Aplicação de algoritmos de clustering na base de dados *Wine Quality*

José R. da Silva[1], Matheus V. A. da Silva[1], Pedro A. S. Patriota[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)

Abstract. *This article is dedicated to analyzing the Wine Quality dataset, made available by the UCI Machine Learning Repository, with the aim of exploring and identifying quality patterns in wines through advanced clustering techniques. We used three of the most renowned machine learning models for cluster analysis: K-means, DBSCAN, and fuzzy c-Means, in order to compare their effectiveness and applicability in distinguishing between different wine qualities based on physicochemical characteristics. To determine the optimum number of clusters in the K-means method, we applied techniques such as the Elbow method and Silhouette analysis, which provided valuable insights for configuring the models. The results obtained reveal significant differences in the performance of the models in terms of clustering accuracy and interpretability of the clusters. K-means proved to be a robust tool for general groupings, while DBSCAN excelled at identifying outliers and more subtle patterns. Fuzzy c-Means, on the other hand, offered a flexible perspective, allowing for a more nuanced interpretation of each sample's belonging to the clusters. This study contributes to the field of wine data analysis by demonstrating how different machine learning techniques can be used to extract valuable knowledge about wine quality, potentially helping producers and consumers to better understand the complexities that define wine preferences.*

Resumo. *Este estudo se dedica à análise do dataset Wine Quality, disponibilizado pelo UCI Machine Learning Repository, com o objetivo de explorar e identificar padrões de qualidade nos vinhos através de técnicas avançadas de clustering. Utilizamos três dos mais renomados modelos de machine learning para análise de clusters: K-means, DBSCAN, e fuzzy c-Means, visando comparar sua eficácia e aplicabilidade na distinção entre diferentes qualidades de vinho baseadas em características fisicoquímicas. Para determinar o número ótimo de clusters no método K-means, aplicamos técnicas como o método Elbow e a análise Silhouette, que forneceram insights valiosos para a configuração dos modelos. Os resultados obtidos revelam diferenças significativas na performance dos modelos em termos de precisão de agrupamento e interpretabilidade dos clusters. O K-means demonstrou ser uma ferramenta robusta para agrupamentos gerais, enquanto o DBSCAN se destacou na identificação de outliers e padrões mais sutis. O Fuzzy c-Means, por sua vez, ofereceu uma perspectiva flexível, permitindo uma interpretação mais nuanceada das pertenças de cada amostra aos clusters. Este estudo contribui para o campo de análise de dados de vinho, demonstrando como diferentes técnicas de machine learning podem ser utilizadas para extrair conhecimentos valiosos sobre a qualidade do vinho, potencialmente auxiliando produtores e consumidores a entenderem melhor as complexidades que definem as preferências de vinho.*

1. Introdução

O dataset Wine Quality [Paulo Cortez and Reis 2009], disponibilizado pela UCI Machine Learning Repository, consiste em um conjunto de dados que relaciona características físico-químicas de vinhos a uma qualidade avaliada por especialistas. Com variáveis que incluem acidez, teor alcoólico, açúcar residual, entre outras, o dataset oferece uma oportunidade única para explorar como esses atributos influenciam a percepção de qualidade do vinho. A relevância desse estudo estende-se da indústria vitivinícola à ciência de dados, fornecendo insights valiosos para vinicultores, sommeliers e consumidores, além de ser um campo fértil para a aplicação e comparação de técnicas de machine learning.

O objetivo específico desta análise é utilizar este dataset para examinar e comparar a eficácia de diferentes modelos de *clustering* na identificação de padrões e agrupamentos intrínsecos, visando não apenas uma melhor compreensão das relações entre as variáveis e a qualidade do vinho, mas também a aplicação prática desses *insights* para melhorar a produção e avaliação dessas bebidas. A análise foca em três modelos de machine learning amplamente reconhecidos por suas capacidades de identificação de clusters: K-means, DBSCAN e Fuzzy c-Means.

A escolha desses modelos baseia-se em suas características únicas e complementares. O K-means é notável por sua simplicidade e eficácia em formar clusters baseados na proximidade de características, sendo ideal para identificar agrupamentos distintos quando o número de clusters é conhecido ou pode ser estimado. O DBSCAN, por outro lado, é eficiente na identificação de clusters com densidades variáveis e na detecção de outliers, adaptando-se bem a datasets com agrupamentos complexos. Finalmente, o fuzzy c-Means oferece uma abordagem mais flexível, permitindo que um ponto pertença a múltiplos clusters com diferentes graus de associação, o que pode revelar nuances importantes na estrutura dos dados que métodos mais rígidos podem ignorar. A combinação dessas técnicas permite uma análise robusta e detalhada do dataset Wine Quality, com o intuito de extrair padrões significativos que possam contribuir para a ciência enológica e a análise de dados.

2. Fundamentos

A análise de dados fundamenta-se no uso de três modelos de clustering distintos, cada um com sua própria filosofia e aplicabilidade em contextos específicos de agrupamento de dados. Esta seção discorre sobre os princípios teóricos que embasam os modelos K-means, DBSCAN e fuzzy c-Means, fornecendo uma compreensão essencial para a interpretação dos resultados obtidos.

2.1. K-means

O K-means é um dos algoritmos de clustering mais simples e amplamente utilizados, caracterizado por sua abordagem de particionamento. O objetivo do K-means é dividir o dataset em um número K de clusters, no qual cada ponto pertence ao cluster que possui o centroide mais próximo. O processo inicia com a seleção aleatória de K centroides e, em seguida, atribui cada ponto ao mais próximo. Após isso, os centros dos clusters são recalculados como a média de todos os pontos atribuídos. Este processo é iterativo e continua até que a posição dos centros dos clusters não mude significativamente, indicando a sua estabilização. O K-means é particularmente útil para identificar padrões claros e distintos nos dados, mas requer que o número de clusters seja definido a priori.

2.2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Diferentemente do K-means, o DBSCAN não requer que o número de clusters seja especificado antecipadamente e é capaz de identificar clusters de formas arbitrárias. O DBSCAN classifica os pontos em três categorias: pontos centrais, pontos de borda e outliers, com base na densidade de vizinhos. Um ponto é considerado central se tiver um número mínimo de outros pontos em sua vizinhança de raio especificado. O algoritmo cresce um cluster ao redor desses pontos centrais, agregando pontos que são alcançáveis na vizinhança densa. DBSCAN é eficaz para datasets com clusters de densidade variável e é robusto à presença de ruído ou outliers nos dados.

2.3. Fuzzy c-Means

O Fuzzy c-Means é uma extensão do algoritmo K-means que introduz a noção de pertencimento parcial de pontos a múltiplos clusters. Ao invés de forçar cada ponto a pertencer exclusivamente a um cluster, o Fuzzy c-Means permite que cada ponto tenha um grau de associação com todos os clusters. Isso é quantificado por coeficientes de pertencimento que variam entre 0 e 1, refletindo a incerteza ou ambiguidade na classificação dos pontos. O algoritmo iterativamente ajusta os centros dos clusters e atualiza os coeficientes de pertencimento para minimizar a função de custo, que é baseada na distância ponderada dos pontos aos centros dos clusters. O Fuzzy c-Means é particularmente útil para datasets onde os limites entre os clusters não são claramente definidos, permitindo uma interpretação mais matizada dos agrupamentos.

Cada um desses modelos traz uma abordagem única para o desafio de identificar agrupamentos naturais nos dados do Wine Quality, permitindo uma análise abrangente que considera tanto as distinções claras quanto as nuances sutis presentes no dataset. A escolha desses modelos reflete o objetivo de explorar a estrutura dos dados de maneira profunda, buscando insights que possam ser aplicados tanto no campo da enologia quanto na ciência de dados.

2.4. Métricas de avaliação

As métricas de avaliação desempenham um papel fundamental na validação e na compreensão dos resultados obtidos por algoritmos de clustering, e algumas delas foram sem dúvidas uma parte essencial deste estudo.

2.4.1. Coeficiente da silhueta

O coeficiente de silhueta é uma métrica de avaliação comumente usada para avaliar a qualidade dos agrupamentos (clusters) em algoritmos de clustering. Ele fornece uma medida da coesão e separação dos clusters, ajudando a determinar se os clusters formados são bem definidos e apropriados para os dados em questão.

2.4.2. Método do cotovelo

O método do cotovelo (Elbow Method) é uma técnica comumente utilizada na análise de algoritmos de clustering, como K-means, para determinar o número ideal de clusters a

serem utilizados para agrupar os dados. O nome ”método do cotovelo”deriva da forma do gráfico gerado pela técnica, que se assemelha a um cotovelo.

A importância do método do cotovelo reside no fato de que, muitas vezes, não sabemos previamente quantos clusters devemos usar para agrupar nossos dados de forma eficaz. Utilizando esse método, podemos encontrar um ponto no gráfico onde a adição de mais clusters não traz uma melhoria significativa na variação intra-cluster, ou seja, não reduz significativamente a distância média entre os pontos e o centróide do cluster ao qual pertencem.

2.4.3. Coeficiente de partição *fuzzy*

O coeficiente de partição fuzzy (FPC) é uma métrica utilizada para avaliar a qualidade dos agrupamentos obtidos por algoritmos de clustering fuzzy, como o Fuzzy C-Means em questão.

O coeficiente de partição fuzzy é uma medida de quanto os clusters sobrepostos são. Ele varia de 0 a 1, sendo que um valor próximo de 1 indica uma boa partição fuzzy, na qual os clusters têm pouca sobreposição, enquanto um valor próximo de 0 indica o contrário.

O cálculo do coeficiente de partição fuzzy considera a pertinência dos pontos de dados em relação aos diferentes clusters. Ele mede a diferença entre a pertinência média e a máxima pertinência para cada ponto. Quanto menor essa diferença, maior é o valor do coeficiente de partição fuzzy e melhor é a partição.

2.5. Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é uma técnica estatística utilizada para reduzir a dimensionalidade de conjuntos de dados complexos, preservando o máximo possível de sua variabilidade original. Ela faz isso transformando os dados originais em um novo conjunto de variáveis não correlacionadas chamadas componentes principais. A PCA é amplamente utilizada para explorar e visualizar dados multidimensionais, como o dataset em questão, de forma mais compreensível e interpretável.

3. Metodologia

A metodologia adotada para analisar o dataset Wine Quality envolveu uma série de etapas cuidadosamente planejadas, desde a preparação dos dados até a aplicação e avaliação dos modelos de clustering K-means, DBSCAN e fuzzy c-Means. Esta seção descreve detalhadamente cada uma dessas etapas, proporcionando uma visão clara do processo experimental.

3.1. Preparação dos Dados

Inicialmente, o dataset Wine Quality foi submetido a um processo de e normalização, eliminando o viés introduzido por variáveis com escalas significativamente diferentes. A normalização foi realizada ajustando os dados para terem média zero e desvio padrão unitário, um procedimento padrão que facilita a comparação entre as características.

3.2. Parâmetros dos Modelos

Para o K-means e o fuzzy c-Means, a seleção do número ótimo de clusters (K) foi uma etapa crítica. Utilizamos o método Elbow e o coeficiente de Silhouette para determinar o valor de K que melhor se adequava aos dados. O método Elbow envolve a plotagem da variação da soma dos quadrados dentro dos clusters em relação ao número de clusters, buscando um ponto onde o declínio da variação se torna menos acentuado (o "cotovelo"). O coeficiente de Silhouette, por outro lado, mede a qualidade do clustering baseando-se na distância entre os clusters e a coesão dentro dos clusters, com valores mais altos indicando uma melhor definição de cluster. Para o DBSCAN, os parâmetros principais incluíram o raio de vizinhança (ϵ) e o número mínimo de pontos ($minPts$) necessários para formar um cluster. Estes parâmetros foram ajustados experimentalmente para maximizar a capacidade do modelo de identificar clusters significativos, ao mesmo tempo em que minimizava a identificação de outliers como clusters independentes.

3.3. Aplicação dos Modelos

Cada modelo de clustering foi aplicado ao dataset Wine Quality seguindo os parâmetros determinados na fase de preparação e seleção. Para o K-means e o fuzzy c-Means, os experimentos foram realizados variando o número de clusters com base nos insights obtidos através do método Elbow e do coeficiente de Silhouette. O DBSCAN foi aplicado com diferentes combinações de ϵ e $minPts$, explorando a capacidade do modelo de adaptar-se à densidade variável dos dados.

3.4. Avaliação dos Modelos

A avaliação dos modelos concentrou-se na análise de diversos critérios, incluindo a interpretabilidade dos clusters formados, a coesão interna e a separação entre clusters. O coeficiente de Silhouette foi novamente utilizado para quantificar a qualidade dos clusters em todos os modelos. Além disso, para o fuzzy c-Means, o Coeficiente de Partição Fuzzy (FPC) foi empregado como uma medida adicional para avaliar a adequação dos clusters, considerando a natureza fuzzy das atribuições de cluster.

3.5. Visualização dos clusters obtidos

De forma a ter uma validação visual dos modelos para além das métricas utilizadas, também foi implementada a visualização dos clusters criados por meio de gráficos de dispersão de pontos. No entanto, é imprescindível salientar que para realizar a plotagem em duas dimensões, antes é necessária a aplicação do algoritmo PCA, o que pode resultar em perda de informações. Portanto, tais gráficos devem ser tratados apenas como uma parte de todas as etapas de validação necessárias.

3.6. Limitações

A metodologia adotada enfrentou limitações inerentes aos modelos de clustering e às técnicas de determinação da quantidade de clusters. Por exemplo, a escolha de parâmetros para o DBSCAN e o fuzzy c-Means pode ser subjetiva e dependente do dataset específico, podendo não generalizar bem para outros conjuntos de dados. Além disso, a eficácia do método Elbow e da análise de Silhouette pode ser comprometida em casos onde a distribuição dos dados não apresenta uma estrutura de cluster clara ou quando os clusters

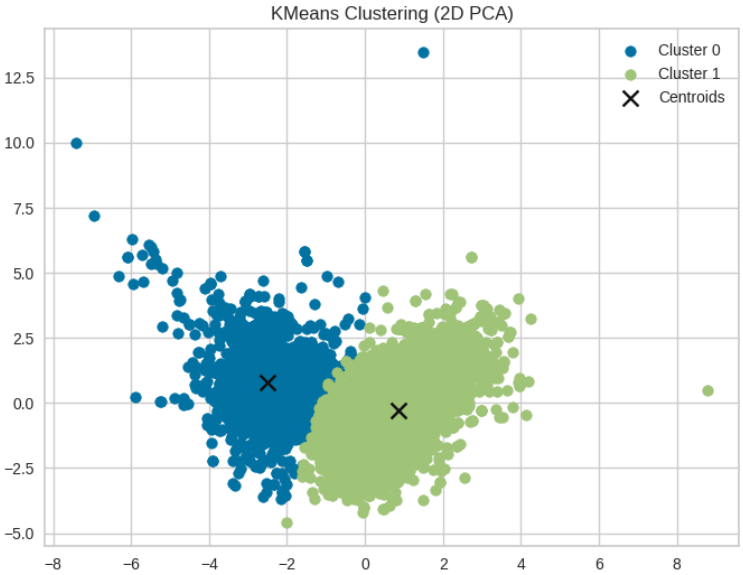
têm densidades variáveis. Essas limitações destacam a importância de uma abordagem holística que considere múltiplas técnicas e a interpretação cuidadosa dos resultados.

Este procedimento metodológico permitiu uma comparação sistemática dos modelos de clustering, levando em conta tanto aspectos quantitativos quanto qualitativos dos agrupamentos. A aplicação cuidadosa dessa metodologia visa não apenas identificar padrões significativos no dataset Wine Quality, mas também avaliar a eficácia de diferentes abordagens de clustering em contextos de dados variados.

4. Resultados

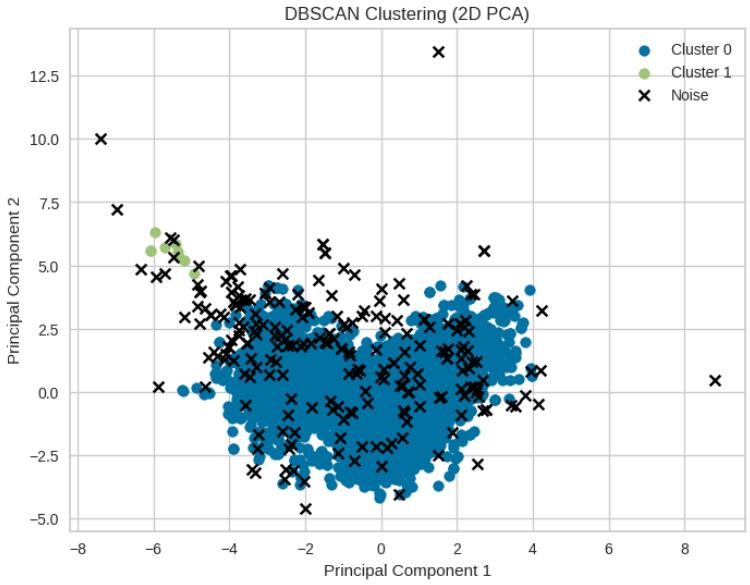
Os experimentos conduzidos com o dataset Wine Quality utilizando os modelos K-means, DBSCAN e fuzzy c-Means revelaram padrões e correlações significativos entre as características dos vinhos e suas qualidades avaliadas. Esta seção resume os resultados chave, com ênfase nas descobertas mais relevantes e na interpretação dos dados visualizados através de tabelas e gráficos.

- K-means: A aplicação do método Elbow e da análise de Silhouette para determinar o número ótimo de clusters no modelo K-means indicou uma clara preferência por um número específico de clusters, refletindo distintas categorias de qualidade dos vinhos. As tabelas resultantes apresentam as médias das características dos vinhos para cada cluster, ilustrando como certas propriedades químicas são predominantes em vinhos de diferentes qualidades. Os gráficos correspondentes visualizam a distribuição dos vinhos nos clusters, mostrando a compactação dos grupos e a separação entre eles, o que facilita a identificação de padrões qualitativos.

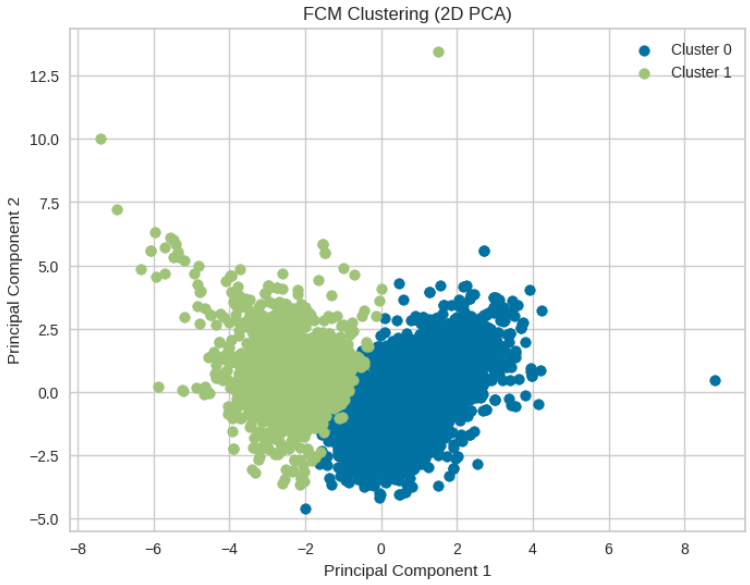


- DBSCAN: O modelo DBSCAN, conhecido por sua capacidade de identificar clusters baseados em densidade sem necessitar de um número pré-definido de grupos, destacou-se pela identificação de agrupamentos não lineares e pela capacidade de manejar outliers. Os resultados incluíram a descoberta de clusters principais que correspondem a grupos de vinhos com características únicas, bem como a identificação de amostras individuais que não se enquadram facilmente em categorias de qualidade padronizadas. A análise visual dos clusters DBSCAN, através

de gráficos de dispersão, oferece insights sobre a estrutura espacial dos dados e a distribuição das amostras.
Entretanto, um aspecto curioso da aplicação desse modelo foi a detecção de *outliers*, pois embora que os pontos claramente identificáveis como outliers possam ser categorizados como tais, muitos pontos pertencentes ao clusters também o são. Porém, não deve-se esquecer da aplicação do método PCA nos dados multidimensionais em questão, o que inviabiliza dizer que o modelo é incapaz de detectar outliers corretamente.



- Fuzzy c-Means: A implementação do fuzzy c-Means proporcionou uma perspectiva única sobre a classificação dos vinhos, permitindo uma associação flexível das amostras a múltiplos clusters. Os resultados destacaram a natureza gradativa das qualidades dos vinhos, com tabelas mostrando a probabilidade de cada vinho pertencer a diferentes clusters. Gráficos de pertencimento ilustram como os vinhos compartilham características entre categorias de qualidade, sugerindo uma abordagem mais nuanciada para a classificação da qualidade.



4.1. Interpretação Conjunta

A análise conjunta dos resultados dos três modelos oferece uma visão compreensiva das relações entre as variáveis físico-químicas dos vinhos e suas qualidades percebidas. Enquanto o K-means e o fuzzy c-Means proporcionam uma visão clara sobre a segmentação baseada em qualidade, o DBSCAN revela a complexidade e a variação dentro dessas categorias. Os gráficos e tabelas, em conjunto, facilitam a interpretação desses padrões, sugerindo que a qualidade do vinho é influenciada por uma combinação de fatores, muitos dos quais transcendem classificações simplistas.

Os resultados experimentais, portanto, não apenas esclarecem a relação entre as características dos vinhos e sua qualidade, mas também demonstram a importância de abordagens analíticas diversificadas na extração de insights profundos a partir de datasets complexos.

5. Conclusão

A análise reforçou a ideia de que a qualidade do vinho não pode ser totalmente explicada por um único conjunto de variáveis físico-químicas, mas é influenciada por uma complexa interação de fatores. Além disso, a utilização de múltiplos métodos de clustering destacou a importância de abordagens analíticas diversificadas para capturar a totalidade dos padrões presentes em dados complexos. As limitações encontradas, incluindo a escolha de parâmetros para os modelos e a interpretação dos métodos de determinação do número de clusters, sublinham a necessidade de cautela e rigor na análise de datasets semelhantes.

5.1. Síntese dos Resultados

Os resultados indicaram que cada modelo de clustering tem suas próprias forças na identificação de padrões nos dados. O K-means provou ser eficaz na segmentação dos vinhos em categorias de qualidade distintas, baseando-se em características químicas específicas. O DBSCAN, por sua vez, destacou-se na detecção de estruturas de dados complexas e na identificação de outliers, desafiando a categorização convencional da qualidade. O fuzzy c-Means ofereceu uma perspectiva flexível, permitindo uma compreensão mais matizada das categorias de qualidade através da associação de amostras a múltiplos clusters. Juntos, esses modelos forneceram uma visão compreensiva e multidimensional da qualidade dos vinhos, evidenciando a interação entre diversos atributos químicos e a percepção sensorial da qualidade.

5.2. Discussões

A análise reforçou a ideia de que a qualidade do vinho não pode ser totalmente explicada por um único conjunto de variáveis físico-químicas, mas é influenciada por uma complexa interação de fatores. Além disso, a utilização de múltiplos métodos de clustering destacou a importância de abordagens analíticas diversificadas para capturar a totalidade dos padrões presentes em dados complexos. As limitações encontradas, incluindo a escolha de parâmetros para os modelos e a interpretação dos métodos de determinação do número de clusters, sublinham a necessidade de cautela e rigor na análise de datasets semelhantes.

5.3. Direções Futuras

Os insights gerados por este estudo abrem várias direções para pesquisas futuras. Primeiramente, a aplicação de modelos de clustering híbridos ou técnicas de aprendizado de

máquina mais avançadas, como redes neurais profundas, pode fornecer novas perspectivas sobre a classificação da qualidade dos vinhos. Além disso, a incorporação de dados sensoriais ou de preferências dos consumidores poderia enriquecer a análise, permitindo uma avaliação mais holística da qualidade. Por fim, estudos futuros poderiam explorar a aplicabilidade desses métodos de clustering em outros domínios da indústria alimentícia, onde a qualidade e a percepção sensorial desempenham papéis cruciais.

Em conclusão, este estudo demonstrou o potencial dos modelos de clustering na exploração de datasets complexos, como o Wine Quality, e na obtenção de insights valiosos sobre a dinâmica entre características químicas e a qualidade percebida. Ao fazê-lo, não apenas contribuiu para o campo da ciência de dados aplicada à enologia, mas também destacou caminhos promissores para pesquisas futuras na interseção da análise de dados, ciência sensorial e percepção de qualidade.

Referências

Paulo Cortez, António Cerdeira, F. A. T. M. and Reis, J. (2009). Wine quality data set. UCI Machine Learning Repository. `https://archive.ics.uci.edu/ml/datasets/wine+quality`.