A close up of a sign

Description automatically generated

REPORT ON

CLASSIFICATION OF RED WINE

A close up of a wine glass

Description automatically generated

NAME : ANIKET CHAVAN

STEVENS CWID : 10449866

# *Table of Contents*

1. ***Introduction and Data Statistics***
2. ***Approach***
3. ***Experimental Design***
4. ***Experimental Results***
5. ***Conclusion***
6. ***References***

***Abstract***

Wine tasting is an art since decades. The main purpose of this study is to predict wine quality based on physicochemical data. These data sets contain around 1600 instances for red wine with 11 features of physicochemical data such as alcohol, chlorides, density, total sulfur dioxide, free sulfur dioxide, residual sugar, and pH. The following five different algorithms were used to classify the quality of red wine: random forests, support vector machines, naïve bayes, logistic regression and decision tree. There are 3 quality classes of red wine. The most successful classification was obtained by using Naïve Bayes Algorithm with 99.25 accuracy rate. In this study, it is also observed that the use of principal component analysis in the feature selection increases the success rate of classification in Naïve Bayes Algorithm.

1. **Introduction**

Today Wine industry is growing and researchers are researching on new technologies for both wine making and selling processes in order to back up this growth. Classification models may be used as part of decision support system in different stages of wine production, hence giving the opportunity for manufacturer to make corrective and additive measure that will result in higher quality wine being produced.

**[1]** For new data mining classifiers this data set has been majorly used as a benchmark

because it is very easy to discriminate. For wine classification according to geographical region; principal component analysis (PCA) was carried out and reported. The data they used in their study includes 33 Greek wines with physicochemical variables. Another work of wine classification depended on the physicochemical information. This information involved in wine aroma chromatograms as measured with a Fast GC Analyzer

1. **Approach**

**Problem Statement and Objective:**

Online articles are different from academic research papers. They cannot be fully trusted, however there are still many valid research papers showing strong evidence that wine tasting is pseudoscience. **[2]** If data mining algorithms can be applied to make predictions with lower errors than random prediction or dummy classifier, some consistency could be added to wine tasting. It is difficult to determine an empirical relation between the subjective quality of a wine and its chemical composition. Wine makers want to know what they can do to their processes to optimize the quality of their wine. **[2]** While attempts have been made to build classifiers for wine from chemical data, not all algorithms have been tested. So, our problem statement is to design and implement the Classification of Red Wine using machine learning algorithms with Python and the tool Scikit-Learn also applying principal component analysis to improve; the objective is to classify the best quality of wine among the existing ones on the Grading class ranging from certain to certain value.

1. **Experimental Design**

**Database Overview**

|  |  |
| --- | --- |
| **Associated Task** | Classification |
| **Data Set Characteristics** | MultiClass |
| **Attribute Characteristics** | Real |
| **Number of Instances** | 1599 |
| **Number of Attributes** | 11 |
| **Missing Values?** | **No** |

**Predicting attribute:** Quality Class of Red Wine Dataset

**Attribute Information:** We have 11 features in this dataset and a target variable class. The dataset includes the following attributes:

* 1. Fixed acidity
* 2. Volatile acidity
* 3. Citric acid
* 4. Residual sugar
* 5. Chlorides
* 6. Free sulfur dioxide
* 7. Total sulfur dioxide
* 8. Density
* 9. pH
* 10. Sulphates
* 11. Alcohol
* Class(Quality):
  + Quality Range 1 to 3
  + Quality Range 4 to 7
  + Quality Range 8 to 10

Since we are trying to find if a wine is good or bad, why not to adapt our target column to reflect this approach? Let’s divide our dataset in 3 classes:

Poor: all wines with rates below 4.

Average: wines with rates 4 to 7.

Excellent: wines with rates higher than 7.

**Import libraries**

A screenshot of a cell phone

Description automatically generated

**Load Dataset**

Pandas is used to load the data.

A screenshot of a cell phone

Description automatically generated

**Summarize the Dataset**

In this step we are going to take a look at the data a few different ways:

1. Dimensions of the dataset.
2. Peek at the data itself.
3. Statistical summary of all attributes.
4. Breakdown of the data by the class variable.

A close up of a logo

Description automatically generated

A close up of a piece of paper

Description automatically generated

A close up of a piece of paper

Description automatically generated

A close up of text on a white background

Description automatically generated

A screenshot of a cell phone

Description automatically generated

## **Experimental Results**

## **Evaluation of Algorithms**

## Creation of validation dataset.

## Test – Train Split

## Build multiple different models

1. Best model Selection

### **Creation of Validation Dataset**

Statistical methods to estimate the accuracy of the models that we create on unseen data. We also want a more concrete estimate of the accuracy of the best model on unseen data by evaluating it on actual unseen data.

That is, we are going to hold back some data that the algorithms will not get to see and we will use this data to get a second and independent idea of how accurate the best model might actually be.

We will split the loaded dataset into two, 80% of which we will use to train, evaluate and select among our models, and 20% that we will hold back as a validation dataset.

### **Test – Train Split**

A screenshot of a cell phone

Description automatically generated

This is a ratio of the number of correctly predicted instances divided by the total number of instances in the dataset multiplied by 100 to give a percentage (e.g. 95% accurate). We will be using the scoring variable when we run build and evaluate each model next.

1. **Build Models**

We'll evaluate these 5 algorithms:

* Logistic Regression (LR)
* Decision Trees (DT)
* Gaussian Naive Bayes (NB)
* Support Vector Machine (SVM)
* Random Forests (RF)

**Logistic Regression**

**A screenshot of a cell phone

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

### **Decision Trees**

A screenshot of a cell phone

Description automatically generated

**Gaussian Naïve Bayes**

A screenshot of a cell phone

Description automatically generated

**Support Vector Machine**

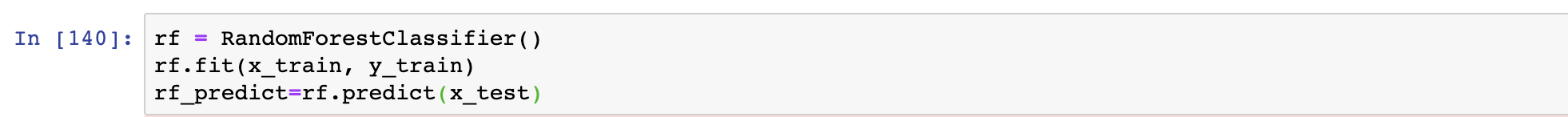
A picture containing screenshot

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Random Forests**

****

**A screenshot of a cell phone

Description automatically generated**

### **Select Best Model**

We now have 5 models and accuracy estimations for each. We need to compare the models to each other and select the most accurate.

From the above results, Gaussian Naïve Bayes proves to be the Best Model out of all with 99.25% accuracy- maximum amongst the all five.

A close up of a clock

Description automatically generated

Finally, the classification report provides a breakdown of each class by precision, recall, f1-score and support showing excellent results.

A screenshot of a cell phone

Description automatically generated

1. **Conclusions**

For each classification model, we analyzed how the results vary whenever test mode is changed. The study includes the analysis of classifiers on red wine data set. The results are described in percentage of correctly classified instances, Precision, Recall, F-Measure, and Support after applying the cross-validation or percentage split mode.

Different classifiers like Logistic Regression, random forests, support vector machines and others are evaluated on datasets. Results from the experiments lead us to conclude that Gaussian Naïve Bayes Algorithm performs better in classification task as compared against the support vector machine, Decision Tree and others. So, in classifying the red wine dataset Gaussian Naïve Bayes Algorithm seemed to be the best in terms of accuracy.

1. **References**

**[1] Yesim Er, Ayten Atasoy The Classification of White Wine and Red Wine According to Their Qualities**

**[2] Noah Spriggs, Murray Dunne, Chris Life, Greg Richardson, Haoyan Xu SENG474: Wine Analysis**