RELATÓRIO PARCIAL DE ATIVIDADES

REFERENTES AOS MESES DE SETEMBRO E OUTUBRO

São Paulo, Outubro de 2020

TRABALHO DE FORMATURA – T.2020154
BMA e BMAC

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: **Prof. Dr. Helder Takashi Imoto Nakaya**
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**

Fabio C.S

_____ _____

Assinatura do aluno Assinatura do orientador

RESUMO

Dada a necessidade de melhorar as medidas dos modelos que foram propostos, as atividades desse bimestre consistiram em realizar diversas etapas de treinos e testes dos modelos, considerando uma vasta gama de possíveis cenários, buscando assim encontrar falhas em algum processo, bem como deixar mais especifico o modelo a ser considerado como aplicável, e que tenha a melhor representatividade de acordo com o problema e a disponibilidade de informações.

1- INTRODUÇÃO

Durante todo o desenvolvimento do *template*, diversos processos foram criados e removidos, com o intuito de melhorar a seleção de informações e parâmetros, de forma que expressassem melhor a classificação dos casos de estudo, contudo um processo que foi muito realizado, consiste na adição e remoção de variáveis, pois as observando, tinha-se situações em que sua importância não era tão relevante ao modelo ou ainda que não estava claro se as mesmas estavam adequadas para uso. Assim para a melhor seleção de informações, foi adicionado um processo de modelagem intermediário denominado *features selection*, por meio da técnica de Lasso, do qual aplicado numa regressão logística se verificava a expressividade de dada variável e assim podia-se remover informações que pouco se relacionavam com a resposta disponível, dessa forma o conjunto de variáveis seria reduzido e poderíamos apresentar um modelo com maiores acertos. Em oposto ao fato de remover o que não era expressivo para o modelo, dados laboratoriais que antes foram de certa forma desconsiderados dada a incerteza das métricas usadas para preenchimento, foram adicionados junto ao processo do Lasso para analisar se as mesmas ainda permaneceriam nas etapas de treino e teste, essa opção foi adicionada novamente pois foi conhecido que um mesmo indivíduo, foi responsável pela alimentação dos dados de Rio Preto e Palmas, que são os dados usados no processo, logo se apresentou uma maior segurança em usar essas informações.

2- ATIVIDADES DESENVOLVIDAS E RESULTADOS PARCIAIS

Como já citado, um novo conjunto de variáveis foi adicionado ao processo, para que fosse possível melhorar a discriminação dos modelos desenvolvidos ampliando assim as possibilidades (Tabela 1), bem como um novo processo para eliminação das informações que se apresentassem pouco relevantes para o modelo (Tabela 2). Com isso foram realizados diversos processos para se analisar qual seria a melhor ou melhores opções para se manter.

TABELA 1: VARIAVEIS INICIAIS
sintoma_febre_7dias_Dor de cabeÃ§a
sintoma_febre_7dias_Dor atrÃ¡s dos olhos
sintoma_febre_7dias_Dor muscular (costas, coxas, panturrilhas, braÃ§os)
sintoma_febre_7dias_Dor nas articulaÃ§Ãµes (juntas)
sintoma_febre_7dias_Manchas na pele
sintoma_febre_7dias_Sinais de artrite (inflamaÃ§Ã£o nas articulaÃ§Ãµes)
sintoma_febre_7dias_Sem sintoma
sintoma_febre_7dias_Conjuntivite
sintoma_febre_7dias_InchaÃ§o
SAQ_Dor intensa e contÃ-nua no abdome (barriga), espontÃ¢nea ou ao apertar
SAQ_VÃ´mitos persistentes

SAQ_Sangramento de mucosas (nariz, gengiva, etc)
SAQ_PressÃ£o baixa, escurecimento da vista, sudorese ou desmaio ao se levantar
SAQ_Muito sono, ideias confusas, fala atrapalhada
SAQ_Queda abrupta de plaquetas
SA_D07_VÃ´mitos persistentes
SA_D07_! 'Muito sono, ideias confusas, fala atrapalhada '
SA_D07_Manchas vermelhas na pele
SA_D07_VÃ´mitos com sangue
SA_D07_Dor intensa e contÃ-nua no abdome (barriga), espontÃ¢nea ou ao apertar
SA_D07_Sem sinais
SA_D07_! 'Sangramento de mucosas (nariz, gengiva, etc) '
SA_D07_Sangramento na urina ou fezes
SA_D07_PressÃ£o baixa, escurecimento da vista, sudorese ou desmaio ao se levantar
SA_D07_Fluxo de sangue persistente e volumoso pelo Ãºtero
SA_D14_Sem sinais
SA_D14_! 'Muito sono, ideias confusas, fala atrapalhada '
SA_D14_Manchas vermelhas na pele
SA_D14_PressÃ£o baixa, escurecimento da vista, sudorese ou desmaio ao se levantar
SA_D14_Dor intensa e contÃ-nua no abdome (barriga), espontÃ¢nea ou ao apertar
SA_D14_VÃ´mitos persistentes
SA_D14_! 'Sangramento de mucosas (nariz, gengiva, etc) '
SA_D14_Sangramento na urina ou fezes
SA_D14_Fluxo de sangue persistente e volumoso pelo Ãºtero
SA_D14_VÃ´mitos com sangue
MARCADOR_SuPAR
HT 1
HB 1
PLAQ 1
TGO 1
TGP 1
LEUCO 1
IGG_nao_reativo
IGM_nao_reativo
PCR_nao_reativo
IGG_reativo
IGM_reativo
PCR_reativo
febre_7dias_Sim
sangramento_NÃ£o
sangramento_Sim
raca_Branca
raca_Parda
raca_Preta
ja_teve_dengue_NÃ£o que eu saiba
ja_teve_dengue_Sim, confirmada por exame de sangue do tipo sorologia
ja_teve_dengue_Sim, por diagnÃ³stico do mÃ©dico
ja_teve_chikungunya_NÃ£o que eu saiba
ja_teve_zika_NÃ£o que eu saiba
tem_alguma_doenca_das_articulacoes_NÃ£o
exantema_NÃ£o
exantema_Sim
edema_NÃ£o
edema_Sim
sinais_artrite_NÃ£o
sinais_artrite_Sim
outros_sinais_alarme_NÃ£o
outros_sinais_alarme_Sem registro
outros_sinais_alarme_Sim
esteve_hospitalizado_D07_NÃ£o
esteve_hospitalizado_D07_Sim
esteve_hospitalizado_agravamento_dengue_D07_NÃ£o sabe
esteve_hospitalizado_agravamento_dengue_D07_Sim
esteve_hospitalizado_D14_NÃ£o
esteve_hospitalizado_agravamento_dengue_D14_NÃ£o sabe
fl_severidade
age_crianca
age_idoso
fl_sexo
age_adolecente
age_adulto

TABELA 2 : VARIAVEIS PARA MODELAGEM APÓS USO DO LASSO
sintoma_febre_7dias_Dor muscular (costas, coxas, panturrilhas, braÃ§os)
sintoma_febre_7dias_Manchas na pele
sintoma_febre_7dias_Sinais de artrite (inflamaÃ§Ã£o nas articulaÃ§Ãµes)
sintoma_febre_7dias_InchaÃ§o
SAQ_Dor intensa e contÃ-nua no abdome (barriga), espontÃ¢nea ou ao apertar
SAQ_VÃ´mitos persistentes
SAQ_PressÃ£o baixa, escurecimento da vista, sudorese ou desmaio ao se levantar
SAQ_Muito sono, ideias confusas, fala atrapalhada
SAQ_Queda abrupta de plaquetas
SA_D07_VÃ´mitos persistentes
SA_D07_! 'Muito sono, ideias confusas, fala atrapalhada '
SA_D07_VÃ´mitos com sangue
SA_D07_Dor intensa e contÃ-nua no abdome (barriga), espontÃ¢nea ou ao apertar
SA_D07_! 'Sangramento de mucosas (nariz, gengiva, etc) '
SA_D07_PressÃ£o baixa, escurecimento da vista, sudorese ou desmaio ao se levantar
SA_D07_Fluxo de sangue persistente e volumoso pelo Ãºtero
SA_D14_! 'Muito sono, ideias confusas, fala atrapalhada '
SA_D14_Manchas vermelhas na pele
SA_D14_Dor intensa e contÃ-nua no abdome (barriga), espontÃ¢nea ou ao apertar
SA_D14_VÃ´mitos persistentes
SA_D14_! 'Sangramento de mucosas (nariz, gengiva, etc) '
SA_D14_VÃ´mitos com sangue
MARCADOR_SuPAR
HT 1
HB 1
PLAQ 1
TGO 1
TGP 1
IGG_nao_reativo
IGG_reativo
sangramento_NÃ£o
sangramento_Sim
raca_Parda
ja_teve_dengue_NÃ£o que eu saiba
ja_teve_zika_NÃ£o que eu saiba
tem_alguma_doenca_das_articulacoes_NÃ£o
edema_NÃ£o
sinais_artrite_NÃ£o
outros_sinais_alarme_NÃ£o
outros_sinais_alarme_Sem registro
esteve_hospitalizado_agravamento_dengue_D07_Sim
esteve_hospitalizado_D14_NÃ£o
esteve_hospitalizado_agravamento_dengue_D14_NÃ£o sabe
fl_severidade
age_crianca
fl_sexo
age_adolecente

2.1- CENÁRIOS E RESULTADOS

Feita a seleção das variáveis que se mostraram relevantes para aplicação nos modelos, foi pensado quais seriam as possibilidades a verificar para cada uma das técnicas de modelagem propostas (Regressão Logistica, Random Forest, Árvore de Decisão e Naive Bayes), e se chegou as seguintes possibilidades que foram testadas:

(a) Modelar todas as variáveis disponíveis.

(b) Modelar apenas as variáveis com informações laboratoriais.

(c) Modelar usando os dados de Rio Preto como treino e aplicar nos dados provenientes de Palmas como etapa de teste, com todas as variáveis.

(d) Modelar usando os dados de Palmas como treino e aplicar nos dados provenientes de Rio Preto como etapa de teste, com todas as variáveis.

(e) Modelar usando os dados de Rio Preto como treino e aplicar nos dados provenientes de Palmas como etapa de teste, com apenas os dados laboratoriais.

(f) Modelar usando os dados de Palmas como treino e aplicar nos dados provenientes de Rio Preto como etapa de teste, com apenas os dados laboratoriais.

Dessa forma é possível observar um conjunto de possibilidades que pode trazer maior certeza sobre o modelo a ser escolhido, pois encontrando um modelo que se comporte adequadamente na maioria as situações, pode ser aplicado com maior segurança e efetividade. Outro ponto a se comentar que para cada técnica em cada cenário foram desenvolvidos 2 modelos, pois se está lidando com processos de *Upsampling* e *Downsampling*, que consistem em equilibrar os dados na etapa de treino e aumentar um pouco a expressão da variável resposta.

2.1.1- RESULTADOS

Dado o grande número de arquivos gerados para os diversos casos, estão aqui comentados os resultados observados, e os mesmos estão disponíveis em anexo para observação, com tabelas de medidas e gráficos de variáveis importantes.

Verificação dos modelos caso *Upsampling*

1A- O cenário onde os modelos foram desenvolvidos com todas as variáveis os modelos *Upsampling* apresentam desempenho bem próximos e com boa classificação com apenas o modelo *random forest* que deve uma leve diferença e apresenta erro em alguns casos.

2A- O cenário onde os modelos foram desenvolvidos considerando apenas os dados laboratoriais temos que o desempenho é levemente pior que do citado acima, mas essa diferença é verificada como leve pois o número de casos que apresentam erro ainda é observado apenas algumas unidades, e podemos considerar usar esses modelos, pois estão classificando bem.

3A- O cenário onde os modelos foram desenvolvidos considerando os dados de Rio preto como treino e Palmas como teste para todas as variáveis, temos que todos os modelos apresentam um erro bem maior do que apresentado nas condições passadas, sendo assim não muito bons para uso, com isso os 2 citados anteriormente se tornam ainda os melhores.

4A- O cenário onde os modelos foram desenvolvidos considerando os dados de Rio preto como teste e Palmas como treino temos que os modelos de regressão logística e *random forest* tem melhor classificação e poderiam ser usados, enquanto que o modelo de árvore não tem uma boa separação para os casos dados como severos, apresentando erro e falso negativo.

5A- O cenário onde os modelos foram desenvolvidos considerando os dados de Rio preto como teste e Palmas como treino apenas usando os dados laboratoriais temos que todos os modelos apresentam uma classificação bem melhor, considerando sua versão com todas as variáveis e podendo chegar ao nível dos 2 casos citados inicialmente, logo podendo sim serem usados.

6A- O cenário onde os modelos foram desenvolvidos considerando os dados de Rio preto como treino e Palmas como teste apenas para dados laboratoriais temos que todos os modelos apresentam treino bom, mas o teste tem resultados muito ruins logo podemos não considerar usar essa opção de modelo.

Verificação modelos caso *Downsampling*

1B- O cenário onde os modelos foram desenvolvidos considerando todas as variáveis temos que os modelos para versão reduzida, são modelos bons mas comparados com os casos *Upsampling* são menos assertivos, logo podemos desconsiderar esses casos.

2B- O cenário onde os modelos foram desenvolvidos considerando apenas as variáveis de laboratório temos que são bons modelos, mas ainda com baixa assertividade se comparados com alguns dos casos *Upsampling*, podendo assim ser desconsiderados para uso.

3B- O cenário onde os modelos foram desenvolvidos considerando os dados de Rio preto como treino e Palmas como teste temos que os modelos apresentam uma assertividade boa para os casos severos, mas apresenta um número de casos elevados para falsos positivos, sendo assim ruim e podemos desconsiderar seu uso.

4B- O cenário onde os modelos foram desenvolvidos considerando os dados de Rio preto como teste e Palmas como treino temos que os modelos apresentam boa classificação, podendo sim ser uma opção para uso.

5B- O cenário onde os modelos foram desenvolvidos considerando os dados de Rio preto como teste e Palmas como treino apenas para dados laboratoriais, temos que os modelos na etapa de treino tem uma separação um pouco ruim, mas que ao serem aplicados no conjunto teste, apresentam uma melhoria, mas fica a critério de observação se deve ser um modelo usado ou não, mas de modo geral seria um bom modelo para classificação, dado que fosse usado em mais conjuntos de dados para validar se o mesmo apresenta resultados constantes dos casos dados como positivos.

6B- O cenário onde os modelos foram desenvolvidos considerando os dados de Rio preto como treino Palmas como teste apenas para dados laboratoriais, temos que os modelos não apresentam boa separação de casos severos, apresentando muitos erros de falso negativo, logo é um modelo que pode ser desconsiderado para uso.

2.1.2- CONCLUSÕES

De forma geral os modelos classificam bem e de certa forma poderíamos usar qualquer um na maioria dos casos, mas alguns apresentam um desempenho que se deve ainda pensar como melhor usar, mas de forma geral podemos observar e decidir entre os modelos *Upsampling* 1A,2A,4A e 5A, valendo verificar as métricas novamente e decidir. Considerando agora os casos *Downsampling* podemos citar que os modelos 1B,4B podem ser observados e podem a depender da análise serem usados, vale ainda citar o 5B que pode sim ser usado também mas ainda verificando outros pontos. Como uma das opções seria usar os modelos com variáveis com dados de fácil acesso (que neste caso seriam os dados laboratoriais), que dentre os citados acima como opções podemos observar em mais detalhes para os modelos presentes nas situações 2A,4A e 5A no caso *Upsampling*. Vale mencionar que os modelos que envolvem a técnica Naive Bayes não foram considerados pois durante todos os testes sua separação não está sendo satisfatória e dessa forma foi optado por não considerar mais como uma técnica válida para esse problema, possivelmente motivados por uma distribuição de informações que não satisfaz os critérios mínimos para uso dessa técnica.

3- ANEXOS

- Arquivos para observação junto aos comentários disponível em:

https://github.com/Fabio343/Projeto-TCC-Machine-Learning-Para-Classificar-Arborivoses/tree/master/Arquivos%20Relat%C3%B3rio/Medidas%20Relatorio%20Setembro-Outubro