XXXII ENCONTRO ANUAL DE INICIAÇÃO CIENTÍFICA

Prognóstico para Pacientes com COVID-19 Usando Técnica de Aprendizado de Máquina

Willian Penteado (PIBIC/Fundação Araucária-UNICENTRO),
Sandra Mara Guse Scós Venske (Orientadora), Carolina Paula de Almeida
e-mail: 57020140062@unicentro.edu.br, ssvenske@unicentro.br,
carol@unicentro.br
Universidade Estadual do Centro-Oeste – UNICENTRO,
Departamento de Ciência da Computação. Guarapuava, Paraná.

ENGENHARIAS - ENGENHARIA DE PRODUÇÃO.

Palavras-chave: Aprendizado supervisionado, Problema de classificação, COVID-19, AdaBoost.

Resumo

A doença do coronavírus 2019 (COVID-19) é uma doença infecciosa altamente contagiosa causada pelo coronavírus, relacionada à síndrome respiratória aguda grave 2 (*Severe Acute Respiratory Syndrome Coronavirus* - SARS-CoV-2). A doença teve um efeito catastrófico no mundo, resultando em milhões de mortes. Este estudo teve como objetivo utilizar Aprendizado de Máquina, que é o processo de indução de uma hipótese a partir de experiências anteriores, para o desenvolvimento de um modelo para apoiar o prognóstico de pacientes com COVID-19. Como conjunto de dados foram utilizados dados públicos pertencentes ao Sistema Único de Saúde (SUS) do Ministério da Saúde do Brasil para o ano de 2022. Os dados do SUS eram altamente ruidosos, inconsistentes, duplicados, ausentes e desbalanceados. Técnicas de pré-processamento foram aplicadas resultando em conjuntos de dados que podem ser utilizados em outros estudos.

Introdução

A Covid-19 é uma grave doença respiratória causada pelo coronavírus SARS-CoV-2, altamente transmissível e globalmente disseminada. Durante a pandemia da Covid-19 muitos dados foram gerados relacionados à doença. O Aprendizado de Máquina (AM) pode ser uma alternativa para utilização destes

dados, pois é a ciência da programação de computadores a fim de que eles possam aprender com base nos dados, sem serem programados para isso (Géron, 2019).

Este trabalho propôs o estudo do aprendizado supervisionado de máquina, mais especificamente do algoritmo AdaBoost (Géron, 2019), a ser aplicado para o desenvolvimento de um modelo para apoiar o prognóstico de pacientes com COVID-19. O conjunto de dados utilizado é público e pertencente ao Sistema Único de Saúde (SUS) do Ministério da Saúde do Brasil[1] correspondente à vigilância da Síndrome Respiratória Aguda Grave (SRAG) no Brasil no ano de 2022. O conjunto de dados é composto por exames clínicos, sintomas, datas, dados demográficos e geográficos, sendo inicialmente composto por 561.242 instâncias e 166 atributos, com uma grande quantidade de valores ruidosos, inconsistentes, duplicados e ausentes, além de ser desbalanceado. Técnicas de pré-processamento foram aplicadas a fim de padronizar o conjunto de dados com instâncias corretas para que o modelo de predição gerado pela técnica de AM Adaboost fosse confiável. O tratamento de dados ajuda a melhorar a qualidade dos dados, eliminando ruídos, *outliers* e valores ausentes. Ao final, diferentes conjuntos de dados tratados foram gerados e disponibilizados, variando nas técnicas de balanceamento e na aplicação ou não da normalização dos dados.

Material e métodos

A metodologia adotada é apresentada na Figura 1, dividida em três estágios: análise e estudo do conjunto de dados, pré-processamento (como foco principal) e modelagem. Algumas das principais técnicas de pré-processamento aplicadas foram (Faceli et al 2021): 1) Delimitação do conjunto de dados para apenas notificações de Covid-19 no estado do Paraná; 2) Eliminação manual de atributos categorizados como Campos Internos, relacionados à localização e dados administrativos; 3) Limpeza de dados redundantes e inconsistentes; 4) Aplicação de um filtro para remover atributos e instâncias que possuíam mais de 50% dos seus dados incompletos; 5) Preenchimento dos dados ausentes com a mediana; 6) Tratamento de datas; 7) Conversão de tipos de dados; 8) Normalização de dados com o método *StandardScaler*[2]; e 9) Balanceamento dos dados.

Para o desenvolvimento deste projeto, foi utilizado o *framework* *Scikit-learn*[3]. O código-fonte do algoritmo desenvolvido está disponível num repositório *online*[4].

[1] https://opendatasus.saude.gov.br/dataset/srag-2021-a-2023/resource/62803c57-0b2d-4bcf-b114-380c392fe825
[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
[3] https://scikit-learn.org/stable/
[4] https://github.com/Willian-P/IniciacaoCientifica2-Covid19

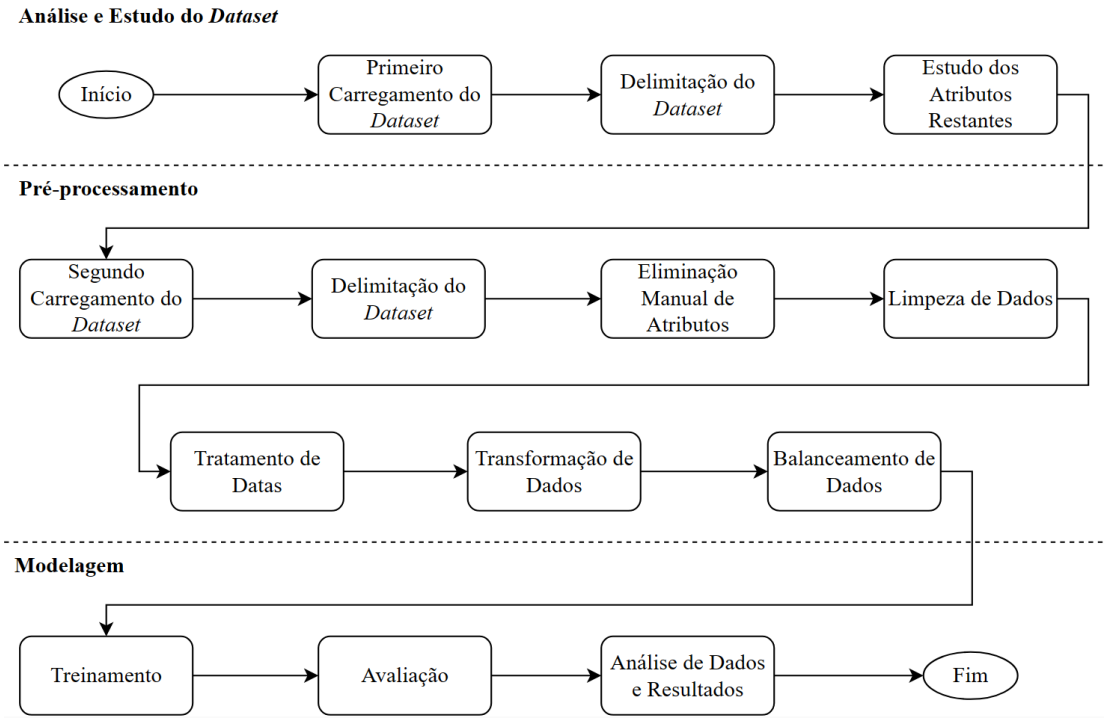

Figura 1 - Fluxograma da metodologia utilizada no projeto.

Resultados e Discussão

Cinco versões foram usadas como conjunto de dados (subconjuntos de dados do SUS): 1) Versão *Nenhum*: sem balanceamento dos dados, mantendo o conjunto original com 17.432 instâncias e classes desbalanceadas; 2) Versão *NearMiss*: a técnica *Near-Miss*[5] foi aplicada para equilibrar as classes, reduzindo o conjunto de dados para 7.822 instâncias; 3) Versão *SMOTE100*: a técnica SMOTE[6] foi aplicada para igualar em 100% a classe minoritária à classe majoritária, resultando em 27.042 instâncias; 4) Versão *SMOTE85*: a técnica de balanceamento SMOTE foi empregada para elevar a classe minoritária a 85% da classe majoritária, resultando em 25.013 instâncias; 5) Versão *DadosSP*: envolveu o uso de dados de São Paulo para expandir a classe minoritária a 86,5% da classe majoritária, elevando o conjunto de dados para 25.210 instâncias. Essas versões foram criadas para avaliar o desempenho dos modelos sob diferentes técnicas de balanceamento de classes.

[5] https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.NearMiss.html
[6] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

A Tabela 1 mostra os resultados em termos de acurácia (mede a proporção de previsões corretas no modelo) e F1-*score* (mede a qualidade geral do modelo), com e sem normalização dos dados para as 5 versões consideradas.

Tabela 1 - Resultados em termos de acurácia e F1-*score*, com e sem normalização dos dados para as 5 versões consideradas.

Balanceamento	Acurácia		F_1-*score*	
	Sem normalizar	Normalizado	Sem normalizar	Normalizado
Nenhum	0.795	0.795	0.328	0.328
NearMiss	0.788	0.800	0.787	0.787
SMOTE100	0.833	0.854	0.834	0.854
SMOTE85	0.828	0.843	0.814	0.832
DadosSP	0.772	0.768	0.762	0.759

Há uma diferença significativa em termos de acurácia e F1-*score* para dados desbalanceados, sendo o F1-*score* mais robusto. Para as outras versões as variações entre as métricas foram sutis. A análise mostra que a versão *SMOTE100 Normalizado* foi superior, atingindo 85,4% em ambas as métricas. Em segundo lugar, a versão *SMOTE85 Normalizado* obteve 83,2% de F1-*score* e a versão *SMOTE100 Sem normalizar* obteve 83,3% e 83,4% de acurácia e F1-*score*, respectivamente.

Agradecimentos

Os autores agradecem à Fundação Araucária pelo apoio financeiro.

Considerações Finais

Este trabalho teve como objetivo o uso de AM com a técnica Adaboost aplicada para o desenvolvimento de modelo de apoio ao prognóstico de COVID-19.

Para o desenvolvimento, aplicaram-se técnicas de pré-processamento nos dados, já que estes apresentaram diversos problemas. Isso resultou em conjuntos de dados tratados para uso em estudos posteriores. Além disso, na criação de um modelo de AM com Adaboost, explorando técnicas de balanceamento.

Como trabalho futuro, pode-se aplicar a metodologia aqui proposta para o pré-processamento dos dados considerando outros Estados, ou até o Brasil todo.

Referências

GÉRON, A. (2019). **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow.** Alta Books.

Faceli, K., Lorena, A. C., Gama, J. e Carvalho, A. (2021). **Inteligência artificial: uma abordagem de aprendizado de máquina.** LTC, 2nd edition.