análise das amostras.

A classificação de risco adotada pela ELN (*European Leukemia Net*), amplamente utilizada por especialistas, combina informações citogenéticas e moleculares e também possui três grupos de risco (favorável, intermediário e adverso). A Tabela 1 apresenta esta classificação e mostra um comparativo com as duas anteriores supracitadas. A expressão $t(x, y)$ representa uma translocação entre o cromossomo x e y; $i(x)$ representa uma inversão do cromossomo x; e, por fim, $m(x)$ significa mutação no gene x. Um cariótipo complexo é dado quando há mudanças em ao menos 3 cromossomos, e, de modo análogo, um cariótipo monossômico apresenta mudanças em ao menos 2 cromossomos.

Em 2022, o grupo ELN propôs uma atualização na sua classificação de risco. A principal mudança está na expressão do gene *FLT3-ITD*: todos os pacientes que apresentam essa expressão e não possuem as demais características do grupo adverso, são classificados no grupo de risco intermediário. Outra alteração foi a inclusão das mutações dos genes *BCOR*, *EZH2*, *SF3B1*, *SRSF2*, *STAG2* e *ZRSR2* na classificação adversa (DÖHNER et al., 2022). Essas mudanças foram feitas em relação a novos aspectos descobertos que influenciam no curso da doença (Cancer Genome Atlas Research Network et al., 2013; ANGENENDT et al., 2019).

A AML é uma doença grave, com alta taxa de mortalidade, que requer um prognóstico de risco assertivo para personalizar terapias (PELCOVITS; NIROULA, 2020). Apesar dos recentes avanços no entendimento da AML, pouco se traduziu na melhoria das classificações de risco, principalmente com respeito à classe de risco intermediário.

Recentemente, Elsayed et al. (2020) analisaram os grupos de risco de AML por meio de expressões gênicas, o que os levaram a um conjunto de seis genes. Especificamente, a análise se deu no contexto de crianças com AML, sendo que os autores compararam seus

Grupos	Classificações de risco		
	Citogenética (DÖHNER et al., 2010)	**Molecular** (DÖHNER et al., 2010)	**ELN** (DÖHNER et al., 2017)
Favorável	$t(8,21)$ ou $t(15,17)$ ou $i(16)$	*m(NPM1)* e *FLT3-ITD* $\leq 0,5$	citogenética, molecular, *m(RUNX1-RUNX1T1)*, *m(CBFB-MYH11)* e *m(CEBPA)*
Intermediário	Não atendem os critérios de favorável e adverso	*m(NPM1)* e *FLT3-ITD* $> 0,5$	citogenética, molecular, $t(9,11)$ e *m(MLLT3-KMT2A)*
Adverso	$t(5,7)$ ou $t(6,9)$ ou $t(9,22)$	*m(NPM1)*, *m(RUNX1)*, *m(ASXL1)*, *m(TP53)* e *FLT3-ITD* $> 0,5$	citogenética, molecular, cariótipo complexo e monossômico, *m(DEK-NUP214)*, *m(KMT2A)* e *m(BCR-ABL1BB)*

Fonte: Döhner et al. (2010) e Döhner et al. (2017)

Tabela 1 – Principais classificações de risco existentes para AML propostas nos anos de 2010 e 2017.

resultados com os obtidos por Ng et al. (2016), que chegaram a um conjunto de dezessete genes (com dados de adultos e crianças). Elsayed et al. (2020) concluíram que o conjunto de seis genes possui um maior poder preditivo em relação a recaídas e análise de grupos de risco nas crianças. Além disso, os autores concluíram que o fator idade exerce grande influência na classificação de risco de pacientes com AML. A recaída é definida quando o paciente apresenta sinais da doença após um período de cura (remissão) (KIM et al., 2021).

Xiaosu et al. (2019) analisaram a expressão gênica no contexto da determinação da necessidade ou não de transplante de células da medula óssea e/ou do sangue periférico. O estudo mostrou que uma expressão maior que 0,2% do gene *CBFβ/MYH11* é um importante fator preditor para transplante. Apesar de pacientes com essa característica serem classificados no grupo de risco favorável, eles ainda possuem uma baixa taxa de sobrevivência com uma alta taxa de recaída para a doença, mesmo após transplantes. Os autores apontaram que a análise do gene *CBFβ/MYH11* é importante para distinguir pacientes dos grupos de risco favorável e adverso. De modo análogo, Wang et al. (2020) investigaram a expressão do RNA *KIAA0125* e Poiani et al. (2021) examinaram o gene *FLT3-ITD*. Ambos concluíram que estes genes são relevantes na classificação de pacientes nos grupos de risco.

2.2 Mapeamento sistemático da literatura

O mapeamento sistemático da literatura (MSL) é um estudo secundário (aquele que busca e analisa trabalhos que propõem/avaliam ou descrevem determinadas soluções) que segue um protocolo de pesquisa bem definido com o intuito de categorizar, analisar e interpretar evidências em relação a um determinado tema de pesquisa científica (KITCHENHAM et al., 2007).

Tal estudo se diferencia de uma revisão sistemática da literatura (RSL) em pequenos aspectos relacionados com o foco dos resultados a serem obtidos. Uma RSL demanda que os seus realizadores tenham um conhecimento prévio sobre o tema analisado e que buscam realizar comparações de métodos/abordagens de solução de determinados problemas. Por sua vez, o objetivo de um MSL é fazer uma busca ampla sem trazer comparações, tendo como intuito produzir categorizações gerais do tema para de certa forma prover um guia de pesquisa futuro.

Nesta seção, é apresentado um MSL guiado por meio de um protocolo baseado nas propostas de Kitchenham et al. (2007) e Petersen, Vakkalanka e Kuzniarz (2015). O objetivo deste mapeamento é caracterizar os métodos existentes de escolha de tratamentos de pacientes com AML, bem como os seus critérios de utilização.

A Figura 1 ilustra o processo do MSL adotado neste trabalho, que possui três fases gerais que se subdividem ao longo de sua execução, planejamento, condução e conclusão, descritas nas subseções abaixo.

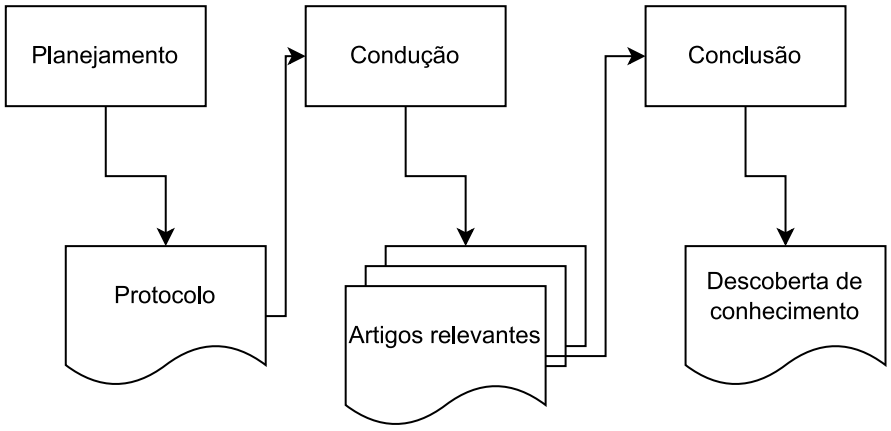

Figura 1 – Processo adotado de Mapeamento Sistemático. Adaptado de Oliveira, Lima e Lóscio (2019).

2.2.1 Planejamento

Esta fase aborda a elaboração geral do protocolo que se torna um guia para as demais fases. Neste contexto, há a definição do objetivo do trabalho, análise de palavras-chave, definição da *string* de busca, fontes de pesquisa e critérios de seleção (inclusão

e exclusão). Para o suporte da construção e execução do protocolo, foram utilizadas as seguintes plataformas online *Parsif.al*[1] e *Google Drive*[2].

2.2.1.1 Questões de pesquisa

A fim de alcançar o objetivo proposto no mapeamento, foram definidas 3 questões de pesquisa descritas na Tabela 2. Estas questões foram elaboradas para oferecer um maior entendimento do tema e, consequentemente, uma melhor compreensão das lacunas da literatura em relação ao uso de técnicas de aprendizado de máquina para auxílio nos guias de tratamento de pacientes com AML.

ID	Questão de pesquisa	Justificativa
QP1	Quais são os planos de escolha de tratamento em pacientes com AML?	Entendimento de quais são os planos de tratamento existentes no contexto de AML.
QP2	Quais são os critérios utilizados para a escolha do tratamento de pacientes com AML?	Entendimento das escolhas do especialista de um tratamento em função de outro.
QP3	Já existem métodos de ML que contribuem para a escolha de tratamentos em pacientes com AML?	Mapeamento de métodos existentes de ML que podem auxiliar na escolha de tratamentos.

Tabela 2 – Questões de pesquisa com suas respectivas justificativas.

2.2.1.2 Estratégia de busca

A fim de facilitar o processo de identificação de palavras-chave para a construção da *string* de busca, foram utilizados os critérios PICOC, proposto por Pai et al. (2004), derivados da medicina. O significado da sigla supracitada é dada a seguir:

- **P** (População, do inglês *Population*): pessoas/elementos afetados pela intervenção.
- **I** (Intervenção, do inglês *Intervention*): critério de investigação do estudo.
- **C** (Comparação, do inglês *Comparison*): parâmetro de referência para comparação de estudos.
- **O** (Resultados, do inglês *Outcomes*): resultados esperados a partir da realização do trabalho.

[1] *Parsif.al* é uma ferramenta online desenvolvida para apoiar pesquisadores na realização de revisões sistemáticas da literatura. Ela está publicamente disponível em: <https://pasif.al/>. Acessado em: 11/07/2022.

[2] Google Drive é um serviço de armazenamento e sincronização de arquivos. Ele está publicamente disponível em: <https://drive.google.com/>. Acessado em: 11/07/2022.

- **C** (Contexto, do inglês *Context*): que se relaciona a uma área de aplicação.

A adaptação dos critérios PICOC para este trabalho é dada a seguir:

- **P**: conjunto de publicações relacionadas a planos de tratamentos para pacientes com AML.
- **I**: métodos de escolha de tratamentos que podem ser melhorados com o uso de ML.
- **C**: não se aplica por se tratar de um mapeamento sistemático onde não há comparações, apenas categorizações de artigos.
- **O**: mapeamento das escolhas de tratamentos utilizadas para pacientes com AML.
- **C**: ambientes clínicos.

Foi definido a estratégia de busca automática pela questão da obtenção de um número abundante de estudos de forma rápida e prática. As bases selecionadas foram a PubMed[3], IEEE Xplore[4] e ACM Digital Library[5] pela razão de indexarem a grande maioria dos trabalhos publicados no campo da medicina e computação.

A definição da *string* de busca foi realizada a partir do critério PICOC e das palavras-chave, em conjunto com os seus sinônimos. O resultado é mostrado abaixo:

(*"acute myeloid leukemia"* OR *"myeloid neoplasms"*) AND (*"treatment"* OR *"therapy"*) AND (*"plan"* OR *"guidelines"*)

2.2.1.3 Critérios de seleção

Os critérios de inclusão e exclusão ajudam a nortear o processo de seleção de documentos relevantes para o contexto do estudo sistemático. A partir disto, foi criado 1 critério de inclusão e 7 critérios de exclusão, descritos abaixo:

Critério de inclusão (CI):

- CI. O artigo propõe ou descreve, ou avalia métodos de tratamento de AML.

Critérios de Exclusão (CE):

[3] PubMed é um motor de busca para artigos médicos. Ele está publicamente disponível em: <https://pubmed.ncbi.nlm.nih.gov/>. Acessado em: 11/07/2022

[4] IEEE Xplore é um motor de busca de artigos científicos voltados para ciência da computação, engenharias e campos relacionados. Ele está publicamente disponível em: <https://ieeexplore.ieee.org/Xplore/home.jsp>. Acessado em: 11/07/2022

[5] ACM Digital Library é um repositório de artigos científicos nos campos de ciência e engenharia da computação. Ele está publicamente disponível em: <https://dl.acm.org/>. Acessado em 11/07/2022

- CE1. Documentos que não estejam no idioma inglês.
- CE2. Artigos que sejam de AML, mas que não abordem tratamentos.
- CE3. Trabalhos que não sejam estudos primários.
- CE4. Artigos que não tenham seu conteúdo completo disponível.
- CE5. Artigos sobre tratamentos de outras doenças.
- CE6. Artigos que são versões resumidas de outros.
- CE7. Estudos que não atendam ao critério de inclusão.

2.2.2 Condução

Esta fase refere-se a execução do protocolo descrito na fase anterior. Ela inclui a identificação e seleção dos estudos primários (trabalhos que proponham/avaliem ou descrevam determinadas soluções), bem como a extração, avaliação e sintetização dos dados encontrados nestes documentos.

O processo de seleção de estudos, sumarizado na Figura 2, foi realizado da seguinte forma:

1. Busca automática por documentos, realizada nos dias 11 e 12 de julho de 2022, usando os motores de busca PubMed, ACM Digital Library e IEEEXplore. A Tabela 3 descreve as *strings* de busca utilizadas em cada repositório com suas respectivas especificações;

2. Inserção dos metadados (título, resumo, ano de publicação, autores e palavras-chave) na plataforma **Parsif.al** dos documentos encontrados. Inicialmente, foram encontrados 92 estudos, sendo 2 na IEEE Xplore, 1 na ACM Digital Library e 89 na PubMed;

3. Análise automática de duplicatas: apenas 2 estudos foram encontrados como duplicados, e, portanto, 90 foram pré-selecionados para análise;

4. Seleção inicial por meio do CI e CEs citados na subseção anterior, a partir da leitura dos metadados. A partir disto, foi atribuído um estado para cada documento: selecionado, removido ou duplicado. Para que um estudo fosse classificado como removido, era apenas necessário que o mesmo se encaixasse em apenas um critério de exclusão. Caso contrário, era preciso que o estudo estivesse adequado ao critério de inclusão. Após esta análise, restaram 42 artigos; e

5. Seleção final dos artigos resultantes da fase anterior. Após a leitura completa dos trabalhos, foi realizada uma nova atribuição de estados para os artigos e, por fim, restaram 12 trabalhos para realizar síntese.

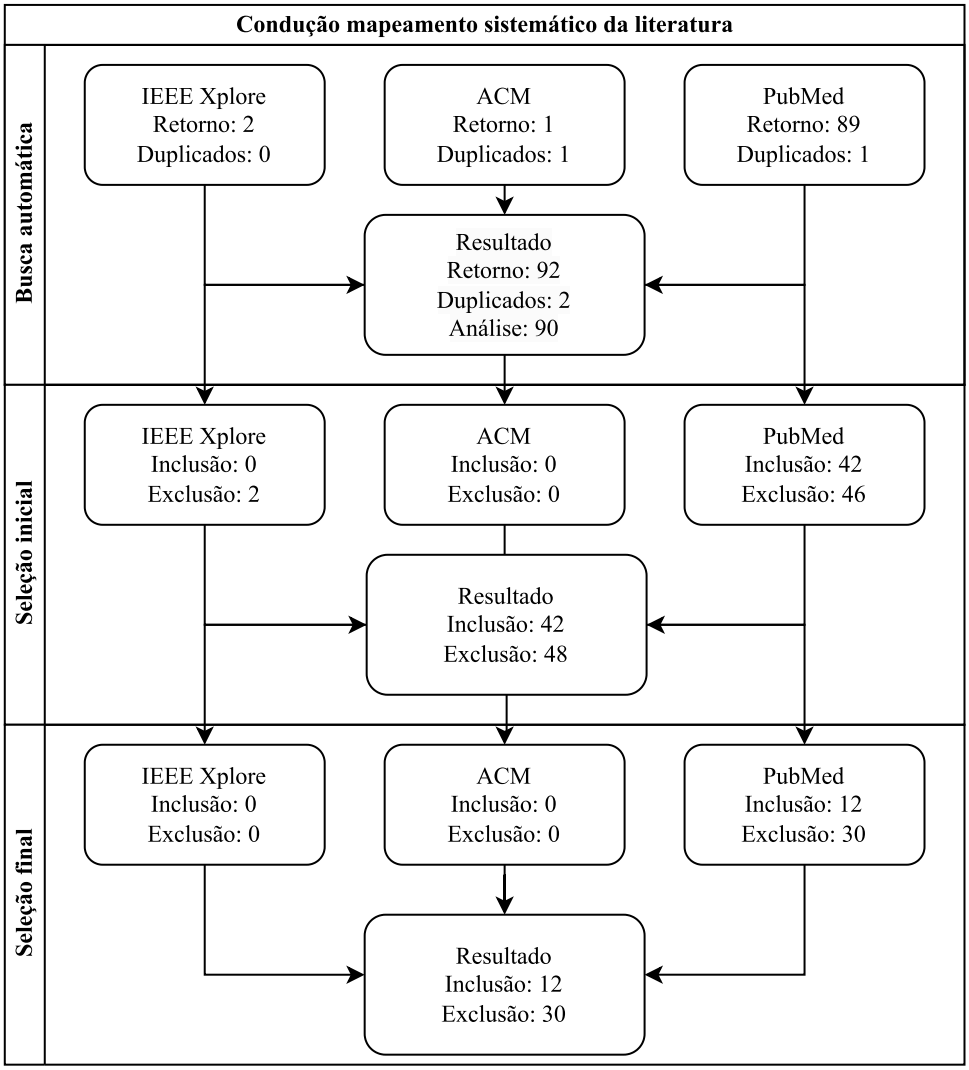

Figura 2 – Processo de seleção de estudos.

2.2.2.1 Seleção inicial

A seleção inicial é aplicada em todos os estudos identificados (92) a partir da leitura dos metadados. As Tabelas 4 e 5 descrevem detalhadamente a aplicação dos critérios de inclusão e exclusão, respectivamente. Não foram selecionados trabalhos das bases IEEE e ACM, pois os artigos encontrados não estão relacionados ao tema ou estão duplicados. A base PubMed retornou apenas 1 artigo duplicado.

Pela razão da PubMed ser especializada no campo da medicina, foi retornada uma maior quantidade de trabalhos. A partir deste repositório, foram analisados 88 artigos,

Repositório	***String* de busca**
PubMed	*(acute myeloid leukemia[Title/Abstract] OR myeloid neoplasms[Title/Abstract]) AND (treatment[Title/Abstract] OR therapy[Title/Abstract]) AND (plan[Title/Abstract] OR guideline[Title/Abstract])*
IEEE Xplore	*(("Abstract":acute myeloid leukemia) OR ("Abstract":myeloid neoplasms)) AND (("Abstract":treatment) OR ("Abstract": therapy)) AND (("Abstract": plan) OR ("Abstract":guideline))*
ACM Digital Library	*[[Abstract: "acute myeloid leukemia"] OR [Abstract: "myeloid neoplasms"]] AND [[Abstract: "treatment"] OR [Abstract: "therapy"]] AND [[Abstract:"plan"] OR [Abstract: "guidelines"]]*

Tabela 3 – *Strings* de busca separadas por base.

dos quais, 42 foram selecionados para leitura completa (Tabela 4) e 46 foram excluídos (Tabela 5), sendo que:

- 1 não estava no idioma inglês (CE1);
- 10 eram sobre AML, mas não estavam no contexto de guias de tratamento (CE2);
- 3 não eram estudos primários (CE3);
- 13 não estavam associados a AML (CE5); e
- 19 não analisavam, não descreviam e não propunham guias de tratamento em AML (CE7).

Repositório	CI
PubMed	42
IEEE Xplore	0
ACM Digital Library	0
Total	42

Tabela 4 – Aplicação do critério de inclusão nos metadados.

Repositório	CE1	CE2	CE3	CE4	CE5	CE6	CE7	Total
PubMed	1	10	3	0	13	0	19	46
IEEE Xplore	0	0	0	0	0	0	2	2
ACM Digital Library	0	0	0	0	0	0	0	0
Total	1	10	3	0	13	0	21	48

Tabela 5 – Aplicação dos critérios de exclusão nos metadados.

2.2.2.2 Seleção final

A seleção final é realizada nos estudos incluídos na seleção inicial por meio da leitura e análise do texto completo. Nesta etapa, foi buscado o conteúdo completo dos trabalhos e, por consequência, foram excluídos 4 estudos por não estarem no idioma inglês (CE1) e, 1 não ter seu conteúdo completo disponível (CE4). Dentre os 37 estudos restantes, 12 foram escolhidos para síntese (Tabela 6), 25 foram excluídos (Tabela 7), pelas seguintes razões:

- 4 se tratavam de AML, mas não eram específicos sobre tratamentos (CE2);
- 11 artigos não se tratavam de estudos primários (CE3);
- 1 estava no contexto tratamentos, mas não de AML (CE5); e
- 9 foram excluídos por não atenderem o CI (CE7).

Repositório	**CI**
PubMed	12
IEEE Xplore	0
ACM Digital Library	0
Total	12

Tabela 6 – Aplicação do critério de inclusão nos textos completos.

Repositório	**CE1**	**CE2**	**CE3**	**CE4**	**CE5**	**CE6**	**CE7**	**Total**
PubMed	4	4	11	1	1	0	9	30
IEEE Xplore	0	0	0	0	0	0	0	0
ACM Digital Library	0	0	0	0	0	0	0	0
Total	4	4	11	1	1	0	9	30

Tabela 7 – Aplicação dos critérios de exclusão nos textos completos.

A Tabela 8 apresenta um resumo dos estudos resultantes da seleção final. Abaixo, a identificação dos estudos selecionados, ordenados de forma decrescente por ano de publicação:

- E1. *Interim results from a postmarketing surveillance study of patients with FLT3-mutated relapsed/refractory AML treated with the FLT3 inhibitor gilteritinib in Japan* (SUGAMORI et al., 2022).
- E2. *Acute cardiotoxicity after initiation of the novel tyrosine kinase inhibitor gilteritinib for acute myeloid leukemia* (KIM et al., 2021).
- E3. *Venetoclax and pegcrisantaspase for complex karyotype acute myeloid leukemia* (EMADI et al., 2021).

- E4. *Understanding Barriers to Oral Therapy Adherence in Adults With Acute Myeloid Leukemia* (BRYANT et al., 2020).
- E5. *Venetoclax + hypomethylating agents combined with dose-adjusted HAG for relapsed/refractory acute myeloid leukemia: Two case reports* (WANG et al., 2020).
- E6. *Validated LC-MS/MS Method for Simultaneous Quantitation of Enasidenib and its Active Metabolite, AGI-16903 in Small Volume Mice Plasma: Application to a Pharmacokinetic Study* (DITTAKAVI et al., 2020)
- E7. *Measurable residual disease monitoring in acute myeloid leukemia with t(8; 21)(q22; q22.1): results from the AML Study Group* (RÜCKER et al., 2019).
- E8. *Comparison of Autologous Stem Cell Transplantation versus Haploidentical Donor Stem Cell Transplantation for Favorable and Intermediate-Risk Acute Myeloid Leukemia Patients in First Complete Remission* (CHEN et al., 2018).
- E9. *Initial Report of a Phase I Study of LY2510924, Idarubicin, and Cytarabine in Relapsed/Refractory Acute Myeloid Leukemia* (BODDU et al., 2018).
- E10. *Improving the Transition to Palliative Care for Patients With Acute Leukemia: A Coordinated Care Approach* (HOPKINS et al., 2017).
- E11. *Selection of elderly acute myeloid leukemia patients for intensive chemotherapy: effectiveness of intensive chemotherapy and subgroup analysis* (KIM et al., 2015).
- E12. *Haplotype mismatched transplantation using high doses of peripheral blood CD34+ cells together with stratified conditioning regimens for high-risk adult acute myeloid leukemia patients: a pilot study in a single Korean institution* (KIM et al., 2005).

Estudo	**Resumo**
E1	Foi avaliado o uso de xospata[6] em pacientes que apresentaram a mutação no gene *FLT3* e que tiveram recaídas para AML ou piora após um tratamento. Destes, 34,3% dos pacientes atingiram remissão. Foram analisados 6 pacientes pediátricos (embora o medicamento não seja recomendado) dos quais 4 responderam ao tratamento e 2 destes atingiram remissão.

[6] O xospata pertence a uma classe de medicamentos chamada inibidores das proteínas quínases. Ele contém a substância ativa gilteritinibe que bloqueia a ação de certas enzimas (quínases) necessárias para a multiplicação e crescimento das células anormais, impedindo, assim, a evolução do câncer.

E2	Foi apresentado um estudo de caso envolvendo uma paciente de 56 anos que teve recaída, fez o uso de xospata e, posteriormente, desenvolveu problemas cardíacos. O tratamento foi interrompido, sendo detectado que o xospata foi responsável pelas disfunções cardíacas. No fim, a paciente teve uma nova remissão de AML com o uso de venetoclax[7] e terapia regular[8].
E3	Foi realizada uma triagem clínica com administração dos medicamentos venetoclax e pegcrisantaspase[9] para pacientes com cariótipo complexo que tiveram recaídas. O estudo combinou os dois medicamentos e também administrou de modo separado. Foi feito o uso de tratamento convencional em alguns pacientes para verificar se haveria um impacto maior nos resultados da combinação dos medicamentos.
E4	A partir de uma coorte de 11 pacientes adultos, foi aplicado um formulário para avaliar as dificuldades dos pacientes em receber suas quimioterapias de forma oral. Os pacientes relataram dificuldade para gerenciar a alta quantidade de medicamentos e o impacto negativo na rotina diária de cada um em relação aos efeitos medicamentosos. O estudo concluiu que os medicamentos intravenosos tem uma perspectiva de melhores resultados.
E5	Estudo de caso que reportou o acompanhamento de 2 pacientes com recaída para AML que fizeram o uso de venetoclax e hipometilantes[10]. Um dos pacientes foi uma mulher com 23 anos que recebeu tratamento de quimioterapia convencional e apresentou recaída para a AML e mutações nos genes *NPM1* e *CEBPA*. Ao todo, houve 4 recaídas, 47 meses de sobrevivência total e 2 meses livres da doença. O segundo paciente foi um homem de 26 anos que teve uma recaída, não resistiu à quimioterapia e interrompeu o tratamento.

[7] O venetoclax é um medicamento inibidor da proteína *BCL-2* presente em células cancerígenas, ele contribui para a morte dessas células.

[8] A terapia regular envolve a aplicação dois medicamentos quimioterápicos, a citarabina e a antraciclina durante um período de 7 dias. Estes medicamento atuam impedindo a divisão celular e o crescimento de células cancerígenas.

[9] A pegcrisantaspase é um medicamento que aumenta a produção de células brancas do sangue (neutrófilos) no organismo. Ele é utilizado para prevenir a neutropenia (baixa contagem de neutrófilos) induzida por quimioterapia e outras terapias.

[10] Os medicamentos hipometilantes atuam reduzindo a metilação do DNA, aumentando a expressão de genes importantes para matar células cancerígenas. A metilação do DNA é um mecanismo importante para regular a expressão gênica e os genes que estão metilados são menos ativos.

E6	O estudo avaliou a eficácia do medicamento enasidenib[11] por amostras de plasma de camundongos e utilizou o método LC-MS/MS[12] para validação. Os resultados sugerem que o enasidenib é seguro e eficaz para o tratamento de pacientes com AML que apresentaram recaída, embora ainda necessite de estudos adicionais com dados de humanos para confirmar esses resultados. O estudo pode contribuir para a inclusão do enasidenib em guias de tratamento para pacientes com AML que apresentaram recaída.
E7	Foram avaliados dados de 155 pacientes com AML com a mutação no gene *RUNX1* e com o prognóstico favorável. Foi empregado regressão linear para prever o impacto da mutação do gene *RUNX1* no tratamento do paciente após o primeiro ciclo de quimioterapia, além da quantidade residual de doença após esse ciclo. Os pacientes com essa mutação foram associados a maiores riscos de recaída da doença. A taxa de menos de 25% de mieloblastos no sangue ou medula óssea após o primeiro ciclo de tratamento e a taxa de menos de 30% após o segundo ciclo foram relacionadas a um risco menor de recaída.
E8	Estudo de caso que acompanhou pacientes após a primeira remissão. Os pacientes receberam a terapia regular e, em sua remissão, foi avaliado qual transplante realizar. Foram acompanhados 195 pacientes entre maio de 2007 e dezembro de 2013. Destes, 88 receberam transplante autólogo[13] e 107 receberam o alogênico[14]. A taxa NRM[15] 3 anos após o transplante autólogo foi significativamente menor do que a observada após o transplante haploidentico[16]. Para pacientes do grupo de risco favorável (classificação de risco da ELN), o transplante alogênico pode alcançar tempos de sobrevida tão promissores quanto o autólogo.

[11] O enasidenib é um inibidor da mutação no gene *IDH1*. Ele atua inibindo a proteína IDH1, impedindo a produção de moléculas que promovem o crescimento e a sobrevivência das células cancerígenas

[12] A LC-MS/MS é uma técnica analítica que combina a separação e identificação de moléculas em amostras biológicas, com alta sensibilidade e especificidade. É composta por duas etapas, a cromatografia líquida para separar as moléculas, e espectrometria de massa para identificar e quantificar essas moléculas.

[13] O transplante autólogo é feito com células-tronco saudáveis do paciente e após o seu tratamento quimioterápico elas são reinseridas de volta no paciente, reduzindo o risco de rejeição.

[14] O transplante alogênico envolve o uso de células-tronco de um doador, presentes no sangue e medula óssea, responsáveis pela produção de células do sangue. O transplante alogênico é realizado quando não é possível usar as células-tronco do próprio paciente.

[15] A NRM é uma abreviação para *nonrelapse mortality*. A NRM refere-se à taxa de mortalidade relacionada a complicações do tratamento, como infecções ou falência de órgãos.

[16] O transplante haploidentico é uma forma de transplante alogênico que utiliza células-tronco de um doador relacionado geneticamente, geralmente um membro da família, que possui metade dos mesmos

E9	Foi avaliado a combinação de um medicamento (LY2510924, inibidor do gene *CXCR4*) com terapia regular em pacientes que apresentaram recaída. A triagem clínica foi realizada com 11 pacientes. Inicialmente, foram tratados apenas com o medicamento supracitado e caso não houvesse melhora no 7º dia de tratamento, foi iniciado a administração da quimioterapia original. Pacientes que apresentaram uma boa resposta receberam mais 4 doses do medicamento em combinação com a terapia regular. Em conclusão, a combinação de LY2510924 com terapia regular é segura em pacientes que apresentaram recaída. A dose de LY2510924 de 10 e 20 mg/dia inibe o gene *CXCR4* em alguns, mas não em todos os pacientes e é necessária uma dose de 30mg/dia de LY2510924 para uma inibição completa do gene *CXCR4*.
E10	Estudo de caso de 23 pacientes com AML que interromperam tratamentos curativos[17]. Com isso, foi proposto um guia de tratamento cujo objetivo foi de melhorar a transição de tratamentos curativos para cuidados paliativos[18] em pacientes idosos, através do desenvolvimento e implementação de um programa de plano de cuidado coordenado. O estudo indicou que a comunicação e o planejamento de cuidados formalizados contribuem para a aceitação dos pacientes e a compreensão pessoal sobre a redefinição dos tipos de cuidados. A individualização dos cuidados no fim da vida também foi reconhecida como importante para os pacientes e famílias.

marcadores HLA (antígenos leucocitários humanos) que o paciente

[17] Os tratamentos curativos visam alcançar uma remissão completa e prolongada, com o objetivo final de curar a doença.

[18] Os tratamentos paliativos pretendem aliviar os sintomas da doença e melhorar a qualidade de vida dos pacientes, sem a cura da doença.

E11	Estudo retrospectivo de 84 pacientes idosos. Eles foram divididos em grupos de tratamentos, 35 receberam quimioterapia convencional, 19 com baixas doses e 30 com tratamento paliativo. Os que tinham idade entre 65 e 70 anos responderam melhor à quimioterapia intensiva. Pacientes com prognóstico favorável foram tratados com terapia regular. Por fim, os com prognóstico adverso receberam apenas citarabina. A taxa de remissão e a sobrevida global do grupo de quimioterapia intensiva foram superiores às do grupo de terapia de baixa intensidade, sugerindo que a quimioterapia intensiva também é eficaz para pacientes idosos com idade entre 65 e 70 anos. Além disso, os benefícios da quimioterapia foram mais altos no grupo sem recaídas do que no grupo com recaídas.
E12	Estudo de caso que acompanhou 11 pacientes com prognóstico de risco adverso, que receberam transplante haploidentico. Dois dos 11 pacientes receberam transplantes de doadores que não eram 100% compatíveis. Os pacientes apresentaram remissão completa inicial e finalizaram o tratamento com transplantes na expectativa de evitar uma nova recaída para a doença. Os pacientes foram dispensados em média no dia 35 de hospitalização após o transplante. Três pacientes apresentaram recaídas e 7 permaneceram vivos até 6 meses após os transplantes.

Tabela 8 – Resumo dos estudos remanescentes após a seleção final.

2.2.3 Conclusão

A síntese gerada a partir deste MSL foi categorizada por meio das respostas das questões de pesquisa geradas neste estudo (Subseção 2.2.1.1). Para tanto, essas respostas são descritas e discutidas nas Subseções 2.2.3.1 a 2.2.3.3.

2.2.3.1 QP1. Quais são os planos de escolha de tratamento em pacientes com AML?

A fim de categorizar as respostas dos guias de tratamento existentes, os especialistas[19] agruparam os tratamentos existentes para AML em 4 grupos de acordo com sua intensidade:

1. *High intensity therapy* (HIT): plano terapêutico que incluiu consolidação com transplante (autólogo ou alogênico).

[19] Grupo de pesquisa do Dr. João Agostinho Machado Neto, professor do Instituto de Ciências Biomédicas da Universidade de São Paulo.

2. *Low intensity therapy* (LIT): plano terapêutico com potencial não curativo/terapia paliativa. Geralmente recomendada para pacientes idosos ou com resposta ruim aos demais tratamentos.

3. *Regular therapy* (RT): plano terapêutico baseado na combinação de citarabina e antraciclina (7+3) e suas variações.

4. *Target therapy* (TT): plano terapêutico que incluiu um fármaco que atua de forma alvo seletiva. Geralmente recomendado para pacientes com uma alteração molecular específica, na qual o fármaco atua.

Dos 12 artigos selecionados, apenas 6 empregaram uma única categorização de tratamento. Destes, 2 (E11 e E10) estão relacionados a guias de tratamentos voltados para abordagens paliativas (LIT), 2 (E4 e E6) estão relacionados a TT e 2 (E8 e E12) são voltados para terapias de alta intensidade com o uso de medicamentos quimioterápicos (HIT). Nenhum dos estudos avaliou o uso isolado de terapia regular, fato que pode apontar para a ineficiência do uso exclusivo de quimioterapias convencionais. Os 6 trabalhos restantes combinam os diferentes tipos de guias de tratamento RT, TT e HIT. A Figura 3 sumariza os resultados da aplicação da categorização do especialista nos estudos finais.

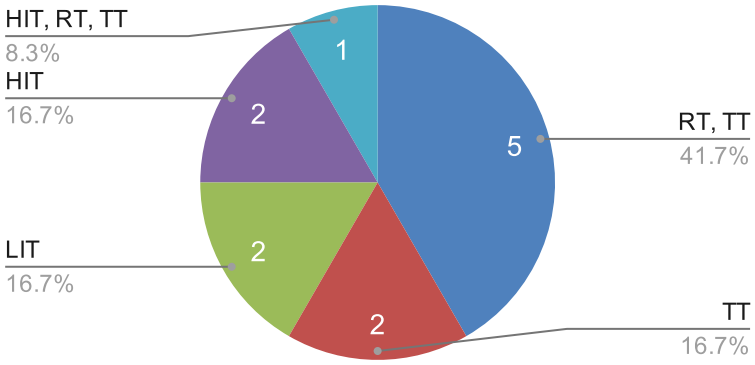

Figura 3 – Categorização dos guias de tratamento conforme a intensidade.

Os trabalhos selecionados também foram classificados com relação a sua atuação nos guias de tratamento, divididos em: proposta (3), avaliação (9) e descrição (0). O estudo E10 propõe um guia de tratamento de transição entre RT e paliativa. Para a geração deste guia de tratamento, foram usados dados clínicos de 26 pacientes idosos (com idade superior a 60 anos), oriundos da Oceania. Hopkins et al. (2017) descrevem que a individualização do programa de plano de cuidados gerou melhor comunicação e resposta do paciente em relação aos objetivos do procedimento paliativo. Os estudos E1 e E5 propuseram guias de tratamento com a aplicação dos medicamentos xospata e venetoclax, respectivamente, com a combinação de TT e RT.

Outra categorização importante para este trabalho é relativa ao tipo de dado utilizado. Os atributos foram divididos em clínicos (CLIN), expressão gênica (EXP) e mutação gênica (MUT). A Figura 4 sumariza os resultados obtidos. Dos 12 trabalhos, 5 (E2, E3, E4, E6 e E10) utilizaram apenas dados clínicos para a definição de guias de tratamento, 5 (E1, E7, E8, E11 e E12) combinaram dados clínicos e genéticos de mutação e apenas 2 (E5 e E9) estudos realizaram a combinação dos três tipos de dados.

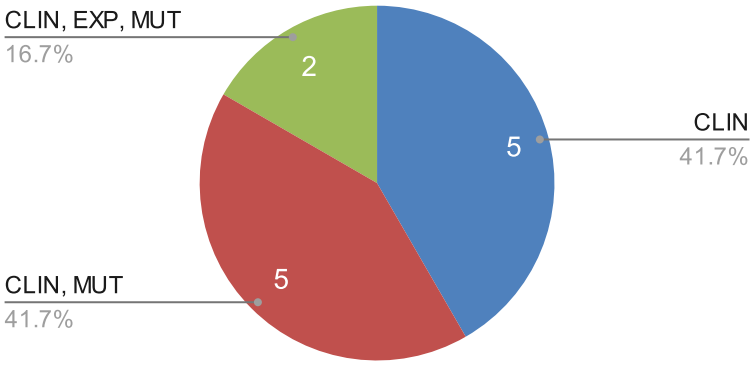

Figura 4 – Categorização dos trabalhos selecionados em relação ao tipo de dado utilizado.

O pequeno número de trabalhos utilizando dados de expressão gênica pode ser explicado pelo alto custo envolvido na coleta e processamento, além de ser complexo de realizar a sua análise de modo individualizado. Essa última deficiência pode ser sanada pelo emprego de técnicas de ML para a análise destes dados, uma vez que a partir destas técnicas é possível descobrir padrões em uma massa de dados de modo automatizado.

Os protocolos de tratamento para AML variam em intensidade e podem afetar diretamente a qualidade de vida dos pacientes de maneira diferente. Definir qual é a melhor escolha neste quesito é um procedimento complexo, uma vez que a manifestação da doença é heterogênea. Portanto, é evidente a necessidade de ferramentas que possam auxiliar a tomada de decisão dos especialistas com respeito a escolha de guias de tratamento em função da predição de sobrevida de cada paciente.

2.2.3.2 QP2. Quais são os critérios utilizados para a escolha do tratamento de pacientes com AML?

Dos 5 trabalhos que combinaram RT e TT, 4 (E2, E3, E5 e E9) consideraram pacientes com recaída para a doença. Estes pacientes foram inicialmente tratados com quimioterapia convencional e, após a recaída, foram submetidos a uma terapia focada na inibição dos genes *FLT3* (E2), *BCL2* (E3 e E5) e *CXCR4* (E9).

Kim et al. (2005) relatam que a expressão do *FLT3* está presente em cerca de 30% dos pacientes com AML e está associada a um curso agressivo da doença, com altas taxas de recaída. Uma alta expressão do gene *BCL2* está relacionado a inibição da morte

programada nas células cancerígenas (WANG et al., 2020). Por sua vez, a desregulação do gene *CXCR5* está relacionada ao descontrole na liberação de células-tronco da médula óssea para o sangue periférico (WANG et al., 2020). O estudo E1 também realizou a combinação dos tratamentos, mas não considerou recaídas e utilizou inibidores do gene *FLT3*.

A idade dos pacientes mostrou consistentemente ser um fator relevante para a escolha dos guias de tratamento. A Figura 5 apresenta a média de idade dos pacientes incluídos nos estudos, com exceção dos estudos E6 e E10 que não disponibilizaram essa informação. De maneira geral, é evidente que a incidência de AML ocorre na maioria das vezes em pacientes adultos. Sugamori et al. (2022) descrevem respostas terapêuticas diferentes de um mesmo medicamento em pacientes pediátricos em relação aos adultos.

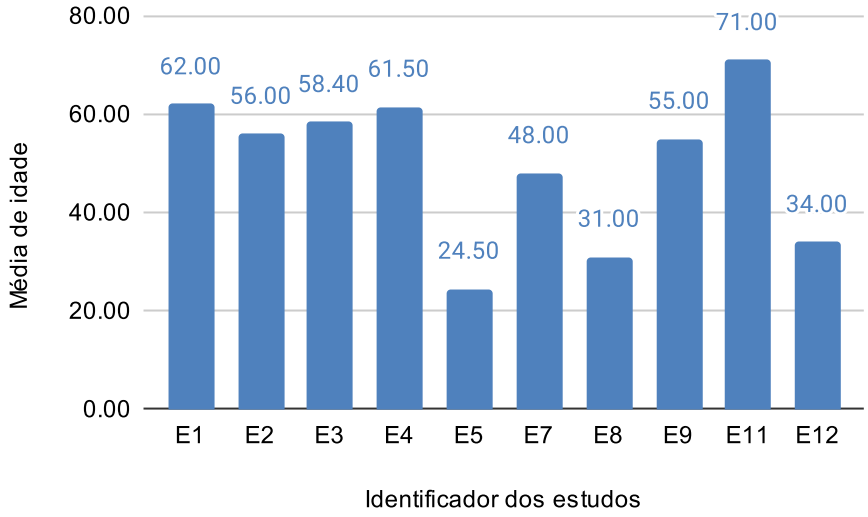

Figura 5 – Idade média dos pacientes acompanhados em cada estudo.

Com respeito aos grupos de idade, foi feita uma divisão de estudos que aceitavam ou não pacientes pediátricos para o guia de tratamento aplicado. Em AML, as divisões etárias são dadas em três grupos: um paciente é pediátrico quando tem idade menor que 18 anos; é declarado adulto quando a idade é maior ou igual a 18 anos e menor ou igual a 60 anos; e, é idoso quando tem idade superior a 60 anos. Apenas 3 trabalhos (E1, E5 e E8) acompanharam a resposta de pacientes pediátricos a diferentes terapias, sendo que 2 (E2 e E5) combinaram RT e TT e 1 (E8) fez apenas uso de HIT. A Figura 6 apresenta o comparativo da idade dos pacientes com os guias de tratamento que acompanharam pacientes que apresentaram recaídas. Apenas 1 estudo (E5) considerou pacientes pediátricos que tiveram recaídas.

A Figura 7 mostra a quantidade de estudos que usaram a classificação de risco da ELN em relação intensidade do tratamento escolhido. A classificação de risco da ELN foi considerada uma parte do guia de tratamento em 10 dos 12 estudos selecionados. Destes,

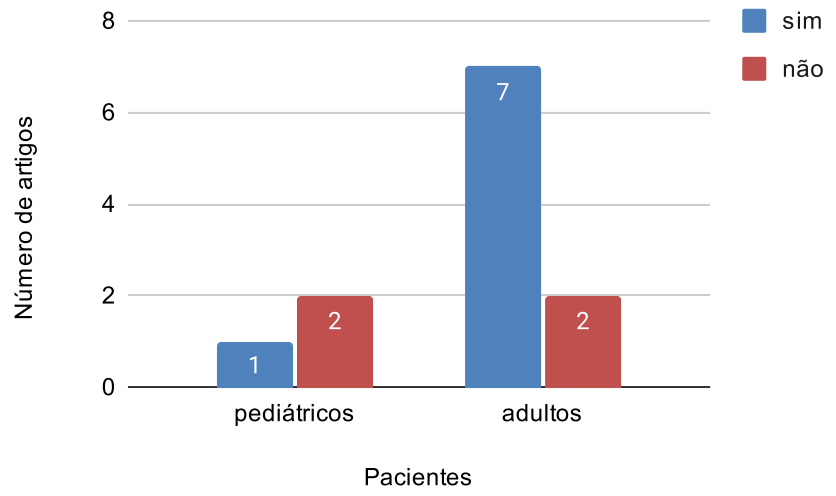

Figura 6 – Estudos que consideraram recaídas em comparação com a idade do paciente.

1 (E4) empregou uma abordagem de TT de modo individualizado, 5 (E1, E2, E3, E5 e E9) combinaram RT e TT, 2 (E8 e E12) aplicaram apenas HIT e 1 (E11) usou LIT para pacientes idosos. Estes estudos indicam que a classificação de risco da ELN não foi considerada isoladamente para a tomada de decisões terapêuticas no curso da AML. Dos 2 estudos que não utilizaram a classificação de risco da ELN, 1 (E10) recorreu a LIT e 1 (E6) a TT.

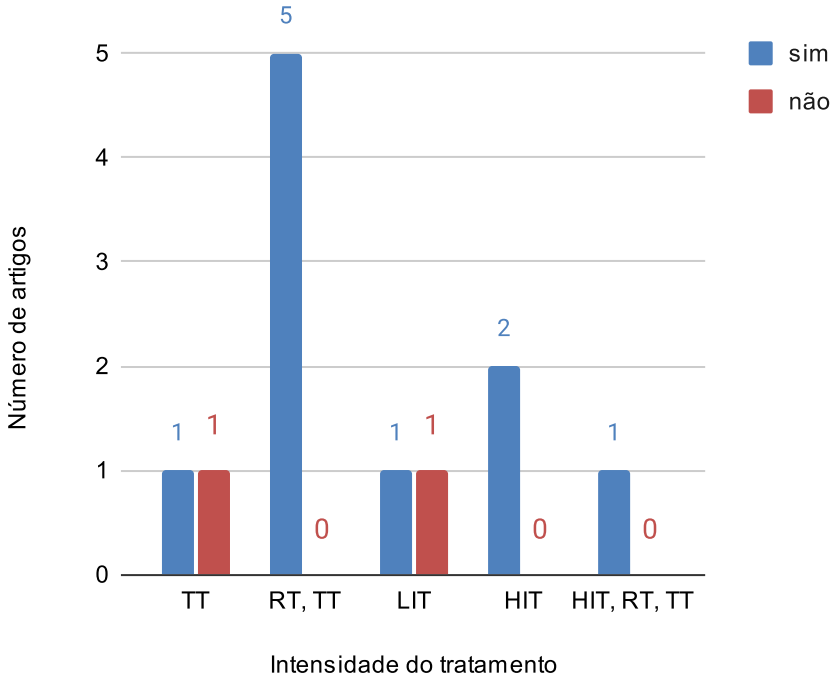

Figura 7 – Comparativo dos estudos que utilizaram a ELN em relação à intensidade do tratamento.

Em conclusão, os critérios utilizados nos guias de tratamento são diversos, como a

idade, a classificação de risco da ELN, recaídas, mutações e expressões de genes específicos e até mesmo subtipos da doença. A variedade de critérios utilizados evidencia o quão difícil é selecionar uma abordagem terapêutica que abrange todos os grupos de pacientes com AML. Neste ponto, estas decisões acabam se tornando cada vez mais personalizadas, dificultando o trabalho dos especialistas, que muitas vezes precisam recorrer a mais exames que, por sua vez, atrasam o início do tratamento e podem agravar o quadro clínico do paciente. Para contornar esse problema, ferramentas computacionais de suporte à decisão tem o potencial de contribuir na escolha da melhor decisão terapêutica, impactando diretamente para o aumento na qualidade e tempo de sobrevida do paciente.

2.2.3.3 QP3. Já existem métodos de ML que contribuem para a escolha de tratamentos em pacientes com AML?

Não foram encontrados estudos que empregaram técnicas de ML para auxiliar a construção de guias de tratamento. Consequentemente, foi realizada uma revisão bibliográfica de forma mais abrangente para fazer o levantamento do uso de ML no contexto de AML e doenças relacionadas. A Seção 2.3. descreve os principais trabalhos encontrados.

2.3 Uso de técnicas de ML como auxílio em análises de doenças

O emprego de ML é relativamente recente no auxílio de descobertas/análises de doenças. Contudo, é notório um aumento expressivo nos últimos anos, provavelmente devido à disponibilidade de dados em alta quantidade e a evolução das próprias técnicas de ML. A seguir, são descritas as tentativas de uso de métodos de ML tanto no contexto de AML quanto em outras doenças relacionadas.

2.3.1 ML e AML

Eckardt et al. (2020) empregaram redes neurais profundas treinadas com imagens de mieloblastos para predizer o diagnóstico de AML e mutações do gene *NPM1*. Foram utilizadas imagens obtidas de 1251 pacientes com AML, dos quais 386 apresentavam mutação no gene *NPM1* (gene utilizado para classificação de risco molecular). Ainda, foram utilizadas 236 imagens obtidas de pacientes saudáveis. No final, foram gerados dois modelos, um para predizer o diagnóstico de AML e outro para predizer a ocorrência de mutação no gene *NPM1*, com 88% e 69% de acurácia, respectivamente. A análise da mutação do gene *NPM1* é parte da predição da classificação de risco molecular e, portanto, realizar a sua predição a partir de imagens reduz custos e tempo de análise, contribuindo para acelerar decisões clínicas.

Utilizando *autoencoders*, Lin et al. (2020) encontraram 14 atributos relevantes para a predição de prognóstico de risco (em relação à taxa de sobrevida ser maior ou não que

730 dias) com uma acurácia de 83%. Os atributos encontrados são: idade, 7 anormalidades genéticas (trissomia do cromossomo 8, deleção do 5 e 7, cariótipo complexo, translocação do 8 e 21, 15 e 17, e, por fim, inversão do 16) e mutações em 6 genes (*FLT3*, *NPM1*, *TP53*, *DNMT3*, *KIT* e *CEBPA*). Os resultados também confirmaram a idade como um fator relevante para o risco da doença, assim como descrito em outros estudos. Os autores relataram que com o uso de ML é possível realizar a personalização de tratamentos de pacientes com AML.

A partir de dados clínicos e genéticos, Orgueira et al. (2021) conseguiram predizer a taxa de sobrevivência de pacientes com AML por meio de algoritmos baseados em florestas aleatórias. O modelo retornou que às três variáveis mais importantes para a predição, em ordem de importância, foram: idade e expressão gênica dos genes *KDM5B* e *LAPTM4B*, respectivamente. Novamente, os modelos de ML indicaram a idade como um grande fator preditivo de prognósticos de risco. Os autores ainda descreveram que a aplicação de ML em dados clínicos e moleculares tem grande potencial preditivo desde o momento do diagnóstico até o auxílio de decisões terapêuticas.

Usando dados extraídos através do sequenciamento de RNA e informações clínicas, Gal et al. (2019) exploraram o uso do ML para predizer casos de remissão completa em pacientes pediátricos com AML. Os autores testaram diversos modelos, sendo que a melhor taxa de acerto foi obtida com um método k-vizinhos mais próximos, que alcançou uma área abaixo da curva ROC de 81%. Os autores relataram haver diferenças significativas nas expressões gênicas dos pacientes em relação aos períodos de pré e pós-tratamento.

Já relacionado ao prognóstico de risco em AML, Fleming et al. (2019) desenvolveram um modelo, baseado em algoritmos de partição recursiva e florestas aleatórias. Os autores usaram a taxa de sobrevida total para dividir os grupos de risco: pacientes com > 3 anos de sobrevida foram classificados como favorável, (2, 3] anos como intermediário, (1, 2] anos como adverso e, por fim, ≤ 1 ano como muito adverso. Os atributos com maior relevância para a classificação foram: cariótipo complexo; inversão dos cromossomos 16 e 13; translocação do 8 e 21; e, 3 com 3, além de mutações dos genes *NPM1* e *TP53*. Os resultados obtidos apresentaram uma taxa de erro 10% menor quando comparados à classificação de risco da ELN, considerada base para a doença.

Os trabalhos relacionados no contexto de ML e AML, apesar de poucos, evidenciam bons resultados. Os estudos relacionados apontam também que a combinação de dados clínicos e genéticos pode ser relevantes para extrair características importantes da doença.

2.3.2 ML e doenças relacionadas

Li et al. (2021) usaram o método de florestas aleatórias para treinar um modelo de classificação para predizer melhorias nas ações sociais de pacientes com esquizofrenia.

Problemas de funções sociais nesses indivíduos podem se tornar fatores relevantes para piora da doença e menor resposta a tratamentos. As predições obtiveram uma taxa de acerto de 79,5%. O estudo destacou 13 importantes atributos, tais como os estabilizadores de humor, o fato de se ter um emprego fixo, o sexo feminino e remédios de proteção ao fígado. Os resultados apontaram que pacientes homens com esquizofrenia que não possuem emprego fixo tendem a ter problemas de funções sociais. Estes resultados, obtidos com o uso de ML, confirmam hipóteses já conhecidas na literatura do tema.

Fitter et al. (2021) descreveram a importância de implementar tratamentos personalizados em relação aos grupos de risco de pacientes com leucemia linfoide aguda (do inglês, *Acute Lymphoblastic Leukemia – ALL*). Os autores apontaram que as taxas de cura da ALL em pacientes pediátricos é superior a 85% em países desenvolvidos. No entanto, indicaram que o conhecimento de recaídas desses pacientes ainda é escasso. Para tanto, um modelo baseado em florestas aleatórias treinado a partir de dados de expressão gênica identificou os genes *CKLF* e *IL1B* como bons preditores. Além disso, os autores descrevem que a classificação de risco em ALL é complexa e não é possível de ser feita no momento do diagnóstico da doença. Eles, concluem que a análise da expressão dos genes *CKLF* e *IL1B* combinado com características clínicas podem gerar um bom modelo de predição de risco (FITTER et al., 2021).

Jing et al. (2020) treinaram um modelo de predição de gravidade do câncer de nasofaringe, baseado em redes neurais profundas, que usou como entrada imagens de ressonância magnética de 1417 pacientes. Os métodos obtiveram uma taxa de acerto de 65,1% e os autores concluíram que pacientes com metástase e recaídas tendem a ir a óbito em um menor tempo em relação aos demais pacientes. Os autores descreveram que os métodos de redes neurais profundas podem contribuir na análise de informações não percebidas pelos seres humanos.

Com o uso de dados clínicos e genéticos, Ris et al. (2019) propuseram um sistema de predição de mortalidade para pessoas com diagnóstico de endocardite infecciosa (EI). A EI é uma doença cardíaca caracterizada pela infecção de membranas do coração e possui alta taxa de mortalidade. Os dados utilizados incluíam informações de 64 pacientes, sendo 41 homens e 23 mulheres, dos quais 51 sobreviveram. Foi utilizado o método de árvores de decisão, que obteve uma taxa de acerto de 91%. A partir desta análise, os autores apontaram que as proteínas *IL-15* e *CCL4* se mostraram importantes preditoras de mortalidade.

O ML também foi aplicado com sucesso na conjuntura de doenças com maior impacto demográfico, como, por exemplo, a COVID-19 (doença pandêmica iniciada em 2019). Por meio de indicadores laboratoriais, algoritmos baseados em gradiente descendente conseguiram descobrir uma relação positiva entre anormalidades na coagulação de pacientes e a presença de sepses (disfunção dos órgãos causada por uma resposta desregulada a uma

infecção), o que pode melhorar prognósticos e reduzir mortalidade (TANG et al., 2021b).

A dengue foi outra doença de grande impacto demográfico que teve aplicações de ML para predição de severidade. Huang et al. (2020) propuseram um modelo de redes neurais profundas para predizer casos severos de dengue, com uma acurácia de 75%. De 5 a 20% dos casos de dengue são passíveis de evolução para um quadro clínico severo, no qual, podem ocorrer sangramentos, falência múltipla de órgãos e até mesmo a morte (HUANG et al., 2020).

Modelos de previsão de risco gerados com base nas características do paciente para predição de desfecho clínico de terapias vêm se tornando cada vez mais importante e frequente na prática da medicina clínica (CRISTOFERI et al., 2018). Estes modelos podem auxiliar na personalização de tratamentos e, consequentemente, contribuir para aumentar a qualidade de vida e sobrevida dos pacientes. É com base nessas evidências e na inexistência de um modelo para AML que este trabalho propõe um sistema inédito de suporte à decisão que objetiva auxiliar os especialistas na escolha de terapias adequadas para pacientes portadores de AML.

3 Pré-processamento e análise dos dados

Este capítulo descreve os processos de obtenção, pré-processamento e análise dos dados utilizados para a geração do sistema de apoio a decisão proposto. Para auxiliar no objetivo deste trabalho, foi utilizado o processo de descoberta de conhecimento em bases de dados (do inglês, *Knowledge Discovery in Databases* – KDD) proposto por Fayyad e Uthurusamy (1996). Trata-se de um processo não trivial de identificação de informações válidas, novas e potencialmente úteis. Ele consiste em um método interativo e iterativo que envolve diversas etapas em conjunto com inúmeras decisões a serem tomadas por um especialista, que possui conhecimento acerca da natureza dos dados.

As etapas do KDD podem ser resumidas em: (1) entendimento do domínio, (2) preparação do conjunto de dados, (3) limpeza e pré-processamento dos dados, (4) redução de dimensionalidade dos dados, (5) emprego de aprendizado de máquina para descoberta de padrões e, por fim, (6) identificação e análise dos padrões encontrados. A adaptação das fases 2, 3 e 4 do KDD para a realização deste estudo é apresentada nas seções subsequentes e na Figura 8. A adaptação da fase 1 é descrita no Capítulo 2. A proposta da adaptação da fase 5, o objetivo deste trabalho, é discutida no Capítulo 4. Finalmente, a adaptação da fase 6 é descrita no Capítulo 5. É importante ressaltar novamente que todas as etapas foram supervisionadas por especialistas no domínio dos dados: o grupo de pesquisa do Dr. João A. Machado-Neto, professor e pesquisador do Departamento de Farmacologia, do Instituto de Ciências Biomédicas da Universidade de São Paulo.

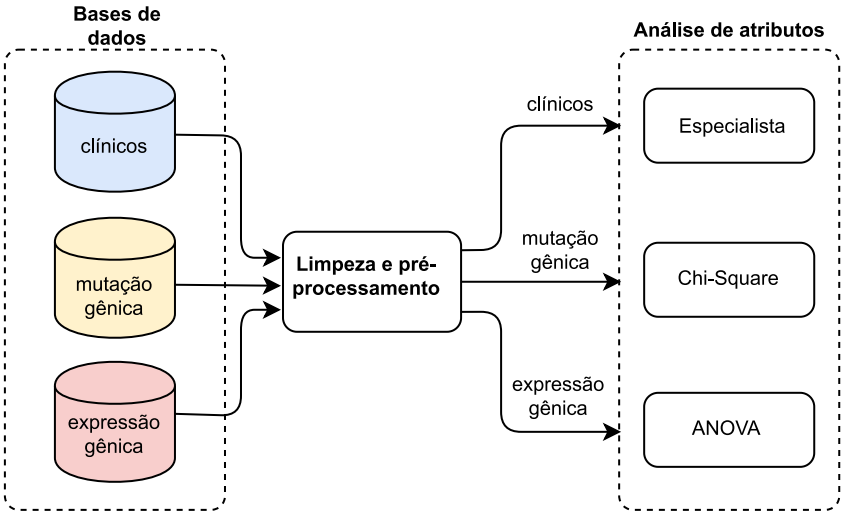

Figura 8 – Processos de obtenção, pré-processamento e análise dos dados.

3.1 Preparação dos conjuntos de dados

Os dados utilizados são provenientes de estudos realizados pelo *The Cancer Genome Atlas Program* (TCGA) e pela *Oregon Health and Science University* (OHSU). O TCGA é um programa financiado pelo Instituto nacional do câncer e pelo Instituto de pesquisa do genoma humano dos EUA. A OHSU é uma universidade pública de pesquisa na área de saúde, localizada no estado de Oregon, EUA.

Os conjuntos de dados empregados são denominados *Acute Myeloid Leukemia* (Cancer Genome Atlas Research Network et al., 2013; TYNER et al., 2018) e compreendem o acompanhamento clínico e genético de pacientes com AML. Ambos são de domínio público e podem ser obtidos no cBioPortal. Foram utilizados três conjuntos de atributos sobre os mesmos pacientes: (i) dados clínicos, (ii) mutações genéticas e (iii) expressão gênica. Os dados de expressão genética em ambas as bases de dados são originalmente padronizados por quartil[1] (AMARATUNGA; CABRERA, 2001). A Tabela 9 sumariza a quantidade original de dados obtida.

			Atributos		
Base de dados	**Amostras**[2]	**Pacientes**	**Clínicos**	**Mutação**	**Expressão**
TCGA	200	200	31	25.000	25.000
OHSU	672	562	97	606	22.825

Tabela 9 – Resumo dos dados utilizados.

3.1.1 Dados do TCGA

A base de dados utilizada do TCGA compreende o acompanhamento clínico e genético de pacientes com AML entre novembro de 2001 e março de 2010 (Cancer Genome Atlas Research Network et al., 2013). Os dados clínicos são compostos por atributos que oferecem informações das classificações de risco citogenética, molecular e da FAB, além de outras informações relativas à doença (Tabela 10). Os dois últimos atributos correspondem ao desfecho do tratamento e, portanto, podem ser usados como atributos-alvo em modelos de predição. Neste estudo, cada linha representa uma amostra de paciente, sendo que não há dados de mais de uma amostra por paciente. Ao todo, há um total de 200 amostras. Além dos dados clínicos, essa base de dados também possui informações de 25.000 mutações e expressões gênicas dos pacientes acompanhados, sendo que os dados de mutação possuem valor 1 caso o paciente tenha apresentado a mutação em um gene e 0 caso contrário. Por

[1] A padronização por quartil é um método de normalização de dados de expressão genética utilizado para remover variações técnicas que podem ocorrer durante o processo de medição. Esse método ajusta os valores de expressão de cada gene de forma que as distribuições dos dados em todas as amostras sejam iguais.

[2] As amostras são das células dos pacientes, colhidas no sangue e medula óssea.

sua vez, os valores dos dados de expressão gênica estão relacionados com a expressão positiva ou negativa de um determinado gene do paciente.

Dados Clínicos	**Descrição**
Study ID	Identificação única do estudo
Patient ID	Identificação única do paciente
Sample ID	Identificação única da amostra
Diagnosis Age	Idade do paciente no momento do diagnóstico de AML
AML in Skin Percentage	Porcentagem de AML encontrada na pele
Bone Marrow Blast Percentage	Porcentagem de células tumorais (blastos) na medula óssea
Cancer Type	Tipo do câncer
Cancer Type Detailed	Detalhamento do tipo do câncer
Cytogenetics	Variações citogenéticas[3] do paciente
Cytogenetic Code (Other)	Categorizações das variações citogenéticas do paciente
DZ Stat	O *Dark Zone Stat*[4] descreve o estado do paciente no momento do transplante
Disease Free (Months)	Tempo de remissão medido em meses
Disease Free Status	Dado binário relacionado ao atributo *Disease Free (Months)*, sendo que o valor 1 corresponde a recaída e 0 caso continue a remissão
FAB	Classificação de risco da FAB
Fraction Genome Altered	Fração do genoma com alteração, referente ao código citogenético
Tumor Other Histologic Subtype	Informações citogenéticas referentes a subtipos de tumores
Induction	Tipo de tratamento
Inferred Genomic Rearangement	Rearranjo cromossômico inferido[5]
Mutation Count	Número de mutações apresentado pelo paciente
Oncotree Code	Dados sobre o tipo de câncer

[3] Para mais informações consulte o Capítulo 2.

[4] A *Dark Zone* (DZ) é utilizada para quantificar a pouca quantidade de blastos no sangue e medula óssea do paciente.

[5] O rearranjo cromossômico inferido é um método de análise genética que permite detectar mudanças estruturais no genoma humano.

PB Blast Percentage	Porcentagem de blastos encontrados no sangue periférico
Race Category	Raça do paciente
Risk (Cyto)	Classificação de risco citogenética
Risk (Molecular)	Classificação de risco molecular
Number of Samples per Patient	Número de amostras[6] coletadas
Sample Type	Tipo da amostra coletada
Sex	Sexo
Somatic Status	Indica se os sintomas dos pacientes coincidem com a doença
Structural Variants	Variantes estruturais[7] encontradas nos pacientes
Subclones	Quantidade de subclones[8]
Transplant Type	Tipo de transplante usado no tratamento dos pacientes, caso o paciente tenha recebido um transplante
WBC	Número de leucócitos no sangue
Overall Survival (Months)	Tempo em meses entre diagnóstico até a morte ou a última avaliação (caso esteja vivo)
Overall Survival Status	Estado de vida do paciente no final do período do estudo, vivo (1) ou morto (0)

Tabela 10 – Descrição dos atributos clínicos da base de dados da TCGA.

3.1.2 Dados da OHSU

A base de dados da OHSU apresenta os dados da coleta de 640 amostras celulares de 535 pacientes. Nesta, uma instância representa uma amostra de um paciente, no qual este pode ter mais de uma amostra colhida. Para o contexto de mutação (606), o valor 1 é atribuído caso o paciente apresente a mutação em um determinado gene e 0 caso contrário. Já a informação de expressão (22.825) está relacionada ao valor da expressão de um determinado gene do paciente, seja ela positiva ou negativa. A Tabela 11 descreve o significado dos dados clínicos (97). O último atributo corresponde ao desfecho do tratamento, portanto, pode ser usado como atributo-alvo em modelos de predição.

Dados Clínicos	**Descrição**
Study ID	Identificação única do estudo

[6] Amostras das células dos pacientes, colhidas no sangue e/ou medula óssea.
[7] Variantes estruturais são alterações na estrutura genética de uma célula.
[8] Os subclones são cópias genéticas diferentes presentes em um conjunto de células.

Patient ID	Identificação única do paciente
Sample ID	Identificação única da amostra
Cancer Study	Identificação única do estudo que gerou a amostra
Cancer Type	Tipo do câncer
Cancer Type Detailed	Detalhamento do tipo do câncer
Number of Samples Per Patient	Número de amostras coletadas
Mutation Count	Número de mutações apresentado no paciente
Sex	Sexo do paciente
Ethnicity Category	Raça do paciente
Platform	Plataforma utilizada na análise dos dados
Age at Diagnosis	Idade do paciente no momento do diagnóstico de AML
Age at Procurement	Idade do paciente quando foram colhidas as amostras para o estudo
Alanine Aminotransferase Level in PB per litre	Níveis de alanina aminotransferase[9] por litro no sangue periférico
Albumin Levels in PB (g/dL)	Níveis de albumina[10] por litro no sangue periférico.
Aspartate Aminotransferase Level in PB per litre	Níveis de aspartato aminotransferase[11] por litro no sangue periférico
Basophils Percent in Peripheral Blood	Porcentagem de basófilos (classe de glóbulos brancos) encontrados no sangue periférico
Bone Marrow Blast Percentage	Porcentagem de células tumorais (blastos) na medula óssea doente
Cause of death source	Causa da morte do paciente, caso tenha morrido
CEBPA Mutation	Se há ou não a mutação no gene *CEBPA*
Chemotherapy	Se o paciente está em tratamento quimioterápico ou não
Cumulative Treatment Regimen Count	Número de regimes de tratamentos realizados no estágio atual da doença

[9] A alanina aminotransferase (ALT) é uma enzima encontrada principalmente no fígado. Quando há dano no fígado, os níveis de ALT no sangue periférico aumentam.

[10] A albumina é uma proteína presente no sangue, produzida pelo fígado.

[11] O aspartato aminotransferase (AST) é uma enzima presente no fígado. Os níveis de AST no sangue periférico podem ser medidos para avaliar as funções do fígado.

Cumulative Treatment Regimens	Regimes de tratamento realizados no estágio atual da doença
Cumulative Treatment Stages	Número de fases dos tratamentos realizados no estágio atual da doença
Cumulative Treatment Type Count	Número de tipos de tratamentos realizados no estágio atual da doença
Cumulative Treatment Types	Tipos de tratamentos realizados no estágio atual da doença
Current Regimen	Regime de tratamento mais recente utilizado
Current Stage	Fase do tratamento mais recente utilizado
Diagnosis	Diagnóstico recebido pelo paciente
Diagnosis at Inclusion	Diagnóstico recebido no momento da obtenção da amostra inicial
DNMT3A Mutation	Se há ou não a mutação no gene *DNMT3A*
Drug Testing data in Analysis	Se os dados de teste do medicamento foram utilizados na análise ou não
Duration of Induction Treatment	Duração da indução do tratamento em dias
ELN 2008 Risk Classification	Primeira classificação de risco proposta pela ELN
ELN 2017 Risk Classification	Segunda classificação de risco proposta pela ELN
Eosinophils Percent in Peripheral Blood	Porcentagem de eosinófilos (tipo de glóbulo branco) no sangue periférico
Ex Vivo Drug Testing	Se o teste de medicamento ex vivo[12] foi realizado na amostra coletada ou não
Exome Seq Analysis	Se os dados da sequenciação do exoma[13] são utilizados na amostra ou não
FAB	Classificação da FAB
FISH Results from Cytogenetics Report	Resultados obtidos pela técnica FISH[14] que identifica anomalias nos cromossomos

[12] O teste de medicamento ex vivo é um tipo de teste, realizado fora do organismo vivo.

[13] A sequenciação do exoma é uma técnica utilizada para identificar mutações genéticas em pacientes com doenças hereditárias.

[14] FISH (*Fluorescence in situ hybridization*) é uma técnica que permite a localização de cromossomos, cromatina, DNA ou RNA em uma amostra.

FLT3-ITD Consensus Call	Se houve consenso ou não em relação à mutação do gene *FLT3-ITD* obtida pelo teste de reação em cadeia da polimerase[15]
Fusion	Gene de fusão (gene híbrido formado por partes de outros genes) utilizado
Group	Divisão de estado do paciente em relação à doença
Hematocrit Levels (%)	Porcentagem de hematócritos[16] no sangue
Hematology serum creatinine laboratory result value in mg dl	Valor do resultado laboratorial de creatinina[17].
Hemoglobin level	Níveis de hemoglobina[18]
IDH1 Mutation	Se há ou não a mutação no gene *IDH1*
IDH2 Negative	Se há ou não a mutação no gene *IDH2*
Immature Granulocytes Percent in Peripheral Blood	Porcentagem de granulócitos imaturos[19] no sangue periférico
JAK2 Negative	Se há ou não a mutação no gene *JAK2*
Karyotype	Resultado do exame de cariótipo[20]
Karyotype and Sample Acquisition Interval Difference in Days	Diferença de dias entre os resultados do cariótipo e a coleta das amostras sanguíneas
KRAS Mutation	Se há ou não a mutação no gene *KRAS*
Laboratory Procedure Lactate Dehydrogenase Result	Resultados do procedimento médico que testa uma amostra para desidrogenase lática[21]

[15] A reação em cadeia da polimerase é uma técnica molecular utilizada para amplificar segmentos específicos de DNA.

[16] Os hematócritos são uma medida dos glóbulos vermelhos no sangue. Eles são geralmente medidos como uma porcentagem do volume total de sangue.

[17] Creatinina é um composto químico produzido pelo metabolismo muscular. É eliminado do corpo através dos rins e sua presença na urina é usada como um indicador da função renal.

[18] A hemoglobina é uma proteína presente nos glóbulos vermelhos cuja função é transportar oxigênio dos pulmões para os tecidos do corpo.

[19] Granulócitos imaturos são células brancas do sangue que ainda não completaram seu desenvolvimento.

[20] O cariótipo é uma técnica utilizada para avaliar a estrutura dos cromossomos em uma célula

[21] A desidrogenase lática (LDH) é uma enzima utilizada para produzir energia para as células. A LDH no sangue, é usada como um marcador de dano celular.

Lymphocytes Percent in Peripherial Blood	Porcentagem de linfócitos[22] no sangue periférico
MDS Two Months Prior to AML Diagnosis	Se houve ou não o diagnóstico de neoplasias mieloproliferativas[23] dois meses antes do diagnóstico de AML
MDS/MPN Diagnosis at Acquisition	Se houve ou não o diagnóstico de neoplasias mieloproliferativas, ou mielodisplásicas[24] no momento do diagnóstico de AML
MDS/MPN Two Months Prior to AML Diagnosis	Se houve ou não o diagnóstico de neoplasias mieloproliferativas, ou mielodisplásicas dois meses antes do diagnóstico de AML
Mean corpuscular volume (fL)	Volume corpuscular médio[25]
Monocytes Percent in Peripherial Blood	Porcentagem de monócitos (tipo de glóbulo branco) no sangue periférico
Most Recent Treatment Duration	Duração do tratamento atual
Most Recent Treatment Type	Tratamento atual realizado pelo paciente
MPN Two Months Prior to AML Diagnosis	Se houve ou não o diagnóstico de neoplasias mielodisplásicas dois meses antes do diagnóstico de AML
Negative for FLT3	Se há ou não a mutação no gene *FLT3*
Negative for KIT	Se há ou não a mutação no gene *KIT*
Neutrophils Percent in Peripheral Blood	Porcentagem de neutrófilos (tipo de glóbulo branco) no sangue
Non AML/MPN/MDS Diagnosis at Acquisition	Se não houve ou houve o diagnóstico de neoplasias mieloproliferativas, ou mielodisplásicas, ou de AML na apresentação dos sintomas

[22] Os linfócitos são uma das principais células do sistema imunológico, responsáveis por combater infecções e doenças.

[23] As neoplasias mieloproliferativas são um grupo de doenças malignas que se originam na medula óssea e se caracterizam pela proliferação anormal de células.

[24] As neoplasias mielodisplásicas são um grupo de doenças hematológicas malignas que afetam a medula óssea e as células sanguíneas. Elas se caracterizam por uma disfunção na maturação das células da medula.

[25] O volume corpuscular médio é um parâmetro utilizado para avaliar o tamanho dos glóbulos vermelhos.

NPM1 Consensus Call	Se houve consenso ou não em relação à mutação do gene *NPM1* obtida pelo teste de reação em cadeia da polimerase
NRAS Mutation	Se há ou não a mutação no gene *NRAS*
Nucleated RBCs Percent in Peripheral Blood	Porcentagem de hemácias nucleadas (glóbulos vermelhos) no sangue periférico
Number of Cumulative Treatment Stages	Número de fases do tratamento no estágio atual da doença
Oncotree Code	Dados sobre o tipo de câncer
PB Blast Percentage	Porcentagem de blastos encontrados no sangue periférico
Platelet count	Contagem de plaquetas (células que ajudam na coagulação sanguínia) na amostra de sangue
Prior Diagnosis of Cancer	Se o paciente recebeu ou não diagnóstico de câncer anterior ao de AML
Prior Malignancy Radiation Therapy	Se o paciente recebeu ou não tratamento radioterápico, para tratamentos de tumores malignos
Prior MDS	Se o paciente recebeu ou não o diagnóstico de neoplasias mielodisplásicas antes do diagnóstico de AML
Prior MDS/MPN	Se o paciente recebeu ou não o diagnóstico de neoplasias mieloproliferativas, ou mielodisplásicas antes do diagnóstico de AML
Prior MPN	Se o paciente recebeu ou não o diagnóstico de neoplasias mieloproliferativas antes do diagnóstico de AML
Response to Induction Treatment	Resposta do paciente em relação à indução do tratamento
RNA Seq Analysis	Análise do sequenciamento do RNA do paciente
RNA Sequenced	Se houve ou não análise do sequenciamento do RNA do paciente
Sample collection center	Centro de coleta da amostra
Sample Timepoint	Tempo (novo diagnóstico ou recaída) do estágio da doença em que a amostra foi coletada
Site of Sample	Região do corpo em que a amostra sanguínea foi coletada
Somatic Status	Estado (normal ou não) da amostra sanguínea colhida
Specific Diagnosis at Acquisition	Subtipo específico de AML diagnosticado no início da doença

Specific Diagnosis at Inclusion	Subtipo específico de AML diagnosticado no momento em que a amostra sanguínea foi colhida
Surface Antigens	Presença ou não do marcador biológico do antígeno de superfície[26]
TMB (nonsynonymous)	Carga mutacional tumoral[27]
Total Protein Levels in the Blood (g/dL)	Quantidade das proteínas albumina e globulina[28] no sangue
TP53 Mutation	Se há ou não a mutação no gene *TP53*
Treatment Type	Tipo de tratamento utilizado
Type of Induction Treatment	Tipo de indução de tratamento recebida
WBC	Número de leucócitos (glóbulos brancos) no sangue
Whole Exome Sequencing	Sequenciação do exoma.
BRAF mutation	Se há ou não a mutação no gene *BRAF*
CEBPA Biallelic Mutation	Se o paciente apresenta ou não mutação em dois alelos (versões diferentes de um gene) relacionados ao gene *CEBPA*
Overall Survival Status	Estado do paciente durante o período do estudo, vivo (1) ou morto (0). Este é o atributo-alvo

Tabela 11 – Descrição dos atributos clínicos da base de dados da OHSU.

3.2 Limpeza e pré-processamento dos dados

Como foram utilizadas bases provenientes de fontes diferentes, os dados foram pré-processados para garantir sua consistência. Com o auxílio dos especialistas da área, foram removidos os seguintes dados espúrios:

1. Amostras que não se encaixam no quadro de AML adulta, observado pela idade do paciente, que não deve ser inferior a 18 anos;

[26] O antígeno de superfície é uma proteína encontrada na superfície da célula. Ele pode ser utilizado para identificar e caracterizar diferentes tipos de células.

[27] A carga mutacional tumoral é uma medida da quantidade de mutações presentes em um tumor.

[28] A globulina é um tipo de proteína encontrada no sangue e no plasma. Elevações ou diminuições nos níveis de globulina no sangue podem indicar problemas de saúde.

2. Amostras que não se encaixam no quadro de AML, observado a quantidade de blastos na medula óssea, que não deve ser inferior a 20% (ARBER et al., 2016);

3. Amostras sem informações sobre o estado do paciente durante o período do estudo (*Overall Survival Status*);

4. Múltiplas amostras sanguíneas do mesmo paciente. Foram excluídas todas as instâncias da base de dados da OHSU em que o valor do atributo *Site of Sample* diferiu de *Bone Marrow Aspirate*[29], isto é, todas as amostras sanguíneas coletadas fora da medula óssea. A escolha de coletar a amostra sanguínea na medula óssea foi justificada pela disponibilidade de dados na base de dados da TCGA, que somente possui amostras sanguíneas coletadas nesta região;

5. Todas as instâncias da base de dados da OHSU em que o valor do atributo *Sample Timepoint* diferiu de *Denovo*[30]. A separação foi realizada devido à limitação dos dados disponíveis na base de dados da TCGA, que possuem apenas amostras sanguíneas de pacientes com *AML Denovo*[31].

6. Atributos de identificação do tipo de câncer, uma vez que todos os pacientes receberam o diagnóstico de AML; e

7. Atributos presentes em apenas uma base de dados.

Para cada paciente, está descrito o tipo de terapia utilizada para tratá-lo, variando de informações relacionadas a tratamentos paliativos, descrição de diversos procedimentos quimioterápicos até transplante de médula óssea. Essas informações são complexas, de difícil interpretabilidade para modelos de ML, que necessitam de simplificação para serem utilizados conforme a proposta deste estudo. Os especialistas no domínio de dados analisaram e agruparam os dados descritos para gerar categorias de tratamento segundo a intensidade de cada terapia, resultando em 4 grupos de tratamentos, descritos a seguir:

1. *High intensity therapy* (HIT): plano terapêutico que inclui consolidação com transplante (autólogo ou alogênico).

2. *Low intensity therapy* (LIT): plano terapêutico com potencial não curativo/terapia paliativa. Geralmente recomendada para pacientes idosos ou com resposta ruim aos demais tratamentos.

[29] A amostra sanguínea é coletada através de uma punção óssea, geralmente do osso do quadril, para obtenção de células da medula óssea.

[30] A expressão *"Sample Timepoint = Denovo"* indica que a amostra sanguínea foi coletada pela primeira vez, sem qualquer histórico de coleta anterior. Esta amostra está relacionada à *"AML Denovo"*.

[31] *"AML Denovo"* é um câncer que não tem relação com nenhuma outra condição pré-existente ou tratamento anterior, ocorrendo espontaneamente.

3. *Regular therapy* (RT): plano terapêutico baseado na combinação de citarabina e antraciclina (7+3) e suas variações.

4. *Target therapy* (TT): plano terapêutico que inclui um fármaco que atua de forma alvo seletiva. Geralmente, é recomendado para pacientes com uma alteração molecular específica, na qual o fármaco atua.

O mesmo procedimento foi realizado para os dados com informações citogenéticas, padronizando a nomenclatura utilizada para descrever cada uma e agrupando quando seu significado era o mesmo. Ao fim, restaram 11 atributos clínicos em comum entre as bases de dados. Em relação às variáveis de caráter genético, foram encontrados 14.712 atributos de expressão em comum entre as duas bases e 318 atributos de mutação. No entanto, foram removidas 37 variáveis genéticas que não apresentaram mutação para nenhum paciente, restando assim 281 atributos sobre as mutações. Posteriormente, foram selecionadas as amostras de pacientes que apresentaram informações tanto clínicas quanto genéticas, por meio das duas bases de dados.

Após o processo de limpeza e pré-processamento, restaram 272 amostras dentre as 872 inicialmente presentes na combinação das bases de dados. A Tabela 12 apresenta a quantidade de dados resultante dessa etapa.

			Atributos		
Base de dados	**Amostras**[32]	**Pacientes**	**Clínicos**	**Mutação**	**Expressão**
TCGA e OHSU	272	272	11	281	14.712

Tabela 12 – Resumo da integração das bases de dados após as fases de limpeza e pré-processamento.

3.3 Preenchimento dos valores faltantes

Após a junção dos dados provenientes das duas bases, foi analisada a existência de valores faltantes. Os dados de natureza genética estavam todos preenchidos. A Figura 9 apresenta um gráfico que ilustra os atributos clínicos com valores faltantes pela amostra de ocorrência. Conforme pode ser observado, somente os atributos *Bone Marrow Blast Percentage*, *Mutation Count*, *PB Blast Percentage* e *WBC* tinham valores faltantes.

O método dos 3-vizinhos mais próximos (do inglês, *3-Nearest neighborhood – 3-NN*) foi utilizado para preencher os valores faltantes. O preenchimento dos dados foi realizado por meio da predição dos modelos gerados com os demais atributos (*i.e.*, aqueles sem valores faltantes), sendo que o atributo com valor faltante foi usado como atributo-alvo. Este processo foi repetido até que todos os atributos tivessem seus dados faltantes preenchidos.

[32] Amostras sanguíneas dos pacientes retiradas da medula óssea.

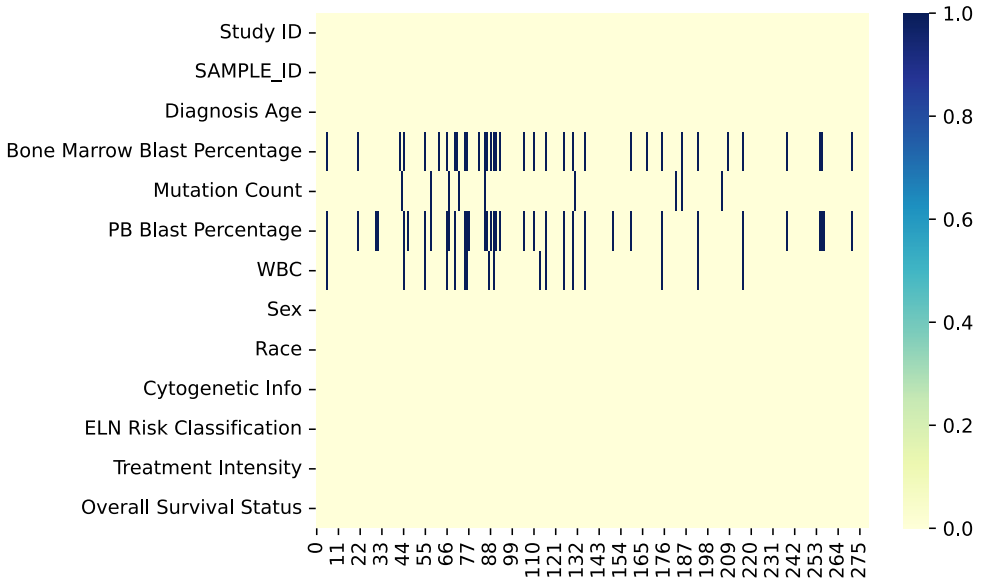

Figura 9 – Valores faltantes nos dados clínicos por amostra.

O método 3-NN foi escolhido em detrimento dos métodos tradicionais (*e.g.*, a mediana da classe na qual a instância pertence ou a predição do valor por meio de regressão linear) porque o resultado do 3-NN apresentou uma distribuição de dados mais próxima da original. As Figuras 10 a 13 ilustram o comparativo das distribuições destes atributos em relação aos métodos supracitados e os dados originais.

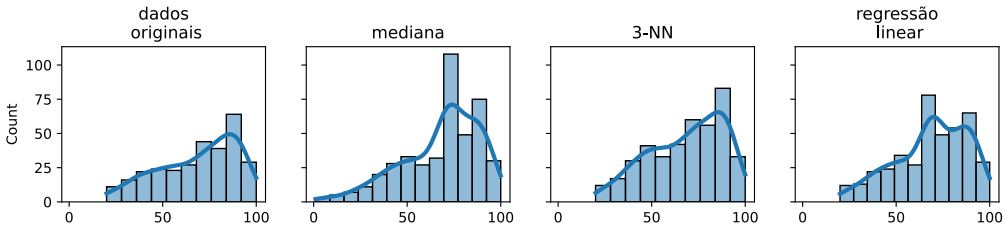

Figura 10 – Comparativo entre diferentes distribuições do atributo *Bone Marrow Blast Percentage* em relação aos diferentes métodos analisados para preencher os valores faltantes.

3.4 Análise dos atributos

Esta seção descreve o processo de análise dos atributos realizado para cada tipo de dado. As seções a seguir descrevem o procedimento de análise realizado para os atributos clínicos, de mutação e expressão gênica.

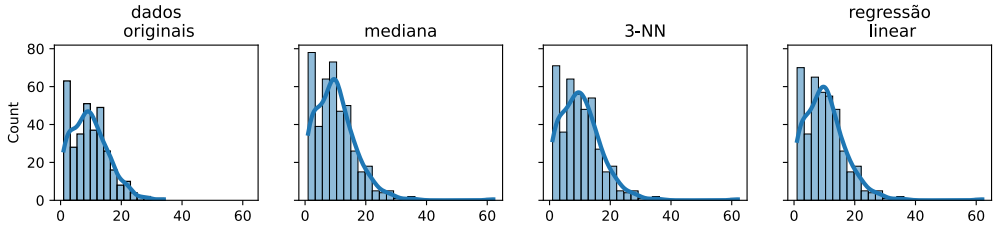

Figura 11 – Comparativo entre diferentes distribuições do atributo *Mutation Count* em relação aos diferentes métodos analisados para preencher os valores faltantes.

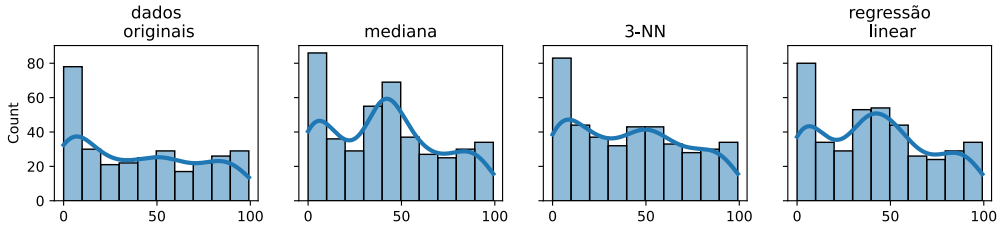

Figura 12 – Comparativo entre diferentes distribuições do atributo *PB Blast Percentage* em relação aos diferentes métodos analisados para preencher os valores faltantes.

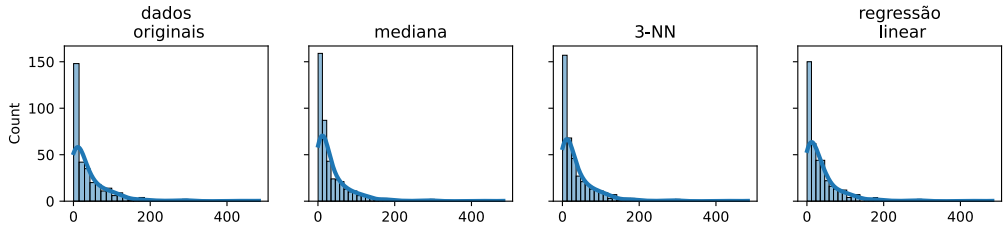

Figura 13 – Comparativo entre diferentes distribuições do atributo *WBC* em relação aos diferentes métodos analisados para preencher os valores faltantes.

3.4.1 Dados clínicos

Dentre os atributos clínicos comuns nas duas bases de dados, foram selecionados 13 atributos segundo a análise de relevância para o prognóstico de risco, realizada pelos especialistas no domínio dos dados. Destes, 2 atributos (*Study ID* e *Sample ID*) são relativos à identificação da amostra do paciente, 10 informações clínicas dos pacientes e 1 atributo-alvo (*Overall Survival Status*). As Figuras 14 e 15 resumem as principais estatísticas referentes a estes dados, divididos pelas duas classes-alvo.

A Tabela 13 descreve a combinação dos atributos das diferentes fontes para a geração da base final. Os valores relacionados ao atributo *Sexo* foram convertidos para números, onde 0 representa masculino e 1 feminino. Os relacionados a *Race Category* foram transformados em um novo atributo *Race*, onde 1 representa a raça branca e 0 as demais. Essa conversão foi motivada pelo fato das pessoas brancas representarem ao

menos 75% dos dados. Os dados referentes à classificação de risco foram combinados para a classificação proposta pela ELN-2017. Os dados provenientes do TCGA não contém a classificação da ELN e, para realizá-la, foram combinados os atributos *Risk (Cyto)* e *Risk (Molecular)* em conjunto com a análise dos especialistas no domínio. Na base de dados da OHSU, há 14 atributos relacionados aos tratamentos realizados e, na TCGA, apenas 1. Esses atributos foram combinados pelos especialistas resultando no atributo *Treatment Intensty*, de acordo com o descrito na Seção 3.2.

TCGA	**OHSU**	**Base final**
Study ID	*Study ID*	*Study ID*
Sample ID	*Sample ID*	*Sample ID*
Diagnosis Age	*Age at Diagnosis*	*Diagnosis Age*
Bone Marrow Blast Percentage	*Bone Marrow Blast Percentage*	*Bone Marrow Blast Percentage*
Mutation Count	*Mutation Count*	*Mutation Count*
PB Blast Percentage	*PB Blast Percentage*	*PB Blast Percentage*
WBC	*WBC*	*WBC*
Sex	*Sex*	*Sex*
Race Category	*Ethnicity Category*	*Race*
Cytogenetic Info	*Karyotype*	*Cytogenetic Info*
Risk (Cyto) + *Risk (Molecular)* + Análise especialista	*ELN 2017 Risk Classification*	*ELN Risk Classification*
Induction + Análise especialista	*Cumulative Treatment Regimen Count* + *Cumulative Treatment Regimens* + *Cumulative Treatment Stages* + *Cumulative Treatment Type Count* + *Cumulative Treatment Types* + *Current Regimen* + *Current Stage* + *Duration of Induction Treatment* + *Most Recent Treatment Duration*+ *Most Recent Treatment Type* + *Number of Cumulative Treatment Stages* + *Response to Induction Treatment* + *Treatment Type* + *Type of Induction Treatment* + Análise especialista	*Treatment Intensity*
Overall Survival Status	*Overall Survival Status*	*Overall Survival Status*

Tabela 13 – Combinação de atributos clínicos das diferentes fontes de dados.

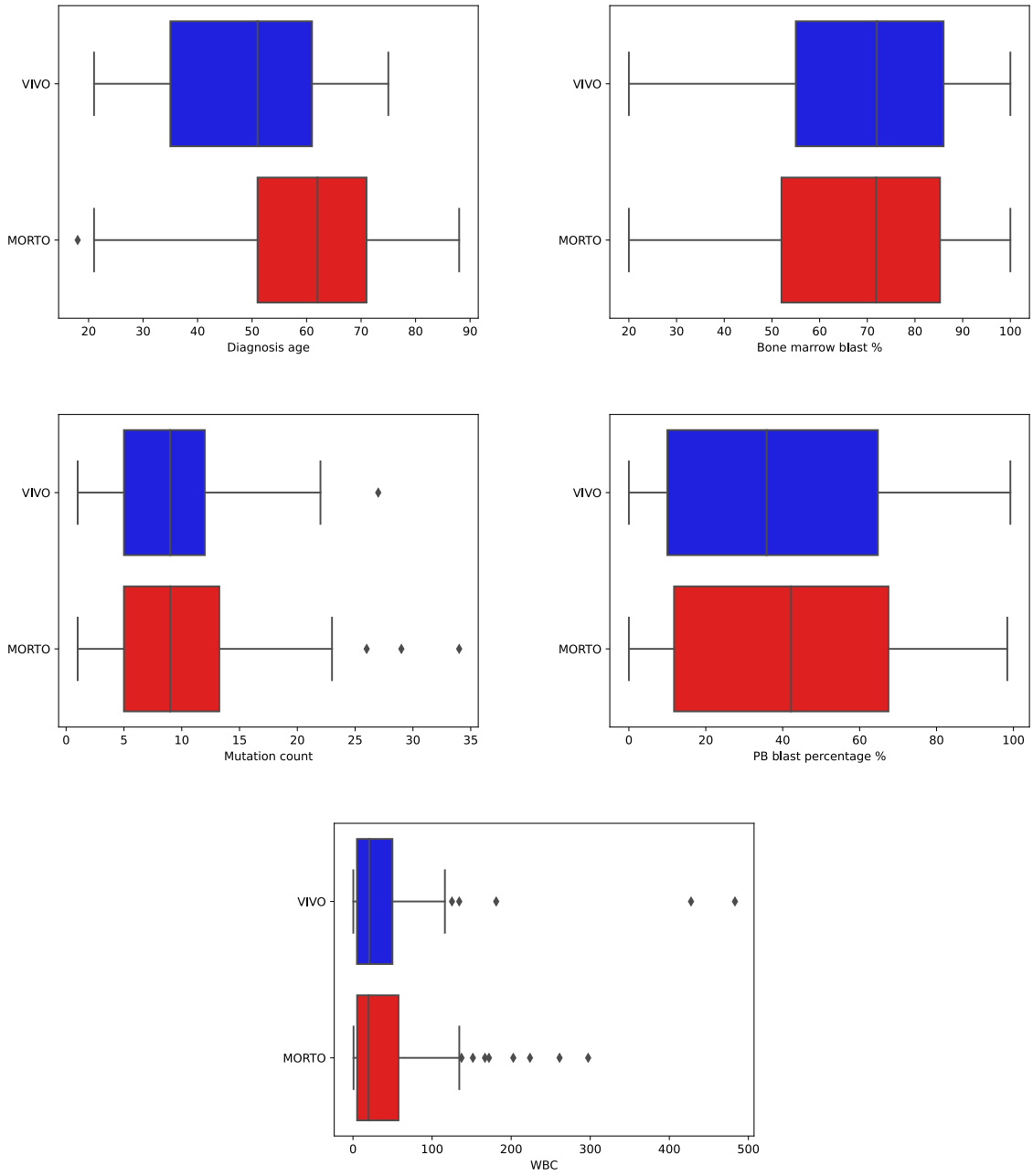

Figura 14 – Boxplots dos atributos clínicos de natureza numérica.

3.4.2 Dados de mutação genética

Foi empregado o método *chi-square* (χ^2) (PEARSON, 1900) para a análise dos 281 atributos de mutação. O χ^2 é um teste estatístico utilizado para avaliar a associação entre duas variáveis categóricas (PEARSON, 1900). No contexto de análise de mutações gênicas, o teste identifica os atributos potencialmente associados com a sobrevida do paciente. O emprego do teste χ^2 foi motivado por estudos anteriores que o aplicaram para analisar a correlação entre mutações gênicas e alguns tipos de câncer (RAHMAN et al., 2019).

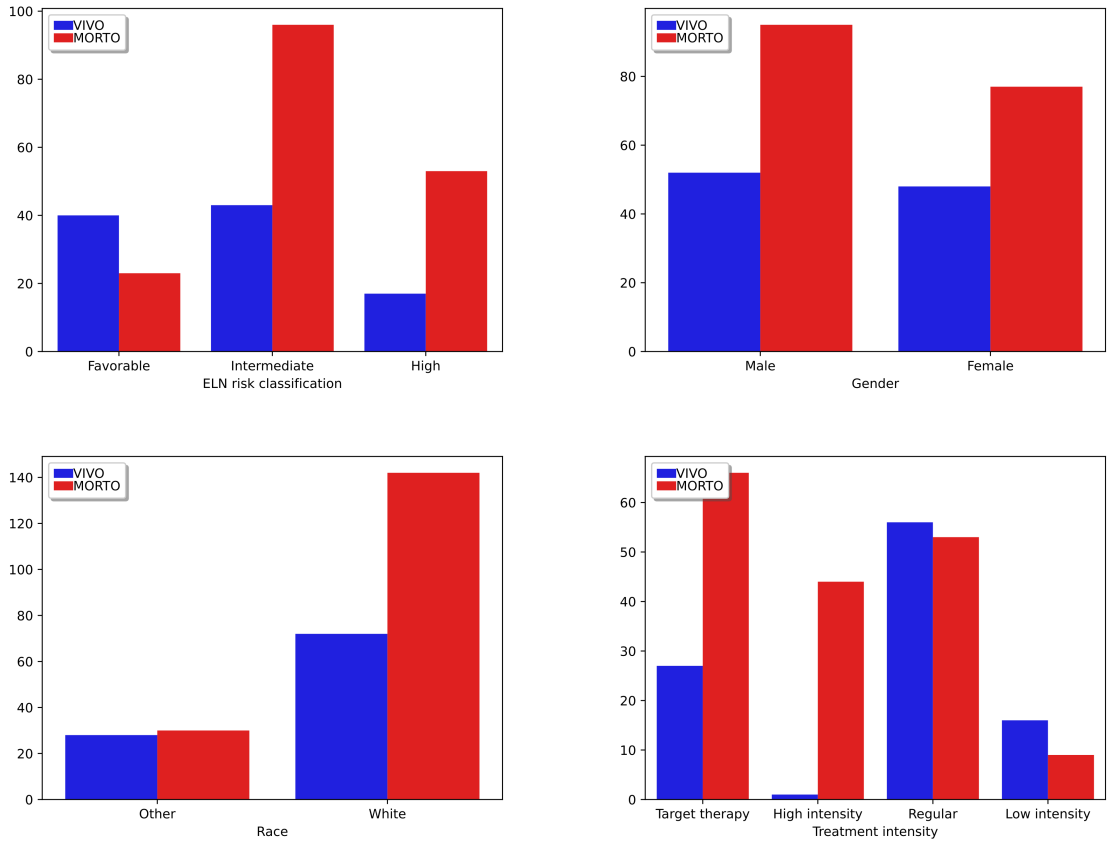

Figura 15 – Gráficos de barra dos atributos clínicos de natureza categórica.

O teste de hipótese foi formulado da seguinte forma:

- H0 – a sobrevivência do paciente é independente da mutação do gene;
- H1 – ambos os grupos são dependentes.

É importante salientar que estes dados apresentam uma natureza esparsa, uma vez que a maioria dos pacientes apresenta poucas mutações em relação à quantidade total de genes e há uma abundância de genes com poucas mutações. A variabilidade genética entre os pacientes pode levar a uma grande heterogeneidade na distribuição de mutações, contribuindo para a esparsidade desse conjunto de dados (KADIA et al., 2015).

Com grau de confiança em 90% e $\alpha = 0,1$, apenas três genes foram selecionados, sendo eles *TP53*, *U2AF1* e *PKD1L2*. Embora não seja usual, esse valor de grau de confiança foi empregado, pois em 95%, valor padrão utilizado na literatura (RAHMAN et al., 2019), apenas o gene *TP53* é selecionado.

O parecer dos especialistas referente aos genes selecionados foi o seguinte:

A mutação *TP53* é considerada a mais importante entre as mutações identificadas. Vários estudos demonstram a relação entre a mutação *TP53* com a resposta terapêutica e o prognóstico. O gene *TP53* é considerado o guardião da estabilidade genômica, pois controla a progressão do ciclo celular e a apoptose em situações de estresse ou danos no DNA. Mutações neste gene são encontradas em cerca de metade dos pacientes com câncer (KASTENHUBER; LOWE, 2017; MONTI et al., 2020). Embora as mutações em *TP53* sejam menos comuns em pacientes com AML (cerca de 10%), elas preveem um mau prognóstico (GROB et al., 2022; PAPAEMMANUIL et al., 2016).

As mutações no gene *U2AF1* são mais comuns em síndromes mielodisplásicas e raras em AML *De novo* (XU et al., 2017; PAPAEMMANUIL et al., 2016), mas já foram associadas a um prognóstico adverso em neoplasias mieloides (ZHU et al., 2021). O *U2AF1* regula os processos de *splicing* de pré-mRNA para gerar mRNAs funcionais, sendo considerado um elemento chave no *spliceossomo* (ZHAO et al., 2022). Por outro lado, a relevância prognóstica das mutações em *PKD1L2* e sua relação com a resposta ao tratamento é desconhecida em AML e requer mais investigação. *PKD1L2* é membro da família de proteínas *poli-cistina-1-like* (receptores acoplados a proteínas G que medem canais de cátions) e alterações na quantidade de cópias desse gene foram associadas à suscetibilidade à carcinogênese (PARK et al., 2017).

3.4.3 Dados de expressão genética

Foi utilizado o método de análise de variância (do inglês, *Analysis of Variance* – ANOVA) (FISHER, 1918) para análise dos 14.172 atributos de expressão. A ANOVA é uma técnica estatística utilizada para testar a hipótese de que as médias de diferentes grupos são iguais. No contexto da análise de dados de expressão gênica, a ANOVA avalia a importância de cada expressão em relação a sobrevida do paciente. Esta técnica foi escolhida para esta análise devido ao seu uso bem-sucedido em diversos estudos anteriores, como de câncer de mama (ARUNASRI et al., 2013), câncer de pulmão (LI et al., 2010) e AML (WILSON et al., 2006).

Inicialmente, foi calculado o valor da métrica-F (variação entre amostras) para cada atributo em relação ao atributo-alvo. O teste de hipótese foi formulado da seguinte forma:

- H0 – as médias dos grupos são iguais;
- H1 – há diferença entre os grupos e, portanto, a hipótese nula pode ser rejeitada.

Após essa etapa, usando $p < 0,05$ (valor padrão utilizado nos demais estudos), restaram 2.169 atributos de expressão gênica que foram ordenados de forma decrescente em relação à métrica-F. Após a ordenação, foram treinados modelos de florestas aleatórias usando 80% dos dados. O primeiro modelo foi treinado com apenas o gene melhor ranqueado, já o segundo modelo foi treinado com os dois genes melhores ranqueados e assim por diante. Em seguida, estes modelos foram comparados com base no desempenho aferido

pela F-medida em um conjunto de validação (10% dos dados). Os modelos que usaram 16 e 19 atributos genéticos tiveram o melhor desempenho, alcançando uma F-medida de 84%. Ao fim, foi escolhido o modelo com 19 atributos por apresentar uma maior quantidade de genes, esta decisão foi guiada pelo especialista.

Os 19 atributos (genes) selecionados foram (em ordem descrente de métrica-F) *MICALL2*, *LTK*, *OSBPL5*, *CFD*, *AGRN*, *PPP1R26*, *SLC29A2*, *ATP13A2*, *LPPR3*, *FNDC3B*, *SLC25A29*, *MPO*, *ITFG1*, *CHD3*, *ELANE*, *ARHGEF10*, *GLB1*, *MAP3K1* e *CHIC1*[33].

A Figura 16 ilustra o processo de avaliação dos modelos gerados por meio das métricas precisão, revocação e F-medida. O gráfico demonstra a evolução do desempenho dos modelos conforme novos atributos foram inseridos. O eixo x foi fixado em 30 pelo fato da inserção dos demais atributos não apresentar melhorias no desempenho.

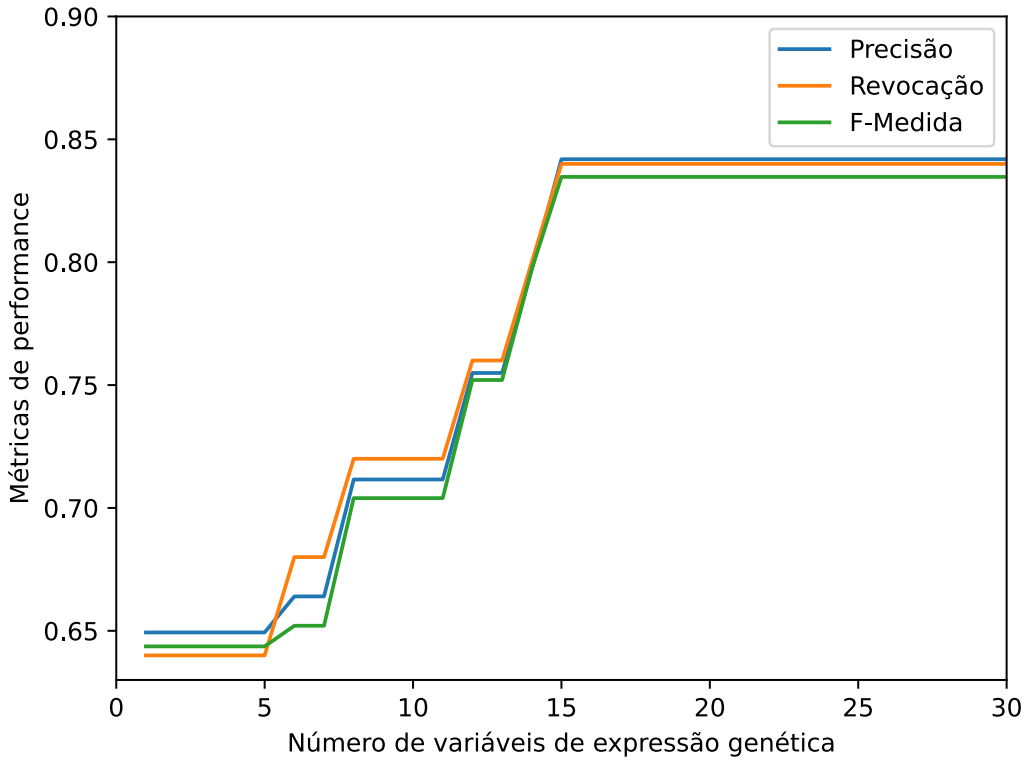

Figura 16 – Evolução da capacidade preditiva conforme dados de expressão são inseridos.

Os especialistas realizaram a análise dos genes selecionados e emitiram o seguinte parecer:

> Entre os genes selecionados cuja expressão foi associada a predição de sobrevida, apenas alguns estão relacionados aos processos relacionados à leucemogênese: *LTK* (ROLL; REUTHER, 2012a; LI et al., 2020),

[33] Mais informações sobre estes genes podem ser encontradas no GeneCards.

LPPR3 (YANG et al., 2018), *FNDC3B* (CK et al., 2017; WANG et al., 2016), *MPO* (OLSSON et al., 1979) e *MAP31* (ROLL; REUTHER, 2012b; NAWATA et al., 2003; TANG et al., 2021a). Assim, os resultados abrem caminhos para a identificação de novos genes potenciais associados à resposta terapêutica na AML.

Em particular, o gene *MICALL2* chamou a atenção. A proteína codificada por este gene regula potencialmente a dinâmica do citoesqueleto, a formação de junções estreitas e o crescimento de neurites, e seu papel em malignidade em tumores sólidos foi descrito. Por exemplo, *MICALL2* está altamente expressa no câncer colorretal (WEN et al., 2022), e, do ponto de vista funcional, a sobre-expressão de *MICALL2* induz a proliferação celular, a migração e a tumorigênese por meio da regulação do caminho *WNT/β-catenina* nesta doença (WEN et al., 2022). Resultados similares foram descritos para cânceres ovarianos, pulmonares e gástricos através da regulação de importantes caminhos oncogênicos (i.e., *MYC*, *EGFR*) (ZHU et al., 2015; MIN et al., 2020; MIN et al., 2019). Em uma análise pan-câncer, *MICALL2* estava altamente expressa em 16 de 33 cânceres comparados com tecidos normais (LIN et al., 2022). Esses achados sugerem que o papel biológico de *MICALL2* na AML pode ser abordado em futuros estudos clínicos e funcionais.

Alguns estudos também associaram os genes *OSBPL5* (HU; YU; WANG, 2022), *AGRN* (WANG et al., 2021; LI et al., 2018), *PPP1R26* (YANG et al., 2022; LI et al., 2018), *SLC25A29* (ZHANG et al., 2018) e *ITFG1* (CK et al., 2017) com o desenvolvimento e/ou progresso de tumores sólidos e não tinham sido associados anteriormente a malignidades hematológicas (incluindo AML).

3.5 Base de dados resultante

Após as fases de limpeza, pré-processamento e análise dos atributos, a base de dados final resultante ficou composta por 272 amostras sanguíneas de pacientes, representados por 11 atributos clínicos (CLIN), 19 expressões de genes (EXP) e 3 mutações de genes (MUT). Os valores dos dados clínicos de natureza numérica (Figura 14) foram padronizados para a distribuição normal padrão, com média 0 e desvio pardão igual à 1. A Tabela 14 sumariza a composição da base de dados resultante, que está publicamente disponível no github.

Conjuntos de dados	**#Atributos**	**Atributos**
Clínicos (CLIN)	11	*Diagnosis age, Bone marrow blast (%), Mutation count, PB blast (%), WBC, Sex, Race, Cytogenetic info, ELN risk classification, Treatment intensity classification, Overall survival status (class)*
Expressão Genética (EXP)	19	*MICALL2*, *LTK*, *OSBPL5*, *CFD*, *AGRN*, *PPP1R26*, *SLC29A2*, *ATP13A2*, *LPPR3*, *FNDC3B*, *SLC25A29*, *MPO*, *ITFG1*, *CHD3*, *ELANE*, *ARHGEF10*, *GLB1*, *MAP3K1* e *CHIC1*
Mutação Genética (MUT)	3	*TP53*, *U2AF1*, e *PKD1L2*

Tabela 14 – Conjuntos de dados resultantes após as etapas de limpeza, pré-processamento e análise dos atributos.

4 Sistema de suporte à decisão

Neste trabalho, foi assumida a hipótese de que o emprego de métodos de ML tem potencial para contribuir na descoberta de novos padrões gênicos e clínicos que possam auxiliar os especialistas em suas decisões terapêuticas. Para esse fim, com a colaboração de especialistas em AML, foi proposto e desenvolvido um sistema de apoio à decisão que visa recomendar automaticamente conjuntos de tratamentos adequados para pacientes com AML em função da predição automática da mortalidade/sobrevivência.

Com o sistema proposto, a subjetividade e o tempo gasto para um especialista selecionar um tratamento apropriado para pacientes com AML poderão ser minimizados. Isso é possível graças ao emprego de técnicas de ML, que permitem personalizar as decisões terapêuticas conforme as características individuais de cada paciente. Além disso, o sistema também considera informações clínicas e genéticas, além das utilizadas pela classificação de risco da ELN, o que pode aumentar a assertividade das recomendações. Como resultado, é esperado que o uso desse sistema possa contribuir para o aumento no tempo de sobrevida e na qualidade de vida dos pacientes.

A Figura 17 sumariza o *pipeline* empregado para o desenvolvimento do sistema proposto. As fases de obtenção dos dados, limpeza, pré-processamento e análise de atributos foram apresentadas no Capítulo 3. A combinação dos conjuntos de dados e o treinamento dos modelos de predição são descritos na Seção 4.1. Por fim, a Seção 4.2 descreve o comitê de classificação usado para selecionar o conjunto de tratamentos mais adequado.

4.1 Treinamento dos modelos de predição

Foram utilizados três métodos consolidados de aprendizado de máquina para a geração dos modelos de predição de sobrevivência (sim/não) baseado na escolha da intensidade do tratamento (*target*, *regular*, *low-intensity*, e *high-intensity*). Os métodos são florestas aleatórias (RF) (BREIMAN, 2001), regressão logística (LR) (CRAMER, 2003) e máquinas de vetores de suporte (SVM) (BOSER; GUYON; VAPNIK, 1992). Estes métodos foram implementados com funções da biblioteca *sickit-learn* (PEDREGOSA et al., 2011), utilizando a linguagem de programação Python.

Não foram utilizadas técnicas de aprendizado profundo por duas razões importantes: (*i*) a pequena quantidade de amostras disponível, uma vez que, estas técnicas demandam grande volume de dados para alcançarem resultados satisfatórios (JAN et al., 2019); e (*ii*) para os especialistas do domínio é muito importante que os resultados da predição sejam interpretáveis. O SVM com *kernel* linear foi escolhido por ser facilmente interpretável, já

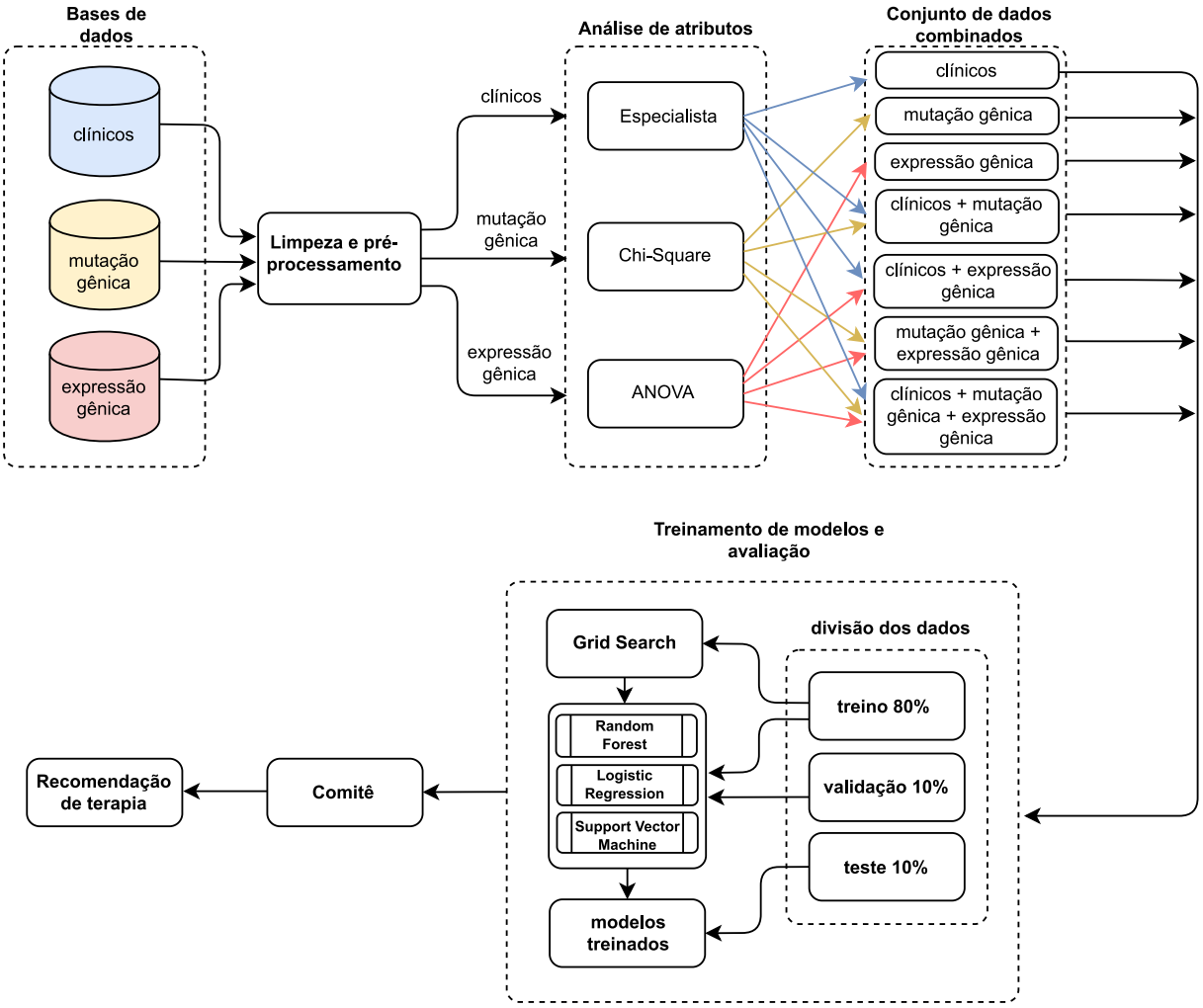

Figura 17 – *Pipeline* do sistema de suporte à decisão proposto.

que sua separação de classes é feita em uma fronteira de decisão linear. A RF também é de certa forma interpretável porque através dela é possível analisar o ranqueamento da importância dos atributos para a predição e a visualização das melhores árvores de decisão geradas.

Os três métodos de classificação (RF, LR e SVM) foram isoladamente empregados para treinar diferentes modelos de predição usando todas as combinações possíveis dos três conjuntos de atributos (CLIN, MUT, EXP) apresentados na Tabela 14. Ao todo, foram treinados 21 modelos de predição (3 métodos de ML $\times$ 7 combinações dos conjuntos de atributos). Este processo é ilustrado nas duas primeiras linhas da Figura 19. Os modelos que obtiveram o melhor desempenho para cada combinação de conjunto de atributos foram combinados na forma de um comitê de classificação, conforme descrito na Seção 4.2.

4.2 Comitê de classificação

A Figura 19 ilustra o processo empregado na criação do sistema de apoio à decisão proposto neste trabalho. Cada fase é apresentada a seguir.

A avaliação inicial dos modelos de classificação para a predição de sobrevivência de pacientes com AML é realizada com base nas combinações dos conjuntos de dados clínicos (CLIN), mutação genética (MUT) e expressão genética (EXP). Ao final do processo de treinamento e avaliação, os modelos que apresentaram o melhor desempenho para cada combinação de conjuntos de atributos foram selecionados para compor o comitê de classificação.

Neste trabalho, foram avaliados dois tipos de comitês de classificação (Figuras 18 e 19): (*i*) somente a partir de dados clínicos e de mutações genéticas (*i.e.*, dados que podem ser obtidos facilmente e apresentados em uma primeira visita clínica); e (*ii*) com a adição de dados de expressões genéticas (*i.e.*, dados mais custosos de serem obtidos). Esta divisão foi motivada para atender a realidade de diferentes ambientes clínicos que, em alguns casos, não dispõem dos dados de expressão genética dos seus pacientes. Estes comitês de classificação são compostos pelos melhores modelos individuais obtidos com as combinações dos conjuntos de atributos (CLIN, MUT e EXP). O comitê (*i*) possui 3 classificadores e o (*ii*) possui 7. A predição final dos comitês é o resultado do voto majoritário dos classificadores individuais.

A recomendação do conjunto de terapias mais adequado é baseada no agrupamento proposto pelos especialistas no qual há 4 grupos que variam conforme as intensidades:

1. *High intensity therapy* (HIT): plano terapêutico que inclui consolidação com transplante (autólogo ou alogênico);
2. *Low intensity therapy* (LIT): plano terapêutico com potencial não curativo/terapia paliativa. Geralmente recomendado para pacientes idosos ou com resposta ruim aos demais tratamentos;
3. *Regular therapy* (RT): plano terapêutico baseado na combinação de citarabina e antraciclina (7+3) e suas variações; e
4. *Target therapy* (TT): plano terapêutico que inclui um fármaco que atua de forma alvo seletiva. Geralmente recomendado para pacientes com uma alteração molecular específica, na qual o fármaco atua.

O sistema proposto recomenda o grupo de terapias com a mais alta probabilidade de sobrevivência do paciente com AML. Para isso, são calculadas as probabilidades de sobrevivência do paciente para cada grupo de tratamento através dos comitês de classificação.

A Figura 20 ilustra um exemplo do processo de escolha de nível de tratamento. Para cada entrada de informações dos pacientes (atributos CLIN, MUT e EXP) são computadas 4 predições de sobrevivência, uma para cada guia de tratamento (HIT, LIT, RT e TT). A partir das saídas de probabilidade de sobrevivência, seja do comitê ou do modelo individual, é escolhido o guia de tratamento que apresenta a maior probabilidade.

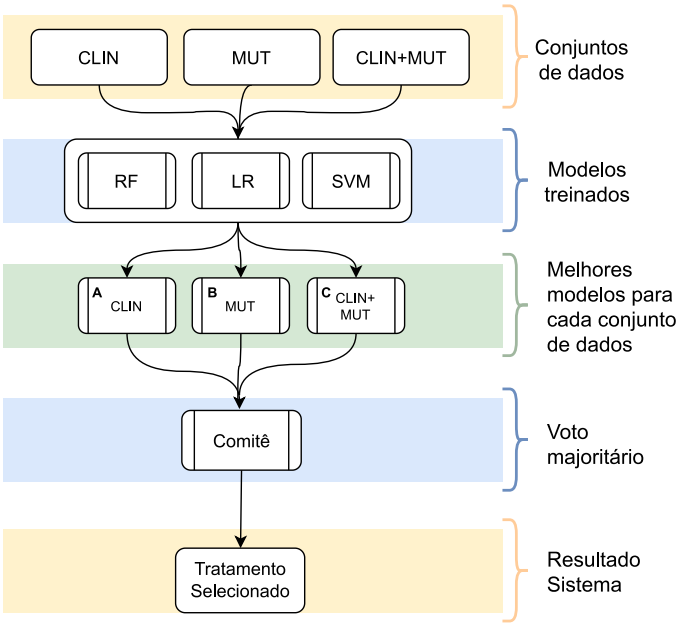

Figura 18 – Processo envolvido para o treinamento do comitê de classificação com os conjuntos de dados clínicos e de mutação genética.

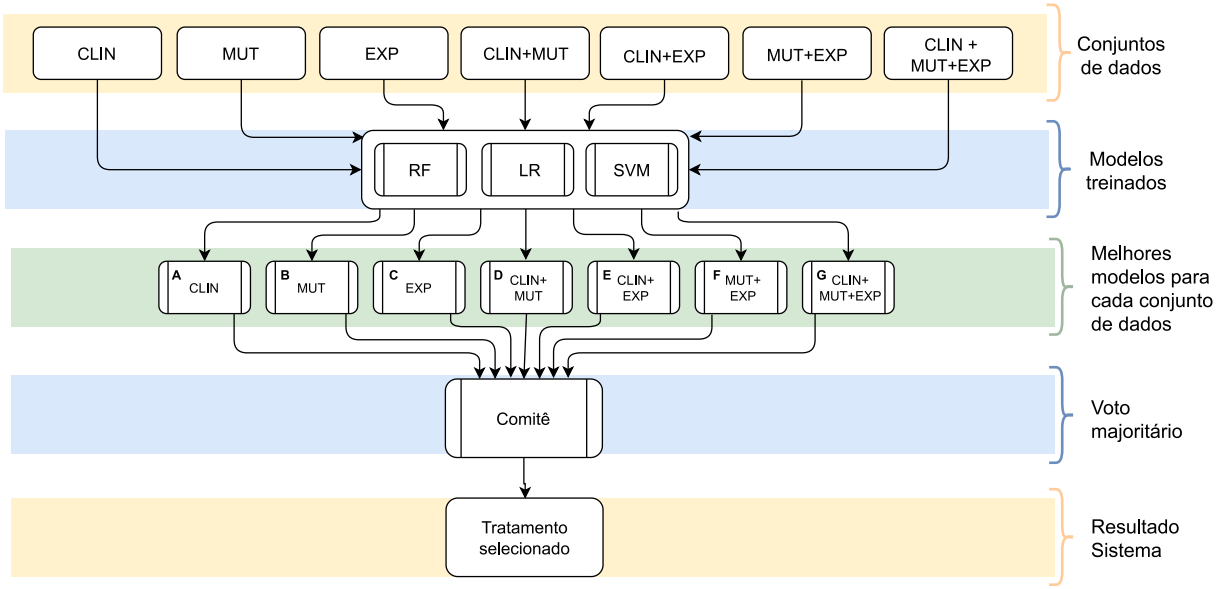

Figura 19 – Processo envolvido para o treinamento do comitê de classificação com todos os conjuntos de dados.

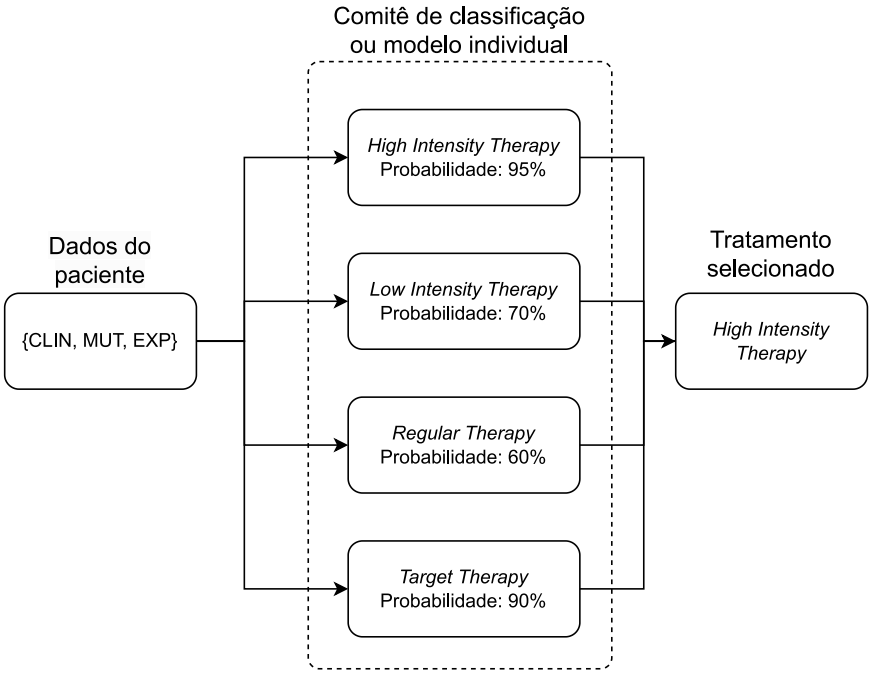

Figura 20 – Processo envolvido para a escolha do conjunto de tratamentos.

5 Experimentos e resultados

Este capítulo detalha os experimentos e resultados obtidos pelos modelos de predição individuais e pelo comitê de classificação. Inicialmente, são descritos os métodos de validação (ver Seção 5.1), a seleção de hiperparâmetros (ver Seção 5.2) e as medidas de desempenho (ver Seção 5.3) utilizadas para avaliar os modelos. A Seção 5.4 apresenta os resultados dos modelos individuais e do comitê de classificação.

5.1 Métodos de validação

Para a avaliação dos modelos, foram utilizados dois métodos tradicionais de divisão de dados, o *hold-out* e a validação cruzada. No *hold-out*, 80% dos dados foram selecionados para compor a partição de treinamento, 10% para validação e os 10% restantes para a partição de teste. Essa divisão, apesar de ter sido aleatória, foi estratificada para preservar a distribuição original das classes em cada partição. A validação cruzada (do inglês, *Cross Validation* – CV) é um método utilizado para avaliar o poder de generalização de modelos de classificação que consiste em dividir os dados em k partes mutuamente excludentes de mesmo tamanho. A cada iteração, uma parte é escolhida para teste, enquanto as outras $(k-1)$ são usadas no treinamento. Este processo é repetido k vezes, sendo que cada partição é usada para teste apenas uma vez (KOHAVI, 1995). O método *Leave-one-out* (LOO) é um caso especial da CV. Ele requer a criação e avaliação de um modelo para cada amostra dos dados. Este método é uma abordagem mais robusta e computacionalmente custosa, recomendado para cenários onde a quantidade de dados é pequena.

5.2 Escolha dos hiperparâmetros

Os hiperparâmetros dos modelos de predição foram selecionados por uma busca em grade (do inglês, *Grid Search*, GS) (MITCHELL, 1997). A GS realiza uma avaliação completa de hiperparâmetros em conjuntos de dados previamente definidos. Ela sofre em espaços de dados de muitas dimensões, uma vez que, o número de avaliações nos conjuntos de busca cresce exponencialmente. No entanto, esta abordagem se mostra simples e eficiente em espaços de busca com poucas dimensões, isto é, é mais prático de se conseguir uma solução ótima na escolha de parâmetros (LAVALLE; BRANICKY; LINDEMANN, 2004).

A Tabela 15 descreve o intervalo de valores utilizados na GS. Os demais parâmetros foram utilizados com os seus valores padrões.

Método	Hiperparâmetros
RF	*class_weight*={balanced, none}, *n_estimators*={10, 50}
LR	*C*={1, 5, 10}, *class_weight*={balanced, none}, *random_state*=1
SVM	*gamma*={scale, auto}, *kernel*={linear}, *C*={1, 5, 10}, *class_weight*={balanced, none}, *random_state*=1

Tabela 15 – Hiperparâmetros avaliados na busca em grade.

5.3 Medidas de desempenho

As medidas utilizadas para avaliar o desempenho dos modelos foram:

Acurácia – taxa de acertos do modelo.

$$\text{Acurácia} = \frac{VP + VN}{VP + FN + VN + FP}$$

Revocação ou *sensibilidade* – taxa de acertos do modelo na classe positiva.

$$\text{Revocação} = \frac{VP}{VP + FN}$$

Precisão – taxa de exemplos positivos classificados corretamente entre todos os preditos como positivos.

$$\text{Precisão} = \frac{VP}{VP + FP}$$

F-medida – média harmônica entre precisão e revocação.

$$\text{F-medida} = 2 * \frac{\text{precisão} * \text{revocação}}{\text{precisão} + \text{revocação}}$$

Nas equações acima, VP é a quantidade de exemplos positivos classificados corretamente, FP é a quantidade de exemplos classificados incorretamente como positivos, VN é a quantidade de exemplos negativos classificados corretamente e FN é a quantidade de exemplos incorretamente classificados como negativos.

Também foi utilizada a área abaixo da curva ROC (do inglês, *area under curve ROC* – AUC). A curva ROC é uma medida de desempenho para classificadores binários que afere o poder de distinção do modelo entre às duas classes (SPACKMAN, 1989). Por sua vez, a AUC é a área entre a curva ROC e o eixo x. A AUC resulta em valores entre 0 e 1, sendo que quanto maior o valor da AUC, melhor é o desempenho do modelo.

A F-medida foi utilizada para escolher o melhor modelo gerado para cada combinação de conjuntos de atributos. Esta escolha foi dada pela natureza dessa métrica, que avalia o poder de predição dos modelos em ambas as classes positiva (sobrevivência) e negativa (morte).

5.4 Resultados

Ao todo, foram avaliados 42 modelos de predição de sobrevivência de pacientes quando expostos a tratamentos de diferentes intensidades. Nestes modelos, diferentes combinações de dados clínicos (CLIN), de mutações gênicas (MUT) e de expressão gênica (EXP) foram utilizadas. Os resultados do desempenho desses modelos são apresentados nas Subseções 5.4.1 a 5.4.7. A Subseção 5.4.8 sumariza os melhores modelos individuais obtidos para cada conjunto de atributos. Finalmente, a Subseção 5.4.9 apresenta o desempenho obtido pelo comitê de classificação.

5.4.1 Modelos treinados com os dados clínicos (CLIN)

A Tabela 16 apresenta o desempenho dos três modelos de classificação treinados apenas com os dados clínicos, usando o método de validação *hold-out*. Valores destacados em negrito indicam o melhor desempenho aferido por cada medida de avaliação.

Método	**F-medida (%)**	**AUC (%)**	**Acurácia (%)**	**Precisão (%)**	**Revocação (%)**
RF	58,03	57,77	57,14	61,22	57,14
LR	**75,23**	**73,88**	**75,00**	**75,66**	**75,00**
SVM	66,95	63,53	67,64	66,75	67,64

Tabela 16 – Resultados obtidos pelos modelos de classificação treinados com os dados clínicos, usando o método de validação *hold-out*.

Conforme pode ser observado, o modelo gerado pelo método de regressão logística foi superior aos demais. A Figura 21 apresenta o nível de importância de cada atributo para a predição deste modelo, de forma decrescente. O eixo x representa a importância dos atributos para a predição, calculada a partir da potência do número de Euler para cada coeficiente da LR.

O atributo *Cytogenetic Info* apresenta distintas categorias citogenéticas, sendo que *Cytogenetic Info_AML-M1* e *M0* estão relacionadas com subtipos de AML, descritas no Capítulo 2. Em seguida, aparece a classificação de risco da ELN, utilizada em muitos casos para determinar guias de tratamento e, consequentemente, influencia no tempo de sobrevida do paciente (SUGAMORI et al., 2022). Por último, aparece, a idade do paciente no momento do diagnóstico, atributo que também é bastante utilizado em decisões terapêuticas (HOPKINS et al., 2017).

Apesar do método RF não ter obtido o melhor desempenho, para fins de análise, também foi gerado o ranqueamento das variáveis mais importantes para a predição (Figura 22). O eixo x da Figura 22 representa a importância das variáveis, dada pela medida de importância da RF, conforme o melhor classificador. As variáveis se mantiveram e preservaram a ordem da LR (Figura 21), reforçando a importância de cada atributo para

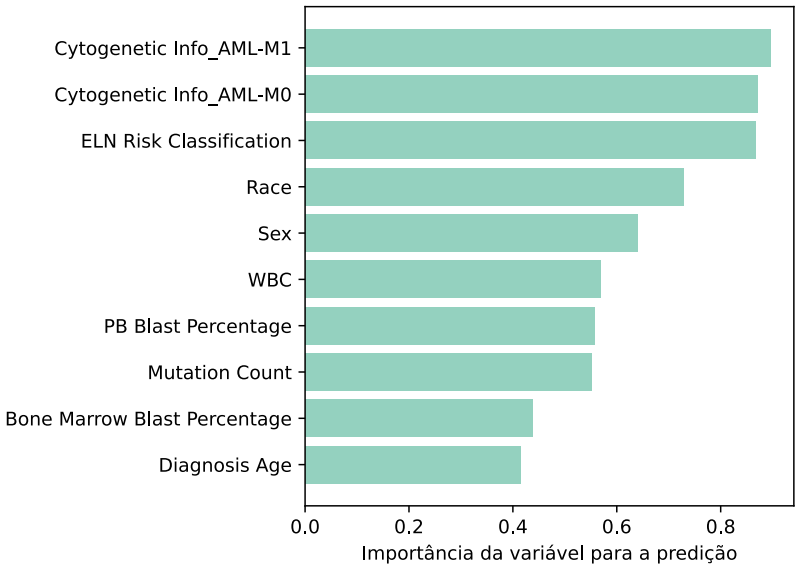

Figura 21 – Importância das variáveis para a predição do modelo gerado pelo treinamento da regressão logística com os dados clínicos.

os modelos preditivos. Contudo, a RF não atribuiu um valor de importância para os 4 últimos atributos apresentados (*PB Blast Percentage*, *Mutation Count*, *Bone Marrow Blast Percentage* e *Diagnosis Age*). Para complementar a análise, as melhores árvores de decisão estão detalhadas no Apêndice A.

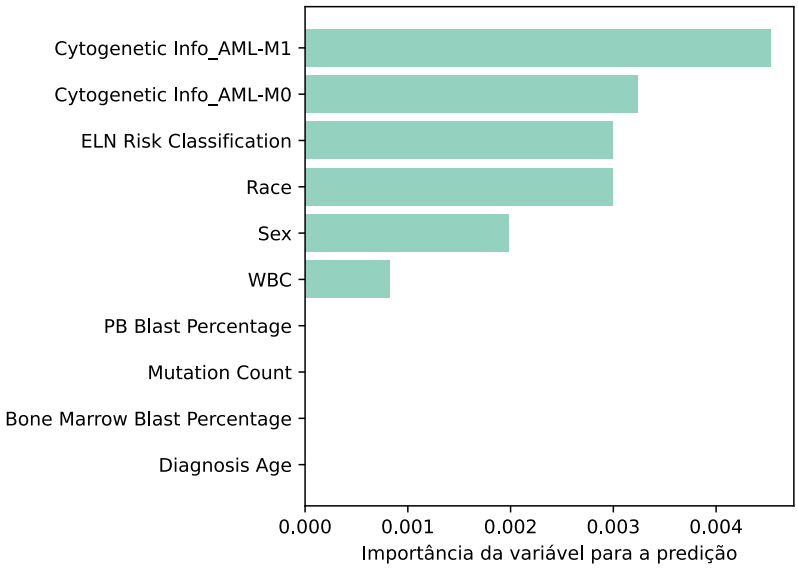

Figura 22 – Importância das variáveis para a predição do modelo gerado pelo treinamento de florestas aleatórias com os dados clínicos.

A Tabela 17 apresenta o desempenho dos três modelos de classificação treinados

com os dados clínicos, usando o método de validação LOO. Valores destacados em negrito indicam o melhor desempenho aferido por cada medida de avaliação. Apesar dos valores terem sido naturalmente menores em relação ao *hold-out* (Tabela 16), o melhor modelo ainda se manteve gerado pela regressão logística.

Método	F-medida (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	63,78	63,23	64,57	63,23
LR	**67,47**	**68,01**	**68,32**	**68,01**
SVM	64,84	65,44	67,53	65,44

Tabela 17 – Resultados obtidos pelos modelos de classificação treinados com os dados clínicos, usando o método de validação LOO.

5.4.2 Modelos treinados com os dados de mutação genética (MUT)

Os dados de mutação são geralmente obtidos logo após os clínicos. Este tipo de dado é comumente utilizado para a classificação de risco proposta pela ELN, melhor descrita no Capítulo 2. A Tabela 18 apresenta o desempenho dos três modelos de classificação treinados apenas com os dados de mutação, usando o método de validação *hold-out*. Valores destacados em negrito indicam o melhor desempenho aferido por cada medida de avaliação.

Método	F-medida (%)	AUC (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**57,92**	**55,00**	**67,85**	**78,57**	**67,85**
LR	**57,92**	**55,00**	**67,85**	**78,57**	**67,85**
SVM	48,99	50,00	63,23	39,98	63,23

Tabela 18 – Resultados obtidos pelos modelos de classificação treinados com os dados de mutação, usando o método de validação *hold-out*.

A Figura 23 exibe o ranqueamento de importância das três mutações gênicas para a predição da LR. As mutações no gene *U2AF1* não são comumente encontradas em pacientes com AML e estão relacionadas a um prognóstico adverso no contexto geral de neoplasias mieloides (PAPAEMMANUIL et al., 2016; XU et al., 2017). Por outro lado, a mutação do gene *TP53* é bem conhecida na literatura e está relacionada a um prognóstico adverso, apesar de ser rara em pacientes com AML (GROB et al., 2022; PAPAEMMANUIL et al., 2016). Já sobre a mutação do gene *PKD1L2*, não há estudos na literatura sobre suas implicações terapêuticas e no prognóstico de risco da doença (PARK et al., 2017). A Figura 24 ilustra o ranqueamento da importância dos atributos para o modelo gerado pela RF. Para complementar a análise, as melhores árvores de decisão estão detalhadas no Apêndice A.

Na validação LOO, todos os modelos gerados obtiveram o mesmo desempenho (Tabela 19), embora um pouco abaixo do obtido na validação *hold-out* para RF e LR.

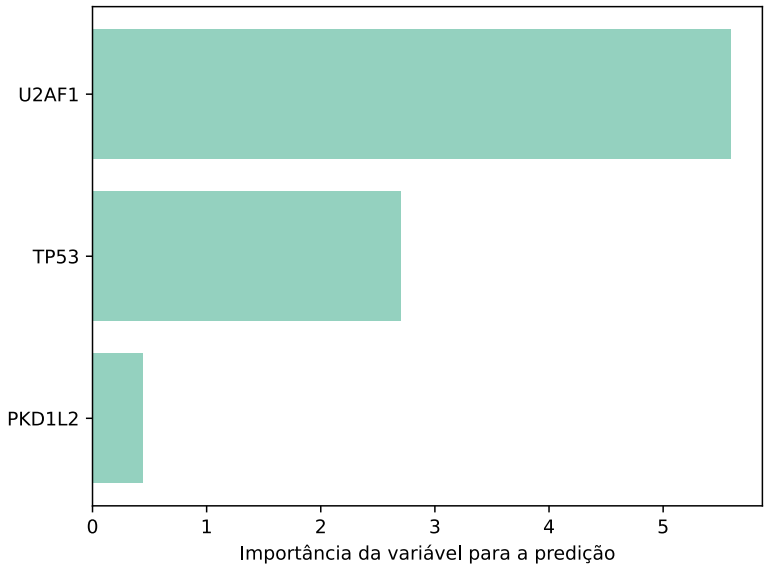

Figura 23 – Importância das variáveis para a predição do modelo gerado pelo treinamento de regressão logística com os dados de mutação genética.

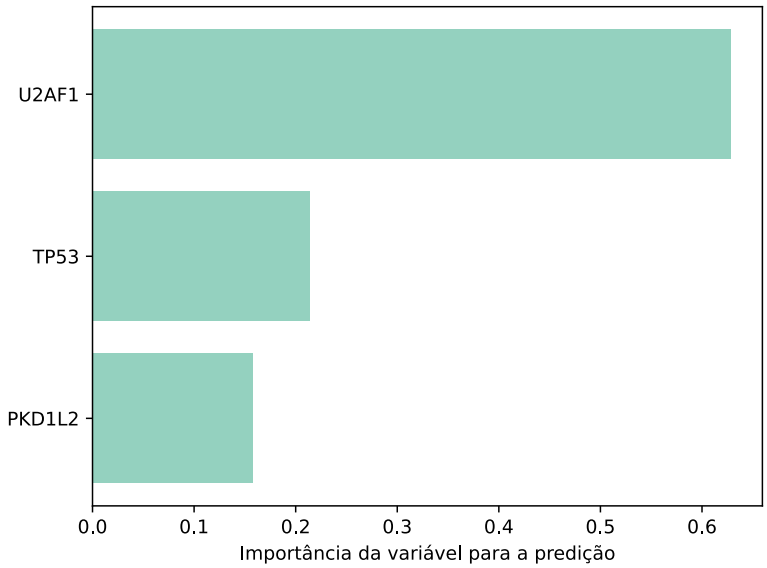

Figura 24 – Importância das variáveis para a predição do modelo gerado pelo treinamento de florestas aleatórias com os dados de mutação genética.

5.4.3 Modelos treinados com os dados de expressão genética (EXP)

Os dados de expressão são normalmente obtidos após uma análise clínica, com o intuito de investigar mais detalhadamente o comportamento da AML em um determinado paciente. A Tabela 20 apresenta o desempenho dos três modelos de classificação treinados apenas com os dados de expressão, usando o método de validação *hold-out*. Valores

Método	F-medida (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**54,67**	**47,05**	**89,94**	**47,05**
LR	**54,67**	**47,05**	**89,94**	**47,05**
SVM	**54,67**	**47,05**	**89,94**	**47,05**

Tabela 19 – Resultados obtidos pelos modelos de classificação treinados com os dados de mutação, usando o método de validação LOO.

destacados em negrito indicam o melhor desempenho aferido por cada medida de avaliação.

Método	F-medida (%)	AUC (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**81,91**	**79,44**	**82,14**	**81,91**	**82,14**
LR	71,42	68,88	71,42	71,42	71,42
SVM	73,44	69,34	75,00	75,00	75,00

Tabela 20 – Resultados obtidos pelos modelos de classificação treinados com os dados de expressão, usando o método de validação *hold-out*.

O melhor resultado foi obtido pela RF. É importante ressaltar que esses modelos foram os que obtiveram os melhores desempenhos dentre os três conjuntos de atributos (clínicos, mutação, expressão) e, portanto, é possível concluir que os dados de expressão genética são bons discriminantes para o curso da doença no paciente.

A Figura 25 apresenta os 10 atributos mais importantes para a predição do modelo gerado pela RF. Os genes *ITFG1* (CK et al., 2017) e *AGRN* (WANG et al., 2021; LI et al., 2018) têm sido associados com o desenvolvimento ou progressão de tumores e dissociados a doenças sanguíneas anteriores. O gene *FNDC3B* (CK et al., 2017; WANG et al., 2016) é o único relacionado com leucemias. Quanto aos demais genes, não há estudos na literatura que tenham associado-os com neoplasias[1] e, portanto, podem representar uma potencial descoberta de padrões, que demanda mais pesquisas e comprovação laboratorial. Para complementar a análise, as melhores árvores de decisão estão detalhadas no Apêndice A.

A Figura 26 ilustra o ranqueamento da importância dos atributos para o modelo gerado pela LR. Embora a LR tenha obtido desempenho inferior a RF, o ranqueamento dos atributos se manteve.

A Tabela 21 apresenta os resultados dos modelos avaliados por meio do LOO. Mesmo com valores naturalmente mais baixos, o comportamento dos modelos gerados se manteve, ou seja, o modelo com o melhor desempenho foi novamente obtido pela RF.

[1] As neoplasias são tumores cancerígenos que se formam a partir de células anormais no corpo.

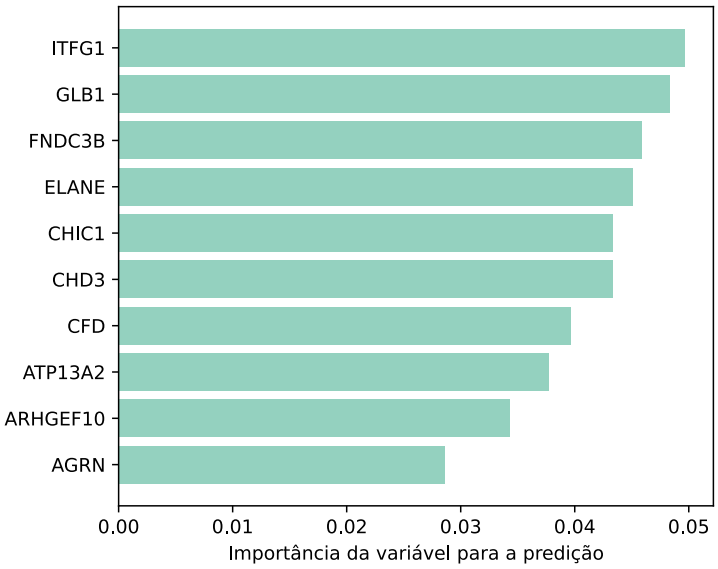

Figura 25 – Importância das variáveis para a predição do modelo gerado pelo treinamento de florestas aleatórias com os dados de expressão genética.

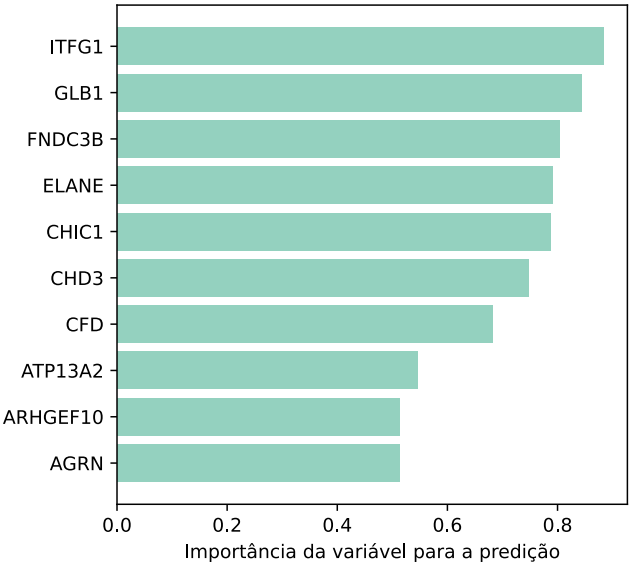

Figura 26 – Importância das variáveis para a predição do modelo gerado pelo treinamento da regressão logística com os dados de expressão genética.

5.4.4 Modelos treinados com os dados clínicos e de mutação genética (CLIN+MUT)

Os atributos clínicos (CLIN) e de mutação genética (MUT) foram combinados (CLIN+MUT) para treinar modelos de predição de sobrevivência com base na exposição de um paciente a um dado grupo de terapias. É importante destacar que os dados clínicos e

Método	F-medida (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**72,34**	**71,32**	**74,43**	**71,32**
LR	69,08	68,48	69,23	69,48
SVM	67,94	68,38	68,13	68,38

Tabela 21 – Resultados obtidos pelos modelos de classificação treinados com os dados de expressão, usando o método de validação LOO.

de mutação são menos custosos para serem obtidos do que os dados de expressão genética. Sendo assim, a combinação CLIN+MUT representa uma opção barata e viável de ser obtida logo após o diagnóstico.

A Tabela 22 apresenta o desempenho dos três modelos de classificação treinados apenas com a combinação dos atributos CLIN+MUT, usando o método de validação *hold-out*. Valores destacados em negrito indicam o melhor desempenho aferido por cada medida de avaliação. É importante observar que LR e SVM obtiveram o mesmo desempenho alcançado quando somente os dados clínicos foram empregados (Tabela 16). Por outro lado, após a adição dos atributos de mutação, a RF apresentou melhoria significativa de desempenho, igualando com a LR.

Método	F-medida (%)	AUC (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**75,23**	**73,88**	**75,00**	**75,66**	**75,00**
LR	**75,23**	**73,88**	**75,00**	**75,66**	**75,00**
SVM	66,95	63,53	67,64	66,75	67,64

Tabela 22 – Resultados obtidos pelos modelos de classificação treinados com a combinação de CLIN+MUT, usando o método de validação *hold-out*.

A Figura 27 apresenta o ranking da importância dos 10 melhores atributos para o modelo gerado pela RF. As variáveis e suas respectivas posições permaneceram idênticas ao modelo gerado pelo conjunto de dados CLIN (Figura 22). Portanto, em geral, os atributos de mutação gênica não contribuíram para a separabilidade dos dados e o ganho de desempenho dos modelos. A Figura 28 apresenta o ranqueamento das 10 variáveis mais importantes para a predição feita pela LR. As variáveis e seus valores são as mesmas descritas nas Figuras 21 e 28. Para complementar a análise, as melhores árvores de decisão estão detalhadas no Apêndice A.

Usando a validação LOO, o melhor desempenho em termos de acurácia foi alcançado pelo modelo gerado pela LR (Tabela 23). Embora os resultados sejam numericamente menores do que os apresentados pelo método de *hold-out*, a qualidade dos modelos foi semelhante, com o modelo gerado pela RF sendo ligeiramente superior ao modelo gerado pela LR quando analisadas as métricas F-medida e precisão.

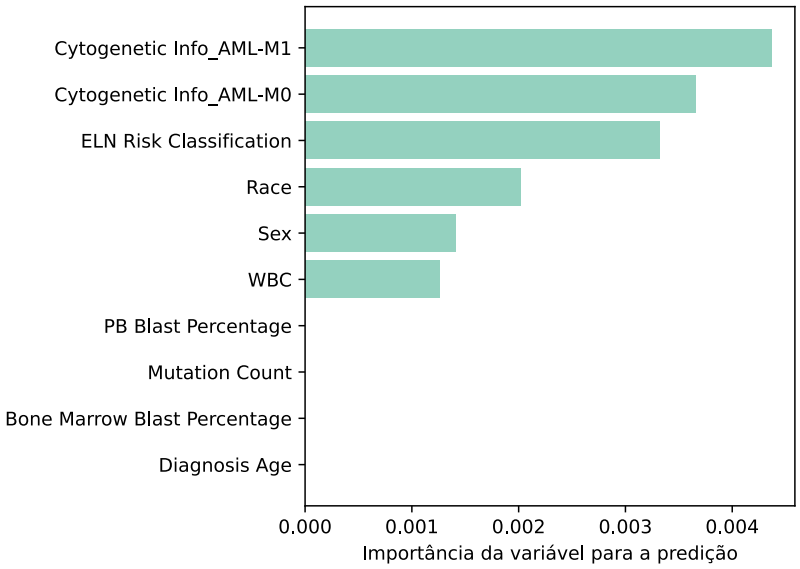

Figura 27 – Importância das variáveis para a predição do modelo gerado pelo treinamento de florestas aleatórias com a combinação dos dados CLIN+MUT.

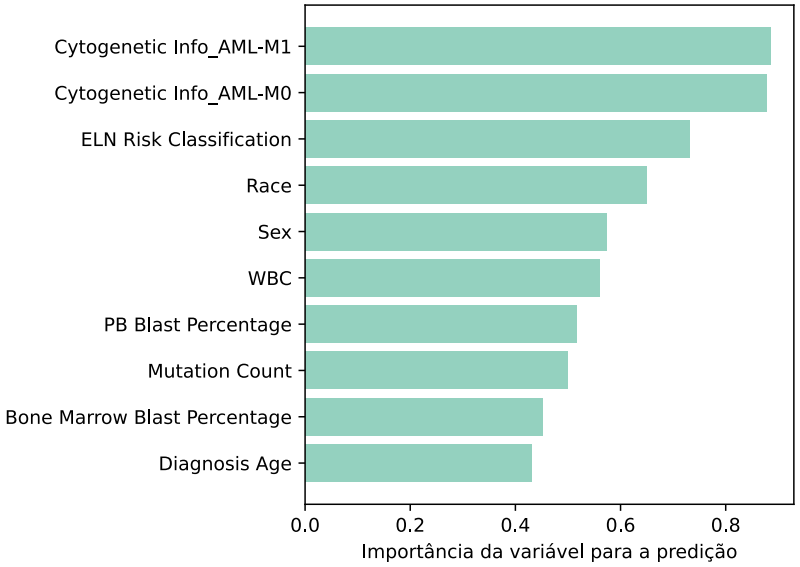

Figura 28 – Importância das variáveis para a predição do modelo gerado pelo treinamento da regressão logística com a combinação dos dados CLIN+MUT.

5.4.5 Modelos treinados com os dados clínicos e de expressão genética (CLIN+EXP)

Esta combinação dos conjuntos de atributos CLIN e EXP foi avaliada para analisar o potencial de predição dos dados de expressão combinados com os dados clínicos. A Tabela 24 mostra os resultados dos modelos utilizando estes conjuntos de dados, com a

Método	F-medida (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**67,61**	66,91	**68,76**	66,91
LR	67,46	**68,01**	68,51	**68,01**
SVM	65,24	65,80	68,47	65,80

Tabela 23 – Resultados obtidos pelos modelos de classificação treinados com a combinação de dados CLIN+MUT, usando o método de validação LOO.

validação feita pelo *hold-out*. O melhor modelo foi o gerado pela RF, com uma melhora de desempenho não tão significativa quando comparada ao modelo gerado somente com o conjunto de dados de expressão genética (Tabela 20).

Método	F-medida (%)	AUC (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**82,31**	**81,66**	**82,14**	**82,69**	**82,14**
LR	79,01	81,11	78,57	82,65	78,57
SVM	68,23	65,03	68,75	68,03	68,75

Tabela 24 – Resultados obtidos pelos modelos de classificação treinados com a combinação de CLIN+EXP, usando o método de validação *hold-out*.

A Figura 29 apresenta o ranqueamento das 10 variáveis mais importantes para o modelo gerado pela RF. Neste caso, os dados de expressão genética perderam destaque em relação aos atributos clínicos. O ranqueamento obtido foi o mesmo do modelo gerado com o conjunto de dados clínicos (Figura 22), porém com valores de importância menores. Além disso, não foi atribuído um valor de importância para os 5 últimos atributos. Para complementar a análise, as melhores árvores de decisão estão detalhadas no Apêndice A.

De maneira análoga, a Figura 30 apresenta a importância das variáveis para a previsão do modelo gerado pela LR. A ordem das variáveis permaneceu a mesma do modelo gerado com base no conjunto de dados clínicos (Figura 21), porém agora com valores maiores de importância. Além disso, o ranqueamento gerado pela RF e LR se manteve, como observado nos casos anteriores.

No que diz respeito aos resultados obtidos na validação LOO (Tabela 25), houve uma diminuição dos valores, o que era esperado. Entretanto, a posição dos modelos foi preservada, sendo o melhor modelo gerado pela RF. O desempenho dos modelos foi satisfatório, mesmo para a validação LOO. Essa confirmação quanto ao desempenho da RF traz perspectivas positivas para sua utilização na tomada de decisões terapêuticas em cenários com dados clínicos e de expressão gênica, além de oferecer uma interpretação das previsões geradas pelo modelo.

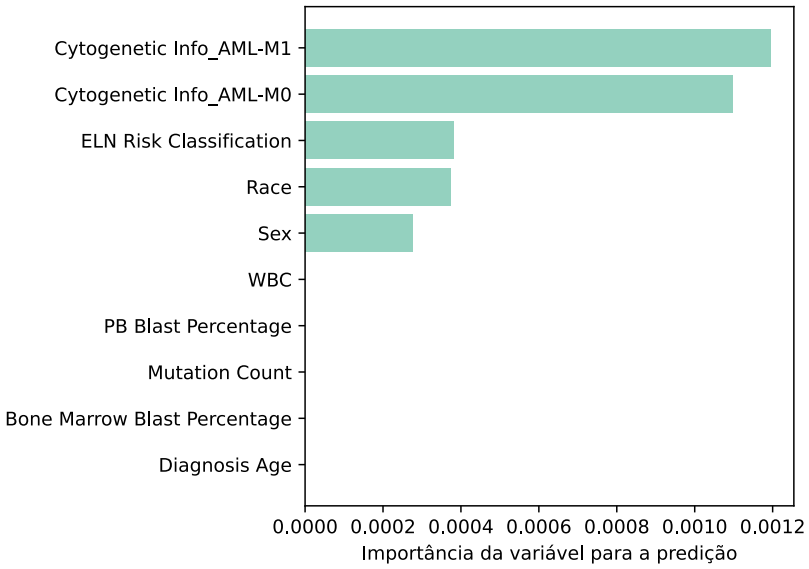

Figura 29 – Importância das variáveis para a predição do modelo gerado pelo treinamento de florestas aleatórias com a combinação dos dados CLIN+EXP.

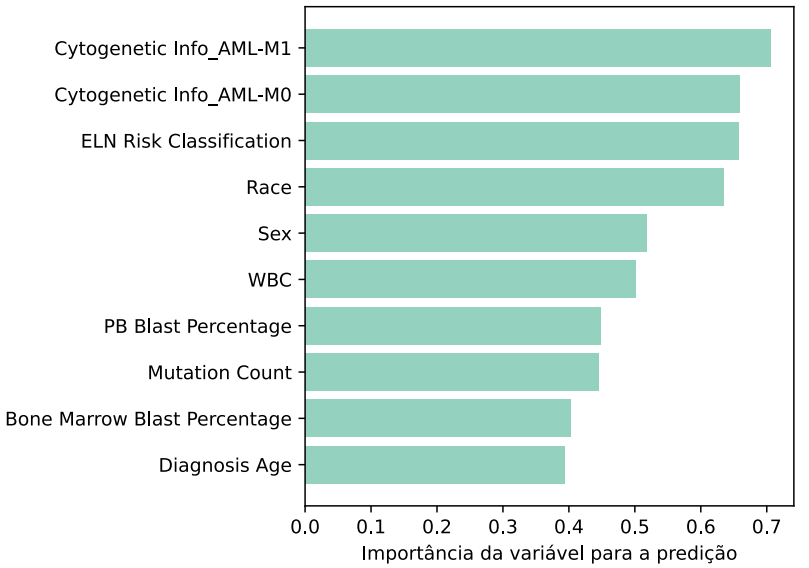

Figura 30 – Importância das variáveis para a predição do modelo gerado pelo treinamento da regressão logística com a combinação dos dados CLIN+EXP.

5.4.6 Modelos treinados com os dados genéticos (MUT+EXP)

A combinação do conjunto de dados MUT e EXP foi gerada com o intuito de avaliar o impacto dos dados genéticos isolados para a predição de sobrevivência. A Tabela 26 apresenta os resultados obtidos pelos modelos na validação *hold-out*. O melhor modelo foi gerado pela RF, apresentando uma melhora significativa em relação aos demais modelos.

Método	F-medida (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**74,98**	**73,89**	**77,49**	**73,89**
LR	69,47	69,85	69,58	69,85
SVM	67,14	67,64	67,56	67,64

Tabela 25 – Resultados obtidos pelos modelos de classificação treinados com a combinação de dados CLIN+EXP, usando o método de validação LOO.

Porém, houve uma queda no desempenho dos modelos quando comparados aos resultados obtidos apenas com o conjunto de dados de expressão (EXP) (Tabela 20), o que pode indicar um impacto negativo dos atributos de mutação gênica na previsão de sobrevida.

Método	F-medida (%)	AUC (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**77,94**	**74,44**	**78,57**	**78,21**	**78,57**
LR	71,42	68,88	71,42	71,42	71,42
SVM	72,00	67,97	73,52	73,14	73,52

Tabela 26 – Resultados obtidos pelos modelos de classificação treinados com a combinação de MUT+EXP, usando o método de validação *hold-out*.

A importância dos atributos para a predição do classificador da RF é apresentada na Figura 31. Entre eles, os atributos de mutação gênica, como *U2AF1*, *TP53* e *PKD1L2*, ocuparam posições relevantes, ficando em 8º, 9º e 10º lugar, respectivamente. Contudo, as mutações tiveram valores de importância consideravelmente menores em comparação às expressões gênicas. Adicionalmente, houve uma alteração na ordem dos atributos comparado ao ranqueamento gerado pelo modelo da RF apenas para o conjunto de dados EXP (Figura 25). Expressões como *ITFG1*, *GLB1* e *FNDC3B* foram removidas e as mutações gênicas foram incluídas na ordenação. Para complementar a análise, as melhores árvores de decisão estão detalhadas no Apêndice A. Na Figura 32 é apresentada as 10 melhores variáveis para a predição do modelo gerado pela LR. A ordenação das variáveis foi a mesma produzida pela RF.

Na validação por LOO (Tabela 27), o melhor modelo foi gerado pela RF, sendo que os modelos seguiram o comportamento apresentado na validação *hold-out* (Tabela 26). O comportamento desses resultados foi similar ao observado somente com o conjunto de dados de expressão (EXP) (Tabela 21) e CLIN+EXP (Tabela 25), o qual pode indicar o bom desempenho deste tipo de classificador em dados genéticos.

5.4.7 Modelos treinados com os dados clínicos e genéticos (CLIN+MUT+EXP)

Os resultados da combinação de todos os conjuntos de dados são apresentados na Tabela 28, gerados pelo método de validação *hold-out*. O melhor desempenho foi alcançado pelo modelo gerado pela LR, com uma pequena diferença para o modelo gerado pela RF.

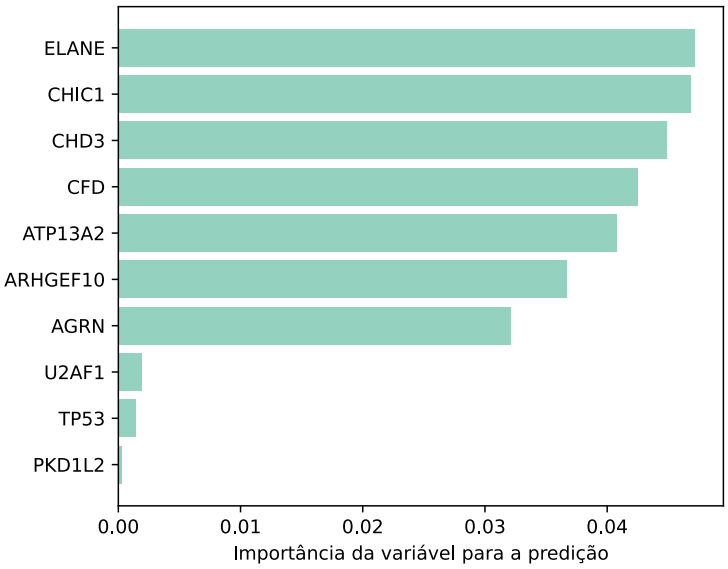

Figura 31 – Importância das variáveis para a predição do modelo gerado pelo treinamento de florestas aleatórias com a combinação dos dados genéticos (MUT+EXP).

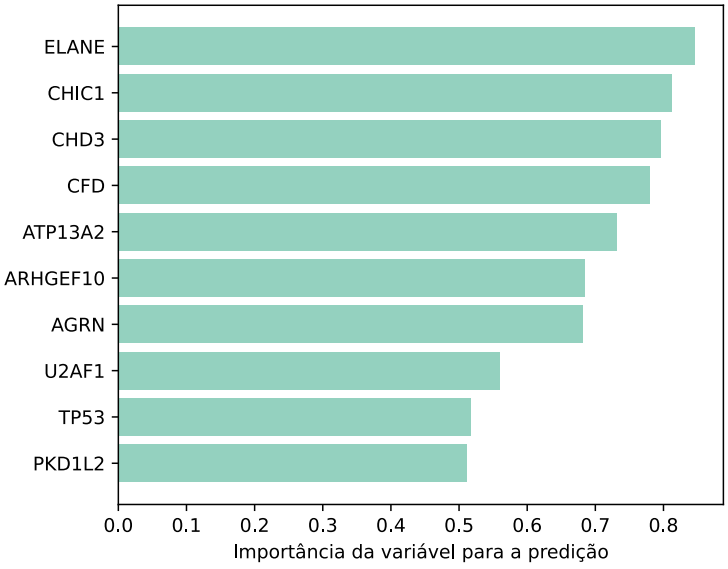

Figura 32 – Importância das variáveis para a predição do modelo gerado pelo treinamento da regressão logística com a combinação dos dados genéticos (MUT+EXP).

A Figura 33 apresenta os 10 melhores atributos para a predição do modelo gerado pela LR. Como descrito com outras combinações que envolvem o conjunto de dados CLIN, as variáveis e suas posições foram preservadas em relação ao modelo gerado pelo conjunto de dados clínicos isolado (Figura 21). É notável que a relevância dos dados genéticos diminui quando combinados com dados clínicos na geração de modelos, como visto nas

Método	F-medida (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	**72,04**	**70,95**	**74,29**	**70,95**
LR	67,51	68,01	68,01	68,01
SVM	67,10	67,64	67,85	67,64

Tabela 27 – Resultados obtidos pelos modelos de classificação treinados com a combinação de dados MUT+EXP, usando o método de validação LOO.

Método	F-medida (%)	AUC (%)	Acurácia (%)	Precisão (%)	Revocação (%)
RF	78,91	78,88	78,57	80,05	78,57
LR	**79,01**	**81,11**	**78,57**	**82,65**	**78,57**
SVM	67,77	65,62	67,64	67,92	67,64

Tabela 28 – Resultados obtidos pelos modelos de classificação treinados com a combinação de CLIN+MUT+EXP, usando o método de validação *hold-out*.

outras combinações de dados clínicos e genéticos. No entanto, os modelos gerados apenas com o conjunto de dados CLIN apresentam desempenho inferior em comparação aos modelos gerados com o conjunto de dados de expressão (EXP).

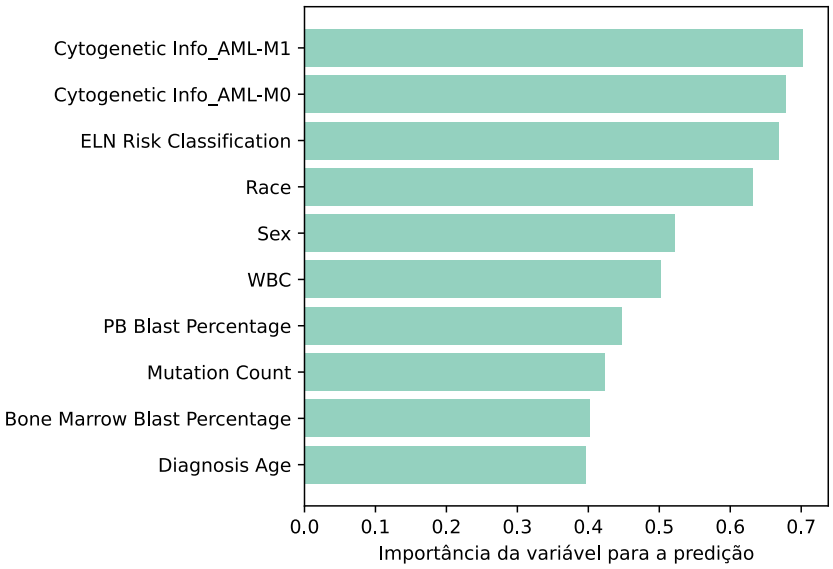

Figura 33 – Importância das variáveis para a predição do modelo gerado pelo treinamento da regressão logística com os dados clínicos e genéticos (CLIN+MUT+EXP).

A preservação das variáveis clínicas no modelo gerado pela RF foi evidenciada na Figura 34. Este ranqueamento é consistente com os modelos gerados por outras combinações de dados que incluem o conjunto CLIN. Contudo, os valores de importância foram consideravelmente menores, o que pode sugerir que a medida de importância da RF é sensível ao número de atributos. É notório descrever que houve apenas atribuição de

valor de importância para o atributo *Cytogenetic Info_AML-M1* referente ao subtipo *M1*[2] de AML. Esse subtipo de AML está geralmente associado a um agravamento da doença (BENNETT et al., 1976). Para complementar a análise, as melhores árvores de decisão estão detalhadas no Apêndice A.

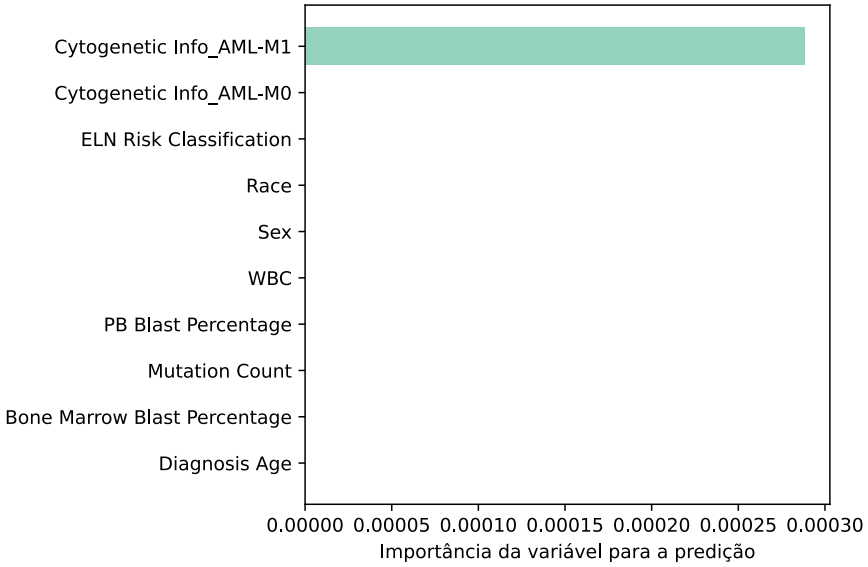

Figura 34 – Importância das variáveis para a predição do modelo gerado pelo treinamento de florestas aleatórias com os dados clínicos e genéticos (CLIN+MUT+EXP).

Conforme apresentado na Tabela 29, o melhor modelo na validação LOO foi gerado pela RF, ao contrário da validação *hold-out* (Tabela 28). Contudo, o desempenho do modelo gerado pela RF foi minimamente inferior. Estes modelos apresentam interpretabilidade, tornando-os atraentes para especialistas. Eles têm potencial para exercer grande impacto na prática clínica, mesmo quando usados de forma individual, graças a esses bons indicadores de qualidade obtidos.

Método	**F-medida (%)**	**Acurácia (%)**	**Precisão (%)**	**Revocação (%)**
RF	**75,28**	**74,26**	**77,63**	**74,26**
LR	70,18	70,58	70,42	70,58
SVM	70,52	70,95	70,96	70,95

Tabela 29 – Resultados obtidos pelos modelos de classificação treinados com a combinação de dados CLIN+MUT+EXP, usando o método de validação LOO.

[2] O subtipo M1 é um tipo específico de AML caracterizado por células imaturas (blastos) no sangue e medula óssea, que não são capazes de se desenvolver e funcionar normalmente. É considerado um dos subtipos mais agressivos de AML e pode requerer tratamento intensivo.

5.4.8 Melhores modelos individuais

A Tabela 30 sumariza os 7 melhores modelos de predição obtidos a partir de cada diferente combinação dos conjuntos de atributos (CLIN, MUT e EXP), usando a validação *hold-out*. A escolha foi feita com base na métrica F-medida, com os melhores desempenhos destacados em negrito. O melhor modelo geral foi obtido pelo treinamento da LR usando a combinação dos dados CLIN+EXP. É importante observar que o modelo gerado somente com o conjunto de dados de expressão genética (Modelo C) obteve um desempenho próximo do melhor modelo (Modelo E). Contudo, este tipo de dado é mais caro de ser produzido em relação aos demais conjuntos de atributos.

Modelos	**Conjuntos de dados**	**Método**	**F-medida (%)**	**AUC (%)**	**Acurácia (%)**	**Precisão (%)**	**Revocação (%)**
A	CLIN	LR	75,23	73,88	75,00	75,66	75,00
B	MUT	LR	57,92	55,00	67,85	78,57	67,85
C	EXP	RF	81,91	79,44	**82,14**	81,91	**82,14**
D	CLIN + MUT	RF	75,23	73,88	75,00	75,66	75,00
E	CLIN + EXP	LR	**82,31**	**81,66**	**82,14**	**82,69**	**82,14**
F	MUT + EXP	RF	77,94	74,44	78,57	78,21	78,57
G	CLIN + MUT + EXP	LR	79,01	81,11	78,57	82,65	78,57

Tabela 30 – Resultados obtidos pelos melhores modelos de classificação treinados com as combinações dos conjuntos de atributos clínicos, de mutação genética e expressão genética, usando o método de validação *hold-out*.

Os classificadores gerados com o conjunto de dados de mutação genética isolados (Modelo B) não apresentaram bons resultados. Este desempenho pode estar atrelado a pequena quantidade de atributos (3 mutações gênicas). O melhor modelo produzido para a combinação CLIN+MUT não apresentou resultados diferentes do gerado somente com os dados clínicos. Portanto, é conclusivo que as mutações genéticas não exercem impacto na predição. Além disso, os resultados obtidos pelo modelo gerado a partir da combinação CLIN+MUT+EXP foram inferiores aos produzidos para a combinação CLIN+EXP.

O desempenho obtido pelos 7 melhores modelos avaliados pela técnica de validação LOO são apresentados na Tabela 31. Os melhores desempenhos são destacados em negrito. Como esperado, houve uma redução nos valores das métricas quando comparadas ao *hold-out*, devido à robustez do método LOO. O melhor foi obtido por RF treinada com a combinação de todos os conjuntos de dados (Modelo G). Mais uma vez, modelos treinados com os dados de mutação genética não obtiveram bons resultados.

Em geral, os melhores modelos de classificação individuais apresentaram resultados satisfatórios na previsão da sobrevivência de pacientes com AML. Seu uso isolado pode auxiliar a tomada de decisão de especialistas, uma vez que os modelos são interpretáveis.

Modelos	Conjuntos de dados	Método	F-medida (%)	Acurácia (%)	Precisão (%)	Revocação (%)
A	CLIN	LR	67,47	68,01	68,32	68,01
B	MUT	LR	54,67	47,05	**89,94**	47,05
C	EXP	RF	72,34	71,32	74,43	71,32
D	CLIN + MUT	RF	67,61	66,91	68,76	66,91
E	CLIN + EXP	RF	74,98	73,89	77,49	73,89
F	MUT + EXP	RF	72,04	70,95	74,29	70,95
G	CLIN + MUT + EXP	RF	**75,28**	**74,26**	77,63	**74,26**

Tabela 31 – Resultados obtidos pelos melhores modelos de classificação treinados com as combinações dos conjuntos de atributos clínicos, de mutação genética e expressão genética, usando o método de validação LOO.

5.4.9 Comitê de classificação

Conforme apresentado na Figura 19, este trabalho propõe um sistema de apoio à decisão formado por um comitê de classificação composto pelos melhores modelos individuais treinados com cada uma das 7 combinações de conjuntos de atributos (Modelos A a G). Como os dados de expressão gênica (EXP) são custosos de serem obtidos, também foi avaliado um segundo comitê composto pelos melhores modelos individuais treinados com cada uma das 3 combinações dos conjuntos de atributos CLIN, MUT e CLIN+MUT (Modelos A, B e D).

A Tabela 32 apresenta os resultados dos dois comitês de classificação validados pelo método *hold-out*. Ambos os contextos de combinações de atributos apresentaram resultados bastante promissores, sugerindo que o comitê pode ser usado como uma ferramenta para auxiliar a escolha de tratamentos para pacientes com AML. Mesmo sem considerar os dados de expressão genética, o comitê obteve bons resultados, sendo importante para realidades clínicas que não têm acesso a esses dados. Contudo, na presença desses dados, o sistema de apoio à decisão foi ainda mais preciso e assertivo, superando 92% de F-medida e acurácia. Observando a tabela de confusão, apenas uma amostra de cada classe foi incorretamente rotulada pelo sistema, alcançando medidas de desempenho impressionantes e animadoras.

Conjuntos de dados	F-medida (%)	AUC (%)	Acurácia (%)	Precisão (%)	Revocação (%)
CLIN e MUT	79,20	77,50	78,57	80,63	78,57
CLIN, MUT e EXP	**92,86**	**92,22**	**92,86**	**92,86**	**92,86**

Tabela 32 – Resultados obtidos pelos comitês de classificação, usando o método de validação *hold-out*.

Para assegurar a robustez dos resultados, os comitês também foram validados usando a técnica LOO (Tabela 33). Novamente, os desempenhos foram consistentes em

ambos os contextos de combinações de atributos e contribuem para reforçar a viabilidade e segurança do emprego dos comitês como um sistema de apoio à decisão para a escolha de tratamentos em pacientes com AML.

Conjuntos de dados	**F-medida (%)**	**Acurácia (%)**	**Precisão (%)**	**Revocação (%)**
CLIN e MUT	77,44	76,83	78,80	76,83
CLIN, MUT e EXP	**83,95**	**83,45**	**83,45**	**83,45**

Tabela 33 – Resultados obtidos pelos comitês de classificação, usando o método de validação LOO.

6 Conclusões

Este trabalho propôs, implementou e avaliou um sistema de suporte à decisão para a escolha de guias de tratamento baseado na predição de sobrevivência de pacientes com AML. O sistema proposto, composto por um comitê de classificadores treinados com dados clínicos e genéticos (mutação e expressão), é capaz de auxiliar os especialistas de modo que eles consigam selecionar os melhores conjuntos de tratamentos (*High intensity therapy*, *Low intensity therapy*, *Regular therapy* e *Target therapy*) baseados na predição de sobrevivência do paciente.

Inicialmente, foram realizados estudos na literatura com o intuito de compreender o comportamento da AML e suas implicações. Foi revisado o modo de diagnóstico da doença, seu prognóstico de risco e sua manifestação nos pacientes. Devido a diferentes apresentações da doença, os especialistas na área apresentam dificuldade em ditar um guia de tratamento e costumeiramente acabam se baseando na classificação de risco da ELN para guiar suas decisões.

Para investigar e classificar os guias de tratamento para AML e seus critérios de utilização, foi realizado um mapeamento sistemático. Dentre os estudos selecionados para síntese, nenhum deles apresentou o uso de ML como ferramenta para os guias de tratamento. Os trabalhos foram categorizados em função da intensidade dos tratamentos, realizada por especialistas do domínio. Como nenhum estudo utilizou citarabina e antraciclina de modo individualizado, é possível inferir que o uso exclusivo dessa abordagem não seja tão eficiente. A metade (6) dos trabalhos combinou duas ou mais intensidades terapêuticas.

As recomendações terapêuticas internacionais de AML precisam adicionar novas técnicas a ELN e a clássica combinação de citarabina e antraciclina. A doença é heterogênea e, portanto, é difícil decidir um curso terapêutico genérico que atenda adequadamente todos os pacientes. Consequentemente, as estratégias terapêuticas têm se tornado cada vez mais personalizadas e isoladas para realidades clínicas individualizadas.

Existem diversos fatores além da ELN que podem ser considerados para personalizar guias de tratamento, como a idade, recaídas e medicamentos inibidores de ações proteicas. A combinação destes e outros critérios é dificultada em uma análise manual e, portanto, é cada vez mais necessário o emprego de ferramentas computacionais que possam realizar análises automatizadas para auxiliar os especialistas na escolha dos melhores protocolos de tratamento.

Com base nas lacunas existentes na literatura e na hipótese de que técnicas de ML podem servir de ferramenta para a escolha de tratamentos, foi proposto um sistema inédito de apoio à decisão. Este sistema foi gerado e avaliado por meio de duas bases de

dados de domínio público, provenientes de estudos de acompanhamento clínico e genético conduzidos pela OSHU e TCGA. Logo após a obtenção dos dados, foram realizados os processos de integração, pré-processamento e seleção, resultando em uma base de dados com 272 amostras de pacientes. Esta base é constituída por 11 atributos clínicos (CLIN), 19 de expressões gênicas (EXP) e 3 de mutações gênicas (MUT).

Ao todo, foram treinados 42 modelos de predição de sobrevivência, sendo a metade (21) treinada e validada usando a técninca *hold-out* e a outra metade por LOO. Em resumo, foram usados três técnicas consolidadas de ML (RF, LR e SVM) para treinar modelos usando todas as possíveis combinações de conjuntos de atributos. Os modelos foram avaliados pelas métricas acurácia, revocação, precisão, F-medida e AUC e, posteriormente, foram selecionados os 7 melhores modelos para cada possível combinação de atributos.

O melhor modelo avaliado por *hold-out* foi gerado por LR treinada com a combinação dos conjunto de atributos CLIN e EXP. Já o melhor modelo obtido na validação LOO foi uma RF treinada com a combinação de todos os conjuntos de atributos (CLIN, MUT e EXP). Os dados de expressão apresentaram um alto poder preditivo para a sobrevivência dos pacientes com AML, contudo, há um alto custo envolvido na obtenção deste tipo de dado. Os modelos gerados pelos dados de mutação não obtiveram bom desempenho, que podem ser explicados pelo fato de ter sido utilizado um pequeno número de genes. Com a adição dos dados clínicos, não foi observado um aumento considerável no poder preditivo dos modelos quando comparado aos que foram produzidos apenas com o conjunto de dados clínicos de modo isolado.

O sistema de suporte à decisão proposto é formado pelos comitês dos melhores modelos com e sem a inclusão de dados de expressão genética. A escolha da classe (sobrevivência/morte) é determinada pelo voto majoritário dos modelos individuais. Isso foi feito porque os dados de expressão podem ser difíceis e dispendiosos de serem obtidos em uma primeira visita clínica.

Os resultados obtidos com o sistema de suporte à decisão proposto são promissores, especialmente evidenciados pela robustez apresentada na avaliação LOO. Os resultados obtidos asseguram o uso deste sistema como uma boa alternativa para auxiliar na tomada de decisão dos especialistas com relação à escolha do melhor guia de tratamento para seus pacientes.

Em resumo, este trabalho oferece as seguintes contribuições diretas e indiretas:

1. Mapeamento sistemático da literatura sobre guias de tratamento, no qual foram extraídos os seguintes resultados:

    - Categorização dos guias de tratamento existentes de acordo com sua intensidade e critérios de aceitação;

- Apesar de diversos estudos clínicos e genéticos, pouco se traduziu para a prática clínica;
- A AML tem diferentes apresentações para os seus pacientes e os especialistas usam diferentes critérios para os seus guias de tratamento. A diversificação destes critérios evidencia a dificuldade em definir abordagens terapêuticas que abranjam todos os grupos de pacientes com AML; e
- Há a necessidade de ferramentas automatizadas que contribuam na escolha dos guias de tratamento de modo a otimizar o tempo de sobrevida e trazer uma melhor qualidade de vida para os pacientes.

2. Seleção e análise da expressão e mutação de genes relacionados ao tempo de sobrevida dos pacientes com AML;

3. Modelos de predição de sobrevivência individuais que, por si só, podem ser utilizados como ferramentas simplificadas de apoio às decisões dos especialistas; e

4. Sistemas de suporte à decisão terapêutica para pacientes com AML baseado na predição de sobrevivência. Este sistema se divide em dois comitês de classificação: (*i*) a partir de dados clínicos e de mutações genéticas, geralmente obtidos em uma primeira visita clínica; e (*ii*) com a adição de dados de expressões gênicas que são mais custosos de serem obtidos.

Apesar dos resultados promissores apresentados neste trabalho, é importante ressaltar algumas limitações:

- A maioria (ao menos 75%) das amostras sanguíneas utilizadas são provenientes de pessoas da raça branca. Isso dificulta a generalização dos resultados para ambientes clínicos em diferentes regiões e países;
- O processo de seleção de atributos genéticos (mutação e expressão) não garante uma combinação ótima de genes relacionados a predição de sobrevivência dos pacientes com AML; e
- A baixa quantidade de dados utilizada para a geração dos modelos de ML.

Sugestões para trabalhos futuros incluem:

1. A criação de uma plataforma pública que facilite e motive o uso do sistema de suporte à decisão proposto;

2. Coleta de mais dados para o treinamento dos modelos de modo que fiquem sempre atualizados em relação à evolução da apresentação da AML no tempo;

3. Análise de novas técnicas de ML interpretáveis para a geração de novos modelos;

4. Refazer os experimentos com a separação das raças dos pacientes e posteriormente compará-los para um melhor entendimento do papel da raça no curso da AML;

5. Empregar *Multi-view learning* (MVL)[1] para a análise dos grupos de guias de tratamento;

6. Análise clínica e laboratorial das expressões e mutações genéticas selecionadas neste trabalho para um melhor entendimento do real impacto no curso da AML;

7. Comparação do método de seleção de atributos da RF com o utilizado neste trabalho para selecionar atributos de expressão genética; e

8. Análise de outras técnicas de seleção de atributos de expressão genética, como LASSO[2] (*Least Absolute Shrinkage and Selection Operator*). Na seleção de expressão genética, o LASSO é aplicado para identificar quais genes estão associados a uma determinada doença ou fenótipo. O objetivo é encontrar um subconjunto de genes que tenham um efeito significativo na doença, enquanto descarta aqueles que não contribuem para a previsão.

[1] O MVL é um campo de estudo em ML que lida com conjuntos de dados que contêm múltiplas visões ou perspectivas do mesmo objeto ou fenômeno. O objetivo do MVL é utilizar as informações fornecidas por todas as vistas para melhorar a precisão e a generalização dos modelos de ML.

[2] O LASSO é um método de regressão linear que pode ser usado para seleção de características em problemas de grande dimensionalidade

Referências

AMARATUNGA, D.; CABRERA, J. Analysis of Data From Viral DNA Microchips. *Journal of the American Statistical Association*, v. 96, n. 456, p. 1161–1170, 2001. Citado na página 52.

ANGENENDT, L. et al. Chromosomal Abnormalities and Prognosis in NPM1-Mutated Acute Myeloid Leukemia: A Pooled Analysis of Individual Patient Data From Nine International Cohorts. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, v. 37, n. 29, p. 2632–2642, 2019. Citado na página 28.

ARBER, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*, v. 127, n. 20, p. 2391–2405, 2016. Citado na página 61.

ARBER, D. A. et al. International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. *Blood*, v. 140, n. 11, p. 1200–1228, 2022. Citado na página 23.

ARUNASRI, K. et al. Effect of Simulated Microgravity on E. coli K12 MG1655 Growth and Gene Expression. *PLOS ONE*, v. 8, n. 3, p. e57860–e57860, 2013. Citado na página 68.

BENNETT, J. M. et al. Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. *British Journal of Haematology*, v. 33, n. 4, p. 451–458, 1976. Citado 2 vezes nas páginas 27 e 94.

BODDU, P. et al. Initial Report of a Phase I Study of LY2510924, Idarubicin, and Cytarabine in Relapsed/Refractory Acute Myeloid Leukemia. *Frontiers in Oncology*, v. 8, p. 369–369, 2018. Citado na página 37.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92*. New York, NY, USA: Association for Computing Machinery, 1992. p. 144–152. Citado na página 73.

BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Citado na página 73.

BRYANT, A. L. et al. Understanding Barriers to Oral Therapy Adherence in Adults With Acute Myeloid Leukemia. *Journal of the Advanced Practitioner in Oncology*, v. 11, n. 4, p. 342–349, 2020. Citado 2 vezes nas páginas 23 e 37.

Cancer Genome Atlas Research Network et al. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *The New England Journal of Medicine*, v. 368, n. 22, p. 2059–2074, 2013. Citado 4 vezes nas páginas 23, 24, 28 e 52.

CHEN, J. et al. Comparison of Autologous Stem Cell Transplantation versus Haploidentical Donor Stem Cell Transplantation for Favorable- and Intermediate-Risk Acute Myeloid Leukemia Patients in First Complete Remission. *Biology of Blood and Marrow Transplantation*, v. 24, n. 4, p. 779–788, 2018. Citado na página 37.

CK, C. et al. FNDC3B is another novel partner fused to RARA in the t(3;17)(q26;q21) variant of acute promyelocytic leukemia. *Blood*, v. 129, n. 19, p. 2705–2709, 2017. Citado 3 vezes nas páginas 70, 85 e 112.

CRAMER, J. The Origins of Logistic Regression. *SSRN Electronic Journal*, 2003. Citado na página 73.

CRISTOFERI, L. et al. Prognostic models in primary biliary cholangitis. *Journal of Autoimmunity*, v. 95, p. 171–178, 2018. Citado na página 49.

DITTAKAVI, S. et al. Validated LC-MS/MS Method for Simultaneous Quantitation of Enasidenib and its Active Metabolite, AGI-16903 in Small Volume Mice Plasma: Application to a Pharmacokinetic Study. *Drug Research*, v. 70, n. 1, p. 41–48, 2020. Citado na página 37.

DÖHNER, H. et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, v. 129, n. 4, p. 424–447, 2017. Citado 2 vezes nas páginas 23 e 29.

DÖHNER, H. et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood*, v. 115, n. 3, p. 453–474, 2010. Citado 5 vezes nas páginas 23, 24, 27, 28 e 29.

DÖHNER, H. et al. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood*, v. 140, n. 12, p. 1345–1377, 2022. Citado 2 vezes nas páginas 23 e 28.

ECKARDT, J.-N. et al. Application of machine learning in the management of acute myeloid leukemia: current practice and future prospects. *Blood Advances*, v. 4, n. 23, p. 6077–6085, 2020. Citado na página 46.

ELSAYED, A. H. et al. A six-gene leukemic stem cell score identifies high risk pediatric acute myeloid leukemia. *Leukemia*, v. 34, n. 3, p. 735–745, 2020. Citado 2 vezes nas páginas 28 e 29.

EMADI, A. et al. Venetoclax and pegcrisantaspase for complex karyotype acute myeloid leukemia. *Leukemia*, v. 35, n. 7, p. 1907–1924, 2021. Citado na página 36.

FAYYAD, U.; UTHURUSAMY, R. Data mining and knowledge discovery in databases. *Communications of the ACM*, v. 39, n. 11, p. 24–26, 1996. Citado na página 51.

FISHER, R. A. 009: The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, v. 52, p. 399–433, 1918. Citado na página 68.

FITTER, S. et al. *CKLF* and *IL1B* transcript levels at diagnosis are predictive of r elapse in children with pre-B-cell acute lymphoblastic leukaemia. *British Journal of Haematology*, v. 193, n. 1, p. 171–175, 2021. Citado na página 48.

FLEMING, S. et al. Use of machine learning in 2074 cases of acute myeloid leukemia for genetic risk profiling. *Blood*, v. 134, n. 1, p. 1392–1392, 2019. Citado na página 47.

GAL, O. et al. Predicting complete remission of acute myeloid leukemia: Machine learning applied to gene expression. *Cancer Informatics*, v. 18, p. 1–5, 2019. Citado na página 47.

GROB, T. et al. Molecular characterization of mutant TP53 acute myeloid leukemia and high-risk myelodysplastic syndrome. *Blood*, v. 139, n. 15, p. 2347–2354, 2022. Citado 2 vezes nas páginas 68 e 83.

HOPKINS, B. et al. Improving the Transition to Palliative Care for Patients With Acute Leukemia: A Coordinated Care Approach. *Cancer Nursing*, v. 40, n. 3, p. E17–E23, 2017. Citado 3 vezes nas páginas 37, 42 e 81.

HU, R.; YU, Y.; WANG, H. The LMCD1-AS1/miR-526b-3p/OSBPL5 axis promotes cell proliferation, migration and invasion in non-small cell lung cancer. *BMC Pulmonary Medicine*, v. 22, n. 1, p. 30, 2022. Citado na página 70.

HUANG, S.-W. et al. Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning. *PLOS Neglected Tropical Diseases*, v. 14, n. 12, p. e0008960–e0008960, 2020. Citado na página 49.

JAN, B. et al. Deep learning in big data Analytics: A comparative study. *Computers & Electrical Engineering*, v. 75, p. 275–287, 2019. Citado na página 73.

JING, B. et al. Deep learning for risk prediction in patients with nasopharyngeal carcinoma using multi-parametric mris. *Computer Methods and Programs in Biomedicine*, v. 197, p. 105684–105684, 2020. Citado 2 vezes nas páginas 24 e 48.

KADIA, T. M. et al. Progress in Acute Myeloid Leukemia. *Clinical Lymphoma Myeloma and Leukemia*, v. 15, n. 3, p. 139–151, 2015. Citado 2 vezes nas páginas 24 e 67.

KASTENHUBER, E. R.; LOWE, S. W. Putting p53 in Context. *Cell*, v. 170, n. 6, p. 1062–1078, 2017. Citado na página 68.

KIM, D. S. et al. Selection of elderly acute myeloid leukemia patients for intensive chemotherapy: effectiveness of intensive chemotherapy and subgroup analysis. *Acta Haematologica*, v. 133, n. 3, p. 300–309, 2015. Citado 2 vezes nas páginas 23 e 37.

KIM, H.-J. et al. Haplotype mismatched transplantation using high doses of peripheral blood CD34+ cells together with stratified conditioning regimens for high-risk adult acute myeloid leukemia patients: a pilot study in a single Korean institution. *Bone Marrow Transplantation*, v. 35, n. 10, p. 959–964, 2005. Citado 2 vezes nas páginas 37 e 43.

KIM, L. et al. Acute cardiotoxicity after initiation of the novel tyrosine kinase inhibitor gilteritinib for acute myeloid leukemia. *Cardio-Oncology*, v. 7, n. 1, p. 36–36, 2021. Citado 2 vezes nas páginas 29 e 36.

KITCHENHAM, B. et al. Guidelines for performing Systematic Literature Reviews in Software Engineering. *Keele Univesity*, 2007. Citado na página 30.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *International joint Conference on artificial intelligence*. [S.l.: s.n.], 1995. v. 14, p. 1137–1145. Citado na página 79.

KOUROU, K. et al. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, v. 13, p. 8–17, 2015. Citado na página 24.

LAVALLE, S. M.; BRANICKY, M. S.; LINDEMANN, S. R. On the Relationship Between Classical Grid Search and Probabilistic Roadmaps. *The International Journal of Robotics Research*, v. 23, p. 673–692, 2004. Citado na página 79.

LI, G. et al. Genomic analysis of biomarkers related to the prognosis of acute myeloid leukemia. *Oncology Letters*, v. 20, n. 2, p. 1824–1834, 2020. Citado na página 69.

LI, X. et al. Oncogenic Properties of NEAT1 in Prostate Cancer Cells Depend on the CDC5L-AGRN Transcriptional Regulation Circuit. *Cancer Research*, v. 78, n. 15, p. 4138–4149, 2018. Citado 2 vezes nas páginas 70 e 85.

LI, Y. et al. A novel gene selection method based on ANOVA and SVM for diagnosis of lung cancer. *Computers in Biology and Medicine*, v. 40, p. 446–456, 2010. Citado na página 68.

LI, Y. et al. A random forest model for predicting social functional improvement in chinese patients with schizophrenia after 3 months of atypical antipsychotic monopharmacy: A cohort study. *Neuropsychiatric Disease and Treatment*, v. 17, p. 847–857, 2021. Citado 2 vezes nas páginas 24 e 47.

LIN, M. et al. Application of deep learning in predicting the prognosis of acute myeloid leukemia using cytogenetics, age, and mutations. *Clinical Oncology and Research*, 2020. Citado na página 46.

LIN, W. et al. Identification of MICALL2 as a Novel Prognostic Biomarker Correlating with Inflammation and T Cell Exhaustion of Kidney Renal Clear Cell Carcinoma. *Journal of Cancer*, v. 13, n. 4, p. 1214–1228, 2022. Citado na página 70.

MIN, P. et al. MICAL-L2 Is Essential for c-Myc Deubiquitination and Stability in Non-small Cell Lung Cancer Cells. *Frontiers in Cell and Developmental Biology*, v. 8, p. 575903, 2020. Citado na página 70.

MIN, P. et al. MICAL-L2 potentiates Cdc42-dependent EGFR stability and promotes gastric cancer cell migration. *Journal of Cellular and Molecular Medicine*, v. 23, n. 6, p. 4475–4488, 2019. Citado na página 70.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997. Citado na página 79.

MONTI, P. et al. Heterogeneity of TP53 Mutations and P53 Protein Residual Function in Cancer: Does It Matter? *Frontiers in Oncology*, p. 593383–593383, 2020. Citado na página 68.

NAWATA, R. et al. MEK kinase 1 mediates the antiapoptotic effect of the Bcr-Abl oncogene through NF-kappaB activation. *Oncogene*, v. 22, n. 49, p. 7774–7780, 2003. Citado na página 70.

NG, S. W. K. et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*, v. 540, n. 7633, p. 433–437, 2016. Citado na página 29.

OLIVEIRA, M. I. S.; LIMA, G. D. F. B.; LÓSCIO, B. F. Investigations into Data Ecosystems: a systematic mapping study. *Knowledge and Information Systems*, v. 61, p. 589–630, 2019. Citado 2 vezes nas páginas 13 e 30.

OLSSON, I. et al. Serum and plasma myeloperoxidase, elastase and lactoferrin content in acute myeloid leukaemia. *Scandinavian Journal of Haematology*, v. 22, n. 5, p. 397–406, May 1979. Citado na página 70.

ORGUEIRA, A. M. et al. Personalized Survival Prediction of Patients With Acute Myeloblastic Leukemia Using Gene Expression Profiling. *Frontiers in Oncology*, v. 11, p. 657191–657191, 2021. Citado na página 47.

PAI, M. et al. Systematic reviews and meta-analyses: an illustrated, step-by-step guide. *The National medical journal of India*, v. 17, n. 2, p. 86–95, 2004. Citado na página 31.

PAPAEMMANUIL, E. et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The New England Journal of Medicine*, v. 374, n. 23, p. 2209–2221, 2016. Citado 2 vezes nas páginas 68 e 83.

PARK, C. et al. A copy number variation in PKD1L2 is associated with colorectal cancer predisposition in korean population. *International Journal of Cancer*, v. 140, n. 1, p. 86–94, 2017. Citado 2 vezes nas páginas 68 e 83.

PEARSON, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Informa UK Limited, v. 50, n. 302, p. 157–175, 1900. Citado na página 66.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011. Citado 2 vezes nas páginas 73 e 111.

PELCOVITS, A.; NIROULA, R. Acute Myeloid Leukemia: A Review. *R I Med J*, v. 103, n. 3, p. 38–40, 2020. Citado 2 vezes nas páginas 23 e 28.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, v. 64, p. 1–18, 2015. Citado na página 30.

POIANI, M. et al. The impact of cytogenetic risk on the outcomes of allogeneic hematopoietic cell transplantation in patients with relapsed/refractory acute myeloid leukemia: On behalf of the acute leukemia working party (alwp) of the european group for blood and marrow transplantation (ebmt). *American Journal of Hematology*, v. 96, n. 1, p. 40–50, 2021. Citado na página 29.

QUINLAN, J. R. Induction of Decision Trees. *Machine Learning*, n. 1, p. 81–106, 1986. Citado na página 111.

RAHMAN, M. M. et al. Association of *p53* Gene Mutation With *Helicobacter pylori* Infection in Gastric Cancer Patients and Its Correlation With Clinicopathological and Environmental Factors. *World Journal of Oncology*, v. 10, n. 1, p. 46–54, 2019. Citado 2 vezes nas páginas 66 e 67.

RIS, T. et al. Inflammatory biomarkers in infective endocarditis: machine learning to predict mortality. *Clinical & Experimental Immunology*, v. 196, n. 3, p. 374–382, 2019. Citado na página 48.

ROLL, J. D.; REUTHER, G. W. ALK-activating homologous mutations in LTK induce cellular transformation. *PloS One*, v. 7, n. 2, p. e31733–e31733, 2012. Citado na página 69.

ROLL, J. D.; REUTHER, G. W. ALK-activating homologous mutations in LTK induce cellular transformation. *PloS One*, v. 7, n. 2, p. e31733, 2012. Citado na página 70.

ROSE-INMAN, H.; KUEHL, D. Acute leukemia. *Emergency Medicine Clinics of North America*, v. 32, n. 3, p. 579–596, 2014. Citado na página 23.

ROSS, E. G. et al. Predicting future cardiovascular events in patients with peripheral artery disease using electronic health record data. *Circulation: Cardiovascular Quality and Outcomes*, v. 12, n. 3, p. e004741–e004741, 2019. Citado na página 24.

RÜCKER, F. G. et al. Measurable residual disease monitoring in acute myeloid leukemia with t(8;21)(q22;q22.1): results from the AML Study Group. *Blood*, v. 134, p. 1608–1618, 2019. Citado na página 37.

SPACKMAN, K. A. Signal detection theory: valuable tools for evaluation inductive learning. In: *6th International Workshop on Machine Learning*. [S.l.: s.n.], 1989. p. 160–163. Citado na página 80.

SUGAMORI, H. et al. Interim results from a postmarketing surveillance study of patients with *FLT3* -mutated relapsed/refractory AML treated with the FLT3 inhibitor gilteritinib in Japan. *Japanese Journal of Clinical Oncology*, v. 52, n. 7, p. 758–765, 2022. Citado 3 vezes nas páginas 36, 44 e 81.

TANG, D. et al. LncRNA KCNQ1OT1 activated by c-Myc promotes cell proliferation via interacting with FUS to stabilize MAP3K1 in acute promyelocytic leukemia. *Cell Death & Disease*, v. 12, n. 9, p. 795, 2021. Citado na página 70.

TANG, G. et al. Prediction of sepsis in covid-19 using laboratory indicators. *Frontiers in Cellular and Infection Microbiology*, v. 10, 2021. Citado 2 vezes nas páginas 24 e 49.

TYNER, J. W. et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*, v. 562, n. 7728, p. 526–531, 2018. Citado na página 52.

WANG, H. et al. Venetoclax + hypomethylating agents combined with dose-adjusted HAG for relapsed/refractory acute myeloid leukemia: Two case reports. *Medicine*, v. 99, n. 47, p. e23265–e23265, 2020. Citado 3 vezes nas páginas 29, 37 e 44.

WANG, H.-Y. et al. Novel FNDC3B and MECOM fusion and WT1 L378fs* 7 frameshift mutation in an acute myeloid leukaemia patient with cytomorphological and immunophenotypic features reminiscent of acute promyelocytic leukaemia. *British Journal of Haematology*, v. 172, n. 6, p. 987–990, 2016. Citado 2 vezes nas páginas 70 e 85.

WANG, Z.-Q. et al. Agrin promotes the proliferation, invasion and migration of rectal cancer cells via the WNT signaling pathway to contribute to rectal cancer progression. *Journal of Receptor and Signal Transduction Research*, v. 41, n. 4, p. 363–370, 2021. Citado 2 vezes nas páginas 70 e 85.

WEN, P. et al. MICALL2 as a substrate of ubiquitinase TRIM21 regulates tumorigenesis of colorectal cancer. *Journal of Cell Communication and Signaling*, v. 20, n. 1, p. 170, Oct 2022. Citado na página 70.

WILSON, C. S. et al. Gene expression profiling of adult acute myeloid leukemia identifies novel biologic clusters for risk classification and outcome prediction. *Blood*, v. 108, n. 2, p. 685–696, 2006. Citado na página 68.

XIAOSU, Z. et al. Classifying aml patients with inv(16) into high-risk and low-risk relapsed patients based on peritransplantation minimal residual disease determined by cbfβ/myh11 gene expression. *Annals of hematology*, v. 98, n. 1, p. 73–81, 2019. Citado na página 29.

XU, F. et al. Exploration of the role of gene mutations in myelodysplastic syndromes through a sequencing design involving a small number of target genes. *Scientific Reports*, v. 7, p. 43113–43113, 2017. Citado 2 vezes nas páginas 68 e 83.

YANG, X. et al. ANP32A regulates histone H3 acetylation and promotes leukemogenesis. *Leukemia*, v. 32, n. 7, p. 1587–1597, 2018. Citado na página 70.

YANG, Y. et al. PPP1R26 drives hepatocellular carcinoma progression by controlling glycolysis and epithelial-mesenchymal transition. *Journal of Experimental & Clinical Cancer Research*, v. 41, n. 1, p. 101, 2022. Citado na página 70.

YU, X. et al. Predicting lung adenocarcinoma disease progression using methylation-correlated blocks and ensemble machine learning classifiers. *PeerJ*, v. 9, p. e10884–e10884, 2021. Citado na página 24.

ZHANG, H. et al. Elevated mitochondrial SLC25A29 in cancer modulates metabolic status by increasing mitochondria-derived nitric oxide. *Oncogene*, v. 37, n. 19, p. 2545–2558, 2018. Citado na página 70.

ZHAO, Y. et al. The Biological and Clinical Consequences of RNA Splicing Factor U2AF1 Mutation in Myeloid Malignancies. *Cancers*, v. 14, n. 18, p. 4406, 2022. Citado na página 68.

ZHU, L.-Y. et al. Silencing of MICAL-L2 suppresses malignancy of ovarian cancer by inducing mesenchymal-epithelial transition. *Cancer Letters*, v. 363, n. 1, p. 71–82, 2015. Citado na página 70.

ZHU, Y. et al. U2AF1 mutation promotes tumorigenicity through facilitating autophagy flux mediated by FOXO3a activation in myelodysplastic syndromes. *Cell Death & Disease*, v. 12, n. 7, p. 655, Jun 2021. Citado na página 68.

APÊNDICE A – Melhores árvores de decisão geradas pelos modelos individuais

A árvore de decisão (do inglês, *Decision trees* – DT) é uma técnica de ML utilizada para classificação e para regressão. Ela foi proposta por Quinlan (1986) no ano de 1986. A construção de cada DT é feita por meio do particionamento dos dados de forma recursiva. A cada iteração, um atributo é escolhido para ser selecionado como nó da árvore por meio de uma função que define o ganho de informação obtido com a escolha do atributo.

Este apêndice apresenta as melhores árvores de decisão obtidas das melhores florestas aleatórias geradas para todas as combinações possíveis dos conjuntos de dados clínicos (CLIN), de mutação genética (MUT) e de expressão genética (EXP).

Para uma melhor visualização das árvores, é possível acessá-las em formato original, no Google Drive. Elas podem ser interpretadas da seguinte forma: o caminho para à esquerda apresenta as instâncias da classe positiva (sobrevivência), e para à direita, a classe negativa (morte). O valor do atributo é descrito no nó da árvore, bem como a quantidade de pacientes em que é coberto naquele nó. Elas foram geradas com o auxílio da biblioteca de ML *sickit-learn* (PEDREGOSA et al., 2011) usando a linguagem de programação Python.

As árvores de decisão apresentadas nas subseções a seguir são parte dos resultados dos modelos individuais obtidos neste trabalho.

A.1 Melhor DT treinada com os dados clínicos (CLIN)

A Figura 35 apresenta a melhor DT, construída pela melhor floresta da RF treinada com o conjunto de dados CLIN. Na raiz da árvore aparece o atributo *Cytogenetic Info* só que com a categoria *NPM1* que diz respeito a uma mutação neste gene. Na árvore é utilizado o valor 0 que diz respeito a falta desta mutação, esta categoria não está presente no ranqueamento apresentado na Figura 22. No próximo nível é apresentado a idade do diagnóstico e o sexo do paciente, que também estão presentes nas variáveis mais importantes para o modelo. A DT gerada apresentou um tamanho considerável, contudo alguns atributos apareceram diversas vezes como a idade do diagnóstico, a classificação da ELN e a intensidade do tratamento. Essa repetição reforça a relevância destas variáveis para a predição.

A.2 Melhor DT treinada com os dados de mutação genética (MUT)

A Figura 36 apresenta a melhor DT gerada a partir do modelo da RF treinado com os dados de mutação genética. Conforme a Figura 24, o gene *U2AF1* é a variável mais relevante. No entanto, a Figura 36 mostra que a mutação do gene *TP53* é o primeiro fator de relevância da DT, superando a importância da mutação do gene *U2AF1*. Por meio da DT é possível extrair regras de decisão ao seguir seus caminhos a partir da raiz. Por exemplo, dentre os 157 pacientes que não apresentavam mutação no gene *TP53*, 10 foram a óbito.

A.3 Melhor DT treinada com os dados de expressão genética (EXP)

A melhor Árvore de Decisão gerada a partir dos dados em questão é apresentada na Figura 37. Na raiz da árvore, é destacada a expressão do gene *CFD*. No nível subsequente, são apresentadas as expressões dos genes *MICALL2* e *LTK*, que, apesar de não terem sido identificados como as variáveis mais importantes pelo modelo da RF (Figura 25), ocupam as primeiras posições no processo de seleção de variáveis (Subseção 3.4.3). Além disso, a expressão do gene *ITFG1*, considerada a mais relevante pelo modelo (Figura 25), é mostrada em diversos nós da árvore com valores diferentes. A expressão desse gene está associada com o desenvolvimento de tumores (CK et al., 2017).

A.4 Melhor DT treinada com os dados clínicos e de mutação genética (CLIN+MUT)

A melhor DT que combina os dados CLIN e MUT é apresentada na Figura 38. Em comparação com a árvore gerada apenas com os dados CLIN (Figura 35), a inclusão dos dados de mutação genética apenas influenciou na estrutura da árvore. A raiz tem como atributo principal *Treatment Intensity* na categoria *HIT*, indicando a importância da intensidade do tratamento para a sobrevivência dos pacientes. Em seguida, o atributo *Cytogenetic Info* com as categorias *CBFB-MYH11* e *NPM1* são consideradas relevantes, sem que as mutações gênicas apareçam diretamente na árvore. O atributo *Treatment Intensity* é destacado em diversos nós da árvore, reforçando a importância da escolha correta do tratamento para a sobrevivência dos pacientes.

A.5 Melhor DT treinada com os dados clínicos e de expressão genética (CLIN+EXP)

A melhor DT para o conjunto de dados CLIN e EXP é apresentada na Figura 39. Nesta árvore, os atributos de expressão gênica são constantemente presentes nos nós, ao contrário da árvore gerada com o conjunto de dados CLIN e MUT (Figura 38). A presença desses dados confirma a importância da utilização da informação de expressão gênica para prever a sobrevivência. Apesar de ser mais caro de obter, essa informação traz ganhos significativos em termos de desempenho e interpretação para os modelos. É interessante notar a presença da expressão *MPO* na raiz, que aparece na 11ª posição na análise de variáveis (Subseção 3.4.3), mas não é apresentada nas Figuras 29 e 25. Além disso, a intensidade do tratamento também é uma presença constante na árvore, como na árvore gerada para o conjunto de dados clínicos (Figura 38).

A.6 Melhor DT treinada com os dados genéticos (MUT+EXP)

A Figura 40 apresenta a melhor árvore de decisão gerada com base no conjunto de dados genéticos. Os atributos de mutação gênica não apareceram na árvore, mas foram responsáveis por alterar os nós da árvore quando comparado a árvore treinada nos dados de expressão genética (Figura 37). Na raiz da árvore há a expressão genética *LTK*, seguida pela presença das expressões *ELANE* e *ITFG1*. Esses atributos foram identificados como importantes para prever a sobrevivência, pois ocuparam as posições 2ª e 13ª, respectivamente, na fase de análise de atributos (Subseção 3.4.3).

A.7 Melhor DT treinada com os dados clínicos e genéticos (CLIN +MUT+EXP)

A Figura 41 traz a melhor árvore de decisão gerada para todos os conjunto de dados (CLIN+MUT+EXP). Novamente, os atributos de mutação genética não aparecem na árvore, como ocorreu nas árvores resultantes da combinação de dados CLIN+MUT (Figura 38) e MUT+EXP (Figura 40). Isso pode validar a hipótese de que este tipo de informação não tem uma significativa contribuição para a previsão de sobrevivência. Na raiz da árvore está presente a expressão gênica *GLB1*, classificada na 17ª posição na etapa de análise de atributos (Subseção 3.4.3). Em geral, as expressões e atributos relacionados a informações citológicas e tipo de tratamento são claramente visíveis nos nós da árvore, representadas por seus diferentes valores. Isso confirma a importância destas informações para realizar a previsão de sobrevivência.

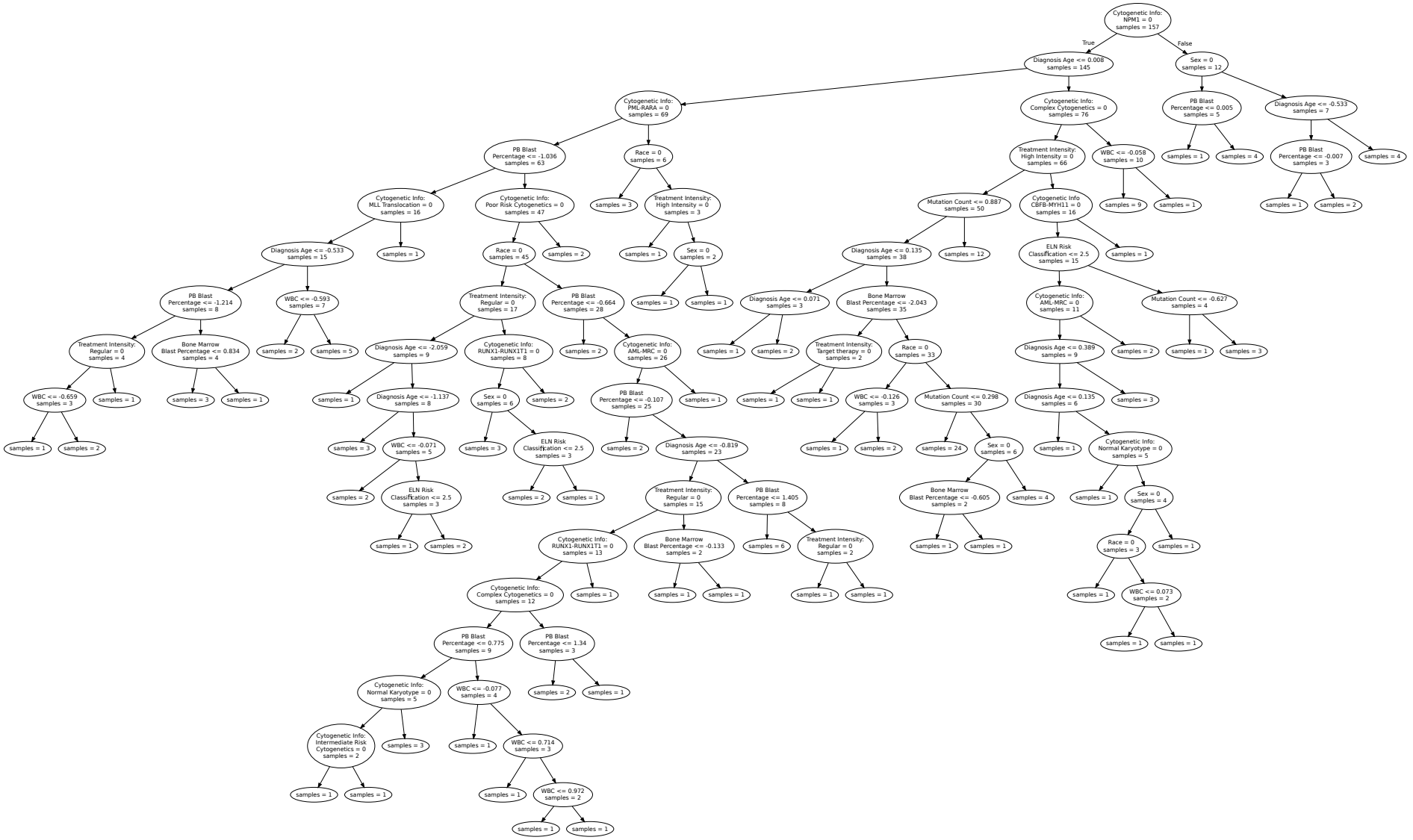

Figura 35 – Melhor árvore de decisão gerada pelo modelo treinado com os dados clínicos.

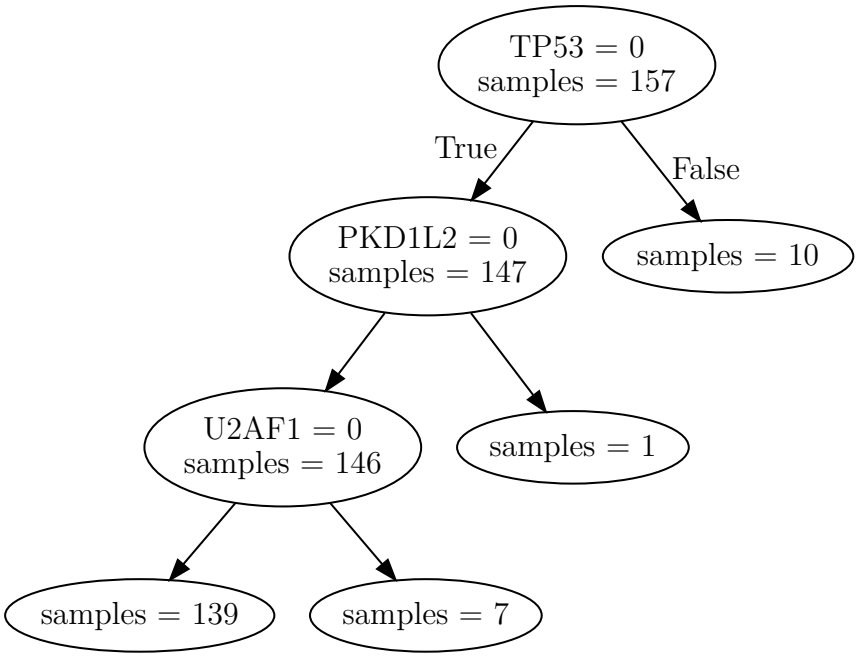

Figura 36 – Melhor árvore de decisão gerada pelo modelo treinado com os dados de mutação genética.

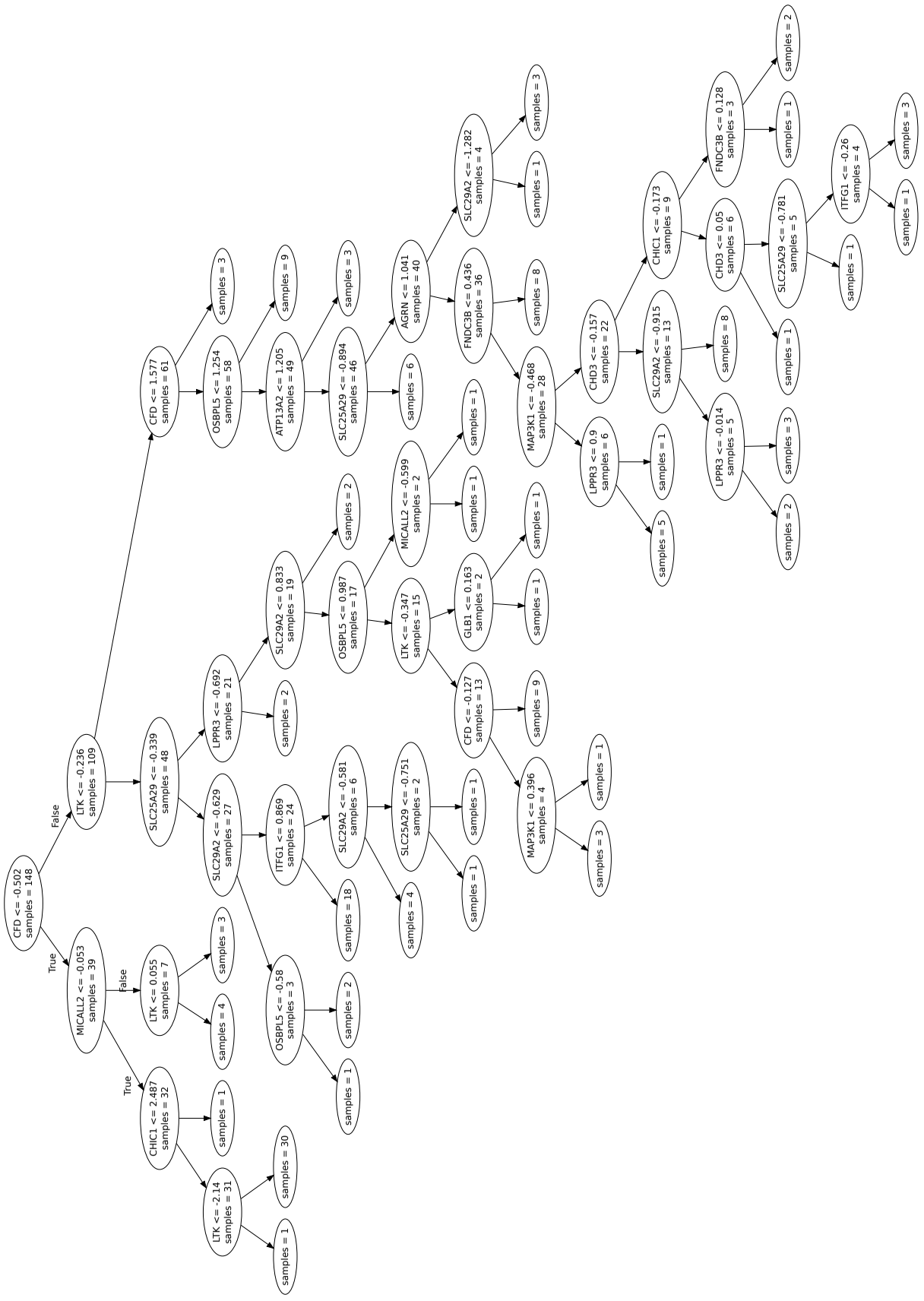

Figura 37 – Melhor árvore de decisão gerada pelo modelo treinado com os dados de expressão genética.

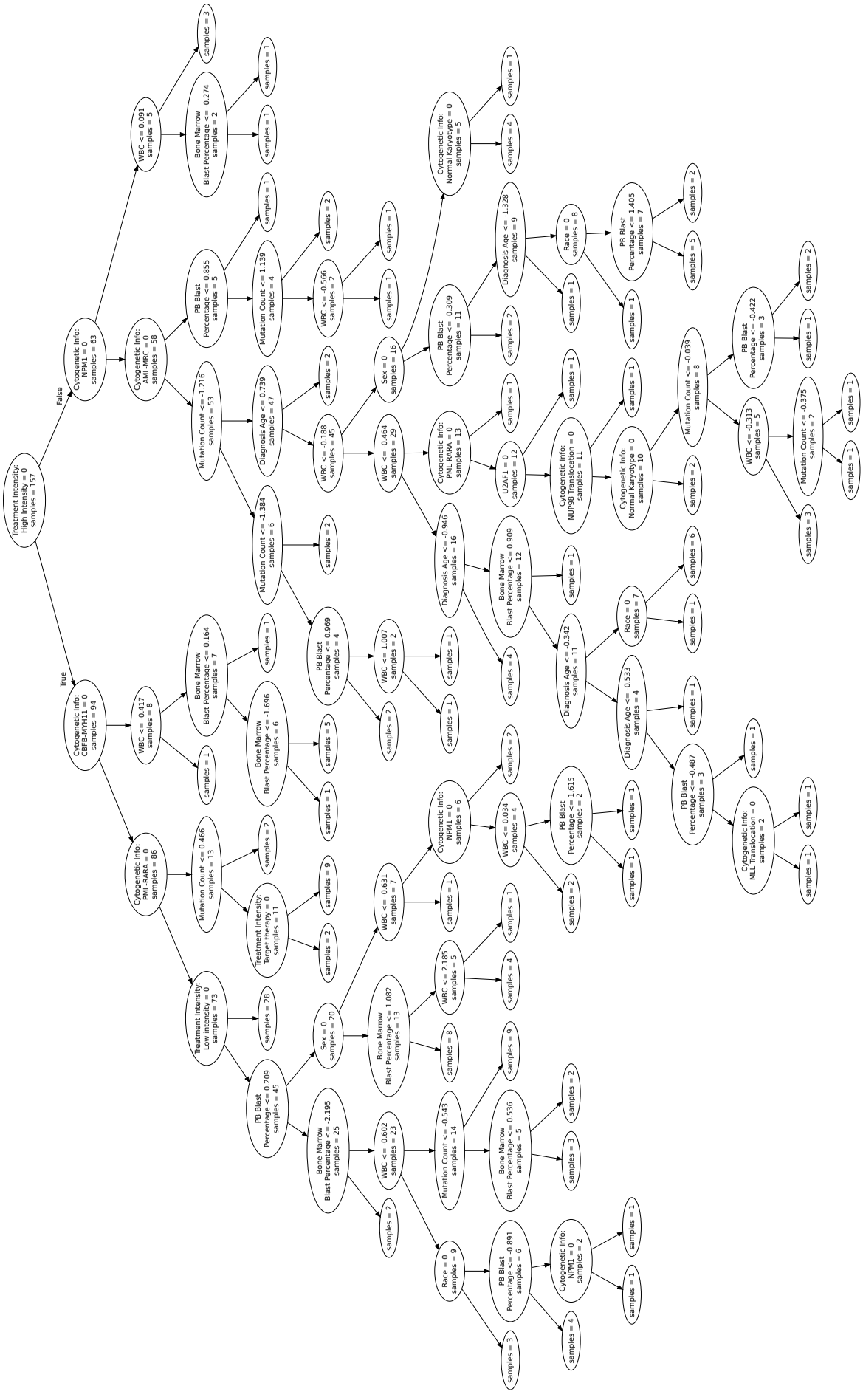

Figura 38 – Melhor árvore de decisão gerada pelo modelo treinado com os dados de clínicos e de mutação genética (CLIN+MUT).

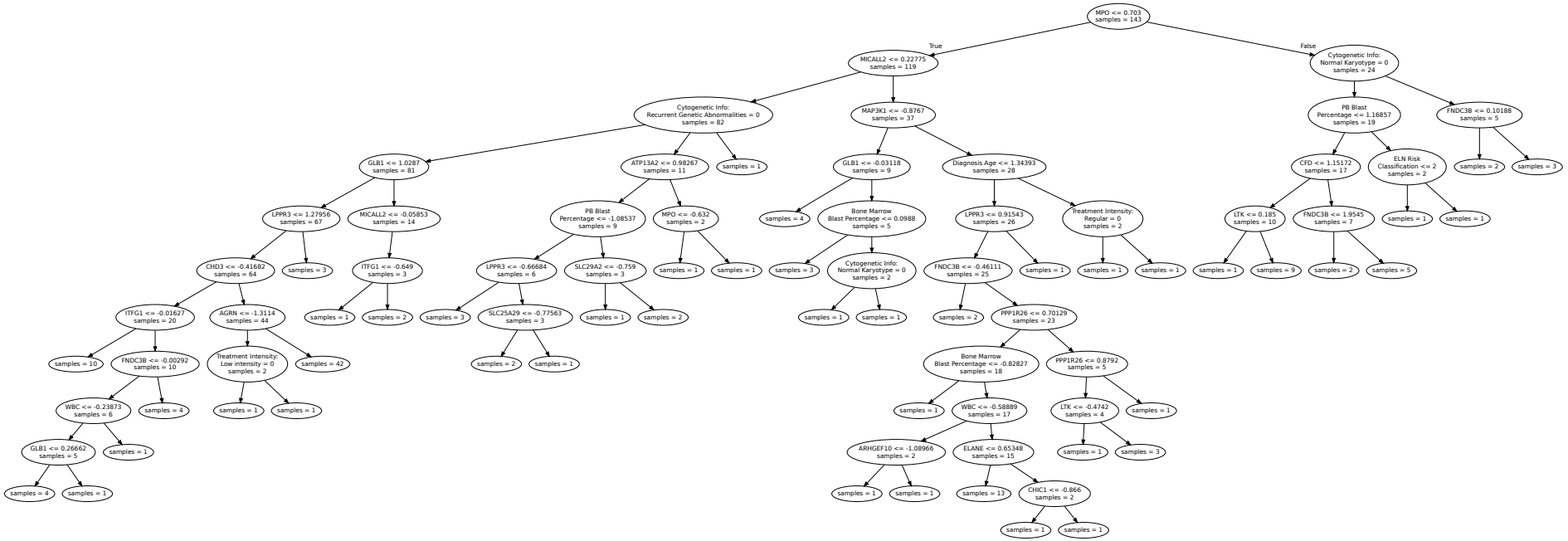

Figura 39 – Melhor árvore de decisão gerada pelo modelo treinado com os dados de clínicos e de expressão genética (CLIN+EXP).

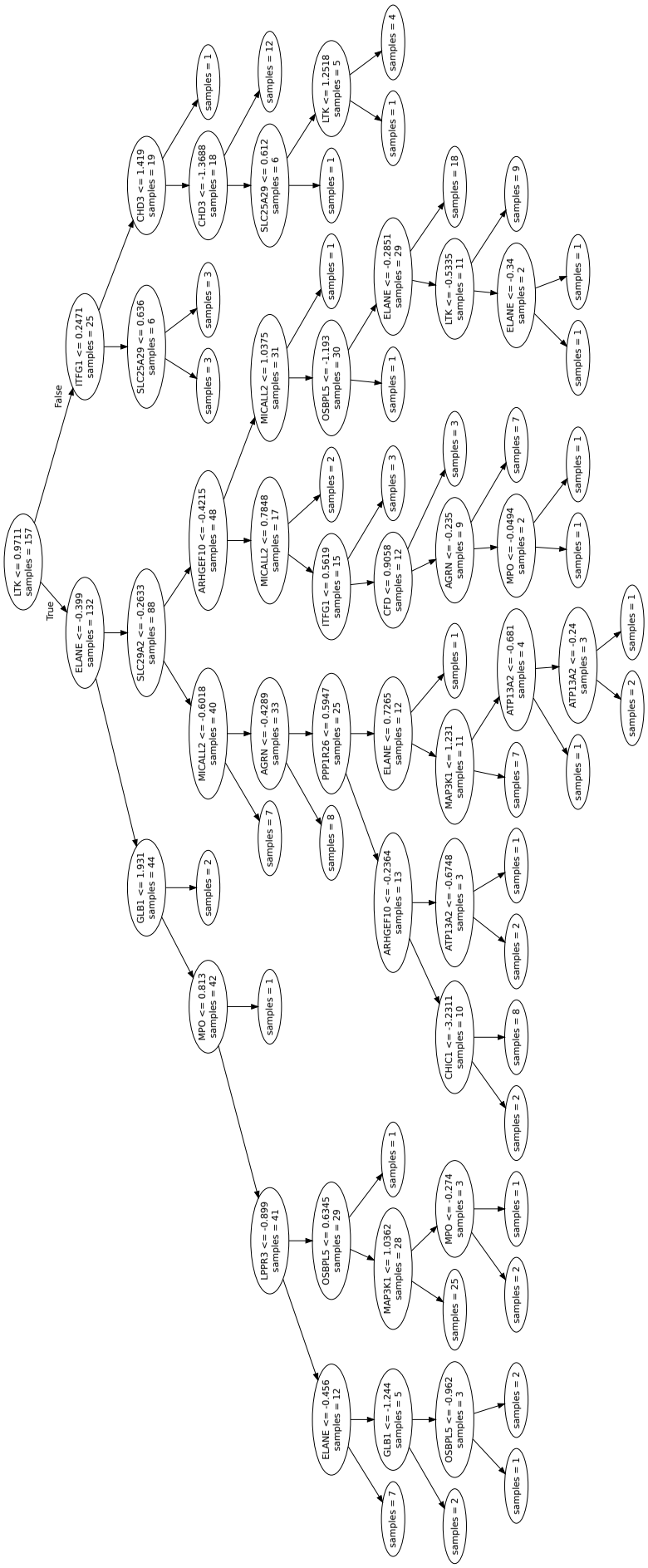

Figura 40 – Melhor árvore de decisão gerada pelo modelo treinado com os dados genéticos (MUT+EXP)

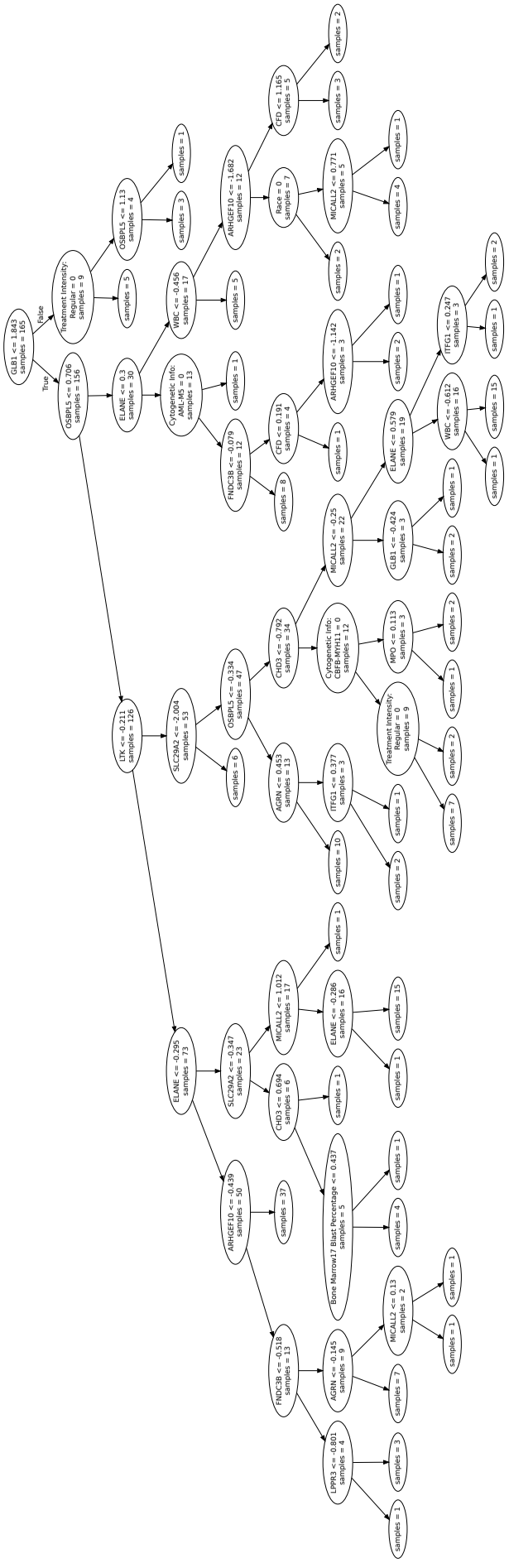

Figura 41 – Melhor árvore de decisão gerada pelo modelo treinado com os dados clínicos e genéticos (CLIN+MUT+EXP).