Классификация цвета вина и качества вина

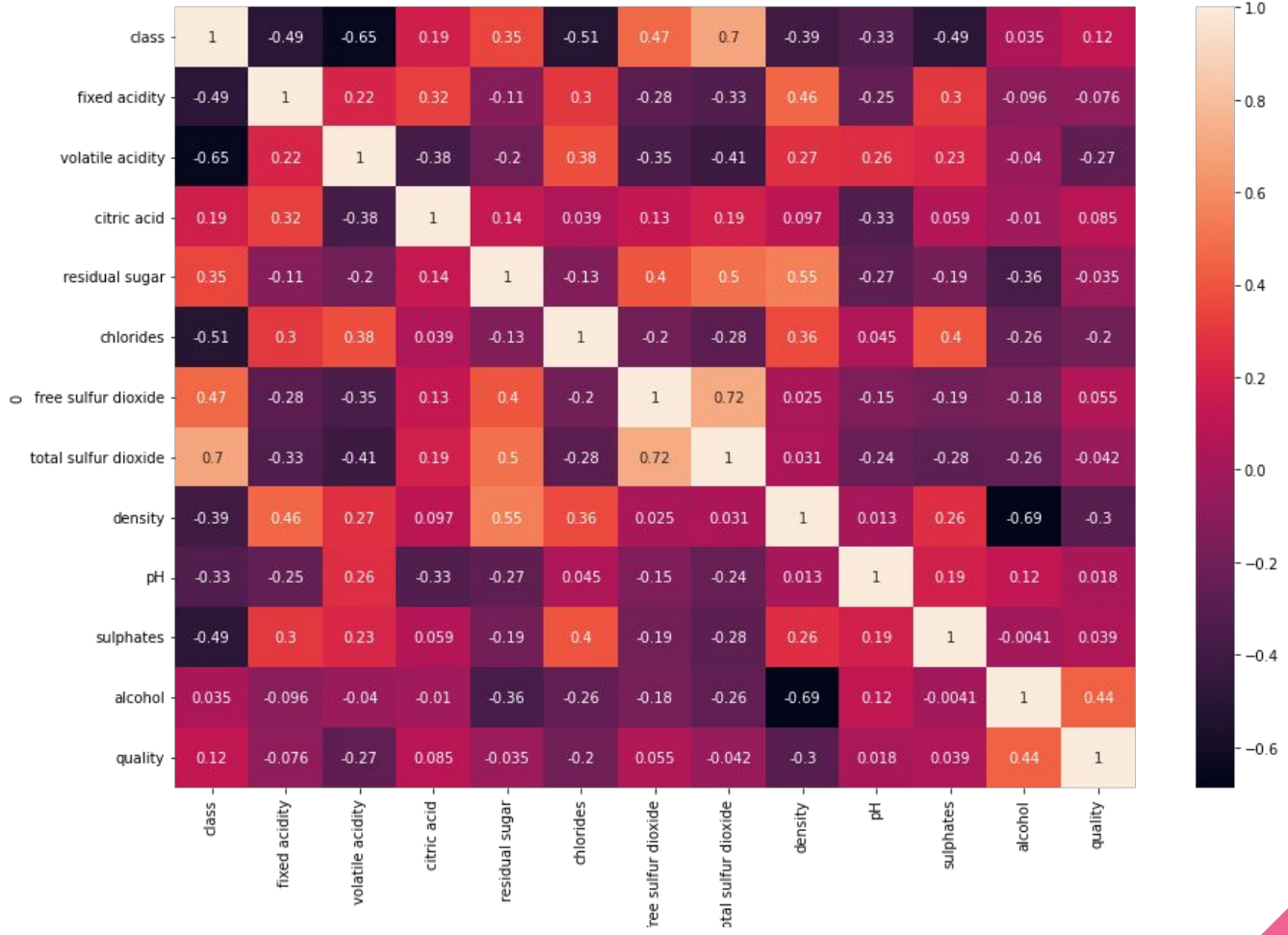
Ходзицкий А. Ф.
Слобожанин А. В.
Сергеев А. В.

Без пропущенных полей

Без выбросов

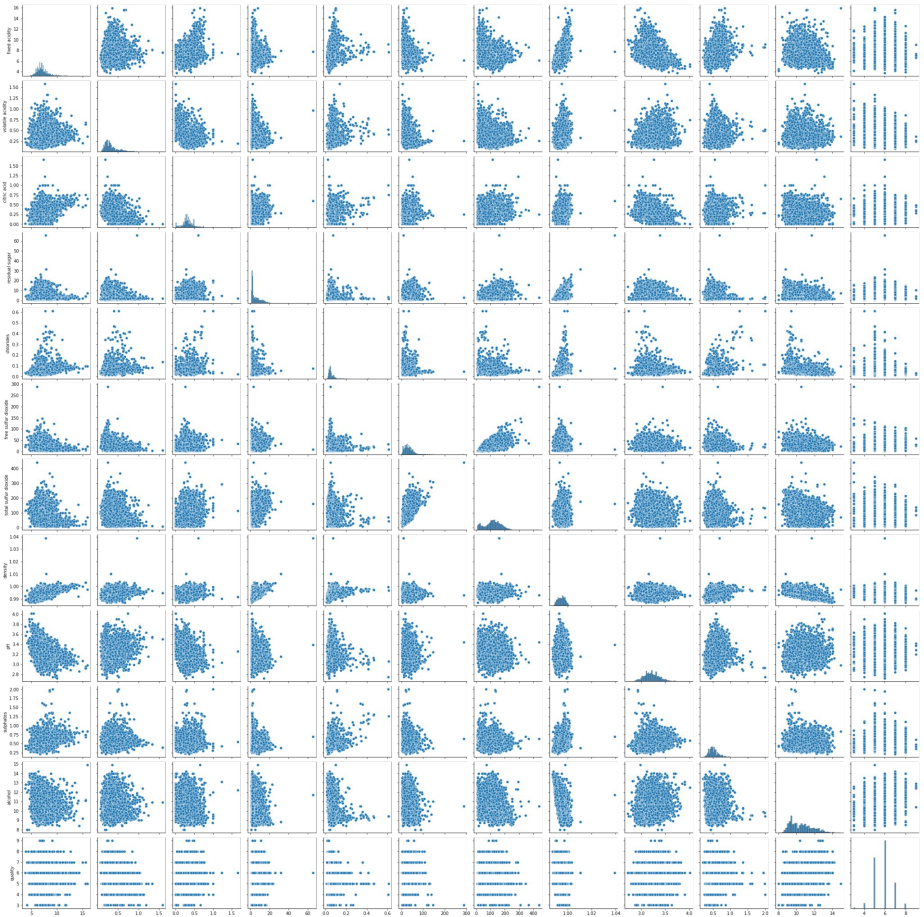
#	Column	Non-Null					
0	class	6497 non-null	object	6463 non-null	string	5975 non-null	int64
1	fixed acidity	6487 non-null	object	6463 non-null	float64	5975 non-null	float64
2	volatile acidity	6489 non-null	object	6463 non-null	float64	5975 non-null	float64
3	citric acid	6494 non-null	object	6463 non-null	float64	5975 non-null	float64
4	residual sugar	6495 non-null	object	6463 non-null	float64	5975 non-null	float64
5	chlorides	6495 non-null	object	6463 non-null	float64	5975 non-null	float64
6	free sulfur dioxide	6497 non-null	object	6463 non-null	float64	5975 non-null	float64
7	total sulfur dioxide	6497 non-null	object	6463 non-null	float64	5975 non-null	float64
8	density	6497 non-null	object	6463 non-null	float64	5975 non-null	float64
9	pH	6488 non-null	object	6463 non-null	float64	5975 non-null	float64
10	sulphates	6493 non-null	object	6463 non-null	float64	5975 non-null	float64
11	alcohol	6497 non-null	object	6463 non-null	float64	5975 non-null	float64
12	quality	6497 non-null	object	6463 non-null	float64	5975 non-null	int64

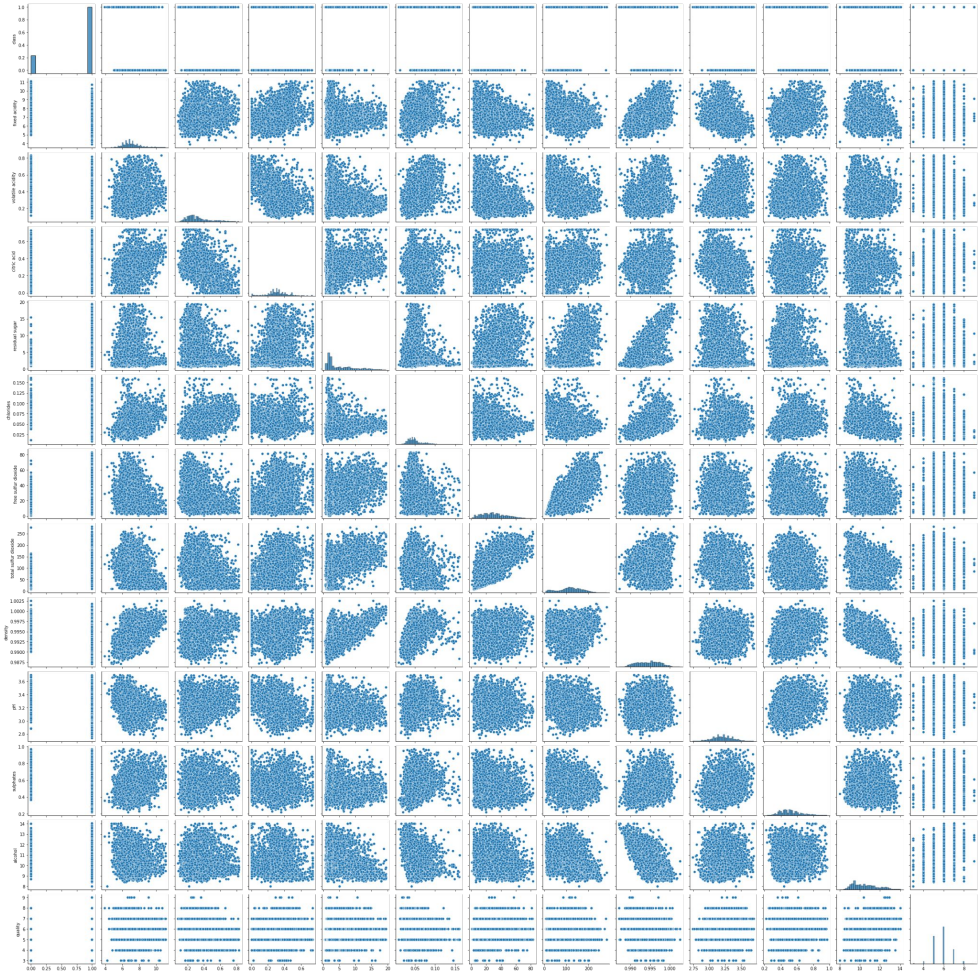
dtypes: object(13)
memory usage: 710.6+ KB



Матрица корреляций

Убрали выбросы





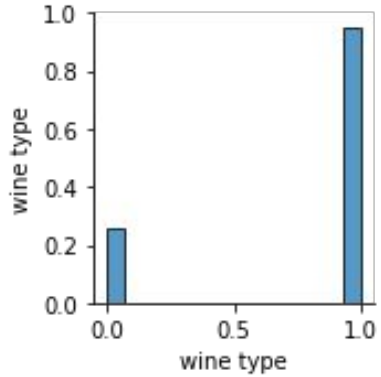
Выводы по данным

1. 34 пропущенных значения (~0.5% данных)
2. 488 выброса (~7.5% данных)
3. Данные представлены в одних и тех же единицах значений
4. Данные распределены без существенных корреляций относительно друг друга, данные распределены близко к нормальному распределению (визуально)
5. Данные были представлены в формате строк

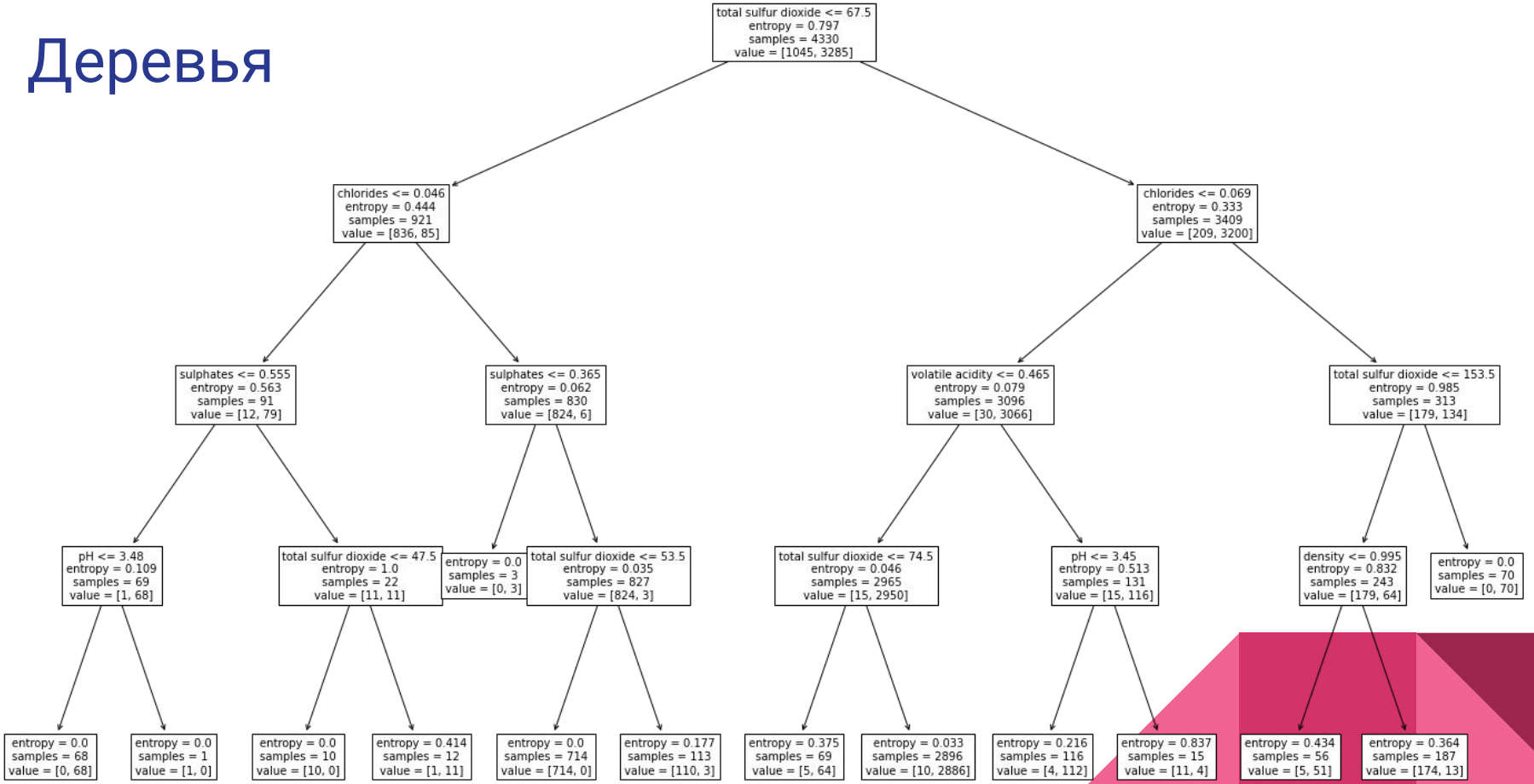
Задача классификации цвета вина

Задача (классификации): определить цвет вина (белый или красный) по его признакам.

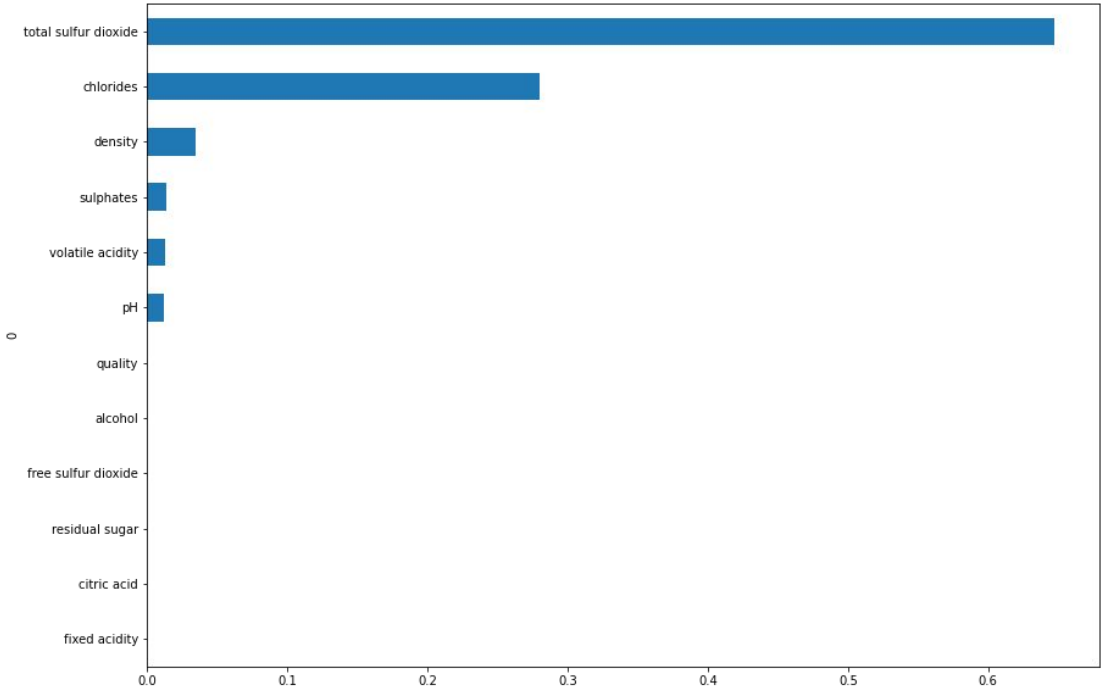
Методы решения задачи: деревья, линейный дискриминант.



Деревья


total sulfur dioxide <= 67.5
entropy = 0.797
samples = 4330
value = [1045, 3285]
chlorides <= 0.046
entropy = 0.444
samples = 921
value = [836, 85]
chlorides <= 0.069
entropy = 0.333
samples = 3409
value = [209, 3200]
sulphates <= 0.555
entropy = 0.563
samples = 91
value = [12, 79]
sulphates <= 0.365
entropy = 0.062
samples = 830
value = [824, 6]
volatile acidity <= 0.465
entropy = 0.079
samples = 3096
value = [30, 3066]
total sulfur dioxide <= 153.5
entropy = 0.985
samples = 313
value = [179, 134]
pH <= 3.48
entropy = 0.109
samples = 69
value = [1, 68]
total sulfur dioxide <= 47.5
entropy = 1.0
samples = 22
value = [11, 11]
entropy = 0.0
samples = 3
value = [0, 3]
total sulfur dioxide <= 53.5
entropy = 0.035
samples = 827
value = [824, 3]
total sulfur dioxide <= 74.5
entropy = 0.046
samples = 2965
value = [15, 2950]
pH <= 3.45
entropy = 0.513
samples = 131
value = [15, 116]
density <= 0.995
entropy = 0.832
samples = 243
value = [179, 64]
entropy = 0.0
samples = 70
value = [0, 70]
entropy = 0.0
samples = 68
value = [0, 68]
entropy = 0.0
samples = 1
value = [1, 0]
entropy = 0.0
samples = 10
value = [10, 0]
entropy = 0.414
samples = 12
value = [1, 11]
entropy = 0.0
samples = 714
value = [714, 0]
entropy = 0.177
samples = 113
value = [110, 3]
entropy = 0.375
samples = 69
value = [5, 64]
entropy = 0.033
samples = 2896
value = [10, 2886]
entropy = 0.216
samples = 116
value = [4, 112]
entropy = 0.837
samples = 15
value = [11, 4]
entropy = 0.434
samples = 56
value = [5, 51]
entropy = 0.364
samples = 187
value = [174, 13]

Деревья



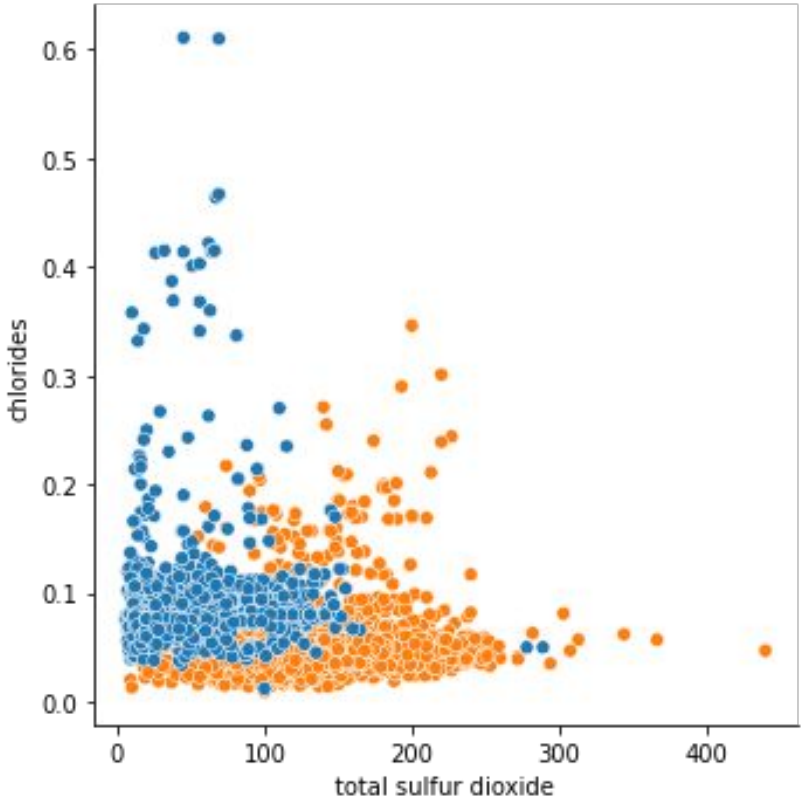

```
Верно определенные ко всем train: 0.9885086185360 98
Верно определенные ко всем test : 0.9827586206896551
```

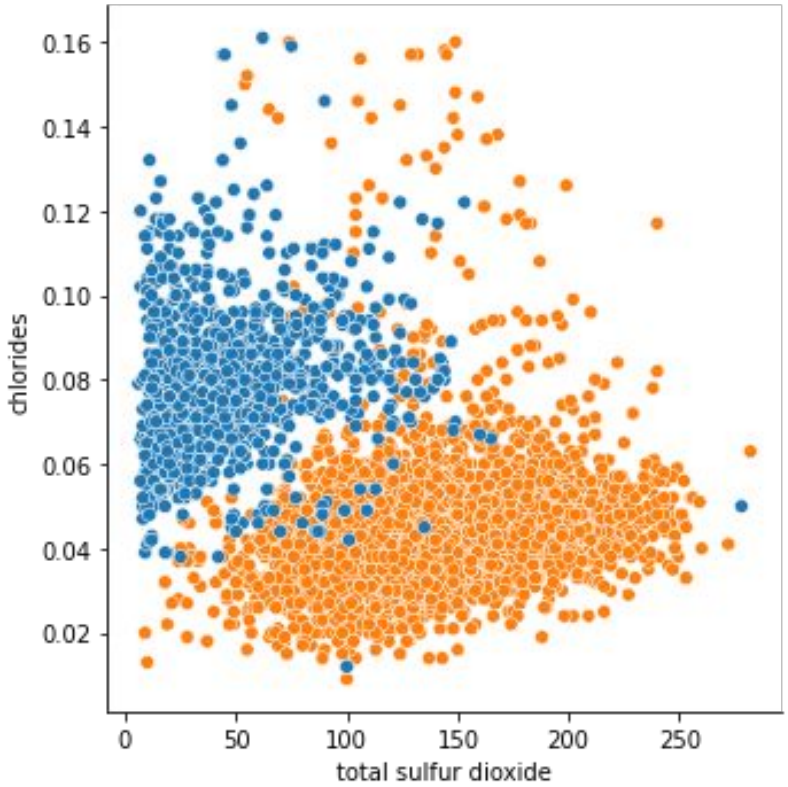
Линейный дискриминант

$$(\vec{x}-\vec{\mu}_0)^T\Sigma_0^{-1}(\vec{x}-\vec{\mu}_0)+\ln|\Sigma_0|-(\vec{x}-\vec{\mu}_1)^T\Sigma_1^{-1}(\vec{x}-\vec{\mu}_1)-\ln|\Sigma_1|$$

```
Верно определенные ко всем train: 0.9960029977516862
Верно определенные ко всем test : 0.9964503042596349
```


ПОЧЕМУ


chlorides
total sulfur dioxide
wine type
0
1


chlorides
total sulfur dioxide
wine type
0
1

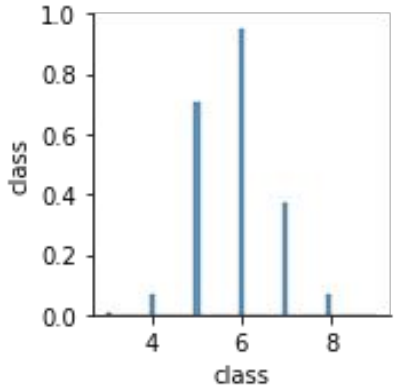
Выводы по задаче классификации цвета

- Это простая задача

Задача классификации качества вина

Задача (классификации): определить оценку вина (от 3 до 9) по его признакам.

Методы решения задачи: деревья, метод k-соседей, дискриминантный анализ, градиентный бустинг.



Линейный дискриминант

Верно определенные ко всем train: 0.5380964276792406

Верно определенные ко всем test : 0.552738336713996

Наш score : 0.9482758620689655

	precision	recall	f1-score	support
3	0.00	0.00	0.00	3
4	0.54	0.11	0.18	63
5	0.61	0.60	0.60	632
6	0.54	0.69	0.60	878
7	0.48	0.31	0.37	340
8	0.00	0.00	0.00	54
9	0.00	0.00	0.00	2
accuracy			0.55	1972
macro avg	0.31	0.24	0.25	1972
weighted avg	0.53	0.55	0.53	1972

Квадратный дискриминант

Верно определенные ко всем train: 0.5400949288033975

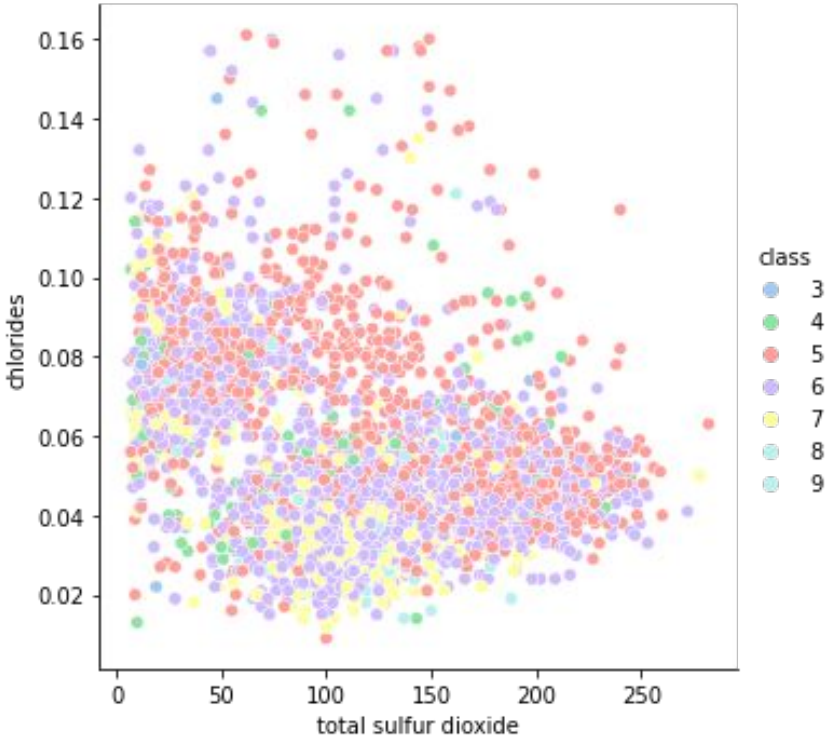
Верно определенные ко всем test : 0.5608519269776876

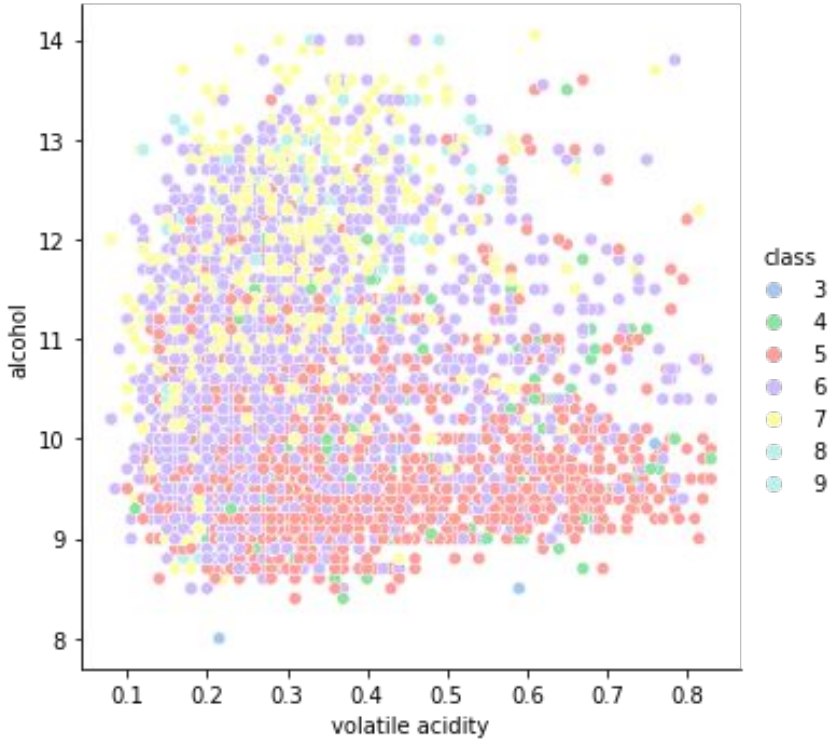
Наш score : 0.9513184584178499

	precision	recall	f1-score	support
3	0.50	0.33	0.40	3
4	0.42	0.21	0.28	63
5	0.61	0.61	0.61	632
6	0.57	0.59	0.58	878
7	0.47	0.55	0.51	340
8	0.38	0.06	0.10	54
9	1.00	1.00	1.00	2
accuracy			0.56	1972
macro avg	0.56	0.48	0.50	1972
weighted avg	0.56	0.56	0.55	1972

Метод ближайших к соседей, accuracy = 0.5241443975621191 , алгоритм auto , количество соседей 8 , параметр метрики Минковского 1

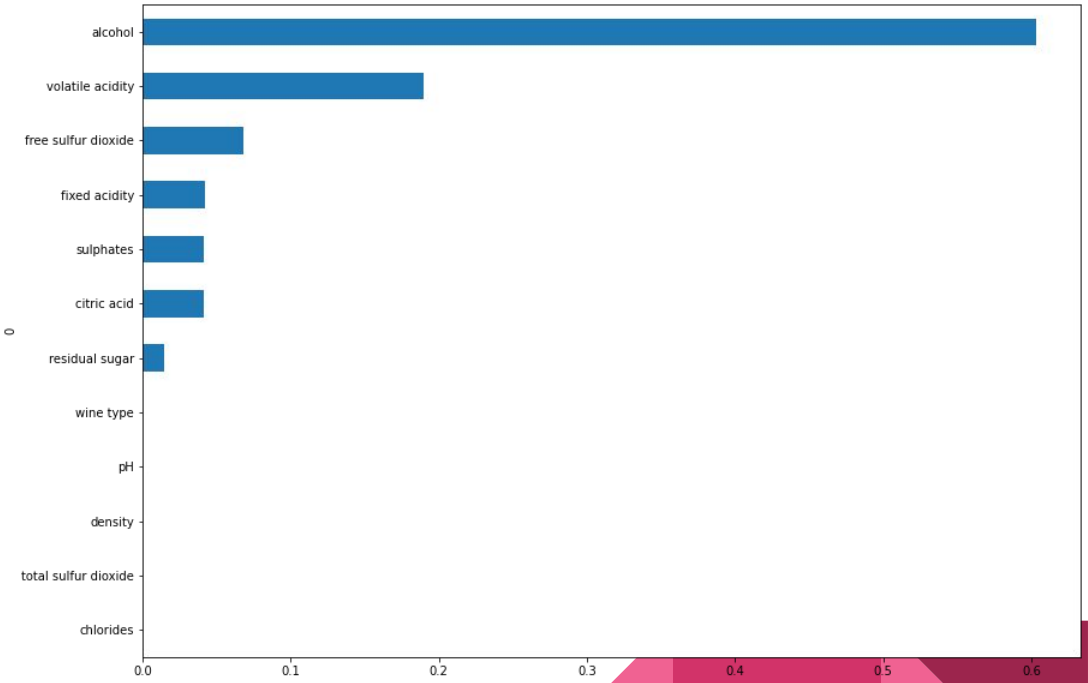
Распределение по качеству вина





Верно определенные ко всем train: 0.5491916859122402
Верно определенные ко всем test : 0.5419596812001876

	precision	recall	f1-score	support
3.0	0.00	0.00	0.00	8
4.0	0.00	0.00	0.00	71
5.0	0.58	0.60	0.59	689
6.0	0.53	0.65	0.58	934
7.0	0.54	0.35	0.42	371
8.0	0.22	0.11	0.14	57
9.0	0.00	0.00	0.00	3
accuracy			0.54	2133
macro avg	0.27	0.24	0.25	2133
weighted avg	0.52	0.54	0.52	2133



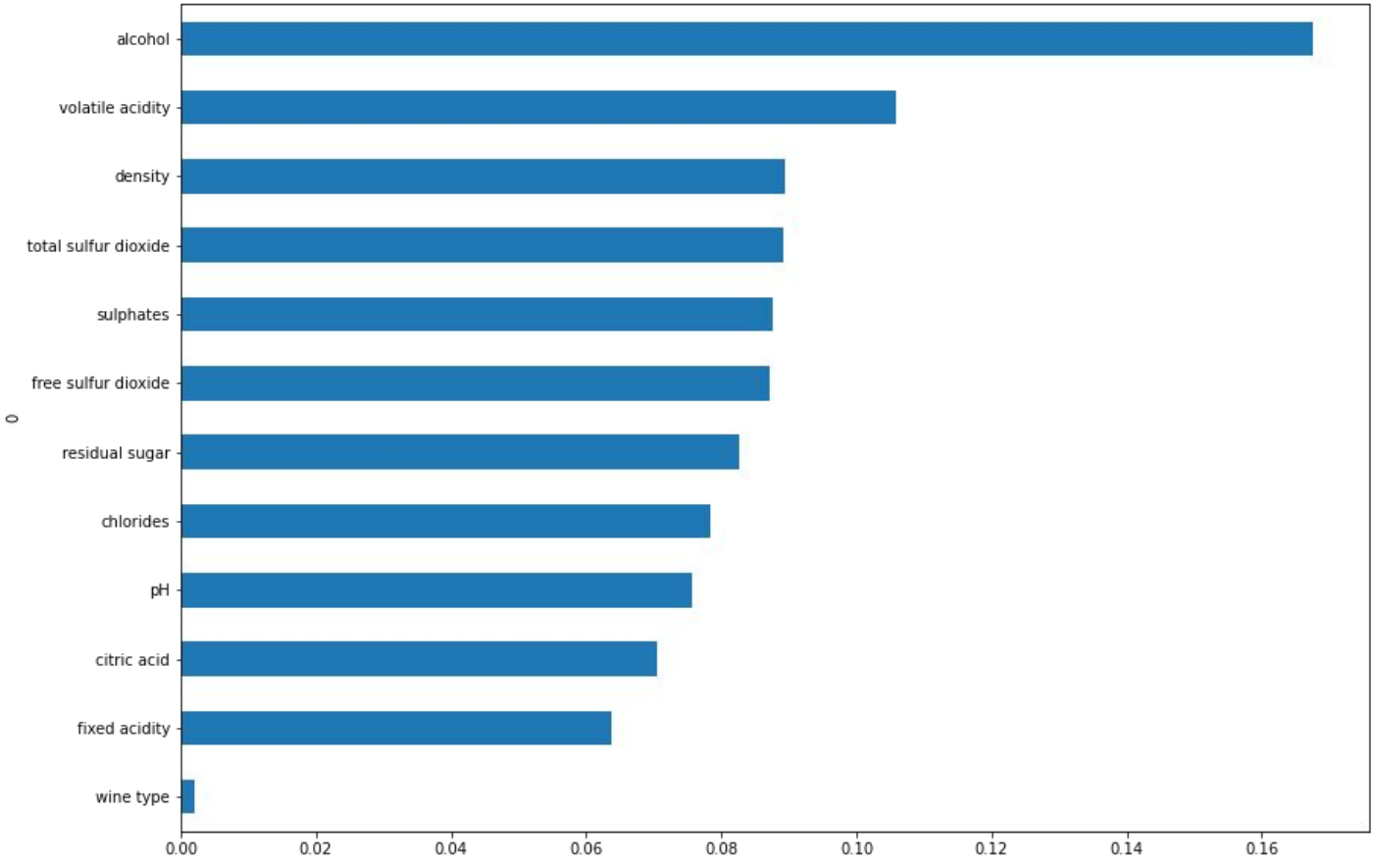
```python
# Поиск оптимальных параметров к деревьям.
# count_trees = 85
# rate = 0.05
# depth = 9

rate_list = [x / 1000.0 for x in range(50, 950, 50)]
best_result = 0
best_param = []

# Всего 10 классов => минимальная глубина 4.
for depth in range(4, 10):
    for count_trees in range(5, 100, 10):
        for rate in rate_list:
            clf = GradientBoostingClassifier(n_estimators = count_trees, learning_rate = rate, max_depth = depth).fit(X_train, y_train)
            if clf.score(X_test , y_test ) > best_result:
                best_result = clf.score(X_test , y_test )
                best_param = [depth, count_trees, rate]

# print(f'Верно определенные ко всем train: {clf.score(X_train, y_train)}')
# print(f'Верно определенные ко всем test : {clf.score(X_test , y_test )}')
print("best result = " + str(best_result), best_param)
# best result = [9, 85, 0.05]
```

Верно определенные ко всем train: 0.997459584295612
Верно определенные ко всем test : 0.6432255039849977
Наш score : 0.9554617909048289

	precision	recall	f1-score	support
3.0	0.00	0.00	0.00	8
4.0	0.44	0.15	0.23	71
5.0	0.67	0.68	0.67	689
6.0	0.63	0.74	0.68	934
7.0	0.68	0.51	0.59	371
8.0	0.43	0.26	0.33	57
9.0	0.00	0.00	0.00	3
accuracy			0.64	2133
macro avg	0.41	0.34	0.36	2133
weighted avg	0.64	0.64	0.63	2133

Верно определенные ко всем train: 0.9987509367974019
Верно определенные ко всем test : 0.6658215010141988
Наш score : 0.9594320486815415

	precision	recall	f1-score	support
3.0	0.00	0.00	0.00	3
4.0	0.59	0.16	0.25	63
5.0	0.68	0.72	0.70	632
6.0	0.65	0.73	0.69	878
7.0	0.70	0.56	0.62	340
8.0	0.66	0.35	0.46	54
9.0	0.00	0.00	0.00	2
accuracy			0.67	1972
macro avg	0.47	0.36	0.39	1972
weighted avg	0.67	0.67	0.66	1972


alcohol
volatile acidity
density
total sulfur dioxide
sulphates
free sulfur dioxide
residual sugar
chlorides
pH
citric acid
fixed acidity
wine type
0
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16

Ресурсы:

- **https://scikit-learn.org**