Aplicação de métodos de aprendizado de máquina para desfecho clínico

Guilherme Fumagali Marques
DComp - Departamento de computação
UFSCar - Campus Sorocaba
Sorocaba, Brasil
guilhermefumagali@estudante.ufscar.br

***Abstract*—No campo de estudo médico existe o problema relacionado a pacientes diagnosticados com Leucemia Mielóide Aguda (LMA). Atualmente o diagnóstico enquadra o paciente em um dos três níveis de risco da doença, no entanto, existem problemas relacionados à diferença entre os pacientes dentro da mesma classe de risco, além de que atualmente não há um estudo claro sobre a classe de risco médio.**

Com isso, com a análise dos dados de paciente com LMA, esse trabalho propõe métodos estatísticos e de Aprendizado de Máquina para se extrair a probabilidade de um paciente sobreviver ao tratamento, considerando suas características físicas e genéticas, no intuito de se trazer novas informações ao profissional médico que irá analisar o paciente.

***Keywords*-Leucemia Mielóide Aguda; Aprendizado de Máquina; Probabilidade de sobrevivência;**

I. Introdução

Doenças cancerígenas pertencem a um grupo patológico caracterizado pelo crescimento anormal de células em um organismo. A Leucemia Mielóide Aguda (LMA) é um tipo de câncer no qual essa proliferação anormal ocorre nas células progenitoras da linhagem mielóide - nome dado a células que dão origem a diferentes corpos que compõem o sistema imunológico do indivíduo - na medula óssea, no sangue e em regiões extramedulares.

Um paciente diagnosticado em um quadro clínico de LMA comumente é enquadrado em um grupo que representa seu nível de risco com base em sua estrutura citogenética, podendo ser favorável, intermediário ou adverso (grupo 1, 2 e 3 respectivamente). Apesar de auxiliar no direcionamento do tratamento do paciente, existe uma dificuldade relacionada à variedade dos pacientes dentro da mesma classe de risco, além de que não há uma definição clara quanto à classe de risco médio.

Dado o contexto, algoritmos de Aprendizado de Máquina podem trazer novas informações a essa área de pesquisa, de modo a ter métricas diferentes além da classificação de risco citadas, para assim ajudar o médico a oferecer os tratamentos mais adequados possíveis a quaisquer pacientes.

II. Dados e Pré-processamento

A base de dados se origina a partir de cadastros clínicos e exames de obtenção de dados citogenéticos dos pacientes. Nela, há referências a pacientes, separados por um código de identificação, que foram diagnosticados com LMA e consequentemente foram submetidos ao sistema tradicional de classificação, no qual foram coletados os dados necessários para classificá-los em um dos três níveis de risco.

Em cada amostra, existe o status geral de sobrevivência do paciente em questão, sendo esse um dos dados mais importantes para o treinamento dos métodos de AM, visto que a partir dele é possível separar as amostras em classe positiva e negativa, sendo respectivamente óbito ou sucesso no tratamento. Com isso, algoritmos de AM podem ser usados para obter deduções sobre como um novo paciente irá reagir ao tratamento com base nas amostras de dados do passado.

Considerando apenas os pacientes com classe definida, existem no total 316 amostras, sendo aproximadamente 65% da classe positiva e 35% da classe oposta, inferindo em um desequilíbrio para a classe positiva, fato esse que deve ser tratado como uma característica do contexto do problema, visto que coletar mais dados desse tipo de dado é muito custoso.

A. Base de Dados

Todo o conjunto de dados é constituído a partir de outros três subconjuntos: dados clínicos, expressões genéticas e mutações gênicas.

O primeiro conjunto engloba informações referentes aos dados cadastrais do paciente no hospital, que inclui a idade, quando foi diagnosticado, sexo, raça, quantidade de mutações genéticas no organismo, porcentagem de blastos na medula óssea e no sangue periférico, quantidade de glóbulos brancos do sistema imunológico, informações da sua estrutura cromossômica, ou citogenética, e por fim, a intensidade do tratamento em que esse paciente foi exposto junto a classificação de risco.

A distribuição desses dados em relação a classe do paciente se mostra equivalente em alguns atributos e desequilibrado em outros. Como exposto na Figura 1, têm-se relativamente quantidades parecidas de amostras de sexo feminino e masculino, por outro lado, existem muitas amostras de uma só raça em comparação as outras, tornando esse atributo potencialmente descartável, considerando também que não há estudos consolidados sobre a relação entre raça e o desenvolvimento de AML.

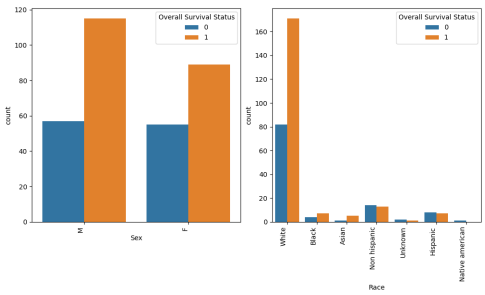

Figure 1. Análise de frequência dos atributos Sexo e Raça

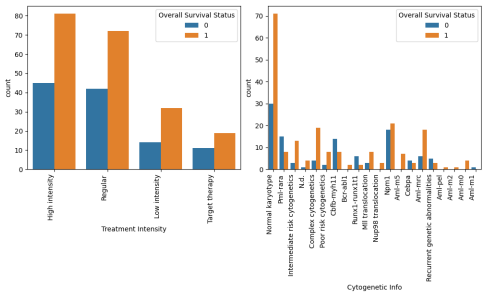

Figure 2. Análise de frequência dos atributos Intensidade do Tratamento e Informação Citogenética

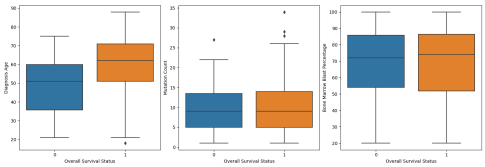

Figure 3. Distribuição dos atributos Idade, Quantidade de Mutações e Blastos na Medula Óssea

Ainda analisando os atributos categóricos, na figura 2 é explicitada a frequência dos dados em relação ao atributo de intensidade de tratamento e informações citogenéticas. Proporcionalmente, repara-se que a intensidade do tratamento não interfere diretamente na classe do paciente. Por outro lado, a estrutura citogenética traz comportamentos diferentes em cada tipo, e associado ao fator idade, essa classificação da estrutura cromossômica se destaca como uma das principais variáveis relacionadas à rotulação de risco do paciente pelo especialista [Velloso et al, 2011].

Nos atributos numéricos, é possível visualizar a influência da idade na taxa geral de sobrevivência dos pacientes, atributo esse considerado o maior determinante da sobrevivência do paciente. [Kumar, C. Chandra, 2011]. Explicitado na primeira imagem da Figura 3, a idade mediana dos pacientes que morrem é aproximadamente 60, enquanto dos que sobrevivem é aproximadamente 50, mas com uma média geral relativamente mais baixa. Ademais, na figura 5, a importância desse atributo é novamente observada ao verificar que a idade é o principal fator que afeta a sobrevivência do

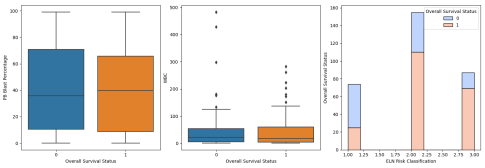

Figure 4. Distribuição dos atributos Blastos no Sangue Periférico, Quantidade de Glóbulos Brancos e Classificação de Risco

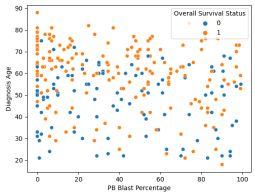

Figure 5. Blastos no sangue periférico X Idade

paciente na relação Idade x Blastos no sangue periférico.

Por fim, outro atributo numérico importante é a classificação de risco. Ao analisar a última imagem da Figura 4, é possível perceber a maior taxa de mortalidade no último nível de classificação, com mais de 75% de mortes.

Nos dados dos dois subconjuntos restantes: Expressões e Mutações Genéticas, têm-se no total informações sobre expressões de 14712 genes, que não foram coletadas completamente em todos os pacientes, resultando em cerca de 20% dos campos sem nenhum valor. Enquanto nos dados de mutações gênicas existem 318 genes, com cerca de apenas 3% dos campos incompletos.

B. Pré-processamento

Primeiramente, antes de qualquer interação ou análises sob os dados, é necessário ter todas as tuplas das tabelas preenchidas e com dados categóricos convertidos à numéricos.

Nos dados clínicos, há 4 atributos categóricos, os mesmos da figura 2 e 3. Como nenhum desses atributos têm características ordinais, ou seja, em que a ordem deve ser preservada, a maneira mais adequada para tratar esses dados é inserindo uma coluna para cada valor de categoria possível, como uma sequência binária, preservando assim a distância entre quaisquer dois valores deste atributo, esse método também é conhecido como codificação one-hot ou método de dummies.

No caso dos campos faltantes, como primeira opção foi considerado permutar o campo ausente pela média da coluna, no entanto, substituir com base em 3 amostras mais próximas traz uma distribuição dos dados mais uniforme, visto que utiliza um valor mais local, e portanto, não muito discrepante daquele espaço da amostra. Além disso, principalmente nos dados de expressões genéticas, no qual 20% dos valores estão ausentes, substituir pela média aumenta

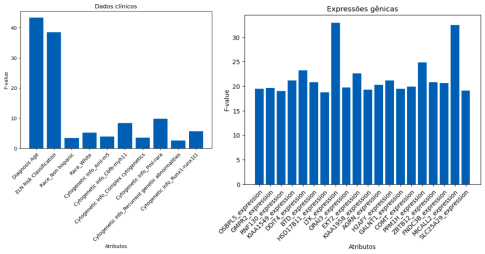

Figure 6. F-valores associados ao teste ANOVA

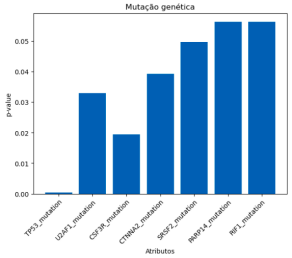

Figure 7. P-valores associados ao teste do qui-quadrado

artificialmente a frequência dos valores daquela coluna no valor da média correspondente.

Após isso, neste trabalho é proposto uma seleção das features mais importantes relativamente à classe, com base em métodos estatísticos. Dois tipos de teste foram realizados: ANOVA e qui-quadrado.

1) ANOVA: "Análise de Variância" ou ANOVA é um teste que consiste em avaliar a relação entre cada atributo com as classes da base de dados. Esse método compara as médias da classe, chamada de variável dependente, em diferentes grupos definidos pelas features, e então calcula a variação total da variável dependente e a divide em duas partes: a variação explicada pelas features e a variação não explicada. Essa variação é expressa através do "F-Valor", onde se a variação explicada for maior que a não explicada, esse valor tende a ser maior, e portanto, mais significativa é a relação entre o atributo e a classe.

2) Qui-quadrado: O teste do qui-quadrado é aplicado a cada feature discreta individualmente e avalia a associação entre a feature e a classe, comparando a frequência observada dos valores da feature com a frequência esperada. A frequência esperada é calculada com base na distribuição da variável de saída e da feature, assumindo que elas são independentes. O resultado do teste do qui-quadrado é dado pelo "p-valor". Esse valor é calculado para cada feature e indica a probabilidade de que a relação entre a feature e a variável de saída seja apenas devido ao acaso. Com isso, o valor de p indica o grau de confiança que se pode ter na relação observada entre uma feature e a variável de saída: quanto menor o valor de p, maior a confiança na relação.

C. Resultados dos testes estatísticos

A escolha da quantidade de features a ser selecionada usualmente é feita por um valor limiar, no qual amostras com resultados inferiores a esse limiar são descartadas. Neste trabalho, o limiar foi definido com base nos valores resultantes de cada uma das três partes da base de dados.

O teste ANOVA foi realizado sobre os dados clínicos e as expressões gênicas, devido às suas características não discretas, e o teste do qui-quadrado com o conjunto de dados restante: mutações genéticas.

Nos dados clínicos, é possível claramente a discrepância dos f-valores dos atributos idade e classificação de risco em relação aos demais no primeiro gráfico da figura 6. Com isso, infere-se que esses dois atributos são bem mais significativos para a taxa geral de sobrevivência do paciente do que todos os outros desse conjunto.

No segundo gráfico da figura 6, destaca-se os 20 maiores f-valores do conjunto de expressões gênicas. Ao contrário do conjunto anterior, os valores desta base de dados estão distribuídos de forma mais uniforme. Definindo um limiar de f-valores maiores que 19, obtêm-se que existem 19 genes com essa faixa de influência na taxa de sobrevivência do paciente.

Por fim, na figura 7, é possível observar os p-valores associados aos atributos das mutações genéticas. Neles, o maior grau de confiança está relacionado com o gene "TP53", mas definindo um limiar de 0.05, também são selecionados os outros quatro primeiros genes da figura.

III. Protocolo Experimental

Os experimentos foram realizados com base em seis diferentes modelos de Aprendizado de Máquina e comparados de modo a utilizar o modelo com métricas mais interessantes para predição da probabilidade de classe positiva nas amostras de teste.

Foram utilizadas as métricas: área da Curva Roc, acurácia, revocação, precisão e F1-score. Ademais, para garantir a melhor precisão desses resultados numéricos, estes foram calculados a partir de um algoritmo de KFold estratificado, no qual as distribuições em cada classe se mantém nas partições de teste e treino.

A. Seleção dos modelos e ajuste de hiperparâmetros

Todos os modelos usados neste trabalho requerem hiperparâmetros a serem ajustados de forma a melhor lidar com os dados. Para realizar esse ajuste, foi utilizado o método de busca em grade, no qual vários hiperparâmetros são testados e o modelo com a combinação de hiperparâmetros com a melhor medida de desempenho é selecionado. No contexto, a métrica de desempenho adotada foi a área sob a curva ROC.

Depois de instanciados os modelos com os melhores hiperparâmetros para o conjunto, estes passam por uma análise de resultados, na qual é selecionado manualmente o modelo com uma maior expectativa de resultado no Kaggle.

Table I
RESULTADOS

Modelo	Score Inicial	Score com seleção de features	Ganho
SVM	0.674432	0.818896	14%
Regressão Logística	0.714919	0.817083	10%
RNA	0.622505	0.799134	17%
Naive Bayes	0.616334	0.795060	18%
Random Forest	0.612478	0.773019	16%
KNN	0.623323	0.762048	14%

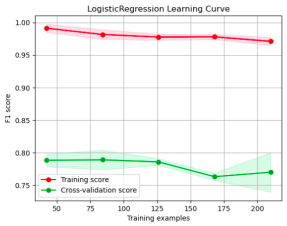

Figure 8. Curva de Aprendizado Regressão Logistíca

IV. RESULTADOS

É sabido que a maior influência para um bom desempenho de um modelo são os dados de treinamento. Com isso, foi-se adotado e testado diversas estratégias de pré-processamento.

Sem nenhuma seleção de features, apenas com o processamento suficiente para se adequar aos modelos, obteve-se resultados da tabela 1.

Entretanto, mesmo com o maior score, ao analisar a curva de aprendizado da Regressão Logística, na figura 8, é possível visualizar um possível overfitting, considerando que o score no conjunto de treino se manteve alto e na validação o resultado tendeu a diminuir, não sendo um modelo com boa generalização, isso mesmo com o hiperparâmetro do custo ajustado para evitar sobreajustamento.

O modelo que apresentou a curva de aprendizado mais próxima do ideal foi o KNN, mas ao analisar as suas métricas de desempenho, infere-se que o modelo rotulou todas as amostras do conjunto de validação para a classe 1, resultando em uma precisão nula para a classe 0.

A. *Pré-processamento com métodos estatísticos*

Diminuindo a quantidade de atributos com base nos valores estatísticos mostrados nas seções anteriores, obteve-se os resultados da tabela 1, que mostra métricas bem mais interessantes, com ganhos de 10% a 18% de acurácia aproximadamente.

Os dois classificadores com melhores resultados apresentam curvas de aprendizado que não indicam o overfitting (Figura 9). Por outro lado, ao comparar as duas, a curva do modelo SVM apresenta a mesma generalização mas com uma taxa maior de acerto nas próprias amostras de treino.

No entanto, ao submeter esses modelos no kaggle, temos que esses dois melhores classificadores ficaram atrás da Rede Neural, que atingiu a pontuação de 0.825, contra 0.6875 da Regressão Logística e do SVM.

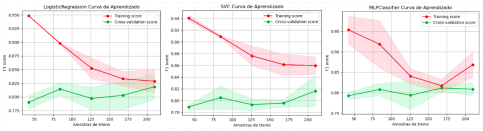

Figure 9. Curva de Aprendizado com seleção de features

Table II
MEDIDAS DE DESEMPENHO DOS MELHORES MODELOS

Modelo	Roc auc	Acurácia	F1	Revocação	Precisão
SVM	0.699405	0.746976	0.810818	0.857143	0.776633
Regressão Logística	0.686147	0.737601	0.805961	0.857143	0.767044
Rede Neural	0.679275	0.705746	0.765776	0.763095	0.78106

Comparando tanto as métricas como a curva de aprendizado, conclua-se que, mesmo com números relativamente piores, a rede neural foi capaz de predizer probabilidades melhores do que os outros modelos, inferindo assim na melhor pontuação.

V. ESTRATÉGIA FINAL

Após seleção de features, o primeiro modelo a ser descartado da estratégia final para o Kaggle foi o Naive-Bayes, por ter apresentado métricas de desempenho baixas até mesmo no conjunto de treino, além de uma revocação baixíssima para a classe negativa, o que indica que o modelo está deixando passar muitos exemplos positivos e classificando-os como negativos, sendo um modelo ruim para o problema. Seguindo essa mesma problemática, o classificador Floresta Aleatória também apresentou o mesmo problema, com boas métricas para a classe positiva, mas ruim para a negativa.

Nos quatro modelos restantes, o primeiro selecionado foi a Rede Neural devido ao seu ótimo desempenho no placar público. O segundo modelo foi escolhido com o critério da área sob a curva roc, e como mostrado na tabela 2, esse modelo foi a Máquina de Vetores de Suporte.

VI. CONCLUSÃO

Infere-se, portanto, que a problemática abordada na introdução foi resolvida. Com o auxílio dos dados, agora há mais uma métrica que pode auxiliar nas decisões médicas além da classificação de risco com base na estrutura genética do paciente. Essa medida, se associada a classificação de risco, possibilita também a segregação dos pacientes dentro do mesmo nível, o que pode ajudar principalmente nas decisões a serem tomadas para os pacientes enquadrados no nível intermediário de risco.

Mesmo não sendo o melhor placar no Kaggle, utilizar de medidas estatísticas traz maior confiança para as decisões, e por ser um problema médico, esse tipo de medida é fundamental para trazer a confiança do próprio profissional que irá analisá-lo.

REFERENCES

[1] Velloso, Elvira Deolinda Rodrigues Pereira, et al. "Alterações citogenéticas e moleculares em leucemia mieloide aguda: revisão e descrição de casos." Einstein (São Paulo) 9 (2011): 184-189.

[2] Kumar, C. Chandra. "Genetic abnormalities and challenges in the treatment of acute myeloid leukemia." Genes & cancer 2.2 (2011): 95-107.