1. Introduction and Background

Wine is one of the most common alcohol beverage served at a party or opened at a special occasion. Typically wine is made from fermented grapes. Yeast is used to consume the sugars in the grapes and convert it into ethanol, carbon dioxide, and heat. Different variation causes complex interaction between biochemical development of the grapes which results in formation of wine. Wine can also be made from fermenting different crops, rice and fruits such as plum, cherry, pomegranate, etc.

There always comes a time, whenever we are invited to someone’s party or some friend’s home where the host pull a rare bottle of recently purchased wine and ask us if this is any good?

Unfortunately, it is not possible to know by looking at the bottle and reading its content, we get to know whether this a wine worth trying. Only a corkscrew, a glass, and your taste buds can do that.

This isn’t enough to say if a wine is good if. We might like a wine because it might be a personal preference, but determining if a wine’s quality is good depends upon multiple factors to determine its objective quality.

Wine industry has shown a significant growth burst as social drinking is on the rise. The price of wine depends on a rather abstract concept of appreciating wine by wine tasters whose opinion may have a high degree of variability. Pricing of wine may depends on this volatile factor to some extent. Many other factor in wine certification and quality assessment is laboratory-based physicochemical tests which takes into account factors like pH level, acidity, the sugar content and many other chemical properties. Given the wine market, it would be of great interest if human quality of tasting can be related to the chemical properties of wine so that quality assessment and can be more controlled.

2. Project Aims

As such, this project aims to use the data provided by UC Irvine Machine Learning Repository on two dataset related to red and white variants of the Portuguese "Vinho Verde" wine to gain insights into the characteristics of the both red and white wine. Can we understand what makes the attribute such as pH level, acidity, alcohol, etc more or less likely to contribute towards the quality of wine? We consider a number of observations on a set of red and white wine varieties involving their chemical properties and ranking by tasters. We have red wine dataset having 1599 different varieties and white has 4898 varieties. Both the wines are being analysed. All wines are produced in a particular area of Portugal.

Data are collected on 12 different properties of the wines having chemical including density, alcohol content etc. and quality, based on sensory data All chemical properties of wines are continuous variables. Quality is an ordinal variable with a possible ranking from 1 (worst) to 10 (best).

1. Literature Review

Previous studies have been conducted to show different approaches for classification of wine grade. One such example is used in gaussian process regression and multi-task learning which Yeo et al [7] found in his study that advanced machine learning algorithms has a potential for prediction of wine price. Ribeiro et al. [8] in his paper predicted wine vinification using data mining tools that predicted on the best accuracies.

In 1991, a “Wine” data-set was donated to UCI repository which contained 178 instances with 13 unique chemical constituents such as magnesium, alcohol. This data set was used to classify three cultivars from Italy[4]. This was one of the benchmark in classifying data because it was easily discriminated. PCA(Principle Component Analysis) and was reported in [5]. 33 Greek wines are used to classify wine using physicochemical information which were to be used in wine aroma chromatogram and was measured with Fast GC Analyser [6]. Cortez et al. [1] has proposed in his paper about taste predicting technique. A support vector machine, multiple regression and neural network were applied to carry chemical analysis on wines for taste prediction.

Cortez et al. [9] presented models like neural network, multiple regression, and support vector machine to measure the quality and the taste of wines. Shanmuganathan’s [10] in his study predicted the relation between the various temperatures on wine crops and quality.

Chen et al. [10] focused more on improving the taste and quality of wine by using the end users feedback and applied hierarchical clustering of wine which can provide an improved framework for wine classification.

1. Data Retrieval

The data utilized for this analysis was provided by UC Irvine Machine Learning Repository. Data retrieval was easy as the dataset were readily available for user download. It contains to csv files for white wine and red wine. The dataset obtained were:

* winequality-red.csv: white variants of the Portuguese "Vinho Verde" wine.
* [winequality-red.csv](http://localhost:8888/edit/Data_Science/Git/WineDataAnalysis/winequality-red.csv): red variants of of the Portuguese "Vinho Verde" wine.

Prior to any data manipulation, it is essential to extract data and transform it into format that can be easily used in the processing stage. The two dataset used were represented in Panda DataFrame. Below is the same of white and red wine dataset. The rest of the dataset is shown in the appendix.

White wine:

A screenshot of a cell phone

Description automatically generated

Red wine:

A screenshot of a cell phone

Description automatically generated

1. Data Representation:

The main language we have chosen to conduct our analysis is Python for its vast data ecosystem and its powerful libraries and packages that are required for our analytical needs.

The following libraries were used:

* 1. **Numpy** : NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
  2. **Pandas:** Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.
  3. **Matplotlib:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It greatly complements pandas by providing visuals to better explore and understand the data.
  4. **Seaborn:** Seaborn is a Python data visualization library based on [matplotlib](https://matplotlib.org/). It provides a high-level interface for drawing attractive and informative statistical graphics.
  5. **Scikit-learn:** Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy

The Secondary language that we have chosen for our data exploration is R for its stunning visualizations and the ability to build aesthetic graphical plots. We have also used R for data exploration to gather key insights about the data and the variables.

The following libraries were used:

1. **ggplot2:** ggplot2 is a [data visualization](https://en.wikipedia.org/wiki/Data_visualization) package for the [statistical programming](https://en.wikipedia.org/wiki/Computational_statistics) language [R](https://en.wikipedia.org/wiki/R_(programming_language)).  ggplot2 can serve as a replacement for the base graphics in R and contains a number of defaults for web and print display of common scales.
2. **Dplyr:** dplyr is a powerful R-package to transform and summarize tabular data with rows and columns. It provides a set of tools for efficiently manipulating datasets in R
3. DataExplorer: This [R](https://cran.r-project.org/) package aims to automate most of data handling and visualization, so that users could focus on studying the data and extracting insights.
4. Data Cleaning and Pre-processing

Data pre-processing involves the transformation of the raw dataset into an understandable format. Pre-processing data is a fundamental stage in data mining to improve data efficiency since the pre-processing methods directly affect the outcomes of any analytic algorithm.

The Data cleaning procedures that we have used are :

* Removing Duplicates :Merging the two datasets to obtain one combined dataset led to a lot of duplicated entries which, if left unchecked will reduce performance and accuracy. Drop\_duplicates() function was used to drop the duplicates.

A screenshot of a cell phone

Description automatically generated

* Removing Missing Values: There were various observations were missing values were present in the dataset. Missing data can have a significant negative impact in the model performance and accuracy that therefore we decided to remove those observations. Dropna() function was used for the same.

A screenshot of a cell phone

Description automatically generated

* Merging Datasets into one dataframe

Since we obtained datasets from two different sources, they had to be combined into one to be processed and worked upon. Pandas merge function was used to merge the two datasets. Pd.merge()

1. Data Exploration and Visualisation :

Data Exploration is the one of the most important step in any data analytics project as it helps in identifying key insights and trends that would not be detected without exploration. These findings in turn play a major role in optimisations and increasing the accuracy and performance of the model. These techniques also enable us to understand our dataset better to perform and illustrate better visualisations of the data.

R has been used since it has powerful packages to conduct analysis.

The Red Wine dataset has :

Observations : 1,599

Variables : 12

The White Wine dataset has :

Observations : 4,898

Variables : 12

Red Wine Dataset Statistics –

A screen shot of a keyboard

Description automatically generated

Quality is the feature that we are predicting, Frequency distribution of quality in the dataset:

A picture containing orange, black, screen, meter

Description automatically generated

A screenshot of a video game

Description automatically generated

White Wine Statistics –

A close up of a keyboard

Description automatically generated

A screenshot of a computer

Description automatically generated

Frequency distribution of quality in the dataset:

A picture containing drawing

Description automatically generated

A screenshot of a video game

Description automatically generated

1. Feature Selection :

We used feature\_importances\_ function from scikit-learn to evaluate the importance of features on the label. Our observation was that 3 features has the most impact, the rest did not.

However it fascinating to find that the features impact on our label Quality is different for both Red and White Wines.

In the Red Wine dataset :

<insert stats and graph for both>

6.1 Models:

SVM classifier:

Support Vector Machine are supervised learning model that analyse wine data which will be used to classify whether the wine is worth drinking(good or bad). Feeding the model with our dataset and features, SVM builds a model with set of rules which will help us to assign into either good or bad class.

Model Implementation:

* Creating 2 bins from our label quality and assigning 0(bad) or 1(good). A wine is good if the quality is between 2.5 to 6.5 and good if the wine is between 6.5 to highest value(8 for red, 9 for white).
* Split the data into train and testing
* x\_train = features obtain from our data frame.
* y = label(quality) obtain from our dataset.
* Build and run the model.
* Use grid-search-cv to find best hyperparameters.
* Train the model using the hyperparameters obtainerd
* Check the accuracy of the model.

To further increase the accuracy of the model, we used Grid Search cv. Grid-search is used to find the best optimal hyperparameters of a model that will results in the best accurate predictions. Below figure gives us the snippet to use Grid-Search. best\_params\_ method gives us the best value of C, kernel and gamma associated with the model.

A screenshot of a social media post

Description automatically generated

**White wine:**

Our classification reports show us that our model was 87% precise predicting our test value with 85% accuracy.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

**Red Wine:**

Our classification reports show us that our model was 90% precise predicting our test value with 90% accuracy.

**A screenshot of a cell phone

Description automatically generated**

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

Extra data Visualization:

White wine stats against quality.

A screenshot of a cell phone

Description automatically generated

Red wine stats against quality:

A picture containing implement

Description automatically generated

Proportion by wine color quality

A screenshot of a social media post

Description automatically generated

# 4. Observations and Conclusion [¶](http://localhost:8888/notebooks/Data_Science/Git/WineDataAnalysis/Wine%20Case%20Study.ipynb#4.-Observations-and-Conclusion-)

* **How many samples of red wine are there?**
  + A - 1559
* **How many samples of white wine are there?**
  + A - 4898
* **How many columns are in each dataset?**
  + A - 12
* **Which features have missing values?**
  + A - None
* **How many duplicate rows are in the white wine dataset?**
  + A - 937
* **Are duplicate rows in these datasets significant/ need to be dropped?**
  + A - Not necessarily
* **How many unique values of quality are in the red wine dataset?**
  + A - 6
* **How many unique values of quality are in the white wine dataset?**
  + A - 7
* **What is the mean density in the red wine dataset?**
  + A - 0.996747
* **Is a certain type of wine (red or white) associated with higher quality?**
  + A - White
* **What level of acidity (pH value) receives the highest average rating?**
  + A - Low
* **Do wines with higher alcoholic content receive better ratings?**
  + A - High
* **Do sweeter wines (more residual sugar) receive better ratings?**
  + A - Yes
* **What level of acidity receives the highest average rating?**
  + A - Low

Bayes Questions:

Question 1: Describe the steps you followed to produce the knowledge-based causal graph. For example, whose and what knowledge did you use? If knowledge was elicited by multiple members of the group, how did you handle disagreements?

Answer 1:

The steps used to generate knowledge-based causal graph for white and red wine are understanding the chemical properties of wine which included

* fixed acidity
* volatile acidity citric acid
* residual sugar
* chlorides
* free sulfur dioxide
* total sulfur dioxide
* density
* pH
* sulphates
* alcohol.

Most of the chemical properties where continuous so, it was decided that we will convert   
these properties into discrete values of 0,1 and 2. Since wine have many properties we took   
the help of some wine connoisseur and google scholar pages to understand which factor   
might be dependent on each other to include into DAGtrue.csv file. We resolved our   
disagreement by continuous trial and error and checking the precision of our graph.

**White Wine: Red Wine:**



Question 2: In your own words, explain the difference between the Phase\_1 and Phase\_2 graphs. How did the algorithm produce the Phase 2 graph from the Phase\_1 graph?

Answer 2:

Phase 1 visualizes an undirected graph(Extending Maximum Spanning Graph) that is based on MeanMax-Marginal-Discrepancy (MMD) associational scores of the connections is established in DAGtrue.csv. The MMD score represents the discrepancy in marginal probabilities between prior and posterior distributions ranging from 0 to 1. Stronger dependency is indicated by higher score.

Phase 2 graph visualizes the EMSG when the required edges have been oriented and the insignificant connection have been pruned. SaiyanH uses MMD scores to perform constraint based learning through conditional independence tests across all the nodes of phase 2 graph and classifies each triple into conditional dependence, independence or insignificance

Question 3: In your own words, explain the difference between the Phase\_2 and Phase\_3 graphs. How did the algorithm produce the Phase\_3 graph from the Phase\_2 graph? If both graphs are identical, how do you explain this result?

Answer 3:

The Phase 3 graph is the result of score-based learning that is built on the top of Phase 2 graph towards the path that maximizes BIC scoring function. SaiyanH make use of hill-climb method to explore neighbouring graphs where an edges is removed, added or reversed, and move along the direction of increasing BIC score. TABU search is put into use as an attempts to escape possible local maxima after max BIC score has achieved and is repeated till TABU cannot discover a graph that will increase BIC score. Graph is the end product which enables full propagation of evidence by design.

The final phase involves a search method that explores neighbouring graphs and a scoring criterion to evaluate each graph. The output of phase 2 serves as the starting graph for search in phase 3

Question 4: List the number of scores generated in each CSV file. For example, if marginalDep.csv has 100 rows of scores, then you should write ‘100’ for that particular file. Discuss the different quantities in scores generated in each file. For example, why do you think there are 100 scores in the file marginalDep.csv and 200 scores in the file conditionalInsignificance.csv?

Answer 4:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| White wine | | | | |
| CSV | conditionalDep | conditionalInsignificance | marginalDep | conditionalIndep |
| Score | 15 | 478 | 55 | 2 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Red wine | | | | |
| CSV | conditionalDep | conditionalInsignificance | marginalDep | conditionalIndep |
| Score | 71 | 423 | 55 | 1 |

For both red and wine data set, we observe that there are more number of scores for conditionalInsignificance.csv than marginalDep.csv. This result is obtained is because of the conditional independence tests on sets of triples (in Phase2) and not­ on pair of nodes for Phase1. Hence, there are a greater combinations of possible variables. More number of variables are classified as conditionally insignificant because we follow the precise classification rules of dependence and independences.

More conditional dependent variable than conditionally independent variables is also seen for both the dataset as we only choose variable that we suspected to have casual connection in our knowledge based graph.

Question 5: Refer to your F1, SHD and BSF scores and compare them to the related scores shown in Fig 2 of the related research paper. Are your scores mostly lower, on par, or higher (in general) compared to those shown in Fig 2 with respect to SaiyanH (ignore results from other algorithms)? Indicate whether this result is in agreement or not with your initial expectations, and explain why.

Answer 5:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **White Wine** | **Red Wine** | **Average case study** |
| **F1** | 0.577 | 0.39 | 0.55 |
| **SHD** | 28.500 | 46 | 69 |
| **BSF** | 0.418 | 0.66 | 0.46 |

Both wines performed better in SHD-Score. This is an indication that it requires fewer edge insertions, arc reversal and deletion to convert learned graph into true graph. We did not expected BSF to be higher since we had 11 chemical properties and to include different arcs was challenging. Higher BSF score suggest we made a better accurate graph.

Question 6: Refer to your elapse time of structure learning and compare it to the runtime shown in Table 2 of the related research paper. Indicate whether your result is consistent or not with the results shown in Table 2, and explain why.

Answer 6: We have observed structural learning of 1 second for both red and white wine which is consistent with result shown in table 2. This is because we have similar number of sample size, nodes and parameter.

A screenshot of a cell phone

Description automatically generated

Question 7: Compare the BIC/MDL score generated at Step 4 with the BIC/MDL score generated at Step 3. What do you understand from the difference in those two scores? Indicate whether this result is in agreement with your initial expectations, and explain why.

Answer 7:

**White Wine:**

|  |  |
| --- | --- |
| Step 3 | Step 4 |
| BIC/MDL score  -109539.972 | BIC/MDL score  -67907.733 |

**Red Wine:**

|  |  |
| --- | --- |
| Step 3 | Step 4 |
| BIC/MDL score  -152523.756 | BIC/MDL score  -20630.979 |

BIC criteria is used to select best from a group of finite set of models. BIC/MDL score higher for both the wines at step 4 in comparison with step 3. Step 3 represents the casual model (knowledge-based graph) made by us with intuition and little knowledge before any constraint-based learning. Step 4 is the model made after Hill Climbing has completed and edges have been modified to maximize the BIC score. Hence Step 4 had better score and was in agreement of expectation.

Question 8: Compare the # of free parameters generated at Step 4 with the # of free parameters generated at Step 3. What do you understand from the difference between these two values? Indicate whether this result is in agreement with your initial expectations, and explain why.

Answer 8:

**White Wine:**

|  |  |
| --- | --- |
| Step 3 | Step 4 |
| # of free parameters  10314 | # of free parameters  302 |

**Red Wine:**

|  |  |
| --- | --- |
| Step 3 | Step 4 |
| # of free parameters  27270 | # of free parameters  302 |

Variables which cannot be predicted precisely or constrained by the model and has to be estimated are represented by free parameters. A lower number is expected in Step 3 because our model is unpruned and has not been optimized using BIC criteria, whereas Step 4 the model has been pruned and optimized, thus requires less number of parameters.

References:

[1] P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, “Modelling wine preferences by data mining from

physicochemical properties,” In Decision Support Systems,Elsevier, 47 (4): 547-553. ISSN: 0167-9236.

[4]A. Asuncion, and D. Newman (2007), UCI MachineLearning Repository, University of California, Irvine,[Online].Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[5]S. Kallithraka, IS. Arvanitoyannis, P. Kefalas, A. El-Zajouli, E. Soufleros, and E. Psarra, “Instrumental and sensory analysis of Greek wines; implementation ofprincipal component analysis (PCA) for classification according to geographical origin,” Food Chemistry, 73(4):501-514, 2001.

[6]N. H. Beltran, M. A. Duarte - MErmound, V. A. S. Vicencio, S. A. Salah, and M. A. Bustos, “Chilean wine classification using volatile organic compounds dataobtained with a fast GC analyzer,”

Instrum. Measurement, IEEE Trans., 57: 2421-2436, 2008.

[7] M. Yeo, T. Fletcher, and J. Shawe-Taylor, “Machine Learning in Fine Wine Price Prediction,” Journal of Wine Economics, Vol. 10, No. 2, pp.151-172, 2015.

[8] J. Ribeiro, J. Neves, J. Sanchez, M. Delgado, J. Machado, and P. Novais, “Wine Vinification Prediction using Data Mining Tools.,” in Proc. Computing and Computational Intelligence, Tbilisi, the Republic of Georgia, pp. 78-85. June 2009.

[9] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, “Modeling Wine Preferences by Data Mining from Physicochemical Properties,” Department of Information Systems/R&D Centre Algoritmi, University of Minho, 4800-058 Guimarães, Portugal, Viticulture Commission of the Vinho Verde Region (CVRVV), 4050- 501 Porto, Portugal.

[10] S. Shanmuganathan, P. Sallis, and A. Narayanan, “Data Mining Techniques for Modeling Seasonal Climate Effects on Grapevine Yield and Wine Quality,” in Proc. IEEE International Conference on Computational Intelligence Communication Systems and Networks, pp. 82-89, July 2010.

[11] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, “Wineinformatics: Applying Data Mining on Sensory Wine Reviews Processed by the Computational Wine Wheel,” in Proc. IEEE International Conference on Data Mining Workshop, pp. 142-149, Dec. 2014.