# Data Preparation For Wine Quality Prediction

**Ivan Kosarevych**

kosarevych@ucu.edu.ua

Applied Sciences Faculty, Ukrainian Catholic University, Lviv, Ukraine

## ABSTRACT

Since the very beginning of human history in civilizations like Sumer, Persia, Ancient Greece wine was in high respect and was used in day to day life. Ancient Greeks called it "The Drink of the Gods" and had it as inseparable part of their life and culture. Today wine consuming culture spread across the countries and nationalities and nowadays it is one of the most popular beverages in the world. However because of such widespread, lots of samples are of poor quality. There is interest whether a machine can distinguish wines better than a human. Within this paper, I am touching mostly the part of dataset preparation for such experiment and in the end present a simple model for wine quality prediction.

## Introduction

Human error takes place in every activity a person participates in. Wine quality classification is not an exception.

Talking about ordinary consumers of wine, their quality prediction is highly influenced by such factors as price and country of origin(COO). In fact, the influence of price and COO was found so powerful as to overwhelm even the taste of poor wine[1].

Even professional sommeliers can not consistently distinguish wines and as result their qualities. Gravel et al. in their work show that the need to conduct replicate tastings when assessing wines for quality as adequate taster repeatability cannot be guaranteed. The combined score of a small team of tasters generally results in more consistent quality assessments[2]. A single even a professional sommelier is very likely to be biased.

Everything mentioned above leads to the attempts to make a machine distinguish wines. One of such experiments was maintained by Frank et al. They applied multivariate regression method Partial Least-Squares(PLS) Regression in order to classify wine quality and country of origin based on chemical measurements. It their work they showed that PLS models can successfully connect inorganic and organic composition with the various sensory parameters.[3].

This work is mostly concentrated on data set preparation for wine quality prediction. Different techniques of data mining are considered and tested. Finally different models for the stated problem are investigated.

## Methods

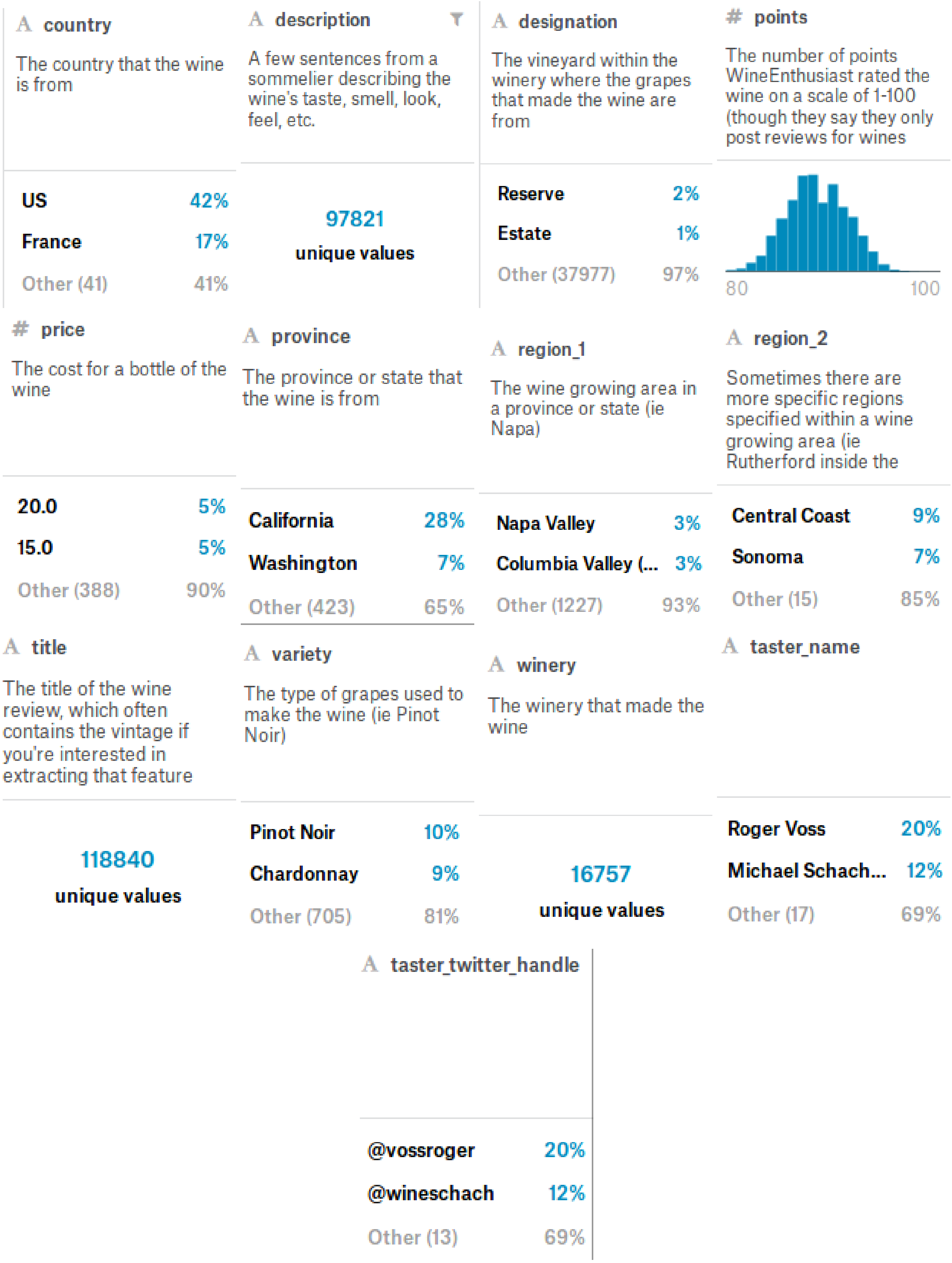
Data set selection was the entry point of the experiment. It was taken from online resource Kaggle and can be found here. The data set originally consisted of three files. One of them is with 150 000 samples and two others have 130 000 samples in CSV and JSON formats. For further analysis the CSV file with 130 thousand samples was selected, considered as the most appropriate as it was more fresh than others and in more easy-to-work format.

The data set has the columns with properties described on Figure 1. As we want to predict wine quality our target column here is column 'points', which describes quality of wine. From Figure 1 we can observe that its distribution is close to normal one.

The process of data preparation consists of several stages:

- Data Cleaning
- Missing Values Imputation
- Outliers investigation
- Data Transformation
- Data Normalization
- Dimension Reduction

And finally we are going to select ML model in order to predict wine quality.



A country

The country that the wine is from

| | |
|---|---|
| US | 42% |
| France | 17% |
| Other (41) | 41% |

A description

A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.

97821

unique values

A designation

The vineyard within the winery where the grapes that made the wine are from

| | |
|---|---|
| Reserve | 2% |
| Estate | 1% |
| Other (37977) | 97% |

# points

The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines

# price

The cost for a bottle of the wine

| | |
|---|---|
| 20.0 | 5% |
| 15.0 | 5% |
| Other (388) | 90% |

A province

The province or state that the wine is from

| | |
|---|---|
| California | 28% |
| Washington | 7% |
| Other (423) | 65% |

A region_1

The wine growing area in a province or state (ie Napa)

| | |
|---|---|
| Napa Valley | 3% |
| Columbia Valley (... | 3% |
| Other (1227) | 93% |

A region_2

Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the

| | |
|---|---|
| Central Coast | 9% |
| Sonoma | 7% |
| Other (15) | 85% |

A title

The title of the wine review, which often contains the vintage if you're interested in extracting that feature

118840

unique values

A variety

The type of grapes used to make the wine (ie Pinot Noir)

| | |
|---|---|
| Pinot Noir | 10% |
| Chardonnay | 9% |
| Other (705) | 81% |

A winery

The winery that made the wine

16757

unique values

A taster_name

| | |
|---|---|
| Roger Voss | 20% |
| Michael Schach... | 12% |
| Other (17) | 69% |

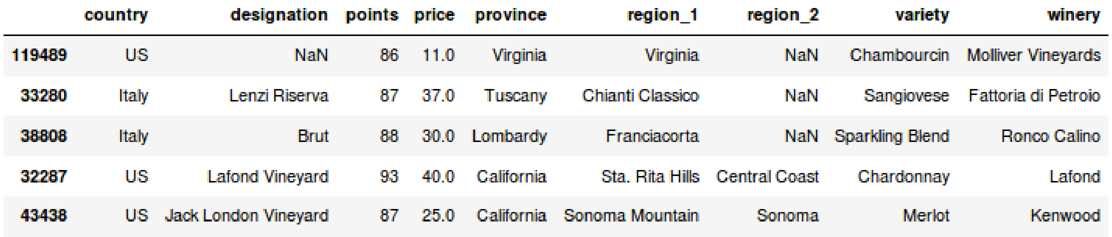
A taster_twitter_handle

| | |
|---|---|
| @vossroger | 20% |
| @wineschach | 12% |
| Other (13) | 69% |

**Figure 1.** Data Set columns

### Data cleaning

Here I investigated data set for duplicates, typos or incorrectly named columns, quotes or trailing whitespaces in data. The data set appeared to be clear from all listed problems. Also on this step I empirically selected meaningful columns which are connected with the target column ’points’. Dropped columns are "taster name", "taster twitter handle", "title", "description". Column "description" in order to have some impact on the result has to be prepared using Natural Language Processing(NLP), which is outside of the scope of this work. Columns "taster name" and "taster twitter handle" are not considered as features for quality prediction as they are relative only for this data set and won’t be appropriate for samples outside of this data. In other words the model for wine quality prediction may overfit on this values. It can overfit on column "title" as well. Other columns are relative to the target column and are considered for further analysis.



| | country | designation | points | price | province | region_1 | region_2 | variety | winery |
|---|---|---|---|---|---|---|---|---|---|
| **119489** | US | NaN | 86 | 11.0 | Virginia | Virginia | NaN | Chambourcin | Molliver Vineyards |
| **33280** | Italy | Lenzi Riserva | 87 | 37.0 | Tuscany | Chianti Classico | NaN | Sangiovese | Fattoria di Petroio |
| **38808** | Italy | Brut | 88 | 30.0 | Lombardy | Franciacorta | NaN | Sparkling Blend | Ronco Calino |
| **32287** | US | Lafond Vineyard | 93 | 40.0 | California | Sta. Rita Hills | Central Coast | Chardonnay | Lafond |
| **43438** | US | Jack London Vineyard | 87 | 25.0 | California | Sonoma Mountain | Sonoma | Merlot | Kenwood |

**Figure 2.** Cleaned data

### Missing Values Imputation

Here I investigated dataset for presence of different sort of missing values. NaN values were found in dataset. They were found in most of the remained columns except for ’points’ and ’winery’ columns. A,so column ’variety’ has only one missing value. However columns ’region 2’ has more than a half values missed. High amount of missings are in columns ’designation’ and ’region 1’. Full description of missing data is in Table 1.

In scope of this work two approaches were considered in dealing with missing values. They were selected based in the assumption that data ws missed at random. In order to test their quality a simple Linear Regression model was trained.

First of them is global most common substitution. Missings in categorical columns were filled with most frequent value(mode). Missed values in numerical columns were imputed with average of this column. This approach scored 0.1801 on train data set and 0.1718 on test one.

The second method is K-Nearest Neighbours Classifier. The columns of the data set were one by one imputed using KNN. Other columns in order to train the classifier were imputed using global most common substitution. Columns which were imputed with KNN on the previous stages were not reimputed with global most common substitution again.

Firstly, all the classifiers were run with default parameters, namely number of components was set to 5, leaf size was equal to 30. Finally, the data set with replaced missing values scored 0.1757 on train and 0.1670 on test, which is worse than simple global most common substitution approach.

Then KNN classifiers were changed. Number of components was set to 30 and leaf size to 100, with regards to high similarity of the data observed on Figure 1. However even this manipulations haven’t improved the score. Moreover to data preparation for the classifiers and training them spend greatly more larger amount of time than the first approach. It can be explained by the fact that values in features are too similar, which is indicated by high percentage of the specific value in the categorical column and ’pillars’ on the histograms of numerical columns.

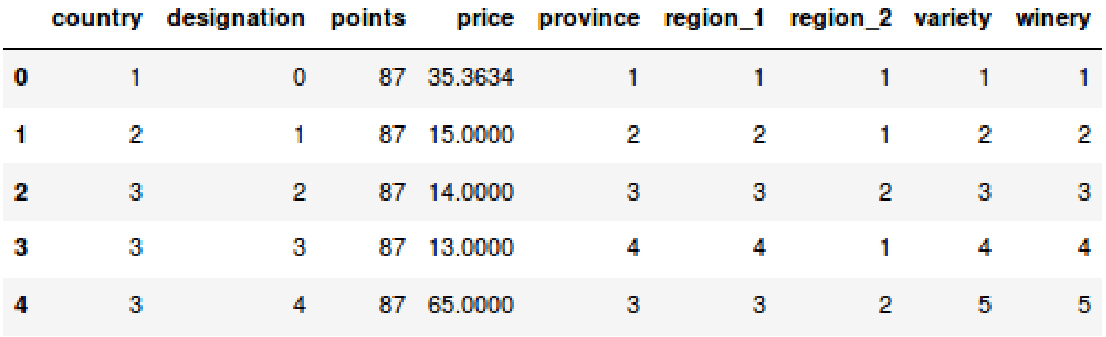
We can conclude that K-Nearest Neighbours approach is not applicable to this data set, therefore data obtained in global most common substitution approach was selected for further analysis.

### Data Transformation

Most of the columns in the data set are categorical, which means values in them are strings. Strings can not be fed into any model for training. Therefore every categorical column was transformed in the following way. For every column all its unique values were obtained. For every such value a unique integer identifier was assigned. Then every value in the column was replaced with its identifier. The transformed data set consisting only from numerical values was considered for further analysis.

### Outliers Detection

Data set investigation for outliers was maintained. Data spread is shown on Figure 4. From that figure one can observe that the data does not have significant number of outliers. More serious problem which is clearly visible is great similarity of data, represented by histogram ’pillars’. No further actions on outliers were taken.



| | country | designation | points | price | province | region_1 | region_2 | variety | winery |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 87 | 35.3634 | 1 | 1 | 1 | 1 | 1 |
| **1** | 2 | 1 | 87 | 15.0000 | 2 | 2 | 1 | 2 | 2 |
| **2** | 3 | 2 | 87 | 14.0000 | 3 | 3 | 2 | 3 | 3 |
| **3** | 3 | 3 | 87 | 13.0000 | 4 | 4 | 1 | 4 | 4 |
| **4** | 3 | 4 | 87 | 65.0000 | 3 | 3 | 2 | 5 | 5 |

**Figure 3.** Transformed data

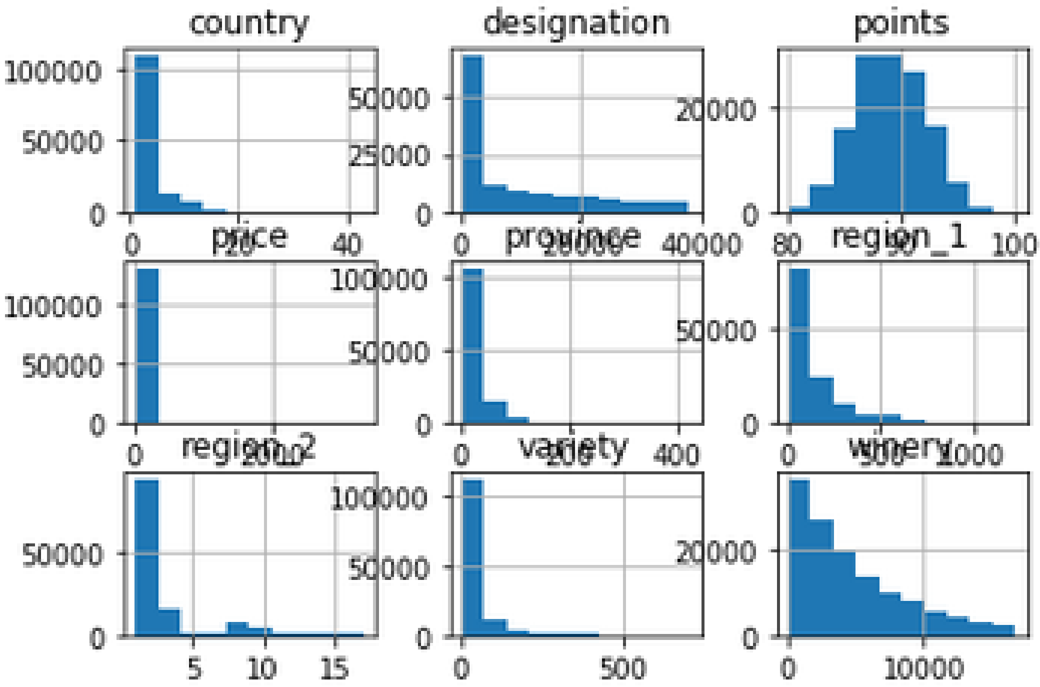


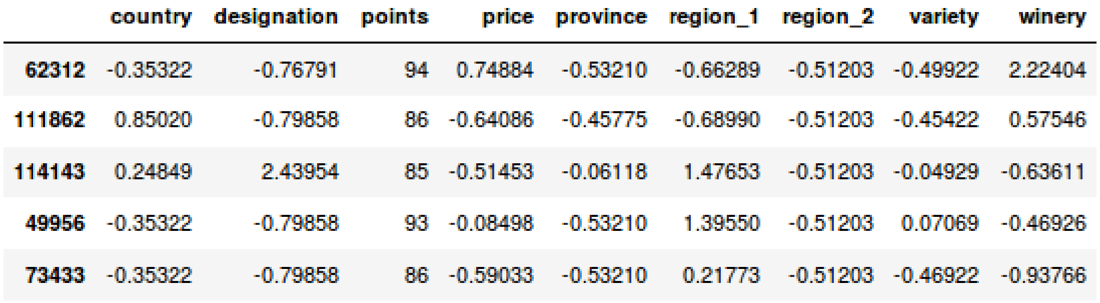
**Figure 4.** Column histograms

| Column | Missing values |
|---|---|
| country | 63 |
| designation | 37465 |
| price | 8996 |
| province | 63 |
| region 1 | 21247 |
| region 2 | 79460 |
| variety | 1 |

**Table 1.** Number of missing values by columns

## Data Normalization

In this part features were normalized in a standard way by substracting the mean and dividing by standard deviation. Result was considered for further analysis and is represented on Figure 5.



| | country | designation | points | price | province | region_1 | region_2 | variety | winery |
|---|---|---|---|---|---|---|---|---|---|
| **62312** | -0.35322 | -0.76791 | 94 | 0.74884 | -0.53210 | -0.66289 | -0.51203 | -0.49922 | 2.22404 |
| **111862** | 0.85020 | -0.79858 | 86 | -0.64086 | -0.45775 | -0.68990 | -0.51203 | -0.45422 | 0.57546 |
| **114143** | 0.24849 | 2.43954 | 85 | -0.51453 | -0.06118 | 1.47653 | -0.51203 | -0.04929 | -0.63611 |
| **49956** | -0.35322 | -0.79858 | 93 | -0.08498 | -0.53210 | 1.39550 | -0.51203 | 0.07069 | -0.46926 |
| **73433** | -0.35322 | -0.79858 | 86 | -0.59033 | -0.53210 | 0.21773 | -0.51203 | -0.46922 | -0.93766 |

**Figure 5.** Normalized data

## Dimension reduction/Feature selection

We observed in previous work that our data has very similar values. We can deal with that by applying dimension reduction to the data set. This approach also decreases training time without significant loss in accuracy.

In order to accomplish it I considered two ways. The quality of the reduction was measured by a simple Linear Regression. Reduction was applied to normalized data set of features. Target column was not touched. The reduction then was also applied to non-normalized features and scored lower.

Principal Component Analysis is a basic approach and was considered at first. PCA models with six, five and four components were trained. The six-component model scored the highest performing 0.1783 on train data set and 0.1704 on test. Moreover, when histograms of the transformed data were considered it was clearly visible that PCA reduced also similarity of data. Results are shown on Figure 6.

The second way is Decision Trees Classifier with 50 estimators. With this model I also found how each feature is important. Results are presented in Table 2. DT model resulted in 4 columns extracted from existing 8. This data set scored 0.1784 on train and 0.1693 on test, which is very similar to PCA. However time for DT training was larger than PCA, also memory consumption was higher.

The later let us say that PCA performs better and it is more applicable for this particular data set. Its result was considered for training the model.

## Model Selection

Finally, we have a data set without missing values and irrelevant features. Its values are transformed into numerical for more broad analysis and normalized. The data set is cleaned from outliers and reduced in dimension for decreasing similarity and faster training.

In this section I consider different ML models to solve the problem of wine quality prediction. It is important to notice that the problem is a classification one with 100 classes. However, as our dataset contains only up to 20 classes (from 81 to 100) the more appropriate is to use one of the regression models. Meanwhile classification models are to be considered as well.

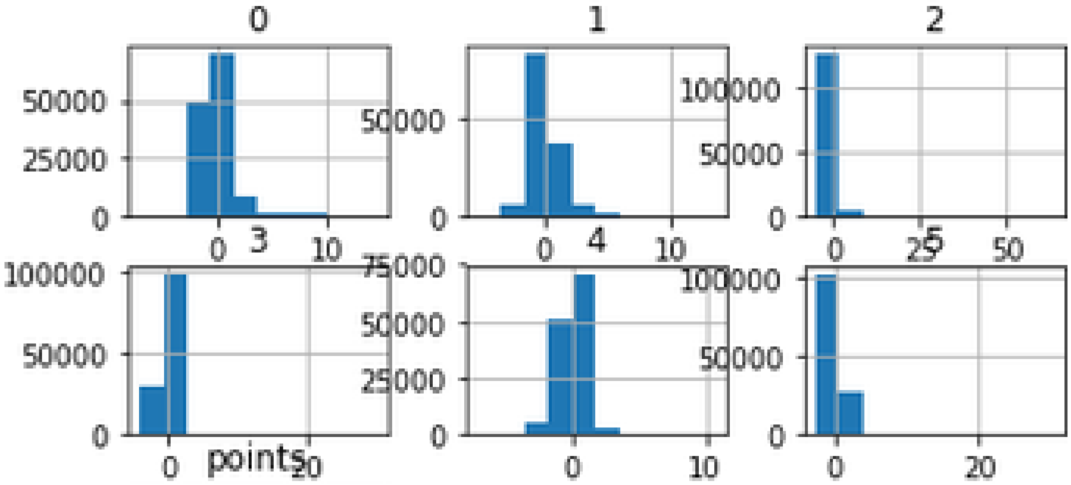


**Figure 6.** PCA reduced data

Firstly, one can observe that simple Linear Regression performed as much as 0.1704 on test data set, while testing the performance of PCA. Then two classification approaches were researched. Support Vector Machines Classifier took large amount of time to train but it only performed nearly as good as Linear Regression, showing 0.1771 on train and 0.1693 on test. Another classifier a simple Logistic Regression scored even worse – only 0.1566 on train and 0.1554 on test.

Still there is one more approach which can be applied here – Regressor based on neural network. Multi-layer Peceptron Regeressor was trained on the data reduced by PCA and performed twice as much as other models, scoring 0.3568 on train and 0.3581 on test data sets.

| Column | Importance |
|---|---|
| country | 0.0245 |
| designation | 0.1792 |
| price | 0.2754 |
| province | 0.0502 |
| region 1 | 0.0895 |
| region 2 | 0.0112 |
| variety | 0.1404 |
| winery | 0.2291 |

**Table 2.** Feature importance

## Results

Approaches applied on different steps are collected here in compact tables from 3 to 8.

### Missing Value imputation

| Approach | Train Accuracy(%) | Test Accuracy(%) |
|---|---|---|
| Global most common substitution | 18.01 | 17.18 |
| K-Nearest Neighbours | 17.57 | 16.70 |

**Table 3.** Missing value imputation prediction accuracy

### Dimension Reduction

### Model Selection

| Approach | Time performance |
|---|---|
| Global most common substitution | 0.35 sec. |
| K-Nearest Neighbours | > 1hr. |

**Table 4.** Missing value imputation time performance

| Approach | Train Accuracy(%) | Test Accuracy(%) |
|---|---|---|
| PCA | 17.83 | 17.04 |
| DT | 17.84 | 16.93 |

**Table 5.** Dimension reduction prediction accuracy

## Conclusion

In this work the process of data preparation was studied for wine quality prediction and finally several models for such purpose were presented. The process is divided in several steps. The study showed that for missing value imputation stage global most common substitution works better than K-Nearest Neighbours approach. Moreover it shows that PCA successfully reduces similarity and dimension of data and does it better than Decision Trees. Finally, neural network model overperformed others in prediction accuracy, which was the most probable.

Still lots of work can be done. More complex approaches can be applied on the every stage. Other data sets can be found, studied and possibly merged with this one. Own multilayer neural network can be composed and trained using more advanced technologies like Tensorflow and Pytorch.

## References

1. Veale R., Q. P. Consumer sensory evaluations of wine quality: The respective influence of price and country of origin. DOI: https://doi.org/10.1017/S1931436100000535 (2008).
2. R.GAWEL, P. G. Evaluation of the consistency of wine quality assessments from expert wine tasters. DOI: https://doi.org/10.1111/j.1755-0238.2008.00001.x (2008).
3. Frank, K. B. R., I. E. Prediction of wine quality and geographic origin from chemical measurements by parital least-squares regression modeling. DOI: https://doi.org/10.1016/S0003-2670(00)84245-2 (1984).

| Approach | Time performance |
|---|---|
| PCA | 0.7816 sec. |
| DT | > 1 min. |

**Table 6.** Dimension reduction time performance

| Approach | Train Accuracy(%) | Test Accuracy(%) |
|---|---|---|
| Linear Regression | 17.83 | 17.04 |
| Logistic Regression | 15.66 | 15.54 |
| SVM Classifier | 17.71 | 16.93 |
| MLP Regressor | 35.68 | 35.81 |

**Table 7.** Selected model prediction accuracy

| Approach | Time performance |
|---|---|
| Linear Regression | 1.28 sec. |
| Logistic Regression | > 26.39 sec. |
| SVM | > 30 min. |
| MLP Regressor | 171.18 sec. |

**Table 8.** Dimension reduction time performance