CBIS'24
XX Congresso Brasileiro de Informática em Saúde
08/10 a 11/10 de 2024 - Belo Horizonte/MG – Brasil

Aprendizado de máquina para auxílio no diagnóstico doença pulmonar obstrutiva crônica

Machine learning to aid in the diagnosis of chronic obstructive pulmonary disease

Aprendizaje automático para ayudar en el diagnóstico de la enfermedad pulmonar obstructiva crónica

Ranier Pereira Nunes de Melo[1], Marco Paulo Soares Gomes[2], Luis Enrique Zárate[2]

1 Bac., Ciência de Dados, PUC Minas, Belo Horizonte (MG), Brasil
2 Dr., Ciência de Dados, PUC Minas, Belo Horizonte (MG), Brasil

Autor correspondente: Prof. Dr. Luis Enrique Zárate
E-mail: zarate@pucminas.br

Resumo

Objetivo: identificar fatores de risco para a doença pulmonar obstrutiva crônica na população brasileira. Método: por meio de um processo para descoberta de conhecimento, e modelos de aprendizado de máquina, identificar fatores de risco para a doença na população brasileira, baseado na Pesquisa Nacional em Saúde 2019. Resultados: o melhor modelo de aprendizado foi alcançado com o algoritmo Floresta Aleatória apresentando uma medida F1 de 75% para o conjunto de teste. Conclusões: a partir da análise do nível de importância dos principais fatores como asma, idade de risco, fumo anterior, índice de massa corpórea, risco domiciliar, dentre outros, destacaram-se os quatro primeiros como principais fatores de risco.

Descritores: Doença pulmonar obstrutiva crônica; Mineração de dados; Descoberta de conhecimento

Abstract.

Objective: to identify risk factors for the Chronic obstructive pulmonary disease in the Brazilian population. Method: through a process for knowledge discovery, and machine learning models, identify risk factors for the disease in the Brazilian population, based on

the 2019 National Health Survey. Results: the best learning model was achieved with the algorithm Random Forest presenting an F1 measure of 75% for the test set. Conclusions: based on the analysis of the level of importance of the main factors such as asthma, age at risk, previous smoking, body mass index, household risk, among others, the first four stood out as the main risk factors.

Keywords: Pulmonary disease, chronic obstructive; Data mining; Knowledge discovery

Resumen

Objetivo: identificar factores de riesgo para la enfermedad pulmonar obstructiva crónica en la población brasileña. Método: a través de un proceso de descubrimiento de conocimiento y modelos de aprendizaje automático, identificar factores de riesgo para la enfermedad en la población brasileña, con base en la Encuesta Nacional de Salud de 2019. Resultados: el mejor modelo de aprendizaje se logró con el algoritmo Random Forest presentando una medida F1 del 75% para el conjunto de prueba. Conclusiones: por médio del análisis del nivel de importancia de los principales factores como asma, edad de riesgo, tabaquismo previo, índice de masa corporal, riesgo del hogar, entre otros, se destacaron los cuatro primeros como principales factores de riesgo.

Descriptores: Enfermedad pulmonar obstructiva crónica; Minería de datos; Descubrimiento del conocimiento

Introdução

De acordo com a Organização Mundial de Saúde (OMS), a Doença Pulmonar Obstrutiva Crônica (DPOC) foi causa de 3,23 milhões óbitos no mundo em 2019, sendo considerada a terceira maior causa para óbitos no planeta[1]. De acordo com a Classificação Internacional de Doenças (https://icd.who.int), a DPOC é uma doença comum e tratável, caracterizada por uma limitação persistente de fluxo de ar, usualmente progressiva, associada a uma resposta inflamatória exacerbada nas vias aéreas e do sistema pulmonar. Essas exacerbações e comorbidades são fatores que podem contribuir para a severidade do portador da doença.

As causas e fatores de risco da doença, de acordo com o Relatório da Iniciativa Global para a Doença Pulmonar Obstrutiva Crônica – 2023[2], é resultado da interação entre a genética e o ambiente, que ocorre ao longo da vida do indivíduo, e que podem produzir danos aos pulmões, e/ou alterarem o seu desenvolvimento normal, ou ainda produzir seu envelhecimento prematuro. Os fatores que são citados não somente pelo relatório, mas também pelo Ministério da Saúde - Brasil (https://www.gov.br/saude/pt-br) e Organização Mundial da Saúde[1], são: o tabagismo, e a inalação de partículas tóxicas e gases de poluição aérea, dentro do domicílio ou em ambientes externos. Fatores biológicos também podem contribuir, como é o caso de desenvolvimento anormal do sistema respiratório. Em relação ao fator genético, que corresponde a uma alteração no gene SERPINA1, pode causar uma deficiência de produção da proteína alpha-1 antitrypsin, que teria como finalidade a proteção dos pulmões.

Quanto aos sintomas da DPOC, de acordo com a OMS, são: a falta de ar, conhecida, pelo jargão médico como dispnéia, dificuldade para respirar, tosse crônica com catarro, e sentimento de cansaço constante[1]. Tendo em vista que são sintomas comuns e que podem ser sintomas compartilhados com outras doenças, como a asma e a gripe, há o diagnóstico clínico chamado Espirometria, aonde verifica-se a capacidade da função pulmonar após expiração forçada, cujos valores são comparados com a média esperada de acordo ao sexo, altura e peso do paciente. Conhecer os principais fatores relacionados (análise multifatorial) que caracterizam pacientes com DPOC é de relevância, pois permitiria um melhor controle e diminuição da doença.

Dentro da área da Ciência de Dados, especialmente na área da saúde, o processo de descoberta de conhecimento em banco de dados é bastante requisitado, pois permite analisar dados acerca de um domínio de problema, numa perspectiva multifatorial, a partir de modelos de aprendizado computacional. A partir desses modelos, é possível identificar fatores relacionados com os pacientes diagnosticados com DPOC. Dentro deste contexto, o estudo[3] realizou uma revisão sistemática e uma meta-análise com objetivo de comparar a performance de modelos de aprendizado de máquina e *deep-learning* nas mais diversas

modalidades de tipos de entradas de dados. Desde modelos voltados para identificação de imagens, até modelos que são construídos com base em dados tabulares. Os autores identificaram inicialmente 3620 trabalhos, mas faram selecionados aqueles que atendem a requisitos estatísticos estabelecidos. Apesar da grande quantidade de estudos encontrados, o que ressalta a importância do tema, os autores identificaram que os modelos de prognóstico não apresentaram evidências suficientes para substituir pontuações pré-existentes de fatores relacionados com a severidade da doença. Os autores também apontaram a dificuldade para lidar com dados desbalanceados e dados ausentes, aspectos que foram levados em consideração durante o pré-processamento e seleção de dados no presente estudo.

A pesquisa[4] considera dados de vigilância da doença na Província de Shanxi, China. Os autores discutem o desafio de lidar com o desbalanceamento de dados ao construir modelos de aprendizado de máquina, e compararam as técnicas de Regressão Logística (RL), Máquina de Vetores de Suporte (SVM), Floresta Aleatória (FA), XGBoost, LightGBM, NGBoost e Stacking. Para lidar com o desbalanceamento consideraram o método SMOTE (*Synthetic Minority Over-sampling Technique*). Os melhores modelos foram apresentados de acordo com a medida G-Mean (Raiz quadrada do produto da sensibilidade e especificidade). Os algoritmos que produziram os melhores resultados foram: a RL com balanceamento de classes, alcançou 0,660; o modelo Stacking (modelo ensemble), utilizando SMOTE, alcançou 0,649; e o LightGBM, fundamentado em Árvores de Decisão, com SMOTE, apresentou resultado de 0.648. Os autores concluem que modelos de aprendizado de máquina baseados em pesquisas por meio de questionários com balanceamento por classe podem auxiliar no diagnóstico inicial da doença.

Por fim, no trabalho[5] foi utilizado uma perspectiva diferente quanto aos dados utilizados para alimentar os modelos de aprendizado de máquina. O estudo teve como escopo a análise de características genéticas dos pacientes diagnosticados com a doença em conjunto com outras variáveis, como o gênero, o fumo e o Índice de Massa Corpórea (IMC). Também utilizaram características genéticas como fatores de risco ou fatores de

proteção à doença. Os autores utilizaram as métricas AU-ROC (Área sob a curva ROC - *Receiver Operating Characteristic*), AU-PRC (Área sob a curva precisão-sensibilidade), sensibilidade (*recall*), Especificidade (*specificity*), Acurácia (*accuracy*), medida F1, Coeficiente de Correlação de Matthews (CCM), Valor Preditivo Positivo (VPP) e o Valor Preditivo Negativo (VPN) como medida de performance dos modelos baseados em: K-Nearest Neighbors (KNN), RL, SVM, Árvore de Decisão (AD), e métodos *ensemble* XGboost e Redes Neurais Multi-camadas Perceptron (MLP). Os autores apontam que os melhores modelos de acordo com as métricas apresentadas foram os modelos baseados nos algoritmos KNN, RL e XGboost. O modelo XGboost obteve as melhores métricas AU-ROC (0,94), AU-PRC (0,97), Acurácia (0,91), Precisão (0,95), Medida F1 (0,94), CCM (0,77) e Especificidade (0,85) além de apontar a idade e IMC do paciente como atributos de maior importância relativa para os modelos de classificação.

O objetivo deste trabalho é, por meio de um processo de mineração de dados, e modelos de aprendizado de máquina, identificar fatores de risco para o DPOC na população brasileira. Para isto será considerado o recente estudo do Instituto Brasileiro de Geografia e Estatística (IBGE), Pesquisa Nacional de Saúde (PNS) 2019, pesquisa realizada por meio de questionários no ano de 2019 em todo o território nacional por meio de amostragem (https://www.pns.icict.fiocruz.br/).

Materiais e Métodos

A PNS 2019, analisa a percepção do estado de saúde, estilos de vida, doenças crônicas e saúde bucal da população brasileira. A PNS constitui uma base sólida para análise de políticas públicas, implementadas pelo Estado Brasileiro, a partir de um retrato da saúde da população brasileira. A base de dados original da PNS-2019 possui 1.088 atributos organizados em 26 módulos, e 293.726 registros devidamente anonimizados. A pesquisa foi aprovada pela Comissão Nacional de Ética em Pesquisa – CONEP pelo parecer 3.529.376 de Agosto de 2019.

Dentro do contexto da PNS 2019, foi realizado um corte de análise voltada para a DPOC, aonde ocorreu uma pré-seleção conceitual de atributos que tenham relação para o diagnóstico da doença. Para este processo foi utilizado tanto do conhecimento tácito proveniente de especialistas de domínio, quanto de conhecimento explícito, proveniente da literatura.

É necessário ressaltar que a doença é gerada pela exposição prolongada à poluição, fumaça e gases. Porém, não há um grau preciso de quanto é necessário de exposição para que se adquira a doença. Infelizmente o estudo PNS 2019 não registra o tempo de exposição a esses elementos. A estratégia para lidar com essa incerteza, foi analisar cada atributo relacionado à doença, dentro de uma perspectiva de risco, analisando-se o risco de exposição para os fatores geradores da doença.

A seguir são apresentadas resumidamente as etapas do processo de preparação de dados e mineração de dados aplicados, de acordo com a metodologia para descoberta de conhecimento, associados ao domínio de problema considerado.

1. **Entendimento do problema:** Esta etapa ressaltada a necessidade de entendimento do domínio da aplicação e da importância do conhecimento a priori, acerca desse domínio, para a obtenção de modelos representativos.

   Devido à grande quantidade de dados disponíveis na base de dados PNS, foi proposto um processo para a seleção conceitual de atributos que possam contribuir para traçar o perfil de indivíduos com DPOC. Para isto, foi considerado o método CAPTO[6], o qual utiliza do conhecimento explícito (baseado na literatura médica) para identificar os principais aspectos relacionados com a doença. Os seguintes aspectos foram identificados: Atividade e ocupação profissional, ambiente de trabalho, Exposição a substâncias tóxicas, Hábitos de fumo, Antropometria, e Doenças crônicas pré-existentes.

2. **Seleção conceitual de atributos:** A partir da base de dados original PNS 2019, foram selecionados conceitualmente os principais atributos, de acordo com as

dimensões e aspectos identificados na etapa anterior. Tendo em vista, por princípio, o risco à doença pela exposição a fatores ambientais, tais como o tabagismo, a inalação de partículas tóxicas e de gases de poluição aérea, dentro do domicílio ou em ambientes externos, cada atributo, foi trabalhado em níveis de categoria, onde quanto maior o nível, maior o risco. Tendo em vista essa estratégia de análise, os dados adquiriram a forma categórica baseada no risco, conforme explicado. Na Tabela I apresenta-se todos os atributos selecionados e a razão de sua seleção.

3. **Pré-processamento:** Por se tratar de uma base de dados tratada e consolidada pelo IBGE, foi considerada a não existência de inconsistências no conjunto de dados. Por tanto, não foi aplicado um processo sistêmico de limpeza, porém um processo de preparação dos dados foi cuidadosamente aplicado.

   Após a seleção conceitual de atributos, foi necessário realizar pré-processamentos para cada atributo: retirada de outliers, a categorização do risco via transformação e fusão de atributos de forma que seja reduzida a dimensionalidade. A seguir, procedimentos adotados para cada atributo são descritos:

   a. **Tipo de situação censitária:** A poluição urbana é um fator ambiental que o indivíduo ao ser exposto, pode gerar danos ao sistema pulmonar à longo prazo. Já a poluição no meio rural é menor, mas ainda existente. Assim o tipo de situação censitária possui dois valores: Morador de área rural, foi atribuído o valor = 1; Morador de área urbana, o valor = 2.
   b. **Destino do lixo:** Caso o destino do lixo seja a queima, dentro da propriedade do entrevistado, o risco foi considerado com valor qualitativo = 1, nas demais hipóteses foi atribuído o valor qualitativo = 0.
   c. **Idade:** Antes de qualquer transformação do valor do atributo, foram retirados outliers por meio de análise de histograma, valores inferiores ao limite inferior (Q1-1,5*IQR) e valores superiores ao limite superior (Q3+1,5*IQR) foram excluídos do conjunto de dados. Em sequência, o atributo <Idade> foi

categorizado em faixas de idade, onde, quanto menor o valor dado para aquela faixa, menor o risco: Idades de 0 anos a 30 anos = 0, de 30 anos a 40 anos = 1, de 40 anos a 50 anos = 2, de 50 anos a 60 anos= 3, 60 anos ou mais = 4.

d. **Ocupação:** Este atributo é apresentado por um código da Classificação nacional de ocupações (de 0 a 9999) para pesquisas domiciliares de 2010. As ocupações com os seguintes códigos foram identificadas por trazer algum risco ocupacional de exposição aos fatores ambientais: 2113, 2114, 2142, 3112, 3116, 3117, 3121, 3122, 7111, 7112, 7126, 7544, 8111, 8112, etc. À título de ilustração, o código 8112 refere-se à "OPERADORES DE INSTALAÇÕES DE PROCESSAMENTO DE MINERAIS E ROCHAS". Os código referentes à ocupações que têm risco de exposição aos fatores da doença, foram atribuídas o valor de 1, onde há risco. Os demais códigos foram atribuídos como 0, já que não há risco em razão da ocupação.
e. **Ambiente de trabalho:** Se o entrevistado trabalha em Ambientes Abertos, recebe o valor qualitativo = 1, se trabalha em ambientes fechados e/ou abertos = 2, se trabalha em ambientes fechados, recebe o valor qualitativo = 3.
f. **Exposição à substâncias químicas na ocupação:** Se tem exposição a esse tipo de substância é atribuído o valor de 1, caso contrário o valor é 0.
g. **Exposição à poeira mineral na ocupação:** Se tem exposição a esse tipo de substância recebe o valor de 1, caso contrário o valor é 0.
h. **Hábito atual de fumar:** Se fuma atualmente, recebe o valor 1, caso contrário 0.
i. **Hábito de fumar no passado:** Se fumou no passado = 1, caso contrário = 0.
j. **Diagnóstico de Asma:** Se o entrevistado foi diagnosticado com Asma por um médico, recebe o valor qualitativo = 1, caso contrário recebe o valor 0.
k. **Diagnóstico de DPOC:** Se o entrevistado foi diagnosticado com DPOC por um médico, recebe o valor qualitativo = 1, caso contrário, o valor 0. O DPOC corresponde ao atributo classe do conjunto de dados.

É importante ressaltar que o processo de transformação dos atributos para dados categóricos ordinais, levando em consideração o risco para o doença, representado pelos níveis da categoria (onde quanto maior o nível, maior o risco) tem por objetivo contribuir para a construção e interpretabilidade de modelos de aprendizado. A categorização de um atributo, escolhida adequadamente, contribui para modular conceitualmente as faixas de valores. Por exemplo, o atributo <Idade> pode ser categorizado em criança, adolescente e adultos. Dependendo da quantidade de valores do atributo idade, algoritmos de aprendizado podem optar por segmentar automaticamente o atributo pela mediana, o que poderia resultar, em por exemplo, valores abaixo e acima de 13 anos (mediana), dificultando a interpretação dos resultados.

Tabela 1 - Atributos selecionados da PNS 2019

Código PNS	Atributo	Razão para a seleção do atributo
V0026	Tipo de situação censitária	O grau de poluição do ar aumenta de acordo com a região aonde o indivíduo reside, onde a área rural tem baixo grau de poluição e a Urbana tem alto grau de poluição.
A016010	Destino do lixo	A queima de lixo libera partículas tóxicas ao ambiente e se inaladas podem gerar danos ao sistema respiratório à longo prazo.
C008	Idade	A idade do indivíduo indica o acúmulo de exposição aos fatores ambientais listado nesse trabalho, seja de forma direta ou indireta.
E01201	Ocupação	A ocupação do indivíduo pode fazer com que seja exposto aos mais diversos fatores ambientais, como poeira, poluição e dentre outros, seja de forma direta ou indireta.
M009	Ambiente de trabalho	O ambiente de trabalho do indivíduo, aberto, fechado ou misto faz com ele esteja menos ou mais exposto aos fatores ambientais.
M011011	Exposição à substâncias químicas na ocupação	Contato direto e constante durante o trabalho é um fator de risco que pode gerar danos ao sistema respiratório
M011071	Exposição à poeira mineral na ocupação	Contato direto e constante durante o trabalho é um fator de risco que pode gerar danos ao sistema respiratório
P00104	Peso	É base para cálculo do Índice de massa corpórea (IMC), aonde indica-se que uma pessoa com alto IMC está vulnerável devido à maior esforço respiratório
P00404	Altura	É base para cálculo do IMC, aonde indica-se que uma pessoa com alto IMC está vulnerável devido à maior esforço respiratório
P050	Hábito atual de fumar	O tabagismo é um dos sinalizadores da doença, já que é a inalação de fumaça para dentro dos pulmões.
P052	Hábito de fumar no passado	O tabagismo é um dos sinalizadores da doença, já que é a inalação de fumaça é agravado pelo hábito passado do fumo.
Q074	Diagnóstico de asma	A Asma tem como característica a fragilização do sistema respiratório, diminuindo a imunidade do indivíduo.
Q11604	Diagnóstico de DPOC	É o atributo alvo, sendo a base de classificação do artigo, também responsável pelo corte de análise da pesquisa.

Dados acessível em: https://github.com/licapLaboratory/DataBase-PNS2019---Chronic-obstructive-pulmonary-disease

I. **IMC:** Os atributos Peso e Altura foram combinados para determinação do IMC. Todavia, antes da aplicação dessa transformação, foram retirados os outliers de ambos os atributos via análise de histograma. Valores inferiores ao limite inferior (Q1-1.5*IQR) e valores superiores ao limite superior (Q3+1.5*IQR) foram retirados do conjunto de dados. Foram definidas faixas de risco baseada no IMC do indivíduo de acordo com a classificação do National Library of Medicine (https://www.nlm.nih.gov/): IMC (Severamente abaixo do peso) menor 16.5 (valor 1), IMC (Abaixo do peso) entre 16.5 e 18.5 (2), IMC (Peso normal) entre 18.5 e 25 (3), IMC (Sobrepeso) entre 25 e 30 (4), IMC (Obesidade) entre 30 e 35 (5), IMC (Obesidade nível I) entre 35 e 40 (6), IMC (Obesidade nível II) entre 40 e 45 (7) e IMC (Obesidade nível III) maior que 45 (8).

3. **Tratamento de dados ausentes/vazios:** Após o processo de transformação, o conjunto de dados possui somente dados categóricos ordinais. Contudo, pela metodologia da pesquisa PNS, algumas perguntas do questionário não foram aplicadas a todos os entrevistados, por essa razão não responderam ao quesito. Tendo em vista esse contexto, e identificado por meio da análise do atributo, que o entrevistado não possui risco à doença, foi imputado o menor valor dentro da escala de risco do atributo, ou valor 0.
4. **Balanceamento do conjunto de dados:** No total foram identificadas 75565 instâncias composta de indivíduos que **não** são diagnosticados com a doença (Classe 0 - **Q11604**) em relação aos indivíduos que foram diagnosticados com DPOC, o número de registros de indivíduos diagnosticados foi de 997 (Classe 1 - **Q11604**). Diante do desbalanceamento de dados, foi aplicado o método de subamostragem aleatória (*Random Under Sampler*), pelo qual, a classe majoritária (Classe 0 - **Q11604**) tem sua amostra diminuída e equalizada aleatoriamente em relação a classe minoritária (Classe 1). Como resultado, foram selecionadas 997 instâncias de pacientes não diagnosticados das 76562 instâncias do conjunto de dados. Assim, o conjunto de dados total constitui-se de 1994 instâncias, com 13

atributos, incluindo o atributo classe (Diagnóstico de DPOC - Q11604).

5. **Modelos de aprendizagem e parametrização:** O objetivo do estudo é caracterizar indivíduo que possuem DPOC. Para isso modelos de classificação foram aplicados para caracterizar os indivíduos diagnosticados ou não com DPOC, valores {1,0} respectivamente. Na literatura[7] são discutidas as vantagens e desvantagens dos métodos caixa branca e caixa preta. Para este trabalho foram escolhidos modelos de aprendizado caixa-branca, por causa da favorecer a interpretabilidade (Árvore de decisão, *Decision-Tree*, Classificador bayesiano, *Naive-Bayes*) e caixa-preta, com menor interpretabilidade, porém normalmente com melhor desempenho (Floresta Aleatória, *Random-Forest*, MLP, e Ada Boost)[7]. Sendo Floresta Aleatória e AdaBoost considerados métodos *ensemble*. É necessário ressaltar que os cinco modelos foram treinados com 90% do conjunto de dados, e 10% dos dados foram destinados para o teste (método Holdout). Foi aplicado validação cruzada de 10 dobras. Além disso, foram calculados os intervalos de confiança para a métrica de F1-Measure, a partir das 10 dobras, com significância de 5%. Foi aplicado otimização randômica de parâmetros nos modelos de Árvore de Decisão, Floresta Aleatória e Ada Boost, utilizando-se a técnica RandomizedSearchCV (scikit-learn - Python) para efetuar uma busca aleatória de parâmetros a partir de uma amostra retirada de uma distribuição de valores possíveis para aquele hiperparâmetro.
   a. **Algoritmo Árvore de Decisão**: Pelo processo de otimização foi alcançado os seguintes hiperparâmetros: critério de medida, sendo Gini; máximo de profundidade, 6; máximo de atributos utilizados, 10; o mínimo de amostras por folha igual a 20; e o mínimo de amostras para divisão sendo 50.
   b. **O algoritmo Floresta Aleatória:** Pelo processo de otimização foi alcançado os seguintes hiperparâmetros: critério de medida, Gini; máximo de profundidade, 10; máximo de atributos, 10; mínimo de amostras por folha igual, 20; e mínimo de amostras para divisão sendo 100 e o número de estimadores igual a 90.

c. **O algoritmo Classificador MLP:** Foi aplicado o teorema de Kolmogorov, onde todo modelo baseado em MLP pode ser projetado com uma camada de entrada que possui a quantidade de neurônios iguais a quantidade de atributos (N), e uma camada escondida com 2*N+1, e função de ativação sigmoide.
d. **O algoritmo AdaBoost:** Pelo processo de otimização foi alcançado os seguintes hiperparâmetros: taxa de aprendizado de 0.1; e número de estimadores igual a 60.

Resultados e discussões

Na Tabela 2, é possível observar que o classificador Floresta Aleatória apresentou intervalos de confiança (com 5% de significância) mais similares, com valores limite muito próximos para a predição do diagnóstico, alcançando valor da medida F1 de 89%.

Tabela 2 - Resultados do Treinamento dos Modelos de Aprendizado

Algoritmo	F1-Measure Diagnosticados com DPOC	F1-Measure Não Diagnosticados DPOC
Floresta Aleatória	[0.66, 0.89]	[0.66, 0.88]
Árvore de decisão	[0.67, 0.73]	[0.70, 0.77]
Classificador Bayesiano	[0.64, 0.70]	[0.67, 0.73]
ADABoost	[0.64, 0.69]	[0.72, 0.78]
Classificador MLP	[0.65, 0.72]	[0.70, 0.76]

De forma a realizar uma análise mais criteriosa, foi aplicado o teste-T nos dados de treinamento, comparando-os entre si, para verificação se a hipótese nula de que as médias obtidas para a medida F1 são equivalentes ou não. A partir dos cálculos comparativos observou-se que o algoritmo Floresta Aleatória (t = 2,455756175, p = 0,02445086292 e valor crítico de 2,10092204) efetivamente apresentou o melhor desempenho no valor médio.

Numa segunda fase dos experimentos, o conjunto de teste foi aplicado aos modelos. O desempenho dos modelos foi calculado pela Precisão, Sensibilidade (recall), e medida F1 de cada modelo. Os resultados obtidos foram compilados no Gráfico 1. A partir dos resultados é possível observar que o melhor resultado preditivo de 75% foi alcançado pelo classificador Naive-Bayes para o conjunto de teste.

É importante ressaltar que embora no processo de teste o classificador Naive-Bayes apresentou melhor desempenho a escolha do melhor modelo deve ser feito considerando várias informações: resultados de treinamento, intervalos de confiança, teste-T, e os resultados do processo de teste. Porém, sempre é importante enfatizar que o desempenho dos modelos de aprendizado de máquina é altamente dependente da qualidade e da representatividade do conjunto de dados considerado. Neste trabalho consideramos que o modelo Floresta Aleatória apresentou resultados mais equilibrados, durante o treinamento, intervalos de confiança, e na comparação do teste-T.

Para interpretação dos modelos foi determinada a importância dos atributos advindas dos modelos baseados em Árvore de decisão, Floresta Aleatória, e AdaBoost. A importância dos atributos foi extraída via algoritmo Feature_importances (scikit-learn - Python) e foi determinada de acordo com o ganho de informação para esses modelos.

Gráfico 1 - Resultados do Modelos de Aprendizado

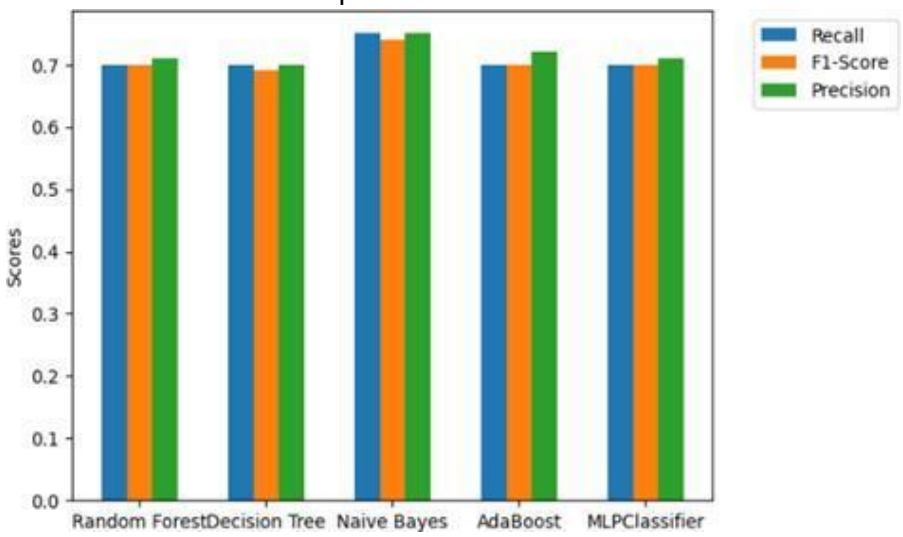
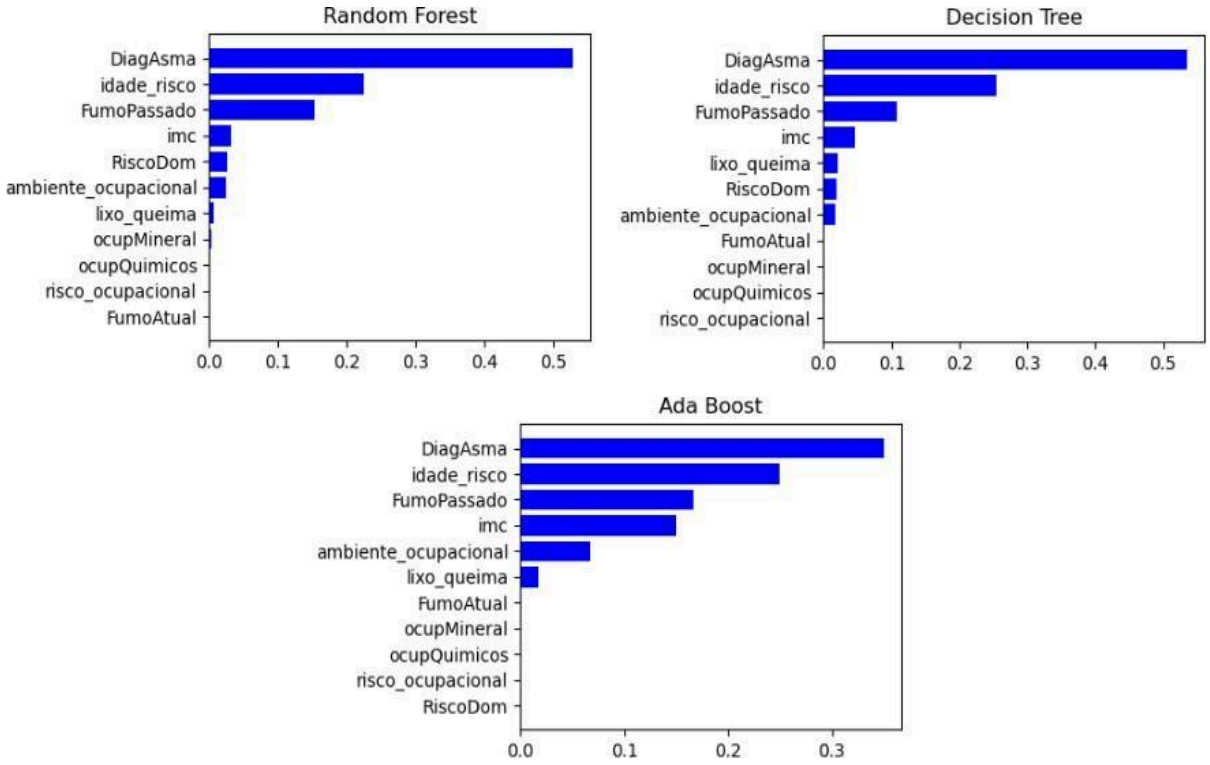


Gráfico 2 - Importância dos atributos no contexto de cada modelo



Os atributos apresentados no Gráfico 2 demonstram o que a literatura médica relacionada à doença já tinha conhecimento, cujo indivíduos que tenham diagnóstico de asma (DiagAsma) são mais propensos a ter o diagnóstico de DPOC. Além disso, a idade mais avançada (idade_risco) e o histórico de fumo (FumoPassado) também são fatores críticos para guiar o profissional médico para a direção do diagnóstico da doença via exame clínico, no caso, a espirometria. É importante ressaltar que os fatores anteriores junto com um IMC desfavorável, pode contribuir para o diagnóstico da doença. Outros fatores que podem auxiliar no diagnóstico é também viver em ambientes próximos de queima de lixo. Para a base de dados PNS 2019 considerada neste trabalho, não foi possível detectar que a exposição a substâncias químicas ou de poeira mineral, e hábitos de fumo possam determinar o diagnóstico para a doença, embora a extensa literatura médica sobre a doença aponte que são fatores de alto risco para adquirir a doença, contradição que pode ser analisada em pesquisas futuras. Em geral, ao dar-se um grau

de importância nesses fatores, pode-se orientar profissionais da área médica a perseguir novos caminhos de pesquisa.

Conclusão

Os modelos obtidos via aprendizado de máquina com base nos dados da PNS 2019 trazem uma forma complementar de analisar tendências e auxiliar em diagnósticos da DPOC. Além disso, pode contribuir para uma análise mais criteriosa para pesquisadores da área médica em determinar quais fatores devem ser mapeados em uma primeira consulta entre um médico e um paciente. Além disso, pode ser um balizador para formulação de políticas públicas ou o incentivo para a produção de pesquisas voltadas exclusivamente para a DPOC, nos mesmos moldes da PNS 2019.

Apesar dos resultados obtidos, também é necessário realizar uma análise crítica dos modelos obtidos, tendo em vista suas limitações e desafios. Como os dados da PNS 2019 não correspondem a estudos longitudinais, há uma limitação ao não considerar o aspecto temporal dos atributos para o diagnóstico mais eficaz da doença, tendo em vista um diagnóstico de maior duração. Além disso, a forma de atribuição de riscos à exposição aos fatores ambientais da doença dá margem para interpretação, tendo em vista a falta de exatidão quanto à forma de como o dado foi coletado durante a pesquisa. É sugerido aos órgãos públicos realizar um estudo longitudinal por longo período de tempo, envolvendo pessoas a risco futuro de contrair DPOC.

Nessa conjuntura, apresentando-se as vantagens e deficiências da abordagem do aprendizado de máquina para diagnóstico de doenças, é importante ressaltar que modelos obtidos via aprendizado de máquina não substituem o trabalho médico, mas sim, tem como finalidade auxiliar, validar e balizar a boa prática médica. Por fim, esse artigo apresentou a possibilidade de uso do aprendizado de máquina na previsão de tendências de DPOC, fornecendo discernimentos para fundamentar iniciativas de saúde pública e auxiliar na identificação da necessidade de um tratamento preventivo e paliativo da Doença Pulmonar Obstrutiva Crônica. Espera-se que a obra produzida tenha sido uma

adição para o combate da doença na população e diminuição do impacto da doença na sociedade.

Agradecimentos

Ao CNPq e à FAPEMIG pelo apoio financeiro recebido e à PUC Minas.

Referências

1. WHO, W. H. O. Chronic obstructive pulmonary disease (copd). https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

2. GOLD-COPD. Global strategy for prevention, diagnosis and management of copd: 2023 report. https://goldcopd.org/ 2023-gold-report-2/, 2023.

3. Smith LA, Oakden-Rayner L, Bird A, Zeng M, To MS, Mukherjee S, Palmer LJ. Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. Lancet Digit Health. 2023 Dec;5(12):e872-e881. doi: 10.1016/S2589-7500(23)00177-2. PMID: 38000872.

4. Wang X, Ren H, Ren J, Song W, Qiao Y, Ren Z, Zhao Y, Linghu L, Cui Y, Zhao Z, Chen L, Qiu L. Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data. Comput Methods Programs Biomed. 2023 Mar;230:107340. doi: 10.1016/j.cmpb.2023.107340. Epub 2023 Jan 6. PMID: 36640604.

5. Ma X, Wu Y, Zhang L, Yuan W, Yan L, Fan S, Lian Y, Zhu X, Gao J, Zhao J, Zhang P, Tang H, Jia W. Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population. J Transl Med. 2020 Mar 31;18(1):146. doi: 10.1186/s12967-020-02312-0. PMID: 32234053; PMCID: PMC7110698.

6. Zarate, L., Petrocchi, B., Maia, C., Felix, C., and Gomes, M. P. CAPTO - A method for understanding problem domains for data science projects. Concilium 23:922–941, 2023.

7. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View, in IEEE Access, vol. 7, pp. 154096-154113, 2019.