Modelling wine quality based on physicochemical tests

Team members:

Abdelrahman Galal Mohamed Anwar Elnahas
Fidan Rustambayli
Kristiina Kulbok

https://github.com/kristiinakulbok/IDS2020-project

Task 2. Business understanding

1- Identifying business goals

- **Background**
  Many companies around the world produce alcoholic beverage. Wine is one of the most popular and classic drinks. Wine comes in many forms, however the most widely-known wines are the red and white wines. According to Harvard business review, the United States wine industry has ballooned from just over $30 billion in 2002 to more than $60 billion today, making it the largest in the world. The number of U.S. wineries has grown by 50% to nearly 10,000 just in the past decade. Some have redefined great wine, gained the loyalty of passionate consumers, and commanded extraordinary prices.
  Wine consists of many chemical components like fixed acidity, density and alcohol content which affect the overall quality of the product. This project is about quantifying the effect of these ingredients and chemical components and measuring how much they affect the wine quality and rank. In order to achieve this goal, different machine learning models will be tested and optimized to have the ability to learn from the exciting data and analyse it. These models could be later used to assess and predict upcoming products before even releasing them to market consumption.

- **Business goals**
  The project's business goal is to help wine producers make highly ranked quality wine by discovering which components and characteristics contribute to the overall quality of the drink. High quality wine will lead to customer satisfaction and more sales and profit as a result.

- **Business success criteria**
  The Business success criteria is to define the 2 components that heavily affect the quality of the wine.

2- Assessing the situation

- **Inventory of resources**
    1. Wine dataset contains samples of wine ingredients for both red and white wine with corresponding quality rank.
    2. Team of three enthusiastic data analysts.
    3. Lecture recordings and PDF documents explaining different technical aspects needed to implement the project.
    4. Support sessions and channels like Piazza with the course instructors in case any help is needed.
    5. Laptops equipped with Python and jupyter notebook as an IDE to provide the computational power needed to complete the analysis and build the machine learning model.
    6. The open library of the internet to search and read about different topics.

- **Requirements, assumptions, and constraints**
    1. The team required to carry out a group project that demonstrates the knowledge of the introduction to data science course applied to practice.
    2. Results of the project are required to be presented in the poster session on Thursday, December 17, 2020 at 14:00-17:00.
    3. We assume that every team member will share the overall project load equally.
    4. Data is obtained through UCI. This dataset is publicly available for research, however citation is still required. So I will include the citation here as "*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. ISSN: 0167-9236.*"
    5. Python, jupyter notebook and scikit learn package and its dependencies are used to carry out the analysis. Those packages do not require a special licence to use.

- **Risks and contingencies**
    1. Data could be huge and more computational power could be needed and as a solution we can use google cloud or ask the instructors to provide support.
    2. Laptops' hard disk failure and potential loss of project code and data. In order to overcome these problems we are using Github repository to store and share the code.

- **Terminology**
    1. Wine Quality: a score between 0 and 10 defines how good the wine is. 0 is the lowest grade while 10 is the highest.
    2. Machine Learning: according to wikipedia, Machine Learning (ML) is the study of computer algorithms that improve automatically through experience. Machine learning algorithms build a model based on sample data, in order to make predictions or decisions without being explicitly programmed to do so.

3. Features: in machine learning, feature is the set of attributes that is used to predict the label. In our case the features are the chemical characteristics of the wine.
4. Labels: in machine learning, labels are the output that you want to predict. It could be a discrete value (classes) or continuous values. In our project the label is the wine quality.

- **Costs and benefits**
  1. The project will cost time to be done.
  2. In case of good implementation, the team will get a high mark.
  3. Demonstrating learned skills with real data is a great benefit.

3-Defining data-mining goals

- **Data-mining goals**
  1. Creating a machine learning model able to predict wine quality.
  2. Demonstrate the results as a poster with high quality visualizations.
  3. Create a GitHub repository to maintain the code.
  4. Increase the team work skills for all the team members.

- **Data-mining success criteria**
  1. Building a machine learning model that is able to predict at least 70 percent of wine quality before releasing them to customers.
  2. Getting the full mark in the project.

Task 3. Data understanding

1- Gathering data

- **Outline data requirements**

  1. A dataset where each row corresponds to one wine sample and contains the results of various physicochemical variables, wine type and a quality rating.
  2. Preferred data format would be csv or any other format which can be converted to a table form.

- **Verify data availability**

  A suitable dataset is provided by the UCI machine learning repository and it is available for public use.

- **Define selection criteria**

  Our source contains two separate csv files, one with data about white wine, the other about red wine. We will be using both datasets and including all the provided attributes in our analysis.

2- Describing data

The data is from *vinho verde* wine samples originating from the north of Portugal. Samples were collected from 2004 to 2007.

Attributes (columns):

Input variables (based on physicochemical tests), represented as a decimal number:

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), represented as an integer:

12. quality

Unlike the rest of the attributes, this is based on a sensory test performed by human experts. It is based on a blind test and represents the median of at least 3 expert ratings.

Number of cases (rows):

Red wine - 1599
White wine - 4898

The data includes everything we would expect and there should be a sufficient amount of cases for analysis.

3- Exploring data

White wine	minimum	maximum	mean	standard deviation
fixed acidity	3.8	14.2	6.854788	0.843868
volatile acidity	0.08	1.1	0.278241	0.100795
citric acid	0	1.66	0.334192	0.121020
residual sugar	0.6	65.8	6.391415	5.072058
chlorides	0.009	0.346	0.045772	0.021848
free sulfur dioxide	2	289	35.308085	17.007137
total sulfur dioxide	9	440	138.360657	42.498065
density	0.987110	1.03898	0.994027	0.002991
ph	2.72	3.82	3.188267	0.151001
sulphates	0.22	1.08	0.489847	0.114126
alcohol	8	14.2	10.514267	1.230621
quality	3	9	5.877909	0.885639

Red wine	minimum	maximum	mean	standard deviation
fixed acidity	4.6	15.9	8.319637	1.741096
volatile acidity	0.12	1.58	0.527821	0.179060
citric acid	0	1	0.270976	0.194801
residual sugar	0.9	15.5	2.538806	1.409928
chlorides	0.012	0.611	0.087467	0.047065
free sulfur dioxide	1	72	15.874922	10.460157
total sulfur dioxide	6	289	46.467792	32.895324
density	0.990070	1.003690	0.996747	0.001887
ph	2.74	4.01	3.311113	0.154386
sulphates	0.33	2	0.658149	0.169507
alcohol	8.4	14.9	10.422983	1.065668
quality	3	8	5.636023	0.807569

At first look, there seem to be no issues with data quality. There are no missing values and the values are consistent across both datasets, meaning there aren’t any outliers or otherwise suspicious entries.

4- Verifying data quality

As pointed out in the previous section, this data is suitable for achieving our goals. There are no obvious quality issues and the data is already clean and formatted. Although it is unlikely to become an issue, one thing to keep in mind is that the dataset for white wines is 3 times larger than the dataset for red wines.

Task 4. Planning your project

As a team we have discussed the distribution of workload among us in several group meetings. We have summed up the whole workload within several tasks. Those are as follows.

Task 1. Analysis of the data.

(All team members will work together on this task to understand future plans properly and at least 2 hours)

- Numbers of instances of each dataset.
- Specifying types of the attributes (nominal / ordinal / numeric))
- Exploring distribution of the values for different feature (using histogram,ranges of attributes)
- Gathering interesting facts
- Exploring and visualisation of relation of attributes.
- Specifying imbalance/balanced dataset.

Task 2. Pre-processing of the data

(All team members will work together on this task approx. 2 hours)

- Cleaning the data (leading and/or trailing spaces , missing values, typos)
- Preparing data for decision tree and the random forest algorithms (Converting strings or categorical features into binary features)
- Getting sample balanced data from imbalance data (The data we use is imbalanced)
- Scaling features of the data.
- Splitting data for test and training set

Task 3. Applying Machine Learning algorithms
(All team members will work seperately on this task approx. 3 hours)

- Training the data with Decision Tree and Random Forest classifiers with different numbers of depth and trees.
- Creating model and training the data with KNN classifier (rescaling data , choosing different k-nearest neighbors)
- Training the data with SVM with different parameters (polynomial,linear and radial kernel)
- Applying models to the test set and predicting labels.

Task 4. Analysing models
(All team members will work seperately on this task approx. 3 hours by reporting the results)

- Analysing overfitting and underfitting (by visualisation of the test set and training set)
- Comparing models (finding the best accuracy)
- Analysing models by visualisation.
- Building confusion matrix (Calculate accuracy, precision, recall)

Task 5. Presenting
(All team members will work together on this task approx. 5 hours)

- Writing a report (by visualisation of results, reporting crucial facts, reporting models and parameters with best accuracies)
- Creating presentation slides
- Training presentation.
- Presenting the project.