Hype
Ciência de Dados

Dados Gerais do Processo Seletivo

ABERTURA DO CASE	18/04/2023
DATA LIMITE DE ENTREGA	09/05/2023 até 23h55
CONTATOS	Grupo no Telegram ou e-mail (hype-each@usp.br) - recomendamos utilizar o Telegram

CASE DE CIÊNCIA DE DADOS

Introdução

Olá candidato(a)! Seja muito bem-vindo(a) ao 1º Processo Seletivo (PS) do Hype de 2023! Nesta primeira fase do PS, você deverá resolver um case técnico. Você recebeu o case de acordo com seu nível de habilidade autodeclarado no formulário de inscrição, que foi o Case de Ciência de Dados.

Este case trata-se de um projeto de Machine Learning e nele será avaliada a capacidade do candidato de realizar tarefas comuns na construção de um classificador. É um projeto bastante flexível em relação aos conhecimentos do candidato, pois há várias possibilidades do que pode ser implementado para atingir um resultado adequado. Recomendamos que realize o projeto com os conhecimentos que já possui e que busque novas ferramentas para atingir um resultado ainda melhor.

É importante que você saiba que envios incompletos também serão considerados. **A nossa maior prioridade com este case é avaliar o esforço e a evolução que você teve.** Sendo assim, não será avaliado somente o resultado final do projeto. Incentivamos que entregue o projeto mesmo que não chegue em resultados satisfatórios ou esteja incompleto.

Caso tenha qualquer dúvida, você também pode perguntar para nós através do nosso grupo no Telegram chamado “Hype - PS 2023” ou chamar os monitores desse grupo no privado (os monitores são os Diretores do Hype). Você também pode tirar dúvidas através do nosso e-mail (hype-each@usp.br) - porém é preferível que envie dúvidas no Telegram, por ser mais rápido de respondermos. Ver que você está se esforçando e tirando suas dúvidas vai nos mostrar seu interesse em entrar no Hype!

Formato de Submissão

Quando terminar seu case, envie-o neste forms. Só aceitaremos uma única submissão. Caso você tenha alguma dúvida de como enviar sua resposta, não hesite em nos contatar pelo grupo do Telegram ou email (hype-each@usp.br).

A resolução deste case deve ser entregue através de um arquivo IPython Notebook (.ipynb). O notebook deve estar nomeado da seguinte forma: "analise_dados.ipynb".

A submissão (entrega) deve ser feita por meio de um único arquivo compactado zip com o nome “nusp_nome_sobrenome.zip”. Dentro desse arquivo .zip deve estar somente o arquivo do seu notebook (.ipynb). Outros arquivos de compactação como .rar, .cab, entre outros, não serão aceitos. Exemplo de estrutura do arquivo zip:

12345678_Felipe_Silva.zip

├ analise_dados.ipynb

Explicação do Case

Seu desafio neste processo seletivo será fazer um projeto de análise e predição utilizando o *dataset* com dados sobre os passageiros do Titanic. O *dataset* apresenta dados sobre as características das pessoas que estavam embarcadas no navio, assim como a informação sobre se cada pessoa sobreviveu ou não ao naufrágio.

A partir desses dados, você deve criar um modelo de classificação que tenha a capacidade de predizer se uma pessoa sobreviveria ou não a partir das variáveis dadas no *dataset*.

Para isso, você deve usar a linguagem de programação Python em alguma plataforma de Notebook (Jupyter Notebook ou Google Colab) e estruturar seu projeto em um arquivo de Notebook (.ipynb).

Sobre o Conjunto de Dados

O conjunto de dados possui 891 linhas e 12 colunas. As variáveis são:

- **PassengerId:** número único que identifica cada passageiro
- **Survived:** coluna que indica se o passageiro sobreviveu. É uma coluna de "zeros e uns", em que o zero significa que o passageiro não sobreviveu e um indica que ele sobreviveu. Essa é a coluna com as labels (classe ou variável alvo) do *dataset*.
- **Pclass**: classe social do passageiro. Pode assumir valor 1, 2 ou 3 (classe A, B e C, respectivamente)
- **Name**: nome do passageiro
- **Sex**: sexo do passageiro
- **Age**: idade do passageiro
- **Sibsp**: número de irmãos / cônjuges embarcados no Titanic
- **Parch**: número de pais / crianças embarcados no Titanic
- **Ticket**: número do ticket
- **Fare**: valor pago na passagem
- **Cabin**: número da cabine que o passageiro estava hospedado
- **Embarked**: local onde o passageiro embarcou. Pode assumir 3 valores: S, C e Q (C = Cherbourg; Q = Queenstown; S = Southampton)

O *dataset* está disponibilizado no github do Hype. É necessário que a importação do *dataset* seja feita pela função pd.read_csv(...) abaixo e que apenas esse conjunto de dados seja importado em seu projeto. Segue uma ilustração de como deve ser feita a importação do *dataset*:

```
[1] import pandas as pd

[2] df = pd.read_csv("https://raw.githubusercontent.com/hype-usp/PS-2023_1/main/Case/Avancado/data/train.csv")
```

Você pode copiar a importação a partir da linha abaixo:

```
df = pd.read_csv("https://raw.githubusercontent.com/hype-usp/PS
2023_1/main/Case/Avancado/data/train.csv")
```

Você pode trocar o nome da variável que armazena o *dataframe*, mas não a forma de importação através do link do github!

Fique atento: Qualquer outra forma de importação dos dados pode acarretar em avaliação negativa ou anulação do case.

Etapas

O projeto feito pelo candidato deve ser organizado de acordo com as seguintes etapas:

Análise dos dados

Faça uma análise exploratória dos dados buscando entender a distribuição dos dados e suas particularidades. Utilize métodos de visualização de dados, por exemplo, a partir da criação de gráficos e/ou tabelas que resumam de alguma forma as características das variáveis. Além disso, é preciso que sejam feitas interpretações sobre os gráficos/tabelas obtidos.

Preparação de dados

Nesta etapa, faça uma preparação dos dados antes de usá-los como entrada para o treinamento do modelo. Dentre as tarefas possíveis de se realizar, recomendamos que faça:

- Tratamento de dados nulos: veja se o *dataset* possui dados nulos e trate-os de alguma forma, seja excluindo as linhas que contêm esses dados ou colocando algum valor no local (mediana, média, moda etc.)
- Transformação de variáveis categóricas em numéricas: caso uma variável categórica for nominal é possível utilizar one-hot encoding, por exemplo. É possível também utilizar outros métodos para transformar variáveis categóricas em numéricas caso o modelo de escolha só aceite dados numéricos na entrada
- Normalização dos dados: talvez seja preciso normalizar os dados para ter uma maior eficácia na construção do modelo. Recomendamos que dê uma olhada no StandandScaler e no RobustScaler caso não saiba por onde começar
- Escolha de *features*/características/atributos: escolha as *features* que serão utilizadas como entrada para seu modelo e justifique a sua escolha

Construção de um modelo de classificação

Utilize algum algoritmo de classificação que seja capaz de predizer a *label* com a entrada de novos valores nas *features*. É possível utilizar dezenas de algoritmos para resolver esse problema, alguns mais complexos, outros mais simples. Recomendamos que utilize o algoritmo KNN em conjunto de algum método para escolher o hiperparâmetro K. Contudo, incentivamos que quem possui mais familiaridade com modelos de Machine Learning teste outros modelos e escolha o melhor deles (relate isso no notebook).

Avaliação do modelo

Utilize métricas de avaliação para o modelo criado na etapa anterior e analise os resultados. É necessário que o candidato utilize pelo menos duas métricas, justifique o

porquê as escolheu e conclua justificando o que o resultado obtido com essa métrica diz sobre o modelo construído. Alguns exemplos de métricas serão acurácia, f1-score, revocação (*recall*), precisão etc. Recomendamos que use o scikit-learn metrics, biblioteca que possui as principais métricas de avaliação, além de que a documentação diz um pouco sobre cada métrica, então pode ser um bom início ler um pouco dessa documentação.

Critérios de avaliação

Primeiramente, será avaliado se o(a) candidato(a) conseguiu cumprir as etapas exigidas acima e a qualidade do notebook criado. A partir disso, será avaliado se o candidato possui certos conhecimentos técnicos sobre interpretação de gráficos, pré-processamento de dados, modelos de Machine Learning, métricas de avaliação, entre outros.

Contudo, não se preocupe caso não consiga desempenho satisfatórios em relação aos resultados das métricas de avaliação do seu modelo ou caso não consiga finalizar o notebook. A nossa avaliação não será feita apenas em cima do resultado da métrica de avaliação obtida. Outros critérios como a organização do código, a lógica do projeto, a autenticidade e as premissas adotadas pelo participante também serão levados em consideração em nossa avaliação. Saiba que cases com bons resultados nas métricas de avaliação mas que não atendam esses critérios podem ser avaliados negativamente ou serem anulados. Por isso, se esforce para alcançar esses outros critérios!

Pensando nisso, é recomendável que o participante tente demonstrar os seus conhecimentos de Ciência de Dados no projeto utilizando técnicas pertinentes e explicando brevemente seu funcionamento e o motivo de seu uso. Ou seja, comente e justifique suas escolhas e seus resultados ao longo do seu notebook.

Saiba que utilizaremos um identificador de plágio, então evite copiar diretamente códigos da Internet ou copiar de outros colegas que estão participando do processo seletivo. Caso use algo da web, deixe a referência anotada no notebook. Incentivamos que os participantes façam buscas na Internet para tentar resolver os problemas que se depararem, ou para encontrar novas alternativas para melhorar seus resultados. Porém, a cópia de resoluções inteiramente prontas será vista como plágio e pode levar à reprovação no processo seletivo. Por isso, é importante que, na medida do possível, você construa sua solução (ou adapte parte de soluções encontradas), citando o que foi "encontrado" e o que foi feito por você e comente seu notebook com suas palavras!