



**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
CAMPUS CAMPINA GRANDE
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO
DISCIPLINA: MINERAÇÃO DE DADOS
PROFESSOR: IGOR BARBOSA DA COSTA**

**THIAGO MATHEUS HONORATO
MATHEUS CARNEIRO DA CUNHA**

RELATÓRIO FINAL

Resumo da proposta da análise:

A ideia do projeto era simples, mas bem interessante: usar os dados de avaliações de filmes para indicar, de forma personalizada, os melhores títulos para cada usuário. Basicamente, queríamos que, a partir do histórico de notas e preferências de um usuário, o sistema conseguisse prever quais filmes ele provavelmente ia curtir, veja o repositório no [github](#). Testamos dois jeitos diferentes: um baseado em filtragem colaborativa (SVD) e outro usando regressão logística para prever a probabilidade de “like”.

- **SVD (Singular Value Decomposition):** técnica de *collaborative filtering* que encontra padrões de gosto entre usuários e filmes. Com isso, consegue gerar um ranking personalizado, listando os filmes que aquele usuário provavelmente vai gostar, que nunca tenha avaliado antes.
- **Regressão Logística:** tratamos o problema como uma classificação binária (*gostou / não gostou*). A partir de variáveis derivadas das avaliações, calculamos a probabilidade de cada usuário gostar de cada filme e recomendamos os que têm maior chance de agradar. Aqui o foco não é prever a nota, mas sim mostrar diretamente quais títulos têm mais probabilidade de entrar para a lista de favoritos do usuário.
- **Rede Neural Profunda (Deep Learning) com TensorFlow/Keras:** Implementamos um modelo que recebe embeddings aprendidos de usuários e filmes, combinados com informações dos gêneros, para prever a probabilidade de um usuário gostar de determinado filme. A rede neural possui múltiplas camadas densas, permitindo capturar relações complexas entre as variáveis.

O que deu certo:

- Conseguimos processar e unir os dados, deixando tudo pronto para os modelos.
- O modelo SVD conseguiu trazer recomendações de filmes conhecidos e relevantes, que fazem sentido olhando de fora.
- A regressão logística rodou sem problemas e gerou uma probabilidade de “gostar” de cada filme.
- Comparando os três modelos deu um panorama legal de como diferentes técnicas se comportam no mesmo cenário.

O que deu errado:

- A regressão logística, apesar de funcionar, acabou recomendando muito filme obscuro e pouco conhecido, que na prática não seria tão útil para um usuário real.
- O dataset utilizado apresentava limitações como dados incompletos e pouca diversidade nas avaliações, prejudicando a qualidade final das recomendações.

O que faria diferente:

- Escolheria outro dataset para esse tipo de projeto, de preferência mais recente, mais limpo e com informações mais ricas sobre os filmes e os usuários.
- O dataset usado funcionou para testes, mas tinha algumas limitações, como muitos filmes antigos, dados incompletos e pouca diversidade de notas para certos títulos.

Aprendizados sobre mineração de dados:

- Aprendemos que o pré-processamento e a limpeza dos dados influenciam diretamente na qualidade da recomendação.
- Entendemos melhor como diferentes algoritmos podem dar respostas muito diferentes mesmo usando os mesmos dados.
- Vimos na prática que métricas de avaliação são essenciais — sem elas, a gente fica no “achismo”.
- E, talvez o mais importante: nem sempre o modelo mais sofisticado é o que dá a melhor experiência para o usuário final.