**INSTITUTO DE MATEMÁTICA E ESTATÍSTICA**

**RELATÓRIO PARCIAL DE ATIVIDADES**

**REFERENTES AOS MESES DE JULHO E AGOSTO**

**São Paulo, Agosto de 2020**

**TRABALHO DE FORMATURA – T.2020154**

**BMA e BMAC**

1. NOME DO ALUNO: **Fabio Carvalho de Souza**

2. NÚMERO USP: **9425125**

3. CURSO: [ ] BMA **[x] BMAC**

4. HABILITAÇÃO:

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

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

5. NOME DO ORIENTADOR: **Prof. Dr. [Helder Takashi Imoto Nakaya](http://csbiology.com/" \t "_blank)**

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

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

UNIDADE DO CO-ORIENTADOR: À Definir

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

Fabio C.S

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Assinatura do aluno Assinatura do orientador

1. **INTRODUÇÃO**

O desenvolvimento de modelos depende muito das métricas/medidas que os mesmos apresentam, pois conseguimos verificar o grau de separação em diversos pontos do processo de modelagem (treino, teste e teste com base total) e assim verificar pontos a melhorar e entender qual o modelo mais adequado para o problema, além de entender a necessidade ou não de realizar outros tipos de tratamentos que possam melhorar essas métricas. E estes são basicamente os pontos a serem observados nessa etapa do desenvolvimento do projeto, com o intuito de possuir uma visão muito mais clara e direcionada das técnicas a serem trabalhadas, dado um conjunto já selecionado de variáveis que podem expressar a resposta do problema proposto, que consiste em definir o que pode provocar num indivíduo a sua chegada em uma condição de dengue severa.

**2- ATIVIDADES DESENVOLVIDAS E RESULTADOS PARCIAIS**

Nesta seção, iremos apresentar um resumo das atividades trabalhadas nesse bimestre, bem como observar os resultados parciais de cada uma das técnicas usadas para modelagem, mas antes de dar início a descrição, é necessário comentar sobre as métricas usadas como observação, pois com base nelas que é possível verificar a melhor opção de modelo para que seja trabalhada em mais detalhes nas fases finais do projeto, sendo elas as seguintes:

**2.1.1- Matriz de Confusão**

É uma tabela que consiste em verificar o total de erros e acertos do modelo em comparação os dados da variável alvo original e que é a base para as demais medidas que são observadas pois dentro dessa tabela podemos ter informação de:

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.

**2.1.2- 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.

**2.1.3- 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.

**2.1.4- 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 que o indivíduo esta saudável.

**2.1.5- 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.

**2.1.6- Curva ROC**

A curva ROC (Receiver Operating Characteristic) é uma métrica de probabilidade, que nos permite verificar o quão bom o modelo é em distinguir duas classes. Ela tem como parâmetros 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 falsos 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 )

**2.1.7- AUC**

O AUC consiste numa métrica para o grau de separabilidade que o modelo apresenta, ou seja, em classificar um grupo como portador de dada característica ou não, sendo basicamente 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 )

**2.2- Tratamentos De Informações:**

Dado ao avanço do projeto, já foram realizados diversos processos para melhorar a qualidade dos dados, sendo desde eliminação e criação de variáveis, alteração de tipos dos dados. Sendo assim os tratamentos nessa etapa são mínimos, apenas para ajuste de medidas, devido aos novos casos adicionados que melhoram a quantidade de casos positivos presentes, mas não de forma suficiente, assim para melhorar sua entrada no processo de modelagem se utilizou duas novas formas de tratamento que consistem basicamente em:

- Classe minoritária Up-sample

Um processo que consiste em duplicar aleatoriamente as observações da classe minoritária para reforçar seu sinal, ou seja, adiciono mais elementos positivos da nossa “flag” até um número próximo ao de elementos negativos para ficar uma proporção mais equilibrada e ajude o modelo a verificar de forma mais adequada as classes existentes e não apenas de forma única a majoritária.

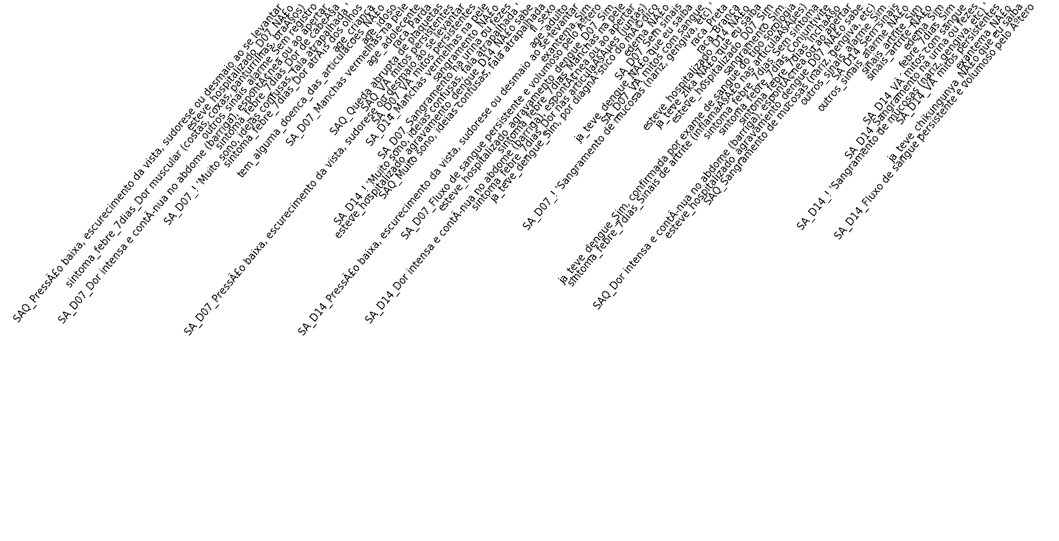
- Classe de maioria reduzida

Este processo consiste do oposto do método anterior mencionado do qual reduz os elementos negativos até um número próximo de elementos positivos, assim podemos verificar tanto o comportamento do modelo com um grande número de informações, bem como verificar seu desempenho com um número pequeno de indivíduos.

**2.3- Resultados Parciais**

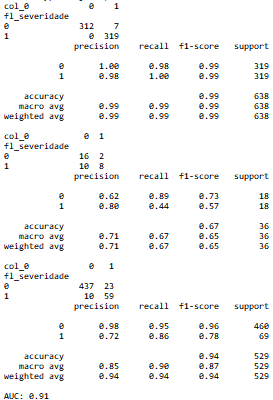
Como mencionado, foram realizadas duas técnicas para melhorar os resultados das 4 propostas de modelagem usadas (Random Forest, Regressão Logistica, Naive Bayes, Árvore de decisão), com um grupo de aproximadamente 530 pacientes que temos marcação de dengue confirmada e dentre eles temos os 69 casos severos, que foram trabalhados com as variáveis da figura 1.

Figura 1: Variáveis Processadas



Assim dividimos os dados em treino (70% dos dados), teste (30% dos dados) e total (teste do modelo com todos os casos) e obtemos algumas das medidas citadas antes e que estão presentes nas imagens abaixo.

Figura 2: Dados Árvore de Decisão



Na figura 2 temos as medidas citadas acima e que se repetem nas demais figuras de cada um dos métodos usados. E podemos verificar que a variavel resposta “fl\_severidade” está expressa como matriz, que é vista da seguinte forma:

- Linha 0 e Coluna 0 são os elementos que são a negativa do objetivo do modelo e que são coerentes com a base original.

- Linha 0 e Coluna 1 são os elementos que originalmente são zeros mas que o modelo errou dizendo ser 1.

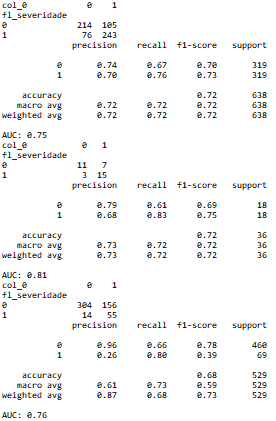
- Linha 1 e coluna 0 são os elementos originalmente como 1, ou seja que são severos na base original e que o modelo errou falando serem não severos.

- Linha 1 e Coluna 1 são os elementos corretos tanto na base original como o que o modelo apontou como severo.

Apenas com essas informações já é possível verificar se o modelo está separando as informações de forma adequada e temos isso em 3 etapas diferentes que são no momento de treino considerando uma amostragem com repetição para aumentar a expressão da variavel alvo, no momento de teste com o balanceamento para reduzir o número de não severos até um número proximo de casos severos e assim analisar o modelo com poucos dados e por fim analisar o modelo considerando a base sem manupulação que seria a situação real do problema e dessa forma verificamos o seguinte:

- Durante o treino com número elevado de casos positivos o modelo consegue dizer o que é um caso severo de não severo o que é bom para analisar grandes volumes de dados de pacientes, mas que , na verificação do teste com poucos individuos o modelo não apresenta boas medidas que podemos apontar ser falta de informação que consiga separar os individuos e por fim na base origem temos medidas consideradas boas com uma separação com um erro não muito elevado mas que precisa ser verificado.

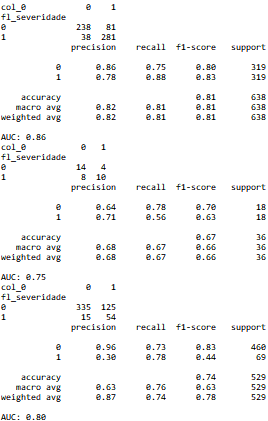
Figura 3: Dados Naive Bayes



Como já citado, todas as figuras com dados apresentam as mesmas medidas o que vai variar é a interpretação do modelo em cada etapa e para o Naive Bayes temos as seguintes:

- No processo de treino verificamos que o modelo tem um erro maior se comparado com o anterior, já se mostrando não muito recomentado pois, ele coloca muitos individuos marcados com positivos como negativos e o mesmo vale para casos negativos como positivos o que é ruim para a proposta, já para o teste verificamos o contrário do modelo anterior ele consegue de melhor forma separar individuos num grupo pequeno, nos dando a ideia se existir a necessidade de um modelo para grandes grupos e um para pequenos grupos, por fim na base origem temos dados equilibrados mas com um erro maior que o apresentado na árvore de decisão.

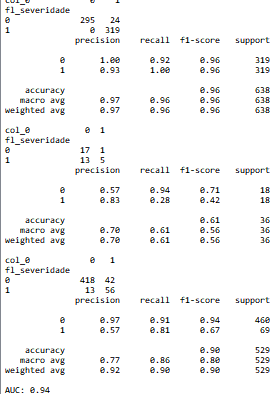
Figura 4: Dados Regressão Logística



Para o modelo de regressão logística em cada etapa temos o seguinte:

- Na etapa de treino o modelo apresenta um erro considerável na separação, mas não tão alto como a técnica anterior, mas consideravelmente maior que o modelo de árvore sendo assim não muito recomentado, mas assim como o modelo de Bayes, para um grupo pequeno conseguimos uma separação boa, já para a etapa da base origem os dados de separação são próximos entre os modelos anteriores.

Figura 5: Dados Random Forest



Por fim a técnica de random forest apresenta uma etapa de treino bem próxima do modelo de árvore de decisão sendo uma opção boa a ser considerada, mas do mesmo modo que a árvore não consegue separar os indivíduos do grupo menor de teste, esta técnica também tem esse problema, e a base origem como nos casos anteriores as medidas estão relativamente próximas sendo de certa forma bons modelos.

Assim podemos concluir que temos dois cenários possíveis sendo o primeiro que teremos um grupo de modelos que tem maior facilidade em separar indivíduos em grupos densos de informações (Random Forest e Árvore de Decisão) pois assim a técnica é capaz de observar um padrão que pelos dados apresentados são bem assertivos, porem temos aqui o segundo cenário que consiste em termos um grupo de modelos para um conjunto menor de indivíduos (Logistica e Naive Bayes). Nesta situação temos que todos os modelos são desenvolvidos da mesma forma, mas aplicados em situações especificas, o que vale como próximos passos é tentar compreender se de fato essa é a forma correta de agir e assim maximizar o desempenho dessas técnicas e entender se é possível encontrar um meio termo em alguma técnica que possa ter a unificação de situações de grandes e pequenos grupos, o que facilitaria sua aplicação ainda mais.

**3- REFERÊNCIAS BIBLIOGRÁFICAS**

BARBOSA WANDERLEY, Maria Fernanda et al.,Seleção de Caracterıs-ticas Baseada em Analise da Área abaixo da Curva ROC de Classificadores KDE-Bayesianos. Julho de 2010

BREIMAN, L.,“Random Forests”, Jan.2001.

GARCIA, S.C.,O uso de árvores de decisão na descoberta de conhecimento na áreada saúde. In: SEMANA ACADÊMICA, 2000. Rio Grande do Sul: UniversidadeFederal do Rio Grande do Sul, 2000.

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

ZHANG, Z.,“Naive Bayes Classification in R”, Ann Transl Med v. 4, n. 12, Jun.2016