Parabéns por ter chegado até esta etapa do processo seletivo da Solvimm! Nesta etapa teremos uma simulação de um possível caso relacionado ao dia a dia do trabalho na empresa. É só seguir as instruções abaixo no cenário proposto. Boa sorte no desafio!

Escopo do Desafio

O "Ponto Quente" tem encontrado desafios na categorização dos produtos após permitir que os usuários cadastrem produtos diretamente no seu site de vendas online.

O sistema do "Ponto Quente" busca permitir que vendedores cadastrem seus produtos diretamente no site. Nos últimos 3 meses foi constatado que os vendedores têm cadastrado a categoria dos produtos de modo errado. Por isso, o "Ponto Quente" selecionou uma amostra de dados que contém as informações do produto e das avaliações feitas pelos compradores e pediu para o seu time de inteligência de dados corrigir todas as categorias dos produtos. Com esse dataset corrigido, desejam desenvolver um modelo para classificar a categoria automaticamente de todos os produtos.

O modelo será utilizado para classificar a categoria dos produtos automaticamente, baseado nas informações que os compradores dizem nas avaliações.

O CTO do "Ponto Quente", Alan Turing, entrou em contato com a Solvimm para construir esse modelo, que usa o dataset criado pelo time de inteligência de dados.

Para treinar o modelo, o time de inteligência de dados do "Ponto Quente" forneceu a base de dados corrigida:

- Base: https://drive.google.com/file/d/1mKhb8Yd-2-aaZjB6UNWaIyLyq0cqs3LP/view?usp=sharing

Para solucionar o desafio, você precisa treinar um modelo de **machine learning** para dizer qual a categoria do produto, baseado nas informações da base de dados.

Você pode enviar dúvidas sobre o escopo de desafio para o email desafio-vagas@solvimm.com dentro do prazo para dúvidas.

ENTREGÁVEIS

- Jupyter Notebook utilizado no desafio (deve ter o nome de **desafio_estagio_em_ciencia_de_dados_solvimm.ipynb**)
    - Devemos conseguir reproduzir o código **sem qualquer tipo** de edição, criação de pastas ou baixar conjunto de dados e colocar em um local específico, **seu notebook deve se resolver sozinho.**
        - Utilizaremos o comando **RUN ALL** para testes no seu Jupyter Notebook
            - **Caso o RUN ALL não funcione, o candidato estará automaticamente eliminado**
    - Documentacão:
        - Seu próprio Jupyter Notebook deve conter a documentação das etapas que você está fazendo, utilizando o **Markdown**
            - https://www.datacamp.com/community/tutorials/markdown-in-jupyter-notebook
    - Utilizaremos a métrica **accuracy_score** para realizar a validação do seu modelo.
        - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
        - **A nota de corte do accuracy_score deste desafio é de 0.65**
    - **Crie uma função chamada validate que recebe como entrada um Pandas DataFrame com as mesmas colunas do dataset de treinamento, sem a coluna product_category.**
        - **Essa função deve realizar a classificação do dataset e retornar o DataFrame com a coluna product_category devidamente preenchida pelo modelo. Exemplo de estrutura:**

```
def validate(model, df_without_product_category):
  """
    Your CODE
  """
  return df_with_product_category
```

    - **Executaremos esta função para classificar o conjunto de testes com o seu modelo, por isso, certifique que ela está funcionando conforme o esperado antes de enviar o desafio.**

OBS:

- O desenvolvimento pode ser feito localmente no seu computador pessoal, no entanto, recomendamos o uso de alguns dos notebooks gratuitos que existem na internet:
    - Google Colaboratory: https://colab.research.google.com/
    - Kaggle Notebooks: https://www.kaggle.com/
- Você pode usar qualquer qualquer algoritmo de Machine Learning
    - Uso de **redes neurais** será considerado um diferencial

Os entregáveis devem ser enviados em um único e-mail para desafio-vagas@solvimm.com com o assunto "**[Nome do Candidato] Desafio Estágio Ciência de Dados**".

PRAZOS

- Dúvidas sobre o desafio: 4 dias
- Entrega do desafio: 7 dias