Revista Real Conhecer - **ISSN 2763-5473**
Periódico Científico Indexado Internacionalmente
www.realconhecer.com.br

MODELO DE FLORESTAS ALEATÓRIAS PARA IDENTIFICAR FALSOS POSITIVOS NA TRIAGEM DE FIBROSE CÍSTICA COM DADOS ARTIFICIAIS

DOI: 10.5281/zenodo.6612605

Paulo Rogério Siqueira Custódio
Minicurrículo do autor: Mestrando, Engenharia Biomédica/Elétrica, paulo55866@gmail.com

Virginia Klausner de Oliveira
Minicurrículo do autor: Doutora, Engenharia Elétrica/Processamento de sinais, virginia@univap.br.

Guilherme Maerschner Ogawa
Minicurrículo do autor: PHd, Biologia, ped@intercientifica.com.br

Rainara Moreno Sanches de Almeida
Minicurrículo do autor: Mestrando, Biomedicina, ped2@intercientifica.com.br.

Resumo: A Fibrose Cística é uma doença letal que é caracterizada por infecções crônicas no pulmão, insuficiência pancreática e elevados níveis de cloro no suor, essa doença é causada pela mutação no gene do Regulador de Condutância Transmembrana da Fibrose Cística (CFTR), essa doença faz com que o organismo produza secreções espessas e viscosas que obstruem os pulmões, pâncreas e no ducto biliar (RIBEIRO ROSA e colab., 2008). Esta doença faz parte do programa de Triagem Neonatal Brasileiro, sendo triada através da quantificação do Tripsinogênio Imunorreativo (IRT), entretanto essa tripsina apresenta um elevado índice de resultados falsos positivos, isso faz com que seja necessário um outro teste para confirmação do diagnóstico. Sabe-se pela literatura que existem fatores que influenciam na alteração da tripsina como a etnia por exemplo devido a doença ser em sua grande maioria na população caucasiana. Então este trabalho propõe a utilização de técnicas de florestas aleatórias para conseguir detectar a probabilidade de um resultado ser falso positivo analisando as informações dos pacientes e as condições de coleta de amostra utilizando dados artificiais.
Palavras-chave: Fibrose Cística. Aprendizado de Máquina. Triagem Neonatal. Falsos Positivos.

Abstract: Cystic Fibrosis is a lethal disease that is characterized by chronic lung infections, pancreatic insufficiency and high levels of chlorine in sweat, this disease is caused by mutation in the gene Cystic Fibrosis Transmembrane Conductance Regulator (CFTR), this disease causes that the organism produces thick and viscous secretions that obstruct the lungs, pancreas and bile duct (RIBEIRO ROSA et al., 2008). This disease is part of the Brazilian Neonatal Screening program, being screened through the quantification of Immunoreactive Trypsinogen (IRT), however this trypsin has a high rate of false positive results, which makes another test necessary to confirm the diagnosis. It is known in the literature that there are factors that influence the change in trypsin, such as ethnicity, for example, due to the disease being mostly in the Caucasian population. So, this work proposes the use of random forest techniques to be able to detect the probability of a false positive result by analyzing patient information and sample collection conditions using artificial data.
Keywords: Cystic Fibrosis. Machine Learning. Newborn Screening. False Positive.

INTRODUÇÃO

A fibrose cística tem capacidade de atacar todos os sistemas do corpo humano, principalmente o respiratório, gastrointestinal e o reprodutor. Ela comumente se manifesta na infância/adolescência e o diagnóstico precoce e tratamento são as melhores formas de se dar sobrevida ao paciente (RIBEIRO e colab., 2021). O IRT – tripsinogênio imunorreativo da enzima pancreática, é encontrado elevadamente em pacientes com fibrose cística (CABELLO e colab., 2003). Portanto é fundamental a realização do teste de quantificação do IRT (tripsinogênio imunorreativo) na triagem Neonatal, no entanto são comuns resultados falsos positivos e muito menos frequente falsos negativos, na triagem o que se deseja é reduzir o número desses resultados errôneos. Na literatura conhece-se mais fatores associados a falsos positivos do que a negativos, também a estudos de impacto psicossocial em falsos positivos (LUMERTZ e colab., 2019). A triagem Neonatal aumenta significativamente a detecção precoce de distúrbios congênitos, embora seja acompanhado de um grande número de falsos positivos como efeito adverso devido à alta sensibilidade exigida no programa para que se evite casos de falsos negativos (KWON e FARRELL, 2000).

Este trabalho propõe a utilização de dados artificiais de pacientes do programa de triagem Neonatal de fibrose cística da APAE de São Luis – MA, para realização de uma análise levando em consideração as características da população triada para determinar um perfil dos recém nascidos testados positivos, isso levando em conta suas características físicas, condições de nascimento, tempo de coleta de amostra e resultados. Com a utilização de técnicas de aprendizado de máquina foi criado um

algoritmo de florestas aleatórias para aprender com esses dados, fazendo com que ele tenha capacidade de fazer previsões para uma nova base de dados e dessa forma conseguir melhorar a previsão do resultado da amostra de ser realmente um verdadeiro positivo ou um falso positivo. Para que o algoritmo possa fazer essa analise com precisão um dado que é de extrema importância é o feedback se a criança realmente foi confirmada para fibrose cística ou não.

Os dados que foram criados para simulação foram baseados nos dados utilizados no trabalho de Peng (PENG e colab., 2020) sendo os mesmos desse estudo acrescentando somente os dados fornecidos pelo laboratório de classificação do paciente e o confirmatório da classificação. A distribuição dos dados foi semelhante ao do trabalho base, feitas em proporções parecidas com a da população do teste do autor.

Com a utilização desse algoritmo de florestas aleatórias será possível que o laboratório atenda com urgência e agilidade os pacientes que tenham maior probabilidade de ser diagnosticado com fibrose cística, e iniciar com mais rapidez do que com as técnicas atuais o tratamento precoce. O alto número de falsos positivos pode tardar um diagnostico após a triagem Neonatal positiva e o aprendizado de máquina empregado nesse campo aumenta a eficiência e agilidade do diagnóstico (PENG e colab., 2020).

METODOLOGIA

Para o desenvolvimento do trabalho foi necessário à criação de um banco de dados artificial e o desenvolvimento de um modelo de aprendizado de máquina. A criação do banco de dados foi desenvolvida com o apoio do laboratório de triagem neonatal, APAE São Luis – MA, que forneceram algumas informações dos seus testes realizados para que a base de dados ficasse com valores reais, na próxima seção está descrito em detalhe como foi desenvolvido o banco de dados. No desenvolvimento do modelo aprendizado de máquina foi escolhido utilizar a técnica de florestas aleatórias para avaliar seu desempenho descrevendo e classificando esse conjunto de dados criados.

BANCO DE DADOS ARTIFICIAL

Para desenvolvimento do banco de dados artificial foi necessário definir as características dos recém-nascidos que serão utilizadas para realizar a previsão se o teste é um falso positivo ou falso negativo. Como Peng et al. (2020) fez um trabalho de proposta parecida na Califórnia, porém para outras doenças, do programa de triagem neonatal local, foi decidido primeiramente testar os mesmos dados dos recém nascidos utilizados pelo autor, as proporções dos dados criados são aproximadamente as mesmas utilizadas pelo autor na população teste de Acidúria Glutárica tipo I (GA-I).

Entretanto como a doença que será analisada é fibrose cística se fez necessário alterar os dados dos resultados para que descreva a doença corretamente, para isso o laboratório de triagem neonatal APAE São Luis – MA colaborou em parceria com o projeto fornecendo alguns dados para elaboração dessa base de dados para que a mesma fosse a mais próxima da realidade possível, os dados fornecidos são de controle do laboratório, não foram fornecidos dados de pacientes. Os dados fornecidos pelo parceiro estão na Tabela – 1, esses dados são referentes ao ano de 2020. É importante dizer que a metodologia adotada pelo laboratório é, caso o primeiro exame de triagem seja dado alterado um segundo exame é realizado e se persistir alterado é realizado o teste do suor para confirmar o diagnóstico.

Tabela 1 - Dados de Resultados de Fibrose Cística 2020 APAE São Luis - MA

Resultado	Quantidade
1° Resultado alterado	63
2° Resultado alterado	10
Confirmado	7
Total de Testes	80.199

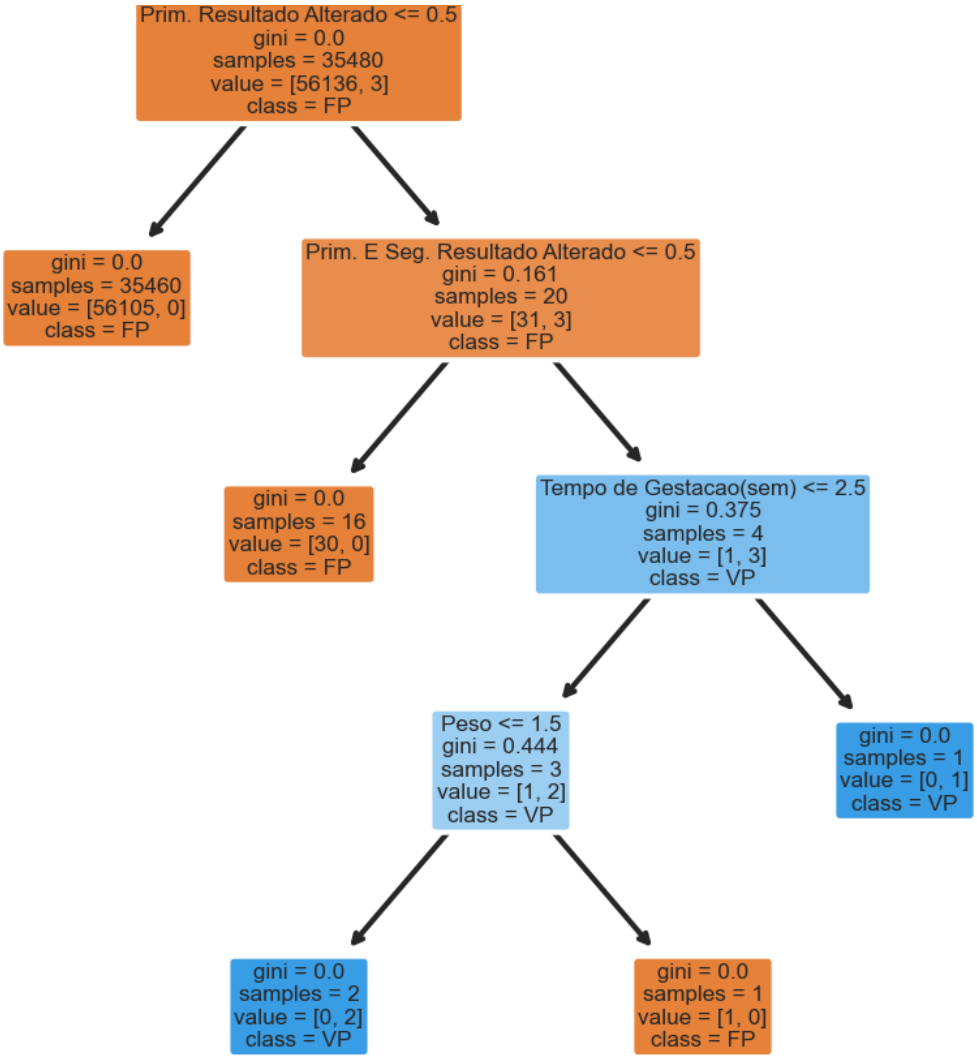
Fonte: o autor.

MODELO DE FLORESTA ALEATÓRIA

Foi decidido utilizar um modelo de floresta aleatória para fazer a previsão da probabilidade de o resultado confirmatório do paciente ser de fato um verdadeiro positivo. Técnicas de florestas aleatórias são ideais para esse tipo de problema devido ao fato de utilizar um conjunto de árvores de decisão que faz com que o modelo tenha um resultado ainda mais promissor do que uma única árvore devido ao fato de o resultado ser uma eleição das classificações de todas as árvores de decisão. Se você agregar as previsões de um conjunto de previsores, muitas vezes obterá melhores previsões do que com o melhor previsor individual (GÉRON, 2019).

As árvores de decisão são algoritmos versáteis, capazes de moldar conjuntos complexos de dados para realizar previsões. A moldagem dos dados é realizada através de buscas de critérios de segmentação dos dados, isso faz com que esse tipo de algoritmo seja intuitivo tendo um alto poder explicativo. Na Figura 1 temos um modelo de uma das árvores da floresta aleatória, onde é possível ver os critérios criados pra fazer a classificação desta árvore e a métrica de impuridade utilizada, o valor tomado para a decisão e a classe para a situação.

Figura 1 - Uma das árvores do modelo de florestas aleatórias



***Fonte**: o autor*

Devido ao fato de se desejar sempre prever entre duas classes, positivo e negativo, o modelo mais recomendado é o de *Ramdon Florest Classifier*, que tem como resultado a moda das previsões. Para desenvolvimento do modelo foi usada a biblioteca de código aberto de aprendizado de máquina em Python *Scikit Learn*. Para realizar a divisão dos nós foi utilizado o critério do índice de gini, que é dado pela equação:

$$Indice\ GINI = 1 - \sum_{i=1}^{c} p_i^2 \quad (1)$$

Onde:

$$p_i = frequencia\ relativa\ de\ cada\ classe\ em\ cada\ nó$$

$$c = numero\ de\ classes$$

O critério de gini foi criado por Conrad Gini em 1912 e mede a impuridade no nó. O que buscamos é um nó puro, isso acontece quando temos um índice igual a zero. Quando nas árvores de decisão se utiliza o critério de gini tende-se isolar no ramo os registros da classe mais frequente (BARBOSA e colab., 2012).

Para avaliar o modelo criado foram utilizados três métricas diferentes, primeiramente a acurácia:

$$Acurácia = \frac{VN + VP}{VN + VP + FN + FP} \quad (2)$$

Que diz o quão próximo às previsões do modelo estão do valor real, porém essa métrica sozinha não é um parâmetro totalmente confiável, pois se a quantidade de dados positivos e negativos tem uma diferença grande entre elas (como esse é o caso do conjunto de dados) pode-se ter uma alta acurácia, porém não quer dizer que o modelo é preciso, então também é avaliada a precisão do modelo:

$$Precisão = \frac{VP}{VP + FP} \quad (3)$$

Esta outra métrica caracteriza a proporção dos valores que são positivos, em relação ao total de valores previstos como positivos. Essa métrica faz com que se tenha um entendimento melhor do desempenho do modelo, porém temos uma terceira métrica que explica a sensibilidade do algoritmo, ou seja, a proporção dos valores que eram realmente positivos, em relação ao todos os casos que de fato eram positivos, essa métrica e chamada de recall:

$$Recall = \frac{VP}{VP + FN} \quad (4)$$

Como sempre temos que prever os casos que de fato são positivos a métrica que tem o maior peso é a recall, entretanto uma alta sensibilidade faz com que o modelo seja menos preciso, então é necessário buscar um equilíbrio entre essas métricas, o modelo deve prever todos os casos que realmente são positivos e ter a maior precisão possível.

RESULTADOS

A Tabela – 2 mostra a distribuição dos dados artificiais criados seguindo a metodologia proposta, como pode-se observar embora a porcentagem de pacientes alterados seja baixa no primeiro teste, quando se compara ela com a de alterados no primeiro e segundo teste vê-se que ela é amplamente superior e essa diferença aumenta ainda mais quando se comparado com os casos confirmados.

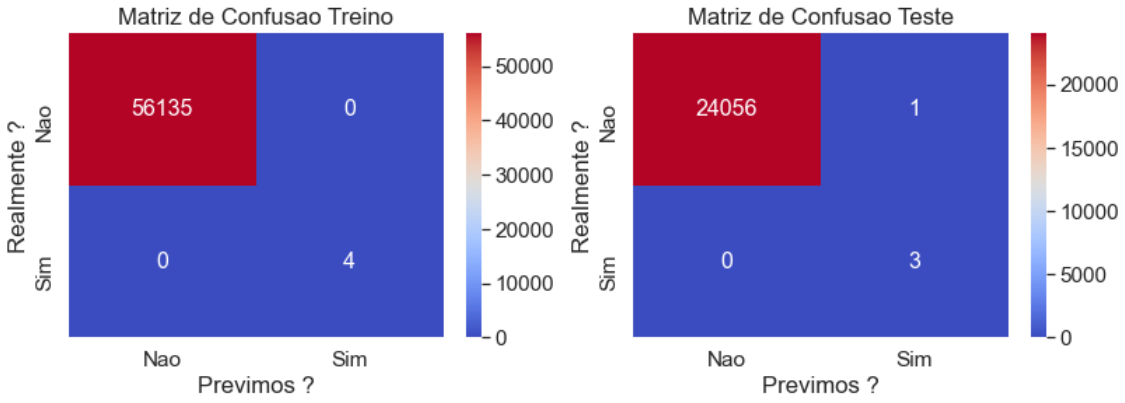
Tabela 2 - Distribuição dos dados artificiais

Dado:	**Observações de distribuição do banco de dados e suas porcentagens**				
Sexo	Masculino = 1 → 60%			Feminino = 2 → 40%	
Etnia	Asiático = 1 → 10%	Negro = 2 → 15%	Latino = 3 → 32%	Branco = 4 → 40%	Outras = 5 → 3%
Idade Gestacional (Sem)	< 37 = 1 → 24%		37-41 = 2 → 73%		> 41 = 3 → 3%
Nutrição Parental	Sim = 1 → 10%			0 = Não → 90%	
Peso ao Nascer (g)	< 2500 = 1 → 20%		2500 – 4000 = 2 → 73%		> 4000 = 3 → 6%
Tempo de coleta (horas)	< 12 = 1 → 18%		12 – 24 = 2 → 63%		> 24 = 3 → 19%
1° Resultado Alterado	Normal = 0 → 99,921%			Alterado = 1 → 0,0786%	
1° e 2° Resultado Alterado	Normal = 0 → 99,987%			Alterado = 1 → 0,013%	
Confirmatório	Normal = 0 → 99,991%			Alterado = 1 → 0,009%	

Fonte: o autor.

O modelo de florestas aleatória foi capaz de fazer um bom aprendizado no conjunto de dados de treinamento, e ao observar as previsões do conjunto de teste podemos ver que não houve um sobre ajuste dos dados. Na Figura 2 está a matriz de confusão dos dois conjuntos de dados. Importante ressaltar que para os testes realizados foi utilizada uma proporção de setenta por cento dos dados artificias no conjunto de treinamento e os outros trinta por cento no conjunto de teste, que é o responsável por fazer a validação do modelo treinado.

Figura 2 - Matiz de confusão conjunto teste e treino



Fonte: o autor.

Na matriz pode-se observar que no conjunto de teste o modelo acertou todas as previsões realizadas e no conjunto teste ele errou somente uma das observações, onde ela não era positiva e o algoritmo a classificou como positiva. Embora o modelo tenha errado essa classificação como olhamos na Tabela - 3 que descreve as métricas obtidas pelo modelo, a métrica *recall* permanece em cem por cento o que é o mais importante no processo de triagem neonatal já que informa a sensibilidade do modelo, temos uma queda na precisão quando se comparado ao conjunto teste.

Tabela 3 - Desempenho do modelo

Métrica	Conjunto Treino (%)	Conjunto Teste (%)
Acurácia	100	100
Precisão	100	75
Recall	100	100

Fonte: o autor.

CONCLUSÃO

Ao analisar os resultados obtidos com o modelo criado e a base de dados artificias, pode-se concluir que com técnicas de florestas aleatórias podem ser obtidos resultados promissores quando este estudo for feito utilizando dados reais e dessa forma diminuir consideravelmente o índice de falsos positivos na triagem neonatal de fibrose cística, dessa forma conseguir realizar um diagnóstico com prematuridade maior do que se tem hoje, assim começando mais cedo o tratamento dos pacientes. Através desse estudo viu-se que é viável utilizar essa metodologia para realizar a

filtragem nos pacientes triados testados positivamente no primeiro e segundo teste, indicando os pacientes que tem maior índice de ser um verdadeiro positivo, isso sem perder a sensibilidade requirida para esse tipo de exame.

Agora se faz necessário fazer um estudo semelhante ao feito abrangendo mais dados, que serão referentes aos anos anteriores a 2020 para ver o desempenho do modelo e ajustá-lo caso se faça necessário. Também deve ser feito um ajuste nos dados obtidos do paciente para que condigam com os dos pacientes triados pelo laboratório.

REFERÊNCIAS

BARBOSA, Juliana Moreira e colab. **Métodos de Classificação por Árvores de Decisão**. Programa de Pós-Graduação em Ciência da Computação, p. 5, 2012.

CABELLO, Giselda M.K. e colab. **Rastreamento da fibrose cística usando-se a análise combinada do teste de IRT neonatal e o estudo molecular da mutação deltaF508**. Jornal Brasileiro de Patologia e Medicina Laboratorial, v. 39, n. 1, 2003. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1676-24442003000100004&lng=pt&nrm=iso&tlng=pt>.

GÉRON, Aurélien. **Mãos à Obra : Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Rio de Janeiro: Copyright © 2019 da Starlin Alta Editora e Consultoria Eireli, 2019.

KWON, Charles e FARRELL, Philip M. **The Magnitude and Challenge of False-Positive Newborn Screening Test Results**. Archives of Pediatrics & Adolescent Medicine, v. 154, n. 7, p. 714, 1 Jul 2000. Disponível em: <http://archpedi.jamanetwork.com/article.aspx?doi=10.1001/archpedi.154.7.714>.

LUMERTZ, Magali Santos e colab. **False-negative newborn screening result for immunoreactive trypsinogen: A major problem in children with chronic lung disease**. Jornal Brasileiro de Pneumologia, v. 45, n. 3, p. 3–4, 2019.

PENG, Gang e colab. **Reducing False-Positive Results in Newborn Screening Using Machine Learning**. International Journal of Neonatal Screening, v. 6, n. 1, p. 16, 3 Mar 2020. Disponível em: <https://www.mdpi.com/2409-515X/6/1/16>.

RIBEIRO, Maria Natália Alves e colab. **Fibrose cística: histórico e principais meios para diagnóstico**. Research, Society and Development, Historia da FC, v. 10, n. 3, p. e11710313075, 8 Mar 2021.

RIBEIRO ROSA, Fernanda e colab. **Cystic fibrosis: a clinical and nutritional approach**. Rev. Nutr, v. 21, n. 6, p. 725–737, 2008.