# ABSTRACT

The wine industry shows a recent exponential growth as social drinking is on the rise. Nowadays, industry players are using product quality certifications to promote their products.

This is a time-consuming process requires the assessment given by human experts, which makes this process very expensive. Also, the price of red wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability.

Another vital factor in red wine certification and quality assessment is physicochemical tests, which are laboratory- based and consider factors like acidity, pH level, sugar, and other chemical properties.

The red wine market would be of interest if the human quality of tasting can be related to wine’s chemical properties so that certification and quality assessment and assurance processes are more controlled. This project aims to determine which features are the best quality red wine indicators and generate insights into each of these factors to our model’s red wine quality.

Wine classification is a difficult task since taste is the least understood of the human senses. A good wine quality prediction can be very useful in the certification phase, since currently the sensory analysis is performed by human tasters, being clearly a subjective approach. An automatic predictive system can be integrated into a decision support system, helping the speed and quality of the performance. Furthermore, a feature selection process can help to analyze the impact of the analytical tests. If it is concluded that several input variables are highly relevant to predict the wine quality, since in the production process some variables can be controlled, this information can be used to improve the wine quality.

# Chapter 1 INTRODUCTION

Wine is a beverage made from fermented grape and other fruit juices with lower amount of alcohol content. Quality of wine is rated on the basis of taste of wine and vintage. This process is time consuming, costly and not so efficient. The aim of this project is to predict the quality of wine on a scale of 1–10 given a set of features as inputs. The dataset used is Wine Quality Data set from UCI Machine Learning Repository. Input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free Sulphur dioxide, total Sulphur dioxide, density, pH, sulphates, alcohol. And the output variable is quality (score between 0 and 10).

We are dealing only with red wine. The higher the value the better the quality. In this project we will treat each class of the wine separately and their aim is to be able and find decision boundaries that work well for new unseen data. In industries, understanding the demands of wine safety testing can be a complex task for the laboratory with numerous analytes and residues to monitor. But, our application’s prediction, provide ideal solution for the solution of wine, which will make this whole process efficient and cheaper with less human involvement.

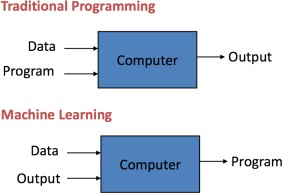
## Introduction to Technology

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

MCA Department, Surana College Page 2

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes.



## Methods of Machine Learning

* + - * **Supervised Machine Learning**

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to “learn” by comparing its actual output with the “taught” outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data.

## Unsupervised Machine Learning

In unsupervised learning, data is unlabelled, so the learning algorithm is left to find commonalities among its input data. As unlabelled data are more abundant than labelled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

## Current Applications of Machine Learning

* + - * **Web Search Engine:** One of the reasons why search engines like google, Bing etc work so well is because the system has learnt how to rank pages through a complex learning algorithm.
      * **Photo tagging Applications:** Be it Facebook or any other photo tagging application, the ability to tag friends makes it even more happening. It is all possible because of a face recognition algorithm that runs behind the application.
      * **Spam Detector:** Our mail agent like Gmail or Hotmail does a lot of hard work for us in classifying the mails and moving the spam mails to spam folder. This is again achieved by a spam classifier running in the back end of mail application.
      * **Database Mining for growth of automation:** Typical applications include Web-click data for better UX (User experience), Medical records for better automation in healthcare, biological data and many more.
      * **Applications that cannot be programmed:** There are some tasks that cannot be programmed as the computers we use are not modelled that way. Examples include Autonomous Driving, Recognition tasks from unordered data (Face Recognition/ Handwriting Recognition), Natural language Processing, computer Vision etc.
      * **Understanding Human Learning:** This is the closest we have understood and mimicked the human brain. It is the start of a new revolution, The real AI. Now, after a brief insight lets come to a more formal definition of Machine Learning

## Why Machine Learning Now

* + - * **Data availability**: Just over 3 billion people are online with an estimated 17 billion connected devices or sensors. [[9]](https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/?gclid=Cj0KCQiAst2BBhDJARIsAGo2ldUX-xWGcOfYDj3owUXkJiGr8af8I265Si0fL-GUC7OpX7Ujq65r4vgaAufCEALw_wcB&_ftn9) That generates a large amount of data which, combined with decreasing costs of data storage, is easily available for use. Machine learning can use this as training data for learning algorithms, developing new rules to perform increasingly complex tasks.
      * **Computing power**: Powerful computers and the ability to connect remote processing power through the Internet make it possible for machine-learning techniques that process enormous amounts of data. [[10]](https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/?gclid=Cj0KCQiAst2BBhDJARIsAGo2ldUX-xWGcOfYDj3owUXkJiGr8af8I265Si0fL-GUC7OpX7Ujq65r4vgaAufCEALw_wcB&_ftn10)
      * **Algorithmic innovation**: New machine learning techniques, specifically in layered neural networks – also known as “deep learning” – have inspired new services, but is also spurring investments and research in other parts of the field. [[11]](https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/?gclid=Cj0KCQiAst2BBhDJARIsAGo2ldUX-xWGcOfYDj3owUXkJiGr8af8I265Si0fL-GUC7OpX7Ujq65r4vgaAufCEALw_wcB&_ftn11)

## Key Elements of Machine Learning

Every machine learning algorithm has three components:

* + - * **Representation**: how to represent knowledge. Examples include decision trees, sets of rules, instances, graphical models, neural networks, support vector machines, model ensembles and others.
      * **Evaluation**: the way to evaluate candidate programs (hypotheses). Examples include accuracy, prediction and recall, squared error, likelihood, posterior probability, cost, margin, entropy k-L divergence and others.
      * **Optimization**: the way candidate programs are generated known as the search process. For example, combinatorial optimization, convex optimization, constrained optimization.

## Machine Learning in Practice

Machine learning algorithms are only a very small part of using machine learning in practice as a data analyst or data scientist. In practice, the process often looks like:

* + - 1. Start Loop
         1. **Understand the domain, prior knowledge and goals**. Talk to domain experts. Often the goals are very unclear. You often have more things to try then you can possibly implement.
         2. **Data integration, selection, cleaning and pre-processing**. This is often the most time consuming part. It is important to have high quality data. The more data you have, the more it sucks because the data is dirty. Garbage in, garbage out.
         3. **Learning models**. The fun part. This part is very mature. The tools are general.
         4. **Interpreting results**. Sometimes it does not matter how the model works as long it delivers results. Other domains require that the model is understandable. You will be challenged by human experts.
         5. **Consolidating and deploying discovered knowledge**. The majority of projects that are successful in the lab are not used in practice. It is very hard to get something used.
      2. End Loop

## When Should You Use Inductive Machine Learning

There are problems where inductive learning is not a good idea. It is important when to use and when not to use supervised machine learning.

4 problems where inductive learning might be a good idea:

* **Problems where there is no human expert**. If people do not know the answer, they cannot write a program to solve it. These are areas of true discovery.
* **Humans can perform the task but no one can describe how to do it**. There are problems where humans can do things that computer cannot do or do well. Examples include riding a bike or driving a car.
* **Problems where the desired function changes frequently**. Humans could describe it and they could write a program to do it, but the problem changes too often. It is not cost effective. Examples include the stock market.
* **Problems where each user needs a custom function**. It is not cost effective to write a custom program for each user. Example is recommendations of movies or books on Netflix or Amazon.

## Objective:

* Our main objective is to predict the wine quality using Machine Learning through Python programming language.
* A large dataset is considered and wine quality is modelled to analyse the quality of wine through different parameters like fixed acidity, volatile acidity and many more.
* All these parameters will be taken into considerations through Machine Learning Algorithms like random forest classifier algorithm which will help to rate the wine on scale of 1-10 or good-bad.
* Output obtained would further be checked for correctness and model will be optimized accordingly.
* It can support the wine expert evaluation and ultimately improve the productivity.

## Methodology

For making automated decisions on model selection, we need to quantify the performance of our model and give it a score. For that reason, for the classifiers, we are using F1 score which combines two metrics: Precision which expresses how accurate the model was on predicting a certain class and Recall which expresses the inverse of the regret of missing out instances which are misclassified. Since we have multiple classes, we have multiple F1 scores. We will be using the unweighted mean of the F1 scores for our final scoring. This is a business decision because we want our models to get optimized to classify instances that belong to the minority side, such as wine quality of 3 or 8 equally well with the rest of the qualities that are represented in a larger number. For the regression task we are scoring based on the coefficient of determination, which is basically a measurement of whether the predictions and the actual values are highly correlated. The larger this coefficient the better. For regressors we can also get F1 score if we first round our prediction.

**Splitting for Testing:**

We are keeping 20% of our dataset to treat it as unseen data and be able and test the performance of our models. We are splitting our dataset in a way such that all of the wine qualities are represented proportionally equally in both training and testing dataset. Other than that, the selection is being done randomly with uniform distribution.

**Preprocessing:**

Label Encoding is used to convert the labels into numeric form so as to convert it into the machine-readable form. It is an important pre-processing step for the structured dataset in supervised learning. We have used label encoding to label the quality of data as good or bad. Assigning 1 to good and 0 to bad.

## Limitations and Discussion

The random forest models built in this study performs the best, however, it is still far from perfect. To further improve my model performance, I could further try the following steps.

1. To have a closer look at misclassified observations from the confusion matrix (in the prediction case, check observations with large differences between predicted quality and real quality), and try to understand why these wines are classified wrongly.
2. Learn some wine quality control knowledge to have a better sense of the prior knowledge in the wine producing business, so that I might be able to properly transform some features or interpret the interaction between features
3. Spend more time to tune model hyper-parameters, and ask colleagues and my manager for further suggestions.
4. If computing power is a limit, I could ask my manager if there is a way to gain more computing power, by either using on-demand cloud computing, or upgrade my computer, or using a computer cluster.
5. Communicate with my clients to know more clearly about their goal: whether to predicting wine quality or to pick up wines with superior quality. Also, try to understand whether they have more tolerance for the type I error or the type II error, so that I can properly set the model cutoff value.
6. Ask clients if they have more data available, especially for the underrepresented classes, such as wines with high or low quality. Using more data, I could build a more complicated model, such as a neural network with more layers.
7. Ask clients if they have other related features. The model could be improved by adding these features in.

# Chapter 2 LETRATURE SURVEY

## Initial Investigation

System Analysis and is an approach towards formalization the analysis and of information system with the objective of improving the system performance by automation.

In the process of planning a new system or to replace or improve the old system, the thorough analysis of the existing system has to be done. So, the system plans a very important role in the process of system development.

## Study Of The System

Wine classification is a difficult task since taste is the least understood of the human senses. A good wine quality prediction can be very useful in the certification phase, since currently the sensory analysis is performed by human tasters, being clearly a subjective approach.

An automatic predictive system can be integrated into a decision support system, helping the speed and quality of the performance. Furthermore, a feature selection process can help to analyze the impact of the analytical tests. If it is concluded that several input variables are highly relevant to predict the wine quality, since in the production process some variables can be controlled, this information can be used to improve the wine quality.

.

## Dataset Description:

The two datasets are related to red wine of the Portuguese "Vino Verde" wine. For more details, consult: [Web Link] or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

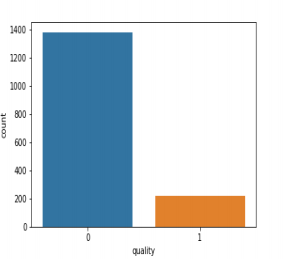
1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide 7)total sulfur dioxide 8)density

9)pH

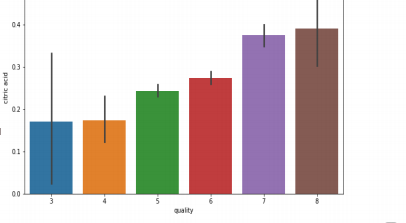
1. sulphates
2. alcohol Output variable (based on sensory data) 12)quality (score between 0 and 10)

## Data Analysis

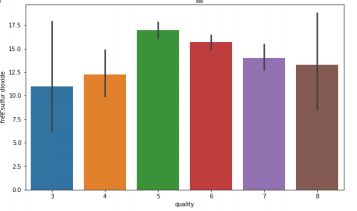
* The dataset contains chemical description of 6499 Portuguese “vino Verde” wines.
* There are 4899 entries for white wine, and 1600 entries for red wine.
* The source of the data is taken from the UCI Machine Learning Repository, provided by Paulo Cortez, from the University of Minho, Portugal.
* We have used several Frameworks like Matplotlib, Scikit-learn, NumPy, Pandas and Seaborn.
* The below bar plot shows the count of data which is good or bad. We can see 80% of the data is classified with good wine quality and 20% with bad quality of wine.



* This bar plot shows a directly proportional relation between citric acid and quality. As the quality of wine increases the amount of citric acid also increases which shows that citric acid is the important feature on which quality of wine depends.



* Free Sulphur dioxide is greatly contributing to the quality of wine, this bar plot gives us a more clear picture.



## Hardware and Software Requirements:

**Hardware Requirements:**

|  |  |
| --- | --- |
| **Hardware Tools** | **Minimum Requirements** |
| Processor | i5 or above |
| Hard Disk | 8GB |
| Monitor | 15” Coloured |
| Mouse | Optical |
| Keyboard | 122 Keys |

**Software Requirements:**

|  |  |
| --- | --- |
| **Software Tools** | **Minimum Requirements** |
| Platform | Windows, Linux or MacOS |
| Operating System | Windows, Linux or MacOS |
| Technology | Windows, Linux or MacOS |
| Scripting Language | Python |
| IDE | PyCharm or Jupiter |

* 1. **Coding**

import numpy as np import pandas as pd import sklearn import seaborn as sns

import matplotlib.pyplot as plt from collections import Counter

from sklearn.preprocessing import StandardScaler from sklearn.decomposition import PCA

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import confusion\_matrix,accuracy\_score from sklearn.linear\_model import LogisticRegression

* + - **Chlorides**

data=pd.read\_csv("winequality-red.csv")

print(data.corr()) print(data.columns) print(data.info()) print(data["quality"].unique()) Counter(data['quality'])

#sns.countplot(x='quality', data=data) sns.boxplot(x='quality',y='chlorides',data=data) plt.show()

* + - **Citric Acid**

data=pd.read\_csv("winequality-red.csv")

print(data.corr()) print(data.columns) print(data.info()) print(data["quality"].unique())

Counter(data['quality']) #sns.countplot(x='quality', data=data) sns.boxplot(x='quality',y='citric acid',data=data) plt.show()

* + - **Density**

data=pd.read\_csv("winequality-red.csv")

print(data.corr())

print(data.columns) print(data.info()) print(data["quality"].unique()) Counter(data['quality'])

#sns.countplot(x='quality', data=data) sns.boxplot(x='quality',y='density',data=data) plt.show()

* + - **Fixed acidity**

data=pd.read\_csv("winequality-red.csv") print(data.corr())

print(data.columns) print(data.info()) print(data["quality"].unique()) Counter(data['quality'])

#sns.countplot(x='quality', data=data) sns.boxplot(x='quality',y='fixed acidity',data=data) plt.show()

* + - **Free-Sulphur-dioxide**

data=pd.read\_csv("winequality-red.csv") print(data.corr())

print(data.columns) print(data.info()) print(data["quality"].unique())

Counter(data['quality']) #sns.countplot(x='quality', data=data)

sns.boxplot(x='quality',y='free sulfur dioxide',data=data) plt.show()

* + - **PH**

data=pd.read\_csv("winequality-red.csv") print(data.corr())

print(data.columns) print(data.info()) print(data["quality"].unique()) Counter(data['quality'])

#sns.countplot(x='quality', data=data) sns.boxplot(x='quality',y='pH',data=data) plt.show()

* **Residual Sugar**

data=pd.read\_csv("winequality-red.csv")

print(data.corr()) print(data.columns) print(data.info()) print(data["quality"].uni que()) Counter(data['quality'])

#sns.countplot(x='quality', data=data) sns.boxplot(x='quality',y='residual sugar',data=data) plt.show()

* **Sulphates**

data=pd.read\_csv("winequality-red.csv")

print(data.corr()) print(data.columns) print(data.info()) print(data["quality"].unique()) Counter(data['quality'])

#sns.countplot(x='quality', data=data) sns.boxplot(x='quality',y='sulphates',data=data) plt.show()

* **Total Sulphur dioxide**

data=pd.read\_csv("winequality-red.csv")

print(data.corr()) print(data.columns) print(data.info()) print(data["quality"].unique()) Counter(data['quality'])

#sns.countplot(x='quality', data=data) sns.boxplot(x='quality',y='total sulfur dioxide',data=data) plt.show()

* **Volatile Acidity**

data=pd.read\_csv("winequality-red.csv")

print(data.corr()) print(data.columns) print(data.info()) print(data["quality"].unique()) Counter(data['quality'])

#sns.countplot(x='quality', data=data)

sns.boxplot(x='quality',y='volatile acidity',data=data) plt.show()

* **Final project**

import os

data=pd.read\_csv("winequality-red.csv")

print(data.corr()) print(data.columns) print(data.info()) print(data["quality"].uni que())

sns.pairplot( data) plt.show()

print(Counter(data['quality']))

sns.countplot(x='quality', data=data) plt.show()

data.describe()

review=[]

for i in data["quality"]: if i<=3 and i>=1:

review.append('1') elif i>=4 and i<=7:

review.append('2') elif i>=8 and i<=10: review.append('3')

data["Reviews"]=review print(data.columns) data["Reviews"].unique() print(Counter(data['Reviews'])) x = data.iloc[:,:11]

y = data['Reviews'] print(x.head(10)) print(y.head(10))

sc = StandardScaler() x=sc.fit\_transform(x) print(x)

pca = PCA()

x\_pca = pca.fit\_transform(x) plt.figure(figsize=(10,10))

plt.plot(np.cumsum(pca.explained\_variance\_ratio\_), 'ro-') plt.grid()

plt.show()

pca\_new = PCA(n\_components=8) x\_new = pca\_new.fit\_transform(x) #print(x\_new)

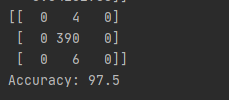
x\_train, x\_test, y\_train, y\_test = train\_test\_split(x\_new, y, test\_size = 0.25) #print(x\_train.shape)

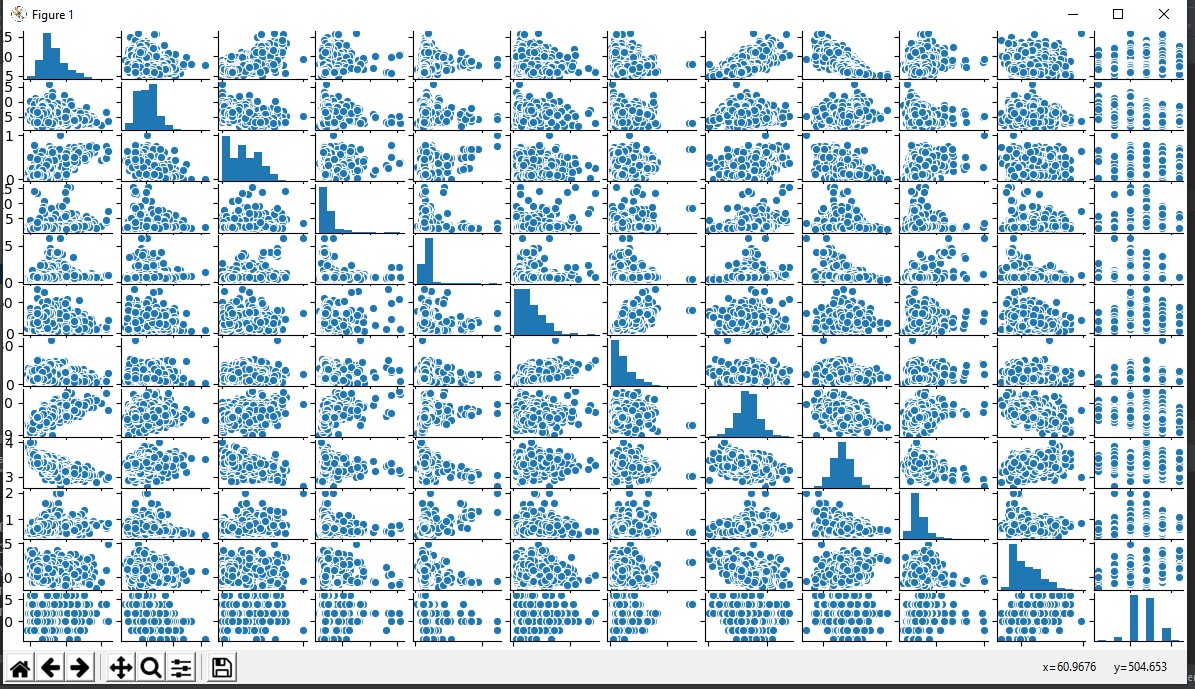
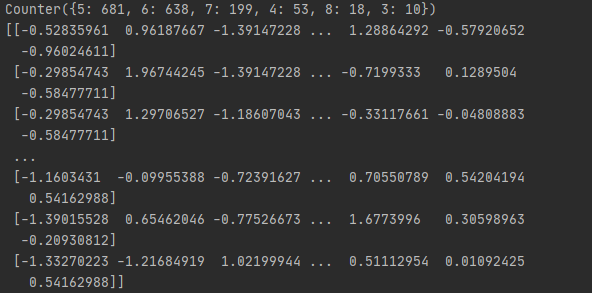
#print(y\_train.shape) #print(x\_test.shape) #print(y\_test.shape)

lr = LogisticRegression() lr.fit(x\_train, y\_train) lr\_predict = lr.predict(x\_test)

#print confusion matrix and accuracy score lr\_conf\_matrix = confusion\_matrix(y\_test, lr\_predict) lr\_acc\_score = accuracy\_score(y\_test, lr\_predict) print(lr\_conf\_matrix) print("Accuracy:",lr\_acc\_score\*100)

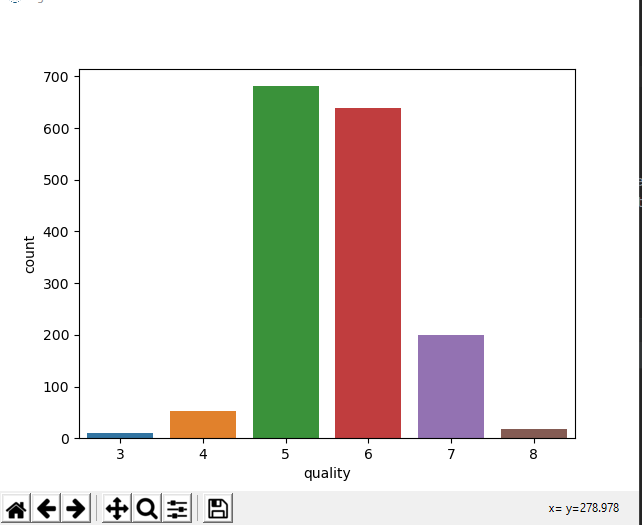
**Output:**



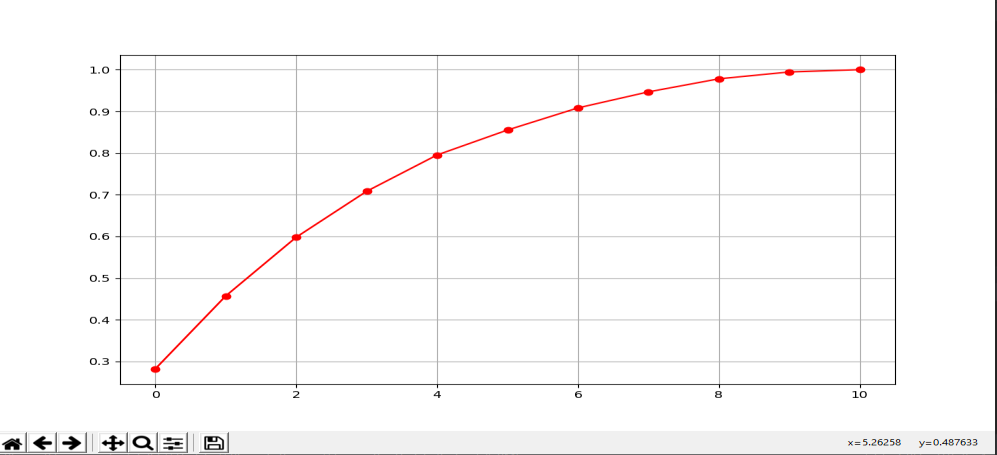


(Shows the correlation between variables)

MCA Department, Surana College Page 24



(Count of the target variable)



(Graph to find the principle components)

**Chapter 3 CONCLUTION AND FUTURE SCOPE**

* 1. **Conclusion**
     + This report uses two datasets of vinos Verde wine to predict the wine type, red or white, and the quality based on physicochemical properties. Quality is a subjective measure, given by the average grade of three experts.
     + Before starting the predictions, the report makes a brief summary of model evaluation, explaining the most common metrics used in categorical problems in machine learning.
     + In data preparation, the datasets are downloaded and imported in R. In this phase, the training and testing sets are created and they will be used during the model building.
     + In data exploration and visualization we look for features that may provide good prediction results. The best predictors have low distribution overlapping area and low correlation among them.
     + Modeling starts explaining very simple models and gradually moves to more complex ones. There’s a brief explanation of the models used in this report. Other important machine learning concepts, such as ensemble and cross validation, are also discussed. Clustering is unsupervised method, which is included for illustrative purposes.
     + The results section presents the modeling results and discusses the model performance. The algorithms are

used to predict wine type; there’s a section for predicting quality and the last part demonstrates clustering.

* + - The physicochemical properties of wine differ between red and white wines, but the difference is not so evident when evaluating red wine quality. Maybe there are other properties not considered in this dataset that are better indicators for quality.

## Future Scope

* + - The machine learning models used in this research aren’t able to predict red wine quality with high accuracy, specificity and sensitivity. This is partially explained by the low prevalence of quality levels 3, 4 and 8 and the large distribution overlapping area stratified by quality.
    - Besides this, the lack of information about how the dataset was created may impact the prediction of quality using the physicochemical properties as predictors. Such examples are the composition of grape varieties in each wine, the mix of experts that evaluated wine quality, or the production year.

# BIBILOGRAPHY

## Bibliography and References:

1. Datasets ( [www.kaggle.com](http://www.kaggle.com/) )
2. Sklearn libraries [( https://www.sciki](http://www.scikit-learn.org/)t[-learn.org/](http://www.scikit-learn.org/) )
3. WHO ( [https://www.who.int/wine](http://www.who.int/wine-qualities/en/)-[qualities/en/](http://www.who.int/wine-qualities/en/) )?
4. UCI Machine Learning Repository ( <https://archive.ics.uci.edu/ml/index.php/>)