Case study- Wine(Sapient)

This analysis is done on the wine dataset shared by Sapient. The dataset contains set of observations for two types of wines-red and white. Also, the quality of wine is defined on a scale of 3-9. Some of the features include quantity of Sulfur, chlorides, sugar, citric acid, alcohol etc.

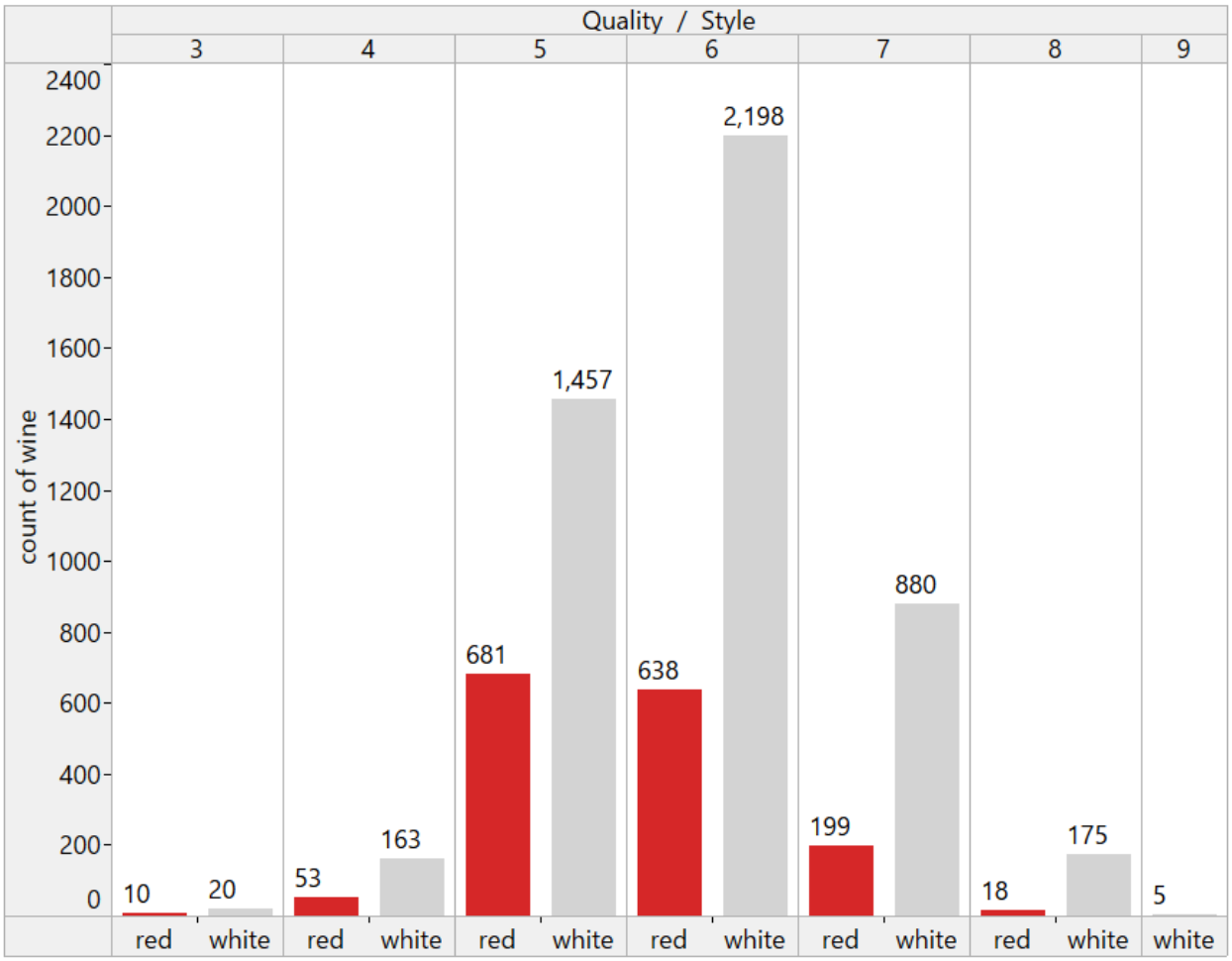
I will perform descriptive and predictive analytics on the data and finally provide recommendations to increase sales and optimize production cost as well as reduce indirect cost thus increasing profit percentage

This report is broken into following components: -

1. *Descriptive analytics*
2. *Model to predict wine color*
3. *Model to predict wine quality*
4. *Recommendations to increase sales*

Descriptive Analytics

Basic Counts: -

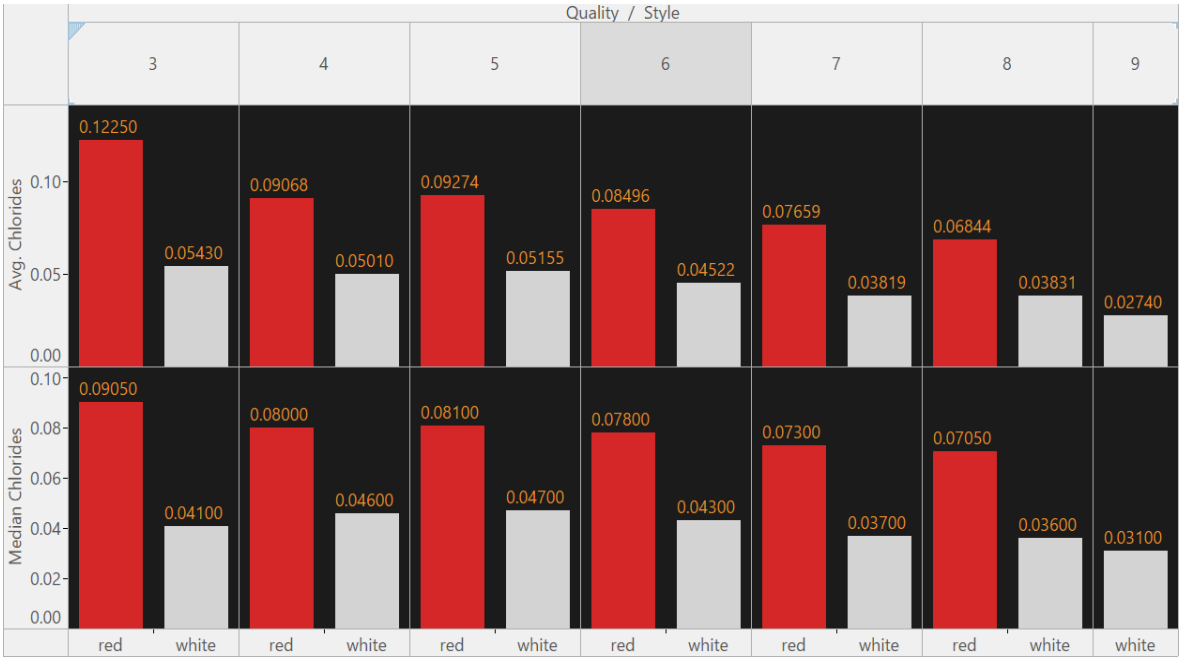


It is clear from above graph that we do have class imbalance in the dataset as white wine has more number of observations for each quality type. For quality 8 and 9 we have very low observations for red wine.

Similarly, for quality 3,4 number of observations for both the wine are extremely low. Most of the data is centered around quality 5 and 6.

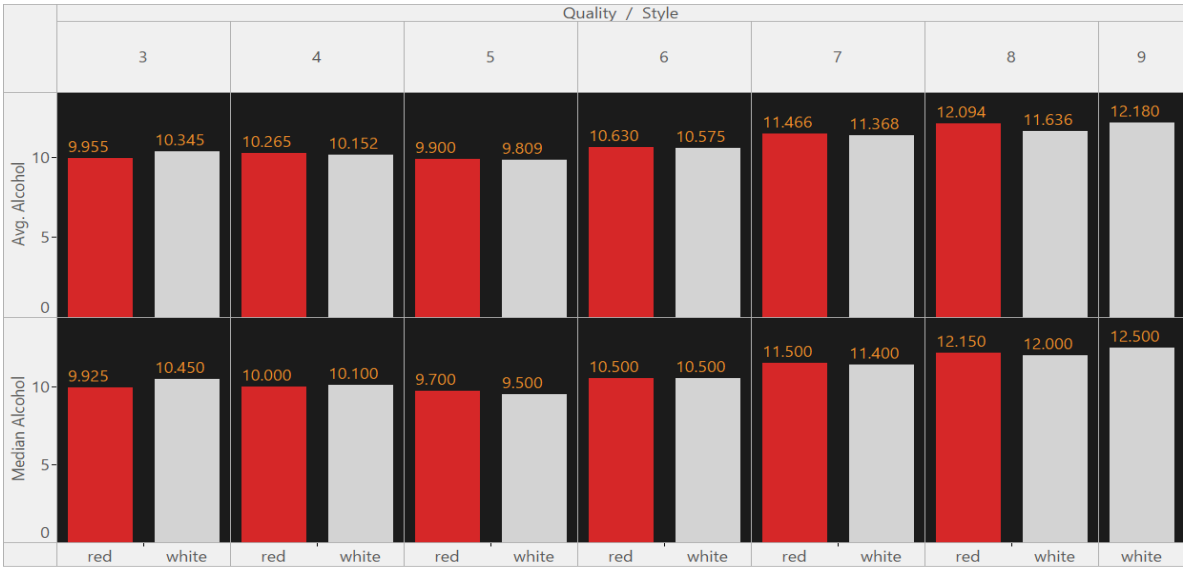
Let us now analyze the various features based on color and quality of wine-

1. Chlorides



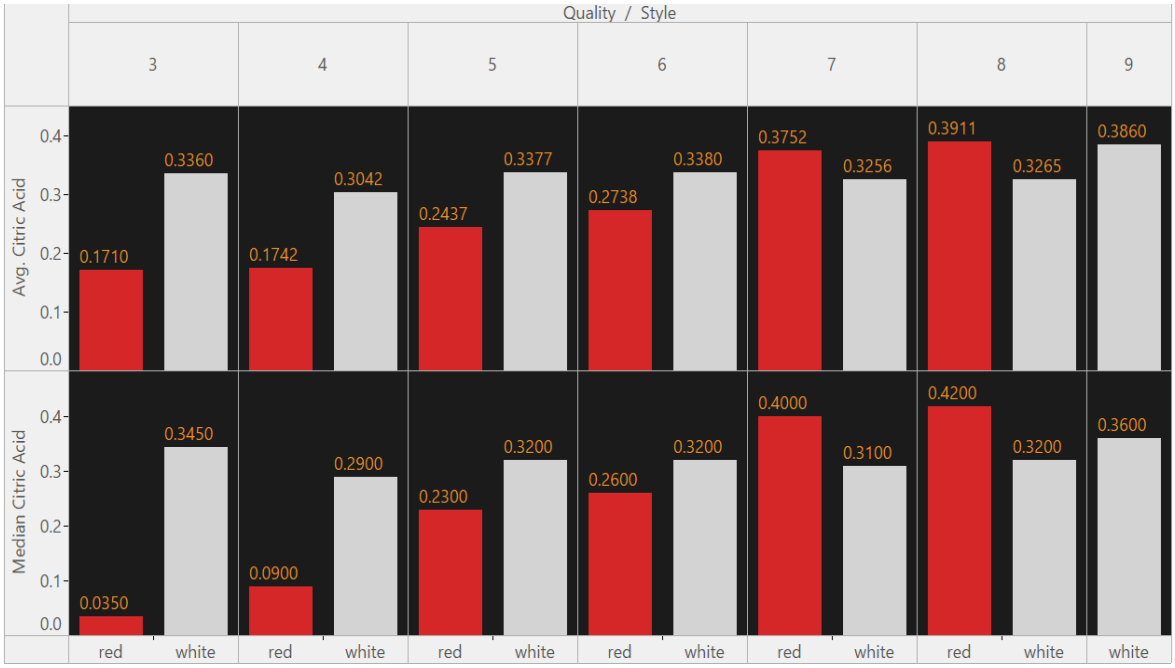
The above graph shows median and average value of chlorides for red and white wine. Clearly, the quantity of chlorides is significantly higher in red wine. This variable could also be a strong predictor for the color of wine. The quantity of chloride decreases with increase in quality for red wine which suggests that poor quality red wine contains more chlorides. However, white wine has same concentration for each quality. Taking both the above observations, it could be inferred that lesser the quantity of chlorides, better is wine.

2. Alcohol



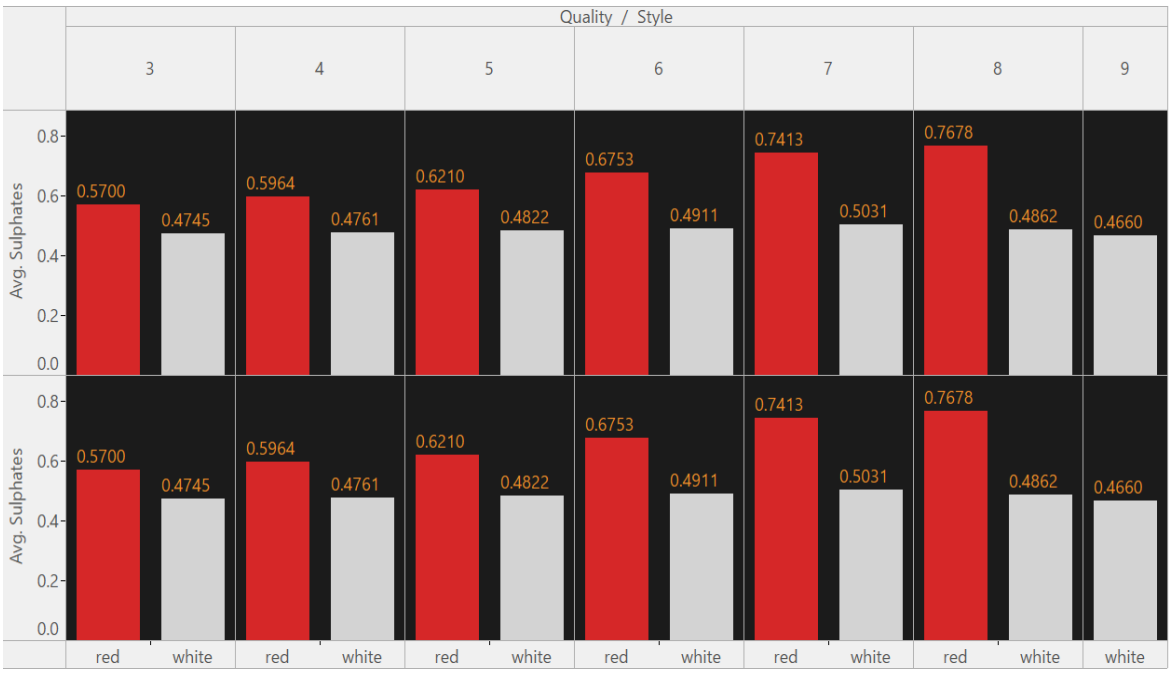
The alcohol quantity is not significantly different for both type of wine. Top quality of red wine has slightly higher alcohol as compared to same quality of white wine.

3. Citric Acid



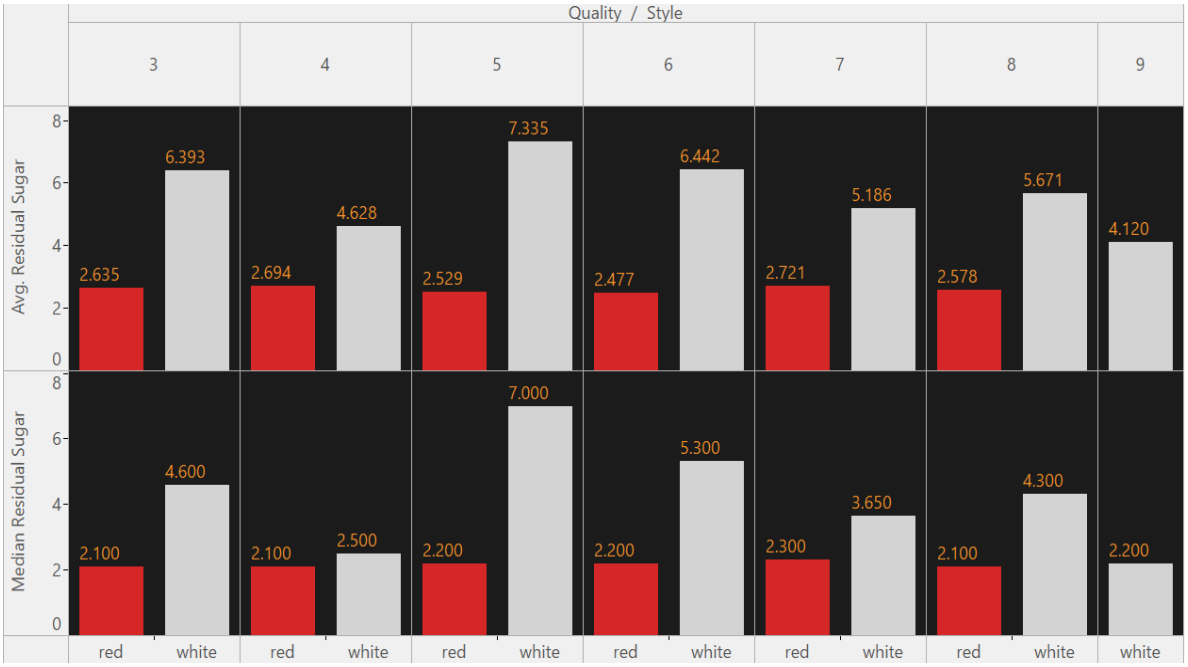
Low Quality white wine contains higher average citric acid as compared to red wine. However, the trend changes for top quality wine. We could see that top-quality red wine contains more average citric acid

4. Sulfur



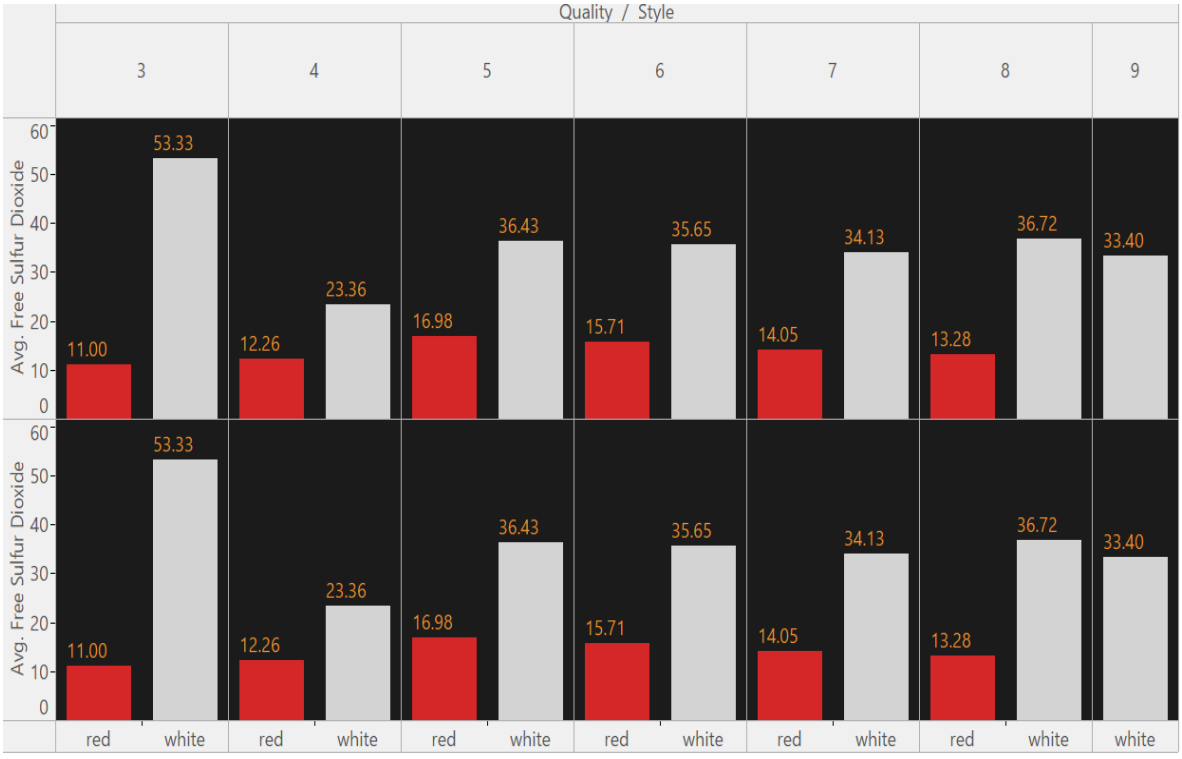
Again, the average Sulphur quantity is higher for red wine across all the qualities. This could be a strong predictor for color of wine.

5. Residual Sugar



The above graph signifies that white wine contains more residual sugar as compared to the red wine for all qualities. If this is related to the sugar content, then white wine is sweeter as compared to red.

6. Free Sulphur dioxide



The average quantity of Free Sulphur dioxide is significantly higher across all qualities for white wine.

Now, I will also perform statistical tests across feature set to determine if the values are statistically significantly different for red and white wine, making sure that assumptions of test are not violated.

Below is an illustration of running Welch two sample t-test on chlorides quantity for two types of wine. The variance of both the dataset that is red and white wine is not similar. This was verified using levenetest that turned out to be significant. Also, to make data normal log transformation was done. Since, variance is not equal we cannot use pooled variance due to which student t-test is not the right choice. The welch t-test is appropriate when variance of datasets are dissimilar. Hence, the summary below illustrates welch t-test on chlorides.

```
        Welch Two Sample t-test

data:  data_red$logchloride and data_white$logchloride
t = -76.594, df = 5058.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.96902 -11.37161
sample estimates:
mean of x mean of y
 12.84282  24.51314
```

In this test we reject null hypothesis in favor of alternate at alpha = 0.05 or 95% confidence interval and conclude that the quantity of chlorides is statistically significantly different for both type of wines

On performing analysis for each feature set, it turns out that results match with descriptive analytics inferences. Most of the features for red and white wine are statistically significantly different as verified by appropriate statistical test.

Predictive model for detecting wine color(Reading-Optional)

Note: -

Red_wine = Class 0

White_wine = Class 1

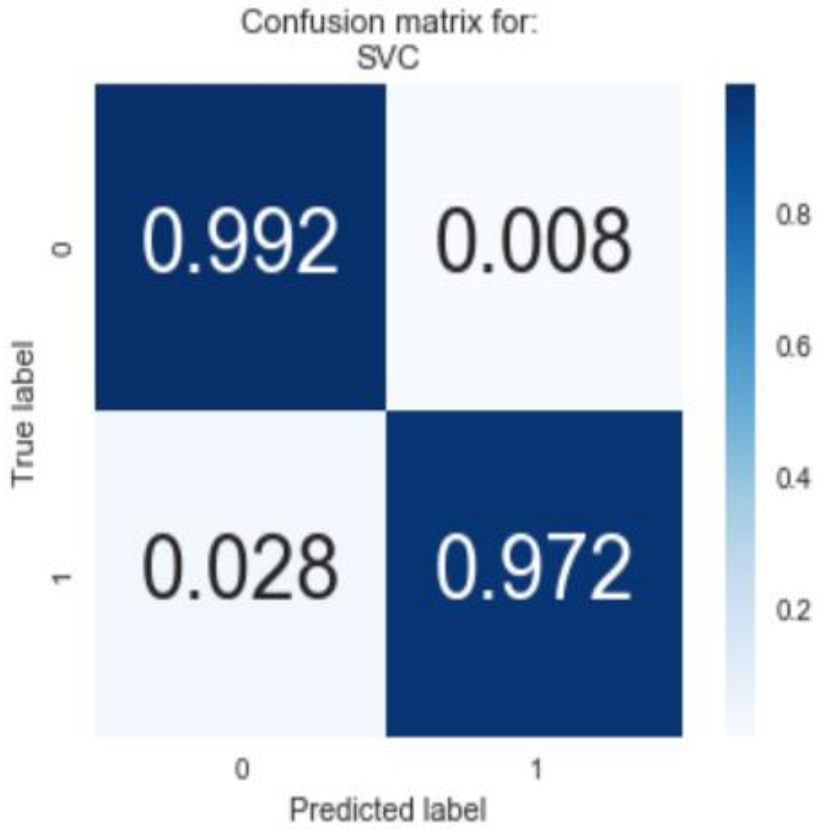
Below is the result of three different classifiers trained on dataset-

1. DecisionTreeClassifier()-

	1%	10%	100%
acc_test	0.897333	0.918667	0.960000
acc_train	0.910000	0.980000	1.000000
f_test	0.884653	0.921599	0.961140
f_train	0.893513	0.984148	1.000000
pred_time	0.000000	0.000000	0.001004
train_time	0.001003	0.001003	0.009024

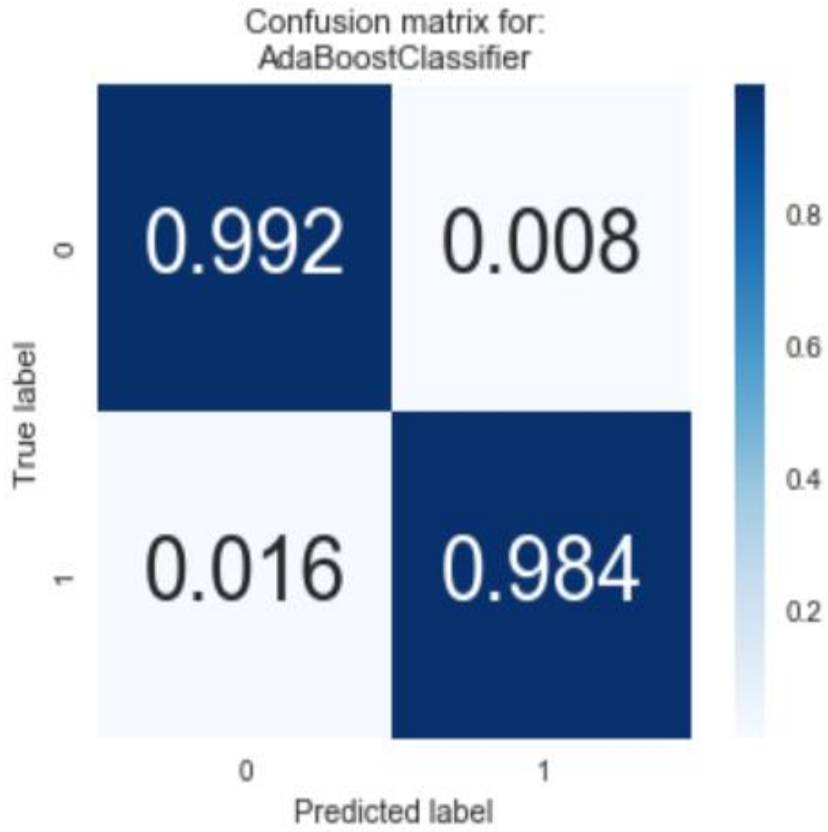
2. SVC

	1%	10%	100%
acc_test	0.514667	0.965333	0.981333
acc_train	0.510000	0.973333	0.986667
f_test	0.569994	0.960754	0.987882
f_train	0.565410	0.966709	0.986928
pred_time	0.001003	0.004010	0.012032
train_time	0.001003	0.001003	0.037098



3. AdaboostClassifier

	1%	10%	100%
acc_test	0.897333	0.949333	0.988000
acc_train	0.910000	0.990000	1.000000
f_test	0.884653	0.950777	0.990615
f_train	0.893513	0.992116	1.000000
pred_time	0.001003	0.011066	0.013036
train_time	0.001003	0.050132	0.118277



Analysis of above classifiers: -

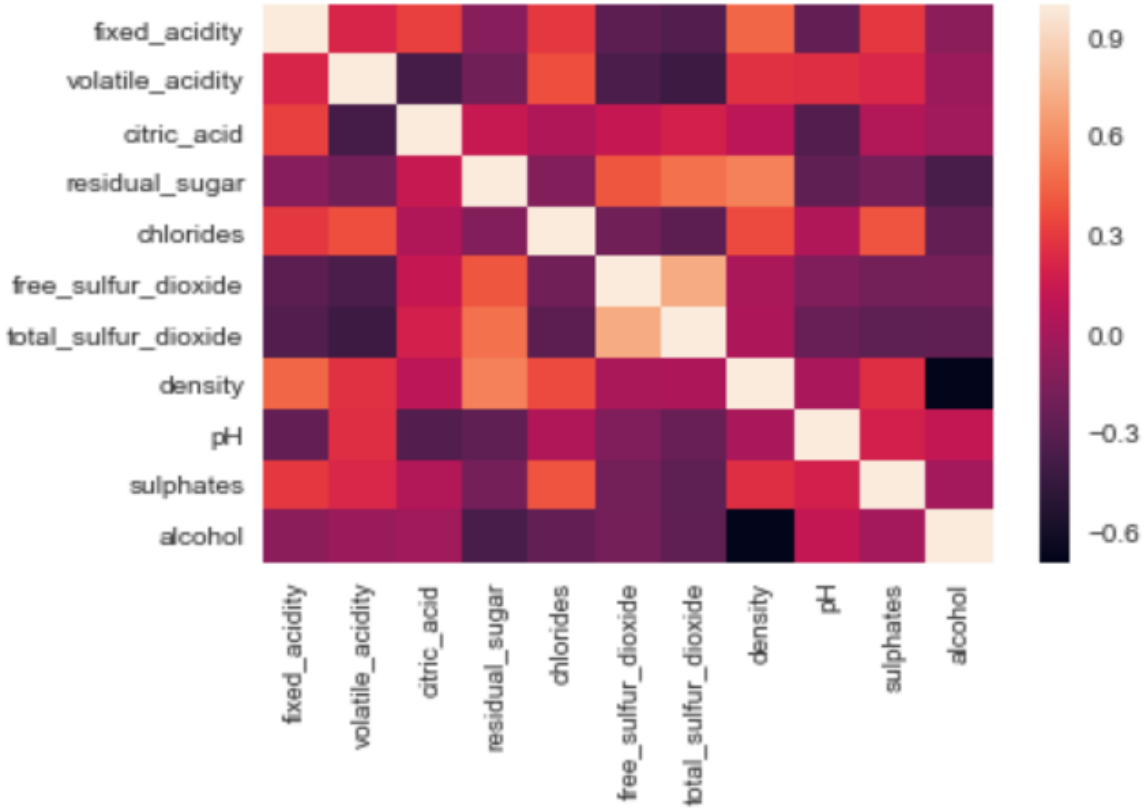
1. All three models do a pretty good job in predicting the wine color achieving accuracy > 95%
2. SVC performs best even on 10% of data giving accuracy greater than 95% and highest f-score
3. In terms of prediction time, Decision Tree performs best compared to other two classifiers
4. When complete data is used ensemble-model gives the best accuracy of 98.8%
5. F-test value is also highest for ensemble-model when 100% data is used

Classification model for detecting wine quality(Reading-Optional)

Data Transformation-

1. Since, the underlying wine dataset has huge class imbalance, data has been transformed into a binary classification problem. Average quality is assumed to be 5.5. Hence, if the quality of wine is above 5.5 it is above average or good quality otherwise below average or bad quality. Keeping this assumption in mind, I have created different classifiers, the detailed analysis of which is given below.
2. There are various methods to handle class imbalance problem such as synthetic oversampling, ensemble classifiers- that use weak classifier and assign greater weight to misclassified data points with each iteration. However, even with all these transformations the maximum accuracy achieved on a multi class classifier was not more than 62%. Finally, the assumption defined above helped to develop a better classifier. But again, the above assumption is completely based on my understanding and it could change according to different problem statements or dataset.

Correlation heatmap



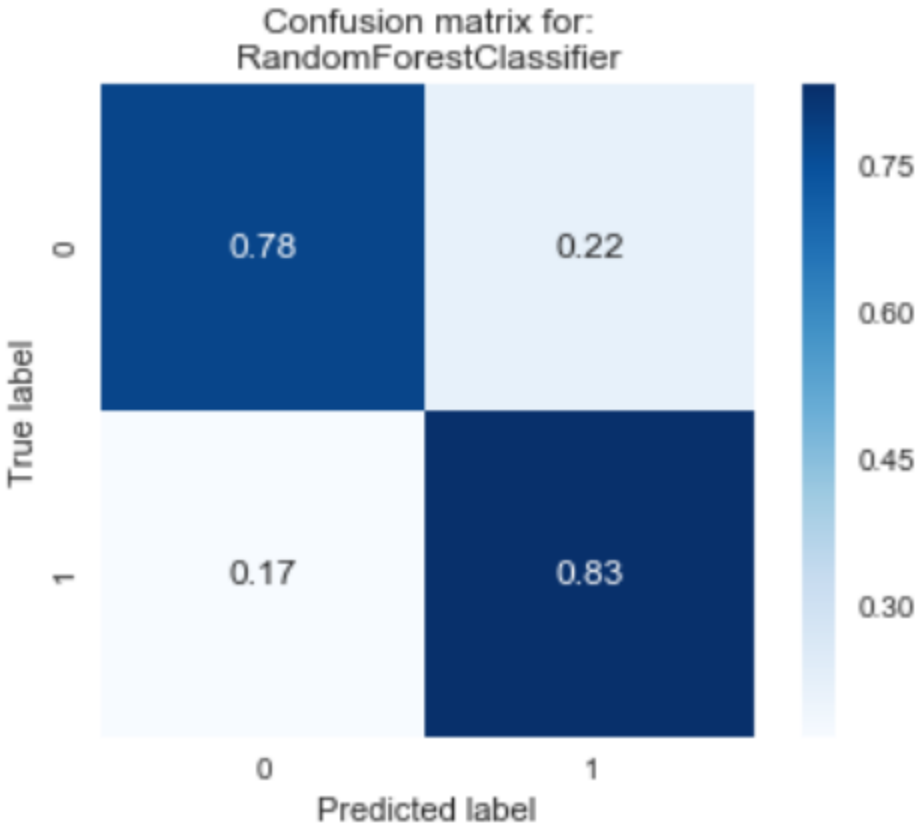
Results of various machine learning algorithms are as follows-

1. RandomForestClassifier()-Ensemble model

```
Training set has 4872 samples.
Testing set has 1625 samples.
```

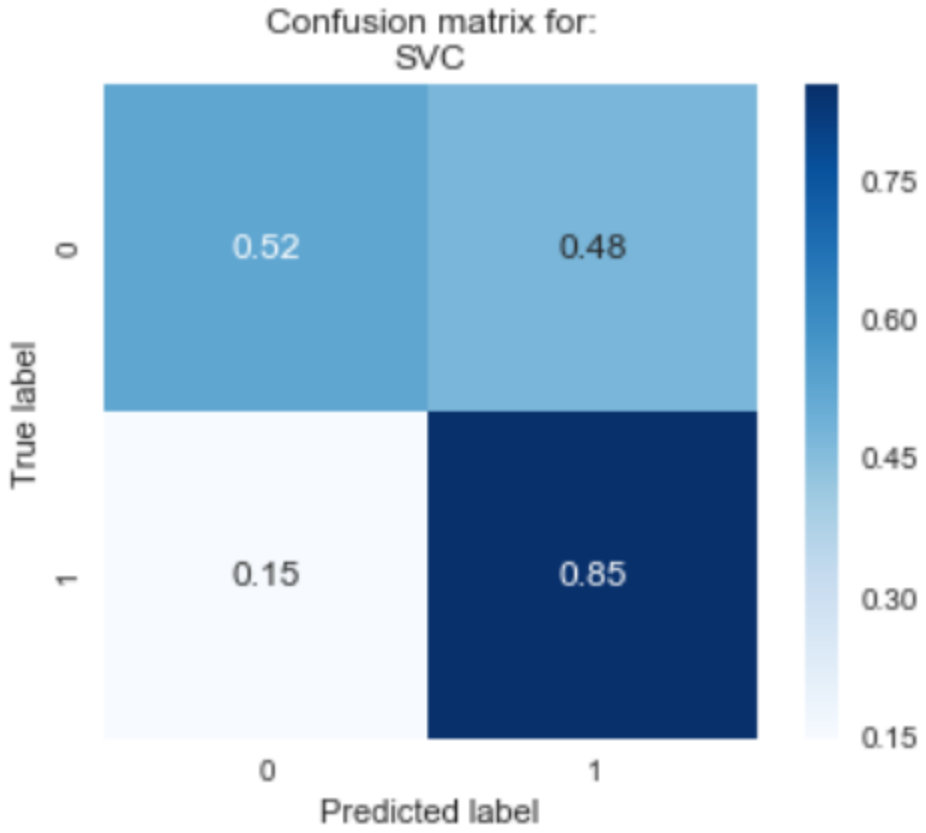
RandomForestClassifier

	1%	10%	100%
acc_test	0.669231	0.711538	0.812308
acc_train	0.753333	0.990000	0.990000
f_test	0.763095	0.784000	0.857683
f_train	0.808642	0.996583	0.993228
pred_time	0.002506	0.004047	0.003540
train_time	0.014540	0.016042	0.109794



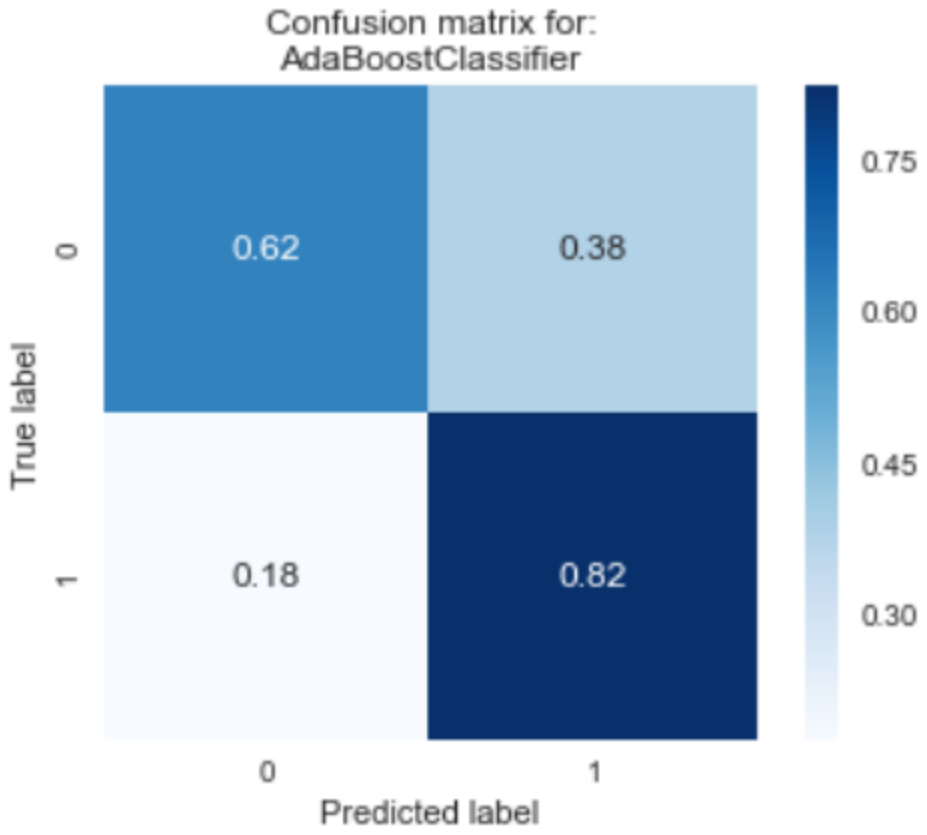
2. SVC

	1%	10%	100%
acc_test	0.629231	0.703077	0.731538
acc_train	0.593333	0.726667	0.760000
f_test	0.679628	0.739437	0.771100
f_train	0.645864	0.735043	0.769231
pred_time	0.001504	0.013536	0.114304
train_time	0.000969	0.006517	0.646721



3. AdaBoostClassifier

	1%	10%	100%
acc_test	0.646154	0.706923	0.747692
acc_train	0.710000	0.870000	0.796667
f_test	0.719195	0.773354	0.793258
f_train	0.746347	0.884615	0.805169
pred_time	0.014039	0.015541	0.014037
train_time	0.043614	0.072192	0.254678



Analysis of above classifiers-

1. It could be inferred that RandomForestClassifier gives best accuracy for our dataset. This also justifies why ensemble models normally have better performance as compared to other machine learning models. Both Adaboost and randomforest are ensemble models however, they differ in terms of implementation. Adaboost could produce a high number of similar classifiers in case of a very strong predictor present in the feature set. However, the implementation of random forest in terms of selecting random sample of features produces distinct classifiers with each iteration providing much better result in this case.
2. However, if we perform hyper parameter tuning of adaboost classifier its accuracy reaches close to that of random forest.

```
Optimized Model
------
Final accuracy score on the testing data: 0.8092
Final F-score on the testing data: 0.8502
```

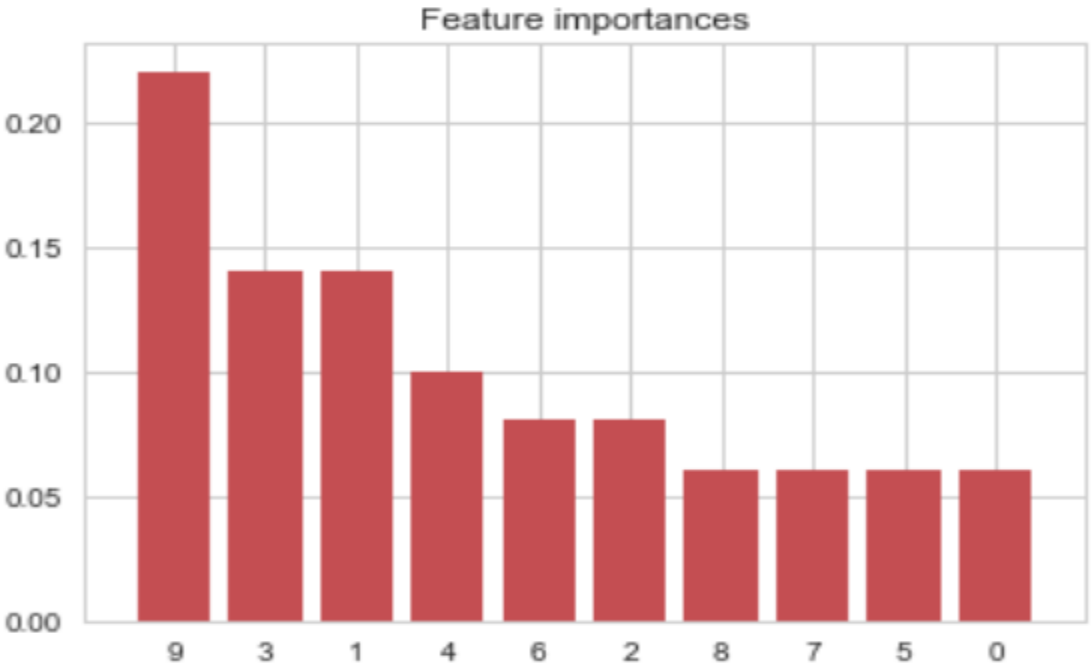
Feature selection and engineering: -

1. The optimal approach towards building any machine learning model is selecting features that explain maximum variance
2. We can use any supervised model that has feature importance method such as adaboost, random forest etc.
3. Feature selection helps to address curse of dimensionality and in building a simple yet powerful generalized model that avoids overfitting. PCA implementation also helps in producing more generalized models with less number of dimensions

4. In tree based models, it is highly important to perform feature engineering as such models are prone to overfitting
5. The below results are for wine quality classifier-

```
['fixed_acidity',
 'volatile_acidity',
 'citric_acid',
 'residual_sugar',
 'chlorides',
 'free_sulfur_dioxide',
 'density',
 'pH',
 'sulphates',
 'alcohol']

Feature ranking:
1. feature 9 (0.220000)
2. feature 3 (0.140000)
3. feature 1 (0.140000)
4. feature 4 (0.100000)
5. feature 6 (0.080000)
6. feature 2 (0.080000)
7. feature 8 (0.060000)
8. feature 7 (0.060000)
9. feature 5 (0.060000)
10. feature 0 (0.060000)
```



Model Insights-

1. Top 4 features are alone responsible for 60% variance in data. Alcohol quantity is best predictor wine quality followed by residual sugar. These two features alone are responsible for 36% variance.
2. Sulphur, SO2 along with some other features explain least variance in the data.

Recommendations to increase wine sales and reduce cost: -

1. It is clear from above predictive and descriptive analysis that few features greatly impact the quality of wine produced. An optimization model could be designed that could be used to play around with quantity of various raw material while keeping cost constant or reducing it depending on the requirement
2. Few constituents such as chlorides and sulfur reduce the quality of wine. Hence, careful quantity of such constituents should be used
3. The above methodologies would help to produce better quality wine. The next step should be collecting customer data from vendors and distributors
4. Using customer data, proper customer segmentation and profiling should be done to offer better recommendations. Also, this segmentation could help to optimize the type of each wine produced based on demographic preferences. This would lead to indirect reduction in cost as inventory would be optimized
5. Using customer data and collaborative filtering techniques, a robust recommendation engine could be designed that would recommend best wine to each customer. This could help in cross sell and up sell
6. Company could also test some promotions and analyze their effect on sales quantity and profit realization. Based on this analysis better promotions and offers could be designed that would lead to better cost realization and increased sales
7. Company could also try to do survey analytics in new cities or states to better understand customer preferences and design better recommendations thus leading to increased sales. An alternative to this could be using census data and designing sales prediction model. This model could then be used to estimate parameters for any new city or area. Lastly, competitor analyses could also prove to be beneficial in increasing sales