Wine Quality Predictive

Elijah M Raffo
Jiazhou Li
Xunyi Jiang

Introduction

- Primary data is from Analytic Vidhya
- Find out what makes good quality wine
- 2 datasets: red and white wine
- Our goal is to create a robust model for classification whose variables are linked to real world methods of testing batches of wine

Descriptive analytics

Variables:

- Fixed acidity,
- volatile acidity,
- citric acid,
- residual sugar,
- chlorides,
- free sulfur dioxide,
- total sulfur dioxide,
- density,
- pH (scale 0 to 14),
- sulphates,
- alcohol

Target Variable:

Quality

(score between 0 to 10)

Descriptive analytics

- Both red wine and white wine

```
In [68]: wine=pd.read_csv('winequality.csv')
         wine.describe()
```

Out[68]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	6497.000000	6497.000000	6497.000000	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	30.525473	115.744728	0.994697	3.218501	0.531268	10.491801
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749313	56.521751	0.002999	0.160787	0.148806	1.192712
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	0.220000	8.000000
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	0.430000	9.500000
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	0.510000	10.300000
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	0.600000	11.300000
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.010000	2.000000	14.900000

Descriptive analytics

- Raw data:

```
In [50]: wine.shape
Out[50]: (6497, 13)
```

- No missing data

```
In [49]: wine.dropna()
```

6486	red	7.2	0.660	0.33	2.50	0.068	34.0	102.0	0.99414	3.27	0.78	12.8	6
6487	red	6.6	0.725	0.20	7.80	0.073	29.0	79.0	0.99770	3.29	0.54	9.2	5
6488	red	6.3	0.550	0.15	1.80	0.077	26.0	35.0	0.99314	3.32	0.82	11.6	6
6489	red	5.4	0.740	0.09	1.70	0.089	16.0	26.0	0.99402	3.67	0.56	11.6	6
6490	red	6.3	0.510	0.13	2.30	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
6491	red	6.8	0.620	0.08	1.90	0.068	28.0	38.0	0.99651	3.42	0.82	9.5	6
6492	red	6.2	0.600	0.08	2.00	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
6493	red	5.9	0.550	0.10	2.20	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
6494	red	6.3	0.510	0.13	2.30	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
6495	red	5.9	0.645	0.12	2.00	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
6496	red	6.0	0.310	0.47	3.60	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

6497 rows × 13 columns

Descriptive analytics

- 2836 datas are in the quality score 6

```
In [98]: wine['quality_reg'].value_counts()

Out[98]: 6    2836
         5    2138
         7    1079
         4     216
         8     193
         3      30
         9       5
         Name: quality_reg, dtype: int64
```

1) Regression

2) Classification

Good >= 6
Bad < 6

3) Red and White

```
KNN4 acc on train: 0.269
KNN4 acc on test: 0.160
LIN acc on train: 0.297
LIN acc on test: 0.291
LOG acc on train: 0.541
LOG acc on test: 0.533
Lasso acc on train: 0.290
Lasso acc on test: 0.285
SVM2 acc on train: 0.297
SVM2 acc on test: 0.290
SVM2 acc on train: 0.321
SVM2 acc on test: 0.282
RF acc on train: 0.931
RF acc on test: 0.527
```

```
KNN of 1 acc on train: 1.000
KNN of 1 acc on test: 0.734
LOR acc on train: 0.742
LOR acc on test: 0.745
LSVM C=.01 acc on train: 0.727
LSVM C=.01 acc on test: 0.718
SVM Gamma=.1 acc on train: 0.912
SVM Gamma=.1acc on test: 0.725
DT Leafs=15 acc on train: 0.765
DT leafs=15 acc on test: 0.738
RF of 50 acc on train: 1.000
RF of 50 acc on test: 0.825
```

	importance
alcohol	0.151985
volatile acidity	0.115596
density	0.097970
chlorides	0.086533
free sulfur dioxide	0.085907
total sulfur dioxide	0.083196
sulphates	0.081278
residual sugar	0.079440
citric acid	0.075513
pH	0.070515
fixed acidity	0.064437
type_white	0.003838
type_red	0.003790

Wine Classification

count	4898.000000
mean	5.877909
std	0.885639
min	3.000000
25%	5.000000
50%	6.000000
75%	6.000000
max	9.000000

White Wine

count	1599.000000
mean	5.636023
std	0.807569
min	3.000000
25%	5.000000
50%	6.000000
75%	6.000000
max	8.000000

Red Wine

Tests	Wine	Red	White
KNN	73.4% [1]	68.8% [30]	68.2% [10]
Logistic Regressor	74.5%	75.7%	73.2%
Lin SVM [C]	71.8% [.01]	74.5% [.1]	70.5% [.01]
Supoort Vector Machine [Gamma]	72.5% [.1]	71.0% [.01]	72.7% [1]
Decision Tree [leaf node]	73.8% [15]	73.5% [20]	75% [30]
Random Forest [n_estimator]	82.5% [50]	81.0% [50]	82.5% [100]

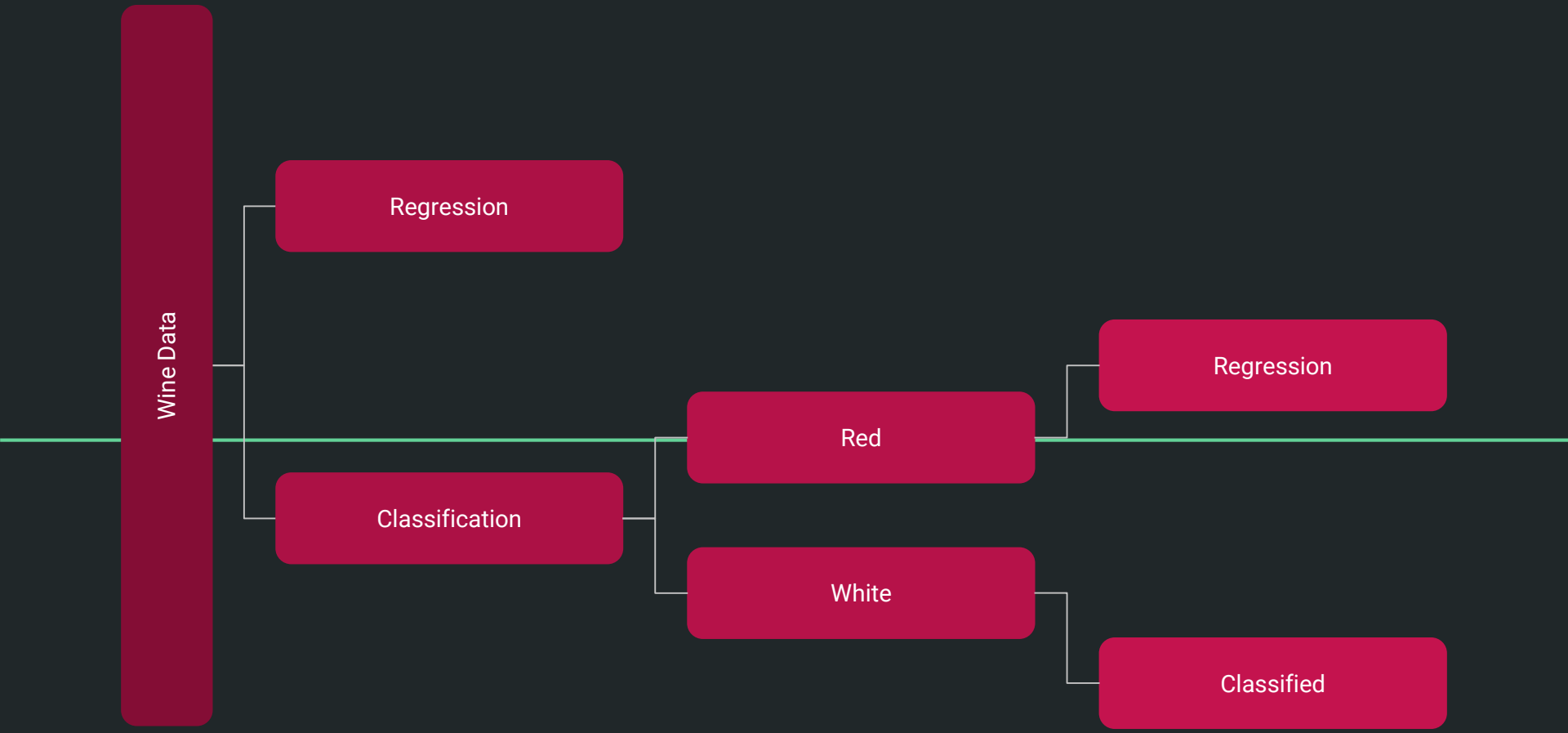
Wine Classification

	importance
alcohol	0.143505
volatile acidity	0.115256
density	0.107909
free sulfur dioxide	0.102896
total sulfur dioxide	0.086073
residual sugar	0.081336
citric acid	0.081118
chlorides	0.080260
pH	0.071636
sulphates	0.065095
fixed acidity	0.064916

White Wine

	importance
alcohol	0.183169
sulphates	0.137991
volatile acidity	0.108503
total sulfur dioxide	0.097881
density	0.088560
pH	0.068507
chlorides	0.068428
citric acid	0.066616
fixed acidity	0.066316
free sulfur dioxide	0.059956
residual sugar	0.054072

Red Wine


Wine Data
Regression
Classification
Red
White
Regression
Classified

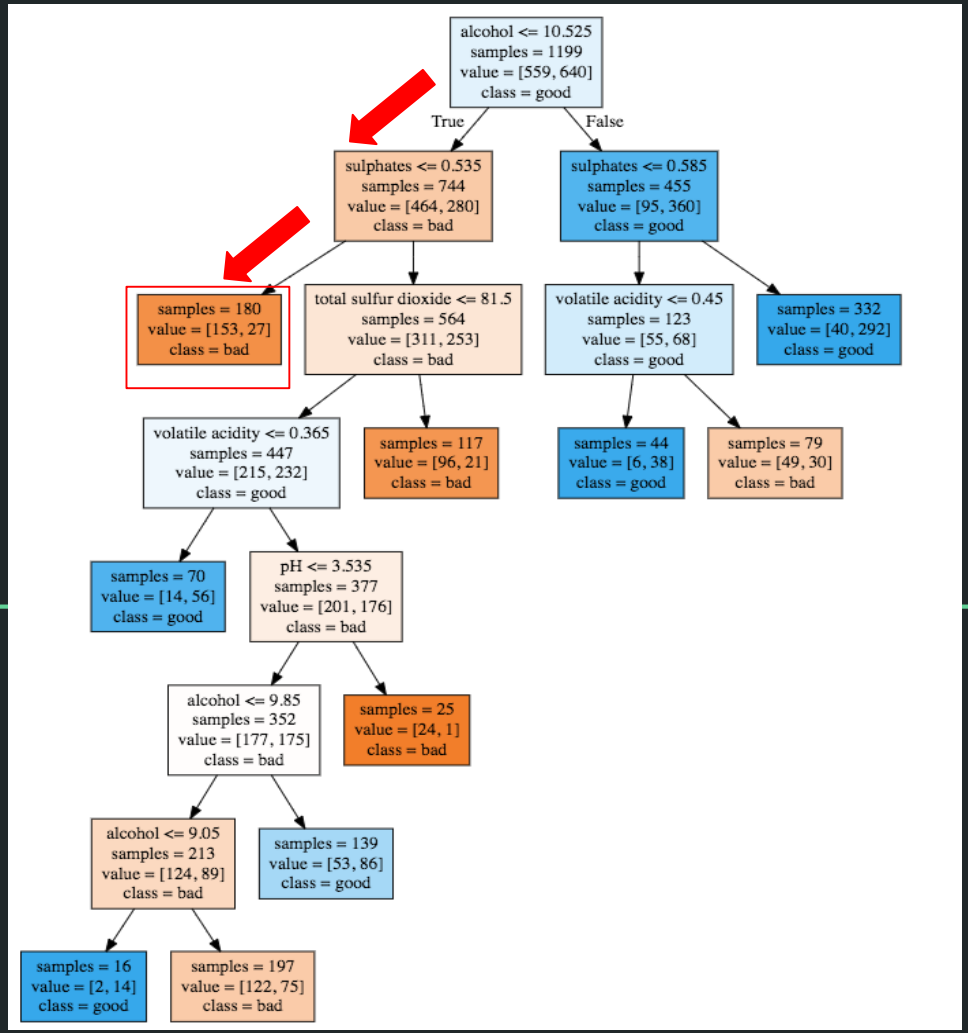
	White Classification		Red Regression
KNN	69.3%	KNN	3.8%
Logistic Regressor	71.9%	Logistic Regressionn	77.1%
Lin SVM	70.4%	Lasso	15.5%
Supoort Vector Machine	79.4.%	Ridge	15.6%
Decision Tree	73.9%	Decision Tree	8.0%
Random Forest	83.8%	Random Forest	33.0%

Decison Tree

- Red wine

Alcohol= 8.8
Sulphates= 0.45

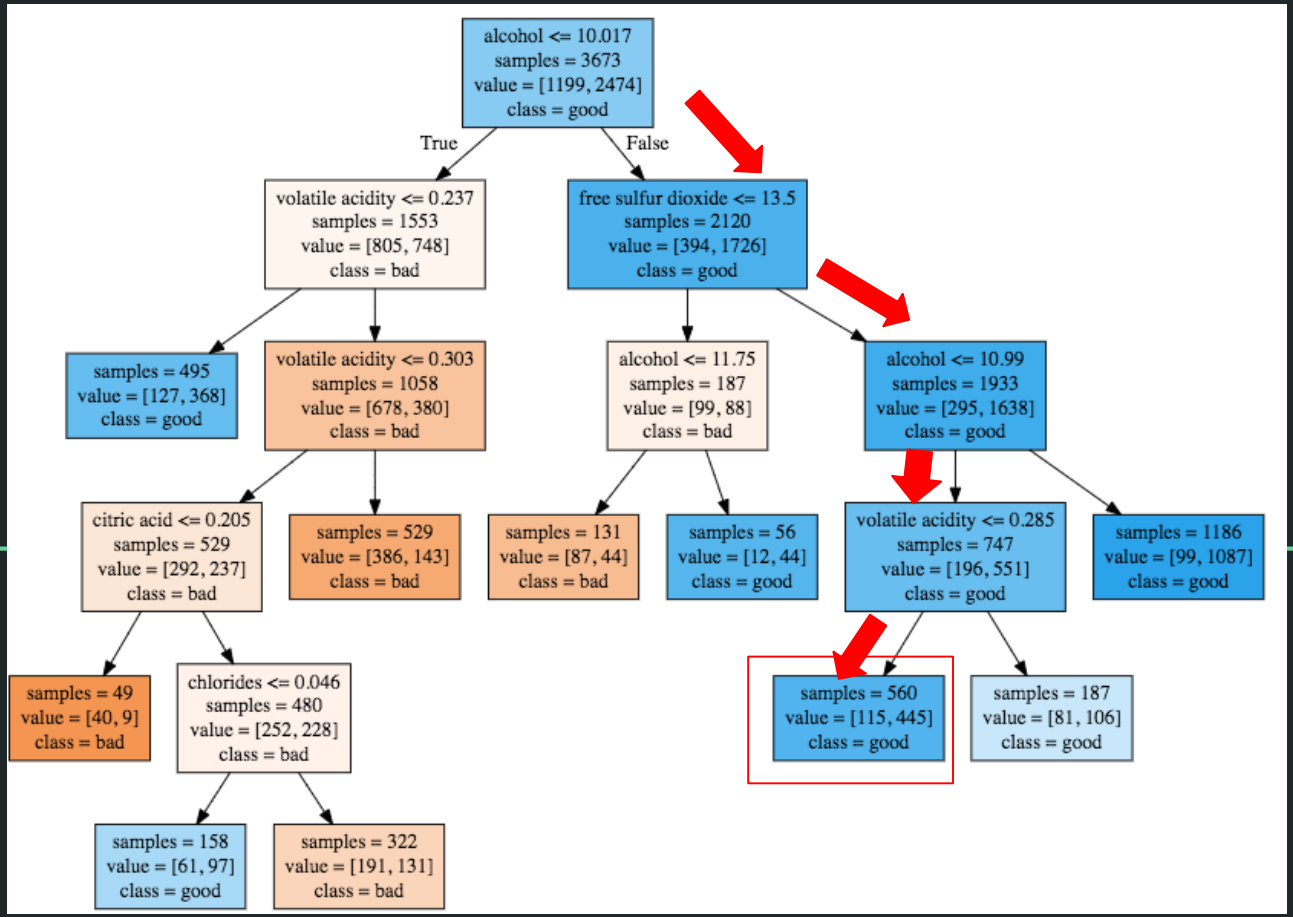
Class bad



- White Wine

alcohol= 10.1
free sulfur dioxide= 30
volatile acidity= 0.28

Class Good



Findings

-Making a “good” wine is possible

-The characteristics of red and white

-Decision Tree is useful and accurate

For a Better Accuracy

-More Descriptive Data

-Domain Knowledge

Using the Findings

-Optimizing wine blends

THANK YOU