Predição do desfecho de pacientes com Tuberculose

Orientador: Prof. Dr. Adelmo Inacio Bertolde
Orientanda: Iara Arruda

1 Introdução

Machine Learning trata-se de um conjunto de técnicas usadas para a análise de dados, onde as máquinas são treinadas a partir de informações pré-existentes identificando padrões e automatizando a criação de modelos analíticos, que serão utilizados para a tomada de decisão. Com o desenvolvimento da tecnologia, a capacidade de gerar esses modelos foi aumentada, podendo ser aplicado em grandes bases de dados (SAS, 2019).

Tendo em vista que com as técnicas adequadas podemos manipular grandes conjuntos de dados, existem algumas preocupações em relação a perda de informações. Sendo o nosso objetivo encontrar fatores que expliquem a causa de determinada situação, se não houver a completude das informações, torna-se inviável a utilização das mesmas. Em algumas situações, é possível utilizar técnicas que tem como objetivo "preencher" os dados faltantes. Porém, é de extrema importância que sejam usadas as técnicas adequadas para não gerar análises viesadas ao final do estudo. Um dos primeiros métodos criados para a solução deste problema, foi a substituição desses dados faltantes pela média ou mediana para se ter o banco completo a fim de realizar as análises. Com o passar do tempo, outras técnicas foram surgindo trazendo melhores resultados. Atualmente, uma das técnicas mais utilizadas para a imputação de dados é a imputação múltipla que consegue corrigir o problema de imputar uma só informação para todo o conjunto faltante (NUNES, 2009).

A aplicação prática desses elementos para estudos reais podem trazer ganhos importantes para diversas áreas. A capacidade de prever situações pode nos levar a ter um olhar mais cuidadoso na prevenção de certos cenários no âmbito da saúde, por exemplo.

2 Metodologia

O banco de dados disponibilizado para o estudo contém informações de pacientes que foram notificados com Tuberculose no Brasil no período de 2016 a 2018 coletado pelo SINAN (Sistema de Informação de Agravos de Notificação). As análises estatísticas serão realizadas no software R, e para as técnicas escolhidas para a imputação e modelagem serão utilizados pacotes específicos no próprio software.

Para a primeira fase do trabalho, foram escolhidas as seguintes técnicas de imputação: Multivariate Imputation via Chained Equations (MICE) e Floresta Aleatória. Para cada uma das técnicas teremos uma base com dados imputados. A partir disso, serão aplicados métodos de Machine Learning em cada uma dessas bases, e usaremos métricas a fim de avaliar os melhores resultados trazidos pela primeira e segunda etapa. Os métodos de Machine Learning que serão usados na modelagem serão: Naive Bayes, Regressão Logística Multinomial, Floresta Aleatória, KNN e Xgboost.

3 Objetivos Gerais e Específicos

Por ser uma doença que possui uma relação direta com grupos populacionais em situação de vulnerabilidade, observamos muitos casos de abandono do tratamento. Em razão disso, verificamos muitas informações faltantes nas fichas de notificação da TB obtidas pelo SINAN. Para o contexto da análise dessas informações, isso pode significar a perda de variáveis importantes na identificação dos fatores que realmente estão associados a causa de um desfecho desfavorável no tratamento destes pacientes. Sabendo disto, o estudo de técnicas apropriadas para a imputação de dados faltantes é de grande relevância para preservar variáveis que antes seriam descartadas. Assim, conseguiríamos utilizar essas informações para obter melhorias na modelagem que será feita posteriomente.

O objetivo geral deste trabalho, consiste em encontrar modelos que possam prever o desfecho dos pacientes com TB utilizando Machine Learning. Os resultados possíves do desfecho do tratamente serão três: cura, abandono do tratamento e óbito. A fins de comparação, serão empregadas as medidas de sensibilidade, especificidade, curva ROC e o índice Kappa para indicar o melhor modelo.

4 Eventuais Resultados

Comparando os resultados de um trabalho de monografia na mesma temática, a Acurácia e o Índice Kappa apresentaram como resultado os valores de 0.731 e 0.312, respectivamente, para o método Naive Bayes (QUINELATO, 2019). No presente trabalho, para a base que foi imputada a partir do método MICE, obtivemos uma acurácia de 0.8126 e o Índice Kappa de 0.322 utilizando também o Naive Bayes. O método de imputação de Floresta Aletória também apresentou resultados um pouco superior. Apesar de ser pouca diferença, esperamos que traga resultados satisfatórios ao final do estudo.

Referências

1. SAS. Machine Learning O que é e qual sua importância? 2019. Disponível em ¡https://www.sas.com/pt$_br/insights/analytics/machine-learning.html > .Acessoem$: 10$Abril$.2020.

2. Nunes LN, Klück MM, Fachel JMG. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. Cad Saúde Pública 2009; 25:268-78.

3. Sauer CM, Sasson D, Paik KE, McCague N, Celi LA, Sánchez Fernández I, et al. Feature selection and prediction of treatment failure in tuberculosis. 2018. PLoS ONE 13(11): e0207491. https://doi.org/10.1371/journal.pone.0207491

4. Ministério da Saúde. Tuberculose: o que é, causas, sintomas, tratamento, diagnóstico e prevenção. Disponível em ¡https://saude.gov.br/saude-de-a-z/tuberculose¿. Acesso em: 10 Abril. 2020

5. Analytics Vidhya. Tutorial on 5 Powerful R Packages used for imputing missing values. Disponível em: ¡https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/¿. Acesso em: 10 Abril. 2020.

6. Quinelato, Luiz Henrique. Aprendizado de Máquina Aplicado ao Estudo da Tuberculose no Brasil. Estatística - UFES (Trabalho de Monografia). 2019.