Mestrado em Engenharia Informática 2021/2022

Machine Learning

Avaliação prática – M1

João Fernando Malva Valente

João Pedro Neves Gonçalves

20 Novembro 2021

Sumário

Este documento apresenta uma descrição completa dos dados e a visualização dos mesmos. Irá ser detalhado todas as alterações feitas ao Dataset escolhido e as devidas justificações.

# Introdução

Esta primeira meta tem como objetivo produzir uma abordagem sistemática e organizada à preparação de um dataset para a sua posterior submissão a algoritmos de Supervised Learning e Unsupervised Learning.

Irá ser mostrado neste documento o dataset escolhido, assim como as alterações necessárias ao mesmo para ficar de acordo com os objetivos propostos. Também será descrito todos os dados relativos ao dataset e a sua visualização.

# Dataset

Decidiu-se pegar num dataset que retrata a qualidade de vinhos. O objetivo desta análise é que, dada um conjunto de características do vinho, se faça uma análise da qualidade do mesmo. Esta qualidade pode variar de 0 a 10, sendo que números acima de 7 se considera um vinho de qualidade, e menos de 7 um vinho com pouca qualidade.

Inicialmente eram dois datasets diferentes, um para vinho tinto, outro para vinho branco, mas devido à necessidade de ir de encontro com as restrições dadas, foi necessário juntar estes dois datasets e criar uma nova coluna, onde diz se o vinho é tinto ou branco.

Pode-se visualizar o dataset na Figura 1.

Uma imagem com mesa

Descrição gerada automaticamente

Figura 1: Dataset

## Descrição dos dados

Podemos ver na Figura 1, os vários dados relativos a este dataset, onde passarei a descrever nos seguintes tópicos:

* **Acidez fixa (Fixed\_acidity):** quantidade de ácidos envolvidos no vinho que sejam fixos ou não voláteis (não evaporam prontamente). É um atributo numérico.
* **Tipo de vinho (Red\_wine):** representa o tipo de vinho, podendo ser ele tinto ou branco. Se for vinho tinto, aparece na tabela com valor 1. É um atributo categórico.
* **Acidez volátil (Volatile\_acidity):** quantidade de ácido acético no vinho, que em níveis muito altos pode levar a um gosto desagradável de vinagre. É um atributo numérico.
* **Ácido cítrico (Citric\_acid):** encontrado em pequenas quantidades, o ácido cítrico pode adicionar “frescor” e sabor aos vinhos. É um atributo numérico.
* **Açúcar residual (Residual\_sugar):** quantidade de açúcar que resta após a fermentação do vinho. Vinhos com mais de 45 gramas/litro são considerados vinhos doces. É um atributo numérico.
* **Cloretos (Chlorides):** quantidade de sal no vinho. É um atributo numérico.
* **Dióxido de enxofre livre (Free\_sulfur\_dioxide):** quantidade de moléculas de SO2 em formas livres, sendo que estas e o ião bissulfito impede o crescimento microbiano e a oxidação do vinho. É um atributo numérico.
* **Dióxido de enxofre total (Total\_sulfur\_dioxide):** quantidade de moléculas de SO2 em formas livres e ligadas. Em baixas concentrações, o SO2 é praticamente indetetável no vinho, mas em concentrações acima de 50 ppm, o SO2 torna-se evidente no cheiro e sabor do vinho. É um atributo numérico.
* **Densidade (Density):** a densidade do vinho é próxima à da água, dependendo da percentagem de álcool e teor de açúcar. É um atributo numérico.
* **pH:** Descreve o quão ácido ou básico é o vinho numa escala de 0 a 14, sendo 0 muito ácido e 14 muito básico. É um atributo numérico.
* **Sulfatos (Sulphates):** Quantidade de aditivo que o vinho leva, que pode influenciar nos níveis de gás de SO2. É um atributo numérico.
* **Álcool (Alcohol):** Percentagem de álcool que o vinho tem. É um atributo ordinal.
* **Qualidade (Quality):** Valor numérico de 1 a 10, que mede a qualidade do vinho, sendo que valor acima de 7 é considerado um vinho com qualidade, e menor que 7 um vinho com pouca qualidade. É um atributo ordinal.

## Limpeza de dados

Realizou-se uma limpeza dos dados e concluiu-se que havia 1177 de dados duplicados, sendo que 240 eram vinhos tintos e 937 vinhos brancos. Além disso, houve uma normalização dos dados, ou seja, limitou-se os dados a um mínimo e um máximo para haver uma diminuição em processamentos futuros.