Aim: To predict whether a wine is red or white, based on the metrics provided

1. Generate the dataset

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
0	6.8	0.56	0.22	1.8	0.074	15.0	24.0	0.99438	3.40	0.82	11.2	6	1
1	6.4	0.30	0.36	2.0	0.052	18.0	141.0	0.99273	3.38	0.53	10.5	6	0
2	5.9	0.17	0.29	3.1	0.030	32.0	123.0	0.98913	3.41	0.33	13.7	7	0
3	7.0	0.24	0.24	1.8	0.047	29.0	91.0	0.99251	3.30	0.43	9.9	6	0
4	6.4	0.45	0.07	1.1	0.030	10.0	131.0	0.99050	2.97	0.28	10.8	5	0

2. a) Check for null values

```
fixed acidity           False
volatile acidity        False
citric acid             False
residual sugar          False
chlorides               False
free sulfur dioxide     False
total sulfur dioxide    False
density                 False
pH                      False
sulphates               False
alcohol                 False
quality                 False
type                    False
dtype: bool
```

2. b) Check for missing values

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
type                    0
dtype: int64


0    4898
```

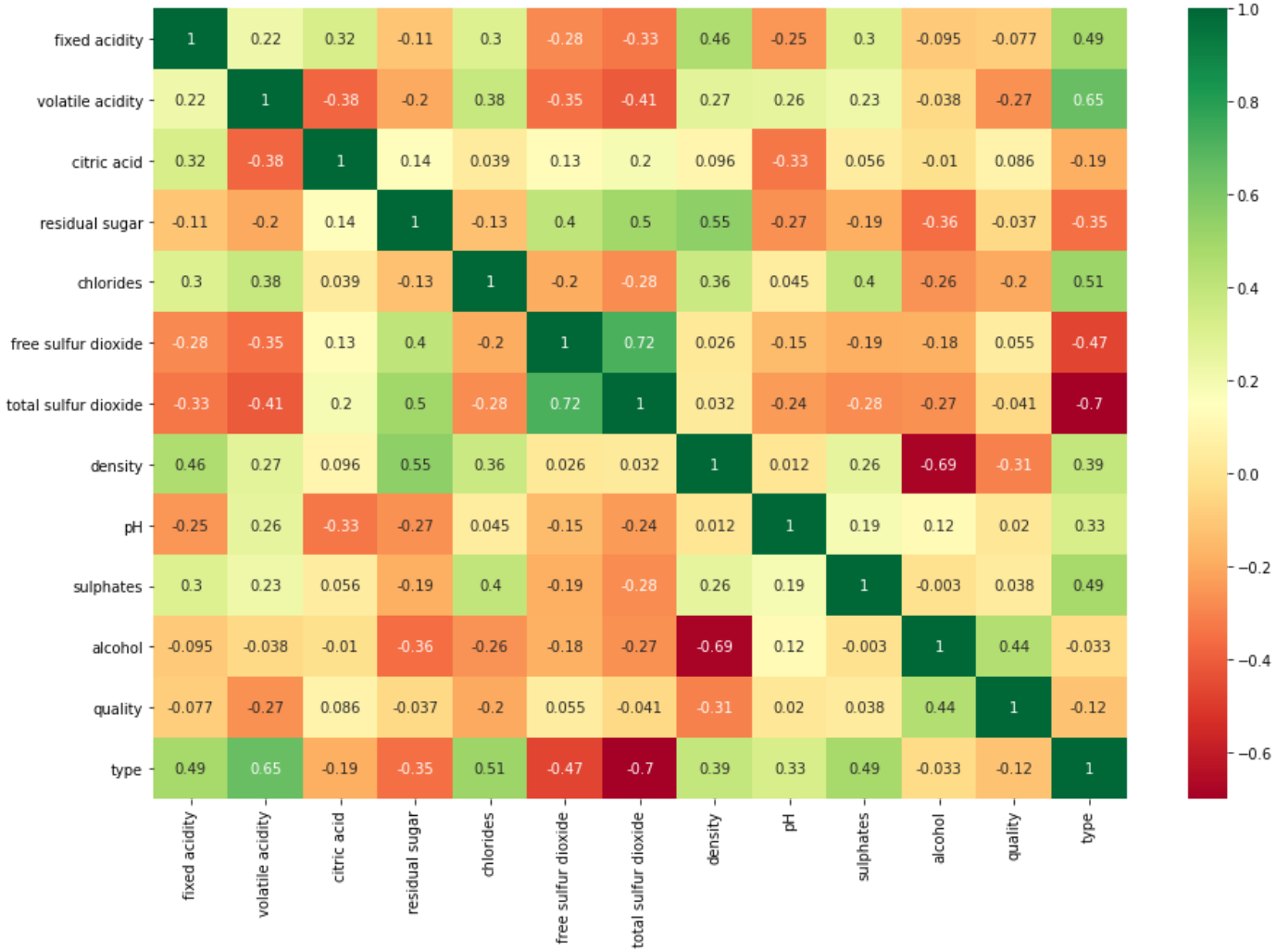
```
1    1599
Name: type, dtype: int64
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	115.744574	0.994697
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	56.521855	0.002999
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980

3. Feature set analysis

Correlations using a heat map

```
<AxesSubplot:>
```

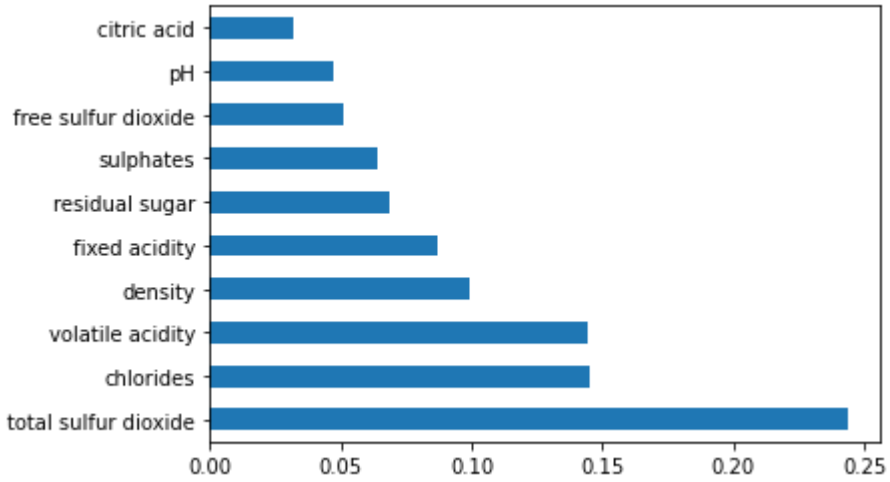


From the heat map above, we see that the positive correlations are seen in the following features: volatile acidity, chlorides, fixed acidity, and sulphates

Highly negative correlations are seen in this feature: total sulfur dioxide

Extra-trees classifier to fit a number of randomized decision trees on sub-samples of the dataset. Using averaging, it improves the predictive accuracy and controls over-fitting.

```
[0.08725983 0.14409015 0.03196323 0.06846117 0.14484194 0.05081832
 0.24356034 0.09896197 0.04713687 0.06417429 0.0187319 ]
```



We see that total sulfur dioxide, volatile acidity, chlorides, density and sulphates have strong importances.

From both these methods, we can say that total sulfur dioxide, volatile acidity, chlorides, and sulphates are important features to consider. Features like fixed acidity and density may also play a role in classification.

4. Creating the training and test datasets with an 80:20 split

Also creating a subset of the wine data consisting only of the top features as per the analysis above

5. Creating classification models

Analyses on both the original wine dataset and the subset with the important features

Decision tree classification on full wine dataset

```
Confusion matrix:
 [[960  10]
 [ 17 313]]

Accuracy score:  0.9792307692307692

Classification report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.99       970
           1       0.97      0.95      0.96       330

    accuracy                           0.98      1300
   macro avg       0.98      0.97      0.97      1300
weighted avg       0.98      0.98      0.98      1300
```

SVM classification on full wine dataset

```
Confusion matrix:
 [[967   3]
 [ 16 314]]

Accuracy score:  0.9853846153846154

Classification report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       970
           1       0.99      0.95      0.97       330

    accuracy                           0.99      1300
   macro avg       0.99      0.97      0.98      1300
weighted avg       0.99      0.99      0.99      1300
```

KNN classification on full wine dataset

```
Confusion matrix:
 [[945  25]
 [ 38 292]]

Accuracy score:  0.9515384615384616

Classification report:
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       970
           1       0.92      0.88      0.90       330

    accuracy                           0.95      1300
   macro avg       0.94      0.93      0.94      1300
weighted avg       0.95      0.95      0.95      1300
```

Decision tree classification on subset of wine dataset

```
Confusion matrix:
 [[962   8]
 [ 15 315]]

Accuracy score:  0.9823076923076923

Classification report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.99       970
           1       0.98      0.95      0.96       330

    accuracy                           0.98      1300
   macro avg       0.98      0.97      0.98      1300
weighted avg       0.98      0.98      0.98      1300
```

SVM classification on subset of wine dataset

```
Confusion matrix:
 [[963   7]
 [ 21 309]]

Accuracy score:  0.9784615384615385

Classification report:
```

```
              precision    recall  f1-score   support

           0       0.98      0.99      0.99       970
           1       0.98      0.94      0.96       330

    accuracy                           0.98      1300
   macro avg       0.98      0.96      0.97      1300
weighted avg       0.98      0.98      0.98      1300
```

KNN classification on subset of wine dataset

```
Confusion matrix:
 [[946  24]
 [ 44 286]]

Accuracy score:  0.9476923076923077

Classification report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       970
           1       0.92      0.87      0.89       330

    accuracy                           0.95      1300
   macro avg       0.94      0.92      0.93      1300
weighted avg       0.95      0.95      0.95      1300
```

Some observations

- We see that all the models perform well on this dataset
- The best performing was the simple Linear Support Vector Machine on the entire dataset, followed by the Decision Tree classification model on the subset of important features
- Choosing a subset of important features via feature selection created comparably accurate models

Using StandardScaler to standardize features by removing the mean and scaling to unit variance

Decision tree classification on scaled input (entire dataset)

```
Confusion matrix:
 [[961   9]
 [ 18 312]]

Accuracy score:  0.9792307692307692

Classification report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.99       970
           1       0.97      0.95      0.96       330

    accuracy                           0.98      1300
   macro avg       0.98      0.97      0.97      1300
weighted avg       0.98      0.98      0.98      1300
```

Support Vectore classification on scaled input (entire dataset)

```
Confusion matrix:
 [[967   3]
 [  6 324]]

Accuracy score:  0.9930769230769231

Classification report:
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       970
           1       0.99      0.98      0.99       330

    accuracy                           0.99      1300
   macro avg       0.99      0.99      0.99      1300
weighted avg       0.99      0.99      0.99      1300
```

KNN classification on scaled input (entire dataset)

```
Confusion matrix:
 [[967   3]
 [  7 323]]

Accuracy score:  0.9923076923076923

Classification report:
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       970
           1       0.99      0.98      0.98       330

    accuracy                           0.99      1300
   macro avg       0.99      0.99      0.99      1300
weighted avg       0.99      0.99      0.99      1300
```