Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Engenharia de Software

Predição de Recidivas em Pacientes de Oncologia Pediátrica com Leucemia Linfoblástica Aguda Usando Aprendizado de Máquina

Autor: Diego Barbosa da Mota França

Orientador: Prof. Dr. Glauco Vitor Pedrosa

Brasília, DF

2021

Diego Barbosa da Mota França

Predição de Recidivas em Pacientes de Oncologia Pediátrica com Leucemia Linfoblástica Aguda Usando Aprendizado de Máquina

Monografia submetida ao curso de graduação em (Engenharia de Software) da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em (Engenharia de Software).

Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Orientador: Prof. Dr. Glauco Vitor Pedrosa

Brasília, DF

2021

Diego Barbosa da Mota França Predição de Recidivas em Pacientes de Oncologia Pediátrica com Leucemia Linfoblástica Aguda Usando Aprendizado de Máquina/ Diego Barbosa da Mota França. – Brasília, DF, 2021- 70 p. : il. (algumas color.) ; 30 cm.
Orientador: Prof. Dr. Glauco Vitor Pedrosa
Trabalho de Conclusão de Curso – Universidade de Brasília - UnB
Faculdade UnB Gama - FGA , 2021.
1. aprendizado-de-máquina. 2. leucemia-linfoblástica-aguda. I. Prof. Dr. Glauco Vitor Pedrosa. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Predição de Recidivas em Pacientes de Oncologia Pediátrica com Leucemia Linfoblástica Aguda Usando Aprendizado de Máquina
CDU 02:141:005.6

Diego Barbosa da Mota França

Predição de Recidivas em Pacientes de Oncologia Pediátrica com Leucemia Linfoblástica Aguda Usando Aprendizado de Máquina

Monografia submetida ao curso de graduação em (Engenharia de Software) da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em (Engenharia de Software).

Trabalho aprovado. Brasília, DF, 10 de novembro 2021:

Prof. Dr. Glauco Vitor Pedrosa
Orientador

Prof. Dr. John Lenon Cardoso Gardenghi
Convidado 1

Dr. José Carlos Martins Córdoba
Convidado 2

Brasília, DF
2021

Este trabalho é dedicado ao meu tio Ségio, que batalho e manteve o bom humor até o fim, mas infelizmente foi mais uma vítima do câncer. Que essa monografia possa de alguma forma ajudar outras pessoas a evitar a tristeza de perder alguém para essa doença.

Nada é fixo e nada é permanente.

Agradecimentos

Agradeço primeiramente à minha família, em especial os meus pais, por terem sempre me dado todo o apoio na vida acadêmica e fora dela. Também agradeço ao meu orientador Glauco pela orientação e por ter embarcado na minha ideia de e feito todo o necessário por parte dele para que este trabalho pudesse acontecer. Por fim, agradeço à equipe do Hospital da Criança de Brasília, em especial ao dr. José Carlos, ao dr. Ricardo Camargo e à toda a equipe do Laboratório de Pesquisa Translacional por terem acolhido a ideia da pesquisa e terem dedicado tempo tanto para compilar os dados quanto para dar esclarecimentos e tirar dúvidas.

Resumo

Durante o tratamento da Leucemia Linfoblástica Aguda (LLA) na infância, a previsão de recidiva, caracterizada pelo retorno do câncer após o paciente passar por um tratamento bem sucedido, é um fator crítico para o sucesso do tratamento e para o planejamento de acompanhamento da doença. O objetivo deste trabalho foi construir um modelo computacional para a previsão de recidiva com base em algoritmos de aprendizado de máquina. A proposta foi construir um modelo preditivo baseado em um conjunto de treinamento obtido a partir de dados laboratoriais e clínicos de pacientes do Hospital da Criança de Brasília (HCB). A construção desse modelo computacional permitiu identificar, com quase 85% de acurácia, pacientes em grupos de risco de recidiva. O principal benefício dessa previsão é possibilitar que a equipe de saúde tome ações que julguem mais adequadas em prol da saúde e visando uma maior chance de sobrevida dos pacientes.

Palavras-chaves: aprendizado de máquina; aprendizado supervisionado; recidiva; leucemia linfoblástica aguda;. oncologia pediátrica.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AUC	*area under the curve*
ROC	*receiver operating characteristics*
TFP	taxa de falsos positivos
DF	Distrito Federal
CEP	Comitê de Ética em Pesquisa
HCB	Hospital da Criança de Brasília
K-NN	*K-Nearest Neighbours*
LLA	leucemia linfoblástica aguda
OSM	Otimização Sequencial Mínima
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Oversampling Technique
VPN	Valor de Predição Negativa
VPP	Valor de Predição Positiva

Sumário

1 Introdução

1.1 Contexto

A leucemia é uma neoplasia maligna que atinge as células do sangue cuja principal característica é o acúmulo, na medula óssea, de células doentes que substituem as células sanguíneas normais (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2019, p. 41). Esse é o tipo de câncer que possui maior incidência entre crianças e adolescentes no mundo (SANTOS, 2018). Entre os jovens com menos de 15 anos, é o câncer que ocorre em cerca de 30% dos casos e em 20% dos casos abaixo de 20 anos (SANTOS, 2018).

Entre os diversos subtipos de leucemia, a Leucemia Linfoblástica Aguda (LLA) é uma das quatro principais. Na população infanto-juvenil, a LLA é a mais frequente entre menores de 5 anos, possuindo maior incidência em crianças entre 2 e 3 anos e sendo mais frequente entre meninos (SANTOS, 2018). Segundo Allemani et al. (2015), no Brasil a sobrevida, por pelo menos 5 anos, para jovens com LLA com idade de 0 a 14 anos caiu de 71,9% entre 1995 e 1999 para 65,80% entre 2005 e 2009. Nesse mesmo artigo é apontado que países como Áustria, Bélgica e Canadá chegaram a ter mais de 90% de sobrevida no mesmo intervalo de 2005 a 2009.

Ainda sobre a LLA, ela pode ser subdividida em dois tipos, B e T. Segundo Farias e Castro (2004), essa divisão ocorre “de acordo com a expressão de antígenos específicos, podendo, inicialmente, essas leucemias ser classificadas de linhagem T ou B, de acordo com as características imunofenotípicas dos linfoblastos”. As LLAs de linhagem B ainda possuem divisões, sendo elas a pró-B, comum, pré-B e B-maduro. Todas essas divisões se referem aos “estágios de diferenciação normal dos progenitores B na medula óssea” (FARIAS; CASTRO, 2004).

Apesar de se ter uma alta taxa de sobrevida entre pacientes pediátricos, ainda há diversos riscos para a criança e o jovem com LLA que recebem tratamento. Um desses riscos é o da recidiva, que consiste no retorno do câncer após o tratamento bem sucedido. Cerca de 20% dos jovens com LLA que sofrem recidiva não possuem um bom prognóstico, chegando a óbito (PAN et al., 2017).

Para evitar a ocorrência da recidiva, vários recursos podem ser utilizados, sendo um deles o uso de protocolos de tratamento mais intensivos (PEDROSA; LINS, 2002). Mas o uso desses protocolos aumenta os efeitos colaterais do tratamento no organismo do paciente. Segundo Hunger e Mullighan (2015), o uso de terapias mais intensivas aumenta o risco de óbito do paciente em virtude dos efeitos

tóxicos dos medicamentos utilizados. Deste modo, estas opções devem ser usadas apenas quando há real necessidade. Por isso, um grande desafio no gerenciamento do tratamento da LLA na infância é classificar os pacientes em grupos de risco apropriados para um melhor acompanhamento clínico. De fato, existe a preocupação em antever o mais cedo possível as ocorrências de recidivas para que seja feito o tratamento na intensidade apropriada para o risco de recidiva (TEACHEY; HUNGER, 2013), e assim mitigar os maus cursos da doença nos pacientes.

A estratificação do tratamento quimioterápico através do reconhecimento precoce de resultados relevantes à determinação do recurso terapêutico a ser utilizado é extremamente importante para mitigar os maus cursos da doença nesses pacientes. Por isso, uma predição precoce de uma recidiva do tratamento é uma fonte valiosa para a equipe médica direcionar ações visando uma maior chance de sobrevida dos pacientes. Isso não é tarefa simples pois, ao longo do tempo, na medida em que novos estudos foram feitos e novos exames criados, muitas variáveis clínicas previamente usadas para previsão não são mais prognósticas (TEACHEY; HUNGER, 2013).

1.2 Problema de Pesquisa

O principal problema a ser investigado neste trabalho pode ser formulado pela seguinte pergunta:

É possível prever com alta confiabilidade a recidiva de pacientes pediátricos com LLA usando algoritmos de aprendizado de máquina a partir de dados clínicos e laboratoriais dos pacientes?

1.3 Objetivos

O objetivo geral deste trabalho é verificar a eficácia com que é possível prever uma recidiva de paciente infanto-juvenil com LLA através de uso de algoritmos de aprendizado de máquina utilizando dados laboratoriais e clínicos de pacientes pediátricos. Para isso serão analisados a evolução de crianças e adolescentes com e sem recidiva da LLA B, diagnosticadas e tratadas pelo Hospital da Criança de Brasília (HCB), e identificada e avaliada a influência de fatores prognósticos para a ocorrência da recidiva e para a sobrevida global.

Os objetivos específicos são:

- Coletar um conjunto de informações laboratoriais e clínicas de pacientes pediátricos para a construção de uma base de dado para o treinamento dos algoritmos de aprendizado de máquina;

- Realizar testes e identificar o desempenho de diferentes algoritmos de classificação na predição de recidivas de pacientes com LLA B;
- Identificar quais dados laboratoriais e clínicos, dentre os coletados, são os melhores preditores de recidiva em pacientes pediátricos com LLA B.

1.4 Organização do Texto

O restante do texto deste trabalho está organizado nos seguintes capítulos:

- No Capítulo 2 - *Algoritmos de Classificação* são apresentados algoritmos a serem utilizados no presente trabalho, assim como conceitos relevantes relacionados a esta área.
- No Capítulo 3 - *Metodologia* é apresentada a metodologia que será utilizada para o desenvolvimento deste trabalho.
- No Capítulo 4 - *Resultados Preliminares* são apresentados os resultados preliminares obtidos a partir do pré-processamento e análise estatística dos dados.

2 Aprendizado de Máquina

Este capítulo apresenta uma revisão teórica sobre a área de aprendizado de máquina – especificamente sobre os algoritmos de classificação supervisionados – que foram utilizados para o desenvolvimento deste trabalho, além das métricas que foram usadas para medir o desempenho dos modelos gerados.

2.1 Considerações Iniciais

Atualmente, a área de aprendizado de máquina tem sido amplamente utilizada na sociedade contemporânea. Por exemplo, diariamente pessoas no mundo todo compartilham informações sobre suas localizações a partir do GPS em seus celulares ou automóveis. Esse tipo de informação é usado em diversas pesquisas que envolvem aprendizado de máquina e relacionadas à mobilidade nas cidades, como para melhorar a sugestão de trajetos a partir do meio de transporte escolhido pela pessoa (YANG et al., 2018) e prever o fluxo urbano (XIE et al., 2020).

Outro exemplo são pesquisadores ao redor do mundo tendo condições de compartilhar imagens de exames para criar bancos de imagens disponibilizados para serem usados em pesquisas. Isto está possibilitando, na pandemia do Covid-19, que o aprendizado de máquina seja utilizado com imagens de exames para ajudar a combater esta doença. Por exemplo, estudos usando imagens de raio-X pulmonares para diagnóstico de Covid-19 (APOSTOLOPOULOS; MPESIANA, 2020) ou de exames de tomografia computadorizada de pulmões de infectados para classificar automaticamente a severidade da doença (TANG et al., 2020).

Em um tempo em que é possível gerar, armazenar e compartilhar diariamente uma enorme quantidade de dados, existe oportunidade para informações úteis que um ser humano não conseguiria. Daí o principal benefício dos algoritmos de classificação da área de aprendizado de máquina.

2.2 Fundamentos e Conceitos Básicos

Ao se trabalhar com algoritmos de aprendizado de máquina, três conceitos básicos precisam ser definidos: o conjunto de dados (*data set*), a instância (*data point*) e os atributos (*features*). Para exemplificar esses conceitos, considere as informações contidas na Figura 1. Ela mostra dados de 4 pessoas, contendo altura, peso e a idade de cada uma delas. Neste exemplo, os atributos (*features*) são a altura, peso e idade,

ou seja, as colunas da tabela. Já as instâncias (*data point*) é o conjunto de dados disponível para uma única pessoa, ou seja, seria a linha da tabela com todos os dados existentes e/ou ausentes. Os dados ausentes nas linhas da tabela (que é o caso do peso do João e a idade da Ana) são chamados de dados faltantes (*missing data*). Por fim, a junção de todos os dados das pessoas, ou seja, de todas as instâncias, é o que chamamos de conjunto de dados (*data set*).

	ALTURA (CM)	PESO (KG)	IDADE (ANOS)
JOÃO	190		50
MARIA	170	60	42
DANIEL	177	70	35
ANA	165	50	

Figura 1 – Exemplo hipotético de um conjunto de dados em uma planilha. As colunas são as características, e as linhas com os dados as instâncias.

O exemplo da Figura 1 é de um conjunto de dados com poucos elementos. Mas nos dias atuais, é cada vez maior o tamanho dos conjuntos de dados existentes, de tal forma que uma pessoa precisa da ajuda de recursos tecnológicos, entre eles de algoritmos baseados em aprendizado de máquina, para entender e usar essas informações.

No aprendizado de máquina, o principal uso dos conjuntos de dados é no treinamento inicial do algoritmo, para que ele possa processar os dados recebidos e, a partir disso, gerar modelos preditivos. O treinamento é o equivalente ao momento para o algoritmo “estudar” os dados e a partir disso ele possa ter um aprendizado. Esses modelos podem ser utilizados então para atividades como classificar uma instâncias dentre uma série de categorias possíveis.

Outro uso comum dos conjuntos de dados é para o teste dos modelos preditivos gerados. Para isso, é usado um conjunto com dados que não foram usados pelo algoritmo durante o treinamento, para que seja testado e, assim, verificar qual é o desempenho do modelo ao lidar com novas instâncias. Esse teste é importante para garantir que o modelo gerado não está tendo o problema de sobre-ajuste, que é quando o desempenho com dados novos é muito pior do que o que era com os dados de treinamento. O sobre-ajuste é o equivalente, no contexto de aprendizado de máquina, a decorar as respostas de um questionário de estudo ao invés de aprender realmente o conteúdo.

No geral, os conjuntos de dados devem passar pelo chamado pré-processamento antes que possam ser usados pelos algoritmos de aprendizado de máquina. Este processo, que pode ser feito várias vezes e em diferentes momentos, tem por objetivo preparar o dado para o uso com o algoritmo. Se isso não for feito

o desempenho do treinamento pode ser pior ou, dependendo do caso, provocar até erros durante a execução do algoritmo.

Alguns procedimentos que costumam ser realizados neste momento são unificação de nomes de categorias (por exemplo, usar apenas o singular da palavra da categoria, ou colocar todas as letras em maiúsculo), remoção de instâncias que por algum motivo não deveriam estar no conjunto de dados e determinar como dados em branco nas características serão tratados.

2.3 Algoritmos Supervisionados

Algoritmos supervisionados pertencem à uma das categorias existentes de algoritmos de aprendizado de máquina. A principal característica dos algoritmos desta categoria é que, para que eles funcionem, é necessário alimentá-los com dados previamente rotulados. Por exemplo, suponha a tarefa de determinar se uma transação bancária é fraudulenta ou não. Para os algoritmos supervisionados funcionarem eles devem ser treinados usando dados de transações previamente rotuladas em fraudulentas e não-fraudulentas. Assim o algoritmo “aprende” com essas informações e, com base nesse aprendizado, ele conseguirá prever futuras transações não rotuladas. Essa predição pode ocorrer através de duas abordagens: classificação e/ou regressão.

Nos algoritmos de classificação, os dados são rotulados em 2 ou mais classes discretas. Já os algoritmos de regressão poderão classificar dados em valores contínuos como, por exemplo, uma temperatura ou um valor monetário. Cabe ressaltar que nada impede de um algoritmo ser capaz de realizar tanto classificações quanto regressões, embora nem sempre esse seja o caso. Neste trabalho, tem-se uma tarefa de classificação binária: classificar se um paciente terá, ou não, recidiva no seu tratamento.

A seguir é detalhado o funcionamento de alguns algoritmos supervisionados que serão utilizados neste trabalho.

2.3.1 K-Nearest Neighbours (K-NN)

O *K-Nearest Neighbours* (K-NN) é um algoritmo criado em 1967 por Cover e Hart (1967) que tem como objetivo classificar um elemento a partir dos rótulos que os k vizinhos mais próximos, já rotulados e pertencentes a um conjunto de dados, possuem.

Para realizar a classificação, o K-NN calcula a distância entre os diferentes elementos. A cada vez que é executado, o K-NN calcula a distância entre as instâncias

que se quer classificar e cada um dos elementos existentes no conjunto de dados. Após o término dos cálculos, é criada uma lista em ordem crescente de proximidade dos elementos do conjunto de dados com o que será classificado. Por não ter uma fase de treinamento para gerar um modelo, o K-NN é classificado como algoritmo preguiçoso (AUNG; NAGAYAMA; TAMAKI, 2017).

Para realizar essa medição, pode-se usar várias fórmulas de cálculo de distâncias como, por exemplo, a euclidiana, a de Chebychev e a Manhattan (MULAK; TALHAR, 2015). A escolha da função de distância deve levar em conta se as características dos dados são qualitativas, quantitativas ou uma mistura de ambos. Isso porque algumas funções funcionam melhor para um ou outro tipo de dado (BATISTA; SILVA et al., 2009).

De modo geral, os seguintes passos são seguidos por algoritmos K-NN:

1. definir o valor de *k* que será usado na execução do algoritmo;
2. para cada item presente no conjunto de dados, calcular a distância vetorial entre ele e a instância a ser classificada;
3. classificar os itens em ordem crescente de distância para a instância a ser classificada;
4. classificar a instância de acordo com a classificação mais frequente entre os *k* primeiros itens da lista.

O principal parâmetro desse algoritmo é o valor de *k*. É ele que determina quantos vizinhos próximos serão usados como base para classificar o elemento. Segundo os testes realizados por Batista, Silva et al. (2009), o *k* escolhido tem importância significativa na acurácia do algoritmo. O melhor valor varia caso a caso, e o peso dele no desempenho do algoritmo é afetado pela função de distância utilizada, assim como pelo uso de funções-peso.

Cabe destacar que existem variações quanto ao critério adotado para classificar a instância. Uma variação comum é colocar peso no voto de cada um dos *k* primeiro elementos da lista. Esse peso costuma ser o inverso da distância do elemento para a instância sendo classificada. Ou seja, quanto mais próximo, maior o poder de voto do item (LAROSE, 2015). Essa técnica ajuda a evitar que ocorram empates nas votações.

Outra variação, proposta por Aung, Nagayama e Tamaki (2017), é que a classificação não seja de acordo com os rótulos mais presentes entre os *k* primeiros itens, mas sim pela média da distância dos elementos pertencentes a cada um dos

n rótulos que apareceram entre os *k* itens, com $n <= k$. Ou seja, ele classifica de acordo com o grupo de rótulo que se encontra mais próximo da instância.

Uma das vantagens desse algoritmo é a facilidade de entendê-lo. A ideia de medir proximidade entre pontos diferentes de um espaço é algo que as pessoas costumam ver desde a escola para espaços com uma, duas ou três dimensões, assim como a noção de classificar elementos de acordo com essa distância medida. Logo, fica fácil entender a ideia geral do funcionamento do K-NN.

Outro aspecto positivo do K-NN é necessitar de poucos parâmetros, tornando a tarefa de ajustá-lo mais simples quando comparado a outros algoritmos de aprendizado de máquina. Isso porque o K-NN, em sua versão mais básica, requer como parâmetros obrigatórios apenas o valor de *k* e a fórmula usada para o cálculo da distância. Opcionalmente, ainda é possível incluir o parâmetro do peso a ser aplicado nos votos.

Apesar da facilidade, o K-NN possui alguns pontos que devem ser levados em conta ao usá-lo. Discrepâncias entre o intervalo de valores das diferentes características podem gerar problemas para o funcionamento desse algoritmo, assim como em outros que dependam de medidas de distância (PIRYONESI; EL-DIRABY, 2020). Isso ocorre pois intervalos maiores acabam tendo mais peso na medição da distância entre as instâncias. Por isso, é necessário normalizar os dados antes de serem processados pelo K-NN, caso discrepâncias estejam presentes (AYYADEVARA, 2018; PIRYONESI; EL-DIRABY, 2020).

Outro ponto que necessita atenção é que, por ser um algoritmo preguiçoso, medidas devem ser tomadas para aumentar a velocidade de execução do k-NN. Isto se deve ao fato que, a cada execução do algoritmo, o cálculo da distância deve ser feito para cada item do conjunto de dados, o que acaba tornando esse algoritmo desaconselhável em contextos que necessitem de análises em tempo real. Uma das alternativas para lidar com esse problema é realizar os cálculos de distância fazendo uso de multiprocessamento (AUNG; NAGAYAMA; TAMAKI, 2017).

2.3.2 Naive Bayes

O algoritmo *Naive Bayes* é um algoritmo supervisionado probabilístico de classificação. O nome do algoritmo pode ser traduzido como Bayes ingênuo, e a "ingenuidade" dele está no fato dele partir do princípio de que as características do conjunto de dados são independentes entre si.

Esse algoritmo se baseia no Teorema de Bayes, que é um teorema estatístico que permite que seja calculada a probabilidade de um evento *A* acontecer, dado que o evento *B* ocorreu. Para aplicar esse teorema, é necessário que sejam conhecidas as

probabilidades de A e de B ocorrerem independentemente. A partir disso, é possível calcular a probabilidade a partir da fórmula:

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}, (1)$$

onde $P(X)$ é a probabilidade de um evento X e $P(X|Y)$ é interpretado como probabilidade desse mesmo evento X, dado que o Y ocorreu.

Com base no exposto acima, podemos entender o porquê do algoritmo ser supervisionado, ou ao menos semi-supervisionado. Deve ser possível calcular a probabilidade das características ocorrerem para que ele funcione corretamente, e para isso é necessário que pelo menos parte dos dados estejam rotulados.

Para o uso no algoritmo, a probabilidade para cada classe é dada por:

$$P(C|X) = P(C) \times P(C|x_1) \times P(C|x_2) \times ... \times P(C|x_n), (2)$$

onde C é uma das classificações possíveis, X é o conjunto das características existentes para o conjunto de dados, x é cada uma das características desse conjunto e n é o número total de características existentes no conjunto.

Ao contrário da fórmula original do Teorema de Bayes, na do algoritmo não há o denominador, pois ele é constante e o mesmo para cada uma dos rótulos, podendo ser retirado da fórmula final. Além disso, usar $P(C|x_i)$, onde $1 \leq i \leq n$, só é possível pois parte-se do princípio que as características do conjunto de dados são independentes entre si.

De maneira geral, o algoritmo *Naive Bayes* funciona da seguinte maneira:

1. para cada rótulo possível, calcula, usando a fórmula 2 acima, a probabilidade de um determinada instâncias pertencer ao rótulo;
2. o algoritmo verifica qual é a maior das probabilidades calculadas;
3. classifica a instância com o rótulo que teve a maior probabilidade dentre todos.

Mesmo quando a premissa das características serem independentes entre si não é verdadeira o *Naive Bayes* costuma ter bom desempenho (HAND; YU, 2001).

Isto se deve ao fato que nestes casos, mesmo a função de perda usando o erro quadrático ter valores altos, ao usar o erro zero-um há um erro baixo. (DOMINGOS; PAZZANI, 1997). Isso faz com que a probabilidade calculada por ele não seja confiável, mas a classificação gerada por ele sim.

Um dos pontos fortes desse algoritmo é ser simples de entender, bastando a pessoa conhecer o Teorema de Bayes. Isso torna mais fácil entender casos onde ele é mais ou menos recomendável. Outra vantagem do *Naive Bayes* está em ele conseguir trabalhar com dados contínuos sem necessitar discretização. No caso deste tipo de dado, existe a opção de trabalhar com atributos numéricos como se tivessem uma distribuição gaussiana ou então usar a estimativa de densidade por Kernel (DOMINGOS; PAZZANI, 1997).

Por outro lado, o *Naive Bayes* possui alguns pontos fracos. O primeiro, que já foi mencionado neste trabalho, está no fato de que muitas vezes não haverá independência entre as características do conjunto de dados. Isso afeta negativamente a confiabilidade das probabilidades calculadas pelo algoritmo e pode afetar significativamente o desempenho do mesmo nas classificações, embora muitas vezes não ocorra (DOMINGOS; PAZZANI, 1997).

Outro ponto fraco, e que deve ser tratado ao usar este algoritmo, é o problema da frequência zero. Ele consiste em, havendo uma característica cujos dados são categorias, ocorrer uma categoria no caso de teste que não havia nas instâncias usadas no treinamento. Caso isso ocorra e nenhuma medida for tomada, a probabilidade para o valor em questão se tornará zero.

2.3.3 Algoritmos baseados em Árvore de Decisão

Antes de falar sobre os próximos 2 algoritmos que foram usados neste trabalho, cabe explicar o conceito de árvore de decisão usado por ambos.

Árvores de decisão são usadas por vários algoritmos de aprendizado de máquina supervisionados. Eles, após processarem o conjunto de dados de treinamento, produzem uma ou mais árvores de decisão que são utilizadas para decidir uma categoria (classificação) ou um valor (regressão) para novas instâncias que estão sendo analisadas.

A Figura 2 mostra um exemplo de diagrama de arvore de decisão. Os retângulos no exemplo são chamados nós, enquanto as linhas que partem dos nós são chamadas de vértices. As árvores de decisão têm um nó de entrada, chamado de nó raiz, a partir do qual a árvore vai crescer e de onde os demais nós e vértices vão sair. Cabe ressaltar que sempre há um vértice entre dois nós, mas nem todos os nós possuem vértices saindo deles.

Nos nós que possuem vértices saindo deles estão o atributo do conjunto de dados que, naquele ponto da árvore de decisão, servirá de pivô para a análise (sendo por isso chamado algumas vezes de nó de decisão). Caso o nó não tenha

[1] Disponível em <https://commons.wikimedia.org/w/index.php?curid=90405437> Acesso em jul. 2021.

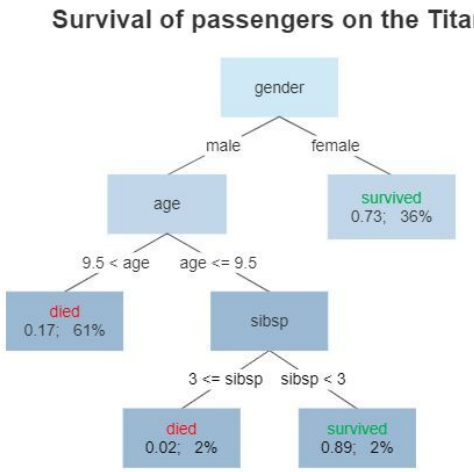

Figura 2 – Exemplo de diagrama de uma árvore de decisão mostrando a chance de sobrevivência de passageiros do Titanic a partir de determinadas características deles. Fonte: Wikimédia.[1]

vértices saindo dele, também denominado de “folha”, ele será um nó que conterá a classificação final da instância em análise. Já nos vértices dos nós com “galhos” estão os valores que irão determinar pra onde seguirá a análise para classificação. Se para aquele determinado atributo, a instância tiver um valor dentro da janela especificada por aquele vértice, a análise continuará pelo nó que está na outra ponta desse vértice.

Usando o diagrama da Figura 2 para exemplificar como se dá uma classificação por meio de uma árvore de decisão, o primeiro atributo a ser analisado para qualquer passageiro é se o gênero é masculino ou feminino. Se for feminino, a classificação será de que ela tem 73% de chances de ter sobrevivido. Já se for masculino, não teremos uma classificação no próximo nó, mas sim um novo atributo para analisar, nesse caso a idade dele. Nesse caso, se o passageiro tiver uma idade menor do que 9,5 anos, ele tem 17% de chance de ter morrido ou, vendo de outro ponto de vista, 83% de chance de ter sobrevivido. Caso a idade seja maior ou igual 9,5 anos, a análise continua com o próximo nó, até chegar a uma classificação.

No geral, a principal vantagem dos algoritmos de árvores de decisão está na facilidade de compreender e analisar o modelo preditivo gerado ao visualizar o diagrama da árvore. Além disso, esses algoritmos costumam poder trabalhar tanto com atributos contínuos quanto discretos.

2.3.3.1 Random Forest

O Algoritmo *Random Forest* é um algoritmo de aprendizado de máquina supervisionado, que serve tanto para fazer classificações como regressões. O primeiro

algoritmo de *random decision forest* foi proposto por Ho (1995) em 1995, e em 2001 o *Random Forest* foi proposto por Breiman (2001) a partir dos trabalhos de Ho.

A “floresta” (*forest*) no nome do algoritmo vem do fato dele ser baseado no uso de várias árvores de decisões ao mesmo tempo, se tornando assim uma floresta de árvores de decisão. Mas usar várias árvores iguais seria inútil, e é nesse ponto que entra a “aleatoriedade” (*random*) do nome.

A primeira aleatoriedade inserida no *Random Forest* é o chamado *bootstrap aggregating* ou simplesmente *bagging*, onde cada árvore usa apenas parte das instâncias existentes no conjunto de dados, sendo estas selecionadas por amostragem aleatória com repetição para cada árvore.

A segunda aleatoriedade inserida é nos atributos que podem ser usados em cada nó de cada árvore, conhecido como *feature bagging*. Basicamente, ao invés de usar todos os atributos existentes para o conjunto de dados, cada nó utiliza apenas um subconjunto destes atributos, selecionados aleatoriamente, para determinar qual será o atributo final daquele nó.

A partir desses subconjuntos de instâncias e de atributos, as árvores são então treinadas. Com as árvores de decisão devidamente treinadas, o processo de classificação ou regressão funciona por meio de votos de cada árvore da floresta. A instância sendo classificada é processada por todas as árvores da floresta, e o resultado obtido em cada uma delas é guardado. No caso de classificação, aquela que for feita pela maior quantidade de árvores será a final. No caso de regressão, o valor final será a média aritmética dos resultados obtidos nas árvores.

O *Random Forest* é um algoritmo muito usado. Isso se deve, em parte, pelo fato dele ter uma vantagem significativa entre as árvores de decisão. Ao contrário de outros algoritmos do tipo, o *Random Forest* não tende a ter o problema de sobre-ajuste mencionado anteriormente neste capítulo. Desse modo, modelos gerado por *Random Forests* tendem a generalizar o aprendizado com o conjunto de dados usado no treinamento para instâncias novas.

Essa vantagem se dá justamente por esse algoritmo não usar uma mas várias diferentes árvores de decisão. Além disso, como já explicado acima, cada uma dessas árvores são criadas com 2 mecanismos que fazem com que cada árvore gerada sejam independentes entre si, fazendo com que elas possam acabar contrabalanceando fraquezas que outras árvores da “floresta” possam ter.

Indo para desvantagens deste algoritmo, a principal é que, por usar o sistema de votação com o conjunto de árvores de decisão, não há uma árvore de decisão final que possa ser visualizada. Isso se deve ao fato de que, ao fim da fase do treinamento, a floresta não terá uma mas sim dezenas, centenas ou até milhares de árvores de

decisão. Desse modo, a capacidade de melhor analisar e entender o resultado do aprendizado mencionada na introdução às árvores de decisão é consideravelmente prejudicada.

2.3.3.2 Árvores de Decisão C4.5

O C4.5 é um algoritmo de classificação que usa árvores de decisão, criado por (QUINLAN, 1993). Ele é considerado uma derivação de um outro algoritmo criado pelo mesmo autor, chamado ID3. A base para o funcionamento desse algoritmo é o conceito de ganho de informação (*information gain*).

O conceito de ganho de informação propõe que, quanto maior a probabilidade de um determinado evento ocorrer, menos informação têm-se a obter caso ele realmente ocorra, pois a entropia será menor. Logo, o ganho possível de informação com a ocorrência de um determinado evento é inversamente proporcional à probabilidade dele ocorrer. Esse conceito foi inicialmente proposto Shannon (1948) em seu trabalho sobre teoria da informação.

No algoritmo C4.5, o ganho de informação é usado como a base para escolher qual o atributo ficará em um nó de decisão. Para isso, o funcionamento básico do algoritmo é primeiramente calcular a entropia de todo o conjunto de dados. Depois disso, a cada nó sendo analisado, é calculado naquele ponto a taxa de ganho de informação para cada atributo do conjunto de dados e o que apresentar a maior taxa de ganho é selecionado e colocado no nó de decisão (HSSINA et al., 2014).

Outra característica do C4.5 é usar uma técnica chamada de *pruning* (HSSINA et al., 2014), que pode ser traduzida como poda. Assim como o nome sugere, essa técnica atua podando galhos da árvore de decisão, diminuindo o tamanho delas e removendo nós que agregam pouco para a tarefa de classificação em questão. No caso do C4.5, a poda é feita de baixo para cima, o que quer dizer que ela parte das folhas e vai subindo até chegar na raiz da árvore.

Com a poda, o C4.5 acaba evitando problemas de sobre-ajuste da árvore de decisão gerada aos conjunto de dados de treinamento, o que se torna uma vantagem desse algoritmo. Outra vantagem é que ele tende a ser rápido e preciso na classificação/regressão dentro da categoria de algoritmos de aprendizado de máquina *main-memory* (RUGGIERI, 2002).

2.4 Métricas para a Medição do Desempenho de Classificadores

Para medir o desempenho dos algoritmos de classificação, existem diversas métricas disponíveis. As que foram usadas neste projeto estão listadas a seguir. Cabe destacar que, em todas as métricas listadas, quanto mais próximo de 100% a medição feita for, melhor terá sido o desempenho do algoritmo naquela métrica.

2.4.1 Acurácia

Esta medida fornece a proporção de acertos nas classificações feitas, independentemente de terem sido de recidiva ou não. Ela é obtida através da fórmula:

$$acurácia = \frac{CC}{TI} \times 100$$

onde CC é o total de classificações corretas e TI o total de instâncias no conjunto.

2.4.2 Sensitividade

Esta medida fornece a proporção de casos reais de recidiva classificados corretamente pelo algoritmo. Ela é obtida através da fórmula:

$$sensitividade = \frac{CCR}{TIR} \times 100$$

onde CCR é o total de classificações corretas de recidiva feitas pelo algoritmo e TIR o total de instâncias que rotuladas como recidiva.

2.4.3 Especificidade

Esta medida fornece a proporção de casos reais de não recidiva classificados corretamente pelo algoritmo. Ela é obtida através da fórmula:

$$especificidade = \frac{CCNR}{TINR} \times 100$$

onde $CCNR$ é o total de classificações corretas de não recidiva feitas pelo algoritmo e $TINR$ o total de instâncias rotuladas como não recidiva.

2.4.4 Valor de Predição Positiva (VPP)

Esta medida fornece, dentre todas as classificações de recidiva feitas pelo algoritmo, com que proporção elas foram corretas. Ela é obtida através da fórmula:

$$VPP = \frac{CCR}{TIR + CER} \times 100$$

onde CCR é o total de classificações corretas de recidiva feitas pelo algoritmo, TIR o total de instâncias rotuladas como recidiva e CER o total de classificações erradas de recidiva feitas pelo algoritmo.

2.4.5 Valor de Predição Negativa (VPN)

Essa medida fornece, dentre todas as classificações de não recidiva feitas pelo algoritmo, com que proporção elas foram corretas. Ela é obtida através da fórmula:

$$VPN = \frac{CCNR}{TINR + CENR} \times 100$$

onde $CCNR$ é o total de classificações corretas de não recidiva feitas pelo algoritmo, $TINR$ o total de instâncias rotuladas como não recidiva e $CENR$ o total de classificações erradas de não recidiva feitas pelo algoritmo.

3 Materiais e Métodos

Este é um trabalho com perfil exploratório e quantitativo que utiliza diferentes técnicas computacionais com o objetivo de explorar um problema, e assim fornecer informações para uma investigação mais precisa. Este trabalho se concentra na descoberta de ideias e pensamentos.

Neste capítulo está descrita a metodologia que foi usada para o desenvolvimento do trabalho. Ela está organizada de acordo com a ordem das atividades realizadas, que foram:

- Levantamento bibliográfico de trabalhos correlatos;
- Obtenção e preparo de dados laboratoriais de pacientes com LLA B que tiveram (ou não) recidiva durante seu tratamento para utilização no treinamento e teste dos algoritmos de aprendizado de máquinas;
- Planejamento para a execução de testes experimentais.

3.1 Levantamento Bibliográfico

Na literatura, alguns trabalhos já procuraram usar o aprendizado de máquina para prever recidivas de pacientes com LLA. O trabalho de Pan et al. (2017), por exemplo, procurou prever recidivas a partir de dados médicos de crianças e adolescentes disponíveis via prontuário eletrônico. No total foram usados dados de 570 pacientes com LLA, com idades entre 0 (menos de 1 ano de vida) até 15 anos e divididos entre 367 meninos e 203 meninas. Entre os 4 algoritmos de aprendizado de máquina testados, o que apresentou o melhor desempenho em prever recidivas foi o *Random Forest*. Ele teve uma acurácia de 82,70% $\pm$ 0,031%, sensitividade de 75,60% $\pm$ 0,051%, especificidade de 89,7% $\pm$ 0,041%, valor de predição positiva de 88,2% $\pm$ 0,040% e valor de predição negativa de 90,20% $\pm$ 0,027%.

Good et al. (2018) realizaram um estudo baseado em células únicas em casos de LLA B usando citometria massiva. Eles buscaram identificar estados dependentes do desenvolvimento da célula que são unicamente associados com recidivas. Para isso aplicaram citometria massiva tanto em células B de pessoas com LLA B (60 pacientes pediátricos dos EUA e da Itália com LLA B) quanto de pessoas saudáveis (5 adultos, sendo 3 mulheres e 2 homens, com idades entre 20 e 44 anos). O aprendizado de máquina aplicado neste estudo, usando a abordagem *elastic net*, permitiu ser construído um modelo preditivo de recidivas a partir de características

proteômicas extraídas de populações de células que tiveram expansão na transição do estágio pré pró-B para o pró-BI. Esse modelo atingiu um AUC dependente do tempo de 0,851 na etapa de teste do desempenho. Além disso, o uso de aprendizado de máquina também permitiu a eles identificarem 6 características celulares relacionadas à predição de recidiva.

Por fim, Fuse et al. (2019) fizeram um estudo para prever a recidiva de pacientes que receberam transplante de medula óssea alogênico depois de 1 ano deste procedimento, usando para isso o algoritmo de aprendizado de máquina ADTree. Para isso eles usaram dados de 217 pessoas com idade entre 10 e 67 anos e diagnósticos de leucemia mielóide aguda (135) e linfoblástica aguda (82), sendo 111 homens e 106 mulheres. Os resultados obtidos durante o treinamento para acurácia, AUC e taxa de falsos positivos foram 78%, 0,746 e 0,508 respectivamente. Quando aplicaram o modelo em um conjunto de dados de validação, naqueles mesmos indicadores usados durante o treinamento eles obtiveram os valores de 71%, 0,667 e 0,216, respectivamente.

3.2 Geração da Base de Dados

Este trabalho se iniciou com a obtenção dos dados que foram usados para o treinamento dos algoritmos de aprendizado de máquina. Esses dados foram obtidos de prontuários de crianças e adolescentes que são ou foram atendidos no Hospital da Criança de Brasília, cujo diagnóstico tenha sido LLA B.

O Hospital da Criança de Brasília é uma instituição pública de saúde, focada em especialidades pediátricas (BRASíLIA, 201-a), sendo uma referência no tratamento de câncer em pacientes pediátricos no Distrito Federal (DF). Ele foi inaugurado em 23 de novembro de 2011, sendo resultado de uma parceria entre a Associação Brasileira de Assistência às Famílias de Crianças Portadoras de Câncer e Hemopatias (Abrace) e o governo do DF (BRASíLIA, 201-b).

Para ter acesso aos dados dos pacientes, primeiramente foi realizado contato com a diretoria de Ensino e Pesquisa do do HCB. A partir do momento que as partes estavam de acordo com a realização da pesquisa dentro da instituição, foi submetido um projeto de pesquisa para análise e aprovação por um Comitê de Ética em Pesquisa (CEP). Essa necessidade de ter aprovação de um CEP se deve ao fato que os dados a serem acessados são provenientes de prontuários médicos de pacientes. Conforme a Carta Circular nº 39 do Comissão Nacional de Ética em Pesquisa (PESQUISA, 2011), pesquisas que envolvam o acesso a esse tipo de informação devem ser aprovadas por um CEP. Deste modo, este projeto foi submetido à Plataforma Brasil, e em março de 2020 ele foi aprovado pelo CEP da Faculdade de Medicina da Universidade de

Brasília.

Após a aprovação pelo CEP, os dados clínicos dos pacientes foram obtidos junto ao Laboratório de Pesquisa Translacional do HCB. Esses dados serão descritos a seguir, mas cabe ressaltar que foram tomadas medidas para proteger a privacidade dos pacientes. Informações pessoais, como nome de paciente e cidade onde mora, foram removidos dos dados originais. Como medida adicional, datas de nascimento e data do diagnóstico tiveram o dia removido, restando apenas o mês e ano.

3.2.1 Tamanho e Perfil da Amostra Coletada

A coleta de dados foi realizada segundo uma abordagem não-probabilística com amostragem por conveniência, ou seja, a seleção dos pacientes foi definida a partir da disposição de acesso aos dados de seus prontuários.

Entre 2011 e 2018, houveram 306 pacientes com LLA sendo tratados pelo HCB. Inicialmente foram recebidos dados de 266 pacientes. Porém, alguns pacientes foram retirados da amostra por não terem informações no prontuário sobre o tratamento feito, pois estavam sendo tratados em outra instituição de saúde e foram ao HCB apenas para tomar algum medicamento. Os casos que não são LLA B também foram retirados devido à baixa quantidade de pacientes. Por fim, foram removidos dados de pacientes que vieram a óbito sem terem recidido, por não terem tido remissão e ao mesmo tempo não terem sobrevivido. Também foram removido dados de pacientes com Síndrome de Down, por estes costumarem ter peculiaridades com relação ao desenvolvimento da doença comparado aos pacientes sem a Síndrome, e ao mesmo tempo serem poucos casos nessa situação. Sendo assim, das 266 instâncias iniciais, serão usado apenas 239, sendo 47 dados de pacientes que tiveram recidivas e 182 que não tiveram recidivas no tratamento.

Para estimar o poder estatístico do tamanho da amostra (n = 229 pacientes), foi utilizada a seguinte fórmula para calcular o erro amostral E:

$$E = \sqrt{\frac{(Z_{a/2}.\delta)^2}{n}} \tag{3.1}$$

em que, n é o tamanho da amostra, $Z_{a/2}$ é o valor crítico para o grau de confiança desejado, usualmente: 1,645 (90%); δ é o desvio padrão populacional da variável. Como este valor é desconhecido para o universo da pesquisa, utiliza-se o valor padrão de δ = 0,5.

Para 90% de confiança a margem de erro da amostra de n = 229 pacientes é de 5,44%. Para 95% de confiança, a margem de erro da amostra n = 229 pacientes é de 6,48%. Para 99% de confiança a margem de erro é de 8,51%.

Muitos trabalhos correlatos utilizaram uma amostra bem menor. Por exemplo, o trabalho de Leite et al. (2007) utilizou uma amostra de 108 pacientes de idade até 18 anos para analisar os fatores prognósticos em crianças e adolescentes com LLA. O trabalho de Souza, Viana e Oliveira (2008) utilizou dados de 95 pacientes com recidiva da LLA tratados no Hospital das Clínicas da UFMG, entre 1988 e 2005, para analisar a evolução de crianças com primeira recidiva da leucemia linfoblástica aguda e identificar fatores prognósticos para segunda recidiva ou óbito. Considerando trabalhos internacionais, o trabalho de Pan et al. (2017) utilizou dados de 570 pacientes, que foi a maior amostra encontrada.

3.2.2 Conjunto de Dados Coletados

A Tabela 1 mostra os atributos do conjunto de dados laboratoriais e clínicos coletados de prontuários de pacientes do HCB. Cada um desses atributos será brevemente discutido nas subseções a seguir.

Tabela 1 – Atributos coletados dos prontuários de pacientes do HCB.

Tipo de Informação	Atributo	Tipo de Dado	Valores	Ocorrência na Base
Clínica	Idade ao Ser Diagnosticado	Numérico	0 - 16	100%
	Recidiva	Booleano	TRUE	47 (20,52%)
			FALSE	182 (79,48%)
	Risco	Categórico	ALTO	113 (49,34%)
			MÉDIO	40 (17,47%)
			BAIXO	63 (27,51%)
			?	13 (5,68%)
Laboratorial	Biologia Molecular Negativo	Booleano	TRUE	97 (42,36%)
			FALSE	85 (37,12%)
			?	47 (20,52%)
	Biologia Molecular T(XX;XX)	Booleano	TRUE	63 (27,51%)
			FALSE	119 (51,97%)
			?	47 (20,52%)
	Biologia Molecular T(12;21)	Booleano	TRUE	39 (17,03%)
			FALSE	143 (62,45%)
			?	47 (20,52%)
	Leucometria	Numérico	300 - 594.000	219 (96 %)
			?	10 (4%)
	SNC	Booleano	POSITIVO	20 (8,73%)
			NEGATIVO	203 (88,65%)
			?	6 (2,62%)
	Celularidade nos MO D15	Categórico	HIPOCELULAR	117 (51,09%)
			NORMOCELULAR	31 (13,54%)
			HIPERCELULAR	11 (4,80%)
			?	70 (30,57%)
	Celularidade nos MO D33	Categórico	HIPOCELULAR	49 (21,40%)
			NORMOCELULAR	72 (31,44%)
			HIPERCELULAR	10 (4,37%)
			?	98 (42,79%)
	Blastos nos MO D15	Categórico	M1	165 (72,50%)
			M2	27 (11,79%)
			M3	11 (4,80%)
			?	26 (11,35%)
	Blastos nos MO D33	Categórico	M1	196 (85,59%)
			M2	6 (2,62%)
			M3	0 (0%)
			?	27 (11,79%)

Fonte: Elaborado pelo autor.

3.2.2.1 Idade ao Ser Diagnosticado

Esse atributo é um valor numérico e discreto que se refere à idade (em anos completos) que o paciente tinha quando foi diagnosticado com LLA B. A Figura 3 mostra o gráfico da distribuição de idade dos pacientes ao serem diagnosticados. Esse valor varia de 0 até 16 anos, sendo a média da idade do conjunto de dados de 5,76 anos.

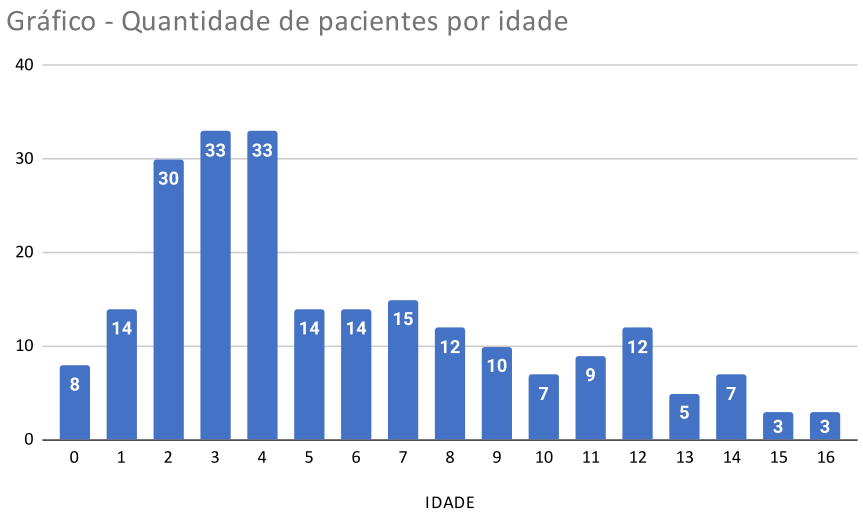

Figura 3 – Gráfico mostrando a quantidade de pacientes por idade (em anos completos) que tinham ao serem diagnosticados. Fonte: Elaborado pelo autor.

3.2.2.2 Recidiva

Esse atributo foi obtido a partir da existência ou não de uma data de recidiva do paciente no conjunto de dados. Se a data estava presente, foi considerado que o paciente recidiu, e o contrário caso a data não tenha sido fornecida.

3.2.2.3 Risco

Esse atributo se refere à classificação de risco do paciente. Os critérios para a classificação de risco de um paciente são determinados pelo protocolo de tratamento usado no caso.

O HCB usou prioritariamente o protocolo GBTLI-93 de 2011 a 2014, passando a adotar como padrão a partir de 2015 o BFM 95. O protocolo GBTLI-93 tem as categorias de baixo e alto risco de recidiva, enquanto que o BFM 95 possui as categorias baixo, médio e alto risco de recidiva.

3.2.2.4 Biologia Molecular

Essa característica de dados categóricos é o resultado do exame de biologia molecular dos pacientes. Como esse atributo é no formato de *strings* (ou seja, formado por conjuntos de caracteres, que podem ou não ter algum significado), optou-se por dividir elas em *substrings* (que são subdivisões da *string*, normalmente seguindo algum critério), para que a presença ou não dessas se tornasse um atributo separado disponibilizado para os algoritmos processarem. No caso do presente trabalho, buscou-se escolher *substrings* que eram recorrentes nos resultados da Biologia Molecular.

Após a criação de cada *substring*, era checado em cada instância se ela possuía o atributo Biologia Molecular e, caso tivesse, era analisado se a *substring* em questão estava presente. Por fim, foi analisada a quantidade de ocorrências delas para ver se eram de pelo menos 24 (10% do total de 239 instâncias). As que ficaram abaixo de 24 ocorrências foram descartadas, o resultando apenas nas *substrings* descritas abaixo:

- Biologia Molecular - “NEGATIVO”: esse atributo indica se o resultado da biologia molecular apresentou o termo “NEGATIVO”;

- Biologia Molecular - “T(XX;XX)”: esse atributo indica se o resultado da biologia molecular apresentou o padrão regex “T\([0-9]?[0-9];[0-9][0-9]\)”. Exemplos de casos reais que se enquadram dentro desse padrão são T(1;19), T(12;21) e T(12;33);

- Biologia Molecular - “T(12;21)”: esse atributo indica se o resultado da biologia molecular apresentou o termo “T(12;21)”.

3.2.2.5 Leucometria

Esse atributo se refere à quantidade de leucócitos encontrados na amostra usada para o exame de leucometria. A Figura 4 contém o gráfico demonstrando a dispersão dos valores de leucometria dos dados coletados. O menor valor encontrado foi de 300 e o maior de 594.000. A média foi de 36.716,93.

3.2.2.6 SNC

Essa característica de dados categóricos informa se houve ou não infiltração de células neoplásicas no sistema nervoso central.

Figura 4 – Gráfico da dispersão da quantidade de leucócitos nos exames de leucometria. Fonte: Elaborado pelo autor.

3.2.2.7 Celularidade nos MO D15 e MO D33

Essa característica de dados categóricos informa qual foi a celularidade encontrada no exame de mielograma. O caso MO D15 se refere ao dado obtido no mielograma feito 15 dias após o início do protocolo de tratamento, e o MO D33 ao obtido 33 dias depois deste início.

Cabe ressaltar que, para o uso desse dado, em casos como hipo/normocelular e normo/hipercelular, o dado foi convertido para normocelular e hipercelular, respectivamente.

3.2.2.8 Porcentagem de Blastos nos MO D15 e MO D33

Essa característica de dados numéricos contínuos informa a porcentagem de blastos no exame de mielograma. O caso MO D15 se refere ao dado obtido no mielograma feito 15 dias após o início do protocolo de tratamento, e o MO D33 ao obtido 33 dias depois deste início.

3.3 Planejamento da Execução dos Testes Experimentais

Esta seção apresenta o planejamento da execução dos testes experimentais, desde a fase de pré-processamento dos dados coletados, como o balanceamento dos dados, até a definição de métricas para a avaliação de desempenho dos algoritmos de classificação.

3.3.1 Balanceamento da Base de Dados

A quantidade de casos de não-recidivas é bem menor que a quantidade de casos com recidivas na base de dados. Tal desbalanceamento pode afetar o desempenho dos algoritmos de classificação. Por isso, para ajustar a distribuição de casos com e sem recidiva no conjunto de dados, foi utilizada a técnica SMOTE para a geração artificial de instâncias de casos com recidiva (classe minoritária), o que é conhecido como *oversampling*.

A técnica chamada SMOTE (*Synthetic Minority Oversampling Technique*), proposta por (CHAWLA et al., 2002), funciona gerando instâncias artificiais da classe que está sub-representada no conjunto de dados, que no caso da pesquisa é a de pacientes com recidiva. Para fazer isso, o algoritmo da SMOTE usa o K-NN. Para cada instância da classe minoritária (pacientes com recidiva, no caso do presente trabalho) são selecionados os 5 vizinhos mais próximos da mesma classe e, a partir deles, é gerado uma nova instância que se encontra no "espaço de atributos" entre o ponto de dado referência e os vizinhos.

Para o presente trabalho, as instâncias de recidiva foram aumentadas em 134 novos dados usando a técnica SMOTE, indo de 47 para 181 e totalizando 363 instâncias no total do conjunto de dados, que foram processadas pelos algoritmos.

3.3.2 Treinamento e Teste Usando Validação Cruzada

Para treinar e validar o desempenho de cada classificador, foi usada a validação cruzada usando o método *k-fold*, com k nesse caso assumindo o valor 10. Usando esse método, o conjunto de dados foi dividido automática e aleatoriamente em 10 subconjuntos de tamanhos próximos, sem que houvesse repetição de instâncias entre os agrupamentos formados. Após a divisão, cada algoritmo foi treinado usando 9 dos subconjuntos enquanto o último – que não foi utilizado para treino – foi usado para testar o desempenho do resultado do treinamento do algoritmo. Esse processo foi repetido 10 vezes, para que todos os subconjuntos pudessem ser usados para testar o algoritmo.

3.3.3 Uso de Penalizações para Testes com a Base Desbalanceada

Em testes iniciais com a base desbalanceada, ao se analisar as matrizes de confusão dos algoritmos de classificação notou-se que os algoritmos estavam tendendo a classificar corretamente casos sem recidiva, mas os casos com recidiva estavam em sua maioria sendo classificados erroneamente como sem.

A Figura 5 contém uma dessas matrizes de confusão mencionadas acima. A letra a representa casos sem recidiva, e a b casos com recidiva. A matriz de confusão

está organizada da seguinte maneira:

- a primeira linha traz os números de casos de não recidiva e, na somatória das 2 colunas que fazem parte dela, o total de casos de não recidiva no conjunto de dados;
- a segunda linha traz os números de casos de recidiva e, na somatória das 2 colunas que fazem parte dela, o total de casos de recidiva no conjunto de dados;
- a primeira coluna traz os números de classificações de não recidiva feitas pelo algoritmo e, na somatória das 2 linhas que fazem parte dela, o total de classificações de não recidiva feitas pelo algoritmo;
- a segunda coluna traz os números de classificações de recidiva feitas pelo algoritmo e, na somatória das 2 linhas que fazem parte dela, o total de classificações de recidiva feitas pelo algoritmo.

Enquanto que na primeira linha 92,82% das classificações foram corretas (classificações na linha a e coluna a da matriz), na segunda apenas 34,04% dos casos de recidiva foram classificados corretamente como tal (classificações na linha b e coluna b da matriz).

```
=== Confusion Matrix ===

   a   b   <-- classified as
 169  13 |   a = FALSE
  31  16 |   b = TRUE
```

Figura 5 – Matriz de confusão gerada pelo algoritmo C4.5 sem o uso de penalizações para erros. Fonte: Elaborado pelo autor.

Para reduzir isso, foi usada a penalização por erros de classificação dos algoritmos, onde classificar erroneamente um caso de recidiva teve maior peso, enquanto classificações erradas de casos de não recidiva tiveram o peso base. A Figura 6 mostra a mudança gerada na matriz de confusão do algoritmo da Figura 5 ao aplicar a penalização.

Ao configurar os pesos dos erros dos algoritmos, a preocupação foi alcançar um balanço entre uma melhor classificação de casos com recidiva sem ao mesmo tempo aumentar excessivamente a classificação errônea dos casos sem recidiva, para não gerar uma distorção do desempenho para o outro lado. Desse modo, a cada algoritmo, o treinamento era realizado mais de uma vez, alterando os pesos

```
=== Confusion Matrix ===

   a   b   <-- classified as
 160  22 |   a = FALSE
  23  24 |   b = TRUE
```

Figura 6 – Matriz de confusão gerada pelo algoritmo C4.5 com o uso de penalizações para erros. Fonte: Elaborado pelo autor.

até encontrar um valor que ao mesmo tempo aumentasse o desempenho deles em classificar recidivas e não diminuísse demais a acurácia geral deles.

Cabe ressaltar aqui que, independente do peso colocado para punir erros, o algoritmo K-NN foi o único que não teve seu desempenho alterado por esse mecanismo das penalizações, com ele servindo como uma referência de como os algoritmos tendiam a desempenhar sem a punição por erro de classificação e sem o balanceamento usando o SMOTE.

4 Resultados e Discussões

Neste capítulo estão apresentados os resultados experimentais realizados usando os materiais e métodos definidos no Capítulo 3. Em suma, dois objetivos nortearam a realização dos testes experimentais:

1. Determinar o algoritmo de classificação com o melhor desempenho na predição das recidivas;
2. Determinar a importância dos atributos na predição dos casos de recidivas.

Os resultados e discussões que serão apresentados a seguir estão organizados nesses dois objetivos.

4.1 Desempenho dos Algoritmos de Classificação

A base de dados disponível para o treinamento e teste dos algoritmos de classificação é desbalanceada. Ou seja, existem mais casos de não-recidivas (classe majoritária) do que casos de recidivas (classe minoritária). Isso pode comprometer o desempenho dos algoritmos de classificação, visto que eles poderão classificar corretamente mais casos da classe majoritária do que da classe minoritária.

Por esse motivo, a análise do desempenho dos algoritmos de classificação foi realizada em duas fases:

- Fase 1: sem balanceamento da base de dados
- Fase 2: com balanceamento da base de dados

Cabe ressaltar que, para a realização da Fase 1, optou-se por penalizar os classificadores quando estes classificavam erroneamente casos de recidivas (classe minoritária), tal como detalhado no Capitulo anterior.

A seguir são apresentados os resultados obtidos em cada uma dessas fases.

4.1.1 Fase 1: sem balanceamento da base

A Tabela 2 traz o desempenho dos algoritmos de acordo com as métricas de desempenho descritas na Seção 2.4. Os valores em negrito em cada coluna são das medições que obtiveram o melhor valor naquela métrica.

Tabela 2 – Desempenho dos algoritmos de classificação usando a base desbalanceada e com penalização de erros na classificação da classe minoritária

Algoritmo	Matriz de confusão	Acurácia	Sensitividade	Especificidade	VPP	VPN
K-NN	153 029 035 012	72,05%	25,53%	84,07%	15,79%	70,51%
Naive Bayes	157 025 022 025	79,48%	53,19%	86,26%	34,72%	76,96%
Random Forest	152 030 024 023	76,42%	48,94%	83,52%	29,87%	73,79%
C4.5	160 022 023 024	**80,35%**	51,06%	**87,91%**	**34,78%**	**78,05%**
Regressão Logística Multivalorada	143 039 021 026	73,80%	**55,32%**	78,57%	30,23%	70,44%
Multi-layer Perceptron	146 036 031 016	70,74%	34,04%	80,22%	19,28%	68,54%
OSM	156 026 029 018	75,98%	38,30%	85,71%	24,66%	73,93%

Fonte: Elaborado pelo autor.

No geral, quando usada a base desbalanceada e a técnica de penalizar erros na classificação de recidivas, os algoritmos apresentaram uma boa acurácia e capacidade em prever casos em que não houve recidivas, mas não apresentaram desempenho superior quanto à previsão de recidivas. De fato, olhando para as matrizes de confusão, pode-se ver a clara diferença entre o desempenho dos classificadores em cada uma dessas classes.

A partir dos dados da Tabela 2, notou-se que todos os classificadores tiveram uma acurácia média de 75,55%. O algoritmo C4.5 apresentou o melhor desempenho com uma acurácia de 80,35% , uma sensitividade de 51,06% e uma especificidade de 87,91%. Analisando a métrica VPP e o VPN do algoritmo C4.5, os valores de ambas foram de 34,78% e 78,05% respectivamente, e uma diferença entre elas de 43,27%. Desse modo, fica clara a tendência dos algoritmos em não desempenharem bem a classificação de casos reais de recidivas, que é a classe minoritária.

De fato, um fator que influenciou na diferença de desempenho foi a distribuição entre casos de recidivas e não recidivas no conjunto de dados. Como mostra a Tabela 1, a amostra tem uma proporção de quase 1 recidiva para cada 5 não recidivas no conjunto de dados, uma proporção semelhante à encontrada no trabalho de (PAN et al., 2017). Associado a isso, o conjunto de dados original teve apenas 239 instâncias, com apenas 47 instâncias de pacientes com recidiva. Então, mesmo usando a penalização de erros de classificação, claramente a capacidade dos modelos

gerados pelos algoritmos foi prejudicada para a classificação de casos de recidiva.

4.1.2 Fase 2: com balanceamento da base

A Tabela 3 mostra o desempenho dos algoritmos de classificação considerando a base de dados balanceada usando a SMOTE, ou seja, com a quase mesma quantidade de casos de recidivas e não-recividas (181 e 182 respectivamente). Os valores em negrito em cada coluna são aqueles que da melhor medição de desempenho obtida naquela métrica.

Tabela 3 – Desempenho dos algoritmos de classificação usando a base balanceada pela técnica SMOTE

Algoritmo	Matriz de confusão	Acurácia	Sensitividade	Especificidade	VPP	VPN
K-NN	134 048 040 141	75,76%	77,90%	73,63%	61,57%	60,36%
Naive Bayes	148 034 041 140	79,34%	77,35%	81,32%	65,12%	66,37%
Random Forest	161 021 035 146	**84,57%**	80,66%	**88,46%**	**72,28%**	**74,19%**
C4.5	146 036 035 146	80,44%	80,66%	80,22%	67,28%	67,28%
Regressão Logística Multivalorada	143 039 036 145	79,34%	80,11%	78,57%	65,91%	65,60%
Multi-layer Perceptron	148 034 034 147	81,27%	81,22%	81,32%	68,37%	68,52%
OSM	133 049 033 148	77,41%	**81,77%**	73,08%	64,35%	61,86%

Fonte: Elaborado pelo autor.

Pode-se observar que, de fato, o uso da técnica SMOTE permitiu mitigar o problema da distribuição desbalanceada dos casos de recidiva e não-recidiva e aumentar o poder de predição dos classificadores. O desempenho dos algoritmos, apresentado na Tabela 3, em termos de acurácia, foi superior a acurácia dos classificadores na Fase 1 dos experimentos. Enquanto na Fase 1 classificação teve acurácia média de 75,55%, a acurácia com o uso da técnica SMOTE para balanceamento dos dados foi de 79,73%. Nesta Fase 2, o classificador com melhor acurácia (84,57%) foi o *Random Forest*.

A melhora dos valores de sensitividade e VPP, duas métricas relacionadas diretamente à capacidade dos algoritmos classificarem corretamente casos de recidiva, foi consideravelmente superior na Fase 2, quando comparado à Fase 1 dos experimentos. Enquanto a média para a sensitividade e a VPP com o uso

da SMOTE para balanceamento dos casos de recidiva ficou em 79,95% e 66,41% respectivamente, na penalização de erros de classificação essas médias ficaram em 43,77% e 27,05% respectivamente. As melhores medições para essas métricas no uso da SMOTE foram sensitividade de 81,77% (SMO) e VPP de 72,28% (*Random Forest*), muito acima de 55,32% (Regressão Logística Multivalorada) e 34,78% (C4.5) obtidos com a penalização de erros.

4.2 Atributos Importantes para a Predição de Recidivas

Para determinar os principais atributos preditores de recidiva, foi escolhido usar a análise de uma árvore de decisão gerada pelo algoritmo C4.5 devido à possibilidade de compreender não apenas como é o processo de classificação usado por ela, mas também pela possibilidade de determinar facilmente quais atributos são usados no processo de classificação. Além disso, o C4.5 apresentou excelente desempenho tanto na Fase 1 quanto na Fase 2 dos experimentos, o que reforça que a árvore de decisão produzida pelo algoritmo é um excelente ativo para avaliar a importância dos atributos na predição de casos de recidiva.

Para fins de comparação, a análise através da árvore de decisão foi realizada considerando duas situações:

- Situação 1) com o atributo Risco
- Situação 2) sem o atributo Risco

Essa diferenciação nas análises se deve ao fato de que o atributo Risco é determinado, entre outras coisas, da combinação de alguns dos atributos presentes no conjunto de dados, como a leucometria e a idade ao ser diagnosticado.

Em ambas as situações, a árvore de decisão utilizada foi a gerada pelo algoritmo C4.5 após ser treinado usando o conjunto de dados balanceado artificialmente usando a técnica SMOTE. Cabe ressaltar que nas árvores geradas e mencionadas mais abaixo, os nós TRUE dentro dos retângulos cinzas significam que a classificação é recidiva enquanto que FALSE é não recidiva.

Nesses mesmos retângulos, o primeiro número que aparece dentro do parênteses se refere ao total ponderado de instâncias que chegaram ali, enquanto o segundo (quando existe) após a barra (/) se refere ao total ponderado de instâncias que acabaram naquele retângulo da árvore mas tinham uma outra classificação quanto à recidiva.

A seguir são apresentadas as análises considerando essas duas situações.

4.2.1 Árvore de Decisão com o Atributo Risco

A árvore gerada considerando o atributo Risco pode ser vista no Apêndice A. A Tabela 4 mostra quantas vezes cada atributo apareceu na árvore de decisão gerada pelo algoritmo C4.5 usando a SMOTE.

Tabela 4 – Atributos presentes na árvore de decisão gerada pelo algoritmo do C4.5 e considerando o atributo Risco na base de dados

Atributo	Ocorrências
Leucometria	3
Bio. Mol. - “Negativo”	2
Celularidade no MO D33	2
Porcent. de Blastos no MO D15	2
Bio. Mol. - “T(XX;XX)”	1
Celularidade no MO D15	1
Risco	1
SNC	1
Bio. Mol. - “T(12;21)”	0
Idade ao ser diagnosticado	0
Porcent. de Blastos no MO D33	0

Fonte: Elaborado pelo autor.

Os atributos que tiveram 0 ocorrências na árvore de decisão final podem ser definidos como aqueles de menor importância preditiva para o algoritmo. Afinal, se no processamento dos dados do classificador C4.5 os dados desses atributos tivessem sido determinados como sendo relevantes, eles teriam sido incluídos em algum ponto da árvore final gerada.

Uma segunda análise que pode ser feita para determinar os atributos de maior poder preditivo é estudar os caminhos percorridos indo das folhas (retângulos) com maior total ponderado de instâncias e com baixas quantidades ponderadas de erros de classificação até a raiz (atributo no topo da árvore). Fazendo isso pode-se ter ideia das melhores combinações entre atributos e seus valores para realizar classificações.

Com o Risco estando entre os atributos, observa-se que o Risco “ALTO” sempre esteve no início dos caminhos analisados que levaram às folhas com maiores totais ponderados de classificações bem sucedidas de recidivas. Por outro lado, vendo a árvore do Apêndice A, pode-se ver que tanto a classificação de Risco “BAIXO” quanto “MÉDIO” tiveram sua relevância para o algoritmo C4.5 (SMOTE). O Risco “BAIXO” em conjunto com uma Leucometria menor ou igual a 17.500 leva à folha com o maior total ponderado de classificações de não recidiva bem-sucedidas, enquanto o Risco “MÉDIO” sozinho leva à segunda folha com melhor desempenho na classificação de não recidivas. Desse modo, dentro do aprendizado o algoritmo considerou que, em conjunto com outros atributos (ou sozinho no caso do valor de Risco “MÉDIO”), as

classificações de risco segundo os protocolos de tratamento têm uma uma relevância como preditores de recidiva e não recidiva.

Focando apenas nos caminhos dentro da árvore que levaram a mais classificações bem-sucedidas de recidivas, na árvore do conjunto de dados que teve o atributo Risco a combinação de atributos Risco “ALTO”, Celularidade no MO D33 “HIPERCELULAR” e Biologia Molecular - “NEGATIVO” FALSO foi o percurso percorrido por 80,73 instâncias, com 7,54 desses não sendo casos de recidiva. O segundo caminho que mais levou à classificações de recidivas foi a combinação de Risco "ALTO", Celularidade no MO D33 “NORMOCELULAR”, Biologia Molecular - “NEGATIVO” FALSO e Biologia Molecular - “T(XX;XX)” FALSO, com 39,05 classificações de recidiva com 7,18 dessas estando erradas.

4.2.2 Árvore de Decisão sem o Atributo Risco

Para fins de comparação, foi gerada uma segunda árvore de decisão do algoritmo C4.5 (SMOTE), mas dessa vez com o conjunto de dados sem o atributo Risco. Esta árvore está no Apêndice B.

A Tabela 5 mostra o número de ocorrências dos atributos nessa nova árvore que não considera o atributo Risco. Nota-se que a Leucometria é de longe o atributo que apareceu mais vezes na árvore.

Tabela 5 – Atributos presentes na árvore de decisão gerada pelo algoritmo C4.5 sem o atributo Risco na base de dados

Atributo	Ocorrências
Leucometria	6
Bio. Mol. - “Negativo”	2
Bio. Mol. - “T(XX;XX)”	1
Celularidade no MO D33	1
Porcent. de Blastos no MO D15	1
Bio. Mol. - “T(12;21)”	0
Celularidade no MO D15	0
Idade ao ser diagnosticado	0
Porcent. de Blastos no MO D33	0
SNC	0

Fonte: Elaborado pelo autor.

Analisando os caminhos da segunda árvore de decisão (sem o atributo Risco) que levam às classificações de recidivas, o que possui o maior total ponderado de instâncias é o Celularidade no MO D3 “HIPERCELULAR”, Biologia Molecular - “NEGATIVO” FALSO e Leucometria acima de 10.044,82, com 84,1 classificações feitas, sendo 8,27 erradas. O segundo caminho com o maior total ponderado foi o Celularidade no MO D3 “NORMOCELULAR”, Leucometria acima de 3.800, Biologia

Molecular - “NEGATIVO” FALSO e Biologia Molecular - “T(XX;XX)” FALSO, tendo 42,31 instâncias que chegaram até essa folha da árvore e 9,36 deles não tinham classificação de recidiva.

Juntando as informações obtidas dos caminhos com maior totais ponderados de classificações de recidiva bem sucedidas das árvores na Situação 1 e na Situação 2, pode-se notar algumas semelhanças. As Celularidades no MO D33 “HIPERCELULAR” e “NORMOCELULAR”estiveram presentes em todos os 4 caminhos analisados (2 aparições de cada valor) e sempre tendo o atributo Biologia Molecular - “NEGATIVO” abaixo dele com o valor FALSO. Além disso, nos caminhos envolvendo a Celularidade no MO D33 “NORMOCELULAR” a Biologia Molecular - “NEGATIVO” FALSO sempre esteve presente também e seguida da Biologia Molecular - “T(XX;XX)” FALSO.

5 Considerações Finais

A partir dos resultados obtidos neste trabalho foi possível identificar o potencial que os algoritmos baseados em aprendizado de máquina possuem como instrumentos no auxílio às equipes médicas que lidam com pacientes pediátricos com LLA-B. Potencial esse que está tanto no desempenho que os algoritmos podem ter na tarefa de classificar casos como sendo recidiva ou não, quanto na questão de identificar potenciais variáveis preditoras de recidivas e como elas podem se relacionar entre si dentro de uma árvore de decisão.

Buscando melhorar ainda mais os resultados obtidos, deve-se determinar como fazer o pré-processamento da *string* por exames de citogenética. Os resultados desse exame têm sido usados por equipes médicas para buscar prever recidivas, tais como o trabalho de (HARRISON; FORONI, 2002), e por isso possibilitar o acesso dele para os algoritmos pode trazer uma melhoria ainda maior na capacidade desses classificarem corretamente casos de recidiva.

Por fim, espera-se que esse trabalho possa contribuir para, de alguma maneira, ajudar equipes médicas e profissionais a, cada vez mais, aumentarem o sucesso do tratamento do câncer e trazer um futuro feliz para os pacientes e seus familiares.

Referências

ALLEMANI, C. et al. Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (concord-2). *The Lancet*, Elsevier, v. 385, n. 9972, p. 977–1010, 2015. Acesso em: 12, out. 2020. Citado na página 19.

APOSTOLOPOULOS, I. D.; MPESIANA, T. A. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, Springer, p. 1, 2020. Acesso em: 31, out. 2020. Citado na página 23.

AUNG, S. S.; NAGAYAMA, I.; TAMAKI, S. Regional distance-based k-nn classification. In: IEEE. *2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*. [S.l.], 2017. p. 56–62. Acesso em: 25, set. 2020. Citado 2 vezes nas páginas 26 e 27.

AYYADEVARA, V. *Pro machine learning algorithms : a hands-on approach to implementing algorithms in Python and R*. Berkeley: Apress, 2018. ISBN 978-1-4842-3564-5. Citado na página 27.

BATISTA, G.; SILVA, D. F. et al. How k-nearest neighbor parameters affect its performance. In: SN. *Argentine symposium on artificial intelligence*. [S.l.], 2009. p. 1–12. Citado na página 26.

BRASíLIA, H. da Criança de. *Histórico do Hospital*. 201–. Acesso em: 04, nov. 2020. Disponível em: <https://www.hcb.org.br/institucional/servicos-1>. Citado na página 36.

BRASíLIA, H. da Criança de. *Histórico do Hospital*. 201–. Acesso em: 04, nov. 2020. Disponível em: <https://www.hcb.org.br/institucional/historico-do-hospital>. Citado na página 36.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 31.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado na página 42.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967. Acesso em: 29, set. 2020. Citado na página 25.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, Springer, v. 29, n. 2-3, p. 103–130, 1997. Acesso em: 16, out. 2020. Citado 2 vezes nas páginas 28 e 29.

FARIAS, M. G.; CASTRO, S. M. d. Diagnóstico laboratorial das leucemias linfóides agudas. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, SciELO Brasil, v. 40, p. 91–98, 2004. Citado na página 19.

Fuse, K. et al. Patient-based prediction algorithm of relapse after allo-hsct for acute leukemia and its usefulness in the decision-making process using a machine learning approach. *Cancer Medicine*, v. 8, n. 11, p. 5058–5067, 2019. Acesso em: 02, dez. 2020. Citado na página 36.

GOOD, Z. et al. Single-cell developmental classification of b cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. *Nature medicine*, Nature Publishing Group, v. 24, n. 4, p. 474, 2018. Acesso em: 02, dez. 2020. Citado na página 35.

HAND, D. J.; YU, K. Idiot's bayes—not so stupid after all? *International statistical review*, Wiley Online Library, v. 69, n. 3, p. 385–398, 2001. Acesso em: 16, out. 2020. Citado na página 28.

HARRISON, C. J.; FORONI, L. Cytogenetics and molecular genetics of acute lymphoblastic leukemia. *Reviews in clinical and experimental hematology*, Wiley Online Library, v. 6, n. 2, p. 91–113, 2002. Citado na página 53.

HO, T. K. Random decision forests. In: IEEE. *Proceedings of 3rd international conference on document analysis and recognition*. [S.l.], 1995. v. 1, p. 278–282. Citado na página 31.

HSSINA, B. et al. A comparative study of decision tree id3 and c4. 5. *International Journal of Advanced Computer Science and Applications*, v. 4, n. 2, p. 13–19, 2014. Citado na página 32.

HUNGER, S. P.; MULLIGHAN, C. G. Acute lymphoblastic leukemia in children. *New England Journal of Medicine*, Mass Medical Soc, v. 373, n. 16, p. 1541–1552, 2015. Citado na página 19.

Instituto Nacional de Câncer José Alencar Gomes da Silva. *Estimativa 2020 : incidência de câncer no Brasil / Instituto Nacional de Câncer José Alencar Gomes da Silva*. [S.l.: s.n.], 2019. 120 p. Acesso em: 10, ago. 2020. ISBN 9788573183887. Citado na página 19.

LAROSE, D. T. *Data mining and predictive analytics*. [S.l.]: John Wiley & Sons, 2015. Citado na página 26.

LEITE, E. P. et al. Fatores prognósticos em crianças e adolescentes com leucemia linfóide aguda. *Revista Brasileira de Saúde Materno Infantil*, SciELO Brasil, v. 7, n. 4, p. 413–421, 2007. Acesso em: 20, fev. 2020. Citado na página 38.

MULAK, P.; TALHAR, N. Analysis of distance measures using k-nearest neighbor algorithm on kdd dataset. *International Journal of Science and Research (IJSR)*, v. 4, n. 7, 7 2015. Disponível em: <https://pdfs.semanticscholar.org/63f9/1b934f3dadec75c12a786f0e99d6df45ff37.pdf>. Acesso em: 10.17.2019. Citado na página 26.

PAN, L. et al. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. *Scientific reports*, Nature Publishing Group, v. 7, n. 1, p. 7402, 2017. Citado 4 vezes nas páginas 19, 35, 38 e 46.

PEDROSA, F.; LINS, M. Leucemia linfóide aguda: uma doença curável. *Rev. bras. saúde mater. infant*, v. 2, n. 1, p. 63–68, 2002. Acesso em: 11, set. 2020. Citado na página 19.

PESQUISA, C. N. de Ética em. Carta circular no. 039/2011/conep/cns/gb/ms. 2011. Acesso em: 04, nov. 2020. Disponível em: <http://conselho.saude.gov.br/images/comissoes/conep/documentos/CARTAS/CartaCircular039.pdf>. Citado na página 36.

PIRYONESI, S. M.; EL-DIRABY, T. E. Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. *Journal of Transportation Engineering, Part B: Pavements*, American Society of Civil Engineers, v. 146, n. 2, p. 04020022, 2020. Acesso em: 29, set. 2020. Citado na página 27.

QUINLAN, J. R. *C4.5: programs for machine learning*. [S.l.]: Morgan Kaufmann Publishers, 1993. Citado na página 32.

RUGGIERI, S. Efficient c4. 5 [classification algorithm]. *IEEE transactions on knowledge and data engineering*, IEEE, v. 14, n. 2, p. 438–444, 2002. Citado na página 32.

SANTOS, M. de O. Incidência, mortalidade e morbidade hospitalar por câncer em crianças, adolescentes e adultos jovens no brasil: Informações dos registros de câncer e do sistema de mortalidade. *Revista Brasileira de Cancerologia*, v. 64, n. 3, p. 439–440, 2018. Acesso em: 10, ago. 2020. Citado na página 19.

SHANNON, C. E. A mathematical theory of communication. *The Bell system technical journal*, Nokia Bell Labs, v. 27, n. 3, p. 379–423, 1948. Citado na página 32.

SOUZA, C. d.; VIANA, M. B.; OLIVEIRA, B. M. d. Recidiva da leucemia linfoblástica na criança: experiência do serviço de hematologia do hospital das clínicas da ufmg (1988-2005). *Rev. méd. Minas Gerais*, v. 18, n. 4, supl. 1, p. S55–S62, 2008. Acesso em: 20, fev. 2020. Citado na página 38.

TANG, Z. et al. Severity assessment of coronavirus disease 2019 (covid-19) using quantitative features from chest ct images. *arXiv preprint arXiv:2003.11988*, 2020. Acesso em: 31, out. 2020. Citado na página 23.

TEACHEY, D. T.; HUNGER, S. P. Predicting relapse risk in childhood acute lymphoblastic leukaemia. *British journal of haematology*, Wiley Online Library, v. 162, n. 5, p. 606–620, 2013. Acesso em: 11, set. 2020. Citado na página 20.

XIE, P. et al. Urban flow prediction from spatiotemporal data using machine learning: A survey. *Information Fusion*, Elsevier, v. 59, p. 1–12, 2020. Acesso em: 31, out. 2020. Citado na página 23.

YANG, X. et al. A review of gps trajectories classification based on transportation mode. *Sensors*, Multidisciplinary Digital Publishing Institute, v. 18, n. 11, p. 3741, 2018. Acesso em: 31, out. 2020. Citado na página 23.

Apêndices

APÊNDICE A – Árvore de Decisão Gerada pelo Algoritmo C4.5 usando a base balanceada

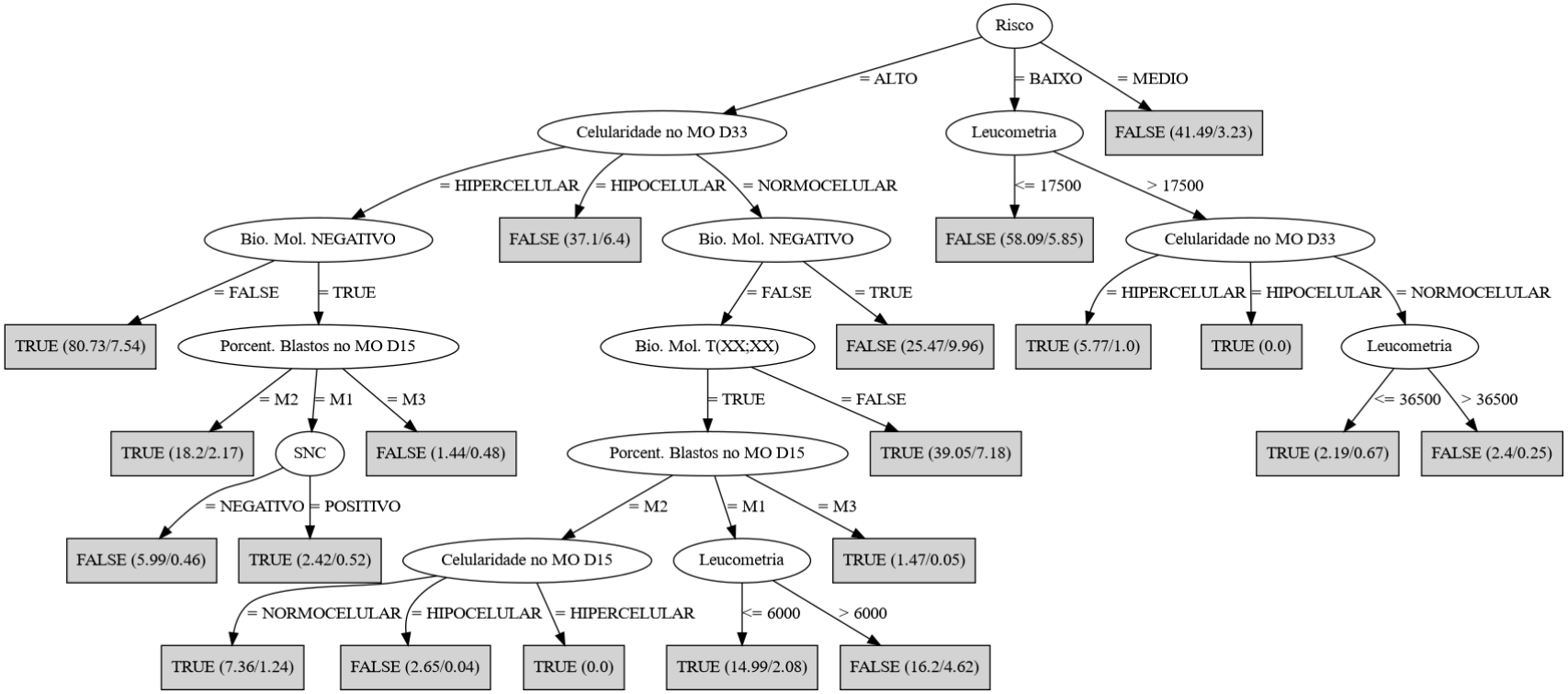

Figura 7 – Árvore de decisão gerada pelo algoritmo C4.5 usando a base de dados balanceada artificialmente usando o SMOTE considerando o atributo Risco. Fonte: Elaborado pelo autor.

APÊNDICE B – Árvore de decisão gerada pelo Algoritmo C4.5 usando a base de dados balanceada e sem considerar o atributo Risco

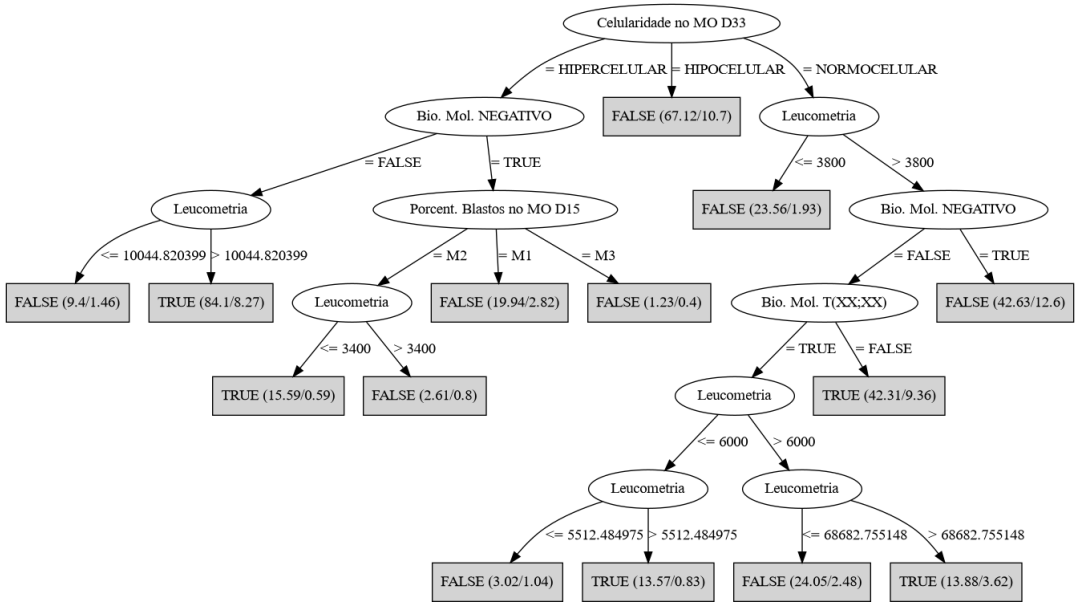

Figura 8 – Árvore de decisão gerada pelo algoritmo C4.5 após treinar com o conjunto de dados balanceado artificialmente usando o SMOTE e sem o atributo Risco. Fonte: Elaborado pelo autor.

Anexos

ANEXO A – Primeiro Anexo

Parecer do CEP informando a aprovação para a realização da pesquisa apresentada neste trabalho.

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Definição de um Modelo Computacional Baseado em Aprendizado de Máquinas para Predição de Recaídas do Tratamento da Leucemia Linfoblástica Aguda em Pacientes de Oncologia Pediátrica

Pesquisador: GLAUCO VITOR PEDROSA

Área Temática:

Versão: 3

CAAE: 28709719.5.0000.5558

Instituição Proponente: Faculdade do Gama

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 4.003.633

Apresentação do Projeto:

Trata-se de projeto de trabalho de conclusão de discente do curso de Engenharia de Software da Universidade de Brasília (UnB/FGA). O alvo de estudo do projeto é a Leucemia Linfoide Aguda (LLA). Os pesquisadores propõe usar informações de pacientes que foram tratados para a doença para construir um modelo computacional baseado em inteligência artificial capaz de prever o risco de recaída da doença.

Objetivo da Pesquisa:

Segundo os pesquisadores, "O objetivo geral deste projeto de pesquisa é o desenvolvimento de um sistema computacional, usando técnicas de aprendizado de máquinas, capaz de realizar a predição de recaídas em crianças com LLA, a partir de dados clínicos obtidos durante o tratamento do paciente."

Avaliação dos Riscos e Benefícios:

No documento de informações básicas do projeto, os pesquisadores avaliam os riscos associados à pesquisa como "Nenhum". Já sobre os benefícios, os pesquisadores afirmam que: "A LLA é o tipo de câncer mais frequente tratado no HCB. Por isso, o desenvolvimento de ferramentas que possam auxiliar no tratamento do paciente com LLA é de suma importância para os pacientes do Hospital."

Comentários e Considerações sobre a Pesquisa:

No parecer anterior, foram levantadas pendências relativas à etapa de pesquisa clínica, para obtenção dos dados de pacientes que serão utilizados, e também ao tamanho amostral. Os pesquisadores sanaram estas pendências.

Considerações sobre os Termos de apresentação obrigatória:

Os pesquisadores solicitam a dispensa de TCLE com os seguintes argumentos: "1. Levantamento retrospectivo de dados em prontuários, o que não interfere no cuidado recebido pelo paciente; 2. Não há riscos físicos e/ou biológicos para o paciente uma vez que o estudo é meramente observacional; 3. População de estudo eventualmente sem seguimento na instituição no presente (pacientes de outras localidades ou falecidos); 4. A confidencialidade da identificação pessoal dos pacientes é garantida pelo pesquisador principal e pelas técnicas de levantamento e guarda dos dados: os pacientes não serão identificados. Esses dados não serão objetos de análise. 5. Não serão utilizadas informações em prejuízo das pessoas e/ou das comunidades, inclusive em termos de autoestima, de prestígio e/ou econômicofinanceiro; Por esses motivos e como o uso e destinação dos dados coletados durante este projeto de pesquisa é puramente acadêmico, solicitamos a dispensa do referido documento."

Conclusões ou Pendências e Lista de Inadequações:

Haja vista que os pesquisadores sanaram as pendências levantadas nos pareceres anteriores, o parecer é pela aprovação do projeto proposto.

Considerações Finais a critério do CEP:

Projeto apreciado em Reunião Ordinária do CEP-FM-UnB-2020. Após apresentação do parecer do (a) Relator (a), aberta a discussão para os membros do Colegiado. O projeto foi Aprovado.

De acordo com a Resolução 466/2012-CONEP/CNS, itens X.1. - 3.b. e XI. -2.d, este Comitê chama a atenção da obrigatoriedade de envio do relatório parcial semestral e final do projeto de pesquisa para o CEP -FM, através de Notificações submetidas pela Plataforma Brasil, contados a partir da data de aprovação do protocolo de pesquisa.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1466721.pdf	14/04/2020 09:03:18		Aceito

Projeto Detalhado / Brochura Investigador	Projeto_Pesquisa.pdf	13/04/2020 13:57:49	GLAUCO VITOR PEDROSA	Aceito
Cronograma	Cronograma.pdf	13/04/2020 13:56:47	GLAUCO VITOR PEDROSA	Aceito
Outros	resposta_parecer.pdf	13/04/2020 13:53:35	GLAUCO VITOR PEDROSA	Aceito
Outros	lattes_diego.pdf	04/02/2020 14:38:23	GLAUCO VITOR PEDROSA	Aceito
Declaração de concordância	termo_concordancia_unb.pdf	04/02/2020 14:37:48	GLAUCO VITOR PEDROSA	Aceito
Outros	termo_concordancia_hcb.pdf	03/02/2020 14:13:23	GLAUCO VITOR PEDROSA	Aceito
Outros	lattes_glauco.pdf	01/02/2020 12:43:43	GLAUCO VITOR PEDROSA	Aceito
Folha de Rosto	folha_rosto.pdf	01/02/2020 12:36:45	GLAUCO VITOR PEDROSA	Aceito
Outros	carta_encaminhamento.pdf	02/12/2019 16:18:19	GLAUCO VITOR PEDROSA	Aceito
Outros	resumo_estruturado.pdf	02/12/2019 15:11:58	GLAUCO VITOR PEDROSA	Aceito
Orçamento	Planilha_orcamentaria.pdf	02/12/2019 15:02:45	GLAUCO VITOR PEDROSA	Aceito
Declaração de Pesquisadores	termo_responsabilidade.pdf	02/12/2019 15:00:56	GLAUCO VITOR PEDROSA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	dispensa_TCLE.pdf	02/12/2019 15:00:23	GLAUCO VITOR PEDROSA	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

BRASILIA, 02 de Maio de 2020

Assinado por:
Antônio Carlos Rodrigues da Cunha
(Coordenador(a))

Endereço: Universidade de Brasília, Campus Universitário Darcy Ribeiro - Faculdade de Medicina
Bairro: Asa Norte **CEP:** 70.910-900
UF: DF **Município:** BRASILIA
Telefone: (61)3107-1918 **E-mail:** cepfm@unb.br