APLICAÇÃO DE MACHINE LEARNING PARA ANÁLISE DE DADOS DE MORTALIDADE NEONATAL NO BRASIL

1
2
3
4
5
6

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

Apresentado no 10° Congresso de Inovação, Ciência e Tecnologia do IFSP
27 e 28 de novembro de 2019- Sorocaba-SP, Brasil

RESUMO: A redução da mortalidade neonatal é de extrema importância no mundo. O período neonatal representa o momento mais suscetível a possíveis complicações que podem levar risco para a sobrevivência da criança e é bastante motivado por condições desfavoráveis de vida e da atenção da saúde. No que diz respeito a predição, apesar da teoria direcionar a orientação de uma análise quantitativa, ela pode ser expandida com outras abordagens. Assim, utilizando técnicas de aprendizado de máquina aplicadas a este contexto é uma proposta inovadora à realidade. Este trabalho, traz a possibilidade de prevenir a mortalidade neonatal utilizando de técnicas de análise e processamento de dados e aprendizado de máquina não supervisionado. Para isso será utilizado o algoritmo *K-Means* juntamente com os dados dos Sistema de Informação sobre Mortalidade e Sistema de Informações de Nascidos Vivos. A análise realizada busca compreender o significado dos dados coletados e além disso tem a finalidade de facilitar o entendimento dos conteúdos através de classificações apresentadas de forma simples, por exemplo, gráficos.

PALAVRAS-CHAVE: mortalidade neonatal; Análise de dados; aprendizado de máquina; SIM, SINASC.

DATA ANALYSIS AND PROCESSING FOR THE PREDICTION OF NEONATAL DEATH

ABSTRACT: Reducing neonatal mortality is of utmost importance in the world. The neonatal period represents the moment most susceptible to possible complications that may pose a risk to the child's survival and is largely motivated by unfavorable living conditions and health care. As for prediction, although theory directs the orientation of a quantitative analysis, it can be expanded with other approaches. Thus, using machine learning techniques applied in this context is an innovative proposal to reality. This work brings the possibility of preventing neonatal mortality using data analysis and processing techniques and unsupervised machine learning. For this, the K-Means algorithm will be used together with the Mortality Information System and Live Birth Information System data. The analysis performed seeks to understand the meaning of the collected data and also has the purpose of facilitating the understanding of the contents through simple presented classifications, for example, graphs.

KEYWORDS: neonatal mortality, data analysis, prediction, pregnancy.

INTRODUÇÃO

Os primeiros 28 dias de vida o período neonatal representa o momento mais vulnerável para a sobrevivência de uma criança e são fortemente influenciados por condições desfavoráveis de vida da

população e da atenção à saúde. Neste sentido, a mortalidade neste período reflete a complexa conjunção de fatores biológicos, socioeconômicos e assistenciais, esses últimos relacionados à atenção à gestante e ao recém-nascido (LANSKY et al., 2014). Cerca de 46% das mortes no mundo acontecem entre os menores de cinco anos e grande parte destas está concentrada nos primeiros dias de vida.

Nos últimos anos um aumento no número de aplicações envolvendo *Machine Learning* (ML) (Aprendizado de Máquina) (BISHOP, 2006) em diversas áreas do conhecimento, inclusive na área da Saúde e Pública. Neste contexto, o presente trabalho tem como objetivo utilizar o métodos de ML para análise de dados de mortalidade neonatal, utilizando abordagens não supervisionados. Com isso espera-se que a partir dos resultados obtidos, aplicados nas bases de dados do Sistema de Informação sobre Mortalidade (SIM) e Sistema de Informações de Nascidos Vivos (SINASC), será possível entender o relacionamento entre as variáveis e posteriormente, os resultados serão analisados em relação a probabilidade de morte neonatal, para entender as causas mais recorrentes, afim de realizar agrupamentos de populações de risco, em um sistema de predição de morte neonatal.

MATERIAL E MÉTODOS

Buscando desenvolver uma solução que auxilie na predição de mortes neonatais, o presente trabalho utiliza de métodos de abordagem não supervisionada de ML para análise. Especificamente neste trabalho está sendo utilizado o algoritmo *K-Means*. O algoritmo será aplicado ao conjunto de dados oriundo de uma associação entre dados do SIM e do SINASC.

Fontes de Dados

O processamento dos dados foi realizado com duas fontes o SIM, desenvolvido pelo Ministério da Saúde, em 1975, e é unificação de diversos modelos e instrumentos que são utilizados para coleta de dados sobre mortalidade no Brasil. Essa base de dados possui variáveis categóricas que permitem, construir atributos que podem eventualmente contribuir para conclusões médicas. A segunda fonte de dados utilizados será SINASC, como o SIM também introduzida pelo Ministério da Saúde em 1990, sua finalidade é adquirir todos os dados de nascimentos em território nacional propiciando uma estatística de nascidos vivos, com variáveis que são de grande valor para a saúde pública, como peso ao nascer, escolaridade da mãe, consultas de pré-natal.

Aprendizado de Máquina

De acordo com as definições de Arthur Samuel (1959), o ML pode ser definido como uma técnica que processa dados de entrada em um sistema de computadores de forma que o computador aprenda a processar e executar rotinas posteriormente sem ser previamente programada e sustentado com novas informações. Essa versatilidade possibilita que o ML atue em praticamente qualquer área e possa nos auxiliares a produzir melhores resultados em menos tempo.

Os algoritmos de ML possuem três fundamentos, independente da variação dos algoritmos, que são: (1) representação, que é o conceito responsável pela forma como o algoritmo ira aprender; (2) avaliação, que basicamente é o componente mais importante, ele define qual será o caminho utilizado para avaliar as condições e chances dos resultados; (3) e a otimização, que é o método no qual as chances e condições dos resultados são combinados criados.

Partindo do conceito de Aprendizado de Máquina apresentando nos parágrafos anteriores o presente trabalho utiliza a abordagem não supervisionada, a qual é utilizada para procurar padrões existentes através de agrupamentos em grandes quantidades de dados não rotulados, a partir dos padrões encontrados é possível descobrir similaridade diferenças entre esses dados, o que pode resultar em conclusões muito relevantes. Existem diversos algoritmos que utilizam a abordagem não supervisionada, entre eles está o *K-Means*, o qual será utilizado nesse trabalho.

K-Means é um dos mais populares algoritmos aprendizado de máquina não supervisionado de "clustering". *K-Means* armazena centróides (pontos que representam o centro do cluster) são usados para definir qual é o padrão dos clusters. Um ponto é considerado em um cluster específico se estiver mais próximo do centróide desse cluster do que de qualquer outro centroide. (PIECH, Chris, 2013)

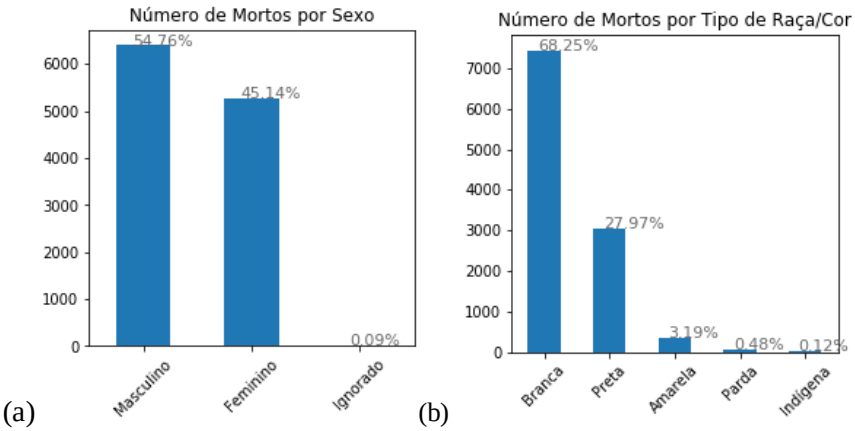
Ele é um algoritmo que é um agrupa de atributos semelhantes. O *K-Means* possui um componente crucial em sua estrutura que é o centróide, que representa o centro do cluster, calculado pela média de todos os atributos que compõe o cluster. A partir disso, os atributos se consolidam de acordo com sua proximidade com centro. De maneira simples pode-se dividir seu processamento em 4 passos: (1) atribuir valores iniciais para os exemplos dos atributos que constituem o cluster; (2) distribuir cada atributo ao cluster que possua a maior semelhança com o exemplo de atributo atribuído ao cluster no primeiro passe; (3) calcular a média dos atributos para obter o centróide; (4) por fim os passo (2) e (3) são repetidos até que se consolidem.

O *K-Means* facilita a análise visual dos dados criando agrupamentos considerando padrões entre os dados. O algoritmo *K-Means* particionar um conjunto de dados em K agrupamentos distintos e não sobrepostos. A execução do algoritmo, é feita primeiramente especificando o número desejado de agrupamentos K; então o algoritmo *K-Means* atribuir a cada um dos atributos a um dos clusters predefinidos. (ARRUDA, N. M, 2019).

‘

RESULTADOS E DISCUSSÃO

Por meio da aplicação do modelo de ML nas bases de dados, levando em consideração principalmente as variáveis como o tempo de gestação, tipo de gravidez, tipo de parto, sexo, peso do recém-nascido, raça e cor entre outras. Foram construídos gráficos que nos viabilizam o conhecimento sobre a relação de determinadas características com mortalidade neonatal.



(a) (b)

FIGURA 1. (a) Gráfico de barras representando a porcentagem de mortos classificados por sexo; (b) Gráfico de barras representando a porcentagem de mortos classificados por tipo de raça/cor.

A Figura dos gráficos acima é uma pequena demonstração de quão relevante as informações e características retiradas das bases de dados podem ser. O gráfico (a) revela que existe um certo equilíbrio entre as mortes registradas classificadas por sexo ocorridas no período neonatal, tendo em vista que a diferença é menor que 10%, agregado à essa informação o gráfico (b) apresenta que a grande maioria dos recém-nascidos que vem a óbito com 68,25% são da cor branca seguidos pelos de cor preta com 27,97%.

CONCLUSÕES

Neste trabalho, foi proposto uma análise nos dados e aplicação de modelos de aprendizado de máquina buscando disponibilizar os dados estatísticos relevantes para auxiliar o entendimento das causas de mortes neonatais. A partir as análises desenvolvidas haverá uma grande possibilidade prevenir a mortalidade neonatal, utilizando os dados gerados, permitindo que os profissionais da área da saúde definam estratégias direcionadas para cada caso, proporcionando um tratamento ou prevenção de possíveis problemas. Como foi proposto, a próxima etapa deste trabalho será a aplicação do modelo de aprendizado de máquina *K-Means*, e assim obtendo mais uma fonte de análise para extrair informações possivelmente ainda mais relevantes.

AGRADECIMENTOS

Agradecemos o apoio do grupo de pesquisa PICAp-IFSP Campinas; ao apoio financeiro da Fundação Bill & Melinda Gates (OPP1201970) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq (443774/2018-8), implementado via projeto institucional IFSP (SUAP: 23305.012088.2018-11); e à NVIDIA pela doação de uma GPU XP Titan, a ser utilizada nesta pesquisa.

REFERÊNCIAS

ARRUDA, N. M. Determinantes da Mortalidade Adulta nas Microrregiões Brasileiras em 2010: Uma análise baseada em modelos de aprendizado de máquina. Dissertação (Mestrado em Demografia) - UNICAMP. Campinas. 2019.

BISHOP, C. M. Pattern Recognition and Machine Learning. Secaucus, NJ, USA: SpringerVerlag New York, Inc., 2006. ISBN 0387310738.

FRIAS, P. G. d. et al. Utilização das informações vitais para a estimação de indicadores de mortalidade no Brasil: da busca ativa de eventos ao desenvolvimento de métodos. Cadernos de Saúde Pública, Scielo, v. 33, 00 2017. ISSN 0102- 311X. LANSKY, S. et al. Pesquisa Nascer no Brasil: perfil da mortalidade neonatal e avaliação da assistência à gestante e ao recém-nascido. Cadernos de Saúde Pública, Scielo, v.30, p. S192 – S207, 00 2014

PIECH, Chris. K MEANS. Disponível em https://stanford.edu/~cpiech/cs221/handouts/kmeans.html Acesso em: 23 de junho de 2019.