Aplicação de métodos de aprendizado de máquina para predição de desfecho clínico

Guilherme Silva de Camargo
Departamento de computação (DComp)
Universidade Federal de São Carlos (UFSCar)
18052-780, Sorocaba, São Paulo, Brasil
guilhermecamargo@estudante.ufscar.br

***Abstract*—A leucemia é um tipo de tumor que afeta a medula óssea, dessa forma, na leucemia mieloide aguda, as células-tronco mieloides, que dão origem às células sanguíneas, sofrem algumas mutações genéticas, cujo entendimento sobre as relações entre os genes causadores, e outros dados, ainda não são totalmente conhecidas. Nesse sentido, este trabalho visa entender através dessas informações (expressão genética, mutações gênicas e dados clínicos) uma relação entre esses dados, para obter uma predição de desfecho clínico, indicando a probabilidade do tratamento resultar em óbito do paciente, ou sobrevivência. Ademais, espera-se também conseguir melhores combinações de tratamentos, essas adaptadas aos dados de cada paciente, tal como idade, auxiliando na escolha do tratamento mais indicado para cada paciente. Por fim, obteve-se resultados na faixa de 75% de acerto na predição, utilizando os métodos de classificação Regressão Logística e Rede Neural Artificial (Perceptron multicamada), assim, a ferramenta mostrou-se útil no âmbito de auxiliar na escolha do tratamento, ou na diminuição da imaterialidade na relação entre os dados clínicos, e genéticos do paciente, sendo promissora para estudos ainda mais aprofundados.**

***Keywords*-aprendizado de máquina, leucemia mieloide aguda; predição clínica.**

I. Introdução

A leucemia mieloide aguda (LMA) é uma doença rara que afeta a medula óssea, as células-tronco dos pacientes com essa doença sofrem uma mutação genética. As células sanguíneas são formadas a partir dessas células-tronco, consistindo de leucócitos, glóbulos vermelhos e plaquetas. Nessa situação, as células ficam doentes, incapazes de se desenvolver e começam a se multiplicar descontroladamente. Há incerteza nos fatores causadores da doença, por isso, ocorre uma dificuldade em indicar tratamentos aos pacientes.

Na maioria dos casos, o médico vale-se de estudos para predizer o tratamento, tal como as recomendações para diagnóstico e tratamento, publicadas pelo grupo de pesquisa European LeukemiaNet (ELN), atualizadas pela última vez em 2022, porém, mesmo estes estudos, não levam em consideração dados como a idade do paciente, o qual são valiosas informações, para determinar um tratamento ao paciente, assim, as respostas terapêuticas a tratamentos por parte dos pacientes, ainda que pertencentes ao mesmo grupo de classificação ELN, podem apresentar diferenças, como, por exemplo, a reação diferente a um mesmo medicamento.

Dessa forma, mostra-se muito valioso apresentar mais uma alternativa, para ajudar na predição da chance de sobrevivência dos pacientes, dado o conjunto de dados clínico e o tratamento prescrito, nesse sentido, este trabalho propõe um sistema para auxiliar o médico na prescrição do tratamento, dada às características de cada caso, através da predição de desfecho clínico, predizendo a probabilidade de sobrevivência e óbito do paciente, visto que, a área de aprendizado de máquina com suas diversas técnicas, vêm mostrando-se eficaz no encontro de padrões em grandes volumes de dados.

Em suma, espera-se reduzir a abstração e subjetividade na predição de tratamentos médicos no contexto da LMA, aumentando a assertividade na indicação de terapias aos pacientes, com isso, ocasionando um aumento na qualidade de vida e sobrevivência destes pacientes.

II. Dados e pré-processamento

A base de dados conta com três diferentes tipos de informação, dados clínicos, de expressão genética e de mutação gênica, inicialmente, constituía-se de 390 amostras com 10 atributos para os dados clínicos, 304 amostras com 14712 atributos para as expressões genéticas e 362 amostras com 318 atributos para as mutações gênicas. Porém, havia atributos com valores faltantes, amostras duplicadas e valores categóricos em alguns atributos.

Além disso, é importante entender quais são estes atributos, e seu significado, nesse sentido, os 10 atributos clínicos são: “Diagnosis Age”, idade de diagnóstico; “Bone Marrow Blast Percentage”, porcentagem de blastos na medula óssea do paciente; “Mutation Count”, contador de mutações gênicas no paciente; “PB Blast Percentage”, porcentagem de blastos em sangue periférico do paciente; “WBC”, leucócitos no sangue do paciente; “Sex”, sexo do paciente; “Race”, raça do paciente; “Cytogenetic Info”, informações citogenéticas do paciente; “ELN Risk Classification”, classificação de risco molecular do paciente; e ”Treatment Intensity”, intensidade do tratamento.

Ademais, na base de dados contendo as expressões genéticas, cada atributo identifica um gene diferente, buscando expressar o quanto aquele gene está presente no paciente, representado por um valor numérico, esta expressão genética pode ser definida como o processo pelo qual a

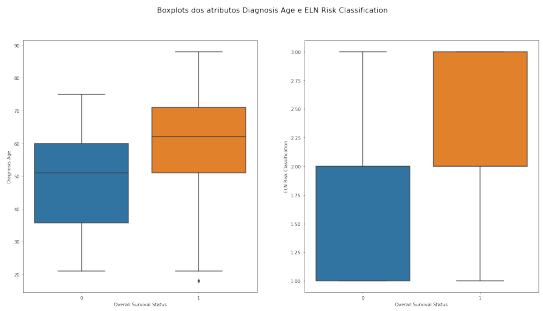

Figure 1. Box plots para os atributos clínicos

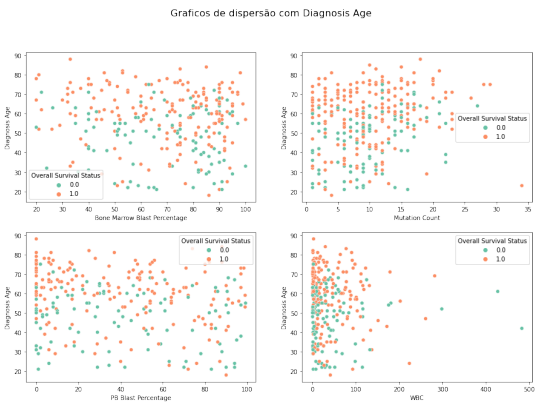

Figure 2. Gráficos de dispersão para os atributos clínicos

informação hereditária contida em um gene, como uma sequência de DNA, é usada para produzir um produto gênico funcional, como proteínas ou RNA.

Por fim, o mesmo ocorre para as mutações gênicas, sendo que cada atributo expressa um gene diferente, além disso, temos que a representação é dada por dois valores, zero para ausência de mutação no gene, e 1 para gene que sofreu mutação.

A. *Análise de dados*

Para definir os pré-processamentos foram utilizados gráficos e medidas descritivas, sendo estes box plots, gráficos de barras para os valores médios dos atributos, histogramas e gráficos de dispersão, além disso, tivemos medidas como a média, desvio padrão, quantidade, mínimo, máximo, mediana, 1° e 3° quartis, e as matrizes de covariância e correlação.

Outrossim, será explicitada apenas as informações que impactam diretamente na escolha dos pré-processamentos, e mais a frente na seleção dos atributos. Nesse sentido, foi possível identificar através dos box plots que a distribuição da classe "Overall Survival Status", para os atributos atributos "Diagnosis Age" e "ELN Risk Classification", está bem distinta para seus dois valores. Portanto, isto indica que as duas classes têm distribuições significativamente diferentes, como pode ser visto na figura 1.

Por fim, utilizando-se dos gráficos de dispersão, constatou-se que para os valores acima de 60 anos do atributo "Diagnosis Age", há uma predominância da classe 1, como pode ser verificado na figura 2.

B. *Pré-Processamentos*

Como visto anteriormente nas análises feitas, há dados ausentes, nesse sentido, ocorreu o primeiro pré-processamento, utilizou-se a média para substituir esses valores. Além disso, ocorreram testes com a mediana, mas essa obteve desempenhos piores que a média. Por isso, optou-se pela média, escolha inicial que se deu pelos seguintes fatores: manutenção da estrutura de dados, a média ajuda a manter a estrutura dos dados originais; melhoria da precisão de análises estatísticas, a média é um bom estimador do valor verdadeiro quando os dados estão distribuídos; e redução da perda de informação.

Table I
QUANTIDADE DE AMOSTRAS E ATRIBUTOS NAS BASES

Base de dados	Amostras (Treino/Teste)	Atributos
Clínica	316 / 28	38
Expressão genética	316 / 28	14713
Mutação gênica	316 / 28	319

Ademais, os atributos categóricos foram convertidos para atributos numéricos, sendo que os atributos receberam valor 0 se a informação não está presente, e 1 indicando que está presente naquela amostra. Além disso, houve agrupamento de valores para "Race", pois havia diferentes valores em formato de texto, contendo a mesma informação, como, por exemplo, "White" e "WHITE", sendo tratados como informações distintas, o que não é o caso.

Por fim, temos a tabela I, indicando a quantidade de amostras e atributos para cada base de dados, logo após os pré-processamentos.

III. PROTOCOLO EXPERIMENTAL

Inicialmente, após a realização dos pré-processamentos, agrupou-se todas as três bases de dados para ocorrer os primeiros experimentos, foram testados os seguintes algoritmos de classificação: k-vizinhos mais próximos (KNN), Naive Bayes Gaussiano, Regressão Logística, Rede Neural Artificial (Perceptron multicamada), Máquina de Vetores de Suporte e Floresta Aleatória. Assim, os primeiros resultados podem ser vistos na tabela II.

Além disso, como será explicitado nos próximos tópicos, para buscar o melhor desempenho, e uma maior generalização dos modelos, utilizando diversas métricas, houve seleção dos melhores atributos para cada base de dados. Dessa forma, obteve-se melhores resultados, aos quais serão explicados na sequência.

Table II
MÉTRICAS DOS ALGORITMOS COM A BASE COMPLETA

Classificador	Acurácia	Desvio Padrão	Curva ROC	F1
KNN	0.622	0.105	0.540	0.736
Naive Bayes	0.467	0.103	0.519	0.344
Regressão Logistica	0.679	0.076	0.621	0.760
Rede Neural Artificial	0.590	0.118	0.544	0.660
Máquina de Vetores de Suporte	0.657	0.101	0.541	0.776
Floresta Aleatoria	0.654	0.090	0.576	0.745

Table III
MÉTRICAS DOS ALGORITMOS PARA A BASE COM 72 ATRIBUTOS

Classificador	Acurácia	Desvio Padrão	Curva ROC	F1
KNN	0.711	0.075	0.630	0.801
Naive Bayes	0.714	0.074	0.629	0.799
Regressão Logistica	0.731	0.102	0.680	0.795
Rede Neural Artificial	0.755	0.072	0.696	0.818
Máquina de Vetores de Suporte	0.746	0.092	0.682	0.813
Floresta Aleatoria	0.708	0.065	0.639	0.790

A. Descrição da forma de avaliação

No treinamento dos métodos de classificação para a seleção de atributos, e na sequência para visualização do desempenho, para cada um dos diversos métodos de classificação, utilizou-se o uso da técnica de validação cruzada com 10 folds, também conhecida como ”cross-fold validation”, esta técnica será particularmente útil, pois o conjunto de dados disponível é pequeno, e precisa ser dividido em conjuntos de treinamento e teste, para a avaliação do modelo, o que ajuda a garantir que o modelo não esteja super ajustando, ou sub ajustando, assim, sendo mais fácil avaliar a capacidade de generalização dos modelos.

Ademais, após a seleção de atributos citada anteriormente, no treinamento dos métodos para a análise de resultados, geração de gráficos, matrizes e curvas, os dados foram divididos em 80% para treinamento e 20% para teste.

Por fim, para predição dos dados de teste submetidos no Kaggle, utilizou-se a base de dados completa.

B. Medidas de desempenho empregadas

Para avaliar os modelos, a princípio utilizou-se as métricas:

- Acurácia, proporção de exemplos classificados corretamente em relação ao número total de exemplos;
- Desvio Padrão, mede a variabilidade dos dados em relação à média;
- Curva ROC, avalia o desempenho de um modelo binário à medida que o limiar de classificação é variado;
- Recall, mede a proporção de exemplos positivos que são classificados corretamente como positivos;
- Precisão, mede a proporção de exemplos classificados corretamente como positivos em relação ao número total de exemplos classificados como positivos;
- F1-score, medida de precisão e recall combinadas, calculada como a média harmônica dessas duas medidas.

Porém, na avaliação dos resultados, também utilizou-se a curva de aprendizado, que mostra a evolução do desempenho do modelo, à medida que o tamanho do conjunto de treinamento é aumentado; e a curva de validação, que avalia o desempenho do modelo em diferentes hiperparâmetros.

C. Ajuste de Parâmetros

Como citado anteriormente, para buscar o melhor desempenho possível, e uma melhor generalização dos dados, empregou-se “seleção de features”, que consiste em selecionar as ”k” melhores features baseado nas pontuações mais altas, para calcular essas pontuações, utilizou-se o valor ANOVA (Análise de variância), para as bases de dados contendo as expressões genéticas, e os dados clínicos, por outro lado, para a base contendo as mutações gênicas, abordou-se o valor de estatísticas de qui-quadrado.

Dessa forma, testaram-se combinações de “k” para as três bases de dados separadamente, agrupando os dados na sequência para treinando com o classificador de Regressão Logística, método ao qual apresentou melhor desempenho inicial, assim, selecionando os valores de “k” que obtiveram maior valor de curva ROC. Ademais, os diversos resultados das métricas para os diferentes valores de “k”, estão link.disponíveis neste link.

Por fim, também foi testado ajuste de parâmetros para cada algoritmo, avaliando o desempenho de cada parâmetro através da curva ROC.

IV. RESULTADOS

Inicialmente, após as seleções de atributos realizadas, foi possível perceber um grande ganho em todos as medidas, como por exemplo, obteve-se 72% em média de acurácia dos algoritmos, demonstrado na tabela III, um ganho expressivo em relação aos valores iniciais, demonstrados na tabela II.

Para a avaliação dos resultados para cada modelo, analisou-se a matriz de confusão, as métricas e as curvas de aprendizado e de validação, dessa maneira, constatou-se melhor desempenho para os seguintes modelos: regressão logística, como já detectado inicialmente, Rede Neural Artificial e Naive Bayes. Assim, como pode ser visto na figura 3, esses modelos citados anteriormente, apresentam uma curva de aprendizado a qual as curvas de pontuação de treinamento e de validação convergem, indicando que o modelo não está sofrendo de overfitting. Ademais, aliado a matriz de confusão, que especifica quantas amostras foram preditas erradas e certas de cada classe, evidenciou-se que os 3 métodos obtiveram uma melhor generalização.

Além disso, com os resultados da curva ROC no treinamento dos modelos, para a base de dados com 72 atributos, confirma-se o bom desempenho dos 3 classificadores, como pode ser visto na tabela III.

Por fim, os resultados de cada um dos 6 modelos de classificação no placar público da competição no Kaggle pode ser visto na tabela IV.

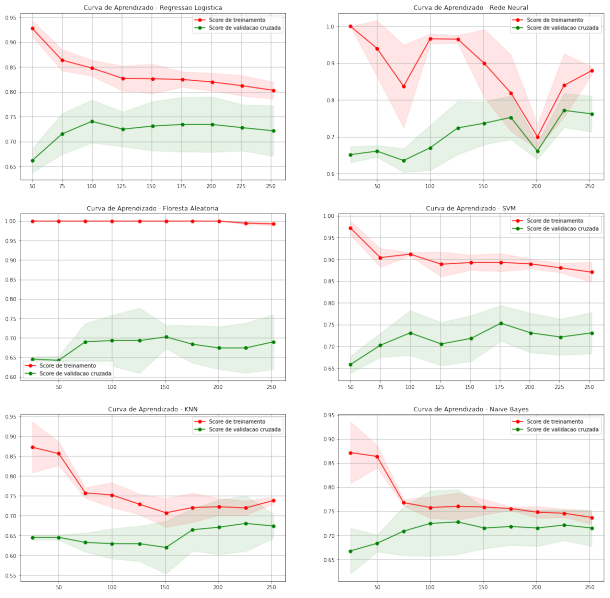

Figure 3. Curva de aprendizado dos classificadores

Table IV
MÉTRICAS DOS ALGORITMOS PARA A BASE COM 72 ATRIBUTOS

Classificador	Pontuação (Placar Público Kaggle)
KNN	0.5
Naive Bayes	0.7
Regressão Logistica	0.75
Rede Neural Artificial	0.725
Máquina de Vetores de Suporte	0.55
Floresta Aleatoria	0.625

V. ESTRATÉGIA FINAL

A solução final escolhida para envio na competição contou com os seguintes métodos, Regressão Logística, apresentou 75% de pontuação no placar público da competição no Kaggle, e Rede Neural Artificial, apresentou 72,5% de pontuação.

A. Base de dados

A base de dados após a realização dos pré-processamentos, e escolha final da seleção de atributos, ficou com 72 atributos, sendo 2 de dados clínicos, 55 de mutação gênica e 15 de expressão genética, elucidados a seguir:

- Clínicos: Diagnosis Age e ELN Risk Classification.

- Expressão Genética: OSBPL5, DDIT4, LTK, EXT2, SLC29A2, CALR, CERCAM, AGRN, H2AFY, ECE1, GALNT1, PPM1H, MICALL2, SLC25A29 e LPPR3.

- Mutação Gênica: FLT3, TP53, U2AF1, PHF6, CALR, THRAP3, PHIP, PKD1L2, CADM2, BICD1, C5, CAPN6, CSF3R, CTNNA2, DIAPH2, DOCK1, EPHB1, HSPA4, KCNJ6, LTBP3, MAX, PKNOX1, PLRG1, PRKAA2, SRSF2, SLC4A3, SOX5, KMT2D, CACNA1G, SLC7A7, AURKB, DDX23, CEP170, ATP6AP2, POSTN, ATG14, MYH15, IGSF9B, GPATCH8, DOCK9, ATRNL1, FILIP1, PCDH11X, SETD2, THEG, LIMA1, NECAB2, PARP14, RIF1, ELMOD1, RALGAPA2, PAPPA2, ARHGEF28, WNK4 e KIAA1109.

Table V
MELHORES PARÂMETROS PARA OS MÉTODOS

Classificador	Melhores Parâmetros
KNN	n_neighbors: 17
Naive Bayes	var_smoothing: 0.01
Regressão Logistica	C: 1
Rede Neural Artificial	hidden_layer_sizes: (10, 10)
Máquina de Vetores de Suporte	C: 1000
Floresta Aleatoria	n_estimators: 100

B. Melhores Parâmetros

Os parâmetros dos métodos de classificação, também foram testados, de forma a buscar o melhor desempenho possível do modelo. Nesse sentido, para os modelos escolhidos, os parâmetros foram os seguintes:

- Rede Neural Artificial: “hidden_layer_sizes”, o i-ésimo elemento representa o número de neurônios na i-ésima camada oculta, com valor (10, 10);

- Regressão Logística: “C”, inverso da força de regularização, com valor 1.

Por fim, os demais parâmetros escolhidos para os outros métodos, podem ser vistos na tabela V.

VI. CONCLUSÃO

Conclui-se que o sistema de auxílio na predição de desfecho clínico, apresenta resultados promissores na ajuda da indicação de melhores tratamentos, para os pacientes acometidos da Leucemia Mieloide Aguda, mostrando-se um campo sujeito a maior exploração no âmbito médico, visto que os métodos de aprendizado de máquina, mostram-se capazes de encontrar relações e padrões, tal como demonstrado neste trabalho. Ademais, os resultados obtidos são satisfatórios, mas apresentam perspectiva de melhora em diversos fatores, caso haja um maior aprofundamento do que foi abordado. Dessa forma, o sistema pode ser adotado como mais uma ferramenta no auxílio de predição de tratamentos médicos, para LMA, acompanhado de um especialista validando a utilidade das informações.

REFERENCES

[1] YU, X. et al. Predicting lung adenocarcinoma disease progression using methylationcorrelated blocks and ensemble machine learning classifiers. PeerJ, v. 9, p. e10884–e10884, 2021.

[2] 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.

[3] CAMPOS, m. g., OLIVEIRA, j. s., CHAUFAILLE, m. l. (2006). Considerações sobre idade e perfil de apresentação em leucemia mielóide crônica. Revista Brasileira de Hematologia Hemoter, 28(Supl.2): Abstr.452.

[4] DO CARMO, c. s., TEIXEIRA, j. r., TEIXEIRA, j. (2014). Leucemia- sociedade em riscos.

[5] MEIRELES, m. f., PAIVA, g. p. (2017). leucemia mieloide. rev brasileira , 12-25