UNIVERSIDADE DE SANTA CRUZ DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E PROCESSOS INDUSTRIAIS – MESTRADO
ÁREA DE CONCENTRAÇÃO EM CONTROLE E OTIMIZAÇÃO DE PROCESSOS INDUSTRIAIS

MAIKEL LUIS KOLLING

MINERAÇÃO DE DADOS APLICADA Á PREDIÇÃO DE CASOS DE ABANDONO NO TRATAMENTO DE TUBERCULOSE EM POPULAÇÕES PRIVADAS DE LIBERDADE

Santa Cruz do Sul
2021

UNIVERSIDADE DE SANTA CRUZ DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E PROCESSOS INDUSTRIAIS – MESTRADO
ÁREA DE CONCENTRAÇÃO EM CONTROLE E OTIMIZAÇÃO DE PROCESSOS INDUSTRIAIS

MINERAÇÃO DE DADOS APLICADA Á PREDIÇÃO DE CASOS DE ABANDONO NO TRATAMENTO DE TUBERCULOSE EM POPULAÇÕES PRIVADAS DE LIBERDADE

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Sistemas e Processos Industriais – Mestrado, Universidade de Santa Cruz do Sul – UNISC

Orientador: Prof. Dr. Leonel Pablo Carvalho Tedesco (UNISC)

Santa Cruz do Sul
2021

AGRADECIMENTOS

Agradeço a minha família, em especial esposa, filho e pais que contribuíram de todas maneiras possíveis durante minha jornada acadêmica na busca deste novo marco do conhecimento. Agradeço também aos amigos, colegas de trabalho e professores do Programa de Pós-Graduação em Sistemas e Processos Industriais que sempre incentivaram essa conquista.

Agradeço ao meu orientador Prof. Dr. Leonel Pablo Carvalho Tedesco que me acolheu em um momento importante desta jornada, contribuindo com sua experiência e sabedoria.

À CAPES – Fundação Coordenação de Aperfeiçoamento Pessoal de Nível Superior – por fomentar a pesquisa através da concessão da bolsa de estudos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Who is the master?!

There is one place that you have not looked.

And it is there, only there, that you shall find the master.”

The Last Dragon, Movie, 1985

Minha primeira experiência com a palavra mestre.

SUMÁRIO

RESUMO

Segundo o Ministério da Saúde Brasileiro (MS, 2021) a tuberculose (TB) é uma doença infectocontagiosa que registra aproximadamente 70 mil novos casos a cada ano no Brasil, levando à morte cerca de 4,5 mil pessoas por ano. Um grande número destas mortes se deve ao alto índice de abandono do tratamento, sendo um agravante os altos índices de abandono do tratamento nos casos incidentes em populações privadas de liberdade (PPLs). Diante deste problema e com a evolução dos algoritmos de mineração de dados, este trabalho propõe a aplicação das etapas do processo de KDD (Knowledge Discovery in Databases) na busca do melhor modelo computacional para previsão de casos de abandono do tratamento de TB de PPLs. Para a realização destas etapas foi utilizada a metodologia SEMMA (Sample, Explore, Modify, Model, Assess) através de técnicas de limpeza, pré-processamento, seleção de atributos e aplicação de algoritmos de mineração de dados sobre um conjunto de dados contendo os registros históricos de pacientes acometidos pela TB registrados no Sistema de Informação de Agravos de Notificação (SINAN). Os resultados obtidos foram avaliados de acordo com a sua acurácia, precisão, sensibilidade e especificidade. Sendo que o principal resultado foi apresentado pelo algoritmo Florestas Aleatórias, o qual obteve 97,07% de acurácia na previsão de abandono no tratamento de TB em PPLs. Para atingir o objetivo deste estudo, foi realizada a escrita de dois manuscritos, sendo o primeiro uma revisão sistemática da literatura, relacionada ao tema de mineração de dados na área da saúde e o segundo a descrição detalhada da aplicação da metodologia SEMMA adotada para a construção dos modelos de predição de abandono de tratamento de TB em PPLs.

Palavras-chave: Mineração de Dados, Tuberculose, Abandono de tratamento, Metodologia/Método SEMMA.

ABSTRACT

According to the Brazilian Ministry of Health (MS, 2021), tuberculosis (TB) is an infectious disease that registers approximately 70 thousand new cases each year in Brazil, leading to the death of approximately 4.5 thousand people per year. A large number of these deaths are due to the high rate of abandonment of treatment, with an aggravating factor being the high rates of abandonment of treatment in cases involving populations deprived of their liberty (PPLs). In view of this problem and with the evolution of data mining algorithms, this work proposes the application of the stages of the KDD process (Knowledge Discovery in Databases) in the search for the best computational model for predicting cases of abandonment of TB treatment of PPLs. To perform these steps, the SEMMA methodology was used (Sample, Explore, Modify, Model, Assess) through cleaning techniques, pre-processing, attribute selection and application of data mining algorithms on a data set containing the records history of patients affected by TB registered in the Notifiable Diseases Information System (SINAN). The results obtained were evaluated according to their accuracy, precision, sensitivity and specificity. The main result was presented by the Random Forest algorithm, which obtained 97.07% accuracy in predicting abandonment in the treatment of TB in PPLs. To achieve the objective of this study, two manuscripts were written, the first being a systematic review of the bibliography related to the topic of data mining in the health area and the second the detailed description of the application of the SEMMA methodology adopted for the construction of the prediction models of TB treatment dropout in PPLs.

Keywords: Data Mining, Tuberculosis, Treatment Abandonment, SEMMA Method / Methodology.

LISTA DE ILUSTRAÇÕES

MANUSCRITO I

MANUSCRITO II

LISTA DE TABELAS

MANUSCRITO I

MANUSCRITO II

LISTA DE ABREVIATURAS

AIDS	*Acquired Immunodeficiency Syndrome*
BK	Bacilo de Koch
BPNA	*Bibliometric Performance and Network Analysis*
CRISP-DM	*CRoss-Industry Standard Process for Data Mining*
DATASUS	Departamento de informática do SUS
DCBD	Descoberta de Conhecimento em Bancos de Dados
FDA	*Food and Drug Administration*
IA	Inteligência Artificial
KDD	*Knowledge Discovery in Databases*
ML	*Machine Learning*
MS	Ministério da Saúde
OMS	Organização Mundial da Saúde
SEMMA	*Sample*, *Explore*, *Modify*, *Model*, *Assess*.
SINAN	Sistema de Informação de Agravos de Notificação
SUS	Sistema Único de Saúde
TB	Tuberculose
TB-MR	Tuberculose Multirresistente
TDO	Tratamento Diretamente Observado
WHO	*World Health Organization*

1. INTRODUÇÃO

A Tuberculose (TB) é uma doença infectocontagiosa causada pela bactéria Mycobacterium Tuberculosis também chamada de Bacilo de Koch (BK), nome dado em homenagem ao cientista Hermann Heinrich Robert Koch que detectou a bactéria pela primeira vez em 1882 estabelecendo uma relação entre o agente microbiano e a TB (GRADMANN, 2001). É uma doença que acomete a humanidade a milênios, tendo em vista que as bactérias causadoras da doença já foram detectadas em corpos sepultados a mais de 3000 anos antes de Cristo. De acordo com relatório da Organização Mundial da Saúde (OMS) a TB matou 45 milhões de pessoas no período de 2000 e 2019 em todo o planeta, sendo a nona causa de morte e a primeira por doença infecciosa no mundo. É estimado que em torno 10 milhões de pessoas tiveram TB no ano de 2019, sendo que 90% dos infectados são adultos e 9% portadores da Síndrome da Imunodeficiência Adquirida (AIDS). Somente no Brasil ocorrem aproximadamente 70 mil novos casos a cada ano causando a morte de aproximadamente 4500 pessoas (MS, 2021), transformando-se nesse contexto em um problema de saúde global.

Uma vez infectado com a tuberculose, o portador além de sofrer com os sintomas da doença, passa a ser um vetor de transmissão da TB, espalhando a bactéria a partir de secreções respiratórias ao respirar, tossir ou espirrar. Normalmente a TB é pulmonar, mas também pode afetar outros órgãos, passando a se denominar TB extrapulmonar. Entre os grupos com mais chances de desenvolver a doença destaca-se pessoas que vivem com HIV/Aids, pessoas privadas de liberdade e moradores de rua (CHURCHYARD, et al., 2017). Mesmo sendo uma doença curável, com um tratamento que varia de 6 a 8 meses com medicações denominadas de primeira linha, o acompanhamento e o tratamento adequado da doença são um grande desafio para as entidades de saúde em todo o planeta, sendo essa dificuldade em receber um atendimento correto o principal motivo que leva os pacientes a não conseguirem concluir o tratamento levando ao abandono do tratamento.

Esse abandono do tratamento propicia o desenvolvimento de formas mais agressivas da doença resistentes aos medicamentos, podendo evoluir para uma forma conhecida como TB Multirresistente (TB-MR), que se caracteriza pela resistência aos medicamentos de primeira linha Isoniazida e Rifampicina (WHO, 2019). Esse abandono do tratamento pode resultar no aparecimento de uma forma mais grave de tuberculose chamada de tuberculose droga resistente e seu tratamento aumenta para 18 meses a 5 anos tornando a cura da doença muito mais traumatizante e com mais efeitos colaterais, tendo

em vista a necessidade da utilização de medicamentos denominados de segunda linha como por exemplo Fluoroquinolona. O maior número de abandono no tratamento ocorre nas populações mais vulneráveis, entre elas a população alvo desta pesquisa que são as populações privadas de liberdade.

Neste cenário, uma melhor compreensão do desfecho do tratamento ou situação de encerramento com abandono do tratamento da TB é um fator importante na busca de estratégias que possam atender as orientações da OMS que recomenda um mínimo de 85% para taxa de cura e máximo de 5% de casos de abandono do tratamento. Desta forma, se no início ou durante o tratamento fosse possível prever com razoável precisão quais pacientes correm o risco de abandonar o tratamento, permitiria que as entidades de saúde criassem programas estratégicos de acompanhamento do paciente que teriam como objetivo aumentar a eficiência da alocação de recursos e esforços buscando melhorar os indicadores de cura do Brasil, que hoje estão em torno de 71,9% dos casos novos e taxa de abandono de 10,4%. Nesse sentido, a OMS em seu programa de combate à TB incentiva o uso de novas tecnologias na busca de novas estratégias que buscam melhorar os indicadores de cura, dentre essas tecnologias destaca-se o Aprendizado de Máquina e a Inteligência Artificial, duas áreas do conhecimento da Ciência da Computação que ganharam destaque nas áreas de pesquisa em saúde.

Inteligência Artificial (IA) é um conceito amplo, que se refere aos sistemas computacionais que poderiam realizar tarefas, tais como calcular, planejar, entender linguagem ou reconhecer objetos da mesma forma que os humanos, mas com maior eficiência na resolução destes problemas (RUSSEL e NORVIG, 2013). Estes sistemas computacionais quando separados denominam-se de Aprendizado de Máquina, ou Machine Learning (ML), que são algoritmos com a capacidade de aprender por meios de padrões e realizar tarefas futuras, baseado no que foi aprendido. Uma das ramificações do ML é a Mineração de Dados (Data Mining) que tem por objetivo a busca de conhecimento aplicando algoritmos que detectam padrões em um grande conjunto histórico de informações. Nas últimas duas décadas demonstraram crescimento constante de pesquisas que envolvem esse assunto junto com a área da saúde, mostrando-se assim um ramo promissor (ALONSO et al. 2018). Para se extrair conhecimento por meio de mineração de dados, diversos são os referenciais metodológicos usados, dentre eles destaca-se o SEMMA (sigla de Sample, Explore, Modify, Model, Assess, ou Extrair, Explorar, Modificar, Modelar, Avaliar) (SANTOS E AZEVEDO, 2005).

Diante do exposto esta dissertação propõe a aplicação da metodologia SEMMA e uso de algoritmos de mineração de dados com o intuito de construir modelos que calculem a probabilidade de um paciente de TB abandonar seu tratamento. Para tanto será utilizado um conjunto de registros provenientes da base de dados do sistema SINAN (Sistema de Informação de Agravos de Notificação). Esses dados estão disponibilizados publicamente pelo departamento de informática do Sistema Único de Saúde do Brasil (DATASUS) para que possam subsidiar análises objetivas da situação sanitária e podendo auxiliar na tomada de decisão baseadas em evidências e elaboração de programas de ações de saúde. Nesta base de dados encontram-se registros dos dados históricos sobre triagem, andamento do tratamento e informações demográficas dos pacientes acometidos pela doença.

Para atingir o objetivo deste estudo, inicialmente foi realizada uma revisão sistemática da literatura sobre mineração de dados na área da saúde com intuito de verificar como estavam as pesquisas relacionadas ao tema, sendo o resultado apresentado no Manuscrito I. Posteriormente, foram realizadas as etapas descritas na metodologia SEMMA para extração e preparação do conjunto de dados, através de técnicas de limpeza de dados, pré-processamento e seleção de atributos na base de dados para então aplicar e comparar a acurácia de diversos algoritmos de mineração de dados encontrados na literatura, apresentado no Manuscrito II.

2. TEMA E PROBLEMA

Segundo a Organização Mundial da Saúde a tuberculose é uma das principais causas de mortalidade relacionadas às doenças infecciosas nos países em desenvolvimento. Dentre os casos registrados no Brasil um dos grupos mais vulneráveis faz parte as populações privadas de liberdade, com alta a incidência de tuberculose. Mesmo existindo políticas públicas relacionadas ao Programa Nacional de Controle da Tuberculose existe uma lacuna no controle, acompanhamento e tratamento adequado deste grande número de portadores da doença, tendo como consequência um número elevado de casos de abandono do tratamento por grande parte dos portadores, podendo chegar a 26% em algumas regiões do país.

Diante deste contexto, o problema de pesquisa é: Como construir modelos de predição com alto grau de acurácia para contribuir na mitigação do problema de abandono

no tratamento da tuberculose da população privada de liberdade e nos programas de acompanhamento desses pacientes?

O objetivo geral é aplicar e validar diferentes algoritmos de mineração de dados buscando modelos computacionais para predição dos casos de abandono no tratamento da tuberculose na base de dados com históricos do sistema SINAN. Os seguintes objetivos específicos são listados:

- Realizar levantamento bibliográfico de pesquisas que utilizem algoritmos de mineração de dados e que busquem por modelos de predição da conclusão do tratamento de doenças.
- Realizar o pré-processamento dos dados históricos, com o uso da Metodologia SEMMA.
- Validar o algoritmo que gere o modelo mais assertivo de previsão de abandono do tratamento da tuberculose, tendo como base de validação os dados históricos obtidos no SINAN

5. CONCLUSÃO

O abandono do tratamento da TB ainda é um obstáculo para o sucesso dos programas de combate à TB incentivados pela OMS. No sistema prisional essa taxa de abandono pode chegar a 26% de abandono em algumas regiões, valor muito acima da meta de 5% sugeridos pela OMS como taxa máxima de abandono. A elaboração de planos de ação e identificação de sintomáticos respiratórios para diagnóstico através de exames e registro no sistema SINAN é de extrema importância, pois fundamentam indicadores e favorecem a aplicação dos programas de combate à doença.

Dentre esses programas destaca-se o Tratamento Diretamente Observado (TDO), onde aproximadamente 50% das PPLs são integrantes do programa, tendo os participantes do programa maior chance de não abandonarem o tratamento. Porém em grande parte dos estados brasileiros o índice de inclusão ao programa não chega aos 25%. Essa baixa inclusão no programa, evidencia a escassez de recursos humanos ou financeiros nas instituições prisionais, forçando que os programas internos de combate a TB façam a seleção empírica dos pacientes que recebem essas sessões de acompanhamento direto, sendo que em casos de progressão de pena ou transferência de unidade prisional o controle e manutenção da situação do tratamento do apenado tam´bem é prejudicada pela falta destes recursos.

Então a seleção e inclusão nos programas TDO dos pacientes com maior chance de abandonar o tratamento, pode ser um fator relevante na diminuição dos casos de abandono, nesse contexto a utilização das tecnologias computacionais na criação de modelos de previsão poderá ser um grande aliado para as equipes de saúde na tomada de decisões em ações de programas de prevenção do abandono do tratamento da TB. Neste estudo, foi proposta a aplicação da metodologia SEMMA para criação de modelos de mineração de dados que possam contribuir para que as equipes responsáveis pelos programas de TB consigam prever de uma forma mais eficiente os pacientes que possam abandonar o tratamento, podendo então focar com mais atenção no acompanhamento direto destes pacientes portadores da doença nas unidades prisionais. Os modelos gerados possuem um alto grau de acurácia, tendo destaque o algoritmo FAs que alcançou a maior taxa de acurácia, alcançando o valor de 97,07%.

Com a obtenção de modelos robustos e sendo o objetivo antecipar e prevenir um possível abandono do paciente, os modelos de predição obtidos durante o desenvolvimento deste estudo poderão ser disponibilizados na forma de software gerando

alertas de possíveis abandonos para as unidades de saúde responsáveis. Neste sentido já foram realizadas reuniões com órgãos da segurança pública com a intenção de formalizar um convênio entre a instituição de ensino e o órgão para o desenvolvimento. Como projeto piloto foi idealizada uma ferramenta de apoio no acompanhamento dos pacientes em tratamento da TB, podendo ser implementada como trabalho futuro e contribuir na tomada de decisões dos especialistas com o objetivo de mitigar o problema do abandono do tratamento de tuberculose.

Tendo em vista que os dados do SINAN são públicos e abrangem inúmeras doenças, a utilização e ferramentas computacionais sobre esse conjunto de informações podem oferecer soluções que possam contribuir para uma maior eficiência no atendimento das pessoas acometidas por alguma doença. Sendo possível também que além da busca por modelos computacionais que façam a previsão do desfecho de tratamento da TB, a aplicação das metodologias de mineração de dados em especial a metodologia SEMMA, podem ser utilizadas em estudos para outros grupos de populações ou doenças registradas no sistema SINAN ou ainda na elaboração de estudos descritivos sobre a evolução das notificações e desfechos no Brasil de doenças com notificação compulsória.

No entanto, curiosamente durante a construção do Manuscrito I nenhum dos clusters se destacou nos estudos relacionados a doenças infecciosas e, portanto, é razoável sugerir a maior exploração de técnicas de mineração de dados neste domínio, especialmente dado o impacto global quando doenças infecciosas acometem a população mundial. Então a realização de trabalhos futuros que possam avançar no campo da mineração de dados na área da saúde seriam de grande de relevância para a sociedade.

Por fim é importante ressaltar que o Manuscrito I foi publicado na revista *International Journal of Environmental Research and Public Health* e o Manuscrito II foi submetido ao Simpósio Brasileiro de Computação Aplicada à Saúde 2021(SBCAS-2021) sendo apresentado como Anexo A a comprovação.

6. REFERÊNCIAS

Russel S, Norvig P. **Inteligência artificial**. Rio de Janeiro: Elsevier; 2013.

Alonso, Susel Góngora et al. Data mining algorithms and techniques in mental health: A systematic review. ***Journal of medical systems***, v. 42, n. 9, p. 1-15, 2018.

Churchyard, Gavin et al. What we know about tuberculosis transmission: an overview. ***The Journal of infectious diseases***, v. 216, n. suppl_6, p. S629-S635, 2017.

Gradmann C. Robert Koch and the Pressures of Scientific Research: Tuberculosis and Tuberculin. **Med Hist**. 2001;45(01):1–32.

MS - Ministério Da Saúde Brasil, Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica, ***Boletim Epidemiológico Tuberculose 2021***.

WHO - World Health Organization, ***Global Tuberculosis Report 2020***. Genebra: OMS. 2020.