**UNIVERSIDADE DE SÃO PAULO**

**INSTITUTO DE MATEMÁTICA E ESTATÍSTICA**

**RELATÓRIO PARCIAL DE ATIVIDADES DA DISCIPLINA MAP2020**

**REFERENTE AOS MESES DE MARÇO E ABRIL**

**SÃO PAULO**

**01/05/2020**

**TRABALHO DE FORMATURA – T.2020154**

**BMA e BMAC**

**Relatório dos meses de Março e Abril**

1. NOME DO ALUNO: Fabio Carvalho de Souza

2. NÚMERO USP: 9425125

3. CURSO: [ ] BMA [ x ] BMAC

4. HABILITAÇÃO:

BMA: [ ]101 [ ] 501 [ ] 611 [ ] 801

BMAC: [ ] 104 [x] 204 [ ] 304 [ ] 404 [ ] 504 [ ] 604 [ ] 704 [ ] 804 [ ] 904 [ ] 1004

5. NOME DO ORIENTADOR: [**Helder Takashi Imoto Nakaya**](http://csbiology.com/)

UNIDADE DO ORIENTADOR: **Faculdade de Ciências Farmacêuticas (FCF)**

6. NOME DO CO-ORIENTADOR (se houver): À Definir

UNIDADE DO CO-ORIENTADOR: À Definir

7. TÍTULO DO PROJETO: **APLICAÇÃO DE MACHINE LEARNING PARA CLUSTERIZAÇÃO DE PACIENTES COM ARBOVIROSES**

**INTRODUÇÃO**

A aplicação de métodos de *machine learning*, em um conjunto de dados biológicos e não biológicos de pacientes que apresentaram alguns tipos de sintoma que estejam relacionados com arboviroses como Dengue (que apresenta 4 tipos DENV1 ao DENV4 sendo este último o caso severo da doença) ou Chikungunya. E com base nos resultados desses métodos, conseguir determinar padrões de sintomas/características que apontem para condições em que os pacientes possam ou não estar aptos a apresentar o caso severo da doença. 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). Dessa forma as atividades iniciais foram definir o alvo das analises, bem como determinar técnicas e métodos para serem aplicados aos dados e dessa forma começar a observar algum significado nas informações.

**ATIVIDADES REALIZADAS**

Inicialmente com o conjunto de aproximadamente 1400 pacientes com informações obtidas por meio do projeto ARBOBIOS, para ambas as doenças, realizou-se a separação dos grupos, ou seja, em primeiro momento se está trabalhando apenas com os pacientes que apresentam marcação para Dengue (aproximadamente 1200 casos), e com eles realizar todo o processo de desenvolvimento do *template* de modelagem e tratamento, que após, em segundo momento será aplicado nos demais casos.

O grupo de verificação inicial, apresenta um conjunto de 703 variáveis das quais precisaram ser analisadas para eliminar dados que claramente não apresentam importância para a modelagem como por exemplo variáveis que envolvem datas (exceto nascimento e entrada no grupo Dengue) e/ou dados que sejam de identificação dos pacientes ou ainda que não apresentam ligação com casos de Dengue, de modo que as mesmas foram retiradas da base, junto com outras mais que apresentaram um valor de constância elevado (variáveis com apenas 1 tipo de informação) e um grau especifico de correlação (Anexo 1),



Anexo 1

assim totalizando uma base com 84 variáveis para serem trabalhadas em outros aspectos, como por exemplo o tratamento de variáveis *strings,* ou seja texto, e para isso se utilizou de uma função em Python para determinar o *targuet encode* das *strings* com base na resposta desejada, ou seja considerando a quantidade de grupos distintos presentes em cada variável , o total de indivíduos da coluna e a quantidade de *targuets* positivos que temos, é possível determinar um valor para cada tipo de *string*, e dessa forma tratar dados que poderiam causar problemas em algumas das técnicas usadas. Vale ressaltar que para obter o grupo *targuet,* foi considerado um conjunto de 300 pacientes previamente classificados de forma manual que apresentam entre eles casos severos e os mesmos foram marcados com *flags* (1 para paciente severo e 0 para paciente não severo) para uso no modelo e aplicar o *targuet encode*.

**TÉCNICAS E MÉTRICAS APLICADAS**

Após a aplicação de todos os tratamentos devidos, foram realizados testes usando a técnica de Random Forest com uma vasta gama de parâmetros (profundidades variando de 3 a 9 e árvores de 100 a 500) de modo a treinar e testar diversos modelos, usando os pacientes que foram classificados manualmente, pois como temos maior certeza que os mesmos tiverem Dengue, eles se tornam um grupo de treino forte e podem ajudar na descriminação de casos com maior precisão. Para validar e verificar o nível de importância das variáveis presentes e o modelo em si se considerou algumas medidas estatísticas de analise como:

AUC:

O valor do **AUC** varia de 0 até 1 e o limiar entre a classe é 0,5. Ou seja, acima desse limite, o algoritmo classifica em uma classe e abaixo na outra classe.

Curva ROC:

A curva **ROC**mostra o quão bom o modelo criado pode distinguir entre duas coisas (já que é utilizado para classificação). Essas duas coisas podem ser 0 ou 1, ou positivoenegativo.

KS:

  Teste Kolmogorov–Smirnov (também conhecido como teste KS) é um teste não paramétrico sobre a igualdade de distribuições de probabilidade contínuas e unidimensionais que pode ser usado para comparar uma amostra com uma distribuição de probabilidade de referência. O KS também quantifica a distância entre a função distribuição empírica da amostra e a função distribuição acumulada da distribuição de referência ou entre as funções distribuição empírica de duas amostras.

Analisando as medidas citadas, verificou-se que o modelo está apresentando overfitting ou seja, ao observar cada medida do modelo de Random Forest para treino e validação os valores se apresentaram muito discrepantes, sendo que em modelos ajustados os valores são consideravelmente próximos em cada etapa, logo se deve observar novos parâmetros e técnicas de modelagem, para ajustar o significado que o modelo apresenta sobre os casos de estudo. Apesar desses pontos observando o gráfico de importância de variáveis (Anexo 2) temos algumas variáveis com grau considerado de expressão e que clinicamente e biologicamente fazem sentido.



Anexo 2

**PASSOS SEGUINTES**

Observando todos estes pontos a continuidade das análises dependem de:

1. Verificar os parâmetros usados no modelo e verificar o intervalo de apresente menor valor de overfitting, assim ao obter as métricas teremos valores de um modelo ajustado, bem como testar novas técnicas que podem em sua estrutura lógica apresentar parâmetros internos padrões que já evitam problemas de overfitting e ainda podem apresentar respostas, que necessitem de tipos diferentes de métricas estatísticas.
2. Analisar as unidades presentes em algumas colunas, pois foi verificado que valores em certas variáveis são muito discrepantes e com alta variação, que podem estar em outras unidades, pelo fato que alguns campos de preenchimento dos formulários da base não terem sido padronizados, e os mesmos foram obtidos de diversas regiões o país e esse “descuido” precisa ser observado para não extrapolar e mascarar informações relevantes ou não.
3. Apresentar uma primeira base final com as correções devidas e que já tenha a marcação que será utilizada para descriminar os casos os 4 tipos de sorotipos para Dengue (DENV1 ao DENV4) e o conjunto de variáveis que mais expressou significância para os pacientes terem a marcação apresentada.

**REFERÊNCIAS BIBLIOGRÁFICAS**

1. BURT, Felicity *et al*. Chikungunya virus: an update on the biology and pathogenesis of this emerging pathogen. **Review**, [*s. l.*], p. 107-117, 2017.
2. Daniel, Wayne W. (30 de junho de 2000). [*Applied Nonparametric Statistics*](https://books.google.com.br/books?id=bCDFAAAACAAJ&dq=Applied+Nonparametric+Statistics&hl=pt-BR&sa=X&redir_esc=y) (em inglês). [S.1] Duxbury ISBN 9780534381943
3. SMITH, Annelies *et al*. Dengue. **Seminar**, [*s. l.*], p. 350-363, 2019.
4. [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.