Modelo de previsão do tempo de internação de pacientes com Síndrome (SARG) pós pandemia.

Grupo Saúde:

- Douglas Onassis
- Icaro Pinheiro

CONTEXTO

Farmácias e hospitais observam alta de casos de Covid-19

Testes rápidos com resultado positivo nas drogarias do país passaram de 9,36% para 15,5% de uma semana para outra

4.nov.2022 às 20h41

BandNews FM

Internação por Covid-19 em UTI aumenta 86,5% em SP

Aumento de casos é observado em outros estados do país, especialistas afirmam que momento é de alerta, mas ainda não de preocupação

BandNews FM
04/11/2022 • 13:29

Contexto

No início da pandemia de covid-19, diversos leitos foram mobilizados e abertos para atendimento dos casos da doença. E seus custos são altos seja nos equipamentos, recursos humanos e manutenção.

Com o avançar da vacinação, o número de casos e internações reduziram substancialmente e os leitos covid19 tornaram-se ociosos. Entretanto, após o surgimento da ômicron antes de 2022 fez aumentar os casos e internações gerando nova preocupação na gestão de leitos pelo sistema de saúde nacional.

Ter ferramentas que identifiquem as características epidemiológicas dos pacientes que possam prever o tempo de internação é fundamental para uma melhor gestão de leitos e para uma boa condução na rotina hospitalar nos diferentes momentos da pandemia.

PROBLEMA DE NEGÓCIO

Problema de Negócio e Impacto

Nosso problema é a dificuldade de estimar os pacientes com Síndrome Aguda Respiratória Grave (SARG) que terão maior tempo de internação no cenário pós Pandemia, sofrendo ou não influências de novas ondas de contaminações seja por Covid, Influenza e outros vírus gripais.

Decidimos então, criar um modelo preditivo que possa ajudar na previsão do tempo de internação de cada paciente a partir de suas características individuais e poder identificar aqueles que terão maior necessidade em leito hospitalar para seu tratamento e cura.

O principal impacto será apoiar nas tomadas de decisões da equipe assistencial e gestores hospitalares visando o melhor tratamento para o paciente, melhor gestão dos leitos acertando na capacidade conforme a demanda, e também para melhor administração dos recursos fármacos e suprimentos, infraestrutura e equipamentos clínicos, humanos e financeiros.


DATASET
THIS IS FINE.

Dataset

Após nosso primeiro contato com o dataset, percebemos que ele possui 166 features e 381.877 linhas.

Primeiro desafio foi encarar o dataset do SUS com dados reais, contendo alto volume de dados a serem tratados:

- Dados nulos, NaN, dados a serem corrigidos e excluídos;
- Análise e seleção das colunas pelo dicionário técnico que esclarece seu significado;
- Substituindo as séries do dataframe adequando com as respostas categóricas (SIM e NÃO);

Outro ponto importante foi a criação da nossa variável resposta: “**PERMANÊNCIA**”

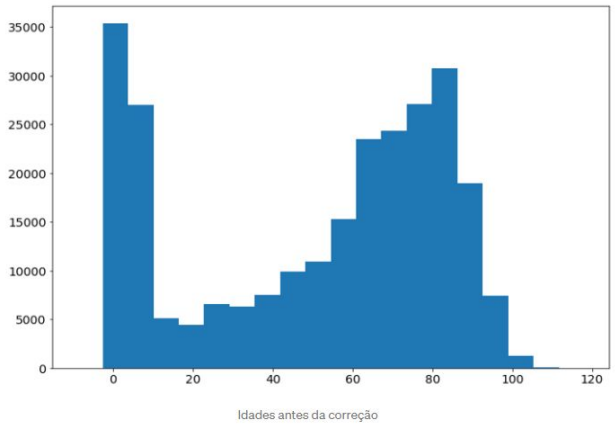
- Subtração da Data de Entrada e da Data de Saída;
- Index da Data de Internação e exclusão da data de saída.

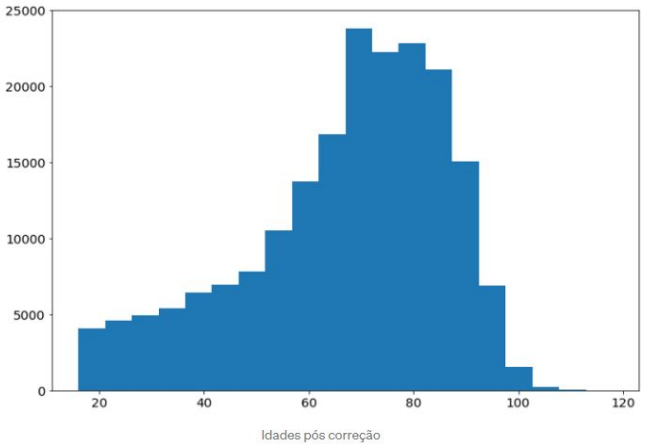
Antes de seguir para EDA, realizamos alguns outros ajustes que fizessem sentido para o nosso problema de negócio.

Dataset

Selecionando a população analisada determinando corte pela idade:

- A presença de crianças no dataset é muito alto, devido a sua imunidade e cuidado maior nas emergências porém de rápida recuperação;
- As razões das internações de menores de idade assim como seu tratamento são diferentes dos adultos, com critérios de avaliação da pediatria;
- Selecionamos a população maior de 16 anos, entendendo ela como ponto principal do impacto de negócio;

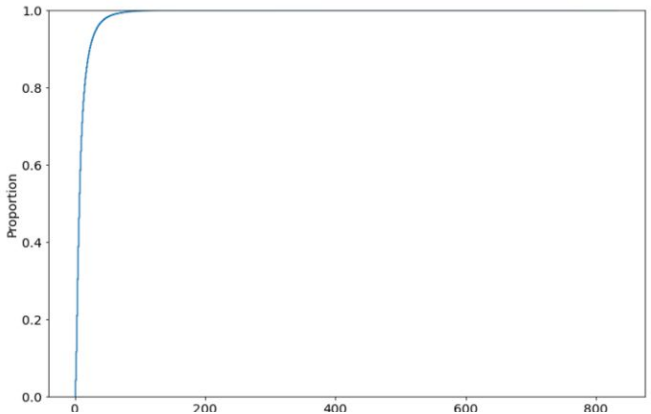


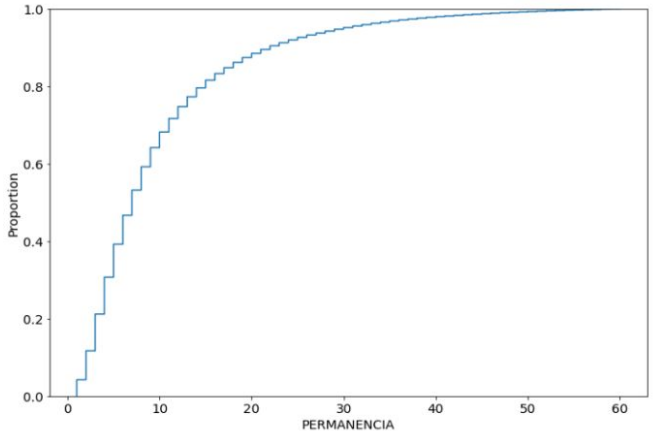


Dataset

Seleção do tempo de permanência com base na proporção de dados no dataset:

- Ajustamos com base na proporção de dados os dias de permanência sem impactar no tamanho do dataset e no modelo;
- Importante para não ser atingido com outliers desproporcionais e com baixo volume de dados;
- Escolhemos então o tempo de permanência de até 60 dias.

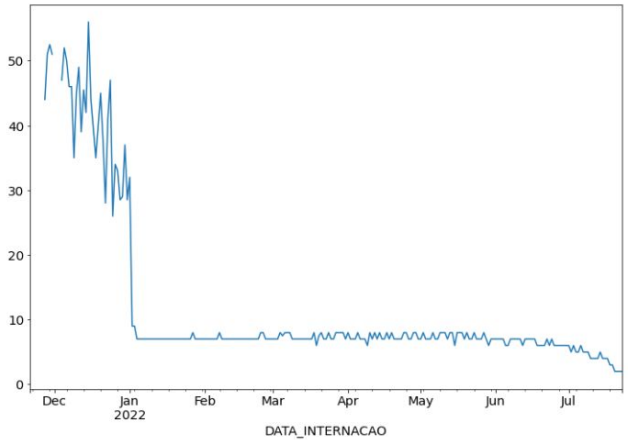


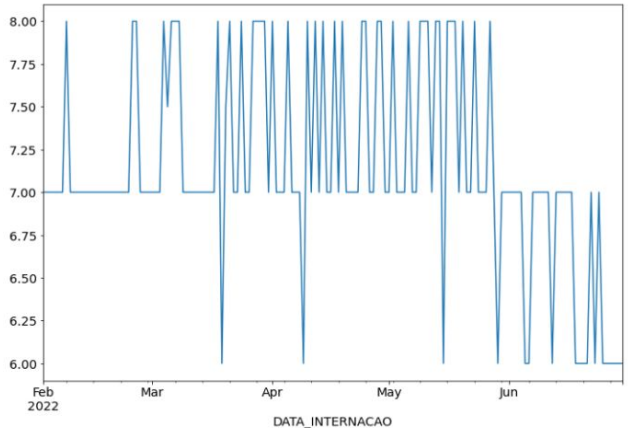


Dataset

Analisando a Influência no tempo: Fizemos uma nova análise temporal com base na data de internação (Indexada).

- Percebemos uma diferença maior na mediana contendo dados antes de 2022 (nosso foco);
- Muito se deve a um cenário com impacto da variante Omicron, diferente do atual. E com queda em julho devido a dados coletados em poucos dias;
- Decidimos dar um novo corte para o modelo, com dados de Fevereiro a Junho de 2022.





EDA

EDA

Finalmente chegamos na EDA!

A partir das nossas análises utilizando Histogramas e Boxplot obtivemos os seguintes insights:

- Analisando a distribuição da nossa target, tem como média 10 dias, mediana de 7 dias e 3º quartil de 13 dias;
- Das variáveis explicativas uma apenas é numérica (Idade) e as demais são consideradas booleanas;
- Algumas features não tiveram influência na permanência como esperávamos;
- Exemplo: asma, vacinação covid19, e os sintomas;

Percebemos porém a influência na permanência entre as features abaixo:

- Apenas saturação como sintoma se mostrou relevante para explicar o tempo maior de internação;
- Comorbidades possuem relevância e algumas delas se destacam mais, por exemplo: cardiopatia, diabetes, e renal;
- UTI, Suporte Ventilatório e Infecção Hospitalar possuem relevância pois explicam o agravamento dos casos;
- Decidimos manter Vacinação Influenza, pois teve maior relevância no tempo de permanência dos não vacinados;
- Retiramos a variável “EVOLUÇÃO” que categoriza a cura ou óbito do paciente, pelo simples fato de ser um evento futuro e assim não faz sentido utilizar para nosso problema de negócios;

MODELAGEM

```
                layer.addTo(group);
                layer.bindPopup(
                    "<p>" + "Species: " + response[i].species + "<
                    "<p>" + "Description: " + response[i].descript
                    "<p>" + "Seen at: " + response[i].latitude + "
                    "<p>" + "On: " + response[i].sighted_at + "</p
            }

            $('select').change(function() {
                species = this.value;
            });
        });
    }
$.ajax({
            url: queryURL,
            method: "GET"
        }).done(function(response) {
            for (var i = 0; i < response.length; i++) {
```

Modelagem

Chegamos agora na parte decisiva do nosso Projeto, a Modelagem!

Fizemos diversos testes e explicaremos os nossos resultados e aprendizados nessa jornada.

Regressão não linear:

- Tentamos aplicar **Decision Tree Regressor** para prever o tempo de internação, com baixos erros residuais;
- Porém, vimos que não foi possível devido a métrica do R2 Score ter como resultado negativo de -0,80;
- Ao pesquisarmos esse resultado, descobrimos que quando temos o R2 Score negativo o modelo não acompanha a tendência dos dados.

Modelagem

Classificação:

- Entendemos assim que modelos de classificação sejam mais adequados para o nosso objetivo;
- Para isso, precisamos transformar a target em intervalos do tempo de internação;
- Optamos em fazer a Classificação Binária por ser mais simples e precisa após testes realizados com multiclasse;
- Na EDA vimos uma redução e tendência de estabilização das internações, tendo representação de 75% dos casos até 13 dias;

Já que os casos não estão tão prolongados, decidimos adotar como corte o 1º quartil (25%) que são internações de até 4 dias como casos razoáveis e acima disso como casos mais críticos (não necessariamente na evolução do paciente, mas sim na sua permanência hospitalar).

Feature Engineering: Q Cut

Através do **q cut** dividimos e criamos as labels "razoável" para casos de até 4 dias, e “crítico” para casos com mais de 4 dias. Isso sendo gerado em uma nova variável resposta, atribuída como “INTERNAÇÃO”. Retiramos a antiga variável resposta “PERMANÊNCIA” após substituição da nova variável “INTERNAÇÃO”.

```
n_df = df_modelo

n_df['INTERNAÇÃO'] = pd.cut(n_df.PERMANENCIA, bins=[-np.inf, 4,
np.inf], labels=['razoavel', 'critico'])
n_df['INTERNAÇÃO'].value_counts()
```

```
critico     76265
razoavel    32690
Name: INTERNAÇÃO, dtype: int64
```

Modelo Ensembles de Classificação

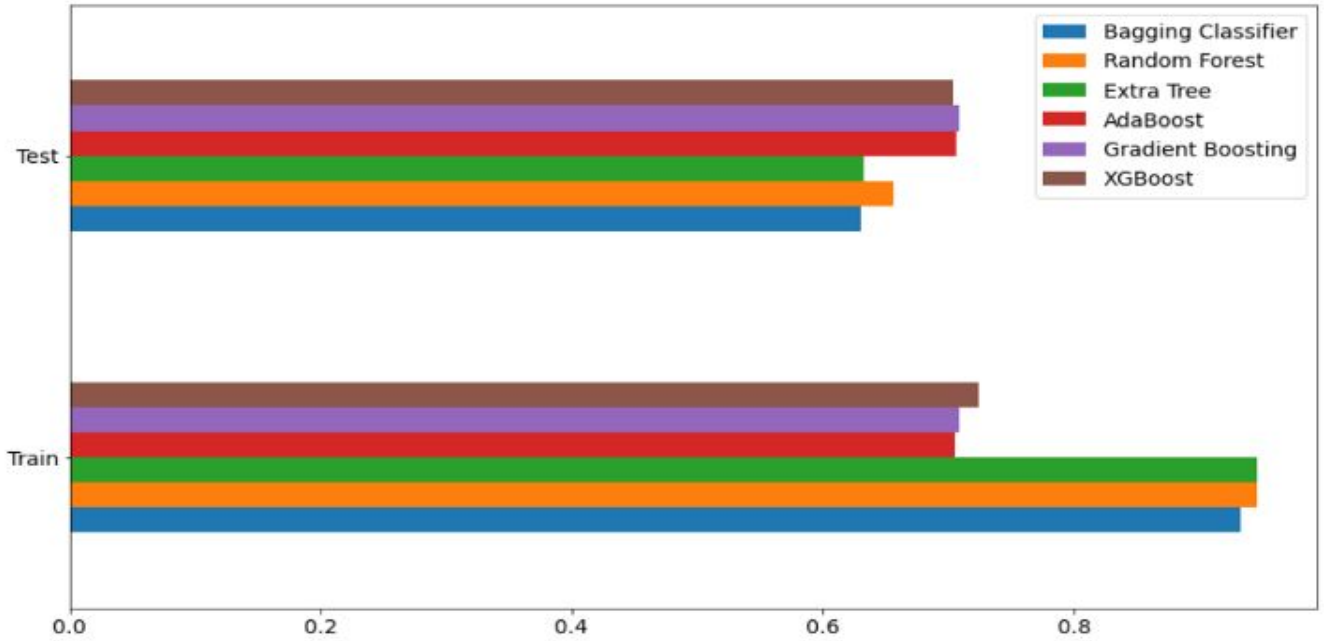
Modelos Ensembles são algoritmos de Machine Learning poderosos para melhorar o desempenho do modelo, que combina as previsões de vários modelos distintos, obtendo resultados melhores do que qualquer modelo individualmente.

Um ensemble é um grupo de classificadores (estimadores) que produzem previsões que são combinadas para produzir previsão agregada.

Com base na complexidade do dataset e do problema de negócio, escolhemos modelos de ensembles por serem mais robustos para acertar as previsões dentro do que desejamos. Abaixo a lista dos modelos:

- Bagged Decision Trees;
- Random Forest;
- Extra Trees;
- Adaboost;
- Gradient Boost;
- XGBoost.

Modelo Ensembles de Classificação



Random Forest Classifier

Comparando a performance dos modelos nos resultados de treino e teste, avaliamos que a **Random Forest Classifier** performou melhor baseada na avaliação das métricas da matriz de confusão. Escolhemos então ela como nosso modelo.

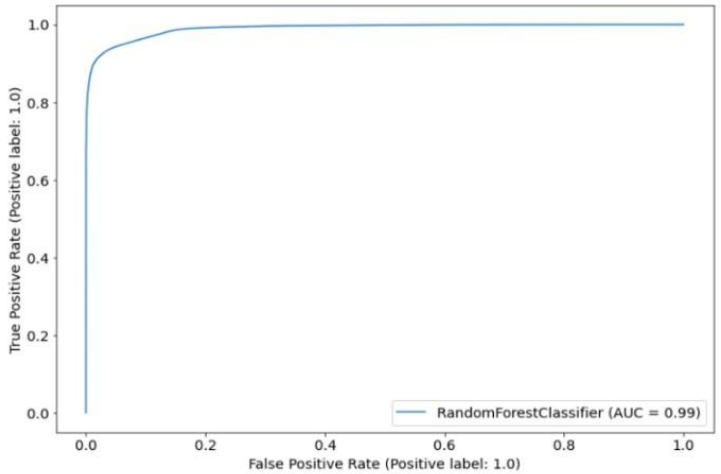
A Random Forest é como coleções (Ensembles) de árvores de decisão aleatórias. Em outras palavras, trabalha construindo múltiplas árvores de decisão durante a fase de treinamento. A decisão da maioria das árvores é escolhida como a decisão final.

A **Matriz de Confusão** podemos avaliar as métricas duas vezes, comparando conforme escolhemos a nossa classe de interesse (razoável e crítico).

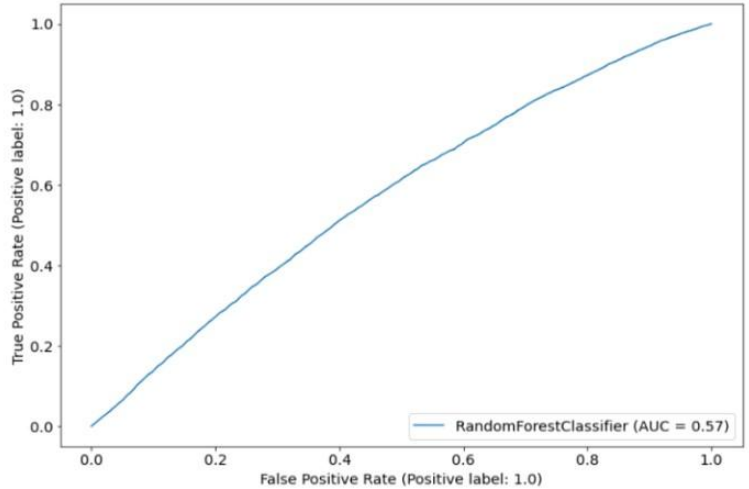
Podemos ver que o nosso modelo performa bem nos resultados de treino, porém não está indo bem nos resultados de teste com Acurácia em 65 % e com AUC em 57%.

Random Forest Classifier

```
Wall time: 11.6 s
TRAINIG RESULTS:
================================
CONFUSION MATRIX:
[[20257  2659]
 [ 1459 51893]]
ACCURACY SCORE:
0.9460
CLASSIFICATION REPORT:
                    0.0           1.0  accuracy     macro avg  weighted avg
precision      0.932815      0.951258  0.946006      0.942036      0.945716
recall         0.883968      0.972653  0.946006      0.928310      0.946006
f1-score       0.907734      0.961836  0.946006      0.934785      0.945581
support    22916.000000  53352.000000  0.946006  76268.000000  76268.000000
```




```
************************************************
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[ 2635  7139]
 [ 4085 18828]]
ACCURACY SCORE:
0.6566
CLASSIFICATION REPORT:
                    0.0           1.0  accuracy     macro avg  weighted avg
precision      0.392113      0.725074  0.656622      0.558594      0.625513
recall         0.269593      0.821717  0.656622      0.545655      0.656622
f1-score       0.319510      0.770376  0.656622      0.544943      0.635559
support    9774.000000  22913.000000  0.656622  32687.000000  32687.000000
```



Random Forest Classifier

Escolhendo a nossa **classe de interesse** como sendo os de pacientes considerados críticos que passaram mais de 4 dias internados (mais importante para nossa estratégia de negócio), temos um bom resultado nas suas métricas:

- Precision: 72% (menor impacto, pois os falsos positivos são menos prejudiciais);
- Recall: 82% (mais importante, pois os falsos negativos são mais prejudiciais);
- F1-Score: 77% (média harmônica entre as duas acima).

Porém, mesmo tendo boa performance acima, nosso modelo não está satisfatório pois “pacientes razoáveis” ele não separa bem nas previsões, classificando boa parte deles como “pacientes críticos”. Se considerarmos essa classe como a de interesse, teremos as métricas abaixo:

- Precision: 39%;
- Recall: 26%;
- F1-Score: 31%.

Balanceamento

Outro fator importante a considerar é o **balanceamento** dos dados. A label "razoável" está com menos dados do que a label “critico”. Fizemos os testes com Oversampling, Undersampling, SMOTE, e ADASYN.

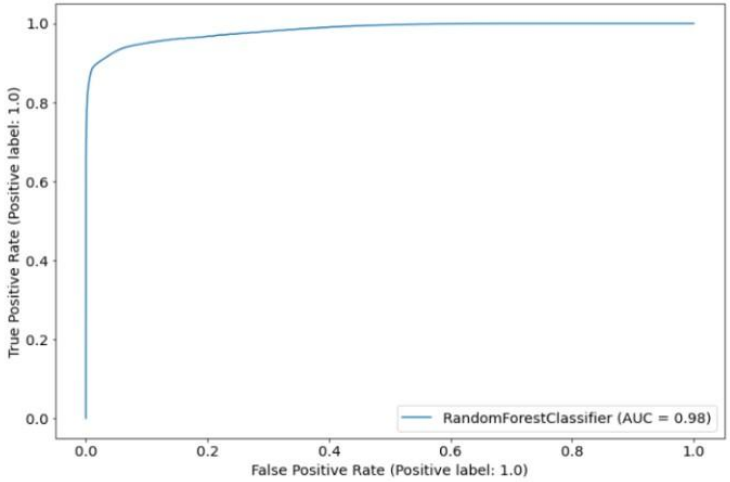
O que escolhemos foi o SMOTE com Acurácia em 64%, AUC em 58% e com uma leve melhora na predição da label "razoável" nas métricas:

- Precision: 36%;
- Recall: 39%;
- F1-Score: 38%.

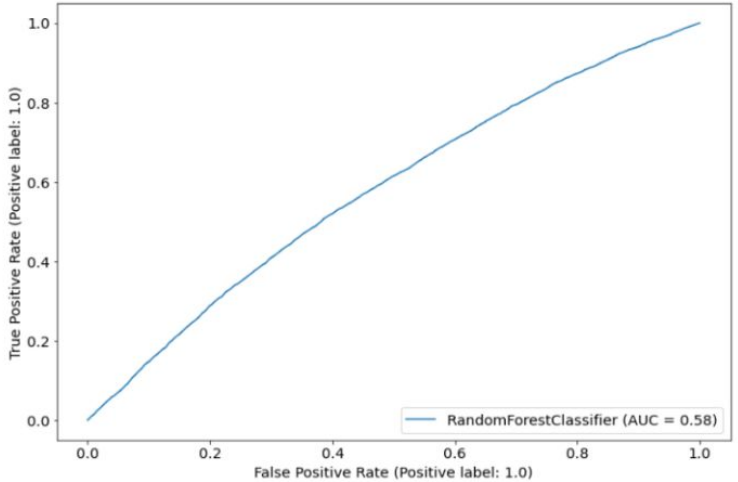
O **Synthetic Minority OverSampling Technique (SMOTE)** ou Técnica de Oversampling Sintética da Minoria, sintetiza elementos da classe minoritária baseado em elementos que já existem de forma randômica, selecionando aleatoriamente observações da classe minoritária e computando pontos através de **K-Nearest Neighbors (KNN)**, ou melhor dizendo, os K Vizinhos. Os pontos sintéticos são adicionados entre os pontos escolhidos e seus vizinhos.

Balanceamento

```
Wall time: 14.1 s
TRAINIG RESULTS:
================================
CONFUSION MATRIX:
[[51619  1733]
 [ 4615 48737]]
ACCURACY SCORE:
0.9405
CLASSIFICATION REPORT:
                    0.0           1.0  accuracy      macro avg   weighted avg
precision      0.917932      0.965663  0.940508       0.941797       0.941797
recall         0.967518      0.913499  0.940508       0.940508       0.940508
f1-score       0.942073      0.938857  0.940508       0.940465       0.940465
support    53352.000000  53352.000000  0.940508  106704.000000  106704.000000
```




```
**************************************************
TESTING RESULTS:
================================
CONFUSION MATRIX:
[[ 3852  5922]
 [ 6596 16317]]
ACCURACY SCORE:
0.6170
CLASSIFICATION REPORT:
                    0.0           1.0  accuracy     macro avg  weighted avg
precision      0.368683      0.733711  0.617034      0.551197      0.624561
recall         0.394107      0.712128  0.617034      0.553118      0.617034
f1-score       0.380971      0.722759  0.617034      0.551865      0.620558
support    9774.000000  22913.000000  0.617034  32687.000000  32687.000000
```



Conclusão

Escolhemos a **Random Forest balanceado em SMOTE** por ter sido o modelo que acertou um pouco mais nas predições dos pacientes razoáveis com métricas próximas de 40%, e mantendo o acerto nas métricas acima de 70% nos pacientes críticos, com acurácia de teste acima de 60%.

Nosso estudo não foi um case de sucesso como gostaríamos que fosse. Mas é importante ressaltar os seguintes pontos:

- Complexidade do dataset: sendo que precisou ser bem tratado e enxugado;
- Qualidade dos dados: sendo ele público e alimentado por diversas pessoas no Brasil;
- Na EDA tivemos dificuldade de encontrar padrões nas variáveis de maior impacto para nosso problema;
- Novo cenário da Pandemia: com a imunização, as comorbidades e sintomas não foram tão determinantes;
- Estabilidade do tempo de internação: com base na média e mediana, indicando redução da permanência;

E agora ?

E agora?

Podemos buscar mais estudos, seja usando outros modelos de Classificação não testados (KNN e SVM), seja ajustando os hiperparâmetros da Random Forest, que possam dar melhor poder de separação na árvore de decisão, assim como avaliar as features que são mais importantes, e mostrar aos gestores do SUS a necessidade de melhorar os inputs dos dados na coleta.

É possível também realizar outros problemas de negócio com o dataset, criando algoritmos de Machine Learning com novas abordagens de classificação. Aqui consideramos novos exemplos ao se ter os dados de triagem do paciente, tendo suas características, sintomas e comorbidades:

- Se o paciente deve ser classificado e direcionado para internar ou não;
- Se o paciente deve ser classificado e direcionado para leito de UTI ou não;
- Se o paciente deve ser classificado com diagnóstico de Covid19, ou Influenza, outros vírus gripais.


Muito obrigado!!!!
OILERS
29
@NHLGIF