**INSTITUTO DE MATEMÁTICA E ESTATÍSTICA**

**UNIVERSIDADE DE SÃO PAULO**

****

**APLICAÇÃO DE MACHINE LEARNING PARA CLUSTERIZAÇÃO DE**

**PACIENTES COM ARBOVIROSES**

**Orientador: Prof. Dr. Helder T. I. Nakaya**

**Aluno: Fabio Carvalho de Souza**

**SÃO PAULO, MARÇO DE 2020**

**INTRODUÇÃO**

Arboviroses, infecções virais transmitidas por vetores artrópodes como mosquitos e carrapatos, estão entre as doenças virais mais difundidas em todo o mundo (GUBLER, 2002; WEAVER et al., 2009). Dentre elas destacam-se Dengue e Chikungunya, ambas transmitidas por mosquitos do gênero *Aedes*. Enquanto a Dengue acomete cerca de 390 milhões de indivíduos todos os anos, a Chikungunya já atingiu cerca de 10 milhões de pessoas desde a reemergência do vírus em 2004 (S. et al., 2013; SUHRBIER, 2019). A ampla distribuição dos mosquitos vetores dessas doenças em áreas tropicais e subtropicais coloca cerca de 3,6 bilhões de pessoas sob risco de infecção (BEATTY et al., 2009). Apesar de serem classificados como arboviroses, os vírus causadores da Dengue e Chikungunya pertencem a famílias virais distintas. O vírus da Dengue (DENV) pertence à família dos *Flavivirus*, na qual estão presentes também os vírus da Febre Amarela e da Zika, e apresenta uma grande diversidade genética com quatro sorotipos distintos (DENV-1 ao DENV-4). Uma vez instalado no organismo esse vírus tem um período de incubação de 4 a 7 dias, após o qual o indivíduo infectado pode apresentar febre, calafrios, mal-estar, vômito e vermelhidão, 75% dos indivíduos infectados são assintomáticos(SMITH et al., 2019). O vírus da Chikungunya (CHIKV), por sua vez, pertence à família dos *Alphavirus*, e apresenta 3 genótipos que se originaram a partir de um evento de reemergência do vírus no ano de 2004. O período de incubação do CHIKV é de 2 a 4 dias, após o qual, o paciente pode apresentar febre, dores de cabeça, vermelhidão, dores musculares e nas articulações. 15 a 25 % dos pacientes infectados são assintomáticos (BURT et al., 2017). Tanto Dengue quanto Chikungunya podem evoluir para casos mais severos. No caso da Dengue entre 2 e 5% dos pacientes podem progredir para a forma mais severa da doença, a qual possui grande associação com infecções secundárias por sorotipos distintos ao da infecção inicial. A forma severa é caracterizada por ocorrência de sangramentos em diversos sistemas, comprometimento de órgãos e extravasamento de plasma podendo evoluir para choque hemodinâmico (Síndrome de Choque da Dengue) e óbito (SMITH et al., 2019). Já a forma crônica da Chikungunya caracteriza-se por manifestações neurológicas graves, como mialgia, por um longo período de tempo e que podem provocar debilitação do paciente. Apesar de ser menos frequente, óbitos podem ocorrer em decorrência da forma severa (BURT et al., 2017).

Por se tratar de doenças que acometem principalmente países subdesenvolvidos e com recursos escassos é imprescindível a existência de protocolos capazes de identificar previamente pacientes com maior ou menor risco de desenvolver as formas mais severas dessas doenças. Essa abordagem, permite o manejo adequado desses pacientes desde o início da infecção e também possibilitando a aplicação adequada de recursos nos casos com sinais claros de progressão para as formas mais severas da doença. Dessa forma, o projeto ARBOBIOS, do qual o Prof Dr. Helder Nakaya faz parte, pretende identificar transcritos biomarcadores ou assinaturas de expressão gênica que possam identificar precocemente pacientes com predisposição ao desenvolvimento de casos severos de Dengue e Chikungunya.

Dentro do projeto ARBOBIOS estão previstas a elaboração de duas coortes\* longitudinais com pacientes acometidos por Dengue e Chikungunya em diferentes cidades do Brasil. No momento da inclusão na coorte, e em duas outras ocasiões no decorrer da doença, amostras sanguíneas são coletadas para posterior extração de RNA e sequenciamento dos transcritos (transcritomica). Isso permitirá a identificação dos genes diferencialmente expressos que possam estar associados ao desenvolvimento dos casos severos da doença e que possam ser utilizados como biomarcadores. Além disso, no momento da inclusão na coorte e nas visitas futuras, são coletadas também diferentes informações clínicas, como os principais sintomas apresentados, necessidade de internação e período, assim como medicamentos utilizados, além de dados laboratoriais, como medidas de concentração de plaquetas e hemácias e medidas de enzimas marcadoras de atividade hepática, entre outros.

Dado o acúmulo de informações clínicas e laboratoriais detalhadas a respeito dos pacientes com Dengue e Chikungunya em diferentes locais do país e em diferentes momentos ao longo do desenvolvimento da doença, existe a possibilidade da exploração dessas informações dentro da área de Ciência de Dados. Uma área multidisciplinar que cresce com o avanço tecnológico e que apresenta como utilidade fundamental o estudo de informações do mais diversos tipos e quantidades, principalmente por meio de técnicas de *Machine Learning*. Tais técnicas consistem em algoritmos com conceitos estatísticos, capazes de distinguir características e definir grupos ou condições para determinada situação a ser compreendida. Além disso permite a exploração e classificação dos dados de forma automatizada, pois em geral os *templates* desenvolvidos são semi-automáticos usando o mínimo de entradas manuais. Tal metodologia permite também a visualização por etapa do andamento do procedimento, sendo este ponto útil em identificar se o insumo usado está fazendo sentido ao problema proposto (ALPAYDIN, 2014). Dessa forma a análise dessas informações de maneira automatizada pode auxiliar na identificação de características chaves associadas a progressão da doença, e dessa forma auxiliar na correta classificação dos pacientes envolvidos.

\*Coortes: Conjunto de indivíduos que têm algo relevante para ser estudado comparativamente.

**JUSTIFICATIVA**

Dada a grande quantidade de informações clínicas e laboratoriais de pacientes incluídos nas coortes de Dengue ou Chikungunya do projeto ARBOBIOS, e a dificuldade de integração dessas informações para correta classificação desses pacientes, iremos aplicar técnicas de aprendizado de máquinas a fim de implementar soluções automatizadas para classificação dos pacientes em grupos de risco bem como para identificação de características chave para tal classificação.

Além de auxiliar no processamento e classificação das informações disponíveis, contribuindo para o desenvolvimento geral do projeto, a presente proposta permitirá uma integração dos conhecimentos em aprendizado de máquinas com informações biológicas reais, possibilitando o amadurecimento da aplicação de tal conhecimento em problemas envolvendo a complexidade de dados biológicos.

**OBJETIVO**

Desenvolvimento de um template (ou biblioteca) de técnicas de *machine learning*, comparando tipos de modelos de classificação e/ou inferência (supervisionados ou não supervisionados), afim definir uma clusterização mais específica de cada uma das doenças apresentadas, com a finalidade de determinar futuros pontos de concentração de casos e o grau de severidade que podem estar sendo expostos os indivíduos ali presentes.

**METODOLOGIA**

Para o desenvolvimento efetivo do *template*, serão consideradas as seguintes etapas:

* Definição do objetivo de interesse para modelagem (resposta).
* Processamento de dados antes da decisão e aplicação das técnicas modelagem, ou seja, desenvolver métodos para tratamento dos dados como limpeza de variáveis com x% de constantes (coluna com o mesmo tipo de informação ou dados nulos), correlação, limite de corte de variáveis para teste e validação (Uso de métricas como AUC e Curva ROC para determinar esses cortes/ limite de variáveis).
* Origem das informações, considerando tipo de bases de dados disponíveis, forma de acesso e necessidade ou não de mascarar os dados, caso a informação não seja de domínio público e/ou a entidade proprietária dos dados solicite visando à Lei Geral de Proteção de Dados (LGPD).
* Determinação prévia de importância das variáveis presentes, ou seja, informações que a *priori* se sabe que devem ser mantidas para dar significado biológico ao trabalho.
* Escolha inicial da(s) seguintes técnica(s) de modelagem (métodos supervisionados ou não) :

1. K-Means que consistem em um método de [Clustering](https://pt.wikipedia.org/wiki/Clustering" \o "Clustering) que objetiva [particionar](https://pt.wikipedia.org/wiki/Parti%C3%A7%C3%A3o_de_um_conjunto) *n* observações dentre *k* grupos onde cada observação pertence ao grupo mais próximo da média.

(BURT et al., 1967)

1. Random Forest que consiste em um método de aprendizado conjunto para classificação e regressão, que consiste de árvores de decisão com atributos aleatórios, que expressão o melhor ganho dado o peso de cada uma das árvores individualmente

(LORENZETT,2016).

1. Gradient Boosting que se caracteriza por se apresentar como uma técnica de aprendizado de máquina para problemas de regressão e classificação de forma aditiva, na qual as árvores de decisão são criadas inicialmente bom os dados de resíduo presentes da primeira decisória obtida dos dados origem. Sua vantagem é dada por em certas situações não ser necessário a manipulação dos dados para o desenvolvimento do modelo final (ZAMPIERI et al., 2015).

* Redução de erros do modelo (evitar overfitting e validação cruzada), ou seja, dada característica está de fato associada ao quadro que o indivíduo apresentou? O quão discrepante é um mesmo modelo aplicado em dois conjuntos de dados (validação e desenvolvimento).

Caso após as etapas exploratórias inicias seja detectado que as técnicas citadas acima não sejam suficientes ou adequadas para realizar a classificação esperada novo métodos de modelagem serão avaliados.

**RESULTADOS ESPERADOS**

Com o desenvolvimento desse projeto esperamos obter um modelo com parâmetros que aplicados nas bases de informações de pacientes com sinais de alarme para Dengue e Chikungunya, permitirá a obtenção de expressões e características, que classifiquem e indiquem a situação de paciente suscetíveis à progressão para casos mais severos das doenças.

**REFERÊNCIAS BIBLIOGRÁFICAS**

1. ALPAYDIN, Ethem. **Introduction to Machine Learning**. 3. ed. rev. e atual. [*S. l.*]: MIT Press, 2014. 613 p.
2. BEATTY, Mark *et al*. Estimating the global burden of dengue. **The American journal of tropical medicine and hygiene**, [*S. l.*], jan. 2009.
3. BURT, Felicity *et al*. Chikungunya virus: an update on the biology and pathogenesis of this emerging pathogen. **Review**, [*s. l.*], p. 107-117, 2017.
4. GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**: conceitos,ferramentas e técnicas para a construção de sistemas inteligentes. [*S. l.*]: Altas Books, 2019. 573 p. v. 1.
5. GUBLER, Dj. The global emergence/resurgence of arboviral diseases as public health problems.  ago. 2002.
6. GRUS, Joel*.* **Data Science do Zero**: Primeiras regras com o python. [*S. l.*]: Altas Books, 2019. 336 p. v. 1.
7. LORENZETT, Cassio Dal *et al*. **Estudo Comparativo Entre Algoritmos de Mineração de Dados Random Forest e J48 na Tomada de Decisão.** 2016
8. MYCKINNEY, Wes. **Python para Análise de Dados**: Tratamento de dados com pandas,numpy e ipython. [*S. l.*]: Novatec Editora, 2018. 616 p. v. 1.
9. MACQUEEN, J. **Methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**. [*S. l.*], 1967. Disponível em: https://projecteuclid.org/euclid.bsmsp/1200512992. Acesso em: 10 mar. 2020.
10. SMITH, Annelies *et al*. Dengue. **Seminar**, [*s. l.*], p. 350-363, 2019.
11. SUHRBIER, Andreas. Rheumatic manifestations of chikungunya: emerging concepts and interventions. **Nature Reviews Rheumatology**, [*S. l.*], v. 15, p. 597-611, 3 set. 2019.
12. S., Bhatt *et al*. The global distribution and burden of dengue. **Nature**, [*S. l.*], p. 504-507, 7 abr. 2013.
13. WEAVER, SC *et al*. Present and future arboviral threats. **Antiviral Res.**, [*S. l.*], 24 out. 2009.
14. ZAMPIERI, Fernando *et al*. Modelo de análise gradiente boosted do impacto do índice de massa corporal nos desfechos em curto prazo de pacientes clínicos gravemente enfermos. **Revista Brasileira de Terapia Intensiva**, [*S. l.*], v. 27, n. 2, 9 abr. 2015.