DESAFIO DO HACKATHON: ANÁLISE E PREDIÇÃO COM POSTGRESQL E H2O
Maelson M. Lima

Os participantes do hackathon terão como objetivo analisar, documentar e criar soluções preditivas a partir de uma base de dados do setor público. O desafio está dividido em quatro fases principais, cada uma focando em habilidades essenciais para o desenvolvimento de uma solução integrada e baseada em dados. A base de dados a ser utilizada será fornecida em um contêiner Docker, incluindo todas as informações necessárias para os participantes explorarem e modelarem seus casos de uso.

1. Explicabilidade da Base por Meio de um Dicionário de Dados

Os participantes deverão criar um **Dicionário de Dados** da base fornecida. O objetivo é garantir que todos compreendam os detalhes da base e possam explicá-la claramente. Eles precisarão identificar e documentar:

- Atributos de cada tabela: tipo de dado, descrição e relações com outras tabelas.
- Papel de cada campo dentro do contexto da base de dados.
- Entidades principais e como elas se conectam.

Essa etapa é essencial para garantir que todos os participantes tenham uma visão clara da base e possam explicá-la de forma compreensível para não-técnicos.

2. Analisar e Documentar a Base de Dados Fazendo o Tratamento dos Dados

Os participantes precisarão realizar uma **análise exploratória dos dados** e, posteriormente, um tratamento adequado da base. As principais tarefas incluem:

- **Limpeza dos Dados**: Identificar e tratar valores ausentes, outliers e inconsistências.
- **Normalização**: Aplicar transformações para padronizar os dados.
- **Documentação**: Descrever cada etapa do tratamento, incluindo as escolhas feitas e os motivos por trás dessas decisões.

Esta fase é fundamental para assegurar que a base esteja apta para a próxima etapa de modelagem e também para garantir a qualidade dos dados que serão usados no modelo.

3. Descrever o Caso de Uso e Elaborar um Modelo de Previsão

Nesta fase, os participantes deverão selecionar um caso de uso específico para implementar uma solução preditiva. Podem optar por:

- **Previsão de Arrecadação**: Utilizando dados de arrecadação tributária, construir um modelo que preveja a arrecadação futura, considerando fatores como região, tipo de imposto e histórico de pagamentos.
- **Previsão de Inadimplência**: Criar um modelo que estime a probabilidade de inadimplência dos contribuintes, identificando padrões e características que podem estar correlacionadas ao atraso nos pagamentos.

Os participantes deverão documentar claramente o objetivo do caso de uso, os dados escolhidos e as métricas utilizadas para avaliar a performance do modelo.

- **Obs.:** Poderão adicionar outros modelos explicativos como agrupamentos ou classificatórios para fundamentar a previsão.

4. Criar um Assistente com H2O GPT que Explique o Projeto

Como desafio final, os participantes deverão criar um **assistente virtual** utilizando o **H2O GPTe**. Este assistente deve ser capaz de responder perguntas sobre o projeto, explicando:

- **O Processo Realizado**: Cada etapa do tratamento dos dados, modelagem e resultados.
- **O Caso de Uso Implementado**: Descrevendo como os dados foram utilizados para chegar à solução preditiva.
- **Resultados e Impacto**: Explicando a precisão do modelo, os resultados obtidos e como isso poderia ser aplicado para gerar valor.

Esse assistente deve ser capaz de proporcionar uma explicação clara, tanto para técnicos quanto para pessoas sem experiência em análise de dados, focando na transparência e na compreensão dos resultados.

Critérios de Avaliação

- **Compreensão da Base de Dados**: Qualidade do dicionário de dados e explicação das entidades e atributos.
- **Tratamento e Limpeza dos Dados**: Qualidade e justificativa das técnicas de tratamento aplicadas.
- **Modelo Preditivo**: Precisão, escolha de métricas e coerência com o caso de uso descrito.
- **Capacidade de Explicação do Assistente**: Eficácia em explicar claramente o projeto, incluindo técnicas, resultados e aplicabilidade.

Ferramentas Permitidas

- **Docker** para a execução do contêiner PostgreSQL com a base de dados.
- **H2O.ai** e **H2O GPTe** para a modelagem preditiva e criação do assistente explicativo.
- Ferramentas adicionais de análise e visualização de dados, como Python ou plataformas de BI