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(a) 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 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 de avaliação 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 Paulo Vasconcellos, ao vivo, na live de revisão de projetos, submeta seu notebook ou a URL do seu projeto público no GitHub até dia 08/08 às 23h59.

Bora mergulhar nesse projeto desafiador!

Olá Cientistas!

Aqui você encontra o conjunto de critérios 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

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, pois você, cientista, precisa mais uma vez, achar o equilíbrio entre: 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 projeto seja bem organizado e estruturado, apresentando uma sequência lógica da análise.

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

- Read Me

O ponto de partida do seu projeto será o documento Read Me, já que sua entrega será feita através de um repositório do GitHub.

O Read Me funciona como um resumo geral do projeto, apresentando o contexto geral, os principais insights, as conclusões e alguns pontos de desenvolvimento futuro, por exemplo.

Dicas para um bom Read Me: - Imagens; - Nome do projeto; - Descrição do projeto; - Apresentação do objetivo do projeto; - Particularidades do projeto em evidência; - Explicação sobre a estrutura dos dados; - Bibliografia.

- Storytelling e conclusões

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

É imprescindível que você pense que seu(ua) interlocutor(a), 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).

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

Lembrete: neste projeto, nosso público é gerente responsável pela modelagem do time de Data Science do Hospital Sírio Libanês. Você precisará persuadir a pessoa de que seu modelo tem os pontos necessários para entrar em produção e ajudar a antever e evitar qualquer colapso no sistema de saúde.

- 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.

- Visualização de dados

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

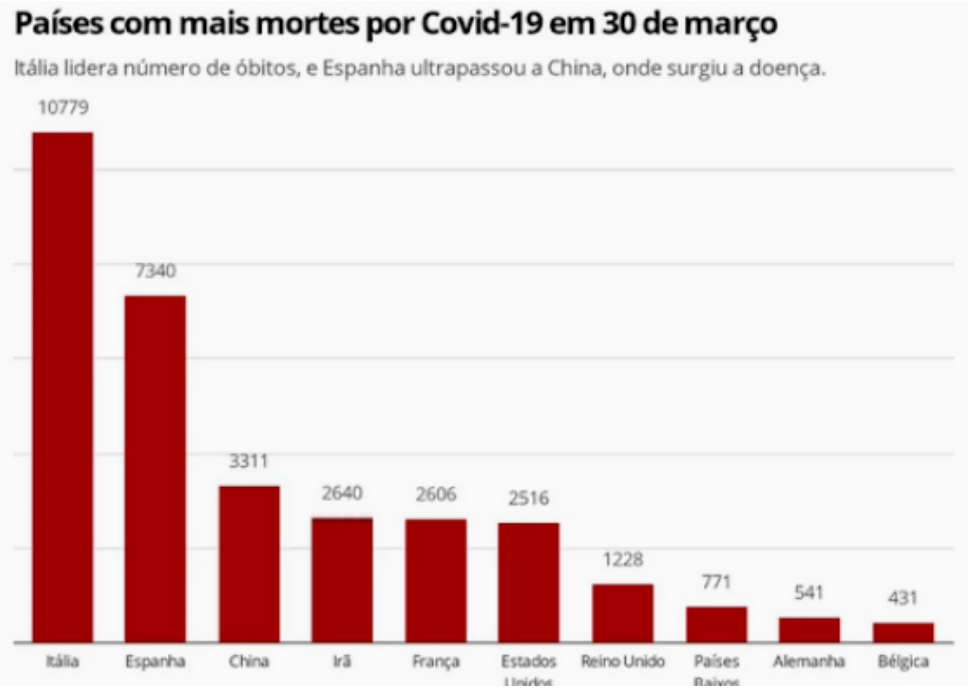
Por isso, gráficos ou tabelas completos e viáveis são indispensáveis (ex: título explicativo, labels nomeadas, no caso específico de gráficos: escala ajustada, início em (0, 0) ou caso não aconteça, apresente justificativa, etc).

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.

Gráfico de linha que apresenta o número de casos notificados de Covid-19 em escala logarítmica em função do dia (de 24/02 a 23/03). Como resultado, temos uma linha central crescente, com pontos que representam os dias.

- 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. Gráfico de barras onde cada barra representa um país e sua respectiva frequência de mortes por Covid-19 em março de 2020. As barras estão ordenadas de maneira decrescente e seguem a seguinte ordem: Itália, Espanha, China, Irã, França, Estados Unidos, Reino Unido, Países Baixos, Alemanha e Bélgica.



- 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.

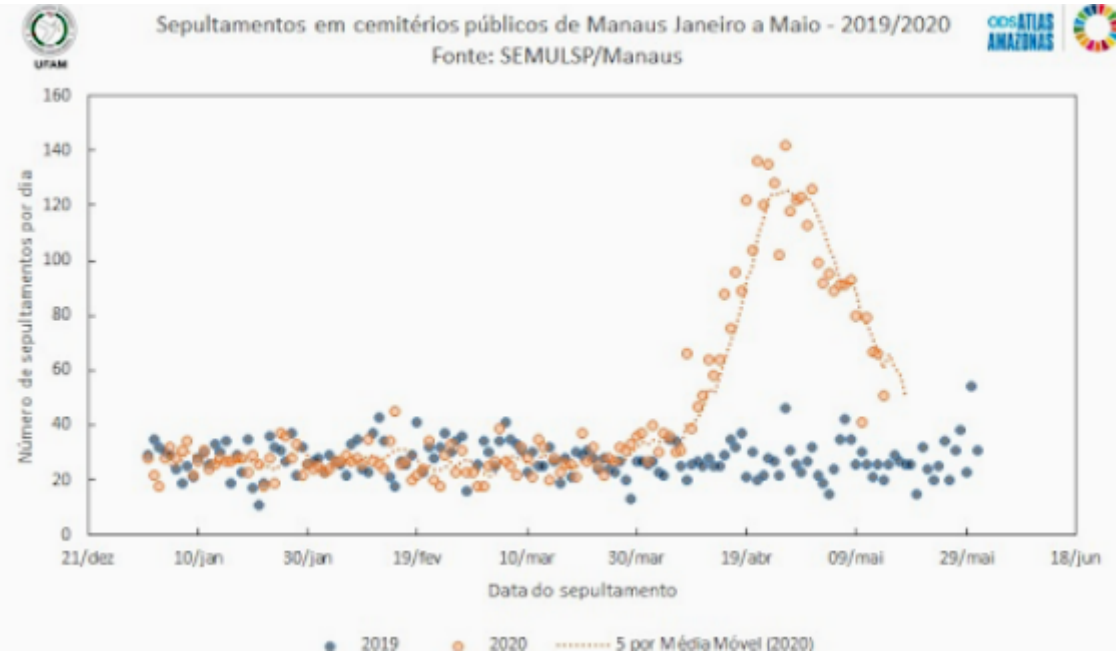


Gráfico de dispersão que apresenta os sepultamentos em cemitérios públicos de Manaus de Janeiro a Maio e compara os dados dos anos de 2019 e 2020. Na parte central do gráfico, temos pontos azuis que representam os dados de 2019 e pontos laranjas que são os dados de 2020. Podemos observar que os dados azuis mantêm uma linearidade em todo o período (21/12 a 18/06) e têm uma grandeza que flutua entre 20 e 40 mortes diárias. Já os dados de 2020, seguem a mesma linearidade e grandeza até 30/03 e então apresenta um pico onde assume maior valor próximo do dia 25/04 e então entra em declínio em direção da grandeza inicial até aproximadamente 19/05.

- 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.

- Projeto inédito

Queremos que esse projeto faça parte do seu portfólio. Para isso é importante que ele seja inédito, ou seja, que você tenha criado ele somente para o nosso curso e que você tenha citado todas as suas fontes de pesquisa, com cuidado para não realizar plágio.

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 Paulo Vasconcellos, ao vivo, na live de revisão de projetos, você precisa submeter seu notebook ou a URL do seu projeto público no GitHub até dia 08/08 às 23h59.

Mergulhe fundo, é apenas o primeiro passo!