



Reviewed Dataset

118

Red Wine Quality

Simple and clean practice dataset for regression or classification modelling



UCI Machine Learning • last updated 6 months ago

1 Files (98.58 KB)

winequality-red.csv

winequality-red.csv

98.58 KB • Updated 6 months ago

About this file

Download

Download All

Context

This datasets is related to red variants of the Portuguese "Vinho Verde" wine. For more details, consult 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.).

The 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).

This dataset is also available from the UCI machine learning repository,

Overview

Data

Kernels

Discussion

Activity

Download (26 KB)

New Kernel

Content

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)

Tips

What might be an interesting thing to do, is aside from using regression modelling, is to set an arbitrary cutoff for your dependent variable (wine quality) at e.g. 7 or higher getting classified as 'good/1' and the remainder as 'not good/0'. This allows you to practice with hyper parameter tuning on e.g. decision tree algorithms looking at the ROC curve and the AUC value. Without doing any kind of feature engineering or overfitting you should be able to get an AUC of .88 (without even using random forest algorithm)

KNIME is a great tool (GUI) that can be used for this. 1 - File Reader (for csv) to linear correlation node and to interactive histogram for basic EDA. 2- File Reader to 'Rule Engine Node' to turn the 10 point scale to dichotome variable (good wine and rest), the code to put in the rule engine is something like this: - \$quality\$ > 6.5 => "good" - TRUE => "bad" 3- Rule Engine Node output to input of Column Filter node to filter out your original 10point feature (this prevent leaking) 4- Column Filter Node output to input of Partitioning Node (your standard train/test split, e.g. 75%/25%, choose 'random' or 'stratified') 5- Partitioning Node train data split output to input of Train data split to input Decision Tree Learner node and 6- Partitioning Node test data split output to input Decision Tree predictor Node 7- Decision Tree learner Node output to input Decision Tree Node input 8- Decision Tree output to input ROC Node.. (here you can evaluate your model base on AUC value)

Inspiration

Use machine learning to determine which physiochemical properties make a wine 'good'!

Acknowledgements

This dataset is also available from the UCI machine learning repository, <https://archive.ics.uci.edu/ml/datasets/wine+quality> , I just shared it to kaggle for convenience. (I am mistaken and the public license type disallowed me from doing so, I will take this down at first request. I am not the owner of this dataset.

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.

Relevant publication



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.

Preview (first 100 rows) Column Metadata Column Metrics 

fixed acidity	most acids involved with wine or fixed or nonvolatile (do not evaporate readily)	Numeric
volatile acidity	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste	Numeric
citric acid	found in small quantities, citric acid can add 'freshness' and flavor to wines	Numeric
residual sugar	the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet	Numeric
chlorides	the amount of salt in the wine	Numeric
free sulfur dioxide	the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine	Numeric
total sulfur dioxide	amount of free and bound forms of SO2; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine	Numeric

 `kaggle datasets download -d uciml/red-wine-quality-cortez-et-al-2009`  

Versions

-  **Version 2** 6 months ago Fixed csv format to use comma as delimiter
-  **Version 1** 6 months ago Initial release

