Processo Seletivo 01/2019

IEEE Computational Intelligence Society, UnB

March 22, 2019

Instruções

1. Esse desafio têm caráter **classificatório e eliminatório** no Processo Seletivo para o CIS. Sua solução deve ser **entregue, em e-mail para ieeecisunb@gmail.com, até às 18h da quinta-feira, 28/3**.

2. Essa etapa visa avaliar as habilidades dos candidatos em manipular, visualizar e interpretar grandes volumes de dados, de forma em que **a classificação para a segunda etapa não depende obrigatoriamente de responder todas as perguntas de forma correta**. O envio é recomendado ainda que o candidato tenha desenvolvido uma solução apenas parcial, visto que todo desenvolvimento será considerado.

3. É esperado que o candidato apresente soluções originais e de própria autoria. Caso seja constatada influência de terceiros na solução apresentada, o candidato será automaticamente desclassificado.

4. Soluções enviadas após o prazo limite poderão ser desconsideradas.

5. Em caso de dúvidas, não hesite em nos contatar por meio do e-mail ieeecisunb@gmail.com

Desafio de Análise de Dados

1. **Introdução**

   Sistemas de bicicletas compartilhadas são comuns nas grandes cidades do mundo, se apresentando como uma alternativa de transporte rápida, barata e sustentável, sobretudo nas regiões centrais, onde congestionamentos e escassez de vagas são frequentes.

   Essa solução se destaca na Região Metropolitana de Los Angeles, onde o sistema de compartilhamento de bicicletas é completamente integrado com o transporte público, permitindo que os usuários que não assinam os planos mensais e anuais do *Metro Bike Share* possam realizar o pagamento de corridas avulsas por meio dos créditos disponíveis em seus *TAP cards*, cartões recarregáveis aceitos em todos os meios de transporte integrados de Los Angeles.

No entanto, o sucesso dessa integração depende da adoção de práticas inteligentes de negócio, sendo fundamental compreender o comportamento dos usuários para assegurar que o sistema conseguirá atender à demanda. Enquanto a tomada de decisões amparada pela análise de dados permitirá um funcionamento fluído do sistema, criando condições para que, por exemplo, as manutenções das bicicletas ocorram em horários de menor demanda e bicicletas sobressalientes em estações de baixa demanda sejam transferidas por operadores do sistema para estações de alta demanda antes dos horários de pico, garantindo uma boa alocação das bicicletas, o emprego de mecanismos ineficientes poderia levar o sistema ao colapso, causando cenários de escassez de bicicletas nas estações e horários mais movimentados mediante à existência de múltiplas bicicletas ociosas nas estações menos movimentadas.

Nessa etapa, você deverá buscar formas de visualização e análise dos dados do *Metro Bike Share* que sejam capazes de ilustrar o perfil de comportamento dos usuários do sistema, bem como eventualmente promover *insights* úteis.

2. **Dataset**

Para isso, utilizaremos uma versão modificada do *dataset Metro Bike Share*, contendo 14 parâmetros de 132 427 corridas realizadas entre os meses de Julho de 2016 e Abril de 2017 na cidade de Los Angeles.

Em cada corrida, são registradas as seguintes informações: *Trip ID*, um identificador único da corrida, *Day of the Week*, o dia da semana em que a corrida foi iniciada, *Month*, o mês no qual a corrida foi iniciada, *Start Date*, a data de início da corrida, *Start Time*, o horário de início da corrida [em um formato de 24h], *Duration*, a duração da corrida, em segundos, *Starting Station ID*, um identificador único da estação de onde a bicicleta foi retirada, *Starting Station Latitude* e *Starting Station Longitude*, as coordenadas geográficas da estação de onde a bicicleta foi retirada, *Ending Station ID*, o identificador único da estação para onde a bicicleta foi devolvida, *Ending Station Latitude* e *Ending Station Longitude*, as coordenadas geográficas da estação onde a bicicleta foi devolvida, *Bike ID*, um identificador único de cada bicicleta, e o *Passholder Type*, contendo a informação sobre a forma de assinatura do cliente.

3. **Análise de Dados**

Por meio dos dados fornecidos, tente responder, da forma mais completa possível, aos seguintes tópicos:

(a) Calcule a média e o desvio padrão da duração das corridas

(b) Calcule a média e o desvio padrão da distância das corridas

(c) Determine a velocidade média das corridas de distância não-nula

(d) Represente, em um gráfico de pizza, a distribuição da quantidade de corridas por tipo de assinatura

(e) Compare o comportamento dos usuários com plano mensal (Monthly Pass) com o comportamento dos usuários com plano flexível (Flex Pass) em termos de horário de utilização. Os histogramas diferem muito? Como podemos classificar a influência do plano do usuário em relação ao horário de utilização das bicicletas?

(f) Comente sobre como a quantidade de viagens varia ao longo do dia. O que pode justificar essa distribuição?

(g) Comente sobre como a distância média das viagens se distribui ao longo do dia. Como justificar esse resultado? Dica: Determine a distância média por corrida em cada hora

(h) Espera-se que bicicletas que rodem maiores distâncias tenham um maior gasto com manutenção. Podemos dizer que as viagens se distribuem de forma uniforme entre as bicicletas disponíveis? Quais bicicletas provavelmente tiveram maior gasto com manutenção? Como essas informações podem ser utilizadas reduzir o gasto com manutenções?

(i) A coluna de distância presente no *dataset* não foi obtida experimentalmente, e sim determinada a partir das coordenadas geodésicas (latitude e longitude) das estações de origem e término de cada viagem por meio da Fórmula de Haversine, que pode ser entendida uma versão do Teorema de Pitágoras para pontos na superfície de uma esfera. A partir dessa informação, determine o significado das corridas de 0km. Comente sobre as limitações dessa forma de obtenção de distância no contexto desse problema.

(j) Uma forma de mitigar essas limitações e obter alguma informação sobre a trajetória das corridas de distância nula é a estimação da distância percorrida por meio da velocidade média de cada corrida e do tempo de duração da corrida. Utilizando a velocidade média calculada no item (c), estime a distância percorrida em cada corrida de 0km. Compare a média das distâncias estimadas com a distância média das corridas de distância não-nula. Elas apresentam valores condizentes? Justifique.

Dica: As distribuições de viagens (quantidade e distância) por horário, bem como a distância total percorrida por cada bicicleta podem ser mais facilmente visualizados por meio de sua representação como histogramas.

4. **Entrega dos Resultados**
   Os dados podem ser analisados utilizando quaisquer ferramentas de sua preferência. Por exemplo:

   - Python;
     (a) Pandas;
     (b) Matplotlib;
     (c) Numpy;
     (d) Sklearn.
   - Matlab;
   - R;
   - Excel;
   - Google Sheets;
   - O que você quiser, really.

   Todos os gráficos utilizados para responder os questionamentos devem ser entregues, bem como todas as linhas de código produzidas. No intuito de facilitar a vida de vocês e agilizar nossa correção, pedimos que o relatório entregue seja preenchido de acordo com o fornecido neste link e entregue por e-mail para ieeecisunb@gmail.com até a data limite de quinta-feira, 28/3, às 18h. Gráficos ilustrativos poderão acarretar em pontos extras.