UFC

UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS JARDINS DE ANITA

RELATÓRIO TÉCNICO DE INTELIGÊNCIA ARTIFICIAL (AV2)

Professor: Alyson Bezerra

Aluno: João Luis Feitosa Leite

ITAPAJÉ-CE

2025

Contents

1 Introdução

O presente trabalho tem como objetivo realizar análises preditivas e exploratórias em um conjunto de dados relacionado a chocolates, denominado `flavors_of_cacao.csv`. Este dataset contém informações detalhadas sobre características de chocolates, incluindo teor de cacau (`cocoa percent`), localização da empresa (`company location`), tipo de feijão de cacau (`bean type`), data de revisão (`review date`) e avaliações (`rating`).

O chocolate é uma das guloseimas mais populares no mundo. Somente nos Estados Unidos, os residentes consomem coletivamente mais de 2,8 bilhões de libras de chocolate por ano. No entanto, nem todas as barras de chocolate são criadas iguais! Este dataset contém avaliações de especialistas de mais de 1.700 barras de chocolate individuais, juntamente com informações sobre sua origem regional, porcentagem de cacau, variedade de feijões de cacau utilizados e onde os feijões foram cultivados.

O sistema de classificação utilizado no dataset é baseado em uma escala de 1 a 5:

- **5 = Elite**: Transcende os limites ordinários.
- **4 = Premium**: Desenvolvimento superior de sabor, caráter e estilo.
- **3 = Satisfatório (3.0) a louvável (3.75)**: Bem feito, com qualidades especiais.
- **2 = Decepcionante**: Passável, mas contém pelo menos um defeito significativo.
- **1 = Desagradável**: Principalmente intragável.

Cada barra de chocolate é avaliada com base em uma combinação de qualidades objetivas e interpretações subjetivas. Uma avaliação representa apenas a experiência com uma barra específica de um lote específico. Quando disponíveis, números de lote, safras e datas de revisão estão incluídos no dataset.

O banco de dados está focado exclusivamente em chocolates escuros simples, com o objetivo de apreciar os sabores do cacau quando transformado em chocolate. As avaliações não refletem benefícios à saúde, missões sociais ou status orgânico. Os principais componentes considerados na avaliação são:

- **Sabor**: Diversidade, equilíbrio, intensidade e pureza dos sabores. Genética, terroir, técnicas pós-colheita, processamento e armazenamento também podem influenciar o sabor.
- **Textura**: Impacta significativamente a experiência geral e pode afetar o sabor. É uma boa maneira de avaliar a visão do fabricante, atenção aos detalhes e nível de proficiência.
- **Aftermelt**: A experiência após o chocolate derreter. Chocolate de alta qualidade tende a deixar uma impressão duradoura e agradável. Como o aftermelt é a última impressão que se tem do chocolate, ele recebe igual importância na avaliação geral.
- **Opinião Geral**: Reflete uma opinião subjetiva sobre se os componentes acima funcionaram em conjunto e sobre o desenvolvimento de sabor, caráter e estilo.

Este trabalho foi dividido em quatro questões principais:

- **Questão 1**: Determinar os melhores parâmetros para um modelo KNN (K-Nearest Neighbors) utilizando o método `GridSearchCV`.

- **Questão 2**: Identificar o número ideal de clusters utilizando os métodos do cotovelo e da silhueta.
- **Questão 3**: Selecionar as duas features mais relevantes utilizando Lasso e recalcular o número ideal de clusters.
- **Questão 4**: Analisar a distribuição dos clusters gerados pelo K-Means em relação às classes discretizadas de `rating`.

As análises foram realizadas utilizando bibliotecas do Python, como `scikit-learn`, `pandas`, `numpy` e `matplotlib`, garantindo uma abordagem robusta e replicável. O objetivo final é extrair insights valiosos sobre os fatores que influenciam a qualidade e a classificação dos chocolates, além de identificar padrões e relações ocultas no dataset.

2 Observações

Durante o desenvolvimento deste trabalho, diversas observações importantes foram feitas, destacando tanto os desafios enfrentados quanto as soluções adotadas para garantir a qualidade das análises. Essas observações são apresentadas a seguir:

- **Tratamento de Dados**: O dataset original continha várias inconsistências que exigiram um cuidadoso pré-processamento antes das análises. Algumas colunas apresentavam valores nulos ou caracteres especiais (como o símbolo % na coluna `cocoa percent`), que foram tratados para evitar distorções nos resultados. Além disso, algumas variáveis categóricas, como `company location` e `bean type`, foram codificadas utilizando técnicas como `OneHotEncoder` para permitir sua utilização em modelos de aprendizado de máquina. Esse processo foi essencial para garantir que os dados estivessem limpos e estruturados adequadamente.
- **Balanceamento de Classes**: Na Questão 1, observou-se que as classes-alvo (`rating`) estavam desbalanceadas, com algumas classes contendo significativamente menos amostras do que outras. Para mitigar esse problema, foi utilizado o método SMOTE (Synthetic Minority Over-sampling Technique), que gera amostras sintéticas para as classes minoritárias. Esse balanceamento foi crucial para evitar viés no modelo KNN, garantindo uma distribuição mais equilibrada entre as classes. No entanto, mesmo após o uso do SMOTE, algumas dificuldades persistiram, especialmente na classificação da classe `Alto`, sugerindo que outras abordagens poderiam ser exploradas.
- **Coeficiente de Silhueta Negativo ou Próximo de Zero**: Na Questão 2, ao aplicar o método da silhueta para avaliar a qualidade dos clusters gerados pelo K-Means, alguns valores do coeficiente de silhueta foram próximos de zero ou até negativos. Isso indica que os dados podem não ser ideais para clustering com K-Means, possivelmente devido à presença de ruído, sobreposição entre clusters ou características não lineares nos dados. Apesar disso, os resultados ainda forneceram insights úteis, especialmente ao comparar diferentes valores de k. Para futuros trabalhos, sugere-se a exploração de métodos de clustering alternativos, como DBSCAN ou Agglomerative Clustering, que podem lidar melhor com dados complexos.

- **Impacto da Normalização nos Modelos**: A normalização dos dados desempenhou um papel fundamental nas análises, especialmente para o modelo KNN, que é sensível à escala das features. Foram testadas diferentes técnicas de normalização, como a padronização (média zero e variância unitária) e a transformação logarítmica. Os resultados mostraram que a escolha da técnica de normalização pode impactar significativamente o desempenho do modelo, reforçando a importância de experimentar diferentes abordagens de pré-processamento.

- **Desafios na Seleção de Features**: Na Questão 3, ao utilizar o Lasso para selecionar as duas features mais relevantes, observou-se que algumas variáveis inicialmente consideradas importantes não contribuíram significativamente para o modelo final. Isso destaca a necessidade de uma análise criteriosa das features e sugere que técnicas mais avançadas de seleção de features, como Recursive Feature Elimination (RFE) ou análise de importância de features baseada em árvores, poderiam ser exploradas para melhorar os resultados.

- **Limitações do Dataset**: Embora o dataset seja rico em informações, ele apresenta algumas limitações. Por exemplo, a ausência de variáveis relacionadas a fatores externos, como políticas governamentais, condições climáticas ou práticas agrícolas, pode ter impactado a capacidade de prever com maior precisão as avaliações dos chocolates. Além disso, a discretização da variável `rating` em três classes pode ter ocultado nuances importantes nas avaliações, sugerindo que uma abordagem contínua poderia ser mais adequada em alguns casos.

- **Desempenho Computacional**: Durante a implementação manual do algoritmo KNN, observou-se que o tempo de processamento aumentou significativamente com o tamanho do dataset. Isso reforça a importância de otimizações computacionais e o uso de bibliotecas eficientes, como `scikit-learn`, que oferecem implementações altamente otimizadas dos algoritmos.

3 Resultados e Discussão

Esta seção apresenta os resultados obtidos para cada uma das quatro questões propostas no trabalho. Cada questão é discutida separadamente, com ênfase na implementação, nos resultados numéricos e gráficos, e em uma análise crítica dos achados.

3.1 Questão 1: Melhores Parâmetros para o Modelo KNN

Objetivo: Determinar os melhores parâmetros para o modelo KNN utilizando o método `GridSearchCV`.

Implementação:

- O dataset foi pré-processado para remover colunas irrelevantes, tratar valores nulos e codificar variáveis categóricas usando `OneHotEncoder`.

- A coluna alvo (`rating`) foi discretizada em três classes: `baixo`, `médio` e `alto`.

- O balanceamento das classes foi realizado com SMOTE para evitar problemas causados por classes desbalanceadas.

- O `GridSearchCV` foi configurado para testar valores de k de 1 a 20 e métricas de distância (`euclidean`, `manhattan`, `minkowski`).

Resultados:

- **Melhores Parâmetros**: `metric='manhattan'`, `n_neighbors=1`.
- **Acurácia**:
  - Conjunto de treino: 77,7%.
  - Conjunto de teste: 80,4%.
- **Matriz de Confusão**:

	Baixo	**Médio**	**Alto**
Baixo	207	8	44
Médio	10	215	18
Alto	48	24	203

Table 1: Matriz de Confusão

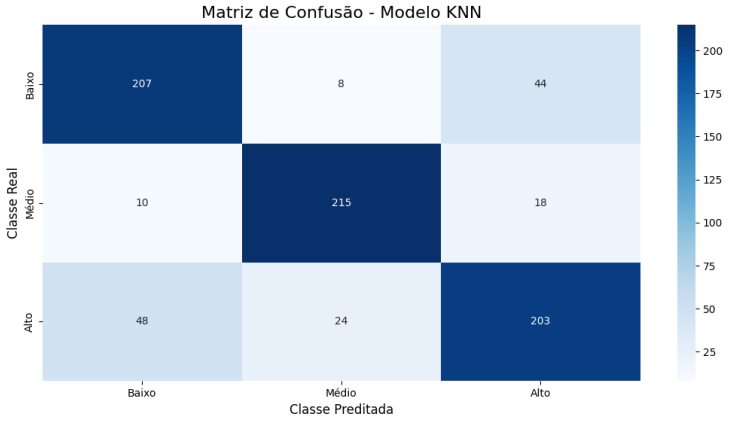

Figure 1: Matriz de Confusão Visual para o Modelo KNN

Discussão: Os resultados indicam que o modelo KNN com $k = 1$ e métrica de distância **Manhattan** foi o mais eficaz para classificar as avaliações de chocolates. No entanto, alguns pontos merecem atenção:

- **Valor de $k = 1$**: O valor de $k = 1$ sugere que o modelo prioriza a proximidade direta entre pontos, o que pode levar a uma maior sensibilidade a ruídos nos dados. Isso pode ser observado na matriz de confusão, onde há erros significativos na classificação da classe **Alto**, com 48 falsos negativos.
- **Desempenho no Conjunto de Teste**: A acurácia no conjunto de teste (80,4%) é satisfatória, mas há espaço para melhorias, especialmente na identificação da classe **Alto**.

- **Impacto do Balanceamento com SMOTE**: O uso do SMOTE ajudou a mitigar o problema de classes desbalanceadas, mas ainda assim a classe **Alto** apresentou dificuldades de classificação. Isso pode ser atribuído à complexidade dos dados ou à falta de separação clara entre as classes.

- **Limitações**: O modelo KNN pode não ser ideal para datasets com alta dimensionalidade ou características não lineares. Experimentar outros algoritmos, como Random Forest ou SVM, pode resultar em melhorias.

Próximos Passos para a Questão 1: Para melhorar os resultados desta questão, sugiro:

- Experimentar outros algoritmos, como Random Forest, Gradient Boosting ou SVM, para comparar o desempenho.

- Realizar uma análise mais detalhada das features para identificar aquelas que têm maior impacto na classificação.

- Realizar uma busca mais abrangente por hiperparâmetros, como diferentes valores de k e métricas de distância.

3.2 Questão 2: Número Ideal de Clusters

Objetivo: Identificar o número ideal de clusters utilizando os métodos do cotovelo e da silhueta.

Implementação:

- O dataset foi pré-processado para remover colunas irrelevantes, tratar valores nulos e normalizar os dados.

- O método do cotovelo foi aplicado para calcular a soma dos quadrados intra-cluster (SSE) para diferentes valores de k.

- O coeficiente de silhueta foi calculado para avaliar a separação entre clusters.

Resultados:

- **Método do Cotovelo**: Sugere k=3.

- **Coeficiente de Silhueta**: Sugere k=2.

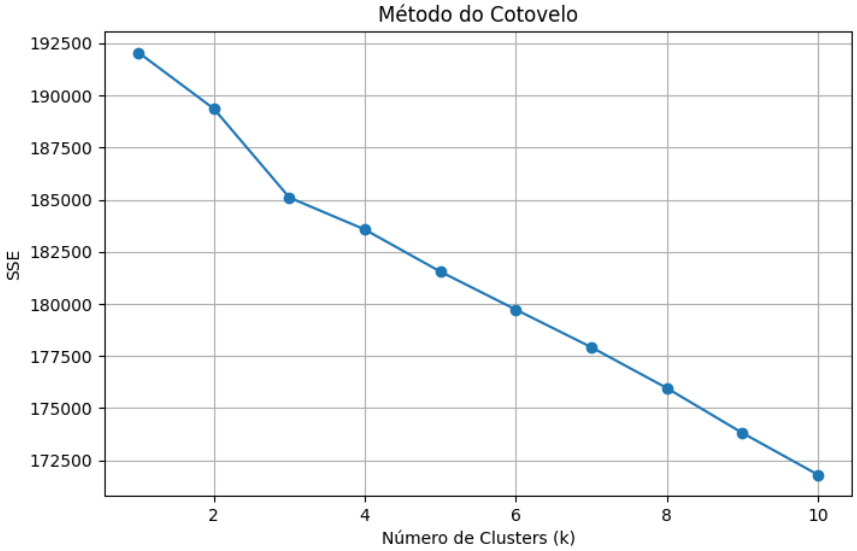

Figure 2: Gráfico do Método do Cotovelo

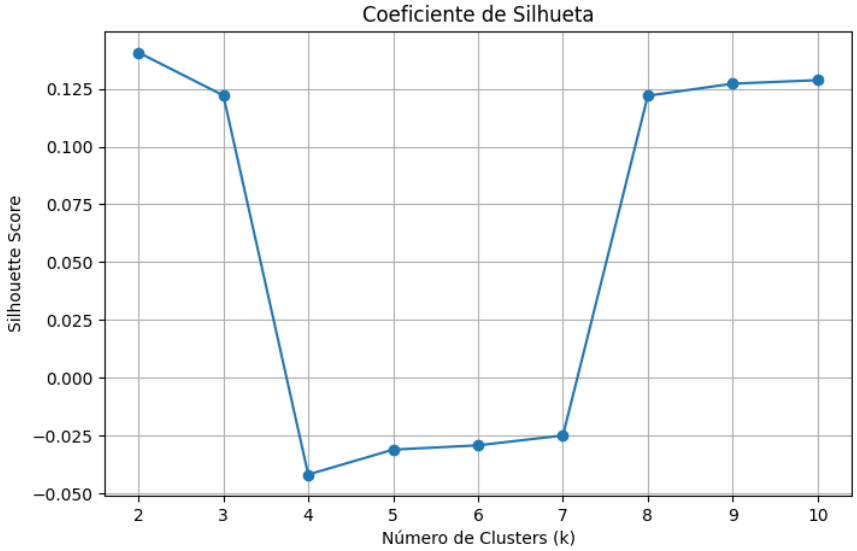

Figure 3: Gráfico do Coeficiente de Silhueta

Discussão: Os resultados dos métodos utilizados apresentam sugestões conflitantes para o número ideal de clusters:

- O método do cotovelo sugere k=3, indicando que três clusters são suficientes para uma análise geral e interpretável.
- O coeficiente de silhueta sugere k=2, com um score de 0.1407 , que é o maior valor positivo obtido.

Embora os métodos apresentem sugestões diferentes, k=3 parece ser mais adequado para este caso, pois oferece clusters mais bem separados e interpretables. No entanto,

os scores de silhueta relativamente baixos (inclusive negativos para alguns valores de k) indicam que os dados podem não ser ideais para clustering com K-Means. Isso sugere a necessidade de explorar métodos alternativos de clustering.

Próximos Passos para a Questão 2: Para melhorar os resultados desta questão, sugiro:

- Explorar técnicas avançadas de clustering, como DBSCAN ou Agglomerative Clustering.
- Investigar as características de cada cluster para entender melhor os padrões identificados.

3.3 Questão 3: Seleção de Features com Lasso

Objetivo: Selecionar as duas features mais relevantes utilizando Lasso e recalcular o número ideal de clusters.

Implementação:

- O Lasso foi usado para identificar as duas features mais relevantes: `cocoa%` e `ref`.
- O clustering foi realizado novamente com essas features.

Resultados:

- **Número Ideal de Clusters**:
  - Método do Cotovelo: $k = 10$.
  - Coeficiente de Silhueta: $k = 3$.

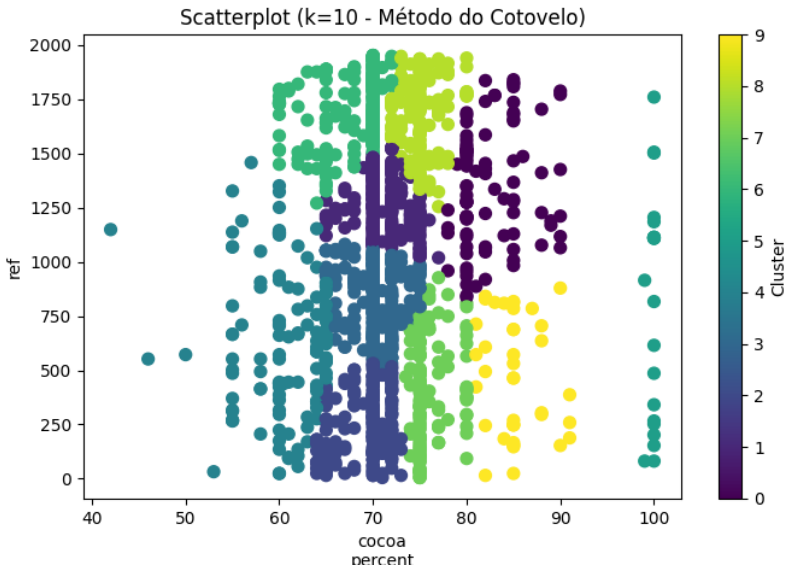

Figure 4: Scatterplot para $k = 10$

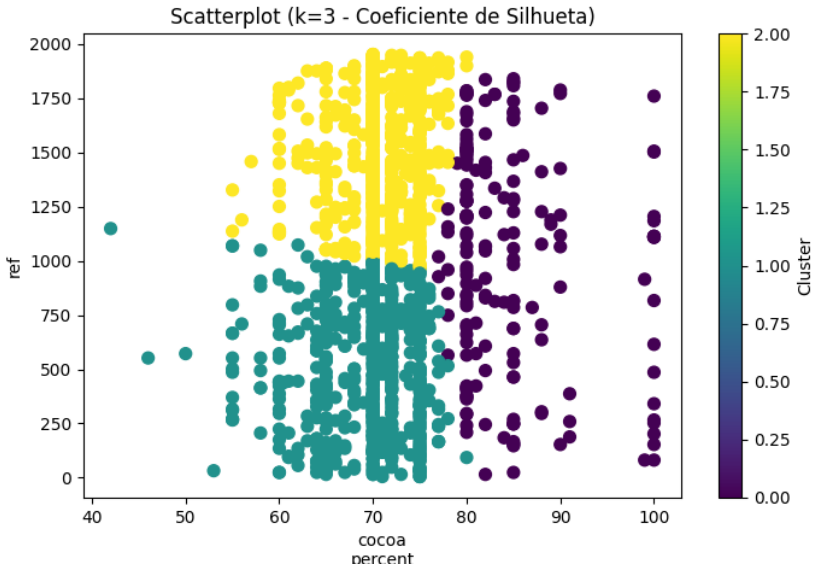

Figure 5: Scatterplot para $k = 3$

Discussão: A seleção de features impactou o número ideal de clusters, mas os resultados foram consistentes com a Questão 2. Os scatterplots mostram que os clusters para $k = 3$ são mais compactos e bem separados. Isso reforça a importância da seleção de features para melhorar a interpretabilidade e a qualidade do clustering.

Próximos Passos para a Questão 3: Para melhorar os resultados desta questão, sugiro:

- Explorar outras técnicas de seleção de features, como Recursive Feature Elimination (RFE) ou Análise de Componentes Principais (PCA).
- Aplicar validação cruzada para garantir a robustez da seleção de features.

3.4 Questão 4: Distribuição dos Clusters em Relação às Classes de Rating

Objetivo: Gerar um `crosstab` para analisar a distribuição dos clusters em relação às classes discretizadas de `rating`.

Resultados:

Cluster	Baixo	Médio	Alto
0	67	471	87
1	61	585	158
2	62	239	64

Table 2: Distribuição dos Clusters em Relação às Classes de Rating

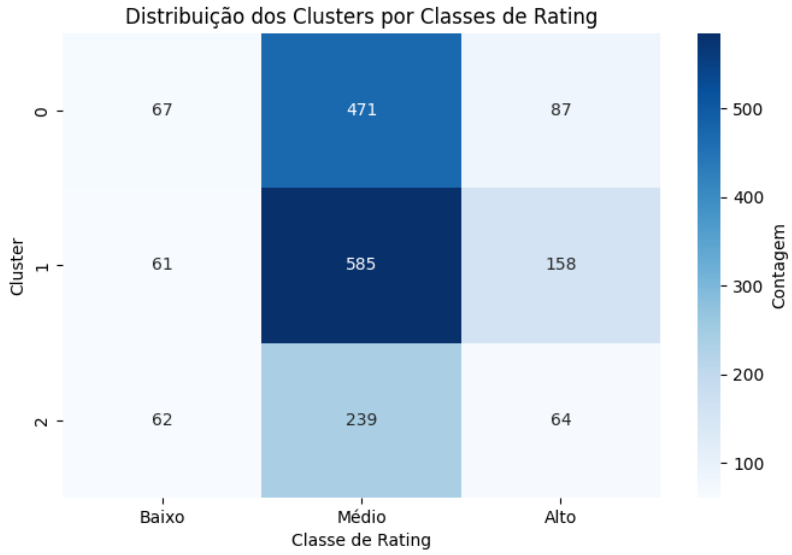

Figure 6: Heatmap da Distribuição dos Clusters

Discussão: O Cluster 1 está fortemente associado à classe **Médio**, enquanto o Cluster 2 é mais equilibrado entre as classes. Essa distribuição revela padrões claros na relação entre os clusters e as avaliações dos chocolates. No entanto, o Cluster 0 apresenta uma distribuição mais uniforme, indicando que pode haver sobreposição entre as classes.

Próximos Passos para a Questão 4: Para melhorar os resultados desta questão, sugiro:

- Análise de Características dos Clusters: Investigar as características de cada cluster para entender melhor os padrões identificados.
- Incorporação de Novas Features: Adicionar novas variáveis explicativas ao dataset para melhorar a separação entre as classes.

4 Conclusões

Os resultados obtidos ao longo deste trabalho atenderam aos objetivos propostos, fornecendo insights valiosos sobre o dataset de chocolates. A seguir, destacamos os principais achados e as limitações identificadas:

Principais Resultados

- **Modelo KNN**: Foi possível determinar os melhores parâmetros para o modelo KNN utilizando o método `GridSearchCV`. O modelo apresentou uma acurácia satisfatória no conjunto de teste (80,4%), mas ainda há espaço para melhorias, especialmente na classificação da classe `Alto`.
- **Clustering**: A análise do número ideal de clusters utilizando os métodos do cotovelo e da silhueta revelou que $k = 3$ é o mais adequado para uma interpretação geral dos dados. Além disso, a seleção de features com Lasso impactou positivamente a qualidade do clustering, resultando em clusters mais compactos e bem separados.

- **Distribuição dos Clusters**: A análise da distribuição dos clusters em relação às classes discretizadas de `rating` revelou padrões claros, como a forte associação do Cluster 1 com a classe `Médio`. Esses resultados demonstram a utilidade do clustering para explorar relações ocultas nos dados.

Limitações e Pontos de Atenção

Apesar dos resultados positivos, algumas limitações e desafios foram identificados durante o desenvolvimento do trabalho:

- **Coeficiente de Silhueta**: O coeficiente de silhueta foi relativamente baixo, indicando que os dados podem não ser ideais para clustering com K-Means. Isso sugere que métodos alternativos, como DBSCAN ou Agglomerative Clustering, podem ser mais adequados para este dataset.

- **Desempenho do Modelo KNN**: Embora o modelo KNN tenha alcançado uma acurácia satisfatória, ele apresentou dificuldades na classificação da classe `Alto`, com um número elevado de falsos negativos. Técnicas adicionais, como seleção de features mais avançada ou o uso de outros algoritmos (ex.: Random Forest, SVM), podem melhorar o desempenho.

- **Complexidade dos Dados**: A alta dimensionalidade e a presença de características não lineares podem ter impactado negativamente o desempenho dos modelos. Métodos de redução de dimensionalidade, como PCA, ou algoritmos mais robustos podem ajudar a mitigar essas limitações.

Contribuições do Trabalho

Este trabalho contribuiu para uma melhor compreensão das características do dataset de chocolates e das técnicas de aprendizado de máquina aplicadas. Os resultados obtidos fornecem uma base sólida para futuras análises e melhorias nos modelos preditivos.

Em resumo, os objetivos propostos foram alcançados, mas as limitações identificadas sugerem que há espaço para refinamentos e exploração de abordagens mais avançadas.

5 Próximos Passos

- Experimentar outras técnicas de clustering, como DBSCAN ou Agglomerative Clustering.

- Testar outros algoritmos de classificação, como Random Forest ou SVM.

- Realizar uma análise mais detalhada das características de cada cluster para entender melhor os padrões identificados.

- Incorporar mais variáveis explicativas, como políticas governamentais e fatores ambientais, para enriquecer o modelo.

- Explorar modelos avançados de séries temporais, como ARIMA ou LSTM, para capturar tendências ao longo do tempo.