EXPLORANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE MORTE POR DOENÇA DE CHAGAS

FILHO, A.F.C[1].; SANTOS, L.I.[2];

[1]Discente do curso superior em Ciência da Computação do IFNMG – *Campus* Montes Claros; [2]Docente do IFNMG – *Campus* Montes Claros.

Introdução

A doença de Chagas (DC) é um problema de saúde pública que vem sendo frequentemente negligenciado na América Latina nos últimos anos. Ferreira et al. (2022) apresentam um estudo que utiliza técnicas de Aprendizado de Máquina (AM) para prever a morte de pacientes acometidos por DC, a partir de dados referentes a sua condição de saúde, em um período de dois anos. Esse estudo reconhece a gravidade da DC e, em especial, a falta de oportunidade das vítimas de receber tratamento ou sequer serem diagnosticadas, o que é comum em áreas cujo acesso aos serviços de saúde de qualidade é difícil.

No contexto da DC, ferramentas de predição de morte podem ser úteis, pois elas têm a capacidade de agilizar prognósticos de agentes de saúde, além de economizar recursos de organizações correlatas, visto que são capazes de identificar pessoas em risco e então orientar atendimento prioritário para elas.

Este estudo tem como objetivo explorar outras abordagens e modelos de AM, sob a justificativa de que foram identificados alguns problemas no conjunto de dados usado, mas que não foram tratados em Ferreira et al. (2022). Portanto, neste estudo foi utilizado uma abordagem alternativa para lidar com o problema de desbalanceamento de dados, o uso da técnica de imputação de dados proposta por Cielen et al. (2016, p. 34) para lidar com o problema de ausência de dados e também, a utilização de variáveis *dummy* (CIELEN et al., 2016, p. 42) em variáveis categóricas.

Material e Métodos

Abordagem computacional

O Aprendizado de Máquina refere-se à capacidade de computadores aprenderem com um grande volume de dados e depois fazer predições e generalizações. Existem vários modelos pré-definidos de AM, mas neste estudo foram usados: Árvore de Decisão (Decision Tree) (QUINLAN, 1990), Floresta Aleatória (Random Forest) (BREIMAN, 2001) e RUSBoost (SEIFFERT et al., 2008).

Para a implementação, foi utilizado a linguagem de programação Python 3.9 junto às bibliotecas: ScikitLearn 1.3, ImbLearn 0.11, Pandas 2 e NumPy 1.25. A primeira para o uso dos modelos de AM, e as últimas para balanceamento e manipulação de dados.

Coleção de dados

Para que o AM seja efetivo, é necessário uma coleção ampla e correta de dados empíricos de onde os computadores possam aprender, isto é, determinar padrões. O conjunto de dados usado neste estudo é o mesmo apresentado e utilizado por Ferreira et al. (2022). Trata-se de um banco de dados proveniente da entrevista e examinação de 1694 pacientes portadores de DC. Foram pesquisadas 48 variáveis, sendo 33 delas adquiridas por entrevista com o paciente e as únicas que foram consideradas neste trabalho. Após o período de dois anos, 134 dessas pessoas faleceram.

Ressalta-se que todos os indivíduos deram consentimento para terem seus dados colhidos e a pesquisa foi devidamente aprovada em comitês de ética competentes (FERREIRA et al., 2022).

Propostas de melhoria

Na coleção de dados dos pacientes de DC, como já mencionado, apenas 134 pessoas faleceram dentre 1694, ou seja, aproximadamente 8 % dos pacientes. Isso indica um desbalanceamento de dados e faz os algoritmos de AM terem maior dificuldade em reconhecer o caso minoritário. Ferreira et al. (2022) utilizaram o Synthetic Minority Oversampling Technique (SMOTE) para fazer o balanceamento. Neste estudo foi utilizado o SMOTE ENN (Edited Nearest Neighbor), um aprimoramento do SMOTE (VIADINUGROHO, 2021).

Um outro problema na coleção de dados é a ausência de alguns valores em certos registros. Uma forma simples de resolver esse problema é ignorando os registros com lacunas (método utilizado em Ferreira et al.), entretanto se pode perder muitos registros fazendo isso. Neste estudo, testou-se métodos de imputação de dados, que consiste em preencher os valores ausentes artificialmente, tendo como base outros registros do banco de dados; foi testada tanto o uso de médias, como o K-Nearest Neighbors (KNN), um algoritmo mais robusto apresentado em (BATISTA et al., 2004).

Um último problema identificado é o uso de variáveis qualitativas nominais que foram tratadas inadequadamente como ordinais. Por exemplo, na coleção de dados, a variável cor de pele é assinalada de maneira contraintuitiva: cada valor dessa variável se associa a alguma etnia arbitrária. Isso potencialmente confunde os algoritmos de AM, contudo essa questão não foi endereçada no modelo original. Um agravante é o fato de que a cor da pele é um fator importante na predição de morte por DC (FERREIRA et al., 2022). Portanto, todas as variáveis com assinalação incorreta foram removidas e, no lugar, foram acrescentadas variáveis auxiliares ou *dummies*. Dessa forma, a cor da pele foi substituída por variáveis representando cada etnia (amarela, branca etc.), evitando a confusão por parte dos métodos de AM.

Medição de eficácia

Para a medição da qualidade e eficácia dos modelos preditivos desenvolvidos, utilizou-se a especificidade, sensibilidade e a *G-mean*. Neste estudo, essas foram as medidas-alvo cujos valores espera-se a maximização. Elas são calculadas por meio do número de pacientes fora de risco (TN), o número de pacientes falsamente classificados como em risco pela máquina (FP), o número real de pacientes em risco (TP) e o número de pacientes falsamente classificados como fora de risco pela máquina (FN). A especificidade (equação 1) está relacionada com a razão de acerto da máquina em classificar indivíduos como fora de risco de morte (LEKHTMAN, 2019), enquanto que a sensibilidade (equação 2) indica a razão de acerto da máquina ao identificar indivíduos que irão falecer nos próximos dois anos (LEKHTMAN, 2019).

$$especificidade = \frac{TN}{TN + FP} \quad (1)$$

$$sensibilidade = \frac{TP}{TP + FN} \quad (2)$$

$$gmean = \sqrt{especificidade \times sensibilidade} \quad (3)$$

Deseja-se alta sensibilidade para se encontrar precisamente pacientes com alto risco de morte, mas também alta especificidade para garantir que não esteja ocorrendo uma superestimação de pacientes em risco. A forma utilizada para se conciliar ambas as medidas é a *G-mean* (equação 3), que consiste na média geométrica desses dois valores.

Resultados e Discussão

As medidas-alvo encontradas para cada modelo estão todas expressas na Tabela 1. A utilização das árvores de decisão (Decision Tree) foi a que apresentou os piores resultados da *G-mean*; apesar de possuir a maior especificidade, tem uma sensibilidade muito baixa, indicando uma tendência de classificar os pacientes como fora de risco mesmo com o balanceamento de dados. Seguido desse modelo, está o Random Forest que foi capaz de gerar um resultado consideravelmente melhor. Finalmente, o RUSBoost, o qual teve um desempenho melhor que o Random Forest não apenas pela *G-mean*, mas também pelo equilíbrio entre a sensibilidade e a especificidade.

Considerações finais

Este trabalho concluiu que — com os modelos, parâmetros e técnicas aqui abordadas — não foi possível ampliar o poder preditivo dos métodos baseados em AM apresentados em estudos anteriores, mesmo com todas as melhorias descritas acima.

Agradecimentos

Os autores agradecem o IFNMG — *Campus* Montes Claros por contemplá-los com uma bolsa de iniciação científica através do Programa Institucional de Bolsas de Iniciação Científica (PIBIC).

Referências

BATISTA, Gustavo EAPA; PRATI, Ronaldo C.; MONARD, Maria Carolina. **A study of the behavior of several methods for balancing machine learning training data.** ACM SIGKDD explorations newsletter, v. 6, n. 1, p. 20-29, 2004.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, p. 5-32, 2001.

CIELEN, Davy; MEYSMAN, Arno D. B.; ALI, Mohamed. **Introducing Data Science.** Shelter Island, NY: Manning Publications, 2016.

FERREIRA, Ariela Mota et al. **Two-year death prediction models among patients with Chagas Disease using machine learning-based methods.** PLoS neglected tropical diseases, v. 16, n. 4, p. e0010356, 2022.

LEKHTMAN, Alon. **Data Science in Medicine - Precision & Recall or Specificity & Sensitivity?**. Medium 2019. Disponível em: <towardsdatascience.com/should-i-look-at-precision-recall-or-specificity-sensitivity-3946158aace1> Acesso em: 12 ago. 2022.

SEIFFERT, Chris et al., **RUSBoost: Improving classification performance when training data is skewed**. 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 2008, pp. 1-4, doi: 10.1109/ICPR.2008.4761297.

QUINLAN, J. Ross. Decision trees and decision-making. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 20, n. 2, p. 339-346, 1990.

Tabela 1. Resultados dos modelos usados.

	Especificidade	Sensibilidade	*G-mean*
Decision Tree	0,87	0,28	0,49
Random Forest	0,75	0,50	0,61
RUSBoost	0,66	0,61	0,63

Fonte: Dados da pesquisa (2022).