Universidade De São Paulo
Instituto de Matemática e Estatística

ABORDAGEM BIOINFORMÁTICA: APLICAÇÃO DE *MACHINE LEARNING* PARA CLUSTERIZAÇÃO DE PACIENTES COM ARBOVIROSES

Fabio Carvalho de Souza

São Paulo

2020

Fabio Carvalho de Souza

ABORDAGEM BIOINFORMÁTICA: APLICAÇÃO DE *MACHINE LEARNING* PARA CLUSTERIZAÇÃO DE PACIENTES COM ARBOVIROSES

Trabalho de Conclusão de Curso apresentado ao Instituto de Matemática e Estatística da Universidade De São Paulo, como parte dos requisitos para obtenção do grau de Bacharel em Matemática Aplicada e Computacional.

Orientador: Prof. Dr. Helder Takashi Imoto Nakaya
Departamento de Análises Clínicas e Toxicológicas-Faculdade de Ciências Farmacêuticas USP

Coorientador: Dr. Felipe ten Caten
Departamento de Moléstias Infecciosas e Parasitárias - Faculdade de Medicina USP

São Paulo 2020

Fabio Carvalho de Souza

ABORDAGEM BIOINFORMÁTICA: APLICAÇÃO DE *MACHINE LEARNING* PARA CLUSTERIZAÇÃO DE PACIENTES COM ARBOVIROSES

BANCA EXAMINADORA

Prof.Dr. Anatoly Yambartsev (IME-USP)

Prof.Dr. Helder Takashi Imoto Nakaya (FCF-USP)

Prof.Dr. Ronaldo Fumio Hashimoto (IME-USP)

Agradecimentos

Agradeço primeiramente a Deus pela força em trilhar essa etapa da vida, cheia de descobertas e que me tornou um indivíduo melhor em alguns aspectos.

Agradeço a minha família pelo apoio durante todo esse processo e principalmente no inicio de tudo com tantas dificuldades.

Agradeço ao Prof. Helder pela orientação e ao Felipe pela coorientação, foi uma ajuda incrível durante o processo, bem como agradecer a oportunidade de conhecer um pouco mais sobre bioinformática e poder aplicar o que conheço de ciência de dados.

Agradeço a todos os colegas que fiz e que ajudaram e proporcionaram momentos durante essa jornada.

Resumo

O grande número de casos de Dengue que ocorrem anualmente em todo o mundo e a ampla distribuição de suas espécies vetores, com expansão para novas áreas devido às mudanças climáticas, colocam essa doença como um grave problema de saúde pública em todo o planeta. Apesar de sua ampla disseminação, ainda não são completamente conhecidos os fatores associados ao desenvolvimento da forma mais grave da doença (Dengue Severa), que em casos extremos pode levar a óbito. Assim como, ainda não existem protocolos capazes de identificar precocemente os indivíduos que possuam maior potencial de agravamento da doença. Dessa forma, nesse trabalho diferentes técnicas de aprendizado de máquinas foram aplicadas para processamento e análise, de forma automatizada, de informações clínicas e laboratoriais de uma coorte de pacientes com sinais de alarme para o desenvolvimento de Dengue Severa provenientes de 5 centros hospitalares do Brasil. A aplicação de tais técnicas permitiu a validação da classificação realizada para parte dos pacientes da coorte, bem como a identificação de variáveis chave associadas com a predição dos pacientes com maior propensão ao desenvolvimento de Dengue Severa, como alterações nas medidas de plaquetas, hematócritos e também nas medidas de marcadores de dano hepático como TGO e TGP. Além disso, o estabelecimento de *templates* padronizados de análise ao longo do trabalho permitirá o fácil processamento dos novos dados disponíveis à medida que novos pacientes sejam incluídos na coorte.

Palavras-chave: Dengue, arboviroses, Ciência de dados, *machine learning*, Bioinformática, Regressão Logística, Árvore de decisão, Naive Bayes, *Random Forest*

Abstract

The large number of Dengue cases that occur annually worldwide and the wide distribution of its vector species, with expansion into new areas due to climate change, place this disease as a serious public health problem globally. Despite its wide occurrence, the factors associated with the development of the most severe form of the disease (Severe Dengue), which in extreme cases can lead to death, are still not completely known. Likewise, there are still no protocols capable of early identification of individuals with potential of bad resolution of disease. Thus, in this work, different machine learning techniques were applied for the automated processing and analysis of clinical and laboratory information from a cohort of patients with warning signs of Severe Dengue from 5 hospital centers in Brazil. The application of such techniques allowed the validation of the classification available for part of the patients in the cohort, as well as the identification of key variables associated with the prediction of patients with greatest potential of progression to Severe Dengue, such as changes in platelet counts, hematocrits and also measures of markers related to liver damage as GOT and GPT. In addition, the establishment of standardized analysis templates throughout the work will allow easy processing of the new data available once new patients are included in the cohort.

Keywords Dengue, arboviroses, Data science, *machine learning*, Bioinformatics,Logistic Regression, Decision Tree, Naive Bayes, *Random Forest*

Lista de Figuras

Lista de Tabelas

Sumário

1 INTRODUÇÃO

1.1 Dengue Aspectos Gerais

A Dengue é uma das doenças virais mais difundidas no mundo, acometendo cerca de 390 milhões de pessoas em todo o planeta todos os anos. (BHATT 2013; SUHRBIER 2019). Seu agente causador, o vírus da Dengue (DENV), é um vírus de RNA de fita simples pertencente ao gênero Flavivirus, da família Flaviviridae, do qual fazem parte também o vírus da Febre Amarela (YFV) e da Zika (ZIKV), e apresenta 4 sorotipos distintos (DENV1 - DENV4). A Dengue é transmitida por mosquitos do gênero *Aedes*, principalmente *Aedes aegypti*, e de forma menos eficiente por *Aedes albopictus* (WILDER SMITH 2019). *A. aegypti* é predominante no Brasil e também é vetor de doenças como Febre Amarela e Chikungunya, enquanto *A. albopictus* se originou na região asiática, mas tem se expandindo para outras regiões, como Europa e os continentes Americanos, elevando assim o número de pessoas sob risco de infecção por DENV (KRAEMER et. al. 2015).

A Dengue faz parte das arboviroses, um grupo de doenças virais transmitidas por vetores artrópodes como moscas, mosquitos e carrapatos (*Arthropod-borne virus*). Apesar de serem transmitidos por vetores de um mesmo grupo, os arbovírus compreendem famílias virais distintas, incluindo as famílias Flaviviridae (Dengue, Zika, Febre Amarela), Togaviridae (Chikungunya, Mayaro), Bunyavirales (Bunyavirus) entre outras. Além de serem um grande problema de saúde pública, principalmente em países em desenvolvimento, esse grupo de vírus pode causar sérios problemas de saúde em espécies de interesse agropecuário, trazendo também grandes prejuízos econômicos (HUBÁLEK et. al. 2014). Fatores como as mudanças climáticas, desmatamento e urbanização sem planejamento têm alterado ecossistemas ambientais e aproximado populações humanas

e os vetores dessas doenças, aumentando sua capacidade de disseminação e o possível surgimento de novas epidemias relacionadas aos arbovírus (GOULD et. al. 2017).

Após a entrada no organismo, o vírus da Dengue apresenta um período de incubação de 4 a 7 dias (ST. JOHN et. al. 2019) . Após esse período 75% dos pacientes desenvolvem a doença de forma assintomática, ou seja, não apresentam qualquer sintoma da infecção, enquanto que 25% dos pacientes apresentam os sintomas característicos da fase aguda da infecção: febre, calafrios, mal-estar, vômito e vermelhidão (WILDER SMITH 2012 & 2019). Após a fase aguda boa parte dos pacientes entra na fase de recuperação, com diminuição da carga viral e controle dos sintomas, no entanto, em alguns casos os pacientes podem evoluir para um quadro mais severo (Dengue Severa) caracterizada principalmente por aumento da permeabilidade vascular e extravasamento de plasma, podendo levar a choque hemodinâmico (caracterizado por taquicardia), pulso fraco ou quase indetectável, com possibilidade de síndrome do desconforto respiratório e ocorrência de sangramentos, o que pode provocar complicações e levar o paciente a óbito. (WILDER SMITH 2019). A evolução para Dengue Severa pode estar associada a infecções secundárias de DENV, uma vez que anticorpos não específicos, previamente existentes no organismo para sorotipos heterólogos teriam ação subótima e poderiam ocasionar um aumento da entrada dos vírus nas células e aumento da resposta imunológica, um mecanismo conhecido como Aumento Dependente de Anticorpos (ST. JOHN et. al. 2019). Além disso, a evolução para a forma grave da doença geralmente está associada a presença de sinais de alarme que incluem dor abdominal, vômito persistente, acúmulo de fluidos, sangramento de mucosas, letargia ou inquietação, hepatomegalia (aumento do fígado), aumento no hematócrito (concentração de hemácias) e queda no número de plaquetas (GUZMAN et. al. 2016).

1.2 Projeto ARBOBIOS

Apesar de ser uma doença amplamente difundida, a Dengue apresenta um sério desafio principalmente em regiões subdesenvolvidas devido seu impacto nos sistemas de saúde. A existência de protocolos com especificações mínimas que sejam capazes de identificar previamente pacientes com maior ou menor risco de desenvolver a forma mais severa dessa doença, permite um manejo adequado desses pacientes desde o início da infecção, aumentando suas chances de recuperação. Além disso, possibilita também a aplicação adequada de recursos nos casos com sinais claros de progressão para as formas mais severas da doença, otimizando a aplicação dos escassos recursos disponíveis (WILDER SMITH 2019). Nesse contexto, o projeto ARBOBIOS, uma parceria entre a FAPESP, Instituto de Medicina Tropical - USP e bioMérieux Brasil Indústria e Comércio de Produtos Laboratoriais Ltda., do qual o nosso laboratório, sob coordenação do Prof Dr. Helder Nakaya, faz parte, pretende identificar genes biomarcadores ou assinaturas de expressão gênica que possam identificar precocemente pacientes com predisposição ao desenvolvimento de casos severos de Dengue e Chikungunya. No âmbito desse projeto foram coletadas informações de indivíduos com sinais de alarme para Dengue Severa em 5 diferentes locais do Brasil, incluindo Campo Grande - MS, São José do Rio Preto - SP, Araraquara - SP, Arcos - MG e Palmas - TO. No momento da inclusão dos pacientes na coorte, foram coletadas informações clínicas, como sintomas e sinais de alarme, e laboratoriais, como medidas de plaquetas, hematócrito, hemoglobina entre outros. Novas informações clínicas foram coletadas 7 e 14 dias após a inclusão, para acompanhamento da evolução dos sinais de alarme. A infecção por DENV foi confirmada por exames de RT-PCR e/ou Elisa IgM. Os pacientes de São José do Rio Preto - SP e Palmas - TO, foram classificados por uma equipe médica, apresentando ou não o quadro de Dengue Severa.

Visto isso, a possibilidade da exploração dessas informações por meio da área de Ciência de Dados é uma opção bem recomendada, pois é 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*, que 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.

1.3 Modelos Propostos

A área de aprendizado de máquinas tem apresentado grande crescimento nos últimos anos, impulsionada principalmente pela facilidade de geração e aquisição de dados e também pelo aumento no desempenho dos computadores, possibilitando o agrupamento e análise de quantidades cada vez maiores de informações. Nesse cenário, diferentes modelos têm sido propostos para os mais variados contextos. A seguir apresentaremos alguns modelos que serão avaliados em relação a sua capacidade de explicar o problema biológico de classificação de pacientes severos e não severos de Dengue.

1.3.1 *Regressão Logística*

A regressão logística é uma das técnicas de modelagem mais usadas em diversas áreas que atuam com dados, como por exemplo o setor financeiro e a área de ciências biológicas. É uma técnica que consiste em classificar um evento em dois grupos sendo em geral um grupo A que apresenta a positiva do evento e um grupo B com a negativa, e com dessa forma determinar a probabilidade de uma nova observação estar mais propensa a ser pertencente ao grupo A ou B. Sua utilização consiste basicamente do uso de dados binários, ou seja, atribuímos 1 para o caso de ocorrência de uma situação e 0 para sua ausência, sendo isto válido tanto para a variável alvo, como para as de insumo (Figura 1). Como queremos lidar com a probabilidade do evento que observamos de estar relacionada ao grupo de interesse, temos que uma dada variável alvo X binária é colocada num processo junto com n variáveis X' que são insumos para analisar a variável alvo, dessa forma temos que a função distribuição da logística é apresentada por:

$$P = P(X = 1|X_i' = X') = \frac{1}{1 + e^{(-g)}}$$

com $g = A_0 + A_1X_1' + ... + A_nX_n'$, que são valores de parâmetros a serem obtidos por meio de estimativas se utilizando do método de máxima verossimilhança, que consistem em estimar os valores dos diferentes parâmetros com o objetivo de maximizar a probabilidade do evento, podemos também observar esses valores como o "peso"que cada variável apresenta no modelo e assim observar o que mais causa descriminação para a ocorrência da situação desejada.(HOSMER & LEMESHOW, 1989)

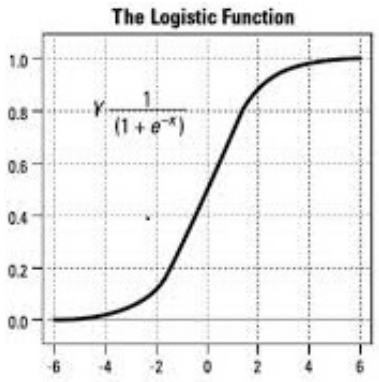


Figura 1: Ilustração Da Função Logística, eixo x representa o conjunto dos números reais e o eixo y a probabilidade que o evento pode assumir.

1.3.2 *Árvore de Decisão*

O método de árvore de decisão é uma técnica de modelagem de dados/ classificação, que se baseia na estratégia de "dividir para conquistar", que consiste em decompor o problema principal em situações menores que são condicionadas para que se possa determinar uma decisão sobre o passo seguinte a se seguir, considerando os atributos que mais se mostrem relevantes (Figura 2). Esta é uma forma de trabalho que se aplica durante todo o processamento da árvore de forma recursiva, sempre decompondo os problemas menores em situações mais simples. Sua estrutura é composta por nós que definem os atributos e ramos que são provenientes dos nós e apresentam um valor que o atributo pode receber, o mesmo valor é considerado no processo de tomada de decisão caso a condição seja satisfeita. Temos também nas árvores nós folhas que são basicamente classes originadas das informações de insumo, desta forma cada nó folha se relaciona com uma classe, e todo o trajeto saindo da raiz (condição base) até um desses nós é uma classe de condições que são aplicadas para a decisão (GARCIA, 2000).

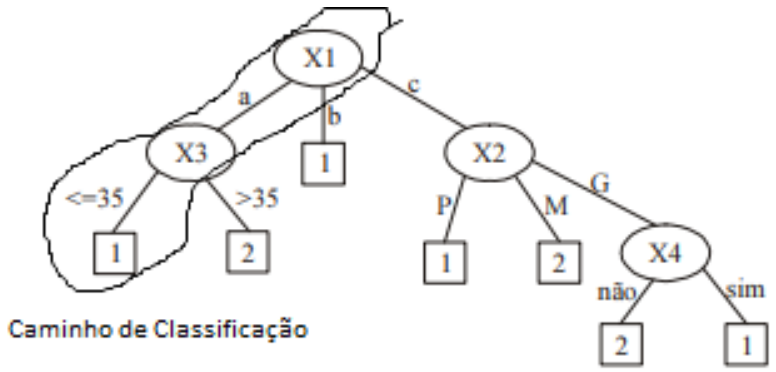


Figura 2: Ilustração-Árvore de decisão, onde os quadrados representam as folhas que são as condições possíveis que os círculos (que representam os nós) podem assumir para continuidade do processo

1.3.3 *Random Forest*

Em *Machine Learning* utilizamos um conjunto de dados para realizar o treinamento do modelo. Neste treinamento são utilizados classificadores combinados em diversas situações, cujo resultado servirá de critério de tomada de decisão dentro do modelo. O resultado dessa combinação promoverá uma melhor explicabilidade. Assim surge a necessidade de utilizar modelos, apresentem diversos classificadores chamados de métodos *ensemble*. Um desses métodos *random forest*, se baseia em diversas árvores de decisão para gerar seus resultados, visando minimizar qualquer tipo de irregularidade do modelo (Figura 3), tomando como ideia, diversas árvores para aprendizado paralelo onde cada modelo é independente e que se utiliza de repetição de amostras, reduzindo as altas variações que os preditores podem apresentar. Como resultado o *random forest* expressa cada árvore como um preditor, assim a decisão é dada por meio da característica que ocorreu em maior frequência durante o processo, ou seja, o quão importante uma variável é, se baseia no total de vezes que a mesma ocorreu entre as mais importantes de cada uma das árvores geradas, sendo assim um tipo de média de

ocorrências (BREIMAN, 2001).

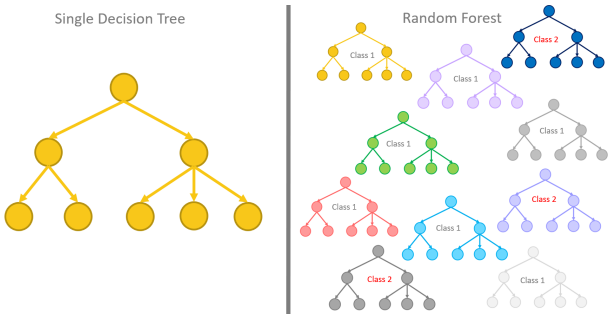


Figura 3: Ilustração- Árvore de decisão x Random Forest, um conjunto de árvores atua como classificador mais sofisticado e preciso.

1.3.4 Naive Bayes

O Naive bayes consiste em determinar uma classe de probabilidade baseada no teorema de Bayes que basicamente consiste em determinar as chances de um evento X ocorrer condicionado a um outro evento X' que ocorreu, conforme a expressão a seguir mostra:

$$P(X|X') = \frac{P(X'|X)P(X)}{P(X')}$$

Onde P(X) e P(X') são as chances de ocorrer X e X' e P(X'|X) é a probabilidade de ocorrer X' condicionado a X ter ocorrido.

O método de implementação deste modelo tem como hipóteses iniciais que cada uma das variáveis de entrada são independentes, o que significa que nenhuma das variáveis tem relação entre si para determinar uma característica e/ou evento ocorrido anteriormente. Temos também que *a priori* das probabilidades usadas pelo método, estão relacionadas com a proporção e similaridade de classes que estejam presentes na base de dados, o que proporciona uma resposta final que seja não relacionada de forma exclusiva com as variáveis de insumo em si, mas também pela proporção da ocorrência

dos casos positivos dentro do modelo (ZHANG, 2016).

2 JUSTIFICATIVA

Dado o grande desafio que as arboviroses, em especial a Dengue, impõem aos sistemas de saúde em muitos países é essencial a existência de protocolos capazes de identificar precocemente os pacientes com maior propensão de evolução para os quadros mais graves dessas doenças. A existência de tais protocolos permite o tratamento dos pacientes ainda na fase preliminar, com maiores chances de boa resolução da doença e também a aplicação adequada dos recursos que muitas vezes são limitados. A análise das informações clínicas e laboratoriais de pacientes com Dengue, coletadas no âmbito do projeto ARBOBIOS, através de abordagens de aprendizado de máquinas possibilitará a identificação de variáveis associadas à evolução das formas mais severa da doença. Além disso, a obtenção de modelos classificadores baseados em informações previamente curadas por equipe médica permitirá a classificação de pacientes com desfecho da doença ainda desconhecido, facilitando as etapas posteriores de identificação de pacientes com as formas severas da doença.

3 OBJETIVO

Este trabalho tem como objetivo principal aplicar técnicas de aprendizado de máquina (*Machine Learning*), para verificar possíveis variáveis e/ou pontos que consigam descriminar o que pode levar um indivíduo a apresentar um caso severo ou não, considerando informações reais de pacientes provenientes do projeto inicial ARBOBIOS. Para tanto, o projeto é composto pelos seguintes objetivos específicos:

- Realizar a seleção e criação de variáveis com determinadas características capazes

de discriminar os casos que poderiam apresentar maiores ou menores chances de evoluir para um caso mais severo de dengue.

- Testar diferentes algoritmos de aprendizado de máquinas com o intuito de identificar os classificadores com melhores resultados na predição de pacientes com quadros severos e não-severos de Dengue que já foram manualmente curados por equipe médica.

- Identificar as principais variáveis associadas à progressão para os casos severos da doença.

- Aplicar o modelo de classificação no conjunto de dados de paciente com desfecho da doença desconhecido. Permitindo assim a identificação dos pacientes com Dengue Severa e direcionando os esforços de análise do grupo de pesquisa.

- Desenvolver um *pipeline* que possa ser utilizado pelo grupo de pesquisa a medida que novos pacientes e dados sejam incluídos na coorte.

4 MATERIAIS E MÉTODOS

4.1 Informações Dos Dados

Até o momento 1159 pacientes com sinais de alarme de Dengue Severa foram incluídos na coorte no âmbito do projeto ARBOBIOS. Destes, 31 foram considerados inelegíveis de acordo com os critérios adotados no estudo (existência de doenças prévias) e 302 apresentaram resultados negativos para infecção por DENV nos exames de RT-PCR e/ou ELISA IgM. Dessa forma, o universo total de pacientes analisados nesse estudo compreende 829 indivíduos. Os pacientes de São José do Rio Preto - SP e Palmas - TO, 409 no total, foram classificados por uma equipe médica entre casos que evoluíram para Dengue Severa, e casos que não evoluíram, totalizando 42 casos de Dengue Severa, que será considerada a variável resposta para o problema. As informa-

ções coletadas no momento de inclusão na coorte, e nos dias 7 e 14 após a inclusão, estão organizadas no dicionário de variáveis disponível para verificação em anexo.

4.2 *Scripts*

Todos os scripts foram implementados na linguagem de programação Python versão 3.7, usando a IDE Spyder versão 4.1.4. Os scripts estão disponíveis para consulta no GitHub (https://github.com/Fabio343/Projeto-TCC-Machine-Learning-Para-Classificar-Arborivoses). Bem como demais análises foram observadas por meio de programas auxiliares como GraphpaD PRISM versão 6.0.

4.3 Tratamentos Específicos

Afim de padronizar os dados disponíveis e o processo de análise nas diferentes etapas desse trabalho os seguintes tratamentos foram realizados nos dados:

4.3.1 Inspeção visual e eliminação de informações

De forma manual observa-se quais variáveis estão presentes na base de dados e com base na proposta de modelagem que é desejada, monta-se uma lista com as variáveis que se deseja remover. Tais variáveis incluem informações como telefone de contato, unidade hospitalar, endereço, entre outras e provavelmente não tem nenhum tipo de relação com a resposta esperada .

4.3.2 Verificação de variáveis Strings

Consiste em verificar via função em Python (BaseDados .selectdtypes(include=['object'])), quais das variáveis que restaram, compreendem a informações não numéricas. Pois seu conhecimento evita erros de execução, visto que as informações não numéricas não podem ser usadas pelos modelos testados, necessitando assim de tratamentos de conversão.

4.3.3 *Target Encoder*

Uso de *Target Encoder*, que consiste em transformar as variáveis Strings encontradas no método anterior em numéricas considerando a variável resposta como métrica, para o cálculo da proporção de casos positivos com cada elemento distinto presente na coluna da variável que apresenta texto.

4.3.4 Transformação em *Dummies*

Transformação em *Dummies*, é um método de manipulação voltado para melhorar as entradas durante o processo de modelagem, e consiste em transformar as informações presentes em cada coluna em múltiplas colunas com valor 0 caso determinada condição seja falsa e 1 caso verdadeira. Dessa forma, dada uma coluna com 3 elementos distribuídos, o algoritmo percorre essa coluna de forma sequencial para cada um deles, e no índice em que ele é correspondente com o elemento do atual percurso, marcamos ele como positivo do contrário recebe zero como valor, sendo assim uma estruturação binaria. Esse método pode ser usado como alternativa ao *Target Encoder*, pois amplia e separa informações que podem ser importantes, mas que estão juntas com dados que

reduzem seu valor quando observados de forma total.

4.3.5 Remoção de variáveis constantes

Com o uso dos métodos de separação de informações citados, pode-se ainda eliminar algumas das novas variáveis criadas, pois seu valor é mínimo para uso em modelo. Por exemplo, uma coluna com 5 informações distintas, que ao realizar-se o processamento verifica-se que uma ou duas dessas informações não tem elementos suficientes para aplicação ao modelo, ou apresentam muitos zeros, podem ser removidas da base de informações e assim refinar os dados. Outra opção para essas variáveis é mantê-las em uma etapa de modelagem e verificar se as mesmas apresentam alguma influência no modelo mesmo sem tantos dados positivos e assim decidir mantê-la ou não.

4.3.6 Balanceamento dos dados

O balanceamento dos dados, é uma técnica que consiste em avaliar se o número de casos que temos presentes como resposta positiva esperada, são em número suficiente, ou seja, queremos evitar que o modelo nos fale com alta precisão apenas os casos negativos e desconsidere a informação de interesse, assim a solução é tentar realizar uma re-amostragem com reposição para aumentar a evidência/presença dos casos que queremos prever. No contexto deste trabalho, os métodos usados de balanceamento foram *Upsampling* que consiste em equilibrar o número casos positivos da variável resposta a um número próximo do total de casos não severos, ou seja, equilibrar para uma proporção 1:1, e o método *Downsampling*, que consistem em equilibrar os dados fazendo o oposto do caso anterior, em que o número de casos não severos será reduzido a um número próximo ao de casos severos, também próximo da proporção de 1:1. Além

desse balanceamento, o conjunto de dados em todas as técnicas aplicadas sofreu um particionamento, gerando subgrupos exclusivos por meio de validação cruzada (*cross validation*), com a finalidade de observar a capacidade que o modelo possui em generalização, buscando analisar a precisão que o modelo apresenta, quando exposto a um novo conjunto de informações durante o treino e teste (KOHAVI, R. 1995).

4.3.7 Processamento de variáveis clínicas e laboratoriais

Para as variáveis relacionadas aos sintomas clínicos (Ex: Exantema, dor abdominal, febre , etc.) foram atribuídos valores binários (0 ou 1) que correspondiam a presença ou não do sintoma. Já para as variáveis relacionadas aos exames laboratoriais (suPAR, TGP, TGO, Plaquetas e Hematócritos), atribuímos duas analises diferentes, sendo uma que baseava-se na presença ou não de valores alterados nos exames (Classificados com base em valores de referencia de normalidade), e a outra que utilizava os dados brutos apresentados nos exames dos pacientes (Unidades laboratoriais de medida). O passo seguinte consistiu-se em separar os dados dos pacientes severos (42 casos) e não-severos (787 casos), realizando um cálculo de porcentagem em relação a presença e/ou alteração da variável dentro de cada grupo. Esses valores foram plotados em um gráfico de distribuição considerando período de avaliação (0-14 dias) e/ou tipo de variável (exame laboratorial ou sintomas), separados por severidade.

4.4 Medidas

A eficiência dos modelos propostos em predizer os caso severos e não severos de Dengue foi avaliada levando com consideração as seguintes métricas:

4.4.1 Matriz de Confusão

Consiste em uma tabela com o total de erros e acertos do modelo em comparação aos dados da variável alvo original e que é a base para as demais medidas que são observadas. Nessa tabela temos os seguintes cenários:

Verdadeiros Positivos: O que o modelo apresenta como 1, é de fato 1 quando observamos a variável alvo.

Falsos Negativos (Erro Tipo II): O modelo verificou elementos como negativos, ou seja 0, mas que na realidade são informações verdadeiras na variável alvo.

Falsos Positivos (Erro Tipo I): O modelo apresenta elementos como verdadeiros com marcação 1 mas que na realidade não constam desta forma na base origem.

Verdadeiros Negativos: O modelo acerta em falar que um elemento não é pertencente ao grupo de interesse quando comparado com a base origem.

4.4.2 Acurácia

Apresenta a proporção entre todas as classificações, quantas o modelo classificou corretamente como verdadeiro positivo ou verdadeiro negativo. Mas pode ser uma métrica enganosa, pelo motivo que o modelo pode estar analisando apenas um valor dentre os todos que podem apresentar para a resposta do modelo, logo deve ser analisada em conjunto a outras medidas para validar de forma mais precisa seu valor.

4.4.3 Precisão

Apresenta a proporção da classificação de elementos positivos que o modelo fez e que estão corretos, é uma métrica útil quando queremos verificar se os falsos positivos são considerados mais danosos que os falsos negativos.

4.4.4 Recall

Apresenta a proporção de todas as situações da classe de positivos com o seu valor esperado correto. É útil para observar os casos de falso negativo, de forma a evitar que, por exemplo, pacientes que estão doentes sejam agrupados na classe em que o indivíduo está saudável.

4.4.5 F1-Score

É a média harmônica entre precisão e recall, é uma forma de analisar somente uma métrica em vez de comparar individualmente a precisão e recall.

4.4.6 Curva ROC

A curva ROC (*Receiver Operating Characteristic*) é uma representação gráfica da sensitividade e sensibilidade do modelo , que nos permite verificar o quão bom o modelo é em distinguir duas classes. Ela tem como parâmetro a taxa de verdadeiros positivos, que é a razão entre o número de verdadeiros positivos dividido pela soma dos verdadeiros positivos e negativos e também apresenta a taxa falsos positivos que consiste na razão entre os falso positivos pela soma dos falsos positivos com verdadeiros negativos. Assim é possível traçar uma curva com "Taxa de verdadeiros positivos X Taxa de falsos positivos"em diferentes limiares de classificação (BARBOSA WANDERLEY et. al.,2010).

4.4.7 AUC

O AUC (*area under the curve*) consiste em uma métrica para o grau de separabilidade que o modelo apresenta, ou seja, classificar um grupo como portador de dada característica ou não. É uma forma de assumir a curva ROC como um único valor, que seria a área sob a curva, de modo que quanto maior seja a área da curva ROC melhor é a separabilidade que teremos no modelo (BARBOSA WANDERLEY et. al.,2010).

4.5 *Feature Selecion*

Após diversos processamentos para remoção e criação de novas informações é possível que ainda exista um grande número de variáveis não relacionadas à característica a ser predita. Logo, como alternativa para refinar ainda mais o conjunto de entrada, recorremos ao uso de métodos conhecidos como *Feature Selection*, que basicamente consistem em técnicas para observar a importância de dada variável dentro do modelo, por meio de processos de modelagem intermediários, gerando subconjuntos de variáveis. Assim aplica-se no conjunto de informações, uma técnica de modelagem mais simples, que no caso deste trabalho foi a técnica LASSO (*Least Absolute Selection and Shrinkage Operator*) com regressão logística, onde as variáveis que não contribuem para o resultado final recebem peso zero. Em resumo consiste basicamente na penalidade aplicada aos parâmetros do modelo, sendo assim permanecendo apenas os valores que apresentam diferenças para a técnica.

5 RESULTADOS

Antes de detalhar os resultados gerais dos modelos, apresentaremos um processo de intermediário de modelagem, com objetivo de identificar as as variáveis a serem de

fato usadas nos modelos, bem como a estrutura final do conjunto de dados.

5.1 Seleção de variáveis pelo método de LASSO

Após a aplicação das etapas iniciais de processamento das variáveis disponíveis, como inspeção manual da importância das variáveis para o problema propostos, saímos de um total inicial de 703 variáveis para um conjunto de 77 variáveis antes do processo de *Feature Selection*. A base de dados completa compreende 829 pacientes positivos para infecção por DENV, destes, 384 foram classificados quanto à presença de Dengue Severa e foram utilizados para a modelagem, dos quais, 42 tiveram classificação para severidade e 342 foram classificados como Dengue não severa. Apesar de se tratar de um conjunto de dados considerado pequeno quando comparado a grandes a bases de dados de larga escala de empresas de diferentes setores, é importante ressaltar que se trata de uma coorte relativamente grande de pacientes em diferentes locais do país e dessa forma torna-se uma base de dados de difícil obtenção. Isso evidencia a importância de se saber lidar adequadamente com a construção da base e da informação nela adicionada. Após a aplicação da abordagem de *Feature Selection* pelo método de LASSO, para identificação das variáveis mais importantes, obtemos um total de 31 variáveis úteis para o processo de modelagem. Entre as variáveis removidas nessa etapa, boa parte estava relacionada à informações verificadas na etapa de entrada do paciente como sexo, idade, etnia, existência de infecções prévias, como Zika, e informações relacionadas aos contatos 7 e 14 dias após a inclusão do paciente na coorte, como necessidade de hospitalização, resultados de exames sem peso para o modelo e sintomas como febre e dores em locais específicos. A fim de identificar se existia algum grau de correlação entre essas variáveis selecionadas pelo método de Lasso que poderia interferir nas etapas posteriores de modelagem calculamos a correlação par-a-par entre todas as 31 variáveis

disponíveis através do método de *Spearman*, sendo possível verificar que nenhuma das variáveis restantes estava altamente correlacionada às demais (Figura 4).

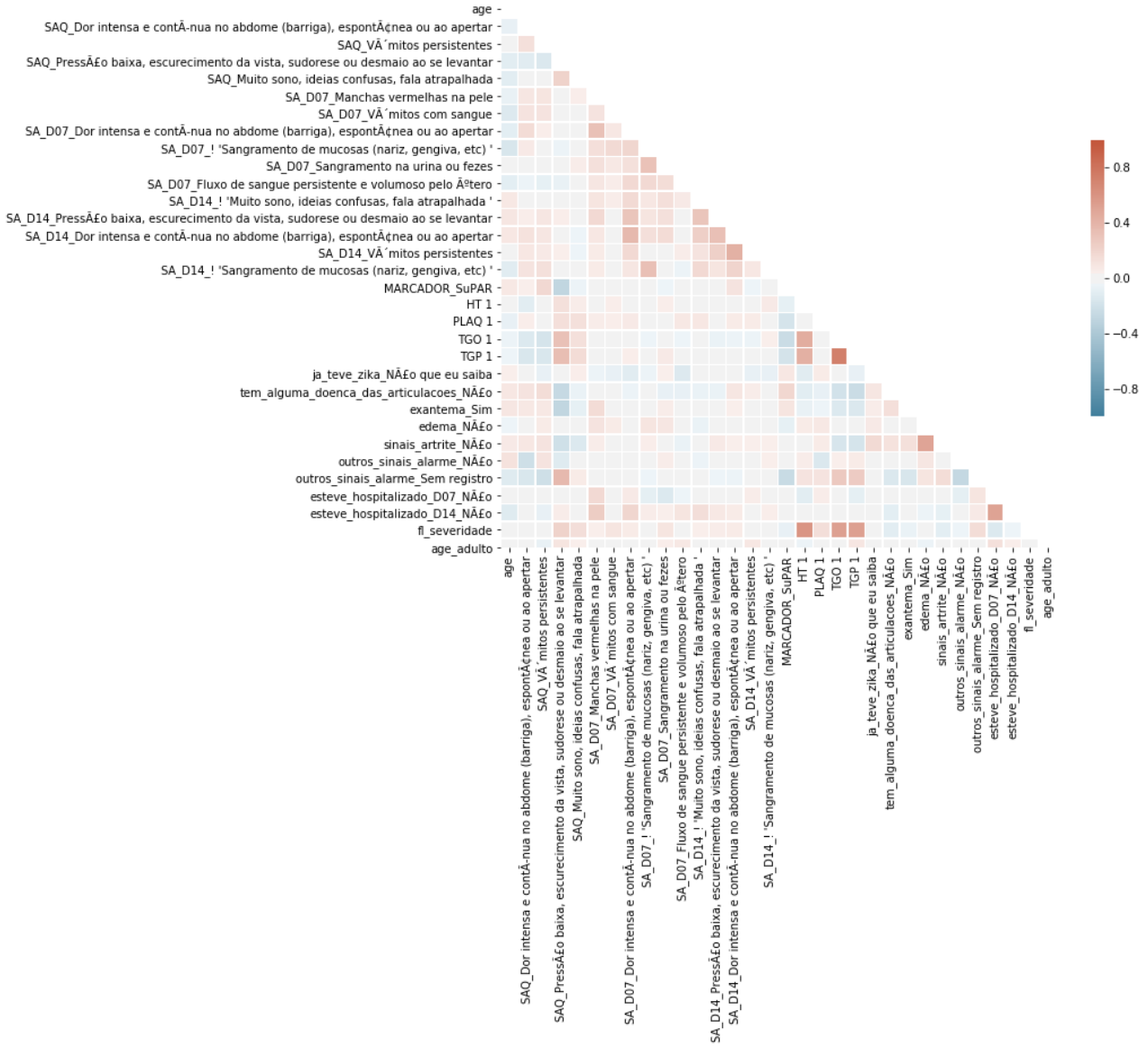


Figura 4: Matriz De Correlação Das Variáveis Selecionadas Pela Técnica De LASSO

5.2 Modelos com todas variáveis provenientes do método Lasso

Afim de identificar o modelo com maior capacidade preditiva dos casos severos e não-severos de Dengue os modelos de Regressão Logística, Árvores de Decisão, Random Forest e Naive Bayes foram aplicados no conjunto de dados de São José do Rio Preto e Palmas utilizando os métodos de *Upsampling* e *Downsampling* da classe a ser predita.

5.2.1 Modelos *Upsampling*

Após a aplicação do método de *Upsampling* o conjunto de dados utilizados para treino dos diferentes modelos foi o seguinte: 259 casos severos, 259 não-severos. É importante ressaltar que o conjunto de dados de teste foi mantido desbalanceado a fim de avaliar a eficiência dos modelos em um contexto mais próximo da realidade, onde a maioria dos pacientes não evoluiu para quadros mais severos de Dengue. O resultado da aplicação dos quatro modelos no conjunto de treino e teste é apresentado no formato de matriz de confusão na Figura 5, onde podemos identificar o número de casos preditos corretamente bem como o número de falsos positivos e negativos obtidos em cada modelo. Dentre os modelos testados, podemos perceber que o modelo Naive Bayes, apresenta os piores resultados , com taxa alta de casos falsos negativos e também falsos positivos (Figura 5). Isso fica evidente também quando comparamos as métricas de maior detalhe, como ROC-AUC (Tabela 1). O modelo Naive Bayes apresentou valores de AUC de 0,65 e 0,62 nos conjuntos de treino e teste respectivamente, bem abaixo dos valores de AUC dos demais modelos, que ficaram acima de 0,95 em todos os cenários . A baixa eficiência do modelo Naive Bayes provavelmente deve as informações de insumo que não satisfazem totalmente os critérios que o modelo necessita, como a total independência das variáveis, logo podemos para esse caso desconsiderar essa técnica devido a ser pouco assertiva.

Para os modelos restantes, observa-se pelas figuras 5 e 6, uma boa separação na etapa de treino, sendo que os modelos de Regressão Logística e Árvore de Decisão são capazes de predizer corretamente todos os indivíduos nesse conjunto. Já na etapa de testes o modelo de Regressão Logística apresentou os melhores resultados, com menores quantidade de falsos positivos e negativos e mair valor geral de AUC (0,97), seguido do modelo de Árvores de Decisão (0,96) e *Random Forest* (0,95).

Com isso, podemos perceber que, com exceção do modelo Naive Bayes, todos os demais modelos tiveram resultados igualmente satisfatórios tanto no conjunto de treino quanto de teste.

Medidas caso *Upsampling*

REGRESSÃO LOG. TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	259	0
SEVERO BASE ORIGEM	0	259

REGRESSÃO LOG. TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	100	8
SEVERO BASE ORIGEM	5	10

NAIVE BAYES TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	226	33
SEVERO BASE ORIGEM	181	78

NAIVE BAYES TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	95	13
SEVERO BASE ORIGEM	12	3

ARVORE DE DECISÃO TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	259	0
SEVERO BASE ORIGEM	0	259

ARVORE DE DECISÃO TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	107	1
SEVERO BASE ORIGEM	4	11

RANDOM FOREST TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	256	3
SEVERO BASE ORIGEM	0	259

RANDOM FOREST TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	103	5
SEVERO BASE ORIGEM	4	11

Figura 5: Matriz De Confusão De Cada Modelo Para Balanceamento *Upsampling*

Tabela 1: Tabela Valores ROC-AUC Modelos *Upsampling*

Técnica	Regressão Logistica	Naive Bayes	Árvore de Decisão	Random Forest
ROC-AUC Treino	1	0.65	0.99	0.98
ROC-AUC Teste	0.97	0.62	0.96	0.95

5.2.2 Modelos *Downsampling*

Para obter um número balanceado de ocorrências positivas e negativas na variável resposta, nesse caso pacientes com e sem severidade para Dengue , testamos também a aplicação do método de *Downsampling*, onde eventos da categoria predominante são removidos randomicamente até atingir a proporção 1:1 entre as categorias. A aplicação desse método no conjunto de treino resultou em um grupo de 27 pacientes severos e 27 não-severos. Testando o poder preditivo dos quatro modelos apresentados e comparando novamente os resultados da predição dispostos na forma de matriz de confusão notamos que, assim como na abordagem de *Upsampling*, o modelo baseado em Naive Bayes apresenta uma taxa de erro elevada tanto no teste como no treino (Figura 6), apresentando-se novamente como a técnica menos eficiente para a modelagem no problema proposto. De forma similar, os valores de AUC para o método Naive Bayes (0,66 e 0,6 treino/teste) se revelaram inferiores aos demais métodos testados. Comparando os resultados de AUC no conjunto de treino podemos observar que o método de Regressão Logística (AUC = 1) apresentou novamente resultados superiores aos modelos de Árvore de Decisão (AUC = 0.94) e *Random Forest* (AUC = 0.95), contudo esses dois últimos tiveram melhores desempenhos quando avaliados na condição de teste, com valor de AUC de 0,9 para o modelo de Árvores de Decisão e 0,93 para *Random Forest*, enquanto o modelo de Regressão Logística apresentou valor de AUC de 0,88 (Tabela 2).

Medidas caso *Downsampling*

REGRESSÃO LOG. TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	27	0
SEVERO BASE ORIGEM	0	27

REGRESSÃO LOG. TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	76	32
SEVERO BASE ORIGEM	2	13

NAIVE BAYES TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	26	1
SEVERO BASE ORIGEM	20	7

NAIVE BAYES TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	95	13
SEVERO BASE ORIGEM	12	3

ARVORE DE DECISÃO TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	27	0
SEVERO BASE ORIGEM	0	27

ARVORE DE DECISÃO TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	98	10
SEVERO BASE ORIGEM	2	13

RANDOM FOREST TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	25	2
SEVERO BASE ORIGEM	0	27

RANDOM FOREST TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	95	13
SEVERO BASE ORIGEM	0	15

Figura 6: Matriz De Confusão De Cada Modelo Para Balanceamento *Downsampling*

Tabela 2: Valores ROC-AUC Modelos *Downsampling*

Técnica	Regressão Logistica	Naive Bayes	Árvore de Decisão	Random Forest
ROC-AUC Treino	1	0.66	0.94	0.95
ROC-AUC Teste	0.88	0.60	0.90	0.93

5.2.3 Modelagem com variáveis laboratoriais

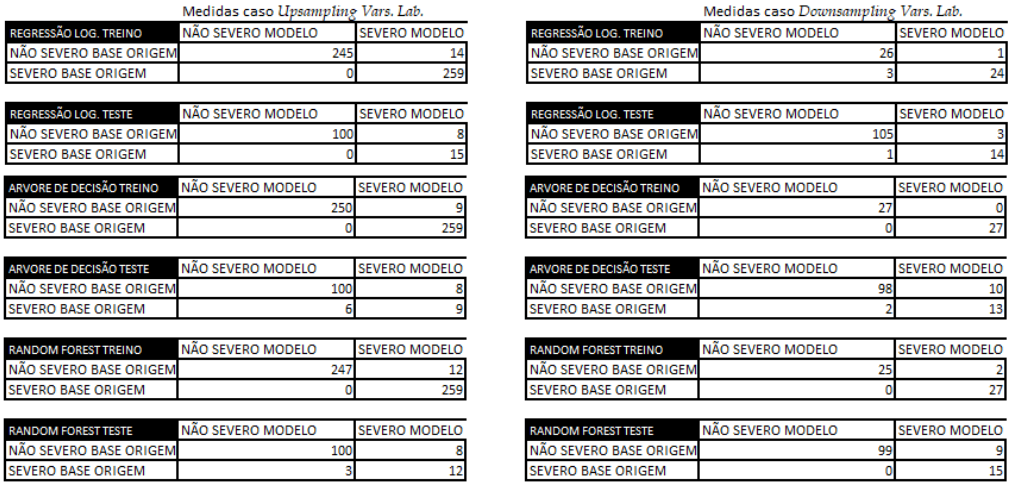
Como pode ser observado anteriormente, as técnicas de Regressão Logística, Árvore de Decisão e *Random Forest*, apresentam uma classificação mais consistente para o problema de severidade, com melhores valores nas métricas de predição. Para identificar o melhor modelo disponível testamos também os modelos apenas com uso de variáveis laboratoriais. Isso se justifica pela maior facilidade de acesso a essas informações, uma

vez que informações de exames são geralmente coletadas nos procedimentos de rotina hospitalar, enquanto que as demais informações, como dados clínicos 7 e 14 dias após inclusão na coorte são de mais difícil obtenção e podem apresentar falhas devido a dificuldade de acompanhamento dos pacientes nas etapas subsequentes.

Utilizando apenas as variáveis laboratoriais podemos observar que os modelos apresentam um desempenho um pouco abaixo dos citados anteriormente, mas que ainda classificam bem e podem ser aplicados ao problema sem perda significativa de desempenho (Figura 7). É possível observar também que a técnica com Regressão Logística, apresenta resultados de AUC ligeiramente superiores aos demais modelos tanto na condição de treino quanto na condição de teste em ambas abordagens de *Upsampling* e *Downsampling* da variável preditora (Tabela 3 e 4). Dessa forma podemos perceber que somente o uso das variáveis laboratoriais já pode ser suficiente para construção de modelos preditivos acerca da severidade de Dengue.

Além disso, levando em consideração os resultados apresentado nos tópicos anteriores, os modelos apresentados usando a Regressão Logística e *Random Forest* se apresentavam como mais assertivos Contudo, analisando um cenário de variáveis mais restritas (Laboratoriais), notamos que a técnica *Random Forest* apresentou resultados inferiores à Regressão Logística. Outro ponto a ser considerado é que embora as técnicas de balanceamento com *Upsampling* e *Downsampling* da variável resposta não apresentaram resultados distintos no cenário de modelagem apenas com as variáveis laboratoriais (Tabela 3 e 4), a técnica de *Upsampling* se mostrou mais eficiente no cenários onde todas as variáveis foram consideradas (Tabela 1 e 2). Dessa forma, tendo em vista os resultados superiores da técnica de Regressão Logística e sua facilidade de implementação e interpretabilidade, ela foi a técnica de escolha para aplicação nas etapas posteriores do trabalho em conjunto com a abordagem de *Upsampling* da variável

resposta.



Medidas caso *Upsampling Vars. Lab.*

REGRESSÃO LOG. TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	245	14
SEVERO BASE ORIGEM	0	259

REGRESSÃO LOG. TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	100	8
SEVERO BASE ORIGEM	0	15

ARVORE DE DECISÃO TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	250	9
SEVERO BASE ORIGEM	0	259

ARVORE DE DECISÃO TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	100	8
SEVERO BASE ORIGEM	6	9

RANDOM FOREST TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	247	12
SEVERO BASE ORIGEM	0	259

RANDOM FOREST TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	100	8
SEVERO BASE ORIGEM	3	12

Medidas caso *Downsampling Vars. Lab.*

REGRESSÃO LOG. TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	26	1
SEVERO BASE ORIGEM	3	24

REGRESSÃO LOG. TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	105	3
SEVERO BASE ORIGEM	1	14

ARVORE DE DECISÃO TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	27	0
SEVERO BASE ORIGEM	0	27

ARVORE DE DECISÃO TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	98	10
SEVERO BASE ORIGEM	2	13

RANDOM FOREST TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	25	2
SEVERO BASE ORIGEM	0	27

RANDOM FOREST TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	99	9
SEVERO BASE ORIGEM	0	15

Figura 7: Matriz De Confusão Com Apenas Variáveis laboratoriais

Tabela 3: Valores ROC-AUC modelos *Upsampling*, uso de variáveis laboratoriais

Técnica	Regressão Logística	Árvore de Decisão	Random Forest
ROC-AUC Treino	0.99	0.96	0.94
ROC-AUC Teste	0.98	0.92	0.90

Tabela 4: Valores ROC-AUC modelos *Downsampling*, uso de variáveis laboratoriais

Técnica	Regressão Logística	Árvore de Decisão	Random Forest
ROC-AUC Treino	0.98	0.94	0.95
ROC-AUC Teste	0.99	0.90	0.93

5.2.4 Verificação da Importância das Variáveis

A fim de demonstrar que as variáveis selecionadas pelo modelo de Regressão Logística apresentam diferenças de distribuição entre os grupos de pacientes com Dengue Severa ou não, resolvemos avaliar a ocorrência de caso positivo dentro de cada variável separadas pela severidade da doença. Isso permite verificar se o conjunto de variáveis apresenta alguma informação que poderá ser usada pelo grupo de estudo, em complemento ao que foi apresentado pelo modelo.

Na Figura 8 podemos observar a porcentagem de casos que apresentaram positividade de sintomas dentro das variáveis selecionadas, e que foram separadas por grupo de pacientes com dengue severa ou não, durante os primeiros 14 dias de acompanhamento. Houve uma tendência clara de aumento de positividade de sintomas no grupo de Dengue Severa dentro do período avaliado. Em seguida, também avaliamos separadamente as variáveis relacionadas a alterações (as alterações correspondem a indivíduos com valores fora os intervalos de medidas padrões, tanto para mais ou para menos) nas medidas laboratoriais de suPAR (receptor do ativador de plasminogênio tipo uroquinase solúvel), de TGP (transaminase glutâmico pirúvica) e TGO (transaminase glutâmico-oxalacética), plaquetas e hematócritos), bem como as medidas relacionadas a hospitalização do paciente (após 7 ou 14 dias) ou ainda sobre a presença de sintomas graves prévios em conjunto com a doença (Sinais de alarme, edema, Exantema, etc.) (Figura 9). Essas variáveis são muito importantes porque dizem respeito ao estado geral de saúde do paciente e influenciam diretamente na progressão da Dengue. suPAR, por exemplo, é considerado um marcador de ativação do sistema imune, cujo nível se encontra elevado (acima 3 ng/ml) em pacientes críticos apresentando infecções virais ou bacterianas (BACKES et. al.,2012). Este marcador está positivamente correlacionado à severidade de diversas doenças , sendo utilizado em diferentes sistemas de

classificação patológica e também como marcador de mal funcionamento de diferentes órgãos (BACKES et. al.,2012). Seguindo nesta mesma linha, os marcadores TGO e TGP , são moléculas que normalmente se elevam no sangue de indivíduos doentes, pois geralmente, essas moléculas se encontram expressas e contidas dentro das células (principalmente células do fígado, rins, coração, cérebro e músculo) e são liberadas em casos específicos de lesão nesses órgãos. Assim, durante as infecções virais ocorre a lise de células nos órgãos afetados, aumentando os níveis dessas enzimas no sangue (SILVA 2008). Segundo Andriolo et.al., são considerados valores de referência normais para essas enzimas, níveis entre 5 U/L e 40 U/L para TGO e entre 7 U/L e 56 U/L para TGP, sendo considerado valores alterados qualquer medida fora desses intervalos. Finalmente, foram considerados também importantes as variáveis que mensuram a quantidade de plaquetas e o nível de hematócritos no sangue. As plaquetas são consideradas elementos do sangue que controlam a coagulação, sendo importante em processos de manutenção da fisiologia dos vasos (PONE 2016). Já o hematócrito está relacionado a porcentagem de hemácias presentes no sangue, sendo encontrado alterado em estados de desidratação ou anemias (PONE 2016). São considerados valores normais considerados entre 150.000-450.000/mm para plaquetas, e valores entre 35%-52% para hematócritos (Andriolo et.al.). Sabemos que é bem conhecido o efeito citopatológico do vírus da Dengue sobre a diminuição dos níveis de plaquetas, podendo ocasionar hemorragias e/ou alterar o volume de líquido sanguíneo, devido a também induzir um quadro crônico de febre, levando a desidratação e consequentemente alterando os níveis de hematócritos do sangue (PONE 2016). Com base nas variáveis selecionadas pelo modelo, separando os pacientes por severidade, podemos observar que os valores de média das variáveis (a média considerada é em relação ao grupo analisado) suPAR, TGP, TGO, Plaquetas e Hematócritos, se encontram elevados no grupo severo em relação ao

não-severo (Figura 10). Assim podemos concluir que este conjunto de variáveis selecionadas pelo modelo apresentam-se mais alteradas para o grupo severo, constituindo-se de um grupo dados que apresentam uma tendência mais clara de separação dos grupos.

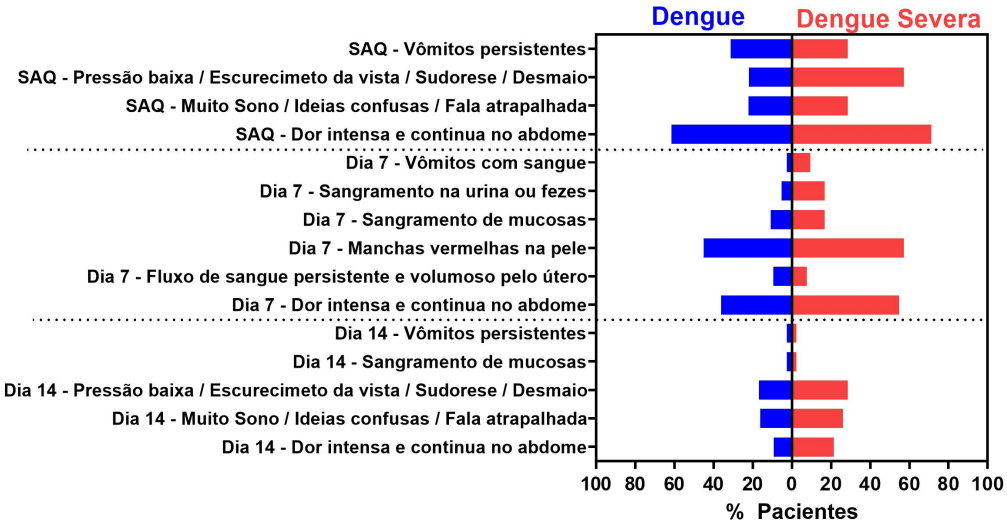


Figura 8: Percentual De Pacientes Em Relação à Severidade Por Período

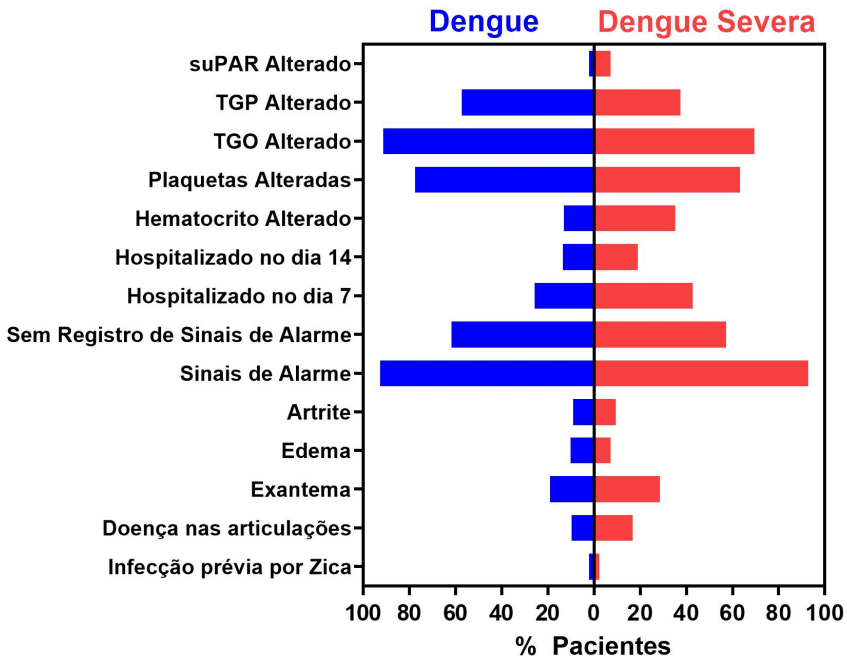


Figura 9: Percentual De Pacientes Por Severidade: Variáveis Clínicas

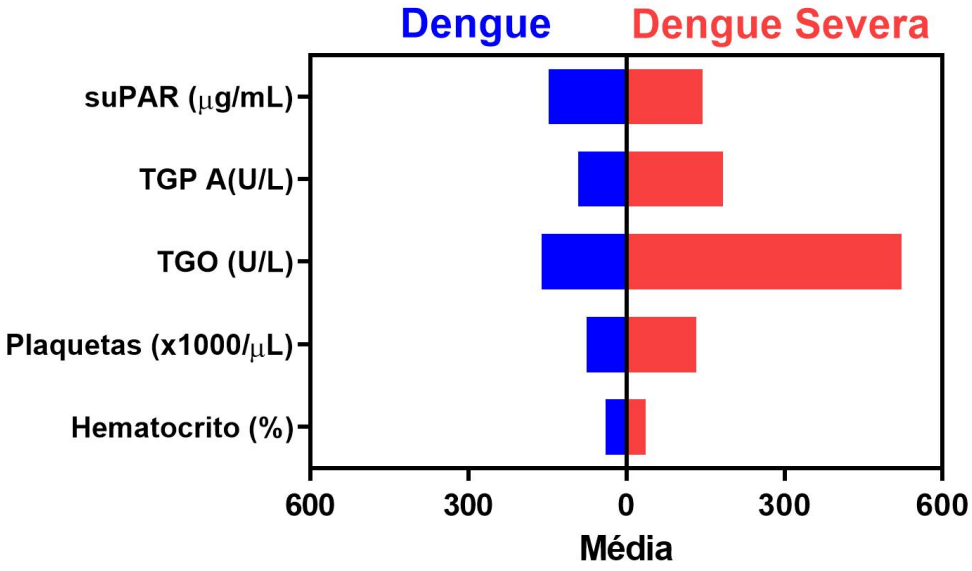


Figura 10: Média Das Variáveis Clínicas Por Severidade

5.2.5 Modelagem dos dados provenientes de diferentes centros hospitalares

Para avaliar a concordância entre a classificação dos casos severos e não severos de Dengue provenientes de diferentes centros hospitalares optamos testar o modelo de Regressão Logística com balanceamento da variável preditora por *Upsampling*, que se revelou a mais eficiente nas etapas anteriores, utilizando os dados de São José de Rio Preto (SJRP) como treino e os dados de Palmas como conjunto de teste, e também o oposto. Sendo que SJRP possui 295 casos confirmados , sendo 15 severos, 277 não-severos e 3 sem classificação, enquanto Palmas apresenta 114 casos confirmados, dos quais 27 severos, 65 não-severos e 22 sem classificação. O cenário utilizando os dados de SJRP como treino e os dados de Palmas como conjunto de teste apresentou valores de AUC no treino e teste de respectivamente 1 e 0,90 (Figura 11), enquanto o cenário desenvolvido com os dados de Palmas no conjunto de treino e testados nos dados de SJRP (Figura 12), apresentou melhores resultados, com significante número de acertos ao comparar com sua contraparte, como pela medida da ROC-AUC onde o novo cenário apresentou etapa de treino com valor 1, e na validação/teste seu valor chegou a 0,98. Em comparação com os resultados obtidos nas primeiras etapas de modelagem, onde os pacientes classificados de São José do Rio Preto e Palmas eram agrupados nos conjuntos de treino e teste (Figuras 5, 6 e 7; Tabelas 1, 2, 3 e 4), os cenário utilizando os dados de SJRP como treino e Palmas como teste de forma separada apresentaram resultados ligeiramente inferiores na etapa de teste (Figura 11). Já o cenário oposto, onde os dados de pacientes classificados de Palmas foram utilizados no conjunto de treino e o modelo foi testado nos dados os pacientes de SJRP apresentou resultados melhores e equiparáveis aos modelos com dados tratados em conjunto. Dessa forma é possível observar que apesar de contarmos com informações coletadas em diferentes centros hospitalares os modelos gerados a partir desses dados são concordantes entre si

e se assemelham aos modelos gerais onde todos os dados são processados em conjunto.

REGRESSÃO LOG. TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	280	0
SEVERO BASE ORIGEM	0	280

REGRESSÃO LOG. TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	70	17
SEVERO BASE ORIGEM	7	20

Figura 11: Matriz De Confusão utilizando Regressão logística com *Upsampling* e dados de SJRP como treino e dados da cidade de Palmas como teste A

Tabela 5: Valores ROC-AUC Treino SJRP - Teste Palmas

Técnica	Regressão Logística
ROC-AUC Treino	1
ROC-AUC Teste	0.90

REGRESSÃO LOG. TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	86	1
SEVERO BASE ORIGEM	0	87

REGRESSÃO LOG. TESTE	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	279	1
SEVERO BASE ORIGEM	3	12

Figura 12: Matriz De Confusão utilizando Regressão logística com *Upsampling* e dados de Palmas como treino e dados da cidade de SJRP como teste B

Tabela 6: Valores ROC-AUC Treino Palmas - Teste SJRP

Técnica	Regressão Logística
ROC-AUC Treino	1
ROC-AUC Teste	0.98

6 CLASSIFICAÇÃO DE NOVOS PACIENTES

Uma vez definido a técnica de Regressão Logística como o modelo que melhor se ajusta a nossa base de dados classificados e verificado que os dados provenientes dos centros hospitalares de São José do Rio Preto e Palmas são concordantes entre si, o próximo passo consiste na aplicação desse modelo na base total de dados, possibilitando assim observar qual seria o conjunto de pacientes com maior propensão de evolução para Dengue Severa entre aqueles ainda não classificados. Nesse cenário, os pacientes classificados de São José do Rio Preto e Palmas foram agrupados como conjunto de treino, totalizando 409 indivíduos, dos quais 367 não apresentaram Dengue Severa e 42 apresentaram. O modelo treinado com essa base foi testado no conjunto de pacientes provenientes de Araraquara, Campo Grande, Nova Serrana e Arcos, totalizando 420 pacientes.

Logo observando a Figura 13, temos que a opção escolhida apresenta uma boa classificação na etapa de treino com a união dos conjuntos de São José do Rio Preto e Palmas, sendo bem assertiva. Para o conjunto total (base com todos os 829 casos), o *template* desenvolvido gera um arquivo .xlsx com todos os pacientes com seu ID de identificação e apresenta 3 colunas adicionais com a marcação original de severidade, a marcação de severidade proposta pelo modelo, onde verificamos quem seriam os demais casos a serem considerados suspeitos de severidade, bem como uma coluna com a probabilidade do paciente ser ou não severo.

Por ser um modelo de regressão logística e observando o arquivo gerado, temos que valores abaixo de 0.5 de probabilidade são os casos que se confirmam como severos, pois coincidem com os casos severos conhecidos (42 pacientes), logo valores dentro do intervalo [0,0.5] são altamente propensos a serem severos, enquanto que fora desse intervalo os casos não devem ser tão propensos, mas que a depender da proximidade com o intervalo, o paciente pode ser um indivíduo que apresenta altas chances de também ser severo em alguma futura infecção, mas que deve ser observado com cuidado. Entretanto, ao analisar a base com aplicação do modelo, notamos que o grupo severo não é expandido, ou seja, o modelo conseguiu classificar os pacientes com quase 100% de acerto, mas sendo os já conhecidos como severos. E dentro do intervalo de probabilidade citado não foi capaz de obter mais casos além desses (vale observar na figura 14 que o modelo coloca apenas 2 casos adicionais como suspeitos de severidade).

Com isso, verificamos que o conjunto pode não apresentar outros pacientes propensos a evolução para Dengue Severa. Para realizarmos uma verificação adicional realizamos a criação de um conjunto "fake"de pacientes (21 casos), tomando como ponto inicial as medidas das variáveis analisadas na seção 5.2.4. Para isso, os valores das variáveis para esses 21 pacientes foram populados com valores aleatórios, mas de forma a não extrapolarem de forma muito elevada (variações acima ou abaixo de 10% aproximadamente) os valores em cada linha . Com isso o modelo desenvolvido foi aplicado nesse conjunto e se observou a severidade sugerida pelo modelo e a probabilidade e notamos que o modelo consegue sugerir que aproximadamente 47% desses casos são altamente severos (10 casos com valores dentro do intervalo [0-0.5]) e que aproximadamente 5% dos casos (1 casos com valor próximo do intervalo) é apontado como suspeito. Sendo assim, é provável que o modelo esteja apresentando a classificação correta para os pacientes ainda não classificados dos demais centros hospitalares,

mas as informações restantes desses pacientes não os apontam como casos altamente suspeito ou suspeito ao desenvolvimento de Dengue Severa.

REGRESSÃO LOG. TREINO	NÃO SEVERO MODELO	SEVERO MODELO
NÃO SEVERO BASE ORIGEM	364	3
SEVERO BASE ORIGEM	0	367

Figura 13: Matriz Correlação Para Base Total Etapa De Treino

resultado modelo	fl_severidade	probabilidade
1	1	0,001813373
1	1	0,001991016
1	1	0,002119758
1	1	0,002393506
1	1	0,002967236
1	1	0,003459934
1	1	0,003841256
1	1	0,004900702
1	1	0,005965371
1	1	0,006236644
1	1	0,006702483
1	1	0,009321743
1	1	0,00021259323778233252
1	1	0,00029721540811078473
1	1	0,0003635636947173504
1	1	0,001156979282411208
1	1	0,001370346799860478
1	1	0,0018314898216514885
1	1	0,002402286596959624
1	1	0,0038177224169820834
1	1	0,005757650874814613
1	1	0,007561213935784905

resultado modelo	fl_severidade	probabilidade
1	1	0,018161354903201143
1	1	0,056862541069786054
1	1	0,07845795008854328
1	1	0,08412203502835125
1	1	0,12449365629233233
1	1	0,16324078840807477
1	1	0,16374124586778394
1	1	0,18270063751342913
1	1	0,20127530992501597
1	1	0,20386664210802707
1	1	0,20489606226655876
1	1	0,22898794393447452
1	1	0,2368011741640632
1	1	0,2854943837029006
1	1	0,33785090112282035
1	1	0,3703029343546169
1	1	0,3993553083900909
1	1	0,44725746600523386
0	1	0,5057029617826909
0	0	0,6687024948526845
0	0	0,7582140696376374
0	1	0,7638415694894684

Figura 14: Comparação Severidade Modelo e Severidade Real Com Probabilidade

7 DISCUSSÃO

Visando compreender a evolução da severidade da dengue, por meio de informações dispostas em conjuntos de dados, foi proposto neste trabalho a utilização de técnicas de aprendizado de máquina, que são capazes de manipular grande quantidade de dados para a aplicação de diferentes métodos e tratamentos de informações, com a finalidade de gerar modelos capazes de classificar adequadamente casos severos e não severos de Dengue entre pacientes participantes do projeto ARBOBIOS etapa inicial necessária consistiu em verificar quais foram os métodos mais adequados para lidar com o problema proposto, tomando como referência a facilidade de construção e balanceamento ou ainda a forma de obtenção de resultados/medidas. Foram considerados 4 métodos (Regressão Logística, Naive Bayes, Árvore de decisão e *Random Forest*), que foram associados a um método intermediário de modelagem (LASSO com Regressão Logística para seleção de variáveis com maior relevância) para analisar o conjunto de informações que discriminasse os indivíduos acometidos por Dengue com maior propensão de desenvolvimento do quadro severo da doença. Para isso, o conjunto de dados clínicos dos pacientes foi pré-processado de maneira a permitir que o maior número de variáveis relevantes fossem mantidas após a etapa de pré-processamento das variáveis e seleção das variáveis com maior contribuição para a classificação final através do método de *Feature Selection* com LASSO, restaram 31 variáveis que foram empregadas nas etapas posteriores de modelagem. Uma vez definido o conjunto de variáveis a ser utilizado, as técnicas utilizadas passaram por diversas etapas de treinos e testes em diversos cenários com o intuito de medir o desempenho geral dos mesmos. Inicialmente, considerando todos os métodos, observou-se que 75% dos modelos propostos apresentaram-se como classificadores eficientes, com boa separação de casos severos e não-severos e medidas de valores como ROC-AUC consistentes em nível bom (0.8-0.9) e excelente(1-0.9)

(Figuras 5,6 e Tabelas 1,2). Dentro das análises, a única técnica com resultado não satisfatório foi a de Naive Bayes, com desempenho bem abaixo dos demais modelos. Outros autores, que utilizaram técnicas de análise semelhantes, também foram capazes de obter métodos de classificação para observar a ocorrência de Dengue. Fathima & Hundewale (2011) usaram SVM (*Support Vector Machine*) + *Random Forest* contra Naive Bayes e notaram que a combinação de técnicas foi mais eficiente que o uso de uma técnica única para observar a precisão do diagnóstico de um caso de Dengue. Seguindo a mesma linha de análise, neste trabalho observamos que mesmo com a combinação de técnicas, o método baseado em Naive Bayes não expressou maior precisão. Isto pode ter ocorrido por que a técnica de Naive Bayes é considerada muito específica, relacionada aos conceitos do teorema de Bayes, que não se aplicam ao problema de classificação proposto aqui.

Com isso notamos que ao longo do processo de modelagem, alguns pontos se tornam mais expressivos no momento de manter ou não uma técnica. Em nosso caso o número de falsos negativos e falsos positivos elevados foi considerado como ponto chave para descartar uma técnica. Para a escolha do modelo final, as 3 técnicas restantes foram expostas ao cenário com insumo apenas de dados de exames laboratoriais (suPAR, TGP, TGO, plaquetas e hematócritos). O resultado do comparativo entre os modelo mostraram-se com leves diferenças (Figura 7 e Tabelas 3 e 4). Observando a separação de pacientes e as medidas de ROC-AUC para um balanceamento positivo (uso de reamostragem para aumentar a expressão da classe severa), o modelo de Regressão Logística apresentou-se com valores acima dos demais métodos e foi considerado como o método definitivo para as análises do conjunto de dados. De modo semelhante, Duarte França (2006), também utilizando o método de Regressão Logística, foram capazes de separar casos de suspeita de dengue de casos graves utilizando informações clínicas pro-

venientes de diferentes bases de dados como prontuários médicos revisados e validados dos pacientes internados e registrados na rede hospitalar do sistema público de saúde de MG e aplicando esses dados ao banco de dados de notificação compulsória do SUS - SINAM. Nesse trabalho, verificou-se também que os casos de dengue registrados no sistema de notificação foram aqueles de evolução mais grave e que não representaram a totalidade de casos internados no sistema público de saúde (identificados pela modelagem de dados) e que acabou superestimando a taxa de letalidade da doença no banco de dados do SUS.

Tendo em vista o modelo escolhido, também avaliamos o seu comportamento quando exposto a dados diversos. Em nosso caso escolhemos a localidade para avaliar a influencia sobre a maneira em que o modelo observa o conjunto de informações de diferentes regiões. Inicialmente os dados para treino e teste foram a união dos conjuntos de dados originados das cidades de Palmas - TO e São José do Rio Preto-SP (SJRP), e posteriormente, treinamos o modelo com os dados de SJRP e aplicamos no conjunto de Palmas. Em seguida verificamos as medidas e separação e comparamos com as medidas do caso inverso de treino e teste. Notamos que de certa forma existe consistência de informações por região, pois percebemos que a localidade de SJRP, gerou um modelo capaz de separar os dados de Palmas com taxa relativamente alta de acertos. Já verificando a situação inversa, o conjunto de Palmas resulta num modelo capaz de separar os dados de SJRP, com casos de falso negativo e falso positivo em menor quantidade, sendo um modelo mais assertivo, expondo que algumas regiões podem proporcionar melhores informações para a modelagem de dados do que outras. Isso reforça a necessidade de um controle maior na coleta de dados para que estes sejam os mais íntegros possíveis. Foi notado neste trabalho desde o início que o conjunto bruto de dados (sem tratamento), apresentava um grande número de variáveis sem preenchimento (dados em branco so-

bre informações de exames laboratoriais e sintomas), o que mostra a necessidade de uma nova formulação do sistema de captação de dados, além de melhor treinamento dos profissionais da saúde, para que sejam capazes de manter uma padronização de insumo, ou seja, capacitação para o preenchimento adequado da base de dados junto ao grupo médico.

Finalmente, o modelo proposto neste trabalho foi validado em diversos cenários e apresenta boas métricas nos mais diversos casos, contudo, o passo principal foi observar se o modelo escolhido seria capaz de identificar os casos de severidade da doença em uma base de dados mais ampla ligada ao projeto ARBOBIOS. Ao final da análise foi gerado um arquivo com a adição de 3 novas colunas (resultado para severo (0 não-severo, 1 severo), resultado original para severidade sem o modelo e a probabilidade que cada paciente apresenta para ser severo ou não (entre 0-0.5 severo; próximo desse intervalo suspeito a severo). Após aplicação desta etapa notamos que o modelo obteve boas medidas de separação e eficiência, porém identificou apenas um caso adicional aos 42 casos severos já conhecidos. Isso nos levou a considerar duas situações: (1) o conjunto de dados total não apresenta valores em nível suficiente para classificar algum indivíduo como severo, ou (2) o conjunto não apresenta mais nenhum indivíduo severo considerando as variáveis apresentadas. Para testar essa última hipótese, rodamos um novo teste onde o modelo foi aplicado à um conjunto de dados aleatórios, onde foi possível observar que ele foi capaz de atribuir probabilidades mais distribuídas para os casos "fakes"e classifica alguns casos como severos, sendo assim o modelo se torna mais uma vez validado e fortalece a hipótese de que pode não existir mais pacientes severos na base dados original.

Ao final verificamos as medidas de percentual e média de ocorrência de casos positivos de sintomas ou exame laboratoriais em cada um dos grupos de pacientes (casos

severos e não-severos) e observamos que, de forma geral, o conjunto de variáveis selecionadas pelo modelo apresenta medidas mais elevadas nos casos severos do que nos não-severos (Figuras 8 e 9). De modo semelhante, um estudo realizado por Gerusa 2012, analisando dados clínicos para sinais de Dengue grave em crianças e adolescentes, observou que um conjunto de variáveis bem semelhante ao usado neste trabalho (como marcadores de tempo de hospitalização, febre, dor abdominal, sangramentos, sonolência / irritabilidade e dificuldade respiratória, entre outros), mostrou-se como um modelo que foi capaz de classificar casos graves da doença.

8 CONCLUSÕES

Com base no apresentado, podemos afirmar que os modelos testados apresentaram bons resultados, e a técnica escolhida como principal, Regressão Logística, foi adequada, pois foi capaz de classificar e definir uma probabilidade para cada paciente em relação a sua propensão em evoluir para quadro de Dengue Severa ou não. Como a classificação da base total não expandiu os casos muito além do grupo de severos conhecidos, apresentamos como opção, o uso das variáveis analisadas em mais detalhes na sessão 5.2.4 em conjunto com o modelo. Dessa forma, observando-se pacientes com alterações em tais variáveis, seria possível para o grupo que analisa os biomarcadores, assumir esta opção como o direcionamento mais adequado para identificar e ampliar o grupo de casos suspeitos de severidade da Dengue facilitando assim a medida e dosagem dos biomarcadores que ainda precisam ser analisados.

Outro ponto a sere considerado como sugestão inclui uma reestruturação da base bruta, ou criação de uma nova base de armazena- mento que siga as variáveis dadas como mais relevantes, ou ainda restringir o conceito de severidade para uma quantidade de informações minímas, de modo a garantir o mínimo de informações de entrada na

nova base, evitando perda de dados importantes. Além disso, essa informações podem orientar os grupos de hospitais parceiros, através do envio de um formulário com os campos e valores a serem preenchidos, além de um guia de capacitação de usuários sobre a alimentação da base, onde no momento do preenchimento das informações dos pacientes, seja possível manter o mais alto grau de qualidade dos insumos.

Por fim, o projeto ARBOBIOS encontra-se na fase de inclusão de pacientes e novos dados poderão ser disponibilizados futuramente. Dessa forma, os *templates* criados para classificação dos pacientes poderão ser empregados no futuro para reavaliação da classificação realizada e possível identificação de novos casos com propensão ao desenvolvimento de Dengue Severa.

Referências

[1] ANDRIOLO, Adagmar et al.,*Intervalos de referência no laboratório clínico; Bras Patol Med Lab • v. 44 • n. 1 • p. 11-16 • fevereiro 2008*

[2] ALPAYDIN, Ethem. , *Introduction to Machine Learning. 3. ed. rev. e atual. [S. l.]: MIT Press, 2014. 613 p.*

[3] BACKES, Yara et al.,*"Usefulness of suPAR as a biological marker in patients with systemic inflammation or infection: a systematic review." Intensive care medicine vol. 38,9 (2012): 1418-28. doi:10.1007/s00134-012-2613-1*

[4] BARBOSA WANDERLEY, Maria Fernanda et al., *Seleção de Caracterısticas Baseada em Analise da Área abaixo da Curva ROC de Classificadores KDE-Bayesianos. Julho de 2010*

[5] BEATTY, Mark et al., *Estimating the global burden of dengue. The American journal of tropical medicine and hygiene, [S. l.], jan. 2009.*

[6] BHATT S, et al.,*The global distribution and burden of dengue. Nature. 2013 Apr 25;496(7446):504-7. doi: 10.1038/nature12060. Epub 2013 Apr 7. PMID: 23563266; PMCID: PMC3651993.*

[7] BURT, FELICITY J ET AL.,*"Chikungunya virus: an update on the biology and pathogenesis of this emerging pathogen." The Lancet. Infectious diseases vol. 17,4 (2017): e107-e117. doi:10.1016/S1473-3099(16)30385-1*

[8] BREIMAN, L., *"Random Forests", Jan.2001.*

[9] DUARTE,Heloisa Helena Pelluci & FRANÇA,Elisabeth Barboza ,*Qualidade dos dados da vigilância epidemiológica da dengue em Belo Horizonte, MG, 2006*

[10] ELITE DATA SCIENCE, *HOW TO HANDLE IMBALANCED CLASSES IN MACHINE LEARNING, Elite Data Science. Disponível em: https://elitedatascience.com/imbalanced-classes. Acesso em 14 de agosto de 2020*

[11] FATHIMA,Shameem & HUNDEWAL,Nisar.,*Comparison of Classification Techniques-SVM and Naives Bayes to predict the Arboviral Disease-Dengue*

[12] GARCIA, S.C., *O uso de árvores de decisão na descoberta de conhecimento na área da saúde. In: SEMANA ACADÊMICA, 2000. Rio Grande do Sul: Universidade Federal do Rio Grande do Sul, 2000.*

[13] 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.*

[14] GIBSON,Gerusa.,*Fatores Associados à Ocorrência de Casos Graves de Dengue: Análise dos Anos Epidêmicos de 2007-2008 no Rio de Janeiro. 2012. Tese (Doutorado) – Escola Nacional de Saúde Pública Sergio Arouca, Rio de Janeiro-RJ 2012.*

[15] GONG, R.; FONSECA, E.; BOGDANOV, D.; SLIZOVSKAIA, O.; GOMEZ, E.; SERRA, X., *Acoustic scene classification by fusing lightgbm and vggnet multichannel predictions. Proc. IEEE AASP Challenge Detection Classification Acoust. scenes events. 2017.*

[16] GOELBEL, M. & GRUENWALD, L., *A Survey of Data Mining and Knowledge Discovery Software Tools; 1999.*

[17] GOULD, Ernest et al.,*Emerging arboviruses: Why today?. One Health, 4 (2017), 1–13.*

[18] GUZMAN, Maria G. et al.,*Dengue infection. Nature Reviews Disease Primers, 2016*

[19] GRUS, Joel., *Data Science do Zero: Primeiras regras com o python. [S. l.]: Altas Books, 2019. 336 p. v. 1.*

[20] HOSMER & LEMESHOW, *"Applied Logistic Regression";2nd ed.*

[21] HUBÁLEK, Zdenek et al.,*Arboviruses Pathogenic for Domestic and Wild Animals; Advances in Virus Research, Volume 89, 2014*

[22] KOHAVI, R.,*A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint Conference on artificial intelligence. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.*

[23] KRAEMER et al.,*The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus 2015.*

[24] KROHLING, Lívia Lima; PAULA, Kely Maria Pereira de; BEHLAU, Mara., *Curva ROC do Protocolo Qualidade de Vida em Voz Pediátrico (QVV-P). CoDAS, São Paulo , v. 28, n. 3, p. 311-313, June 2016.*

[25] MOREIRA MAIA, Cynthia et al., *Estudo Sobre o Uso de Árvores de Decisão na Área da Saúde. Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA, p. 23-30, jun. 2017.*

[26] MYCKINNEY, Wes. , *Python para Análise de Dados: Tratamento de dados com pandas,numpy e ipython. [S. l.]: Novatec Editora, 2018. 616 p. v. 1.*

[27] PEDREGOSA ET AL., *Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.*

[28] PONE, Sheila Moura et al .,*Sinais clínicos e laboratoriais para o dengue com evolução grave em crianças hospitalizadas. J. Pediatr. (Rio J.), Porto Alegre , v. 92, n. 5, p. 464-471, Oct. 2016.*

[29] SILVA, Ana Maria da.,*Estudo de cinética de viremia do vírus dengue sorotipo 3 em formas clínicas da dengue com diferentes níveis de gravidade. / Ana Maria da Silva. — Recife: A. M. da Silva, 2008. Dissertação (Mestrado em Saúde Pública) - Centro de Pesquisas Aggeu Magalhães, Fundação Oswaldo Cruz.*

[30] ST. JOHN, Ashley L. et al.,*Adaptive immune responses to primary and secondary dengue virus infections. Nature Reviews Immunology, (2019), –. doi:10.1038/s41577-019-0123-x*

[31] SUHRBIER, A.,*Rheumatic manifestations of chikungunya: emerging concepts and interventions. Nat Rev Rheumatol 15, 597–611 (2019). https://doi.org/10.1038/s41584-019-0276-9*

[32] WILDER SMITH A. ET AL.,*Dengue. Lancet. 2019 Jan 26;393(10169):350-363. doi: 10.1016/S0140-6736(18)32560-1. PMID: 30696575.*

[33] WILDER SMITH A. ,*Dengue Infections. Challenges in Infectious Diseases, 2012*

[34] ZHANG, Z., *"Naive Bayes Classification in R", Ann Transl Med v. 4, n. 12, Jun. 2016*

9 ANEXOS

9.1 Dicionário De Variáveis

Variável	Descrição	Valor
age	Idade	Anos
age_adulto	Faixa de Idade	Anos Entre 20 e 60 anos
ja_teve_zika	Paciente relatou já ter apresentado Zika	0 (não) – 1 (sim)
tem_alguma_doenca_das_articulacoes	Paciente apresenta doença nas articulações	0 (não) – 1 (sim)
exantema	Presença de exantema	0 (não) – 1 (sim)
edema	Presença de edema	0 (não) – 1 (sim)
sinais_artrite	Presença de sinais de artrite	0 (não) – 1 (sim)
outros_sinais_alarme	Presença de algum outro sinal de alarme	0 (não) – 1 (sim)
esteve_hospitalizado_D07	Paciente relatou hospitalização no período até 7 dias após inclusão	0 (não) – 1 (sim)
esteve_hospitalizado_D14	Paciente relatou hospitalização no período até 14 dias após inclusão	0 (não) – 1 (sim)
SAQ_Dor intensa e contínua no abdome (barriga), espontânea ou ao apertar	Apresentou o sinal de alarme "dor intensa e contínua no abdomen" na inclusão	0 (não) – 1 (sim)
SAQ_Vômitos persistentes	Apresentou o sinal de alarme "vomitos persistentes" na inclusão	0 (não) – 1 (sim)
SAQ_Pressão baixa, escurecimento da vista, sudorese ou desmaio ao se levantar	Apresentou o sinal de alarme "Pressão baixa, escurecimento da vista, sudorese ou desmaio ao se levantar" na inclusão	0 (não) – 1 (sim)
SAQ_Muito sono, ideias confusas, fala atrapalhada	Apresentou o sinal de alarme "Muito sono, ideias confusas, fala atrapalhada" na inclusão	0 (não) – 1 (sim)
SA_D07_Manchas vermelhas na pele	Apresentou o sinal de alarme "Manchas vermelhas na pele" 07 dias após a inclusão	0 (não) – 1 (sim)
SA_D07_Vômitos com sangue	Apresentou o sinal de alarme "Vômitos com sangue" 07 dias após a inclusão	0 (não) – 1 (sim)
SA_D07_Dor intensa e contínua no abdome (barriga), espontânea ou ao apertar	Apresentou o sinal de alarme "Dor intensa e contínua no abdome (barriga), espontânea ou ao apertar" 07 dias após a inclusão	0 (não) – 1 (sim)
SA_D07_! 'Sangramento de mucosas (nariz, gengiva, etc) '	Apresentou o sinal de alarme "Sangramento de mucosas" 07 dias após a inclusão	0 (não) – 1 (sim)
SA_D07_Sangramento na urina ou fezes	Apresentou o sinal de alarme "Sangramento na urina ou fezes" 07 dias após a inclusão	0 (não) – 1 (sim)
SA_D07_Fluxo de sangue persistente e volumoso pelo útero	Apresentou o sinal de alarme "Fluxo de sangue persistente e volumoso pelo útero" 07 dias após a inclusão	0 (não) – 1 (sim)
SA_D14_! 'Muito sono, ideias confusas, fala atrapalhada '	Apresentou o sinal de alarme "Muito sono, ideias confusas, fala atrapalhada" 14 dias após a inclusão	0 (não) – 1 (sim)
SA_D14_Pressão baixa, escurecimento da vista, sudorese ou desmaio ao se levantar	Apresentou o sinal de alarme "Pressão baixa, escurecimento da vista, sudorese ou desmaio ao se levantar" 14 dias após a inclusão	0 (não) – 1 (sim)
SA_D14_Dor intensa e contínua no abdome (barriga), espontânea ou ao apertar	Apresentou o sinal de alarme "Dor intensa e contínua no abdome (barriga), espontânea ou ao apertar" 14 dias após a inclusão	0 (não) – 1 (sim)
SA_D14_Vômitos persistentes	Apresentou o sinal de alarme "vomitos persistentes" na inclusão	0 (não) – 1 (sim)
SA_D14_! 'Sangramento de mucosas (nariz, gengiva, etc) '	Apresentou o sinal de alarme "Sangramento de mucosas" 07 dias após a inclusão	0 (não) – 1 (sim)
MARCADOR_SuPAR	Marcador suPAR (ativador de plasminogênio tipo uroquinase solúvel)	pg/ml
HT 1	Hematócrito	%
PLAQ 1	Plaquetas	mil/mm³
TGO 1	Transaminase glutâmico oxalacética (TGO) – aspartato aminotransferase (AST)	U/L
TGP 1	Transaminase glutâmico pirúvica (TGP) - alanina aminotransferase (ALT) – U/L	U/L

9.2 Medidas Brutas De Cada Técnica

Medidas Brutas Regressão Logística

```
Treinamento AUC-ROC:1.0
Validacao AUC-ROC:0.9691358024691358
col_0              0    1
fl_severidade
0                259    0
1                  0  259
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       259
           1       1.00      1.00      1.00       259

    accuracy                           1.00       518
   macro avg       1.00      1.00      1.00       518
weighted avg       1.00      1.00      1.00       518

AUC: 1.00
col_0              0   1
fl_severidade
0                100   8
1                  5  10
              precision    recall  f1-score   support

           0       0.95      0.93      0.94       108
           1       0.56      0.67      0.61        15

    accuracy                           0.89       123
   macro avg       0.75      0.80      0.77       123
weighted avg       0.90      0.89      0.90       123

AUC: 0.97
```

```
Treinamento AUC-ROC:1.0
Validacao AUC-ROC:0.8820987654320988
col_0              0   1
fl_severidade
0                 27   0
1                  0  27
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        27
           1       1.00      1.00      1.00        27

    accuracy                           1.00        54
   macro avg       1.00      1.00      1.00        54
weighted avg       1.00      1.00      1.00        54

AUC: 1.00
col_0              0   1
fl_severidade
0                 76  32
1                  2  13
              precision    recall  f1-score   support

           0       0.97      0.70      0.82       108
           1       0.29      0.87      0.43        15

    accuracy                           0.72       123
   macro avg       0.63      0.79      0.63       123
weighted avg       0.89      0.72      0.77       123

AUC: 0.88
```

Medidas Brutas Naive Bayes

```
Treinamento AUC-ROC:0.6565793592820 62
Validacao AUC-ROC:0.6222222222222222
col_0              0   1
fl_severidade
0                226  33
1                181  78
              precision    recall  f1-score   support

           0       0.56      0.87      0.68       259
           1       0.70      0.30      0.42       259

    accuracy                           0.59       518
   macro avg       0.63      0.59      0.55       518
weighted avg       0.63      0.59      0.55       518

AUC: 0.66
col_0              0   1
fl_severidade
0                 95  13
1                 12   3
              precision    recall  f1-score   support

           0       0.89      0.88      0.88       108
           1       0.19      0.20      0.19        15

    accuracy                           0.80       123
   macro avg       0.54      0.54      0.54       123
weighted avg       0.80      0.80      0.80       123

AUC: 0.62
```

```
Treinamento AUC-ROC:0.6639231824417009
Validacao AUC-ROC:0.604320987654321
col_0              0  1
fl_severidade
0                 26  1
1                 20  7
              precision    recall  f1-score   support

           0       0.57      0.96      0.71        27
           1       0.88      0.26      0.40        27

    accuracy                           0.61        54
   macro avg       0.72      0.61      0.56        54
weighted avg       0.72      0.61      0.56        54

AUC: 0.66
col_0              0   1
fl_severidade
0                 95  13
1                 12   3
              precision    recall  f1-score   support

           0       0.89      0.88      0.88       108
           1       0.19      0.20      0.19        15

    accuracy                           0.80       123
   macro avg       0.54      0.54      0.54       123
weighted avg       0.80      0.80      0.80       123

AUC: 0.60
```

Medidas Brutas Árvore de Decisão

```
col_0              0    1
fl_severidade
0                259    0
1                  0  259
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       259
           1       1.00      1.00      1.00       259

    accuracy                           1.00       518
   macro avg       1.00      1.00      1.00       518
weighted avg       1.00      1.00      1.00       518

col_0              0   1
fl_severidade
0                107   1
1                  4  11
              precision    recall  f1-score   support

           0       0.96      0.99      0.98       108
           1       0.92      0.73      0.81        15

    accuracy                           0.96       123
   macro avg       0.94      0.86      0.90       123
weighted avg       0.96      0.96      0.96       123
```

```
col_0              0   1
fl_severidade
0                 27   0
1                  0  27
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        27
           1       1.00      1.00      1.00        27

    accuracy                           1.00        54
   macro avg       1.00      1.00      1.00        54
weighted avg       1.00      1.00      1.00        54

col_0              0   1
fl_severidade
0                 98  10
1                  2  13
              precision    recall  f1-score   support

           0       0.98      0.91      0.94       108
           1       0.57      0.87      0.68        15

    accuracy                           0.90       123
   macro avg       0.77      0.89      0.81       123
weighted avg       0.93      0.90      0.91       123
```

Medidas Brutas Random Forest

```
col_0              0    1
fl_severidade
0                256    3
1                  0  259
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       259
           1       0.99      1.00      0.99       259

    accuracy                           0.99       518
   macro avg       0.99      0.99      0.99       518
weighted avg       0.99      0.99      0.99       518

col_0              0   1
fl_severidade
0                103   5
1                  4  11
              precision    recall  f1-score   support

           0       0.96      0.95      0.96       108
           1       0.69      0.73      0.71        15

    accuracy                           0.93       123
   macro avg       0.83      0.84      0.83       123
weighted avg       0.93      0.93      0.93       123
```

```
col_0              0   1
fl_severidade
0                 25   2
1                  0  27
              precision    recall  f1-score   support

           0       1.00      0.93      0.96        27
           1       0.93      1.00      0.96        27

    accuracy                           0.96        54
   macro avg       0.97      0.96      0.96        54
weighted avg       0.97      0.96      0.96        54

col_0              0   1
fl_severidade
0                 95  13
1                  0  15
              precision    recall  f1-score   support

           0       1.00      0.88      0.94       108
           1       0.54      1.00      0.70        15

    accuracy                           0.89       123
   macro avg       0.77      0.94      0.82       123
weighted avg       0.94      0.89      0.91       123
```

Todas Variáveis

Medidas Brutas Regressão Logística

```
Treinamento AUC-ROC:0.9936718295791654
Validacao AUC-ROC:0.9870370370370370
col_0            0    1
fl_severidade
0              245   14
1                0  259
              precision    recall  f1-score   support

           0       1.00      0.95      0.97       259
           1       0.95      1.00      0.97       259

    accuracy                           0.97       518
   macro avg       0.97      0.97      0.97       518
weighted avg       0.97      0.97      0.97       518

AUC: 0.99
col_0            0   1
fl_severidade
0              100   8
1                0  15
              precision    recall  f1-score   support

           0       1.00      0.93      0.96       108
           1       0.65      1.00      0.79        15

    accuracy                           0.93       123
   macro avg       0.83      0.96      0.88       123
weighted avg       0.96      0.93      0.94       123

AUC: 0.99
```

```
Treinamento AUC-ROC:0.9835390946502057
Validacao AUC-ROC:0.9907407407407407
col_0            0   1
fl_severidade
0              26   1
1               3  24
              precision    recall  f1-score   support

           0       0.90      0.96      0.93        27
           1       0.96      0.89      0.92        27

    accuracy                           0.93        54
   macro avg       0.93      0.93      0.93        54
weighted avg       0.93      0.93      0.93        54

AUC: 0.98
col_0            0   1
fl_severidade
0              105   3
1                1  14
              precision    recall  f1-score   support

           0       0.99      0.97      0.98       108
           1       0.82      0.93      0.87        15

    accuracy                           0.97       123
   macro avg       0.91      0.95      0.93       123
weighted avg       0.97      0.97      0.97       123

AUC: 0.99
```

Medidas Brutas Naive Bayes

```
Treinamento AUC-ROC:0.8885004695815506
Validacao AUC-ROC:0.8512345679012345
col_0            0   1
fl_severidade
0              240  19
1              191  68
              precision    recall  f1-score   support

           0       0.56      0.93      0.70       259
           1       0.78      0.26      0.39       259

    accuracy                           0.59       518
   macro avg       0.67      0.59      0.54       518
weighted avg       0.67      0.59      0.54       518

AUC: 0.89
col_0            0   1
fl_severidade
0              97  11
1              12   3
              precision    recall  f1-score   support

           0       0.89      0.90      0.89       108
           1       0.21      0.20      0.21        15

    accuracy                           0.81       123
   macro avg       0.55      0.55      0.55       123
weighted avg       0.81      0.81      0.81       123

AUC: 0.85
```

```
Treinamento AUC-ROC:0.8885004695815506
Validacao AUC-ROC:0.8512345679012345
col_0            0   1
fl_severidade
0              240  19
1              191  68
              precision    recall  f1-score   support

           0       0.56      0.93      0.70       259
           1       0.78      0.26      0.39       259

    accuracy                           0.59       518
   macro avg       0.67      0.59      0.54       518
weighted avg       0.67      0.59      0.54       518

AUC: 0.89
col_0            0   1
fl_severidade
0              97  11
1              12   3
              precision    recall  f1-score   support

           0       0.89      0.90      0.89       108
           1       0.21      0.20      0.21        15

    accuracy                           0.81       123
   macro avg       0.55      0.55      0.55       123
weighted avg       0.81      0.81      0.81       123

AUC: 0.85
```

Medidas Brutas Árvore de Decisão

```
col_0            0    1
fl_severidade
0              250    9
1                0  259
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       259
           1       0.97      1.00      0.98       259

    accuracy                           0.98       518
   macro avg       0.98      0.98      0.98       518
weighted avg       0.98      0.98      0.98       518

col_0            0  1
fl_severidade
0              100  8
1                6  9
              precision    recall  f1-score   support

           0       0.94      0.93      0.93       108
           1       0.53      0.60      0.56        15

    accuracy                           0.89       123
   macro avg       0.74      0.76      0.75       123
weighted avg       0.89      0.89      0.89       123
```

```
col_0            0   1
fl_severidade
0              26   1
1               0  27
              precision    recall  f1-score   support

           0       1.00      0.96      0.98        27
           1       0.96      1.00      0.98        27

    accuracy                           0.98        54
   macro avg       0.98      0.98      0.98        54
weighted avg       0.98      0.98      0.98        54

col_0            0   1
fl_severidade
0              100   8
1                1  14
              precision    recall  f1-score   support

           0       0.99      0.93      0.96       108
           1       0.64      0.93      0.76        15

    accuracy                           0.93       123
   macro avg       0.81      0.93      0.86       123
weighted avg       0.95      0.93      0.93       123
```

Medidas Brutas Random Forest

```
col_0            0    1
fl_severidade
0              247   12
1                0  259
              precision    recall  f1-score   support

           0       1.00      0.95      0.98       259
           1       0.96      1.00      0.98       259

    accuracy                           0.98       518
   macro avg       0.98      0.98      0.98       518
weighted avg       0.98      0.98      0.98       518

col_0            0   1
fl_severidade
0              100   8
1                3  12
              precision    recall  f1-score   support

           0       0.97      0.93      0.95       108
           1       0.60      0.80      0.69        15

    accuracy                           0.91       123
   macro avg       0.79      0.86      0.82       123
weighted avg       0.93      0.91      0.92       123
```

```
col_0            0   1
fl_severidade
0              25   2
1               0  27
              precision    recall  f1-score   support

           0       1.00      0.93      0.96        27
           1       0.93      1.00      0.96        27

    accuracy                           0.96        54
   macro avg       0.97      0.96      0.96        54
weighted avg       0.97      0.96      0.96        54

col_0            0   1
fl_severidade
0              99   9
1               0  15
              precision    recall  f1-score   support

           0       1.00      0.92      0.96       108
           1       0.62      1.00      0.77        15

    accuracy                           0.93       123
   macro avg       0.81      0.96      0.86       123
weighted avg       0.95      0.93      0.93       123
```

Variáveis Laboratório

Medidas Brutas Regressão Logística

```
Treinamento AUC-ROC:1.0
Validacao AUC-ROC:0.9876190476190476
col_0            0   1
fl_severidade
0               86   1
1                0  87
              precision    recall  f1-score   support

           0       1.00      0.99      0.99        87
           1       0.99      1.00      0.99        87

    accuracy                           0.99       174
   macro avg       0.99      0.99      0.99       174
weighted avg       0.99      0.99      0.99       174

AUC: 1.00
col_0            0   1
fl_severidade
0              279   1
1                3  12
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       280
           1       0.92      0.80      0.86        15

    accuracy                           0.99       295
   macro avg       0.96      0.90      0.93       295
weighted avg       0.99      0.99      0.99       295

AUC: 0.99
```

```
Treinamento AUC-ROC:1.0
Validacao AUC-ROC:0.9538095238095238
col_0            0   1
fl_severidade
0               27   0
1                0  27
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        27
           1       1.00      1.00      1.00        27

    accuracy                           1.00        54
   macro avg       1.00      1.00      1.00        54
weighted avg       1.00      1.00      1.00        54

AUC: 1.00
col_0            0   1
fl_severidade
0              223  57
1                0  15
              precision    recall  f1-score   support

           0       1.00      0.80      0.89       280
           1       0.21      1.00      0.34        15

    accuracy                           0.81       295
   macro avg       0.60      0.90      0.62       295
weighted avg       0.96      0.81      0.86       295

AUC: 0.95
```

Medidas Brutas Naive Bayes

```
Treinamento AUC-ROC:0.574448407979918
Validacao AUC-ROC:0.46714285714285714
col_0            0   1
fl_severidade
0               47  40
1               38  49
              precision    recall  f1-score   support

           0       0.55      0.54      0.55        87
           1       0.55      0.56      0.56        87

    accuracy                           0.55       174
   macro avg       0.55      0.55      0.55       174
weighted avg       0.55      0.55      0.55       174

AUC: 0.57
col_0            0   1
fl_severidade
0              213  67
1               10   5
              precision    recall  f1-score   support

           0       0.96      0.76      0.85       280
           1       0.07      0.33      0.11        15

    accuracy                           0.74       295
   macro avg       0.51      0.55      0.48       295
weighted avg       0.91      0.74      0.81       295

AUC: 0.47
```

```
Treinamento AUC-ROC:0.5514403292181069
Validacao AUC-ROC:0.5204761904761904
col_0            0   1
fl_severidade
0               19   8
1               17  10
              precision    recall  f1-score   support

           0       0.53      0.70      0.60        27
           1       0.56      0.37      0.44        27

    accuracy                           0.54        54
   macro avg       0.54      0.54      0.52        54
weighted avg       0.54      0.54      0.52        54

AUC: 0.55
col_0            0   1
fl_severidade
0              256  24
1               13   2
              precision    recall  f1-score   support

           0       0.95      0.91      0.93       280
           1       0.08      0.13      0.10        15

    accuracy                           0.87       295
   macro avg       0.51      0.52      0.52       295
weighted avg       0.91      0.87      0.89       295

AUC: 0.52
```

Medidas Brutas Árvore de Decisão

```
col_0            0   1
fl_severidade
0               87   0
1                0  87
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        87
           1       1.00      1.00      1.00        87

    accuracy                           1.00       174
   macro avg       1.00      1.00      1.00       174
weighted avg       1.00      1.00      1.00       174

col_0            0   1
fl_severidade
0              277   3
1                3  12
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       280
           1       0.80      0.80      0.80        15

    accuracy                           0.98       295
   macro avg       0.89      0.89      0.89       295
weighted avg       0.98      0.98      0.98       295
```

```
col_0            0   1
fl_severidade
0               27   0
1                0  27
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        27
           1       1.00      1.00      1.00        27

    accuracy                           1.00        54
   macro avg       1.00      1.00      1.00        54
weighted avg       1.00      1.00      1.00        54

col_0            0   1
fl_severidade
0              277   3
1                4  11
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       280
           1       0.79      0.73      0.76        15

    accuracy                           0.98       295
   macro avg       0.89      0.86      0.87       295
weighted avg       0.98      0.98      0.98       295
```

Medidas Brutas Random Forest

```
col_0            0   1
fl_severidade
0               87   0
1                0  87
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        87
           1       1.00      1.00      1.00        87

    accuracy                           1.00       174
   macro avg       1.00      1.00      1.00       174
weighted avg       1.00      1.00      1.00       174

col_0            0   1
fl_severidade
0              277   3
1                2  13
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       280
           1       0.81      0.87      0.84        15

    accuracy                           0.98       295
   macro avg       0.90      0.93      0.91       295
weighted avg       0.98      0.98      0.98       295
```

```
col_0            0   1
fl_severidade
0               27   0
1                0  27
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        27
           1       1.00      1.00      1.00        27

    accuracy                           1.00        54
   macro avg       1.00      1.00      1.00        54
weighted avg       1.00      1.00      1.00        54

col_0            0   1
fl_severidade
0              251  29
1                0  15
              precision    recall  f1-score   support

           0       1.00      0.90      0.95       280
           1       0.34      1.00      0.51        15

    accuracy                           0.90       295
   macro avg       0.67      0.95      0.73       295
weighted avg       0.97      0.90      0.92       295
```

Treino Palmas- Teste Rio Preto

Medidas Brutas Regressão Logística

```
Treinamento AUC-ROC:0.9999777264661553
Validacao AUC-ROC:0.9999697464754643
col_0            0    1
fl_severidade
0              364    3
1                0  367
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       367
           1       0.99      1.00      1.00       367

    accuracy                           1.00       734
   macro avg       1.00      1.00      1.00       734
weighted avg       1.00      1.00      1.00       734

AUC: 1.00
col_0            0   1
fl_severidade
0              784   3
1                0  42
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       787
           1       0.93      1.00      0.97        42

    accuracy                           1.00       829
   macro avg       0.97      1.00      0.98       829
weighted avg       1.00      1.00      1.00       829

AUC: 1.00
```

Treino Palmas+ Rio Preto – Teste Base Total