PREDICTION OF WINE QUALITY

Petr Feilhauer

Doctoral Degree Programme (1), FEEC BUT

E-mail: xfeilh00@stud.feec.vutbr.cz

Supervised by: František Zezulka

E-mail: zezulka@feec.vutbr.cz

Abstract: This article deals with the design and creation of models for predicting wine quality. The basis for creating models for predicting wine quality is the measured physic-chemical properties. The aim of this experimental project is to create the models capable of prediction quality wines with as much accuracy as possible using the minimal measured properties. Component of this article is validation of obtained predictions, or apply appropriate adjustments to achieve similar or better prediction quality wines than the available studies dealing with the same examined data set.

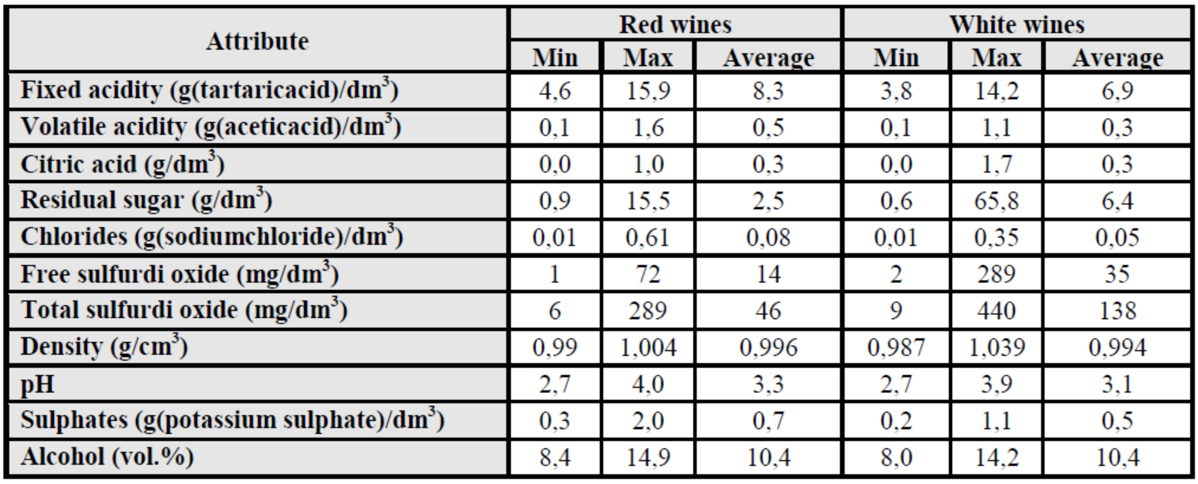
Keywords: MAD, prediction, rapid miner, support vector machine

1. INTRODUCTION

For the purposes of creation the prediction models was used data measured on real samples of wines, which were evaluated by experts. The measured data are available for download on the Internet [4]. In the article [3], the authors refer to the basic statistical data (input attributes) of measured data, which are listed in Table 2.1. To verify that the data available on the Internet are identical with the data described in article, were made the same statistical calculations and their results are shown in Table 2.2. When comparing the values in these tables shows that the data are consistent with the data on which predictions were made in a technical article [3].

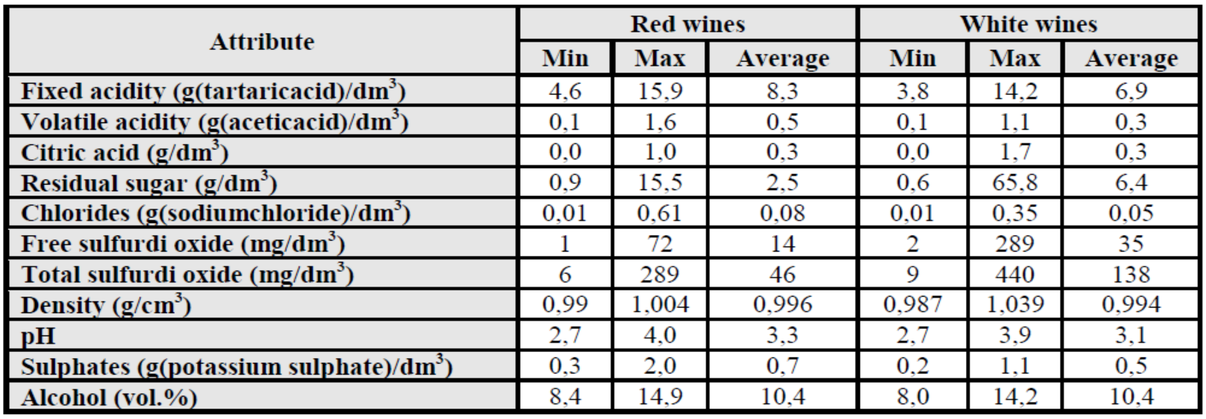
2. INPUT DATA

For creation the models, we have two data stacks. The first data stack describes the red wine - 1599 samples of red wines. The second data stack describes the white wines - 4898 samples of white wines. For both types is for each sample is measured eleven physic-chemical parameters (inputs of model) and recorded its quality (output of model). The quality of wines is evaluated using a scale from 0 to 10 by integers.



Attribute	Red wines			White wines		
	Min	Max	Average	Min	Max	Average
Fixed acidity (g(tartaricacid)/dm^3)	4,6	15,9	8,3	3,8	14,2	6,9
Volatile acidity (g(aceticacid)/dm^3)	0,1	1,6	0,5	0,1	1,1	0,3
Citric acid (g/dm^3)	0,0	1,0	0,3	0,0	1,7	0,3
Residual sugar (g/dm^3)	0,9	15,5	2,5	0,6	65,8	6,4
Chlorides (g(sodiumchloride)/dm^3)	0,01	0,61	0,08	0,01	0,35	0,05
Free sulfurdi oxide (mg/dm^3)	1	72	14	2	289	35
Total sulfurdi oxide (mg/dm^3)	6	289	46	9	440	138
Density (g/cm^3)	0,99	1,004	0,996	0,987	1,039	0,994
pH	2,7	4,0	3,3	2,7	3,9	3,1
Sulphates (g(potassium sulphate)/dm^3)	0,3	2,0	0,7	0,2	1,1	0,5
Alcohol (vol.%)	8,4	14,9	10,4	8,0	14,2	10,4

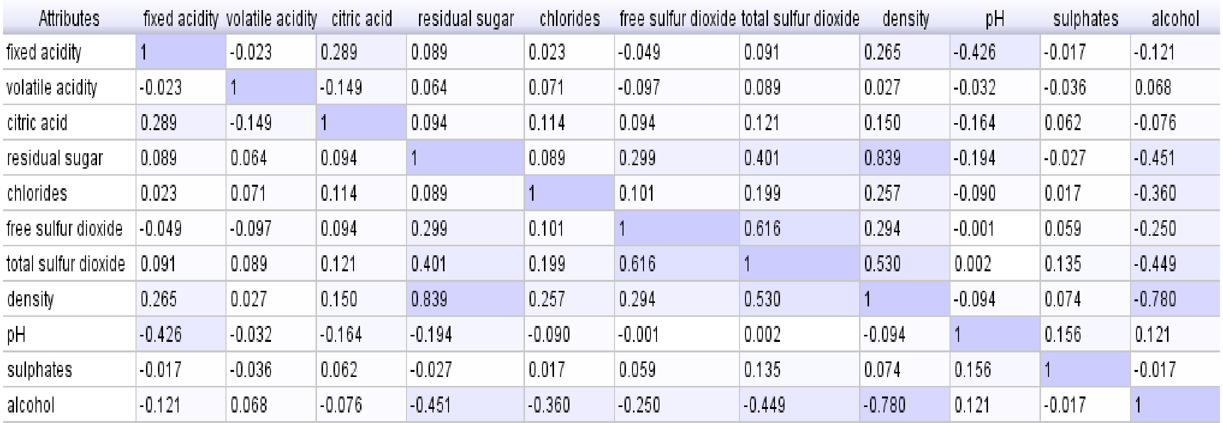
Table 2.1: Basic statistics of inputs attributes referred in article [3]



Attribute	Red wines			White wines		
	Min	Max	Average	Min	Max	Average
Fixed acidity (g(tartaricacid)/dm^3)	4,6	15,9	8,3	3,8	14,2	6,9
Volatile acidity (g(aceticacid)/dm^3)	0,1	1,6	0,5	0,1	1,1	0,3
Citric acid (g/dm^3)	0,0	1,0	0,3	0,0	1,7	0,3
Residual sugar (g/dm^3)	0,9	15,5	2,5	0,6	65,8	6,4
Chlorides (g(sodiumchloride)/dm^3)	0,01	0,61	0,08	0,01	0,35	0,05
Free sulfurdi oxide (mg/dm^3)	1	72	14	2	289	35
Total sulfurdi oxide (mg/dm^3)	6	289	46	9	440	138
Density (g/cm^3)	0,99	1,004	0,996	0,987	1,039	0,994
pH	2,7	4,0	3,3	2,7	3,9	3,1
Sulphates (g(potassium sulphate)/dm^3)	0,3	2,0	0,7	0,2	1,1	0,5
Alcohol (vol.%)	8,4	14,9	10,4	8,0	14,2	10,4

Table 2.2: Basic statistics of inputs attributes of downloaded data stack [4]

For the identification of dependencies between attributes of processed data, were created correlation matrices for data stack of the red and white wines data that describe the interdependence of individual attributes. Mutual correlation dependencies are shown for example only for red wines - Table 2.3. Correlation of the same attributes (elements on the main diagonal) is always one.



Attributes	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
fixed acidity	1	-0.023	0.289	0.089	0.023	-0.049	0.091	0.265	-0.426	-0.017	-0.121
volatile acidity	-0.023	1	-0.149	0.064	0.071	-0.097	0.089	0.027	-0.032	-0.036	0.068
citric acid	0.289	-0.149	1	0.094	0.114	0.094	0.121	0.150	-0.164	0.062	-0.076
residual sugar	0.089	0.064	0.094	1	0.089	0.299	0.401	0.839	-0.194	-0.027	-0.451
chlorides	0.023	0.071	0.114	0.089	1	0.101	0.199	0.257	-0.090	0.017	-0.360
free sulfur dioxide	-0.049	-0.097	0.094	0.299	0.101	1	0.616	0.294	-0.001	0.059	-0.250
total sulfur dioxide	0.091	0.089	0.121	0.401	0.199	0.616	1	0.530	0.002	0.135	-0.449
density	0.265	0.027	0.150	0.839	0.257	0.294	0.530	1	-0.094	0.074	-0.780
pH	-0.426	-0.032	-0.164	-0.194	-0.090	-0.001	0.002	-0.094	1	0.156	0.121
sulphates	-0.017	-0.036	0.062	-0.027	0.017	0.059	0.135	0.074	0.156	1	-0.017
alcohol	-0.121	0.068	-0.076	-0.451	-0.360	-0.250	-0.449	-0.780	0.121	-0.017	1

Table 2.3: Correlation dependence of the input attributes – red wines [4]

From Table 2.3 shows that the correlation of certain attributes reach values outside the interval (-0.5, 0.5), it indicating considerable interdependence. Highly correlated input attributes may be omitted to simplify the model, improvement of prediction or acceleration of model. In this table are marked size interdependencies of attributes with intensity of blue color.

3. EVALUATION OF THE QUALITY PREDICTION

From the reason to compare the quality of proposed models, were comparison of models made using the average absolute deviation (MAD) for each model, which is obtained from the difference between actual and predicted values of quality wines.

The average absolute deviation is given by:

$$MAD = \sum_i^N |y_i - \hat{y}_i| / N \qquad (3.1)$$

$\hat{y}_i$ - estimated quality for the *k*-th sample of wine

y_i - real quality for the *k*-th sample of wine

N - number of submitted samples of wine

The models were then compared with the accuracy of classification in each classes of wine samples. In order to compare with [3], so classes were chosen with a tolerance of T = 0.25, 0.5, 1. If T = 0.25 and the value of good quality wine is 4, the correct prediction of the class has a range of 3.75 to 4.25, poor prediction is a prediction outside this range. For T = 1 the intervals are overlap. Since quality is reported in whole numbers, is our greatest information benefit of predicting tolerance of size 0.5 (classes are follow each other and do not create "dead" space) and 1. In the case of a tolerance is 1 within a ten-value scale do not affect (over-estimation or underestimation) the quality of the sample.

For T = 0.5 were created for all models the cost matrices. These matrices are not listed given the scope of this document.

4. PREDICTIVE MODELS

Due to reproduction models used in [3], were created three types of models:

- MR – linear regression
- NN – artificial neural network
- SVM – support vector machine

When creating all models were used according to [1] method of estimating error the cross-validation, which was performed the same settings.

Input attributes of models for each of the 17 different models are different:

- All attributes used without modification
- Before entering into the model was performed normalization, backward selection, resp. forward selection with implemented SVM method or linear regression
- Removal of the input attributes based on weight from genetic algorithms
- Removal of the input attributes based on the correlation matrix (Table 2.3)
- Random removal of input attributes

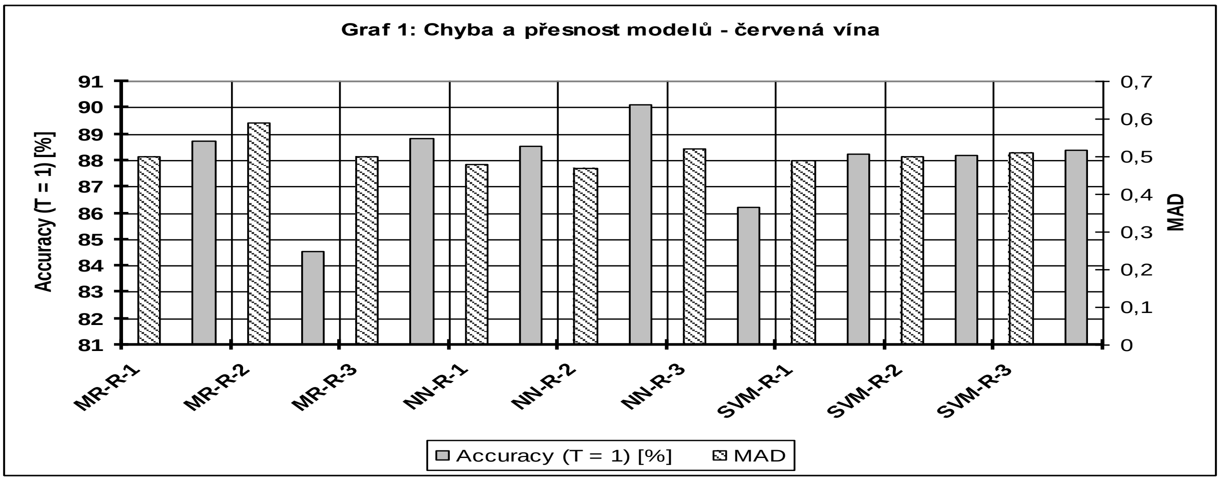
Train and test data are always same for the one model (always use the entire suite of samples), only may vary by the deleted attributes. Test data is used only in order to obtain predictions for the external vector processing in MS Excel.

Error MAD and accuracy of individual models for tolerance T = 1 is for better clarity and evaluation plotted in Graph 4.1 (red wine). Values for white wine are not given because the scope of this document is limited. From this graph it is clear that the MAD error is higher, the accuracy of prediction models is lower and vice versa.

The best accuracy of prediction (90.12 ± 0.44%) for red wine follows from the model NN-R-2, which uses eight input attributes. The best accuracy of prediction (90.51 ± 0.46%) for white wine follows the model NN-W-1, which uses all eleven input attributes.

In annex - Comparison of different models - Table 6.1, are listed the errors and accuracy of prediction models for different tolerances. The values listed in the part "article" cannot accurately determine whether the accuracy and errors are from diameter of 20 designed models or from the best designed model.

As stated in the introduction, is not necessary that the accuracy at T = 0.25 was too high. In terms of determining the quality of wine is important T = 0.5 or T = 1, because these values have no significant effect on wine quality - within the ten-scale is the minimum deviation.



Graph 4.1: Accuracy and error of models – red wines (drawn up by RapidMiner)

Attributes	MR-R-1	MR-R-2	MR-R-3	NN-R-1	NN-R-2	NN-R-3	SVM-R-1	SVM-R-2	SVM-R-3
Fixed acidity	-	-	-	1	1	1	1	1	1
Volatile acidity	1	1	1	1	1	1	1	1	1
Citric acid	-	-	-	1	-	-	1	-	-
Residual sugar	-	-	-	-	1	-	-	-	-
Chlorides	1	-	1	1	-	-	1	-	-
Free sulfur dioxide	1	-	-	-	1	-	1	1	-
Total sulfur dioxide	1	-	1	-	1	-	1	1	-
Density	-	-	-	1	-	-	1	-	-
pH	1	1	1	1	1	1	1	-	-
Sulphates	1	1	1	1	1	1	1	1	1
Alcohol	1	-	1	1	1	1	1	1	1

Table 4.1: Attributes used for creating the model – red wines (drawn up by RapidMiner)

*Note.: **1** = attribute is used; **-** = attribute is not used or eliminated*

The quality of the prediction each models cannot be assessed only in terms of maximum accuracy, but also in terms of number the parameters, which were used for this accuracy. In Table 4.1 are individual attributes, that were used for designed the models.

5. CONCLUSION

The aim of this project was to reproduce the experimental models used in [3], verify the quality of obtained predictions, or appropriate adjustments to achieve similar or better prediction quality wines in the examined data set. By studying the article [3] and the processing of the data we have shown that our obtained data sets are consistent with data sets that were used to build models in [3]. To determine the interdependencies between the inputs attributes for the red and white wines were calculated correlation matrices. Some perceived dependence was applied in reduction of input attributes. Models for predicting wine quality were used by linear regression, artificial neural networks and support vector machine. For all models was applied the method of cross-validation according to [3]. In total were described 17 models with the best or with most interesting results.

Comparison of our proposed models and models published in [3] is problematic, because in [3] is not stated whether the indicated errors and accuracy of models are the average values of the number of simulations or the best value achieved - see above. Evaluation of the quality of models and comparison with [3] was performed using the average absolute deviation (MAD) and the accuracy of classification into different classes for different tolerances.

In Tables 6.1 and 4.1 are the errors, the accuracy of models and selected attributes that were used in design of models. These tables show that for red wine we get better prediction using the same or a smaller number of inputs attributes than for white wines. This situation probably occurred by that for the white wine wasn't found the best combination of attributes. The attributes that we applied had a low information gain or dependency param. on the quality of red wine is simpler.

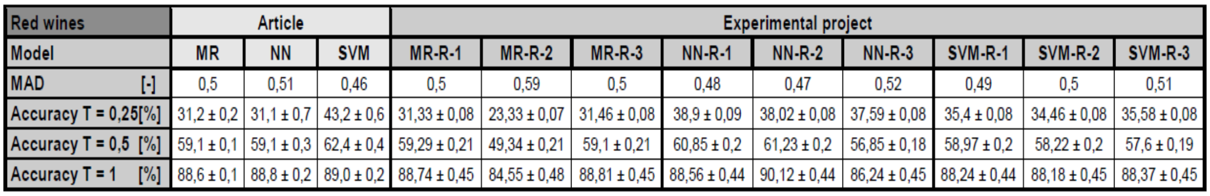
For models with linear regression and with artificial neural networks for red and white wines were obtained similar or better prediction accuracy than in [3] (improving the quality of prediction is probably not statistically significant). It is necessary to point out that in [3] the quality of predictions was achieved for the average number of input attributes from 9 to 10, but this project for similar results obtained from 3 to 8 input attributes (with the exception of the NN-W-1).

The most interesting model can give a linear regression model MR-R-2, which, while it reaches a higher MAD error and tolerance is less than for T = 0.25 and T = 0.5 such accuracy as other models, but for predicting quality of wine it have only three measurements physic-chemical attributes. Accuracy of prediction for T = 1 is 84.55 ± 0.48%, other models reach values up to 90%, but using a larger number of attributes. This model due to the small number of input attributes needs a low cost for measuring physic-chemical parameters of each samples and is thus very suitable for determining the quality of wine, which at T = 1 there is no significant overestimation or underestimation of the quality of wine.

The model with the highest prediction accuracy and with the lowest MAD error for red wine is an artificial neural network NN-R-2 (white NN-W-1), which for tolerance T = 1 achieves an accuracy of prediction of 90.12 ± 0.44% (white 90.51 ± 0.46%). For other tolerances also achieves the best prediction accuracy compared to other models. For the design of this model was used 8 attributes (white 11).

All models were designed and tuned in the program RapidMiner version 5.0.003.

6. ANNEX – COMPARISON OF DIFFERENT MODELS



Red wines	Article			Experimental project								
Model	MR	NN	SVM	MR-R-1	MR-R-2	MR-R-3	NN-R-1	NN-R-2	NN-R-3	SVM-R-1	SVM-R-2	SVM-R-3
MAD [-]	0,5	0,51	0,46	0,5	0,59	0,5	0,48	0,47	0,52	0,49	0,5	0,51
Accuracy T = 0,25[%]	31,2 ± 0,2	31,1 ± 0,7	43,2 ± 0,6	31,33 ± 0,08	23,33 ± 0,07	31,46 ± 0,08	38,9 ± 0,09	38,02 ± 0,08	37,59 ± 0,08	35,4 ± 0,08	34,46 ± 0,08	35,58 ± 0,08
Accuracy T = 0,5 [%]	59,1 ± 0,1	59,1 ± 0,3	62,4 ± 0,4	59,29 ± 0,21	49,34 ± 0,21	59,1 ± 0,21	60,85 ± 0,2	61,23 ± 0,2	56,85 ± 0,18	58,97 ± 0,2	58,22 ± 0,2	57,6 ± 0,19
Accuracy T = 1 [%]	88,6 ± 0,1	88,8 ± 0,2	89,0 ± 0,2	88,74 ± 0,45	84,55 ± 0,48	88,81 ± 0,45	88,56 ± 0,44	90,12 ± 0,44	86,24 ± 0,45	88,24 ± 0,44	88,18 ± 0,45	88,37 ± 0,45

Table 6.1: Comparison of models – red wines (drawn up by RapidMiner)

REFERENCES

[1] HONZÍK, P.: Strojové učení. Brno: VUT, FEKT, 2006. 85s.

[2] BERRY M.J.A., LINOFF G.S.: Data Mining Techniques, Wiley Publishing, Inc., 2004. ISBN 0-471-47064-3.

[3] Modeling wine preference by data mining from physicochemical properties,, Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, José Reis, Received 28 July 2008, Available online 6 June 2009, Portugal (http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V8S-4WGK6HF-2&_user=10&_coverDate=11%2F30%2F2009&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_searchStrId=1238790273&_rerunOrigin=google&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=7bbe8ce3d0bd6f00b84e2a1d12726a1e)

[4] *Wine Quality Data Set*, Paulo Cortez, University of Minho, Guimarães, Portugal, 2009-10-07 (archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/)