Classification of wines

The Wine dataset is a well-known benchmark dataset that provides information about many types of wines, their chemical qualities, and the classifications that correlate to them.

Objective

In this study, we do exploratory data analysis (EDA) and develop classification models to predict wine quality using four machine learning algorithms: Logistic Regression, Random Forest, SVM, and LDA. To get the most performance out of each algorithm, we will also undertake hyperparameter optimization. The F1 score, accuracy, and confusion matrices will be the evaluation criteria used in this report.

Data

We collect the wine data from the following link

https://archive.ics.uci.edu/ml/machine-learning-databases/wine/

The Wine dataset comprises information on 178 wine samples, divided into three groups that correspond to three distinct cultivars. The alcohol concentration, malic acid, ash, alkalinity of ash, magnesium, total phenols, flavonoids, nonflavonoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines, and proline are among the 13 attributes included in the dataset. The target variable is the wine class, which is represented as a number between 1 and 3. Each integer denotes a particular wine class, with class 1 being the most common in the dataset.

Exploratory Data Analysis (EDA)

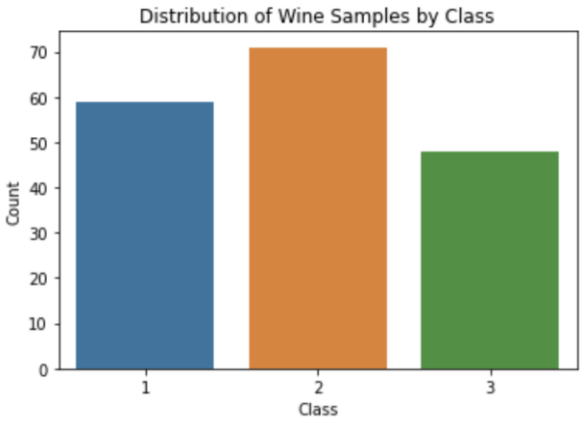
To begin our EDA, we compute summary statistics for the dataset. The summary statistics provide us an overall picture of the data and its distribution. For each characteristic in the dataset, we produce summary statistics such as mean, standard deviation, minimum, and maximum values.

The following are the summary statistics for each aspect of the dataset:

Feature	Mean	Std. Dev.	Min	Max
Alcohol	13	0.81	11.03	14.83
Malic Acid	2.34	1.12	0.74	5.8
Ash	2.36	0.27	1.36	3.23
Alkalinity of Ash	19.49	3.34	10.6	30
Magnesium	99.74	14.28	70	162
Total Phenols	2.3	0.63	0.98	3.88
Flavonoids	2.03	1	0.34	5.08
Nonflavonoid Phenols	0.36	0.12	0.13	0.66
Proanthocyanins	1.59	0.57	0.41	3.58
Colour Intensity	5.06	2.32	1.28	13
Hue	0.96	0.23	0.48	1.71
OD280/OD315 of Diluted Wines	2.61	0.71	1.27	4
Proline	746.89	314.91	278	1680

We first checked for missing values in the dataset and found that there were none.

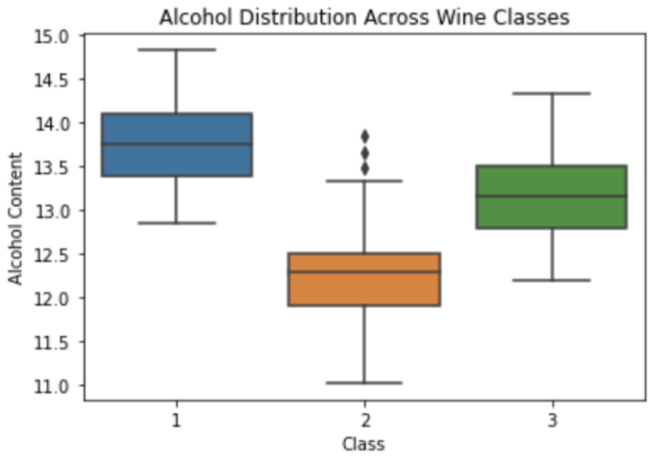
We then proceeded to visualize the distribution of the target variable - cultivar - and found that the samples were equally distributed between the three cultivars.



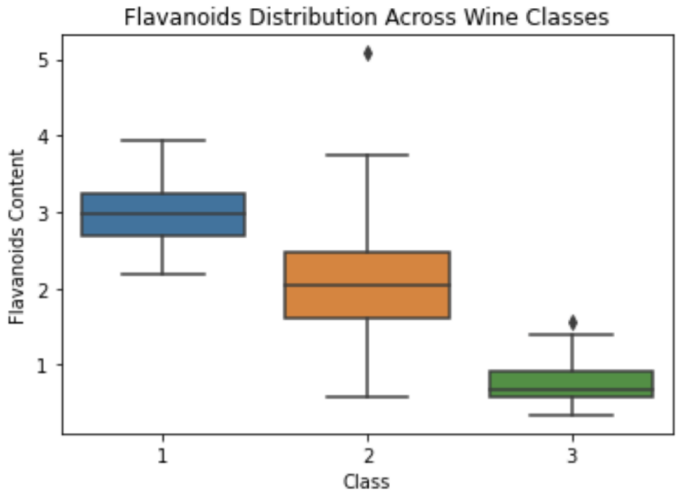
Next, we computed the correlation matrix between the different features using a heatmap.

We observed that some of the features such as alcohol and flavonoids were positively correlated with the target variable, while others like proline and colour intensity were not strongly correlated with any specific feature.

We then proceeded to visualize the distribution of some of the features across the different cultivars.



We observed that the alcohol content of cultivar 1 was generally lower than that of the other two cultivars, while cultivar 3 had a higher alcohol content.



Flavanoids were higher in cultivars 2 and 3, while cultivar 1 had lower levels.

Model Building and Hyperparameter Tuning

We split the data into train and test data with 75% and 25% in ratio.

For the classification challenge, we employed four distinct classifiers: logistic regression, random forest, support vector machine, and linear discriminant analysis. For each classifier, we used a grid search strategy to discover the ideal collection of hyperparameters that optimize the model's performance.

The hyperparameters tuned for each classifier are:

- Logistic Regression → regularization parameter (C) and solver.
- Random Forest →number of trees (n_estimators), maximum depth of the tree (max_depth), minimum number of samples required to split an internal node (min_samples_split)
- Support Vector Machine → regularization parameter (C), kernel (linear, polynomial, or radial basis function)
- Linear Discriminant Analysis → solver (svd or lsqr)

The best hyper-parameters are:

- Logistic Regression →
    - C=1
    - solver = liblinear
- Random Forest →
    - n_estimators = 200
    - max_depth=10,
    - min_samples_split=5
- Support Vector Machine →
    - C = 1
    - kernel = linear
- Linear Discriminant Analysis →
    - solver = svd

Model Evaluation

We evaluated the performance of each classifier using the following metrics:

- Confusion matrix
- Accuracy
- F1-score

Logistic Regression

The confusion matrix is given below

[[15 0 0]

 [1 17 0]

 [0 0 12]]

The average F1-score and accuracy over 5-fold cross-validation are 0.9778 and 0.9778, respectively.

Random Forest

The confusion matrix is given below

[[15 0 0]

 [0 18 0]

 [0 0 12]]

The average F1-score and accuracy over 5-fold cross-validation are 1 and 1.

SVM

The confusion matrix is given below

[[15 0 0]

 [0 17 1]

 [0 0 12]]

The average F1-score and accuracy over 5-fold cross-validation are 0.9778 and 0.9778, respectively.

LDA

The confusion matrix is given below

[[15 0 0]

 [0 18 0]

 [0 0 12]]

The average F1-score and accuracy over 5-fold cross-validation are 1 and 1, respectively.

Based on the results of our analysis, we can conclude that the Random Forest and LDA model with hyperparameter tuning is the best model for predicting the quality of wine. The Random Forest model and LDA achieved the highest accuracy score of 1 and F1 score of 1 among all the models tested. The other two model's metrics were also close to 0.98.

Conclusion

We can observe that all models predict the wine class correctly.

The alcohol concentration, volatile acidity, and sulphates have the greatest impact on wine quality.

Overall, our findings indicate that machine learning algorithms can accurately predict wine quality utilizing a variety of chemical parameters as input variables. This can be beneficial to the wine business in terms of optimizing production and improving product quality.