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

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 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.

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>

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.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Inference-based evaluation \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

BIC/MDL score -152523.756

# of free parameters 27270

#############################################################

#############################################################

#############################################################

run:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Training data info \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Variables: 11

Sample size: 1449

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Knowledge-based constraints \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Temporal constraints specified: 0

Directed constraints specified: 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Structure learning \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Running SaiyanH with settings:

a) Associational score: MeanMax [Absolute]

b) Conditional independence pruning: true

c) Faithfulness condition pruning: true

d) TABU search max escape attempts: V(V-1)

Entering Phase 1 [EMST graph]...

marginalDep.csv saved.

Phase 1 completed.

Entering Phase 2 [constraint-based learning]...

conditionalDep.csv saved.

conditionalIndep.csv saved.

conditionalInsignificance.csv saved.

Phase 2 completed.

Entering Phase 3 [score-based learning]...

fixedacidity->volatileacidity[label="0.078"];fixedacidity->residualsugar[label="0.072"];fixedacidity->chlorides[label="0.078"];fixedacidity->density[label="0.217"];fixedacidity->pH[label="0.234"];fixedacidity->sulphates[label="0.074"];fixedacidity->alcohol[label="0.073"];volatileacidity->sulphates[label="0.112"];volatileacidity->alcohol[label="0.072"];citricacid->fixedacidity[label="0.230"];citricacid->volatileacidity[label="0.144"];citricacid->residualsugar[label="0.055"];citricacid->pH[label="0.161"];citricacid->sulphates[label="0.103"];citricacid->alcohol[label="0.072"];freesulfurdioxide->citricacid[label="0.077"];freesulfurdioxide->totalsulfurdioxide[label="0.275"];totalsulfurdioxide->citricacid[label="0.080"];density->residualsugar[label="0.137"];density->chlorides[label="0.140"];pH->chlorides[label="0.086"];alcohol->density[label="0.170"];graph[fontname=Arial, fontsize = 10, label="SaiyanH\_Phase\_3 graph (final). \lTotal arcs: 22 \l"]

Phase 3 completed.

Arcs randomised during phase 2 constraint-based learning: 0

Structure learning elapsed time: 0 seconds total (Phase 1 = 0 secs, Phase 2 = 0 secs).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Evaluation \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Nodes: 11

Sample size: 1449

TrueDAG arcs: 60

TrueDAG independencies: -5

LearnedDAG arcs: 22

LearnedDAG independencies: 33

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Confusion matrix stats \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Arcs discovered (TP): 12.0

Partial arcs discovered (TP\*0.5): 8.0

False dependencies discovered (FP): 2.0

Independencies discovered (TN): -7.0

Dependencies not discovered (FN): 44.0. [NOTE: # of edges missed is 40.0]

\_\_\_\_\_\_\_\_\_\_\_\_\_ Stats from metrics and scoring functions \_\_\_\_\_\_\_\_\_\_\_\_\_\_

Precision score: 0.727

Recall score: 0.267

F1 score: 0.390

SHD score: 46.000

DDM score: -0.500

BSF score: 0.667

# of independent graphical fragments: 1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Inference-based evaluation \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

BIC/MDL score -20630.979

# of free parameters 302

BUILD SUCCESSFUL (total time: 32 seconds)