Paulo Rogério Siqueira Custódio

ANÁLISE DE DADOS DO PROGRAMA DE TRIAGEM NEONATAL DE FIBROSE CÍSTICA ATRAVÉS DA QUANTIFICAÇÃO DE IRT – TRIPSINOGÊNIO IMUNORREATIVO UTILIZANDO APRENDIZADO DE MÁQUINA PARA REDUZIR FALSOS POSITIVOS

DATA ANALYSIS OF NEONATAL SCREENING PROGRAM OF CYSTIC FIBROSIS THROUGH THE QUANTIFICATION OF IRT – TRYPSINOGEN IMMUNORREATIVE USING MACHINE LEARNING TO REDUCE FALSE POSITIVES

São José dos Campos - SP

2023

Paulo Rogério Siqueira Custódio

ANÁLISE DE DADOS DO PROGRAMA DE TRIAGEM NEONATAL DE FIBROSE CÍSTICA ATRAVÉS DA QUANTIFICAÇÃO DE IRT – TRIPSINOGÊNIO IMUNORREATIVO UTILIZANDO APRENDIZADO DE MÁQUINA PARA REDUZIR FALSOS POSITIVOS

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia Biomédica da Universidade do Vale do Paraíba, como parte dos requisitos necessários para a obtenção do título de Mestre em Engenharia Biomédica.

Universidade do Vale do Paraíba - UNIVAP

Instituto de Pesquisa e Desenvolvimento - IP&D

Programa de Pós-Graduação em Engenharia Biomédica

Orientador: Dr. Virginia Klausner de Oliveira

São José dos Campos - SP

2023

Fundação Valeparaibana de Ensino

BIBLIOTECA

TERMO DE AUTORIZAÇÃO DE DIVULGAÇÃO DA OBRA

Ficha catalográfica

Custódio, Paulo Rogério Siqueira
Análise de dados do programa de triagem neonatal de fibrose cística através da quantificação de IRT – tripsinogênio imunorreativo utilizando aprendizado de máquina para reduzir falsos positivos / Paulo Rogério Siqueira Custódio; orientadora, Virginia Klausner de Oliveira. - São José dos Campos, SP, 2023.
1 CD-ROM, 105 p.

Dissertação (Mestrado Acadêmico) - Universidade do Vale do Paraíba, São José dos Campos. Programa de Pós-Graduação em Engenharia Biomédica.

Inclui referências

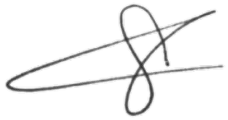
1. Engenharia Biomédica. 2. Fibrose cística. 3. Aprendizado do computador. 4. Triagem neonatal. I. Oliveira, Virginia Klausner de , orient. II. Universidade do Vale do Paraíba. Programa de Pós-Graduação em Engenharia Biomédica. III. Título.

Eu, Paulo Rogério Siqueira Custódio, autor(a) da obra acima referenciada:

Autorizo a divulgação total ou parcial da obra impressa, digital ou fixada em outro tipo de mídia, bem como, a sua reprodução total ou parcial, devendo o usuário da reprodução atribuir os créditos ao autor da obra, citando a fonte.

Declaro, para todos os fins e efeitos de direito, que o Trabalho foi elaborado respeitando os princípios da moral e da ética e não violou qualquer direito de propriedade intelectual sob pena de responder civil, criminal, ética e profissionalmente por meus atos.

São José dos Campos, 14 de Março de 2024.



Autor(a) da Obra

Data da defesa: 07 / 03 / 2023

PAULO ROGÉRIO SIQUEIRA CUSTÓDIO

"ANÁLISE DE DADOS DO PROGRAMA DE TRIAGEM NEONATAL DE FIBROSE CÍSTICA ATRAVÉS DA QUANTIFICAÇÃO DE IRT – TRIPSINOGÊNIO IMUNORREATIVO UTILIZANDO APRENDIZADO DE MÁQUINA PARA REDUZIR FALSOS POSITIVOS."

Dissertação aprovada como requisito parcial à obtenção do grau de Mestre, do Programa de Pós-Graduação em Engenharia Biomédica, do Instituto de Pesquisa e Desenvolvimento da Universidade do Vale do Paraíba - Univap, pela seguinte banca examinadora:

Prof.ª Dr.ª Lucia Vieira	Lucia Vieira
Prof.ª Dr.ª Virginia Klausner de Oliveira	Virginia Klausner de Oliveira
Prof.ª Dr.ª Kumiko Koibuchi Sakane	Kumiko Koibuchi Sakane
Prof. Dr. Guilherme M. Ogawa	Guilherme Maerschner Ogawa
Prof. Dr. Carlos M. Gurjão de Godoy	Carlos Marcelo Gurjão de Godoy

Prof.ª Dr.ª Lúcia Vieira

Diretora do IP&D – Univap

São José dos Campos, 07 de março de 2023.

AGRADECIMENTOS

Gostaria de agradecer a todos que estiveram ao meu lado me apoiando de todas as formas, durante todo o período do mestrado. Em especial agradeço a minha orientadora Virgínia que me apoiou e deu todo o suporte para realização da pesquisa. A Intercientifica por ter me apoiado me liberando para assistir as aulas, dando apoio em conseguir parcerias, com todo suporte que precisei. Aos laboratórios que ouviram a pesquisa e tiveram sua contibuição nesta pesquisa, em especial a APAE de São Luiz do Maranhão e a APAE de Salvador.A Danielle minha namorada que viveu todas as fases dessa pesquisa sendo minha maior ouvinte. A meus pais que sempre me apoiaram e sempre estiveram comigo. Também gostaria de agradecer a banca aqui presente por terem aceitado a fazer parte desse momento. A todos muito obrigado!

Este trabalho é dedicado a todos que contribuíram, direta ou indiretamente, para sua realização. Seu apoio, colaboração e inspiração foram fundamentais ao longo desta jornada.

RESUMO

Essa pesquisa trata-se da utilização de florestas aleatórias para criação de uma metodologia para desenvolvimento de um modelo capaz de fazer previsão de casos verdadeiros positivos na triagem neonatal de fibrose cística, utilizando dados sintéticos para treinamento do modelo e variando os parâmetros do modelo para buscar aquele que retorne o melhor poder preditivo, assim tendo uma prova de conceito para utilização dessa metodologia para encontrar um modelo capaz de fazer a previsão de verdadeiros positivos para fibrose cística em um banco de dados real. A fibrose cística é uma doença, que faz parte do programa nacional de triagem neonatal brasileiro que é causada por mutações do gene de Condutância Transmembrana da Fibrose Cística (CFTR, do inglês Cystic Fibrosis Transmembrane Conductance Regulator). Esta doença é caracterizada pela produção de um muco espesso podendo causar problemas respiratórios, gastrointestinais, complicações metabólicas entre outras enfermidades que variam de acordo com as mais de 2000 mutações existentes. Já a floresta aleatória é um algoritmo comum de aprendizado de máquina que consiste em utilizar um banco de dados para treinar um modelo de inúmeras árvores de decisão capazes de criar critérios para tentar explicar um dado alvo baseando-se em atributos, para poder fazer previsões em um novo banco de dados desconhecido utilizando somente os atributos. Esse tipo de tecnologia vem ganhando espaço na área da saúde principalmente na área de diagnóstico por conta de seu alto poder preditivo. A triagem dessa doença faz parte do programa de tiragem neonatal brasileiro através do teste do pezinho, que é feita com a quantificação da tripsina imunorreativa, este exame possui alta incidência de falsos positivos. Se esta prova de conceito for positiva pode-se fornecer atendimento precoce a esses pacientes, aumentando suas expectativas de vida. Para isso, foi utilizado índices gerais (número de pacientes triados e número de exames alterados) de triagens e diagnóstico do laboratório de triagem neonatal APAE (Associação de Pais e Amigos dos Excepcionais) de São Luís do Maranhão, para desenvolver uma metodologia que buscasse sempre a melhor sensibilidade para o modelo. Os resultados obtidos nesta dissertação com dados sintéticos mostram que essa metodologia pode permitir alcançar o objetivo devido a melhora que ela trouxe na sensibilidade do modelo mesmo que utilizando dados sintéticos para o treinamento, o que tende a melhorar quando utilizar dados reais de pacientes, pois a correlação desses dados será maior o que fará com que o modelo tenha melhor ajuste sobre os dados sendo capaz de explica-los com a sensibilidade e precisão superior à obtida com os dados sintéticos.

Palavras-chave: ; Fibrose cística; florestas aleatórias; aprendizado de máquina; triagem neonatal; falsos positivos; tripsina imunorreativa.

DATA ANALYSIS OF NEONATAL SCREENING PROGRAM OF CYSTIC FIBROSIS THROUGH THE QUANTIFICATION OF IRT - TRYPSINOGEN IMMUNORREATIVE USING MACHINE LEARNING TO REDUCE FALSE POSITIVES

ABSTRACT

This research is about the use of Random Forests to predict true positive cases in neonatal screening for cystic fibrosis disease. Cystic fibrosis is a disease, which is part of the Brazilian national neonatal screening program that is caused by mutations in the CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) gene. This disease is characterized by the production of a dense mucus that can cause respiratory, gastrointestinal, metabolic complications among other diseases that can vary according to the more than 2000 existing mutations. Random forests are a field of study of machine learning that consists of using a database to train an algorithm of multiple decision trees that are capable of creating criteria to try to explain a given target based on attributes, in order to be able to do predictions in a new unknown database using attributes only. This type of technology has been gaining ground in the health area, mainly in the area of diagnosis due to its high predictive power. So this research proposes the development of a methodology to generate a model capable of learning from an artificial database of patients screened for cystic fibrosis and predict which of them have the greatest chance of being a true positive. Screening for this disease is part of the Brazilian neonatal screening program through the tootsy test, which is screened through the quantification of immunoreactive trypsin, which have a high incidence of false positives, thus being able to provide early care to these patients, increasing their life expectancy. For this, general screening and diagnostic indexes from the neonatal screening laboratory APAE (Association of Parents and Friends of the Handicapped) in São Luís do Maranhão were used to develop a methodology that always sought the best sensitivity for the algorithm. The results obtained in this dissertation with artificial data show that this technique can allow reaching the objective due to the improvement it brought in sensitivity even when using artificial data for model training, which tends to improve when using real patient data, since the correlation of these data will be greater, which will make the model have a better fit on the data, being able to explain them with sensitivity and precision superior to that obtained with artificial data.

Keywords: Cystic fibrosis; random forests; machine learning; neonatal screening; false positives; immunoreactive trypsin.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A Fibrose Cística (FC) é uma doença genética 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 et al., 2008). A FC tem capacidade de atacar todos os sistemas do corpo humano, principalmente, os sistemas respiratório, gastrointestinal e reprodutor. A FC comumente se manifesta na infância/adolescência. O diagnóstico precoce e tratamento são as melhores formas de se dar sobrevida ao paciente (RIBEIRO et al., 2021). O IRT – tripsinogênio imunorreativo da enzima pancreática, é encontrado elevadamente em pacientes com fibrose cística (CABELLO et al., 2003). Portanto, é fundamental realizar o teste de quantificação do IRT (tripsinogênio imunorreativo) na triagem Neonatal. No entanto, é comum resultados falsos positivos, e muito menos frequentes, os falsos negativos. Em uma triagem, o que mais se deseja é reduzir o número desses resultados errôneos. Na literatura, conhece mais fatores associados a falsos positivos do que a negativos, esses resultados errôneos podem causar impactos psicossociais na família do recém-nascido devido à preocupação com o diagnostico falso positivo (LUMERTZ et al., 2019). O tardar do diagnóstico correto devido aos resultados inexatos faz com que o tratamento nutricional que proporciona a sobrevida ao paciente se atrase, já que a o tempo para se ter o diagnóstico é um dos fatores que levam a desnutrição do paciente com FC (FARRELL et al., 1997). Segundo um estudo de realizado em Ontário, Canadá outro fator que atrasa o diagnóstico de recém nascidos com resultados falsos positivos é o fato de apresentarem maior uso de serviços ambulatoriais (HAYEEMS et al., 2017). A triagem Neonatal aumenta significavelmente a detecção precoce de distúrbios congênitos, embora ainda seja acompanhada de um grande número de falsos positivos. Esse fato é um efeito adverso advindo da alta sensibilidade exigida no programa para que se evite casos de falsos positivos (KWON; FARRELL, 2000).

Visando ajudar a mitigar esse tipo de problema no diagnóstico da FC, esta dissertação propõe a aplicação de técnica de aprendizado de máquina (Machine Learning) utilizando para sua aprendizagem um banco de dados sintético que simule um banco de pacientes triados para fibrose cística, levando em consideração suas características para determinar um perfil dos recém nascidos testados como positivos. Foram consideradas características físicas, condições de nascimento, tempo de coleta de amostra (sangue em papel filtro), entre outras informações dos recém nascidos que estivessem acessíveis e se mostrassem necessárias a partir do estudo da literatura. Com a utilização de técnicas aprendizado de máquina (*Machine Learning*) será criado um modelo de florestas aleatórias

para aprender com esses dados artificiais, fazendo com que se tenha capacidade de fazer previsões para uma nova base de dados, e dessa forma conseguir melhorar a previsão da amostra testada ser realmente um verdadeiro positivo ou um falso positivo. Para que o modelo possa fazer essa análise com precisão um dado que é de extrema importância é o *feedback*, resultado confirmatório que mostra se a criança realmente foi diagnosticada para a doença em questão ou não.

Para desenvolvimento do modelo de validação e da metodologia para melhorar o seu desempenho, foi utilizado um banco de dados gerado artificialmente a partir de índices de triagens e diagnósticos referente a um período de 5 anos (entre os anos de 2017 e 2021) de pacientes triados pelo laboratório de triagem neonatal da APAE de São Luis do Maranhão. O estudo que foi feito nessa dissertação baseou-se na criação do banco de dados artificial contendo caraterísticas clínicas dos pacientes e os resultados de triagem e diagnóstico, utilizando-os para treinar e testar o modelo de florestas aleatórias, variando seus parâmetros até encontrar o melhor sensibilidade possível no conjunto de teste, assim tendo uma metodologia que futuramente possa ser aplicada ao banco de dados real.

Com a utilização da metodologia proposto nesta pesquisa mantendo resultados com alto índice de sensibilidade quando aplicada aos dados reais, será possível o desenvolvimento de um software que auxilie o laboratório conseguir dar maior atenção e um atendimento ágil a pacientes que tenham maior probabilidade de ser diagnosticado com fibrose cística. Assim, o paciente poderá iniciar o tratamento com uma rapidez ainda maior da que se tem hoje. O alto número de falsos positivos podem tardar um diagnóstico após a triagem Neonatal positiva e o aprendizado de máquina empregado nesse campo aumenta a eficiência e agilidade do diagnóstico (PENG et al., 2020).

1.1 Referencial Teórico

A ideia desta dissertação surgiu após perceber que os resultados do teste de quantificação do IRT dos clientes da empresa INTERCIENTIFICA apresentavam um alto índice de falsos positivos quando comparado aos outros exames que a empresa também fornece kits de triagem. Assim em conversas com clientes, detectou-se que haviam alguns fatores como o peso do recém-nascido, que era ser um fator que poderia causar esse resultado errado, ou até mesmo o tempo de coleta da amostra em papel filtro, poderiam influenciar o resultado. Então, foi feito uma pesquisa sobre falsos positivos relacionados a quantificação do IRT, e foi verificado que pode haver outros fatores que influenciam no resultado FP, como a etnia, por exemplo. Afim de abordar esse problema, foi decidido utilizar técnicas de aprendizado de máquina para ajudar a diminuir os índices de falsos positivos.

Então para elaboração desta dissertação, foi realizada uma pesquisa bibliográfica sobre o tema relacionando a triagem neonatal com aprendizado de máquina. Onde

encontrou-se uma pesquisa realizada nos Estados Unidos da América no estado da Califórnia, na qual foi aplicada um modelo de florestas aleatória para reduzir o número de falsos positivos na determinação de doenças metabólicas do programa de triagem neonatal da Califórnia. Então, foi idealizado a replicação de um estudo similar para reduzir o índice de falsos positivos de quantificação do IRT no programa de triagem neonatal brasileiro (PENG et al., 2020).

A grande motivação para realização desta dissertação, encontra-se no fato que o parâmetro IRT é utilizado para determinação da triagem de fibrose cística, uma doença que exige um tratamento rápido para dar maior sobrevida ao paciente portador da doença (CHEILLAN et al., 2005).

1.2 Justificativa

Atualmente a triagem da fibrose cística faz parte do programa nacional de triagem neonatal oferecido pelo SUS juntamente com o teste do pezinho, por meio da quantificação do tripsinogênio imunorreativo. O teste do suor, que consiste na quantificação de eletrólitos no suor, é realizado para confirmação do diagnóstico, caso o teste do suor seja duvidoso, o paciente é encaminhado para pesquisa de mutações genéticas (BONFIM et al., 2019). Porém, a quantificação dessa enzima apresenta um alto índice de falsos positivos, por exemplo, foi constado que a etnia é um fator que pode levar a um falso positivo (CHEILLAN et al., 2005). Como a fibrose cística é uma doença que afeta muito a vida do paciente e de sua família, a FC necessita de um diagnóstico preciso o mais cedo possível. O resultado falso positivo do exame de fibrose cística pode levar a família a problemas psicossociais devido a angústia gerada até se ter uma confirmação de fato do diagnóstico (HAYEEMS et al., 2016).

1.3 Objetivos

Essa dissertação propõe uma prova de conceito, onde foi desenvolvido um modelo utilizando técnicas de Aprendizado de Máquina (AM) para diminuir os índices de falsos positivos na triagem de fibrose cística utilizando dados sintéticos, o modelo deve fazer previsões de pacientes verdadeiros e falsos positivos baseando sua análise nas características físicas, de coleta de amostra e nascimento. O modelo deve ser capaz de criar critérios e encontrar relações entre essas características e o diagnóstico do paciente para realizar a previsão.

Para atingir o objetivo, foram realizados estudos bibliográficos dos fatores que podem levar a um falso positivo na triagem da fibrose cística. Laboratórios que realizam esse exame foram contados para uma possível parceria onde possa se obter os dados dos pacientes, e esses dados foram utilizados para treinar o modelo de aprendizado de máquina que será desenvolvido e que deverá ter a capacidade de fazer as previsões do paciente

ser um verdadeiro ou falso positivo para fibrose cística reduzindo a taxa atual de falsos positivos na triagem neonatal brasileira sem influenciar negativamente na sensibilidade atual, ou seja sem aumentar o número de falsos negativos.

Parte I

FIBROSE CÍSTICA

2 INTRODUÇÃO A FIBROSE CÍSTICA

Por milhares de anos, humanos morriam de fibrose cística pelo mundo, mas a doença não tinha esse nome em seu princípio. Em antigos registros populares folclóricos do norte da Europa, foi registrado um dito que garantia em que, se você beijasse uma criança com gosto salgado, ela seria "enfeitiçada"e morreria prematuramente (FRÍAS et al., 2019). Hoje se sabe que a grande quantidade de sal no suor é um sinal clássico de FC. Geralmente a família percebe esse sinal ao beijar o recém-nascido (RC) e ele apresenta um gosto salgado (CUNNINGHAM; TAUSSIG, 2013).

Em 1936, Guido Feanconi estabeleceu relação entre a doença de fibrose cística no pâncreas e a doença celíaca (FRÍAS et al., 2019) e em 1938 Dorothy H. Andersen fez análises em pessoas que foram diagnosticadas com FC no pâncreas e doença celíaca, onde ele constatou que a lesão pancreática não implicava na causa da doença celíaca, mas indicava que todos os pacientes com FC no pâncreas que sobreviveram pelo menos um ano eram clinicamente indistinguíveis de pacientes com a doença celíaca, assim estabelecendo que a doença celíaca pode ser estabelecida por essa patologia (ANDERSEN, 1938).

Em 1943, Sidney Farber e colaboradores deu o nome alternativo ‘mucovidose’ devido ao engrossamento do muco obtido do duodenal de paciente com FC quando comparado a outros pacientes estudados(FARBER; SHWACHMAN; MADDOCK, 1943). Neste trabalho, Farber, Shwachman e Maddock (1943) estenderam a análise de seu relatório preliminar da atividade enzimática no duodenal para 150 pessoas e estabeleceram uma relação entre a medida da atividade enzimática no duodenal com o diagnóstico da doença (doença celíaca idiopática, fibrose cística pancreática, esteatorréia idiopática, espru tropical). Neste estudo, foi constato que pacientes com lesão pancreática não necessariamente apresentavam características clinicas da síndrome celíaca, doença, a qual era atribuída anteriormente esse quadro clinico de lesão pancreática. Farber, Shwachman e Maddock (1943) detectaram em seu estudo que em 4 RN que apresentavam muitas das características clínicas da síndrome celíaca e mostravam atividade enzimática no duodenal reduzida ou ausente, no exame pós morte foram encontradas alterações obstrutivas no ducto-acinar e graus variáveis de fibrose no pâncreas. Como na doença pancreática, os pacientes morriam quase invariavelmente e os paciente com a doença celíaca idiopática respondiam bem ao tratamento dêitico, se fazia necessária a diferenciação entre essas duas doenças, já que eram confundidas clinicamente. Durante esse estudo, também foi medido o nível de tripsina e viu-se que ela por si só pode ser utilizada para um diagnóstico preciso da fibrose pancreática (FARBER; SHWACHMAN; MADDOCK, 1943).

O pesquisador Sant’Agnese em 1948 após uma onda de calor em Nova York descobriu que os pacientes que tem FC produziam um alto volume de sal em seu suor

(DI SANT'AGNESE et al., 1953). Em 1951, houve a conexão entre o transporte de sal pelo corpo e a FC. Kessler e Andersen, em um estudo realizado no Hospital de Bebês, perceberam que crianças estavam sendo internadas com os mesmos sintomas agudos, prostração por calor, vômitos e sinais de choque, devido a alguma infecção. Após a reidratação, somente uma criança não apresentou melhora (De Almeida Matos; MARTINS, 2019). Em 1954, Sant'Agnese e seus colaboradores submeteram recém nascidos com FC a um estresse térmico leve, onde eles foram mantidos em um quarto a 32°C a uma umidade de 50%, e assim foi notada uma anormalidade na função das glândulas sudoríparas, as quais produziram em excesso sódio e cloreto de potássio em comparação com os pacientes controle (SHWACHMAN; ANTONOWICZ, 1962). Com essa descoberta foi possível o desenvolvimento do teste do suor para diagnosticar a fibrose cística pancreática. Em 1959, fazia um teste no qual era medida a concentração dos eletrólitos no suor utilizando pilocarpina por iontoforese (GIBSON; COOKE, 1959).

Em 1955, foi instituída a Fundação da Fibrose Cística por um grupo de pais determinados em salvar seus filhos. Em 1961, esse grupo formou uma rede de centro de atendimentos credenciados e criou dois centros de tratamento especializados no tratamento da FC. No ano de 1966, foi lançado um registro de dados de pacientes que coletava suas informações de saúde em centros de atendimentos credenciados. Este registro foi o que impulsionou o progresso no cuidado da FC e na pesquisa, se tornando um modelo para outros registros de doença (FOUNDATION, 2022d).

Em 1979, Crossley mediu o nível de tripsina imunorreativa (IRT) em pacientes com fibrose cística em seus primeiros meses de vida, estabelecendo esse ensaio como um potencial exame de triagem neonatal para FC nos recém nascidos. O teste foi realizado utilizando sangue seco em papel filtro e para determinar o nível da tripsina imunorreativa. Foi realizado um imunoensaio, no qual teve como amostra sangue coletado de 26 crianças com FC entre um 1 mês e 18 anos, e detectou-se que essas crianças quando nos primeiros meses de vida demonstravam um alto nível de IRT (CROSSLEY; ELLIOTT; SMITH, 1979).

O especialista em fluido epitelial, Paul Quinton, em 1983, fez um estudo com pacientes controle e com FC onde micro perfundiram ductos sudoríparos isolados e viu em seus resultados que a permeabilidade anormalmente baixa de cloreto na FC leva a baixa absorção de NaCl no ducto sudoríparo (QUINTON, 1983). Neste mesmo ano, Knowles e colaboradores constatou que pacientes com FC apresentam um reabsorção de íons de sódio nas via aéreas, maior do que os pacientes controle (KNOWLES et al., 1983). Esse estudo vai de encontro com o trabalho de Boucher e colaboradores onde ele tentou aumentar a absorção de Cl- do epitélio de pacientes com FC e concluiu que os epitélios das vias aéreas desse pacientes absorvem potássio (Na) em uma taxa acelerada (BOUCHER et al., 1986).

A fibrose cística foi reconhecida como a mais importante doença hereditária,

potencialmente letal. O gene da FC foi identificado em 1989, clonado e sequenciado, possibilitando o conhecimento dos mecanismos bioquímicos responsáveis pela fisiopatogenia da doença, o aconselhamento genético e o tratamento de suas complicações. Essa descoberta foi feita por Lap-Chee Tsui e colaboradores onde ele fez uma análise genética com pacientes com FC e seus pais (TSUI et al., 1989b; Ribeiro ROSA et al., 2008). Originalmente, a FC era caracterizada como um conjunto de síndromes que se relacionavam. Hoje, ela é reconhecida como uma doença única, onde seus diversos sintomas derivam-se da ampla distribuição no tecido do produto gênico que é defeituoso para FC, o canal iônico e regulador CFTR - Regulador de Condutância Transmembrana da Fibrose Cística (LYCZAK; CANNON; PIER, 2002). O CFTR é uma glicoproteína essencial na membrana apical das células epiteliais para manter a homeostase iônica e fluida, ela é a única integrante da família de proteínas do cassete de ligação de nucleotídeo de adenina (ABC), conhecido por ser um canal iônico. A ausência dessa atividade precisamente regulada do canal iônico resulta na falha da homeostase iônica e da água nas superfícies epiteliais exócrinas. Isso ocorre na maioria dos tecidos exócrinos, mas com as consequências mais graves no pâncreas, onde a deficiência de líquido rico em bicarbonato e a secreção de enzimas prejudicam a digestão e absorção intestinal e nas vias aéreas do pulmão, nesse local, o acúmulo de muco viscoso e a colonização por microrganismos causam respostas inflamatórias prejudiciais e perda de função (RIORDAN, 2008).

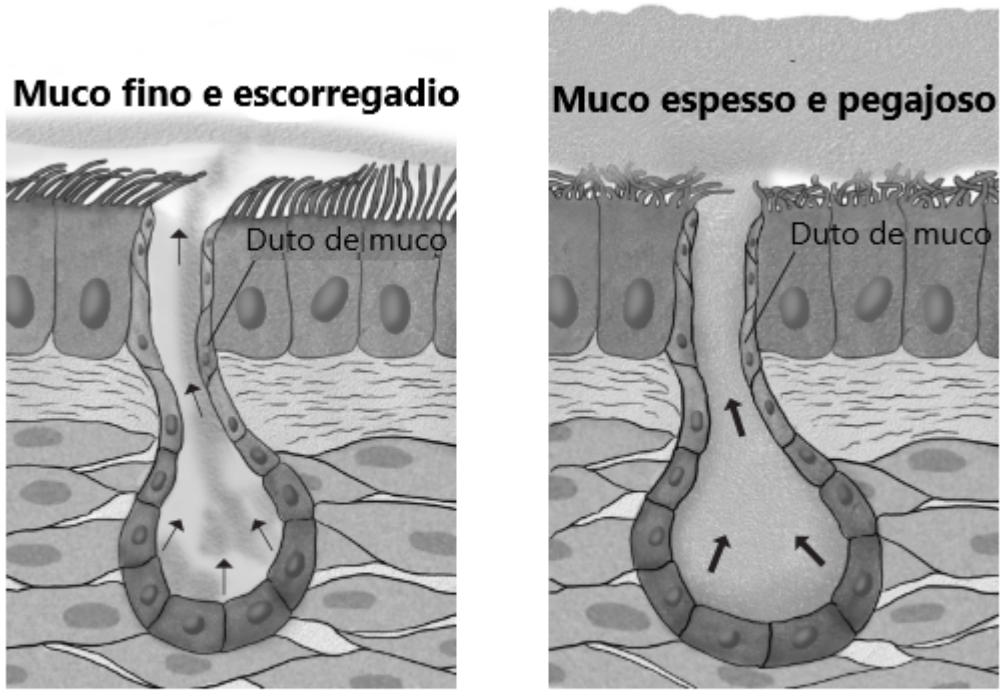
A FC é uma doença hereditária que faz com que algumas glândulas do corpo não funcionem corretamente, as glândulas exócrinas que são responsáveis pela excreção de secreções. Elas normalmente produzem secreções finas e lúbricas, incluindo suor, muco, lágrimas, saliva e sucos digestivos. Essas secreções se movem através de dutos para a superfície corporal ou para órgãos ocos como intestinos ou vias aéreas. Elas ajudam para que o corpo funcione normalmente, porém para pessoas diagnosticadas com FC estas secreções produzidas são densas, assim obstruindo os dutos e outras passagens (1). Esta doença também afeta o nível de sal (sódio e cloreto) e potássio produzido no suor fazendo com que esses níveis sejam muito altos, assim podendo causar problemas em períodos de aumento de transpiração (CUNNINGHAM; TAUSSIG, 2013).

O diagnóstico sintomático de fibrose cística está associado a complicações de curto e longo prazo, incluindo deficiência de crescimento, atrofiamento, definhamento, deficiências de vitaminas e minerais, infecções pulmonares recorrentes associadas à diminuição da função pulmonar e hospitalizações recorrentes (BOROWITZ et al., 2009).

Esta doença atinge com maior frequência a população branca sem diferença entre sexos, tendo maior prevalência no povo caucasiano. Tem como países mais afetados os Estados Unidos, Europa e Canadá, com variação de 1:2.000 a 3.500 RN caucasianos portadores do gene (De Almeida Matos; MARTINS, 2019). Os índices no Brasil estão estimados em 1:10.000 nascidos vivos, entretanto, como se trata de um país com raça

miscigenada, esse índice pode variar de acordo com a região, tendo a região sul, índices próximos ao encontrados na Europa, cerca de 1:2.500 nascidos vivos (SANTOS et al., 2005) (MARIANO; CONDE, 2017).

Figura 1 – Células pulmonares com FC e sem FC



sem FC
em uma glândula mucosa, as secreções se movem facilmente através do muco

com FC
FC faz com que o muco seja espesso e pegajoso

Fonte: (CUNNINGHAM; TAUSSIG, 2013).

3 CARACTERÍSTICAS DA FIBROSE CÍSTICA

Nesta seção, serão abordados de forma aprofundada os aspectos históricos, clínicos e característicos de pacientes com FC, baseando-se em estudos anteriores sobre o tema, guias e diretrizes que regulam os procedimentos relacionados a FC, para descrever as causas dessa doença, os sintomas que ela causa em pacientes afetados, as formas de realização dos diagnósticos e tratamentos.

3.1 Causa

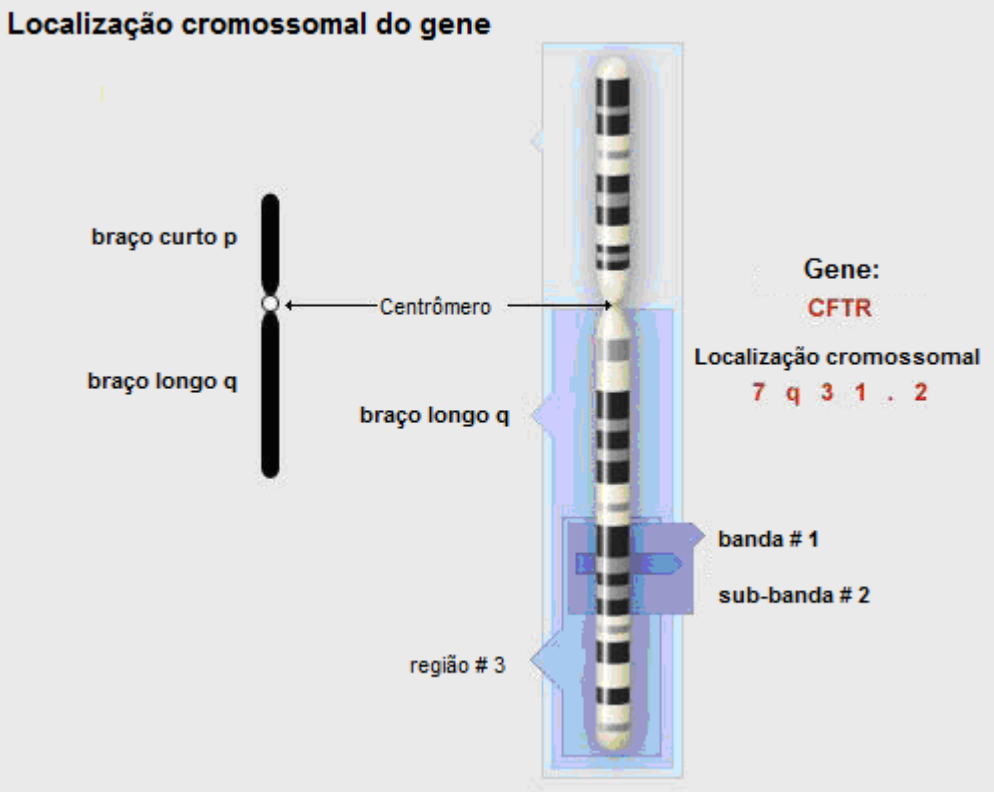
Com a descoberta de Paul Quinton que viu que a impermeabilidade do epitélio dos pacientes com FC aos cloretos, conclui-se que o causador dessa doença era algum canal transportador de íon que deveria estar com mal funcionamento nesses pacientes. Foi então que os cientistas da época se engajaram para encontrar o gene responsável por causar a fibrose cística (FIRMIDA; MARQUES; COSTA, 2011).

A identificação do gene da fibrose cística foi descrita pela primeira vez em setembro de 1989 por Lap-Chee Tsui, Francis Collins e colaboradores, onde publicaram na revista Science a análise genética, a clonagem e caracterização do DNA complementar e a caminhada e salto cromossômico, sobre o gene causador da FC (COLLINS et al., 1989; TSUI et al., 1989b; TSUI et al., 1989a).

O geneticista Lap-Chee Tsui e colaboradores fizeram uma análise genética afim de identificar o gene responsável pela FC, para isso foi realizado uma pesquisa genética nos membros de famílias de paciente com FC, onde nesse estudo forneceu evidencias da existência de um único lócus de FC no braço longo do cromossomo 7 humano na região q31.2 (2) e que marcadores de DNA intimamente ligados indicavam que podiam haver várias mutações de FC. Para estudar as famílias foram utilizados RFLPs, Polimorfismo de Comprimento de Fragmentos de Restrição, isso para determinar cada um dos segmentos de DNA isolados dos experimentos de caminhada e salto cromossômico (TSUI et al., 1989b).

No estudo da clonagem e caracterização do DNA complementar Tsui, viu fazendo a análise da sequência de cDNA sobrepostos, um polipeptídio de 1480 aminoácidos com uma massa molecular de 168.138 Daltons.Nesta análise, percebeu-se também que está proteína apresentava características da glicoproteína P de resistência a múltiplas drogas em mamíferos e várias outras proteínas associadas a membrana, prevendo assim que o produto do gene da FC está relacionado com o transporte de íons através da membrana, assim sendo provavelmente um membro da família de super proteínas de membrana, o produto desse gene foi denominado CFTR – Regulador de condutância transmembrana da fibrose cística (TSUI et al., 1989a).

Figura 2 – Localização do gene CFTR



Fonte: https://bit.ly/3KwWCs6

Em 1993, Welsh estudou os mecanismo moleculares da disfunção do CFTR, onde constatou que o compreendimento na época sobre esse gene ainda era superficial, se fazendo importante entender como as mutações da disfunção do CFTR e relacionar esses dados com os fenótipos clínicos, podendo fornecer *insights* sobre esse problema (WELSH; SMITH, 1993). Já foram descritas mais de 2000 mutações do gene CFTR que apresentam distribuição de forma diferente. Esse conhecimento do perfil genético contribuiu para o entendimento da relação fenótipo-genótipo, principalmente em populações mistas (MOTA et al., 2018).

3.2 Sintomas

A fibrose cística é uma doença caracterizada pelo aumento da produção de muco em alguns órgãos ocasionando doenças pulmonares por obstrução crônica, insuficiências pancreática e uma alta concentração de eletrólitos no suor dos pacientes portadores de FC (FURTADO; LIMA, 2003). Embora a FC seja uma doença multissistêmica, a principal causa de morbidade e mortalidade causada é representada pela doença pulmonar. Sendo um ciclo vicioso de acúmulo de muco nas vias aéreas, infecções recorrentes que levam a danos epiteliais, inflamações crônicas, remodelação do tecido e deterioração do tecido pulmonar. Esses sintomas fazem com que os pacientes necessitem do uso de antibióticos orais em eventos leves e em casos mais rígidos podendo levar a hospitalização com antibiótico em

via intravenosa (VENDRUSCULO; Fagundes Donadio; Araújo Pinto, 2021).

Infecções recorrentes ou crônicas no sistema respiratório geralmente se manifestam nos RN com tosse, produção de escarro e respiração ruidosa. A tosse acaba sendo o sintoma mais desconfortável e, frequentemente, é acompanhada de saliva, sufocação, vômito e distúrbios de sono. Outro sintoma pode ser o íleo mecônial, decorrente da obstrução do íleo por mecônio viscoso, podendo ser um primeiro sinal e presente em cerca de 13% a 18% dos pacientes com FC. Em pacientes que não apresentam esse sintoma a doença pode ser precedida por atraso na recuperação de peso e ganho inadequado na quarta à sexta semana de vida. Outro sintoma é o suor do RN salgado que dava o nome a doença no seu início de doença do beijo salgado (ROSENSTEIN, 2019). Na Tabela 1, estão descritas as características da FC, ressaltando que com o passar do tempo e melhora nos tratamentos dando maior sobrevida aos pacientes acarreta em mais comorbidades (BELL et al., 2019).

Tabela 1 – Características fenotípicas da FC

Respiratória:	Bronquite com infecção crônica
	Pneumotórax
	Hemoptise
	Insuficiência respiratória
	Rinosinusite crônica e pólipo nasal
Gastrointestinal (Luminal):	Íleo meconial
	Doença de refluxo gastresofágico
	Síndrome de obstrução intestinal distal
	Constipação crônica
	Prolapso Retal
	Intussuscepção
	Câncer colorretal e polipose colônica
	Outras malignidades gastrointestinais
Gastrintestinal (hepatobiliar):	Insuficiência pancreática
	Pancreatite aguda recorrente (em pacientes com insuficiência pancreática)
	Lodo biliar ou colelitíase
	Cirrose biliar
Complicações metabólicas:	Diabetes relacionado à fibrose cística: complicações microvasculares ($\geq$ 10 anos a partir do diagnóstico)
	Doença óssea relacionada à fibrose cística ou osteoporose: aumento do risco de fratura
	Cálculos ureterais
	Oligomenorreia
Infertilidade masculina:	Ausência bilateral congênita do ducto deferente

Adaptado de (BELL e colab., 2019)

Hasiak fez um estudo sobre os pacientes com mortes relacionadas a FC entre os anos 1999 a 2017, onde ele relacionou as causas associadas a morte dos 2384 pacientes em que a

FC foi identificada como causa básica da morte. Na Tabela 2, tem-se as causas associadas que o autor identificou em seu estudo, tendo como as doenças mais associadas a morte as doenças respiratórias com 77,0% das causas associadas, as doenças infeciosas com 31,0% e as causas mal definidas com 24,5%, salientando que a somatórias das porcentagens não é 100%, pois alguns pacientes podiam ser acometidos por mais de uma causa (HASIAK; VICENTE; FERREIRA, 2021).

Tabela 2 – Causas de morte relacionadas a FC no Brasil de 1999-2017

Causa da Morte	**N**	**%**
Doenças Infecciosas e parasitarias	788	33,05%
Neoplasias [tumores]	23	0,96%
Doenças do sangue e dos órgãos hematopoiéticos e alguns transtornos imunitários	63	2,64%
Doenças endócrinas, nutricionais e metabólicas	359	15,06%
Transtornos mentais e comportamentais	24	1,01%
Doenças do sistema nervoso	34	1,43%
Doenças do aparelho circulatório	362	15,18%
Doenças do aparelho respiratório	1836	77,01%
Doenças do aparelho digestivo	129	5,41%
Doenças da pele e do tecido subcutâneo	3	0,13%
Doenças do sistema osteomuscular e do tecido conjuntivo	20	0,84%
Doenças do aparelho geniturinário	150	6,29%
Gravidez, parto e puerpério	2	0,08%
Algumas afecções originadas no período perinatal	84	3,52%
Malformações congênitas, deformidades e anomalias cromossômicas	32	1,34%
Sintomas, sinais e achados anormais de exames clínicos e de laboratório, não classificados em outra parte	585	24,54%
Lesões, envenenamento e algumas outras consequências de causas externas	50	2,10%
Causas externas de morbidade e de mortalidade	92	3,86%
Fatores que influenciam o estado de saúde e o contato com os serviços de saúde	2	0,08%

Fonte: Adaptada (HASIAK e colab., 2021).

3.3 Diagnóstico

Alguns métodos foram desenvolvidos ao decorrer da evolução histórica da fibrose cística, dentre eles o mais utilizado para a confirmação é o teste do suor. Esse é um método de teste clássico que mesmo com testes mais elaborados como o teste genético, ele demonstra ser o mais utilizado, mesmo assim ele não se livra de discussões em relação ao seu uso e sua complexidade. Além do teste do suor e do estudo genético do indivíduo, outra forma de diagnóstico da FC é através do teste do pezinho e de uma anamnese minuciosa(RIBEIRO et al., 2021).

A FC é diagnosticada de pelo menos um achado de fenótipo, histórico de FC na família, ou resultado positivo na triagem neonatal, acompanhada de uma evidencia laboratorial. Realizando o teste do suor, onde é coletado o suor corporal para uma análise da concentração de cloreto e sódio, para auxiliar no diagnóstico. A diferença de potencial nasal (DPN), um método para detectar anormalidades no transporte iônico no epitélio respiratório, esse exame fundamente o diagnóstico em um quadro positivo. Além desses exames tem-se como exame complementar o diagnóstico a identificação das mutações causadoras da FC nos genes do CFTR, avaliação da função pulmonar através de radiografias ou teste do esforço, azoospermia obstrutiva e pela dosagem da tripsina imunorreativa (IRT), marcador da insuficiência pancreática que auxilia na triagem neonatal e teste de "Screening" realizado com o teste do pezinho (SILVA et al., 2015).

A concentração de tripsina imunorreativa pode ser medida utilizando sangue em papel filtro, sua concentração é de duas a cinco vezes maior em RN portadores de fibrose cística. Porém esse nível cai entre um a dois meses e então se perde a confiabilidade do resultado. Esse tipo de teste é acompanhado de uma alta taxa de falsos positivos e não é usado para diagnóstico somente para triagem em uma abordagem onde é mensurado duas vezes (WALLIS, 1997).

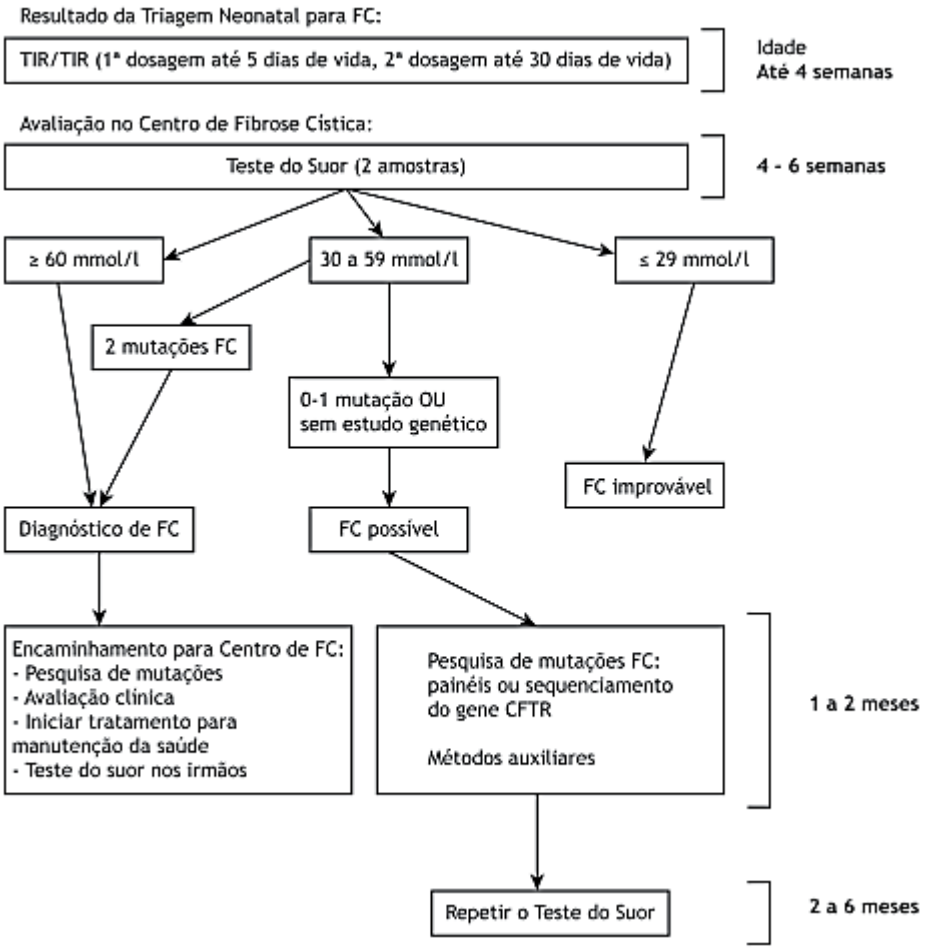
No Brasil em 1992 o Sistema Único de Saúde (SUS) incorporou o popular Teste do Pezinho e em 2001 criou o Programa Nacional de Triagem Neonatal ampliando-o, incorporando o diagnóstico precoce para FC. Tendo como metodologia para triagem baseada na dosagem do IRT. A sensibilidade desse teste se situa ao redor de 95%, porém sua especificidade é baixa girando em torno de 32% a 74% dependendo dos níveis de cortes adotados pelos laboratórios (DAMASCENO, 2010).

O programa de triagem neonatal brasileiro se baseia na quantificação dos níveis de TIR - Tripsina Imunorreativa (do Inglês IRT - *Immunoreactive Trypsinogen*) em duas dosagens, sendo a segunda feita em até 30 dias de vida, tendo as duas dosagens positivas faz-se o teste do suor para confirmação ou a exclusão de FC. Na 3 tem-se um fluxograma que resume o processo de condução em caso de uma triagem neonatal positiva (ATHANAZIO et al., 2017).

O teste do suor é um termo geral que se refere a análise quantitativa e qualificativa do suor para determinar os níveis de concentração dos eletrólitos. Esse teste é indicado em casos de triagem neonatal positiva e sinais clínicos sugestivos para FC. Esse teste geralmente é baseado em duas técnicas, o método de Gibson e Cooke que é realizada em um sistema "in house" ou pelo método comercial Wescor, nesses dois métodos se possuem três estágios: estimulação, coleta e análise do suor (MISHRA; GREAVES; MASSIE, 2005).

O método de Gibson-Cooke se baseia na produção de suor que é estimulada através da iontoforese da pilocarpina, o suor é coletado em papel filtro ou gaze depois levados para análise. Sistema de teste do suor Wescor Macroduct que também utiliza a

Figura 3 – Condução dos casos com triagem neonatal positiva para fibrose cística.



Fonte: (ATHANAZIO e colab., 2017)

iontoforese da pilocarpina, mas a ela é incorporada em gel e o suor coletado em tubos microbore. Alguns centros além da medição de cloreto no suor também fazem medem a condutividade dos eletrólitos no suor, coletando-o em um copo de metal ou em tubo microbore Wescor (COAKLEY et al., 2009). Porém esses métodos podem carregar erros devido a sua complexidade e número de passos, Michael J. Rock e colaboradores realizaram um estudo comparativo entre esses dois métodos com o método de teste do suor CF Quantum e obtive que essa metodologia quantitativa é mais simples e rápida, porém essa nova tecnologia requer aprimoramento para diminuir resultados errôneos em valores de alta concentração de cloreto (ROCK; MAKHOLM; EICKHOFF, 2014).

O teste do suor embora seja considerado padrão ouro para o diagnóstico de FC, tem como fator determinante para sua confiabilidade depende de uma condução do exame por técnicos experientes seguindo as diretrizes rígidas. Então Maria Fátima Servidoni e colaboradores conduziram um estudo transversal com um questionário sobre como era feito o teste em 14 centros que realizam o teste no estado de São Paulo, que possui a maior frequência de FC no Brasil. Nesse estudo foi constado que não havia uniformidade entre

os procedimentos realizados pelos laboratórios, tendo sua maioria divergente das diretrizes internacionais, principalmente em relação a coleta do suor. Os resultados mostram que se faz necessário a padronização dos procedimentos e o treinamento para que os profissionais se qualifiquem, assim tendo um diagnóstico confiável (SERVIDONI et al., 2017)

3.4 Tratamento

Para o tratamento da FC deve-se ter uma equipe multidisciplinar composta por médicos, enfermeiros, nutricionistas, fisioterapeutas e terapeutas em respiração, consultores, farmacêuticos e assistentes sociais. Deve se ter como objetivo terapêutico a manutenção do estado nutricional adequado, prevenção ou tratamento agressivo pulmonar e outras complicações, estimulação de atividade física e proporcionar suporte psicossocial adequado. Esses cuidados promovem sucesso no trabalho e na relação familiar, fazendo com que os pacientes se adaptem a ambientes doméstico, escolar, trabalho, de acordo com sua idade (ROSENSTEIN, 2019). Um trabalho realizado por Tonello e colaboradores detectou que a presença do profissional farmacêutico no time multidisciplinar favorece evitando erros com os medicamentos, evitando discrepâncias, custo de atendimento e quando aliada a tecnologia pode reduzir problemas relacionados a conciliação (TONELLO et al., 2017).

Em 2002 José Dirceu Ribeiro e colaboradores fizeram um estudo levantando os pontos controversos no tratamento da FC, tendo como tratamento padrão para a doença a utilização de medicamentos ou procedimentos, como o uso de antibióticos, prevenção de infecção cruzada, anti-inflamatórios, broncodilatadores, mucolíticos, fisioterapia, tratamento cirúrgico, suporte nutricional, enfoque psicológico e social e terapia genética, higiene das vias aéreas e exercícios, suplementação de oxigênio. Seus resultados demonstram que apesar da doença não ter cura, o conhecimento que sem tem adquirido sobre sua etiologia e fisiopatologia da FC proporcionam uma nova abordagem no tratamento, que tem se mostrado efetiva aumentando a sobrevida dos pacientes (DALCIN; SILVA, 2008)(RIBEIRO; RIBEIRO; RIBEIRO, 2002).

Tem-se como o mais indicado o tratamento pré-sintomático para pacientes com FC, esse tipo de abordagem visa adiar as infecções pulmonares e controlar as deficiências enzimáticas (Ribeiro ROSA et al., 2008). Com o aumento da sobrevida dos pacientes com FC e do aprimoramento no aprendizado da patologia da doença viu-se que quanto mais cedo se começa o tratamento para doença maiores são as chances de preveni-la, já que os danos graves causados pelas doenças pulmonares começam na juventude, na maioria da vezes antes dos sintomas óbvios aparecerem (PROESMANS, 2017).

Geralmente no tratamento do paciente é definido um cuidador para ele que tem a responsabilidade principal sobre o cuidado do paciente, essa pessoa é a que o ajudara no dia-dia com os medicamentos, fisioterapia, atividades físicas e no contato com a equipe multidisciplinar, cuidador deve ser o maior aliado da equipe para que o tratamento, sendo

assim é de suma importância o conhecimento dos profissionais sobre os aspectos sociais, econômicos e emocionais da família. Em seu estudo Stella Alves e colaboradores analisaram o perfil dos cuidadores, onde foi visto que na maioria dos pacientes a responsabilidade caia sobre a mãe e que a mais da metade não trabalhava fora de casa, o que mostra a dependência e sobrecarga causada sobre o cuidador (ALVES; BUENO, 2018).

De acordo com a Fundação de Fibrose Cística podem ser prescritos muitos medicamentos no tratamento da doença, sendo medicamentos para limpeza dos pulmões, prevenção e combate contra infecções, para alguns pacientes medicamentos para ajudar a corrigir a causa a subadjacente da doença. Os portadores da doença tem uma maior propensão ao desenvolvimento de infecções bacterianas devido ao acumulo de muco, os antibióticos são utilizados com uso diário regular para evitar estas infecções (FOUNDATION, 2022a). São utilizados também broncodilatadores que podem auxiliar no aumento das via aéreas assim relaxando os músculos abrangentes, assim permitindo que entre mais ar pelas vias aéreas o que favorece na eficácia de alguns medicamentos, alguns pacientes também podem tomar este tipo de medicamento para auxiliar na realização de exercícios físicos, um outro medicamento utilizado são os diluentes de muco que geralmente são utilizados juntamente dos broncodilatadores, pois devido ao aumento das vias aéreas a movimentação do muco é mais fácil já que ele está mais fino e menos pegajoso (FOUNDATION, 2022b)(FOUNDATION, 2022a).

Existe também terapias para corrigir o mal funcionamento do gene CFTR, entretanto como existem diversas mutações desse os medicamentos desenvolvidos atendem a casos de mutações especificas, atualmente existem quatro medicamentos, Kalydeco® (ivacaftor), Orkambi ® (lumacaftor/ivacaftor), Symdeko® (tezacaftor /ivacaftor) e Trikafta® (elexacaftor /tezacaftor/ivacaftor), mais medicamentos vem sendo desenvolvidos para abordar outras causas subadjacentes (FOUNDATION, 2022c). Pacientes que exigem terapia intravenosa podem vir a utilizar dispositivos para acesso vascular, esses dispositivos permitem acesso repetido e a longo prazo na corrente sanguínea (FOUNDATION, 2022e).

Um ponto a se ressaltar é que embora seja comprovada a eficácia do tratamento para FC, muitas das vezes é difícil a aderência dos pacientes ao tratamento já que esse causa uma sobrecarga ao paciente, afetando sua qualidade de vida devido à complexidade dos regimes terapêuticos. Assim se faz necessário que o time multidisciplinar faça um plano estratégico para que o paciente possa vencer as barreiras e intervenções psicossociais. A adesão ao tratamento já que às ações inerentes a doença está relaciona a benefícios clínicos expressivos (ATHANAZIO et al., 2017). Além da dificuldade na aderência do programa existe também fatores que atrapalham no tratamento da doença que estão apresentados na Tabela 3. Então isso faz com que o diagnóstico precoce seja ainda mais importante já que ele favorece em um melhor desfecho clinico (BELL et al., 2019)(TRAVERT; HEELEY; HEELEY, 2020).

Tabela 3 – Problemas comuns que complicam a fibrose cística e seu tratamento

Condições de saúde mental:	Depressão
	Ansiedade
Complicações do acesso vascular:	Risco de trombose com dispositivos de acesso vascular
Complicações medicamentosas:	Reações de hipersensibilidade a antibióticos e intolerância
	Distúrbio vestíbulo auditivo incluindo zumbido
	Doença renal crônica
Complicações metabólicas:	Sobrepeso e obesidade (especialmente em pacientes com função pancreática exócrina residual)
Complicações pós-transplante (relevantes para fibrose cística):	Doença renal crônica e insuficiência renal (em pessoas com ou sem diabetes relacionado à fibrose cística pré-transplante)
	Organismos multirresistentes que contribuem para complicações das vias aéreas
	Câncer em sobreviventes de longo prazo (incluindo câncer gastrointestinal, de pele e urogenital)

Fonte: (BELL e colab., 2019)

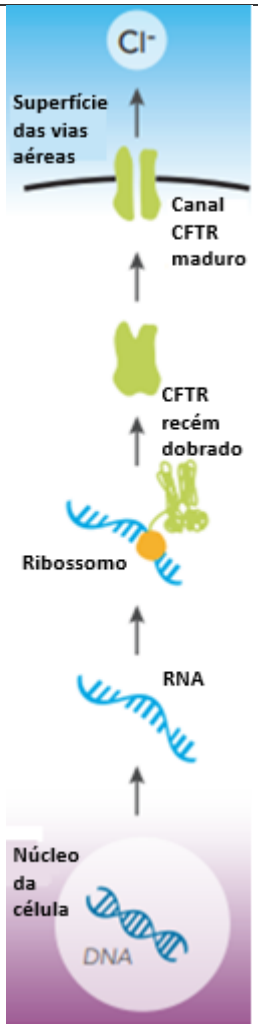
4 MUTAÇÕES DA FIBROSE CÍSTICA

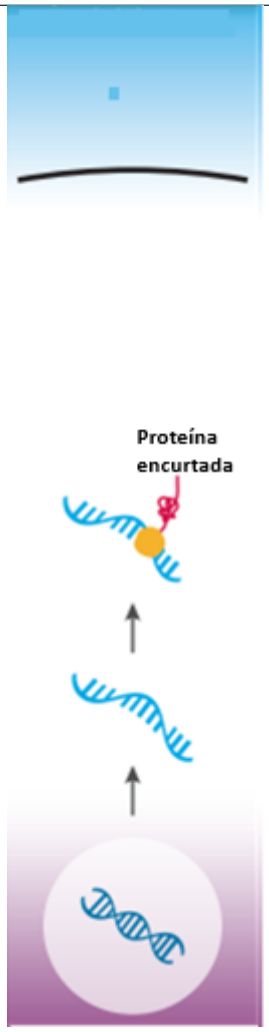
Em 1989 com a descoberta do gene CFTR, foi formado um consorcio de pesquisadores genéticos com o intuito de catalogar e identificar o grande e crescente número de mutações nesse gene. O intuito do consorcio é facilitar e aumentar a comunicação entre pesquisadores de FC. De acordo com a base de dados para mutações de fibrose cística fornecida pelo consorcio hoje são listados 2109 tipos de mutações no gene CFTR (CFMD, 2022). Sabe-se que a incidência de diversas mutações da FC se dá por conta da composição étnica da população (CHILD, 2001).

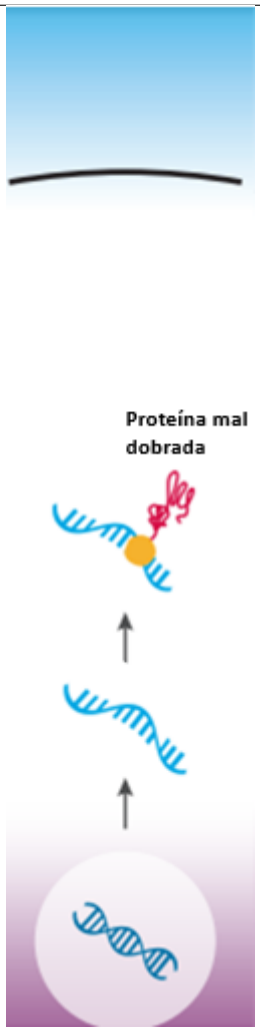
Após a identificação do gene causador da FC em 1989 por Tsui, Collins e colaboradores, primeiramente observaram que a maioria dos pacientes possuíam uma alteração em 3 nucleotídeos que resultavam na deleção in-frame de um resíduo de fenilalanina na posição 508 da proteína (DF508). A proteína CFTR possui dois domínios transmembranares (MSDs – Membrane Spanning Domains), cada um contendo 6 segmentos hidrofóbicos, que são provavelmente responsáveis por formar os poros de passagem de íons, a proteína também possui dois domínios de ligação de nucleotídeos, NFB1 e NFB2 (Nucleotide-binding fold), que também participam da do transporte de íons. Tendo também um domínio regulador, R, que parece funcionar como uma porta que regula a abertura do poro. Esses domínios são importantes para o funcionamento e para estrutura da proteína (CABELLO, 2011).

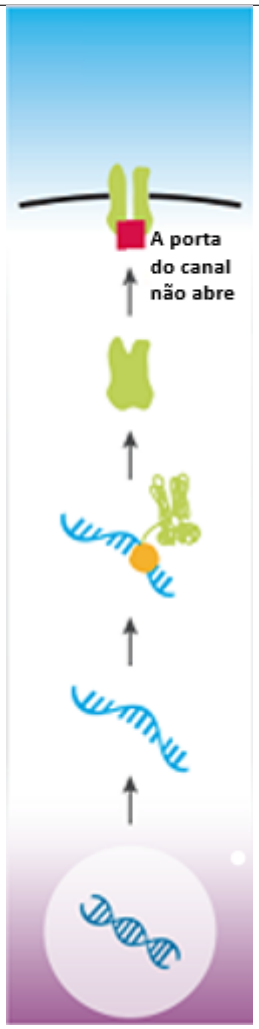
Os cientistas já utilizaram diversas maneiras de agrupar as mutações de FC em diferentes classes, atualmente se utilizam cinco classes de mutações do CFTR que as agrupam de acordo com os efeitos causadores de problemas na produção de proteína CFTR, sendo elas produção de proteínas, processamento de proteínas, gating, condução e proteína insuficiente (FOUNDATION, 2022a). Na Tabela 4 tem-se um descritivo de cada classe de mutação mostrando como ela age nos pacientes e como elas são caracterizadas.

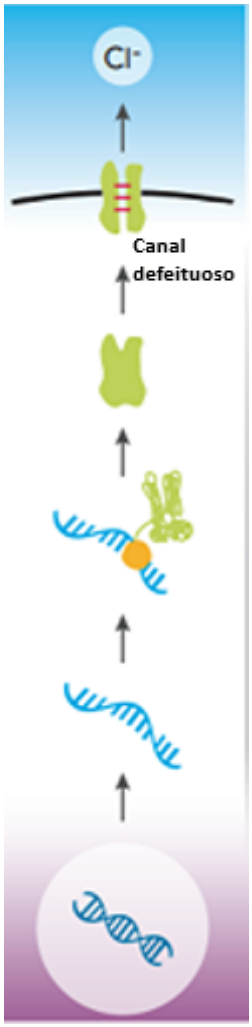
Tabela 4 – Classes de mutação do CFTR

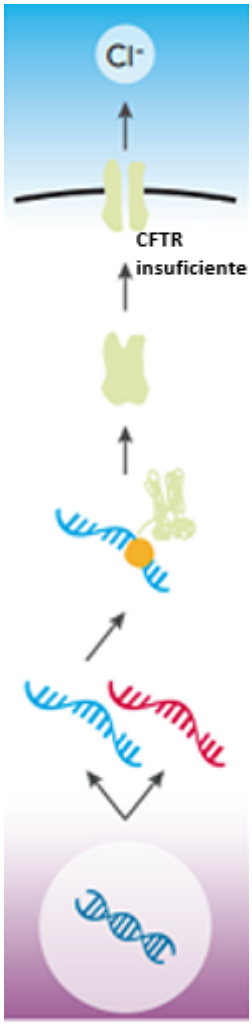
Descrição	% de pessoas com FC que tem pelo menos uma mutação nessa classe	Exemplos de mutação	O que está acontecendo na célula	Terapias possíveis
A proteína CFTR é criada, move-se para superfície da célula e permite a transferência de cloreto e água	Sem mutação	Sem mutação		Sem mutação

Nenhum CFTR funcional é criado	22%	G542X W1282X R553X Também conhecida como “mutações de produção”		Compostos de leitura podem permitir a produção de CFTR completo para mutações sem sentido

A proteína CFTR é criada, mas dobra mal impedindo-o se se mover para superfície da célula	88%	F508del N1303K I507del Também conhecida como “mutações de processamento”		Corretores como Lumacaftor ou Tezacaftor ajudam o CFTR defeituoso a dobrar corretamente

A proteína CFTR é criada, move-se para superfície da célula, mas a porta do canal não abre corretamente	6%	G551D S549N Também conhecida como “mutação de portão”		Potenciadores como o Ivacaftor ajudam a abrir o canal CFTR e também ajudam a aumentar a função do CFTR normal

A proteína CFTR é criada, move-se para superfície da célula, mas a função do canal está com defeito	6%	D1152H R347P R117H Também conhecida como “mutação de condutância”	

A proteína CFTR é criada, move-se para superfície da célula, mas em quantidade insuficiente	5%	3849+10kbC →T 2789+5G →A A455E Incluindo algumas mutações de emenda		

A proteína CFTR é responsável pela regulação do fluxo de íons de cloreto e agua dentro e fora da célula, e em geral nas mutações onde não há produção desta proteína estão concentrados os casos mais graves da doença (INC., 2021). No entanto essa correlação não é perfeita e o conhecimento sobre a mutação do paciente nem sempre pode dizer sobre a gravidade da doença (FOUNDATION, 2017).

5 FALSOS POSITIVOS (FP) E FALSOS NEGATIVOS (FN)

Em 1989 Michael Rock e colaboradores viu que a sensibilidade, a especificidade e o declínio no tempo dos valores de IRT ainda eram fatores desconhecidos e que precisavam ser avaliados de forma abrangente. Então foi realizada um analise sobre os pacientes falsos positivos de uma triagem para fibrose cística de 87000 recém nascidos. Desses 92 tiveram um resultado elevado para o IRT e desses pacientes apenas 13 tiveram o diagnóstico positivo para FC através do teste do suor, sendo assim os outros pacientes foram classificados como falsos positivos. Propondo avaliar a hipótese de influência de estresse perinatal sobre os valores elevados de IRT, Michael utilizou o índice de Apgar, e o índice de RN falsos positivos eram significavelmente menores (P=0,0004 e P = 0,0102 em 1 e 5 minutos, respectivamente), em comparação com os pacientes da população geral. E mesmo que a asfixia perinatal seja um fator associado responsável pelos altos valores de IRT, a maioria dos RN sem FC com IRT elevado possuem um índice de Apgar normal (ROCK et al., 1989).

Na França em 2002 Cheillan e colaboradores fizeram um estudo com 35.141 recém nascidos e percebeu-se que na região de Rhône-Alpes, se tinha uma incidência de 0.65% contra 0.50% de incidência de triagens positivos, porém não se teve um aumento na incidência de FC. Após a análise da população e dos resultados chegou-se à conclusão que essa diferença se dá devido a etnia da população da região, foi constato que recém nascidos de etnia norte africana apesar de terem casos positivos na quantificação do IRT a incidência de FC é bem menor e tendo que na maioria dos casos positivos através do teste de tripsinogênio imunorreativo são “falsos positivos” (CHEILLAN et al., 2005).

A concentração de IRT varia de acordo com a idade do recém-nascido na coleta do sangue em papel filtro, onde ela cai acentuadamente a partir da terceira semana de vida. Com isso diferentes estratégias e protocolos de triagem neonatal foram estabelecidos, além deste fator tem-se a baixa especificidade desse exame de triagem que faz com que se aumente os resultados falsos positivos. (ARRUDI-MORENO et al., 2021).

De acordo com o trabalho de Brockow e colaboradores acredita-se que o aumento de IRT em pacientes com FC se dá devido a um vazamento na parte anterior do ducto pancreático exócrino. Entre os fatores que se associam aos resultados falsos positivos estão a etnia, a condição do portador e a saúde perinatal, já em relação aos falsos negativos se tem a idade do lactante e à presença se íleo meconial. Um estudo americano mostrou que o índice de falsos positivos é cerca de três vezes maior em afro-americanos do que em brancos, e até onde se sabe nenhuma outra etnia apresentou um efeito semelhante (BROCKOW; NENNSTIEL, 2018).

Como o valor preditivo somente de um exame de IRT era baixo, um protocolo de dois estágios (IRT+IRT) foi amplamente adotado para melhorar o poder preditivo positivo. Além disso os centros laboratoriais podem utilizar um valor de corte menor no primeiro estágio para evitar falsos negativos extras. Tendo ainda outras metodologias como IRT + DNA, IRT + DNA + IRT, IRT + meconomium + IRT, essas metodologias são adotadas sempre visando aumentar o poder preditivo do exame e evitar ansiedade e estresse nos neonatos (BROCKOW; NENNSTIEL, 2018).

Sabe-se que em geral pode se ter resultados falsos negativos na triagem através de IRT devido a erros de laboratório (eluição da amostra de sangue em papel filtro não eficaz), mudança em procedimentos (uso de cut-off muito alto), ou fatores biológicos (íleomeconial ou outra condição envolvendo obstrução intestinal que parece favorecer a alta incidência de IRT falso negativo). O maior problema dos resultados falsos negativos é que eles não serão detectados e escaparão da triagem molecular, por isso deve-se ter cautela na condução do procedimento de triagem (CABELLO et al., 2003).

Ao longo de 26 anos (1992 – 2018) Tacceti e colaboradores na Toscana, Itália, avaliou os diagnósticos de FC, identificando os pacientes que tiveram resultados falsos negativos. Constatando que a introdução do protocolo com analise de DNA melhorou a sensibilidade do teste e assim reduziu o número de casos FN. Em pelo menos 8.7% dos casos de FC avaliados tiveram-se resultados FN, tendo o diagnóstico tardiamente com media em 6.6 anos devido aos sintomas característicos da doença como problemas respiratórios e síndromes de perda de sal (TACCETTI et al., 2020).

É de suma importância testes para detectar a fibrose cística em casos que crianças apresentem doenças pulmonares crônicas, Lumertz e colaboradores identificaram 4 pacientes com resultados falsos negativos na triagem para FC através da concentração de IRT, os quatro pacientes eram caucasianos e do sexo masculino e dois desses pacientes apresentaram íleomeconial ao nascimento, esses pacientes tiveram concentração de IRT no exame de triagem abaixo do valor estabelecido pelo laboratório, 110ng/ml e tiveram o diagnóstico positivo para FC depois do exame do suor. O terceiro paciente teve o diagnóstico positivo devido a repetitivas infecções respiratórias já no primeiro ano de vida e baixo ganho de estatura tendo também sido detectado através do teste do suor devido a idade avançada onde o teste de quantificação do IRT não é mais eficiente. Já o quarto e último paciente apresentava tosse, esteatorreia, baixo crescimento e sibilância recorrentes, também tendo o resultado positivo através do teste do suor. Isso mostra que os pediatras devem se preocupar com sintomas como estes mesmo se o exame de triagem primário se mostrar negativo (LUMERTZ et al., 2019).

6 EFEITOS DE RESULTADOS FALSO POSITIVOS

Uma pesquisa feita com pais de recém nascidos com fibrose cística em Leeds entre 1998 a 2002, levantou que as mães descreveram diversas emoções no decorrer do processo de um segundo exame de quantificação de IRT, sendo elas angustia, ansiedade e aborrecimento, sendo o tempo de espera do segundo resultado emocionalmente mais difícil, com os pais dizendo que a diminuição desse atraso ou informa-los com um diagnóstico definitivo melhoraria o processo. Um outro ponto levantado por esta pesquisa é que a forma em que os profissionais abordam os pais influenciavam na contingência da ansiedade, sendo que eles se sentiam mais confortáveis com profissionais especialistas em FC (MORAN et al., 2007).

Um estudo similar foi realizado na China após a expansão do programa de triagem neonatal, onde foi avaliado o impacto do estresse causado nos pais e a percepção da saúde das crianças em relação aos familiares dos pacientes com resultados falsos positivos. Tendo que sim o resultado errôneo pode causar o estresse nos pais e causar uma sensação de preocupações futuras com os filhos. Esses sintomas podem ser reduzidos também com uma melhor e educação e comunicação com os pais sobre os resultados falsos positivos (TU et al., 2012).

Parte II

APRENDIZADO DE MÁQUINA

7 INTRODUÇÃO AO APRENDIZADO DE MÁQUINA

Em 1959, Samuel (1959) fez um estudo aplicando técnicas de aprendizado de máquina para ensinar um computador a jogar "Dama", onde definiu aprendizado de máquina (AM) como campo do estudo que dá aos computadores a habilidade de apreender sem ser explicitamente programado. Como conclusão, Samuel (1959) verificou que seria possível conceber esquemas de aprendizado, superarão a média das pessoas e futuramente poderão ser viáveis economicamente podendo ser aplicados na vida real.

Desde que os computadores foram inventados, a pergunta feita é se eles podem aprender de acordo com suas experiências. Com o passar dos anos e com a compreensão dos computadores amadurecendo, parece que é inevitável que o aprendizado de máquina desempenhe cada vez mais um papel central perante ciência da computação e na tecnologia computacional. Então, Mitchell (1997) definiu o aprendizado de máquina onde o computador melhore seu desempenho em uma tarefa por meio de uma experiência, precisamente sendo:

"Diz-se que um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho em tarefas em T, medido por P, melhora com a experiência E." - Tom Mitchell, 1997

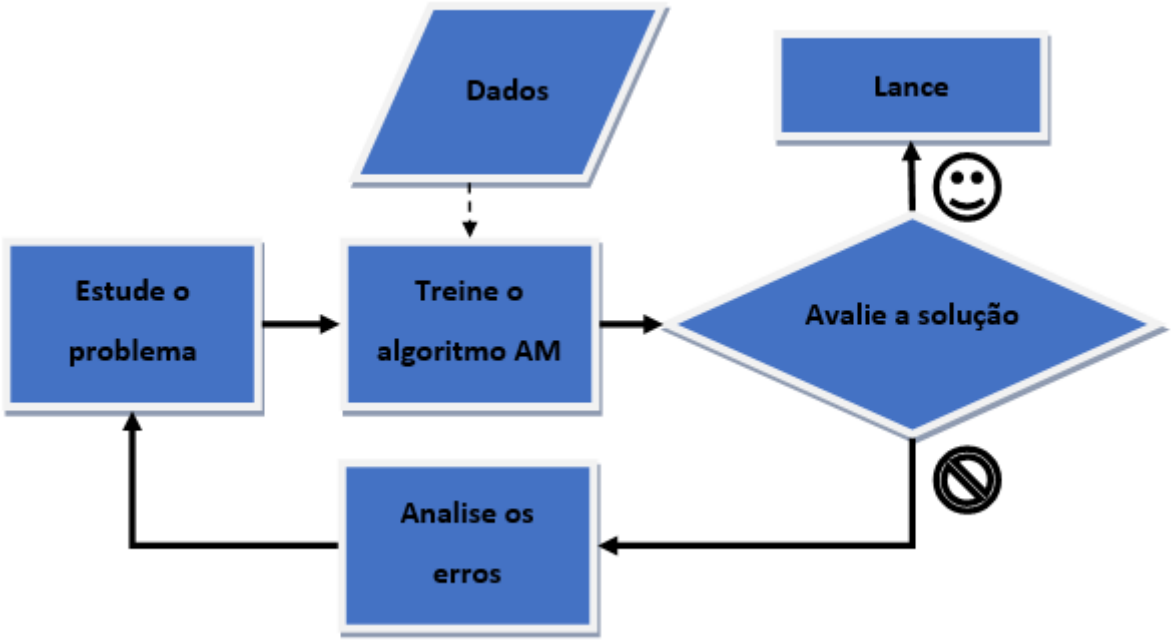
O campo de AM cresceu a partir de estatísticas tradicionais e comunidades de inteligências artificiais. Esforços de megacorporações como Google, Microsoft, Facebook, Amazon, etc. se tornando um dos tópicos de ciência computacional mais quentes da última década. Isso proporcionou uma oportunidade de revigorar as abordagens estatísticas e computacionais para gerar automaticamente modelos uteis a partir de um conjunto de dados (W. Edgar; O. Manz, 2017).

O aprendizado de máquina trata-se de extrair conhecimento a partir de um banco de dados. É um campo da pesquisa resultante da estatística, inteligência artificial e ciência da computação, sendo reconhecido pelo análise preditiva ou pelo aprendizado estatístico. A aplicação dessa tecnologia está presente cada vez mais no cotidiano desde recomendações de filmes, a procura de novos planetas ou até mesmo na análise de sequências de DNA. Antigamente os aplicativos inteligentes utilizavam operações lógicas, "se" e "se não", porém dependendo da quantidade de dados ou casos em que as hipóteses não compreendidas com clareza pelo especialista humano se tornam inviável a utilização dessa metodologia, no entanto com um bom conjunto de dados o modelo de aprendizado de máquina é capaz de determinar padrões entre eles e classificá-los ou fazer previsões (GUIDO; MULLER, 2016).

Um exemplo clássico que se tem para o entendimento de aprendizado são os modelos de filtro de SPAM (Sending and Posting Advertisement in Mass, traduzindo em português

Enviar e Postar Publicidade em Mass) onde na metodologia tradicional usa-se regras condicionais para definir os possíveis e-mails que sejam spams, como por exemplo palavras ou frases frequentes nesses e-mails. Já uma abordagem que utiliza o aprendizado de máquina é passada um conjunto de e-mails onde diz-se quais são spam e quais não são, o filtro apreende automaticamente as frases ou características dos spams de acordo com o banco de dados, na Figura 4 tem-se um fluxograma básico desse processo de classificação. Dessa forma a manutenção do algoritmo se torna mais simples já que o programa fica menor e provavelmente mais preciso (GÉRON, 2019).

Figura 4 – Fluxograma básico de um algoritmo de classificação AM



Fonte: Adaptado de Géron, A. (2019)

Os modelos de AM são divididos entre três subgrupos principais, aprendizado supervisionado (o que será utilizado neste trabalho), não supervisionado e aprendizado por reforço. Objetivamente explicando esses três subgrupos temos o aprendizado supervisionado que se refere a um conjunto de algoritmos que a partir de uma entrada visa prever dados rotulados (dados que possuem classificação específica) a partir de suas características. O aprendizado não supervisionado ao contrário do supervisionado visam extrair informações de dados não rotulados (nenhuma classificação específica é conhecida ou necessária). E por tem-se o aprendizado por reforço onde o foco é ensinar um agente a interagir com um ambiente com base em uma observação de sua condição atual (BONETTO; LATZKO, 2021).

7.1 Aprendizado supervisionado

No aprendizado supervisionado, os dados fornecidos ao algoritmo incluem as soluções desejadas, os rótulos. Podendo ser utilizado em casos de classificação ou previsão de um valor numérico. Os principais modelos de aprendizado supervisionado que estão listados e com suas características principais resumidas na Tabela 5. Nesse trabalho será utilizada a metodologia de Florestas aleatória e árvore de decisão (GÉRON, 2019).

Tabela 5 – Características básicas dos algoritmos supervisionados

Modelo	Principais características
K-vizinhos	Esse modelo consiste apenas em armazenar o conjunto de treinamento. Para a previsão de um novo dado, ele encontra o ponto mais próximo no conjunto de treinamento e atribui esse rótulo para classificá-lo (GUIDO e MULLER, 2016).
Regressão Linear	É uma adequação da fórmula, $y = ax + b$, onde o intercepto b da um valor de base para a previsão e o coeficiente a, o ajuste para cada entrada. Esse modelo pressupõe que a previsão é uma combinação linear dos dados (HARRISON, 2020).
Regressão Logística	Esse modelo estatístico utiliza a função sigmóide, $\Omega(z) = 1/1 + e - z$, (função logística) para estimar uma probabilidade, essa técnica pode ser utilizada tanto para problemas de classificação quanto regressão, porém é mais comum em classificação (SARKER, 2021).
Máquinas de vetores de suporte (SVM)	É um modelo versátil e poderoso capaz de realizar classificações lineares e não lineares, de regressão e até mesmo *outliers*, pode ser considerado como o preenchimento da via mais larga possível entre duas classes distintas (GÉRON, 2019).
Árvores de decisão e florestas aleatórias	Como o modelo SVM (Support Vector Machine, traduzindo em português Maquina de Vetores de Suporte) estes também são muito versáteis, podendo ser aplicado em diversas situações. São algoritmos muito poderosos muito poderosos capazes de moldar os conjuntos complexo de dados. Esses dois modelos serão melhor explicados no próximo capítulo por serem os utilizados nesse trabalho (GÉRON, 2019).
Redes Neurais (Deep Learning)	São modelos lineares que realizam vários estágios em camadas ou multicamadas de processamento para chegar em uma decisão. Este tipo de algoritmo são adaptados com muito cuidado para cada caso especifico (GUIDO e MULLER, 2016).

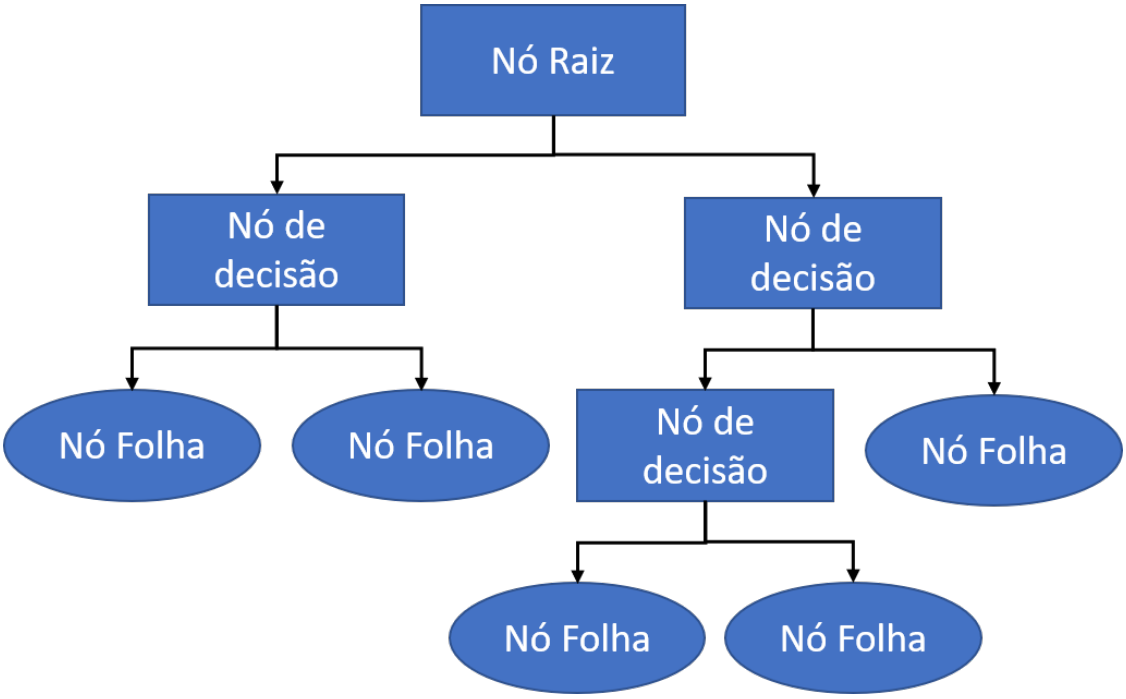
7.2 Florestas Aleatórias

Modelos de florestas aleatórias (FA) são um conjunto de árvores de decisão (AD). Esses dois métodos estão entre os modelos mais poderosos disponíveis atualmente. Elas funcionam com o treinamento de muitas árvores de decisão em subconjuntos aleatórios das características e em seguida é calculada a média das suas previsões. Construir um modelo a partir de um conjunto de outros modelos é chamado de "Ensemble Learning" e muitas vezes é uma ótima maneira de aumentar o poder do algoritmo (GÉRON, 2019).

A ideia das florestas aleatórias é criar uma "floresta" de árvores de decisão treinadas em diferentes colunas de dados de treinamento. Se cada árvore tiver uma chance melhor de 50% de fazer uma classificação correta, você deve incorporar essa predição. Trata-se ótima ferramenta tanto para classificação como regressão (HARRISON, 2020).

Ao ordenar uma árvore da raiz para alguns nós folhas, como mostrado na Figura 5, a AD classifica as instâncias. Estas são classificadas verificando o atributo definido por aquele nó, começando pelo nó raiz e depois descendo para os ramos correspondentes ao valor do atributo. Para fazer a divisão entre os nós o critério com maior utilização é o "gini" para impureza de Gini e de "entropia" para o ganho de informação (SARKER, 2021).

Figura 5 – Estrutura de uma árvore de decisão



Fonte: Adaptado de Sarker, Iqbal H (2021)

A regressão ou a classificação nesse tipo de algoritmo são baseadas em sucessivas divisões binárias de subconjuntos do conjunto de treinamento até que a divisão final gere a classificação desejada com a menor impureza possível. Para se entender, a estrutura de uma árvore de decisão diz-se que uma árvore é definida por um nó que não tem um pai,

então, ele é a raiz da árvore (que só tem uma raiz). Cada nó só tem um pai, cada nó que tem filho é uma decisão binária (galhos) e se um nó não tem filho, ele é uma folha. Cada nó de decisão representa uma decisão binária em relação a um único recurso tendo dois filhos identificados como filho esquerdo ou direito e vão sendo divididos em novas decisões até que se tornem folhas que são as classe nas quais o conjunto de dados será classificado (BONETTO; LATZKO, 2021).

O critério de Gini utilizado como critério nas divisões, 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 gini, dado pela Equação 7.1, igual a zero. Quando nas árvores de decisão se utiliza o critério de Gini tende-se isolar num ramo os registros da classe mais frequente (BARBOSA et al., 2012). Esta é a fórmula mais comum de quantificar o desvio de uma distribuição uniforme. O índice varia entre 0 e 1, sendo 0 quando todos os membros da sociedade investigada são iguais na quantidade relevante e 1 se um membro estiver monopolizando a totalidade dos recursos disponíveis (BIRÓ; NÉDA, 2020).

$$IndiceGINI = 1 - \sum_{i=1}^{c} p_i^2 \tag{7.1}$$

onde P_i = frequência relativa de cada classe em cada nó e c = número de classes.

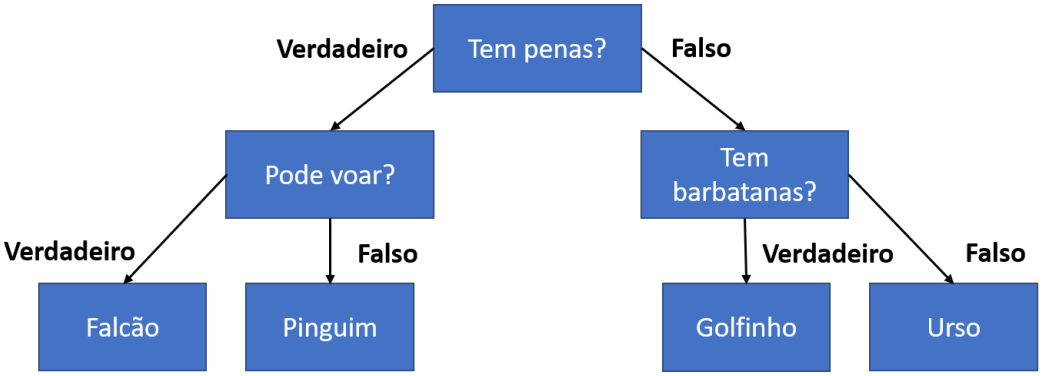
O ganho de informação mede a redução da entropia após a transformação dos dados. Isso é calculado comparando a entropia do conjunto de dados antes e depois da transformação dos dados. A entropia é a medida de homogeneidade da amostra. A entropia é dada pela Equação 7.2 (SAUD; SHAKYA; NEUPANE, 2021).

$$Entropia = -\sum_{i=1}^{c} p_i log_2 p_i \tag{7.2}$$

Árvores de decisão são muito rápidas, não precisam de dimensionamento dos dados e podem ser facilmente explicadas, são amplamente utilizados para tarefas de classificação e regressão. Essencialmente esse tipo de modelo aprende uma hierarquia de perguntas “if/else”, levando a uma decisão. As perguntas são muito semelhantes a perguntas feitas por exemplo para distinguir animais (ursos, falcões, pinguins e golfinhos) onde o objetivo é chegar a resposta certa com a menor quantidade de perguntas. Na Figura 6 temos um exemplo desta árvore para fazer essa decisão na prática (GUIDO; MULLER, 2016).

O crescimento de um conjunto de árvores de decisão e a permissão para que elas votem na classe mais popular, resultaram em melhorias significativas na precisão de classificação. Breiman (2001) definiu florestas aleatórias como um classificador que consiste em uma coleção de classificadores estruturados em árvore onde o são vetores aleatórios independentes identicamente distribuídos e cada arvore lança um voto único na classe

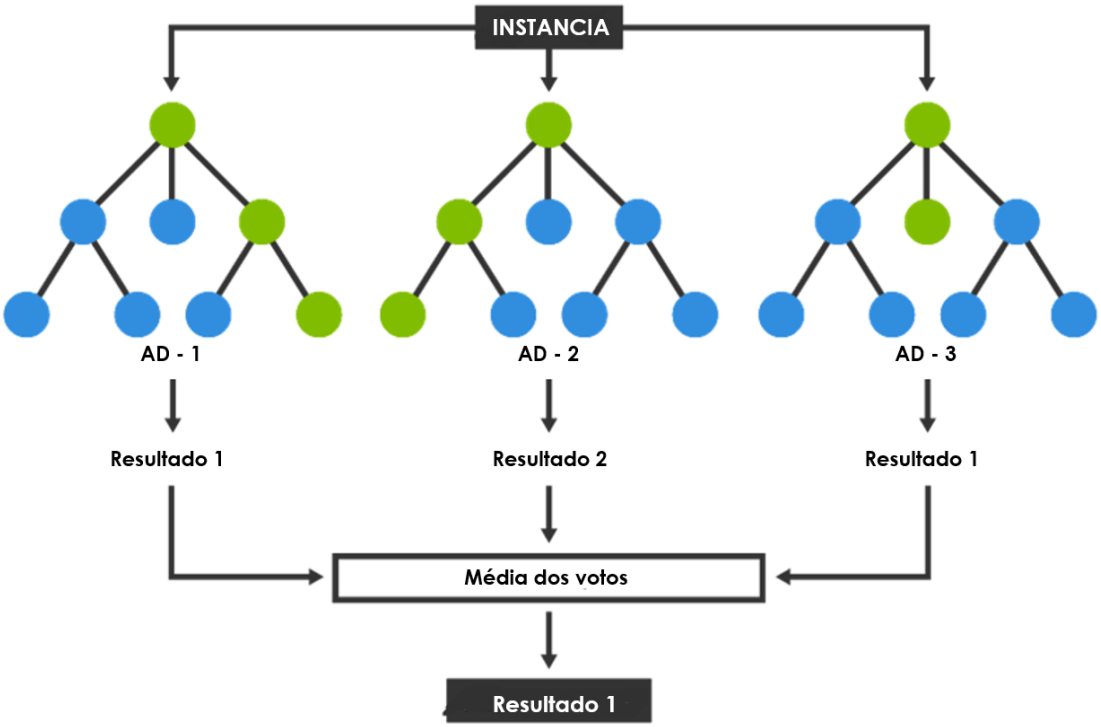
Figura 6 – Exemplo de árvores de decisão na pratica



Fonte: Adaptado de S. Guido et. al (2016)

mais popular na entrada X, como exemplifica a Figura 7. As florestas aleatórias é uma ferramenta eficaz na previsão, por conta da Lei dos Grandes Números, eles não se ajustam excessivamente.

Figura 7 – Exemplo básico de uma floresta aleatória



Fonte: O autor

As florestas aleatórias surgiram como concorrentes sérios para métodos de última geração como *boosting* e SVMs, podendo lidar com um alto número de variáveis de entrada sem *overfitting*, sendo consideradas umas da técnicas mais precisas disponíveis (BIAU, 2012), Para florestas aleatórias as amostras "N" do conjunto de dados de treinamento são divididas aleatoriamente em "n" amostras retiradas a partir dos dados originais, onde "n"

« ”N”. Esses subconjuntos são utilizados para o cultivo de cada árvore da floresta (ALI et al., 2012).

Numa literária, feita por Sarica, Cerasa e Quattrone (2017), foram levantados importantes vantagens de modelos de florestas aleatórias em termos de robustez e *overfitting*, capacidade de lidar com dados altamente não lineares, estabilidade na presença de *outliers*, e oportunidade de processamento paralelo eficiente. Isso por conta da sua metodologias que utiliza um conjunto de coleção de Árvores de Classificação e Regressão (CART) que são treinadas em conjuntos de dados do mesmo tamanho que o conjunto de treinamento, chamado de *bootstraps*. Esses *bootstraps* criados a partir de uma reamostragem aleatória do próprio conjunto de treinamento (SARICA; CERASA; QUATTRONE, 2017).

7.3 RamdomForestClassifier - ScikirLearn

Para realização dos testes, é utilizada a biblioteca para a linguagem de programação Python, Scikit Learn que fornece diversos recursos para aprendizado de máquina. A classe utilizada para realizar a criação dos modelos de florestas aleatórias é a “RamdomForestClassifier”, na qual vários classificadores (árvores de decisão) utilizam de subamostras de um conjunto de dados para poder melhorar a precisão preditiva e o ajuste excessivo aos dados. Na tabela 6 estão descritos os parâmetros de funcionamento dessa classe e uma breve explicação de suas características (AL., 2011).

Tabela 6 – Parâmetros *Ramdom Florest*

Parâmetro	Funcionalidade	Tipo	Valor padrão
n_estimadores	Número de árvores na floresta	int	100
criterion	Função que define a qualidade da divisão	“gini”, “entropia” e “log_loss”	Gini
max_depth	Profundidade máxima de cada árvore	int	None
min_samples _split	Número mínimo de amostras para dividir um nó interno	Int/float	2
min_samples _leaf	Número mínimo de amostra necessárias para estar em um nó folha	Int/float	1
min_weight_ fraction_leaf	A fração ponderada mínima da soma total dos pesos (de todas as amostras de entrada) necessária para estar em um nó folha.	float	0.0

max_features	Número de recursos a serem considerados ao procurar a melhor divisão, podendo ser “sqrt”, “log2” e nenhum	Int/ float	Sqrt
max_leaf_ nodes	Define a quantidade máxima de nós folhas	Int	None
min_impurity _decrease	Um nó será dividido se esta divisão induzir uma diminuição da impureza maior ou igual a este valor	float	0.0
bootstrap	Se as amostras bootstrap são usadas ao construir árvores. Se False, todo o conjunto de dados é usado para construir cada árvore.	Bool	True
oob_score	Se deve usar amostras prontas para estimar a pontuação de generalização. Disponível apenas se bootstrap=True.	Bool	False
n_jobs	O número de trabalhos a serem executados em paralelo.	Int	None
random_state	Controla tanto a aleatoriedade do bootstrap das amostras usadas na construção das árvores (if bootstrap=True) quanto a amostragem dos recursos a serem considerados ao procurar a melhor divisão em cada nó (if)	Int	None
verbose	Controla a verbosidade ao ajustar e prever	Int	0
warm_start	Quando definido como True, reutilize a solução da chamada anterior para ajustar e adicionar mais estimadores ao conjunto, caso contrário, apenas ajuste toda uma nova floresta	Bool	False
class_weight	Pesos associados as classes. Podendo ser “balanced”, “balanced_subsample”	Dict ou lista de dicts	None
ccp_alpha	Parâmetro de complexidade usado para redução de complexidade de custo mínimo	Float não negativo	0.0
Max_samples	Se bootstrap for True, o número de amostras a serem extraídas de X para treinar cada estimador de base.	Int/float	None

Serão abordados alguns dos parâmetros e descritos brevemente quais seus impactos no modelo de uma forma geral. Iniciaremos abordando o $n_estimators$ que tem forte ligação a robustez e variância do modelo, quanto mais estimadores se usam, entretanto, o tempo de treinamento aumenta. O uso de muitas árvores auxilia a não ter superajustes

aos dados, então, a escolha do número de árvores está altamente relacionada a capacidade de processamento do computador utilizado e a melhora de sua capacidade preditiva, assim como ı$max_features$ȷ e ımin_sample_leafȷ. O número máximo de características diz quantos recursos uma árvore individual pode tentar utilizar, geralmente aumentando esse parâmetro tem-se uma melhora no desempenho, pois a cada nó tem-se uma maior opção de características a serem consideradas, isso também pode comprometer o tempo de processamento também se fazendo necessário encontrar um parâmetro de equilíbrio. Já a quantidade mínima de folhas, ou seja, tamanho do nó final de uma árvore, quando muito pequeno, torna o modelo propenso a ruídos se fazendo necessário encontrar o tamanho mínimo de acordo com a base de dados utilizada (James Thorn, 2020)(Tavish Srivastava, 2015).

Outro importante parâmetro que se utiliza é o de ımax_depthȷ, esse parâmetro pode ser definido como o maior caminho entro o nó raiz e o nó folha, sendo então o número de divisões que uma árvore pode fazer. Tendo uma relação onde se for pequeno pode subajustar os dados e se for muito grande sobreajustar. Um método desenvolvido por Milos Simic demonstra os efeitos relacionados a profundidade da árvores de decisão em uma FA onde viram que os efeitos da número de camadas dependem dos dados que são utilizados (SIMIC, 2022)(RAM, 2020).

Outros dois parâmetros muito importantes são o “criterion” que define qual será o critério usado para fazer a divisão do nós, por padrão se tem o critério de Gini, mas também podem ser utilizados o de entropia, gini tem a vantagem de ser mais rápido, pois o cálculo de entropia é uma função logarítmica que tem um custo maior de processamento. Entretanto, como existem somente esses dois critérios, é valido compará-los para ver a diferença dos desempenhos, já que cada um leva a um modelo diferente. O outro parâmetro que exige ajuste para melhorar o desempenho e a velocidade do modelo é o ı$random_state$ȷ, e, a melhor forma de encontrá-lo é utilizando uma pesquisa para achar o melhor valor de acordo com o banco de dados. Isso fará com que os dados sejam separados sempre da mesma forma (James Thorn, 2020).

7.4 Dificuldades em modelos de aprendizado de máquina

A resolução de problemas complexos de AM necessita de design automático, eficiente e correto do sistema. Escolhas eficientes de pré-processamento de dados, seleção da família apropriada de algoritmos, escolha dos hiper-parâmetros, seleção dos atributos e pós-processamento ajudam a evitar problemas com AM. As máquinas, por exemplo, necessitam de um número muito maior de exemplos para aprender uma atividade do que os humanos. E não existem bons simuladores de da vida real para gerarem dados de treinamento. Além de que os conjuntos de dados fornecidos aos modelos podem não ter uma boa representação do mundo real e estarem enviesados (LUDERMIR, 2021).

Na Tabela 7, tem-se os principais desafios na criação de modelos de AM, num geral duas coisas podem dar errado, são elas, “algoritmos ruins” e “dados ruins” (GÉRON, 2019). Na sequência desse capítulo, serão abordados com mais detalhes essas possíveis dificuldades.

Tabela 7 – Dificuldade no aprendizado de maquina

Dados ruins	Quantidade insuficiente de dados para treinamento
	Dados não representativos
	Dados de baixa qualidade
	Características irrelevantes
Algoritmos ruins	Sobreajustando os dados de treinamento
	Subajustando os dados de treinamento

Fonte: O autor

7.4.1 Dados Ruins

Segundo estudo feito pela Microsoft, os cientistas de dados acham que o principal motivo de problemas de qualidade geral de um trabalho de AM está ligado a análise de dados. Isso se dá por conta, por exemplo, da falta de qualidade de dados como valores ausentes, tipos de dados inconsistentes e verificação de suposições. Se faz necessária a análise dos dados descobrir possíveis vieses ou mudanças inesperadas em suas distribuições nestes casos (PALEYES; URMA; LAWRENCE, 2022).

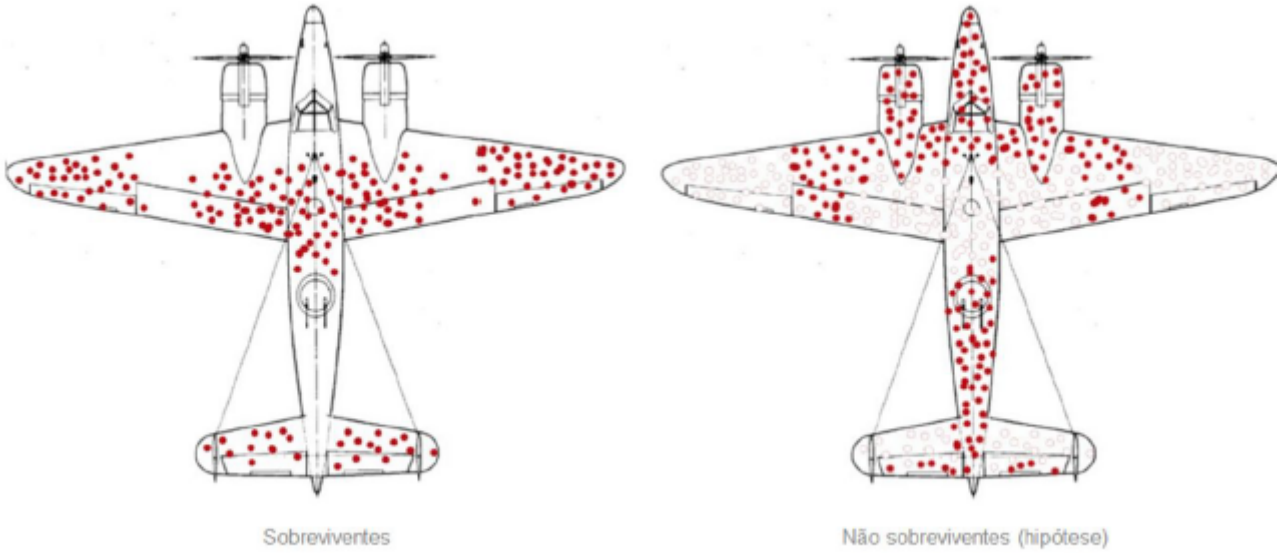
Os modelos de AM geralmente apresentam um desempenho melhor com um grande conjunto de dados, por isso é importante e enfatizar que eles são abordagens orientas por dados e seu desempenho está totalmente atrelado a eles. Portanto, garantir um grande número de dados de treinamento e a identificação de *outliers* (dados atípicos ou aberrantes) é de suma importância (EL et al., 2018).

Outro problema bem comum encontrado é sobre a representatividade dos dados, trazendo problemas de vieses, que podem ser compreendidos como, tendência associada ou determinada por fatores externos. Ou seja, é a capacidade de tomar decisões baseadas em decisões passadas. Por tanto, se treinar um modelo com dados que não representam o campo a ser aplicado como um todo, é muito provável que a saída de seu modelo seja enviesada (AVILA; CANTERO; FRANCISCO, 2021).

Um grande exemplo de viés de seleção ocorreu durante a Segunda Guerra. Abraham Wald, trabalhando em como diminuir a vulnerabilidade das aeronaves, verificou onde as aeronaves que retornavam eram mais atingidas e reforçou as respectiva partes identificadas. Entretanto, isso não surtiu efeito. Quando analisou novamente para ver o porquê não funcionou, chegou à conclusão de que deveria analisar as aeronaves não sobreviventes, e, viu que os pontos que essas aeronaves eram atingidas eram opostos das que voltavam, chegando à conclusão que aquela era de fato a área frágil da aeronave. Na Figura 8, tem o

diagrama dos locais atingidos das aeronaves sobreviventes e não sobreviventes, esse é um exemplo clássico como os dados analisado devem ser representativos (MANGEL; CRUZ; SAMANIEGO, 1984).

Figura 8 – Aeronaves estudadas por Abraham Walds



Fonte: Fonte: <https://miro.medium.com/max/1400/1*A78v2NOlNoLnnbzOqtfP_A.png>

O desempenho do modelo também dependerá muito da escolha e da representação dos atributos. O modelo depende indiretamente das atributos do conjunto de dados. Portanto, selecionar características que de fato sejam proeminentes faz com que o método seja robusto (ZHOU et al., 2017). Existe um ditado que diz: entra lixo, sai lixo. Ou seja, o sistema será capaz de aprender somente se os atributos selecionados forem relevantes suficientes e poucas características irrelevantes (GÉRON, 2019).

7.4.2 Algoritmos Ruins

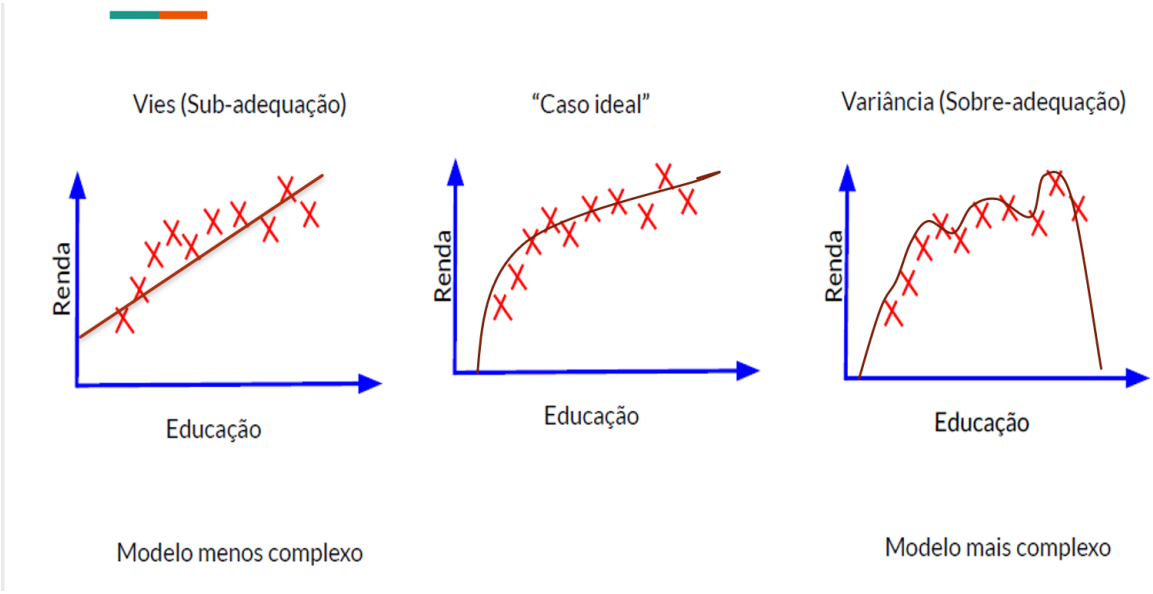
Compreender um modelo é importante para entender o motivo de uma baixa precisão, pode-se dizer que um modelo está subajustando ou sobreajustando os dados de treinamento, analisando o erro das previsões tanto no conjunto de treinamento ou de avaliação. Um dos motivos que pode levar a isso é um modelo pouco complexo ou até mesmo complexo demais, podendo até ter um terceiro motivo ligado a falta de dados para o treinamento (SERVICES, 2022).

O objetivo do aprendizado supervisionado é correlacionar entradas X a saídas Y. E o sobreajuste (do inglês *overfitting*) ocorre por meio de uma memorização dos exemplos fornecidos no conjunto de treinamento com características ruidosas ou irrelevantes, ao invés de apreender a verdadeira correlação entre X e Y. Já o subajuste (do inglês *underfitting*)

ocorre quando um modelo não tem capacidade suficiente para apreender completamente a relação, seja por memorização ou não (BASHIR et al., 2020).

Intuitivamente sempre espera-se que modelos simples extrapolem bem para novos dados. Porém, um modelo muito simples leva ao *underfitting* por não capturar todos os aspectos e variabilidade dos dados. Em contrapartida, um modelo muito complexo pode se ajustar muito bem as particularidades do conjunto de treinamento e, com isso, ele extrapola mal para dados desconhecidos ocasionando o *overfitting*. Então, o que se busca como modelo ideal é um modelo balanceado que não seja muito complexo e nem muito simples, que explique bem os dados de treinamento e, ao mesmo tempo, generalize com precisão dados novos. Na Figura 9 tem-se uma exemplificação de um modelos sobreajustado, outro subajustado e um balanceado, utilizando a renda em relação ao nível de educação (GUIDO; MULLER, 2016).

Figura 9 – Ajuste de modelos sobre renda versus educação



Fonte: O autor

Existem algumas formas de aumentar a flexibilidade do modelo, em caso de subajuste, pode-se adicionar novos recursos específicos de domínio, mais produtos cartesianos de recursos, alterar o tipo de processamento ou diminuir a quantidade de regularizadores. Já em caso de sobreajuste, deve-se diminuir a flexibilidade do modelo, utilizando a seleção de recursos, diminuição do tamanho de processamento, diminuir o número de atributos e aumentar a quantidade de regularizadores (SERVICES, 2022).

7.5 Métricas de avaliação de algoritmos

Existem muitas medições de desempenho para modelos de AM entre as principais estão a acurácia, matriz de confusão (nome dado devido a facilidade de ver se o sistema esta confundindo duas classes), precisão, *recall*, f1-score, a curva ROC, validação cruzada,

o erro quadrado médio, o coeficiente de correlação, entre outras. Nesta seção, algumas dessas métricas serão abordadas explicando como funcionam e o que elas representam (GÉRON, 2019).

A métrica de avaliação de desempenho de um classificador ideal tem um papel crítico durante o treinamento do classificador. Por isso, a seleção de uma métrica adequada é importante para se obter um ótimo classificador. Para problemas de classificação binária, a melhor solução para avaliação do modelo é a utilização de uma matriz de confusão como mostrado na Tabela 8. Na matriz de confusão se tem quatro classes de onde se tira quatro resultados, o de verdadeiros positivos (VP) que ocorre quando o classificador faz uma previsão positiva e ela é positiva, verdadeiro negativo (VN) quando o classificador prevê que é negativo e é negativo, a de falso positivo (FP) quando o modelo prevê como positivo e a amostra é negativa e, por fim, se tem a falsa negativa (FN) onde o modelo prevê como negativo e ela é positiva (M; M.N, 2015).

Tabela 8 – Matriz de confusão para um classificador binário

	Real classe positiva	**Real classe negativa**
Classe de previsão positiva	Verdadeiro positivo (VP)	Falso Positivo (FP)
Classe de previsão negativa	Falso Negativo (FN)	Verdadeiro negativo (VN)

Fonte: O autor

A primeira métrica ser abordada será a de acurácia que mede a proporção de previsões corretas sobre o número total de instâncias avaliadas que é expressa pela Equação 7.3. A acurácia é uma boa métrica, porém, em casos, onde a hipótese de previsão é muito rara ela pode ser muito alta mesmo que erre muitas previsões devido a raridade da hipótese ocorrer. Então se faz necessário utilizar essa métrica aliada a outras métricas e fazer uma análise do custo de resultados falsos positivos e negativos, assim ajudando a determinar um modelo razoável (HARRISON, 2020).

$$acc = \frac{VP + VN}{VP + VN + FP + FN} \tag{7.3}$$

A precisão é uma outra métrica importante que mostra a quantidade de valores previstos positivamente em relação ao total de todas as classes positivas, sendo expresso pela Equação 7.4. A precisão também nos dá a medida dos pontos de dados relevantes, por exemplo, no caso dessa dissertação, não se quer tratar um paciente que não tem a doença (HUILGOL, 2020).

$$p = \frac{VP}{VP + FP} \tag{7.4}$$

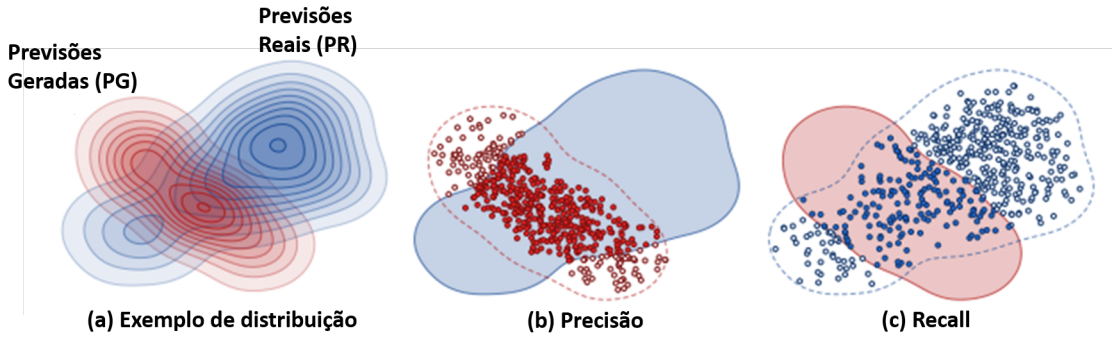
Para se saber a proporção de casos previstos positivamente entre todos que realmente são positivos, utiliza-se a métrica *recall* descrita pela Equação 7.5. O *recall* seria a

sensibilidade do modelo, e em um contexto médico, um bom *recall* é de extrema importância, pois o objetivo é identificar todos os casos positivos (POWERS, 2011).

$$r = \frac{VP}{VP + FN} \tag{7.5}$$

Da Figura 10, consegue-se fazer uma análise para entender melhor as duas métricas anteriores, precisão e *recall*. Na Figura 10(a), temos a distribuição das previsões reais (PR) em azul e a distribuição das previsões geradas pelo classificador (PG) em vermelho. Analisando Figura 10(b), consegue-se ver que a precisão é a probabilidade de que a previsão aleatória gerada caia dentro da previsão real. E, na Figura 10(c), temos a probabilidade de que a previsão real caia dentro da previsão gerada (KYNKÄÄNNIEMI et al., 2019).

Figura 10 – Definição de precisão e *recall* para distribuições



Fonte: Adaptado de T. Kynkäänniemi e colaboradores (2019)

A medida f1-score, descrita na Equação 7.6, é uma média harmônica entre a precisão e a sensibilidade do modelo, ou seja, ela penaliza valores extremos por ambas métricas. Ela não é simétrica entre as classes, ou seja, depende de qual é negativa e positiva. Essa medida varia entre [0,1] assumindo 1 quando tem máxima precisão e sensibilidade e zero em caso oposto (HICKS et al., 2021). Ou seja, um modelo terá um alto f1-score se a precisão e o *recall* forem altos, médio em caso de um deles ser alto e o outro baixo e no pior dos casos terá uma média baixo se as outras duas métricas forem baixas (JASKARI et al., 2020).

$$f1 = 2 \times \frac{precisao \times recall}{precisao + recall} = \frac{2 \times VP}{2 \times VP + FP + FN} \tag{7.6}$$

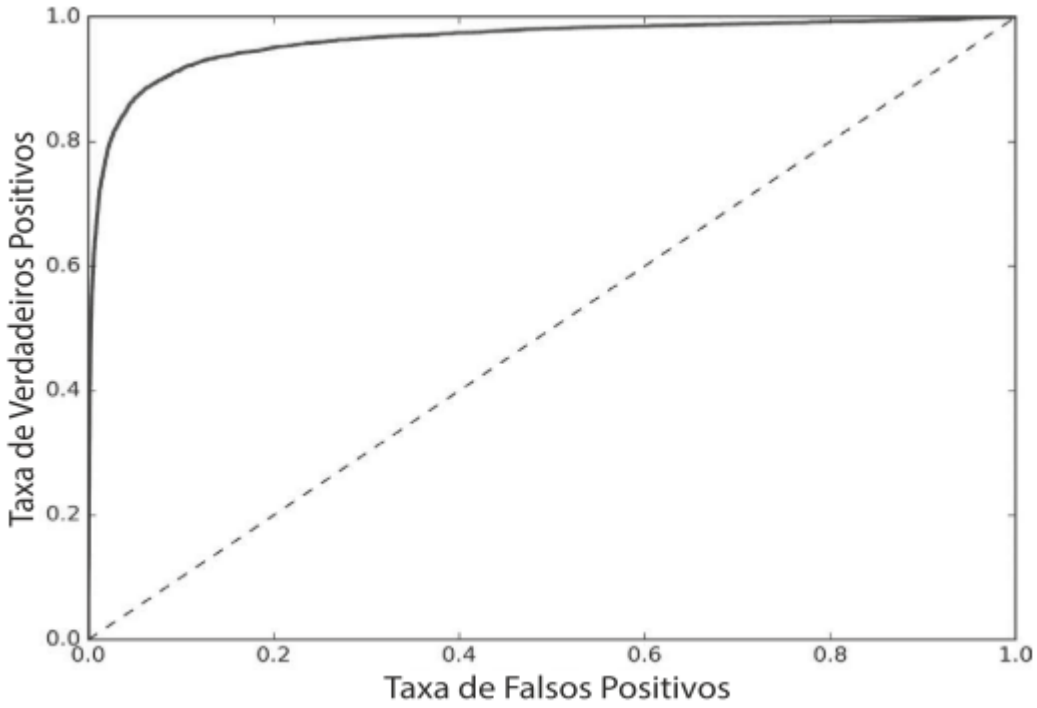
As medidas anteriores distinguem a classificação correta de rótulos dentro de classes, se concentrando em uma classe. O *recall* é uma função que exemplifica os casos classificados corretamente (VP) e os mal classificados (FN). A função de precisão exemplifica de verdadeiros positivos e casos de falsos positivos e o f1-score é uniformemente balanceado. Uma abordagem com maior abrangência pode ser obtida através da curva ROC (Curva Característica Operacional do Receptor) expressa pela equação 7.7, onde P(x|C) denota

uma probabilidade condicional de uma entrada de dados tenha uma classe C. A curva ROC gera os resultados de classificação, da classificação mais positiva para a mais negativa, sendo amplamente utilizada na análise de custo/benefício. O ROC e a Área Sobre a Curva (AUC) encontraram aplicações em casos com funções custos assimétricos e conjuntos de dados desequilibrados (SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006).

$$ROC = \frac{P(x|positivo)}{P(X|negativo)} \tag{7.7}$$

A curva ROC (Figura 11) plota a taxa de verdadeiros positivos (do inglês TPR – *True Positive Rate*) no eixo Y, também é um outro nome para *recall*, pela taxa de falso positivos (do inglês FPR – *False Positive Rate*) no eixo X. FPR é a razão entre os falsos positivos pelo total real de negativos, que é igual a um menos a especificidade (GÉRON, 2019; HUILGOL, 2020). Esta medida fornece *insights* geométricos sobre a natureza das medidas e sua capacidade para enviesar, além de permitir a comparação visual entre classificadores, entretanto, as condições para escolha do modelo ideal depende da minimização da AUC (POWERS, 2011).

Figura 11 – Curva ROC (Curva Característica de Operador Remoto)



Fonte: (GÉRON, 2019)

A AUC é uma das métricas mais populares, sendo aplicada amplamente na área biomédica e de estudos médicos. Essa medida foi utilizada para criação de um modelo de AM otimizado e para comparação de modelos. A AUC reflete o desempenho geral da classificação podendo ser obtido pela equação 7.8. Embora, tenha sido excelente para

processo de avaliação e descriminação, seu custo computacional é alto principalmente em problemas multi-classe (M; M.N, 2015).

$$AUC = \frac{\frac{s_p - n_p(n_n+1)}{2}}{n_p \times n_n} \tag{7.8}$$

onde s_p é a soma de todos os eventos positivos classificados e n_p, n_n são os números de exemplos positivos e negativos, respectivamente.

A AUC mede toda a área bidimensional abaixo de toda a curva ROC de [0,0] a [1,1]. Uma maneira de se interpretá-la é como a probabilidade de que o modelo classifique um exemplo positivo aleatório mais alto que um exemplo negativo aleatório (DEVELOPERS, 2020). Um excelente modelo apresenta AUC próximo a 1, o que indica que tem uma boa medida de separabilidade, já um modelo ruim tem um valor próximo a 0, que descreve que tem a pior medida de separabilidade e quando ele está na metade mostra que o modelo não tem capacidade de separação de classe presente (DEY, 2021; NARKHEDE, 2018) .

Pode-se interpretar a AUC observando a Figura 11, no ponto mais baixo em [0,0] o limite é definindo em 1, no ponto mais alto em [1,1] tem se definido o limite em 0. Para os outros pontos da curva tem-se os valores limiares, para FPR próximo 0 TRP está próximo a 1, quando isso acontece o modelo prevê todos os casos positivos quase que perfeitamente (HUILGOL, 2020). O estudo feito por P.Bradley (1997) chegou à conclusão que a área sobre a curva ROC mostrou propriedades desejáveis para uma métrica de desempenho de classificação como, aumento de sensibilidade nos testes de analise de variância, a métrica não depende de um limiar de decisão escolhido, indica o quão bem estão separadas as classes negativas e positivas para o índice de decisão, é invariante a relação de probabilidades de classes anteriores, dá uma indicação da quantidade de "trabalho feito" por um esquema de classificação.

7.6 Aprendizado de máquina aplicado na triagem e diagnóstico

Os diagnósticos médicos se enquadram em uma categoria de exames utilizados para detectar infecções, condições e doenças, sendo chamados de diagnóstico *in vitro* (IVD), onde são utilizadas amostras biológicas isoladas do corpo humano, como sangue ou tecido, para fornecer os resultados. Um estudo feito pelo Instituto de Medicina das Academias Nacionais de Ciências, Engenharia e Medicina mostra que cerca de 10% da morte de pacientes são causadas por erros de diagnóstico, e de 6% a 17% das complicações hospitalares. Então, a utilização de aprendizado de máquina aparece como um suporte para melhorar os diagnósticos médicos (FAGGELLA, 2020).

O aprendizado de máquina oferece uma abordagem que se baseia em princípios para desenvolvimento de algoritmos sofisticados, automáticos e objetivos para análise de

dados biomédicos multimodais e de alta densidade. Este tipo de método se baseia na capacidade dos algoritmos aprenderem ou adaptarem suas estruturas pela observação de um conjunto de dados, com uma adaptação feita por otimização através de uma função custo ou objetivo, sendo de grande interesse na comunidade biomédica por conta de sua capacidade de poder aumentar o poder preditivo ou de especificidade no diagnóstico e detecção de doenças na mesma forma em que aumentam o a objetividade do processo de tomada de decisão (SAJDA, 2006).

Algoritmos de inteligência artificial (IA) podem ajudar a prover uma variedade de cuidados com pacientes e para um sistema de saúde inteligente. As técnicas de IA vão desde o AM ao "Deep learning" e são predominantes no diagnóstico de doenças, descoberta de medicamentos e identificação de grau de risco de pacientes. E para isso se faz necessário uma grande base de dados médicos, para que o modelo leve a um diagnóstico perfeito usando este tipo de técnica. As abordagens de IA no sistema de saúde essências principalmente para o diagnóstico de doenças, como por exemplo muitos estudos a utilizam para diagnosticar doenças como diagnóstico de Alzheimer, câncer, diabetes, doenças crônicas, doenças cardíacas, acidente vascular cerebral e doença cerebrovascular, hipertensão, doenças de pele e doenças do fígado (KUMAR et al., 2022).

A utilização de computadores na pré-triagem de doenças é conhecida como ferramenta de pré-triagem auxiliada por computador (CAPST) e servem para interpretar as informações médicas ajudando a melhorar a precisão do diagnóstico. O desenvolvimento significativo da tecnologia de IA nos últimos anos mostrando sucesso em aplicações de diversos domínios, tendo como um ponto estimulante sua aplicação no diagnóstico e pré triagem de doenças e quais dados o modelo precisa pra ser treinado e aprender. Hoje existem muitos exemplos de modelos que utilizam AM no campo da medicina e seus resultados mostram uma melhora significativa na precisão de classificação (KUMAR, 2019).

À necessidade global de diagnosticar doenças de forma eficaz não é atendida, a complexidade de diferentes mecanismos de doenças e de sintomas subadjacentes são desafios no desenvolvimento de uma ferramenta de diagnóstico precoce eficaz. E o AM permite que alguns desses desafios possam ser superados. Em seu trabalho, (AHSAN; LUNA; SIDDIQUE, 2022) levantaram pontos desafiadores na utilização de AM no diagnóstico de doenças expressos na Tabela 9. Os autores concluem que a AM é uma ferramenta promissora para melhorar o diagnóstico de doenças, mas é necessário abordar esses desafios para garantir que ela seja usada de forma eficaz.

Tabela 9 – Desafios de AM no diagnóstico de doenças

Campo de desafio	Desafio	Explicação
Relacionados a dados	Escassez de dados	Embora os dados de muitos pacientes tenham sido registrados por diferentes hospitais e serviços de saúde, devido à lei de privacidade de dados, os dados do mundo real geralmente não estão disponíveis para fins de pesquisa global.
	Dados ruidosos	Frequentemente, os dados clínicos contêm ruídos ou valores ausentes; portanto, esse tipo de dado leva um tempo razoável para torná-lo treinável.
	Ataque adversário	O ataque adversário é um dos principais problemas no conjunto de dados de doenças. Ataque adversário significa a manipulação de dados de treinamento, dados de teste ou modelo de aprendizado de máquina para resultar em saída errada do AM.
Relacionados a diagnóstico de doenças	Classificação incorreta	Embora o modelo de aprendizado de máquina possa ser usado para desenvolver um modelo de diagnóstico de doenças, qualquer classificação incorreta para uma doença específica pode causar danos graves. Por exemplo, se um paciente com câncer de estômago for diagnosticado como um paciente sem câncer, isso terá um enorme impacto.
	Segmentação de imagem errada	Um dos principais desafios do modelo de AM é que o modelo geralmente identifica a região errada como uma região infectada.
	Confusão	Algumas das doenças como COVID-19, pneumonia, edema no peito muitas vezes apresentam sintomas semelhantes; nesses casos particulares, muitos modelos CNN detectam todas as amostras de dados em uma classe, ou seja, COVID-19.

Relacionados ao algoritmo	Supervisionados X não supervisionado	A maioria dos modelos de AM (regressão linear, regressão logística) teve um desempenho muito bom com os dados rotulados. No entanto, o desempenho de algoritmos semelhantes foram significativamente reduzido com os dados não rotulados. Por outro lado, algoritmos populares que podem ter bom desempenho com dados não rotulados, como clustering K-means, SVM e desempenho de KNNs, também degradaram com dados multidimensionais.
	Caixa preta	Um dos algoritmos de AM mais usados são as redes neurais convulsionais. No entanto, um dos principais desafios associados a esse algoritmo é que muitas vezes é difícil interpretar como o modelo ajusta parâmetros internos, como taxa de aprendizado e pesos. Na área da saúde, a implementação de um modelo relacionado a algoritmos precisa de explicações adequadas.

Fonte: (AHSAN; LUNA; SIDDIQUE, 2022)

Florestas aleatórias são capazes de lidar com dados ruidosos e ausentes, transparência do conhecimento do diagnóstico e a capacidade de explicar decisões. Por isso são frequentemente utilizadas em ferramentas de diagnósticos e triagens de doenças. Uma das formas de otimizar ainda mais essa capacidade desse tipo de algoritmo lidar com esses desafios é controlando o número de AD dentro da floresta (TRIPOLITI; FOTIADIS; MANIS, 2012).

Modelos de FA estão se mostrando extremamente representativos, tendo grande sucesso como método de regressão e classificação e está sendo amplamente aderido à indústria médica em casos como previsão de antígenos de células tumorais. Atualmente foi utilizado para previsão de COVID-19 nos Estados Unidos, além do diagnóstico e triagem de doenças (VLACHAS et al., 2022).

Peng et al. (2020) treinaram um classificador de florestas aleatórias, para melhorar a classificação de resultados falsos positivo e verdadeiros positivos, de 39 analitos metabólicos que são detectados na espectrometria de massa conjuntamente de um características clinicas. Nesse estudo, eles conseguiram reduzir o número de falsos positivos para quatro distúrbios, acidemia Glutárica tipo 1 (redução de 89%), acidemia metilmalônica (redução de 45%), deficiência de ornitina transcarbamilase (redução em 98%) e deficiência de acil-CoA desidrogenase de cadeia muito longa (redução de 2%).

Parte III

METODOLOGIA

8 MÉTODO

Primeiramente para a realização da dissertação foi feita uma revisão bibliográfica sobre a fibrose cística para averiguar os fatores físicos, regionais, condições de nascimento, tempo de coleta de amostra entre outras características. Para ter um banco de dados foi necessário estabelecer parceria com centro de referência que realize a triagem de FC para saber o índice de número dos pacientes que eram triados erroneamente. Com os índices será criado um banco de dados sintético para realizar a prova de conceito para criar uma metologia de desenvolvimento de modelo de floresta aleatória que possa fazer previsão de falsos positivos nesse banco de dados e depois aplica-lo em um trabalho futuro a dados reais. A importância dos índices foi tornar os dados sintéticos minimamente representativos ao que se diz em relação a taxa de falsos positivos.

Felizmente, foi obtida uma parceria com laboratório de triagem neonatal APAE São Luís do Maranhão. O laboratório forneceu os índices qualitativos de pacientes triados positivamente no primeiro e segundo exame de triagem através da quantificação de IRT e dos casos realmente confirmados entre os anos de 2017 a 2021. Com esses dados em mãos, foi criado o banco de dados sintético e criada uma metodologia em busca dos melhores parâmetros pros modelos utilizando os dados das características dos pacientes gerados sinteticamente e quando se tiver os dados reais essa mesma metodologia será empregada a eles e será avaliada as métricas obtidos vendo de fato a viabilização de um software que auxilie na triagem de FC.

Serão analisadas diversos treinamentos do modelo, comparando entre cada um seus desempenhos em diversos métricas como sensibilidade, precisão, acurácia, entre outras métricas de avaliação de modelo de AM. Os dados serão divididos em dois para avaliação geral do modelo: (1) dados de treinamento no qual vai ser a maior parte dos dados e servirão para treinar o programa desenvolvido, e (2) dados de teste, o qual os dados são desconhecidos pelo modelo e servirão para avaliar a capacidade do modelo fazer previsões em um dado que ele não conhece previamente. A análise de desempenho da previsão será utilizada para determinar a probabilidade de a amostra ser um falso positivo ou não, e o poder preditivo do modelo.

Na primeira etapa foi criado um modelo simples de FA para validar a teoria proposta de utilizar um modelo de FA para fazes previsões de FP e em seguida esse modelo foi tendo parâmetros variados com o intuito de melhorar a capacidade desse modelo fazer previsões em um banco de dados desconhecido, para se obter os melhores valores de parâmetros eles foram iterados "N"vezes, um de cada vez, e foi registrado sempre o que retornava a melhor sensibilidade no conjunto de dados de teste.

8.1 Modelo de validação da teoria

Um modelo de validação da teoria da linha de pesquisa foi realizado utilizando dados sintéticos para fazer a previsão de pacientes verdadeiros e falsos positivos para fibrose cística. O objetivo desse modelo foi verificar se com dados criados artificialmente se baseando em distribuições de índices do laboratório de triagem neonatal APAE de São Luís – MA, que forneceu para os estudos a quantidade total de testes realizados, o número de casos positivos na triagem, e os realmente confirmados positivos. Os dados sintéticos foram criados se baseando nessas informações juntamente de características dos pacientes utilizados no trabalho referência e sua mesma porcentagem em um dos grupos de estudo de Peng e colaboradores (PENG et al., 2020).

Para isso essa etapa foi dividida em duas, criação do banco de dados e desenvolvimento do modelo de aprendizado de máquina.

8.1.1 Banco de dados artificial

Para desenvolvimento do banco de dados sintético, 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 fez um trabalho de proposta parecida no programa de triagem neonatal local da Califórnia, porém para outras doenças, foi decidido primeiramente testar os mesmos dados dos recém nascidos utilizados por este autor. As porcentagens de distribuição dos dados criados foram aproximadamente as utilizadas pelo autor na população teste de Acidúria Glutárica tipo I (GA-I) (PENG et al., 2020).

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 sintética para que a mesma fosse mais próxima da realidade. Os dados fornecidos são de índices do laboratório, não foram fornecidos dados de pacientes. Os dados fornecidos pelo parceiro estão na Tabela 10. É importante dizer que a metodologia adotada pelo laboratório é que em caso de 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. Esses dados são importantes, pois são os rótulos buscados pelo modelo para determinar se o paciente é um verdadeiro positivo ou não.

Então foram utilizados somente os dado referentes ao ano de 2020 mais as características usadas no trabalho de Peng para criar uma base de dados sintéticos de 2020, que será utilizado para treinar um modelo que retorne métricas que demonstrem uma previsão que represente uma taxa de precisão de acima de 50% e sensibilidade de 100%. Esse banco de dados foi dividido para treinar e testar o modelo, na etapa de teste o banco de dados

foi utilizado para testar o modelo de FA em um banco de dados desconhecido.

Tabela 10 – Dados de Resultados de Fibrose Cística 2020 APAE São Luis - MA

Ano	Resultado	Quantidade
2021 (até julho)	1° Resultado alterado	22
	2° Resultado alterado	11
	Confirmado	1
	Total de Testes	48.503
2020	1° Resultado alterado	63
	2° Resultado alterado	10
	Confirmado	7
	Total de Testes	80.199
2019	1° Resultado alterado	44
	2° Resultado alterado	11
	Confirmado	11
	Total de Testes	95.815
2018	1° Resultado alterado	59
	2° Resultado alterado	2
	Confirmado	2
	Total de Testes	89.396
2017	1° Resultado alterado	43
	2° Resultado alterado	2
	Confirmado	1
	Total de Testes	82.767

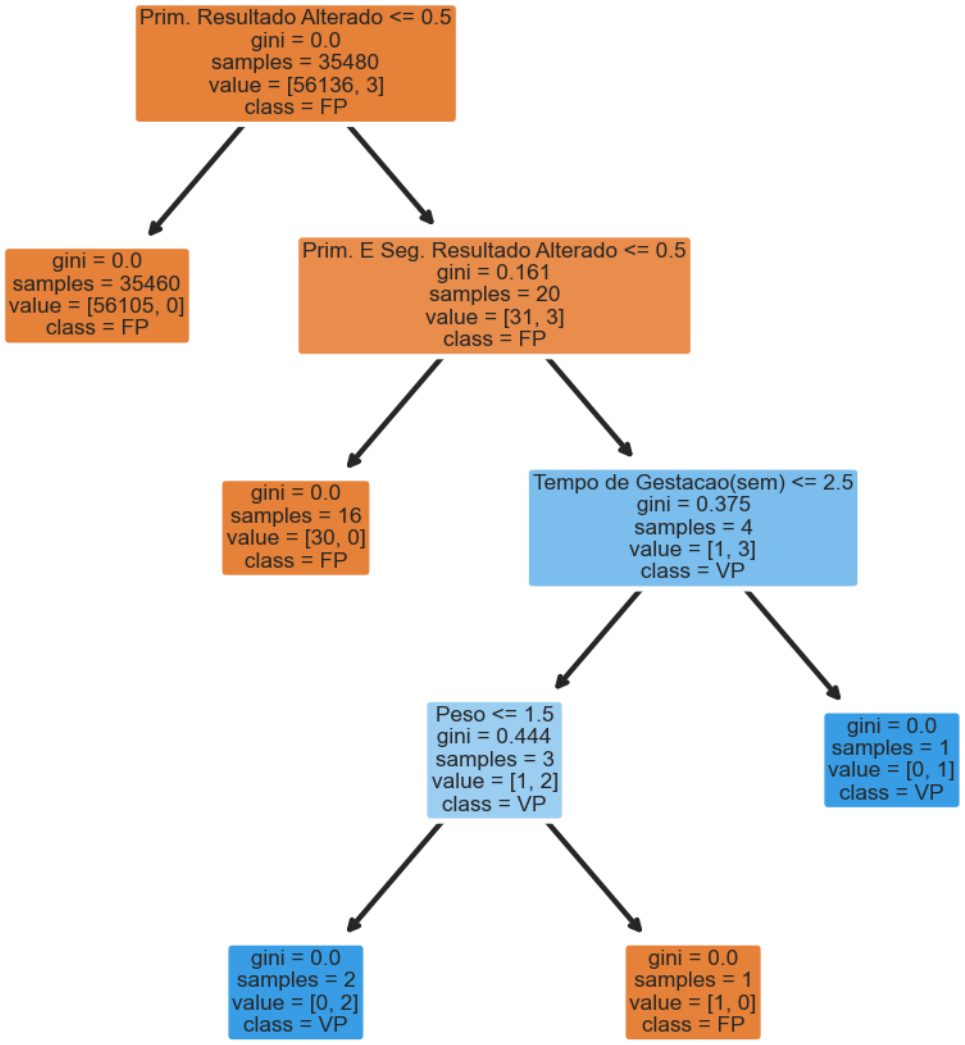
8.1.2 Modelo de aprendizado de máquina

Foi decidido utilizar um modelo de floresta aleatória para fazer a previsão da probabilidade de o resultado confirmatório do paciente para ser de fato um verdadeiro positivo. Modelos de florestas aleatórias são ideais para esse tipo de problema devido ao fato de utilizarem um conjunto de árvores de decisão que faz com que o modelo tenha um resultado ainda mais promissor do que com uma única árvore devido ao fato do resultado ser uma eleição das classificações de todas as árvores de decisão. Se você agregar as previsões de um conjunto de previsores para ter como resultado de previsão de um modelo, 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 12 pode ser visto um modelo de uma das árvores da floresta aleatória, onde é possível observar os critérios gerados para fazer a classificação desta árvore e a métrica de impuridade utilizada, o valor tomado para a decisão e a classe para a situação.

Neste exemplo de árvore no nó raiz tem-se o primeiro critério de divisão onde é analisado se as 35.480 amostras tem o primeiro resultado de triagem alterado, onde se o atributo for menor que 0,5 (cada atributo recebeu um rotulo que é um número inteiro) dividem para um novo nó criando novos critérios até se obter um gini igual a zero onde se tem o nó folha, ou seja aquele que só tem uma classe.

Figura 12 – 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 *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.

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

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 modelo, ou seja, a proporção dos valores que eram realmente positivos, em relação ao todos os casos que de fato eram positivos, essa métrica é chamada de *recall*.

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 pode fazer 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.

8.2 Banco de dados para treinar e testar novos modelos

Após a validação da teoria proposta com o banco de dados criado a partir dos índices da APAE de São Luís do Maranhão do ano de 2020 e as características com valores sintéticos e sua distribuição percentual a mesma de (PENG et al., 2020), foram feitos mais dois bancos de dados diferentes para que fosse possível comparar o aprendizado dos modelos nestas duas novas bases, sendo eles denominados pelo autor como "Banco de dados expandido"e "Banco de dados sintéticos 2017-2021".

Os dois banco de dados foram criados seguindo a mesma sistemática utilizada para gerar o banco de dados na fase de validação da teoria. Para o banco de dados expandido foi feita uma adição no conjunto de teste da etapa de validação da teoria com dados gerados a partir do anos de 2017, 2018, 2019 e 2021, onde esse conjunto foi utilizado para teste do modelo treinado com os dados de treinamento sintéticos de 2020. Já o conjunto de dados de 2017-2021 foi criado utilizando os dados de todos os anos e esta base foi utilizada tanto para treinamento quanto para teste, sendo dividida em uma proporção de 70% e 30% respectivamente. Todos os conjuntos de dados utilizados estão disponíveis em: https://doi.org/10.5281/zenodo.8351811.

8.2.1 Novos modelos para melhora de performance

Com os novos banco de dados foi iniciada a etapa de treinamento e teste dos modelos utilizando esses dados. No primeiro teste foi testada a mesma estrutura do modela da validação da teoria nessas bases comparando os resultados com os obtidos na primeira base de dados que implicava somente os dados sintéticos de 2020. Essa primeira estrutura foi denominada como "M1"e ela era estruturada com 100 árvores de decisão e 10 camadas máximas de profundidade cada uma das árvores.

Em seguida visando buscar a melhora de performance de cada um dos modelos foram variados parâmetros em busca daquele parâmetro que retorne a melhor métrica, para isso o parâmetro de profundidade e numero de arvores será iterado em um *loop* de 1000 ciclos e armazenado a melhor métrica registrada e tendo esse como melhor parâmetro, já para os outros parâmetros serão 250 ciclos com ressalva ao critério de divisão e exclusão de características que suas aplicações não cabem iterações. Essa redução de ciclos ocorreu devido ao alto tempo de processamento do modelo que chegou a demorar mais de 14 horas para encontrar cada um dos primeiros parâmetros que fora iterados 1000 vezes.

Na tabela 11 tem se todos os modelos criados e quais foram os parâmetros variados em cada um desses modelos, ao todo foram criados 8 modelos diferentes e foram denominados de M1 a M8, essa variação foi aplicada em treinamento e teste para os dois bancos de dados, na seção de resultados serão apresentados os resultados obtidos com cada um desses modelos. Importante dizer que os modelos vão utilizando os parâmetros obtidos nas etapas anteriores.

Tabela 11 – Modelos de aprimoramento da performance

Modelo	Parâmetro variado
M1	Modelo padrão (100 árvores + 10 camadas max. de profundidade)
M2	Número de árvores
M3	Camadas max. de profundidade
M4	Máximo de atributos
M5	Aleatoriedade dos dados
M6	Critério de divisão
M7	Exclusão de atributos menos importantes
M8	Exclusão do atributo de segundo resultado de triagem

Parte IV

RESULTADOS, DISCUSSÃO E CONCLUSÃO

9 RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados todos os resultados que foram obtidos no decorrer do desenvolvimento desta dissertação.

9.1 Modelo de validação com dados artificiais referentes a 2020

Primeiramente os dados referentes a 2020 foram distribuídos em uma tabela CSV (Comma-separated values, traduzindo em português valores separados por vírgula) em uma proporção como é possível ver na Tabela 12 que mostra a distribuição dos dados sintéticos criados seguindo a metodologia proposta para validação da teoria, onde os dados referentes aos resultados de triagem e diagnósticos confirmatórios fornecidos pela APAE de São Luís do Maranhão foram relacionados com características dos pacientes com valores sintéticos, sendo essas características as mesmas utilizadas no trabalho de (PENG e colab., 2020), as distribuições percentuais de cada característica também foram as mesmas utilizadas pelo autor. Para facilitar a implementação do modelo cada uma das classes de cada característica recebeu um valor de rotulo que esta indicado na tabela logo a frente do sinal de igual e o valor percentual que essa classe representa está logo após a seta.

Tabela 12 – Distribuição dos dados artificiais 2020

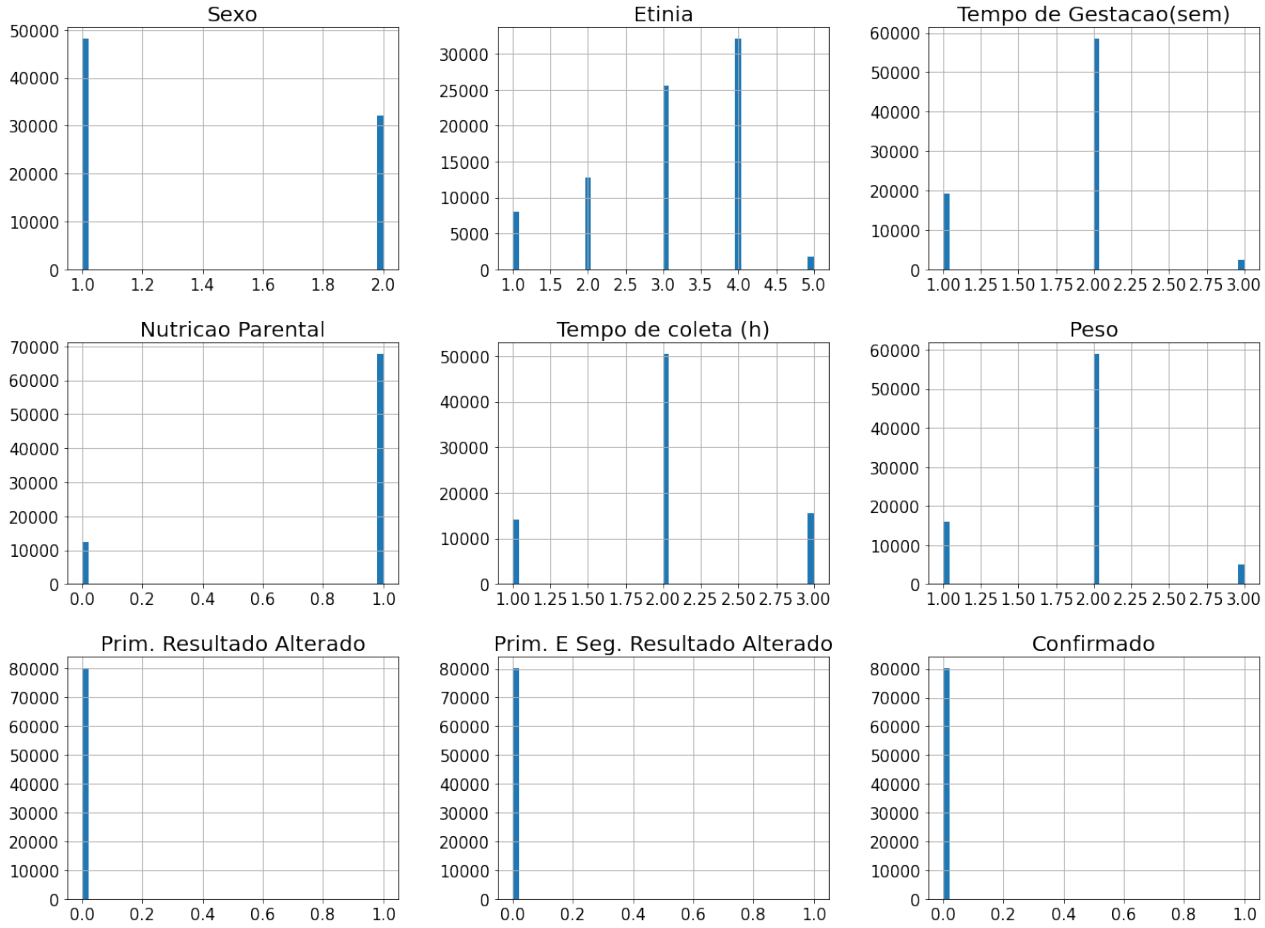
Dado:	Observações de distribuição do banco de dados e suas porcentagens				
Sexo	Masculino = 1 → 60%		Feminino = 2 → 40%		
Etnia	Asiático = 1 → 10%	Negro = 2 → 15%	Latino = 3 → 32%	Branco = 4 → 40%	Outras = 5 → 3%
Idade Gestacional (Sem)	<37 = 1 → 24%	37 - 41= 2 → 73%		>41 = 3 → 3%	
Nutrição Parental	Sim = 1 → 10%		0 = Não → 90%		
Peso ao nascer (g)	<2500 = 1 → 20%	2500 – 4000 = 2 → 73%		>4000 = 3 → 6%	
Tempo de coleta da amostra (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

Na figura 13 tem-se um histograma da distribuição dos dados sintéticos criados, para ajudar na visualização. E como pode-se observar na tabela juntamente ao gráfico, embora a porcentagem de pacientes alterados seja baixa no primeiro teste de quantificação

do IRT, quando se compara a porcentagem destes resultados de pacientes alterados com a de alterados no segundo teste de quantificação do IRT vê-se que esta porcentagem é amplamente superior e essa diferença aumenta ainda mais quando se comparado com os casos confirmados através do teste do suor.

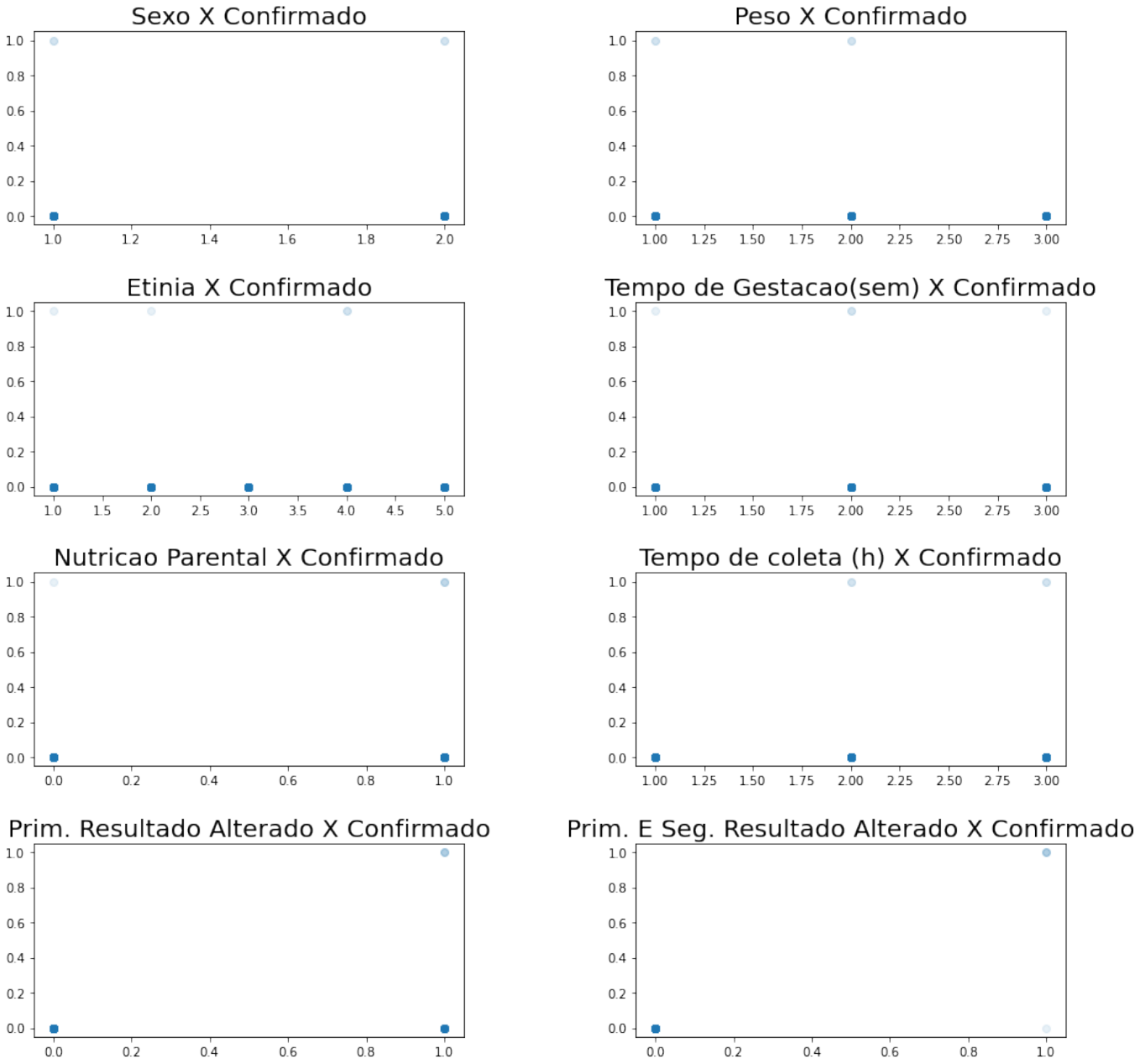
Figura 13 – Histograma de distribuição dos dados artificiais referentes a 2020



Fonte: O autor.

Os dados foram divididos em dois conjuntos, o conjunto de teste e de treinamento no qual eles têm respectivamente 30% e 70%, do conjunto original de dados artificiais referentes ao ano de 2020. No primeiro momento para desenvolvimento do modelo se faz necessário a divisão dos dados de treino, onde foi definido que o *target* seria os pacientes "Confirmados". Na Figura 14 pode-se ver a relação entre cada característica e o *target*, ou seja onde a marcação esta mais escura representa que a frequência daquele atributo estar relacionado com o *target* é de maior incidência. Onde pode se observar o azul mais intenso no gráfico representa aquela hipótese que ocorre por mais vezes.

Figura 14 – Dispersão dos atributos em relação ao *target*



Fonte: O autor.

Outro parâmetro utilizado para analisar os dados do conjunto de treinamento foram os índices de correlação padrão ou de Pearson (r) entre as características e o *target*, foram criados também subgrupos juntando as características para medir as suas respectivas correlações com o *target*, as correlações obtidas estão descritas na Tabela 13, em sequência das que mais se correlacionam positivamente até as que se correlacionam negativamente. Através da correlação padrão e possível ter noção de como os atributos se relacionam com o *target* e nos subgrupos onde teve a junção de características pode trazer uma combinação de características que pode ajudar o modelo explicar ainda melhor o conjunto de dados, lembrando sempre que se faz necessário a avaliação dos atributos, pois atributos que tem

baixa correlação com o alvo podem trazer sobreajuste ou subajuste ao modelo.

Tabela 13 – Correlação padrão entre atributos e *target* (Confirmados) no conjunto de treino

Atributo	r de Pearson
Confirmado	1
Prim. E Seg. Resultado Alterado	0.89442
Nutricao Parental_per_Prim. E Seg. Resultado Alterado	0.612372
Prim. Resultado Alterado	0.298047
Tempo de coleta (h)_per_Prim. Resultado Alterado	0.170531
Tempo de Gestacao(sem)_per_Prim. Resultado Alterado	0.17002
Sexo_per_Prim. Resultado Alterado	0.07874
Nutricao Parental_per_Prim. Resultado Alterado	0.0268
Tempo de coleta (h)	0.006274
Sexo_per_Etinia	0.005528
Tempo de Gestacao(sem)	0.003475
Sexo	0.001634
Tempo de Gestacao(sem)_per_Prim. E Seg. Resultado Alterado	0
Nutricao Parental_per_Tempo de Gestacao(sem)	-0.00056
Peso_per_Etinia	-0.00067
Sexo_per_Nutricao Parental	-0.00098
Sexo_per_Tempo de Gestacao(sem)	-0.00123
Nutricao Parental	-0.00211
Etinia	-0.00258
Sexo_per_Tempo de coleta (h)	-0.00288
ID	-0.00297
Tempo de Gestacao(sem)_per_Tempo de coleta (h)	-0.00335
Peso_per_Tempo de Gestacao(sem)	-0.0035
Etinia_per_Tempo de Gestacao(sem)	-0.00366
Nutricao Parental_per_Tempo de coleta (h)	-0.00385
Peso_per_Sexo	-0.00472
Etinia_per_Nutricao Parental	-0.00549
Peso	-0.00582
Etinia_per_Tempo de coleta (h)	-0.00592
Peso_per_Tempo de coleta (h)	-0.00651
Peso_per_Nutricao Parental	-0.00799
Etinia_per_Prim. E Seg. Resultado Alterado	-0.08575
Etinia_per_Prim. Resultado Alterado	-0.18021
Peso_per_Prim. Resultado Alterado	-0.31155
Peso_per_Prim. E Seg. Resultado Alterado	-0.40825
Sexo_per_Prim. E Seg. Resultado Alterado	-0.40825
Tempo de coleta (h)_per_Prim. E Seg. Resultado Alterado	-0.40825
Prim. Resultado Alterado_per_Prim. E Seg. Resultado Alterado	NaN

Com a análise dos dados de treinamento feita, utilizou-se esses dados para criar o modelo de florestas aleatórias, para isso foi necessário fazer uma análise nos dados para remover valores indesejados com valores nulos ou que tendem ao infinito, isso para os

valores gerados a partir da combinação de dados pois devidos as divisões os resultados podem cair nesta situação, esses valores foram substituídos por 0. Então foi criado um modelo de FA com 100 arvores de decisão com no máximo 10 camadas de profundidade, utilizando o critério de gini, esse modelo será chamado de M1. Onde forma obtidos os resultados da tabela 14 no conjunto de treinamento, esses valores foram obtidos através de cálculos feitos dentro do próprio modelo.

Tabela 14 – Performance do modelo de validação com dados de 2020 no conjunto de treinamento

Métrica	Score (%)
R2	100
Acurácia	100
Precisão	100
Recall	100
F1-Score	100

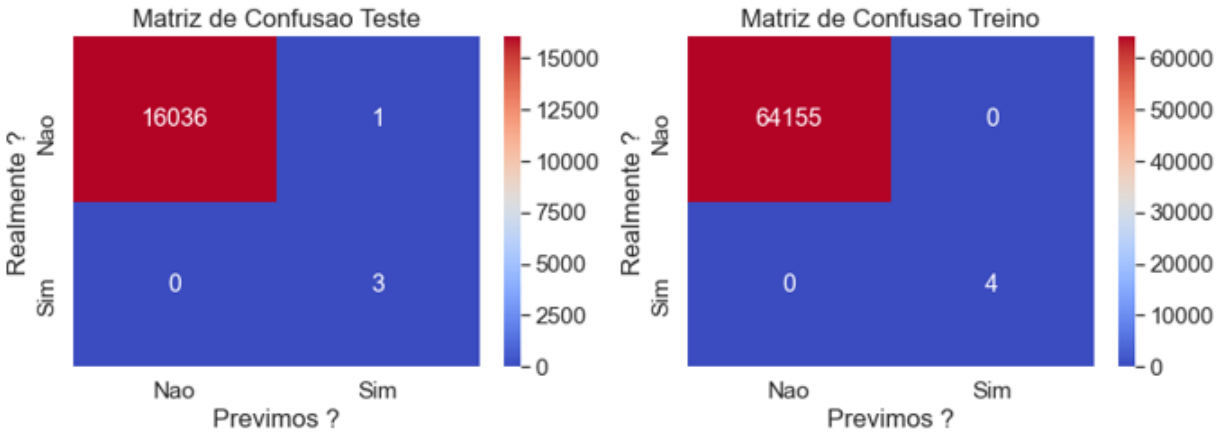
O modelo de florestas aleatória foi capaz de fazer um bom aprendizado no conjunto de dados de treinamento prevendo todos os valores de *target* do conjunto corretamente. Esse mesmo modelo foi utilizado no conjunto de dados teste para poder avaliar seu comportamento preditivo em um banco de dados desconhecido pelo modelo e as métricas obtidas estão descritas na Tabela 15. Com as métricas é possível dizer que o modelo de avaliação nesse primeiro momento de validação pode fornecer uma melhora significativa com 75% de melhora da precisão e mantendo a sensibilidade do exame em 100%.

Tabela 15 – Performance do modelo de validação com dados de 2020 no conjunto de teste

Métrica	Score (%)
R2	66.7
Acurácia	100
Precisão	75
Recall	100
F1-Score	85.7

Na Figura 15 está a matriz de confusão dos dois conjuntos de dados, onde é possível ter de forma visual os acertos e erros do modelo nos 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 equivalentes a 64.159 dados e trinta por cento no conjunto de teste equivalentes a 16.040 dados, que é o responsável por fazer a validação do modelo treinado.

Figura 15 – Matriz de confusão conjunto teste e treino dados artificiais 2020



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 modelo a classificou como positiva. Embora o modelo tenha errado essa classificação como mostrado nas tabelas acima que descreve as métricas obtidas pelo modelo no conjunto de dados de 2020. 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, tem-se uma queda na precisão quando se comparado ao conjunto teste.

9.2 Modelo M1 aplicado em banco de dados expandido

Esse modelo criado foi testado no conjunto de dados simulados utilizando a mesma sistemática referente aos anos de 2017 até metade de 2021, sem considerar o ano de 2020, sendo ao todo 316.481 dados para analisar como o modelo iria se comporta dentro de um banco de dados com maior incidência de casos positivos, se a sua eficácia se manterá a mesma que no conjunto de teste que era um conjunto menor de dados. Na tabela 16 estão descritas as métricas obtidas quando se foi aplicado o modelo no conjunto maior de dados.

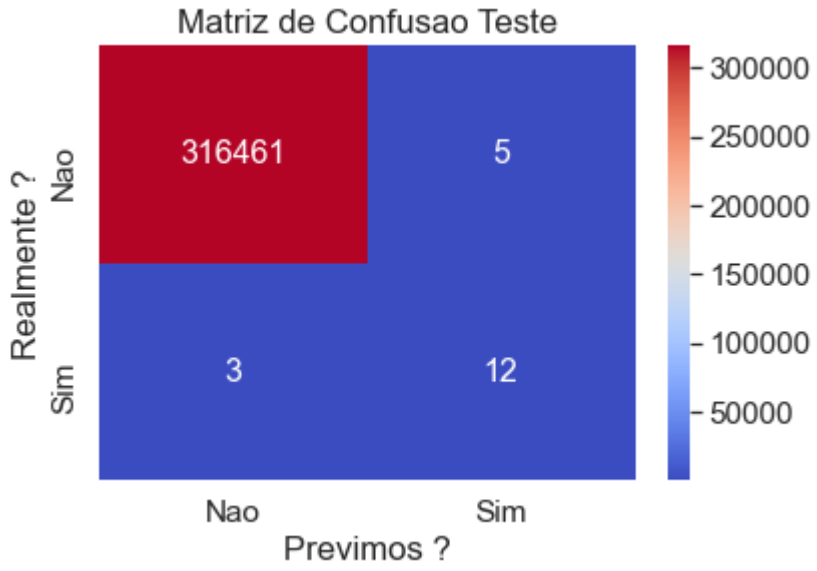
Tabela 16 – Performance do modelo de validação com dados de 2020 no conjunto de teste expandido

Métrica	Score (%)
R2	46.7
Acurácia	100
Precisão	70.6
Recall	80
F1-Score	75

Ao analisar os resultados obtidos é possível entender que esse modelo apresentou uma queda no seu poder preditivo onde a *recall* e precisão caíram em relação ao *scores*

percentuais obtidos no teste anterior, isso pode ser explicado por um sobreajuste dos dados no qual ele está representando muito bem os dados no conjunto de treino atingindo pontuação máxima em todas as métricas, já em um banco de dados desconhecido ele não conseguiu ter previsões com a mesma precisão piorando ainda mais quando o conjunto desconhecido contém mais dados, perdendo até mesmo a sensibilidade tendo casos de falsos negativos como pode-se ver na matriz de confusão Figura 16.

Figura 16 – Performance do modelo de validação com dados artificiais de 2020 no conjunto de teste expandido



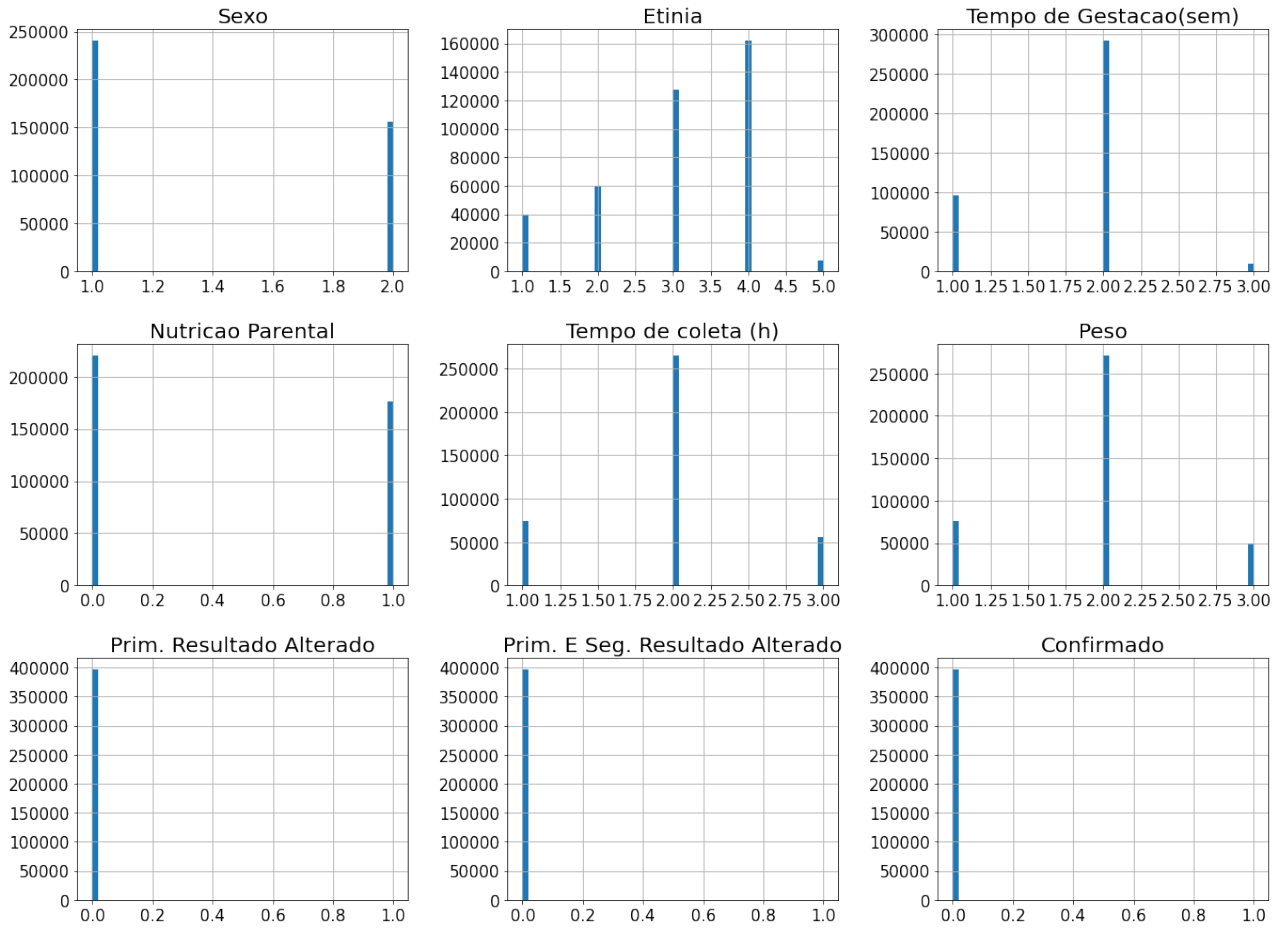
Fonte: O autor.

Como foi visto que o modelo utilizando somente o recorte de 2020 não explicou com a sensibilidade desejada um banco de dados desconhecido, ou seja não atingiu a métrica de recall de 100%. Então foi desenvolvido um modelo utilizando todos os dados simulados (396.680 dados) referente ao recorte dos anos de 2017 a metade de 2021, dividindo-os em uma proporção de 80% e 20% para os conjuntos de treino e teste respectivamente, assim podendo analisar se é possível obter melhora nas previsões do modelo aumentando o tamanho das amostras dos conjuntos.

9.3 Aplicando estrutura M1 em banco de dados simulados entre 2017 a 2021

Os dados referentes aos anos de 2017 a 2021 (ao todo 396.895 dados) foram divididos em dois conjuntos, o conjunto de teste e de treinamento no qual eles têm respectivamente % e 70%, do conjunto original de dados artificiais da mesma maneira em que foi feito no conjunto de dados somente do ano de 2020. Na Figura 17 pode-se ver a relação entre a tendência de cada atributo, onde no eixo X tem-se os rótulos para cada classe do atributo e no eixo Y tem-se o rótulo do *target*, o resultado confirmatório.

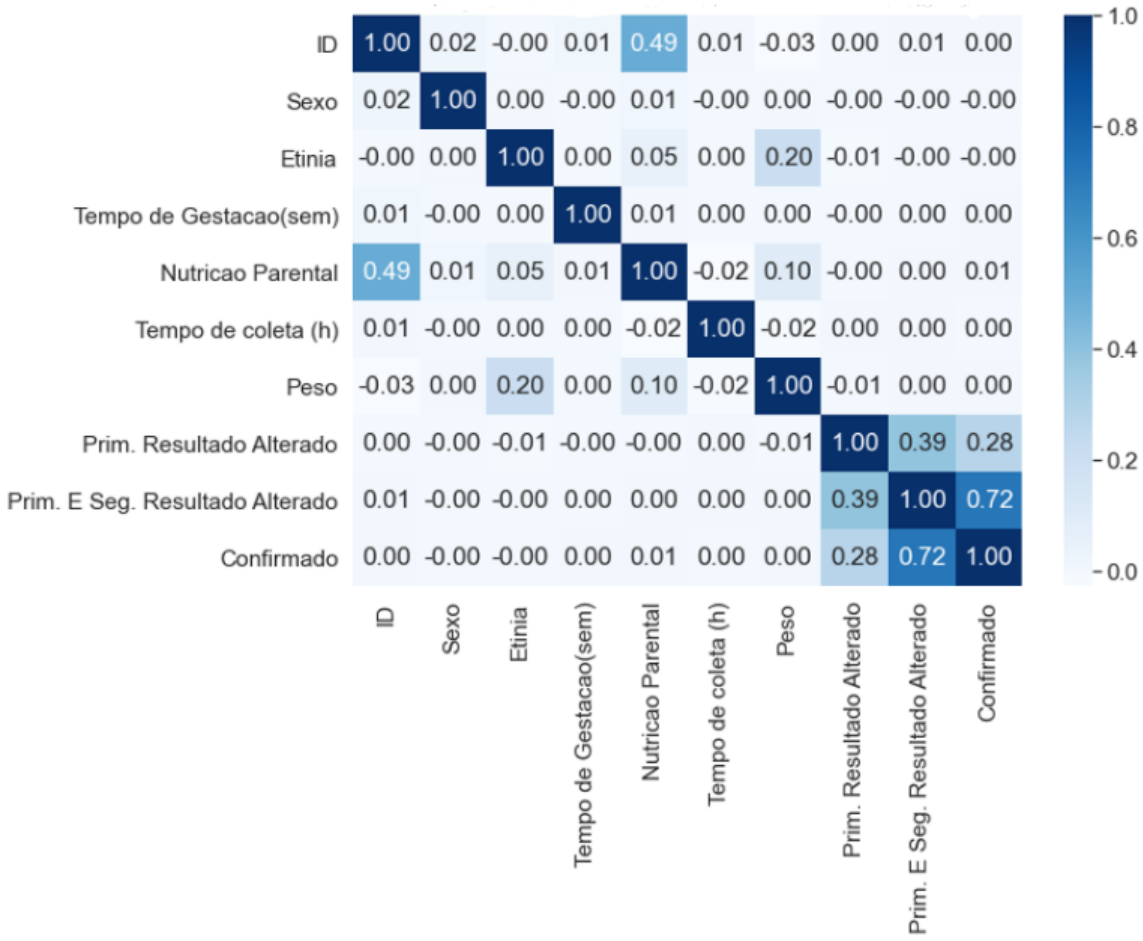
Figura 17 – Histograma de distribuição dados artificiais 2017 - 2021



Fonte: O autor.

Para entender um pouco mais como os dados de treino estão correlacionados a Figura 18 mostra o quanto cada variável se correlaciona com o *target* assim tornando simples a compreensão dos atributos com maior influência sobre o banco de dados. E mostrando que como no banco de dados artificial com os dados somente do ano de 2020 os que tem maior correlação são dos dados fornecidos pelo parceiro, pacientes confirmados, e alterados nas triagens.

Figura 18 – Correlação dados 2017 - 2021



Fonte: O autor.

Utilizando um modelo de FA de mesma estrutura que o utilizado no modelo que foi treinado e testado com os dados de 2020 o "M1", para verificar como o aumento do banco de dados de treinamento iria influenciar nas previsões. Teve-se como resultado obtidos as métricas expressas na Tabela 17 para o conjunto de teste, já no conjunto de treinamento obteve-se a mesmas métricas obtidas com o modelo com os dados se 2020.

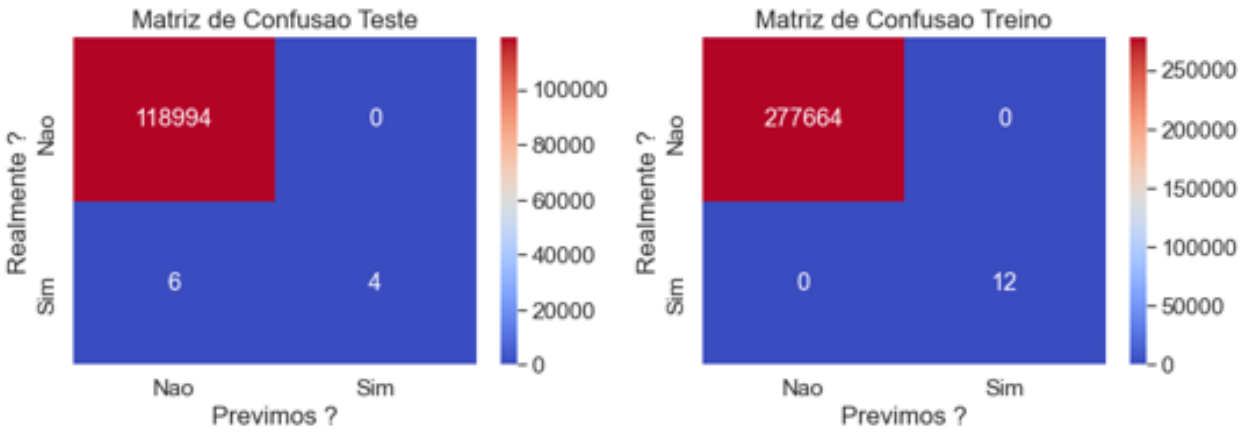
Tabela 17 – Performance do modelo M1 de validação com dados de 2017 a 2021 conjunto teste

Métrica	Score (%)
R2	40
Acurácia	100
Precisão	100
Recall	40
F1-Score	57.1

Os resultados obtidos não trouxeram a melhora esperada, tendo efeitos contrários do que se espera, a métrica mais importante que caracteriza a sensibilidade do sistema piorou, na matriz de confusão da Figura 19, é possível ver que mesmo com a maior

quantidade de casos positivos num banco de dados maior o modelo não conseguiu prevê-los com uma alta sensibilidade no conjunto teste, prevendo 6 dos 10 casos como FN e como no de treinamento ele acertou todos os casos indica o sobreajuste do modelo aos dados.

Figura 19 – Matriz de Confusão modelo M1 gerado a partir de dados artificias 2017 - 2021



Fonte: O autor.

Como com a alteração do tamanho dos bancos de dados não surtiu efeito positivo no poder preditivo do modelo, testes com modelos com estruturas diferentes foram realizados e testados também nesses dois cenários anteriores para se chegar em um modelo que tenha resposta preditiva satisfatória.

9.4 Novos modelos para melhorar a performance

Para melhorar o poder preditivo do modelo e manter a *recall* em 100% foram desenvolvidos diversos métodos fazendo diversos testes até que se obtivesse um modelo que apresentasse a precisão superior a 50% e a *recall* máxima também, isso levando em consideração os dois cenários em que se foram feitos os testes com o modelo de estrutura M1, utilizando o conjunto de dados de 2020 para treinamento mais o conjunto de dados expandido para teste e o conjunto de dados totais para treinamento e teste. Os testes realizados e seus resultados serão apresentados nessa seção individualmente e no final será feito uma análise geral comparando todos os modelos.

9.4.1 Alterando o número de árvores do modelo – M2

Foi utilizado um modelo em *loop* para encontrar o melhor número de árvores para estimar o melhor resultado de *recall*, isso mantendo a profundidade máxima de cada árvore de 10 camadas. Para isso o modelo variou o número de estimadores de 1 a 1.000 para encontrar o melhor parâmetro nos dois bancos de dados artificias usados nos testes anteriores.

O resultado obtido pelo modelo para ter o modelo com o melhor número de estimadores para o banco de dados que utiliza os dados de 2020, para criar o modelo e testá-lo ele em um banco de dados estendido, foi de 253 estimadores e as métricas obtidas estão descritas na tabela 18. Sendo importante ressaltar que foi utilizado o *loop* até 1000 estimadores devido ao tempo longo de processamento, para esse número levou-se cerca de 6 horas para encontrar o melhor valor, se fossem utilizados um range maior esse tempo cresceria exponencialmente.

Tabela 18 – Métrica melhor quantidade de estimadores - modelo M2 gerado a partir de dados artificiais de 2020

Conjunto de dados	Métrica	Score (%)
Conjunto de treino	R2	100
	Acurácia	100
	Precisão	100
	Recall	100
	F1-Score	100
Conjunto de Teste	R2	66.7
	Acurácia	100
	Precisão	75
	Recall	100
	F1-Score	85.7
Conjunto de teste expandido	R2	46.7
	Acurácia	100
	Precisão	70.6
	Recall	80
	F1-Score	75

O mesmo teste realizado no banco de dados descrito acima foi repetido para o banco de dados que considera todos os dados simulados entre os anos de 2017 a 2021. O melhor número de estimadores encontrados foi de 16 tendo as métricas da tabela 19 obtidas.

Tabela 19 – Métrica melhor quantidade de estimadores modelo M2 gerado a partir de dados artificiais de 2017-2021

Conjunto de dados	Métrica	Score (%)
Conjunto de treino	R2	100
	Acurácia	100
	Precisão	100
	Recall	100
	F1-Score	100
Conjunto de Teste	R2	60
	Acurácia	100
	Precisão	100
	Recall	60
	F1-Score	75

Comparando se os dois resultados com os obtidos nos testes do modelo M1 tem-se que o número de estimadores teve influência maior sobre o modelo gerado a partir dos dados totais onde teve-se um aumento de 20% no *recall* do conjunto de teste, também é notável o melhor desempenho em todos as outras métricas que aumentaram sua pontuação percentual significativamente.

9.4.2 Alterando o número de camadas máximas de profundidade – M3

Metodologia similar à utilizada no modelo M2, porém foram testados diversos valores de camadas máximas diferentes para encontrar a que resultasse no melhor poder preditivo ao modelo. Para isso foi colocado o programa em um *loop* de 1 a 1000 a cada iteração o valor do número máximo de camadas das arvores de decisão era o mesmo do *loop* que estava sendo executado e mantendo o número de arvores igual ao melhor valor encontrado para o modelo M2. Tendo como melhor valor encontrado 526 camadas de profundidade para o modelo treinado na base de dados referente ao ano de 2020 que gerou resultado as métricas de avaliação descritas na tabela 20. Teve-se mantido o recall encontrado no teste anterior de busca de estimadores, porém teve-se uma melhora nos outros parâmetros como precisão, R2 e F1-score, para o conjunto de teste expandido.

Tabela 20 – Métrica para melhor quantidade de camadas de profundidade modelo M3 gerado a partir de dados artificiais de 2020

Conjunto de dados	Métrica	Score (%)
Conjunto de treino dados 2020	R2	100
	Acurácia	100
	Precisão	100
	Recall	100
	F1-Score	100
Conjunto de Teste dados 2020	R2	66.7
	Acurácia	100
	Precisão	75
	Recall	100
	F1-Score	85.7
Conjunto de teste expandido	R2	53.3
	Acurácia	100
	Precisão	75
	Recall	80
	F1-Score	77.4

Esse mesmo teste foi feito criando um modelo a partir de todos os dados (de 2017 a 2021) tendo os resultados das métricas obtidas descritos na tabela X. Onde foi possível comparar com o resultado nesse mesmo banco de dados onde foi encontrado o melhor número de estimadores e viu-se que em com o melhor valor de camadas de profundidade encontrado foi em 272 camadas máximas de profundidade. Tendo as métricas iguais ao do

conjunto de teste M2 o que demonstra baixa influência da quantidade utilizada no modelo M2 de 10 camadas máximas para a encontrada como a melhor, quando se olha para o conjunto de treinamento teve-se uma queda no poder descritivo desse conjunto, porém não interferiu na caracterização de um banco dedados desconhecido.

Tabela 21 – Métrica para melhor quantidade de camadas de profundidade modelo M3 gerado a partir de dados artificiais de 2017-2021

Conjunto de dados	Métrica	Score (%)
Conjunto de treino	R2	83.3
	Acurácia	100
	Precisão	91.7
	Recall	91.7
	F1-Score	91.7
Conjunto de Teste	R2	60
	Acurácia	100
	Precisão	100
	Recall	60
	F1-Score	75

9.4.3 Modelo testado com outros hiperparâmetros – M4, M5, M6

Existem diversos parâmetros que podem influenciar no desempenho de um modelo de aprendizado de máquina, até o momento foram variados somente os valores do número de estimadores e do número máximo de profundidade de cada arvore, nesta seção serão abordados alguns outros parâmetros e serão avaliados sua influencias sobre o modelo. Porém como serão reduzidos os números de iterações do loop de 1000 para 250 devido ao alto custo de tempo de processamento do modelo até encontrar o melhor parâmetro.

O primeiro parâmetro a ser definido para seguir com os testes será o gerador aleatório da FA, esse foi o modelo M4 e o resultados obtidos estão descritos nas tabelas **??** e 23 , o valor encontrado para o modelo aplicado ao conjunto de dados do ano de 2020 foi de 214 e para o outro banco de dados que engloba 2017 a 2021 foi de 19.

Para o modelo criado a partir dos dados de 2020 teve uma melhora de 13% no recall do conjunto de teste expandido um aumento significativo e chegando próximo do objetivo de tem uma pontuação de 100%, mesmo que os parâmetros de precisão e R2 tenham caído. Já no modelo criado com o outro banco de dados os resultados se mantiveram os mesmos para o conjunto de treino, mas com melhora no conjunto de treinamento.

Tabela 22 – Métrica melhor valor para $Random_state$ modelo M4 gerado a partir de dados artificiais de 2020

Conjunto de dados	Métrica	Score (%)
	R2	100
	Acurácia	100

Conjunto de treino dados 2020

	Precisão	100
	Recall	100
	F1-Score	100
Conjunto de Teste dados 2020	R2	66.7
	Acurácia	100
	Precisão	75
	Recall	100
	F1-Score	85.7
Conjunto de teste expandido	R2	40
	Acurácia	100
	Precisão	63.6
	Recall	93.3
	F1-Score	75.7

Tabela 23 – Métrica melhor valor para $Random_state$ modelo M4 gerado a partir de dados artificiais de 2017-2021

Conjunto de dados	Métrica	Score (%)
Conjunto de treino	R2	91.7
	Acurácia	100
	Precisão	92.3
	Recall	100
	F1-Score	96
Conjunto de Teste	R2	60
	Acurácia	100
	Precisão	100
	Recall	60
	F1-Score	75

Depois de encontrado o melhor valor do parâmetro de ı$ramdom_state$ȷ, a variável responsável por selecionar os atributos para divisão ı$max_features$ȷ, para isso foi repetida a mesma sistemática dos passos anteriores, porém foi feito um rangem de 250 posições entre 0.004 e 1 o número que se pode alterar o parâmetro, esse será o modelo M5.

Os resultados obtidos com a busca do melhor valor do parâmetro para os bancos de dados estão descritos nas tabelas 24 e 25 onde teve para o conjunto de dados de 2020 o valor do parâmetro encontrado de 0.112 e para o outro conjunto de 0.084. É possível ver que não houve alteração nas métricas obtidas na etapa anterior no modelo com os dados de todos os anos, porém no modelo com os dados de 2020 teve melhora para os parâmetros de precisão, F1-score e R2.

Tabela 24 – Métrica melhor valor para $max_features$ modelo M5 gerado a partir de dados artificiais de 2020

Conjunto de dados	Métrica	Score (%)
Conjunto de treino dados 2020	R2	100
	Acurácia	100
	Precisão	100
	Recall	100
	F1-Score	100
Conjunto de Teste dados 2020	R2	66.7
	Acurácia	100
	Precisão	75
	Recall	100
	F1-Score	85.7
Conjunto de teste expandido	R2	53.3
	Acurácia	100
	Precisão	70
	Recall	93.3
	F1-Score	80

Tabela 25 – Métrica melhor valor para $max_features$ modelo M5 gerado a partir de dados artificiais de 2017-2021

Conjunto de dados	Métrica	Score (%)
Conjunto de treino	R2	91.7
	Acurácia	100
	Precisão	92.3
	Recall	100
	F1-Score	96
Conjunto de Teste	R2	60
	Acurácia	100
	Precisão	100
	Recall	60
	F1-Score	75

O último parâmetro que foi testado nessa avaliação será a alteração do critério de avaliação de gini para o de entropia e compara-los para ver qual apresenta melhor performances nos dois bancos de dados com os parâmetros obtidos pelos modelos anteriores. E como é possível observar nas tabelas 26 e 27 os resultados obtidos para os dois critérios foram exatamente o mesmo, sugerindo que nesse caso é indiferente a escolha dos dois critérios.

Tabela 26 – Comparação entre critérios de divisão dados artificiais de 2020 - M6

Conjunto de dados	Métrica	Score Gini (%)	Score Entropia (%)
Conjunto de treino dados 2020	R2	100	100
	Acurácia	100	100
	Precisão	100	100
	Recall	100	100
	F1-Score	100	100
Conjunto de Teste dados 2020	R2	66.7	66.7
	Acurácia	100	100
	Precisão	75	75
	Recall	100	100
	F1-Score	85.7	85.7
Conjunto de teste expandido	R2	53.3	53.3
	Acurácia	100	100
	Precisão	70	70
	Recall	93.3	93.3
	F1-Score	80	80

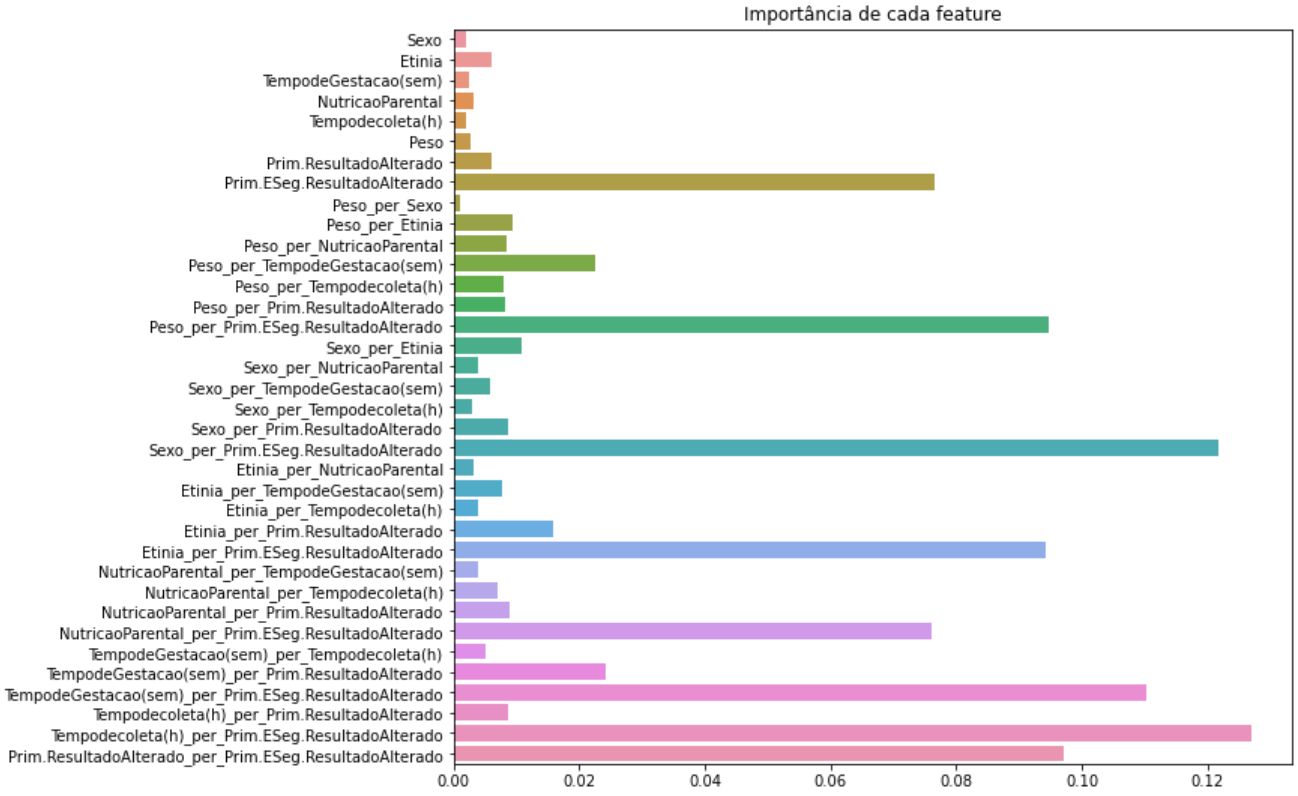
Tabela 27 – Comparação entre critérios de divisão dados artificiais de 2017-2021 – M6

Conjunto de dados	Métrica	Score Gini (%)	Score Entropia (%)
Conjunto de treino	R2	91.7	91.7
	Acurácia	100	100
	Precisão	92.3	92.3
	Recall	100	100
	F1-Score	96	96
Conjunto de Teste	R2	60	60
	Acurácia	100	100
	Precisão	100	100
	Recall	60	60
	F1-Score	75	75

9.4.4 Removendo os atributos menos significantes – M7

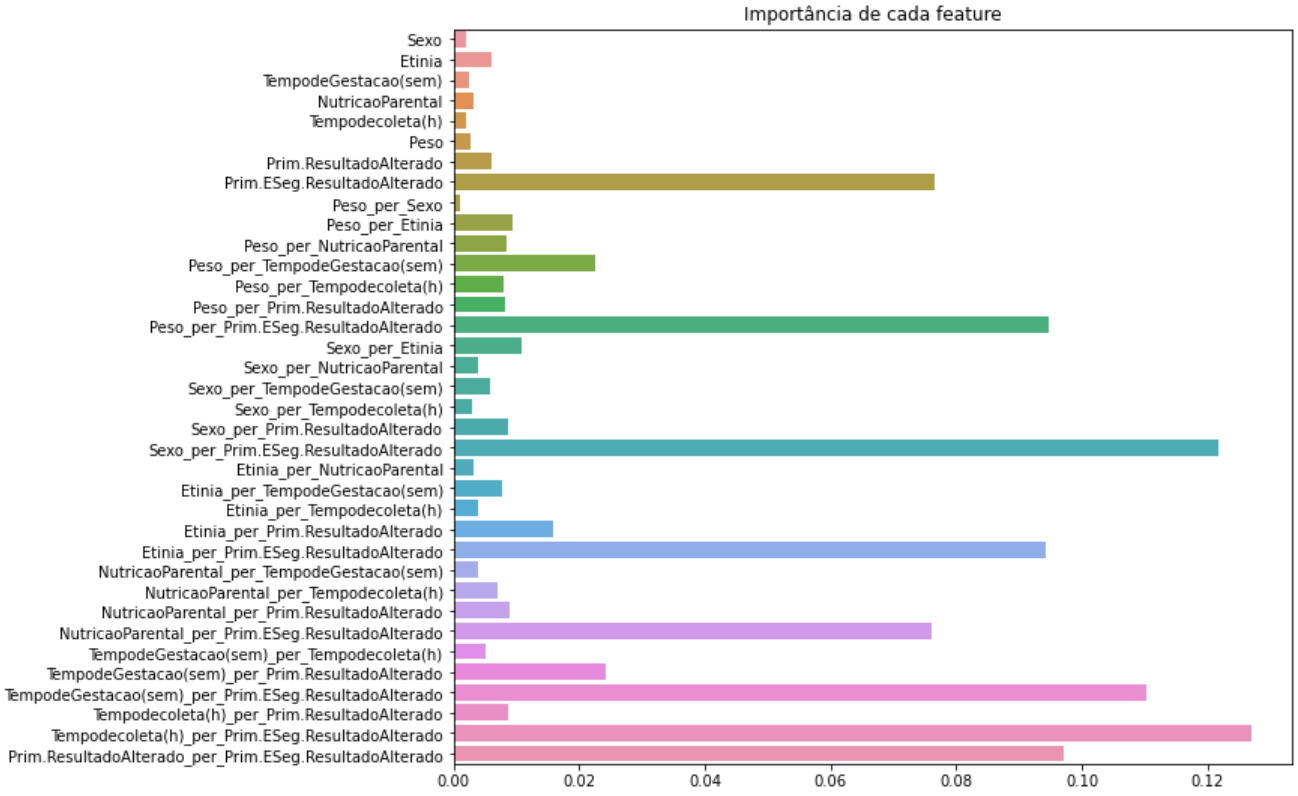
Outra metodologia utilizada para tentar obter o melhor modelo possível foi a de exclusão dos atributos menos significantes do banco de dados, ou seja, aqueles que tem a menor capacidade de explicar ou ter relação com o *target*. Para saber quais são esses dados foi utilizado a função $\llbracket feature_importances \rrbracket$ para saber os atributos mais importantes para o modelo, os resultados obtidos para o banco de dados de 2020 estão na figura 20 e do modelo de dados totais estão na figura 21, sendo importante ressaltar que por se tratar de dados sintéticos esses índices não expressam a realidade encontrada nos recém nascidos.

Figura 20 – Importância dos atributos modelo de dados artificial 2020



Fonte: O autor.

Figura 21 – Importância dos atributos modelo de dados artificial de 2017-2021



Fonte: O autor.

Para criar o modelo foram mantidos os dados apresentados na Tabela 28. O modelo executado teve como estrutura base os melhores parâmetros encontrados nos testes anteriores e rodado novamente sem essas características que não tem tenta relevância para o modelo. Foram desconsiderados os atributos com índice menor que o de etnia para os dados de 2020 e para os dados de 2017-2021 os menores que o primeiro resultado alterado.

Tabela 28 – Dados relevantes usados no primeiro teste

Banco de dados	**Dados relevantes utilizados**
Dados artificiais 2020	Etnia
	Etnia por 1° e 2° Resultado Alterado
	Etnia por Tempo de coleta (h)
	Etnia por Tempo de Gestação(sem)
	Etnia por 1° Resultado Alterado
	Nutrição Parental por 1° e 2° Resultado Alterado
	Nutrição Parental por 1° Resultado Alterado
	Nutrição Parental por Tempo de coleta(h)
	Peso por Etnia
	Peso por 1° e 2° Resultado Alterado
	Peso por Tempo de Gestação(sem)

	Peso por Nutrição Parental
	Peso por 1° Resultado Alterado
	Peso por Tempo de coleta (h)
	1° Resultado Alterado por 1° e 2° Resultado Alterado
	1° e 2° Resultado Alterado
	1° Resultado Alterado
	Sexo per 1° Resultado Alterado
	Sexo per Etnia
	Sexo por 1° e 2° Resultado Alterado
	Tempo de coleta(h) per 1° Resultado Alterado
	Tempo de coleta (h) por 1° e 2° Resultado Alterado
	Tempo de Gestação(sem) por 1° e 2° Resultado Alterado
	Tempo de Gestação (sem) per 1° Resultado Alterado
Dados artificiais 2017 - 2021	1° e 2° Resultado Alterado
	1° Resultado Alterado
	Etnia
	Etnia por Tempo de coleta (h)
	Etnia por Nutrição Parental
	Etnia por 1° e 2° Resultado Alterado
	Etnia por Tempo de Gestação(sem)
	Nutrição Parental
	Nutrição Parental por Tempo de coleta(h)
	Nutrição Parental por 1° e 2° Resultado Alterado
	Nutrição Parental por 1° Resultado Alterado
	Peso por Tempo de coleta(h)
	Peso por 1° Resultado Alterado
	Peso por Etnia
	Peso por 1° e 2° Resultado Alterado
	Peso por Tempo de Gestação (sem)
	Sexo por 1° e 2° Resultado Alterado
	Sexo por Etnia
	Sexo por Nutrição Parental
	Sexo por Tempo de Gestação (sem)
	Tempo de coleta (h) por 1° e 2° Resultado Alterado
	Tempo de coleta (h)
	Tempo de Gestação(sem)
	Tempo de Gestação(sem) por Tempo de coleta(h)
	Tempo de Gestação(sem) por 1° e 2° Resultado Alterado

	Tempo de Gestação (sem) por 1° Resultado Alterado

Executando o modelo descrito anteriormente para os modelos gerados a partir do dois banco de dados utilizando os dados descritos acima, foi obtido os mesmos resultados obtidos com o modelo utilizando todos os dados, então foi decido retirar os atributos com pontuação de importância menor que a de nutrição parental por tempo de coleta para o modelo com o banco de dados 2017-2021, obtendo resultados piores que com mais dados, os resultados estão na tabela 29. E para o banco de dados de 2020 temos a tabela 30 que mostra as métricas obtidas, para obtenção foi desconsiderado valores menores que o da pontuação da etnia obtivemos uma melhora na recall que obteve pontuação máxima para o conjunto de teste estendido, mesmo tendo uma piora na métrica de precisão e R2, foi atingido o objetivo de ter a sensibilidade de 100% para o conjunto de dados estendido.

Tabela 29 – Métricas sem atributos com pontuação de importância abaixo de nutrição parental por tempo de coleta gerado a partir de dados artificiais de 2017-2021

Conjunto de dados	Métrica	Score (%)
Conjunto de treino	R2	100
	Acurácia	100
	Precisão	100
	Recall	100
	F1-Score	100
Conjunto de Teste	R2	20
	Acurácia	100
	Precisão	100
	Recall	20
	F1-Score	33.3

Tabela 30 – Métricas sem atributos com pontuação de importância abaixo de etnia gerado a partir de dados artificiais de 2020

Conjunto de dados	Métrica	Score (%)
Conjunto de treino dados 2020	R2	100
	Acurácia	100
	Precisão	100
	Recall	100
	F1-Score	100
Conjunto de Teste dados 2020	R2	66.7
	Acurácia	100
	Precisão	75
	Recall	100
	F1-Score	85.7
Conjunto de teste expandido	R2	33.3
	Acurácia	100
	Precisão	60
	Recall	100
	F1-Score	75

9.4.5 Testes dos modelos sem utilizar o segundo resultado de triagem – M8

Para testar se o modelo apresentaria métricas iguais ou próximas dos testes onde foi considerado o resultado do segundo resultado de exame de triagem neonatal, foram testadas as versões dos modelos M6 de cada banco de dados. Isso foi feito, pois se as métricas atenderem os objetivos proposto do projeto reduzirá em um exame no processo de identificação dos pacientes falsos positivos já que só seria necessário realizar o exame uma única vez e com esse resultado ser capaz de prever se o paciente é um verdadeiro positivo ou não, fornecendo um diagnóstico de maior agilidade, já que precisaria um exame a menos para o encaminhamento para o diagnóstico.

Tabela 31 – Métricas sem dado do segundo exame de triagem gerado a partir de dados artificiais de 2017-2021

Conjunto de dados	Métrica	Score (%)
Conjunto de treino	R2	66.7
	Acurácia	100
	Precisão	100
	Recall	66.7
	F1-Score	80
Conjunto de Teste	R2	-20
	Acurácia	100
	Precisão	0
	Recall	0
	F1-Score	0

Como é possível observar nas Tabelas 31 e 32 as métricas dos modelos sem considerar

Tabela 32 – Métricas sem dado do segundo exame de triagem gerado a partir de dados artificiais de 2020

Conjunto de dados	Métrica	Score (%)
Conjunto de treino dados 2020	R2	75
	Acurácia	100
	Precisão	100
	Recall	75
	F1-Score	85.7
Conjunto de Teste dados 2020	R2	0
	Acurácia	100
	Precisão	0
	Recall	0
	F1-Score	0
Conjunto de teste expandido	R2	0
	Acurácia	100
	Precisão	0
	Recall	0
	F1-Score	0

o dado de segundo exame de triagem para o banco de dados gerado artificialmente teve métricas ruins, com discrepância dos objetivos propostos mostrando que nesse caso é de extrema importância considerar esse exame para se obter uma previsão com menor taxa de erro.

9.5 Comparando e analisando os modelos

Neste capitulo será analisado as melhoras e evolução do modelo com a variação dos parâmetros, e testes realizados nos capítulos anteriores, para facilitar a análise e compreensão da análise na Tabela 33 está o resumo das métricas obtidas para cada um dos modelos. A análise desses dados facilita entender o que levou aos resultados obtidos e o que deve se buscar no banco de dados real para obtenção do melhor modelo. Importante dizer que o modelo M6 não foi considerado na tabela devido ele apresentar as mesmas métricas para o modelo M5 com os dois critérios de avaliação testados.

Tabela 33 – Resumo métricas dos modelos gerado a partir de dados sintéticos

		Base 2020			Base 2017 - 2021	
		Treino	Teste 2020	Teste Ext.	Treino	Teste
M1	R2	100	66.7	46.7	100	40
	Acurácia	100	100	100	100	100
	Precisão	100	75	70.6	100	100
	Recall	100	100	80	100	40
	F1-Score	100	85.7	75	100	57.1
M2	R2	100	66.7	46.7	100	60
	Acurácia	100	100	100	100	100
	Precisão	100	75	70.6	100	100
	Recall	100	100	80	100	60
	F1-Score	100	85.7	75	100	75
M3	R2	100	66.7	53.3	83.3	60
	Acurácia	100	100	100	100	100
	Precisão	100	75	75	91.7	100
	Recall	100	100	80	91.7	60
	F1-Score	100	85.7	77.4	91.7	75
M4	R2	100	66.7	40	91.7	60
	Acurácia	100	100	100	100	100
	Precisão	100	75	63.6	92.3	100
	Recall	100	100	93.3	100	60
	F1-Score	100	85.7	75.7	96	75
M5	R2	100	66.7	53.3	91.7	60
	Acurácia	100	100	100	100	100
	Precisão	100	75	70	92.3	100
	Recall	100	100	93.3	100	60
	F1-Score	100	85.7	80	96	75
M7	R2	100	66.7	33.3	100	20
	Acurácia	100	100	100	100	100
	Precisão	100	75	60	100	100
	Recall	100	100	100	100	20
	F1-Score	100	85.7	75	100	33.3
M8	R2	75	0	0	66.7	-20
	Acurácia	100	100	100	100	100
	Precisão	100	0	0	100	0
	Recall	75	0	0	66.7	0
	F1-Score	85.7	0	0	80	0

Ao analisar o desempenho dos modelos que utilizaram o banco de dados sintético com dados de 2020 para treinamento, observamos que o modelo M1 e M2 obtiveram o mesmo desempenho, independentemente da profundidade máxima de camadas utilizada. Isso indica que o número de camadas do primeiro modelo foi suficiente para alcançar o mesmo resultado obtido com o melhor parâmetro encontrado para obter a melhor métrica de recall no conjunto de teste estendido. A medida que avançamos até o modelo M7, notamos que a recall evoluiu e que a exclusão de alguns dados irrelevantes para o modelo

foi um fator determinante para atingir o objetivo. No entanto, a base de dados com casos positivos é bastante reduzida em comparação com os casos negativos, o que indica que seria ideal ter mais dados para validar o modelo e obter mais parâmetros de avaliação.

Já os modelos criados utilizando os dados sintéticos totais, referentes aos anos de 2017 a 2021 não teve o objetivo atingido ou próximo do recall desejado, o mais próximo que ele chegou foi o score de 60% que foi obtido depois de encontrar o melhor parâmetro de camadas de profundidade das árvores. Depois mesmo variando os outros parâmetros os índices do conjunto testem não se alteraram até a alteração de remover os atributos menos importantes, quando se diminuiu até ter alteração das métricas viu se uma piora na classificação do modelo.

As dificuldades em se obter um modelo com altos índices preditivos, principalmente ao modelo com mais dados, provavelmente se dê por conta da baixa correlação entre os dados. Isso fica mais evidente quando removemos um dos dados com maior grau de correlação que foi o de segundo exame de triagem feito no M8, onde os resultados obtidos foram péssimos. Entretanto uma metodologia eficiente mesmo que com dados com baixo índice de correlação mostrou um efeito evolutivo no poder preditivo do modelo foi idealizada. O ideal será a aplicação desses mesmos testes em dados reais para ver com todos os dados tendo correlação e não como foi feito nestes testes com a grande maioria deles gerados sinteticamente devido à dificuldade de se obtê-los, vendo então como será o poder preditivo do modelo, para assim viabilizar de fato uma ferramenta que auxilie na triagem da fibrose cística.

9.6 Obtenção dos dados reais

Essa foi uma das etapas mais desafiadoras do projeto, num primeiro momento foi estabelecido um contato com o laboratório de triagem neonatal de São Luis do Maranhão, onde eles começaram colaborando com o projeto fornecendo dados genéricos, porem para adquirir os dados reais dos pacientes no momento de obter autorização da diretoria do laboratório para entrar com a solicitação e aprovação do projeto na plataforma Brasil, não se teve retorno, e logo após foi encerrado os contatos com o laboratório.

Com isso alguns outros laboratórios foram contatados, onde foi feita uma apresentação do projeto de pesquisa, entre eles a Fundação Ecumênica de Pais e Expedicionários (FEPE) e Diagnóstico do Brasil (DB) de Curitiba e a APAE de Salvador, este último é o mais promissor onde as conversas estão avançadas e está em processo de negociação ente instituição e a INTERCIENTIFICA, a empresa que está intermediando as negociações e apoiando a pesquisa.

Uma grande dificuldade que se encontra ao obter esses dados com o laboratório são as de um meio automático de se obter as informações dos pacientes, muitos laboratórios

não tem essas informações obtidas de forma automatizada, as vezes até mesmo só existem nos cartões de coleta. Outra informação que muitos laboratórios também acabam não tendo é sobre o exame confirmatório da doença, que muitas vezes são realizados por outros laboratórios.

Inclusive um dos pontos que estão sendo negociados com a APAE de Salvador é a viabilização de um bolsista para auxiliar na catalogação dos dados devido a falta de alguns resultados confirmatórios, e muitas vezes dados faltantes, devido a coleta das amostras ser feita em anotações nos cartões de coleta de amostra pelos municípios e as vezes tem-se dados perdidos ou faltantes.

Os contatos realizados até o momento mostram que a negociação tem grandes chances de acontecer devido a importância e relevância do projeto ainda mais para o laboratório que é um dos maiores centros de referência do país e apresentam uma grande taxa de falsos positivos devido a heterogenia da população e que acaba se tornando um problema devido ao tamanho dos estados o que dificulta muitas vezes a reconvocação de pacientes em municípios distantes, para estarem realizando novos exames.

10 CONCLUSÃO

Dado as pesquisas feitas nesta área tem se a relevância desta dissertação, onde se utiliza de uma tecnologia que está ganhando muito espaço no mercado de diagnósticos, devido ao seu poder de análise de dados e previsões. A doença abordada também é muito relevante devido a sua alta taxa de falsos positivos e a importância do diagnóstico precoce, já que com a triagem teve um grande aumento na sobrevida dos paciente, e conseguindo utilizar da metodologia para diminuir ainda mais o tempo do diagnóstico sem perder a sensibilidade do exame será de grande benefício aos pacientes com FC e aos familiares também, já que estudos apontam efeitos psicossociais negativos a família devido ao estresse causado pela possível doença no recém-nascido, e os exames de confirmação também que acabam gerando um estresse em toda família e especialmente na criança.

Devido à grande dificuldade em se obter dados reais dos pacientes para ter um modelo realista, foi necessário utilizar dos dados que se tinham para validar a teoria proposta e criar uma metodologia para prova de conceito de utilização de modelo de florestas aleatórias, que pudesse ser aplicada aos dados reais quando disponíveis mostrando que é possível utilizar o modelo de aprendizado de máquina para realizar as previsões dos pacientes verdadeiramente positivos.

Os primeiros testes para validar a teoria proposta mostraram que mesmo que com dados sintéticos seria possível realizar previsões dos falsos positivos. Esses resultados mostraram que o modelo de treino conseguiu aprender muito bem com os dados tem acertado todos os casos e no conjunto de teste uma sensibilidade de 100% e uma precisão de 75% em uma base de dados desconhecida. Isso sugere que seria possível fazer as previsões a partir dos atributos atingindo uma sensibilidade máxima e aumentando com uma precisão acima de 50%.

Como o banco de dados utilizado para validar a teoria era um banco de dados relativamente pequeno, foi repetido a aplicação desse mesmo modelo fazendo previsões em um banco de dados maior, onde a incidência de casos positivos era muito maior. Então, o modelo não conseguiu manter o mesmo desempenho apresentado anteriormente. Isso já era esperado, já que comparando o conjunto de dados em que o modelo foi treinado e o que ele era testado era amplamente maior.

Em busca da melhora da performance do modelo, foram feitos testes alterando diversos valores dos parâmetros de criação do modelo de florestas aleatórias, sempre em busca do parâmetro que fornecesse o melhor retorno em relação a *recall*, ou seja era escolhido como parâmetro aquele que gerasse a maior sensibilidade. Nesses testes foram criados dois bancos de dados afim de comparar o desempenho do programa quando treinado

com um banco de dados que representava a parcela somente de um ano e outro em que ele utilizava dados de todos os anos para treinar o modelo. No fim de todos os testes, foi possível obter uma melhora significativa em ambos cenários de treinamento, sendo que os melhores parâmetros foram obtidos a partir do que aprendeu com os dados somente de um ano, atingindo a sensibilidade de 100% e precisão de 60%, quando testado fazendo previsão dos dados de todos os anos.

Embora os resultados obtidos nesses testes utilizando essa metodologia tenham sido extremamente positivos, se faz necessário aplicar essa mesma tecnologia em dados reais, pois provavelmente pela falta de correlação da grande maioria dos dados por terem sido gerados sintéticamente fez com que esse modelo seja enviesado somente a esse banco de dados.

Com um banco de dados real onde todas as informações estão correlacionadas o poder preditivo tende a ser muito maior, levando em consideração que quando se removeu o dado de maior correlação dos testes os resultados pioraram drasticamente. Como ponto positivo, tem-se que se a metodologia foi capaz de trazer melhora em um banco de dados onde o grau de correlação era baixo ela tende a ter resultados muito melhores nos dados reais.

Podendo-se concluir que a pesquisa realizada tem um grande aspecto social podendo trazer maior sobrevida a pacientes com fibrose cística mantendo o desempenho encontrado com dados sintéticos em dados reais, sendo possível utilizar esta prova de conceito em pesquisas relacionadas a aplicação de *Random Forest* na identificação de falsos positivos. Sendo o maior desafio encontrado a obtenção desses dados devido os processos burocráticos e a formação de parcerias com os centros de referência. Os resultados obtido mostram que existem grandes chances de que se obtenha previsões com um alto índice de sensibilidade com dados com maior correlação e tornando possível utilizar do modelo criado para desenvolvimento de um software capaz de auxiliar na triagem a partir do preenchimento de um formulário com os dados do paciente e fornecendo se ele tem maior chance de ser um verdadeiro positivo, possibilitando assertividade e agilidade para os pacientes que tem a maior chance de possuírem a doença.

Tendo como possibilidade de trabalhos futuros a implementação dessa metodologia em dados reais, que pode possibilitar a criação de um software que seja capaz de fazer previsões baseado nas características do pacientes, possibilitando também a utilização dessas mesmas técnicas para fazer previsões para outras doenças que o programa de triagem neonatal abrange e que possam ter outros fatores associados a resultados falsos positivos.

REFERÊNCIAS

AHSAN, M. M.; LUNA, S. A.; SIDDIQUE, Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare*, v. 10, n. 3, p. 541, mar 2022. ISSN 2227-9032. Disponível em: <https://www.mdpi.com/2227-9032/10/3/541>. Citado 2 vezes nas páginas 46 e 48.

AL., P. et. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, v. 12, p. 2825—-2830, 2011. Citado na página 36.

ALI, J. et al. Random Forests and Decision Trees. *IJCSI Int. J. Comput. Sci. Issues*, v. 9, n. 5, p. 272–278, 2012. Citado na página 36.

ALVES, S. P.; BUENO, D. The profile of caregivers to pediatric patients with cystic fibrosis. *Cienc. e Saude Coletiva*, v. 23, n. 5, p. 1451–1457, 2018. ISSN 16784561. Citado na página 17.

ANDERSEN, D. H. CYSTIC FIBROSIS OF THE PANCREAS AND ITS RELATION TO CELIAC DISEASE. *BMJ*, v. 56, n. 2, p. 344–399, apr 1938. ISSN 0959-8138. Disponível em: <https://www.bmj.com/lookup/doi/10.1136/bmj.1.2988.413-a>. Citado na página 6.

ARRUDI-MORENO, M. et al. Cribado neonatal de fibrosis quística: análisis y diferencias de los niveles de tripsina inmunorreactiva en recién nacidos con cribado positivo. *An. Pediatría*, Asociación Española de Pediatría, v. 95, n. 1, p. 11–17, jul 2021. ISSN 16954033. Disponível em: <https://doi.org/10.1016/j.anpedi.2020.04.029https://linkinghub.elsevier.com/retrieve/pii/S1695403320302253>. Citado na página 26.

ATHANAZIO, R. A. et al. Brazilian guidelines for the diagnosis and treatment of cystic fibrosis. *J. Bras. Pneumol.*, v. 43, n. 3, p. 219–245, 2017. ISSN 18063756. Citado 2 vezes nas páginas 14 e 17.

AVILA, S.; CANTERO, L.; FRANCISCO, S. Vieses no Aprendizado de Máquina e suas Implicações Sociais: Um Estudo de Caso no Reconhecimento Facial. *Work. SOBRE AS IMPLICAÇÕES DA Comput. NA Soc.*, v. 2, p. 90–101, 2021. Citado na página 39.

BARBOSA, J. M. et al. Métodos de Classificação por Árvores de Decisão. *Programa Pós-Graduação em Ciência da Comput.*, p. 5, 2012. Citado na página 34.

BASHIR, D. et al. An Information-Theoretic Perspective on Overfitting and Underfitting. *33rd Australas. Jt. Conf. Artif. Intell.*, v. 2, n. 6, p. 1–15, 2020. Disponível em: <https://arxiv.org/abs/2010.06076>. Citado na página 41.

BELL, S. C. et al. The future of cystic fibrosis care: a global perspective. *Lancet. Respir. Med. Comm.*, p. 1–60, jan 2019. ISSN 2213-2619. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S2213260019303376http://www.ncbi.nlm.nih.gov/pubmed/31570318>. Citado 2 vezes nas páginas 12 e 17.

BIAU, G. Analysis of a Random Forests Model. *J. Mach. Learn. Res.*, v. 13, p. 1063–1095, 2012. Citado na página 35.

BIRÓ, T. S.; NÉDA, Z. Gintropy : Gini Index Based Generalization of Entropy. *Entropy*, v. 2, n. 1, p. 1–13, 2020. Citado na página 34.

BONETTO, R.; LATZKO, V. Machine learning. In: *Comput. Commun. Networks*. Elsevier Inc., 2021. cap. 8, p. 135–167. ISBN 9780128204887. Disponível em: <https://doi.org/10.1016/B978-0-12-820488-7.00021-9>. Citado 2 vezes nas páginas 31 e 34.

BONFIM, I. M. et al. Perfil dos pacientes com fibrose cística atendidos no centro de referência pediátrico do Espírito Santo. *Rev. Bras. Pesqui. em Saúde/Brazilian J. Heal. Res.*, v. 21, n. 1, p. 80–85, jul 2019. ISSN 2446-5410. Disponível em: <http://periodicos.ufes.br/rbps/article/view/26471>. Citado na página 3.

BOROWITZ, D. et al. Cystic fibrosis Foundation evidence-based guidelines for management of infants with cystic fibrosis. *J. Pediatr.*, Mosby, Inc., v. 155, n. 6 SUPPL., p. S73–S93, 2009. ISSN 10976833. Disponível em: <http://dx.doi.org/10.1016/j.jpeds.2009.09.001>. Citado na página 8.

BOUCHER, R. C. et al. Na+ transport in cystic fibrosis respiratory epithelia. Abnormal basal rate and response to adenylate cyclase activation. *J. Clin. Invest.*, v. 78, n. 5, p. 1245–1252, nov 1986. ISSN 0021-9738. Disponível em: <http://www.jci.org/articles/view/112708>. Citado na página 7.

BREIMAN, L. E. O. *Random Forests*. [S.l.: s.n.], 2001. v. 45. 5–32 p. Citado na página 34.

BROCKOW, I.; NENNSTIEL, U. Newborn screening for cystic fibrosis. *Padiatr. Prax.*, v. 90, n. 4, p. 559–567, 2018. ISSN 21981698. Citado 2 vezes nas páginas 26 e 27.

CABELLO, G. M. et al. Rastreamento da fibrose cística usando-se a análise combinada do teste de IRT neonatal e o estudo molecular da mutação deltaF508. *J. Bras. Patol. e Med. Lab.*, v. 39, n. 1, 2003. ISSN 1676-2444. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1676-24442003000100004{&}lng=pt{&}nrm>. Citado 2 vezes nas páginas 1 e 27.

CABELLO, G. M. K. Avanços da Genética na Fibrose Cística. *Rev. Hosp. Univ. Pedro Ernesto, UERJ*, Rio de Janeiro, p. 36–45, dec 2011. Citado na página 19.

CFMD. *Cystic Fibrosis Mutation Database*. 2022. Disponível em: <http://www.genet.sickkids.on.ca/cftr/Home.html>. Citado na página 19.

CHEILLAN, D. et al. False-positive results in neonatal screening for cystic fibrosis based on a three-stage protocol (IRT/DNA/IRT): Should we adjust IRT cut-off to ethnic origin? *J. Inherit. Metab. Dis.*, v. 28, n. 6, p. 813–818, dec 2005. ISSN 0141-8955. Disponível em: <http://doi.wiley.com/10.1007/s10545-005-0067-0>. Citado 2 vezes nas páginas 3 e 26.

CHILD, A. D. Neonatal screening for cystic fibrosis. *Arch. Dis. Child.*, v. 84, n. 5, p. 449–449, may 2001. ISSN 00039888. Disponível em: <https://adc.bmj.com/lookup/doi/10.1136/adc.84.5.449>. Citado na página 19.

COAKLEY, J. et al. Sweat testing for cystic fibrosis: Standards of performance in Australasia. *Ann. Clin. Biochem.*, v. 46, n. 4, p. 332–337, 2009. ISSN 00045632. Citado na página 15.

COLLINS, F. S. et al. Identification of the Cystic Fibrosis Gene: Chromosome Walking and Jumping. *Science (80-.).*, v. 245, n. 4922, p. 1059–1065, sep 1989. ISSN 0036-8075. Disponível em: <https://www.science.org/doi/10.1126/science.2772657>. Citado na página 10.

CROSSLEY, J. R.; ELLIOTT, R. B.; SMITH, P. A. DRIED-BLOOD SPOT SCREENING FOR CYSTIC FIBROSIS IN THE NEWBORN. *Lancet*, p. 472–474, 1979. Citado na página 7.

CUNNINGHAM, J. C.; TAUSSIG, L. M. *An Introduction to Cystic Fibrosis for Patients and Their Families*. 6°. ed. Bethesda: Apitalis, 2013. 25–27 p. Citado 2 vezes nas páginas 6 e 8.

DALCIN, P. D. T. R.; SILVA, F. A. D. A. Fibrose cística no adulto: Aspectos diagnósticos e terapêuticos. *J. Bras. Pneumol.*, v. 34, n. 2, p. 107–117, 2008. ISSN 18063713. Citado na página 16.

DAMASCENO, N. *Triagem neonatal para fibrose cística*. 2010. 6 p. Disponível em: <https://www.spsp.org.br/2010/03/15/triagem{_}neonatal{_}para{_}fibrose>. Citado na página 14.

De Almeida Matos, B.; MARTINS, R. C. Fibrose Cística: Uma Revisão De Literatura Cystic Fibrisis: a Literature Review. *Brazilian J. Surg. Clin. Res.*, v. 29, n. 2, p. 114–119, 2019. ISSN 2317-4404. Disponível em: <http://www.mastereditora.com.br/bjscr>. Citado 2 vezes nas páginas 7 e 8.

DEVELOPERS, G. *Classification : ROC Curve and AUC*. 2020. Disponível em: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>. Citado na página 45.

DEY, V. *Understanding the AUC-ROC Curve in Machine Learning Classification*. 2021. Disponível em: <https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/>. Citado na página 45.

DI SANT'AGNESE, P. A. et al. ABNORMAL ELECTROLYTE COMPOSITION OF SWEAT IN CYSTIC FIBROSIS OF THE PANCREAS. *Pediatrics*, v. 12, n. 5, p. 549–563, nov 1953. ISSN 0031-4005. Disponível em: <https://publications.aap.org/pediatrics/article/12/5/549/39262/ABNORMAL-ELECTROLYTE-COMPOSITION-OF-SWEAT-IN>. Citado na página 7.

EL, I. et al. Machine learning and modeling: Data, validation, communication challenges. *Med Phys*, v. 45, n. 10, p. 1–7, 2018. Citado na página 39.

FAGGELLA, D. *Machine Learning for Medical Diagnostics – 4 Current Applications*. 2020. Disponível em: <https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current-applications/>. Citado na página 45.

FARBER, S.; SHWACHMAN, H.; MADDOCK, C. L. PANCREATIC FUNCTION AND DISEASE IN EARLY LIFE. I. PANCREATIC ENZYME ACTIVITY AND THE CELIAC SYNDROME 1. *J. Clin. Invest.*, v. 22, n. 6, p. 827–838, nov 1943. ISSN

0021-9738. Disponível em: <http://www.jci.org/articles/view/101456>. Citado na página 6.

FARRELL, P. M. et al. Nutritional Benefits of Neonatal Screening for Cystic Fibrosis. *N. Engl. J. Med.*, v. 337, n. 14, p. 963–969, oct 1997. ISSN 0028-4793. Disponível em: <http://www.nejm.org/doi/abs/10.1056/NEJM199710023371403>. Citado na página 1.

FIRMIDA, M. D. C.; MARQUES, B. L.; COSTA, C. H. da. Fisiopatologia e Manifestações Clínicas da Fibrose Cística. *Rev. Hosp. Univ. Pedro Ernesto*, v. 10, n. 4, p. 46–58, 2011. ISSN 1983-2567. Citado na página 10.

FOUNDATION, C. F. *A CFTR Mutation Classes*. [S.l.], 2017. 2 p. Citado na página 25.

FOUNDATION, C. F. *Antibiotics*. 2022. Disponível em: <https://www.cff.org/managing-cf/antibiotics>. Citado 2 vezes nas páginas 17 e 19.

FOUNDATION, C. F. *Bronchodilators*. 2022. Disponível em: <https://www.cff.org/bronchodilators>. Citado na página 17.

FOUNDATION, C. F. *CFTR Modulator Therapies*. 2022. Disponível em: <https://www.cff.org/managing-cf/cftr-modulator-therapies>. Citado na página 17.

FOUNDATION, C. F. *Our History*. 2022. Disponível em: <https://www.cff.org/about-us/our-history>. Citado na página 7.

FOUNDATION, C. F. *Vascular Access Devices : PICCs and Ports*. 2022. Disponível em: <https://www.cff.org/vascular-access-devices-piccs-and-ports>. Citado na página 17.

FRÍAS, J. P. et al. The History of Cystic Fibrosis. *Open J. Pediatr. Child Heal.*, v. 4, n. 1, p. 001–006, mar 2019. ISSN 26407612. Disponível em: <http://dx.doi.org/10.17352/ojpch.000015https://www.peertechz.com/articles/OJPCH-4-115.php>. Citado na página 6.

FURTADO, M. C. d. C.; LIMA, R. A. G. de. O cotidiano da família com filhos portadores de fibrose cística: subsídios para a enfermagen pediátrica. *Rev. Lat. Am. Enfermagem*, v. 11, n. 1, p. 66–73, 2003. ISSN 01041169. Citado na página 11.

GÉRON, A. *Mãos à Obra : Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Rio de Janeiro: Copyright © 2019 da Starlin Alta Editora e Consultoria Eireli, 2019. 577 p. ISBN 9788550809021. Citado 8 vezes nas páginas 31, 32, 33, 39, 40, 42, 44 e 52.

GIBSON, L. E.; COOKE, R. E. A TEST FOR CONCENTRATION OF ELECTROLYTES IN SWEAT IN CYSTIC FIBROSIS OF THE PANCREAS UTILIZING PILOCARPINE BY IONTOPHORESIS. *Pediatrics*, v. 23, n. 3, p. 545–549, mar 1959. ISSN 0031-4005. Disponível em: <https://publications.aap.org/pediatrics/article/23/3/545/29475/A-TEST-FOR-CONCENTRATION-OF-ELECTROLYTES-IN-SWEAT>. Citado na página 7.

GUIDO, S.; MULLER, A. C. *Introduction to Machine Learning with Python*. Oreilly. Sebastopol - CA: [s.n.], 2016. 392 p. ISBN 9781449369415. Citado 3 vezes nas páginas 30, 34 e 41.

HARRISON, M. *Machine Learning - Guia de Referência Rápida*. São Paulo: Novaltec Editora Ltda., 2020. 286 p. ISBN 9781492047544. Citado 2 vezes nas páginas 33 e 42.

HASIAK, A.; VICENTE, L.; FERREIRA, R. Tendências de mortalidade relacionada à fibrose cística no Brasil no período de 1999 a 2017 : um estudo de causas múltiplas de morte. *J Bras Pneumol. 2021;47(2)e20200166*, v. 47, n. 2, p. 1–8, 2021. Citado na página 13.

HAYEEMS, R. Z. et al. Parent experience with false-positive newborn screening results for cystic fibrosis. *Pediatrics*, v. 138, n. 3, 2016. ISSN 10984275. Citado na página 3.

HAYEEMS, R. Z. et al. False-positive newborn screening for cystic fibrosis and health care use. *Pediatrics*, v. 140, n. 5, 2017. ISSN 10984275. Citado na página 1.

HICKS, S. A. et al. On evaluation metrics for medical applications of artificial intelligence. *medRxiv - Prepr. Serv. Heal. Sci.*, p. 1–10, 2021. Citado na página 43.

HUILGOL, P. *Precision vs. Recall – An Intuitive Guide for Every Machine Learning Person*. 2020. Disponível em: <https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/>. Citado 3 vezes nas páginas 42, 44 e 45.

INC., B. *Types of CFTR Mutations*. 2021. Disponível em: <https://cysticfibrosisnewstoday.com/types-of-cftr-mutations/>. Citado na página 25.

James Thorn. *Random Forest: Hyperparameters and how to fine-tune them*. 2020. Disponível em: <https://towardsdatascience.com/random-forest-hyperparameters-and-how-to-fine-tune-them-17aee785ee0d>. Citado na página 38.

JASKARI, J. et al. Machine Learning Methods for Neonatal Mortality and Morbidity Classification. *IEEE Access*, Institute of Electrical and Electronics Engineers Inc., v. 8, p. 123347–123358, 2020. ISSN 2169-3536. Disponível em: <https://ieeexplore.ieee.org/document/9131772/>. Citado na página 43.

KNOWLES, M. R. et al. Abnormal Ion Permeation Through Cystic Fibrosis Respiratory Epithelium. *Science (80-.).*, v. 221, n. 4615, p. 1067–1070, sep 1983. ISSN 0036-8075. Disponível em: <https://www.science.org/doi/10.1126/science.6308769>. Citado na página 7.

KUMAR, U. Applications of Machine Learning in Disease Pre-screening. In: *Adv. Med. Diagnosis, Treat. Care B.* Hershey PA, USA: IGI Global, 2019. cap. 10, p. 278–320. ISBN 9781522571315. Disponível em: <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-5225-7131-5.ch010>. Citado na página 46.

KUMAR, Y. et al. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J. Ambient Intell. Humaniz. Comput.*, Springer Berlin Heidelberg, 2022. ISSN 18685145. Disponível em: <https://doi.org/10.1007/s12652-021-03612-z>. Citado na página 46.

KWON, C.; FARRELL, P. M. The magnitude and challenge of false-positive newborn screening test results. *Arch. Pediatr. Adolesc. Med.*, v. 154, n. 7, p. 714–718, 2000. ISSN 10724710. Citado na página 1.

KYNKÄÄNNIEMI, T. et al. Improved Precision and Recall Metric for Assessing Generative Models. *Adv. Neural Inf. Process. Syst.*, v. 32, n. NeurIPS, apr 2019. ISSN 10495258. Disponível em: <http://arxiv.org/abs/1904.06991>. Citado na página 43.

LUDERMIR, T. B. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. *Estud. AVANÇADOS*, v. 35, n. 101, p. 85–94, 2021. Citado na página 38.

LUMERTZ, M. S. et al. False-negative newborn screening result for immunoreactive trypsinogen: A major problem in children with chronic lung disease. *J. Bras. Pneumol.*, v. 45, n. 3, p. 3–4, 2019. ISSN 18063756. Citado 2 vezes nas páginas 1 e 27.

LYCZAK, J. B.; CANNON, C. L.; PIER, G. B. Lung Infections Associated with Cystic Fibrosis. *Clin. Microbiol. Rev.*, v. 15, n. 2, p. 194–222, apr 2002. ISSN 0893-8512. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC118069/https://journals.asm.org/doi/10.1128/CMR.15.2.194-222.2002>. Citado na página 8.

M, H.; M.N, S. A Review on Evaluation Metrics for Data Classification Evaluations. *Int. J. Data Min. Knowl. Manag. Process*, v. 5, n. 2, p. 01–11, mar 2015. ISSN 2231007X. Disponível em: <http://www.aircconline.com/ijdkp/V5N2/5215ijdkp01.pdf>. Citado 2 vezes nas páginas 42 e 45.

MANGEL, M.; CRUZ, S.; SAMANIEGO, F. J. Abraham Wald ' s Work on Aircraft Survivability. *J. Am. Stat. Assoc.*, v. 79, n. 386, p. 259–267, 1984. Citado na página 40.

MARIANO, T.; CONDE, C. R. CHILD TO THE NURSE ' S ASSISTANCE WITH CYSTIC FIBROSIS. *Rev. UNINGÁ*, Maringa, v. 52, p. 144–150, 2017. Citado na página 9.

MISHRA, A.; GREAVES, R.; MASSIE, J. The relevance of sweat testing for the diagnosis of cystic fibrosis in the genomic era. *Clin. Biochem. Rev.*, v. 26, n. 4, p. 135–53, nov 2005. ISSN 0159-8090. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16648884http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1320177>. Citado na página 14.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill Science, 1997. 432 p. ISBN 9780071154673. Disponível em: <https://books.google.com.br/books?id=EoYBngEACAAJ>. Citado na página 30.

MORAN, J. et al. Newborn screening for CF in a regional paediatric centre: The psychosocial effects of false-positive IRT results on parents. *J. Cyst. Fibros.*, v. 6, n. 3, p. 250–254, 2007. ISSN 15691993. Citado na página 28.

MOTA, L. R. et al. Description of rare mutations and a novel variant in Brazilian patients with Cystic Fibrosis: a case series from a referral center in the Bahia State. *Mol. Biol. Rep.*, Springer Netherlands, v. 45, n. 6, p. 2045–2051, 2018. ISSN 15734978. Disponível em: <http://dx.doi.org/10.1007/s11033-018-4361-y>. Citado na página 11.

NARKHEDE, S. *Understanding AUC - ROC Curve*. 2018. Disponível em: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>. Citado na página 45.

PALEYES, A.; URMA, R.-G.; LAWRENCE, N. D. Challenges in Deploying Machine Learning : a Survey of Case Studies. *ACM Comput. Surv*, 2022. Citado na página 39.

P.BRADLEY, A. THE USE OF THE AREA UNDER THE ROC CURVE IN THE EVALUATION OF MACHINE LEARNING ALGORITHMS. *Pattern Recognit.*, v. 30, n. 7, p. 1145–1159, 1997. Citado na página 45.

PENG, G. et al. Reducing false-positive results in newborn screening using machine learning. *Int. J. Neonatal Screen.*, MDPI Multidisciplinary Digital Publishing Institute, v. 6, n. 1, 2020. ISSN 2409515X. Citado 5 vezes nas páginas 2, 3, 48, 51 e 54.

POWERS, D. EVALUATION : FROM PRECISION , RECALL AND F-MEASURE TO ROC , INFORMEDNESS , MARKEDNESS & CORRELATION. *Int. J. Mach. Learn. Technol.*, v. 2, n. 1, p. 37–63, 2011. Citado 2 vezes nas páginas 43 e 44.

PROESMANS, M. Best practices in the treatment of early cystic fibrosis lung disease. *Ther. Adv. Respir. Dis.*, v. 11, n. 2, p. 97–104, 2017. ISSN 17534666. Citado na página 16.

QUINTON, P. M. Chloride impermeability in cystic fibrosis. *Nature*, v. 301, n. 5899, p. 421–422, feb 1983. ISSN 0028-0836. Disponível em: <http://www.nature.com/articles/301421a0>. Citado na página 7.

RAM, S. *Dominando florestas aleatórias: um guia completo*. 2020. Citado na página 38.

RIBEIRO, J. D.; RIBEIRO, M. Â. G. D. O.; RIBEIRO, A. F. Controvérsias na fibrose cística – do pediatra ao especialista Controversies in cystic fibrosis – from pediatrician to specialist. v. 78, p. 171–186, 2002. Citado na página 16.

RIBEIRO, M. N. A. et al. Fibrose cística: histórico e principais meios para diagnóstico. *Res. Soc. Dev.*, Research, Society and Development, v. 10, n. 3, p. e11710313075, mar 2021. ISSN 2525-3409. Citado 2 vezes nas páginas 1 e 13.

Ribeiro ROSA, F. et al. Cystic fibrosis: a clinical and nutritional approach. *Rev. Nutr*, Campinas, v. 21, n. 6, p. 725–737, 2008. Citado 3 vezes nas páginas 1, 8 e 16.

RIORDAN, J. R. CFTR function and prospects for therapy. *Annu. Rev. Biochem.*, v. 77, p. 701–726, 2008. ISSN 00664154. Citado na página 8.

ROCK, M. J.; MAKHOLM, L.; EICKHOFF, J. A new method of sweat testing: The CF Quantum®sweat test. *J. Cyst. Fibros.*, European Cystic Fibrosis Society., v. 13, n. 5, p. 520–527, 2014. ISSN 18735010. Disponível em: <http://dx.doi.org/10.1016/j.jcf.2014.05.001>. Citado na página 15.

ROCK, M. J. et al. Immunoreactive trypsinogen screening for cystic fibrosis: Characterization of infants with a false-positive screening test. *Pediatr. Pulmonol.*, v. 6, n. 1, p. 42–48, 1989. ISSN 10990496. Citado na página 26.

ROSENSTEIN, B. J. *Fibrose cística*. 2019. Disponível em: <https://www.msdmanuals.com/pt/profissional/pediatria/fibrose-c{í}stica-fc/fibrose-c>. Citado 2 vezes nas páginas 12 e 16.

SAJDA, P. Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.*, v. 8, p. 537–565, 2006. ISSN 15239829. Citado na página 46.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.*, v. 3, n. 3, p. 210–229, 1959. Citado na página 30.

SANTOS, G. P. C. et al. Programa de triagem neonatal para fibrose cística no estado do Paraná: avaliação após 30 meses de sua implantação. *J. Pediatr. (Rio. J).*, v. 81, n. 3, p. 240–244, jun 2005. ISSN 0021-7557. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S0021-75572005000400011{&}lng=pt{&}nrm>. Citado na página 9.

SARICA, A.; CERASA, A.; QUATTRONE, A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer ' s Disease : A Systematic Review. *Front. Aging Neurosci*, v. 9, n. October, p. 1–12, 2017. Citado na página 36.

SARKER, I. H. Machine Learning : Algorithms, Real - World Applications and Research Directions. *SN Comput. Sci.*, Springer Singapore, v. 2, n. 3, p. 1–21, 2021. ISSN 2661-8907. Disponível em: <https://doi.org/10.1007/s42979-021-00592-x>. Citado na página 33.

SAUD, A. S.; SHAKYA, S.; NEUPANE, B. Analysis of Depth of Entropy and GINI Index Based Decision Trees for Predicting Diabetes. *Indian J. Comput. Sci.*, v. 6, n. 6, p. 19–28, 2021. Citado na página 34.

SERVICES, A. W. *Amazon Machine Learning - Developer Guide*. 2022. Disponível em: <https://docs.aws.amazon.com/machine-learning/latest/dg/machinelearning-dg.pdf{\#}model-fit-underfitting-vs-overfitt>. Citado 2 vezes nas páginas 40 e 41.

SERVIDONI, M. et al. Sweat test and cystic fibrosis: Overview of test performance at public and private centers in the state of São Paulo, Brazil | Teste do suor e fibrose cística: Panorama da realização do teste em centros públicos e privados do estado de São Paulo. *J. Bras. Pneumol.*, v. 43, n. 2, p. 121–128, 2017. ISSN 18063756. Citado na página 16.

SHWACHMAN, H.; ANTONOWICZ, I. THE SWEAT TEST IN CYSTIC FIBROSIS. *Ann. N. Y. Acad. Sci.*, v. 93, n. 12, p. 600–624, aug 1962. ISSN 00778923. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1111/j.1749-6632.1962.tb30495.x>. Citado na página 7.

SILVA, A. G. et al. HISTÓRICO , AVANÇOS NO DIAGNÓSTICO E TRATAMENTO DA FIBROSE CÍSTICA. 2015. Citado na página 14.

SIMIC, M. *Os efeitos da profundidade e do número de árvores em uma floresta aleatória*. 2022. Disponível em: <https://www.baeldung.com/cs/random-forest-tuning>. Citado na página 38.

SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond Accuracy , F-Score and ROC : A Family of Discriminant Measures for Performance Evaluation. *Adv. Artif. Intell.* ·, v. 4304, p. 1015–1021, 2006. Citado na página 44.

TACCETTI, G. et al. Clinical and genotypical features of false-negative patients in 26 years of cystic fibrosis neonatal screening in Tuscany, Italy. *Diagnostics*, v. 10, n. 7, p. 1–10, 2020. ISSN 20754418. Citado na página 27.

Tavish Srivastava. *Tuning the parameters of your Random Forest model*. 2015. 3–7 p. Citado na página 38.

TONELLO, M. L. et al. Discrepâncias Entre Os Registros De Prontuários Acerca Da Farmacoterapia De Pacientes Pediátricos Com Fibrose Cística. *Clin. Biomed. Res.*, v. 37, n. 3, p. 181–186, 2017. Citado na página 16.

TRAVERT, G.; HEELEY, M.; HEELEY, A. History of Newborn Screening for Cystic Fibrosis — The Early Years. *Int. J. Neonatal Screen.*, v. 6, n. 8, p. 1–7, 2020. Citado na página 17.

TRIPOLITI, E. E.; FOTIADIS, D. I.; MANIS, G. Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm. *IEEE Trans. Inf. Technol. Biomed.*, v. 16, n. 4, p. 615–622, jul 2012. ISSN 1089-7771. Disponível em: <http://ieeexplore.ieee.org/document/6080732/>. Citado na página 48.

TSUI, L.-c. et al. Identification of the Cystic Fibrosis Gene: Genetic Analysis. *Science (80-.).*, v. 245, n. 4922, p. 1073–1080, sep 1989. ISSN 0036-8075. Disponível em: <https://www.science.org/doi/10.1126/science.2570460>. Citado na página 10.

TSUI, L.-C. et al. Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA. *Science (80-.).*, v. 245, n. 4922, p. 1066–1073, sep 1989. ISSN 0036-8075. Disponível em: <https://www.science.org/doi/10.1126/science.2475911>. Citado 2 vezes nas páginas 8 e 10.

TU, W.-J. et al. Psychological Effects of False-Positive Results in Expanded Newborn Screening in China. *PLoS One*, v. 7, n. 4, p. e36235, apr 2012. ISSN 1932-6203. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0036235>. Citado na página 28.

VENDRUSCULO, F. M.; Fagundes Donadio, M. V.; Araújo Pinto, L. Cystic fibrosis in Brazil: achievements in survival. *J. Bras. Pneumol.*, v. 47, n. 2, p. e20210140, apr 2021. ISSN 1806-3756. Disponível em: <http://www.jornaldepneumologia.com.br/details/3517/en-US/cystic-fibrosis-in-brazil--achievements-in-survival>. Citado na página 12.

VLACHAS, C. et al. Random forest classification algorithm for medical industry data. *SHS Web Conf.*, v. 139, n. 03008, p. 1–6, may 2022. ISSN 2261-2424. Disponível em: <https://www.shs-conferences.org/10.1051/shsconf/202213903008>. Citado na página 48.

W. Edgar, T.; O. Manz, D. Machine Learning. In: *Res. Methods Cyber Secur.* [S.l.]: Syngress, 2017. cap. 6, p. 153–173. ISBN 9780128053492. Citado na página 30.

WALLIS, C. Diagnosing cystic fibrosis: blood, sweat, and tears. *Arch. Dis. Child.*, v. 76, n. 2, p. 85–88, feb 1997. ISSN 0003-9888. Disponível em: <https://www.bmj.com/lookup/doi/10.1136/bmj.1.3398.303-ahttps://adc.bmj.com/lookup/doi/10.1136/adc.76.2.85>. Citado na página 14.

WELSH, M. J.; SMITH, A. E. Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis. *Cell*, v. 73, n. 7, p. 1251–1254, jul 1993. ISSN 00928674. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/009286749390353R>. Citado na página 11.

ZHOU, L. et al. Machine Learning on Big Data : Opportunities and Challenges. *Neurocomputing*, v. 237, p. 350–361, 2017. Citado na página 40.