Relatório do Projeto de Ciência de Dados para a Disciplina de Proficiência

Este arquivo contém informações sobre o projeto que está sendo executado (nesse caso, inferir a probabilidade de óbito de indivíduos hospitalizados com Síndrome Respiratória Aguda, utilizando machine learning). Ele é organizado de acordo com o Processo da Team Data Science (TDSP).

# 1. Entendimento do Negócio

Este é um trabalho de conclusão do curso de Pós-Graduação, portanto, o escopo não corresponde necessariamente a um projeto real de ciência de dados que aborda uma questão comercial específica. Em um projeto real, é provável que a seção de definição de problema, escopo, plano e pessoal seja muito mais detalhada, com base em discussões com o cliente (ou proprietário da empresa), a estrutura da equipe de ciência de dados etc.

## Definição do Problema

O objetivo do projeto é inferir a probabilidade de óbito de indivíduos hospitalizados com Síndrome Respiratória Aguda (SAR) utilizando ferramentas de machine learning.

O dataset para este projeto se encontra no site do Ministério da Saúde (http://plataforma.saude.gov.br/coronavirus/dados-abertos/). Foram utilizados dados dos anos de 2019 e 2020 (atualizados até 04/05/2020) referentes a Síndrome Respiratória Aguda de aproximadamente 148 mil indivíduos.

Mais informações sobre os dados baixados e tratados podem ser encontrados no repositório deste projeto no GitHub (https://github.com/tiagoacc/PosDS-sars)

## Escopo

* O escopo deste projeto é criar um modelo de aprendizado de máquina de classificação binária que resolva o problema acima.
* Utilizamos da ferramenta Pentaho com objetivo de tratar os dados, realizar a normalização das features.
* Utilizamos da linguagem de programação Python juntamente com as bibliotecas Pandas para exploração dos dados e a biblioteca scikit-learn para criação dos modelos de machine learning.
* Utilizamos o Power BI para entendimento dos dados, análise gráfica e insights.

## Planejamento

Seguimos os estágios do ciclo de vida do TDSP, a documentação sobre o trabalho e as descobertas de cada um dos estágios do ciclo de vida está incluída abaixo.

## Equipe

Este projeto foi executado pelos alunos:

DAVID COSTA TABOSA - POS19100198

IGOR KONOVALOFF LACERDA - POS19100100

TIAGO SOUSA DE ALENCAR - POS19100438

ALEXANDRE AUGUSTO DESTRO DE CAMPOS - POS19100459

## Métricas

O desempenho dos modelos de aprendizado de máquina será avaliado realizando a validação cruzada para estimar a habilidade de predição de cada modelo, serão avaliados os modelos Árvore de Decisão, Gradient Boosting e Random Forest.

# 2. Aquisição e Entendimento dos Dados

## Dados Brutos

Para informações mais detalhadas sobre os dados utilizados, acessar o Dicionário de Dados(http://plataforma.saude.gov.br/coronavirus/dados-abertos/sivep-gripe/dicionario-dados-SRAG-hospitalizado-Sivepgripe.pdf), providenciado pelo Ministério da Saúde na sua plataforma de dados abertos.

Existe um total de 148.085 registros, antes que qualquer filtro.

TARGET: Evolução - Óbito e Cura.

FEATURES:

CS\_SEXO, NU\_IDADE\_N, CS\_GESTANT, CS\_RACA, SURTO\_SG, NOSOCOMIAL, AVE\_SUINO, 35-FEBRE, 35-TOSSE, 35-GARGANTA, 35-DISPNEIA, 35-DESC\_RESP, 35-SATURACAO, 35-DIARREIA, 35-VOMITO, 35-OUTROS, 36-PUERPERA, 36-CARDIOPATI, 36-HEMATOLOGI, 36-SIND\_DOWN, 36-HEPATICA, 36-ASMA, 36-DIABETES, 36-NEUROLOGIC, 36-PNEUMOPATI, 36-IMUNODEPRE, 36-RENAL, 36-OBESIDADE, 36-OUTROS, VACINA, ANTIVIRAL, HOSPITAL, 59-POS\_IF\_FLU, 59-IF\_VSR, 59-IF\_PARA1, 59-IF\_PARA2, 59-IF\_PARA3, 59-IF\_ADENO, 59-IF\_OUTRO, 63-POS\_PCRFLU, 63-PCR\_VSR, 63-PCR\_PARA1, 63-PCR\_PARA2, 63-PCR\_PARA3, 63-PCR\_PARA4, 63-PCR\_ADENO, 63-PCR\_METAP, 63-PCR\_BOCA, 63-PCR\_RINO, 63-PCR\_OUTRO, CLASSI\_FIN, EVOLUCAO, 63-PCR\_SARS2.

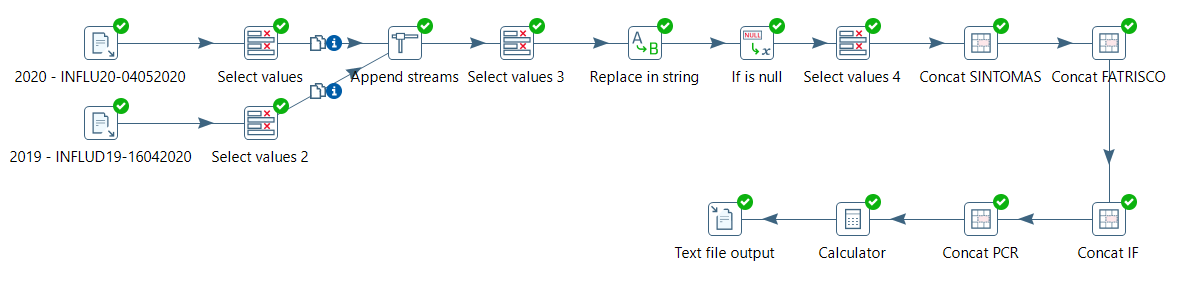
A exploração dos dados foi feita utilizando das ferramentas Pentaho, Python e Power BI.

# 3. Modeling

## Feature Engineering

### Limpeza dos dados

Utilizando o pentaho realizamos o tratamento dos dados e carga das features utilizadas no modelo, conforme a figura abaixo e explicação das etapas aplicadas.



Na primeira etapa realizamos a leitura dos arquivos CSV dos anos de 2019 e 2020 e por terem um pequena diferença nas colunas, fizemos a leitura separadamente para somente após aplicação da segunda etapa de seleção das colunas escolhidas, juntarmos para formar o dataset na terceira etapa.

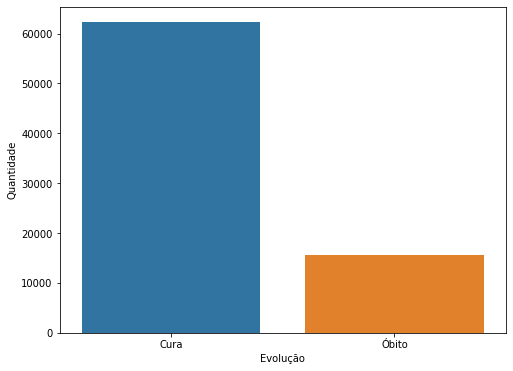
Foi realizado a etapa Replace in String e If is Null para substituir os valores originais do dataset para o padrão que escolhemos de 0 (não) e 1 (sim).

Em seguida aplicamos o concatenar para agrupar os Sintomas, Fatores de Risco (comorbidades), resultados dos exames IF (Imunofluorescência) e exames PCR, mas sem excluir as colunas originais. O objetivo foi testar se utilizando uma coluna com todos os sintomas agrupados em 0 e 1 (One Hot Encoding) seria melhor.

Por fim, geramos um novo arquivo csv, com os dados de 2019 e 2020, que foi levado para o Jupyter Notebook para uma nova análise e ajustes no dataset.

Após aplicação da análise exploratória dos dados, decidimos por excluir algumas colunas, remover os registros SRAG com situação de outro agente Etiológico ('CLASSI\_FIN'), remover os registros com valor 'Ignorado' para Sexo ('CS\_SEXO') e remoção dos registros com valor 'Ignorado' para Evolução ('EVOLUCAO').

Fizemos a avaliação do equilíbrio das classes



Quantidade de valores possíveis para classe evolução:

* Cura: 68983
* Óbito: 19745

Foram excluídos os registros com algum valor nulo.

### Label encoding

O recurso categórico CS\_SEXO foi codificado com Label Encoding usando o pré-processamento do Scikit-learn. A função LabelEncoder() foi aplicada. Portanto, as mesmas codificações são mantidas no conjunto de testes e validação.

### Padronização das features numéricas

### Dados de validação

O dataset foi embaralhado de forma aleatória retirando 20% dos dados para validação (15578)

### Over sampling

Este recurso foi codificado usando o método sample do pandas e equalizou as classes de Óbito e Cura pela quantidade de registros da maior classe, no caso Cura. Este método foi utilizado para que o algoritmo consiga resolver melhor a classe de Óbito.

Separação dos atributos (features) da classe

### Seleção do melhor modelo

Foi executada a validação cruzada para estimar a habilidade de predição de cada modelo, foram avaliados os modelos Árvore de Decisão, Gradient Boosting e Random Forest obtendo os seguintes resultados:

-------------------------------

SCORE DOS DADOS DE TREINO/TESTE , MODELO : DTC

--------------------------------

acurácia : 0.9026175910139405

roc : 0.9026175910139403

f1-score : 0.8958042708445111

recall score : 0.8372279610871527

precision score : 0.9631937233183339

[[8348 1623]

[ 319 9652]]

--------------------------------

SCORE DOS DADOS DE TREINO/TESTE , MODELO : GBC

--------------------------------

acurácia : 0.9022665730618794

roc : 0.9022665730618795

f1-score : 0.8953107374979856

recall score : 0.8358238892789088

precision score : 0.963913948646773

[[8334 1637]

[ 312 9659]]

--------------------------------

SCORE DOS DADOS DE TREINO/TESTE , MODELO : RFC

--------------------------------

acurácia : 0.9025674455922175

roc : 0.9025674455922172

f1-score : 0.8957562100971082

recall score : 0.8372279610871527

precision score : 0.9630826026765114

[[8348 1623]

[ 320 9651]]

Os resultados obtidos foram muito próximos nos deixando livres para escolha do modelo utilizado.

## Treinamento do Modelo

Após a escolha do melhor algoritmo o restante do conjunto de dados foi separado em treino e teste na proporção de 70/30, sendo 70% para treino (69797) e 30% para teste (29913).

## Avaliação do Modelo

Foram avaliados e configurados os hiperparâmetros do modelo e o score final aplicado nos dados de validação segue abaixo:

--------------------------------

SCORE DOS DADOS DE VALIDAÇÃO

MODELO : Random Forest Classifier PASSO: 0

--------------------------------

acurácia : 0.80215688791886

roc : 0.7140736382849666

f1-score : 0.8743169398907102

recall score : 0.861459337833494

precision score : 0.8875641662526909

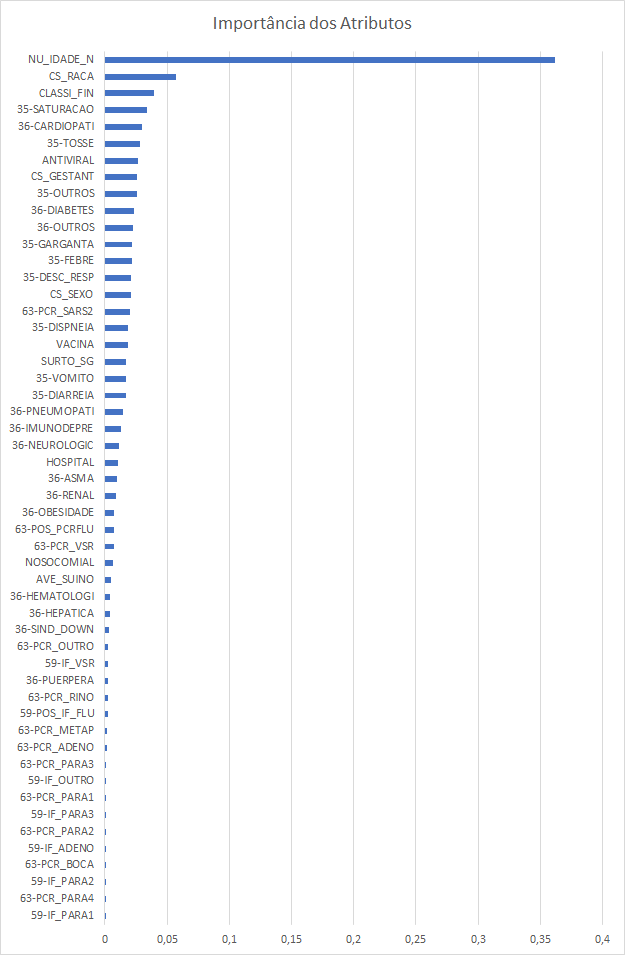
[[10720 1724]

[ 1358 1776]]

Este modelo foi considerado insuficiente para atender a resolução do problema. Os scores de recall, que considera a capacidade do modelo de discernir as classes foi abaixo do esperado. Em modelos que dizem respeito a saúde, pode haver um dano muito maior em não identificar uma doença, do que identificá-la em pacientes saudáveis. Sempre que precisarmos minimizar os falsos negativos devemos buscar uma maior pontuação nessa métrica.

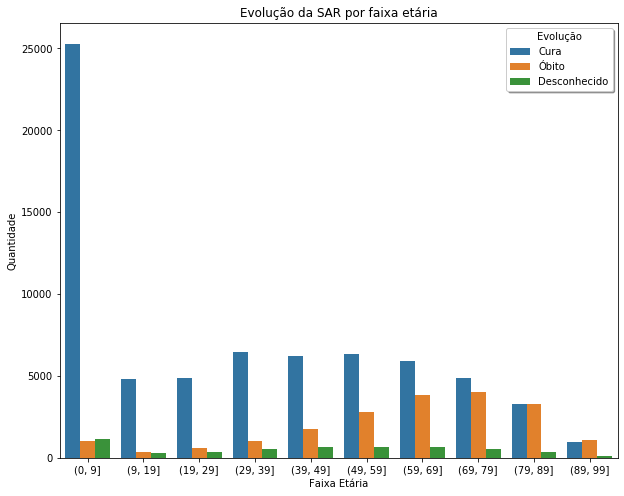
Portanto o modelo foi recusado.

A importância dos atributos do modelo Random Forest é mostrada a seguir:

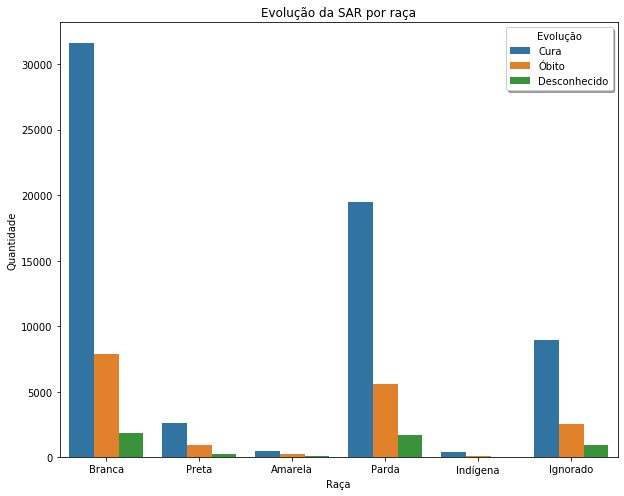


Nos gráficos a seguir, foi realizada uma análise dos atributos mais relevantes para o modelo, comparados com a evolução do quadro dos indivíduos da amostra.

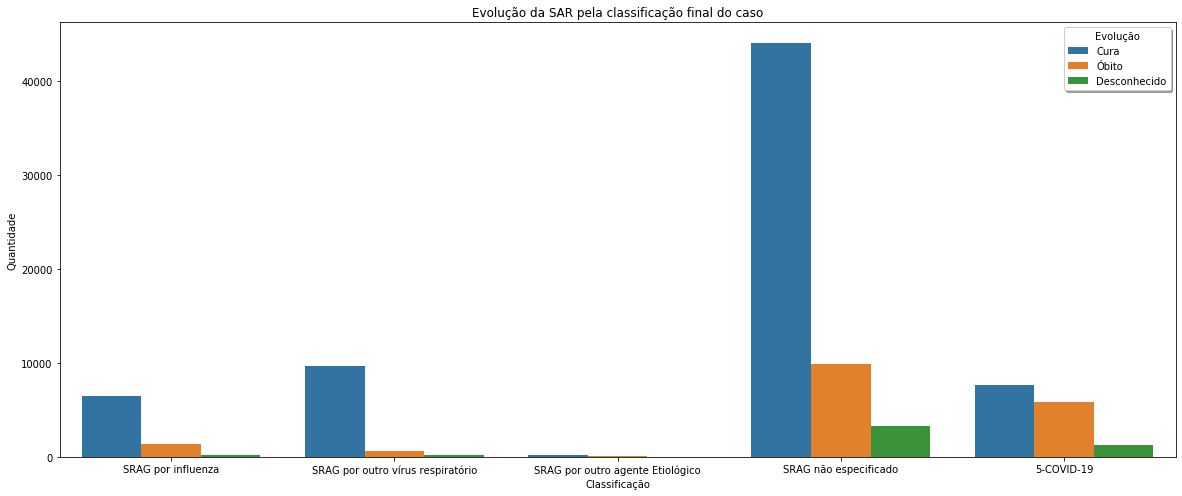
A faixa etária foi o atributo de maior importância para o modelo, no gráfico a seguir podemos observar que o aumento dos indivíduos que vieram a óbitos em relação aos curados aumenta conforme a faixa etária, nota-se que de 0 a 9 anos o índice de óbitos para os indivíduos afetados pela SAR é muito baixo, este índice mantém uma estabilidade para a faixa que abrange dos 9 ao 29 anos, a partir dos 29 anos até a faixa de 79 anos observamos uma ascendente nos casos de óbitos com uma estabilidade nos casos de cura, na faixa abrangida entre os 79-89 anos inicia-se a inversão desses índices, com a quantidade de óbitos ultrapassando as dos indivíduos que foram curados da SAR.

****

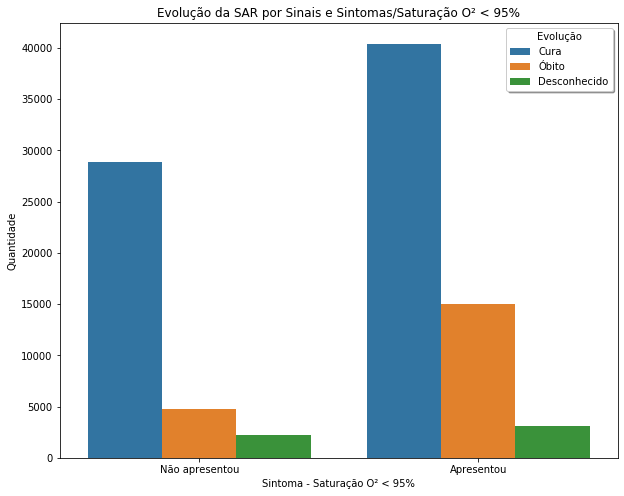
O atributo Raça mostra uma maior quantidade de afetados de cor branca tendo também o maior número de curados, o número de indivíduos declarados de cor parda apresenta uma maior relação na quantidade de óbitos por indivíduos curados comparados com os da cor branca.



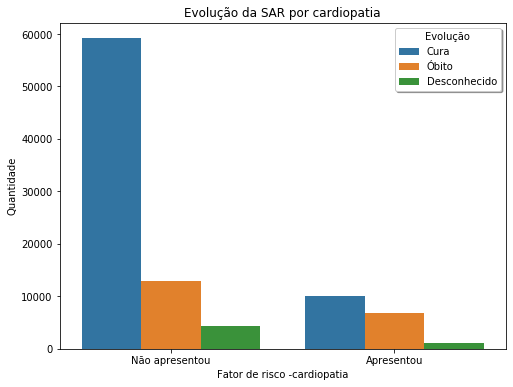
A evolução dos casos de SAR pela classificação final do caso nos mostra que a quantidade de curados é alta em relação ao número de indivíduos que vieram a óbito, destaca-se aqui os casos de COVID-19, onde a quantidade de indivíduos que vieram a óbito é quase a mesma dos indivíduos curados.



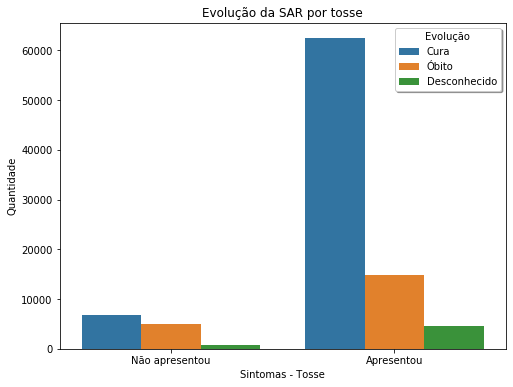
O gráfico a seguir exibe a relação do atributo “Sintoma de Saturação de O² “, entre os casos que obtiveram a mensuração da saturação abaixo dos 95% e dos que apresentaram a saturação normal, observa-se que a quantidade de óbitos, em relação aos curados, dos indivíduos que apresentação a saturação abaixo de 95% é aproximadamente 2x maior dos que apresentaram o índice normal.



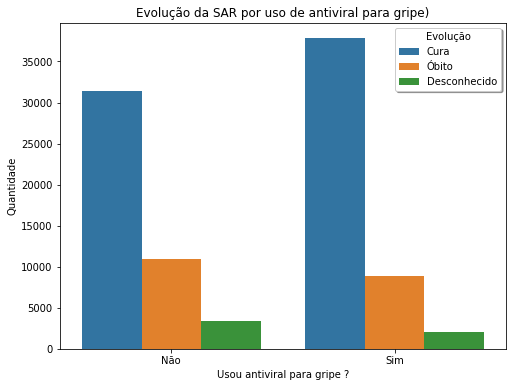
Para o atributo referente ao indivíduo possuir cardiopatia mostrado no gráfico a seguir, podemos observar que é um fator relevante a presença desta comorbidade, pois, é alta a quantidade de óbitos comparado com os indivíduos curados que apresentam este quadro.



Um dos atributos reportados é a presença de tosse, no gráfico a seguir, vemos que a grande maioria dos indivíduos que apresentaram tosse evoluíram para a cura, mas também é interessante observar que a taxa dos indivíduos que vieram a óbito, dos que não apresentaram tosse, é bastante elevada.



A utilização de antivirais para gripe, mostrado no gráfico abaixo, indica que os indivíduos que fizeram uso deste tipo de medicamento apresentaram uma melhor evolução do quadro para cura.



# 4. Deployment

Após a verificação do melhor modelo, decidimos por implementar o modelo de Árvore de Decisão, no qual será executado pelo Jupyter Notebook através de uma máquina local para geração dos resultados.

## Architecture, Environments, Code Execution

### Development

O projeto foi desenvolvido em ambiente Windows, com 6 CPU’s físicas e 12 CPU’s lógicas e com 16GB de RAM. Utiliza a versão 3.7 do python, versão 8.1 do Pentaho e a versão 2.82.5858.641 do Power BI.

Um repositório no Github foi utilizado para controlar o conteúdo deste projeto.

Para implantação, copiamos os seguintes arquivos no diretório raiz:

Arquivos csv’s compactados (input dados brutos.rar) com os dados de entrada obtidos do portal da saúde.

O arquivo de transformação do pentaho (SaudeGov.ktr).

O arquivo csv compactado com o output dos dados do tratamento do pentaho (output dado tratado pentaho.rar).

O dicionário dos dados obtidos do portal da saúde do governo em pdf.

O arquivo do power bi com a visualização dos dados.

O arquivo sars\_v4.1.py com o código do desenvolvimento do trabalho.

# Referências

https://github.com/tiagoacc/PosDS-sars

http://plataforma.saude.gov.br/coronavirus/dados-abertos/

http://plataforma.saude.gov.br/coronavirus/dados-abertos/sivep-gripe/dicionario-dados-SRAG-hospitalizado-Sivepgripe.pdf

https://github.com/Azure/Microsoft-TDSP/blob/master/Docs/lifecycle-detail.md