UNIVERSIDADE FEDERAL DO CEARÁ - CAMPUS CRATEÚS

RELATÓRIO:
[T3] APRENDIZADO DE MÁQUINA E PROCESSAMENTO DE LINGUAGEM NATURAL

DISCENTES:
ALYCIA ALVES ANDRADE - 536593
ANNA IWINY ALVES ANDRADE - 540134
JOSE WYTALO ADRIANO PACÍFICO - 542305
WILDNEY KESNEY RODRIGUES DE SOUSA - 535767

CRATEÚS
2025

SUMÁRIO

1. INTRODUÇÃO

O presente trabalho foi desenvolvido a partir dos conhecimentos adquiridos na disciplina de Ciência dos Dados, com o objetivo de aplicar, de forma prática, conceitos relacionados ao **Aprendizado de Máquina e ao Processamento de Linguagem Natural** (PLN). Ao longo da atividade, buscou-se consolidar a compreensão dos fluxos completos de análise de dados, desde o pré-processamento até a avaliação e comparação de diferentes modelos.

Para a realização do estudo, foi selecionada uma base de dados adequada às tarefas propostas, dessa forma foi possível estudar problemas de classificação, **clusterização** em dados tabulares, bem como a classificação de textos no contexto de PLN. A partir dessa base, foi realizada análise exploratória e experimentos, variando hiperparâmetros e utilizando métricas apropriadas para cada tipo de problema.

Além disso, o trabalho enfatiza a importância da validação adequada dos modelos, do tratamento de dados faltantes e desbalanceados, e da interpretação dos resultados obtidos, por meio de métricas quantitativas e visualizações gráficas. Dessa forma, o estudo contribui para o desenvolvimento de uma visão crítica sobre o desempenho dos modelos e suas limitações, reforçando a aplicação prática dos conteúdos abordados em sala de aula.

2. DADOS UTILIZADOS

A base de dados selecionada para a realização deste trabalho refere-se a um conjunto de títulos disponíveis na plataforma de streaming Netflix, contemplando tanto filmes quanto séries. O *data set* é intitulado como “**Netflix Movies and TV Shows**” e está presente na plataforma *Kaggle.* Além disso, as informações foram adicionadas no período dos anos de 2007 à 2021.

No que diz respeito ao seu formato, trata-se de um arquivo do tipo CSV, o que o torna adequado para análises de dados tabulares e para a aplicação de técnicas de Aprendizado de Máquina e Processamento de Linguagem Natural. Quantos aos seus atributos, há presença de dados textuais, temporais e categóricos, o que possibilita a aplicação de diferentes abordagens analíticas, como classificação, análise de tendências e exploração semântica de descrições.

São exemplos de *features* aplicadas na base de dados escolhida:

- **show_id**: identificador único atribuído a cada título;
- **type**: indica se o conteúdo é um filme ou uma série;

- **title**: nome do título disponibilizado na plataforma;
- **director**: nome do diretor ou diretores responsáveis pela obra;
- **cast**: elenco principal associado ao título;

Além de informações estruturadas, a base se destaca pela presença de atributos textuais, como *listed_in* e *description*, que permitem a aplicação de técnicas de vetorização de texto, como *Bag of Words* e *TF-IDF*, bem como a construção de modelos de classificação textual.

3. OBJETIVO

O objetivo deste trabalho é analisar um conjunto de dados referente aos títulos disponibilizados na plataforma Netflix, utilizando técnicas de Ciência dos Dados, Aprendizado de Máquina e Processamento de Linguagem Natural. A partir dessa análise, é possível entender características gerais do catálogo, identificar padrões relacionados aos tipos de conteúdo, gêneros, períodos de lançamento bem como descrições textuais. Além disso, é possível avaliar o desempenho de diferentes modelos aplicados aos dados.

4. PRINCIPAIS ACHADOS

Ainda na etapa de aprendizado de máquina, a clusterização dos dados revelou a existência de grupos distintos de títulos da Netflix, indicando que características como duração, ano de lançamento e tipo de conteúdo contribuem para a formação de padrões naturais dentro do catálogo. A utilização de métricas como o método do cotovelo e índices de avaliação de clusters auxiliou na escolha de configurações mais adequadas para os algoritmos testados, como *K-Means, Agglomerative Clustering* e DBSCAN.

No que se refere ao Processamento de Linguagem Natural, a análise das descrições textuais dos títulos demonstrou que tanto a abordagem Bag of Words quanto TF-IDF são capazes de representar adequadamente os textos para tarefas de classificação. O modelo baseado em TF-IDF apresentou maior acurácia e precisão, enquanto o Bag of Words obteve melhor equilíbrio entre *precision* e *recall*, refletido em um F1-score superior.

Considerando o desbalanceamento das classes, o modelo com Bag of Words mostrou-se mais adequado para o problema estudado, sendo escolhido como a melhor representação textual neste contexto. Esse resultado evidencia a importância de avaliar múltiplas métricas e não apenas a acurácia ao lidar com dados desbalanceados.

5. DESAFIOS E APRENDIZADOS

5.1) Desafios encontrados:

Durante o desenvolvimento do trabalho, um dos principais desafios enfrentados esteve relacionado ao pré-processamento dos dados, especialmente no tratamento de valores ausentes e na padronização de atributos categóricos e textuais. A presença de campos incompletos exigiu decisões cuidadosas para evitar a perda excessiva de informações ou a introdução de vieses nos modelos.

Outro desafio relevante foi o desbalanceamento das classes em algumas tarefas de classificação, o que impactou diretamente a interpretação das métricas de desempenho. Esse cenário exigiu uma análise mais criteriosa dos resultados, indo além da acurácia e considerando métricas como *precision, recall* e F1-score.

Além disso, a escolha e ajuste de hiperparâmetros para diferentes modelos de aprendizado de máquina representou um desafio técnico, uma vez que demandou experimentação sistemática e validação cruzada para garantir comparações justas entre os algoritmos avaliados.

5.2 Aprendizados adquiridos

A realização deste trabalho proporcionou um aprendizado significativo sobre a aplicação prática dos fluxos de Aprendizado de Máquina e Processamento de Linguagem Natural, reforçando a importância de etapas como pré-processamento, validação e avaliação de modelos.

Dessa forma foi possível entender, na prática, como diferentes modelos e representações dos dados influenciam diretamente os resultados obtidos. Isso evidencia que não há uma abordagem única ideal, mas sim soluções mais adequadas de acordo com o problema e as características do conjunto de dados.

Outro aprendizado importante foi a relevância de utilizar múltiplas métricas de avaliação, especialmente em cenários com dados desbalanceados, garantindo uma análise mais completa e confiável do desempenho dos modelos. Por fim, o trabalho contribuiu para o desenvolvimento de uma visão crítica sobre os resultados obtidos, destacando a necessidade de interpretar.