THE ANALYSIS OF SLAVONIAN WINES WITH APPLICATION OF CLUSTER ANALYSIS

Jelčić, Ana

University of Zagreb

Faculty of Economics and Business

Master in Managerial informatics

THE ANALYSIS OF SLAVONIAN WINES WITH APPLICATION OF CLUSTER ANALYSIS

Master thesis

Ana Jelčić

Zagreb, September 2020

University of Zagreb

Faculty of Economics and Business

Master in Managerial informatics

THE ANALYSIS OF SLAVONIAN WINES WITH THE APPLICATION OF CLUSTER ANALYSIS

Master thesis

Student: Ana Jelčić

Student ID: 0067532379

Mentor: Mirjana Pejić Bach, Ph.D.

Zagreb, September 2020

STATEMENT ON THE ACADEMIC INTEGRITY

I hereby declare and confirm by my signature that the final thesis The analysis of Slavonian wines with the application of cluster analysis is the sole result of my own work based on my research and relies on the published literature, as shown in the listed notes and bibliography.

I declare that no part of the thesis has been written in an unauthorized manner, i.e., it is not transcribed from the non-cited work, and no part of the thesis infringes any of the copyrights.

I also declare that no part of the thesis has been used for any other work in any other higher education, scientific or educational institution.

_____

(personal signature of the student)

_____

(place and date)

Contents

1. INTRODUCTION

1.1. Subject and Purpose

There is ten millennia where people all over the world enjoy in various delicious wines. There are also numerous evidences that ancient people made and drank wine. Wine from our territories was so good that even Greek writer Agatarid in the second century BC wrote:" On Vis island in Adriatic Sea, a wine is produced that is better than all others if compared with it". There are many details and things that people deal with, but wine falls into the category of creative art in which the winemaker not only creates pleasure for himself but also for others to enjoy the perfect wine. It is a love that involves a lot of care, time and attention, but once you discover it, you become addicted to it.

In this thesis two completely unrelated fields are combined - informatics and agronomy. Since they do not have any similar segment, the work is a very different and instructive examination. The purpose of this master thesis was to examine and analyse the wines in Slavonia but with the application of cluster analysis. The original inspiration came by an interest in the breadth of the possibility of testing through cluster analysis and since Slavonia is familiar and offering a large assortment of wines and winemakers, the goal become to do the analysis in that matter.

1.2. Explanation of Methodology

The methodological application is physical as well as online. Physical involving in-depth literature research (both primary and secondary) and data collection from conference call interviews and therefore direct communication with a producers and other industry experts. The data have been collected in a way of sending a table with necessary data, and mostly sent by email, that had to be fulfilled by the experts. Research was able to proceed after all the data was gathered.

The other part of literature was accessible online through official websites of wine producers and other sites that are as closely linked to the topic as possible. Relevant online books and articles were also used which as fulfilled and enriched this master thesis research as possible. Most of the results are described by author's own words or supported by already mentioned sources. Graphs, figures, and tables within master thesis are largely made by the author itself.

1.3. Structure and Content

This master thesis is divided into 9 parts and introduction is embodied and positioned as the first of them. The next unit of the thesis is dedicated to Slavonian wines which contains interesting and important facts from the books by Croatian authors on the related topic. Also, it explains everything important about Slavonian wines such as what Slavonian wines are, what are the defects, and all about the white and red wines that have been specifically researched and analysed for this master. After that comes a whole that is dedicated to knowledge discovery in databases and are explained the most important parts about it like the importance of classification or grouping, what is cluster analysis - what is segmentation method, and what are the types of the algorithms used for the analyses. The fourth part is devoted to the process of knowledge discovery in databases which includes overall explanation of the process itself based on the example on the this master thesis's topic about wines in Slavonia, and how the data have been collected for the analysis. The fifth part is about overview of the dataset which includes all the applicable data for the analysis, the explanations and descriptions of the data, and some important additions such as what is ARFF file, how it looks like in a this master thesis matter, and the difference between nominal and numeric quantities. The unit is finished by the explanation of two normalization techniques.

The next part is dedicated to a pictorial presentation of the analysed attributes, that are inserted into Weka pre-process area, reviewed and compared for this thesis in a pictorial format. This part also includes eleven subunits, and each is dedicated to one of the attributes. The seventh part is about the official results of the Slavonian white and then red wine cluster analyses and are supported by the pictorial representations of cluster analyses and their visualizations. The eight part of the thesis is dedicated to graphical presentations on the comparisons of the number of iterations and squared errors in a cluster of white and red wines in Slavonia when number of either clusters or clusters and seeds changes. The last part of the thesis is conclusion in which are official conclusions on obtained results.

2. PLEASANT FACTS ABOUT WINE

According to a book from Benašić Z. (2001) there are some interesting facts that are going to be translated and described in a matter of this master thesis's topic about wines in Slavonia. Croatia is the country that is on the list of groups of countries with highest consumption of wine. There are 2 wine-growing regions in Croatia - continental and coastal, and Slavonia belongs to the continental. Slavonia is one of the largest and most important Croatian wine-growing subregions in terms of wine production.[1]In Slavonia wine thrives up to 300m above sea level. The average lifespan of vines in today's vineyards is 25 to 30 years, but before the onset of the disease the lifespan of the vines was even longer on average 40 to 50 years. One of the most popular diseases is Gray mold (or noble mold) which appeared in France in 1904 and is also known as grape rot or botrytis (cinerea). This fungus can cause great damage to grapes, but in dry years (usually in the fall) it can also give “noble” effects.

The vine is a modest plant and thus it is satisfied by poor and very poor soils, while rich soils give higher income but poorer wine quality. The vine is planted in autumn or spring and winemakers are not allowed to plant the varieties they want because it is prescribed by the certain laws and regulations. Many years of experience and scientific research have shown that grape varieties produce different wines in different regions. For example, Graševina - which is most represented in the continental vineyard region and produces excellent wines if planted in Dalmatia would give wines with completely different analytical properties and thus it would not be the same Graševina wine. So, the law and regulations on wine prescribed which grape varieties may be planted in certain areas, and two groups of varieties have been identified - permitted and recommended. In Slavonia are allowed varieties for planting and the most common white grape varieties are Graševina, Pinot bijeli, Traminac, Rizling rajnski, Silvanac zeleni, Chardonnay, Pinot sivi, and Sauvignon, while red grape varieties are Frankovka, Pinot crni, and Portugizac[2] (most wines are analysed in this master thesis). Slavonia is home not only to fine wines but also to one of the most highly rated oak types that are commonly used by Croatian winemakers, especially for bigger barrels. Within this part of Croatia hot summers and cold winters still make wines rich and mature with lots of floral tones

[1] Ljubljanović S. (1996). *Hrvatski vinski vodić.* Zagreb: vlastita naklada.
[2] Benašić Z. (2001). *Što ljubitelji vina žele i vole znati*: *100 odgovora na 101 pitanje.* Đakovo: vlastita naklada.

in white wines, and sugared fruity tastes in the reds. The very last interesting fact about wines is that Cognac is actually destilled wine. It is type of a brandy that is produced in a specially defined area of the town Cognac on the west coast of France.

2.1. Analysed Wines

The wines that are used for this master thesis's cluster analysis are explained in the next few sentences. Here would be explained the origin of a wines, the soil mortality that defines the quality of wine, the vine resistance to a low and high temperatures, the resistance on a grey mold (fungi), and the amount of sugar that grapes are containing. It is important to highlight that there are wines as G* point white by Galić winery, or Graševina de Ghoto by Kutjevo winery that are named by wineries itself, i.e. do not contain original title but creative or imaginative one.

2.1.1. White wines in Slavonia

- **Chardonnay** – wine from France and is also known as *Pinot Blanc Chardonnay* or *Morillon*, gives a mediocre soil mortality for quality and premium wines. Resistant to low temperatures, but sensitive to Gray mold and gives 22% of sweetener (sugar).
- **Graševina** – the original name is *Reisling Italian*, but in Croatia was originally called *Grašica*, because the berries resembled peas, and later became *Graševina*. wine is also from France, gives high stable soil fertility for quality and often top-quality wines. Resistant to low temperatures, but moderately sensitive to Gray mold and gives 22% of sweetener (sugar).
- **Pinot bijeli** – is a wine from France, from the province of Burgundy so the name Burgundac is protected by the French for wines from that province (in our country it is used to be called *Burgundac white*). Soil fertility is small to medium for quality and premium wines. It is resistant to low temperatures, but is very sensitive to Gray mold, and gives high percentage of sweetener (sugar).
- **Pinot sivi** – also wine from France and known as *Pinot Gris* and *Tokaj de Alsace*. Gives a small soil fertility for quality and top-quality wine, provides from 18% to 22% sugar. Resistant is to low temperatures and is less resistant to Gray mold than Pinot Noir.

- **Rizling Rajnski** – is originally from Germany, from the Rhine river area. Soil fertility are low that give quality wines and contain sugar from 18% to 22%. It is very resistant to low temperatures, and moderately resistant to Gray mold.
- **Sauvignon** – wine from France, gives low soil fertility for quality and premium wines, well tolerates low temperatures, but is less resistant to Gray mold.
- **Traminac** - It originates from South Tyrol, province in northern Italy. Gives low soil fertility for quality and premium wines, and 18% to 22% sugar. It is resistant to low temperatures and quite resistant to Gray mold.
- **Pošip** - autochthonous variety from Dalmatia. It has a medium yield for high-quality and dessert white wines.

2.1.2. Red wines in Slavonia

- **Pinot crni** - also known as *Pinot noir*, wine from France, from the province of Burgundy (in our country it used to be called Burgundac red). Gives low soil fertility for quality wines, contains a lot of sugar. Well tolerates low temperatures and is poorly resistant to Gray mold.
- **Cabernet sauvignon** - more represented in coastal Croatia. from the French province of Bordeaux, and there is also known as *Petit cabernet and Petit vidure*. Soil fertility is low but yields high-quality red wines. Well resistant to low temperatures and Gray mold.
- **Merlot** – wine from the French province of Bordeaux. Fertility of a land is mediocre and yields high-quality red wines. contains 20% or more sugar, resistant to low temperatures and Gray mold.

2.2. What is Wine?

"According to our, and thus world regulations, wine is a product obtained by complete or partial alcoholic fermentation of mash or must from a grapes".[3] Benašić Z. in his book (2001) also defined the white and red wines. White wine is produced from white and pink grapes. After the grapes have been mulched, the mash is obtained and should be isolated from the s solid parts by pressing. To kill the micro-organisms and avoid oxidation of the must, sulfur

[3] Benašić Z. (2001). *Što ljubitelji vina žele i vole znati*: *100 odgovora na 101 pitanje. Đakovo: vlastita naklada.*

should be inserted in the obtained must. It is then left to settle for 24 hours and then after being poured, chosen yeasts are infused in it – they are turning the sugar into the alcohol, i.e. the must into the wine.

If this method is used to process red grapes, we would get Rose wine. However, it is strongly permitted to mix white and red wines to produce Rose wines.

Red wine is produced by processing red grapes so that mash stays in the tubs for a few days to extract the red colour from the skin of the berries by boiling. The must, which is already drained during boiling is pressed, and the hook is pressed to separate the liquid part that it is placed in barrels for a further boiling.[4] According to an age, wines are divided into two groups of young and old wines. If wine is to 3 years old it is considered as young, if more then 3 years it is considered as old wine. [5]

2.3. The Importance of Wine Labels

Controlled origin label is one that is established for control of the origin (provenance) of a wine by legal regulations in order to protect consumers and winemaker in all major wine-growing countries. According to quality and production process wines are classified as:

- **Table wines** – they are produced from one or more varieties of grapes. They are not under any approved body's authority, and therefore must not bear the mark of controlled origin. This means that any grapes produced in Slavonia and Croatia can be used for table wine without restrictions.
- **Table wines with a label of controlled origin** – under approved body's authority.
- **Quality wines with a label of controlled origin** – these are the wines under that label that originates only from one wine growing subregion or vineyard and the yield of grapes is lower than for table wine with controlled origin.
- **Premium wines with a label of controlled origin** - the best wine that can be obtained in one vineyard. Premium wines shall not be repaired as sweetened, acidified or deacidified unlike other types of wines of regulated origin.
- **Predicate wines** - wines that in exceptional years as well as in special conditions of ripening, methods of harvesting and processing of grapes give a special quality of wine.

[4] Benašić Z. (2001). *Što ljubitelji vina žele i vole znati*: *100 odgovora na 101 pitanje. Đakovo: vlastita naklada.*
[5] Ljubljanović S. (1996). *Hrvatski vinski vodić.* Zagreb: vlastita naklada.

They are known as *Late harvest* (must contains 19% of sugar), *Selective harvest* (must contains 21% of sugar), *Selective harvest of berries* (must contains 29% of sugar), *Selective harvest of dried berries* (must contains 36% of sugar), and *Ice wine* which is obtained from grapes harvested at a temperature of -7°C and contains high percentage of alcohol 14 to 16% (must contains 29% of sugar).

- **Special wines** - these are: *Dessert wine* which must contains at least 15% of alcohol and for that wine's alcohol is added to it which increases the percentage of alcohol by 8%, then *Liqueur wine* to which concentrate or alcoholic must is added, has a very sweet taste and contains 15-22% of alcohol, and *Aromatized wine* which contains 70% natural wine to which wine distillate, fragrant bitter parts of plants or other substances of plant origin are added, it must contain 15-22% alcohol.
- **Dry wines** - a wine in which there is no or only a little residual sugar because by fermenting of the must all the sugar is converted into alcohol, and no liquid sugars are given. They are: *Dry wine* (contain at least 4g/l of residual sugar), *Semi-dry wine* (contain from 4g/l to 12g/l of residual sugar), Semi-sweet wine (contain from 12g/l to 50g/l of residual sugar) and Sweet wine (contain more than 50g/l of residual sugar)
- **Bio wine** – wine that is produced without chemical additions. Manure, compost, earthworms, etc. are methods that returned grape production for bio wine to the old way of fertilization and disease protection.
- **Sparkling wines** - wines that along with other prescribed ingredients in the bottle are under pressure and when opened they cause a bang, and when poured into glasses foam. These foams are actually beads of carbon dioxide gas that came into the wine in two ways: by subsequent fermentation or by injection under pressure. *Natural sparkling wine* is one that has at least 3,5 and at the most 7 bars, while *Natural pearl wine* has even smaller pressure of 1 to 2,5 bars. If added CO_2 in wine, then it is named *Carbonated sparkling wine* and has at least 3,5 and at the most 7 bars.
- **Champagne –** it is sparkling wine that is produced by classic method i.e. by subsequent fermentation in a bottle. It is made in the Champagne province of France, so it is the only sparkling wine in the world which can be called "Champagne". In this province it

is produced from 3 types of grapes: Pinot noir which gives fullness and long life, Pinot Meunier which gives freshness and Chardonnay which gives finesse and elegance.[6]

2.4. What does Wine Contain?

For this master thesis, the question „what does wine contain?“ was a primary one to start and finish overall analysis. The data about wine have been collecting within longer period of time and was aimed to be as transparent and original as possible. Under thesis are examined and analysed 11 ingredients, but wine contains more than 700 different ingredients according to book by Benašić Z. (2001) and most important ones are: Alcohol (most common is ethanol) and interesting is that percentage of alcohol in a wine can be calculated before wine is even made by a simple calculation: grape's sweetener times 0,6 would give approximate volume of alcohol percentage that wine will contain. Sugar - a must have to contain at least 11% of sugar for table wine and 18% for top-class wines, and most common sugar is glucose, i.e. grape sugar and fructose i.e. fruit sugar. It is strongly forbidden to add sweeteners into must or wine except the law and regulations say opposite and it is the case when due to bad weather in a year the sugar content is lower than the average in the previous five years. Moreover, important ingredients are also tannin and tannin compounds, nitrogen compounds (most common are proteins), vitamins, minerals (wine contains from 1,5 to 3 g/l and the most common are potassium, sodium, calcium and magnesium), phenolic substances (it is confirmed wine is good for health according to a findings and examinations of certain doctors), organic acids (most common are non-volatile known as wine, apple and dairy, and volatile known as vinegar and ant acids), and last ingredient is water which is in the percentage of 70% to 80% in wine. [7]

2.5. Wine Defects

According to a book (1997) by Licul R. and Premužić D. wine defects occur under external interference and are caused by both physical and chemical processes. The most important disadvantages are: *smell of a wine on hydrogen sulfide* (H_2S) smell like rotten eggs but it is removed by ventilation and the addition of sulfur dioxide, then *browning of wine or oxidation*

[6] Benašić Z. (2001). *Što ljubitelji vina žele i vole znati*: *100 odgovora na 101 pitanje. Đakovo: vlastita naklada.*
[7] Benašić Z. (2001). *Što ljubitelji vina žele i vole znati*: *100 odgovora na 101 pitanje. Đakovo: vlastita naklada.*

of wine but this is prevented by adding the required amount of sulfur dioxide (SO_2) and filling the containers with wine to the end to prevent the entry of oxygen, and last is *strange smell* that occurs if wine is improper storage of wine and the most common smells are the smells of wood and of mold. Wine absorbs foreign smells very easily, so there must be no substances in the room where the wine is located that can transmit unpleasant smells to the wine.[8]

[8] Licul R. and Premužić D. (1977). *Praktično vinogradarstvo i podrumarstvo.* Zagreb: Nakladni zavod Znanje.

3. KNOWLEDGE DISCOVERY IN DATABASES

Knowledge discovery in databases (KDD) is an automatic, exploratory analysis and modelling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets.[9] Knowledge discovery in databases is newly discovered methodology for lighting on valuable data in databases of companies. It is a study and analysis of huge amounts of data using automatic or partially automatic methods with the purpose of discovering significant regularities. Several other names that are commonly used for knowledge discovery in databases are data mining, pattern analysis, data dredging, information harvesting, and knowledge extraction.

According to Pejić-Bach M. presentation (2020), there are three main goals of knowledge discovery in data bases. The first one observes how to enlarge the probability of right decisions by installing the right analysis which can be either indexes (e.g. price increase probability index), forecasting (e.g. shares value), classification (e.g. cat, dog and fish are animals), or grouping (e.g. which Slavonian wine is of the best quality). The next goal helping us managing business rules by developing tools (e.g. tool for detecting risky clients), and the last goal is to develop applications which are integrated into information system, with which are automated functions of tools (e.g. application which warming sales department that customer switch to company's competitor). She is also mentioned that knowledge discovery in databases contains two types of KDD applications – horizontal and vertical. Horizontal is most useful, collected and shared information of certain sector in the context of business rules, place graphical user interfaces and industry-specific terms, while vertical is data analysis method for general use.[10]

3.1. Classification or Grouping

In essence, data mining or KDD is a situation of growing need for summarizing or classifying data sets due to increasing number of large databases that happening in all industries and

[9] Adhikari A. (2015). *Advances in Knowlege Discovery in Databases* [online]. Switzerland. Springer International Publishing. Available at: https://books.google.hr/books?id=KLPzBQAAQBAJ&printsec=frontcover&dq=Knowledge+discovery+in+databases&hl=en&sa=X&ved=2ahUKEwjepvCn2srrAhXhkYsKHTFWDE0Q6AEwAnoECAYQAg#v=onepage&q=Knowledge%20discovery%20in%20databases&f=false [02.09.2020.].

[10] Pejić-Bach M. (2020). *Presentation on process discovery in KDD.* Zagreb.

sciences, and for investigation of such databases are used cluster analysis and other similar analysis. It is especially important to highlight the term classification or grouping due to its enormous importance in our lives, in business, in science, and IT. In this master thesis, for example, classified are wines which have many features in common, thus named are pH, density, producers, etc. It is essential to classify everything (by similarities and differences) including large data sets because it becomes more precisely organized so that it could be more understandable and easier for use.

3.2. Cluster Analysis

Most important knowledge discovery in databases methods are Decision trees, Association rules and Cluster analysis. Cluster analysis is a numerical classification and the most preferred term for grouping data procedures nowadays and is used for analysis of Slavonian white and red wines.

Decision tree is a predictive model which can be used to represent both classifiers and regression models, in operations research decision trees refer to a hierarchical model of decisions and their consequences.[11] To determine the strategy that will most likely reach its goal, the decision maker uses decision tree.

On the other hand there are association rules which are used to forecast any attribute and combinations of attributes as well, and they are not predetermined to be used together (compared to classification rules that forecasting only class and are used together as a set). Dissimilar association rules convey dissimilar regularities that underlying the dataset, and typically they forecast dissimilar items.

Within cluster analysis single cluster represents individual while set of clusters represent all individuals. It also must be pointed out that data grouping is not justified, in other words once data is grouped, the one who observes data should know how to describe the results. The cluster analysis is a generic name for a variety of mathematical methods, numbering in the

[11] Maimon Z O. And Rokach L. (2015). *Data Mining With Decision Trees: Theory And Applications* [online]. Singapore. Word Scientific Publishing Co. Pte. Ltd. Available at: https://books.google.hr/books?id=OVYCCwAAQBAJ&printsec=frontcover&dq=decision+tree&hl=en&sa=X&ved=2ahUKEwjb4b7_4crrAhVjwIsKHcSeAzIQ6AEwAHoECAEQAg#v=onepage&q=decision%20tree&f=false [02.09.2020.]

hundreds, that can be used to find out which objects in a set are similar.[12] It is mathematical method which collect attributes mathematically with similar descriptions into the same cluster.

3.2.1. Cluster Analysis – Segmentation method

Segmentation method is used to split the examples into a set of groups (i.e. clusters) under two conditions. The first condition is that each cluster represents a homogeneous set in which examples of the same category are identical to one another. The second condition is that each cluster have to be distinct from other groups, or in other words examples belonging to a given group vary significantly from other examples belonging to other groups.

3.2.2. Cluster Analysis – Types of Algorithms

There are five types of algorithms used for Cluster analysis:

- Self-organizing neutral networks (Kohonen network)
- Probabilistic methods (AutoClass algorithm)
- K-mean value algorithm (K-means)
- Joining tree clustering
- Agglomeration cluster analysis (Cobweb)

3.2.3. K-mean Value Algorithm (K-means)

For this master thesis only K-mean value algorithm is used and thus will be explained. It is an incremental approach to clustering that dynamically adds one cluster centre at a time through a deterministic global search procedure consisting of N (with N being the size of the data set) executions of the k-means algorithm from suitable initial positions.[13]

In other words, in a multidimensional space that represents a „centre“ or „average“ position of a specific collection of examples centroid exists as an artificial point. K points are always

[12] Romesburg C. (2004). *Cluster Analysis for Ressearches* [online]. North Carolina. Lulu press. Available at: https://books.google.hr/books?id=ZuIPv7OKm10C&printsec=frontcover&dq=cluster+analysis&hl=en&sa=X&ved=2ahUKEwiJy8Wzh8vrAhUKtYsKHaD-AFIQ6AEwAnoECAAQAg#v=onepage&q=cluster%20analysis&f=false [03.09.2020.]

[13] Likas A. (2003). *The global k-means clustering algorithm* [online]. Patter recognition, Volume 36, Pg. 451-461. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0031320302000602 [03.09.2020.]

randomly chosen and can vary in size depending on how large data is so would be desirable to experiment with a different number of groups. Also, for k-mean value algorithm is important that nominal attributes are transformed into numeric values.

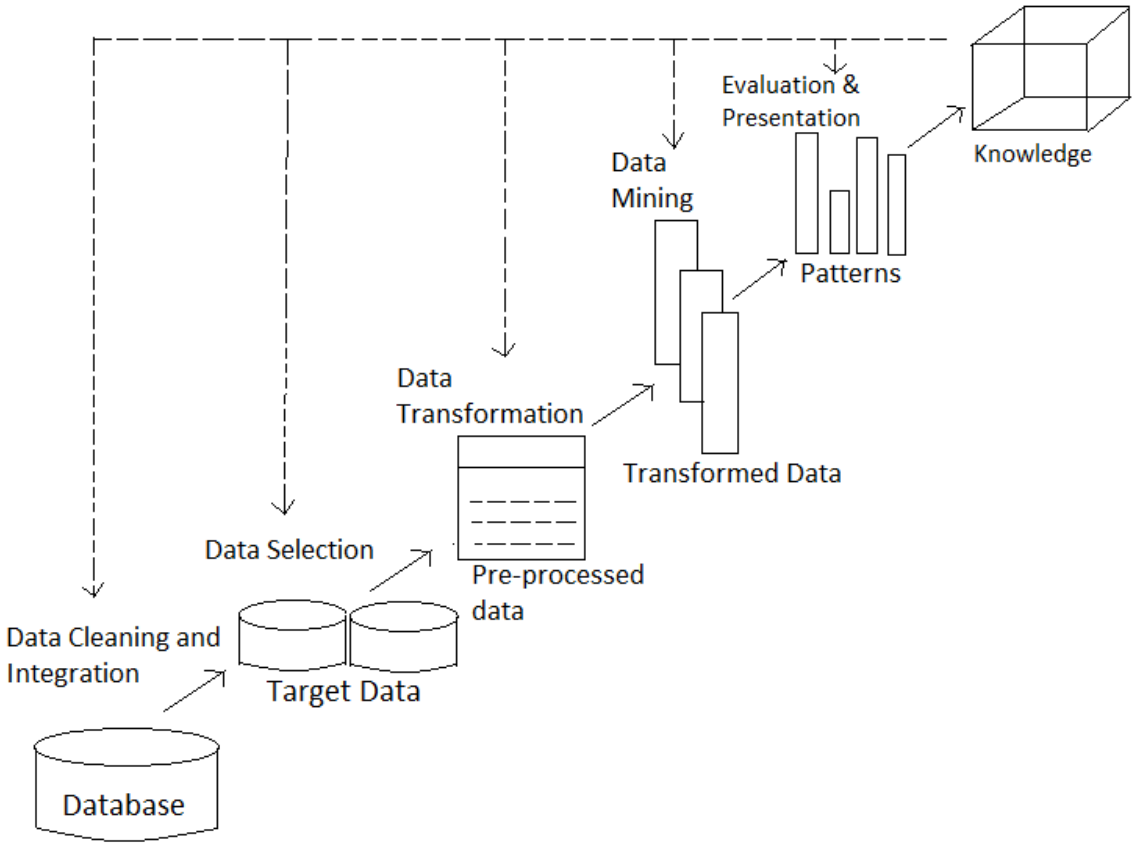
3.2.4. Sum of Squared Errors

Sum of squared errors is an important part of clustering and shows the sum of the differences in squares between each measurement and the mean of it's group. It may be used within a cluster as an indicator of the variance. If all cases within a cluster are similar and comparable then sum squared errors would be 0. For example, if there is 53 instances and chosen are 53 or more then 53 clusters that need to be analysed, then sum of squared errors would be 0 permanently.

4. PROCESS OF KNOWLEDGE DISCOVERY IN DATABASES

There are several steps of process of KDD known as defining, preparing, modelling and implementing the data set.

Figure 1 Process of knowledge discovery in databases



Source: Internet. Available at: https://techblogmu.blogspot.com/2018/05/describe-various-functionalities-of.html

Figure represent process of KDD and its basic task which is to extract knowledge from a lower level database. As can be seen, starting point of this process, once data is collected, is to define a business problem and then setting project target and selection of the data. Next step is pre-processing or setting needed data which leads us to transformation and valuation of data. The next step is choosing a mining technique to model data, and final step is implementation of data through interpretation of the results. In the next few subtopics all steps will be explained in depth and the base example is going to be the topic of this master thesis.

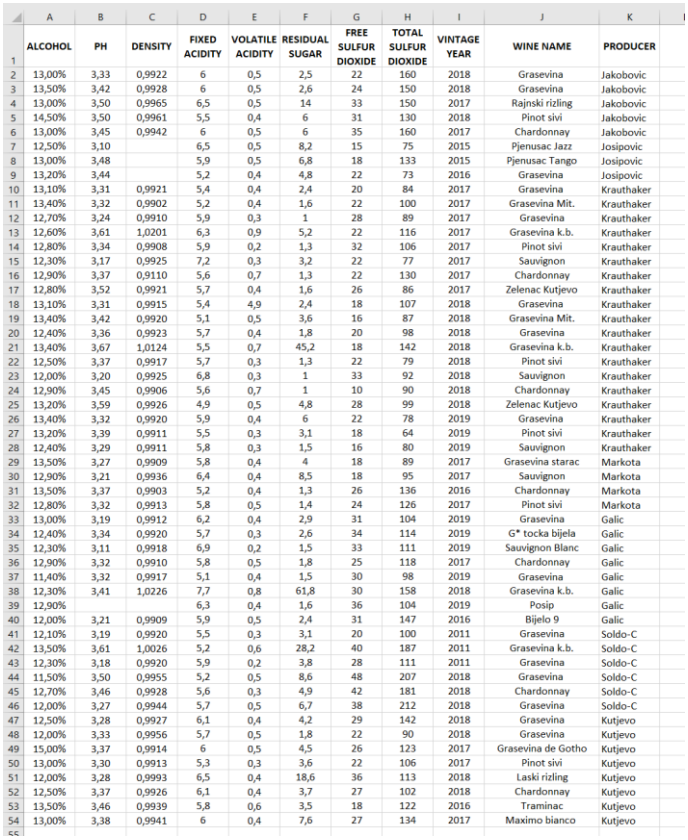
4.1. Comprehensive Data on Slavonian wines

Data about Slavonian wine is gained directly from the best wine producers in Slavonia. Data have been collected within longer period due to as the best transparency of data as possible. Data is collected from 7 different producers (Jakobović winery, Josipović winery, Krauthaker winery, Markota winery, Soldo-Čamak winery, Galić winery, and Kutjevo winery) and once it was all together, it was inserted into two .xlsx formats that can be seen in two figure 2 and figure 4 below, one for white wine (saved as WW.xlsx) and the other for red wine (saved as RW.xlsx) data. Moreover, it was essential to choose which data is going to be the target data and therefore 11 most important attributes of wine are selected: alcohol volume of wine, pH, density, fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, vintage year, wine name, and producer name. It is sorted in 11 rows, one for each of the attributes, and the main idea was to examine the best quality and to choose the best wine in Slavonia. There are 79 instances in total, 53 of white and 26 of red wine.

4.2. Pre-processed Data on Slavonian wines

After data sets in tables were placed it was essential to transform data from ordinal .xlsx to CSV (MS-DOS) file not separated with columns but commas. Firstly data was opened in Notepad (to replace all commas *, with semicolons *;) so it can be correctly shown in Weka that is the program later used for transformed data. Once data is saved as WW.csv and RW.csv as illustrated in figure 3 and figure 5 it was ready to be transformed. The differences between .xlsx and .csv excel data types can also be seen in figures illustrated below – for white wine difference is shown in figures 2 and 3, while for red wine in figures 4 and 5.

Figure 2 Data on Slavonian white wine saved as WW.xlsx



	A	B	C	D	E	F	G	H	I	J	K
1	ALCOHOL	PH	DENSITY	FIXED ACIDITY	VOLATILE ACIDITY	RESIDUAL SUGAR	FREE SULFUR DIOXIDE	TOTAL SULFUR DIOXIDE	VINTAGE YEAR	WINE NAME	PRODUCER
2	13,00%	3,33	0,9922	6	0,5	2,5	22	160	2018	Grasevina	Jakobovic
3	13,50%	3,42	0,9928	6	0,5	2,6	24	150	2018	Grasevina	Jakobovic
4	13,00%	3,50	0,9965	6,5	0,5	14	33	150	2017	Rajnski rizling	Jakobovic
5	14,50%	3,50	0,9961	5,5	0,4	6	31	130	2018	Pinot sivi	Jakobovic
6	13,00%	3,45	0,9942	6	0,5	6	35	160	2017	Chardonnay	Jakobovic
7	12,50%	3,10		6,5	0,5	8,2	15	75	2015	Pjenusac Jazz	Josipovic
8	13,00%	3,48		5,9	0,5	6,8	18	133	2015	Pjenusac Tango	Josipovic
9	13,20%	3,44		5,2	0,4	4,8	22	73	2016	Grasevina	Josipovic
10	13,10%	3,31	0,9921	5,4	0,4	2,4	20	84	2017	Grasevina	Krauthaker
11	13,40%	3,32	0,9902	5,2	0,4	1,6	22	100	2017	Grasevina Mit.	Krauthaker
12	12,70%	3,24	0,9910	5,9	0,3	1	28	89	2017	Grasevina	Krauthaker
13	12,60%	3,61	1,0201	6,3	0,9	5,2	22	116	2017	Grasevina k.b.	Krauthaker
14	12,80%	3,34	0,9908	5,9	0,2	1,3	32	106	2017	Pinot sivi	Krauthaker
15	12,30%	3,17	0,9925	7,2	0,3	3,2	22	77	2017	Sauvignon	Krauthaker
16	12,90%	3,37	0,9110	5,6	0,7	1,3	22	130	2017	Chardonnay	Krauthaker
17	12,80%	3,52	0,9921	5,7	0,4	1,6	26	86	2017	Zelenac Kutjevo	Krauthaker
18	13,10%	3,31	0,9915	5,4	4,9	2,4	18	107	2018	Grasevina	Krauthaker
19	13,40%	3,42	0,9920	5,1	0,5	3,6	16	87	2018	Grasevina Mit.	Krauthaker
20	12,40%	3,36	0,9923	5,7	0,4	1,8	20	98	2018	Grasevina	Krauthaker
21	13,40%	3,67	1,0124	5,5	0,7	45,2	18	142	2018	Grasevina k.b.	Krauthaker
22	12,50%	3,37	0,9917	5,7	0,3	1,3	22	79	2018	Pinot sivi	Krauthaker
23	12,00%	3,20	0,9925	6,8	0,3	1	33	92	2018	Sauvignon	Krauthaker
24	12,90%	3,45	0,9906	5,6	0,7	1	10	90	2018	Chardonnay	Krauthaker
25	13,20%	3,59	0,9926	4,9	0,5	4,8	28	99	2018	Zelenac Kutjevo	Krauthaker
26	13,40%	3,32	0,9920	5,9	0,4	6	22	78	2019	Grasevina	Krauthaker
27	13,20%	3,39	0,9911	5,5	0,3	3,1	18	64	2019	Pinot sivi	Krauthaker
28	12,40%	3,29	0,9911	5,8	0,3	1,5	16	80	2019	Sauvignon	Krauthaker
29	13,50%	3,27	0,9909	5,8	0,4	4	18	89	2017	Grasevina starac	Markota
30	12,90%	3,21	0,9936	6,4	0,4	8,5	18	95	2017	Sauvignon	Markota
31	13,50%	3,37	0,9903	5,2	0,4	1,3	26	136	2016	Chardonnay	Markota
32	12,80%	3,32	0,9913	5,8	0,5	1,4	24	126	2017	Pinot sivi	Markota
33	13,00%	3,19	0,9912	6,2	0,4	2,9	31	104	2019	Grasevina	Galic
34	12,40%	3,34	0,9920	5,7	0,3	2,6	34	114	2019	G* tocka bijela	Galic
35	12,30%	3,11	0,9918	6,9	0,2	1,5	33	111	2019	Sauvignon Blanc	Galic
36	12,90%	3,32	0,9910	5,8	0,5	1,8	25	118	2017	Chardonnay	Galic
37	11,40%	3,32	0,9917	5,1	0,4	1,5	30	98	2019	Grasevina	Galic
38	12,30%	3,41	1,0226	7,7	0,8	61,8	30	158	2018	Grasevina k.b.	Galic
39	12,90%			6,3	0,4	1,6	36	104	2019	Posip	Galic
40	12,00%	3,21	0,9909	5,9	0,5	2,4	31	147	2016	Bijelo 9	Galic
41	12,10%	3,19	0,9920	5,5	0,3	3,1	20	100	2011	Grasevina	Soldo-C
42	13,50%	3,61	1,0026	5,2	0,6	28,2	40	187	2011	Grasevina k.b.	Soldo-C
43	12,30%	3,18	0,9920	5,9	0,2	3,8	28	111	2011	Grasevina	Soldo-C
44	11,50%	3,50	0,9955	5,2	0,5	8,6	48	207	2018	Grasevina	Soldo-C
45	12,70%	3,46	0,9928	5,6	0,3	4,9	42	181	2018	Chardonnay	Soldo-C
46	12,00%	3,27	0,9944	5,7	0,5	6,7	38	212	2018	Grasevina	Soldo-C
47	12,50%	3,28	0,9927	6,1	0,4	4,2	29	142	2018	Grasevina	Kutjevo
48	12,00%	3,33	0,9956	5,7	0,5	1,8	22	90	2018	Grasevina	Kutjevo
49	15,00%	3,37	0,9914	6	0,5	4,5	26	123	2017	Grasevina de Gotho	Kutjevo
50	13,00%	3,30	0,9913	5,3	0,3	3,6	22	106	2017	Pinot sivi	Kutjevo
51	12,00%	3,28	0,9993	6,5	0,4	18,6	36	113	2018	Laski rizling	Kutjevo
52	12,50%	3,37	0,9926	6,1	0,4	3,7	27	102	2018	Chardonnay	Kutjevo
53	13,50%	3,46	0,9939	5,8	0,6	3,5	18	122	2016	Traminac	Kutjevo
54	13,00%	3,38	0,9941	6	0,4	7,6	27	134	2017	Maximo bianco	Kutjevo
55											

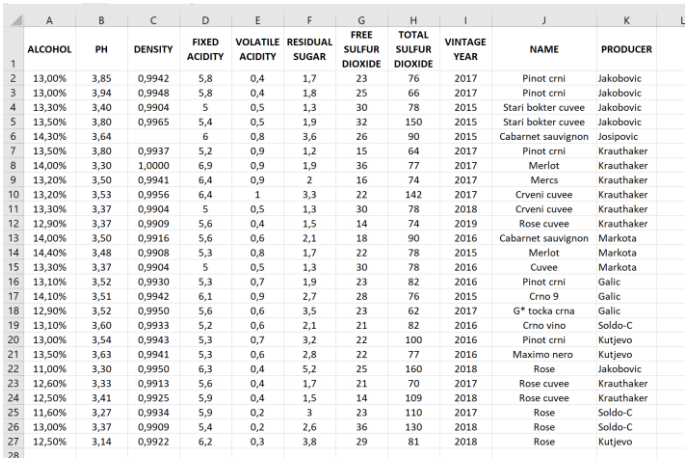
Source: Excel

Figure 3 Data on Slavonian white wine saved as WW.csv

	A
1	ALCOHOL ,PH,DENSITY,FIXED ACIDITY,VOLATILE ACIDITY,RESIDUAL SUGAR,FREE SULFUR DIOXIDE,TOTAL SULFUR DIOXIDE,VINTAGE YEAR,WINE NAME,PRODUCER
2	13.00%,3.33,0.9922,6,0.5,2.5,22,160,2018,Grasevina ,Jakobovic
3	13.50%,3.42,0.9928,6,0.5,2.6,24,150,2018,Grasevina ,Jakobovic
4	13.00%,3.5,0.9965,6.5,0.5,14,33,150,2017,Rajnski rizling,Jakobovic
5	14.50%,3.5,0.9961,5.5,0.4,6,31,130,2018,Pinot sivi,Jakobovic
6	13.00%,3.45,0.9942,6,0.5,6,35,160,2017,Chardonnay,Jakobovic
7	12.50%,3.1,,6.5,0.5,8.2,15,75,2015,Pjenusac Jazz,Josipovic
8	13.00%,3.48,,5.9,0.5,6.8,18,133,2015,Pjenusac Tango,Josipovic
9	13.20%,3.44,,5.2,0.4,4.8,22,73,2016,Grasevina ,Josipovic
10	13.10%,3.31,0.9921,5.4,0.4,2.4,20,84,2017,Grasevina ,Krauthaker
11	13.40%,3.32,0.9902,5.2,0.4,1.6,22,100,2017,Grasevina Mit.,Krauthaker
12	12.70%,3.24,0.991,5.9,0.3,1,28,89,2017,Grasevina,Krauthaker
13	12.60%,3.61,1.0201,6.3,0.9,5.2,22,116,2017,Grasevina k.b.,Krauthaker
14	12.80%,3.34,0.9908,5.9,0.2,1.3,32,106,2017,Pinot sivi,Krauthaker
15	12.30%,3.17,0.9925,7.2,0.3,3.2,22,77,2017,Sauvignon,Krauthaker
16	12.90%,3.37,0.911,5.6,0.7,1.3,22,130,2017,Chardonnay,Krauthaker
17	12.80%,3.52,0.9921,5.7,0.4,1.6,26,86,2017,Zelenac Kutjevo,Krauthaker
18	13.10%,3.31,0.9915,5.4,4.9,2.4,18,107,2018,Grasevina ,Krauthaker
19	13.40%,3.42,0.992,5.1,0.5,3.6,16,87,2018,Grasevina Mit.,Krauthaker
20	12.40%,3.36,0.9923,5.7,0.4,1.8,20,98,2018,Grasevina,Krauthaker
21	13.40%,3.67,1.0124,5.5,0.7,45.2,18,142,2018,Grasevina k.b.,Krauthaker
22	12.50%,3.37,0.9917,5.7,0.3,1.3,22,79,2018,Pinot sivi,Krauthaker
23	12.00%,3.2,0.9925,6.8,0.3,1,33,92,2018,Sauvignon,Krauthaker
24	12.90%,3.45,0.9906,5.6,0.7,1,10,90,2018,Chardonnay,Krauthaker
25	13.20%,3.59,0.9926,4.9,0.5,4.8,28,99,2018,Zelenac Kutjevo,Krauthaker
26	13.40%,3.32,0.992,5.9,0.4,6,22,78,2019,Grasevina ,Krauthaker
27	13.20%,3.39,0.9911,5.5,0.3,3.1,18,64,2019,Pinot sivi,Krauthaker
28	12.40%,3.29,0.9911,5.8,0.3,1.5,16,80,2019,Sauvignon,Krauthaker
29	13.50%,3.27,0.9909,5.8,0.4,4,18,89,2017,Grasevina starac,Markota
30	12.90%,3.21,0.9936,6.4,0.4,8.5,18,95,2017,Sauvignon,Markota
31	13.50%,3.37,0.9903,5.2,0.4,1.3,26,136,2016,Chardonnay,Markota
32	12.80%,3.32,0.9913,5.8,0.5,1.4,24,126,2017,Pinot sivi,Markota
33	13.00%,3.19,0.9912,6.2,0.4,2.9,31,104,2019,Grasevina,Galic
34	12.40%,3.34,0.992,5.7,0.3,2.6,34,114,2019,G* tocka bijela,Galic
35	12.30%,3.11,0.9918,6.9,0.2,1.5,33,111,2019,Sauvignon Blanc,Galic
36	12.90%,3.32,0.991,5.8,0.5,1.8,25,118,2017,Chardonnay,Galic
37	11.40%,3.32,0.9917,5.1,0.4,1.5,30,98,2019,Grasevina ,Galic
38	12.30%,3.41,1.0226,7.7,0.8,61.8,30,158,2018,Grasevina k.b.,Galic
39	12.90%,,,6.3,0.4,1.6,36,104,2019,Posip,Galic
40	12.00%,3.21,0.9909,5.9,0.5,2.4,31,147,2016,Bijelo 9,Galic
41	12.10%,3.19,0.992,5.5,0.3,3.1,20,100,2011,Grasevina,Soldo-C
42	13.50%,3.61,1.0026,5.2,0.6,28.2,40,187,2011,Grasevina k.b.,Soldo-C
43	12.30%,3.18,0.992,5.9,0.2,3.8,28,111,2011,Grasevina,Soldo-C
44	11.50%,3.5,0.9955,5.2,0.5,8.6,48,207,2018,Grasevina ,Soldo-C
45	12.70%,3.46,0.9928,5.6,0.3,4.9,42,181,2018,Chardonnay,Soldo-C
46	12.00%,3.27,0.9944,5.7,0.5,6.7,38,212,2018,Grasevina,Soldo-C
47	12.50%,3.28,0.9927,6.1,0.4,4.2,29,142,2018,Grasevina,Kutjevo
48	12.00%,3.33,0.9956,5.7,0.5,1.8,22,90,2018,Grasevina,Kutjevo
49	15.00%,3.37,0.9914,6,0.5,4.5,26,123,2017,Grasevina de Gotho,Kutjevo
50	13.00%,3.3,0.9913,5.3,0.3,3.6,22,106,2017,Pinot sivi,Kutjevo
51	12.00%,3.28,0.9993,6.5,0.4,18.6,36,113,2018,Laski rizling,Kutjevo
52	12.50%,3.37,0.9926,6.1,0.4,3.7,27,102,2018,Chardonnay,Kutjevo
53	13.50%,3.46,0.9939,5.8,0.6,3.5,18,122,2016,Traminac,Kutjevo
54	13.00%,3.38,0.9941,6,0.4,7.6,27,134,2017,Maximo bianco ,Kutjevo
55	

Source: Excel

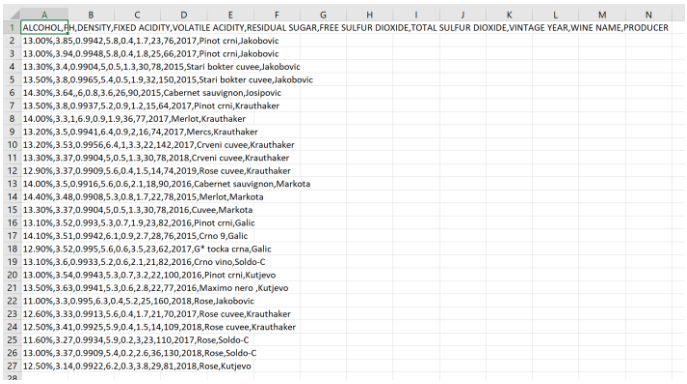
Figure 4 Data on Slavonian red wine saved as RW.xlsx



	A	B	C	D	E	F	G	H	I	J	K
1	ALCOHOL	PH	DENSITY	FIXED ACIDITY	VOLATILE ACIDITY	RESIDUAL SUGAR	FREE SULFUR DIOXIDE	TOTAL SULFUR DIOXIDE	VINTAGE YEAR	NAME	PRODUCER
2	13,00%	3,85	0,9942	5,8	0,4	1,7	23	76	2017	Pinot crni	Jakobovic
3	13,00%	3,94	0,9948	5,8	0,4	1,8	25	66	2017	Pinot crni	Jakobovic
4	13,30%	3,40	0,9904	5	0,5	1,3	30	78	2015	Stari bokter cuvee	Jakobovic
5	13,50%	3,80	0,9965	5,4	0,5	1,9	32	150	2015	Stari bokter cuvee	Jakobovic
6	14,30%	3,64		6	0,8	3,6	26	90	2015	Cabarnet sauvignon	Josipovic
7	13,50%	3,80	0,9937	5,2	0,9	1,2	15	64	2017	Pinot crni	Krauthaker
8	14,00%	3,30	1,0000	6,9	0,9	1,9	36	77	2017	Merlot	Krauthaker
9	13,20%	3,50	0,9941	6,4	0,9	2	16	74	2017	Mercs	Krauthaker
10	13,20%	3,53	0,9956	6,4	1	3,3	22	142	2017	Crveni cuvee	Krauthaker
11	13,30%	3,37	0,9904	5	0,5	1,3	30	78	2018	Crveni cuvee	Krauthaker
12	12,90%	3,37	0,9909	5,6	0,4	1,5	14	74	2019	Rose cuvee	Krauthaker
13	14,00%	3,50	0,9916	5,6	0,6	2,1	18	90	2016	Cabarnet sauvignon	Markota
14	14,40%	3,48	0,9908	5,3	0,8	1,7	22	78	2015	Merlot	Markota
15	13,30%	3,37	0,9904	5	0,5	1,3	30	78	2016	Cuvee	Markota
16	13,10%	3,52	0,9930	5,3	0,7	1,9	23	82	2016	Pinot crni	Galic
17	14,10%	3,51	0,9942	6,1	0,9	2,7	28	76	2015	Crno 9	Galic
18	12,90%	3,52	0,9950	5,6	0,6	3,5	23	62	2017	G* tocka crna	Galic
19	13,10%	3,60	0,9933	5,2	0,6	2,1	21	82	2016	Crno vino	Soldo-C
20	13,00%	3,54	0,9943	5,3	0,7	3,2	22	100	2016	Pinot crni	Kutjevo
21	13,50%	3,63	0,9941	5,3	0,6	2,8	22	77	2016	Maximo nero	Kutjevo
22	11,00%	3,30	0,9950	6,3	0,4	5,2	25	160	2018	Rose	Jakobovic
23	12,60%	3,33	0,9913	5,6	0,4	1,7	21	70	2017	Rose cuvee	Krauthaker
24	12,50%	3,41	0,9925	5,9	0,4	1,5	14	109	2018	Rose cuvee	Krauthaker
25	11,60%	3,27	0,9934	5,9	0,2	3	23	110	2017	Rose	Soldo-C
26	13,00%	3,37	0,9909	5,4	0,2	2,6	36	130	2018	Rose	Soldo-C
27	12,50%	3,14	0,9922	6,2	0,3	3,8	29	81	2018	Rose	Kutjevo

Source: Excel

Figure 5 Data on Slavonian red wine saved as RW.csv



```
ALCOHOL,PH,DENSITY,FIXED ACIDITY,VOLATILE ACIDITY,RESIDUAL SUGAR,FREE SULFUR DIOXIDE,TOTAL SULFUR DIOXIDE,VINTAGE YEAR,WINE NAME,PRODUCER
13.00%,3.85,0.9942,5.8,0.4,1.7,23,76,2017,Pinot crni,Jakobovic
13.00%,3.94,0.9948,5.8,0.4,1.8,25,66,2017,Pinot crni,Jakobovic
13.30%,3.4,0.9904,5,0.5,1.3,30,78,2015,Stari bokter cuvee,Jakobovic
13.50%,3.8,0.9965,5.4,0.5,1.9,32,150,2015,Stari bokter cuvee,Jakobovic
14.30%,3.64,,6,0.8,3.6,26,90,2015,Cabernet sauvignon,Josipovic
13.50%,3.8,0.9937,5.2,0.9,1.2,15,64,2017,Pinot crni,Krauthaker
14.00%,3.3,1,6.9,0.9,1.9,36,77,2017,Merlot,Krauthaker
13.20%,3.5,0.9941,6.4,0.9,2,16,74,2017,Mercs,Krauthaker
13.20%,3.53,0.9956,6.4,1,3.3,22,142,2017,Crveni cuvee,Krauthaker
13.30%,3.37,0.9904,5,0.5,1.3,30,78,2018,Crveni cuvee,Krauthaker
12.90%,3.37,0.9909,5.6,0.4,1.5,14,74,2019,Rose cuvee,Krauthaker
14.00%,3.5,0.9916,5.6,0.6,2.1,18,90,2016,Cabernet sauvignon,Markota
14.40%,3.48,0.9908,5.3,0.8,1.7,22,78,2015,Merlot,Markota
13.30%,3.37,0.9904,5,0.5,1.3,30,78,2016,Cuvee,Markota
13.10%,3.52,0.993,5.3,0.7,1.9,23,82,2016,Pinot crni,Galic
14.10%,3.51,0.9942,6.1,0.9,2.7,28,76,2015,Crno 9,Galic
12.90%,3.52,0.995,5.6,0.6,3.5,23,62,2017,G* tocka crna,Galic
13.10%,3.6,0.9933,5.2,0.6,2.1,21,82,2016,Crno vino,Soldo-C
13.00%,3.54,0.9943,5.3,0.7,3.2,22,100,2016,Pinot crni,Kutjevo
13.50%,3.63,0.9941,5.3,0.6,2.8,22,77,2016,Maximo nero ,Kutjevo
11.00%,3.3,0.995,6.3,0.4,5.2,25,160,2018,Rose,Jakobovic
12.60%,3.33,0.9913,5.6,0.4,1.7,21,70,2017,Rose cuvee,Krauthaker
12.50%,3.41,0.9925,5.9,0.4,1.5,14,109,2018,Rose cuvee,Krauthaker
11.60%,3.27,0.9934,5.9,0.2,3,23,110,2017,Rose,Soldo-C
13.00%,3.37,0.9909,5.4,0.2,2.6,36,130,2018,Rose,Soldo-C
12.50%,3.14,0.9922,6.2,0.3,3.8,29,81,2018,Rose,Kutjevo
```

Source: Excel

4.3. Transformed Data of Slavonian wines

From this point on data was ready to be transformed to and analysed in the Weka program. Weka is tried and tested open source machine learning software that can be accessed through a graphical user interface, standard terminal applications, or a Java API, and is widely used for teaching, research, and industrial applications.[14] Data was separately transformed and loaded in two Weka windows, and then saved as .arff file (WW.arff and RW.arff) which is Weka's default file format, and which was used for overall examination.

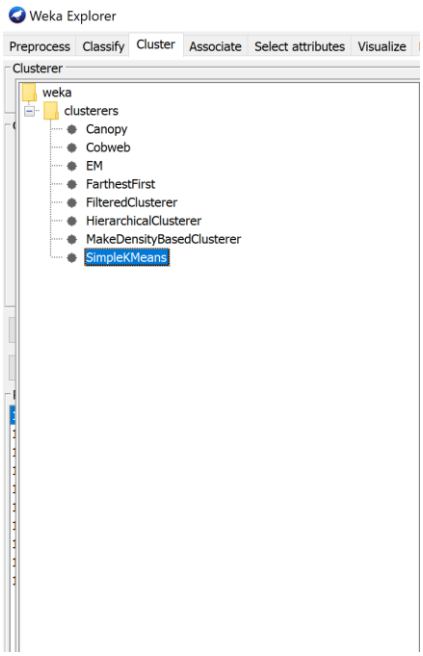
[14] Weka site. Available at: https://www.cs.waikato.ac.nz/ml/weka/ [01.09.2020.]

4.4. Patterns of Slavonian Wines

It is literally identification of the genuinely fascinating forms of information dependent on fascinating steps. The patterns of data is possible to reach only with the help of data mining technique, and for this thesis it was SimpleKMeans technique within cluster analysis. It is important to point out that Weka program contains three main parts known as Classify, Cluster, and Associate. Cluster is one that interests us and it offers eight (8) different clusterers that are lined up in the next order: Canopy, Cobweb, EM, FarthestFirst, FilteredClusterer, HierarchicalClusterer, MakeDensistyBasedClusterer and SimpleKMeans, and all of them are shown in a figure 6 below.

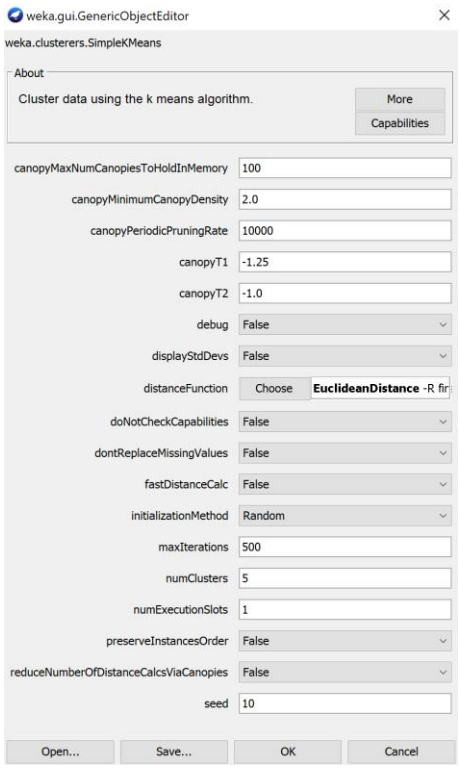
SimpleKMeans automatically handles mixtures of categorical and numerical attributes, and It can use either the Euclidean distance or Manhattan distance. For this analysis Euclidean distance is used where centroids are computed as the mean distance between instances and clusters, but if for instance Manhattan distance was used, the centroids would be computed as the component-wise meridian. On a figure 7 below is illustrated how Weka generic object editor looks like which is important while choosing number of clusters and seeds for cluster analysis, and it also corroborate that Euclidean distance function was chosen.

Figure 6 Types of Weka Clusterers



Source: Weka

Figure 7 Weka generic object editor



Source: Weka

4.5. Knowledge about Slavonian wines

The methods, techniques and representation of information are used to display the knowledge extracted to consumers. It is overall knowledge gained about Slavonian wines, their attributes, quality and producers, and visualisation of knowledge is going to be in form of graphs, figures and tables later in the text. All of the attributes examined and compared to each other are randomly chosen, not automatically by Weka program. Also, all gained results are not interpreted, described or explained by program, but should be understood by the observer.

5. DATASET OVERWIEV

For this master thesis two datasets were developed in which one includes samples of red and the other samples of white wine. Number of instances is seventy-nine (79) in total for both wine datasets where red wine dataset is smaller and counts twenty-six (26) instances while white wine dataset is something bigger and counts for fifty three (53) instances. Moreover, the main idea of this clustering was to explore the attributes of as many as possible Slavonian wines and understand how different ingredients affect their quality.

Number of attributes in this cluster analysis is 11 in total. They are identified as follows:

1. **Alcohol volume**
2. **pH**
3. **Density**
4. **Fixed acidity**
5. **Volatile acidity**
6. **Residual sugar**
7. **Free sulfur dioxide**
8. **Total sulfur dioxide**
9. **Vintage year**
10. **Wine name**
11. **Producer**

It should also be pointed out that there was one more attribute “bottle size” that was included in examination but was sufficient for this analysis.
The inputs include objective tests that are made by wine experts responsible for examining chemical part of a wine analysis in a laboratory specially designed for that area.

5.1. The List and Descriptions of the Analysed Attributes

Table 1 Description of the analysed attributes

No.	Attribute name	Description of the attribute	Attribute format (nominal, numeric, binary)
1.	**Alcohol volume**	The wine's percental alcohol content	Nominal
2.	**pH**	Defines whether wine is acid or basic. How acid or basic a wine is at a range from 0 (very acidic) to 14 (very basic). Most wines are between 3-4 on the pH scale.	Numeric
3.	**Density**	Based on the amount of alcohol and sugar content it can be determined whether it is water or wine – content of a water composition is similar to that of a wine.	Numeric
4.	**Fixed acidity**	The most acids mixed into wine that are fixed or non-volatile (do not evaporate instantly)	Numeric
5.	**Volatile acidity**	The volume of acetic acid in a wine, which may contribute to an undesirable taste of vinegar if exists at very high amounts	Numeric
6.	**Residual sugar**	The quantity of sugar remaining after ferment ation ends. Wines are considered too sweet if they have either less than 1 gram / litre or more than 45 grams / litre of residual sugar.	Numeric
7.	**Free sulfur dioxide**	It is eliminating microbial development and wine oxidation. Also, it is free source of SO2 and occurs in a equilibrium between molecular SO2 (as a dissolved gas) and bisulphite ion.	Numeric

8.	**Total sulfur dioxide**	It is a volume of free and bound sources of SO2. SO2 is often undetectable in wine at low concentrations, although at free concentrations of SO2 above 50 ppm, SO2 becomes apparent in the taste and smell of a wine.	Numeric
9.	**Year of production**	This actually represent the year of grape harvest and not the year of wine production.	Numeric
10.	**Name**	The official name of a wine sorts.	Nominal
11.	**Producer**	It is a name of a producer of several wine types or sorts.[15]	Nominal

Source: Master thesis author

Table 1 above describes attributes used for the cluster analysis. Each individual, independent instance that provides the input to machine learning is characterized by its values on a fixed, predefined set of features or attributes.[16] First column shows the numerical presentation of analysed wine attributes and the second column presents all of the 11 attributes that are used for cluster analysis. The third column explains the meaning of each of the attributes and thus definitions are positioned in the same row as attributes are. The last, forth column presents the type (in general can be numeric, nominal, string, date, relational[17]) of each attribute which are in this cluster case either numeric or nominal.

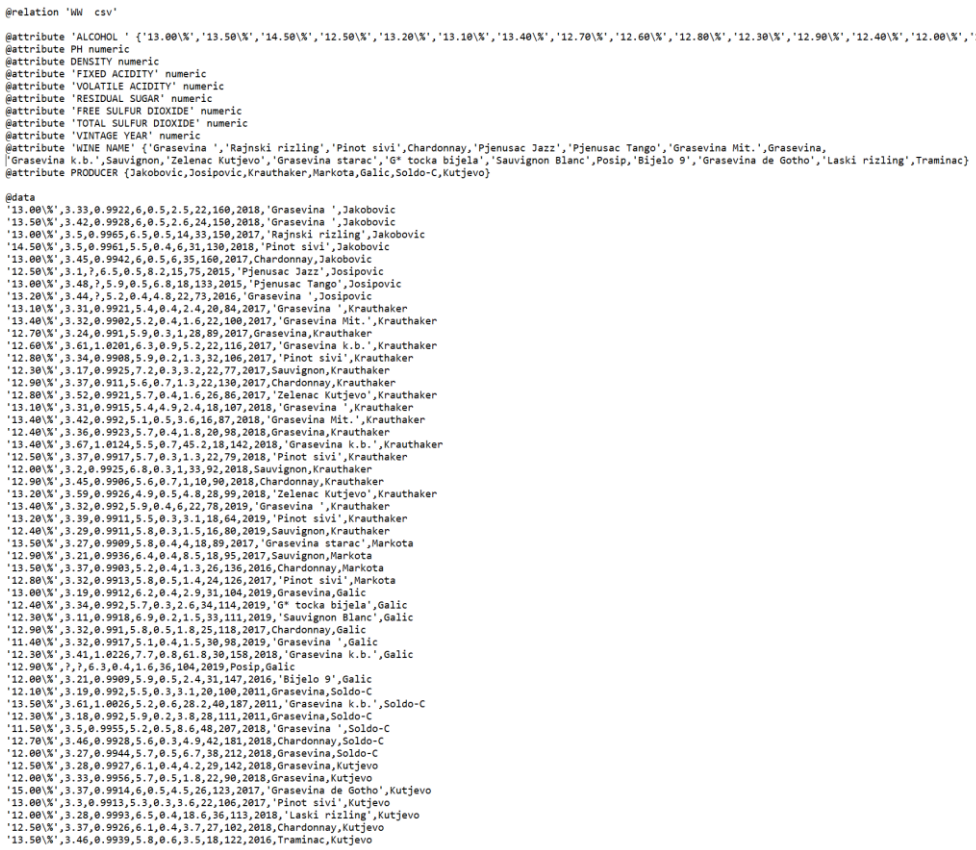
5.2. The ARFF Format

The ARFF file stands for Attribute-Relation file format and is a text file representing a collection of instances that share a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.

[15] Dua, D. and Graff, C. (2019). *Wine data set* [online]. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml [25.08.2020.]
[16] Ian H. Witten, (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* [online]. United States of America: Academic press. Available at: https://books.google.hr/books?id=6lVEKlrTq8EC&pg=PA294&dq=weka&hl=en&sa=X&ved=2ahUKEwiu6qiBv7PrAhXMyaQKHQf7DosQ6AEwB3oECAkQAg#v=onepage&q=types%20of%20attributes&f=false [24.08.2020.].
[17] Pejić-Bach, M. (2019) *Data analysis in Weka.* Prezentacija. Zagreb: Ekonomski fakultet

Ian H. Witten in his book (2000) described ARFF format as one that takes a look at a standard way of presenting dataset where instances are independent and are not in any relationship.[18]

Figure 8 ARFF file for a white wine data



```
@relation 'WW  csv'

@attribute 'ALCOHOL ' {'13.00\%','13.50\%','14.50\%','12.50\%','13.20\%','13.10\%','13.40\%','12.70\%','12.60\%','12.80\%','12.30\%','12.90\%','12.40\%','12.00\%','11.40\%','12.10\%','11.50\%','15.00\%'}
@attribute PH numeric
@attribute DENSITY numeric
@attribute 'FIXED ACIDITY' numeric
@attribute 'VOLATILE ACIDITY' numeric
@attribute 'RESIDUAL SUGAR' numeric
@attribute 'FREE SULFUR DIOXIDE' numeric
@attribute 'TOTAL SULFUR DIOXIDE' numeric
@attribute 'VINTAGE YEAR' numeric
@attribute 'WINE NAME' {'Grasevina ','Rajnski rizling','Pinot sivi',Chardonnay,'Pjenusac Jazz','Pjenusac Tango','Grasevina Mit.',Grasevina,
'Grasevina k.b.',Sauvignon,'Zelenac Kutjevo','Grasevina starac','G* tocka bijela','Sauvignon Blanc',Posip,'Bijelo 9','Grasevina de Gotho','Laski rizling',Traminac}
@attribute PRODUCER {Jakobovic,Josipovic,Krauthaker,Markota,Galic,Soldo-C,Kutjevo}

@data
'13.00\%',3.33,0.9922,6,0.5,2.5,22,160,2018,'Grasevina ',Jakobovic
'13.50\%',3.42,0.9928,6,0.5,2.6,24,150,2018,'Grasevina ',Jakobovic
'13.00\%',3.5,0.9965,6.5,0.5,14,33,150,2017,'Rajnski rizling',Jakobovic
'14.50\%',3.5,0.9961,5.5,0.4,6,31,130,2018,'Pinot sivi',Jakobovic
'13.00\%',3.45,0.9942,6,0.5,6,35,160,2017,Chardonnay,Jakobovic
'12.50\%',3.1,?,6.5,0.5,8.2,15,75,2015,'Pjenusac Jazz',Josipovic
'13.00\%',3.48,?,5.9,0.5,6.8,18,133,2015,'Pjenusac Tango',Josipovic
'13.20\%',3.44,?,5.2,0.4,4.8,22,73,2016,'Grasevina ',Josipovic
'13.10\%',3.31,0.9921,5.4,0.4,2.4,20,84,2017,'Grasevina ',Krauthaker
'13.40\%',3.32,0.9902,5.2,0.4,1.6,22,100,2017,'Grasevina Mit.',Krauthaker
'12.70\%',3.24,0.991,5.9,0.3,1,28,89,2017,Grasevina,Krauthaker
'12.60\%',3.61,1.0201,6.3,0.9,5.2,22,116,2017,'Grasevina k.b.',Krauthaker
'12.80\%',3.34,0.9908,5.9,0.2,1.3,32,106,2017,'Pinot sivi',Krauthaker
'12.30\%',3.17,0.9925,7.2,0.3,3.2,22,77,2017,Sauvignon,Krauthaker
'12.90\%',3.37,0.911,5.6,0.7,1.3,22,130,2017,Chardonnay,Krauthaker
'12.80\%',3.52,0.9921,5.7,0.4,1.6,26,86,2017,'Zelenac Kutjevo',Krauthaker
'13.10\%',3.31,0.9915,5.4,4.9,2.4,18,107,2018,'Grasevina ',Krauthaker
'13.40\%',3.42,0.992,5.1,0.5,3.6,16,87,2018,'Grasevina Mit.',Krauthaker
'12.40\%',3.36,0.9923,5.7,0.4,1.8,20,98,2018,Grasevina,Krauthaker
'13.40\%',3.67,1.0124,5.5,0.7,45.2,18,142,2018,'Grasevina k.b.',Krauthaker
'12.50\%',3.37,0.9917,5.7,0.3,1.3,22,79,2018,'Pinot sivi',Krauthaker
'12.00\%',3.2,0.9925,6.8,0.3,1,33,92,2018,Sauvignon,Krauthaker
'12.90\%',3.45,0.9906,5.6,0.7,1,10,90,2018,Chardonnay,Krauthaker
'13.20\%',3.59,0.9926,4.9,0.5,4.8,28,99,2018,'Zelenac Kutjevo',Krauthaker
'13.40\%',3.32,0.992,5.9,0.4,6,22,78,2019,'Grasevina ',Krauthaker
'13.20\%',3.39,0.9911,5.5,0.3,3.1,18,64,2019,'Pinot sivi',Krauthaker
'12.40\%',3.29,0.9911,5.8,0.3,1.5,16,80,2019,Sauvignon,Krauthaker
'13.50\%',3.27,0.9909,5.8,0.4,4,18,89,2017,'Grasevina starac',Markota
'12.90\%',3.21,0.9936,6.4,0.4,8.5,18,95,2017,Sauvignon,Markota
'13.50\%',3.37,0.9903,5.2,0.4,1.3,26,136,2016,Chardonnay,Markota
'12.80\%',3.32,0.9913,5.8,0.5,1.4,24,126,2017,'Pinot sivi',Markota
'13.00\%',3.19,0.9912,6.2,0.4,2.9,31,104,2019,Grasevina,Galic
'12.40\%',3.34,0.992,5.7,0.3,2.6,34,114,2019,'G* tocka bijela',Galic
'12.30\%',3.11,0.9918,6.9,0.2,1.5,33,111,2019,'Sauvignon Blanc',Galic
'12.90\%',3.32,0.991,5.8,0.5,1.8,25,118,2017,Chardonnay,Galic
'11.40\%',3.32,0.9917,5.1,0.4,1.5,30,98,2019,'Grasevina ',Galic
'12.30\%',3.41,1.0226,7.7,0.8,61.8,30,158,2018,'Grasevina k.b.',Galic
'12.90\%',?,?,6.3,0.4,1.6,36,104,2019,Posip,Galic
'12.00\%',3.21,0.9909,5.9,0.5,2.4,31,147,2016,'Bijelo 9',Galic
'12.10\%',3.19,0.992,5.5,0.3,3.1,20,100,2011,Grasevina,Soldo-C
'13.50\%',3.61,1.0026,5.2,0.6,28.2,40,187,2011,'Grasevina k.b.',Soldo-C
'12.30\%',3.18,0.992,5.9,0.2,3.8,28,111,2011,Grasevina,Soldo-C
'11.50\%',3.5,0.9955,5.2,0.5,8.6,48,207,2018,'Grasevina ',Soldo-C
'12.70\%',3.46,0.9928,5.6,0.3,4.9,42,181,2018,Chardonnay,Soldo-C
'12.00\%',3.27,0.9944,5.7,0.5,6.7,38,212,2018,Grasevina,Soldo-C
'12.50\%',3.28,0.9927,6.1,0.4,4.2,29,142,2018,Grasevina,Kutjevo
'12.00\%',3.33,0.9956,5.7,0.5,1.8,22,90,2018,Grasevina,Kutjevo
'15.00\%',3.37,0.9914,6,0.5,4.5,26,123,2017,'Grasevina de Gotho',Kutjevo
'13.00\%',3.3,0.9913,5.3,0.3,3.6,22,106,2017,'Pinot sivi',Kutjevo
'12.00\%',3.28,0.9993,6.5,0.4,18.6,36,113,2018,'Laski rizling',Kutjevo
'12.50\%',3.37,0.9926,6.1,0.4,3.7,27,102,2018,Chardonnay,Kutjevo
'13.50\%',3.46,0.9939,5.8,0.6,3.5,18,122,2016,Traminac,Kutjevo
```

Source: Notepad

The figure 8 above show how ARFF file for a white wine data. The file lines beginning with the name of the file @relation (white wine) and the part below it interpreting each @attribute (alcohol, pH, density, fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, vintage year, wine name, and producer). Numeric values are ones that are standing before keyword numeric. Attributes are always explained in one line per attribute, but for this case the attribute „NAME“ is broken into two lines so the pictorial presentation can be more readable.

[18] Ian H. Witten, (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* [online]. United States of America: Academic press. Available at: https://books.google.hr/books?id=6lVEKlrTq8EC&pg=PA294&dq=weka&hl=en&sa=X&ved=2ahUKEwiu6qiBv7Pr [24.08.2020.].

Figure 9 ARFF file with a data for red wine

```
@relation 'RW '

@attribute ALCOHOL {'13.00\%','13.30\%','13.50\%','14.30\%','14.00\%','13.20\%','12.90\%','14.40\%','13.10\%','14.10\%','11.00\%','12.60\%','12.50\%','11.60\%'}
@attribute PH numeric
@attribute DENSITY numeric
@attribute 'FIXED ACIDITY' numeric
@attribute 'VOLATILE ACIDITY' numeric
@attribute 'RESIDUAL SUGAR' numeric
@attribute 'FREE SULFUR DIOXIDE' numeric
@attribute 'TOTAL SULFUR DIOXIDE' numeric
@attribute 'VINTAGE YEAR' numeric
@attribute 'WINE NAME' {'Pinot crni','Stari bokter cuvee','Cabernet sauvignon',Merlot,Mercs,'Crveni cuvee','Rose cuvee',
'Cabarnet sauvignon',Cuvee,'Crno 9','G* tocka crna','Crno vino','Maximo bianco ','Maximo nero ',Rose}
@attribute PRODUCER {Jakobovic,Josipovic,Krauthaker,Markota,Galic,Soldo-C,Kutjevo}

@data
'13.00\%',3.85,0.9942,5.8,0.4,1.7,23,76,2017,'Pinot crni',Jakobovic
'13.00\%',3.94,0.9948,5.8,0.4,1.8,25,66,2017,'Pinot crni',Jakobovic
'13.30\%',3.4,0.9904,5,0.5,1.3,30,78,2015,'Stari bokter cuvee',Jakobovic
'13.50\%',3.8,0.9965,5.4,0.5,1.9,32,150,2015,'Stari bokter cuvee',Jakobovic
'14.30\%',3.64,?,6,0.8,3.6,26,90,2015,'Cabernet sauvignon',Josipovic
'13.50\%',3.8,0.9937,5.2,0.9,1.2,15,64,2017,'Pinot crni',Krauthaker
'14.00\%',3.3,1,6.9,0.9,1.9,36,77,2017,Merlot,Krauthaker
'13.20\%',3.5,0.9941,6.4,0.9,2,16,74,2017,Mercs,Krauthaker
'13.20\%',3.53,0.9956,6.4,1,3.3,22,142,2017,'Crveni cuvee',Krauthaker
'13.30\%',3.37,0.9904,5,0.5,1.3,30,78,2018,'Crveni cuvee',Krauthaker
'12.90\%',3.37,0.9909,5.6,0.4,1.5,14,74,2019,'Rose cuvee',Krauthaker
'14.00\%',3.5,0.9916,5.6,0.6,2.1,18,90,2016,'Cabarnet sauvignon',Markota
'14.40\%',3.48,0.9908,5.3,0.8,1.7,22,78,2015,Merlot,Markota
'13.30\%',3.37,0.9904,5,0.5,1.3,30,78,2016,Cuvee,Markota
'13.10\%',3.52,0.993,5.3,0.7,1.9,23,82,2016,'Pinot crni',Galic
'14.10\%',3.51,0.9942,6.1,0.9,2.7,28,76,2015,'Crno 9',Galic
'12.90\%',3.52,0.995,5.6,0.6,3.5,23,62,2017,'G* tocka crna',Galic
'13.10\%',3.6,0.9933,5.2,0.6,2.1,21,82,2016,'Crno vino',Soldo-C
'13.00\%',3.38,0.9941,6,0.4,7.6,27,134,2017,'Maximo bianco ',Kutjevo
'13.00\%',3.54,0.9943,5.3,0.7,3.2,22,100,2016,'Pinot crni',Kutjevo
'13.50\%',3.63,0.9941,5.3,0.6,2.8,22,77,2016,'Maximo nero ',Kutjevo
'11.00\%',3.3,0.995,6.3,0.4,5.2,25,160,2018,Rose,Jakobovic
'12.60\%',3.33,0.9913,5.6,0.4,1.7,21,70,2017,'Rose cuvee',Krauthaker
'12.50\%',3.41,0.9925,5.9,0.4,1.5,14,109,2018,'Rose cuvee',Krauthaker
'11.60\%',3.27,0.9934,5.9,0.2,3,23,110,2017,Rose,Soldo-C
'13.00\%',3.37,0.9909,5.4,0.2,2.6,36,130,2018,Rose,Soldo-C
'12.50\%',3.14,0.9922,6.2,0.3,3.8,29,81,2018,Rose,Kutjevo
```

Source: Notepad

This figure 9 above showing ARFF file for a red wine data. The attributes of the red wine dataset are the same as ones to white wine, but the data entrances are quantitatively smaller and completely different. In this case the attribute „NAME“ is broken into two lines as well, so the pictorial presentation can be more readable.

It is also important to highlight that the wine issue is to forecast the class value producer, the class attribute is not separated in any way from the actual values of the other attributes in the data file. The ARFF model simply includes a dataset which means that it does not define which of the attributes should be expected. It ensures that the same file will be used to analyse how good each attribute can be predicted from the others, or to find association rules, or as in this case - for clustering.

The next part of the ARFF file describing overall @data or all instances that exist within dataset. Instances are written one per line, with values divided by commas for each attribute. In a white wine case are some missing values and thus empty places are replaced with a question mark.

5.3. Nominal and Numeric quantities

According to Ian H. Witten, the ARFF format accommodates two basic data types, nominal and numeric. It is important to point out that there is a huge difference between quantities that are nominal and ones that are numeric. Numeric attributes measure numbers (they are also called continuous attributes) and the numbers can be either real-valued or integer-valued (these are not continuous in a mathematical sense). On the other hand, nominal attributes are more like prespecified values or finite set of prospects (also called categorical). But how these data types are interpreted depends on the learning scheme being used.[19]

In this master thesis learning scheme that was being used is normalization and thus numeric attributes are measured on ratio scales. Normalization include two techniques, finding minimum and maximum, and calculating statistical mean and standard deviation of the attribute values.

5.4. Two normalization techniques

5.4.1. Finding minimum and maximum

In the first normalization technique attributes are normalized to exist within a fixed range, usually from zero to one, by dividing all values by the highest value observed, or by subtracting the lowest value and splitting the amount between the highest (maximum) and lowest (minimum) values.

5.4.2. Calculating statistical mean and standard deviation

In the second normalization technique, where statistical mean and standard deviation have been measured in a way of subtracting the mean from each value and thus dividing the outcome with the standard deviation. This process is called standardizing a statistical variable, and results in a set of values whose mean is zero and standard deviation is one.[20]

---

[19] Ian H. Witten, (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* [online]. United States of America: Academic press. Available at: https://books.google.hr/books?id=6lVEKlrTq8EC&pg=PA294&dq=weka&hl=en&sa=X&ved=2ahUKEwiu6qiBv7PrAhXMyaQKHQf7DosQ6AEwB3oECAkQAg#v=onepage&q=types%20of%20attributes&f=false [24.08.2020.].

[20] Ian H. Witten, (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* [online]. United States of America: Academic press. Available at: https://books.google.hr/books?id=6lVEKlrTq8EC&pg=PA294&dq=weka&hl=en&sa=X&ved=2ahUKEwiu6qiBv7PrAhXMyaQKHQf7DosQ6AEwB3oECAkQAg#v=onepage&q=types%20of%20attributes&f=false [24.08.2020.].

6. PICTORIAL REPRESENTATIONS AND DESCRIPTIONS OF THE ANALYSED ATTRIBUTES

Once all the data was set in a Weka program, it was necessary to take another look to it so its confirmed and checked. In a Weka's pre-process area overall data was accessible and ready for the following processes. The first step of the process was to compare attributes in order to obtain general guidelines about white and red wines in Slavonia.

6.1. Alcohol Volume

Figure 10 Alcohol volume in Slavonian white wines

Selected attribute

Name: ALCOHOL Missing: 0 (0%) Distinct: 18 Type: Nominal Unique: 6 (11%)

No.	Label	Count	Weight
1	13.00%	7	7.0
2	13.50%	5	5.0
3	14.50%	1	1.0
4	12.50%	4	4.0
5	13.20%	3	3.0
6	13.10%	2	2.0
7	13.40%	4	4.0
8	12.70%	2	2.0
9	12.60%	1	1.0
10	12.80%	3	3.0
11	12.30%	4	4.0
12	12.90%	5	5.0
13	12.40%	3	3.0
14	12.00%	5	5.0
15	11.40%	1	1.0
16	12.10%	1	1.0
17	11.50%	1	1.0
18	15.00%	1	1.0

Source: Weka

Figure 11 Alcohol volume in Slavonian red wines

Selected attribute

Name: ALCOHOL Missing: 0 (0%) Distinct: 14 Type: Nominal Unique: 6 (23%)

No.	Label	Count	Weight
1	13.00%	4	4.0
2	13.30%	3	3.0
3	13.50%	3	3.0
4	14.30%	1	1.0
5	14.00%	2	2.0
6	13.20%	2	2.0
7	12.90%	2	2.0
8	14.40%	1	1.0
9	13.10%	2	2.0
10	14.10%	1	1.0
11	11.00%	1	1.0
12	12.60%	1	1.0
13	12.50%	2	2.0
14	11.60%	1	1.0

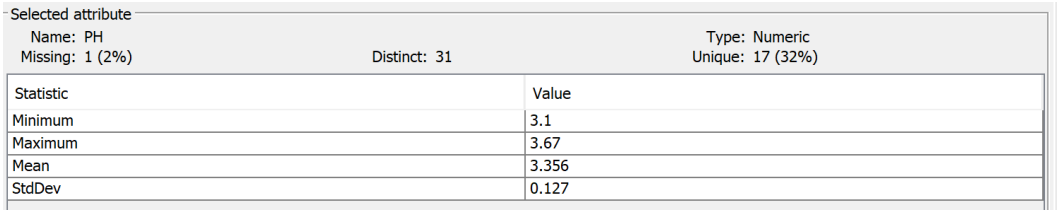
Source: Weka

Figures 10 and 11 are presenting the amount of the alcohol in a Slavonian white and red wine. The alcohol amount for each wine is expressed in percentages and is mostly different for every wine type. For both wine attribute type is nominal, and there is no missing values. White wine counts for 18 different percentages of alcohol that are ranked from the lowest percentage of 11.4% to the highest of 15%, while the biggest number of wine labels (7 of them) contains 13% of alcohol.

Red wine counts for 14 different percentages of alcohol due to smaller number of wines, and are ranked from the lowest percentage of 11% to the highest of 14.4%. The biggest number of wine labels (4 of them) contains 13% of alcohol which is the same as for the white wine.

6.2. pH

Figure 12 pH of Slavonian white wines



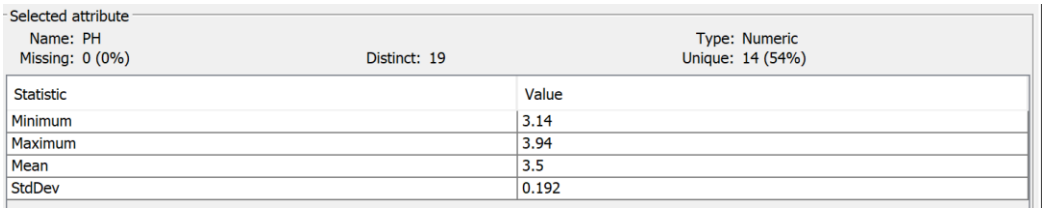
Selected attribute

Name: PH | Missing: 1 (2%) | Distinct: 31 | Type: Numeric | Unique: 17 (32%)

Statistic	Value
Minimum	3.1
Maximum	3.67
Mean	3.356
StdDev	0.127

Source: Weka

Figure 13 pH of Slavonian red wines



Selected attribute

Name: PH | Missing: 0 (0%) | Distinct: 19 | Type: Numeric | Unique: 14 (54%)

Statistic	Value
Minimum	3.14
Maximum	3.94
Mean	3.5
StdDev	0.192

Source: Weka

Figures 12 and 13 above are presenting numeric attribute of pH scale and what is the pH of the Slavonian white and red wines. Describes how acidic or basic wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale.[21] Statistic presents that minimum amount of pH in a white wine is 3.1 and in red wine is 3.14, while the maximum volume is 3.67 and 3.94 respectively. The mean of 3.5 and standard deviation of

[21] Dua, D. and Graff, C. (2019). *Wine data set* [online]. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml [25.08.2020.]

0.192 confirmed that Slavonian red wines are little bit higher on a pH scale than Slavonian white wines, meaning they are healthier as well.

6.2.1. Alcaline vs. Acid food

The reason it is used term "healthier" is due to the constant issue whether alkaline or acid food is healthy or not. Sally Robertson wrote very interesting review (2019) about alkaline diet: "The most important aspect was the explanation on why acid products are highly beneficial where she stated that the alkaline (basic) food is heavy on fruit, vegetables and grains and cuts back on red meat, added sugar and processed and junk foods. Some people find they have more energy, experience fewer gut problems and lose weight while consuming mostly alkaline food. However, this could be a result of cutting back on processed and junk foods rather than being linked to reduced acidity. Unlike some other diets, the alkaline diet does not eliminate entire food groups. People can still consume acid-forming foods; they simply eat less of them."[22]

6.3. Density

Figure 14 Density of Slavonian white wines

Selected attribute

Name: DENSITY | Missing: 4 (8%) | Distinct: 35 | Type: Numeric | Unique: 25 (48%)

Statistic	Value
Minimum	0.911
Maximum	1.023
Mean	0.993
StdDev	0.014

Source: Weka

Figure 15 Density of Slavonian red wines

Selected attribute

Name: DENSITY | Missing: 1 (4%) | Distinct: 19 | Type: Numeric | Unique: 14 (52%)

Statistic	Value
Minimum	0.99
Maximum	1
Mean	0.993
StdDev	0.002

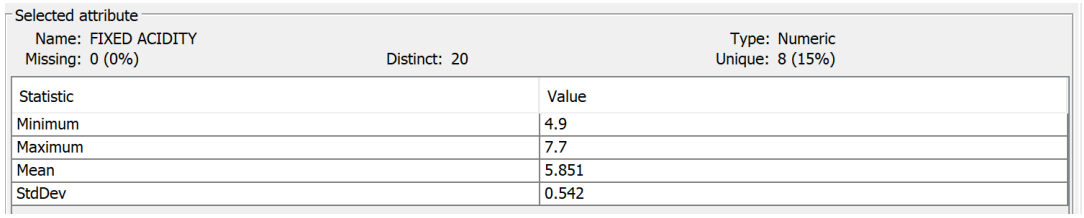
Source: Weka

[22] Robertson S. (2019). *Alcaline diet: Pros and cons* [online]. US: News medical life science. Available at: https://www.news-medical.net/health/Alkaline-Diet-Pros-and-Cons.aspx [26.08.2020.]

Figure above are explaining the numeric attribute of density of the Slavonian white and red wine. The density of wine is close to that of water depending on the percent alcohol and sugar content.[23] Statistic indicate that minimum is 0.911 in a white wine and 0.99 in a red wine, whereas maximum is 1.023 and 1 respectively. The mean of 0.993 and standard deviation of 0.002 are both something higher in a red wine confirming that red wine in Slavonia is denser than Slavonian white wine.

6.4. Fixed Acidity

Figure 16 Fixed acidity in the Slavonian white wines

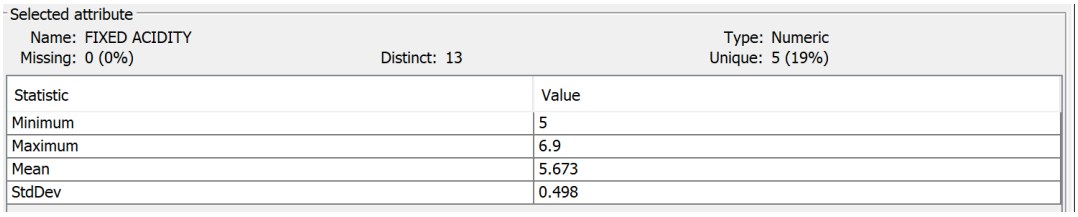

Selected attribute

Name: FIXED ACIDITY Type: Numeric
Missing: 0 (0%) Distinct: 20 Unique: 8 (15%)

Statistic	Value
Minimum	4.9
Maximum	7.7
Mean	5.851
StdDev	0.542

Source: Weka

Figure 17 Fixed acidity in the Slavonian red wines


Selected attribute

Name: FIXED ACIDITY Type: Numeric
Missing: 0 (0%) Distinct: 13 Unique: 5 (19%)

Statistic	Value
Minimum	5
Maximum	6.9
Mean	5.673
StdDev	0.498

Source: Weka

Figures 16 and 17 are presenting the amount of fixed acidity in Slavonian white and red wines and is shown as numeric attribute. The minimum for white wine is 4.9 and for red wine is 5, while maximum is 7.7 and 6.9 respectively. The mean of 5.851 and standard deviation of 0.542 are both higher in white wine affirming that white wine in Slavonia is stronger or tartaric in taste than Slavonian red wine.

[23] Dua, D. and Graff, C. (2019). *Wine data set* [online]. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml [25.08.2020.]

6.5. Volatile Acidity

Figure 18 Volatile acidity in Slavonian white wines

Selected attribute

Name: VOLATILE ACIDITY | Missing: 0 (0%) | Distinct: 9 | Type: Numeric | Unique: 3 (6%)

Statistic	Value
Minimum	0.2
Maximum	4.9
Mean	0.525
StdDev	0.629

Source: Weka

Figure 19 Volatile acidity in Slavonian red wines

Selected attribute

Name: VOLATILE ACIDITY | Missing: 0 (0%) | Distinct: 9 | Type: Numeric | Unique: 2 (8%)

Statistic	Value
Minimum	0.2
Maximum	1
Mean	0.581
StdDev	0.228

Source: Weka

Figure 18 and 19 are showing there is as existence of second part of total acidity - volatile acidity in Slavonian white and red wine. Minimum amount of volatile acidity is 0.2 in both red and white wine, while maximum amount of volatile acidity is 4.9 in white wine and 1 in red wine. Mean is something higher for red wine (0.581), but standard deviation is higher for white wine (0.629). The conclusion is that white wine contains higher amount of citric acid, but since it is not too high, wine has no cheap, vinegar taste.

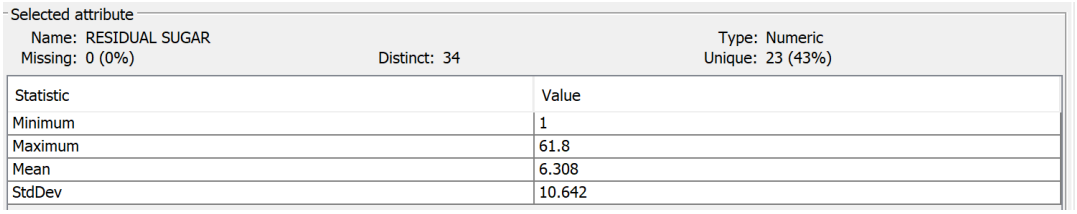
6.5.1. Total Acidity

Doug Nierman in his article (2004) described acids as they are significant constituents of wine and contribute greatly to its taste. In addition, acids impart softness or tartness which is a fundamental feature in wine taste. Wines that are low in acid are “flat”. Chemically the acids control titrable acidity that affects the flavour and pH that affects colour, oxidation quality, and therefore the total lifespan of a wine. He also explained how acidity arises – in the grapes themselves and then carry over into the wine, but there are some acids that arise as a result

of the fermentation process from either yeast and/or bacteria. In general, total acidity is separated into two groups, volatile and non-volatile or fixed acids. [24]

6.6. Residual Sugar

Figure 20 Residual sugar in Slavonian white wines



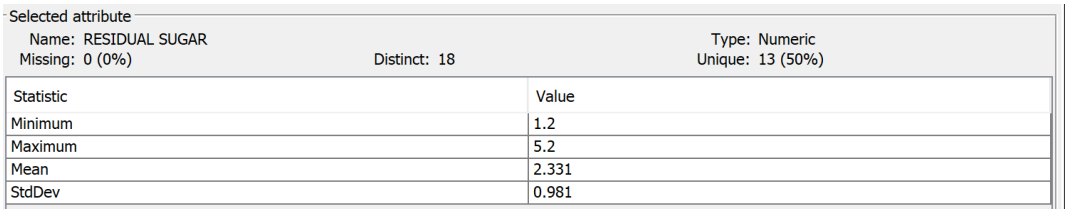
Selected attribute

Name: RESIDUAL SUGAR | Type: Numeric
Missing: 0 (0%) | Distinct: 34 | Unique: 23 (43%)

Statistic	Value
Minimum	1
Maximum	61.8
Mean	6.308
StdDev	10.642

Source: Weka

Figure 21 Residual sugar in Slavonian red wines



Selected attribute

Name: RESIDUAL SUGAR | Type: Numeric
Missing: 0 (0%) | Distinct: 18 | Unique: 13 (50%)

Statistic	Value
Minimum	1.2
Maximum	5.2
Mean	2.331
StdDev	0.981

Source: Weka

Figures 20 and 21 are presenting the numeric attribute of residual sugar of the Slavonian white and red wine. Fermentation is the age-old mechanism by which yeast converts sugar-laden grape juice into wine, while residual sugar is the quantity of sugar remaining after fermentation ends. Wines are considered too sweet if they have either less than 1 gram/litre or more than 45 grams/litre of residual sugar.[25] The minimum in this case is 1 for white wines and 1.2 for red wines, while maximum is 61.8 for white wines and 5.2 for red wines. Both the mean (6.308) and standard deviation (10.642) are obviously higher in Slavonian white wine confirming that it is sweeter than red wine but can be less sweet as well based on it's

[24] Nierman D. (2004). *Fixed acidity* [online]. Department of Viticulture and Enology, University of California. Davis, CA 95616 USA. Available at: https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity#:~:text=The%20predominant%20fixed%20acids%20found,2%2C000%20mg%2FL%20succinic%20acid [25.08.2020.]

[25] Nierman D. (2004). *Fixed acidity* [online]. Department of Viticulture and Enology, University of California. Davis, CA 95616 USA. Available at: https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity#:~:text=The%20predominant%20fixed%20acids%20found,2%2C000%20mg%2FL%20succinic%20acid [26.08.2020.]

minimum. In addition, reason for too "sweet" wines are late harvest ice wines or sweet dessert wines.

6.6.1. The difference between late harvest and ice wines

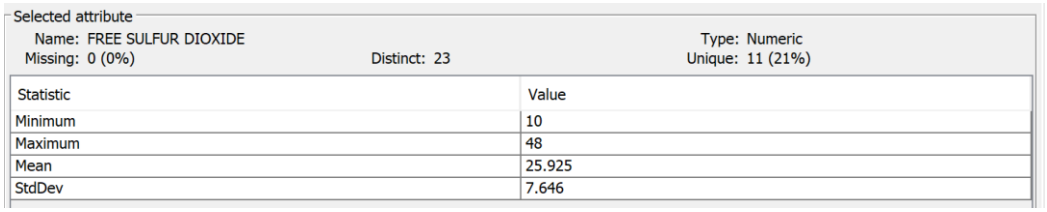
The difference between late harvest and ice wines perfectly explained dr. Vinny in his reply (2007) on question what is distinction between ice wine and late harvest wine? All ice wines are also late-harvest wines, but not all late-harvest wines are ice wines too. Both ice and late harvest wines are made in an extremely sweet style. Late-harvest wines are produced longer than usual from the grapes left on the vine, allowing them to ripen and reap. This makes the grapes oxidize naturally, concentrating their flavours as they take on raisin-like, sweet qualities. On the other hand, if grapes are left on the vine so far past the typical growing season and are frozen before being picked, then ice wine is made. The result is an even more concentrated and sweet wine because the water inside the grapes freezes – while sugar and other solids do not.[26]

Indeed, as explained earlier in the text, real ice wine is extremely hard to find and expensive to produce and to buy. One of the reasons it is expensive for producers is because it only happens during bizarre years when a vineyard is freezing, and then while grapes are still frozen ice-wine must be harvested and pressed. Moreover, the reason it is expensive for buyers is because it is produced in a small quantity due to smaller amount of juice obtained from frozen grapes, and then price goes up abnormally.

[26] Dr. Vinny (2007). *What is distinction between ice wine and late-harvest wine?* [online]. Wine Spectator. Available at: https://www.winespectator.com/articles/whats-the-distinction-between-ice-wine-and-late-harvest-wine-5295#:~:text=Late%2Dharvest%20wines%20are%20made%20from%20grapes%20left%20on%20the,to%20get%20riper%20and%20riper.&text=Ice%20wine%20is%20a%20type,froze%20before%20they%20were%20picked [24.08.2020.].

6.7. Free Sulfur Dioxide

Figure 22 Free sulfur dioxide in Slavonian white wines



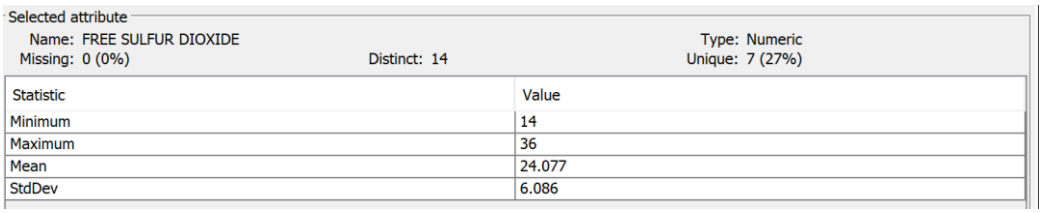
Selected attribute

Name: FREE SULFUR DIOXIDE — Type: Numeric
Missing: 0 (0%) — Distinct: 23 — Unique: 11 (21%)

Statistic	Value
Minimum	10
Maximum	48
Mean	25.925
StdDev	7.646

Source: Weka

Figure 23 Free sulfur dioxide in Slavonian red wines



Selected attribute

Name: FREE SULFUR DIOXIDE — Type: Numeric
Missing: 0 (0%) — Distinct: 14 — Unique: 7 (27%)

Statistic	Value
Minimum	14
Maximum	36
Mean	24.077
StdDev	6.086

Source: Weka

Figures 22 and 23 are presenting the amount of free sulfur dioxide in Slavonian white and red wines in a form of numeric attributes. The minimum is 10 for white wine and 14 for red wine, while maximum is 48 and 36 respectively. Both the mean (25.925) and standard deviation (7.646) are higher for Slavonian white wine meaning they usually comprise higher amount of free sulfur dioxide compared to Slavonian red wines. Free sulfur dioxide is eliminating microbial development and wine oxidation. Also, it is free source of SO2 and occurs in a equilibrium between molecular SO2 (as a dissolved gas) and bisulphite ion.[27]

[27] Nierman D. (2004). *Fixed acidity* [online]. Department of Viticulture and Enology, University of California. Davis, CA 95616 USA. Available at: https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity#:~:text=The%20predominant%20fixed%20acids%20found,2%2C000%20mg%2FL%20succinic%20acid [26.08.2020.]

6.8. Total Sulfur dioxide

Figure 24 Total sulfur dioxide in Slavonian white wines

Selected attribute

Name: TOTAL SULFUR DIOXIDE Type: Numeric
Missing: 0 (0%) Distinct: 42 Unique: 31 (58%)

Statistic	Value
Minimum	64
Maximum	212
Mean	116.321
StdDev	33.783

Source: Weka

Figure 25 Total sulfur dioxide in Slavonian red wines

Selected attribute

Name: TOTAL SULFUR DIOXIDE Type: Numeric
Missing: 0 (0%) Distinct: 18 Unique: 12 (46%)

Statistic	Value
Minimum	62
Maximum	160
Mean	90.538
StdDev	26.922

Source: Weka

Figures 24 and 25 are showing the numerical attributes of total sulfur dioxide of both Slavonian white and red wine. The minimum value is 64 for white and 62 for red wine, while maximum value is 212 for white and 160 for red wine. The mean of 116.321 and standard deviation of 33.783 are both higher for white wine confirming that white wine in Slavonia (together with free sulfur dioxide) comprise higher amount of total sulfur dioxide.

6.8.1. Why total and free sulfur dioxide are important?

Tanya M. Monro with colleagues emphasized why total and free sulfur dioxide are important for wines. Two classes of sulfites are found in wine: free and bound. The free sulfites are those available to react and thus exhibit both germicidal and antioxidant properties. The bound sulfites are those that have reacted (both reversibly and irreversibly) with other molecules within the wine medium. The sum of the free and bound sulfites defines the total sulfite concentration.[28]

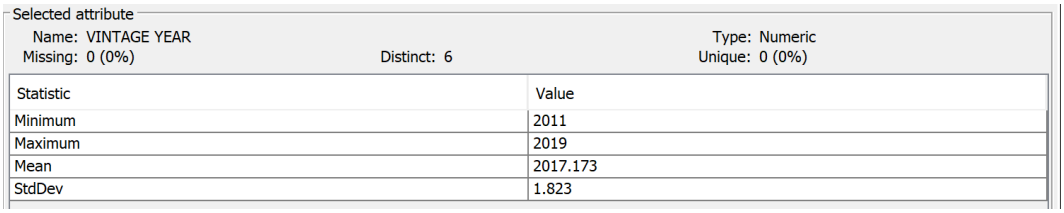
They also explained the connection between SO_2 (sulfur dioxide) and pH. SO_2 is applied as a preservative to the must and juice to avoid bacterial growth and delay the oxidation process

[28] Tanya M. Monro (2012). *Sensing free sulfur dioxide in wine* [online]. Sensors, Basel. US National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3472855/ [26.08.2020.]

by preventing oxidative enzymes. SO_2 also boosts the taste and retains the fruity aromas and freshness of the wine. They said that after introducing, SO_2 is usually applied to either potassium or sodium metabisulfite producing a pH-dependent speciation in solution. The prevalent species is molecular sulfur dioxide (SO_2) that expresses germicidal properties. However, the major species at the pH of the wine (between 3.0 and 3.8) is the bisulfite anion (HSO_3^-) which acts as an antioxidant.[29]

6.9. The Vintage Year

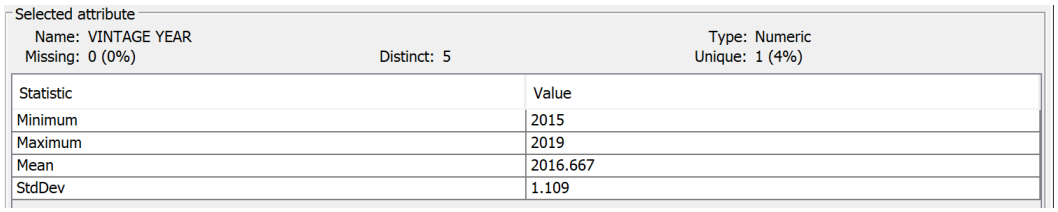
Figure 26 Vintage year of Slavonian white wines


Selected attribute
Name: VINTAGE YEAR Type: Numeric
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

Statistic	Value
Minimum	2011
Maximum	2019
Mean	2017.173
StdDev	1.823

Source: Weka

Figure 27 Vintage year of Slavonian red wines


Selected attribute
Name: VINTAGE YEAR Type: Numeric
Missing: 0 (0%) Distinct: 5 Unique: 1 (4%)

Statistic	Value
Minimum	2015
Maximum	2019
Mean	2016.667
StdDev	1.109

Source: Weka

Figures 26 and 27 are presenting numerical attributes of the year of production of Slavonian white and red wine. The minimum or the first year that appears in white wine cluster is 2011, while for red wine cluster is 2015. The maximum year that appears in both clusters is 2019. Mean and standard deviation do not have specific meaning. One fun fact about wines is that the year of production or vintage year on a wine label represents the grape's harvest year from which the wine was made. Also, the quality of wine highly depends on the year grapes were harvested.

[29] Tanya M. Monro (2012). *Sensing free sulfur dioxide in wine* [online]. Sensors, Basel. US National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3472855/ [26.08.2020.]

Grainger K. and Tattersall H. described it in their book (2016) as: "Wine is an agricultural product and as such is at mercy of the weather, diseases and other natural forces that can result in considerable annual variation in both quantity and quality."[30]

6.10. The Wine Name

Figure 28 The names of Slavonian white wines

Selected attribute

Name: WINE NAME | Type: Nominal
Missing: 0 (0%) | Distinct: 20 | Unique: 12 (23%)

No.	Label	Count	Weight
2	Rajnski rizling	1	1.0
3	Pinot sivi	6	6.0
4	Chardonnay	7	7.0
5	Pjenusac Jazz	1	1.0
6	Pjenusac Tango	1	1.0
7	Grasevina Mit.	2	2.0
8	Grasevina	8	8.0
9	Grasevina k.b.	4	4.0
10	Sauvignon	4	4.0
11	Zelenac Kutjevo	2	2.0
12	Grasevina starac	1	1.0
13	G* tocka bijela	1	1.0
14	Sauvignon Blanc	1	1.0
15	Posip	1	1.0
16	Bijelo 9	1	1.0
17	Grasevina de Gotho	1	1.0
18	Laski rizling	1	1.0
19	Traminac	1	1.0
20	Maximo bianco	1	1.0

Source: Weka

Figure 29 The names of Slavonian red wines

Selected attribute

Name: WINE NAME | Type: Nominal
Missing: 0 (0%) | Distinct: 13 | Unique: 6 (23%)

No.	Label	Count	Weight
1	Pinot crni	5	5.0
2	Stari bokter cuvee	2	2.0
3	Cabernet sauvignon	2	2.0
4	Merlot	2	2.0
5	Mercs	1	1.0
6	Crveni cuvee	2	2.0
7	Rose cuvee	3	3.0
8	Cuvee	1	1.0
9	Crno 9	1	1.0
10	G* tocka crna	1	1.0
11	Crno vino	1	1.0
12	Maximo nero	1	1.0
13	Rose	4	4.0

Source: Weka

[30] Grainger K. and Tattersall H. (2016). *Wine production and quality* [online]. 2nd edition. Oxford : Wiley. Available at: https://books.google.hr/books?id=9KjLCgAAQBAJ&pg=PA266&dq=year+of+wine+production&hl=en&sa=X&ved=2ahUKEwjR75Sf37rrAhUJ3qQKHW2bA-EQ6AEwAXoECAAQAg#v=onepage&q=year%20of%20wine%20production&f=false [27.08.2020.]

Figures 28 and 29 are representation of official list of all Slavonian wine types from most important Slavonian producers and thus relevant for this master thesis. The upper figure shows the nominal attribute of white wine's names while the lower figure shows the nominal attribute of red wine's names.

There are 20 different labels for Slavonian white wine, while the most common name in Slavonia is „Graševina“ for white wine which count 24 labels in total. 16 of them are of original name and can be seen at positions 1 and 8, and there are 8 additional Graševina labels – „Graševina Mit.“ (meaning surname Mitrovic which its name comes from) counts 2 labels at position 7, „Graševina k.b.“ (meaning kasna berba or late harvest wine) counts 4 labels at position 9, „Graševina starac“ counts 1 label at position 12, and „Graševina de Gotho“ counts 1 label at position 17. The next very represented labels are „Chardonnay“ with 7 labels under position 4, „Pinot sivi“ with 6 labels under position 3, „Sauvignon“ with 4 labels under position 10, and „Zelenac Kutjevo“ with 2 labels under position 11. There are also per 1 label from different producers of „Rajnski rizling“ under position 2, „Pjenušac Jazz“ and „Pjenušac Tango“ under positions 5 and 6 respectively (pjenušac in translation means champagne), then „G* točka bijela“ under position 13, „Sauvignon blanc“ under position 14, „Pošip“ under position 15, „Bijelo 9“ under position 16, „Laski rizling“ under position 18, „Traminac“ under position 19, and very last „Maximo Bianco“ under position 20.

Moreover, in Slavonia red wine counts for 13 different labels, while the most common name is „Pinot crni“ and counts for 5 labels under position 1. Popular are also „Stari bokter cuvee“ which counts for 2 labels under position 2, „Cabernet sauvignon“ with 2 labels under position 3, „Merlot“ with 2 labels under position 4, and „Crveni cuvee“ with 2 labels under position 6. There are also per 1 label from different producers of „Mercs“ under position 5, „Cuvee“ under position 8, „Crno 9“ under position 9, „G* točka crna“ under position 10, „Crno vino“ under position 11, and very last „Maximo Nero“ under position 12. Under red wine cluster can be found „Rose wines“ – they are made from a red grapes in a way that red grape skin only touch wine for a short period of time and then they ferment few hours to 2 or 3 days, depending on winemaker's decision (if compare to some red wines which ferment for weeks). Almost any red wine can be used to make rose wines, but only several common types and grapes are

preferred for rose. Here are 2 types „Rose cuvee“ which counts for 3 labels under position 7 and „Rose“ with 4 labels under position 13.

6.11. The Producers

Figure 30 The producers of Slavonian white wines

Selected attribute

Name: PRODUCER Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count	Weight
1	Jakobovic	5	5.0
2	Josipovic	3	3.0
3	Krauthaker	19	19.0
4	Markota	4	4.0
5	Galic	8	8.0
6	Soldo-C	6	6.0
7	Kutjevo	8	8.0

Source: Weka

Figure 31 The producers of Slavonian red wines

Selected attribute

Name: PRODUCER Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 1 (4%)

No.	Label	Count	Weight
1	Jakobovic	5	5.0
2	Josipovic	1	1.0
3	Krauthaker	8	8.0
4	Markota	3	3.0
5	Galic	3	3.0
6	Soldo-C	3	3.0
7	Kutjevo	3	3.0

Source: Weka

Figure 30 and 31 representing the official list of the best producers in Slavonia that have been examined and analysed for this master thesis. Once again, there are 80 different wines of which 54 are white and 26 are red wines. The list of producers is nominal and count for 7 Slavonian producers for both white and red wine. As can be seen, winemakers are the same for both red and white wine and producing specific amount of wines. „Jakobović“ is producing 5 white and 5 red wines, „Josipovic“ 3 white and 1 red wine, „Krauthaker“ is the largest one with 19 white and 8 red wines, then „Markota“ is producing 4 white and 3 red wine, „Galić“ 8 white and 3 red wines, „Soldo-Čamak“ 6 white and 3 red wines, and „Kutjevo“ is producing 8 white and 3 red wines.

7. THE RESULTS OF THE CLUSTER ANALYSIS

Again, here is seventy-nine (79) instances in total of which 54 are from Slavonian white wine and 26 are from Slavonian red wine, and they are going to be explained in every relevant detail. There is 7 main wine producers in Slavonia from which data have been collected and thus they represent 7 different clusters that are going to be analysed. It is also important to highlight that there have been 4 missing values within white wine dataset and 1 missing value within red wine dataset.

7.1. The results of the Slavonian white wine research

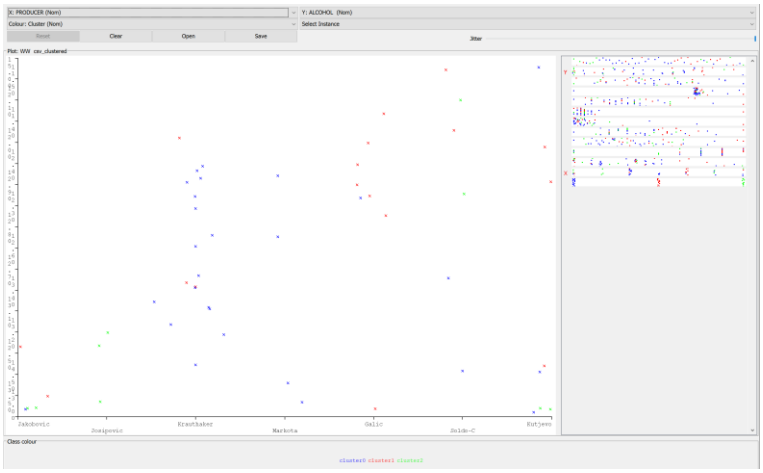
For the results of the Slavonian white wine research SimpleKMeans with 3 clusters and 10 seeds was chosen in a Weka program, whereas model and evaluation are based on a training set. 3 clusters were randomly chosen. There is 4 missing values under this dataset.

Figure 32 The results of the Slavonian white wine – alcohol volume and vintage year

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDist
Relation:     WW  csv
Instances:    53
Attributes:   11
              ALCOHOL
              VINTAGE YEAR
              WINE NAME
              PRODUCER
Ignored:
              PH
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              RESIDUAL SUGAR
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 106.81284041394335

Initial starting points (random):

Cluster 0: '12.90\%',2017,Chardonnay,Krauthaker
Cluster 1: '12.40\%',2019,'G* tocka bijela',Galic
Cluster 2: '13.00\%',2015,'Pjenusac Tango',Josipovic

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                           (53.0)     (27.0)     (17.0)      (9.0)
=============================================================================
ALCOHOL                    13.00%     13.50%     12.00%     13.00%
VINTAGE YEAR           2017.1698  2017.1481  2018.2353  2015.2222
WINE NAME              Grasevina  Chardonnay Grasevina  Grasevina
PRODUCER               Krauthaker Krauthaker      Galic  Josipovic



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      27 ( 51%)
1      17 ( 32%)
2       9 ( 17%)
```

Source: Weka

Figure 33 Visualisation of alcohol volume of a white wines by different wine producers in Slavonia



Source Weka Clusterer Visualize

Figure 32 representing that 3 clusters were chosen to analyse which 3 Slavonian white wines, of which producers, and within which vintage year contain the highest percentage of alcohol. The results shows that within 53 cluster instances, there are 27 or 51% samples in a cluster 0, 17 or 32% samples in a cluster 1, and 9 or 17% samples in a cluster 2. There are 7 ignored attributes and they are pH, density, fixed acidity, volatile acidity, residual sugar, free sulfur dioxide and total sulfur dioxide. Number of iterations is 5, while within cluster sum of squared errors is 106.81284041394335.

The highest percentage of alcohol (13.5%) has white wine Chardonnay, wines from 2017 vintage year, and Krauthaker winery based on 0th cluster centroid. The full data indicates that 13% of alcohol volume is the mean between 53 different white wines in Slavonia where Krauthaker is the wine producer that mostly produces wines that contain this alcohol percentage. However, the highest percentage of all has Graševina de Gotho by Kutjevo winery (15%), while the lowest has Graševina by Galić winery (11,4%) but this information is not visible here. Krauthaker is the producer of the largest scale of white wine types in Slavonia and the most gifted year was 2017, while the most produced wine with mean percentage is Graševina. The results are illustrated in a figure 33 where on a x-axis lies all 7 producers and the clusters they belong to, while on y-axis is alcohol volume that jittering stronger if a

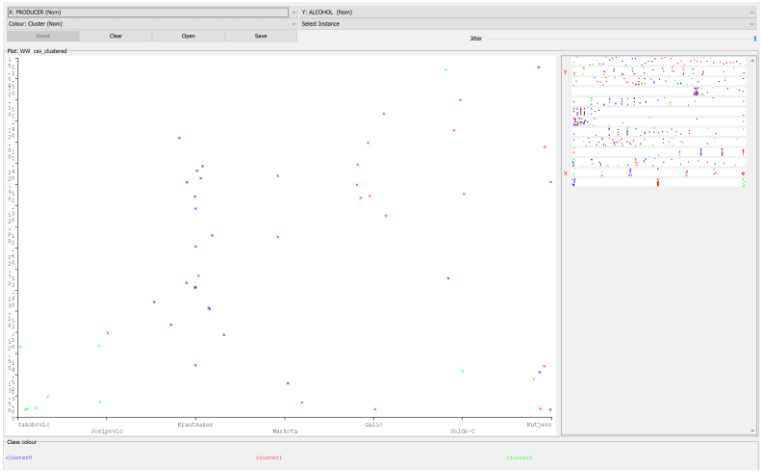
producer holds for bigger scale of produced wines and if most of them are within specific alcohol percentage.

Figure 34 The results of the Slavonian white wine – pH

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     WW  csv
Instances:    53
Attributes:   11
              PH
              WINE NAME
              PRODUCER
Ignored:
              ALCOHOL
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              RESIDUAL SUGAR
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 60.501068061313205

Initial starting points (random):

Cluster 0: 3.37,Chardonnay,Krauthaker
Cluster 1: 3.34,'G* tocka bijela',Galic
Cluster 2: 3.48,'Pjenusac Tango',Josipovic

Missing values globally replaced with mean/mode


Final cluster centroids:
                                    Cluster#
Attribute               Full Data          0          1          2
                           (53.0)     (23.0)     (20.0)     (10.0)
=============================================================================
PH                          3.3562     3.3896     3.2613      3.469
WINE NAME                Grasevina  Chardonnay  Grasevina Grasevina
PRODUCER                 Krauthaker Krauthaker      Galic  Jakobovic



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      23 ( 43%)
1      20 ( 38%)
2      10 ( 19%)
```

Source: Weka

Figure 35 Visualisation of pH volume of a white wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Figure 34 is representing that three clusters were chosen again to analyse which of 3 Slavonian white wines, by which producer are the most acid or contain lowest pH worth. The results are showing that within 53 cluster instances, there are 23 or 43% samples in a cluster 0, 20 or 38% samples in a cluster 1, and 10 or 19% samples in a cluster 2. There are 9 ignored attributes and they are alcohol, density, fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, and vintage year. Number of iterations is 4, while within cluster sum of squared errors is 60.501068061313205.

Based on a cluster number 2 the highest pH mean of 3.469 has Graševina wine meaning colour is stronger, oxidation quality is stronger, and the total lifespan of Graševina is longer compared to other white wines in Slavonia. However, the information that is not shown here is that Graševina k.b. or late harvest white wine by Krauthaker is the highest on a pH scale (3,67) meaning it is the least sour wine in Slavonia and healthiest as well, while Pjenušac Jazz by Josipovic winery is the lowest on a pH scale (3,1) and thus most acid wine in Slavonia. Krauthaker winery is the producer of most acid wines in Slavonia based on a full data, while on a pH mean in Slavonia mostly lies Graševina wine.

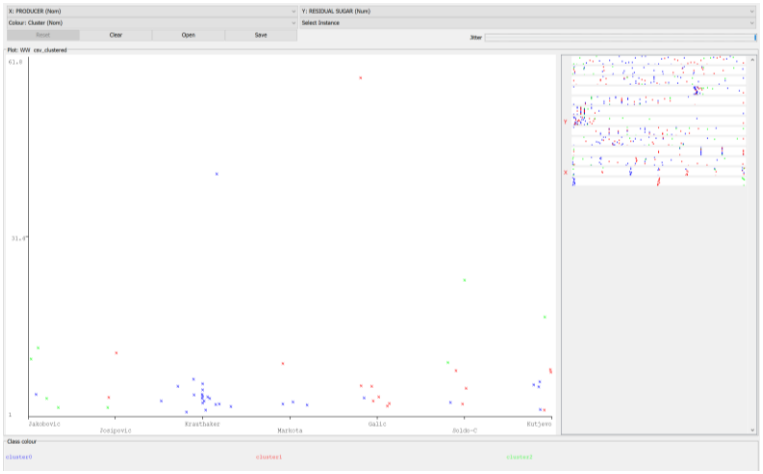
Figure 35 representing visualisation of the results. On a x-axis lies all 7 producers and the clusters they belong to, while on y-axis is pH scale that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific acidity level.

Figure 36 The results of the Slavonian white wine – residual sugar

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     WW  csv
Instances:    53
Attributes:   11
              RESIDUAL SUGAR
              WINE NAME
              PRODUCER
Ignored:
              ALCOHOL
              PH
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 60.50992505449496

Initial starting points (random):

Cluster 0: 1.3,Chardonnay,Krauthaker
Cluster 1: 2.6,'G* tocka bijela',Galic
Cluster 2: 6.8,'Pjenusac Tango',Josipovic

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                           (53.0)     (29.0)     (16.0)      (8.0)
=============================================================
RESIDUAL SUGAR            6.3075     4.2759     7.8125    10.6625
WINE NAME              Grasevina  Chardonnay  Grasevina Grasevina
PRODUCER               Krauthaker Krauthaker      Galic  Jakobovic



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      29 ( 55%)
1      16 ( 30%)
2       8 ( 15%)
```

Source: Weka

Figure 37 Visualisation of residual sugar in a white wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

As can be seen from figure 36, 3 clusters were chosen to analyse which producers and which 3 Slavonian white wines contain the highest amount of residual sugar. The results are showing that within 53 cluster instances, there are 29 or 55% samples in a cluster 0, 16 or 30% samples in a cluster 1, and 8 or 15% samples in a cluster 2. There are 8 ignored attributes and they are alcohol, pH, density, fixed acidity, volatile acidity, free sulfur dioxide, total sulfur dioxide, and vintage year. Number of iterations is 4, while within cluster sum of squared errors is 60.50992505449496.

The original data indicates the highest residual sugar has Graševina k.b. or late harvest by Galić winery (61,8), while the lowest amount has Chardonnay and Sauvignon by Krauthaker (0,91). It is also confirmed by the full data and clusters number 1 and 2 that the white wine that contains the highest amount of residual sugar in Slavonia is Graševina (even if it does not illustrates which one), while the cluster number 0 confirms Chardonnay by Krauthaker is the one that contains lowest amount. On a figure 37 can be seen that on a x-axis lies all 7 producers and the clusters they belong to, while on y-axis is residual sugar that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific level of sweetness.

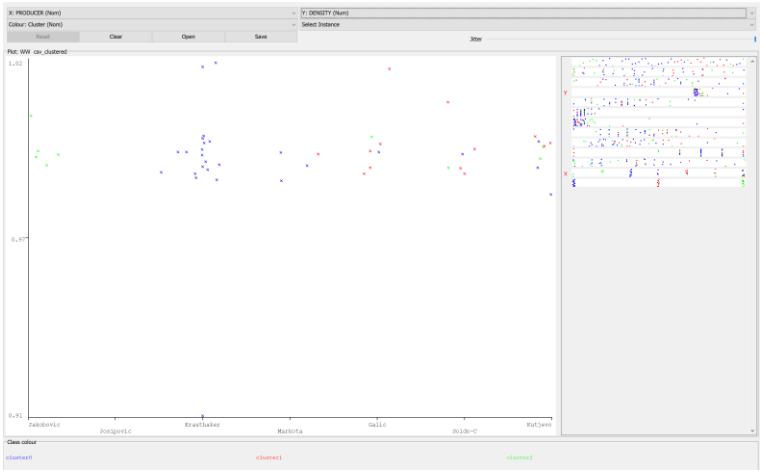
Figure 38 The results of the Slavonian white wine – density, alcohol, and residual sugar

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     WW  csv
Instances:    53
Attributes:   11
              ALCOHOL
              DENSITY
              RESIDUAL SUGAR
              WINE NAME
              PRODUCER
Ignored:
              PH
              FIXED ACIDITY
              VOLATILE ACIDITY
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 100.18941895216265

Initial starting points (random):

Cluster 0: '12.90\%',0.911,1.3,Chardonnay,Krauthaker
Cluster 1: '12.40\%',0.992,2.6,'G* tocka bijela',Galic
Cluster 2: '13.00\%',0.992631,6.8,'Pjenusac Tango',Josipovic

Missing values globally replaced with mean/mode
```

```
Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                          (53.0)     (27.0)     (14.0)     (12.0)
=============================================================================
ALCOHOL                   13.00%     13.40%     12.00%     13.00%
DENSITY                   0.9926     0.9905      0.996     0.9935
RESIDUAL SUGAR            6.3075      4.237      10.55     6.0167
WINE NAME               Grasevina  Chardonnay  Grasevina Grasevina
PRODUCER                Krauthaker Krauthaker      Galic  Jakobovic


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      27 ( 51%)
1      14 ( 26%)
2      12 ( 23%)
```

Source: Weka

Figure 39 Visualisation of density of a white wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

To analyse which of the 3 Slavonian wines are the densest, 3 clusters were chosen as can be seen on figure 38. For this cluster analysis residual sugar and alcohol are included since density depends on these attributes. The results are showing that within 53 cluster instances, there are 27 or 51% samples in a cluster 0, 14 or 26% examples in a cluster 1, and 12 or 23% examples in a cluster 2. There are 6 ignored attributes and they are pH, fixed acidity, volatile acidity, free sulfur dioxide, total sulfur dioxide, and vintage year. Number of iterations is 6, while within cluster sum of squared errors is 100.18941895216265.

The highest density mean of 0.996 is Graševina because it's alcohol volume is 12% (the lowest compared to other analysed wines), and the amount of residual sugar is 10.55 (the highest compared to other analysed wines). It can be seen in a cluster number 1 and is confirmed by the full data that Graševina is the mean densest of all analysed Slavonian wines. The full data

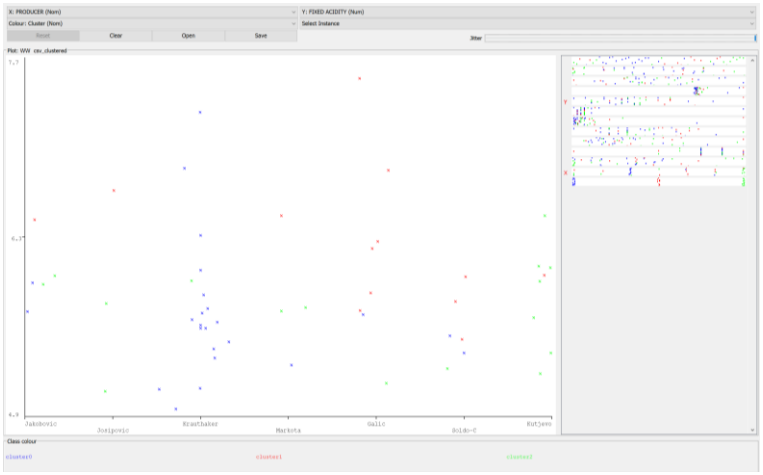
also explaining that Krauthaker winery produces most quality wines that contains the highest alcohol percentage, lowest residual sugar and are the least dense (can be seen from cluster number 0). However, the real data that is not visible here but is original one that indicates the highest density has Graševina k.b. or late harvest by Galić vinery (1,023), while the lowest amount of the density has Chardonnay by Krauthaker (0,91). This is also confirmed by clusters 1 and 0. On a figure 39 can be seen that on a x-axis lies all 7 producers and the clusters they belong to, while on y-axis is density that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific density level.

Figure 40 The results of the Slavonian white wine – fixed and volatile acidity

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     WW  csv
Instances:    53
Attributes:   11
              FIXED ACIDITY
              VOLATILE ACIDITY
              WINE NAME
              PRODUCER
Ignored:
              ALCOHOL
              PH
              DENSITY
              RESIDUAL SUGAR
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 60.490038972232604

Initial starting points (random):

Cluster 0: 5.6,0.7,Chardonnay,Krauthaker
Cluster 1: 5.7,0.3,'G* tocka bijela',Galic
Cluster 2: 5.9,0.5,'Pjenusac Tango',Josipovic

Missing values globally replaced with mean/mode


Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                           (53.0)     (24.0)     (13.0)     (16.0)
=============================================================================
FIXED ACIDITY              5.8509     5.6875     6.2538     5.7687
VOLATILE ACIDITY           0.5245     0.6333     0.4154       0.45
WINE NAME              Grasevina  Chardonnay  Grasevina Grasevina
PRODUCER               Krauthaker Krauthaker      Galic    Kutjevo



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      24 ( 45%)
1      13 ( 25%)
2      16 ( 30%)
```

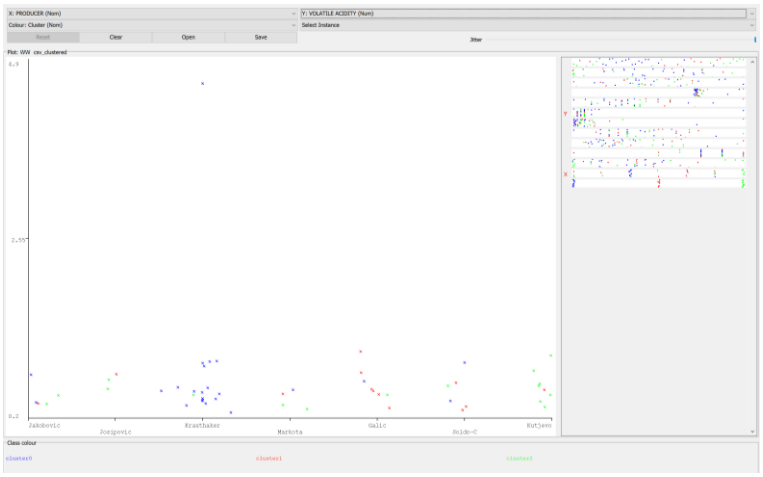
Source: Weka

Figure 41 Visualisation of fixed acidity white wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Figure 42 Visualisation of volatile acidity of white wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Three clusters were chosen to analyse which of 3 Slavonian wines, by which producer contain highest volume of fixed and volatile acidity as can be seen from a figure 40 . The results are showing that within 53 cluster instances, there are 24 or 45% samples in a cluster 0, 13 or 25% samples in a cluster 1, and 16 or 30% samples in a cluster 2. There are 7 ignored attributes and they are alcohol, pH, density, residual sugar, free sulfur dioxide, total sulfur dioxide, and vintage year. Number of iterations is 5, while within cluster sum of squared errors is 60.490038972232604.

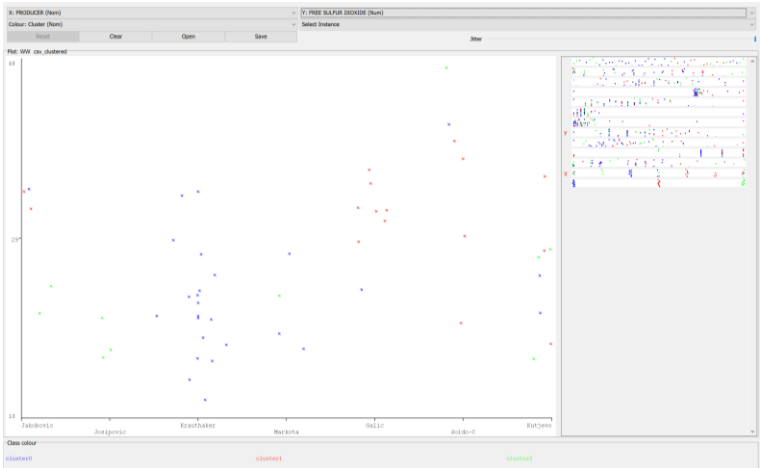
The original data indicates the highest fixed acidity has Graševina k.b. or late harvest by Galić winery (7,7), while the lowest amount has Zelenac Kutjevo by Krauthaker winery (4,9). The highest volatile acidity has Graševina by Krauthaker winery (4,9), while the lowest volatile acidity have Graševina by Soldo-Čamak winery, Sauvignon blanc by Krauthaker winery, and Pinot sivi by Krauthaker winery (0,2 for all of them). It is also confirmed by the full data and cluster number 0 that in Slavonia Krauthaker producing wines that evaporate the longest, while full data and clusters number 1 and 2 confirms Graševina is the white wine in Slavonia that would least lead to unpleasant taste of vinegar. The results are illustrated above where on x-axis lies all 7 producers and the clusters they belong to, while on y-axis are fixed acidity (figure 41) and volatile acidity (figure 42) that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific either fixed or volatile acidity level.

Figure 43 The results of the Slavonian white wine – free sulfur dioxide and total sulfur dioxide

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     WW  csv
Instances:    53
Attributes:   11
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              WINE NAME
              PRODUCER
Ignored:
              ALCOHOL
              PH
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              RESIDUAL SUGAR
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 63.91468867100471

Initial starting points (random):

Cluster 0: 22,130,Chardonnay,Krauthaker
Cluster 1: 34,114,'G* tocka bijela',Galic
Cluster 2: 18,133,'Pjenusac Tango',Josipovic

Missing values globally replaced with mean/mode


Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                           (53.0)     (27.0)     (16.0)     (10.0)
=============================================================================
FREE SULFUR DIOXIDE        25.9245    23.2593     31.375       24.4
TOTAL SULFUR DIOXIDE      116.3208   103.3704   129.4375       130.3
WINE NAME                Grasevina  Chardonnay  Grasevina Grasevina
PRODUCER                 Krauthaker Krauthaker      Galic  Josipovic



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      27 ( 51%)
1      16 ( 30%)
2      10 ( 19%)
```

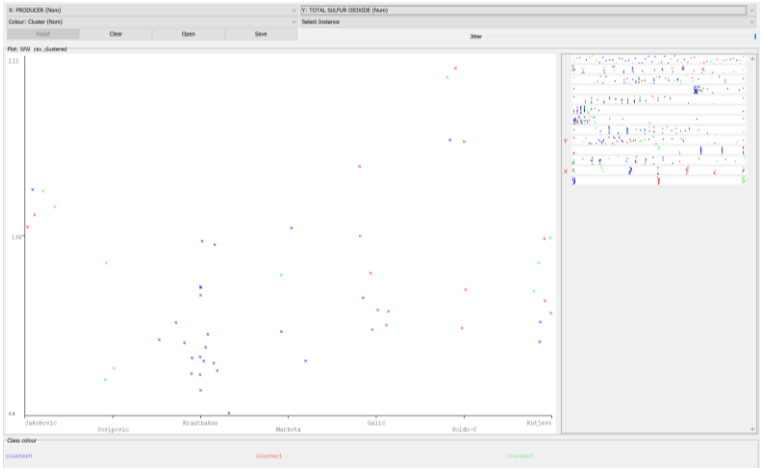
Source: Weka

Figure 44 Visualisation of free sulfur dioxide of a white wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Figure 45 Visualisation of total sulfur dioxide of a white wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Three clusters were chosen to analyse which producer and which of 3 Slavonian wines contain the highest volume of free and total sulfur dioxide, as can be seen from figure 43 above. The results are showing that within 53 cluster instances, there are 27 or 51% samples in a cluster 0, 16 or 30% samples in a cluster 1, and 10 or 19% samples in a cluster 2. There are 7 ignored attributes and they are alcohol, pH, density, residual sugar, fixed acidity, volatile acidity and vintage year. Number of iterations is 6, while within cluster sum of squared errors is 63.91468867100471.

The original data indicates the highest amount of free sulfur dioxide has Graševina by Soldo-Čamak winery (48), while the lowest amount has Chardonnay by Krauthaker winery (10). The highest amount of total sulfur dioxide has Graševina by Soldo-Čamak winery (212), while the lowest amount has Pinot sivi by Krauthaker winery (64). On a figure 43, regarding free sulfur dioxide is also confirmed by full data and clusters number 1 and 2 that Graševina is the white wine in Slavonia that have the most eliminated microbial development and wine oxidation, while Krauthaker is the producer of white wines which have the lowest total sulfur dioxide and thus least are intense in taste and aroma because SO_2 is near and below 50ppm (below that is undetectable).

The results are illustrated above where on x-axis lies all 7 producers and the clusters they belong to, while on y-axis are free sulfur dioxide (figure 44) and total sulfur dioxide (figure 45)

that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific either free or volatile sulfur dioxide level.

Figure 46 The results of the Slavonian white wine – no attribute ignored

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     WW  csv
Instances:    53
Attributes:   11
              ALCOHOL
              PH
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              RESIDUAL SUGAR
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
              WINE NAME
              PRODUCER
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 110.7212392255042

Initial starting points (random):

Cluster 0: '12.90\%',3.37,0.911,5.6,0.7,1.3,22,130,2017,Chardonnay,Krauthaker
Cluster 1: '12.40\%',3.34,0.992,5.7,0.3,2.6,34,114,2019,'G* tocka bijela',Galic
Cluster 2: '13.00\%',3.48,0.992631,5.9,0.5,6.8,18,133,2015,'Pjenusac Tango',Josipovic

Missing values globally replaced with mean/mode


Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                           (53.0)     (25.0)     (17.0)     (11.0)
=============================================================================
ALCOHOL                    13.00%     13.40%     12.00%     13.00%
PH                          3.3562       3.39      3.248     3.4464
DENSITY                     0.9926     0.9904     0.9946     0.9946
FIXED ACIDITY               5.8509        5.7     6.1647     5.7091
VOLATILE ACIDITY            0.5245      0.636        0.4     0.4636
RESIDUAL SUGAR              6.3075      4.496     7.7176     8.2455
FREE SULFUR DIOXIDE        25.9245       22.4    28.9412    29.2727
TOTAL SULFUR DIOXIDE      116.3208      104.8        115   144.5455
VINTAGE YEAR             2017.1698    2017.52  2017.0588  2016.5455
WINE NAME                Grasevina  Chardonnay  Grasevina Grasevina
PRODUCER                 Krauthaker Krauthaker      Galic  Jakobovic



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      25 ( 47%)
1      17 ( 32%)
2      11 ( 21%)
```

Source: Weka

The figure 46 illustrates overall results of the Slavonian white wine with no attribute ignored. This show that if we decide to change the number of clusters and if we play with the number of ignored attributes, we can get as many combinations and results as possible, and thus can make many conclusions. However, the full data that is most relevant and remain unchanged

permanently (whether we decide to change the attributes, number of clusters or seeds), except we make a change in an inserted dataset.

7.2. The results of the Slavonian red wine research

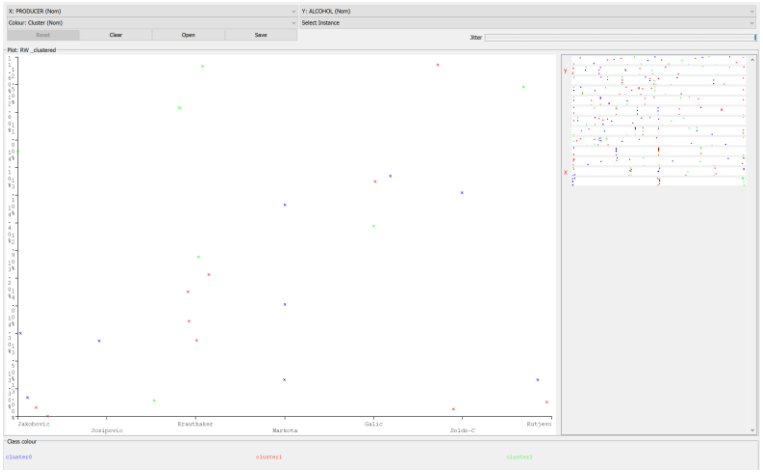
For the results of the Slavonian red wine research SimpleKMeans with 3 clusters and 10 seeds was chosen in a Weka program, whereas model and evaluation are based on a training set. 3 clusters were randomly chosen. There is 1 missing values under this dataset.

Figure 47 The results of the Slavonian red wine – alcohol volume and vintage year

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDist
Relation:     RW
Instances:    26
Attributes:   11
              ALCOHOL
              VINTAGE YEAR
              WINE NAME
              PRODUCER
Ignored:
              PH
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              RESIDUAL SUGAR
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 49.49871031746032

Initial starting points (random):

Cluster 0: '14.10\%',2015,'Crno 9',Galic
Cluster 1: '13.50\%',2017,'Pinot crni',Krauthaker
Cluster 2: '12.60\%',2017,'Rose cuvee',Krauthaker

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                Full Data             0                1                2
                            (26.0)           (9.0)           (10.0)           (7.0)
=============================================================================================
ALCOHOL                     13.00%          13.30%           13.00%           12.90%
VINTAGE YEAR             2016.6538       2015.4444           2016.9        2017.8571
WINE NAME                Pinot crni Stari bokter cuvee      Pinot crni       Rose cuvee
PRODUCER                 Krauthaker            Markota      Krauthaker       Krauthaker



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       9 ( 35%)
1      10 ( 38%)
2       7 ( 27%)
```

Source: Weka

Figure 48 Visualisation of alcohol volume of a red wines by different wine producers in Slavonia



Source Weka Clusterer Visualize

Figure 47 is representing that 3 clusters were chosen to analyse which 3 Slavonian red wines, of which producers, and within which vintage year contain the highest percentage of alcohol. The results shows that within 26 cluster instances, there are 9 or 35% samples in a cluster 0, 10 or 38% samples in a cluster 1, and 7 or 27% samples in a cluster 2. There are 7 ignored attributes and they are pH, density, fixed acidity, volatile acidity, residual sugar, free sulfur dioxide and total sulfur dioxide. Number of iterations is 3, while within cluster sum of squared errors is 49.49871031746032.

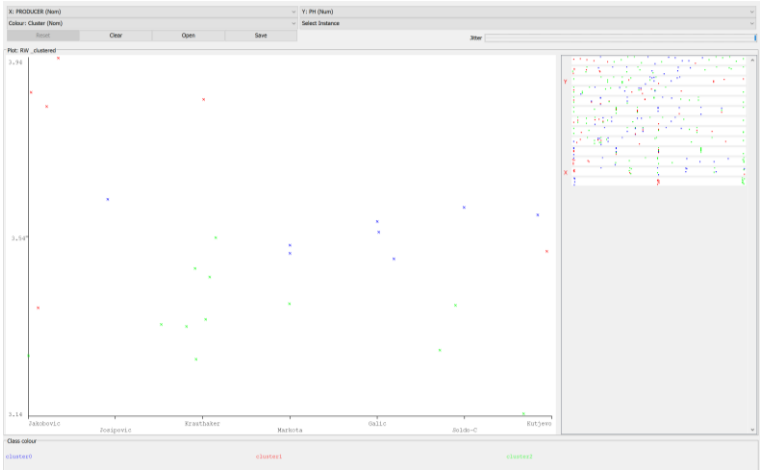
The highest percentage of alcohol (13.3%) has wine Stari bokter cuvee, wines from 2015 vintage year, and Markota winery based on 0^{th} cluster centroid. However, the highest percentage actually has Merlot wine by Markota winery (14.4%), while the lowest has Rose by Jakobović winery (11%) but this information is not visible here. The full data indicates that 13% of alcohol volume is the mean between 26 different red wines in Slavonia where Pinot crni wine is the one that mostly contains referred alcohol percentage. Krauthaker is the producer of largest scale of red wine types in Slavonia and the most gifted year was 2016 (can be seen on figure 47). The figure 48 illustrates that on a x-axis lies all 7 producers and the clusters they belong to, while on y-axis is alcohol volume that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific alcohol percentage.

Figure 49 The results of the Slavonian red wine – pH

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     RW
Instances:    26
Attributes:   11
              PH
              WINE NAME
              PRODUCER
Ignored:
              ALCOHOL
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              RESIDUAL SUGAR
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 28.556848958333333

Initial starting points (random):

Cluster 0: 3.51,'Crno 9',Galic
Cluster 1: 3.8,'Pinot crni',Krauthaker
Cluster 2: 3.33,'Rose cuvee',Krauthaker

Missing values globally replaced with mean/mode


Final cluster centroids:
                                        Cluster#
Attribute                 Full Data             0                1                2
                             (26.0)           (8.0)            (6.0)           (12.0)
=============================================================================================
PH                          3.4996            3.55           3.7217            3.355
WINE NAME                Pinot crni Cabernet sauvignon      Pinot crni             Rose
PRODUCER                 Krauthaker          Galic          Jakobovic       Krauthaker



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 31%)
1       6 ( 23%)
2      12 ( 46%)
```

Source: Weka

Figure 50 Visualisation of pH volume of a red wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

To analyse which of 3 Slavonian red wines, by which producer are the most acid or contain lowest pH worth, 3 clusters were chosen as presented on a figure 49. The results are showing that within 26 cluster instances, there are 8 or 31% samples in a cluster 0, 6 or 23% samples in a cluster 1, and 12 or 46% samples in a cluster 2. There are 9 ignored attributes and they are alcohol, density, fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, and vintage year. Number of iterations is 3, while within cluster sum of squared errors is 28.556848958333333.

The highest pH mean of 3.7217 has Pinot crni wine and is located in a cluster number 1 meaning colour is stronger, oxidation quality is stronger, and the total lifespan of Pinot crni is longer compared to other red wines in Slavonia. Krauthaker winery is actually producer of most red wines with the mean pH of 3,4996 in Slavonia based on a full data. However, the information that is not shown here is that Pinot crni wine by Krauthaker is the highest on a pH scale (3,94) meaning it is the least sour wine in Slavonia and healtiest as well, while Rose by Josipovic winery is the lowest on a pH scale (3,14) and thus most acid wine in Slavonia(can be seen on figure 49). Figure 50 representing visualisation of the results. On a x-axis lies all 7 producers and the clusters they belong to, while on y-axis is pH scale that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within acidity level.

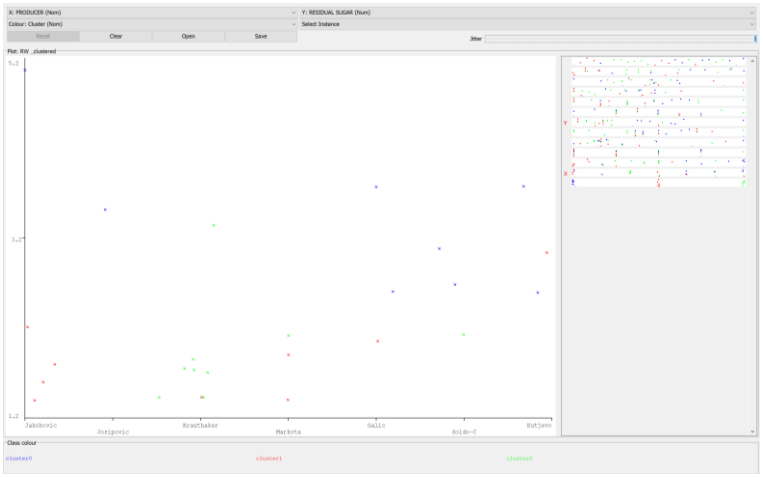
Figure 51 The results of the Slavonian red wine – residual sugar

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     RW
Instances:    26
Attributes:   11
              RESIDUAL SUGAR
              WINE NAME
              PRODUCER
Ignored:
              ALCOHOL
              PH
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 27.66972222222222

Initial starting points (random):

Cluster 0: 2.7,'Crno 9',Galic
Cluster 1: 1.2,'Pinot crni',Krauthaker
Cluster 2: 1.7,'Rose cuvee',Krauthaker

Missing values globally replaced with mean/mode
```

```
Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                          (26.0)       (8.0)      (9.0)      (9.0)
==========================================================================
RESIDUAL SUGAR             2.3308        3.4     1.7778     1.9333
WINE NAME              Pinot crni       Rose Pinot crni Rose cuvee
PRODUCER                Krauthaker      Galic  Jakobovic Krauthaker


Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 31%)
1       9 ( 35%)
2       9 ( 35%)
```

Source: Weka

Figure 52 Visualisation of residual sugar of a red wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Figure 51 represents that 3 clusters were chosen again to analyse which producers and which 3 Slavonian red wines contain the highest amount of residual sugar. The results are showing that within 26 cluster instances, there are 8 or 31% samples in a cluster 0,9 or 35% samples in a cluster 1, and 8 or 9% samples in a cluster 2. There are 8 ignored attributes and they are alcohol, pH, density, fixed acidity, volatile acidity, free sulfur dioxide, total sulfur dioxide, and vintage year. Number of iterations is 3, while within cluster sum of squared errors is 27.66972222222222.

The original data indicates the highest residual sugar has Rose by Jakobović winery (5,2), while the lowest amount has Pinot crni by Krauthaker winery (1,2). It is also confirmed by the full clusters number 0 and 2 that the red wine that contains the highest amount of residual sugar in Slavonia is Rose, while the full data and cluster number 1 confirms Pinot crni by Krauthaker

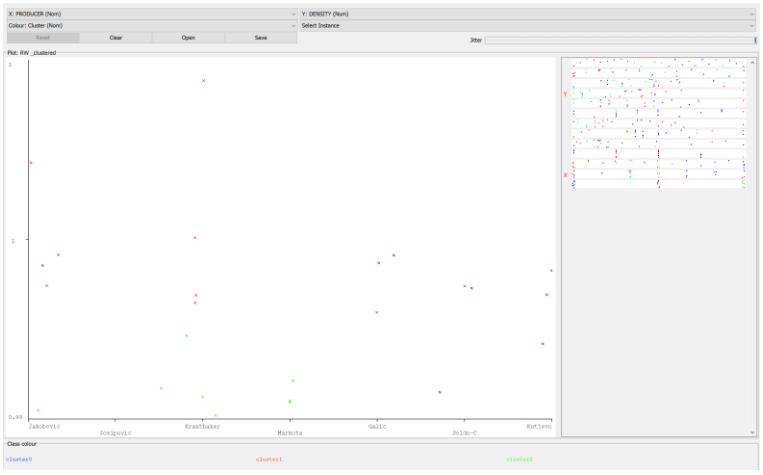
is the one that contains lowest amount (can be seen on figure 51). Figure 52 illustrates that on a x-axis lies all 7 producers and the clusters they belong to, while on y-axis is residual sugar that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific level of sweetness.

Figure 53 The results of the Slavonian red wine – density, alcohol, and residual sugar

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     RW
Instances:    26
Attributes:   11
              ALCOHOL
              DENSITY
              RESIDUAL SUGAR
              WINE NAME
              PRODUCER
Ignored:
              PH
              FIXED ACIDITY
              VOLATILE ACIDITY
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 50.19808608940972

Initial starting points (random):

Cluster 0: '14.10\%',0.9942,2.7,'Crno 9',Galic
Cluster 1: '13.50\%',0.9937,1.2,'Pinot crni',Krauthaker
Cluster 2: '12.60\%',0.9913,1.7,'Rose cuvee',Krauthaker

Missing values globally replaced with mean/mode


Final cluster centroids:
                                  Cluster#
Attribute               Full Data          0          1          2
                           (26.0)      (8.0)     (10.0)      (8.0)
=============================================================================
ALCOHOL                    13.00%     13.00%     13.00%     13.30%
DENSITY                    0.9933     0.9935      0.995      0.991
RESIDUAL SUGAR             2.3308        3.4        2.1       1.55
WINE NAME              Pinot crni       Rose Pinot crni Rose cuvee
PRODUCER                Krauthaker      Galic Krauthaker Krauthaker



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 31%)
1      10 ( 38%)
2       8 ( 31%)
```

Source: Weka

Figure 54 Visualisation of density of a red wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Figure 53 shows that to analyse which of the 3 Slavonian wines are the densest, 3 clusters were chosen. For this cluster analysis residual sugar and alcohol are included since density depends on these attributes. The results are showing that within 26 cluster instances, there are 8 or 31% samples in a cluster 0, 10 or 38% examples in a cluster 1, and 8 or 31% examples in a cluster 2. There are 6 ignored attributes and they are pH, fixed acidity, volatile acidity, free sulfur dioxide, total sulfur dioxide, and vintage year. Number of iterations is 3, while within cluster sum of squared errors is 50.19808608940972.

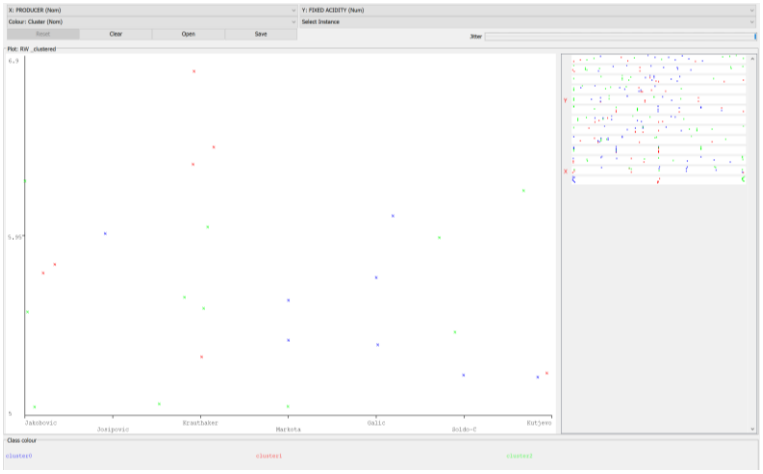
The density mean of 0.991 placed Rose cuvee wine on the first place with the 13.3% of alcohol (the highest compared to other analysed red wines) and 1.55 of the residual sugar (the lowest compared to other analysed red wines). The full data explaining that Krauthaker winery produces most quality red wines that contains the highest alcohol percentage, lowest residual sugar and are the least dense. However, the real data that is not visible here but is original one indicates the highest density has Merlot by Krauthaker winery (1), while the lowest amount of the density has Rose by Soldo-Čamak (0,99). Krauthaker is also confirmed by clusters 1 and full data, while Rose is confirmed by cluster number 0 and 1 (can be seen on figure 53). Figure 54 is representing visualisation of the results. On x-axis lies all 7 producers and the clusters they belong to, while on y-axis is density that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific density level.

Figure 55 The results of the Slavonian red wine – fixed acidity and volatile acidity

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     RW
Instances:    26
Attributes:   11
              FIXED ACIDITY
              VOLATILE ACIDITY
              WINE NAME
              PRODUCER
Ignored:
              ALCOHOL
              PH
              DENSITY
              RESIDUAL SUGAR
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 33.42538538331475

Initial starting points (random):

Cluster 0: 6.1,0.9,'Crno 9',Galic
Cluster 1: 5.2,0.9,'Pinot crni',Krauthaker
Cluster 2: 5.6,0.4,'Rose cuvee',Krauthaker

Missing values globally replaced with mean/mode


Final cluster centroids:
                                        Cluster#
Attribute                 Full Data             0                1                2
                             (26.0)          (8.0)            (7.0)           (11.0)
=============================================================================================
FIXED ACIDITY                5.6731           5.55           5.9714           5.5727
VOLATILE ACIDITY             0.5808            0.7           0.7429           0.3909
WINE NAME                Pinot crni Cabernet sauvignon      Pinot crni             Rose
PRODUCER                 Krauthaker             Galic      Krauthaker       Krauthaker



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       8 ( 31%)
1       7 ( 27%)
2      11 ( 42%)
```

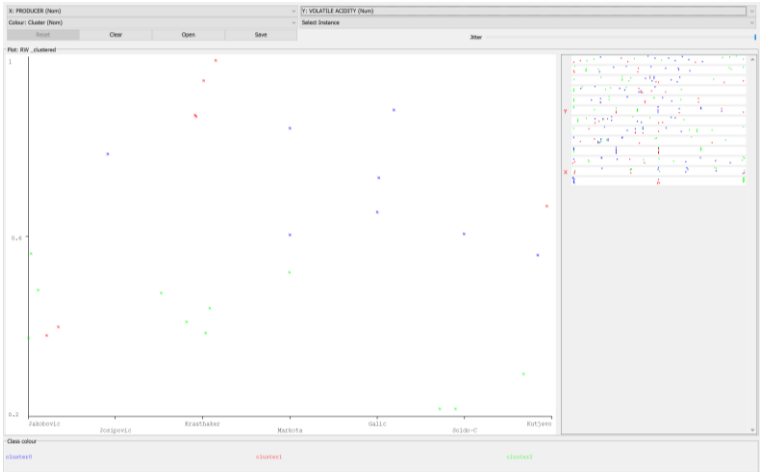
Source: Weka

Figure 56 Visualisation of fixed acidity of red wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Figure 57 Visualisation of volatile acidity of red wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Three clusters were chosen to analyse which of 3 Slavonian wines and by which producer contain highest volume of fixed and volatile acidity, as shown on a figure 55. The results are showing that within 26 cluster instances, there are 8 or 31% samples in a cluster 0, 7 or 27% samples in a cluster 1, and 11 or 42% samples in a cluster 2. There are 7 ignored attributes and they are alcohol, pH, density, residual sugar, free sulfur dioxide, total sulfur dioxide, and vintage year. Number of iterations is 7, while within cluster sum of squared errors is 33.42538538331475.

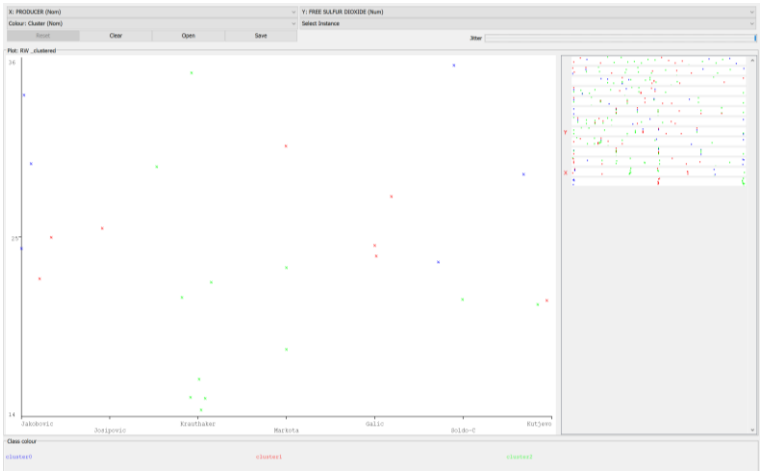
The original data indicates the highest fixed acidity has Merlot by Krauthaker winery (6,9), while the lowest amount have Crvenu cuvee by Krauthaker and Cuvee by Markota winery (5). The highest volatile acidity has Crveni cuvee by Krauthaker winery (1), while the lowest volatile acidity has Rose by Soldo-Čamak winery. It is also confirmed by the full data and clusters number 1 and 2 (can be seen on figure 55) that in Slavonia Krauthaker producing wines that evaporate the longest, while cluster number 2 confirms Rose is the red wine in Slavonia that would least lead to unpleasant taste of vinegar. The results are illustrated above where on x-axis lies all 7 producers and the clusters they belong to, while on y-axis are fixed acidity (figure 56) and volatile acidity (figure 57) that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific of either fixed or volatile acidity level.

Figure 58 The results of the Slavonian red wine – free sulfur dioxide and total sulfur dioxide

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDist
Relation:     RW
Instances:    26
Attributes:   11
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              WINE NAME
              PRODUCER
Ignored:
              ALCOHOL
              PH
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              RESIDUAL SUGAR
              VINTAGE YEAR
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 8
Within cluster sum of squared errors: 29.579879767727633

Initial starting points (random):

Cluster 0: 28,76,'Crno 9',Galic
Cluster 1: 15,64,'Pinot crni',Krauthaker
Cluster 2: 21,70,'Rose cuvee',Krauthaker

Missing values globally replaced with mean/mode


Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                           (26.0)      (6.0)      (8.0)     (12.0)
=============================================================================
FREE SULFUR DIOXIDE       24.0769    29.1667         25    20.9167
TOTAL SULFUR DIOXIDE      90.5385   118.1667      78.75    84.5833
WINE NAME              Pinot crni       Rose Pinot crni Rose cuvee
PRODUCER                Krauthaker  Jakobovic      Galic Krauthaker



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       6 ( 23%)
1       8 ( 31%)
2      12 ( 46%)
```

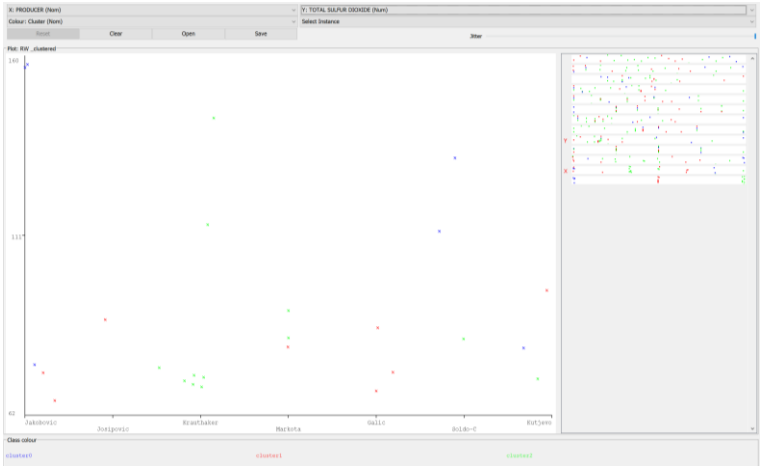
Source: Weka

Figure 59 Visualisation of free sulfur dioxide of a red wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Figure 60 Visualisation of total sulfur dioxide of a red wines by different wine producers in Slavonia



Source: Weka Clusterer Visualize

Figure 57 is presenting that 3 clusters were chosen to analyse which producer and which of 3 Slavonian wines contain the highest volume of free and total sulfur dioxide. The results are showing that within 26 cluster instances, there are 6 or 23% samples in a cluster 0, 8 or 31% samples in a cluster 1, and 12 or 46% samples in a cluster 2. There are 7 ignored attributes and they are alcohol, pH, density, residual sugar, fixed acidity, volatile acidity and vintage year. Number of iterations is 8, while within cluster sum of squared errors is 29.579879767727633. The original data indicates the highest amount of free sulfur dioxide have Rose and Merlot by Soldo-Čamak winery and Krauthaker winery respectively (36), while the lowest amount has Rose cuvee by Krauthaker winery (14). The highest amount of total sulfur dioxide has Rose by Jakobović winery (160), while the lowest amount has G* point by Galić winery (62). On a figure 58, regarding free sulfur dioxide is also confirmed by clusters number 0 and 2 that Rose wine is the red wine in Slavonia that have the most eliminated microbial development and wine oxidation, while full data and clusters number 1 and 2 confirmed that Krauthaker and Galić are the producers of red wines which have the lowest total sulfur dioxide and thus least are intense in taste and aroma because SO_2 is near and below 50ppm (below that is undetectable). The results are illustrated above where on x-axis lies all 7 producers and the clusters they belong to, while on y-axis are free sulfur dioxide (figure 59) and total sulfur dioxide (figure 60) that jittering stronger if a producer holds for bigger scale of produced wines and if most of them are within specific either free or volatile sulfur dioxide level.

Figure 61 The results of the Slavonian red wine – no attribute ignored

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDista
Relation:     RW
Instances:    26
Attributes:   11
              ALCOHOL
              PH
              DENSITY
              FIXED ACIDITY
              VOLATILE ACIDITY
              RESIDUAL SUGAR
              FREE SULFUR DIOXIDE
              TOTAL SULFUR DIOXIDE
              VINTAGE YEAR
              WINE NAME
              PRODUCER
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 59.50169840774666

Initial starting points (random):

Cluster 0: '14.10\%',3.51,0.9942,6.1,0.9,2.7,28,76,2015,'Crno 9',Galic
Cluster 1: '13.50\%',3.8,0.9937,5.2,0.9,1.2,15,64,2017,'Pinot crni',Krauthaker
Cluster 2: '12.60\%',3.33,0.9913,5.6,0.4,1.7,21,70,2017,'Rose cuvee',Krauthaker

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                           (26.0)      (5.0)     (10.0)     (11.0)
============================================================================
ALCOHOL                    13.00%     14.30%     13.00%     13.30%
PH                         3.4996       3.49      3.658       3.36
DENSITY                    0.9933     0.9947     0.9936     0.9924
FIXED ACIDITY              5.6731       5.98       5.39     5.7909
VOLATILE ACIDITY           0.5808        0.8       0.59     0.4727
RESIDUAL SUGAR             2.3308       2.68          2     2.4727
FREE SULFUR DIOXIDE       24.0769         27       23.1    23.6364
TOTAL SULFUR DIOXIDE      90.5385       76.6       86.5   100.5455
VINTAGE YEAR            2016.6538     2015.8     2016.1  2017.5455
WINE NAME              Pinot crni     Merlot Pinot crni       Rose
PRODUCER               Krauthaker      Galic  Jakobovic Krauthaker



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       5 ( 19%)
1      10 ( 38%)
2      11 ( 42%)
```

Source: Weka

The figure 61 illustrates overall results of the Slavonian red wine with no attribute ignored. This confirms one more time that if we decide to change the number of clusters and if we play with the number of ignored attributes, we can get as many combinations and results as possible, and thus can make many conclusions. However, the full data that is most relevant and remain unchanged permanently (whether we decide to change the attributes, number of clusters or seeds), except we make a change in an inserted dataset.

8. GRAPHICAL REPRESENTATIONS

Under this unit will be explained what is happening with the number of iterations and number of cluster sum of squared errors if we change either number of clusters while number of seeds remain unchanged, or the number of clusters with parallel change in the number of seeds. Firstly will be explained what is happening with Slavonian white wine clusters and then with red wine clusters.

8.1. Graphical presentations of the results from the changing number in clusters of white wine from Slavonia

Table 2 Comparison of the number of iterations and squared errors in a clusters of white wines in Slavonia

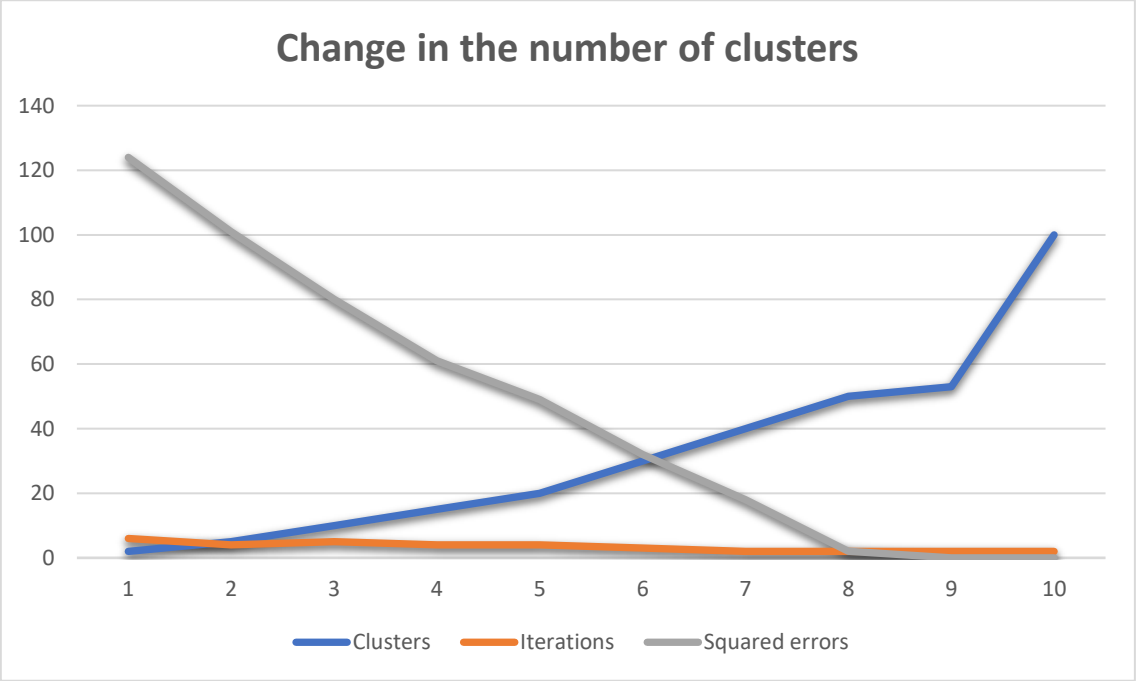
Clusters	Iterations	Squared errors	Clusters	Seeds	Iterations	Squared errors
2	6	124	2	2	4	126
5	4	101	5	5	3	97
10	5	80	10	10	5	80
15	4	61	15	15	5	63
20	4	49	20	20	3	58
30	3	32	30	30	3	31
40	2	18	40	40	3	16
50	2	2	50	50	2	3
53	2	0	53	53	2	0
100	2	0	100	100	2	0

Source: Master thesis author

Table 2 above presents the comparison of the number of iterations and squared errors in the cluster of white wines in Slavonia. Clusters have been chosen on a scale from minimum number of 2 clusters and maximum number of 100 clusters. On the left side of the table (left side from the middle line) is represented what is happening with the number of iterations and sum of squared errors if the number of clusters indeterminately increase, while the number of seeds is set on 10 and remain unchanged endlessly. As can be seen both number of iterations and sum of squared errors decreasing smoothly. However, on the right side of the table (right side from the middle line) is represented what is happening with the number of iterations and sum of squared errors if the number of seeds increase in parallel with the increase in the number of clusters. There is small increase in the number of iterations from 4

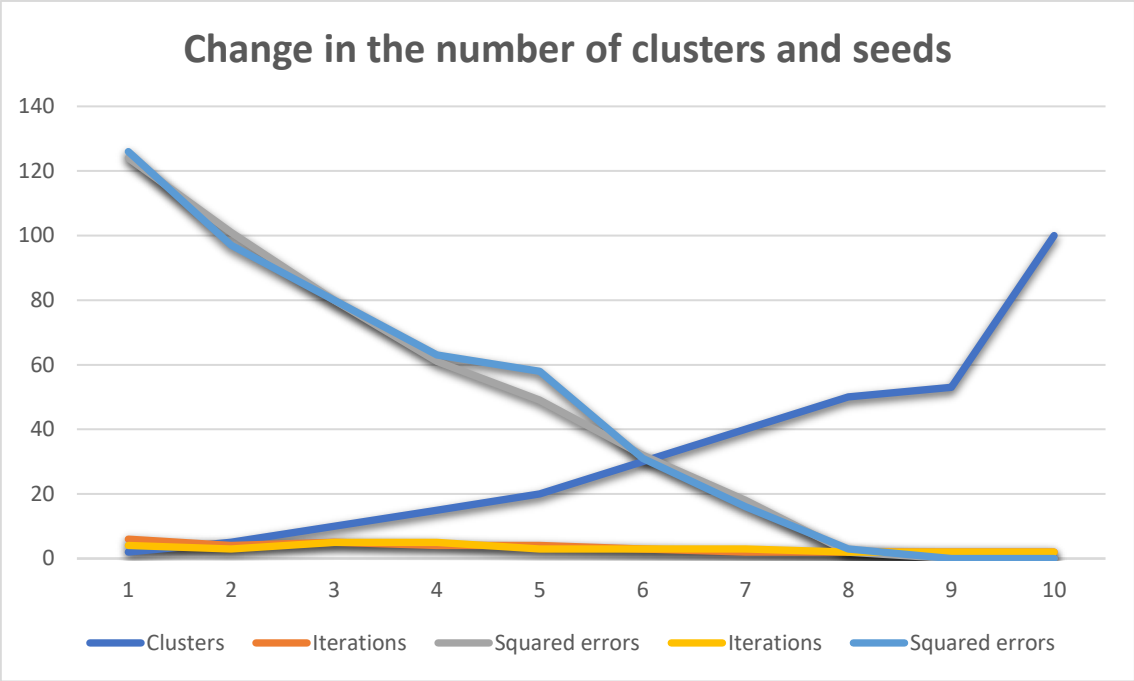
to 5 as the number of clusters increase from 2 to 15, and then decrease from 5 to 2 as the number of clusters increase from 15 to 100. The interesting thing is that no matter for how much number of clusters continue to increase (above the original number of 53 clusters) the number of iterations will continue to be 2 permanently. Moreover, the number of sum squared errors is dropping as well but more sharply compared to sum squared errors from the left side of a table.

Figure 62 Graphical presentation on the change in the number of iterations and sum of squared errors if number of clusters of Slavonian white wine changes



Source: Master thesis author

Figure 63 Graphical presentation on the change in the number of iterations and sum of squared errors if number of clusters and seeds of Slavonian white wine changes



Source: Master thesis author

First graphical presentation illustrates a first change that is set down on a left side of a table, while second graphical presentation illustrates both changes together. Dark Blue line represents number of both clusters, from the first case with constant number of seeds (10) and from the second case where number of seeds increase in parallel with the number of clusters, considering the numbers are the same. Grey line on both graphs represents sum of squared errors from the first case, while light blue line represents sum of squared errors from the other case. Iterations are represented by orange and yellow lines.

8.2. Graphical presentations of the results from the changing number in clusters of red wine from Slavonia

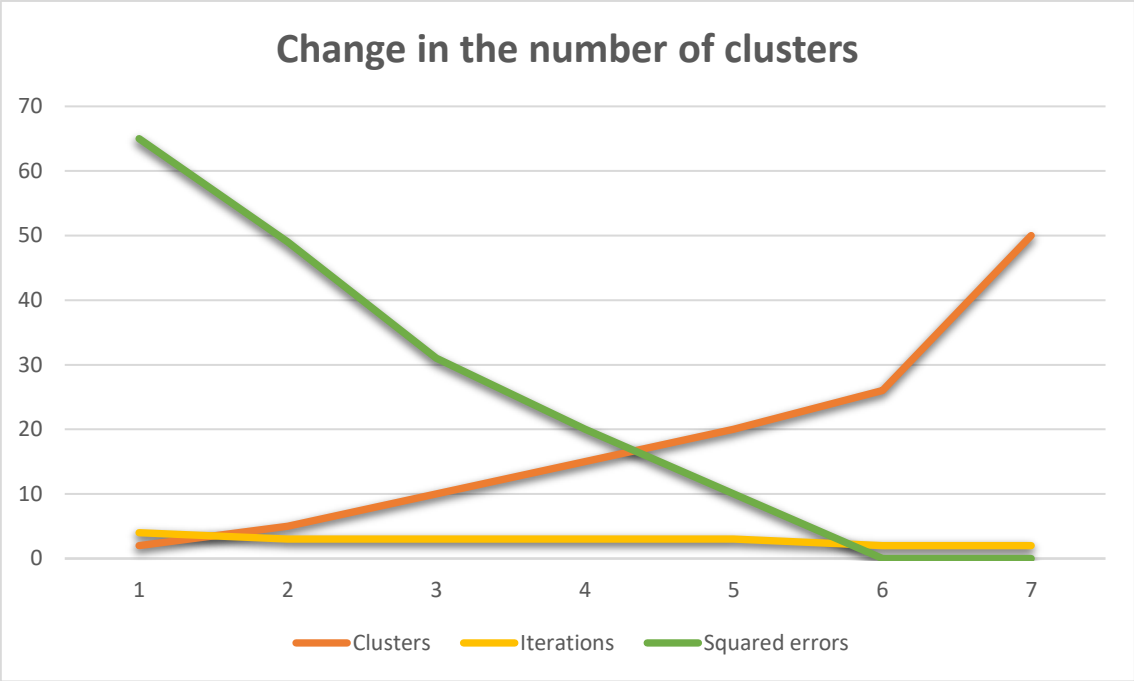
Table 3 Comparison of the number of iterations and squared errors in a clusters of red wines in Slavonia

Clusters	Iterations	Squared errors	Clusters	Seeds	Iterations	Squared errors
2	4	65	2	2	2	64
5	3	49	5	5	4	50
10	3	31	10	10	3	31
15	3	20	15	15	2	19
20	3	10	20	20	3	10
26	2	0	26	26	2	0
50	2	0	50	50	2	0

Source: Master thesis author

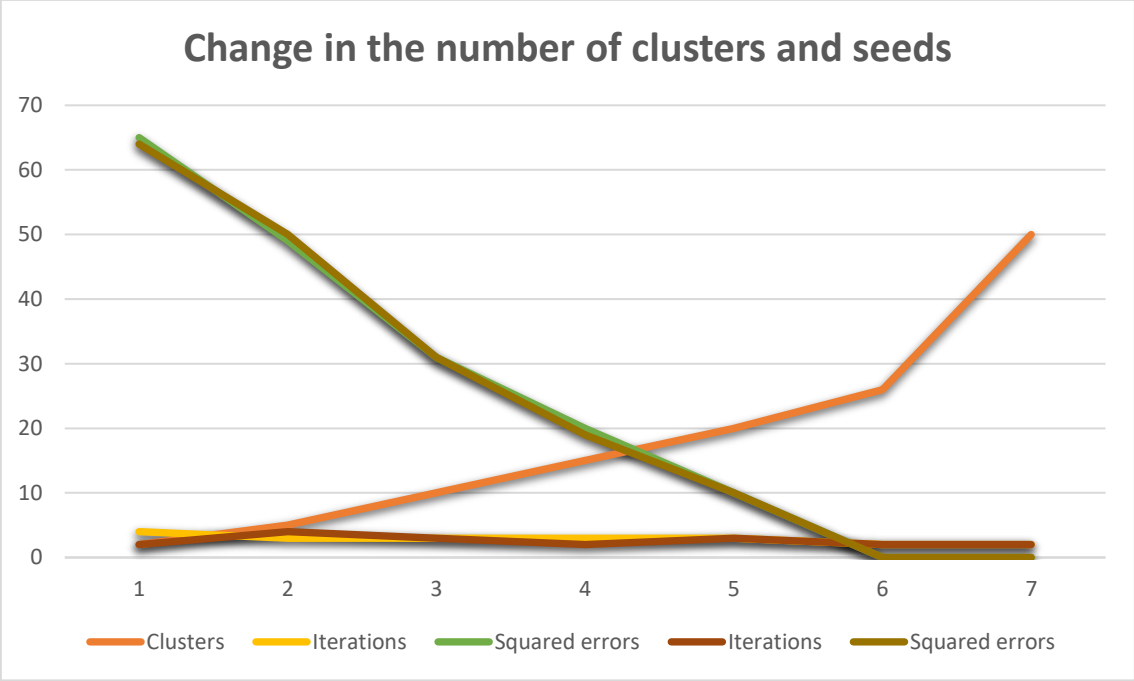
Table number 3 presents the comparison of the number of iterations and squared errors in the cluster of red wines in Slavonia. Clusters have been chosen on a scale from minimum number of 2 clusters and maximum number of 50 clusters which is less then those for white wines because there is smaller number if instances of a red wines. On the left side of the table (left side from the middle line) is represented what is happening with the number of iterations and sum of squared errors if the number of clusters indeterminately increase, while the number of seeds is set on 10 and remain unchanged endlessly. As can be seen both number of iterations and sum of squared errors decreasing smoothly as was the case with white wines. However, on the right side of the table (right side from the middle line) is represented what is happening with the number of iterations and sum of squared errors if the number of seeds increase in parallel with the increase in the number of clusters. There is initial increase in the number of iterations from 2 to 4 as the number of clusters increase from 2 to 5, and then wavy decrease from 4 to 2 as the number of clusters increase from 5 to 50. The interesting thing is that no matter for how much number of clusters continue to increase (above the original number of 26 clusters) the number of iterations will continue to be 2 permanently. Moreover, the number of sum squared errors is dropping as well but less sharply compared to sum squared errors from the left side of a table.

Figure 64 Graphical presentation on the change in the number of iterations and sum of squared errors if number of clusters of Slavonian white wine changes



Source: Master thesis author

Figure 65 Graphical presentation on the change in the number of iterations and sum of squared errors if number of clusters and seeds of Slavonian white wine changes



Source: Master thesis author

Graphical presentation 3 illustrates a first change that is set down on a left side of a table, while graphical presentation 4 illustrates both changes together. Orange line represents number of both clusters, from the first case with constant number of seeds (10) and from the second case where number of seeds increse in parallel with the number of clusters, considering the numbers are the same. Green line on both graphs represents sum of squared errors from the first case, while brown line represents sum of squared errors as well but from the other case. Iterations are represented by red and yellow lines.

To conclude, if we finally compare the results it show that differences between number of iterations and sum of squared errors are smaller for red wine clusters than for white wine clusters.

9. CONCLUSION

Croatia is a country that is rich in vine seedlings and thus the wines themselves. Slavonia is a part of Croatia that is one of the largest and richest in the same. There are many types of wines that are planted in Slavonia, but the most important and widespread are analysed in this master thesis. Wine has many positive properties, but it can also be very endangered (mostly due to diseases such as Gray mold). Moreover, Knowledge discovery in databases and its processes explain to us various ways of working and modelling very large datasets, much larger than this one about wines. The whole series of processes is explained in this master on the example of the topic of the thesis, and the most prominent part is the cluster analysis itself.

Analysing the data, following conclusions were made. There are more sorts of Slavonian white wines than reds. Krauthaker is the producer of the largest scale of white and red wines in Slavonia, while Graševina (white) and Pinot crni (red) are wines with the best means in the overall analysis meaning they are the best quality wines i.e. these are the varieties that are the most cultivated in Slavonia, most processed, most appreciated and enjoyed.

Alcohol percentage mean is something higher in Slavonian red wines by cluster analysis meaning red grapes are sweeter than white grapes in Slavonia i.e. since sulfur should be inserted in the wine to obtain must and then after 24 hours of stillness, yeasts are infused in it to turn sugar into alcohol. However, full data specified that both white and red wines in Slavonia are of the same and thus perfect alcohol percentages. The most gifted year was 2017 for white wines and 2016 for red wines. The pH value is between 3-4 on a pH scale for both white and red wines which indicates they are both acid. However, based on a full data of both analyses it seems red wines are higher on a pH scale meaning they are less acid than whites in Slavonia. Regarding Residual sugar means it seems that white wines in Slavonia are much sweeter than reds. The reason for that are mostly predicate wines known as late harvest, selective harvest of berries, selective harvest of dried berries and ice wines, and all of them must contain more than 19% of alcohol - at least.

Full data also indicates that both white and red wines are of quite same density, but red wines are little bit denser than whites are due to higher amount of alcohol and lower amount of sugar of some Slavonian red wines. White wines contain higher amount of fixed acidity which is the evidence that white wines in Slavonia are stronger or tartaric in taste than reds. Red wines have higher amount of volatile acidity which affirms that red wines in Slavonia contain higher amount of citric acid, but since it is not too high, wines have no cheap, vinegar taste. If compared to red wine, based on a full data white wine in Slavonia is protected more from microbial growth and the oxidation because of a higher free sulfure dioxide content, and is more intense in taste and aroma because of higher total sulfur dioxide content.

For the end, cluster analysis seems to be so individual that there are many ways to analyse the simplest dataset. No matter how much data is entered, all results ends with different changes. So, it seems, even in small changes in clusters or seeds, differences arising and thus changing the picture of the overall observer's conclusions.

Bibliography

1. Ljubljanović S. (1996). *Hrvatski vinski vodić.* Zagreb: vlastita naklada.

2. Benašić Z. (2001). *Što ljubitelji vina žele i vole znati*: *100 odgovora na 101 pitanje.* Đakovo: vlastita naklada.

3. Licul R. and Premužić D. (1977). *Praktično vinogradarstvo i podrumarstvo.* Zagreb: Nakladni zavod Znanje.

4. Adhikari A. (2015). *Advances in Knowlege Discovery in Databases* [online]. Switzerland. Springer International Publishing. Available at: https://books.google.hr/books?id=KLPzBQAAQBAJ&printsec=frontcover&dq=Knowledge+discovery+in+databases&hl=en&sa=X&ved=2ahUKEwjepvCn2srrAhXhkYsKHTFWDE0Q6AEwAnoECAYQAg#v=onepage&q=Knowledge%20discovery%20in%20databases&f=false [02.09.2020.].

5. Pejić-Bach M. (2020). *Presentation on process discovery in KDD.* Zagreb.

6. Maimon Z O. And Rokach L. (2015). *Data Mining With Decision Trees: Theory And Applications* [online]. Singapore. Word Scientific Publishing Co. Pte. Ltd. Available at: https://books.google.hr/books?id=OVYCCwAAQBAJ&printsec=frontcover&dq=decision+tree&hl=en&sa=X&ved=2ahUKEwjb4b7_4crrAhVjwIsKHcSeAzIQ6AEwAHoECAEQAg#v=onepage&q=decision%20tree&f=false [02.09.2020.]

7. Romesburg C. (2004). *Cluster Analysis for Ressearches* [online]. North Carolina. Lulu press. Available at: https://books.google.hr/books?id=ZuIPv7OKm10C&printsec=frontcover&dq=cluster+analysis&hl=en&sa=X&ved=2ahUKEwiJy8Wzh8vrAhUKtYsKHaD-AFIQ6AEwAnoECAAQAg#v=onepage&q=cluster%20analysis&f=false [03.09.2020.]

8. Likas A. (2003). *The global k-means clustering algorithm* [online]. Patter recognition, Volume 36, Pg. 451-461. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0031320302000602 [03.09.2020.]

9. Weka site. Available at: https://www.cs.waikato.ac.nz/ml/weka/ [03.09.2020.]

10. Ian H. Witten, (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* [online]. United States of America: Academic press. Available at: https://books.google.hr/books?id=6lVEKlrTq8EC&pg=PA294&dq=weka&hl=en&sa=X&ved=2ahUKEwiu6qiBv7PrAhXMyaQKHQf7DosQ6AEwB3oECAkQAg#v=onepage&q=types%20of%20attributes&f=false [24.08.2020.]

11. Pejić-Bach, M. (2019) *Data analysis in Weka.* Prezentacija. Zagreb: Ekonomski fakultet

12. Ian H. Witten, (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* [online]. United States of America: Academic press. Available at: https://books.google.hr/books?id=6lVEKlrTq8EC&pg=PA294&dq=weka&hl=en&sa=X&ved=2ahUKEwiu6qiBv7Pr

13. Dua, D. and Graff, C. (2019). *Wine data set* [online]. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml [25.08.2020.]

14. Robertson S. (2019). *Alcaline diet: Pros and cons* [online]. US: News medical life science. Available at: https://www.news-medical.net/health/Alkaline-Diet-Pros-and-Cons.aspx [26.08.2020.]

15. Nierman D. (2004). *Fixed acidity* [online]. Department of Viticulture and Enology, University of California. Davis, CA 95616 USA. Available at: https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity#:~:text=The%20predominant%20fixed%20acids%20found,2%2C000%20mg%2FL%20succinic%20acid [26.08.2020.]

16. Dr. Vinny (2007). *What is distinction between ice wine and late-harvest wine?* [online]. Wine Spectator. Available at: https://www.winespectator.com/articles/whats-the-distinction-between-ice-wine-and-late-harvest-wine-5295#:~:text=Late%2Dharvest%20wines%20are%20made%20from%20grapes%20left%20on%20the,to%20get%20riper%20and%20riper.&text=Ice%20wine%20is%20a%20type,froze%20before%20they%20were%20picked [26.08.2020.]

17. Tanya M. Monro (2012). *Sensing free sulfur dioxide in wine* [online]. Sensors, Basel. US National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3472855/ [26.08.2020.]

18. Grainger K. and Tattersall H. (2016). *Wine production and quality* [online]. 2nd edition. Oxford : Wiley. Available at: https://books.google.hr/books?id=9KjLCgAAQBAJ&pg=PA266&dq=year+of+wine+production&hl=en&sa=X&ved=2ahUKEwjR75Sf37rrAhUJ3qQKHW2bA-EQ6AEwAXoECAAQAg#v=onepage&q=year%20of%20wine%20production&f=false [27.08.2020.]

Table of figures

List of Tables