DATA SCIENCE USE CASE

Table of Contents

1.1 Task: Creation of a full Data Science Use-Case

1.1.1 Conception phase

The goal of this proposed document is to determine a case study to develop an application for gaining insight into the origin of wines. The production of wines ends with several processes which are complex in nature . The key factors are depending on the quality of the grapes and needs more practice in making wines. A better understanding of the origin of wines is playing a significant role for controlling a good quality of wines and the protection of brand is highly demanded. This analysis will provide a comprehensive information about the composition of wines of different classes. The analysis will be done using the python platform to get a desired result.

Description of the use case:

This use case follows an effective analysis for determining the origin of wines. The work aims to develop a reliable methodology to determine the wines by plotting graphs of the chemical composition of the wines. A number of analytical techniques will be employed for analysing wines and gaining a better understanding.

The dataset for the study will be obtained from "kaggle" to ensure the best outcome as the result. After collecting the dataset the alcoholic components of the wines will be analysed. Using various techniques the identification of key compounds of wines will be undergone. The collected data will be implemented and processed by utilising a various statistical methods and algorithms for recognising patterns. Also, multivariate analysis will be executed for identifying the relationships between the characteristics of the components and the correlation will be explored as well. The detailed information of the analysis will be provided into the study therefore.

Despite the fact that chemical analysis techniques are now more sophisticated, there are still specific understanding gaps that pose challenges in precisely identifying the country of origin of wines. The following methods will be employed to fill these gaps:

- Adding more wine samples from different locations and vintages will increase the data's representativeness and boost the analysis's statistical strength (Robles *et al.* 2019).

- Studies that span several harvests and vintages will take into consideration the variances resulting from yearly climate changes and grape methods of management.

- Knowledge sharing includes the utilisation of methods and the analytical data through the research organisations and the interaction between a number of farms will be helpful for making an advancement of analysis of the origin of wines (Sivaraman *et al.* 2022).

- A better understanding about the chemical compositions and the connection between the components of the wines will be gained by summing up the knowledge of analytical chemistry and related fields.

- The origin of wines can be determined more precisely and effectively by using some cutting edge technologies which includes the users of machine learning, artificial intelligence. The patterns recognition techniques will be helpful to get an effective and accurate wine origin.

1.1.2 Development phase

Problem-Solving Techniques

Capturing the Problem:

The subjective and time-consuming nature of determining the provenance of wines is the issue at hand. Traditional approaches rely on expert judgements, which might produce discrepancies and mistakes. A system that makes use of chemical analysis and machine learning to deliver impartial and trustworthy results in identifying the origin of wines is required to get around these restrictions.

Clear Problem Definition:

The main goal is to create a model of machine learning that can be precisely predict the origin of wines based on the chemical makeup. This is an objective of clear problem definition. A large and varied dataset made up of wines from several areas and their accompanying chemical profiles should used to train this algorithm. The model hopes to accomplish these goals with the help of offering a method of automation and unbiased for identifying the region of origin of wines, boosting quality assurance, and facilitating market transparency in the wine business (Moghimi *et al.* 2020).

Understandable Concept:

The idea is to evaluate the chemical makeup of wines and predict the origin through the use of the algorithms of machine learning and evaluation of the chemical. The model of the machine learning can be uses the profiles of the chemical of wines as the features of the input, which contains the characteristics like acidity levels, sugar content, alcohol percentage, and volatile components. The programme teaches the patterns and correlations between the characteristics of the chemical and the relevant origins with the help of training on a dataset of wines with known origins. Once trained, the model can used to properly identify the origin of new wines.

Methodology/Ideas/Procedure

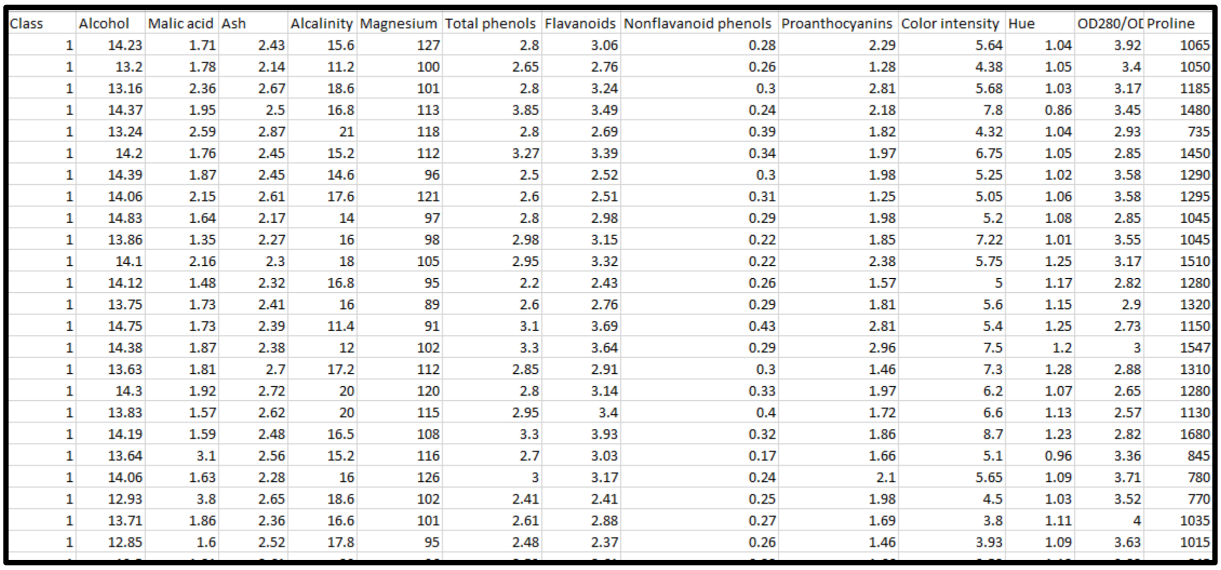
There are several approaches and methodologies are there for machine learning, the approaches are:

1. **Data Collection and Preprocessing:** Data on the chemical makeup of wines from various areas, as well as a with the help of the dataset of those wines, must gathered. This dataset ought to broad and representative, containing a range of grape types and vintages. For making sure the data is suitable for training machine learning models, it needs go through preprocessing, which can consist of cleaning, normalisation, and feature engineering (Wu *et al.* 2019).

2. **Identification and Extract of Significant Features:** It is very important to discover the features of the chemical or physical that are most useful in order to determine the provenance of the wine. The approaches of statistics and the expertise of the domain can used to select or extract the dataset's most useful aspects. The proportions of the models are reduced using this process to focus on its most elements which are distinct.

3. **Selection and Training of Machine Learning Models:** Various methods of machine learning, like random forests, support vector machines, and neural networks, can evaluated and contrasted for the prediction in order to determine the provenance of wines. In this project, logistic regression and SVM will be done with the help of Machine Learning. The dataset can be split into training and validation sets in order to optimise and train the selected model. Cross-validation and tuning of hyper Parameters are techniques that can be used to improve the generality and accuracy of the model (Wu *et al.* 2021).

4. **Model Validation:** The model that ate trained must extensively validated and evaluated in order to determine the performance. The precision, recall, and other pertinent metrics of the model can be evaluated in the real world through a different collection of wines with known origins. For implementation and to compare the performance of the model against predetermined standards, statistical tools like confusion matrices and receiver operating characteristic (ROC) curves can be utilised.



Class	Alcohol	Malic acid	Ash	Alcalinity	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	OD280/OD	Proline
1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
1	14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480
1	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
1	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
1	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
1	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295
1	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045
1	13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045
1	14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510
1	14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280
1	13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320
1	14.75	1.73	2.39	11.4	91	3.1	3.69	0.43	2.81	5.4	1.25	2.73	1150
1	14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	7.5	1.2	3	1547
1	13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310
1	14.3	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280
1	13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	1.13	2.57	1130
1	14.19	1.59	2.48	16.5	108	3.3	3.93	0.32	1.86	8.7	1.23	2.82	1680
1	13.64	3.1	2.56	15.2	116	2.7	3.03	0.17	1.66	5.1	0.96	3.36	845
1	14.06	1.63	2.28	16	126	3	3.17	0.24	2.1	5.65	1.09	3.71	780
1	12.93	3.8	2.65	18.6	102	2.41	2.41	0.25	1.98	4.5	1.03	3.52	770
1	13.71	1.86	2.36	16.6	101	2.61	2.88	0.27	1.69	3.8	1.11	4	1035
1	12.85	1.6	2.52	17.8	95	2.48	2.37	0.26	1.46	3.93	1.09	3.63	1015

Figure 1: Dataset

The above figure is the dataset of the chemical analysis that determines the origin of wines, this dataset has been collected from Kaggle.

Implementation

The execution of ***"exploratory data analysis"*** or EDA for the chemical analysis of wine is adequate in obtaining insights into this dataset. EDA assists in recognising patterns, relations, and also outliers, assessing in realising the specific wine attributes through statistical summaries or visualizations. This offers researchers in making informed decisions concerning data preprocessing, feature selection, and also selection of the model. EDA acts a critical role in extracting significant information, discovering trends, and also enhancing the accuracy of the machine-learning models for wine analysis. The implementations of this EDA on the wine dataset are illustrated below.

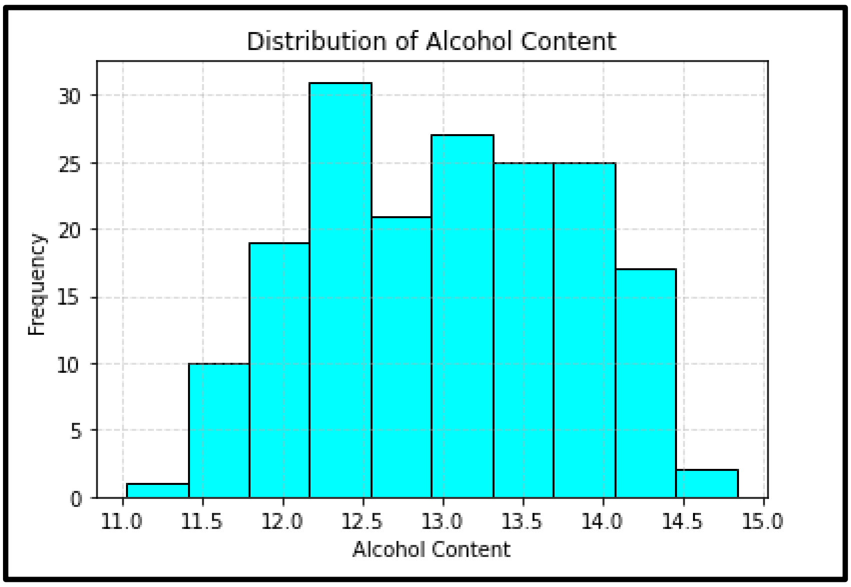


Figure 2: The distribution of the alcohol content

(Source: Obtained from the Python Environment)

This is the histogram plot of the distribution of the alcohol content. The x-axis denotes the distribution of the alcohol content and the y-axis denotes the frequency. This can be realised from the figure that the alcohol content of 12.5 has the maximum percentage.

This plot portrays the distribution of the alcohol content in wine as crucial in chemical analysis as this offers a visual representation of the overall frequency distribution. This assists in reliaising the central variability and tendency of the alcohol content in this dataset. This particular information is important for assessing the wine quality, as alcohol content relevantly influences its aroma, taste, and entire composition. Researchers can be able to make significant decisions about the quality assessment, wine classification, and also formulation of particular chemical analysis approaches by assessing the distribution (Astray *et al.* 2019).

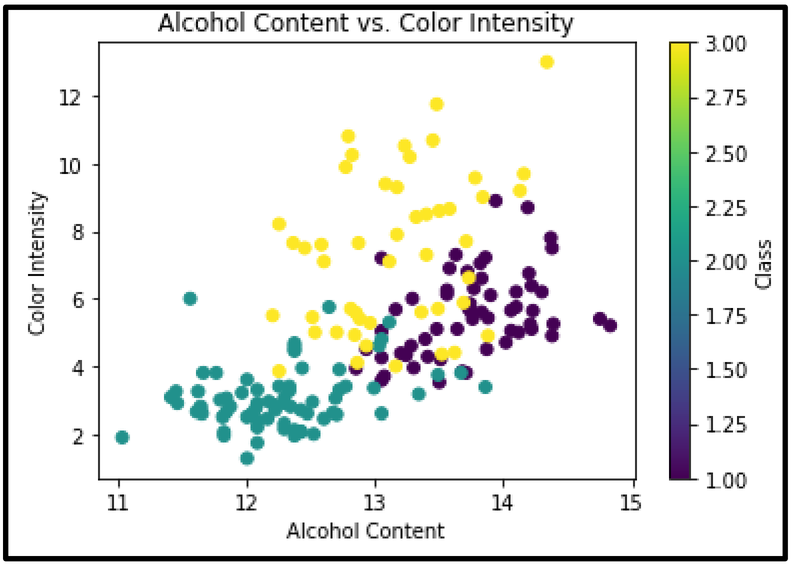
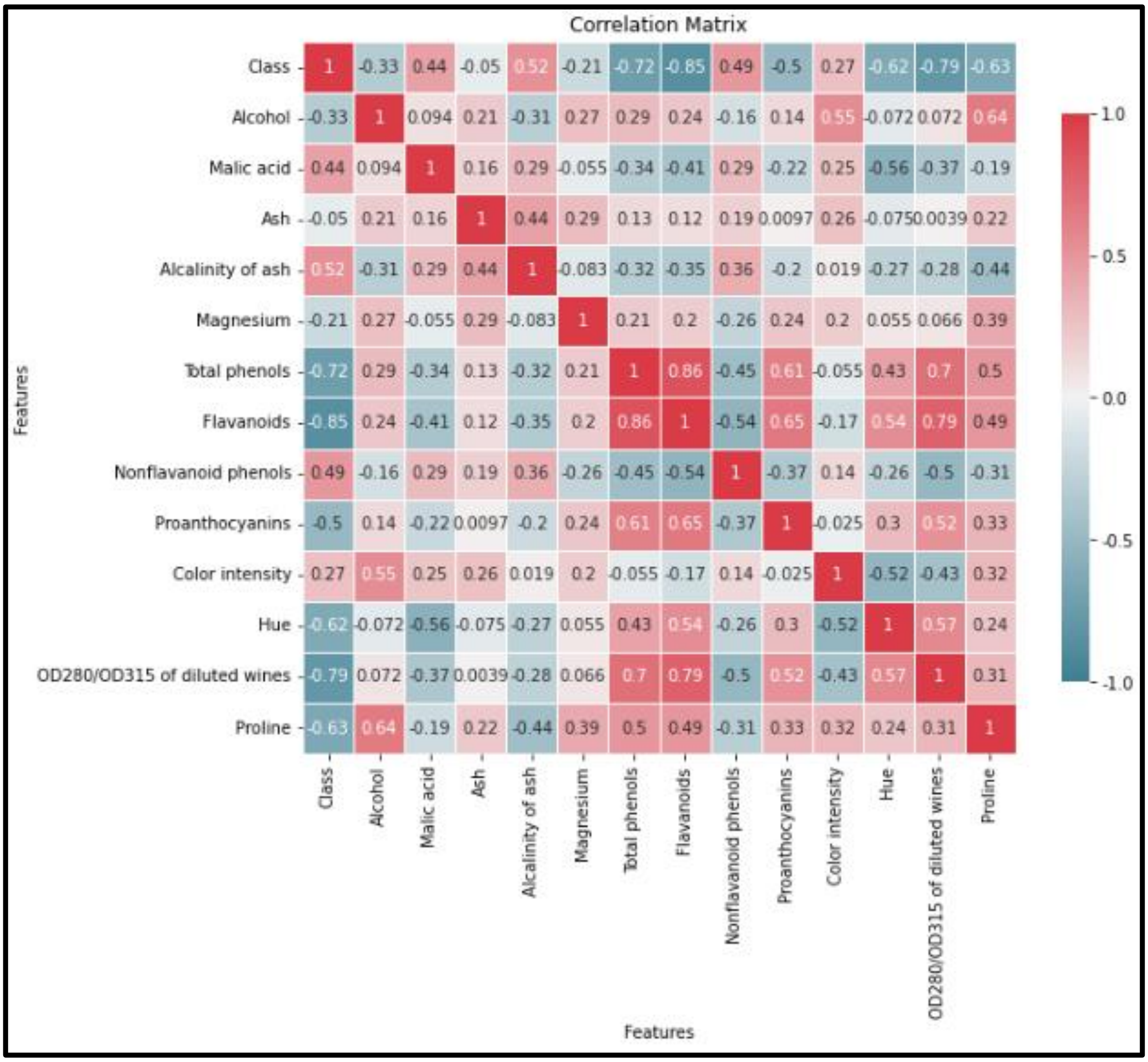


Figure 3: Showing alcohol content based on colour intensity

(Source: Obtained from the Python Environment)

The alcohol content regarding the colour intensity is visualised in the scatter plot for different classes. This can be identified that the distribution of colour intensity of class 3 has the maximum percentage. This assists in recognising any patterns or correlations between these two particular attributes. This information is significant in realising how alcohol content can improve the colour intensity of the wine, assessing in classification and quality assessment.



	Class	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	OD280/OD315 of diluted wines	Proline
Class	1	-0.33	0.44	-0.05	0.52	-0.21	-0.72	-0.85	0.49	-0.5	0.27	-0.62	-0.79	-0.63
Alcohol	-0.33	1	0.094	0.21	-0.31	0.27	0.29	0.24	-0.16	0.14	0.55	-0.072	0.072	0.64
Malic acid	0.44	0.094	1	0.16	0.29	-0.055	-0.34	-0.41	0.29	-0.22	0.25	-0.56	-0.37	-0.19
Ash	-0.05	0.21	0.16	1	0.44	0.29	0.13	0.12	0.19	0.0097	0.26	-0.075	0.0039	0.22
Alcalinity of ash	0.52	-0.31	0.29	0.44	1	-0.083	-0.32	-0.35	0.36	-0.2	0.019	-0.27	-0.28	-0.44
Magnesium	-0.21	0.27	-0.055	0.29	-0.083	1	0.21	0.2	-0.26	0.24	0.2	0.055	0.066	0.39
Total phenols	-0.72	0.29	-0.34	0.13	-0.32	0.21	1	0.86	-0.45	0.61	-0.055	0.43	0.7	0.5
Flavanoids	-0.85	0.24	-0.41	0.12	-0.35	0.2	0.86	1	-0.54	0.65	-0.17	0.54	0.79	0.49
Nonflavanoid phenols	0.49	-0.16	0.29	0.19	0.36	-0.26	-0.45	-0.54	1	-0.37	0.14	-0.26	-0.5	-0.31
Proanthocyanins	-0.5	0.14	-0.22	0.0097	-0.2	0.24	0.61	0.65	-0.37	1	-0.025	0.3	0.52	0.33
Color intensity	0.27	0.55	0.25	0.26	0.019	0.2	-0.055	-0.17	0.14	-0.025	1	-0.52	-0.43	0.32
Hue	-0.62	-0.072	-0.56	-0.075	-0.27	0.055	0.43	0.54	-0.26	0.3	-0.52	1	0.57	0.24
OD280/OD315 of diluted wines	-0.79	0.072	-0.37	0.0039	-0.28	0.066	0.7	0.79	-0.5	0.52	-0.43	0.57	1	0.31
Proline	-0.63	0.64	-0.19	0.22	-0.44	0.39	0.5	0.49	-0.31	0.33	0.32	0.24	0.31	1

Figure 4: Depicting the correlation matrix

(Source: Obtained from the Python Environment)

The mentioned figure shows the correlation matrix of the particular attributes of this wine dataset. The different values of the respective features are demonstrated in the correlation matrix.

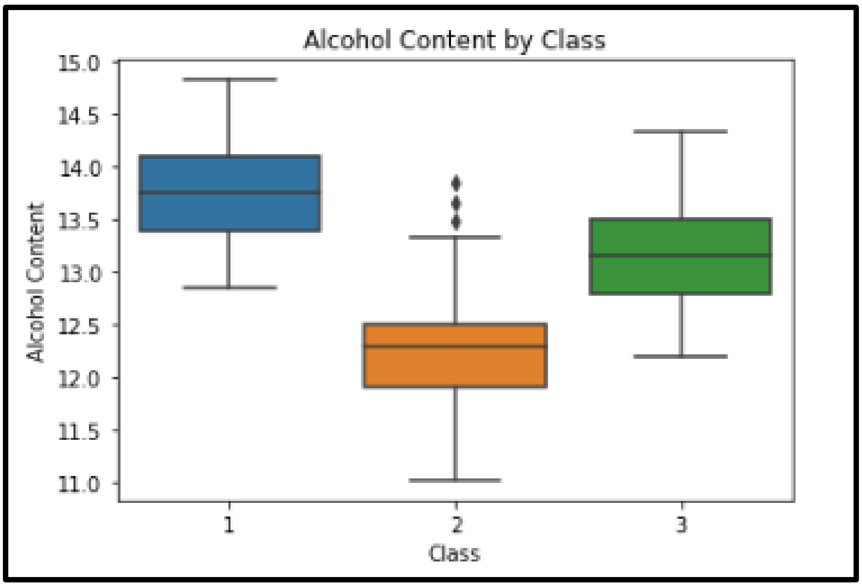


Figure 5: Displaying alcohol content with respect to class

(Source: Obtained from the Python Environment)

In the above figure the specific alcohol content by class is plotted in the boxplot. It can be seen from this figure that the class 1 has the maximum value of the alcohol content.

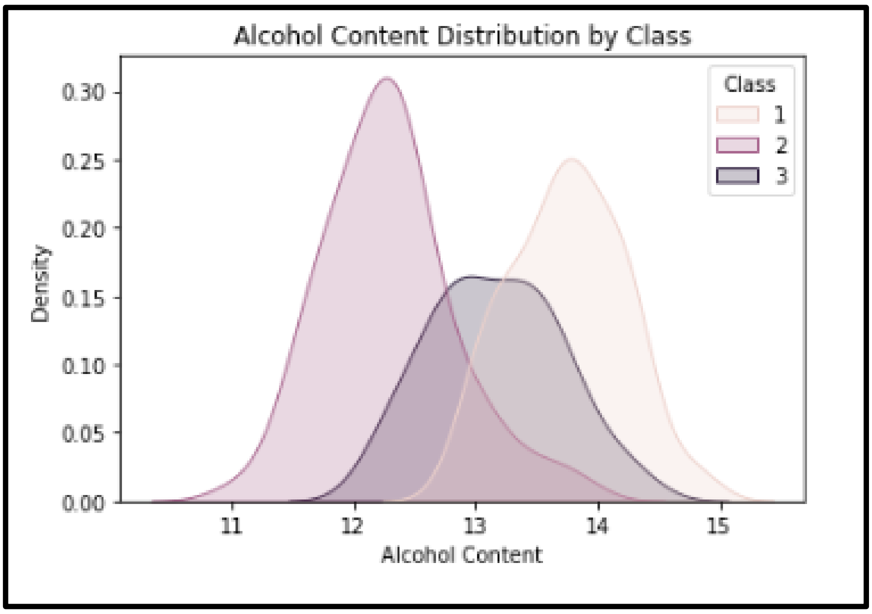


Figure 6: Showing Alcohol content distribution by class

(Source: Obtained from the Python Environment)

The respective distribution of the different alcohol contents by different classes is demonstrated here. The particular class 2 has the highest values of density and also the highest alcohol content. This portrays the trends and variations in the alcohol content across these wine classes. This particular analysis is effective for distinguishing wine classes regarding the alcohol content and also can help in the specific quality assessment and also classification of the wines regarding their chemical composition.

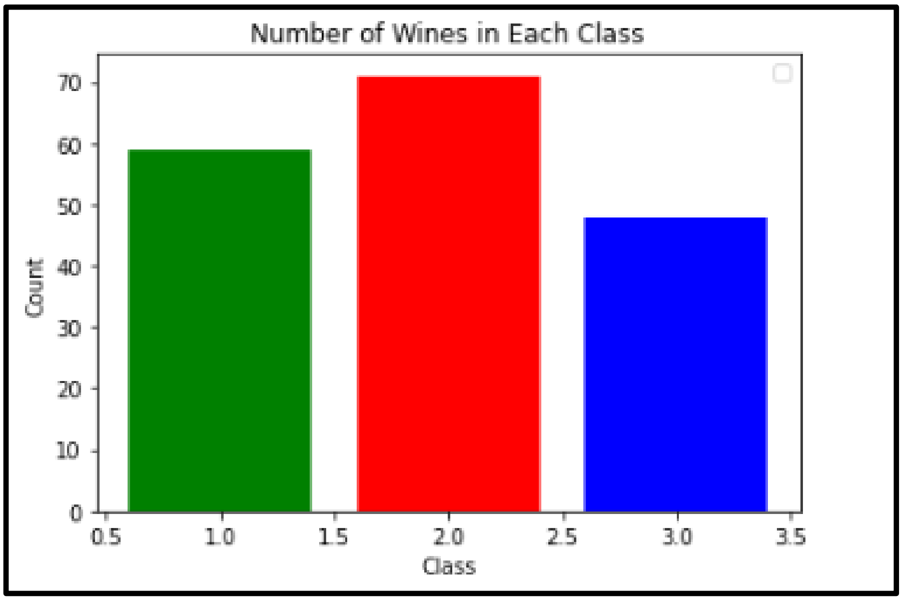


Figure 7: Demonstrating the number of wines in each class

(Source: Obtained from the Python Environment)

The overall number of wines in every class is depicted here. It can be determined from the figure that class 2 has the maximum number of wines. This offers insights into the distribution of dissimilar wine classes. This particular information assists in realising the composition of the dataset and can assess decision-making regarding model training, data sampling, along with analysis concerning specific wine classes.

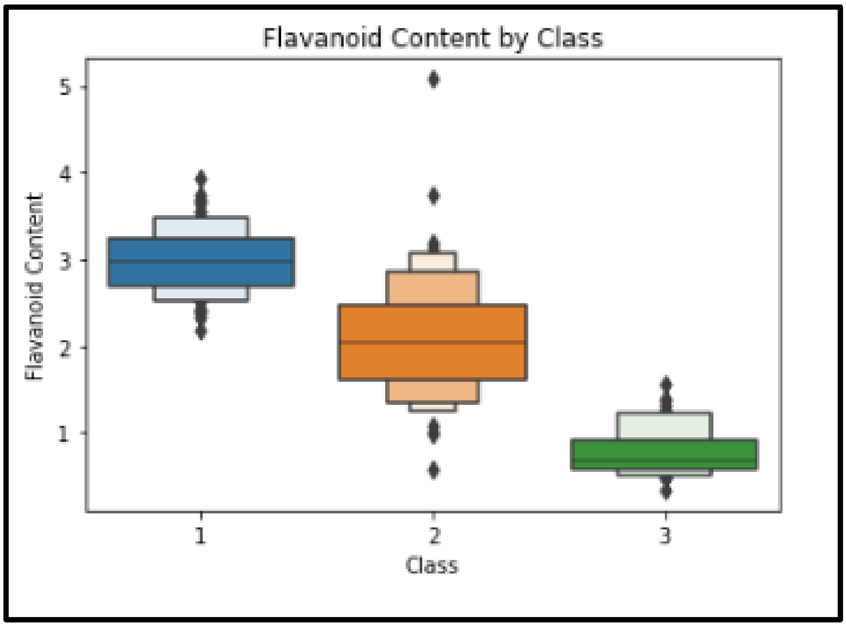


Figure 8: Flavanoid content by the class

(Source: Obtained from the Python Environment)

The above-mentioned figure shows the flavanoid content by the particular class. The specific flavonoid content for the respective classes 1,2 and 3 are represented here.

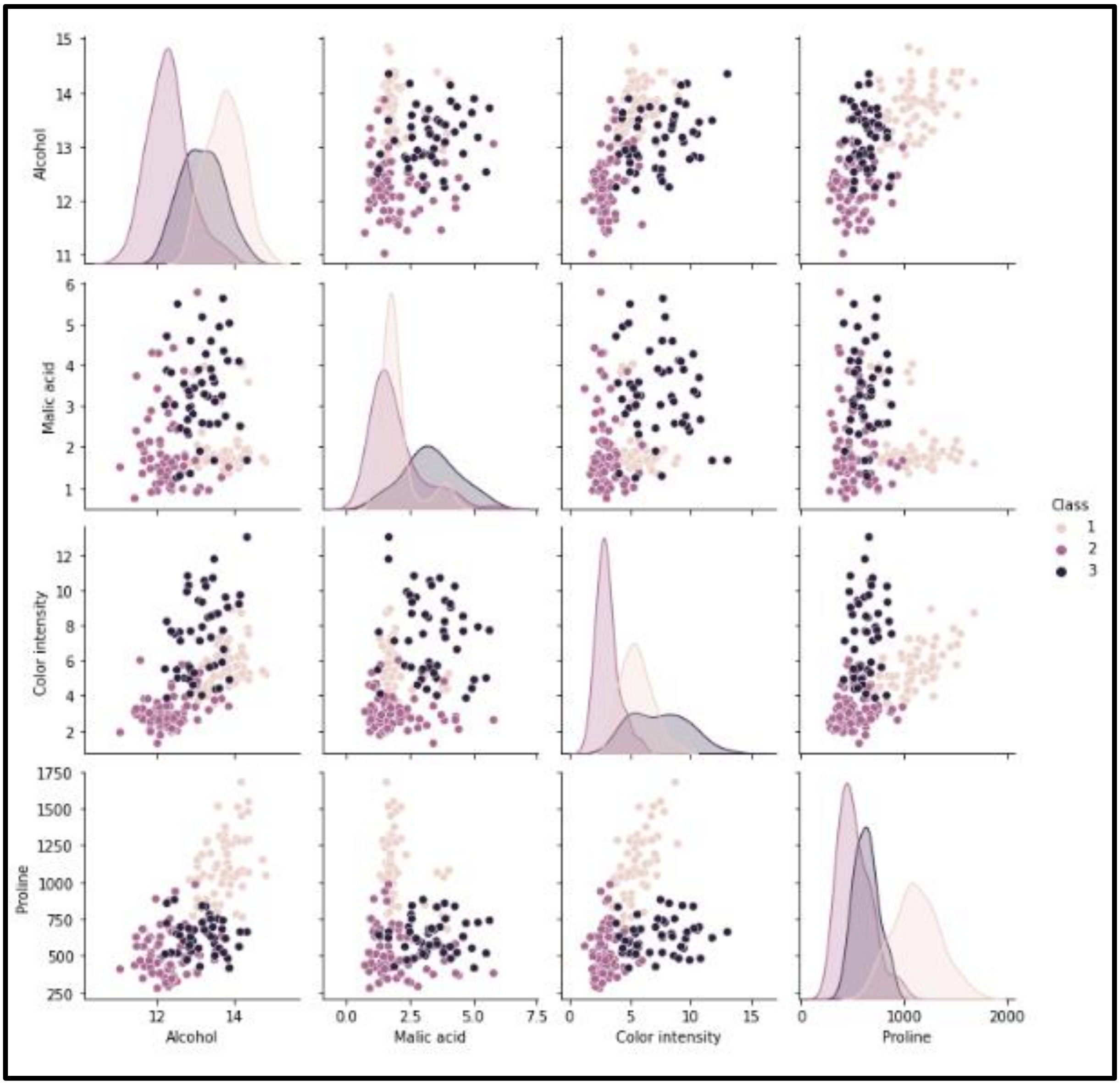


Figure 9: Showing Scatter matrix plot

(Source: Obtained from the Python Environment)

It is the scatter matrix plot for the respective attributes. Here the graphs are plotted for the different classes for different attributes.

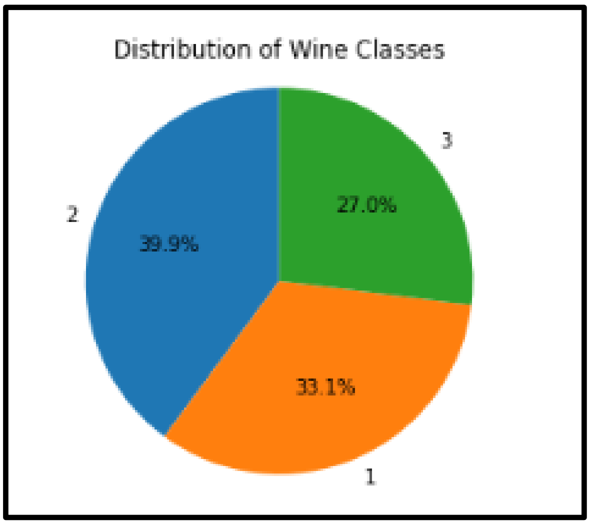
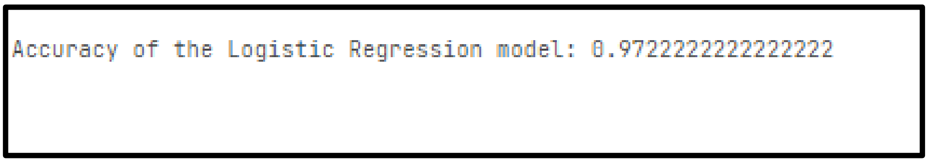


Figure 10: Depicting distribution of the wine classes

(Source: Obtained from the Python Environment)

The distribution of the wine classes is visualised in the particular pie chart. It can be realised from the figure that class 2 has the maximum percentage of wine. The particular distribution of class 2 wine is 39.9%. This offers a precise representation of the following proportion of every wine class. This particular information assesses in realising the class balance of the dataset and can address further analysis along with modelling decisions concerning particular wine classes.

1.1.3 Finalization phase

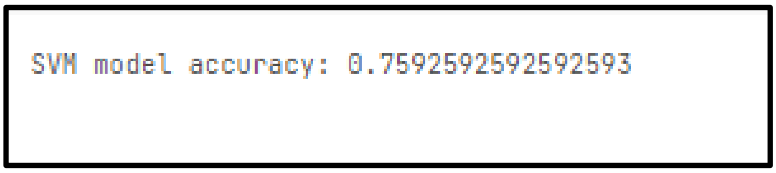


```
Accuracy of the Logistic Regression model: 0.9722222222222222
```

Figure 11: Showing accuracy score of the Logistic Regression

(Source: Obtained from the Python Environment)

The particular accuracy score of the following “Logistic Regression” model is shown here. The particular modules for this model are imported first. In train test splitting this dataset is split into 20% for testing sets and 80% for the training sets. It can be seen here that the accuracy score of this model is about 97%.

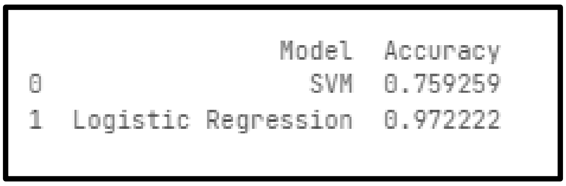


```
SVM model accuracy: 0.7592592592592593
```

Figure 12: Depicting the accuracy score of the SVM model

(Source: Obtained from the Python Environment)

The respective accuracy score of the ***"Support Vector Machines"*** or SVM model is depicted here. The required modules for this model are also imported first. In train test splitting for this model, the dataset is split into 30% for the testing sets and 70% for the training sets. It can be identified from the mentioned figure that the accuracy score fhte SVM model is about the 76%.



```
                 Model  Accuracy
0                  SVM  0.759259
1  Logistic Regression  0.972222
```

Figure 13: Comparison table of the models

(Source: Obtained from the Python Environment)

It is the comparison table of these two models that are performed for the chemical analysis in a better way. The accuracy of the SVM is about 76% and the accuracy of the logistic regression is about 97%. Hence, this can be told that the logistic regression is better for this dataset in performing the chemical analysis using machine learning approaches.

This offers a quantitative approach of the performance of the logistic regression and SVM models. Researchers can evaluate which model is more adequate for the particular dataset by comparing these accuracy scores,. In this particular case, logistic regression outperforms the SVM with a greater accuracy of 97%, signifying its suitability for the chemical analysis tasks utilising machine learning and assessing in the choosing of model for the wine analysis.

Reference

Sivaraman, V., Wu, Y. and Perer, A., 2022, March. Emblaze: Illuminating machine learning representations through interactive comparison of embedding spaces. In 27th International Conference on Intelligent User Interfaces (pp. 418-432). https://dl.acm.org/doi/abs/10.1145/3490099.3511137

Moghimi, A., Pourreza, A., Zuniga-Ramirez, G., Williams, L.E. and Fidelibus, M.W., 2020. A novel machine learning approach to estimate grapevine leaf nitrogen concentration using aerial multispectral imagery. Remote Sensing, 12(21), p.3515. https://www.mdpi.com/870050

Wu, H., Tian, L., Chen, B., Jin, B., Tian, B., Xie, L., Rogers, K.M. and Lin, G., 2019. Verification of imported red wine origin into China using multi isotope and elemental analyses. Food Chemistry, 301, p.125137. https://www.sciencedirect.com/science/article/pii/S0308814619312439

Wu, H., Lin, G., Tian, L., Yan, Z., Yi, B., Bian, X., Jin, B., Xie, L., Zhou, H. and Rogers, K.M., 2021. Origin verification of French red wines using isotope and elemental analyses coupled with chemometrics. Food Chemistry, 339, p.127760. https://www.sciencedirect.com/science/article/pii/S0308814620316228

Astray, G., Mejuto, J.C., Martínez-Martínez, V., Nevares, I., Alamo-Sanza, M. and Simal-Gandara, J., 2019. Prediction models to control aging time in red wine. Molecules, 24(5), p.826. https://www.mdpi.com/417994

Robles, A., Fabjanowicz, M., Chmiel, T. and Płotka-Wasylka, J., 2019. Determination and identification of organic acids in wine samples. Problems and challenges. TrAC Trends in Analytical Chemistry, 120, p.115630. https://www.sciencedirect.com/science/article/pii/S0165993619303498