Aplicação de Inteligência Artificial para Extração de Características e Predição a Partir de Dados Hospitalares no Diagnóstico de Sequelas Pós-Aguda da Covid-19

Ruann C. C. Farrapo[1], Sara D. de Souza[2],Márcio A. B. Amora[3], Iális C. de P. Júnior[4]

[1,3,4]Programa de Pós Graduação em Engenharia Elétrica e Computação – PPGEEC
Universidade Federal do Ceará – (UFC), Sobral - CE, CEP 62.010-560, Brasil

[2]Hospital Regional Norte, Sobral - CE, CEP 62031-305, Brasil

{ruann.campos_01@alu,marcio@dee,ialis@sobral}.ufc.br, saraenf83@hotmail.com

Abstract. *The Post-Acute Sequelae of COVID-19 (PASC) characterize a global crisis. Thus, it is important to apply an extractive and predictive analysis of the SPAC's. Considering what was exposed above, this work was developed from the application of Artificial Intelligence (AI) techniques. Predictions were built using Decision Tree (DA), Random Forest (FA) and Artificial Neural Network (ANN). The database consists of real data from patients registered from laboratory tests and questionnaires. The data were divided into training, validation and testing. Results were obtained using the AD, FA and ANN models. Some of these results are: 93% for Accuracy, 88% for Precision and 91% for F1.*

Resumo. *As Sequelas Pós-Aguda da COVID-19 (SPAC) caracterizam uma crise global. Assim, torna-se importante a aplicação de uma análise extrativa e preditiva das SPAC's. Considerando o que foi exposto acima, este trabalho foi desenvolvido à partir da aplicação de técnicas de Inteligência Artificial (IA). As predições foram construídas utilizando Árvore de Decisão (AD), Floresta Aleatória (FA) e Rede Neural Artificial (RNA). O banco de dados é composto por dados reais de pacientes registrados a partir de exames laboratoriais e questionários. Os dados foram divididos em treino, validação e teste. Foram obtidos resultados com a utilização dos modelos AD, FA e RNA. Sendo alguns desses resultados: 93% de Acurácia, 88% de Precisão e 91% de F1.*

1. Introdução

As enfermidades respiratórias são umas das razões mais pertinentes no quesito morbidade e mortalidade em todo o mundo, sendo que sua maioria é causado por um patógeno viral. Associado a isso, com o passar dos anos, tornou-se crescente a descoberta de vírus causadores de doenças respiratórias. Entre elas, estão os coronavírus (Covs), uma linhagem de vírus mutagênicos que infectam tanto humanos como animais silvestres, como algumas espécies de macacos, morcegos e entre outros [de Albuquerque et al. 2020].

No fim de 2019, um novo beta coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2), o COVID-19, se espalhou pela China inicialmente e depois para todas as regiões do planeta. O patógeno viral provocava pneumonia agressiva e uma aguda insuficiência pulmonar. A doença era tão agressiva que em muitos indivíduos

a situação se agravava rapidamente, levando à morte em um curto período de tempo [de Albuquerque et al. 2020].

De acordo com [Bragatto et al. 2021], o SARS-CoV-2 possui material genético do tipo RNA com a superfície incrustada de proteínas *spike* (S) que possibilitam a entrada na célula. As manifestações clínicas estão relacionadas à resposta imune do hospedeiro e se caracterizam por acometimentos em diversos órgãos e sistemas. A afecção causada pelo coronavírus pode afligir fortemente o indivíduo infectado, trazendo sintomas leves, moderados ou também bastante graves. Por consequência, é possível ocorrerem diversas sequelas, as quais podem manter-se por tempo indefinido.

As Sequelas Pós-Aguda do COVID-19 (SPAC) representam uma crise global emergente. Entre 31% a 69% dos pacientes com COVID-19 sofrem de sequelas [Groff et al. 2021]. Há também o COVID longo, que é definido como uma série de problemas de saúde novos, recorrentes ou contínuos em que os enfermos podem experimentar quatro ou mais semanas após o início da infecção por SARS-CoV-2. Associado a isso, o SPAC pode incluir perda de memória, problema gastrointestinal, fadiga, anosmia, falta de ar e outros sintomas [Huang et al. 2021].

O crescimento atual da Inteligência Artificial (IA) através de suas subáreas como por exemplo, o aprendizado de máquina, fez essa tecnologia tornar-se importante no diagnóstico e tratamento médico. As técnicas de aprendizado de máquina podem ser usadas para classificar os resultados de testes para COVID-19 por meio da análise conjunta de exames.

Neste trabalho é proposto a construção de uma metodologia extrativa e preditiva que analisa os aspectos das sequelas em pacientes infectados com COVID-19 fazendo tratamento dos dados e aplicando técnicas de aprendizado de máquina. Com isso, busca-se predições construídas empregando AD, FA e RNA.

O artigo está divido em seções citadas a seguir. Na seção 2 é apresentada uma revisão bibliográfica de métodos preditivos e analíticos a partir de dados de COVID-19 e suas sequelas. Na seção 3 é apresentada a metodologia do trabalho, em que: é exposta a estratégia de pesquisa, é apresentada a base de dados utilizada, são apresentados o tratamento e divisão dos dados, também são fornecidos os conceitos de AD, FA, RNA e além disso, são apresentadas as métricas de resultados utilizadas. Na seção 4 são apresentados e discutidos os resultados obtidos da AD, FA e da RNA desenvolvidas. Na seção 5 são apresentadas as conclusões deste trabalho e suas perspectivas futuras.

2. Trabalhos Relacionados

O trabalho de [Mueller et al. 2022], propõe uma abordagem de aprendizado de máquina usando soro pró-inflamatório, anti-inflamatório e medições de anticorpos anti-SARS-CoV-2 como dados de entrada. É fornecido um esquema baseado no tipo imunológico para estratificar pacientes com COVID-19 na admissão hospitalar em categorias clínicas de alto e baixo risco com perfis distintos de citocinas e anticorpos que podem orientar a terapia personalizada. Foi aplicado um modelo hierárquico de aprendizado não supervisionado com 83% de acerto na identificação dos grupos através da aplicação de *clusters*.

No trabalho de [Sayed et al. 2021], é feito um modelo que prever diferentes níveis de riscos em pacientes com COVID-19 baseando-se em imagens de raios-X com técnicas

de aprendizado de máquina. Um modelo pré-treinado profundo *CheXNet* e técnicas artesanais híbridas foram usadas para extrair recursos, sendo elas: Análise de Componentes Principais (PCA) e Eliminação de Recurso Recursivo (RFE). Como resultado, o classificador *XGBoost* obteve o melhor desempenho com os recursos mesclados (PCA + RFE), onde alcançou 97% de acurácia, 96% f1-score e 100% roc-auc. Com uso de *Support Vector Machine* (SVM) foram obtidos 97% de acurácia, 95% f1-score e 99% ROC-AUC.

3. Metodologia

3.1. Base de Dados

A base de dados utilizada foi construída e apresentada no trabalho de [Su et al. 2022]. Os autores disponibilizam de forma pública a base de dados com a permissão dos hospitais e pacientes, para que assim, essas informações clínicas e hospitalares possam ser usados em pesquisas que constroem novos modelos de predições e análises sobre as SPAC's.

A base é composta por dados de pacientes que foram identificados em hospitais e clínicas afiliadas localizadas em regiões da cidade de *Seattle*, nos Estados Unidos. Os pacientes internados foram hospitalizados no *Harborview Medical Center, UW Medical Center Montlake,* ou *UWMedical Center Northwest* sendo inscritos durante a internação hospitalar. Os pacientes ambulatoriais foram identificados por meio de um sistema de alerta de laboratório. Posteriormente, todos os participantes foram solicitados a retornar 60 ou 90 dias depois para acompanhamento, onde foram entrevistados sobre os sintomas. Além disso, as coletas de sangue foram tomadas durante esses retornos.

O banco de dados utilizado é composto por atributos de entradas que trazem informações à partir das entrevistas e exames feitos para 525 pacientes. As classes de saídas para a construção dos modelos, são quatro sequelas manifestadas nos pacientes após infecção por covid, sendo elas: Asma, Hipertensão, Insuficiência cardíaca congestiva e Doença arterial coronária. As sequelas que servirão como classes de saídas, são comorbidades que trazem muitos problemas para aqueles que convivem com as mesmas.

Considerando a explanação acima sobre o tamanho da base de dados e sobre as classes de saída, destaca-se que 174 pacientes foram diagnosticados com Hipertensão, 181 com Asma, 182 com Insuficiência cardíaca congestiva e 190 com Doença arterial coronária.

Os atributos de entrada para a construção das predições, são características e informações coletadas a partir dos exames clínicos, exames sanguíneos e entrevistas feitos com os pacientes. Na Tabela 1, são apresentados os atributos de entrada para a construção dos modelos preditivos das classes de saídas, sendo essas, as sequelas estudadas nesta pesquisa.

Tabela 1. Tabela dos Atributos de entrada das 4 classes.

Atributos	Classes
Idade, IMC, Cortisol, Nicotinamida, Fosfato, Glicose, Creatina, Tiroxina, Início dos sintomas, Dias de observação, Temperatura, Pulso,Colesterol,Serotonina,Colesterol,Leucina, Glicina, Glicocolato, Citrulina, Lisina, Metionina, Ornitina, Orotato, Serina, Urato, Colina.	Hipertensão, Asma, Doença Arterial Coronária, Insuficiência Cardíaca Congestiva

3.2. Tratamento e Divisão dos Dados

Inicialmente as classes de saídas são variáveis categóricas, tendo como saída *Yes*, que são pacientes que constataram a sequela e *No* para os pacientes que não constataram. Associado a isso, para possibilitar a construção dos modelos, o primeiro tratamento foi transformar as saídas categóricas em binárias, com 1 para sequela constatada e 0 para não constatada. Associado a isso, também foram tratadas linhas nulas da base de dados. Além de excluir algumas linhas que apresentavam boa parte dos atributos nulos, também foram aplicadas técnicas como por exemplo, a média dos valores de determinada coluna em linhas importantes e que normalmente só apresentavam um atributo com valor nulo.

Ressalta-se que foi aplicado uma técnica para balanceamento dos dados entre as classes. As mesmas apresentavam um desbalanceamento entre as suas saídas. O balanceamento foi feito utilizando o método *sampling* que é um pré-processamento que visa minimizar as discrepâncias entre as classes por meio de uma reamostragem. Para gerar esse balanceamento, foi utilizado a técnica *oversampling*, que cria novas observações da classe minoritária à partir das informações contidas nos dados originais [Liberty et al. 2016].

Para a criação dos modelos, o banco de dados foi dividido em 70% para treinamento, 15% para validação e 15% para teste. Na validação, foi utilizado um método de busca randômica chamado *random search* para a definição dos melhores hiperparâmetros dos modelos.

O *random search* se configura como uma grade de valores de hiperparâmetros e seleciona combinações aleatórias para treinar o modelo. Isso permite controlar explicitamente o número de combinações de parâmetros que são tentadas. O número de iterações de pesquisa é definido com base no tempo ou nos recursos [Liberty et al. 2016].

Depois do processo de treinamento e validação, os modelos já treinados são testados com elementos de teste que não foram utilizados em suas criações.

3.3. Árvore de Decisão

Uma árvore de decisão é um algoritmo de aprendizado de máquina supervisionado que é utilizado para classificação e para regressão. Assim como um fluxograma, a árvore de decisão estabelece nós de decisão que se relacionam entre si por uma hierarquia. Existe o nó-raiz, que é o mais importante, e os nós-folha, que são os resultados finais. No contexto do aprendizado de máquina, o raiz é um dos atributos da base de dados e o nó-folha é a classe ou o valor que será gerado como resposta [Wang et al. 2019].

Ressalta-se que foi utilizado tanto para a AD, como todos os outros modelos, a aplicação da otimização de hiperparâmetro *random search* e o *framework* da linguagem *Python* scikit learn, que serviu através de seus métodos como base para a estruturação de todos os modelos construídos.

3.4. Floresta Aleatória

Floresta Aleatória (*random forest*) é um algoritmo de aprendizagem supervisionada. A "floresta" criada é uma combinação (*ensemble*) de árvores de decisão, na maioria dos casos treinados com o método de *bagging*. O método de *bagging* é um meta- algoritmo em que é possível construir classificadores agregados. O método gera subconjuntos de exemplos através de sorteios aleatórios simples com reposição, sobre o conjunto de dados de treinamento original [Gupta et al. 2021].

3.5. Redes Neurais

Rede Neural Artificial (RNA) pode ser definida como uma estrutura complexa interligada por elementos de processamento simples, chamados de neurônios, que possuem a capacidade de realizar operações como cálculos em paralelo, para processamento de dados e representação de conhecimento [Gupta et al. 2013].

A Rede Neural utilizada para a construção dos modelos foi a *Perceptron* Multicamadas (MLP — *Multi Layer Perceptron*) é uma rede neural que, como pode-se ver na Figura 1, adaptada de [Batista 2012], possui no mínimo três camadas: a camada de entrada, a camada escondida que podem ser múltiplas, e a camada de saída. A propagação dos dados, ou sinal de entrada, é na ordem direta desde a entrada até a saída da rede, camada a camada. As conexões entre os neurônios são retratadas por pesos que indicam força ou importância das conexões dos neurônios. O aprendizado da rede baseia-se nos ajustes iterados dos pesos e dos valores de *bias* nos neurônios [Batista 2012].

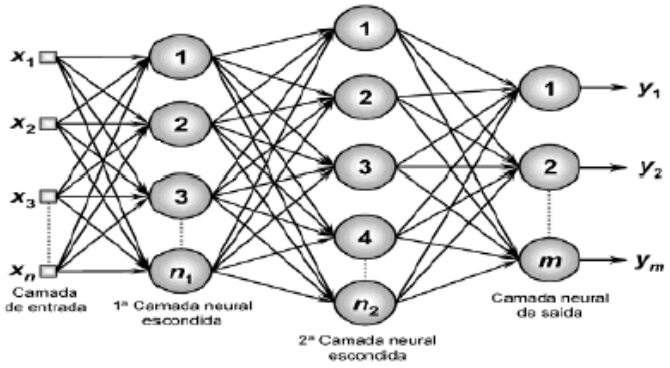

Figura 1. Arquitetura de uma rede neural multicamada perceptron.

3.6. Métricas para os Resultados

Como métricas para os resultados, foram adotados: Acurácia, Precisão e a Média Harmônica - F1 (entre Precisão e *Recall*). Para calculá-los, os resultados foram primeiramente analisados e rotulados com os seguintes nomes: Verdadeiro Positivo - VP (quando o modelo declara que a classe é positiva e, ao verificar a resposta, vê-se que a classe era realmente positiva), Falso Positivo - FP (quando o modelo declara que a classe é positiva, mas ao verificar a resposta, vê-se que a classe era negativa), Verdadeiro Negativo - VN (quando o modelo declara que a classe é negativa e, ao verificar a resposta, vê-se que a classe era realmente negativa) e Falso Negativo - FN (quando o modelo declara que a classe é negativa, mas ao verificar a resposta, vê-se que a classe era positiva) [Kamei et al. 2012].

A Acurácia avalia simplesmente o percentual de acertos, ou seja, ela pode ser obtida pela razão entre a quantidade de acertos e o total de entradas, equação 1, [Kamei et al. 2012]:

$$Acurácia : \frac{VP + VN}{VP + VN + FP + FN}. \tag{1}$$

A Precisão é uma métrica que avalia a quantidade de verdadeiros positivos de uma determinada classe, sobre a soma de todos os valores positivos, equação 2, [Kamei et al. 2012]:

$$Precisão : \frac{VP}{VP + FP}. \tag{2}$$

A métrica *Recall* é utilizada no cálculo da Média Harmônica - F1. O *Recall* é a divisão entre casos que foram erros e a previsão que estava correta (VP), pela soma dos casos VP com os casos falsos negativos (FN), equação 4, [Kamei et al. 2012]:

$$Recall : \frac{VP}{VP + FN}. \tag{3}$$

A Média Harmônica - F1 representa a média harmônica entre Precisão e *Recall*, equação 4, [Kamei et al. 2012]:

$$F1 : \frac{2 * RECALL * PRECISÃO}{RECALL * PRECISÃO}. \tag{4}$$

4. Resultados

Nesta seção são apresentados os resultados de acurácia, Precisão e F1 dos modelos para cada uma das sequelas (classes de saída). Além disso, nesta seção são apresentadas os valores médios dos resultados alcançados e também comparações dos resultados com os de outros trabalhos relacionados.

Os resultados de acurácia obtidos com os modelos para as quatro sequelas, estão na Tabela 2, elaborada pelos autores. Fazendo a comparação dos resultados de Acurácia dos modelos construídos e apresentados na Tabela 2, observa-se que: para a sequela de hipertensão, o modelo de AD apresentou um melhor resultado; para a Asma e Insuficiência Cardíaca, o modelo de RNA MLP apresentou melhores resultados; para a sequela de Doença Arterial Coronária, o modelo de FA apresentou um melhor resultado. Ao se analisar a média geral de acurácia, considerando todas as sequelas, nota-se que os modelos FA e RNA apresentam os maiores valores de médias, tendo os dois tipos de modelos uma média de acurácia de 88%.

Tabela 2. Resultados Acurácia

Sequelas	Árvore de Decisão	Floresta Aleatória	Rede Neural
Hipertensão	83%	81%	75%
Asma	82%	87%	93%
Doença Arterial Coronária	83%	91%	90%
Insuficiência Cardíaca C.	89%	92%	93%
Média Geral	84%	88%	88%

Os resultados de precisão obtidos com os modelos para as quatro sequelas estão na Tabela 3, elaborada pelos autores. Fazendo a comparação dos resultados de Precisão dos modelos construídos e apresentados na Tabela 3, observa-se que: para a sequela de hipertensão, o modelo de AD apresentou um melhor resultado; para a Asma e Insuficiência Cardíaca, o modelo de RNA MLP apresentou melhores resultados; para a sequela de Doença Arterial Coronária, o modelo de FA apresentou um melhor resultado. Ao se observar a média geral de precisão, considerando todas as sequelas, nota-se que o modelo RNA apresenta o maior valor de média, tendo uma média de precisão de 86%.

Tabela 3. Resultados Precisão

Sequelas	Árvore de Decisão	Floresta Aleatória	Rede Neural
Hipertensão	80%	78%	76%
Asma	79%	80%	90%
Doença Arterial Coronária	79%	90%	88%
Insuficiência Cardíaca C.	88%	89%	90%
Média Geral	81%	84%	86%

Os resultados de F1 obtidos com os modelos para as quatro sequelas estão na Tabela 4, elaborada pelos autores. Fazendo a comparação dos resultados de F1 dos modelos construídos e apresentados na Tabela 4, observa-se que: para a sequela de hipertensão, o modelo de AD apresentou um melhor resultado; para a Asma e Insuficiência Cardíaca, o modelo de RNA MLP apresentou melhores resultados; para a sequela de Doença Arterial Coronária, o modelo de FA apresentou um melhor resultado. Ao se analisar a média geral de F1, considerando todas as sequelas, percebe-se que o modelo RNA apresenta o maior valor de média, tendo uma média de F1 de 86%.

Tabela 4. Resultados F1

Sequelas	Árvore de Decisão	Floresta Aleatória	Rede Neural
Hipertensão	78%	77%	74%
Asma	77%	79%	89%
Doença Arterial Coronária	80%	89%	89%
Insuficiência Cardíaca C.	88%	90%	91%
Média Geral	81%	84%	86%

5. Conclusão e Trabalhos Futuros

Foi possível demonstrar neste trabalho, a partir dos modelos AD, FA e RNA construídos, taxas de resultados de Acurácia que variaram de 75% à 93%, resultados de Precisão que variaram entre 76% à 90% e resultados de F1 que variaram de 74% e 91%. Sendo assim, apresentou-se diferentes resultados que proporcionam embasamento e referencial prático para a base de dados utilizada neste trabalho, já que a mesma ainda não possuía estudos com esse tipo de objetivo preditivo. Logo, ressalta-se que o trabalho apresentou análises relevantes e importantes que podem auxiliar no estudo e no combate das SPAC's.

Como perspectivas futuras, estão a melhora dos modelos já construídos, além da busca de melhores hiperparâmetros e também outros atributos de entradas de exames dos pacientes que possam ser analisados e utilizados para uma construção ainda melhor e mais robustas destes resultados. Além disso, é sempre importante buscar outras métricas de resultados que possam dá cada vez mais robustez e confiança para a solução preditiva e analítica construída.

Referências

Batista, B. C. F. (2012). Soluções de equações diferenciais usando redes neurais de múltiplas camadas com os métodos da descida mais íngreme e levenberg-marquardt. *Universidade Federal do Pará*.

Bragatto, M. G., de Almeida, B. M., de Sousa, G. C., Silva, G. A., Pessoa, L. d. S. G., Silva, L. K., Amorim, L. B., Bar, S. F., and de Sousa, V. T. (2021). Estudo das sequelas neuroanatômicas associadas à síndrome pós-covid-19. *Revista Eletrônica Acervo Saúde*, 13(12):e8759–e8759.

de Albuquerque, L. P., da Silva, R. B., and de Araújo, R. M. S. (2020). Covid-19: origin, pathogenesis, transmission, clinical aspects and current therapeutic strategies. *Revista Prevenção de Infecção e Saúde*, 6.

Groff, D., Sun, A., Ssentongo, A. E., Ba, D. M., Parsons, N., Poudel, G. R., Lekoubou, A., Oh, J. S., Ericson, J. E., Ssentongo, P., et al. (2021). Short-term and long-term rates of postacute sequelae of sars-cov-2 infection: a systematic review. *JAMA network open*, 4(10):e2128568–e2128568.

Gupta, N. et al. (2013). Artificial neural network. *Network and Complex Systems*, 3(1):24–28.

Gupta, V. K., Gupta, A., Kumar, D., and Sardana, A. (2021). Prediction of covid-19 confirmed, death, and cured cases in india using random forest model. *Big Data Mining and Analytics*, 4(2):116–123.

Huang, C., Huang, L., Wang, Y., Li, X., Ren, L., Gu, X., Kang, L., Guo, L., Liu, M., Zhou, X., et al. (2021). 6-month consequences of covid-19 in patients discharged from hospital: a cohort study. *The Lancet*, 397(10270):220–232.

Kamei, Y., Shihab, E., Adams, B., Hassan, A. E., Mockus, A., Sinha, A., and Ubayashi, N. (2012). A large-scale empirical study of just-in-time quality assurance. *IEEE Transactions on Software Engineering*, 39(6):757–773.

Liberty, E., Lang, K., and Shmakov, K. (2016). Stratified sampling meets machine learning. In *International conference on machine learning*, pages 2320–2329. PMLR.

Mueller, Y. M., Schrama, T. J., Ruijten, R., Schreurs, M. W., Grashof, D. G., van de Werken, H. J., Lasinio, G. J., Álvarez-Sierra, D., Kiernan, C. H., Castro Eiro, M. D., et al. (2022). Stratification of hospitalized covid-19 patients into clinical severity progression groups by immuno-phenotyping and machine learning. *Nature communications*, 13(1):915.

Sayed, S. A.-F., Elkorany, A. M., and Mohammad, S. S. (2021). Applying different machine learning techniques for prediction of covid-19 severity. *Ieee Access*, 9:135697–135707.

Su, Y., Yuan, D., Chen, D. G., Ng, R. H., Wang, K., Choi, J., Li, S., Hong, S., Zhang, R., Xie, J., et al. (2022). Multiple early factors anticipate post-acute covid-19 sequelae. *Cell*, 185(5):881–895.

Wang, X.-L., Cao, J.-B., Li, D.-D., Guo, D.-X., Zhang, C.-D., Wang, X., Li, D.-K., Zhao, Q.-L., Huang, X.-W., and Zhang, W.-D. (2019). Management of imported malaria cases and healthcare institutions in central china, 2012–2017: application of decision tree analysis. *Malaria Journal*, 18:1–11.