Configuration Manual

MSc Research Project
Data Analytics

Swapnil Chinchwalkar
Student ID: x21106681

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Swapnil Chinchwalkar
Student ID:	x21106681
Programme:	Data Analytics
Year:	2022/23
Module:	MSc Research Project
Supervisor:	Mohammed Hasanuzzaman
Submission Due Date:	15/12/2022
Project Title:	Configuration Manual
Word Count:	2140
Page Count:	14

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Swapnil Chinchwalkar
x21106681

1 Introduction

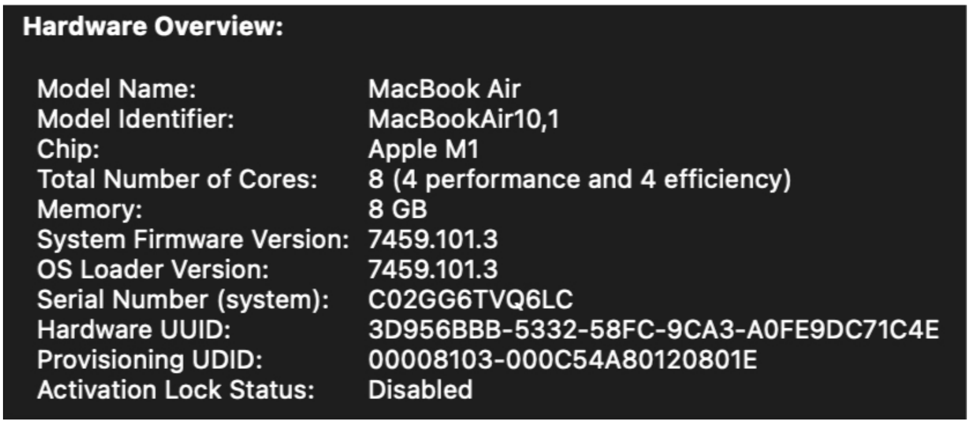
The setup documents detail how to launch the scripts that were written for this study. This will guarantee the code is error-free and runs efficiently. In addition to the same minimal recommended configuration, this also includes details on the hardware configuration of the computer on which the scripts are executed. Reproducing the project's results will be simpler if these steps are adhered to. Further investigation and analysis are therefore simplified by having this information at hand.

2 System Requirement

This section dives into the precise needs, both for hardware and software, that must be met in order to put the study into practice.

2.1 Hardware Configuration

The required hardware specifications are shown in Figure 1.



```
Hardware Overview:

  Model Name:                 MacBook Air
  Model Identifier:           MacBookAir10,1
  Chip:                       Apple M1
  Total Number of Cores:      8 (4 performance and 4 efficiency)
  Memory:                     8 GB
  System Firmware Version:    7459.101.3
  OS Loader Version:          7459.101.3
  Serial Number (system):     C02GG6TVQ6LC
  Hardware UUID:              3D956BBB-5332-58FC-9CA3-A0FE9DC71C4E
  Provisioning UDID:          00008103-000C54A80120801E
  Activation Lock Status:     Disabled
```

Figure 1: System Hardware Specification

- Device Name - Mac Book Air
- Operating System - macOS Monterey
- Display - Build in Retina (13.3-inch (2560 × 1600))
- Operating Version - 12.3.1

2.2 Software Configuration

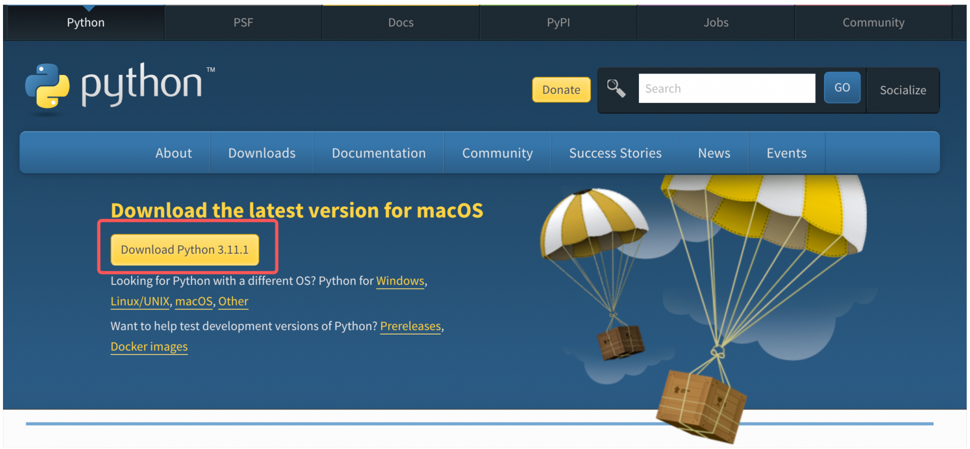
- Anaconda Navigator for MacOs (Version 2.1.4)
- Jupyter Notebook (Version 6.3.0)
- Python (Version 3.8)

2.3 Code Execution

Both the Jupyter Notebook and Google Collaboratory are capable of running the code. The Jupyter Notebook is included with Anaconda 3, and it may be started directly from the Anaconda 3 starting screen. The Jupyter Notebook will open in the specified browser. Once the organisational structure of the system has been shown by the web browser, go to the folder in which the code file is stored. Access the code file located inside the folder, and then, to execute the code, choose ”run all cells” from the Kernel menu. In a similar manner, putting the data on Google Drive and linking it to Google Collab is a way to get the code to execute after downloading the packages shown in figure 2.

3 Downloads and Installation

- Python



Download the latest version for macOS

Looking for Python with a different OS? Python for Windows, Linux/UNIX, macOS, Other

Want to help test development versions of Python? Prereleases, Docker images

Figure 2: Python Download Page

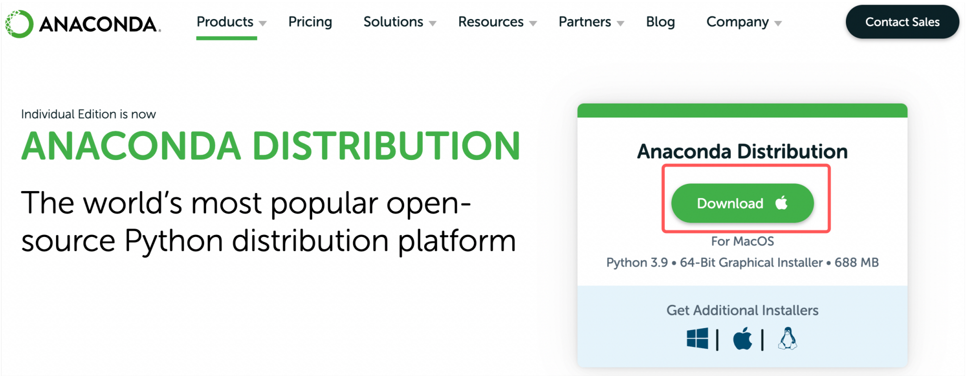
Python is the programming language that is used to carry out this study. It features an impressively large number of models that can be used to help Deep - Learning and Machine - learning. It also features a number of libraries and various modules, all of which contribute to its simplicity of use and implementation, as well as to the seamless processing of pictures and their modifications. Therefore, having the most recent edition of Python downloaded is the very first thing that has to be done in order to execute the code on the laptop. This may be done by going to the download page on the Python website [1] and installing the software installation for the version one wants, depending

[1] Python Download: `https://www.python.org/downloads/`

on the MacBook's operating system that will be running Python. A snapshot of the official Python website, where the most recent version may be downloaded and installed, is shown in Figure 2. The file has to be installed once it has been downloaded by carefully following the installation instructions.

- Anaconda

Anaconda is the next item to be downloaded. It provides a variety of Python-based IDEs that can be used for coding and result viewing. Jupyter Note book & Spyder are the most prominent IDEs offered by Anaconda Navigator by default. It is downloaded from the official website [2] for several operating systems, requiring the installation of the OS-specific installer. Multiple IDEs are presented upon successful download and installation of Anaconda Navigator, which may be picked based on the developer's needs. This research study uses Jupyter IDE, one of the various integrated development environments (IDEs). Figure 3 depicts a screenshot of the authorized Anaconda website, where the most latest edition can be downloaded and installed.



Individual Edition is now

ANACONDA DISTRIBUTION

The world's most popular open-source Python distribution platform

Anaconda Distribution

For MacOS

Python 3.9 • 64-Bit Graphical Installer • 688 MB

Figure 3: Anaconda Download Page

- Data Source

This investigation makes use of the datasets titled "Wine Review" and "Red Wine Quality." The meta data & data sets are accessible online [3]. As there are a large number of different types of wine accessible, the author has included information on the responses of wine drinkers and the chemicals found in wine. The study uses quality and description characteristics as its goal variables, which are derived from the total number of wines that are presently available. This project makes use of both the source data as well as the meta data that is presently available for wine.

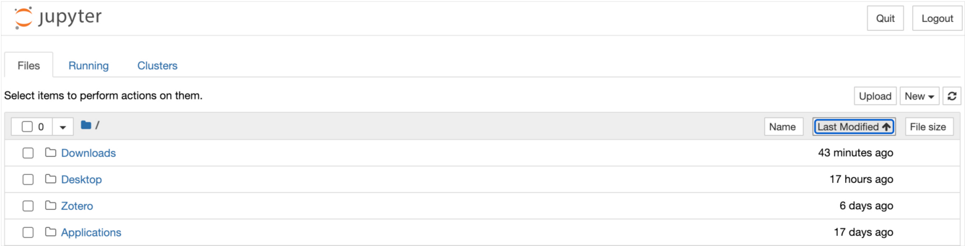
4 Project Development

It is essential that Jupyter Notebook be started from the navigator that was installed, as seen in Figure 4. When you start the Jupyter Studio, a new tab will automatically appear in your web browser. You are able to initiate the creation of a fresh Python 3

[2] Anaconda Download: `https://www.anaconda.com/products/distribution`

[3] Meta Data: `https://www.winemag.com/?s=&drink_type=wine`

notebook and provide an appropriate name for the associated file. This is the very first step of the coding process. The.ipynb file extension will be used for the new file that will be created by us. In order for the project to continue with the construction of Machine Learning algorithms, some more libraries of Python will need to be installed as needed. Installing these libraries is as simple as using the pip command either on the command line or from the Jupyter Notebook.



Select items to perform actions on them.

Downloads	43 minutes ago
Desktop	17 hours ago
Zotero	6 days ago
Applications	17 days ago

Figure 4: Jupyter Home Page

For instance:

1. To type at the command prompt: pip install numpy
2. With Jupyter Notebook: !pip install numpy

To begin, some of the basic libraries that must be present before a model can be constructed for sentiment analysis have been installed. These libraries have been updated to their most recent versions, and a few of the modules installed are as follows:

- sklearn
- Keras
- Transformers
- NLTK
- Numpy
- Pandas
- Seaborn
- Lazy Predict
- Collections
- Os

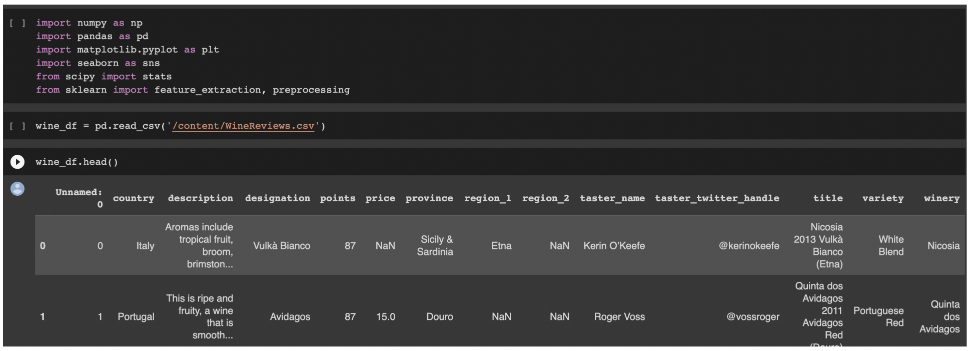
After the coding is finished, the script may be executed using the jupyter command or by executing the code in chunks. Both of these methods are described below. In the event that there are any errors in the coding, they will be shown underneath the code block so that the errors may be fixed or debugged. In order to get started with the method of running the model, it is necessary to transform the data and then retrieve it into a pandas dataframe.

4.1 Data Collection

The data was retrieved from a Kaggle source that is available to the general public; the URLs to the dataset can be found here [4] [5]. The information consists of more than 130k rows and more than 1600 rows across both data sets, with 12 and 14 attributes, respectively.

4.2 Importing Libraries and Reading File

As shown in Figure 5 & Figure 6, the required libraries are installed and the data from file in read and validated.



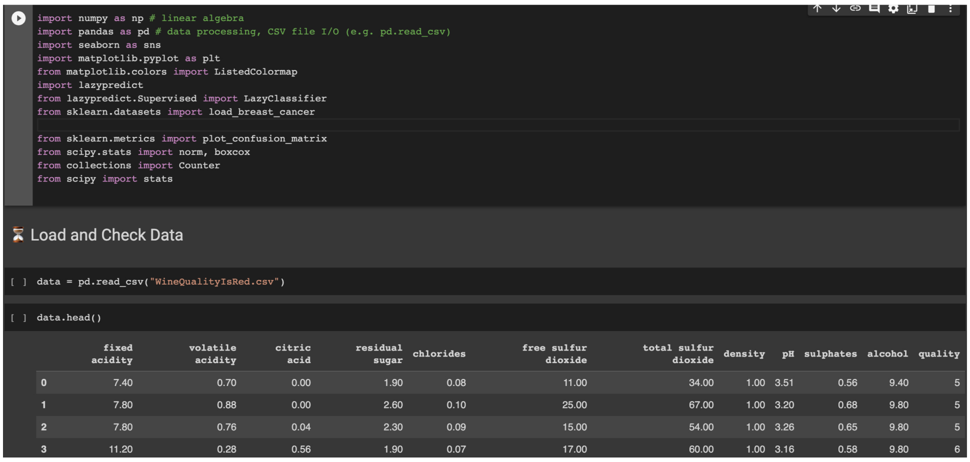
```
[ ] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    from scipy import stats
    from sklearn import feature_extraction, preprocessing
```

```
[ ] wine_df = pd.read_csv('/content/WineReviews.csv')
```

```
wine_df.head()
```

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos

Figure 5: Import Library and Read Data for Customer Reviews



```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap
import lazypredict
from lazypredict.Supervised import LazyClassifier
from sklearn.datasets import load_breast_cancer

from sklearn.metrics import plot_confusion_matrix
from scipy.stats import norm, boxcox
from collections import Counter
from scipy import stats
```

⌛ Load and Check Data

```
[ ] data = pd.read_csv("WineQualityIsRed.csv")
```

```
[ ] data.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.40	0.70	0.00	1.90	0.08	11.00	34.00	1.00	3.51	0.56	9.40	5
1	7.80	0.88	0.00	2.60	0.10	25.00	67.00	1.00	3.20	0.68	9.80	5
2	7.80	0.76	0.04	2.30	0.09	15.00	54.00	1.00	3.26	0.65	9.80	5
3	11.20	0.28	0.56	1.90	0.07	17.00	60.00	1.00	3.16	0.58	9.80	6

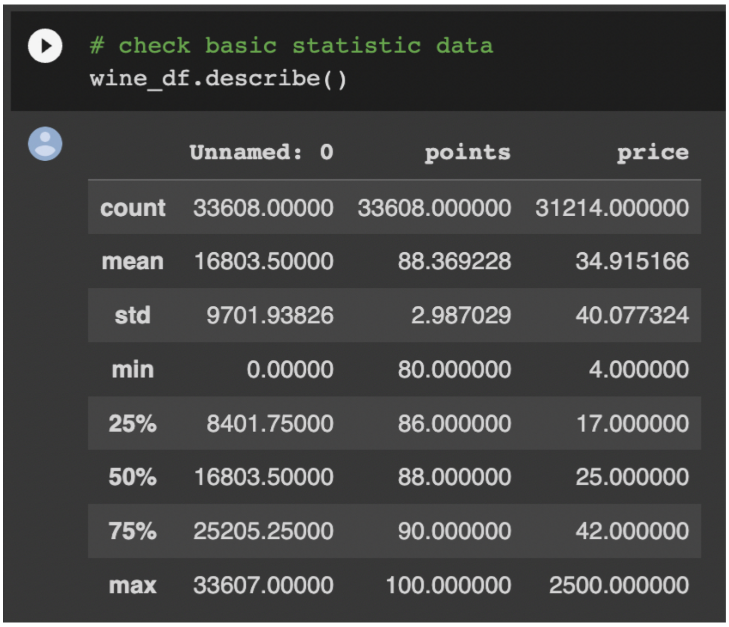
Figure 6: Import Library and Read Data for Chemical Components

4.3 Exporatory Data Analysis

To check the mean, minimum, maximum and count related to the data; describe method is used. Where as the information about the attributes is checked with the help of info method. This can be seen in Figure 7 & Figure 8.

[4] Data Set 1: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

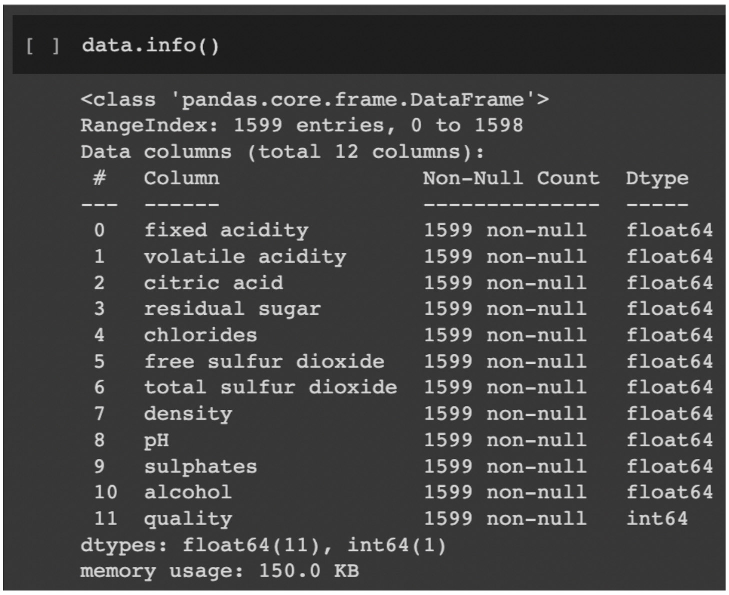
[5] Data Set 2: https://www.kaggle.com/datasets/zynicide/wine-reviews



```
# check basic statistic data
wine_df.describe()
```

	Unnamed: 0	points	price
count	33608.00000	33608.000000	31214.000000
mean	16803.50000	88.369228	34.915166
std	9701.93826	2.987029	40.077324
min	0.00000	80.000000	4.000000
25%	8401.75000	86.000000	17.000000
50%	16803.50000	88.000000	25.000000
75%	25205.25000	90.000000	42.000000
max	33607.00000	100.000000	2500.000000

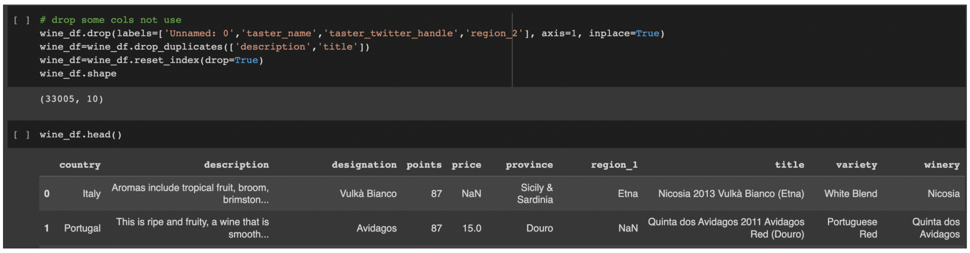
Figure 7: Data Description



```
[ ] data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

Figure 8: Data Information

Removing several columns from the code since they serve no purpose and contain a significant amount of null values. Also, removing any duplicate entries from the description and title columns as depicted in Figure 8.



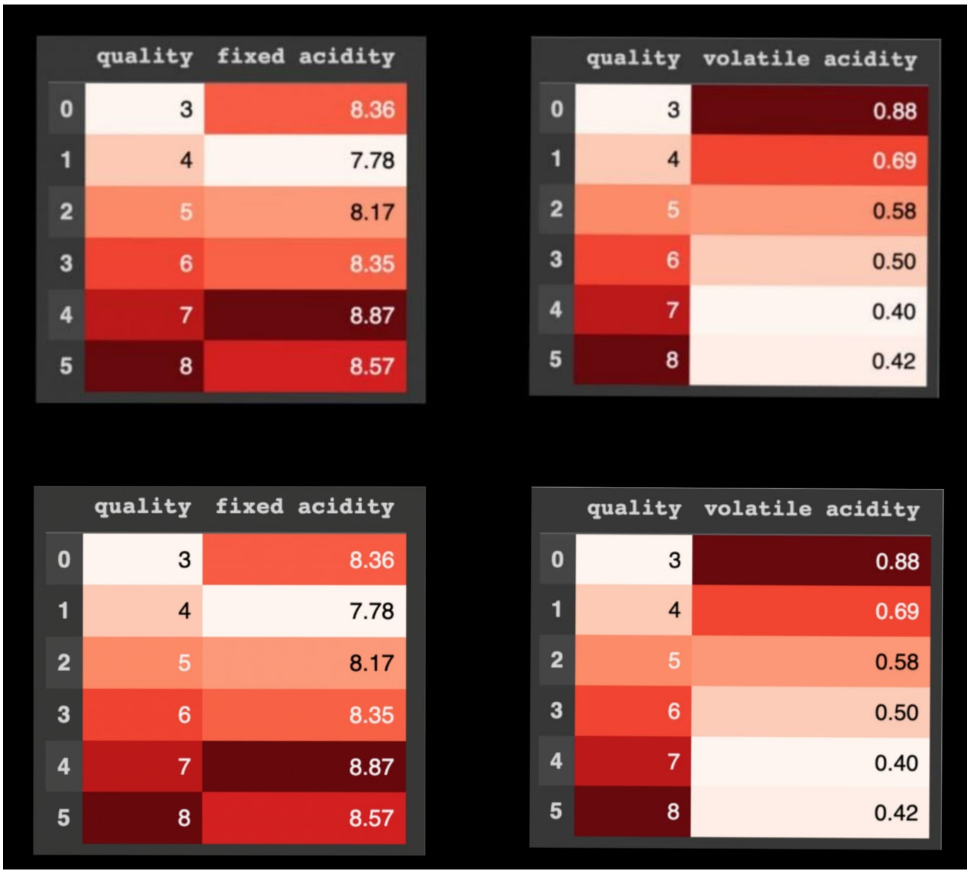
```
[ ] # drop some cols not use
    wine_df.drop(labels=['Unnamed: 0','taster_name','taster_twitter_handle','region_2'], axis=1, inplace=True)
    wine_df=wine_df.drop_duplicates(['description','title'])
    wine_df=wine_df.reset_index(drop=True)
    wine_df.shape

    (33005, 10)

[ ] wine_df.head()
```

	country	description	designation	points	price	province	region_1	title	variety	winery
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos

Figure 9: Data Filteration

- Effect of chemical properties on Quality

The effect of various attributes on the quality can be seen in figure Figure 10, Figure 11 & Figure 12



	quality	fixed acidity
0	3	8.36
1	4	7.78
2	5	8.17
3	6	8.35
4	7	8.87
5	8	8.57

	quality	volatile acidity
0	3	0.88
1	4	0.69
2	5	0.58
3	6	0.50
4	7	0.40
5	8	0.42

	quality	fixed acidity
0	3	8.36
1	4	7.78
2	5	8.17
3	6	8.35
4	7	8.87
5	8	8.57

	quality	volatile acidity
0	3	0.88
1	4	0.69
2	5	0.58
3	6	0.50
4	7	0.40
5	8	0.42

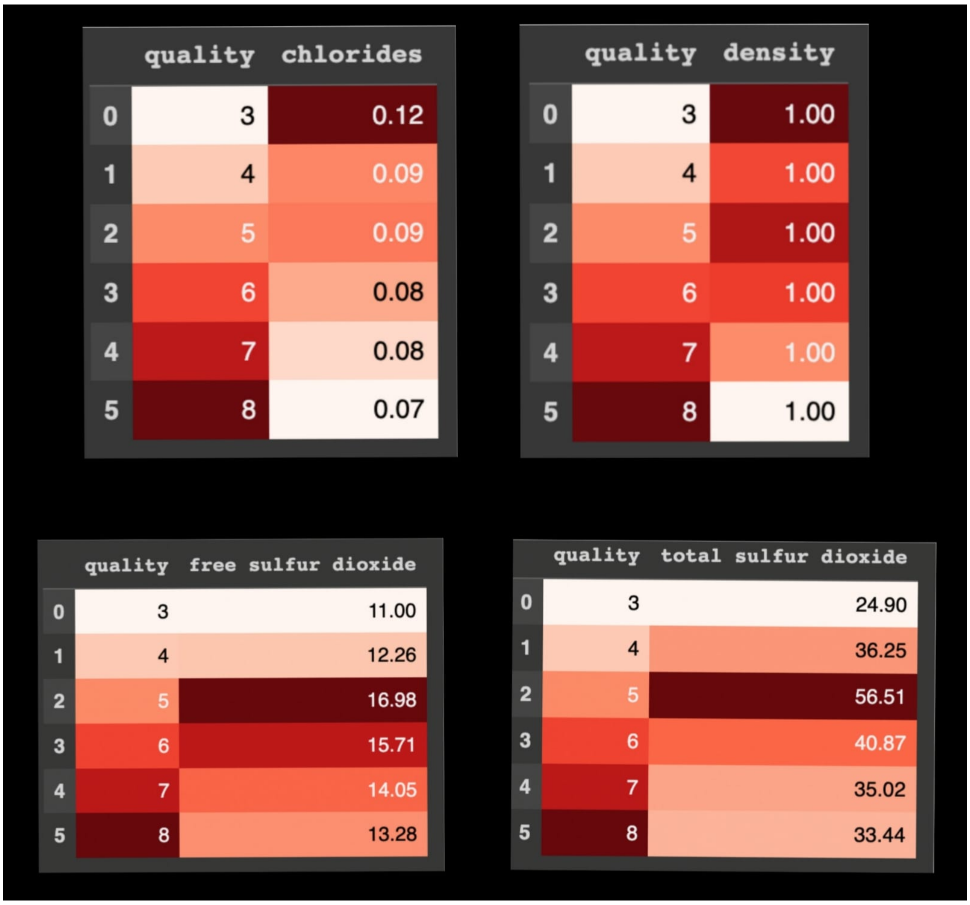
Figure 10: Atrributes effect on Quality - Case 1

1. Fixed Acidity - Quality: The increase in fixed acidity positively affects the given vote but it is difficult to make a full conclusion at this stage.

2. Volatile Acidity - Quality: Here, more precise results can be seen according to the comment made by customer. The decrease in volatile acidity affects the votes positively.

3. Citric Acid - Quality: From here, it can be seen that the increase in citric acid positively affected the votes.

4. Residual Sugar - Quality: It seems difficult to make an inference about residual sugar from the graph.



	quality	chlorides
0	3	0.12
1	4	0.09
2	5	0.09
3	6	0.08
4	7	0.08
5	8	0.07

	quality	density
0	3	1.00
1	4	1.00
2	5	1.00
3	6	1.00
4	7	1.00
5	8	1.00

	quality	free sulfur dioxide
0	3	11.00
1	4	12.26
2	5	16.98
3	6	15.71
4	7	14.05
5	8	13.28

	quality	total sulfur dioxide
0	3	24.90
1	4	36.25
2	5	56.51
3	6	40.87
4	7	35.02
5	8	33.44

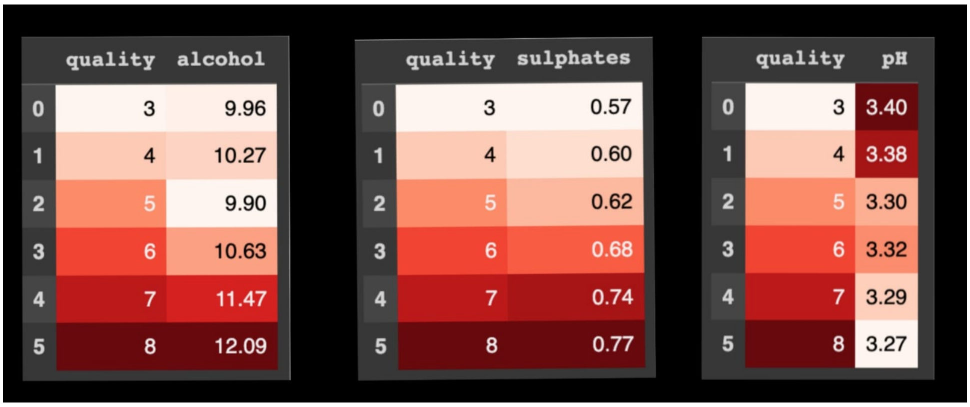
Figure 11: Atrributes effect on Quality - Case 2

1. Chlorides - Quality: The decrease in chlorides positively affects the votes cast.

2. Free Sulfur Dioxide - Quality: It doesn't seem easy to comment on free sulfur dioxide either. The 11, 12 levels seem to have been voted well.

3. Total Sulfur Dioxide - Quality: It doesn't make sense to comment on total sulfur dioxide. It changed according to the votes. There is no order.

4. Density - Quality: The intensity is almost the same in all votes, but there is only a slight decrease. Density drop seems to have a positive effect on the votes.



	quality	alcohol
0	3	9.96
1	4	10.27
2	5	9.90
3	6	10.63
4	7	11.47
5	8	12.09

	quality	sulphates
0	3	0.57
1	4	0.60
2	5	0.62
3	6	0.68
4	7	0.74
5	8	0.77

	quality	pH
0	3	3.40
1	4	3.38
2	5	3.30
3	6	3.32
4	7	3.29
5	8	3.27

Figure 12: Atrributes effect on Quality - Case 3

1. pH - Quality: The decrease in pH affects the votes positively.
2. Sulphates - Quality: It is clear from here that the increase in sulphates has a positive effect.
3. Alcohol - Quality: In general, it can be said that the increase in alcohol ratio affects the votes well.

4.4 Visualization

1. Customer Reviews

   Figure 13 shows the boxplot of different varieties of wine a customer can consume. The pints for the wine category ranges from 80 to 98 and just a minute skewness can be observed in the graph, which can be ignored or skipped. Almost all the the mentioned varieties of wine can be seen close to the 90 mark point.

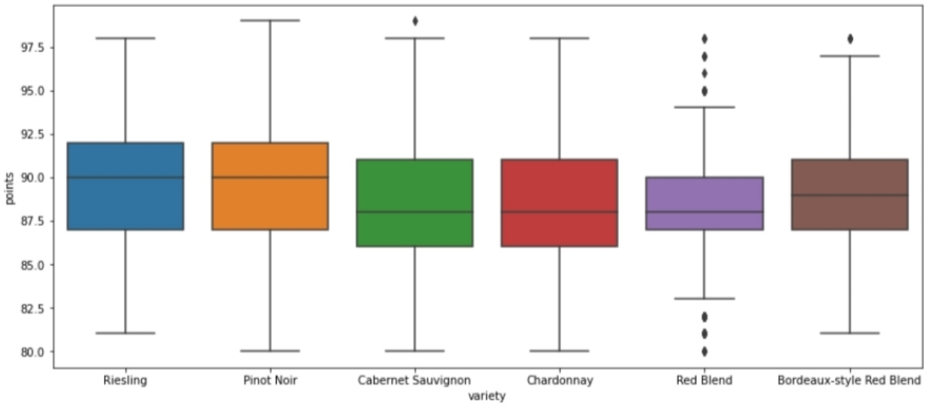


Figure 13: Wine Varieties Box Plot

The distribution of wine with respect to province can be seen in Figure 14. The top twenty countries with highest wine consumption is sorted in descending order as per the graph. Also, the graph shows the number of grapes used in the wine which is expensive.

The data preparation for the wine begins after this. The Nltk packages are imported for pre-processing, all the texts are converted to lower case and if any special characters are present in the comments - all of them are removed. Before beggining with the creation of model, the tokenization is invoked along with stemming and lemmatization process.

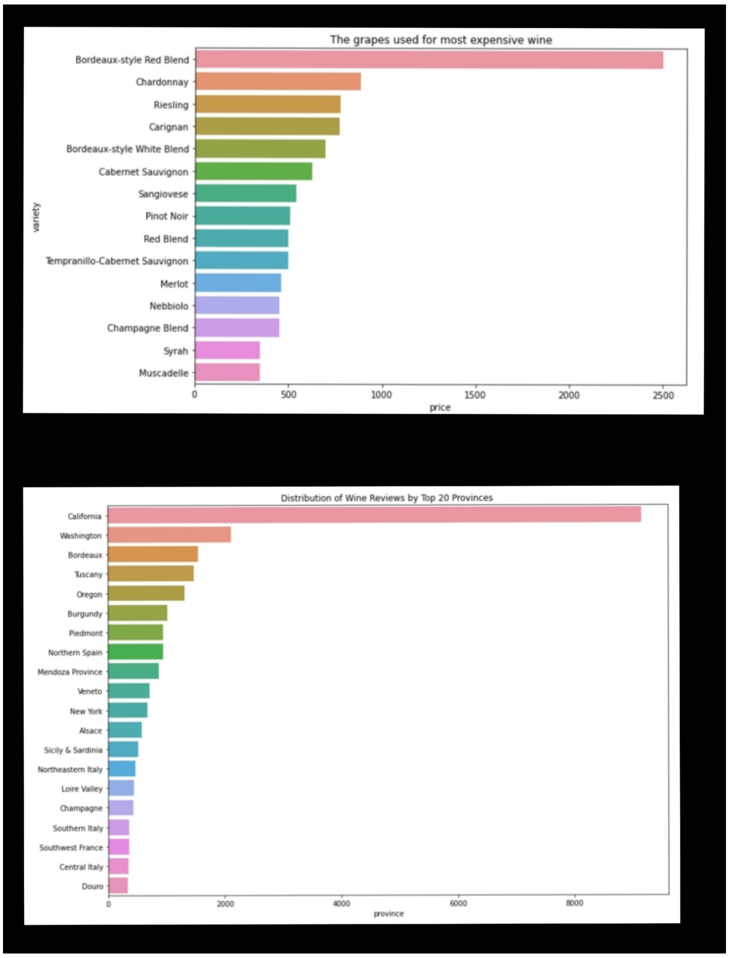


Figure 14: Province Distribution VS Grapes in Wine

2. Chemical Components

   Figure 15, Figure 16 & Figure 17 helps to check for the skewness in the graphs.

   - Fixed Acidity - It has very little of an influence on the way votes are divided, and the skewness visible to the right is something that has to be rectified.
   - Volatile Acidity - The reduction in the amount of volatile acidity appears to have a favourable effect on the votes. When we take a look at the second graph, one can observe that the dispersion is satisfactory.
   - Citric Acid - The rise in citric acid has a beneficial impact on the results of the vote.
   - Residual Sugar - The estimation of the amount of sugar that is still there does not appear to have much of an impact. Based on the normal distribution, there is a tendency toward the right that may be described as a skewness.

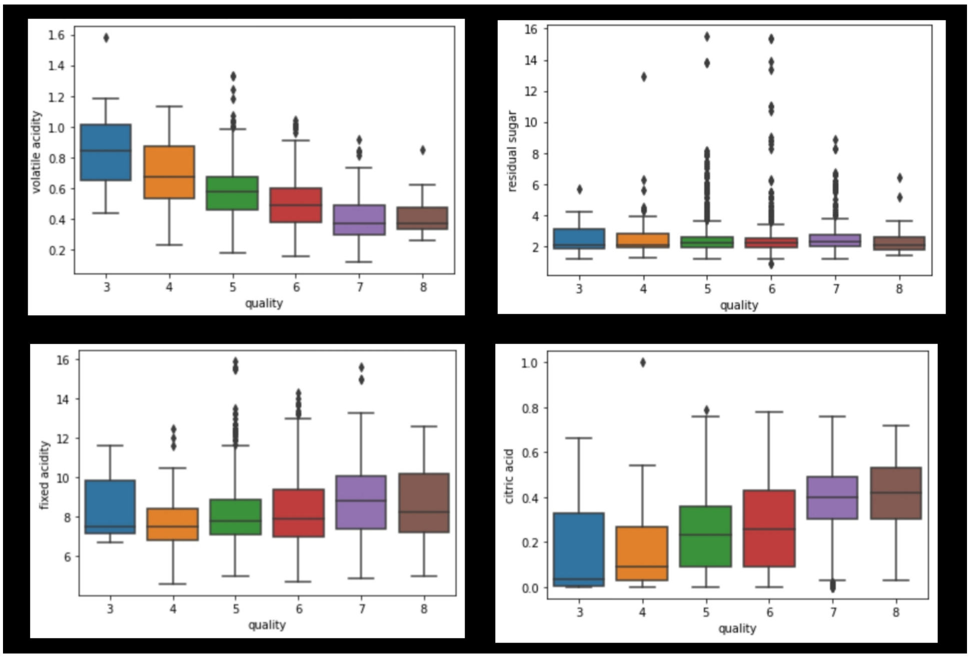


Figure 15: Detecting Skewness - Case 1

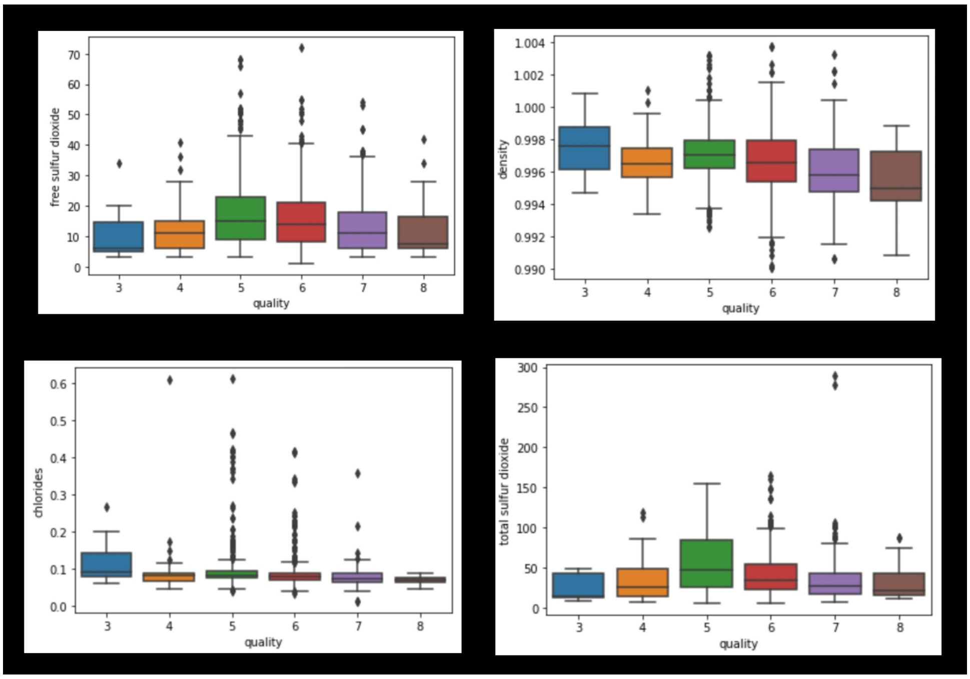


Figure 16: Detecting Skewness - Case 2

- Chlorides - The decrease in chlorides has a positive effect on the votes. There are too many outliers which needs to be fixed. Otherwise, they will negatively affect the model.
- Free Sulfur Dioxide - When one examine the graphs here, the graph is tailing to the right, which needs to be corrected. There are too many outliers which needs to be fixed else they will negatively affect the model.
- Total Sulfur Dioxide - When one looks at these charts, there is no discernible pattern of influence on the variable being studied. Coming to any kind of decision is not an easy task. A skewness towards the right may be seen in this example.
- Density - It seems to be rather challenging to provide an accurate prediction of the influence on the variable being targeted. If one looks closely at the scatterplot, one can see that the data follows a normal distribution.

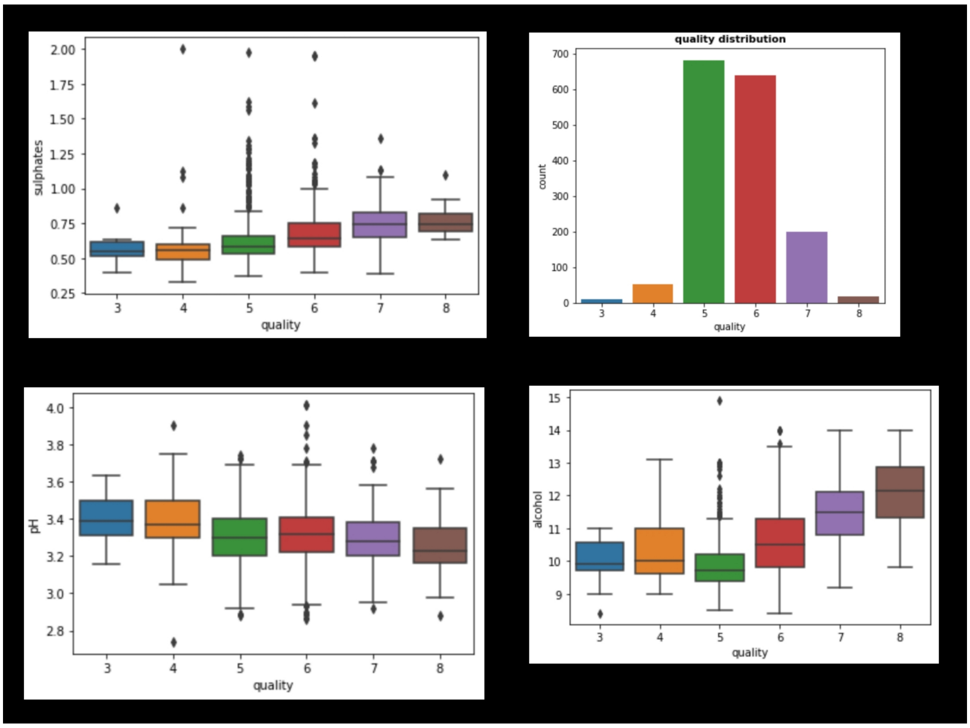


Figure 17: Detecting Skewness - Case 3

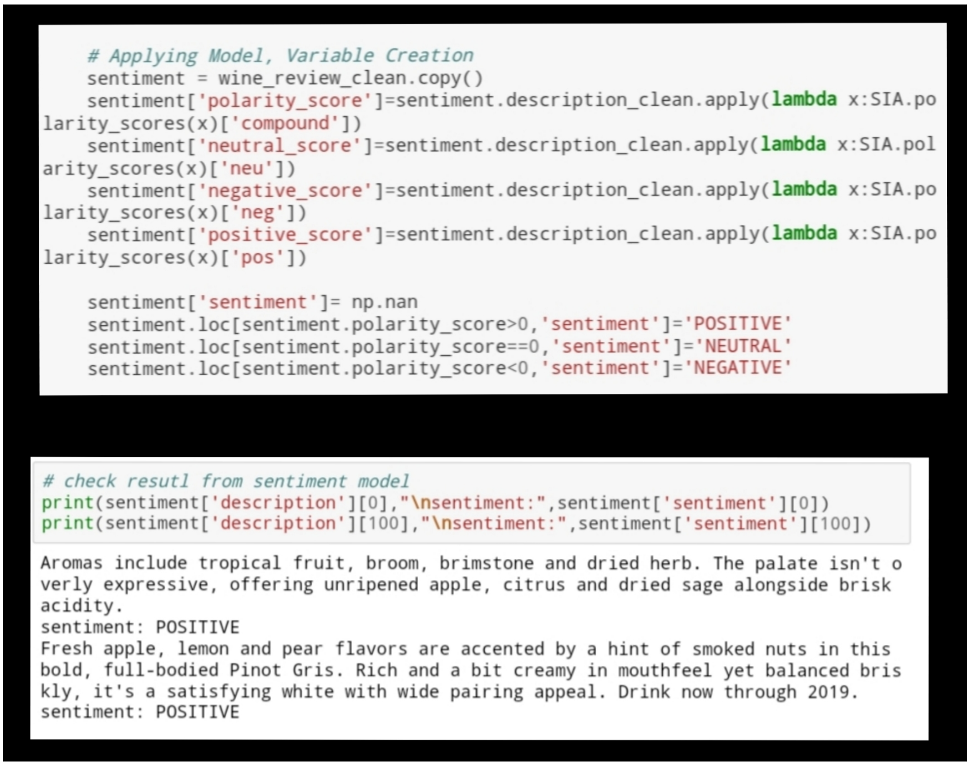
- pH - It is clear that the drop in ph level has a beneficial impact on the votes, as seen by the fact that they have increased. The scatterplot exhibits both normal and outlier values; hence, it is necessary to locate and eliminate outliers.
- Sulphates - When the quantity of the sulphates is higher, the votes tend to be more positive. It is necessary to eliminate outliers.
- Alcohol - When the chart is evaluated in its whole, it is possible to draw the conclusion that the rise in the use of alcohol has a favourable impact on the votes. The graph has skewness, which has to be corrected before it can be used.

- Quality - Votes are distributed in this chart with respect to count

Hence, the important attributes whose skewness needs to be corrected and those which effects the wine quality are: Free Sulfur Diocide, Fixed Acidity Residual Sugar, Alcohol and Total Sulfur Dioxide

5 Modelling

Figure 18 shows how the sentiments of the customers are classified as positive, negative and neutral based on the compound score of polarity. The two comments from customers shows that the response is been positive according to the VADER implementation.



```
    # Applying Model, Variable Creation
    sentiment = wine_review_clean.copy()
    sentiment['polarity_score']=sentiment.description_clean.apply(lambda x:SIA.po
larity_scores(x)['compound'])
    sentiment['neutral_score']=sentiment.description_clean.apply(lambda x:SIA.pol
arity_scores(x)['neu'])
    sentiment['negative_score']=sentiment.description_clean.apply(lambda x:SIA.po
larity_scores(x)['neg'])
    sentiment['positive_score']=sentiment.description_clean.apply(lambda x:SIA.po
larity_scores(x)['pos'])

    sentiment['sentiment']= np.nan
    sentiment.loc[sentiment.polarity_score>0,'sentiment']='POSITIVE'
    sentiment.loc[sentiment.polarity_score==0,'sentiment']='NEUTRAL'
    sentiment.loc[sentiment.polarity_score<0,'sentiment']='NEGATIVE'
```

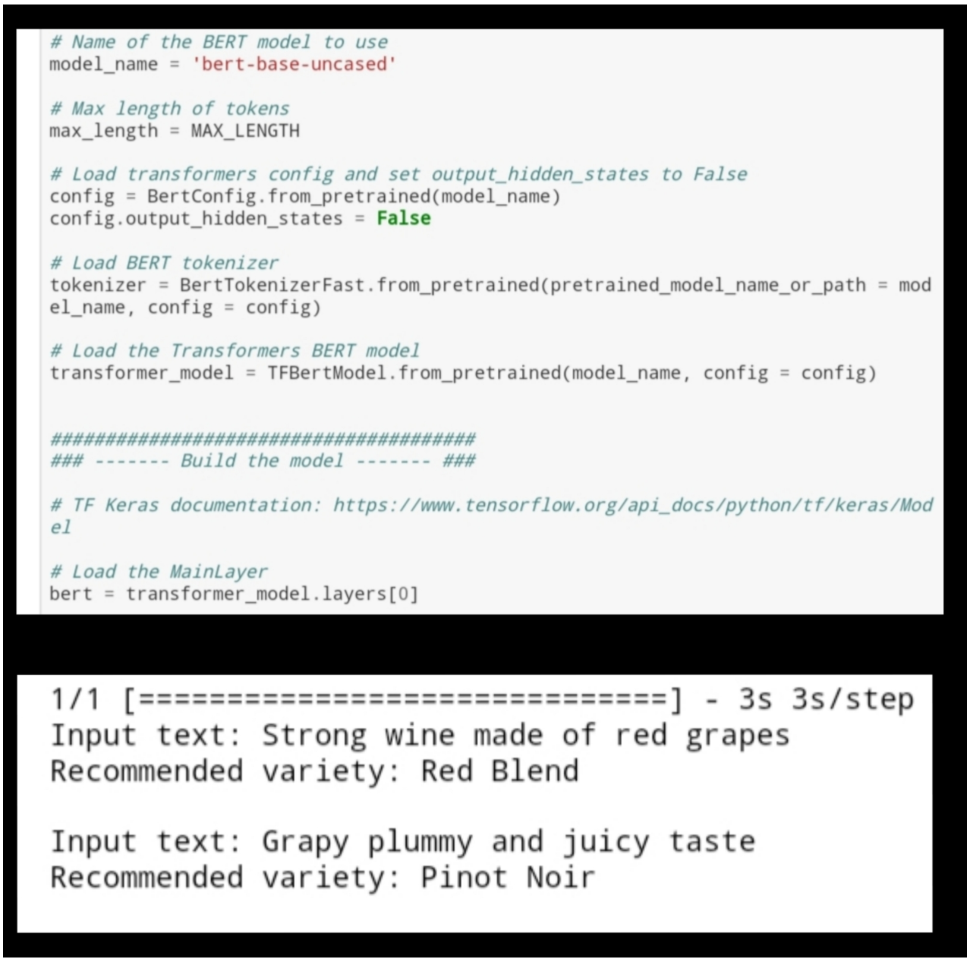
```
# check resutl from sentiment model
print(sentiment['description'][0],"\nsentiment:",sentiment['sentiment'][0])
print(sentiment['description'][100],"\nsentiment:",sentiment['sentiment'][100])
```

```
Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't o
verly expressive, offering unripened apple, citrus and dried sage alongside brisk
acidity.
sentiment: POSITIVE
Fresh apple, lemon and pear flavors are accented by a hint of smoked nuts in this
bold, full-bodied Pinot Gris. Rich and a bit creamy in mouthfeel yet balanced bris
kly, it's a satisfying white with wide pairing appeal. Drink now through 2019.
sentiment: POSITIVE
```

Figure 18: Sentiment Classification and Prediction using VADER

Following Figure 19, the image shows the setup of BERT. It uses the BERT tokenizer to load the input into the trained models with value of epochs ranging from five to ten. Also, the two inputs given to model as customer feedback provides the output as recommended wine which is of best quality according to the customers taste.

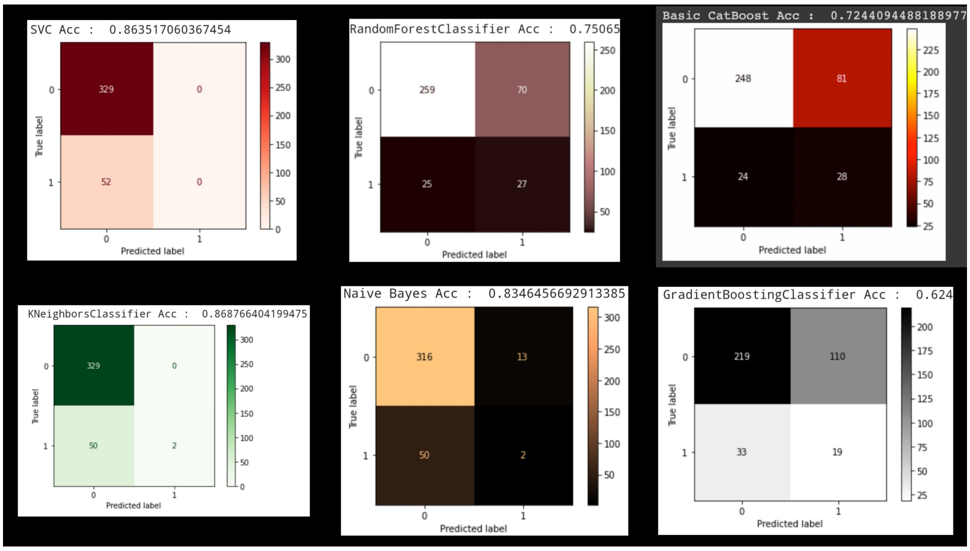
The final figure, Figure 20 shows the confusion matrix of all the models that have been used in this project. It can be seeen that the highest accuracy obtained is for SVC and KNN while Gradient Boosting has the lowest accuracy.



```
# Name of the BERT model to use
model_name = 'bert-base-uncased'

# Max length of tokens
max_length = MAX_LENGTH

# Load transformers config and set output_hidden_states to False
config = BertConfig.from_pretrained(model_name)
config.output_hidden_states = False

# Load BERT tokenizer
tokenizer = BertTokenizerFast.from_pretrained(pretrained_model_name_or_path = mod
el_name, config = config)

# Load the Transformers BERT model
transformer_model = TFBertModel.from_pretrained(model_name, config = config)


#######################################
### ------- Build the model ------- ###

# TF Keras documentation: https://www.tensorflow.org/api_docs/python/tf/keras/Mod
el

# Load the MainLayer
bert = transformer_model.layers[0]
```

```
1/1 [==============================] - 3s 3s/step
Input text: Strong wine made of red grapes
Recommended variety: Red Blend

Input text: Grapy plummy and juicy taste
Recommended variety: Pinot Noir
```

Figure 19: Wine Recommendation and Setup of BERT



SVC Acc : 0.863517060367454

RandomForestClassifier Acc : 0.75065

Basic CatBoost Acc : 0.7244094488188977

KNeighborsClassifier Acc : 0.868766404199475

Naive Bayes Acc : 0.83464566929133858

GradientBoostingClassifier Acc : 0.624

Figure 20: Confusion Matrix and Accuracy of various Models