SEGMENTAÇÃO DE VINHOS POR PERFIL QUÍMICO

INTRODUÇÃO

O Problema

A avaliação de vinhos é predominantemente subjetiva. Isso gera inconsistência na classificação e, mais grave, falhas na detecção de compostos nocivos em larga escala.

Objetivo

Desenvolver um sistema preditivo baseado em propriedades físico-químicas para classificar vinhos tintos de forma objetiva.

Solução

Utilizar o dataset Wine Quality para treinar modelos de Machine Learning capazes de identificar padrões de excelência e anomalias químicas.

JUSTIFICATIVA

- **Motivação de Saúde Pública**: o controle químico rigoroso não é apenas estético; é vital. Falhas no processo ou adulterações podem gerar altos níveis de Acidez Volátil ou contaminações por Metanol.

- A análise sensorial humana é incapaz de detectar certos riscos químicos sem expor o provador ao perigo.

Como Itália superou 'escândalo do vinho' adulterado com metanol e revolucionou mercado do país

GAMMA-RAPHO VIA GETTY IMAGES

O DATASET

Estrutura dos Dados:

11 Variáveis de Entrada (Features): Acidez Fixa, Acidez Volátil, Ácido Cítrico, Açúcar Residual, Cloretos, Dióxido de Enxofre Livre, Dióxido de Enxofre Total, Densidade, pH, Sulfatos e Álcool.

1 Variável de Saída (Target): Qualidade (score sensorial entre 0 e 10)

Características Importantes:

Dados multivariados com escalas distintas.

Classes desbalanceadas (muitos vinhos "médios", poucos "excelentes" ou "ruins").

Ausência de valores nulos (Dataset limpo).

PRÉ PROCESSAMENTO

Técnica 1: Padronização (StandardScaler)

O que faz: Ajusta os dados para média 0 e desvio padrão 1.

Justificativa: As variáveis possuem grandezas incompatíveis. O "Dióxido de Enxofre" varia na casa das centenas, enquanto "Cloretos" são decimais. Sem padronização, algoritmos baseados em distância (como K-Means) seriam enviesados pela variável de maior valor numérico.

PRÉ PROCESSAMENTO

Técnica 2: PCA (Análise de Componentes Principais)

O que faz: Reduz as 11 dimensões originais para 2 componentes principais.

Justificativa: Elimina a multicolinearidade (variáveis que dizem a mesma coisa, ex: pH e Acidez) e permite a visualização gráfica dos grupos em um plano 2D, facilitando a interpretação humana.

K-MEANS

O que é?

O **k-means** é um algoritmo de aprendizado não supervisionado (Unsupervised Learning) que busca formar grupos dentro de um conjunto de dados. Ele funciona iterativamente para atribuir cada ponto de dados a um de **k grupos**, minimizando a soma do quadrado das distâncias entre os pontos e o centroide (média) de seu cluster.

K-MEANS

Natureza do Objetivo

O objetivo principal do projeto era descobrir e classificar grupos de vinhos com perfis químicos semelhantes sem conhecimento prévio de suas categorias. Essa tarefa de agrupamento é intrinsecamente um problema de aprendizado não supervisionado, sendo o k-means o algoritmo mais fundamental e utilizado.

Simplicidade e Eficiência

O k-means é conhecido por sua simplicidade conceitual e alta eficiência computacional, especialmente em grandes conjuntos de dados. Isso torna a exploração inicial de agrupamentos rápida e eficaz.

K-MEANS

Interpretabilidade

Os centroides dos clusters (k-means) são fáceis de interpretar, pois representam os valores médios das características químicas para cada grupo. Isso permitiu a fácil caracterização dos perfis de vinho encontrados

Relevância para a Aplicação

A segmentação por perfil químico é uma forma de análise exploratória que busca padrões latentes. O k-means é excelente para isso e se alinha perfeitamente com a aplicação sugerida como um "filtro de controle de qualidade", onde vinhos podem ser rapidamente classificados em um dos perfis químicos estabelecidos.

K-MEANS

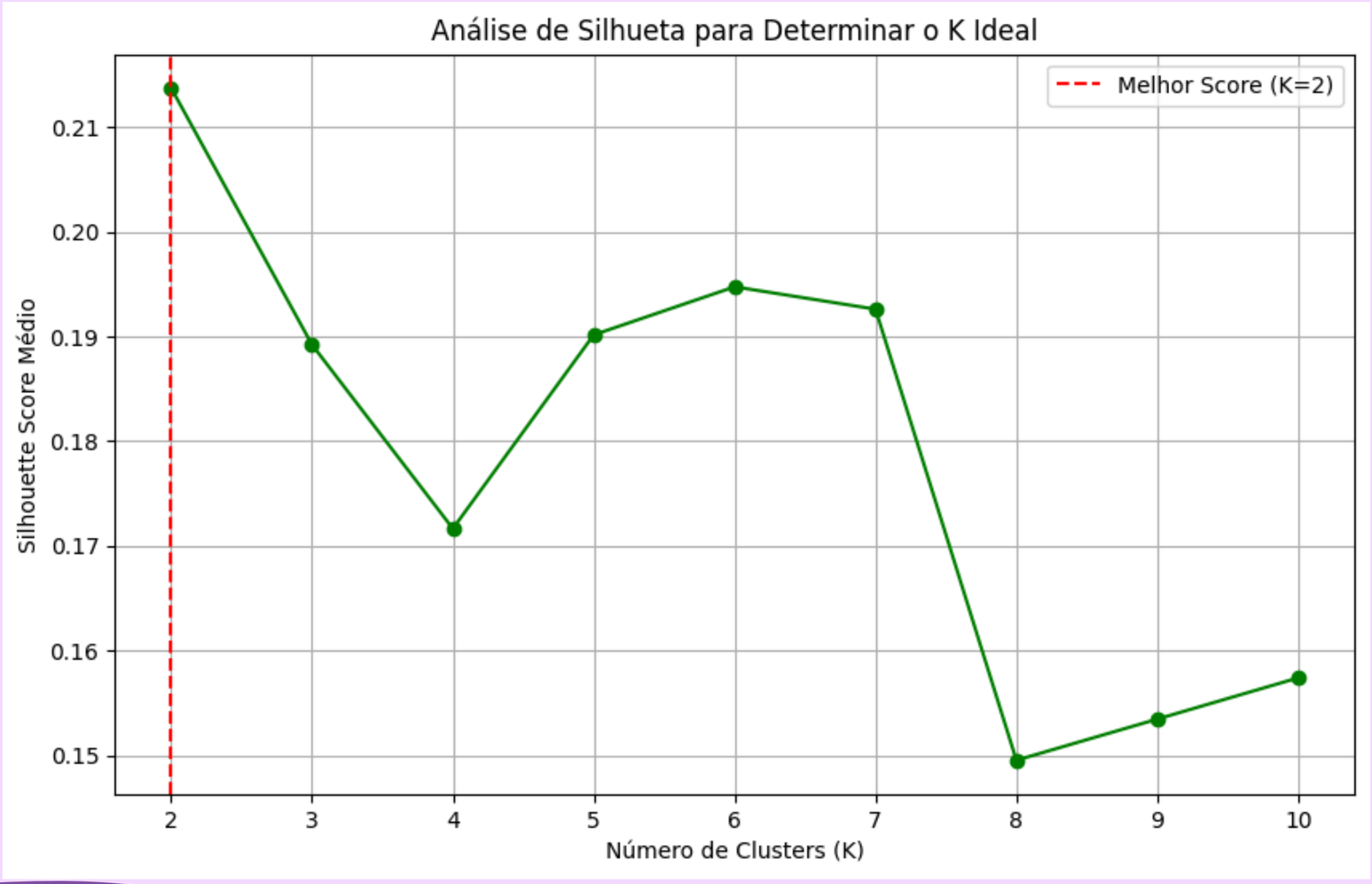
Configuração: K=2 (definido via método silueta).

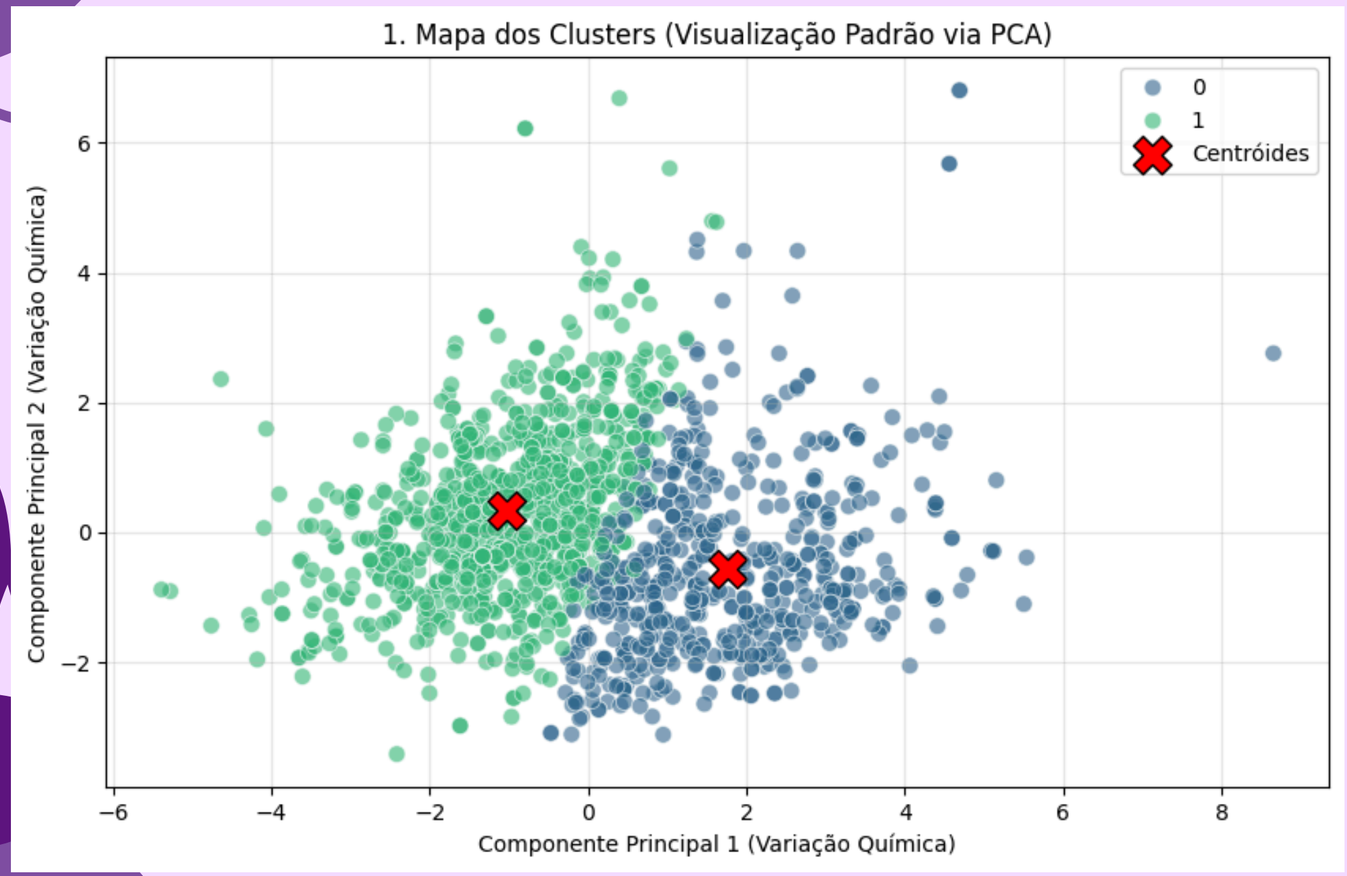
Cluster 0:

Maior teor alcoólico (Média: 10.61%)
Menor acidez volátil (Média: 0.41) = QUALIDADE MÉDIA MAIOR

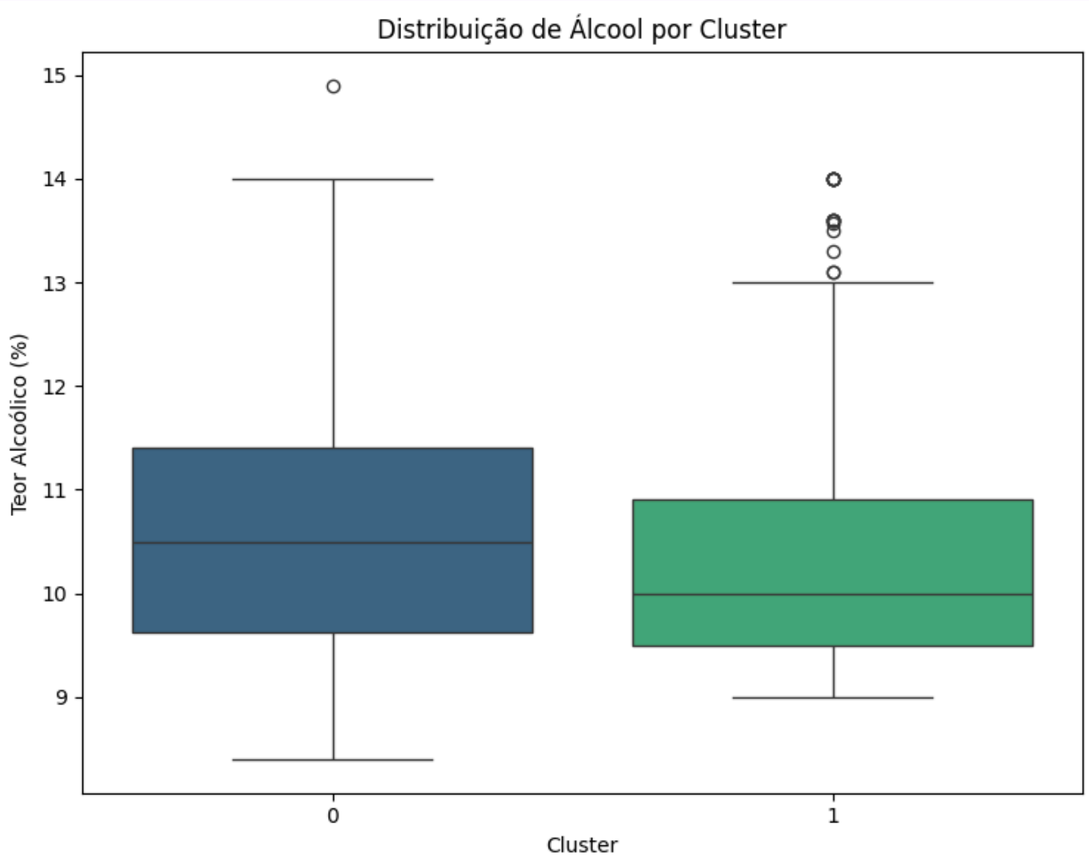
Cluster 1:

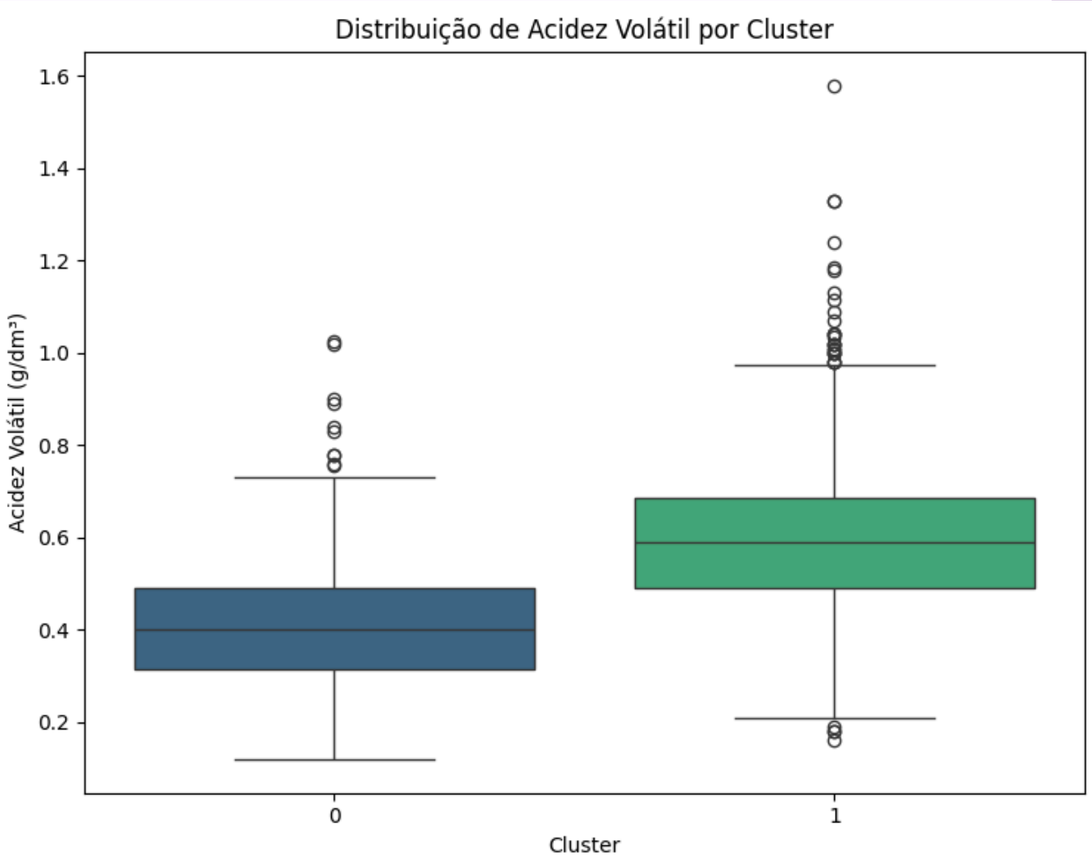
Menor teor alcoólico (Média: 10.32%)
Maior acidez volátil (Média: 0.59) = QUALIDADE MÉDIA MENOR


Análise de Silhueta para Determinar o K Ideal
Melhor Score (K=2)
Silhouette Score Médio
0.21
0.20
0.19
0.18
0.17
0.16
0.15
2
3
4
5
6
7
8
9
10
Número de Clusters (K)



- Cluster 0 (Azul): 590 elementos
- Cluster 1 (Verde): 1009 elementos–
- Total de 1599 elementos


Distribuição de Álcool por Cluster
Teor Alcoólico (%)
15
14
13
12
11
10
9
0
1
Cluster

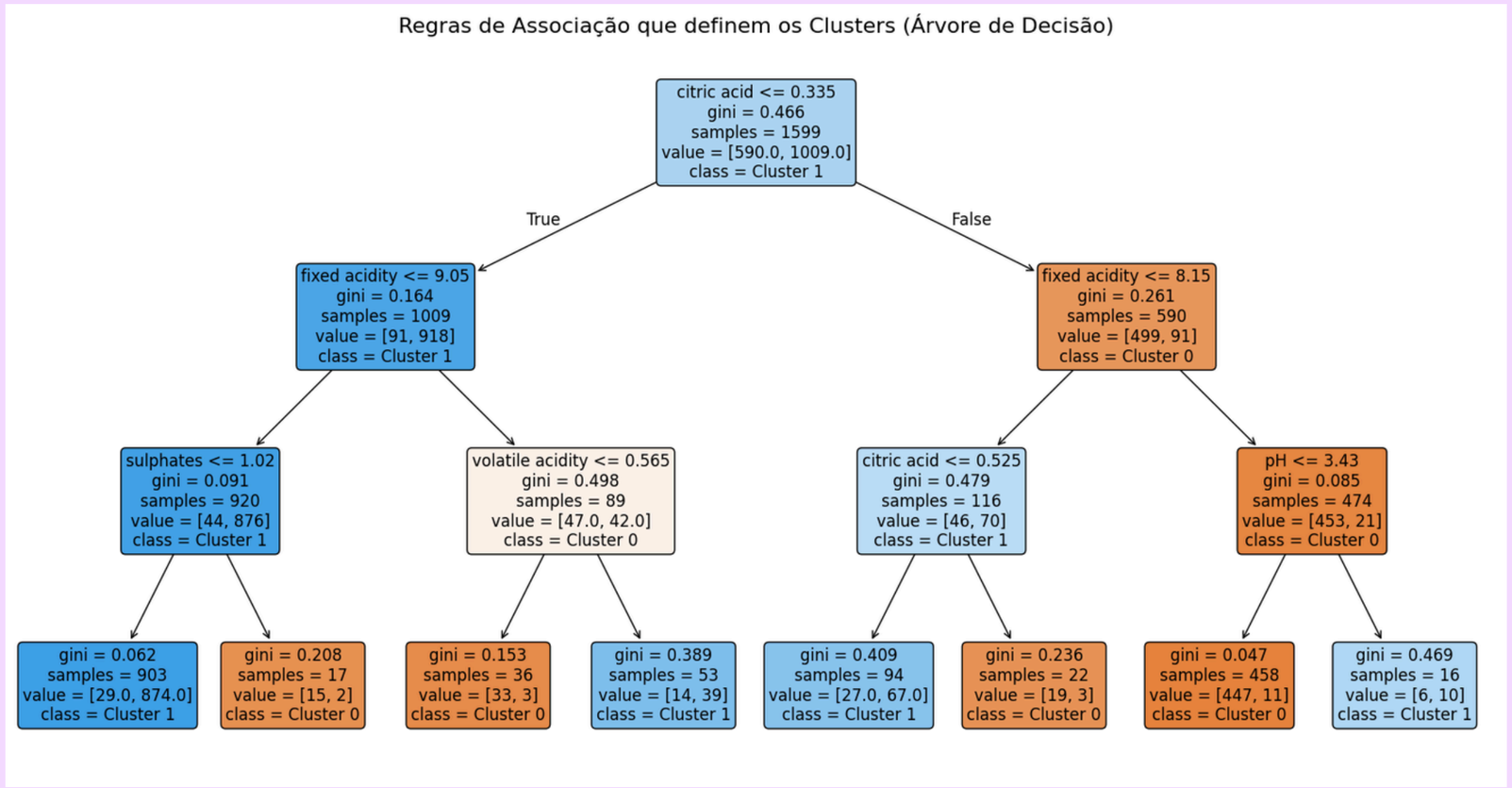

Distribuição de Acidez Volátil por Cluster
Acidez Volátil (g/dm³)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
1
Cluster

ASSOCIAÇÃO

SE Teor Alcoólico >= 10.5% **E** Acidez Volátil < 0.65 g/dm³ **ENTÃO** classificar como Perfil Superior (Cluster 1)

SE Teor Alcoólico < 10.5% **OU** Acidez Volátil >= 0.70 g/dm³ **ENTÃO** classificar como Perfil de Entrada/Risco (Cluster 0).

SE o Álcool estiver entre 9.8% e 10.5% **E** os Sulfatos forem maiores que 0.60 g/dm³ **ENTÃO** o vinho ainda pode ser classificado no Perfil Superior.


Regras de Associação que definem os Clusters (Árvore de Decisão)
citric acid <= 0.335
gini = 0.466
samples = 1599
value = [590.0, 1009.0]
class = Cluster 1
True
False
fixed acidity <= 9.05
gini = 0.164
samples = 1009
value = [91, 918]
class = Cluster 1
fixed acidity <= 8.15
gini = 0.261
samples = 590
value = [499, 91]
class = Cluster 0
sulphates <= 1.02
gini = 0.091
samples = 920
value = [44, 876]
class = Cluster 1
volatile acidity <= 0.565
gini = 0.498
samples = 89
value = [47.0, 42.0]
class = Cluster 0
citric acid <= 0.525
gini = 0.479
samples = 116
value = [46, 70]
class = Cluster 1
pH <= 3.43
gini = 0.085
samples = 474
value = [453, 21]
class = Cluster 0
gini = 0.062
samples = 903
value = [29.0, 874.0]
class = Cluster 1
gini = 0.208
samples = 17
value = [15, 2]
class = Cluster 0
gini = 0.153
samples = 36
value = [33, 3]
class = Cluster 0
gini = 0.389
samples = 53
value = [14, 39]
class = Cluster 1
gini = 0.409
samples = 94
value = [27.0, 67.0]
class = Cluster 1
gini = 0.236
samples = 22
value = [19, 3]
class = Cluster 0
gini = 0.047
samples = 458
value = [447, 11]
class = Cluster 0
gini = 0.469
samples = 16
value = [6, 10]
class = Cluster 1

RESULTADOS

Quimica VS Paladar

O estudo confirmou que o teor alcoólico é o maior "driver" de qualidade percebida, enquanto a acidez volátil é o maior detrator.

Segurança

O modelo aprendeu sozinho a penalizar vinhos com alta acidez volátil. Isso valida o uso da IA como ferramenta de triagem para evitar que vinhos estragados cheguem ao mercado.

Analise Critica

Acurácias perfeitas são impossíveis devido à subjetividade humana nas notas originais, mas o modelo químico é mais consistente e reproduzível.

CONCLUSÃO

Sintese

A mineração de dados provou ser eficaz para transformar a enologia em uma ciência de dados exata.

Contribuição Pratica

Para produtores: Ajuste fino da fermentação para atingir os clusters "Premium".
Para segurança: Detecção automática de lotes anômalos antes do envase.

FONTES

- https://archive.ics.uci.edu/dataset/186/wine+quality
- https://www.bbc.com/portuguese/articles/c620lqp7v76o
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

OBRIGADO