UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Luis Sávio Gomes Rosa

1. Introdução

1.1. Sobre o dataset 1:

O conjunto de dados contém várias informações que afetam as previsões, como idade, sexo, pressão arterial, níveis de colesterol, proporção de Na para potássio e, finalmente, o tipo de medicamento.

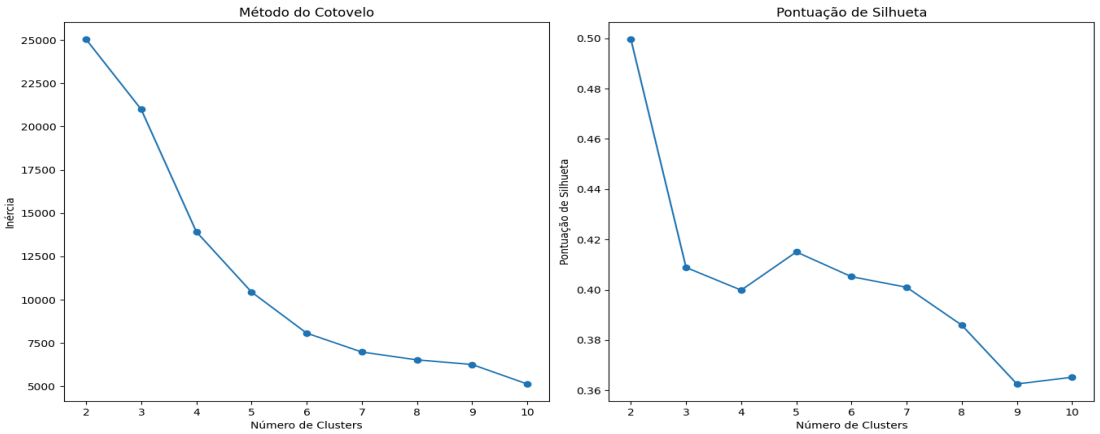
1.2. Dataset e suas variáveis:

O dataset possui 5 colunas, sendo elas:

- **Age;**
- **Sex;**
- **Blood Pressure Levels (BP);**
- **Cholesterol Levels**;
- **Na to Potassium Ration.**

Diante das informações citadas, o objetivo se concentra na aplicação do algoritmo K-Means com o intuito de retornar a label que cada ponto do conjunto de dados pertence, com base na quantidade de cluster definida.

1.3. Resultados e discussões sobre o dataset1:



A escolha de 2 clusters é sustentada pela pontuação de silhueta mais alta, o que indica que, apesar do método do cotovelo sugerir que 4 clusters poderiam ser uma opção, 2 clusters proporcionam a melhor definição de grupos. Portanto, o algoritmo escolheu 2 clusters como a opção mais robusta e significativa para o seu conjunto de dados.

O resultado de 2 clusters como o número ideal sugere que o conjunto de dados possui duas categorias ou grupos distintos que podem ser bem separados com base nas características fornecidas. O processo de determinação do número de clusters foi suportado tanto pelo método do cotovelo quanto pela pontuação de silhueta, garantindo que a divisão seja robusta e significativa.

2. Introdução ao dataset 2:

O dataset trata sobre as avaliações escritas pelos clientes, as avaliações divididas em duas partes: avaliações negativas e críticas positivas, e as avaliações são importantes para cada restaurante. O conjunto de dados consiste em 6 colunas e linhas.

Descrição das colunas:

- **Country:** O país onde o restaurante está localizado.
- **Restaurant Name:** Nome do restaurante que está sendo avaliado.
- **Sentiment:** sentimento (positivo/negativo) da avaliação.
- **Review Title:** O título ou título da revisão.
- **Review Date:** quando a revisão foi publicada.
- **Review:** Descrição do conteúdo da revisão.

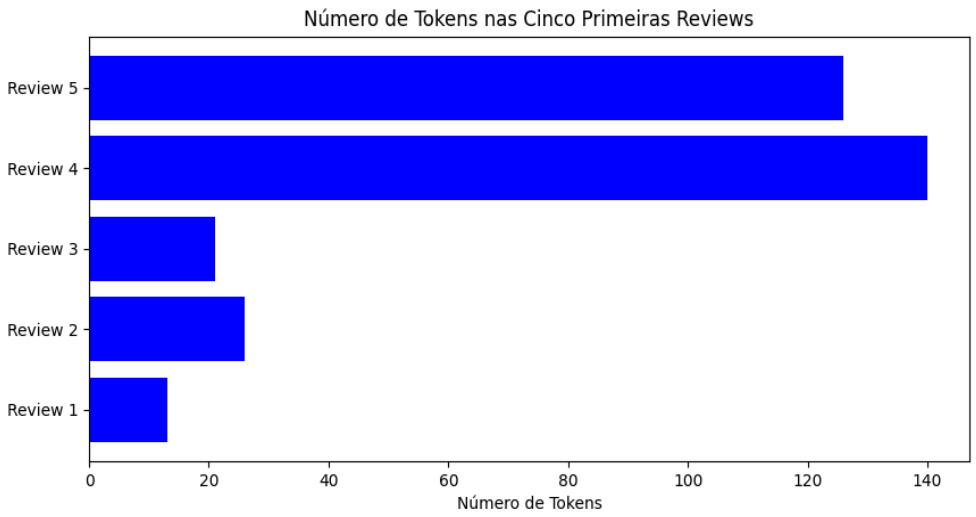
2.1. Pré Processamento:

O pré-processamento realizado na base de dados foi realizado realizando a retirada de caracteres e símbolos que não são palavras e posteriormente, foi-se retirado as “stop words”.

O objetivo é plotar as cinco primeiras listas de tokens geradas.

2.2. Resultados:

2.2.1. Gráfico:



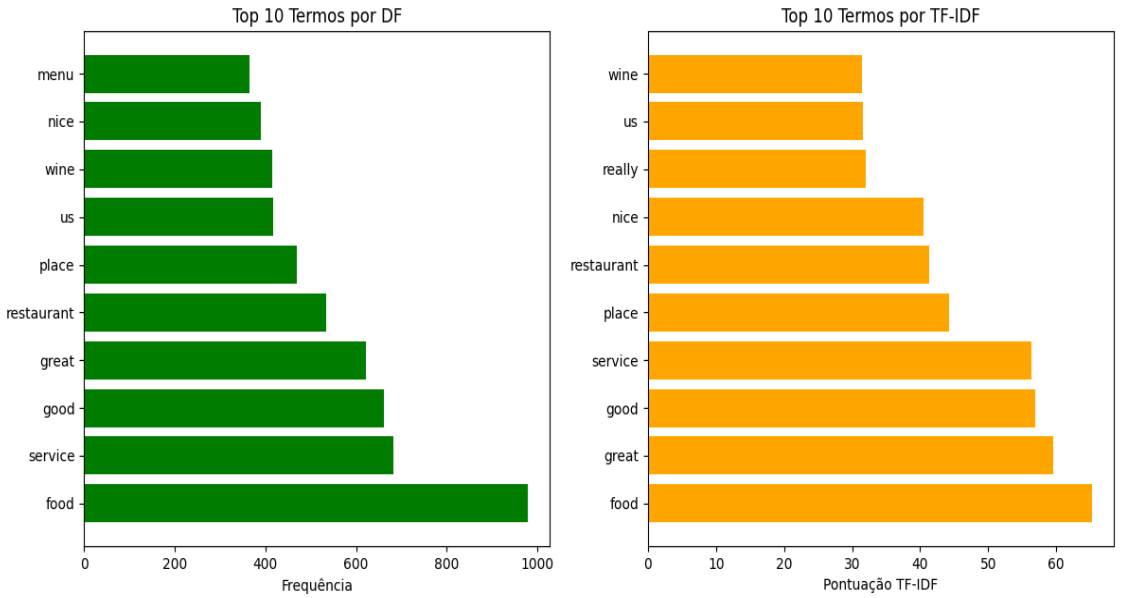
O gráfico indica uma variação significativa no comprimento das reviews, o que pode refletir diferentes níveis de detalhamento e engajamento dos revisores. Reviews mais longas tendem a fornecer mais informações, o que pode ser útil para uma análise mais aprofundada do sentimento ou do conteúdo. Por outro lado, reviews curtas podem ser mais diretas, mas podem não capturar todos os aspectos da experiência do usuário.

3. Top 10 termos com maior TF-IDF

O próximo objetivo é gerar dois conjuntos de atributos numéricos. O primeiro, baseado no DF e outro baseado no TFIDF. Plote os 10 termos com maior TF-IDF e os 10 termos com maior DF.

3.1. Resultados:

3.1.1. Gráfico de barras com 10 termos que mais se repetem:



No gráfico de Document Frequency (DF), termos como "food", "service", "good", e "great" aparecem com muita frequência nas reviews, indicando que esses são tópicos comumente mencionados pelos usuários. O termo "food" é o mais frequente, aparecendo em quase 1000 documentos, seguido por "service" e "good". Isso sugere que os usuários tendem a falar muito sobre a qualidade da comida, serviço e experiências positivas em suas avaliações.

No caso do gráfico do TD-IDF, os termos como "food", "great", "good", e "service" ainda são muito relevantes, mas a métrica TF-IDF coloca mais ênfase na importância desses termos dentro de documentos específicos. Note que termos como "really" e "nice" também aparecem entre os principais quando se usa TF-IDF, indicando que, embora possam não ser os mais frequentes em geral, eles são particularmente importantes em contextos específicos. O termo "wine", por exemplo, aparece em ambas as listas, mas sua importância relativa é mais destacada pela TF-IDF do que pelo DF simples.

4. Classificação usando os algoritmos DF e TFIDF:

O objetivo seguinte é aplicar uma classificação com o algoritmo apropriado, comparando o desempenho com as duas formas de extração de atributos implementadas.

4.1. Tabela com os resultados:

4.1.1. Classification Report CountVectorizer:

	Precision	recall	f1-score	support
Negative	0.95	0.73	0.83	49
Positive	0.95	0.99	0.97	252
accuracy			0.95	301
macro avg	0.95	0.86	0.90	301
weighted avg	0.95	0.95	0.95	301

O CountVectorizer demonstrou um desempenho superior em termos de acurácia, recall e f1-score, especialmente para a classe "Negative". A acurácia do modelo com CountVectorizer foi de 95%, enquanto com TFIDF foi de 87%. Além disso, o CountVectorizer mostrou uma maior consistência na identificação das duas classes, com uma performance equilibrada entre precisão e recall.

4.1.2. Classification Report TFIDF:

	Precision	recall	f1-score	support
Negative	1.00	0.20	0.34	49
Positive	0.87	1.00	0.93	252
accuracy			0.87	301
macro avg	0.95	0.86	0.63	301
weighted avg	0.89	0.87	0.83	301

Por outro lado, o modelo com TFIDF apresentou um recall significativamente baixo para a classe "Negative" com apenas 20%, indicando que ele teve dificuldades em identificar corretamente as instâncias negativas, resultando em um f1-score baixo para essa classe. Em contraste, a precisão para a classe "Negative" foi alta, mas isso não se refletiu em uma melhor classificação geral.

4.2. Conclusão:

Com base nos resultados dos relatórios de classificação apresentados, podemos observar que o modelo treinado utilizando o CountVectorizer demonstrou um desempenho superior em termos de acurácia, recall e f1-score, especialmente para a classe "Negative". A acurácia do modelo com CountVectorizer foi de 95%, enquanto com TFIDF foi de 87%. Além disso, o CountVectorizer mostrou uma maior consistência na identificação das duas classes, com uma performance equilibrada entre precisão e recall.

Em resumo, o CountVectorizer foi mais eficaz para este conjunto de dados e contexto específico, proporcionando uma maior precisão na classificação das duas classes e uma melhor generalização do modelo. Isso sugere que, para este cenário, a abordagem baseada em frequências simples é mais adequada do que a ponderação por frequência inversa de documentos.

5. Referências

GORTOROZYANNNN. **European Restaurant Reviews**. Disponível em: <https://www.kaggle.com/datasets/gororororororo23/european-restaurant-reviews>. Acesso em: 14 ago. 2024.

Drug Classification. Disponível em: <https://www.kaggle.com/datasets/prathamtripathi/drug-classification>.