

Task:

In this folder you will find a CSV file with around 1600 data sets, which contain, on the one hand, chemical or physical properties of red wines (various acids, chlorides, etc.) and, on the other hand, their alcohol content in per mille as well as their quality (scale of 0 ("Fusel") ) to 10 ("top wine").

Please implement a mini-machine learning project that results in a decision tree model to explain which properties make up:

a) a high quality wine

b) a wine with a very high alcohol content

Please note the following information:

Project phases:

-> Data Understanding (inspect, load, visualize data)

-> Data Cleansing (clean data, address quality problems)

-> Control: Can I make hypotheses about connections based on my understanding of the data?

-> Modeling: Is successful when the model answers the client's questions and confirms or refutes hypotheses

Please cover all project phases in detail.

The task is deliberately formulated in general terms; this is very typical of similar requests from companies to data analysts. Only as the project progresses is it often possible to define the goal more clearly. Therefore, write comments defining the general goals in detail at the end.

Don't forget to test (fitting analysis): Please divide your data appropriately into training and test data. Check how well your model predicts wine quality or alcohol content for the test data.

Does this roughly correspond to the training results or is there overfitting?

Due to the open topic, there is no standard solution; rather, you receive additional individual feedback with the points on the eLearning platform

Comments: Please write a comment in each code block that explains your thoughts and intentions. What is assessed is their understanding of the task and the intermediate results, not the code.