Olá Cientistas!

Bem-vinda e bem-vindo ao último mega desafio do Bootcamp!

Neste desafio, usaremos a base de dados da Covid-19, disponibilizada pelo Hospital Sírio Libanês - São Paulo e Brasília, no Kaggle.

Nela, encontramos diversos tipos de informações que foram separadas em 4 grupos:

1. Informação demográfica - 3 variáveis
2. Doenças pré-existentes - 9 variáveis
3. Resultados do exame de sangue - 36 variáveis
4. Sinais vitais - 6 variáveis

Sabemos que há urgência na obtenção e manipulação de dados para melhorar a previsão e assim, conseguir preparar o sistema de saúde, evitando colapsos.

**Nosso objetivo será prever quais pacientes precisarão ser admitidos na unidade de terapia intensiva e assim, definir qual a necessidade de leitos de UTI do hospital, a partir dos dados clínicos individuais disponíveis.**

Quando conseguimos definir a quantidade de leitos necessários em um determinado hospital, conseguimos evitar rupturas, visto que, caso outra pessoa procure ajuda e, eventualmente, precise de cuidados intensivos, o modelo preditivo já conseguirá detectar essa necessidade e, desta forma, a remoção e transferência deste paciente pode ser organizada antecipadamente.

Queremos que você aplique tudo o que aprendeu durante toda sua trajetória no Bootcamp e construa um modelo com as técnicas de Machine Learning que busquem a nossa variável-resposta.

Tenha em mente que este projeto será apresentado, de maneira fictícia, para o gerente responsável pela modelagem de dados do time de Data Science do Hospital Sírio Libanês. Você precisará persuadi-lo de que seu modelo tem os pontos necessários para entrar em produção e ajudará a antever e evitar qualquer ruptura.

Como a entrega é obrigatória para certificação, montamos um conjunto de critérios mínimos para avaliação que vocês poderão usar como um guia para montar seu estudo.

**Temos dois blocos a serem considerados:**

1. Técnico
2. Prático

Na seção de critérios mínimos deste projeto, você encontra quais são os aspectos que compõem estes blocos e suas respectivas descrições.

Para que o seu projeto seja avaliado pelo Thiago G. Santos e Átila Iamarino, ao vivo, na live de revisão de projetos, **submeta seu notebook ou a URL do seu projeto público no GitHub até dia 24/02 às 23h59.**

Bora mergulhar nesse projeto desafiador!

Olá Cientistas!

Aqui você encontra o conjunto de critérios mínimos que serão levados em consideração na análise do seu estudo, conforme foi mencionado na descrição deste projeto.

Mas fique tranquila e tranquilo, usem este documento como um guia para a construção da sua pesquisa e mergulhem fundo!

Os critérios foram divididos em 2 blocos lógicos:

1. Técnico
2. Prático

Sugerimos que os pontos mencionados abaixo, sejam destacados.

**1. Técnico**

**- Escopo do Projeto**

Delimitar qual será o escopo do seu projeto e colocá-lo, de fato, em prática pode ser bastante desafiador pois é um equilíbrio entre a criatividade/entusiasmo e o tempo.

Por isso, começar, desenvolver e finalizar todas as frentes abertas em um estudo é valioso.Você, cientista, precisa mais uma vez, ponderar: explorar pouco as possibilidades e ter um estudo raso ou explorar muitas possibilidades e não ser capaz de fechar dentro do elemento limitador, o tempo.

**- Estrutura do projeto**

É necessário que seu estudo seja bem organizado e estruturado, apresentando uma sequência lógica da análise.

O projeto precisa expressar e justificar qual a linha de raciocínio foi criada e seguida durante o processo de elaboração.

**- Storytelling e conclusões**

Parte da entrega de um estudo, é mostrar para a comunidade qual o seu valor, ou seja, contextualizar e trazer o interlocutor para o mesmo ponto de partida é vital.

É imprescindível que você pense que seu interlocutor, muitas vezes, não sabe do que aquele estudo se trata e/ou nem tem familiaridade com tecnologia e programação. Por isso, o notebook precisa ser explicativo de forma que a informação seja acessível para todos.

As conclusões parciais e a conclusão final são ótimos artifícios para que a informação você extraiu dos dados, seja mais facilmente entregue ao leitor (lembre-se: resultados podem ser inconclusivos, também).

**Lembrete**: este projeto será apresentado, de maneira fictícia, para o gerente responsável pela modelagem de dados do time de Data Science do Hospital Sírio Libanês. Você precisará persuadi-lo de que seu modelo tem os pontos necessários para entrar em produção e ajudará a antever e evitar qualquer ruptura.

**- Boas práticas de programação**

Parte essencial de Data Science é a construção de código fundamentada nas boas práticas de programação.

Uma boa documentação do código, nomes significativos para as variáveis, a reutilização de funções, podem ser exemplos de como colocar esse conceito em prática.

Por isso, durante a correção, será dada uma atenção especial a esse cuidado que deve ser dado ao notebook.

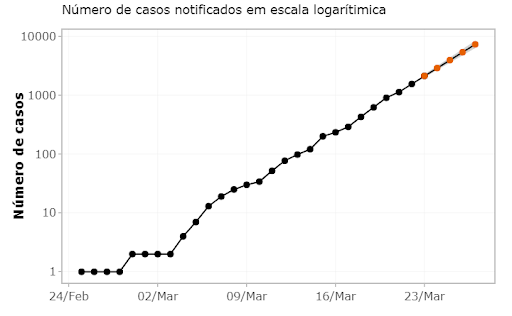
**- Gráficos e Tabelas**

A organização dos dados em gráficos e/ou tabelas é fundamental para a construção de uma boa visualização dos dados, ou seja, precisamos entender como eles estão distribuídos e se comportam ao longo do tempo.

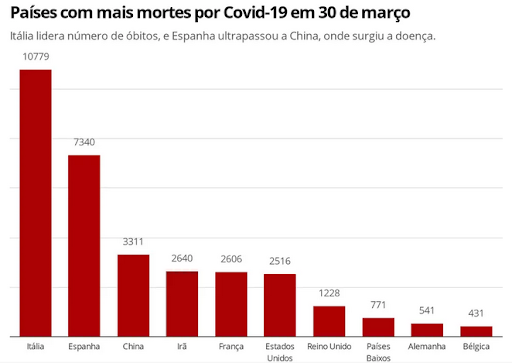
Por isso, gráficos ou tabelas completos e viáveis são indispensáveis (ex: título explicativo, labels nomeadas. No caso específico de gráficos: escala ajustada, início em (0, 0) ou caso não aconteça, apresente justificativa, por exemplo).

**Dicas para a construção de gráficos**: existem diversos tipos que podem ter diferentes aplicações. Abaixo, mostramos as principais.

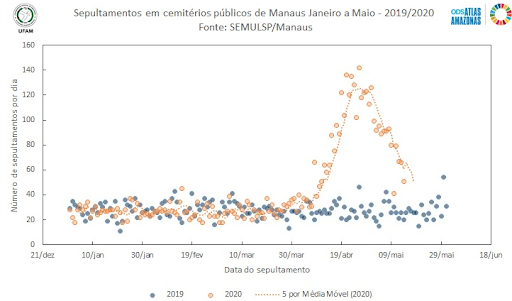
- O **gráfico de linhas** é normalmente utilizado para mostrar a evolução de algum evento ao longo do tempo, ou seja, o eixo x está, geralmente, relacionado à dados temporais e o eixo y, a dados quantitativos.



- O **gráfico de barras** é utilizado para comparar valores, sejam eles absolutos ou percentuais. Por isso, é bastante usado para comparar as categorias de uma variável.



- O **gráfico de dispersão** pode ser uma boa ferramenta para entender a qualidade da correlação entre variáveis já que, nesta visualização, é possível identificar se existe ou não relação causa-efeito entre as variáveis e o grau dessa relação.



- E o gráfico de pizza? Pizza só para acompanhar o Netflix! Brincadeiras a parte, conforme já foi discutido em alguns momentos, opte por outras visualizações.

**- Pesquisas externas e cruzamento de dados**

Do ponto de vista do estudo, é muito enriquecedor que outras fontes de informações sejam usadas para agregar valor e corroborar na construção da argumentação do projeto. E do ponto de vista técnico, isso mostra adaptabilidade e pensamento sempre um passo à frente, isso porque o cruzamento de dados é um passo muito importante no seu amadurecimento enquanto Data Scientist.

Porém, é preciso tomar bastante cuidado ao fazer essa junção: será avaliado o valor agregado à pesquisa, não somente o cruzamento em si.

**Dica**: os dados do DataSUS podem ser uma boa fonte de inspiração para os cruzamentos. Além disso, você pode expandir suas análises feitas durante os projetos do módulo 01 e módulo 04, visto que, a partir das nossas conclusões, conseguimos justificar a implantação de um modelo que visa monitorar a evolução do quadro pandêmico no Brasil que apresenta alta nos casos.

**2. Prático**

Os critérios mínimos práticos são bastante objetivos e claros, cientista. Use como um lembrete sobre o conteúdo que deve produzir.

Os dados estão dentro do escopo? (É obrigatório o uso da base de dados da Covid-19, disponibilizada pelo Hospital Sírio Libanês - São Paulo e Brasília, no Kaggle)

Ao rodar o notebook, ele apresenta erros? (Warnings serão desconsiderados)

Quando necessário, as variáveis foram tratadas?

Se houve criação de variáveis, as mesmas foram descritas?

Ficou claro qual foi o modelo final escolhido e o que motivou a escolha?

Quais testes foram aplicados? Foi justificado?

O modelo foi testado e validado adequadamente?

O notebook tem uma narrativa convincente e coerente?

O projeto contém meios para visualizar dados (gráficos ou tabelas) que ajudam na argumentação dos pontos principais do cientista?

A bibliografia e fontes de dados alternativas foram citadas?

**Não se esqueça que para sua pesquisa ser avaliada pelo Thiago G. Santos e Átila Iamarino, ao vivo, na live de revisão de projetos, você precisa submeter seu notebook ou a URL do seu projeto público no GitHub até dia 24/02 às 23h59.**

**Mergulhe fundo, é apenas o primeiro passo!**