|  |  |
| --- | --- |
| **Wine Quality Data Set**  *Download*: [Data Folder](https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/), [Data Set Description](https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.names)  **Abstract**: Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], [[Web Link]](http://www3.dsi.uminho.pt/pcortez/wine/)). | https://archive.ics.uci.edu/ml/assets/MLimages/Large186.jpg |
|  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set Characteristics: | Multivariate | Number of Instances: | 6.497 | Area: | Business |
| Attribute Characteristics: | Real | Number of Attributes: | 12 | Date Donated | 2009-10-07 |
| Associated Tasks: | Classification, Regression | Missing Values? | N/A | Number of Web Hits: | 381176 |

**Source:**

Paulo Cortez, University of Minho, Guimarães, Portugal, <http://www3.dsi.uminho.pt/pcortez>   
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

**Data Set Information:**

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [[Web Link]](http://www.vinhoverde.pt/en/) or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).   
  
These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

**Attribute Information:**

For more information, read [Cortez et al., 2009].   
Input variables (based on physicochemical tests):

**Relevant Papers:**

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.   
  
Available at: [[Web Link]](http://dx.doi.org/10.1016/j.dss.2009.05.016)

**Citation Request:**

Please include this citation if you plan to use this database:   
  
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.

|  |  |  |
| --- | --- | --- |
| **Característica** | **Descrição** | |
| * **fixedacidity**  *(Acidez fixa)* | | a maioria dos ácidos presentes no vinho ou fixos ou não voláteis (não evaporaram prontamente) |
| * **volatileacidity** *(Acidez volátil)* | | a quantidade de ácido acético no vinho, que em níveis muito altos pode levar a um gosto desagradável de vinagre. |
| * **Citricacid** (Ácido cítrico) | | encontrado em pequenas quantidades, o ácido cítrico pode adicionar “frescura” e sabor aos vinhos. |
| * **Residualsugar**   *(Açúcar residual)* | | a quantidade de açúcar restante depois que o processo de fermentação (em contato com a levedura, o açúcar da uva vai se transformando em álcool, ou seja, a fermentação vai ocorrendo) |
| * **Chlorides**  *(*Cloretos) | | a quantidade de sal no vinho. |
| * **freesulfurdioxide**   *(Dióxido de enxofre livre)* | | a forma livre de SO2 existe em equilíbrio entre o SO2 molecular (como gás dissolvido) e o íon bissulfito; impede o crescimento microbiano e a oxidação do vinho. Garante condições melhores para os processos de vinificação da bebida, elimina bactérias e leveduras frágeis e indesejáveis, o que permite que apenas as melhores prossigam com o processo fermentativo. Além do mais, melhora o aroma e afina a cor da bebida |
| * **totalsulfurdioxide**   *(Dióxido de enxofre total)* | | quantidade de formas livres e encadernadas de S02; em baixas concentrações, o SO2 é quase indetectável no vinho, mas nas concentrações de SO2 livre acima de 50 ppm, o SO2 se torna evidente no nariz e no sabor do vinho. |
| * **Density**  (*Densidade)* | | a densidade é próxima à da água, dependendo do percentual de álcool e teor de açúcar. |
| * **pH** (*PH)* | | descreve como o vinho é acídico ou básico numa escala de 0 (muito ácido) a 14 (muito básico); a maioria dos vinhos tem entre 3-4 na escala de pH. |
| * **Sulphates** *(Sulfatos)* | | um aditivo de vinho que pode contribuir para os níveis de gás de dióxido de enxofre (S02), que age como um antimicrobiano e antioxidante. |
| * **alcohol** *(Álcool)* | | o teor alcoólico percentual do vinho. |
| * **quality** *(Qualidade)* | | variável de saída (com base em dados sensoriais) que poderiam ser de 0 a 10 sendo zero muito ruim e 10 muito excelente. |

Disciplina: ***IAeML***

Trabalho: ***Conceitos Estatísticos para IA***

Utilizando a base descrita e disponibilizada de vinhos desta região de Portugal com as variáveis de características (composição) dos vinhos.

**Etapa1 1:**

Faça um algoritmo que estime a variável “Quality” em função das características físico-químicas dos vinhos;

1. Análise Exploratória dos dados
2. Exclusão de *outliers*, caso necessário (sempre explicando a opção)
3. Treinamento dos modelos e Validação

Em cada opção de modelo colocar a saída do modelo e sua respectiva métrica validação.

1. Para cada etapa anterior colocar comentários sobre a técnica utiliza e análise sobre as variáveis utilizadas e seus respectivos “achados”

* Técnicas: Regressão Linear, Árvore de Regressão

1. Com a métrica correta faça a comparam entre os diferentes os algoritmos e qual foi o melhor entre eles.

**Etapa 2:**

Sabendo que os vinhos com notas>=6 são considerados vinhos de boa qualidade faça um algoritmo que classifique os vinhos em “Bom” ou “Ruim” em função de suas características físico-químicas;

1. Definição da variável resposta
2. Análise Exploratória dos dados
3. Treinamento dos modelos e Validação

Em cada opção de modelo colocar a saída do modelo e sua respectiva métrica validação.

1. Para cada etapa anterior colocar comentários sobre a técnica utilizada e análise sobre as variáveis utilizadas e seus respectivos “achados”

* Técnicas: Regressão logística, Árvore de decisão

1. Com a métrica correta faça a comparam entre os diferentes os algoritmos e qual foi o melhor entre eles.

**Etapa 3:**

Faça um algoritmo que a partir das características físico-químicas dos vinhos defina grupos para auxiliar na importação.

1. Definição das variáveis segmentadoras
2. Análise Exploratória dos dados
3. Critérios para definição da quantidade de grupos (clusters)
4. Resultados encontrados – Caracterização dos Clusters
5. Para cada etapa anterior colocar comentários sobre a técnica utiliza e análise sobre as variáveis utilizadas e seus respectivos “achados”
6. Utilizando componentes principais (PCA), explique sobre a quantidade de componentes e a interpretação dos componentes.
7. Defina os clusters com os componentes encontrados.
8. Quantos clusters utilizando os componentes encontrados
9. Resultados encontrados – Caracterização dos Clusters gerado com os componentes
10. Cruzar os clusters obtidos a partir das variáveis e clusters obtidos pelos componentes principais. O que aconteceu?

* Técnicas: K-means, Cluster Hierárquico, Componentes Principais