Aplicação de Técnicas de Pré-processamento e Machine Learning para Diagnóstico de Covid-19

William de Souza Gomes
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC
Santo André, Brasil
gomes.souza@aluno.ufabc.edu.br

Matheus Costa Damasceno
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC
Santo André, Brasil
matheus.costa@aluno.ufabc.edu.br

***Resumo*—Neste estudo é apresentado o uso de técnicas de pré-processamento de dados e machine learning para aplicação no diagnóstico de Covid-19 a partir de análises em resultados de exames clínicos em pacientes com suspeita de contaminação pelo o vírus. Para tal, foram aplicadas técnicas para preparação dos dados e modelos de predição para determinar a presença ou não do vírus SARS-CoV-2 em pacientes com sintomas suspeitos.**

***Palavras-chave*—processamento de dados, machine learning, diagnóstico de Covid-19**

I. Introdução

Em tempos de pandemia, o diagnóstico adequado da doença que em sua forma grave é capaz de levar o paciente a óbito, pode significar a diferença entre a vida e a morte do paciente. A Covid-19, síndrome respiratória causada pela ação do vírus SARS-Cov-2, apresentou altas taxas de transmissão se espalhando por todo o mundo. Alguns de seus principais sintomas envolvem febre, tosse, fadiga e dor de cabeça, e em sua forma mais grave, a doença pode causar a morte do paciente por insuficiência respiratória em alguns dias [2]. Além disso, enquanto não há uma vacina disponível para prevenir a população de contrair o vírus, além do isolamento social e cuidados extras de higiene, a realização de testes mostra-se essencial para que seja possível adotar medidas eficazes no combate à pandemia [1].

A partir da análise de um *dataset* contendo informações de testes de Covid-19 em conjunto com resultados de outros exames coletados simultaneamente, esse trabalho propõe a utilização de técnicas de *machine learning* para realizar o diagnóstico dos pacientes que se submeterem a testes de exames laboratoriais, visando apresentar uma solução que auxilie na identificação de contágio pelo novo coronavírus. Essa abordagem se mostra de relevante importância ao considerar que, em casos de pandemia, o rápido diagnóstico da doença permite a adoção de políticas por parte dos governos e ministérios da saúde para evitar que a situação se torna mais crítica.

II. Metodologia

A. Seleção do dataset

O estudo foi realizado a partir do dataset *Diagnosis of COVID-19 and its clinical spectrum* [3]. Tal dataset é proveniente de dados anônimos de pacientes do Hospital Albert Einstein (São Paulo, Brasil), a partir de amostras coletadas para verificação do SARS-CoV-2 que contêm dados de diversos testes laboratoriais, além do resultado para a Covid-19, podendo este ser positivo positivo ou negativo. O dataset é formado por um total de 5644 amostras e 111 características, incluindo o alvo a ser predito. Além disso, este já foi disponibilizado com média 0 e desvio padrão unitário.

B. Análise preliminar dos dados

Foi realizada uma análise preliminar do *dataset* para identificação de quais dados estavam presentes, seus tipos, bem como a estruturação e organização destes. Devido à ausência significativa de dados, foi executada a remoção de atributos cuja incompletude poderia afetar de forma negativa as análises posteriores. Para tal, inicialmente foram removidas as características que apresentavam mais de 85% dos valores faltantes do total de dados. Além disso, dentre aquele cujo alvo foi positivo, também foram removidas as características com mais de 70% de dados faltantes.

C. Pré-processamento

Conforme será melhor descrito na seção Resultados, a análise preliminar não apresentou resultados formidáveis para a posterior análise dos dados. A partir de então, optou-se por uma nova abordagem, que consistiu em remover inicialmente os atributos que possuíam valor nulo ou constante para todas as amostras. Na sequência, decidiu-se remover as amostras que possuíam menos de 50% do total de atributos restantes. Essa escolha foi adotada para atenuar um problema característico do *dataset*, no qual uma parcela significativa do total de amostras possuíam apenas o rótulo do resultado, ou então poucos atributos, o que configuraria um problema nas etapas seguintes.

Um outro tratamento realizado foi a conversão de dados categóricos em números, de modo que os termos “positive” e “detected” foram substituídos por “1”, enquanto os termos “negative”, “not_detected” e “not_done”, por “0”. Essa etapa apresenta suma importância para a criação de modelos preditivos a partir de dados numéricos.

Após a compreensão do dataset e exclusão das amostras e características faltantes em consideráveis quantidades, a etapa seguinte consistiu em aplicar técnicas de imputação simples de dados, de modo que cada dado faltante foi preenchido com a média da coluna. [6]. Estas foram realizadas visando preencher os dados faltantes, de modo que o dataset pré-processado fosse apresentado de maneira consistente, formado por uma matriz quadrada de tamanho $N_a \times N_c$, onde N_a corresponde ao total de amostras e N_c o total de características. Além disso, foi realizada uma nova normalização, por meio da função *MinMaxScaler* do Scikit-Learn, sendo esta dada pela Equação 1 abaixo:

$$\frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

D. Treinamento

Após feito o processamento dos dados, foi necessário realizar a etapa de treinamento para que fosse possível a identificação da Covid-19. Para tal, foi utilizada a técnica *GridSearchCV*, implementada na biblioteca Scikit-Learn [4], o que permitiu combinar os hiper parâmetros dos classificadores a fim de obter os melhores. Neste trabalho, o *GridSearchCV* foi executado com os seguintes classificadores:

- Linear Discriminant Analysis (LDA);
- Suport Vector Machine (SVM);
- Random Forest;
- Multilayer Perceptron (MLP).

Além disso, esse algoritmo também realiza validação cruzada para cada combinação de parâmetros utilizando o *K-Fold* com $k = 5$ [5].

Utilizou-se como critério de escolha dos parâmetros a métrica sensibilidade. Vale ressaltar que a escolha desse parâmetro – ao invés da comumente utilizada acurácia – se deu partindo da hipótese de que um algoritmo com uma sensibilidade maior pode trazer resultados mais significativos, uma vez que neste contexto, o acerto de verdadeiros positivos é vital para a saúde do paciente, enquanto falsos negativos podem levá-lo à morte. A Equação 2 abaixo mostra a sensibilidade:

$$sensibilidade = \frac{VP}{VP + FN}\ , \tag{2}$$

onde VP corresponde ao Verdadeiro Positivo e FN ao Falso Negativo.

III. Resultados e Discussão

A. Tratamento dos dados inicial

O dataset faz parte de uma iniciativa do Hospital Albert Einstein para explorar melhor o Covid-19 e vale ressaltar que vem com um número imenso de dados faltantes. Isso se explica devido ao fato que das 111 características, nem todas foram aplicadas a todos os pacientes, portanto, para alguns pacientes há a presença de n indicadores e para outros, outros m indicadores, podendo esses indicadores terem sobreposição ou não. Seguindo a análise observa-se que se trata de um problema de dados desbalanceados, sendo 90% resultados negativos contra 10% de positivos, o que também deve ser levado em conta.

```
Monocytes                                  5043
Red blood cell distribution width (RDW)    5042
Serum Glucose                              5436
Respiratory Syncytial Virus                4290
Influenza A                                4290
Influenza B                                4290
Parainfluenza 1                            4292
CoronavirusNL63                            4292
Rhinovirus/Enterovirus                     4292
Mycoplasma pneumoniae                      5644
Coronavirus HKU1                           4292
Parainfluenza 3                            4292
                                           ...
Urine - Sugar                              5644
Urine - Leukocytes                         5574
Urine - Crystals                           5574
Urine - Red blood cells                    5574
Urine - Hyaline cylinders                  5577
Urine - Granular cylinders                 5575
Urine - Yeasts                             5574
Urine - Color                              5574
Partial thromboplastin time (PTT)          5644
Relationship (Patient/Normal)              5553
```

Fig. 1. Quantidade de dados faltantes por característica.

Para lidar com o número alto de dados faltantes, conforme se observa na Figura 1, o primeiro critério de remoção desses dados baseou-se na remoção de características com dados faltantes maior ou igual a 80%, o que levou à diminuição em 82% das características, restando apenas 19 do mesmo. Como trata-se de um *dataset* desbalanceado, o próximo critério de remoção foi pensado apenas no *dataset* filtrado por resultado positivo do Covid19. Realizado esse filtro, esperava-se remover características nais quais dados faltantes fossem maior que 70%, porém 90% dessas apresentavam tal porcentagem. Como esses altos índices e levando em conta apenas 19 características, não foi possível remover tais atributos.

```
Patient age quantile           int64
SARS-Cov-2 exam result        object
Respiratory Syncytial Virus   object
Influenza A                   object
Influenza B                   object
Parainfluenza 1               object
CoronavirusNL63               object
Rhinovirus/Enterovirus        object
Coronavirus HKU1              object
Parainfluenza 3               object
Chlamydophila pneumoniae      object
Adenovirus                    object
Parainfluenza 4               object
Coronavirus229E               object
CoronavirusOC43               object
Inf A H1N1 2009               object
Bordetella pertussis          object
Metapneumovirus               object
Parainfluenza 2               object
```

Fig. 2. Características após primeiro abordagem.

A Figura 2 apresenta as características que permaneceram após a primeira abordagem do processamento de dados. Esta não foi uma abordagem bem sucedida, umas vez que as características restantes, em sua maioria, foram referentes à presença de outros tipos de vírus no organismo do paciente, apresentando assim uma baixa significância dessas características na determinação do Covid-19.

B. Tratamento de dados nova abordagem

Na segunda abordagem para o tratamento do *dataset*, foi feita a suposição de que o uso das características de presença de outros vírus não é a melhor maneira de se determinar a presença do Covid-19, o que permitiu a remoção de 19 características.

Com a eliminação das características que apresentavam valor nulo ou constante para todas as amostras, o *dataset* foi reduzido para 464 amostras com 77 características, de modo que o desbalanceamento foi melhorado, ficando com 85% para negativos e 15% positivos. Por fim, ao considerar somente os atributos com 50% ou mais de dados presentes o *dataset* final foi reduzido à 22 características com 464 amostras.

C. Imputação e normalização dos dados

A Figura 3 apresenta as características que serão utilizadas para treinar o modelo e o número de valores faltantes. Conforme é possível observar, os valores faltantes, há uma presença significativa de valores faltantes, os quais foram substituídos pela média em cada coluna.

Patient age quantile	0
SARS-Cov-2 exam result	0
Hematocrit	1
Hemoglobin	1
Platelets	1
Mean platelet volume	4
Red blood Cells	1
Lymphocytes	1
Mean corpuscular hemoglobin concentration (MCHC)	1
Leukocytes	1
Basophils	1
Mean corpuscular hemoglobin (MCH)	1
Eosinophils	1
Mean corpuscular volume (MCV)	1
Monocytes	2
Red blood cell distribution width (RDW)	1
Neutrophils	90
Urea	71
Proteina C reativa mg/dL	70
Creatinine	47
Potassium	93
Sodium	95

Fig. 3. Características e dados faltantes.

Com a normalização, os valores foram reduzidos a um intervalo de -1 à 1. O uso do algoritmo utilizado pela função *MinMaxScaler* apresentou resultados preliminares ligeiramente superiores em relação ao uso da *Standard Scaler*, o que se mostra um resultado esperado, dado que a presença ou não de *outliers* implica na sensibilidade de cada um desses métodos.

D. Treinamento

A Figura 4 mostra as acurácias e o desvio padrão para cada classificador. Observa-se que a acurácia do Random Forest apresentou o melhor desempenho, seguida pela do LDA. Além disso, em termos de complexidade, o LDA apresentou bom desempenho, apesar do menor custo computacional, o que implicou em uma velocidade significativamente maior na etapa de treinamento.

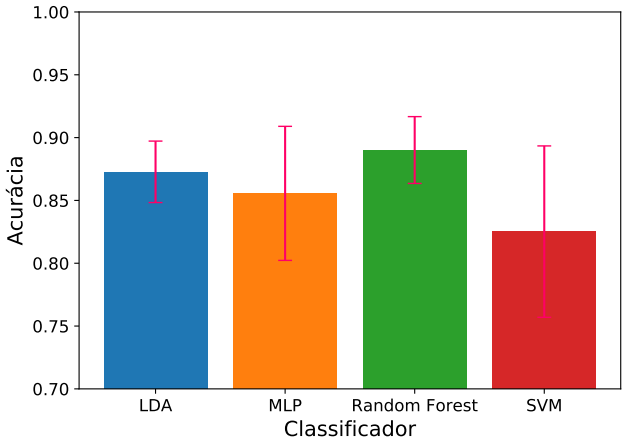

Fig. 4. Acurácia dos Classificadores.

Para o *Random Forest* os hiper parâmetros obtidos pelo *GridSeach* foram:

- criterion='entropy';
- max_features=None;
- min_samples_split=8;
- n_estimators=10.

Quanto a sensibilidade, os resultados podem ser vistos na Figura 5 e observa-se que o MLP teve a melhor sensibilidade com o valor aproximado de 76% e em segundo o *Random Forest* com aproximadamente 75%.

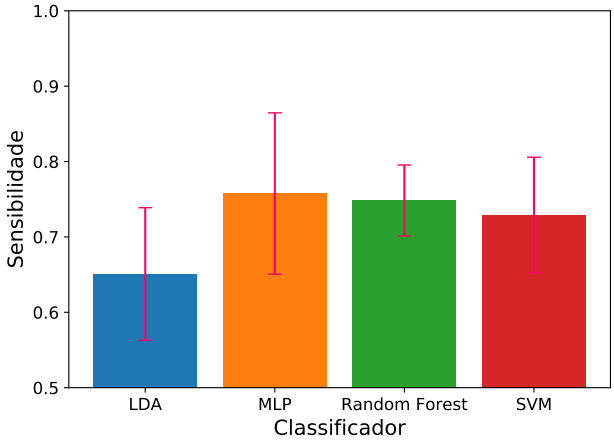

Fig. 5. Sensibilidade dos Classificadores.

O MLP demorou mais em relação aos outros algoritmos, dado o número maior de hiper parâmetros para combinar,

porém foi o que apresentou os melhores resultados de sensibilidade.

IV. Conclusão

O *dataset*, apesar da alta quantidade de características, mostrou que muitas dessas não valiam para todos os pacientes o que dificultava no tratamento dos dados, além do fato de haverem muitos valores faltantes – característica muito presente em bases de dados médicas. Além disso, o desbalanceamento dos dados também são comuns, dado que se tratam de dados obtidos a partir de um cenário real.

Mostrou-se importantíssimo nesse trabalho o tratamento dos dados antes de aplicar os algoritmos, desde a avaliação das características, como sua distribuição, seus dados faltantes, sua normalização e uma técnica muito interessante de lidar com dados faltantes que é de imputação para a inserção de dados. Vale ressaltar que é imprescindível a compreensão a respeito do tipo de dados com o qual se trabalha, pois tal entendimento pode tornar o problema mais claro para quem o analisa e de maior simplicidade para ser executado pelo algoritmo.

Como comentado nos resultados, a métrica escolhida para a escolha dos melhores hiper parâmetros foi a sensibilidade, para uma aplicação real o resultado ainda deve ser melhorado, como todos os classificadores usados deram um resultado próximo, tudo indica que para trabalhos futuros deve-se melhorar o processamento no *dataset* afim de obter uma sensibilidade maior, porém não somente no processamento do *dataset*, mas sim a aquisição de novos dados e a completude dos mesmos.

O trabalho proposto aqui alcançou uma sensibilidade um pouco maior do que de outros trabalhos feitos com o mesmo *dataset* [7], mostrando que, em estudos futuros, existe a possibilidade de se alcançar resultados com uma sensibilidade maior. Além disso também há espaço para melhorar a acurácia dos resultados obtidos.

Referências

[1] Salathé, Marcel, et al. “COVID-19 epidemic in Switzerland: on the importance of testing, contact tracing and isolation,” Swiss medical weekly, 150.11-12, 2020.

[2] Rothan, Hussin A., and Siddappa N. Byrareddy. “The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak.”, Journal of autoimmunity, 2020.

[3] Einstein Data4u. “Diagnosis of COVID-19 and its clinical spectrum”. Kaggle. 27 mar. 2020, https://www.kaggle.com/einsteindata4u/covid19.

[4] Pedregosa, F. and Varoquaux e et al. “Scikit-learn: Machine Learning in Python,”Journal of Machine Learning Research, vol. 12, pp 2825 2830, 2011.

[5] Dougherty, Geoff. Pattern recognition and classification: an introduction. Springer Science Business Media, 2012.

[6] R. S. Somasundaram, and R. Nedunchezhian, “Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values,” International Journal of Computer Applications, vol. 21, pp 14-19, 2011.

[7] Moda, Lucas.“COVID-19: Optimizing Recall with SMOTE”. Kaggle. 31 Mar. 2020, https://www.kaggle.com/lukmoda/covid-19-optimizing-recall-with-smote