Wine Classification

Damir Zunic

Abstract

- We analyzed the physicochemical attributes of wine and their relationships and significance with wine quality and types classifications. Two datasets (one for red and one for white wines) were downloaded from UCI Machine Learning Repository, combined and prepared for Exploratory Data Analysis.
- We followed standard machine learning and data mining workflow for data analysis and for predicting a) the type of wine (red or white) and b) the quality of wine (low, medium, high).
- For Exploratory Data Analysis, description statistics, correlations and Seaborn visualization (several plot types) were used.
- We have used cross-validation pipeline that included standardization and classifiers. Logistic Regresion was used for wine type prediction while Decision Tree, Random Forest and Support Vector Machine for quality prediction.
- Heatmaps of normalized confusion matrices were used for better visual performance evaluation.
- In the first task (type of wine classification) a simple model was able to obtain an excellent result with 99.49% accuracy.
- In the second challenge, wine quality classification (low, medium, high), simple models could not obtain better results due to variation in data, especially of high quality wines. The best accuracy, 81.13%, was obtained with Random Forest Classifier.

Motivation

- Wine is an alcoholic beverage made from fermented grapes. It is a seemingly simple beverage that becomes more complex the more you study it. The good thing is, it doesn’t matter how much you know, nearly everyone can appreciate wine.
- It will definitely be interesting to analyze the physicochemical***** attributes of wine and understand their relationships and significance with wine quality and types classifications.
- We will follow standard machine learning and data mining workflow to try to predict the quality of a wine (low, medium, high) and its type (red, white) from the physicochemical attributes
- Such insight could be useful to support the oenologist****** wine tasting evaluations and help winemakers to look for wines of better qualities and improve wine production. Wine stores and large distributors could qualify new wines, even before acquiring them. That way they could better evaluate their purchase cost and their opportunity to sell.

* *<u>physicochemical</u> - relating to both physical and chemical properties*

****** *<u>oenology</u> - is the science and study of wine and winemaking; distinct from viticulture, the agricultural endeavors of vine-growing and of grape-harvesting.*

Dataset(s)

I worked with the “Quality Wine Data set” from UCI ML repository. Two datasets are included, related to red and white wine varieties of the Portuguese "Vinho Verde" wine. The datasets can be viewed as classification or regression tasks. The classes are ordered but not balanced. There are much more normal wines than excellent or poor ones.

Wine Quality Data Sets

Source: UCI Machine Learning Repository - https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Data sets:

- winequality-red.csv , 1599 x 12
- winequality-white.csv, 4898 x 12

Input variables based on physicochemical tests (1 – 11); Output variable based on sensory data (12)		
1. fixed acidity	5. chlorides	9. pH
2. volatile acidity	6. free sulfur dioxide	10. sulphates
3. citric acid	7. total sulfur dioxide	11. alcohol
4. residual sugar	8. density	12. quality (target variable)

Datasets: Description of attributes

1. **fixed acidity:** most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2. **volatile acidity:** the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. **citric acid:** found in small quantities, citric acid can add 'freshness' and flavor to wines
4. **residual sugar:** the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
5. **chlorides:** the amount of salt in the wine
6. **free sulfur dioxide:** the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. **total sulfur dioxide:** amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. **density:** the density of water is close to that of water depending on the percent alcohol and sugar content
9. **pH:** describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. **sulphates:** a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
11. **alcohol:** the percent alcohol content of the wine
12. **quality** (score between 0 and 10): Output variable (based on sensory data)

Data Preparation and Cleaning

- Loaded both data sets and checked them for nulls. Both data sets were clean, no noise nor missing data. They had the same features.
- In each dataframe created a new column **“type”** with wine type labels, **“red”** or **“white”**
- Merged both dataframes and named the new one **“wines”**
- Grouped the records into three buckets, based on the quality ratings, and created the new column **“quality class”** :

  low for quality <= 5; **medium** for quality 6 and 7; **high** for quality > 7
- Reshuffled the records to randomize data points
- Created the column **“color”** with encoded labels (red : 0, white : 1), this is the target for the 1st research question
- Converted non-numeric **“quality class”** labels to numeric labels, according to the dictionary's mapping (low : 0, medium : 1, high : 2). This series “y_qclass” is the target for the 2nd research question
- There was no problems with datasets

Research Question(s)

We will combine both downloaded datasets and analyze data for both, red and white wines together.

There are two main questions that we want to answer:

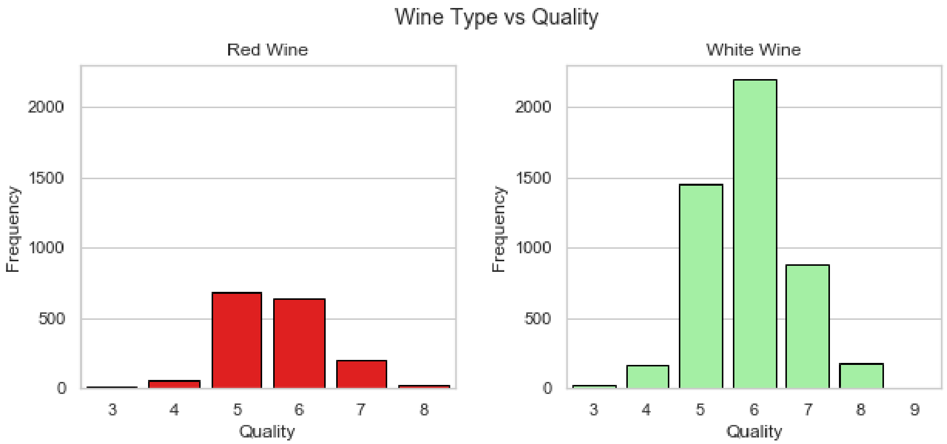
1. Can we predict the type of a wine (red or white) from the physicochemical attributes in the data set? This prediction has only educational foundation.
2. Can we predict the quality of a wine from the physicochemical attributes in the data set? This prediction could have some practical applications.

We will use one or more Machine Learning models for each question.

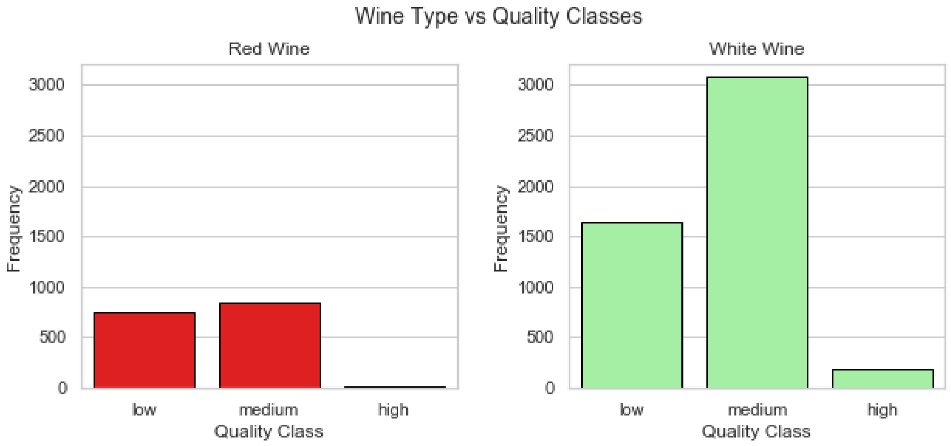
Methods

- For **exploratory data analysis** I used descriptive statistics, correlations and Seaborn visualization (pairplots, box plots, violin plots, lmplots, heatmaps …).
- **Predicting wine type**: this is a binary classification and I used Logistic Regression
- For **predicting quality class of wines** I used Decision Tree Classifier, Random Forest Classifier and Support Vesctor Machine Classifier
- I used standardization, hyperparameters tuning and cross-validation
- To perform cross-validation I used GridSearchCV and included modeling pipeline and tunable hyperparameters inside a cross-validation loop. The modeling pipeline first transforms data using StandardScaler() and then fits a model using one of above mentioned classifiers
- Having data preprocessing included inside the cross-validation loop, prevents accidentally tainting training sets with influential data from the test data.
- To evaluate classification results I used heatmap of confusion matrix. I had to perform normalization by class support size to have a more visual interpretation of which class is being misclassified.

Exploratory Data Analysis: Data Distributions



- Distributions of the data by wine types and quality ratings are shown on the left.
- We can see that they are normally distributed. Most of the wines are rated 5 to 7, while very few wines are rated “very good” (8 – 9) and “very poor” (3 – 4).



- Distributions of the data by wine types and quality classes are shown on the right.
- It confirms the imbalance between classes, especially with few cases in high quality.

Exploratory Data Analysis: Descriptive Statistics - 1

Features	Red Wine								White Wine							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
fixed acidity	1599	8.32	1.74	4.60	7.10	7.90	9.20	15.90	4898	6.85	0.84	3.80	6.30	6.80	7.30	14.20
volatile acidity	1599	0.53	0.18	0.12	0.39	0.52	0.64	1.58	4898	0.28	0.10	0.08	0.21	0.26	0.32	1.10
citric acid	1599	0.27	0.19	0.00	0.09	0.26	0.42	1.00	4898	0.33	0.12	0.00	0.27	0.32	0.39	1.66
residual sugar	1599	2.54	1.41	0.90	1.90	2.20	2.60	15.50	4898	6.39	5.07	0.60	1.70	5.20	9.90	65.80
chlorides	1599	0.09	0.05	0.01	0.07	0.08	0.09	0.61	4898	0.05	0.02	0.01	0.04	0.04	0.05	0.35
free sulfur dioxide	1599	15.87	10.46	1.00	7.00	14.00	21.00	72.00	4898	35.31	17.01	2.00	23.00	34.00	46.00	289.00
total sulfur dioxide	1599	46.47	32.90	6.00	22.00	38.00	62.00	289.00	4898	138.36	42.50	9.00	108.00	134.00	167.00	440.00
density	1599	1.00	0.00	0.99	1.00	1.00	1.00	1.00	4898	0.99	0.00	0.99	0.99	0.99	1.00	1.04
pH	1599	3.31	0.15	2.74	3.21	3.31	3.40	4.01	4898	3.19	0.15	2.72	3.09	3.18	3.28	3.82
sulphates	1599	0.66	0.17	0.33	0.55	0.62	0.73	2.00	4898	0.49	0.11	0.22	0.41	0.47	0.55	1.08
alcohol	1599	10.42	1.07	8.40	9.50	10.20	11.10	14.90	4898	10.51	1.23	8.00	9.50	10.40	11.40	14.20
quality	1599	5.64	0.81	3.00	5.00	6.00	6.00	8.00	4898	5.88	0.89	3.00	5.00	6.00	6.00	9.00

The highlights from descriptive statistics of wine types:

- Mean total sulfur dioxide and residual sugar content seems to be much higher in white wines that in red wines
- Citric acid is more present in white wines, while fixed acidity, volatile acidity and sulphates are more present in red wine
- Red wines have double concentration of chlorides then white wines

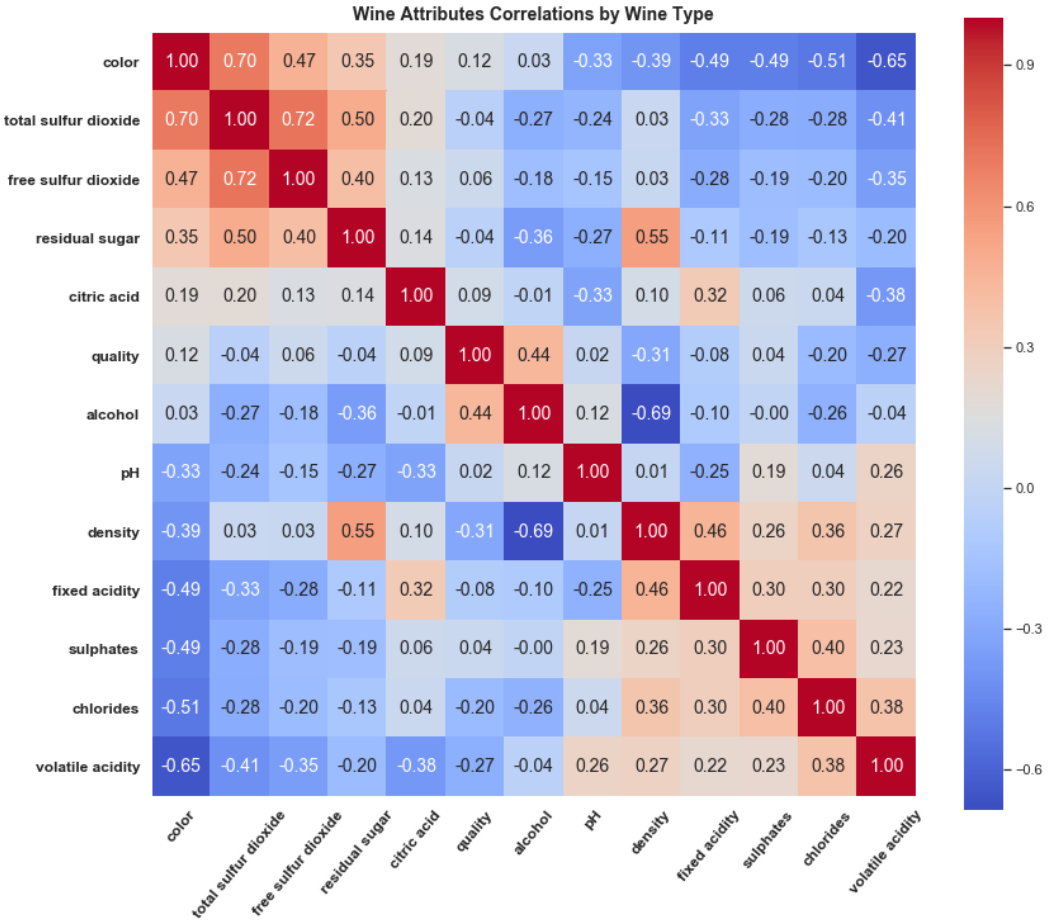
Exploratory Data Analysis: Descriptive Statistics - 2

		fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
Low Quality Wine	**count**	2384	2384	2384	2384	2384	2384	2384	2384	2384	2384	2384
	mean	7.33	0.40	0.30	5.65	0.06	29.48	119.28	1.00	3.21	0.52	9.87
	std	1.27	0.19	0.16	4.92	0.04	19.84	61.89	0.00	0.16	0.14	0.84
	min	4.20	0.10	0.00	0.60	0.01	2.00	6.00	0.99	2.74	0.25	8.00
	25%	6.50	0.26	0.21	1.80	0.04	14.00	68.00	0.99	3.11	0.44	9.30
	50%	7.10	0.34	0.30	2.90	0.05	26.00	124.00	1.00	3.20	0.50	9.60
	75%	7.80	0.50	0.40	8.52	0.08	42.00	167.00	1.00	3.31	0.58	10.40
	max	15.90	1.58	1.00	23.50	0.61	289.00	440.00	1.00	3.90	2.00	14.90
Medium Quality Wine	**count**	3915	3915	3915	3915	3915	3915	3915	3915	3915	3915	3915
	mean	7.16	0.31	0.33	5.32	0.05	30.96	113.51	0.99	3.22	0.54	10.81
	std	1.31	0.14	0.13	4.68	0.03	16.33	53.57	0.00	0.16	0.15	1.20
	min	3.80	0.08	0.00	0.70	0.01	1.00	6.00	0.99	2.72	0.22	8.40
	25%	6.40	0.21	0.26	1.80	0.04	19.00	81.00	0.99	3.11	0.43	9.80
	50%	6.90	0.27	0.31	3.00	0.04	29.00	116.00	0.99	3.21	0.51	10.80
	75%	7.60	0.36	0.39	7.90	0.06	41.00	150.00	1.00	3.33	0.61	11.70
	max	15.60	1.04	1.66	65.80	0.42	112.00	294.00	1.04	4.01	1.95	14.20
High Quality Wine	**count**	198	198	198	198	198	198	198	198	198	198	198
	mean	6.85	0.29	0.33	5.35	0.04	34.51	117.48	0.99	3.23	0.51	11.69
	std	1.14	0.12	0.10	4.16	0.02	17.08	41.69	0.00	0.16	0.16	1.27
	min	3.90	0.12	0.03	0.80	0.01	3.00	12.00	0.99	2.88	0.25	8.50
	25%	6.20	0.21	0.28	2.00	0.03	24.00	96.00	0.99	3.13	0.38	11.00
	50%	6.80	0.28	0.32	4.05	0.04	34.00	118.50	0.99	3.23	0.48	12.00
	75%	7.30	0.35	0.37	7.57	0.04	43.00	145.00	0.99	3.33	0.60	12.60
	max	12.60	0.85	0.74	14.80	0.12	105.00	212.50	1.00	3.72	1.10	14.00

The highlights from descriptive statistics of wine quality classes:

- Alcohol concentration increases with quality of wines.
- Higher quality wines have less volatile acidity and the chlorides.
- Fixed acidity is lower with higher quality wines.
- The free sulfur dioxide is higher with high quality wines.

Exploratory Data Analysis: Correlations - 1



Wine Attributes Correlations by Wine Type

	color	total sulfur dioxide	free sulfur dioxide	residual sugar	citric acid	quality	alcohol	pH	density	fixed acidity	sulphates	chlorides	volatile acidity
color	1.00	0.70	0.47	0.35	0.19	0.12	0.03	-0.33	-0.39	-0.49	-0.49	-0.51	-0.65
total sulfur dioxide	0.70	1.00	0.72	0.50	0.20	-0.04	-0.27	-0.24	0.03	-0.33	-0.28	-0.28	-0.41
free sulfur dioxide	0.47	0.72	1.00	0.40	0.13	0.06	-0.18	-0.15	0.03	-0.28	-0.19	-0.20	-0.35
residual sugar	0.35	0.50	0.40	1.00	0.14	-0.04	-0.36	-0.27	0.55	-0.11	-0.19	-0.13	-0.20
citric acid	0.19	0.20	0.13	0.14	1.00	0.09	-0.01	-0.33	0.10	0.32	0.06	0.04	-0.38
quality	0.12	-0.04	0.06	-0.04	0.09	1.00	0.44	0.02	-0.31	-0.08	0.04	-0.20	-0.27
alcohol	0.03	-0.27	-0.18	-0.36	-0.01	0.44	1.00	0.12	-0.69	-0.10	-0.00	-0.26	-0.04
pH	-0.33	-0.24	-0.15	-0.27	-0.33	0.02	0.12	1.00	0.01	-0.25	0.19	0.04	0.26
density	-0.39	0.03	0.03	0.55	0.10	-0.31	-0.69	0.01	1.00	0.46	0.26	0.36	0.27
fixed acidity	-0.49	-0.33	-0.28	-0.11	0.32	-0.08	-0.10	-0.25	0.46	1.00	0.30	0.30	0.22
sulphates	-0.49	-0.28	-0.19	-0.19	0.06	0.04	-0.00	0.19	0.26	0.30	1.00	0.40	0.23
chlorides	-0.51	-0.28	-0.20	-0.13	0.04	-0.20	-0.26	0.04	0.36	0.30	0.40	1.00	0.38
volatile acidity	-0.65	-0.41	-0.35	-0.20	-0.38	-0.27	-0.04	0.26	0.27	0.22	0.23	0.38	1.00

Most of correlations are week but there are some exceptions:

- **Total (0.70)** and **free (0.47)sulfur dioxides** have the highest correlation with white wines.
- **Free sulfur dioxide** is a part of **total sulfur dioxide** and thus they have the highest correlation (0.72). That represents a collinearity that could be a problem for some models and free sulfur dioxide might need to be dropped later.

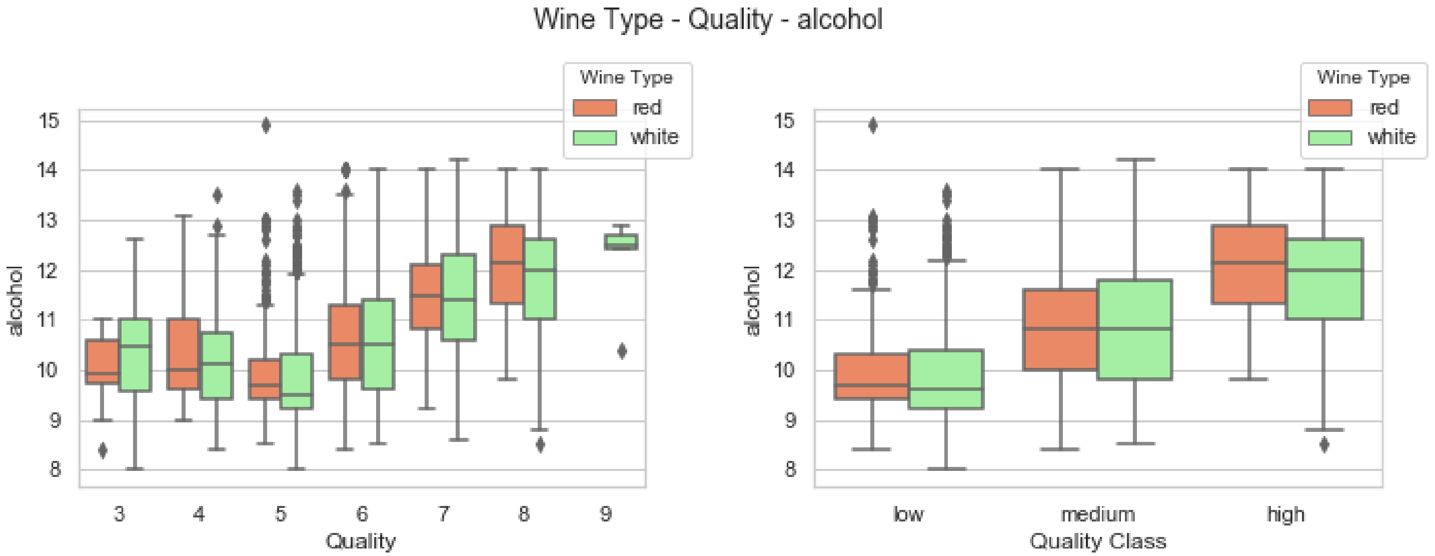
(continues in the next slide)

Exploratory Data Analysis: Correlations - 2

(continued from the previous slide)

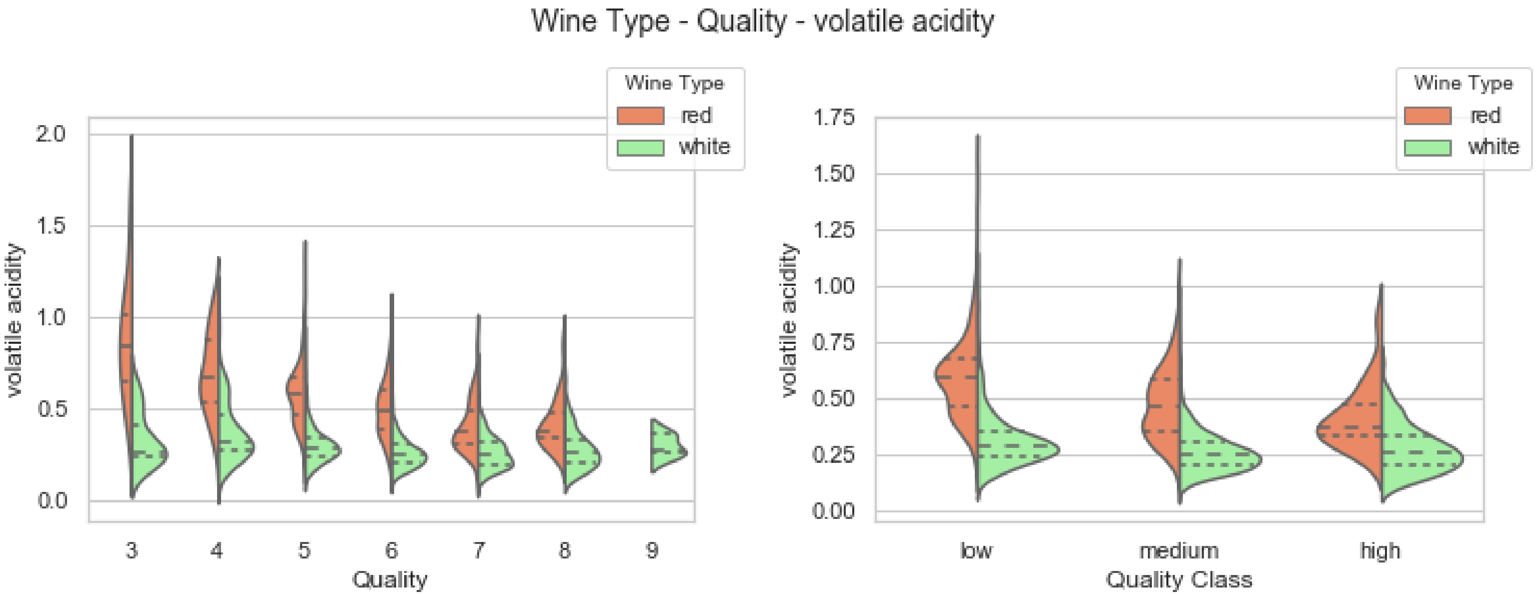
- The **volatile acid** (-0.65) and **chlorides** (-0.51) have negative correlation with **color**. This indicates a tendency to red wines classification.
- The **residual sugar** has 0.50 relation with **total sulfur dioxide** and 0.40 with **free sulfur dioxide**. This is an indication that more sulfur dioxide is added to wines with higher sugar content to prevent secondary fermentation of remaining sugar.
- **Density** has a relatively high negative correlation to **alcohol** (-0.69). This is confirmed by the decreasing linear trend from left to right. **Density** has also relatively high positive correlation to **residual sugar** (0.55), which is reinforced by two white wine outliners.
- **Sulphates**, **chlorides**, **fixed acidity** and **volatile acidity** in red wine seem to be higher than in white wine.
- **Residual sugar**, **total sulfur dioxide** and **citric acid** seem to be higher in white wines as compared to red wine.

Exploratory Data Analysis: Alcohol – Quality - Type



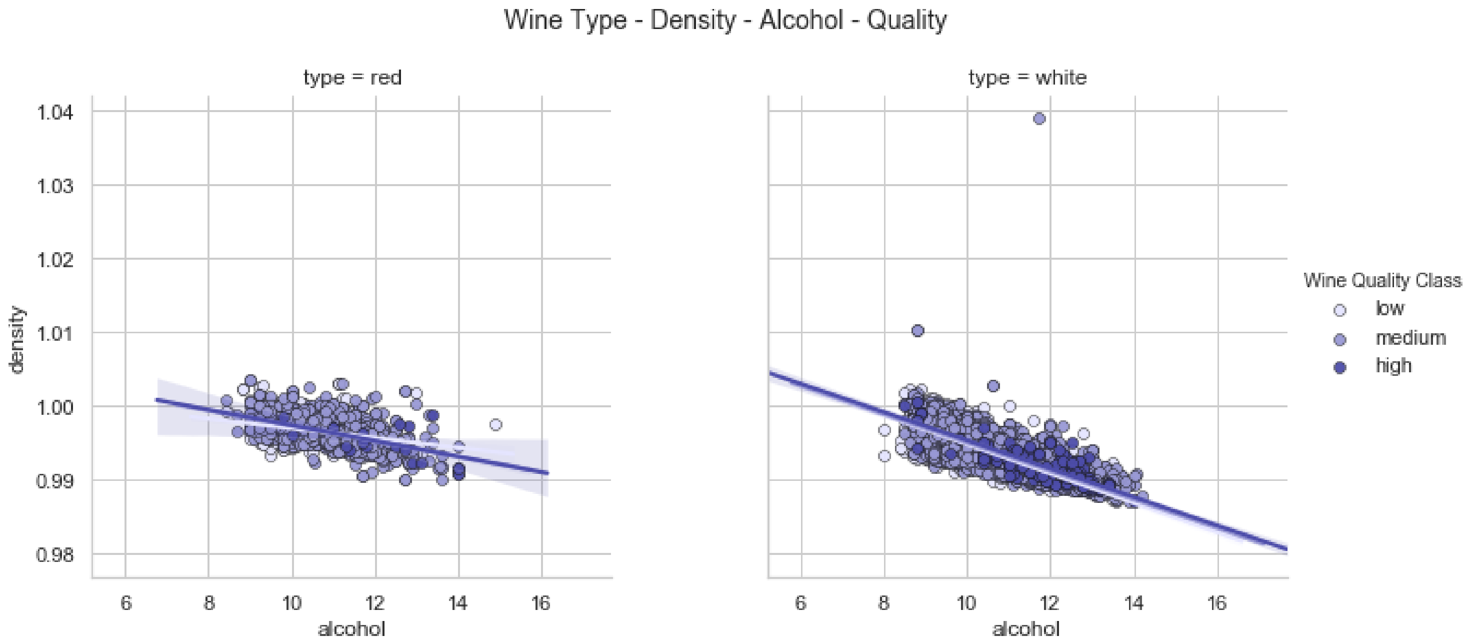
- This plot confirms what was stated before: alcohol concentration increases with the quality of wines.
- There is no big difference in alcohol concentration between red and white wines in the same quality class.

Exploratory Data Analysis: Volatile Acidity – Quality - Type



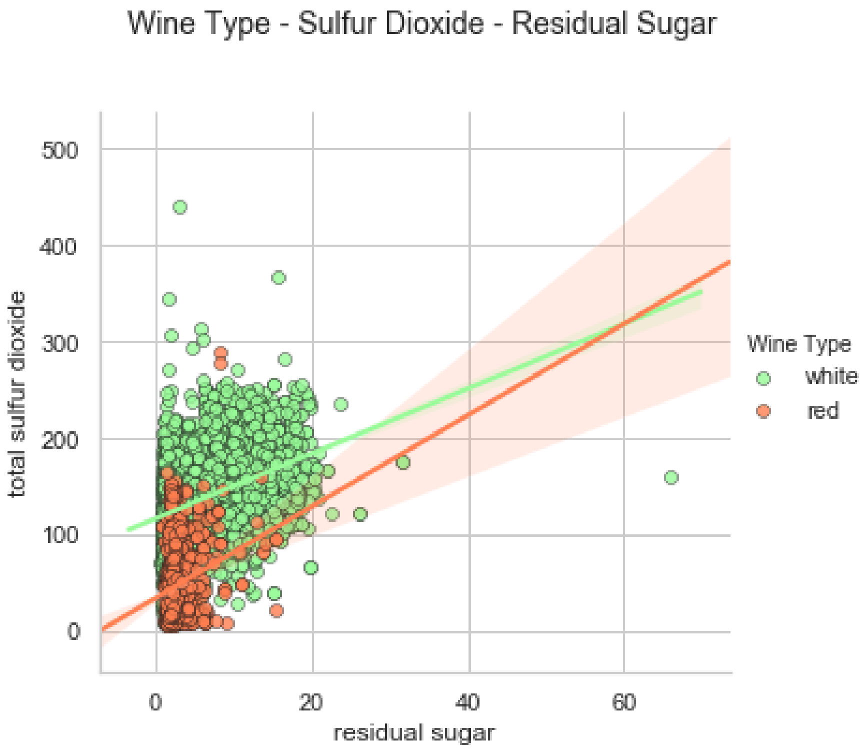
- Volatile acidity is more present in red wines than in white wines
- Higher quality wines have less volatile acidity

Exploratory Data Analysis: Density - Alcohol



- This plot confirms correlations results: density has a relatively high negative correlation to alcohol (linear trend is decreasing from left to right).

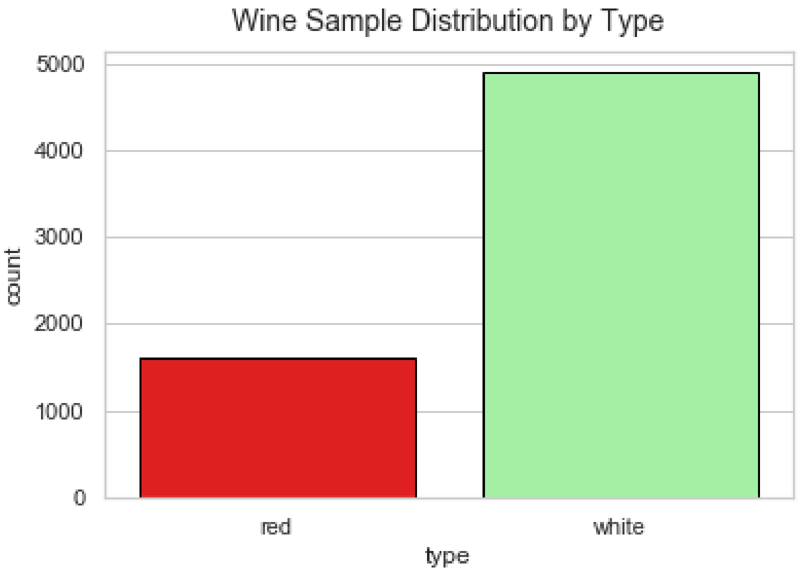
Exploratory Data Analysis: Residual Sugar – Total Sulfur Dioxide

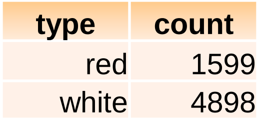


- The plot confirms previously stated results: total sulfur dioxide and residual sugar content seems to be much higher in white wines than in red wines
- The residual sugar has 0.50 correlation with total sulfur dioxide. This is an indication that more sulfur dioxide is added to wines with higher sugar

Predicting Wine Type (Red or White)

Distribution of wine samples





type	count
red	1599
white	4898

- The number of red wine samples is about 1/3 of the number of white wine samples.
- It is still pretty big number and it should be good enough for our prediction.

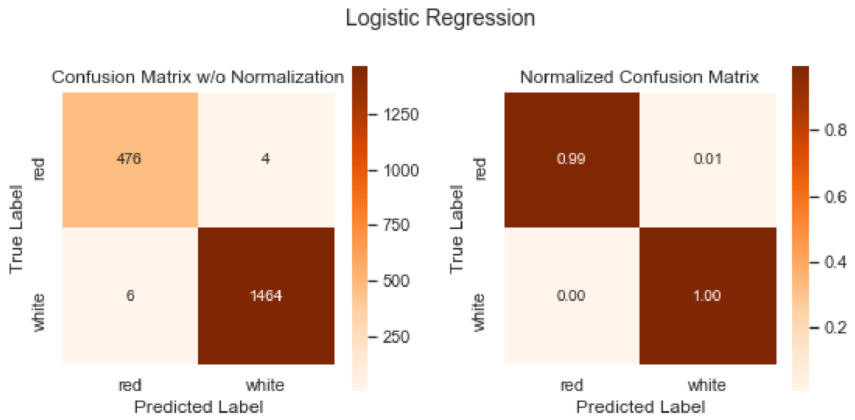
Predicting Wine Type (Red or White)

Logistic Regression

Classification Report				
	precision	recall	f1-score	support
red	0.99	0.99	0.99	480
white	1	1	1	1470
micro avg	0.99	0.99	0.99	1950
macro avg	0.99	0.99	0.99	1950
weighted avg	0.99	0.99	0.99	1950

Confusion Matrix				
		Predicted:		
		red	white	
Actual:	red	476	4	
	white	6	1464	

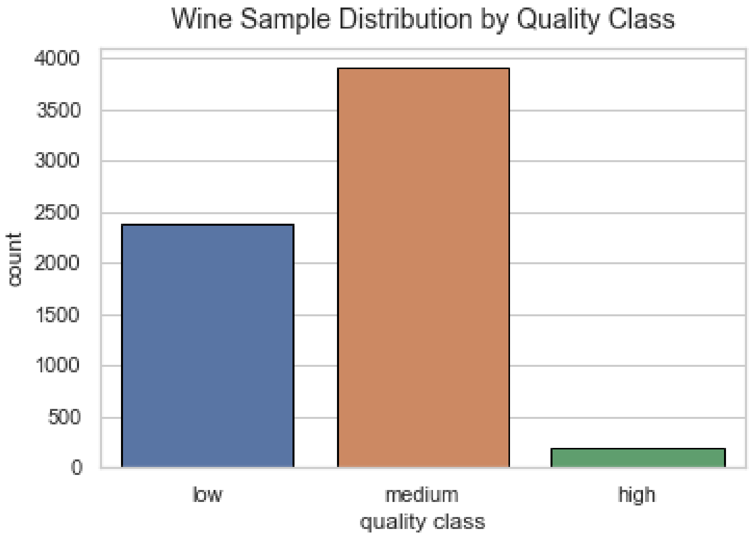
Accuracy Score:	0.994872



- Since this is a binary classification task we used logistic regression.
- As you can see, we got a wonderful result with 99.49 % accuracy. Only 10 of 1950 wine samples were misclassified. This result is really sufficient and obtained through a very simple model.
- This is also confirmed with the normalized confusion matrix on the left. The diagonal elements are close to 1 (dark) and off-diagonal elements are close to 0 (bright).

Predicting Quality Class of Wine (low, medium, high)

Distribution of wine samples



quality class	quality attribute	count
low	0 - 5	2384
medium	6 - 7	3915
high	8 - 10	198

- We grouped wine quality scores into three qualitative buckets as per the table above.
- The distribution data by wine quality classes is confirming the imbalance between classes.
- The number of samples in high quality class is pretty small and we could expect that could cause problems to our models.

Predicting Quality Class of Wine (low, medium, high)

Decision Tree Classifier

Classification Report

	precision	recall	f1-score	support
low	0.67	0.66	0.66	716
medium	0.76	0.79	0.78	1175
high	0.28	0.12	0.17	59
micro avg	0.72	0.72	0.72	1950
macro avg	0.57	0.52	0.54	1950
weighted avg	0.71	0.72	0.72	1950

Confusion Matrix

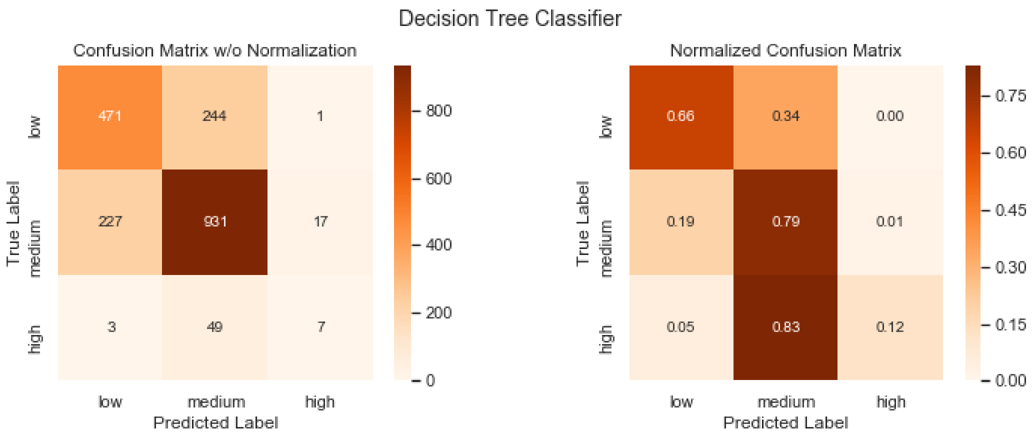
		Predicted:		
		low	medium	high
Actual:	low	471	244	1
	medium	227	931	17
	high	3	49	7

Accuracy Score:	0.722564

- Not as good results as for the wine type classification.
- From the class based statistics we can see that the recall for high quality wines is pretty bad. A lot of them have been misplaced into low and especially medium ratings.
- This was expected since there are not a lot of training samples for high quality wines.

Predicting Quality Class of Wine (low, medium, high)

Decision Tree Classifier – Confusion Matrix



- The diagonal elements in the confusion matrix represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier.
- The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

- Due to heavy class imbalance in our data, we had to perform normalization by class support size to have a more visual interpretation of which class is being misclassified.
- In our case it is obvious that lack of high quality wine samples caused their misplacement mostly into medium class.

Predicting Quality Class of Wine (low, medium, high)

Random Forest Classifier

Classification Report				
	precision	recall	f1-score	support
low	0.80	0.73	0.76	716
medium	0.82	0.88	0.85	1175
high	0.83	0.34	0.48	59
micro avg	0.81	0.81	0.81	1950
macro avg	0.82	0.65	0.70	1950
weighted avg	0.81	0.81	0.81	1950

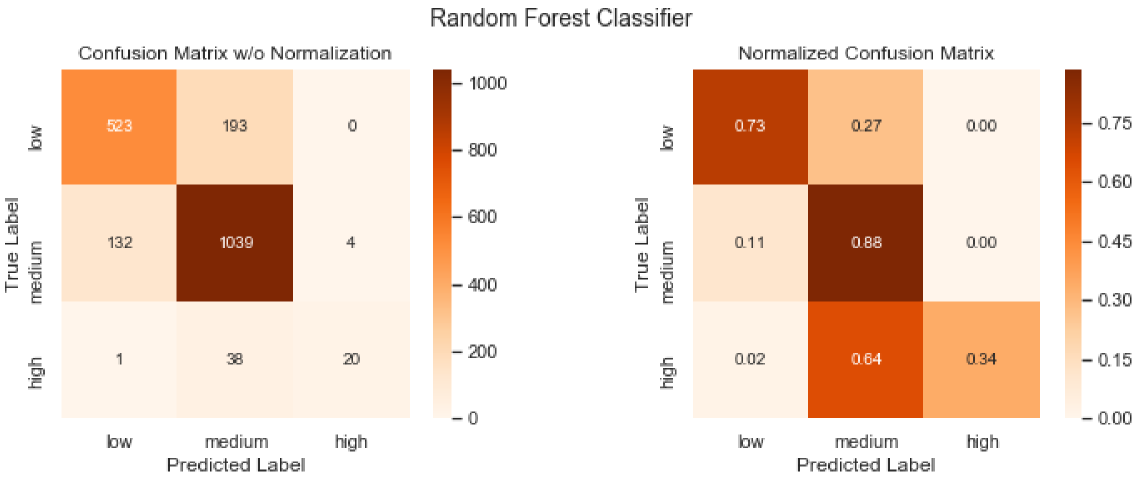
Confusion Matrix				
		Predicted:		
		low	medium	high
Actual:	low	523	193	0
	medium	132	1039	4
	high	1	38	20

Accuracy Score:	0.811282

- This model prediction results are quite good with an improvement in accuracy of about 9% from the decision tree model. Recall for high quality wines is considerably higher as well.
- Also there is no low quality wine sample has been misclassified as high, and only one high quality wine sample has been misclassified as low.
- We can see a considerable overlap between medium and high/low quality wine samples. Given the nature of the data and class distribution, that is expected.

Predicting Quality Class of Wine (low, medium, high)

Random Forest Classifier – Confusion Matrix



- In the perfect scenario diagonal elements in our normalized confusion matrix will be the darkest (~1) while the off-diagonal elements will be the brightest (~0)
- We can see visual improvements comparing to the decision tree model. For comparison see the 3rd slide bellow.

Predicting Quality Class of Wine (low, medium, high)

Support Vector Machine Classifier

Classification Report				
	precision	recall	f1-score	support
low	0.76	0.59	0.66	716
medium	0.75	0.88	0.81	1175
high	0.78	0.12	0.21	59
micro avg	0.75	0.75	0.75	1950
macro avg	0.76	0.53	0.56	1950
weighted avg	0.75	0.75	0.74	1950

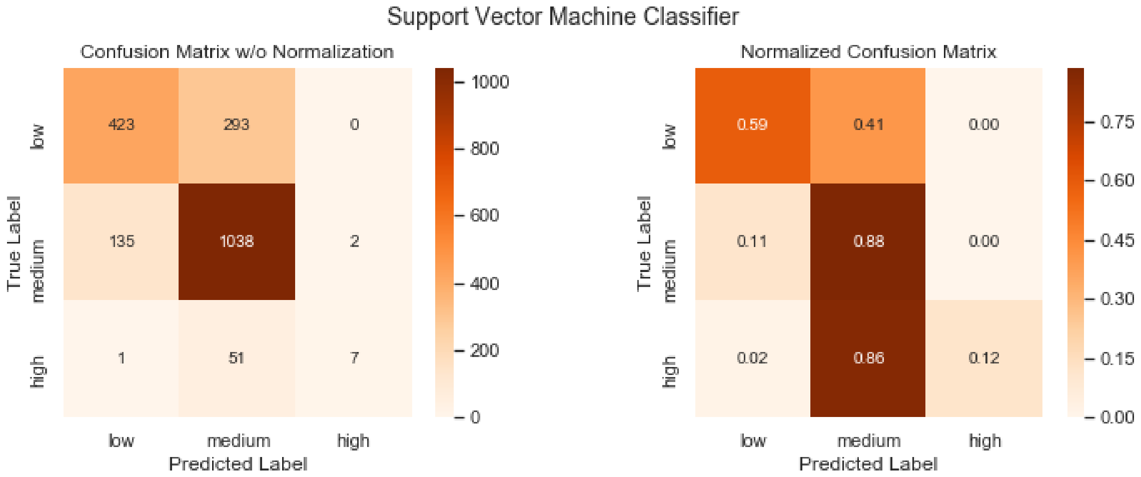
Confusion Matrix				
		Predicted:		
		low	medium	high
Actual:	low	423	293	0
	medium	135	1038	2
	high	1	51	7

Accuracy Score:	0.752821

- This model behaves almost the same as Random Forest Classifier for the medium quality samples.
- It fails for low quality wines and especially for high quality wines. Recall for high quality wines is very bad, the same as for Decision Tree Classifier.
- It looks that results in this model depends more on the number of samples than for the other 2 models.

Predicting Quality Class of Wine (low, medium, high)

Support Vector Machine Classifier – Confusion Matrix

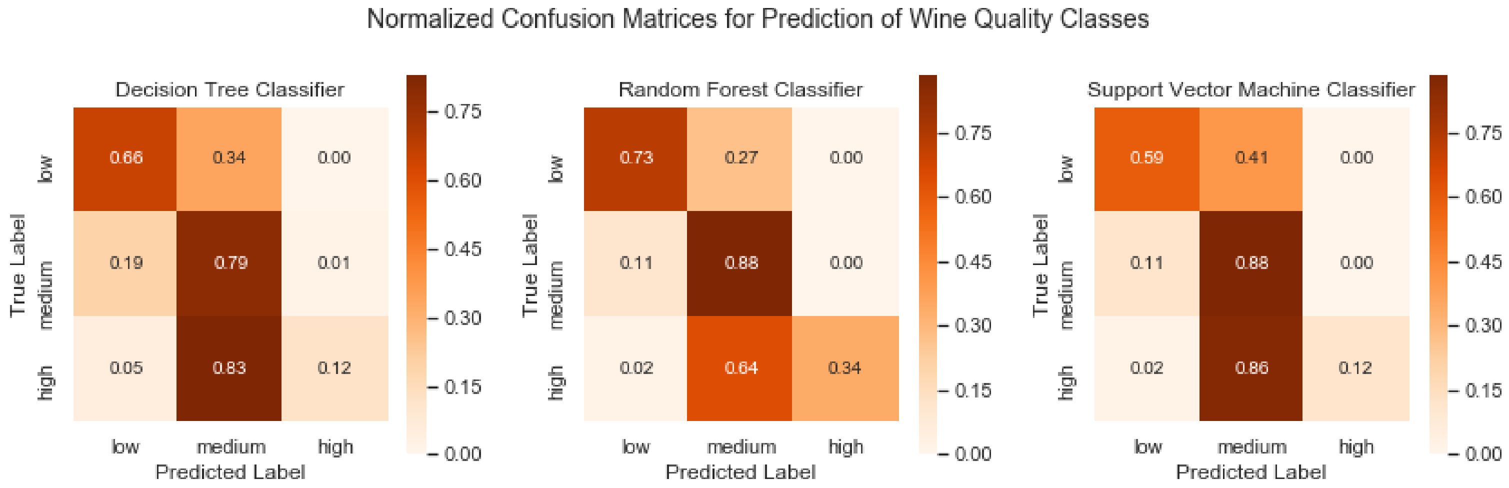


Accuracy Score	
DTC	0.722564
RFC	0.811282
SVC	0.752821

- The very dark off-diagonal element and bright diagonal element for high quality wines in the normalized matrix, represent a failure of this model.
- We can notice visually the similar behavior for low quality wines.
- The accuracy score places this model between Decision Tree Classifier, but worse than Random Forest Classifier.

Predicting Quality Class of Wine (low, medium, high)

Comparing Confusion Matrices



Normalized Confusion Matrices for Prediction of Wine Quality Classes

Decision Tree Classifier

Random Forest Classifier

Support Vector Machine Classifier

- From the normalized confusion matrices above, we can easily see which model performs better
- The diagonal elements for Random Forest Classifier are darker and off-diagonal elements brighter than for the other two models
- Due to better performance for medium quality wines, Support Vector Machine Classifier slightly outperforms Decision Tree Classifier

Limitations

- The following Notes by UCI about these datasets are describing main limitations which we were facing throughout this project:

  - The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

  - Outlier detection algorithms could be used to detect the few excellent or poor wines.

  - Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

- Applying some sort of feature selection could be useful also due to the fact that some of the attributes might be correlated.
- To correct some of the above listed limitations, we should perform more tuning on the used models and/or use other classifiers and see how they will behave.

Conclusions

- We have achieved both of our goals with relatively good results.
- In the Exploratory Data Analysis step we were able to discover several relationships between physicochemical attributes and wine types and they could be used for red or white wine classification. There were only few but week relations with wine quality and they could not help so much with wine quality classification. This was all confirmed with visualization and predictive models.
- In the first classification challenge (type: red or white) a simple Logistic Regression model was able to obtain an excellent result with 99.49% accuracy.
- In the second classification challenge (quality: low, medium, high) we tried three classifiers. Due to imbalance in data distribution (very few samples for high quality wines) they were not able to bring us the desired results. Random Forest was the best classifier with the accuracy score of 81.13%.
- We can conclude that it is possible to predict the type of a wine and its quality from the physicochemical attributes, but for the quality classification some improvements are needed by employing other methods and/or fine tuning used models.

Acknowledgements

No one gave me feedback

References

- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.
  In Decision Support Systems, Elsevier, 47(4):547-553, 2009. Available at: [Web Link]
- Vinho Verde: http://www.vinhoverde.pt/en/promotional-materials#!
- Wine Folly: https://winefolly.com/blog/
- Seaborn Visualization: https://jovianlin.io/data-visualization-seaborn-part-1/
- Confusion Matrix: http://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html#sphx-glr-auto-examples-model-selection-plot-confusion-matrix-py
- Cross- Validation: https://machinelearningmastery.com/k-fold-cross-validation/
- Machine Learning with Scikit-Learn: https://elitedatascience.com/python-machine-learning-tutorial-scikit-learn#step-5