UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
INSTITUTO DE QUÍMICA DE SÃO CARLOS

ANA CLARA DE ANDRADE MIOTO

Estudo e aplicações de técnicas de aprendizado de máquina na análise de desfechos inesperados de tuberculose

São Carlos
2023

ANA CLARA DE ANDRADE MIOTO

Estudo e aplicações de técnicas de aprendizado de máquina na análise de desfechos inesperados de tuberculose

Dissertação apresentada ao Programa de Pós-Graduação Interunidades em Bioengenharia da Escola de Engenharia de São Carlos – Faculdade de Medicina de Ribeirão Preto e Instituto de Química de São Carlos da Universidade de São Paulo, como requisito para a obtenção do Título de Mestra em Ciências.

Orientador: Prof. Dr. Domingos Alves

VERSÃO CORRIGIDA

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Mioto, Ana Clara de Andrade
M669e Estudo e aplicações de técnicas de aprendizado de máquina na análise de desfechos inesperados de tuberculose / Ana Clara de Andrade Mioto; orientador Domingos Alves. São Carlos, 2023.

Dissertação (Mestrado) - Programa de Pós-Graduação Interunidades em Bioengenharia e Área de Concentração em Bioengenharia -- Escola de Engenharia de São Carlos; Faculdade de Medicina de Ribeirão Preto; Instituto de Química de São Carlos, da Universidade de São Paulo, 2023.

1. Tuberculose. 2. Machine Learning. 3. Desfechos Ruins. 4. Saúde Pública. 5. Clustering. 6. Ensemble Learning. 7. Automated Learning. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE JULGAMENTO

Candidato(a): Ana Clara de Andrade Mioto

Título: "Estudo e aplicações de técnicas de aprendizado de máquina na análise de desfechos inesperados de tuberculose"

Data da defesa:01/11/2023

Comissão Julgadora	Resultado
Prof(a). Dr(a). Domingos Alves Faculdade de Medicina de Ribeirão Preto - FMRP/USP - **Orientador**	Não Votante
Prof(a). Dr(a). Antonio Ruffino Netto Faculdade de Medicina de Ribeirão Preto - FMRP/USP	Aprovada
Dr(a). Mariane Barros Neiva Rede Nacional de Doenças Raras	Aprovada
Prof(a). Dr(a). Rafael Mello Galliez Universidade Federal do Rio de Janeiro/UFRJ	Aprovada

Prof. Dr. Domingos Alves - Orientador

Prof. Dr. Adair Roberto Aguiar - Presidente da Comissão de Pós-Graduação:

AGRADECIMENTOS

Antes de tudo agradeço a Deus, por ter me guiado e iluminado por todo o caminho percorrido no desenvolvimento deste projeto. E me dá forças para superar todas as dificuldades.

Aos meus pais Daniela e Carlos, obrigada por tudo. Não há palavras suficientes para dizer como são importantes para mim. São e sempre serão meu alicerce na vida e minha maior fonte de inspiração.

A minha irmã Rebecca, por sempre estar comigo em todos os momentos e obrigada por acreditar em mim.

Ao meu marido Fabrício, por ser meu confidente, me aguentar nos momentos de desabafo e por participar de todas minhas conquistas, comemorando junto comigo. Além disso, obrigada por me ajudar sempre e estar presente.

Ao Prof. Dr. Domingos Alves por me receber de braços abertos para orientação, pela ajuda, aprendizados compartilhados e por dar todo o suporte neste projeto.

Ao Filipe Andrade Bernardi pela paciência em sanar todas as minhas dúvidas, ter disponibilidade e por toda a dedicação quando necessário durante o desenvolvimento deste trabalho.

Obrigada à toda equipe do Laboratório de Inteligência em Saúde (LIS) por todo o carinho e conhecimento adquirido e compartilhado no período de mestrado, além das amizades criadas, em especial à Isabelle Carvalho, Mariane Neiva, Vinicius Costa Lima, Victor Cassão, Giovane Thomazini, Renan Barbieri, Pedro Emilio Andrade, Mariana Mozini.

Obrigada à todos os meus amigos e amigas de fora, em especial ao Guilherme Gomes, Karin Targas, Jenifer Vieira, Cinthia Lavanhini, Raphael Mariano, pelo companheirismo e carinho, pelos almoços, risadas, choros e os desesperos juntos, mas principalmente pela amizade neste último ano de projeto.

À todos os meus familiares e familiares do meu marido, em especial a Denise e Dario (meus tios), que sempre apoiaram e desejaram meu sucesso.

Aos Profs. Drs. Antonio Ruffino Netto e Rafael Galliez, que deram suporte essencial durante o desenvolvimento deste projeto.

Á Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo suporte financeiro ao desenvolvimento deste projeto, sob o número 2021/01961-0, que tornou possível a experiência, conhecimento e oportunidade de publicar e apresentar diversos artigos dos resultados aqui obtidos.

RESUMO

MIOTO, A. C. A. **Estudo e aplicações de técnicas de aprendizado de máquina em desfechos inesperados de tuberculose.** 2023. Dissertação (Mestrado) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2023.

A tuberculose (TB) continua sendo uma das doenças infecciosas mais mortais globalmente, com milhões de casos e mortes relatados a cada ano. Este problema é agravado quando associado a comorbidades, como o HIV, tornando-se ainda mais letal. Além disso, fatores socioeconômicos e culturais desempenham um papel importante na prevalência da TB, indicando uma estreita ligação entre a doença e o desenvolvimento social precário. Com o Brasil, sendo um país significantemente afetado pela Tuberculose, vem trabalhando em ações e tratamentos que possam ser implantados para o controle e prevenção da TB e redução da vulnerabilidade dos pacientes. Um aspecto crucial para realização destas intervenções, é a disponibilidade de dados de saúde abrangentes e a aplicação de técnicas de análise de dados, como o aprendizado de máquina (AM), para melhorar a qualidade do atendimento e as decisões médicas. Inclusive, estudos têm mostrado que o AM é uma área emergente na saúde, pois pode aprender com dados históricos e identificar padrões que podem levar a evitar um desfecho inesperado no tratamento da TB, como o abandono do tratamento, óbito e resistência medicamentosa. Neste contexto, esta pesquisa visa utilizar técnicas de descoberta de conhecimento em bases de dados (KDD) e aprendizado de máquina para analisar e identificar padrões desconhecidos que possam relacionar fatores sociodemográficos e clínicos e a probabilidade de um certo desfecho negativo do tratamento da TB ocorrer com um paciente. Além disso, a crescente disponibilidade de dados de pacientes no campo da saúde torna o uso de técnicas como o AM ainda mais relevante para melhorar o manejo dos pacientes com TB.

Palavras-chave: Tuberculose; KDD; Aprendizado de Máquina; Desfechos inesperados.

ABSTRACT

MIOTO, A. C. A. **Study and applications of machine learning techniques in tuberculosis bad outcomes.** 2023. Dissertation (Master's degree) – São Carlos School of Engineering, University of São Paulo, 2023.

Tuberculosis (TB) continues to be one of the deadliest infectious diseases globally, with millions of cases and deaths reported every year. This problem is exacerbated when associated with comorbidities such as HIV, making it even more lethal. Furthermore, socioeconomic and cultural factors play a significant role in the prevalence of TB, indicating a close link between the disease and poor social development. Brazil, significantly affected by Tuberculosis, has been working on actions and treatments that can be implemented for TB control and prevention, as well as reducing patient vulnerability. A crucial aspect for the implementation of these interventions is the availability of comprehensive health data and the application of data analysis techniques such as machine learning (ML) to improve the quality of care and medical decisions. In fact, studies have shown that ML is an emerging area in healthcare because it can learn from historical data and identify patterns that can help avoid unexpected outcomes in TB treatment, such as treatment abandonment, death, and drug resistance. In this context, this research aims to use knowledge discovery in databases (KDD) techniques and machine learning to analyze and identify unknown patterns that may relate sociodemographic and clinical factors to the likelihood of a certain negative outcome in TB treatment occurring with a patient. Furthermore, the increasing availability of patient data in the healthcare field makes the use of techniques like ML even more relevant to enhance the management of TB patients.

Keywords: Tuberculosis; KDD; Machine learning; Bad outcomes.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LIS - Laboratório de Inteligência em Saúde

TB - Tuberculose

OMS - Organização Mundial da Saúde

DSS - Decision Support System

AM - Aprendizado de Máquina

BVS - Biblioteca Virtual em Saúde

TBWEB - Sistema de Notificação e Acompanhamento de Casos de TB

LPA - Line Probe Assay

KDD - Knowledge Discovery in Database

AUTO-ML - Automated Machine Learning

EMQ - Erro Médio Quadrático

KNN - K-vizinhos mais próximos

SUMÁRIO

1 APRESENTAÇÃO

O projeto de mestrado em questão desfrutou de uma bolsa FAPESP (processo número: 021/01961-0) e estava em execução no âmbito da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (FMRP-USP). Particularmente no Laboratório de Inteligência em Saúde (LIS), grupo de pesquisa na área de Sistemas de Informação para Gestão em Saúde Pública, liderado pelo prof. Domingos Alves. O LIS é composto por alunos do curso de graduação em Informática Biomédica da USP, bem como alunos de pós-graduação de várias áreas envolvidos com a FMRP, sejam do programa de pós-graduação em Saúde Pública do Departamento de Medicina Social ou do programa Interunidades em Bioengenharia.

Este projeto também está vinculado a um projeto mais geral, financiado pela FAPESP, dentro do programa e-Science: Saúde Digital Humana (processo número: 2020/01975-9). Este projeto tem uma sinergia importante ao projeto de Doutorado de Verena Hokino Yamaguti orientada pelo Prof. Antonio Ruffino Netto (um dos pesquisadores principais do projeto mais geral citado) e também financiado pela FAPESP (processo número: 2018/23963-2) intitulado *Estudo de modelo de predição de abandono ao tratamento da tuberculose (TB),* bem como outros projetos, sendo um de mestrado e dois de iniciação científica em andamento atualmente no âmbito do laboratório, a saber: *Estudo de um modelo de predição de resistência a drogas para tuberculose (TB)* do aluno Victor Cassão (bolsa CAPES de Mestrado), pertencente ao programa de pós-graduação Interunidades em Bioengenharia, *DSS-TB: Sistema de Suporte à Decisão para Tuberculose* da aluna Mariana Tavares Mozini (bolsa FAPESP, processo número: 2022/00020-0) e *Estudo comparativo do desempenho da modelagem baseada em Machine Learning e Regressão Logística aplicada a Tuberculose* do aluno Pedro Emílio Andrade Martins (bolsa FAPESP, processo número: 2022/03477-1), ambos fazem graduação em Informática Biomédica.

Além disto, a presente mestranda deste projeto estava liderando as iniciativas do LIS na área de análises e aplicações de machine learning, que hoje conta com 1 doutorando, 1 mestrando e 3 alunos de iniciação Científica.

2 INTRODUÇÃO E MOTIVAÇÃO

A tuberculose (TB) ainda está entre uma das doenças infecciosas mais fatais mundialmente. Apesar de ser uma doença curável, causada pela *Mycobacterium tuberculosis complex* (GLAZIOU et al., 2018), em apenas um ano houve 1.4 milhões de mortes e cerca de 10 milhões de pessoas ficaram doentes (WHO, 2020). Quando associado com outras comorbidades, como HIV, é ainda mais letal. Em 2019, TB contribuiu para mais de 208.000 mortes entre pessoas infectadas com HIV (WHO, 2020).

Outro fator relevante associado com esta doença são as questões socioeconômicas e culturais. Os determinantes da pobreza, a fragilidade das organizações dos serviços de saúde e as limitações quanti-qualitativas da gestão, veem mostrando indícios de que o destaque da TB no país, reflete a situação em que se encontra o desenvolvimento social neste território e a dificuldade por trás de ações de controle (SILVA; ANJOS; NOGUEIRA, 2014). A Organização Mundial da Saúde (OMS) e diversos autores também estabeleceram a relação entre a pobreza, que pode ser um resultado das condições baixas de saúde, e TB, que pode gerar a pobreza ao restringir a subsistência e as possibilidades de se ter um emprego (GLAZIOU et al., 2018).

O Brasil está entre os 30 países com maior prevalência de TB, somente em 2019 apresentou cerca de 96.000 casos, sendo estes 11.000 com pacientes contendo HIV-positivo (WHO, 2020), e diversos estudos estão sendo conduzidos (NERY et al., 2017; SANTOS et al., 2007), analisando como as condições socioeconômicas, intervenções neste sentido e as diferentes possibilidades de desfechos na TB estão relacionados. Estes podem auxiliar ao melhor entendimento da situação aqui no Brasil, encaminhando as melhores tomadas de decisão em relação às ações e tratamentos que precisam ser implantados para o controle e prevenção da TB, e os objetivos estabelecidos pela OMS.

Dentre estes objetivos estabelecidos pela OMS no programa End TB Strategy, temos o pilar da Pesquisa e Inovação (WHO, 2017), como caminho para a descoberta e desenvolvimento de novas ferramentas de apoio à decisão, por exemplo, que são imprescindíveis para orientar a prática de cuidados em saúde e sustentar as decisões de médicos e gestores que irão influenciar diretamente na qualidade do atendimento prestado à um paciente ou população (TANAKA; TAMAKI, 2012).

Um importante aspecto a ser considerado, é a disponibilização de um conjunto de dados de referência como base para um Sistema de Suporte à Decisão (DSS, Decision Support System), considerando integração, análise, comparação e visualização de dados de saúde por meio da integração de bases de dados heterogêneas e dispersas. Além disso, aplicar técnicas

que considerem dados com significado e que sejam contextualizados produzem informação em saúde que, por sua vez, auxilia na geração de conhecimentos; no desenvolvimento de políticas; planejamento; vigilância; monitoramento; avaliação de ações implementadas; orientação, alocação e gerenciamento de recursos; e melhoria na resolução e no atendimento de problemas e demandas da saúde da população (TANAKA; TAMAKI, 2012).

Sobretudo, os sistemas de apoio à decisão (DSS) incluem sistemas baseados no conhecimento que ajudam na tomada de decisões, provendo suporte para a tomada de ações/intervenções pelo gestor/profissional de saúde. A taxonomia do DSS consiste em: DSS orientado por modelo, que auxilia na tomada de decisões através da análise de dados e parâmetros fornecidos pelo utilizador; DSS orientado a dados, que auxilia na tomada de decisões, recuperando e manipulando dados; DSS orientado a documentos, que auxilia na tomada de decisões analisando o documento não estruturado; DSS orientado ao conhecimento, que auxilia na tomada de decisão real com a ajuda de fatos, regras, procedimentos etc.

Entre as técnicas computacionais sendo utilizadas por trás dos DSS, o aprendizado de máquina ou *machine learning* (AM) está sendo amplamente aplicado e já estão funcionando apropriadamente para a área médica, como alguns estudos mostram (SILVA, ANJOS, NOGUEIRA, 2014; ZENG; KALHORI, 2013). Este método baseia-se no aprendizado automatizado da máquina, em que é treinado com dados preliminares e inferido um modelo. A partir deste conhecimento inicial, a máquina irá classificar, prever ou padronizar os novos dados que forem apresentados, utilizando o modelo. Sendo um ciclo de aprendizado constante, contendo validação e otimização do modelo ao trabalhar com novas inserções de informações (MOHRI; AFSHIN; AMEET, 2018).

Este aprendizado pode ser de duas formas, a supervisionada e não supervisionada. A primeira já se sabe o resultado que espera obter com o modelo, enquanto a não supervisionada, procura identificar aspectos desconhecidos (latentes) do conjunto de dados de interesse.

Diante deste cenário, esta pesquisa possui como objetivo o uso de técnicas de ciência de dados e aprendizado de máquina, para a análise e identificação de padrões desconhecidos que possam associar fatores sociodemográficos, clínicos com os diferentes desfechos no tratamento de tuberculose, por exemplo o abandono, óbito e resistência. O número alto de desfechos ruins ainda se destaca em diversas regiões do país, como as taxas de abandono ao tratamento, variando de 4.5 a 20.3% (PAIXÃO; GONTIJO, 2007), fazendo com que proporcionem aumento na resistência medicamentosa e sejam fatores de efeito negativo no controle da TB (YAMAGUTI et al, 2020). Além disso, o domínio da saúde apresenta uma complexidade e um constante aumento na quantidade disponível de dados do paciente para serem explorados, sendo

mais um fator positivo para o uso de novas técnicas, como o AM, para extrair novos conhecimentos e melhores decisões quanto ao manejo do paciente com TB.

3 OBJETIVO

O objetivo deste projeto é a aplicação de técnicas de aprendizado de máquina supervisionado e não supervisionado em dados de tratamento de tuberculose. Com isso será possível desenvolver um sistema de apoio à decisão, ou seja, uma ferramenta para auxiliar médicos e especialistas a identificarem com antecedência a possibilidade de um paciente vir a ter um desfecho ruim e tomar medidas e cuidados para evitá-lo.

4 MATERIAIS E MÉTODOS

4.1 Revisão da Literatura

A revisão da literatura seguiu as diretrizes do Joanna Briggs Institute Manual for Evidence Synthesis (AROMATARIS; MUNN, 2020) no primeiro momento para *scoping review* e depois para *umbrella reviews* (também conhecido como revisões de revisões). É importante relatar que decidimos seguir por este último caminho devido ao grande número de revisões sistemáticas e sínteses de pesquisas recentes que foram identificadas para o tópico principal abordado neste projeto: modelos preditivos para desfechos de tuberculose.

Dessa forma, o foco da revisão bibliográfica em curso é a busca por publicações de métodos de algoritmos de aprendizado de máquinas utilizados no tratamento da TB, para explorarmos e entendermos melhor os métodos e ferramentas em uso, e obtivemos diversos avanços. A começar pelas hipóteses levantadas para orientar a busca, vale destacar que começamos com perguntas mais gerais e depois fomos afunilando para nosso objetivo:

● O que há na literatura de aplicações de aprendizado de máquina não supervisionado com dados clínicos?

● O que há na literatura de tomada de decisões em saúde guiadas por aprendizado de máquina supervisionado em TB?

● O que há na literatura de reconhecimento de padrões em dados clínicos de TB?

● O que há na literatura do uso de statistical learning e machine learning em dados clínicos de TB?

Em seguida foi feito um levantamento de descritores, utilizando a base do DeCS - Descritores em Saúde, criado pela Biblioteca Regional de Medicina (BIREME) e disponibilizado na Biblioteca Virtual em Saúde (BVS). A seleção destes foi baseada no título, palavras importantes no resumo e na área de pesquisa do projeto:

1. Tuberculose
2. Tuberculose Extensivamente Resistente a Medicamentos
3. Tuberculose Latente
4. Ciência de Dados
5. Aprendizado de Máquina Não Supervisionado
6. Inteligência Artificial
7. Computação Matemática
8. Tomada de Decisões
9. Avaliação de Processos e Resultados em Cuidados de Saúde

10. Interpretação Estatística de Dados
11. Análise de Dados
12. Reconhecimento Automatizado de Padrão
13. Dados de Saúde Gerados pelo Paciente
14. Interpretação Estatística de Dados
15. Análise por Conglomerados

Logo após este levantamento, começamos a busca pelos artigos na base de dados do PubMed, não somente utilizando destes descritores, mas de suas combinações também, sendo neste primeiro momento estes critérios de inclusão. Esta base contém mais de 34 milhões de citações e *abstracts* da literatura na área da saúde, e está disponível online desde 1996.

Já para o refinamento, foi definido incluir apenas os artigos que: (1) possuem apenas adultos nas amostras; (2) artigos que estejam completos, removendo os que contenham somente *abstracts*; (3) utilizam de algum método de AM; (4) uso de dados clínicos e sociodemográficos associados à TB, removendo os que trabalhem somente com imagens ou dados genéticos da TB; (4) que contenham pelo menos um desfecho envolvido no estudo, principalmente os ruins e (5) que estejam escritos em algum desses idiomas: português e inglês.

4.2 Base de dados utilizadas

A principal base de dados que trabalhamos consiste dos dados dos pacientes cadastrados no Sistema de Notificação e Acompanhamento de Casos de TB (TBWEB) da Secretaria de Estado da Saúde em parceria com a PRODESP. O objetivo do TBWEB é a vigilância epidemiológica e monitoramento de casos da TB no Estado de São Paulo. O sistema foi construído on-line, permitindo o registro de notificações de TB em todo o estado, e em tempo real, onde novos casos podem ser submetidos e seus dados podem ser recuperados através da internet durante todo o processo do tratamento da TB em curso. O sistema fornece uma única entrada para cada paciente e um histórico de tratamentos anteriores a esta. Além de atuar como banco de dados centralizado para casos de TB, o TBWEB fornece ferramentas para obtenção de informações gerenciais, como relatórios de dados de pacientes, análise de coorte de dados e monitoramento do tratamento, além de proporcionar uma melhor comunicação entre os vários níveis de monitoramento epidemiológico e seus serviços de assistência (SECRETARIA DE SAÚDE DO ESTADO DE SÃO PAULO, 2008).

Logo no começo do projeto tivemos acesso aos dados relacionados aos períodos de 2006 a 2016. Em agosto de 2021, a partir da aquisição dos dados complementares até o período de 2019, iniciamos o processo de entendimento e limpeza dessa base. Nesta fase começamos por

fazer uma análise exploratória dessa base de dados para entender adequadamente os dados contidos nela.

Além do TBWEB, também estamos trabalhando em colaboração com o projeto de validação e custo da performance do *Line Probe Assay* (LPA) como método de diagnóstico rápido para tuberculose resistente em centros de referência no Brasil, coordenado pelo Prof. Dr. Afrânio, organizado pelo Ministério da Saúde e Secretaria de Vigilância em Saúde. Os dados estão armazenados na plataforma RedCap e é constituído por 14 formulários, contendo 679 variáveis (MINISTÉRIO DA SAÚDE, 2021).

Os objetivos deste estudo são: (1) analisar a acurácia diagnóstica do teste LPA1 (resistência a rifampicina e isoniazida); (2) analisar a acurácia diagnóstica do teste LPA2 (resistência a capreomicina, amicacina e quinolona); (3) analisar o tempo decorrido entre a triagem e a detecção de resistência e início do tratamento anti-TB; (4) analisar custos em cada centro de referência; (5) analisar o sistema de gestão de qualidade laboratorial (MINISTÉRIO DA SAÚDE, 2021). Destacando os objetivos 1, 2 e 3, que vão de encontro aos objetivos deste projeto de mestrado e auxiliam no entendimento do desfecho associado à resistência, e seus dados em sua maior parte são originados do formulário 5, que iremos focar nas análises.

4.3 Definição de desfecho

Um importante aspecto que tem de ser levado em consideração, para este projeto como um todo, é a definição de desfecho de maneira padronizada e que sejam mutuamente exclusivos. Para isso vamos seguir as recomendações da OMS para cada tipo de desfecho trabalhado nesta pesquisa, como mostrado na Tabela 1 (LASERSON et al., 2005)(PEETLUK et al., 2020)(WHO, 2014).

As definições foram projetadas para atender à ampla gama de regimes de tratamento e durações atualmente em uso no mundo todo. De maneira simples, os desfechos ruins podem ser vistos como: 1. Óbito, 2. Internação, 3. Abandono ao tratamento, 4. Desenvolvimento de resistência (MDR e XDR), 5. Baixa adesão ao tratamento.

Tabela 1 - Definição da OMS para desfechos no tratamento de TB

Desfechos	**Definição**
Cura	Tratamento realizado de forma completa e sem evidência de falha e/ou paciente confirmado de TB no início do tratamento que pelo menos no último mês ou em ocasião anterior tenha

	resultado negativo em baciloscopia ou cultura.
Tratamento completo	Paciente de TB que realizou de forma completa o tratamento, mas não há registro de baciloscopia ou cultura negativa no último mês ou em ocasião anterior (testes não realizados ou resultados indisponíveis).
Sucesso do tratamento	Uma junção da cura e tratamento completo.
Falha do tratamento	Paciente de TB que possui baciloscopia ou cultura positiva no mês 5 ou mais tardar do tratamento.
Perda de seguimento (abandono do tratamento)	Paciente de TB que não iniciou o tratamento ou que foi interrompida por 2 meses consecutivos ou mais.
Não avaliado	Paciente de TB que não possui um desfecho definido. Aqui inclui casos de transferência para outra unidade de tratamento ou que é desconhecido o seu desfecho.
Óbito	Paciente de TB que morre por qualquer razão, antes ou durante o tratamento

Fonte: (WHO, 2014)

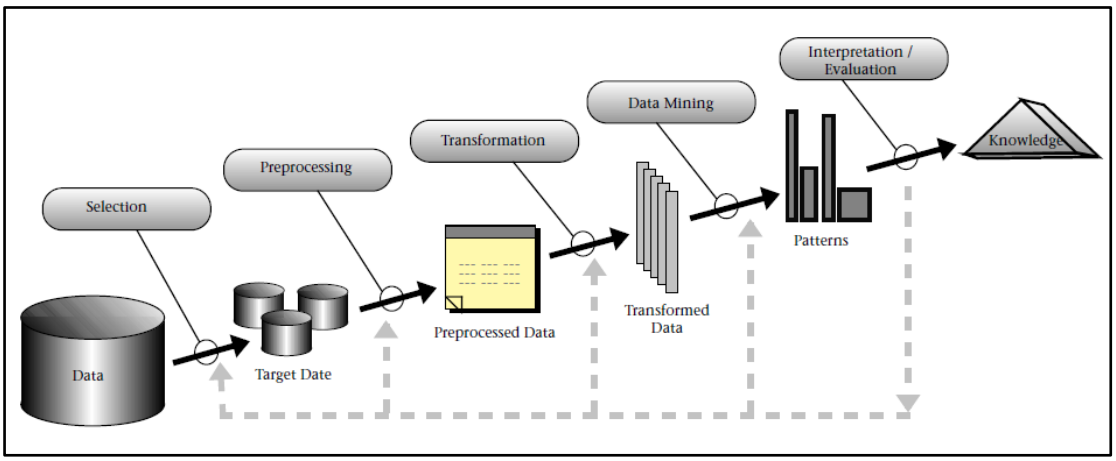
4.4 Processo KDD

Uma metodologia já muito utilizada na literatura (MCCLEAN et al., 2004)(DEGRUY, 2000) para o processo de descoberta de conhecimento é o KDD (Knowledge Discovery in Database), desenvolvido com intuito de "dar sentido" ao grande volume de dados que vêm sendo gerados e trazer praticidade, para aplicações (descoberta de padrões) que antes eram feitas de forma manual, ocorrendo atrasos, gastos elevados e análises supérfluas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Antes de desenvolvermos uma solução contendo modelos de AM neste processo, é necessário o entendimento do problema e o contexto em que se insere. Depois vamos para a etapa de preparação e pré-processamento dos dados, seleção de características (*feature selection*) e redução de dimensionalidade, desenvolvimento de modelos, validação e uso do

conhecimento, ou seja, colocar em produção para o uso no dia a dia dos médicos e especialistas em TB. Na figura 1, podemos visualizar estas etapas.

Figura 1 - Visão geral do processo KDD



Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

Estes passos são interativos e geralmente cíclicos, com intuito de sempre manter atualizado e otimizado o modelo de AM e corretamente alinhado ao problema que busca resolver. Assim, abordaremos brevemente as principais etapas deste ciclo.

4.4.1 Preparação e pré-processamento dos dados

Nesta etapa é preparado os dados para análises futuras e para a construção dos modelos, bem como para identificar e corrigir problemas relacionados à qualidade de dados, e estes objetivos podem ser divididos da seguinte forma:

Identificação dos dados: refere-se à caracterização das bases de dados disponíveis, nomeadamente a descrição das variáveis, o volume de dados, formato armazenado. Todas as variáveis relevantes no contexto dos problemas clínicos serão identificadas e categorizadas (preditoras ou *outcomes*);

Aquisição e filtragem dos dados: Os dados obtidos das bases serão recolhidos e reunidos para remoção de dados corrompidos ou de variáveis irrelevantes para a condução das análises;

***Data profiling*:** Inclui análises para validação de metadados, detecção de anomalias /*outliers* para posterior correção através da limpeza dos dados (*data cleaning*), entender a estrutura e os relacionamentos existentes nos dados e antecipar eventuais erros/problemas antes do início da construção dos modelos. Em geral, a metodologia para o *data profiling* envolve as

seguintes etapas: (1) análise de colunas individuais, em que cada variável é avaliada para verificar se apresenta valores válidos; (2) análise da estrutura da base de dados, em que o relacionamento entre as colunas de uma mesma tabela e o relacionamento entre as tabelas são verificados e validados; (3) uso de regras de validação simples que avaliam a consistência e credibilidade dos dados do ponto de vista clínico; (4) uso de regras complexas, que também avaliam os dados mediantes do ponto de vista clínico, porém através de condições que só podem ser verificadas com o processamento de grande quantidade de dados, geralmente comparando múltiplos registros de pacientes, presentes em múltiplas fontes; e (5) análise estatística e quantitativa dos dados para a detecção de problemas que não podem ser identificados através das regras de validação simples e complexas (OLSON, 2003).

Balanceamento dos dados: Uma característica comum dos conjuntos de dados médicos é a presença de classes desbalanceadas. Apesar de extensivamente pesquisados, os problemas de classes desbalanceadas ainda são uma questão em aberto. Detalhes sobre os algoritmos que poderão ser utilizados para lidar com este problema serão descritos na próxima seção**.**

Tratamento de *missings*: Dados com valores em falta (*missings*) podem reduzir o desempenho dos modelos e levar a erros de previsão. Três métodos para tratamento de *missings* poderão ser explorados nesta proposta: (1) deleção, em que os casos que apresentam *missings* são eliminados da base ou, se forem numerosos, estes casos podem ser mantidos, porém apenas as variáveis de interesse que se encontrarem presentes serão consideradas para as análises: (2) Imputação, em que os valores em falta são preenchidos por valores estimados com base nos outros dados e em técnicas estatísticas (médias, modas etc.) (GAMA et al., 2012).

Tratamento de *outliers*: Casos que apresentam desvios substanciais em relação aos padrões gerais dos dados (*outliers*) podem introduzir um viés considerável durante a construção dos modelos preditivos. Desta forma, *outliers* serão identificados através de métodos tradicionais na literatura (e.x. técnicas de visualização de dados, como análise de *box-plot* e histogramas, identificar casos que apresentam valores fora do 5º ou 95º percentis ou que apresentam três ou mais desvios padrões em relação à média, distância de Mahalanobis, distância de Cook, etc.). Quanto ao tratamento dos *outliers*, o método dependerá da quantidade dos mesmos. Em situações com poucos *outliers,* pode-se recorrer à eliminação direta destes dados, do contrário os *outliers* podem ser tratados separadamente durante a construção dos modelos. (GAMA et al., 2012).

Amostragem: Algoritmos preditivos nem sempre precisam utilizar a totalidade dos dados, principalmente quando o volume de dados requer grande capacidade de processamento. Nesta fase, deverá ser selecionada uma amostra representativa dos dados de modo a aumentar a eficiência e ainda obter modelos com bom desempenho.

Seleção de características: Existem vários tipos de estratégias para selecionar atributos, que podem ser classificadas em três abordagens principais: filtros, *wrappers* e métodos incorporados. Filtros são técnicas que consideram a seleção de atributos como uma etapa de pré-processamento independente do algoritmo indutor. Eles são computacionalmente mais rápidos do que os outros e a seleção de atributos é realizada com base em características de dados intrínsecas, sem considerar um algoritmo de aprendizado. Algoritmos de *wrappers* consideram um algoritmo de aprendizado para guiar a busca pelo melhor subconjunto de atributos e cada subconjunto é avaliado por um algoritmo indutor. Por outro lado, os métodos incorporados utilizam a seleção de recursos no processo de treinamento, otimizando as taxas de classificação ao selecionar o melhor subconjunto. Diferentemente das abordagens anteriores, em métodos incorporados, o processo de aprendizado não pode ser separado da seleção de recursos.

4.4.2 Técnicas de aprendizado de Máquina

O aprendizado de máquina ou *machine learning*, é uma subárea da Inteligência Artificial, em que a partir de algoritmos computacionais, a máquina é capaz de aprender e aperfeiçoar de forma automatizada. Para que isso ocorra, são altamente dependentes da disponibilidade e da qualidade dos dados. Os dados são utilizados não só para treinamento inicial, teste, validação, mas também para contínuo aprimoramento dos algoritmos.

Indo mais a fundo nas técnicas de AM, temos duas que são mais aplicadas tanto na literatura quanto no mercado: aprendizado supervisionado e não supervisionado. A primeira técnica, retorna seus resultados com o sistema de tomada de decisão tendo base em exemplos inseridos no momento de treino do modelo, em que a partir dos dados de entrada já sabemos os dados de saída correspondente. Já a segunda técnica, funciona de uma forma diferente, em que não sabemos quais são os dados de saída correspondentes, temos apenas os dados de entrada, então a máquina irá analisar essas informações e utilizar de padrões e características semelhantes para agrupar os dados e obter uma possível saída.

Além disso temos outros problemas, principalmente encontrados em bases de dados de saúde, que podem utilizar de algoritmos de AM para solucioná-los, como:

Bases de dados desbalanceadas: O problema das bases de dados desbalanceadas ainda é um problema em aberto no aprendizado supervisionado. Uma característica que define o

problema é que a amostra de uma classe supera significativamente o número da outra ou de outras. Geralmente, as tentativas de resolver esse problema solucionam o problema no nível dos dados, no nível do modelo ou na combinação de ambos (BOUGHORBEL, 2017; LÓPEZ, 2013). As soluções em nível de dados lidam com técnicas de distribuição e amostragem de dados. Soluções baseadas em modelo, referem-se a projetos de algoritmos para resolver o problema, como algoritmos de aprendizado sensíveis a custos e por conjuntos.

Análise de agrupamento – *Clustering*: Este tipo de abordagem faz parte do aprendizado não supervisionado, em que buscamos padrões e relações presentes nas características dos dados e a partir destas semelhanças estes são agrupados em *clusters*. Vem sendo muito utilizado na saúde para segmentação de pacientes, ao prover novos entendimentos em relação a quem são estes pacientes (características sociodemográficas, genéticas, clínicas) e do que precisam (comportamento a determinados tratamentos e exames) (KOO et al., 2022; YAO et al., 2023). Os algoritmos mais conhecidos são o K-means, C-Means e hierarchical.

Aprendizado sensível aos custos – *Cost-Sensitive learning*: Nos problemas de classificação, especialmente no contexto médico, existem diferenças entre os erros de uma classificação incorreta. O impacto do erro tipo II pode ser muito mais sério do que o impacto do erro do tipo I. Por esse motivo, a aprendizagem sensível ao custo tem recebido muita atenção em ML, pois é capaz de ponderar as instâncias de treinamento de acordo com sua importância (LÓPEZ, 2013; SUN, 2007). O algoritmo sensível ao custo pode ser projetado para lidar diretamente com o problema ou envolver um classificador. Geralmente, a técnica recebe o nome de classificador de meta-aprendizado com sensibilidade a custos. O algoritmo permite definir uma matriz de custos para fornecer custos associados à matriz de confusão. A matriz de custos altera os valores para penalizar os algoritmos quando classifica incorretamente uma instância como um dos tipos de erro. Nesse sentido, é possível definir um peso maior para erros falsos negativos (LI, 2010; HAIXIANG, 2017).

Aprendizado baseado em agrupamento – *Ensemble learning:* Os métodos de conjunto são algoritmos que empregam um conjunto de classificadores para fornecer uma votação ou votação ponderada para pontos de dados, fornecendo uma previsão. Recentemente, vários métodos de agrupamento estão disponíveis. Os métodos mais famosos são florestas aleatórias (*random forest*), *bagging* e votação (BREIMAN, 2001).

Aprendizado de máquina automatizado - *Automated Learning*: Atualmente, uma grande quantidade de algoritmos de classificação está disponível e eles também foram utilizados para fins muito diversos, às vezes por usuários com pouca ou nenhuma experiência na área de aprendizado de máquina. Esses algoritmos têm vários hiperparâmetros, que exigem

ajustes muito específicos para maximizar a precisão. Nesse contexto, o aprendizado automatizado de máquinas (Auto-ML) é um tópico crescente de pesquisa e está chamando atenção para tentar resolver o problema de encontrar o melhor algoritmo de classificação (incluindo a melhor configuração ou definição de parâmetros) para qualquer conjunto de dados (FEURER, 2015). Alguns pacotes ML já estão disponíveis para esse fim e os resultados obtidos são significativos.

4.4.3 Validação

A partir da construção do modelo, iremos tentar validar e classificar os resultados gerados por tais algoritmos fazendo simulações rápidas e pré-testes em uma segunda etapa. Isso será feito inicialmente com uma validação interna ao aferir a capacidade de predição do algoritmo na base de dados legada, criando pacientes "virtuais" dessas bases. Também realizaremos pré-testes com pacientes reais estudados na segunda fase. De uma maneira geral, pretendemos relatar todas as etapas de modelagens e análises com detalhes suficientes para maximizar a transparência e a reprodutibilidade, inclusive aderindo às diretrizes do TRIPOD (COLLINS et al., 2015).

Também serão utilizadas as medidas de avaliação típicas para essa área que são: acurácia, precisão, sensibilidade, erro médio quadrático (EMQ), matriz de confusão e área ROC (aprendizado supervisionado) e para os modelos não supervisionados: coeficiente da silhueta e método do cotovelo. Vale lembrar que os modelos de AM são tipicamente avaliados em termos de desempenho de discriminação (por exemplo, precisão, área sob a curva ROC).

4.5 Ferramentas Utilizadas

Para aplicar todo o ciclo da descoberta de conhecimento nos dados, foi utilizado a linguagem de programação Python. É gratuita e uma ferramenta fácil de usar, orientada a objetos e compatível com diversos sistemas operacionais como Windows e Linux (MCKINNEY, 2010). Também possui bibliotecas que fornecem recursos para aplicações como matemática científica, inteligência artificial e análises biomoleculares.

Depois do problema e do conjunto de dados estarem bem definidos, foi acessado de forma remota o banco de dados, usando a biblioteca SQLAlchemy, que permite usar a linguagem de banco de dados SQL adaptada para Python, sem perder performance e eficiência (BAYER, 2012). Para o pré-processamento e preparação dos dados, as bibliotecas Pandas e Numpy são essenciais. Pandas fornece ferramentas para análise de dados, então temos recursos para ler diferentes tipos de conjuntos de dados (como SQL e excel), para manipular e limpar os dados, e aplicar cálculos matemáticos e estatísticos, juntamente com Numpy (MCKINNEY,

2010), também será utilizado as bibliotecas Matplotlib e Seaborn, focadas na visualização das informações.

Na etapa de aplicações das técnicas de aprendizado de máquina, é usado a biblioteca Scikit-Learn (PEDREGOSA, 2011) e PyCaret (AutoML)(ALI, 2020), e para o balanceamento a biblioteca Imbalanced-learn (que possui sua base na Scikit-learn). A interpretação dos resultados pode ser feita com auxílio de diferentes gráficos, utilizando novamente o Matplotlib e o Seaborn. Para a validação, a própria biblioteca dos modelos fornece formas de validar e comparar diferentes modelos de AM (HUNTER, 2007).

5 RESULTADOS E DISCUSSÃO

5.1 Revisão da Literatura

Até o momento obtivemos alguns resultados interessantes. Quando buscamos no PubMed os principais descritores, obtivemos respectivamente:

A. Tuberculose: 272.542 artigos, entre 1848 e 2021;

B. Aprendizado de Máquina Não Supervisionado: 2.463 artigos, entre 1990 e 2021;

C. Ciência de dados: 572.158 artigos, entre 1880 e 2021.

D. Análise de dados: 2.114.464 artigos, entre 1913 e 2021.

Agora, ao utilizar conectores como AND (deve conter no artigo os dois termos) e OR (deve conter ao menos um dos dois termos), os resultados foram filtrados drasticamente, tanto em quantidade, quanto no período em que houve estudos. Como podemos ver nos seguintes exemplos:

A. Ciência de Dados AND Tuberculose: 3.244 artigos, entre 1947 e 2021;

B. Aprendizado de Máquina Não Supervisionado AND Tuberculose: 5 artigos, entre 2017 e 2021;

C. Análise de Dados AND Tuberculose: 13.426 artigos, entre 1926 e 2021.

D. Aprendizado de Máquina Não Supervisionado OR Tuberculose: 275.019 artigos, entre 1848 e 2021.

Entre os resultados encontrados nestes exemplos, algo bem interessante de ser observado é, primeiro que o item B, apenas 2 artigos que realmente eram relacionados à TB e quando buscamos Aprendizado de Máquina ou Aprendizado de Máquina Não Supervisionado AND tuberculose, em sua grande maioria são pesquisas com imagens médicas ou relacionadas a resistência a medicamentos específicos. Outra curiosidade é quando buscamos Aprendizado de Máquina AND Tuberculose AND Desfecho, há 34 artigos, entre os anos de 2014 e 2021 suas publicações, em que temos também artigos relacionados à biologia molecular da TB.

Para trabalhos futuros, iremos definir um checklist do novo caminho baseando-se nos conceitos estabelecidos no JBI Manual for Evidence Synthesis (AROMATARIS; MUNN, 2020). Também iremos aplicar estas buscas em outras bases de dados de artigos, como o Scopus e começar a refinar os resultados para selecionar os artigos que são relevantes para o objetivo final desta revisão.

5.2 Processo KDD aplicado ao TBWEB

5.2.1 Análise exploratória dos dados

No primeiro momento foram realizadas análises e a auditoria dos dados do sistema TBWEB. Foi possível identificar cerca de 277.870 registros de pacientes e 114 variáveis, listadas no Apêndice A. Um fato relevante desta tabela de variáveis é que o banco de dados não continha um dicionário de dados. Nesse sentido, construímos um tal dicionário com a ajuda da equipe da Secretaria Estadual de Saúde relacionada ao TBWEB. Como pode ser visto na tabela a despeito de a maioria das variáveis ali contidas serem relativamente transparentes em sua interpretação, é ainda importante termos estabelecido a descrição efetiva de cada uma.

Neste processo de entendimento dos dados, obtivemos alguns resultados interessantes. Alguns valores que estavam vazios, foram marcados com '?' ou ' ', foi necessário utilizar da biblioteca matemática numpy e substituir por 'NaN' (*Not a Number*), para que pudéssemos analisar os valores faltantes de forma mais precisa. Depois verificamos a quantidade de valores vazios e sua respectiva porcentagem em cada variável, analisando individualmente, e depois considerando apenas acima de 70% vazio.

As seguintes primeiras variáveis chamaram a atenção e que foi necessário cuidado na sua forma de tratar: 'Forma_Clínica3' com 99.77% das informações vazias, 'Motivo_Mudanca_Esquema' com 99%, 'Esquema_Atual' com 98.64% e 'Tipo_Saida_3' com 97.86%. Além dessas, pode ser verificado outras variáveis importantes também, como 'Resistência' contendo 80.57% *missings*.

Com relação aos desfechos, pode ser observado na tabela 2, que o TBWEB utiliza de uma variedade de desfechos (Situacao_Atual, SIT1 até SIT6) e com distinções relevantes entre os pacientes, podendo haver mais de um desfecho para um paciente em particular. Isso mostra que os dados como organizados no TBWEB, não formam conjuntos mutuamente exclusivos (MINISTÉRIO DA SAÚDE, 2016). Nesse sentido esses dados foram reorganizados de maneira a obter um conjunto de desfechos mais plausíveis para a análise e alinhados com a definição da OMS (WHO, 2014).

Tabela 2 - Desfechos na base de dados TBWEB

Desfecho	**Definição**

Cura	Paciente que apresentar duas baciloscopias negativas, sendo uma em qualquer mês de acompanhamento e outra ao final do tratamento (5º ou 6º mês). Para os casos com necessidade de ampliar o tempo de tratamento, serão considerados os 2 últimos meses. A alta por cura também será dada ao paciente que completou o tratamento sem evidência de falência, e teve alta com base em critérios clínicos e radiológicos, por impossibilidade de realizar exames de baciloscopia ou cultura.
Abandono	Paciente que fez uso da medicação por 30 dias ou mais e interrompeu o tratamento por mais de 30 dias consecutivos.
Abandono Primário, Faltoso	Paciente que fez uso da medicação por menos de 30 dias e interrompeu por mais de 30 dias consecutivos, ou quando o paciente diagnosticado não iniciou o tratamento.
Óbito Não TB	Por ocasião do conhecimento da morte do paciente por qualquer causa básica que não seja tuberculose, mesmo que a tuberculose esteja constando como causa associada no SIM. A causa do óbito deve estar de acordo com as informações contidas no SIM.
Óbito TB	Quando o óbito foi causado pela tuberculose. A causa do óbito deve estar de acordo com as informações contidas no SIM
Mudança de diagnóstico	Quando ocorrer alteração no diagnóstico e for elucidado que não se tratava de um caso de tuberculose.
Falência/Resistência	Será registrada nas seguintes situações: - persistência da baciloscopia de escarro positiva ao final do tratamento; - doentes que no início do tratamento apresentavam baciloscopia fortemente positiva (+ + ou + + +) e mantiveram

	essa situação até o 4º mês; - baciloscopia positiva inicial seguida de negativação e de novos resultados positivos por 2 meses consecutivos, a partir do 4º mês de tratamento
Transferência, Transferência Outro Estado/País	Quando o doente for transferido para outro serviço de saúde. A transferência deve ser processada por meio de documento que contenha informações sobre o diagnóstico e o tratamento realizado até aquele momento. É de responsabilidade da unidade de origem a confirmação de que o paciente compareceu à unidade para a qual foi transferido.
Em tratamento ambulatorial, Em tratamento internado	Quando o paciente ainda segue em tratamento (ainda não realizou de forma completa e sem evidências de falha e/ou abandono); Quando o paciente ainda segue em tratamento, porém houve agravamento de sua situação e foi internado, seja por causa da TB ou não (ainda não realizou de forma completa e sem evidências de falha e/ou abandono).
Mudança de Esquema por Intolerância/Toxicidade	Quando o paciente necessitar da adoção de regimes terapêuticos diferentes do esquema básico, seja por intolerância e/ou por toxicidade medicamentosa.

Fonte: (MINISTÉRIO DA SAÚDE, 2016; 2019)

Inclusive foram observados 206.765 (74.41%) desfechos de cura, 33.127 de abandono somado a abandono primário (11.92%), 21.186 (7.62%) de óbitos totais e 2136 (0.76%) de resistências. Do total do número de abandonos, 77.53% são homens, 35.51% têm entre 20 e 29 anos, foram classificados com TB pulmonar 87.95% destes casos, 55.71% é negativo para HIV e aids, e cerca 20% possuem baixa escolaridade (4 a 7 anos).

Em relação às resistências, pode ser observado que 73% são homens e 26% mulheres, 25% do total destes estão entre 30 e 39 anos, 33% possuem baixa escolaridade (4 a 7 anos), 92% foram classificados com TB pulmonar e 36% com multirresistência, ou seja, resistentes a pelo menos dois medicamentos anti-TB.

Outra característica relevante observada nesta base de dados é que 31.756 pacientes são detentos (11.43%), destes 11.374 são de etnia Branca (35.81%) e 11.284 são Pardos (35.53%),

27.059 pacientes possuem entre 20 e 39 anos (85.20%), 31028 (97.70%) são do gênero masculino. Do ponto de vista clínico, estes pacientes possuem o desfecho Cura em grande maioria (84.70%), porém em relação aos desfechos ruins, o abandono é o que mais se destaca com 3153 pacientes (9.93%), mais detalhes podemos observar na tabela 3.

Tabela 3 – Distribuição das características básicas

Base de dados total (n=277870)			
Faixa Etária		**Diabetes**	
Menor de 1 ano	831	Sim	15878
01-04	2156	Não	259707
05-09	1707	**Doença mental**	
10-14	3015	Sim	4446
15-19	14265	Não	271139
20-29	71285	**Aids**	
30-39	63809	Sim	28273
40-49	50899	Não	247312
50-59	36762	**HIV**	
60-69	18897	Sim	30399
70-79	8433	Não	205293
Maior de 80 anos	3172	**Outra doença autoimune**	
Sexo		Sim	2497
Feminino	79845	Não	273088

Masculino	195740	**Classificação anatômica**	
Raça		Pulmonar	224338
Branco	106752	Extrapulmonar	40972
Pardo	78721	Pulmonar + Extrapulmonar	9074
Preto	25709	Disseminado	1083
Amarelo	2058	**Baciloscopia (Status Microbiológico)**	
Indígena	1083	Positivo	146181
Escolaridade		Negativo	69561
Nenhuma	9158	**Descoberta do caso**	
De 1 a 3 anos	23574	Demanda ambulatorial	135484
De 4 a 7 anos	78010	Urgência/ Emergência	59636
De 8 a 11 anos	76852	Elucidação Diagnóstica em Internação	45644
De 12 a 14 anos	14926	Busca Ativa em Instituição	11620
15 anos ou mais	5921	Investigação de contatos	7991
Detento		Busca Ativa na Comunidade	5130

Sim	31756	Descoberta após o óbito	2810
Não	243805	**Desfechos**	
Uso de Álcool		Cura	206765
Sim	44741	Abandono (abandono, primário e faltoso)	33185
Não	230844	Óbito não TB	12527
Uso de drogas		Óbito TB	8659
Sim	33822	Mudança diagnóstica	6893
Não	241763	Falência/ Resistência	2136
Tabagismo		Em tratamento ambulatorial	1706
Sim	33056	Transferência (estado, país, hospital)	1856
Não	242529	Em tratamento internado	163
		Mudança de Esquema Intolerância/ toxicidade	442

Fonte: Própria autora (2023)

5.2.2 Pré-processamento

Após o entendimento, realizamos a etapa de limpeza e foi objetiva, com a eliminação dos valores duplicados, inconsistentes e os vazios sem forma de tratar, ou seja, os casos em que ‘sem informação’, ‘ignorado’ ou ‘outros’ não foi possível de utilizar como imputação. Também foi padronizada as linhas e colunas com letras maiúsculas, uma vez que a linguagem de

programação python é case sensitive, e foi realizada a transformação de variáveis que constavam como objeto, porém são numéricas. Para a variável ´Codigo_Tratamento_Anterior', temos o código 9 indicando sem informação, então fiz a imputação nos casos vazios deste valor. Já para as variáveis 'TOTCOMUNIC', 'COMUNICEXA' e 'COMUNICDOE', optei por colocar 0 como valor de imputação por não termos essa informação e para não causar um enviesamento ao utilizar outro valor numérico, no mesmo caso para o 'nro_doses_pri' e 'nrp_doses_seg'.

Como determinado pela OMS e decidido seguir neste projeto, os valores identificados na base da TBWEB, na variável 'Situacao_Atual', foram substituídos conforme a tabela 4 e seguindo suas definições apresentadas anteriormente.

Tabela 4 – Desfechos TBWEB X OMS

Tipo de Desfecho TBWEB	**Tipo de Desfecho OMS**
Cura	Cura
Abandono	Perda de seguimento
Óbito NTB	Óbito
Óbito TB	Óbito
Mudança de diagnóstico	Não avaliado
Falência/resistência	Falha do tratamento
Em tratamento ambulatorial	Não avaliado
Transferência outro estado/país	Não avaliado
Abandono primário	Perda de seguimento
Mudança de esquema intolerância/toxicidade	Falha no tratamento

Transferência	Não avaliado
Em tratamento internado	Não avaliado
Faltoso	Perda de seguimento
Sem informação	Não avaliado
Outra	Não avaliado

Fonte: Própria autora (2022).

Em seguida, foi realizada a seleção das variáveis, baseando-se na literatura e na sua importância para o desenvolvimento dos modelos de AM (YAMAGUTI, 2018; YAMAGUTI, 2020). Como resultado, obtivemos um novo conjunto com 54 variáveis, listadas no Apêndice C com seus respectivos valores possíveis, e 181.605 pacientes com TB. Como uma forma de avaliação destas variáveis, a seguir temos diferentes análises de correlação, mostrando a associação ou relação presente entre elas e os desfechos (principalmente).

Para que os próximos passos fossem realizados, foi necessário transformar as variáveis categóricas em valores numéricos. Levando em consideração a cardinalidade destas, estipulei um limiar de 5 rótulos para que o tratamento fosse feito de forma a transformar cada categoria em uma coluna nova e seus valores seriam 0 (categoria não presente) e 1 (sim categoria presente)(função get_dummies - biblioteca Pandas). Caso fosse maior que 5 o número de rótulos, o tratamento seria de forma a criar uma codificação de 0 até *n*-1, sendo *n* o número de categorias presentes (função LabelEncoder() - biblioteca scikit-learn). Na tabela 5, podemos ver quais variáveis passaram por cada tipo de tratamento.

Tabela 5 - LabelEncoder() X Dummies

LabelEncoder()		**Dummies**	
'RACA_COR'	'MUNICIPIO_TRATAMENTO'	'SEXO'	'AIDS'
'FAIXA_ETARIA'	'ESQUEMA_ATUAL'	'GESTANTE'	'DIABETES'

'NATURALIDADE'	'MOTIVO_INTERNACAO_1'	'TIPO_CASO'	'ALCOOLISMO'
'ESCOLARIDADE'	'TIPO_SAIDA_1'	'CLASSIFICACAO'	'DOENCA_MENTAL'
'TIPO_OCUPACAO'	'MOTIVO_INTERNACAO_2'	'BACILOSCOPIA_ESCARRO'	'USO_DROGAS'
'MUNICIPIO_RESIDENCIA'	'TIPO_SAIDA_2'	'BACILOSCOPIA_OUTRO_MATERIAL'	'OUTRAS_DOENCAS_IMUNO'
'FORMA_CLINICA_1'	'MOTIVO_INTERNACAO_3'	'CULTURA_ESCARRO'	'TABAGISMO'
'FORMA_CLINICA_2'	'TIPO_SAIDA_3'	'CULTURA_OUTRO_MATERIAL'	'ESQUEMA_INICIAL'
'FORMA_CLINICA3'		'HISTOPATOLOGIA'	'MUDANCA_ESQUEMA'
'DESCOBERTA'		'NECROPSIA'	'MOTIVO_MUDANCA_ESQUEMA'
'RX_TORAX'		'HIV'	'RESISTENCIA'
'RX_OUTRO'		'TESTE_SENSIBILIDADE'	'TIPO_TRATAMENTO'

Fonte: (Própria autora, 2023)

Por fim, foi realizada a normalização destas variáveis, ou seja, foram transformadas para uma mesma escala, possibilitando realizar as análises de correlação e aplicar os modelos de machine learning. Esta normalização foi realizada por meio da função MinMaxScaler() - biblioteca scikit-learn.

5.2.3 Análise de Correlação

A correlação é definida no campo da Estatística, como um método para medir a associação linear entre variáveis contínuas e sua medida é chamada de coeficiente de correlação, que demonstra a força desta associação, podendo ser um valor entre -1 e 1. Vale

destacar que quanto mais próximo de 1, maior o grau de correlação positiva entre as variáveis e quanto mais próximo do -1, maior o grau de correlação negativa. No caso de ser igual a ambas, demonstra uma relação perfeita e igual a 0, não há nenhuma correlação linear (MUKAKA, 2012). No entanto, podem existir correlações não lineares ou complexas que este tipo de cálculo, muitas vezes não detecta.

Neste cenário, apliquei a Correlação de Pearson, em que dada duas variáveis em um população, é calculada por meio da razão entre a covariância dessas duas variáveis e o produto dos seus desvios padrão (**σ**) como pode ser visto na equação 1 abaixo.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \qquad (1)$$

Os resultados foram que nenhuma variável possui correlação linear significativa com nossa variável alvo que representa os desfechos (‘Situacao_Atual’). Como pode ser observado no Apêndice D, os valores ficaram entre 30 < 0 < -30, indicando ‘correlação insignificante’.

Isso nos mostra que a relação entre nossas variáveis preditoras e a variável alvo não é linear, como na grande maioria das aplicações reais, não indicando necessariamente que estas variáveis não possuem nenhum tipo de relação ou padrão. Vamos explorar a seguir o reconhecimento de padrões na abordagem não supervisionada e começar a identificar os perfis dos desfechos inesperados de Tuberculose entre os pacientes.

5.2.4 Abordagem não supervisionada

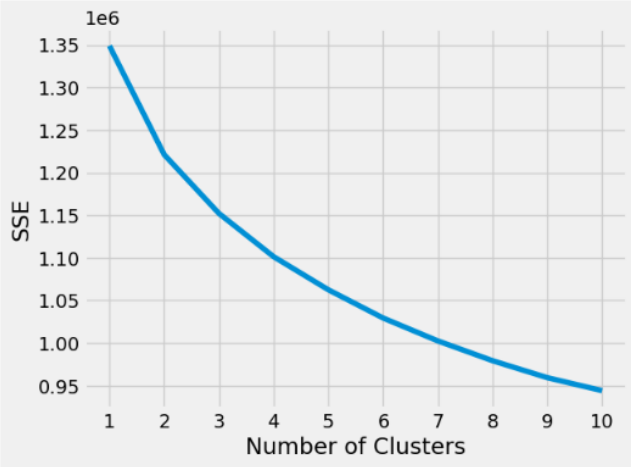
5.2.4.1 K-Means

O principal objetivo da escolha deste tipo de abordagem é buscar semelhanças nas características sociodemográficas e clínicas que levam a agrupar os dados, de forma a identificar essas semelhanças com os desfechos de TB e verificarmos aqueles grupos com maior risco de desenvolver um desfecho ruim.

Para agrupar os pacientes, o método de clusterização K-means, necessita que o número de clusters (k) seja especificado antes de aplicar o algoritmo. Dessa forma, definimos um número k = 5, considerando o número de desfechos. Depois o algoritmo irá iniciar os centróides em posições aleatórias e associar o dado mais próximo (menor distância) a cada centróide e irá recalcular a posição do centróide considerando a média das distâncias de todos os dados até ele. Estes passos irão se repetir até que a média comece a ficar constante.

Após este modelo preliminar, foi realizada a validação utilizando o método do cotovelo (Elbow Method), em que é feito vários testes com diferentes valores de k e calculado a soma do erro quadrático (SSE) por cluster, e então ambos são comparados e a diferença mais extrema, formando o ângulo do cotovelo mostra o melhor número de cluster. Os resultados mostram um leve ângulo no k=4, como visto na figura 2. Uma última validação neste resultado, utilizei a biblioteca Kneed e a função KneeLocator(), retornando realmente o valor 4.

Figura 2 - Resultado do Método do cotovelo



Fonte: (Própria autora, 2023)

Dessa forma, o modelo foi re-treinado com k=4 e foram observados 72.152 pacientes no **cluster 0** (39.73%), 50.079 no **cluster 1** (27.57%), 40.064 no **cluster 3** (22.06%) e 19.310 no **cluster 2** (10.63%). No **cluster 0**, foram observados 21.34% de pacientes com desfechos ruins, sendo eles 'Perda de seguimento','Óbito', 'Não avaliado' e 'Falha no tratamento', no **cluster 1** foram 25.67%, no **cluster 2** cerca de 48.42% e no **cluster 3** foram 18.46%.

Recordando nosso objetivo, temos como destaque o **cluster 3** com a menor porcentagem de desfechos ruins, indicando um possível perfil a ter sucesso no tratamento e o **cluster 2** com a maior porcentagem e consequentemente o maior risco de desenvolvimento de desfechos ruins no tratamento.

Na tabela 6, é apresentado algumas informações relevantes para o entendimento do perfil de maior risco. Nesta segmentação, foi identificada que 12.752 pacientes estão entre 30 e 49 anos (66.03%), destes em grande maioria são do sexo masculino (74.05%), possuem resultado positivo para Aids (98.98%) e HIV (99.67%). Outro fato interessante desta amostra de pacientes é que parte significativa estava em situação grave, com a descoberta do caso sendo por elucidação diagnóstica em internação (38.03%) e Urgência/Emergência (20%), e quando analisamos os desfechos negativos presentes mais de 20% é óbito.

Tabela 6 – Cluster 2 (n=19310)

Faixa Etária		**Diabetes**	
Menor de 1 ano	35	Sim	362
01-04	19	Não	18948
05-09	34	**Doença mental**	
10-14	75	Sim	300

15-19	262	Não	19010
20-29	3287	**Aids**	
30-39	6898	Sim	19111
40-49	5854	Não	199
50-59	2293	**HIV**	
60-69	473	Sim	19238
70-79	75	Não	17
Maior de 80 anos	5	**Outra doença autoimune**	
Sexo		Sim	106
Feminino	5087	Não	19204
Masculino	14223	**Classificação anatômica**	
Raça		Pulmonar	11839
Branco	7926	Extrapulmonar	4669
Pardo	5297	Pulmonar + Extrapulmonar	2323
Preto	2213	Disseminado	479
Amarelo	56	**Baciloscopia (Status Microbiológico)**	
Indígena	24	Positivo	6587
Escolaridade		Negativo	7467
Nenhuma	360	**Descoberta do caso**	

De 1 a 3 anos	1348	Demanda ambulatorial	6984
De 4 a 7 anos	5636	Urgência/ Emergência	3853
De 8 a 11 anos	5580	Elucidação Diagnóstica em Internação	7443
De 12 a 14 anos	970	Busca Ativa em Instituição	365
15 anos ou mais	389	Investigação de contatos	103
Detento		Busca Ativa na Comunidade	97
Sim	1566	Descoberta após o óbito	303
Não	17745	**Desfechos**	
Uso de Álcool		Cura	9960
Sim	3037	Perda de seguimento	3706
Não	16273	Falha tratamento	271
Uso de drogas		Óbito	4124
Sim	3598	Não avaliado	1249
Não	15712	**Tabagismo**	
		Sim	1818
		Não	17492

Fonte: Própria autora (2023)

Em relação ao **cluster 3**, conforme podemos ver na tabela 7, há a presença de pacientes mais jovens com idade entre 20 e 39 anos (45.46%), destes 39.464 são do sexo feminino (98.50%), ao contrário do cluster 2, também observamos poucos casos com Aids (0.26%) e HIV (0.82%) positivos. Em relação aos desfechos, mais de 80% são Cura e no lado dos desfechos negativos, o foco desta busca, a perda de seguimento se sobressai com 8.13%, apesar de ser um valor baixo, ainda é relevante para identificarmos com antecedência os pacientes futuros com características semelhantes, para que não aumente este número.

Tabela 7 – Cluster 3 (n = 40064)

Faixa Etária		**Diabetes**	
Menor de 1 ano	208	Sim	2932
01-04	596	Não	37132
05-09	487	**Doença mental**	
10-14	891	Sim	686
15-19	3191	Não	39378
20-29	10210	**Aids**	
30-39	8003	Sim	106
40-49	6075	Não	39958
50-59	5067	**HIV**	
60-69	2972	Sim	330
70-79	1646	Não	34121
Maior de 80 anos	718	**Outra doença autoimune**	
Sexo		Sim	764

Feminino	39464	Não	39300
Masculino	600	**Classificação anatômica**	
Raça		Pulmonar	30915
Branco	18789	Extrapulmonar	7879
Pardo	11862	Pulmonar + Extrapulmonar	1184
Preto	4071	Disseminado	86
Amarelo	479	**Baciloscopia (Status Microbiológico)**	
Indígena	308	Positivo	19299
Escolaridade		Negativo	9794
Nenhuma	1899	**Descoberta do caso**	
De 1 a 3 anos	3282	Demanda ambulatorial	21210
De 4 a 7 anos	9543	Urgência/ Emergência	8202
De 8 a 11 anos	12916	Elucidação Diagnóstica em Internação	6834
De 12 a 14 anos	3422	Busca Ativa em Instituição	482
15 anos ou mais	1792	Investigação de contatos	1905
Detento		Busca Ativa na Comunidade	995
Sim	351	Descoberta após o óbito	284
Não	39713	**Desfechos**	

Uso de Álcool		Cura	32667
Sim	2157	Perda de seguimento	3261
Não	37907	Falha tratamento	416
Uso de drogas		Óbito	1971
Sim	2313	Não avaliado	1749
Não	37751	**Tabagismo**	
		Sim	3303
		Não	36761

Fonte: Própria autora (2023)

5.2.5 Abordagem supervisionada

5.2.5.1 AutoML vs. Modelos tradicionais

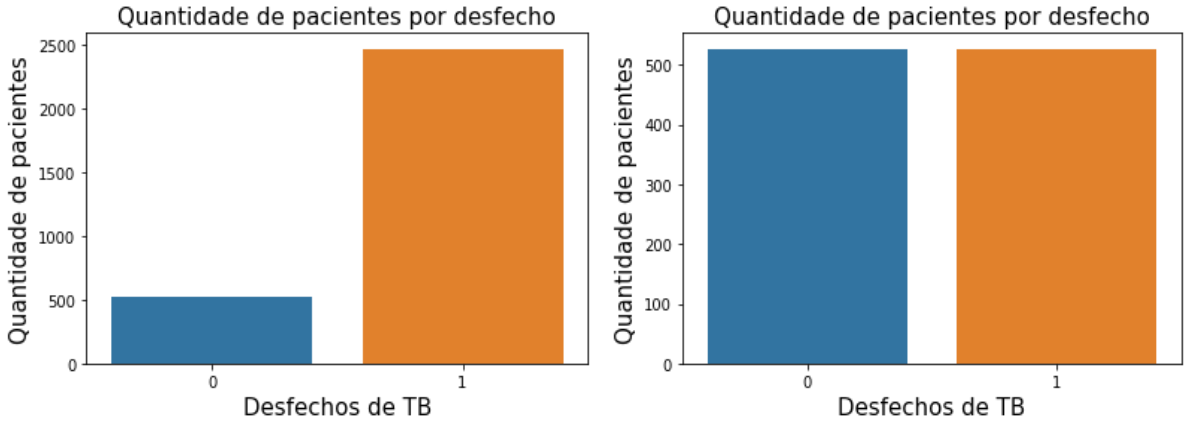
Em um estudo preliminar, foram testados métodos clássicos de ML em relação a uma abordagem de AutoML para detectar resultados de TB, tentando avaliar e discutir como o AutoML pode ser útil no contexto de dados de TB. O principal objetivo desta análise é melhorar a previsão de alguns desfechos de TB. Com isso, espera-se ser capaz de detectar o abandono do tratamento antes que ele aconteça e prever riscos de morte por TB, considerando apenas as informações precoces do paciente, coletadas no início do tratamento.

Foi realizada uma amostra aleatória, do banco de dados do TBWEB tratado, contendo 3000 pacientes para utilizarmos como teste nestes modelos. Dando continuidade, os desfechos foram agrupados em ‘Bons’ (1 = cura, em tratamento ambulatorial e em tratamento internado) e ‘Ruins’ (0 = abandono, óbito, mudança de diagnóstico, transferência, resistência, mudança de esquema, faltoso e sem informação), neste primeiro momento para ajudar no desenvolvimento de uma variável de saída binária, facilitando o entendimento da máquina, por estar em valores numéricos e demandando menos custo computacional.

Após a amostragem, observou-se que nosso atributo classificador - desfecho está desbalanceado, ou seja, apenas 18% dos pacientes havia desfecho não desejável, como pode ser visto na figura 3. Para evitarmos que o aprendizado da máquina seja enviesado e seu

desempenho afetado, este desbalanceamento foi tratado com a técnica Near Miss (método *under-sampling),* que reduz a variável com maior valor para a mesma quantidade da de menor valor, por meio do método dos K-vizinhos mais próximos, em seguida os dados foram normalizados (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017).

Figura 3 - Antes e depois da técnica Near Miss



Fonte: (Própria autora, 2022)

Nesta etapa foi utilizada a abordagem supervisionada no desenvolvimento dos modelos, então utilizamos a tarefa de classificação, em que queremos a partir de dados de novos pacientes, prever seu possível desfecho. Foram testados os modelos KNN (K-vizinhos mais próximos) e Random Forest (floresta aleatória), ambos são destaque na literatura (ALANAZI, ABDUL, KASHIF, 2017; DEGRUY, 2000) e têm diversas vantagens em relação ao seu desempenho e sensibilidade aos dados, e comparados seus desempenhos com a abordagem automatizada (*Automated Machine Learning - AutoML*). Todos os modelos foram treinados com 70% do conjunto e 30% restante foi para teste e validação, essa divisão foi aleatória e garantida a reprodutibilidade pelo parâmetro ‘random_state = 42’.

Os resultados foram bem interessantes, a acurácia do modelo KNN foi relativamente baixa, com 69%, enquanto do Random Forest foi de 78%, mas em relação a sensibilidade (métrica recall) e F1-score, métricas importantes quando estamos trabalhando com dados de saúde, apresentou 81% e 82% para KNN, 83% e 79% para o Random Forest, respectivamente, como pode ser visto na tabela 8.

Tabela 8 - Resultados abordagem supervisionada

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-Score**

K-vizinhos mais próximos (KNN)	0.69	0.65	0.81	0.72
Random Forest	0.78	0.75	0.83	0.79

Fonte: (Própria autora, 2022)

A biblioteca PyCaret (ALI, 2020), utilizada para criar o modelo automatizado, tem uma opção de balanceamento de dados, basta colocar como 'True' no código. Dessa forma, o modelo escolheu utilizar a técnica SMOTE (*over-sampling*), diferente do que realizamos com os modelos anteriores e apresentou os 3 melhores modelos com resultados acima do esperado, como indicado na tabela 9.

Tabela 9 - Resultados abordagem AutoML

Melhores Modelos	**Acurácia**	**Precisão**	**Recall**	**F1-Score**
Random Forest Classifier	0.8499	0.8564	0.9839	0.9157
Extra Trees Classifier	0.8437	0.8561	0.9753	0.9118
Light Gradient Boosting Machine	0.8552	0.8864	0.9465	0.9154

Fonte: (Própria autora, 2022)

5.3 LPA

Entre os meses de fevereiro de 2021 e março de 2022 foram registrados no estudo 607 pacientes, sendo este valor resultante do total de 1ª amostra realizada. Destes, temos ainda 100 pacientes sem informação e apenas 15 (2,47%) com 2ª amostra realizada.

Considerando apenas os pacientes da 1ª amostra, 254 possuem baciloscopia positiva (41,85%), 153 negativa (25,20%), 109 não realizada (17,95%) e 91 sem informação (15%), ou seja, temos ainda 32,95% dos pacientes sem resultados para baciloscopia. Na ocorrência dos casos positivos de baciloscopia, obtivemos que 95 pacientes possuem 1º grau de baciloscopia positiva, 50 pacientes com 2º grau e 36 pacientes com 3º grau.

Ao aplicarem o teste de resistência Xpert, padrão ouro, nestes 607 pacientes, foram confirmados 64 pacientes resistentes à rifampicina, enquanto no teste LPA1, foram obtidos apenas 486 pacientes com TB confirmada, e destes apenas 49 foram positivos para resistência.

Em relação aos sensíveis temos 521 no Xpert e 395 no LPA1. Em relação a Isoniazida, o LPA1 identificou 42 pacientes resistentes e 28 pacientes foram identificados como resistentes a ambos os tratamentos.

Já para o LPA2, apenas 71 pacientes foram confirmados com TB e 87,80% (533) dos 607 pacientes ainda não possuem informação referente a este teste, porém considerando os casos confirmados, obtivemos 9 com resistência a fluoroquinolonas e 3 à amicacina.

A pesquisa segue em andamento com os seguintes passos sendo desenvolvidos e aperfeiçoados: (i) aumentar o período da análise do estudo, com intuito de aumentar a nossa amostra de pacientes, (ii) analisar por centros clínicos e laboratoriais os resultados obtidos do LPA1 e LPA2, (iii) verificar a concordância (Kappa) do teste Xpert com o LPA no geral e por centro.

6 CONCLUSÃO

É muito importante destacar que ainda existem diversas barreiras para que tenhamos o modelo perfeito, uma das possíveis razões é que os atributos aqui usados podem não ser preditivos, ou seja, os dados do paciente registrados no início do tratamento podem não ser suficientes para prever um desfecho ruim, outra razão é a complexidade e qualidade que estão envolvidos no processo de obtenção destes dados, uma vez que há todo um caminho para o paciente percorrer antes de fato ter um desfecho definido e consequentemente estes dados vão o acompanhando (SANTOS et al., 2012). Um ponto também relevante, é que estamos trabalhando com uma amostra referente a pacientes do Estado de São Paulo, infelizmente não sendo uma quantidade representativa dos dados mundiais.

Apesar disso, nossos resultados mostram que é possível prever o desfecho final do tratamento de TB de forma satisfatória, destacando-se o AutoML como ferramenta promissora, e que podemos trabalhar em cima destas barreiras para buscar um excelente modelo. Destaca-se que nas tabelas, os melhores modelos são *ensembles*, considerando o recall como principal métrica e assim entramos numa questão importante em relação a interpretabilidade do modelo, que é extremamente relevante no domínio da saúde (CARUANA et al., 2015; FREITAS, 2019). Os resultados do modelo não supervisionado, também mostraram ser promissores, trazendo diversos insights e entendimentos sobre os perfis de pacientes que tendem a ter um desfecho ruim durante o tratamento de TB, como por exemplo a caracterização do cluster 3 que tende a ser o de menor risco a ter desfechos inesperados.

Para trabalhos futuros deste projeto, é sugerido a implantação em formato de Teste A/B, ou seja, grupos de pacientes selecionados de forma aleatória, para acompanhamento e validação dos modelos aqui desenvolvidos e assim verificarmos o uso entre profissionais da saúde em Tuberculose e quanto a probabilidade está sendo próxima do real.

7 PRODUÇÕES CIENTÍFICAS

1. Ana Clara de Andrade Mioto, Newton Shydeo Brandão Miyoshi, Filipe Andrade Bernardi, Victor Cassão, Domingos Alves. *Unsupervised machine learning techniques in the analysis of bad outcomes in the treatment of tuberculosis: a research protocol*. Artigo aceito para publicação no **BOOK OF INDUSTRY PAPERS, POSTER PAPERS AND ABSTRACTS OF HCIST 2021** – International Conference on Health and Social Care Information Systems and Technologies. No Anexo B, segue a cópia do artigo.

2. Victor Cassão, Domingos Alves, Ana Clara de Andrade Mioto, Filipe Andrade Bernardi, Newton Shydeo Brandão Miyoshi. Unsupervised analysis of COVID-19 evolution in brazilian states. **Procedia Computer Science**, v. 196, p. 525-532, 2022. http://dx.doi.org/10.1016/j.procs.2021.12.061.

3. Isabelle Carvalho, Newton Shydeo Brandão Miyoshi, Mariane Barros Neiva, Nathalia Yukie Crepaldi, Filipe Andrade Bernardi, Vinícius Costa Lima, Ketlin Fabri dos Santos, Ana Clara de Andrade Mioto, Mariana Tavares Mozini, Rafael Mello Galliez, Mauro Niskier Sanchez, Domingos Alves. Knowledge Discovery in Databases: Comorbidities in Tuberculosis Cases. **Lecture Notes in Computer Science**. v. 13352, p. 3-13, 2022. https://doi.org/10.1007/978-3-031-08757-8_1.

4. Ana Clara de Andrade Mioto, Mariana Tavares Mozini, Renan Barbieri Segamarchi, Giovane Thomazini Soares, Pedro Emilio Andrade Martins, Victor Cassão, Luís Gustavo Barichello Ferrassini, Newton Shydeo Brandão Miyoshi, Domingos Alves, Lariza Laura de Oliveira. Preliminary Results to Predict Tuberculosis Outcomes Applying Traditional and Automated Machine Learning Models. **Procedia Computer Science.** v. 219, p. 1365-1372, 2023. https://doi.org/10.1016/j.procs.2023.01.424.

5. Giovane Thomazini Soares, Diego Bettiol Yamada, Filipe Andrade Bernardi, Mariane Barros Neiva, Luiz Pedro Lombardi Junior, André Luiz Teixeira Vinci, Ana Clara de Andrade Mioto, Domingos Alves. Scaling laws and spatial effects of Brazilian health regions: a research protocol. **Procedia Computer Science**, v. 219, p. 1325-1332, 2023. https://doi.org/10.1016/j.procs.2023.01.417.

6. Pedro Emilio Andrade Martins, Márcio Eloi Colombo Filho, Ana Clara de Andrade Mioto, Filipe Andrade Bernardi, Vinícius Costa Lima, Têmis Maria Félix, Domingos Alves. Supervised Machine Learning Techniques Applied to Medical Records Toward the Diagnosis of Rare Autoimmune Diseases. **Lecture Notes in Computer Science.** v. 10475, p. 170–184, 2023. https://doi.org/10.1007/978-3-031-36024-4_13.

7. Victor Cassão, Filipe Andrade Bernardi, Vinícius Costa Lima, Giovane Thomazini Soares, Newton Shydeo Brandão Miyoshi, Ana Clara de Andrade Mioto, Afrânio Kritski, Domingos Alves. A Web Portal for Real-Time Data Quality Analysis on the Brazilian Tuberculosis Research Network: A Case Study. **Lecture Notes in Computer Science.** v. 10475, p. 300–312, 2023. https://doi.org/10.1007/978-3-031-36024-4_24.

8. Ana Clara de Andrade Mioto, Pedro Emilio Andrade Martin, Gabriel Modina, Domingos Alves, Vinicius Costa Lima, Mariane Neiva, Filipe Andrade Bernardi. Quality analysis and study of tuberculosis diagnostic data. Submissão e aceite Apresentação Oral e publicação no HCist - International Conference on Health and Social Care Information Systems and Technologies. Apresentação à ser realizada em Novembro de 2023.

9. Mariana Tavares Mozini, Raul Rothschild, Ana Clara de Andrade Mioto, Filipe Andrade Bernardi, Vinicius Costa Lima, Giovane Thomazini Soares, Renan Barbieri Segamarchi, Domingos Alves. OUTB: Application for decision-support in the outcomes of Tuberculosis. Submissão e aceite Apresentação Oral e publicação no HCist - International Conference on Health and Social Care Information Systems and Technologies. Apresentação à ser realizada em Novembro de 2023.

REFERÊNCIAS

AROMATARIS E., MUNN Z. **JBI Manual for Evidence Synthesis**. JBI, 2020. Disponível em: <https://synthesismanual.jbi.global/>

ARROYO, L. H; RAMOS, A. C. V.; YAMAMURA, M.; BERRA, T. Z.; ALVES, L. S.; BELCHIOR, A. S., et al. **Predictive model of unfavorable outcomes for multidrug-resistant tuberculosis.** Rev Saude Publica. p. 53-77, 2019.

BAYER, M. **The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks**. 2016. Disponível em: <https://www.aosabook.org/en/sqlalchemy.html>

BOUGHORBEL, S.; JARRAY, F.; EL-ANBARI, M. **Optimal classifier for imbalanced data using matthews correlation coefficient metric.** PloS one, n.12, 2017. DOI: https://doi.org/10.1371/journal.pone.0177678

BREIMAN, L. **Random forests**. Mach. Learn. n.45, p.5–32, 2001. DOI: 10.1023/A:1010933404324.

CARUANA, R.; LOU, Y.; GEHRKE, J.; KOCH, P.; STURM, M.; ELHADAD, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *In:* INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING**. Proceedings of the 21th ACM SIGKDD**. 2015. p. 1721–1730.

COLLINS G.S.; REITSMA J.B.; ALTMAN D.G.; MOONS K.G. **Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis** (TRIPOD): the TRIPOD statement. Ann Intern Med.; v. 162, n. 1, p. 55-63, 2015.

FEURER, M.; KLEIN, A.; EGGENSPERGER, K.; SPRINGENBERG, J.; BLUM, M.; HUTTER, F. Efficient and robust automated machine learning. *In:* PROC. ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2015, p. 2962-2970.

FREITAS, A. A. Automated machine learning for studying the trade-off between predictive accuracy and interpretability. *In:* INTERNATIONAL CROSS-DOMAIN CONFERENCE FOR MACHINE LEARNING AND KNOWLEDGE EXTRACTION. Springer, 2019. p. 48–66.

FUNDAÇÃO DE AMPARO À PESQUISA DO ESTADO DE SÃO PAULO. Programa FAPESP de Pesquisa em eScience e Data Science. 2013. Disponível em: <https://fapesp.br/escience/>

GAMA, J., CARVALHO, A.P.L., FACIOLI, K., LORENA, A.C., OLIVEIRA, M. Extração de Conhecimento de Dados - Data Mining. Edições Sílabo, 2012.

GLAZIOU, P.; FLOYD, K.; RAVIGLIONE, M. C. Global epidemiology of tuberculosis. **SEMINARS IN RESPIRATORY AND CRITICAL CARE MEDICINE**. v. 39, n. 3, p. 271-285, 2018.

HAIXIANG, G. et al. Learning from class-imbalanced data: Review of methods and applications. **Expert. Syst. with Appl.** n. 73, p. 220–239, 2017.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. **Computing in Science & Engineering.** v. 9, n. 3, p. 90-95, 2007.

KALHORI, S.R.N., ZENG, X.J. Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course. **Journal of Intelligent Learning Systems and Applications**, n. 5, p. 184-193, 2013.

KOO, H.K., Min, J., Kim, H.W. et al. Cluster analysis categorizes five phenotypes of pulmonary tuberculosis. **Scientific Reports**, n. 12, 10084, 2022.

LASERSON, K. F. et al. Speaking the same language: treatment outcome definitions for multidrug-resistant tuberculosis. **The International Journal of Tuberculosis and Lung Disease**, v. 9, n. 6, p. 640-645, 2005.

LI, D.-C., LIU, C.-W. & HU, S. C. A learning method for the class imbalance problem with medical data sets. **Comput. biology medicine**. n. 40, p. 509–518, 2010.

MUKAKA, M.M. Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research. **Malawi Medical Journal.** n. 24, p. 69-71, 2012.

MCKINNEY, WES. Data Structures for Statistical Computing in Python. *In:* **Proceedings of the 9th Python in Science Conference**. n. 445, p. 56-61, 2010.

MOHRI, M.; AFSHIN, R.; AMEET, W.T. **Foundations of machine learning.** MIT press, 2018.

MINISTÉRIO DA SAÚDE; SECRETARIA DE VIGILÂNCIA EM SAÚDE; DEPARTAMENTO DE VIGILÂNCIA DAS DOENÇAS TRANSMISSÍVEIS. Manual de Recomendações e Controle da Tuberculose no Brasil. n. 2, p. 320-330, 2019. Disponível em:<file:///C:/Users/anacl/Downloads/Manual%20de%20Recomendacoes%20e%20Controle%20da%20Ttuberculose%20no%20Brasil%202%C2%AA%20ed.pdf>

MINISTÉRIO DA SAÚDE; SECRETARIA DE VIGILÂNCIA EM SAÚDE; DEPARTAMENTO DE VIGILÂNCIA DAS DOENÇAS TRANSMISSÍVEIS. Vigilância epidemiológica da tuberculose: Análise de indicadores operacionais e epidemiológicos a partir da base de dados do Sinan versão 5.0. 2016. Disponível em: <http://portalsinan.saude.gov.br/images/documentos/Agravos/Tuberculose/Apostila_Curso_Sinan_2016.pdf>

MINISTÉRIO DA SAÚDE; SECRETARIA DA SAÚDE; SECRETARIA DE VIGILÂNCIA EM SAÚDE. Validação e Custo da performance do *Line Probe Assay* como método de diagnóstico rápido para tuberculose resistente em centros de referência no Brasil. **VIII ENCONTRO CIENTÍFICO DE PESQUISAS APLICADAS À VIGILÂNCIA EM SAÚDE**. 2021.

NERY, J. S., et al. Effect of Brazil's conditional cash transfer programme on tuberculosis incidence. **The international journal of tuberculosis and lung disease**. v. 21, n. 7, p. 790-796, 2017.

OLIOSI, J. G. N., et al: Effect of the Bolsa Familia Programme on the outcome of tuberculosis treatment: a prospective cohort study. **The Lancet**, p. 219–226, 2019.

OLSON, J. Data Quality -The Accuracy Dimension. **Morgan Kaufmann**, São Francisco, 2003.

PAIXÃO, L.M.M.; GONTIJO, E.D. Perfil de casos de tuberculose notificados e fatores associados ao abandono, Belo Horinzonte, MG. **Revista de Saúde Pública**, v. 41, n. 2, p. 205-213, 2007.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**. n. 12, p. 2825-2830, 2011.

PEETLUK L.S., RIDOLFI F.M., REBEIRO P.F., et al. Systematic review of prediction models for pulmonary tuberculosis treatment outcomes in adults. **BMJ Open**, 2021. DOI:10.1136/bmjopen-2020-044687

PINHEIRO, C. A. R.; PATETTA, M. **Introduction to Statistical and Machine Learning Methods for Data Science**. SAS Institute Inc. 2021.

SANTOS, M. L. S. G.; et al. Poverty: socioeconomic characterization at tuberculosis. **Revista latino-americana de enfermagem.** v. 15, p. 762-767, 2007.

SECRETARIA DE SAÚDE DO ESTADO DE SÃO PAULO; CENTRO DE VIGILÂNCIA EPIDEMIOLÓGICA PROF ALEXANDRE VRANJAC. Manual de utilização do TBWEB. 2008. Disponível em: <http://www.saude.sp.gov.br/resources/cve-centro-de-vigilanciaepidemiologica/areas-de-vigilancia/tuberculose/manuais tecnicos/dvtbc_tbweb_2008.pdf.>

SILVA, E.A., ANJOS, U.U., NOGUEIRA, J.A. Modelo preditivo ao abandono do tratamento da tuberculose. **Revista Saúde Debate**. Rio de Janeiro, v. 38, n. 101, p. 200-209, 2014.

SUN, Y., KAMEL, M. S., WONG, A. K. & WANG, Y. Cost-sensitive boosting for classification of imbalanced data. **Pattern Recognit.** v. 40, p. 3358–3378, 2007.

TORRENS, A. W. et al. Effectiveness of a conditional cash transfer programme on TB cure rate: a retrospective cohort study in brazil.Transactions of the Royal Society of Tropical Medicine and Hygiene, v. 110, n. 3, p. 199–206, mar 2016. Disponível em: <http://dx.doi.org/10.1093/trstmh/trw011>

TANAKA, O. Y.; TAMAKI, E. M. O papel da avaliação para tomada de decisão na gestão de serviços de saúde. **Ciência & Saúde Coletiva**. v. 17, n. 4, p. 821-828, 2012.

WORLD HEALTH ORGANIZATION. Global tuberculosis report 2020: executive summary, 2020.

WORLD HEALTH ORGANIZATION. Definitions and reporting framework for tuberculosis - 2013 revision. *In:* **Annex 2**, TB case and treatment outcome definitions. Geneva, 2014.

YAMAGUTI, V. H. et. al. Development of CART model for prediction of tuberculosis treatment loss to follow up in the state of São Paulo, Brazil: A case–control study. International **Journal of Medical Informatics**. n. 141, 2020.

YAMAGUTI, V. H. et. al. Charlson Comorbidities Index importance evaluation as a predictor to tuberculosis treatments outcome in the state of São Paulo, Brazil. **Procedia Computer Science.** n. 138, p. 258-263, 2018.

YAO, Y. et. al. Identification of spinal tuberculosis subphenotypes using routine clinical data: a study based on unsupervised machine learning. **Annals of Medicine.** n. 55, p. 2, 2023.

APÊNDICE A – Tabela de Variáveis do TBWEB

Id
SINAN
Raca_Cor
Idade
Faixa Etária
Sexo
Gestante
Naturalidade
Escolaridade
Tipo_Ocupacao
Ocupacao
GVE_Residencia
Municipio_Residencia
Unidade_Notificacao
Municipio_Notificacao
GVE_Notificacao
Codigo_Tratamento_Anterior
Tempo_Tramento_Anterior
Data_Notificacao
Data_Inicio_Tratamento
Unidade_Atendimento
Municipio_Atendimento
GVE_Atendimento

Situacao_Atual
Tipo_Caso
Tipo_Encerramento
Data_Encerramento
Forma_Clinica_1
Forma_Clinica_2
Forma_Clinica3
Classificacao
Descoberta
Data_Primeiros_Sintomas
Recebido_de
Baciloscopia_Escarro
Baciloscopia_Outro_Material
Cultura_Escarro
Cultura_Outro_Material
Rx_Torax
Rx_Outro
Histopatologia
Necropsia
HIV
Teste_Sensibilidade
AIDS
DIABETES
ALCOOLISMO
DOENCA_MENTAL

USO_DROGAS
OUTRAS_DOENCAS_IMUNO
TABAGISMO
TOTCOMUNIC
COMUNICEXA
COMUNICDOE
Instituto_Tratamento
Endereco_Tratamento
Nro_Endereco_Tratamento
CEP_Tratamento
Estado_Tratamento
Bairro_Tratamento
Area_Residencia_Tratamento
Municipio_Tratamento
GVE_Endereco_Tratamento
Esquema_Inicial
Mudanca_Esquema
Esquema_Atual
Motivo_Mudanca_Esquema
Resistencia
Tipo_Tratamento
Nro_Doses_Pri
Nro_Doses_Seg
Unidade_Sup
SIT1, SIT2, SIT3, SIT4, SIT5, SIT6

BAC1, BAC2, BAC3, BAC4, BAC5, BAC6, BAC7, BAC8, BAC9, BAC10, BAC11, BAC12
Unidade_Tratamento_1, Unidade_Tratamento_2, Unidade_Tratamento_3,
Unidade_Tratamento_4, Unidade_Tratamento_5, Unidade_Tratamento_6
Data_Solicitacao_Ultima_Transf
Data_Aceitacao_Ultima_Transf'
Origem_Ultima_Transf
Destino_Ultima_Transf
Codigo_Hospital_1
Motivo_Internacao_1
Data_Saida_Hospital_1
Tipo_Saida_1
Data_Internacao_2
Codigo_Hospital_2
Motivo_Internacao_2
Data_Saida_Hospital_2
Tipo_Saida_2
Data_Internacao_3
Codigo_Hospital_3
Motivo_Internacao_3
Data_Saida_Hospital_3
Tipo_Saida_3

APÊNDICE B – Variáveis Formulário 5 LPA

Instrument: **5 Resultados Laboratório** (resultados_laboratrio) — Collapse

	413	record_id_aux5	Record ID Aux	calc Calculation: concat([record_id],'') Field Annotation: @HIDDEN
	414	data_preench_result_lab	Data de preenchimento no REDCap	text (date_dmy), Required Field Annotation: @READONLY @TODAY-SERVER
	415	resp_preench_result_lab	Responsável pelo preenchimento no REDCap	text, Required Field Annotation: @READONY @USERNAME
	416	resp_receb_amostra_lab	Responsável pelo recebimendo da amostra no laboratório	text
	417	numero_amostra_gal	Número de identificação da amostra no GAL	text (integer), Identifier
	418	nro_requisicao_gal	Número de requisição do exame no GAL	text, Identifier
	419	id_lab	N° interno do laboratório	text, Identifier
	420	nro_amostra	Section Header: *Amostra respiratório* Amostra recebida	radio, Required 1_amostra: 1ª Amostra 2_amostra: 2ª Amostra
	421	amostra_recebida	Tipo de amostra respiratória	radio, Required escarro: Escarro escarro_induzido: Escarro induzido lavado_broncoalveolar: Lavado broncoalveolar outro: Outro
	422	outro_met_coleta_am_resp Show the field ONLY if: [amostra_recebida] = 'outro'	Especificar (Tipo de amostra respiratória)	text
	423	data_coleta	Data da coleta	text (date_dmy), Required
	424	data_laboratorio	Data de entrada no laboratório	text (date_dmy), Required
	425	bacilloscopia_lab	Bacilloscopia	radio, Required positiva: Positiva negativa: Negativa nao_realizada: Não realizada na: Não se aplica
	426	grau_posit_bacilloscopia Show the field ONLY if: [bacilloscopia_lab] = 'positiva'	Grau de posititividade da bacilloscopia	radio, Required escasso_1_a_9: Escasso (1 a 9) 1_positivo: 1+ 2_positivo: 2+ 3_positivo: 3+
	427	xpert	Xpert MTB RIF	radio, Required nao_detectado: Não detectado detectado_tracos: Detectado (Traços) - RIF indeterminada detectado_sensivel_rif: Detectado (Sensível à RIF) detectado_resistente_rif: Detectado (Resistente à RIF) detectado_resistente_rif_indeterminada: Detectado (Resistente à RIF indeterminada) invalido: Inválido (para resultados que deram erros ou "no result")
	428	quant_dna_xpert Show the field ONLY if: [xpert] = 'detectado_resistente_rif_indeterminada'	Quantificação de DNA dada pelo Xpert	radio, Required muito_baixo: Muito baixo baixo: Baixo medio: Médio alto: Alto
	429	extracaodna_lpa1	Section Header: *LPA1* Data da Extração do DNA	text (date_dmy), Required
	430	pcr_data_lpa1	Data da PCR	text (date_dmy), Required
	431	dnahibridizacao_lpa1	Data da Hibridização do DNA	text (date_dmy), Required
	432	tb_confirmada_lpa1	Complexo Mycobacterium tuberculosis confirmado?	radio, Required sim: Sim nao_detectado: Não

	433	checkbox_rif_band_lpa1	Padrão de bandas de Rifampicina	checkbox, Required

rpoB_locus	checkbox_rif_band_lpa1___rpob_locus	rpoB locus
rpoB_WT1	checkbox_rif_band_lpa1___rpob_wt1	rpoB WT1
rpoB_WT2	checkbox_rif_band_lpa1___rpob_wt2	rpoB WT2
rpoB_WT3	checkbox_rif_band_lpa1___rpob_wt3	rpoB WT3
rpoB_WT4	checkbox_rif_band_lpa1___rpob_wt4	rpoB WT4
rpoB_WT5	checkbox_rif_band_lpa1___rpob_wt5	rpoB WT5
rpoB_WT6	checkbox_rif_band_lpa1___rpob_wt6	rpoB WT6
rpoB_WT7	checkbox_rif_band_lpa1___rpob_wt7	rpoB WT7
rpoB_WT8	checkbox_rif_band_lpa1___rpob_wt8	rpoB WT8
rpoB_MUT1	checkbox_rif_band_lpa1___rpob_mut1	rpoB MUT1
rpoB_MUT2A	checkbox_rif_band_lpa1___rpob_mut2a	rpoB MUT2A
rpoB_MUT2B	checkbox_rif_band_lpa1___rpob_mut2b	rpoB MUT2B
rpoB_MUT3	checkbox_rif_band_lpa1___rpob_mut3	rpoB MUT3

	434	rif_resistencia_lpa1	Interpretação da resistência à Rifampicina	radio, Required

sensivel	Sensível
resistente	Resistente
indeterminado	Indeterminado
invalido	Inválido
resistencia_inferida	Resistência Inferida
heteroressistente	Heterorresistente
nao_realizado	Não Realizado

	435	checkbox_iso_band_lpa1	Padrão de bandas de Isoniazida	checkbox, Required

katG_locus	checkbox_iso_band_lpa1___katg_locus	katG locus
katG_WT	checkbox_iso_band_lpa1___katg_wt	katG WT
katG_MUT1	checkbox_iso_band_lpa1___katg_mut1	katG MUT1
katG_MUT2	checkbox_iso_band_lpa1___katg_mut2	katG MUT2
inhA	checkbox_iso_band_lpa1___inha	inhA locus
inhA_WT1	checkbox_iso_band_lpa1___inha_wt1	inhA WT1
inhA_WT2	checkbox_iso_band_lpa1___inha_wt2	inhA WT2
inhA_MUT1	checkbox_iso_band_lpa1___inha_mut1	inhA MUT1
inhA_MUT2	checkbox_iso_band_lpa1___inha_mut2	inhA MUT2
inhA_MUT3A	checkbox_iso_band_lpa1___inha_mut3a	inhA MUT3A
inhA_MUT3B	checkbox_iso_band_lpa1___inha_mut3b	inhA MUT3B

	437	resistencia_inf_iso Show the field ONLY if: [iso_resistencia_lpa1(resistencia_inferida)] = '1'	Resistência inferida à isoniazida	checkbox, Required katg, resistencia_inf_iso___katg, katG inha, resistencia_inf_iso___inha, inhA
	438	lpa1_comentarios	Comentários	notes Custom alignment: LH
	439	dnaextracao_lpa2	Section Header: *LPA2* Data da Extração do DNA	text (date_dmy), Required
	440	pcr_data_lpa2	Data da PCR	text (date_dmy), Required
	441	dnahibridizacao_lpa2	Data da Hibridização do DNA	text (date_dmy), Required
	442	tb_confirmada_lpa2	Complexo Mycobacterium tuberculosis confirmado?	radio, Required sim, Sim nao_detectado, Não
	443	checkbox_fluor_band_lpa2	Padrão de bandas de Fluoroquinolonas (Moxifloxacina, Levofloxacina)	checkbox, Required gyrA_locus, checkbox_fluor_band_lpa2___gyra_locus, gyrA locus gyrA_WT1, checkbox_fluor_band_lpa2___gyra_wt1, gyrA WT1 gyrA_WT2, checkbox_fluor_band_lpa2___gyra_wt2, gyrA WT2 gyrA_WT3, checkbox_fluor_band_lpa2___gyra_wt3, gyrA WT3 gyrA_MUT1, checkbox_fluor_band_lpa2___gyra_mut1, gyrA MUT1 gyrA_MUT2, checkbox_fluor_band_lpa2___gyra_mut2, gyrA MUT2 gyrA_MUT3A, checkbox_fluor_band_lpa2___gyra_mut3a, gyrA MUT3A gyrA_MUT3B, checkbox_fluor_band_lpa2___gyra_mut3b, gyrA MUT3B gyrA_MUT3C, checkbox_fluor_band_lpa2___gyra_mut3c, gyrA MUT3C gyrA_MUT3D, checkbox_fluor_band_lpa2___gyra_mut3d, gyrA MUT3D gyrB_locus, checkbox_fluor_band_lpa2___gyrb_locus, gyrB locus gyrB_WT1, checkbox_fluor_band_lpa2___gyrb_wt1, gyrB WT gyrB_MUT1, checkbox_fluor_band_lpa2___gyrb_mut1, gyrB MUT1 gyrB_MUT2, checkbox_fluor_band_lpa2___gyrb_mut2, gyrB MUT2
	444	fluor_resistencia_lpa2	Interpretação da resistência à Fluoroquinolonas (Moxifloxacina, Levofloxacina)	checkbox, Required sensivel, fluor_resistencia_lpa2___sensivel resistente, fluor_resistencia_lpa2___resistente indeterminado, fluor_resistencia_lpa2___indeterminado invalido, fluor_resistencia_lpa2___invalido resistencia_inferida, fluor_resistencia_lpa2___resistencia_inferida heteroressistente, fluor_resistencia_lpa2___heteroressistente nao_realizado, fluor_resistencia_lpa2___nao_realizado

	445	resistencia_inf_fluor Show the field ONLY if: [fluor_resistencia_lpa2(resistencia_inferida)] = '1'	Resistência inferida à Fluoroquinolonas (Moxifloxacina, Levofloxacina)	checkbox, Required gyra, resistencia_inf_fluor___gyra, GyrA gyrb, resistencia_inf_fluor___gyrb, GyrB
	446	checkbox_amino_band_lpa2	Padrão de bandas de Aminoglicosídeos (Amicacina)	checkbox, Required rrs_locus, checkbox_amino_band_lpa2___rrs_locus, rrs locus rrs_WT1, checkbox_amino_band_lpa2___rrs_wt1, rrs WT1 rrs_WT2, checkbox_amino_band_lpa2___rrs_wt2, rrs WT2 rrs_MUT1, checkbox_amino_band_lpa2___rrs_mut1, rrs MUT1 rrs_MUT2, checkbox_amino_band_lpa2___rrs_mut2, rrs MUT2 eis_locus, checkbox_amino_band_lpa2___eis_locus, eis locus eis_WT1, checkbox_amino_band_lpa2___eis_wt1, eis WT1 eis_WT2, checkbox_amino_band_lpa2___eis_wt2, eis WT2 eis_WT3, checkbox_amino_band_lpa2___eis_wt3, eis WT3 eis_MUT1, checkbox_amino_band_lpa2___eis_mut1, eis MUT1
	447	amino_resistencia_lpa2	Interpretação da resistência à Aminoglicosídeos (Amicacina)	radio, Required sensivel, Sensível resistente, Resistente indeterminado, Indeterminado invalido, Inválido resistencia_inferida, Resistência Inferida heteroressistente, Heterorresistencia nao_realizado, Não realizado
	448	lpa2_comentarios	Comentários	notes Custom alignment: LH
	449	data_result_cultura	Section Header: *Cultura* Data do resultado da cultura	text (date_dmy), Required

	450	result_cultura_mgit	Resultado da cultura no MGIT	radio, Required negativa, Negativa positiva, Positiva contaminada, Contaminada sem_dado, Sem dado
	451	id_cultura_microbac Show the field ONLY if: [result_cultura_mgit]='positiva'	Identificação da cultura de micobactérias	radio, Required CMTB, Complexo Mycobacterium tuberculosis - CMTB MNT, Micobactéria não Turbercolosa - MNT Contaminada, Contaminada cult_mista, Cultura Mista - CMTB + MNT
	452	temp_positiv_mgit_hr	Tempo para positividade no MGIT (horas)	text (integer), Required
	453	temp_positiv_mgit_dias	Tempo para positividade no MGIT (dias)	text (integer), Required
	454	data_result_tsa_1linha	Section Header: *TSA 1ª Linha* Data do resultado	text (date_dmy), Required
	455	tsa_rifampicina	Rifampicina - 1,0	radio, Required sensivel, Sensível resistente, Resistente contaminada, Contaminado invalido, Inválido nao_realizado, Não realizado
	456	tsa_rifamp_meio	Rifampicina - 0,5	radio, Required sensivel, Sensível resistente, Resistente contaminada, Contaminado invalido, Inválido nao_realizado, Não realizado

	457	tsa_isoniazida	Isoniazida	radio, Required sensivel, Sensível resistente, Resistente contaminada, Contaminado invalido, Inválido nao_realizado, Não realizado
	458	data_result_tsa_2line	Section Header: *TSA 2ª Linha* Data do resultado do TSA 2ª linha	text (date_dmy), Required
	459	tsa_levofloxacino	Levofloxacino	radio, Required sensivel, Sensível resistente, Resistente contaminada, Contaminado nao_realizado, Não realizado
	460	tsa_moxifloxacino	Moxifloxacino	radio, Required sensivel, Sensível resistente, Resistente contaminada, Contaminado nao_realizado, Não realizado
	461	tsa_amicacina	Amicacina	radio, Required sensivel, Sensível resistente, Resistente contaminada, Contaminado nao_realizado, Não realizado
	462	discordante_sensibilidade	Houve resultados discordantes entre o resultado do LPA e no teste de sensibilidade fenotípico?	radio, Required sim, Sim nao, Não pendente, Pendente - Aguardando resultado
	463	drogas_discordantes Show the field ONLY if: [discordante_sensibilidade]= 'sim'	Fármaco(s) discordante(s):	checkbox, Required rifampicina, drogas_discordantes___rifampicina, Rifampicina isoniazida, drogas_discordantes___isoniazida, Isoniazida moxifloxacina, drogas_discordantes___moxifloxacina, Moxifloxaci levofloxacina, drogas_discordantes___levofloxacina, Levofloxaci amicacina, drogas_discordantes___amicacina, Amicacina
	464	cim_realizado	Section Header: *Concentração Inibitória Mínima (CIM)* CIM realizado?	radio, Required sim, Sim nao, Não
	465	data_teste Show the field ONLY if: [cim_realizado]= 'sim'	Data da realização do teste:	text (date_dmy), Required
	466	data_cultura Show the field ONLY if: [cim_realizado]= 'sim'	Data do recebimento da cultura:	text (date_dmy), Required
	467	rifampicin_cim Show the field ONLY if: [cim_realizado]= 'sim'	Rifampicina	radio, Required sensivel, Sensível resistente, Resistente indeterminado, Indeterminado nao_realizado, Não realizado contaminada, Contaminado
	468	cim_valor_rif Show the field ONLY if: [cim_realizado]= 'sim'	Valor - Rifampicina	text, Required
	469	isoniazid_cim Show the field ONLY if: [cim_realizado]= 'sim'	Isoniazida	radio, Required sensivel, Sensível resistente, Resistente indeterminado, Indeterminado nao_realizado, Não realizado contaminada, Contaminado

	470	cim_valor_isoniazida Show the field ONLY if: [cim_realizado]= 'sim'	Valor - Isoniazida	text, Required
	471	amikacin_cim Show the field ONLY if: [cim_realizado]= 'sim'	Amicacina	radio, Required sensivel, Sensível resistente, Resistente indeterminado, Indeterminado heteroressistente, Heterorresistente nao_realizado, Não Realizado
	472	cim_valor_amica Show the field ONLY if: [cim_realizado]= 'sim'	Valor - Amicacina	text, Required
	473	levofloxacin_cim Show the field ONLY if: [cim_realizado]= 'sim'	Levofloxacino	radio, Required sensivel, Sensível resistente, Resistente indeterminado, Indeterminado nao_realizado, Não realizado contaminada, Contaminado
	474	cim_valor_levoflo Show the field ONLY if: [cim_realizado]= 'sim'	Valor - Levofloxacino	text, Required
	475	moxifloxacin_cim Show the field ONLY if: [cim_realizado]= 'sim'	Moxifloxacino	radio, Required sensivel, Sensível resistente, Resistente indeterminado, Indeterminado nao_realizado, Não realizado contaminada, Contaminado
	476	cim_valor_moxi Show the field ONLY if: [cim_realizado]= 'sim'	Valor - Moxifloxacino	text, Required
	477	seq_sanger	**Section Header:** *Sequenciamento Sanger* Foi realizado sequenciamento Sanger?	radio, Required sim_total, Sim, total sim_parcial, Sim, Parcial nao, Não
	478	isolado_criopreservado	**Section Header:** *Armazenamento* O isolado foi criopreservado?	radio, Required sim, Sim nao, Não
	479	data_cripreservado Show the field ONLY if: [isolado_criopreservado] = 'sim'	Data do cripreservado	text (date_dmy)
	480	obs_5	Observações	notes Custom alignment: LH

	481	alerta_preench_lpa_cult Show the field ONLY if: [rif_resistencia_lpa1] = 'invalido' or [rif_resistencia_lpa1] = 'indeterminado' or [amino_resistencia_lpa2] = 'invalido' or [amino_resistencia_lpa2] = 'indeterminado' or ([iso_resistencia_lpa1(invalido)] = '1' or [iso_resistencia_lpa1(indeterminado)] = '1') or ([fluor_resistencia_lpa2(invalido)] = '1' or [fluor_resistencia_lpa2(indeterminado)] = '1')	ATENÇÃO* Realizar o preenchimento do Formulário LPA1 e LPA2 - Cultura (no REDCap) devido a interpretação inválida e/ou indeterminada em LPA1/LPA2 no preenchimento deste formulário.	descriptive
	482	resultados_laboratrio_complete	**Section Header:** *Form Status* Complete?	dropdown 0, Incomplete 1, Unverified 2, Complete

APÊNDICE C - Descrição das variáveis selecionadas

Atributo	Descrição	Valores possíveis
SINAN	Número da notificação	Qualquer valor numérico
Raca_Cor	Raça	Branco, Pardo, Preto, Amarelo, Indigena, Ignorado
Faixa_Etaria	Faixa de idade	Menor de 1 ano, 01-04, 05-09, 10-14, 15-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, Maior de 80 anos
Sexo	Sexo	F, M
Gestante	Informação sobre a possibilidade de ser gestante	S, N, Ignorado
Naturalidade	Continente de origem do paciente	Brasil – SP, Brasil – Outros Estados, América do Sul, África, Ásia, Oceania, Europa, América do Norte/ Central
Escolaridade	Anos que frequentou a escola	De 1 a 3 anos, De 4 a 7 anos, De 8 a 11 anos, De 12 a 14 anos, 15 anos ou mais, Nenhuma, Ignorado
Tipo_Ocupacao	Qual o tipo da profissão que exerce (geral)	Outra, Desempregado, Detento, Dona de Casa, Aposentado, Profissional de Saúde, Profissional Sistema Penitenciario

Municipio_Residencia	Qual município de residência do paciente	Qualquer município
Codigo_Tratamento_Anterior	Código do tratamento realizado anteriormente	1 - Não tratou (caso novo); 2 - Sim, alta cura (recidiva); 3 - Sim, alta abandono (retratamento); 4 - Não sabe; 5 - Tratamento anterior encerrado por resistência; 6 - (*em estudo*); 9 - Sem informação
Tempo_Tratamento_Anterior	Faz quanto tempo que realizou o tratamento	Valores em anos completos
Situacao_Atual	Desfechos associados ao tratamento	Cura, Abandono, Obito NTB, Obito TB, Mudança de diagnóstico, Falencia/Resistencia, Em tratamento ambulatorial, Tranferencia para outro Estado/País, Abandono Primário, Mudança de esquema intoler/toxicidade, tranferencia, em tratamento internado, faltoso, sem informação, outra
Tipo_Caso	Tipo de caso deste paciente	Novo, Recidiva, Retratamento Abandono, Retratatamento após falencia/resistencia, retratamento após mudança de esquema int/tox

Forma_Clinica_1, Forma_Clinica_2 e Forma_Clinica3	Forma clínica da TB	Pulmonar, pleural, ganglionar periferica, meningea, miliar, oftalmica, ossea, multiplos órgãos, vias urinárias, pele, intestinal, genital, laringea, outras
Classificacao	Classificação da TB	Pul, Ext, P+E, Dissem
Descoberta	Forma de descoberta da TB	Demanda Ambulatorial, Urgencia/Emergencia, Elucidadcao Diagn. em Internacao, Busca Ativa em Instituicao, Investigacao de Contatos, Busca Ativa na Comunidade, Descob. Apos Obito, S/inf
Baciloscopia_Escarro	Baciloscopia	Pos, Neg, N/realiz, S/inf, And
Baciloscopia_Outro_Material	Baciloscopia de outro material	Pos, Neg, N/realiz, S/inf, And
Cultura_Escarro	Cultura de escarro	Pos, Neg, N/realiz, S/inf, And
Cultura_Outro_Material	Cultura de outro material	Pos, Neg, N/realiz, S/inf, And
Rx_Torax	Raio X do tórax	Susp TB, N/realiz, Susp c/cavid, Normal, Outra Patologia, S/inf, Pos
Rx_Outro	Raio X de outra parte do corpo	Susp TB, N/realiz, Susp c/cavid, Normal, Outra Patologia, S/inf, Pos

Histopatologia	Resultado do teste histopatológico	N/realiz, Sugestivo TB, S/inf, BAAR pos, Outra Patologia
Necropsia	Resultado da necrópsia	N/realiz, Sugestivo TB, S/inf, BAAR pos, Outra Patologia
HIV	Teste de HIV	Neg, Pos, N/realiz, And, S/inf
Teste_Sensibilidade	Teste de sensibilidade	S, N, S/inf
AIDS	Teste de Aids	S, N, Pos
DIABETES	Se há Diabetes	S, N
ALCOOLISMO	Se há alcoolismo	S, N
DOENCA_MENTAL	Se há alguma doença mental	S, N
USO_DROGAS	Se há uso de drogas	S, N
OUTRAS_DOENCAS_IMUNO	Se há presença de outras doenças imunológicas	S, N
TABAGISMO	Se há tabagismo	S, N
TOTCOMUNIC	Total de pessoas que residem com o paciente	Qualquer valor numérico
COMUNICEXA	Número de pessoas que residem com o paciente que compareceram para o exame	Qualquer valor numérico
COMUNICDOE	Quantas pessoas que residem com o paciente e adoeceram de tuberculose	Qualquer valor numérico

Municipio_Tratamento	Município que deu continuidade ao tratamento	
Esquema_Inicial	Esquema medicamentoso incial	RHZE, RHZ, MR, SZEET, OUTROS
Mudanca_Esquema	Se houve mudança no esquema	N, S
Esquema_Atual	Esquema medicamentoso atual	RHZE, RHZ, MR, SZEET, OUTROS
Motivo_Mudanca_Esquema	No caso de haver mudança, qual o motivo	Intolerancia/Toxicidade, Resistência Medicamentosa, Outro Motivo
Resistência	Verificação da resistência	SENS, AND, TB R, TB MR
Tipo_Tratamento	Tipo de tratamento	Supervisionado, Auto-Administrado, S/inf
Nro_Doses_Pri	Número de doses até o 2º mês do medicamento	Qualquer valor numérico
Nro_Doses_Seg	Número de doses do 3º ao 6º mês do medicamento	Qualquer valor numérico
Motivo_Internacao_1, Motivo_Internacao_2, Motivo_Internacao_3	Qual o motivo que levou o paciente a ser internado durante o tratamento	Elucidação diagnóstica, Insuficiência Respiratória Aguda, Outros, Causas Sociais, Aids, Caquexia, Hemoptise, Não adesão ao tratamento, Meningite, Intolerância Medicamentosa, TB miliar, Abscesso, Diabetes, Alta para tratamento ambulatorial, S/Inf

Tipo_Saida_1, Tipo_Saida_2, Tipo_Saida_3	Qual foi o tipo de desfecho da internação durante o tratamento	Alta para tratamento ambulatorial, Óbito por outra causa, Óbito por TB, Transferência para outro hospital, Evadiu-se, Cura, A pedido, Mudança Diagnóstica, Disciplinar

APÊNDICE D - Correlação de Pearson

SITUACAO_ATUAL	1.000000
CODIGO_TRATAMENTO_ANTERIOR	0.116957
TEMPO_TRAMENTO_ANTERIOR	0.002256
TOTCOMUNIC	-0.079414
COMUNICEXA	-0.092830
COMUNICDOE	-0.031230
NRO_DOSES_PRI	-0.202628
NRO_DOSES_SEG	-0.305435
RACA_COR	-0.007719
FAIXA_ETARIA	0.136084
NATURALIDADE	0.005214
ESCOLARIDADE	0.039373
TIPO_OCUPACAO	-0.047570
MUNICIPIO_RESIDENCIA	0.067712
FORMA_CLINICA_1	-0.046093
FORMA_CLINICA_2	-0.070178
FORMA_CLINICA3	-0.024021
DESCOBERTA	0.071616
RX_TORAX	0.000411
RX_OUTRO	0.038843
MUNICIPIO_TRATAMENTO	0.072668
ESQUEMA_ATUAL	0.015234
MOTIVO_INTERNACAO_1	-0.201810
TIPO_SAIDA_1	-0.107785
MOTIVO_INTERNACAO_2	-0.132439
TIPO_SAIDA_2	-0.112038
MOTIVO_INTERNACAO_3	-0.091367

TIPO_SAIDA_3	-0.083051
SEXO_F	-0.020355
SEXO_M	0.020355
GESTANTE_IGNORADO	0.018341
GESTANTE_N	-0.018319
GESTANTE_S	-0.000866
TIPO_CASO_NOVO	-0.105586
TIPO_CASO_RECIDIVA	0.042670
TIPO_CASO_RETR ABAND	0.088000
TIPO_CASO_RETRAT APOS FALENCIA/RESISTENCIA	0.055975
TIPO_CASO_RETRAT APOS MUD ESQUEMA INT/TOX	0.011326
CLASSIFICACAO_DISSEM	0.052618
CLASSIFICACAO_EXT	0.016754
CLASSIFICACAO_P+E	0.059559
CLASSIFICACAO_PUL	-0.052441
BACILOSCOPIA_ESCARRO_AND	0.023176
BACILOSCOPIA_ESCARRO_N/REALIZ	0.036166
BACILOSCOPIA_ESCARRO_NEG	0.047437
BACILOSCOPIA_ESCARRO_POS	-0.079857
BACILOSCOPIA_ESCARRO_S/INF	0.040835
BACILOSCOPIA_OUTRO_MATERIAL_AND	0.008833
BACILOSCOPIA_OUTRO_MATERIAL_N/REALIZ	-0.071874
BACILOSCOPIA_OUTRO_MATERIAL_NEG	0.064058
BACILOSCOPIA_OUTRO_MATERIAL_POS	0.030011
BACILOSCOPIA_OUTRO_MATERIAL_S/INF	0.016642
CULTURA_ESCARRO_AND	0.032146
CULTURA_ESCARRO_N/REALIZ	0.009162
CULTURA_ESCARRO_NEG	-0.021416
CULTURA_ESCARRO_POS	-0.013204

CULTURA_ESCARRO_S/INF	0.036313
CULTURA_OUTRO_MATERIAL_AND	0.034980
CULTURA_OUTRO_MATERIAL_N/REALIZ	-0.085415
CULTURA_OUTRO_MATERIAL_NEG	0.060297
CULTURA_OUTRO_MATERIAL_POS	0.052563
CULTURA_OUTRO_MATERIAL_S/INF	0.019204
HISTOPATOLOGIA_BAAR POS	-0.002114
HISTOPATOLOGIA_N/REALIZ	0.005384
HISTOPATOLOGIA_S/INF	0.028789
HISTOPATOLOGIA_SUGESTIVO TB	-0.028183
NECROPSIA_BAAR POS	0.021794
NECROPSIA_N/REALIZ	-0.070422
NECROPSIA_S/INF	0.019976
NECROPSIA_SUGESTIVO TB	0.123900
HIV_AND	0.030558
HIV_N/REALIZ	0.137612
HIV_NEG	-0.244853
HIV_POS	0.170043
HIV_S/INF	0.049814
TESTE_SENSIBILIDADE_N	0.002631
TESTE_SENSIBILIDADE_S	0.013052
TESTE_SENSIBILIDADE_S/INF	-0.010523
AIDS_N	-0.171532
AIDS_S	0.171532
DIABETES_N	-0.015111
DIABETES_S	0.015111
ALCOOLISMO_N	-0.067558
ALCOOLISMO_S	0.067558
DOENCA_MENTAL_N	-0.009036

DOENCA_MENTAL_S	0.009036
USO_DROGAS_N	-0.050147
USO_DROGAS_S	0.050147
OUTRAS_DOENCAS_IMUNO_N	-0.042457
OUTRAS_DOENCAS_IMUNO_S	0.042457
TABAGISMO_N	-0.012088
TABAGISMO_S	0.012088
ESQUEMA_INICIAL_MR	0.023376
ESQUEMA_INICIAL_OUTROS	0.146104
ESQUEMA_INICIAL_RHZ	-0.041335
ESQUEMA_INICIAL_RHZE	-0.038439
ESQUEMA_INICIAL_SZEET	0.007047
MUDANCA_ESQUEMA_N	-0.053596
MUDANCA_ESQUEMA_S	0.053521
MUDANCA_ESQUEMA_S/INFO	0.004225
MOTIVO_MUDANCA_ESQUEMA_INTOLERANCIA/TOXICIDADE	0.012864
MOTIVO_MUDANCA_ESQUEMA_OUTRO MOTIVO	-0.057406
MOTIVO_MUDANCA_ESQUEMA_RESISTENCIA MEDICAMENTOSA	0.084580
RESISTENCIA_AND	0.006535
RESISTENCIA_SENS	-0.144692
RESISTENCIA_TB MR	0.178488
RESISTENCIA_TB R	0.118974
TIPO_TRATAMENTO_AUTO-ADMINISTRADO	0.087652
TIPO_TRATAMENTO_S/INF	0.129790
TIPO_TRATAMENTO_SUPERVISIONADO	-0.161155

ANEXO A – DECLARAÇÃO SOBRE COMITÊ DE ÉTICA

São Carlos, 2 de junho de 2022..

Ilmo.(a) Sr.(a)
Prof.(a) Dr.(a) Adair Roberto Aguiar
Presidente da Comissão de Pós-Graduação do Programa de Pós-Graduação Interunidades em Bioengenharia
EESC/FMRP/IQSC-USP

Senhor Presidente,

A título de esclarecimento, com relação a qualificação de mestrado da minha orientada – Ana Clara de Andrade Mioto, 10277252, referente ao projeto intitulado "*Estudo e Aplicação de técnicas de aprendizado de máquina na análise de desfechos não desejáveis de tuberculose*".

Declaro que este mesmo projeto desfruta de uma bolsa FAPESP (processo número: 021/01961-0 está vinculado a um projeto mais geral, financiado pela FAPESP, dentro do programa e-Science: Saúde Digital Humana (processo número: 2020/01975-9)

Declaro também que o projeto de pesquisa mais geral, coordenado por mim obteve aprovação do Comitê de Ética do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da USP, cadastrado sob o nº 4.749.122. Assim, o projeto de mestrado citado está incluído no mesmo parecer do Comitê de Ética.

Atenciosamente,

Domingos Alves