|  |  |
| --- | --- |
| **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: | 4898 | 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 munch 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):

|  |  |
| --- | --- |
| 1 - fixed acidity  2 - volatile acidity  3 - citric acid  4 - residual sugar  5 - chlorides  6 - free sulfur dioxide  7 - total sulfur dioxide  8 - density  9 - pH  10 - sulphates  11 - alcohol  Output variable (based on sensory data):  12 - quality (score between 0 and 10) | 1 - acidez fixa  2 - acidez volátil  3 - ácido cítrico  4 - açúcar residual  5 - cloretos  6 - dióxido de enxofre livre  7 - Dióxido de enxofre total  8 - densidade  9 - pH  10 - sulfatos  11 - álcool |

**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.

Disciplina: **Conceitos Estatísticos para IA**

Trabalho: **Uso e análise dos conceitos e técnicas estatísticas**

Utilizando a base descrita e disponibilizada em aula o objetivo do trabalho é mensurar a variável “Quality” dos vinhos desta região de Portugal com as variáveis de características (composição) dos vinhos.

Algumas etapas importantes:

1. Base

Reconhecimento da Base: Um descrito do tamanho da base, quantidade de variáveis, formatos, e alguns descritivos (“achados”) sobre a base/variáveis.

Avaliar se a análise será feita com os dois tipos de vinhos juntos ou se separaria por tipo para analisá-los.

Análise exploratória de dados: Detecção de outliers, gráficos e análise sobre os dois tipos de vinhos. Correlações entre elas (numéricas e gráficos).

* No fechamento desta etapa colocar qual a opção seguirá sobre os tipos de vinhos , sobre os outliers (caso tenha) e o uso de Componentes Principais

1. Algoritmos para explicar a variável “Quality”

* Modelo1: Regressão Linear
* Modelo 2: Árvore de Regressão
  + Para cada modelo fazer as análises adequadas como:
    - explicar a técnica
    - qual a variável dependente,
    - quais são as variáveis independentes,
    - relações entre elas (numéricas e gráficos) (verificar se todas já foram efetuadas adequadamente na parte 1.
    - saída do modelo (análise)
    - qualidade do modelo
    - O que cada modelo gerou de resultados?
  + Comparação entre os modelos:

Utilizando as métricas adequadas para comparação de modelos façam um resumo sobre a qualidade dos modelos e indiquem qual o modelo/ técnica que vocês recomendariam

1. Algoritmos para explicar a variável “Quality” : Vinhos Bons e Ruins

* Modelo 1: Árvore de Decisão
* Modelo 2: Regressão Logística
  + Para cada modelo fazer as análises adequadas como:
    - explicar a técnica
    - qual a variável dependente,
    - quais são as variáveis independentes,
    - relações entre elas (numéricas e gráficos) (verificar se todas já foram efetuadas adequadamente na parte 1.
    - saída do modelo (análise)
    - qualidade do modelo
    - O que cada modelo gerou de resultados?
  + Comparação entre os modelos:

Utilizando as métricas adequadas para comparação de modelos façam um resumo sobre a qualidade dos modelos e indiquem qual o modelo/ técnica que vocês recomendariam

1. Nesta atividade usamos somente algumas técnicas supervisionadas,
2. quais outras técnicas supervisionadas vocês indicariam como adequadas para esta análise?
3. e, das técnicas Não Supervisionadas, quais?