Datasheet for Wine Quality Dataset*

Zijun Meng

March 20, 2024

1 Introduction

This datasheet accompanies the Wine Quality dataset, documenting key aspects such as motivation, composition, collection process, and recommended uses. The dataset includes physicochemical properties and quality ratings for red and white wines from the Vinho Verde region in Portugal (C. Cortez Paulo and Reis 2009).

2 Motivation

The primary goal behind the creation of the Wine Quality dataset was to provide a detailed analysis of physicochemical tests as predictors of wine quality. This dataset was developed to support research in machine learning and data science, specifically targeting the agricultural and food industries. It aims to foster advancements in predictive modeling techniques for wine quality, contributing to the broader field of food quality assessment (P. Cortez et al. 2009).

3 Composition

The dataset is composed of two separate files for red and white wines, containing 1,599 and 4,898 instances, respectively. Each instance represents a single wine sample and includes the following 11 physicochemical attributes:

Fixed acidity Volatile acidity Citric acid Residual sugar Chlorides Free sulfur dioxide Total sulfur dioxide Density pH Sulphates Alcohol Additionally, each wine sample is assigned a quality score ranging from 0 (very poor) to 10 (excellent), based on sensory data.

*Code and data are available at: https://github.com/FrankMengZJ/datasheet-wine.git, reviewed by Kuiyao Qiao

4 Collection Process

The physicochemical data were collected through standardized laboratory tests performed on wine samples. The quality ratings were assigned by wine experts, who evaluated the wines according to sensory characteristics. The specific methodology for collecting the sensory evaluations and physicochemical tests has not been detailed in the dataset documentation.

5 Recommended Uses

The Wine Quality dataset is particularly suited for the following tasks:

Regression Analysis: Predicting the quality score of a wine based on its physicochemical properties. Classification: Categorizing wines into predefined quality classes (e.g., low, medium, high) based on their physicochemical attributes. Data Exploration: Analyzing relationships between different physicochemical properties and how they influence the sensory quality of wine. This dataset is recommended for educational purposes, as well as for researchers and practitioners in the fields of food science, chemistry, and machine learning.

6 Distribution and Access

The Wine Quality dataset is freely available for download from the UCI Machine Learning Repository (C. Cortez Paulo and Reis 2009). It is distributed under a license that permits non-commercial use, ensuring that researchers and educators can utilize the dataset without restrictions. Users are encouraged to cite the original source when using the dataset in their work.

7 Conclusion

The Wine Quality dataset serves as a valuable resource for exploring the relationship between physicochemical properties and wine quality. It offers a practical case study for various machine learning tasks, promoting research and education at the intersection of data science and food technology.

References

Cortez, Cerdeira, Paulo, and J. Reis. 2009. “Wine Quality.” UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

Cortez, P., Antonio Luíz Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. “Modeling Wine Preferences by Data Mining from Physicochemical Properties.” *Decis. Support Syst.* 47: 547–53. https://api.semanticscholar.org/CorpusID:2996254.