# A DATA MINING APPROACH TO WINE QUALITY PREDICTION

**Dragana Radosavljević**
*Faculty of Technical Sciences of University in Pristina, Kosovska Mitrovica, Serbia*

**Siniša Ilić**
*Faculty of Technical Sciences of
University in Pristina,
Kosovska Mitrovica, Serbia*

**Stefan Pitulić**
*Faculty of Technical Sciences of
University in Pristina,
Kosovska Mitrovica, Serbia*

**Abstract**

*Many people choose the wine based on its rating. The quality of the wine is set by organoleptic (tasting-sensory) and chemical analysis. Organoleptic characteristics are of primary importance for wine quality, while chemical analysis is performed for market control of the wine. The basic idea presented in this paper is to categorize wine only on the basis of its physicochemical properties. The research is performed using the vinho verde wine database from the northwest of Portugal, taken from the Center for Machine Learning and Intelligent Systems (UCI Machine Learning) website. The following classification algorithms were used to estimate the quality of the wine: Decision Tree, Random Forest, Algorithm k star, Support Vector Machine, Multilayer perceptron, and Naïve Bayes Classifier. Comparing the results of these algorithms, it can be seen that the Random Forest algorithm provides promising results that could potentially lead to conclusions that would be useful for future application in wine quality evaluation.*

**Keywords:** Data Mining, Wine, Classification, Decision Tree, Random Forest, k star, Support Vector Machine, Neural Network, Naïve Bayes.

## INTRODUCTION

Wine occupies an important place in human life from the era of the ancient Greeks and Romans until today. It is an integral part of religious life. It was made immortal by poets, historians, philosophers, artists. Describing the taste of wine, and therefore assessing its quality, is a real challenge, but also a science.

It is difficult to define the quality of the wine, as it is a multi-faceted construct, lacking a uniform and generally accepted definition [1]

To determine the quality of the wine sensory tests are used, which rely on human experts' knowledge, but physicochemical properties of wine can also be used.

The experts (sommeliers) use their vast experience to evaluate the quality of the wine. Depending on the scoring system used by wine judges, wine can be ranked on scales of 0-10, 0-20 and 0-100 points [2]. Their grades are subjective and it is very difficult or almost impossible to define it precisely. Research shows that only about 10% of judges can repeat their wine rating within one group of medals [3].

Also, the relationships between physicochemical and sensory analysis are complex and not yet fully understood, but significant correlations can be found between the quality of wine and some of its physicochemical properties [1, 4]

Investing in new technologies in the wine production process allows wineries to maintain the quality of production, and thus to secure their place in the wine market. Sensors are already being used in the wine industry to collect and monitor different data, from those related to the condition of the vine plant to those that monitor the entire wine production process. The question is: How to extract some useful knowledge from such large-scale, often very complex, datasets?

Data mining (DM) techniques are a powerful tool that allows to easily analyze relationships between different attributes of datasets. These techniques can be used for classification, clustering, forecasting, optimization and summarization.

In the wine industry, DM is used to make recommendations on the purchase of wine, based on expert wine ratings, consumer criticisms and wine prices. There are a large number of websites and mobile applications that make recommendations for choosing wines based on that information (e.g. www.go-wine.com, www.cnbc.com, www.wine-searcher.com, the Vivino app, etc.).

Despite its potential to “predict” wine quality based on the physicochemical properties, DM techniques are not often used in this task.

In this paper we present a classification of wines based on their physicochemical properties that are easily measurable and accessible. This analysis can be valuable to wine producers (to improve the production process), to consumers (to select wine), but it can also be used by experts to support their evaluation of wine and to potentially improve the speed and quality of their decisions.

## MATERIALS AND METHODS

### Dataset and features

The data set is a wine quality dataset that is publicly available for research purposes from UCI – Machine Learning [5]. The dataset contains real data on vinho verde wines from the northwest of Portugal. The dataset consists of 4898 white wine and 1599 red wine samples. Data were collected from May 2004. to February 2007 using an iLab computerized system that automatically manages the process of wine samples testing starting from the manufacturer's requirements to laboratory and sensory analysis [4].

Table 1 presents the physicochemical statistics for each dataset.

*Table 1: The physicochemical data statistics per wine type*

| | All wine (6497 instances) | | White wine (4898 instances) | | Red wine (1599 instances) | |
|---|---|---|---|---|---|---|
| | Min | Max | Min | Max | Min | Max |
| fixed acidity (*g/dm*$^3$) | 3.80 | 15.9 | 3.80 | 14.2 | 4.60 | 15.9 |
| volatile acidity (*g/dm*$^3$) | 0.08 | 1.58 | 0.08 | 1.10 | 0.12 | 1.58 |
| citric acid (*g/dm*$^3$) | 0.00 | 1.66 | 0.00 | 1.66 | 0.00 | 1.00 |
| residual sugar (*g/dm*$^3$) | 0.6 | 65.8 | 0.60 | 65.8 | 0.9 | 15.5 |
| chlorides | 0.01 | 0.61 | 0.01 | 0.35 | 0.01 | 0.61 |
| free sulfur dioxide (*mg/dm*$^3$) | 1 | 289 | 2 | 289 | 1 | 72 |
| total sulfur dioxide (*mg/dm*$^3$) | 6 | 440 | 9 | 440 | 6 | 289 |
| density (*g/cm*$^3$) | 0.99 | 1.04 | 0.99 | 1.04 | 0.99 | 1.00 |
| ph | 2.72 | 4.01 | 2.72 | 3.82 | 2.74 | 4.01 |
| sulphates (*g/dm*$^3$) | 0.22 | 2.00 | 0.22 | 1.08 | 0.33 | 2.00 |
| alcohol *(% vol.*) | 8.0 | 14.9 | 8.0 | 14.2 | 8.4 | 14.9 |
| quality | 3.00 | 9.00 | 3.00 | 9.00 | 3.00 | 8.00 |

## DATA MINING TECHNIQUES

The WEKA open source software was used for data processing (preprocessing and classification). The following classification algorithms were used to solve the classification tasks: Decision Tree, Random Forest, Algorithm k star, Support Vector Machine, Multilayer perceptron and Naïve Bayes Classifier.

**Decision Tree (J48) -** is an open source Java implementation of C4.5 decision tree algorithm. Decision tree is constructed in a top-down recursive divide-and-conquer manner, where the internal nodes of a decision tree denote the different attributes, the branches between the nodes tell us the possible values that these attributes can have in the observed samples and the terminal nodes tell us the final value (class).

This type of algorithm is very robust in case of missing data and allows combining numerical and categorical attribute values [6].

**Random forest (RF)** is a classifier based on many decision trees. It is based on a simple idea: 'the wisdom of the crowd'. Aggregate of the results of multiple predictors gives a better prediction than the best individual predictor. Random Forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [7].

**Naïve Bayes (NB) -** The Naïve Bayesian classifier is a classification technique based on Bayes’ Theorem with an assumption of independence between predictors. It is called naive because the assumption of conditional independence in practice is generally not valid. Parameter estimation for naive Bayes models uses the method of maximum likelihood. Some of the major advantages of this method is that it needs a small amount of data and it is resistant to noise.

**Multilayer perceptron (MLP)** is the most commonly used architecture of artificial neural networks. MLP is a feedforward Neural Network model that maps the random data set to a set of corresponding outputs. It contains a large number of nodes – neurons. Neurons are arranged in multiple layers (input layer, hidden layer and output layer), where each neuron of one layer is linked by weighted connections with all neurons of the next layer. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples [8].

**Support vector machine (SVM)** is based on a simple idea: to define a hyperplane (in boundless dimensional space) that separates data into the appropriate classes. In addition to performing linear classification, SVM can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces where they can be linearly separated and the optimal hyperplane be determined [9].

Sequential Minimal Optimization (SMO) implements John Platt's sequential minimal optimization algorithm for training a support vector classifier [10].

**K star (K*)** is an Instance-Based (IB) classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function.[10] The principal difference of K* against other IB algorithms is the use of the entropy concept for defining its distance metric, which is calculated by mean of the complexity of transforming an instance into another; so it is taken into account the probability of this transformation occurs in a “random walk away” manner [11].

## DATA MINING ESTIMATION

There is no fixed algorithm to provide high accuracy; this is called No Free lunch theorem [13].

Several experiments were conducted to evaluate the performance of the selected tool used on a given dataset. Evaluation is done in two ways:

- using the 10-fold cross-validation method
- using the test sample method. A complete set of samples was divided into training and test set (2/3: 1/3) using the RESAMPLE filter (implemented in the WEKA software package)

The test of significance was taken as 0.05. To evaluate the effects of algorithms, some standard performance measures are also calculated: percent Correctly Classified Instances, Kappa statistics (kappa), Mean absolute error (MAE), Root mean squared error (RMSE), Recall and ROC Area.

## EXPERIMENT AND RESULTS

In our study, 6 classification algorithms were used to wine samples classification. The model was built using each of the methods and is applied to:

1. a complete wine dataset
2. white wines data set
3. red wine data set.

All datasets are categorized in 2 ways following different ways of wine quality evaluating [14,15] (Table 2).

*Table 2: The ways of wine categorizing according to the given quality rating*

| quality | Categorization 1 | Categorization 2 |
|---|---|---|
| 3 | Poor | Poor |
| 4 | | |
| 5 | fair | Commended |
| 6 | Commended | |
| 7 | Bronze | Medal |
| 8 | Silver | |
| 9 | Gold | |

The classification results of all the above algorithms are evaluated in both test methods: 10-way cross-validation and test sample method. To evaluate the performance of the classifiers, standard performance measures are calculated. Some of these are shown in the following tables.

*Table 3: Performance of wine classification results - Categorization 1 in cross-validation mode*

| Categorization 1 | | J48 | RF | NB | MP | SMO | k* |
|---|---|---|---|---|---|---|---|
| Correctly Classified Instances (%) | All wine | 69.91 | 70.57 | 48.59 | 55.61 | 53.29 | 65.91 |
| | White wine | 58.53 | 70.05 | 49.04 | 55.12 | 52.08 | 65.50 |
| | Red wine | 61.60 | 70.17 | 59.41 | 60.54 | 58.04 | 65.98 |
| Kappa statistic | All wine | 0.53 | 0.54 | 0.24 | 0.29 | 0.22 | 0.49 |
| | White wine | 0.38 | 0.54 | 0.27 | 0.28 | 0.19 | 0.48 |
| | Red wine | 0.40 | 0.52 | 0.35 | 0.36 | 0.29 | 0.47 |
| Mean absolute error | All wine | 0.15 | 0.15 | 0.19 | 0.19 | 0.24 | 0.12 |
| | White wine | 0.15 | 0.15 | 0.19 | 0.19 | 0.24 | 0.12 |
| | Red wine | 0.16 | 0.17 | 0.19 | 0.19 | 0.26 | 0.14 |
| Root mean squared error | All wine | 0.26 | 0.26 | 0.34 | 0.31 | 0.33 | 0.30 |
| | White wine | 0.35 | 0.26 | 0.34 | 0.31 | 0.33 | 0.30 |
| | Red wine | 0.37 | 0.28 | 0.33 | 0.33 | 0.35 | 0.34 |
| Recall | All wine | 0.70 | 0.71 | 0.49 | 0.56 | 0.53 | 0.66 |
| | White wine | 0.59 | 0.70 | 0.49 | 0.55 | 0.52 | 0.66 |
| | Red wine | 0.62 | 0.70 | 0.59 | 0.61 | 0.58 | 0.66 |
| ROC Area | All wine | 0.88 | 0.88 | 0.69 | 0.73 | 0.68 | 0.86 |
| | White wine | 0.72 | 0.87 | 0.70 | 0.72 | 0.67 | 0.85 |
| | Red wine | 0.72 | 0.87 | 0.76 | 0.75 | 0.70 | 0.85 |

*Table 4: Performance of wine classification results - Categorization 2 in cross-validation mode*

| Categorization 2 | | J48 | RF | NB | MP | SMO | k* |
|---|---|---|---|---|---|---|---|
| Correctly Classified Instances (%) | All wine | 78.03 | 85.59 | 67.64 | 77.52 | 74.62 | 82.56 |
| | White wine | 78.36 | 85.73 | 72.91 | 77.52 | 74.62 | 82.56 |
| | Red wine | 83.74 | 87.24 | 82.99 | 83.86 | 82.49 | 84.49 |
| Kappa statistic | All wine | 0.36 | 0.58 | 0.32 | 0.28 | 0.00 | 0.55 |
| | White wine | 0.43 | 0.59 | 0.35 | 0.28 | 0.00 | 0.55 |
| | Red wine | 0.40 | 0.48 | 0.40 | 0.37 | 0.00 | 0.45 |
| Mean absolute error | All wine | 0.19 | 0.16 | 0.24 | 0.20 | 0.28 | 0.12 |
| | White wine | 0.17 | 0.16 | 0.20 | 0.20 | 0.28 | 0.12 |
| | Red wine | 0.14 | 0.13 | 0.13 | 0.14 | 0.26 | 0.11 |
| Root mean squared error | All wine | 0.34 | 0.27 | 0.40 | 0.33 | 0.37 | 0.31 |
| | White wine | 0.35 | 0.27 | 0.37 | 0.33 | 0.37 | 0.31 |
| | Red wine | 0.31 | 0.25 | 0.30 | 0.29 | 0.34 | 0.30 |
| Recall | All wine | 0.78 | 0.86 | 0.68 | 0.78 | 0.75 | 0.83 |
| | White wine | 0.78 | 0.86 | 0.73 | 0.78 | 0.75 | 0.83 |
| | Red wine | 0.84 | 0.87 | 0.83 | 0.84 | 0.83 | 0.83 |
| ROC Area | All wine | 0.76 | 0.90 | 0.73 | 0.77 | 0.51 | 0.89 |
| | White wine | 0.75 | 0.90 | 0.76 | 0.77 | 0.51 | 0.89 |
| | Red wine | 0.70 | 0.88 | 0.80 | 0.79 | 0.54 | 0.44 |

*Table 5: Performance of wine classification results - Categorization 1 in test sample mode*

| Categorization 1 | | | J48 | RF | NB | MP | SMO | k* |
|---|---|---|---|---|---|---|---|---|
| Correctly Classified Instances (%) | All wine | Training | 91.16 | 100.00 | 50.67 | 57.07 | 53.53 | 100.00 |
| | | Test | 57.85 | 68.00 | 48.87 | 55.64 | 53.23 | 64.72 |
| | White wine | Training | 90.02 | 100.00 | 50.88 | 56.39 | 51.63 | 100.00 |
| | | Test | 57.14 | 66.80 | 48.98 | 57.01 | 54.42 | 61.16 |
| | Red wine | Training | 87.22 | 100.00 | 64.25 | 64.52 | 58.80 | 100.00 |
| | | Test | 60.63 | 68.33 | 59.38 | 62.92 | 57.08 | 61.67 |
| Kappa statistic | All wine | Training | 0.87 | 1.00 | 0.27 | 0.32 | 0.22 | 1.00 |
| | | Test | 0.37 | 0.51 | 0.24 | 0.31 | 0.23 | 0.47 |
| | White wine | Training | 0.85 | 1.00 | 0.30 | 0.29 | 0.19 | 1.00 |
| | | Test | 0.35 | 0.48 | 0.27 | 0.28 | 0.23 | 0.42 |
| | Red wine | Training | 0.80 | 1.00 | 0.42 | 0.43 | 0.30 | 1.00 |
| | | Test | 0.39 | 0.49 | 0.35 | 0.41 | 0.28 | 0.40 |
| Mean absolute error | All wine | Training | 0.04 | 0.06 | 0.18 | 0.19 | 0.24 | 0.00 |
| | | Test | 0.15 | 0.15 | 0.19 | 0.19 | 0.24 | 0.12 |
| | White wine | Training | 0.05 | 0.06 | 0.18 | 0.19 | 0.24 | 0.00 |
| | | Test | 0.15 | 0.15 | 0.19 | 0.19 | 0.24 | 0.13 |
| | Red wine | Training | 0.08 | 0.06 | 0.17 | 0.19 | 0.26 | 0.00 |
| | | Test | 0.17 | 0.18 | 0.19 | 0.19 | 0.26 | 0.16 |
| Root mean squared error | All wine | Training | 0.15 | 0.10 | 0.33 | 0.31 | 0.33 | 0.01 |
| | | Test | 0.35 | 0.27 | 0.34 | 0.31 | 0.33 | 0.31 |
| | White wine | Training | 0.16 | 0.10 | 0.33 | 0.31 | 0.33 | 0.01 |
| | | Test | 0.35 | 0.27 | 0.34 | 0.31 | 0.33 | 0.32 |
| | Red wine | Training | 0.20 | 0.11 | 0.32 | 0.31 | 0.35 | 0.00 |
| | | Test | 0.36 | 0.29 | 0.34 | 0.33 | 0.35 | 0.36 |
| ROC Area | All wine | Training | 0.99 | 1.00 | 0.71 | 0.75 | 0.68 | 1.00 |
| | | Test | 0.71 | 0.86 | 0.69 | 0.74 | 0.69 | 0.84 |
| | White wine | Training | 0.98 | 1.00 | 0.73 | 0.74 | 0.67 | 1.00 |
| | | Test | 0.71 | 0.85 | 0.71 | 0.74 | 0.68 | 0.83 |
| | All wine | Training | 0.96 | 1.00 | 0.80 | 0.80 | 0.71 | 1.00 |
| | | Test | 0.72 | 0.85 | 0.75 | 0.76 | 0.70 | 0.81 |

*Table 6: Performance of wine classification results - Categorization 2 in test sample mode*

| Categorization 2 | | | J48 | RF | NB | MP | SMO | k* |
|---|---|---|---|---|---|---|---|---|
| Correctly Classified Instances (%) | All wine | Training | 91.62 | 99.98 | 74.91 | 80.34 | 76.80 | 100.00 |
| | | Test | 79.13 | 85.38 | 72.67 | 78.10 | 76.00 | 83.54 |
| | White wine | Training | 89.91 | 100.00 | 73.66 | 79.03 | 74.24 | 100.00 |
| | | Test | 77.07 | 83.61 | 72.65 | 77.55 | 75.51 | 80.41 |
| | Red wine | Training | 93.12 | 100.00 | 84.36 | 89.45 | 82.93 | 100.00 |
| | | Test | 82.71 | 87.29 | 84.38 | 85.21 | 81.46 | 83.96 |
| Kappa statistic | All wine | Training | 0.76 | 1.00 | 0.35 | 0.31 | 0.00 | 1.00 |
| | | Test | 0.41 | 0.55 | 0.29 | 0.25 | 0.00 | 0.54 |
| | White wine | Training | 0.73 | 1.00 | 0.37 | 0.33 | 0.00 | 1.00 |
| | | Test | 0.36 | 0.51 | 0.34 | 0.25 | 0.00 | 0.48 |
| | Red wine | Training | 0.75 | 1.00 | 0.39 | 0.59 | 0.00 | 1.00 |
| | | Test | 0.41 | 0.50 | 0.44 | 0.47 | 0.00 | 0.45 |
| Mean absolute error | All wine | Training | 0.09 | 0.06 | 0.19 | 0.19 | 0.28 | 0.00 |
| | | Test | 0.16 | 0.16 | 0.20 | 0.20 | 0.28 | 0.12 |
| | White wine | Training | 0.10 | 0.06 | 0.20 | 0.20 | 0.29 | 0.00 |
| | | Test | 0.18 | 0.17 | 0.20 | 0.20 | 0.28 | 0.14 |
| | Red wine | Training | 0.08 | 0.05 | 0.12 | 0.10 | 0.26 | 0.00 |
| | | Test | 0.14 | 0.13 | 0.13 | 0.12 | 0.27 | 0.12 |
| Root mean squared error | All wine | Training | 0.21 | 0.10 | 0.36 | 0.31 | 0.36 | 0.00 |
| | | Test | 0.34 | 0.28 | 0.37 | 0.32 | 0.36 | 0.30 |
| | White wine | Training | 0.23 | 0.11 | 0.36 | 0.32 | 0.37 | 0.00 |
| | | Test | 0.36 | 0.28 | 0.36 | 0.33 | 0.37 | 0.33 |
| | Red wine | Training | 0.20 | 0.10 | 0.28 | 0.24 | 0.34 | 0.00 |
| | | Test | 0.32 | 0.25 | 0.28 | 0.28 | 0.34 | 0.31 |
| ROC Area | All wine | Training | 0.94 | 1.00 | 0.77 | 0.80 | 0.52 | 1.00 |
| | | Test | 0.73 | 0.88 | 0.74 | 0.77 | 0.52 | 0.87 |
| | White wine | Training | 0.94 | 1.00 | 0.78 | 0.81 | 0.51 | 1.00 |
| | | Test | 0.74 | 0.87 | 0.77 | 0.76 | 0.51 | 0.86 |
| | All wine | Training | 0.93 | 1.00 | 0.82 | 0.84 | 0.54 | 1.00 |
| | | Test | 0.77 | 0.89 | 0.83 | 0.78 | 0.54 | 0.84 |

Considering the results presented in the previous tables, it can be concluded that the most successful classification result for all three sets of wines was obtained with the RF algorithm for both test models and in both methods of wine categorization.

The accuracy of each cross-validation is greater than 70% for categorization of wines in 6 categories (Categorization 1) and greater than 85% for categorization of wines in 3 categories (Categorization 2). When using the test sample method, if we look only at the results of the test sets, the accuracy for both methods of categorization in all observed sets is greater than 83%.

K * algorithm in the worst case lags the RF by a maximum of 5%. The worst results in terms of accuracy are reported by NB and SVM.

Identical conclusions can be obtained from other measures: Kappa statistic, Mean absolute error, Root mean squared error and ROC Area.

Based on all of the above, it can be concluded that RF is the most appropriate algorithm for wine classification, whether we look at the whole set of wines or we separately distinguish only white and red wines, regardless of which of these methods of wine categorization is used.

## CONCLUSION

In this paper a method to predict affiliation of wine to a certain category based on measured physical and chemical properties is presented.

The experiment is performed using a set of 6497 wine samples from the northwest of Portugal. The classification was made on the whole dataset but also on the two subsets containing the data about white and red wine in particular. All datasets are categorized in two different ways.

It was used 6 classification algorithms: J48, RF, MLP, SVM, NB, and k *. After analyzing the results by comparing statistical indicators (percent Correctly Classified Instances, Kappa statistics (kappa), Mean absolute error (MAE), Root mean squared error (RMSE), Recall and ROC Area), it can be concluded that the best results were achieved with Random Forest algorithm. With this algorithm wines could be classified into one of the appropriate categories with an accuracy greater than 85%.

A few percents worse results than RF were achieved by the k * algorithm, while the worst results were achieved by SVM.

In this paper we have presented the ability to provide a virtual sommelier using classification algorithms that would allow categorization of wines based on their physicochemical properties only.

## Acknowledgment

This work has been supported by the Ministry of Science and Technological Development of the Republic of Serbia under Projects No.TR-35026 and III47016.

## REFERENCE

[1] HopferH., Nelson J., Ebeler S. E. and Heymann H., Correlating Wine Quality Indicators to Chemical and Sensory Measurements, Molecules 2015, 20, 8453-8483.

[2] https://www.ucdavis.edu

[3] Hodgson R. T. An Examination of Judge Reliability at a major U.S. Wine Competition, Journal of Wine Economics, Vol. 3, Issue 2, Fall 2008, Pages 105–113

[4] 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.

[5] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[6] Quinlan, J. Ross. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA 1993.

[7] Livingston F., Implementing Breiman's Random Forest Algorithm into Weka, ECE591Q Machine Learning Conference Papers. November 27, 2005

[8] Goodman, Rodney M., and Zheng Zeng. "A learning algorithm for multi-layer perceptrons with hard-limiting threshold units." Neural Networks for Signal Processing [1994] IV. Proceedings of the 1994 IEEE Workshop. IEEE, 1994

[9] Han J., Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2006

[10] Keerthi S.S., Shevade S.K., Bhattacharyya C., Murthy K.R.K., Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 2001, 13(3):637-649

[11] http://weka.sourceforge.net/doc.stable/weka/classifiers/lazy/KStar.html

[12] John G. Cleary, Leonard E. Trigg: K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, 108-114, 1995

[13] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas."Supervised machine learning: A review of classificationtechniques." Emerging artificial intelligence applicationsin computer engineering 160, 2007.

[14] www.winespectator.com

[15] www.decanter.com