RELATÓRIO E RESPOSTAS DO DESAFIO INDICIUM

1. Faça uma análise exploratória dos dados (EDA), demonstrando as principais características entre as variáveis e apresentando algumas hipóteses relacionadas. Seja criativo!

Resposta: Ao analisar a tabela através do EDA, percebe-se que alguns valores são nulos, e os mesmos se encontram nas colunas Meta_score, Gross e Certificate. A decisão tomada nos valores faltantes da coluna Gross foi a exclusão de todas as linhas que não contenham valores, pois adicionar valores atuais em um data frame desatualizado traria inconsistência para o data frame visto que os dados de faturamento não condizem com os dados encontrados atualmente na internet. Podemos apontar mais um argumento para a exclusão, que seria a particularidade de cada filme. O filme Hamilton na 18º linha era uma peça da Broadway, posteriormente comprada pela Disney e lançado em uma plataforma de streaming, e só então lançado nos cinemas, gerando dificuldades de mensurar o faturamento do filme. Nas colunas Gross, Runtime e Release_Year foram necessários ajustes de formatação para números inteiros ao invés de objetos. Após isso, foi necessário realizar uma análise da coluna Meta_score, e a decisão mais sensata para o tratamento de dados é a de estimar uma mediana por gênero e adicionar aos dados faltantes. Para isso, separei uma tabela por gênero e calculei a mediana dos filmes para cada categoria. Não seria sensato a exclusão desses dados visto que perderia mais dados além das linhas perdidas anteriormente no campo de faturamento. Além disso, adicionei a moda dos valores faltantes na coluna Certificate.

Após a base de dados ser tratada adequadamente, dei início de fato ao EDA, onde propus retirar os seguintes insights:

- os artistas que mais aparecem
- os gêneros que mais aparecem
- os diretores que mais aparecem
- os diretores mais bem colocados por imdb_rating dentro dos diretores mais repetidos
- quais gêneros tem melhores notas
- qual a média de Runtime do top 20 de cada gênero
- a média de votos do top 20 de cada gênero
- as décadas que mais se repetem nos 100 melhores filmes

Obs: todas as respostas obtidas do EDA estão no notebook “EDA_BASE_IMDB_02.ipynb”

2. Responda também às seguintes perguntas:
    a. Qual filme você recomendaria para uma pessoa que você não conhece?

    Resposta: Com base nos resultados do EDA, o correto foi filtrar da seguinte maneira:

    - filtrar apenas filmes da década de 1990
    - filtrar apenas filmes de drama
    - necessário conter Christopher Nolan ou Steven Spielberg como diretores
    - necessário conter Robert De Niro , Tom Hanks, Al Pacino ou Brad Pitt como atores

    Após executar o código presente no tópico “PERGUNTA 2” do arquivo “EDA_BASE_IMDB_02.ipynb”, a resposta gerada foi: “Saving Private Ryan”

    b. Quais são os principais fatores que estão relacionados com alta expectativa de faturamento de um filme?

    Resposta: Para responder essa pergunta criei um novo dataframe com os 50 filmes que mais faturaram e selecionei três variáveis para análise: gênero, atores e diretores. Constatei que os gêneros aventura e ação são significativamente mais rentáveis do que os outros, também foi perceptível a repetição de alguns atores (Robert Downey Jr, Daniel Radcliffe e Rupert Grint) e diretores (Peter Jackson, Pete Docter e Anthony Russo) específicos. Não selecionei variáveis como ano de lançamento e nota de IMDB pois são variáveis que não se aplicariam à um modelo de predição futura, visto que não podemos lançar filmes em anos anteriores e não podemos selecionar as notas de IMDB e Meta Score antes de um lançamento. Variáveis como gênero, diretores e atores podem ser selecionadas no planejamento de um filme.
    Obs: código presente no tópico “PERGUNTA 2” do arquivo “EDA_BASE_IMDB_02.ipynb”

c. Quais insights podem ser tirados com a coluna *Overview?* É possível inferir o gênero do filme a partir dessa coluna?

3. Explique como você faria a previsão da **nota do imdb** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?

   Resposta: Inicialmente houve uma tentativa de usar KNN regressivo para desenvolver o modelo de ML, mas a pouca variedade de material disponível em modelos de regressão e a disponibilidade de modelos melhores para esse projeto me levaram a usar RandomForest. As variáveis escolhidas foram “Gross”, “No_of_Votes”, “Runtime”, “Released_Year”, “Meta_Score”, “Genre” e “Director”. A razão pela qual não atribuí às variáveis o nome do filme, os atores a sinopse foi a formatação em string e a alta quantidade de variáveis. Variáveis como gênero e diretor mostraram ter forte influência na nota do filme, o que validou a inclusão das mesmas no modelo. Por fim utilizei uma valideação de erro quadrático médio pois o modelo aponta os erros de maneira exponencial, uma das razões do porquê o MSE é amplamente utilizado em problemas de regressão

4. Supondo um filme com as seguintes características:

```
{'Series_Title': 'The Shawshank Redemption',
 'Released_Year': '1994',
 'Certificate': 'A',
 'Runtime': '142 min',
 'Genre': 'Drama',
 'Overview': 'Two imprisoned men bond over a number of years,
finding solace and eventual redemption through acts of common
decency.',
 'Meta_score': 80.0,
 'Director': 'Frank Darabont',
 'Star1': 'Tim Robbins',
 'Star2': 'Morgan Freeman',
 'Star3': 'Bob Gunton',
 'Star4': 'William Sadler',
 'No_of_Votes': 2343110,
```

```
'Gross': '28,341,469'}
```

Qual seria a nota do IMDB?

Resposta: de acordo com o modelo de machine learning desenvolvido ao longo do projeto, a nota do filme seria de 8.77.