ИТМО

Лабораторная работа
Анализ данных о качестве вина

Выполнила:
Замерова Полина

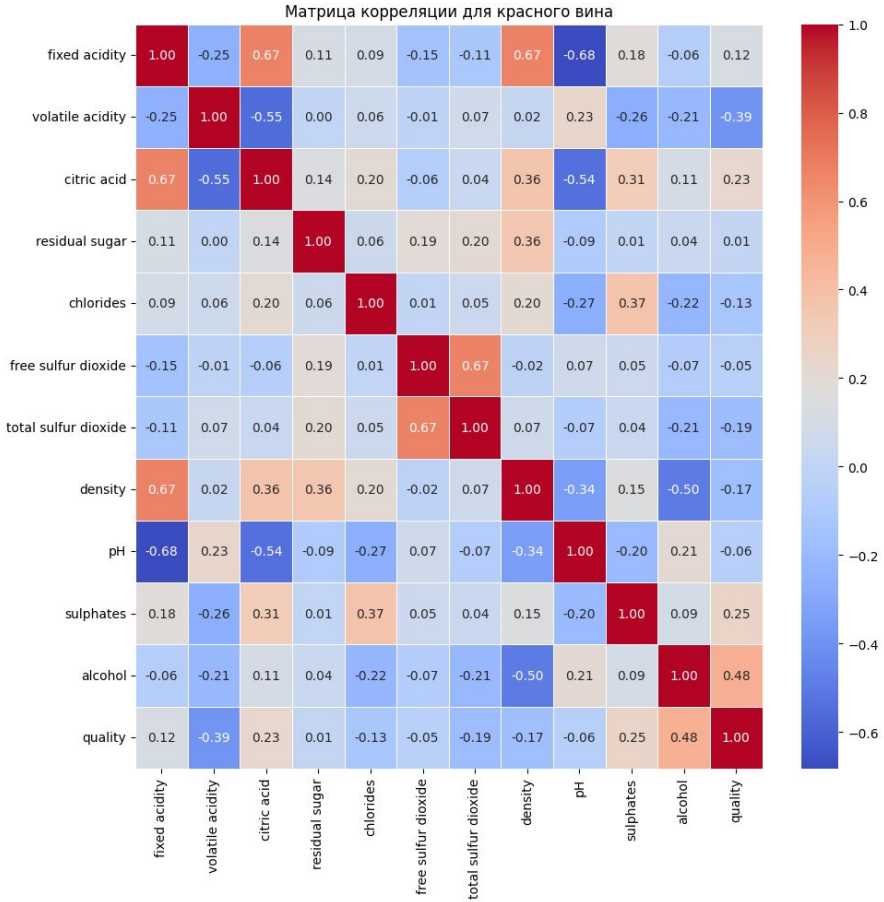
Цель и задачи

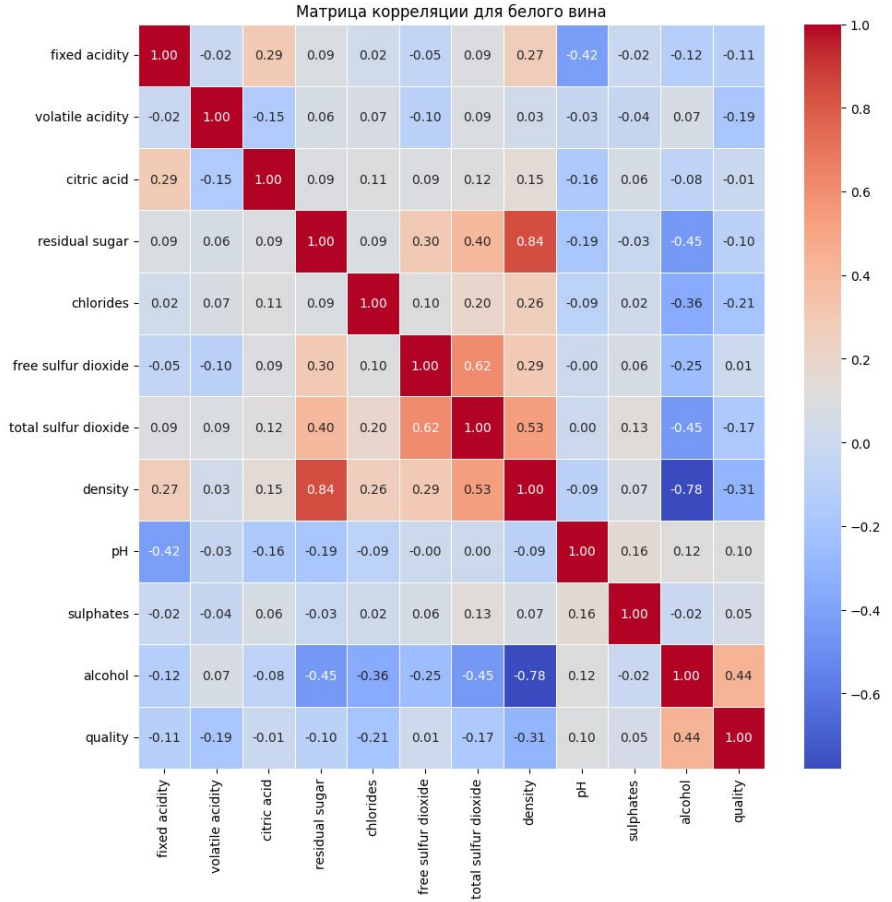
Цель:
Использовать методы анализа данных и машинного обучения для изучения химических характеристик вин и их связи с качеством. Выполнить задачи по прогнозированию, кластеризации и визуализации данных о вине.

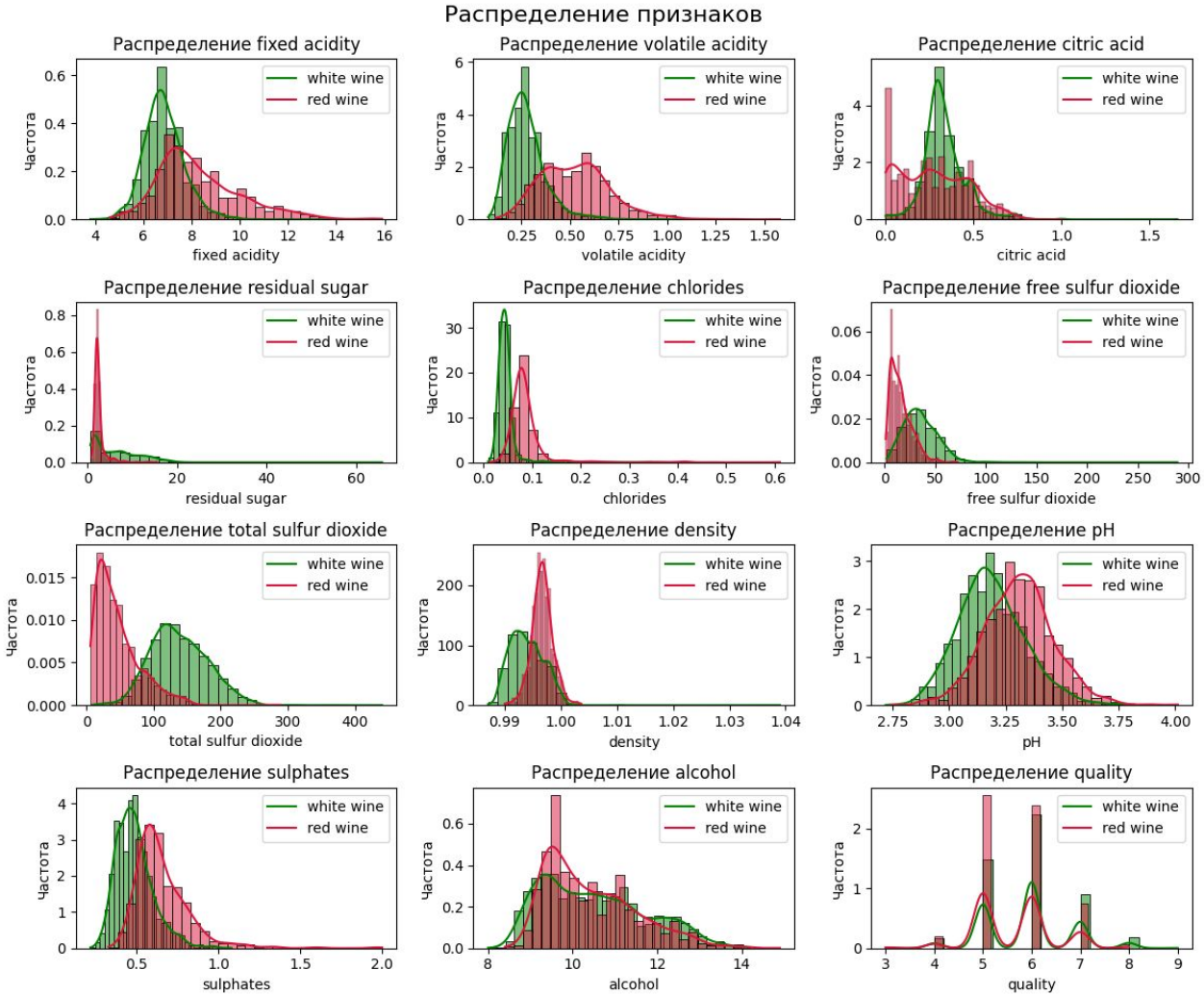
Задачи:

1. Изучение структуры данных и предварительная обработка.
2. Визуализация влияния химических характеристик на качество вина.
3. Классификация вина по его качеству.
4. Прогнозирование рейтинга вина (регрессия).
5. Кластеризация вин по их химическим характеристикам.
6. Оптимизация рецепта вина с учетом целевого качества.
7. Проведение экспериментов для оценки моделей.

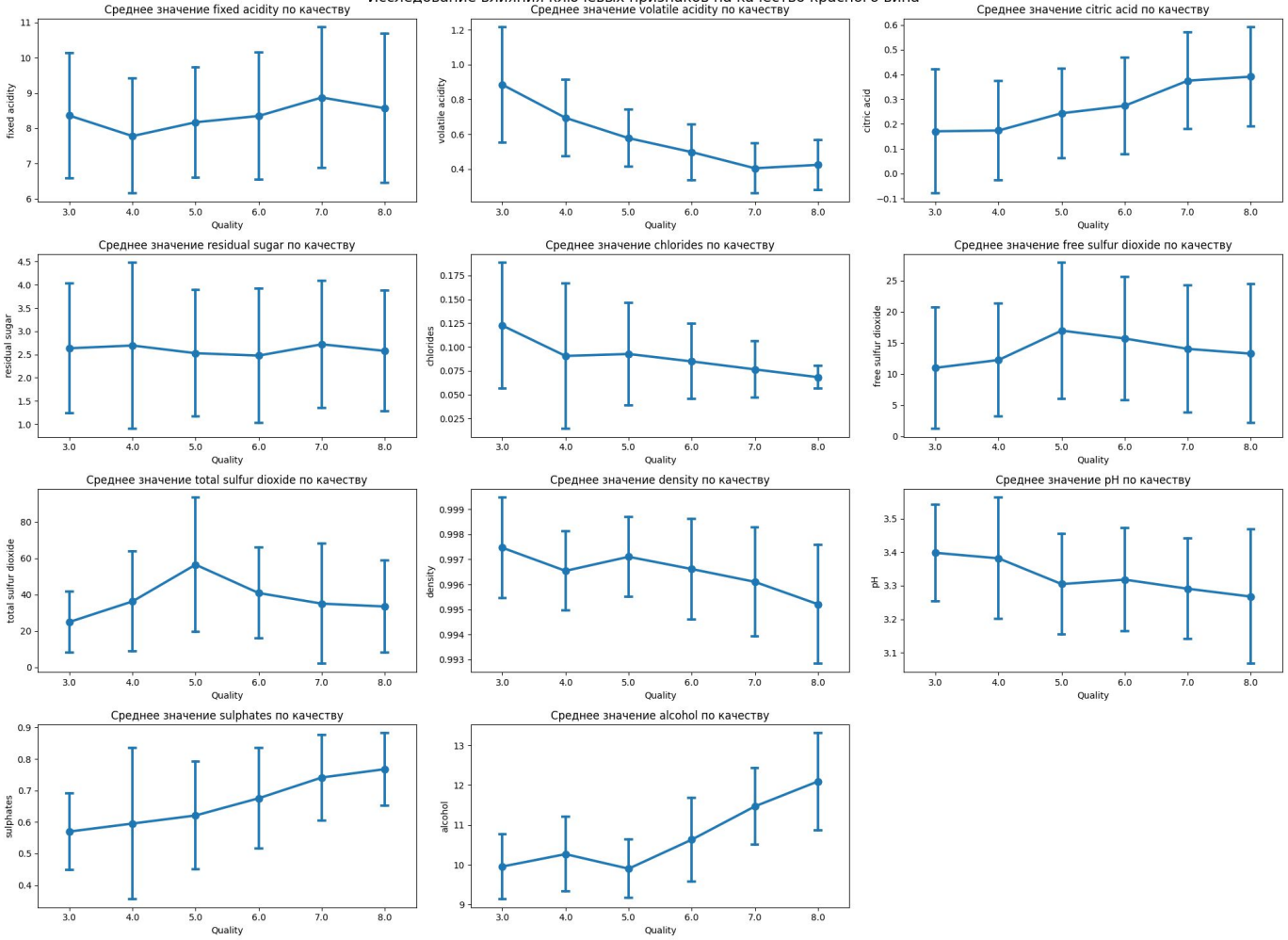
Корреляционные матрицы



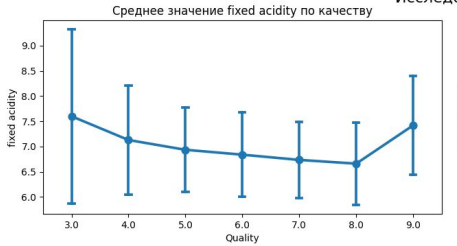


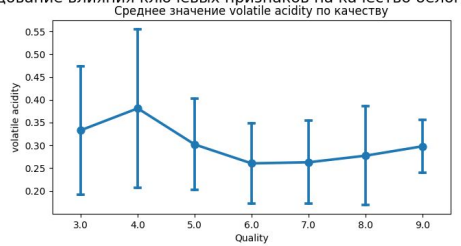

Распределение признаков
Распределение fixed acidity
Распределение volatile acidity
Распределение citric acid
Распределение residual sugar
Распределение chlorides
Распределение free sulfur dioxide
Распределение total sulfur dioxide
Распределение density
Распределение pH
Распределение sulphates
Распределение alcohol
Распределение quality
white wine
red wine
Частота
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality

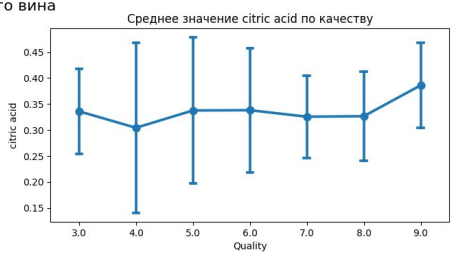
Исследование влияния ключевых признаков на качество красного вина

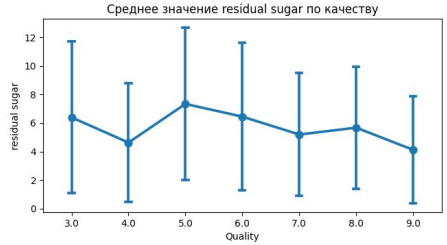


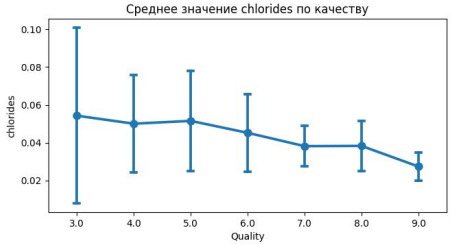
Исследование влияния ключевых признаков на качество белого вина

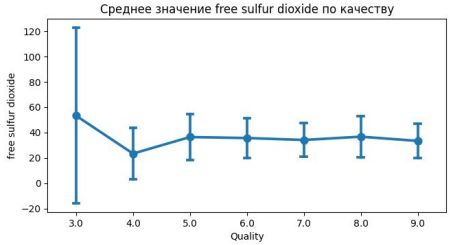


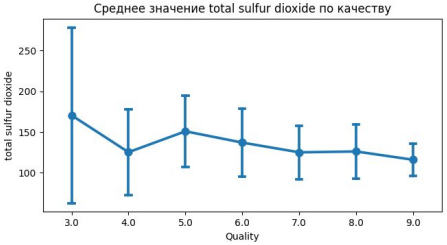


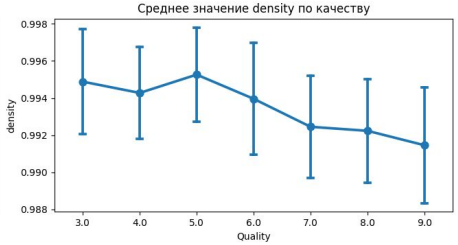


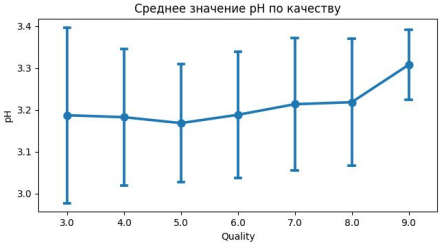


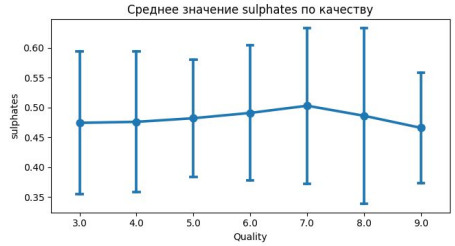


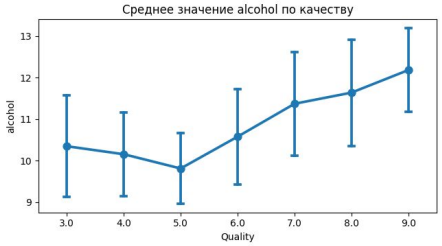












Классификация вина по его качеству

- Разделили выборку по типу вин: красные и белые
- Для каждого типа обучили модель RandomForest
- Подобрали гиперпараметры с помощью GridSearchCV

Лучшие значения гиперпараметров:

- ❏ max_depth: 20
- ❏ min_samples_split: 2
- ❏ n_estimators: 100

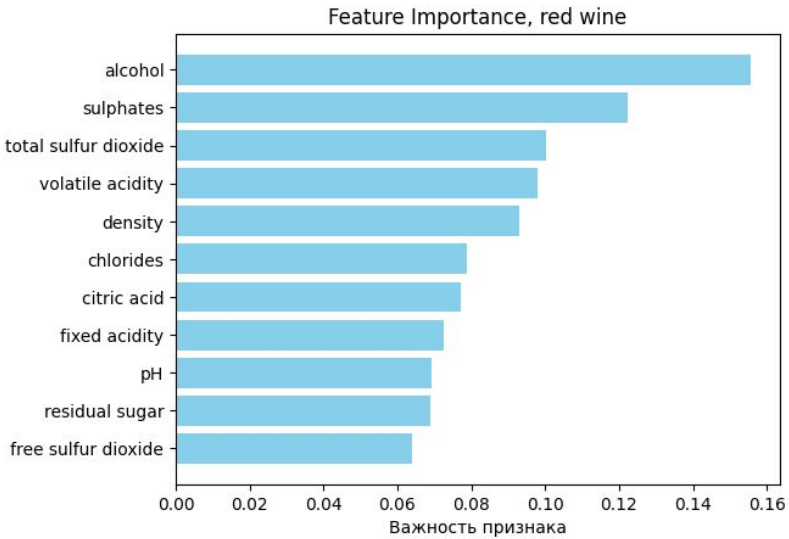
Отчет классификации красного вина:

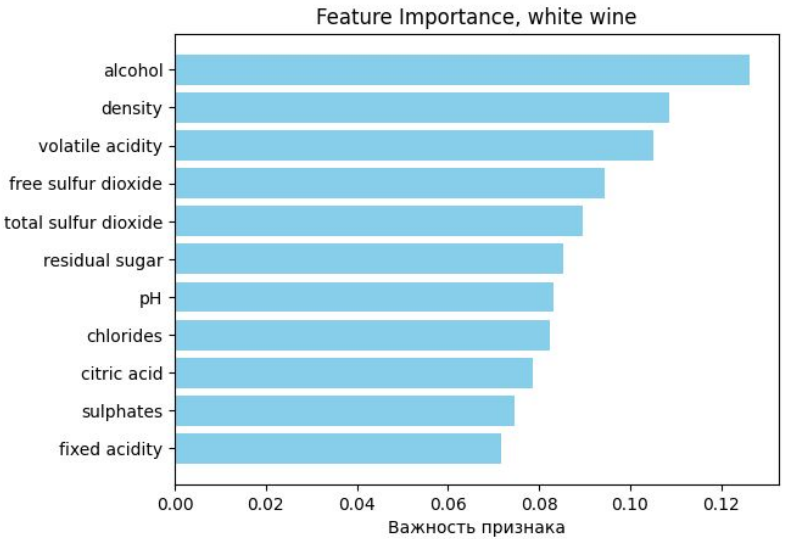
	precision	recall	f1-score	support
high	0.67	0.51	0.58	47
low	0.77	0.81	0.79	141
medium	0.64	0.65	0.64	132
accuracy			0.70	320
macro avg	0.69	0.66	0.67	320
weighted avg	0.70	0.70	0.70	320

Отчет классификации белого вина:

	precision	recall	f1-score	support
high	0.83	0.68	0.75	227
low	0.76	0.74	0.75	321
medium	0.70	0.78	0.74	432
accuracy			0.74	980
macro avg	0.77	0.74	0.75	980
weighted avg	0.75	0.74	0.75	980

Важность признаков





Прогнозирование рейтинга вина (регрессия)

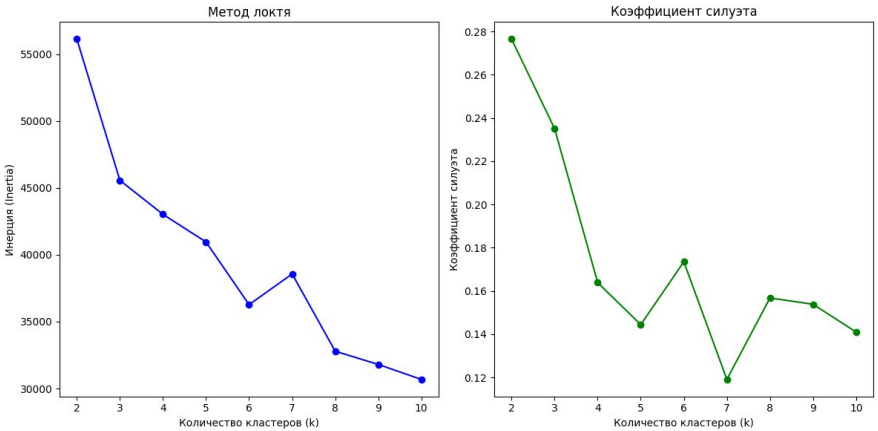
Ошибка предсказания градиентного бустинга

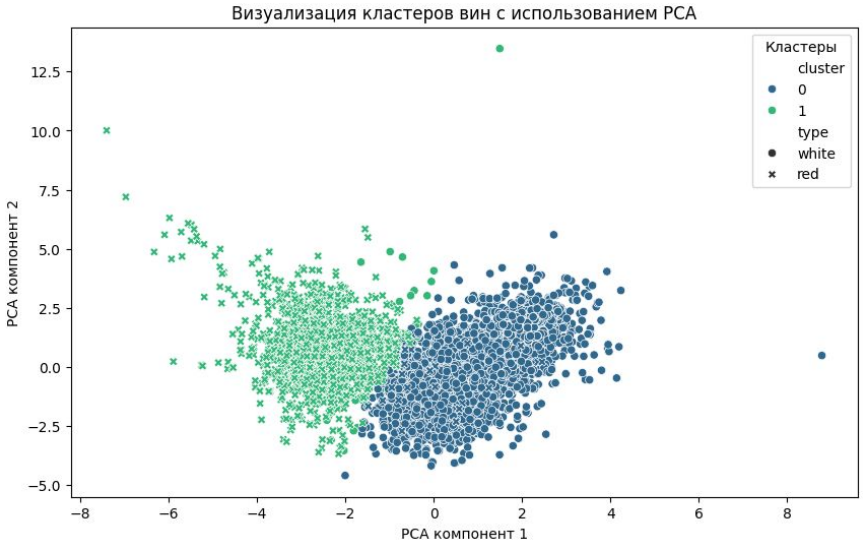
	MAE	RMSE
Red Wine	0.48	0.60
White Wine	0.53	0.68

В результате подбора гиперпараметров градиентного бустинга удалось уменьшить ошибку предсказания

	MAE	RMSE
Gradient Boosting (tuned)	0.42	0.58
Gradient Boosting	0.51	0.65
Linear Regression	0.55	0.69

Кластеризация вин по химическим характеристикам





- Для разных значений k (количество кластеров) была вычислена инерция и коэффициент силуэта
- Выбрано количество кластеров k=2
- Данные разделяются на две естественные группы: красные и белые вина

Оптимизация рецепта вина

В качестве оптимизируемой функции мы использовали модель градиентного бустинга, предсказывающую рейтинг вина по его характеристикам.

Методы оптимизации:

1. scipy.optimize.minimize (L-BFGS-B)
2. Дифференциальная эволюция

Ограничения аргументов были выбраны как максимальные и минимальные значения, наблюдаемые в датасете.

Оптимальные характеристики для белого вина:

	result
fixed acidity	7.828519
volatile acidity	0.512841
citric acid	0.080381
residual sugar	11.201799
chlorides	0.018568
free sulfur dioxide	63.204572
total sulfur dioxide	127.537141
density	0.987899
pH	3.819427
sulphates	0.768919
alcohol	12.502202
type_white	1.000000

Рейтинг белого вина с оптимальными характеристиками: **10.58**

Оптимальные характеристики для красного вина:

	result
fixed acidity	9.184025
volatile acidity	0.125134
citric acid	0.310193
residual sugar	3.567025
chlorides	0.176101
free sulfur dioxide	60.913562
total sulfur dioxide	99.692642
density	0.990545
pH	4.007609
sulphates	1.872693
alcohol	8.635467
type_white	0.000000

Рейтинг красного вина с оптимальными характеристиками: **10.16**

Веб-интерфейс для модели

Deploy

Характеристики вина:

Fixed Acidity
7.22
3.80 — 15.90

Volatile Acidity
0.34
0.08 — 1.58

Citric Acid
0.32
0.00 — 1.66

Residual Sugar
5.44
0.60 — 65.80

Chlorides
0.06
0.01 — 0.61

Free Sulfur Dioxide
66.36
1.00 — 289.00

Total Sulfur Dioxide
115.74
6.00 — 440.00

Density
0.99

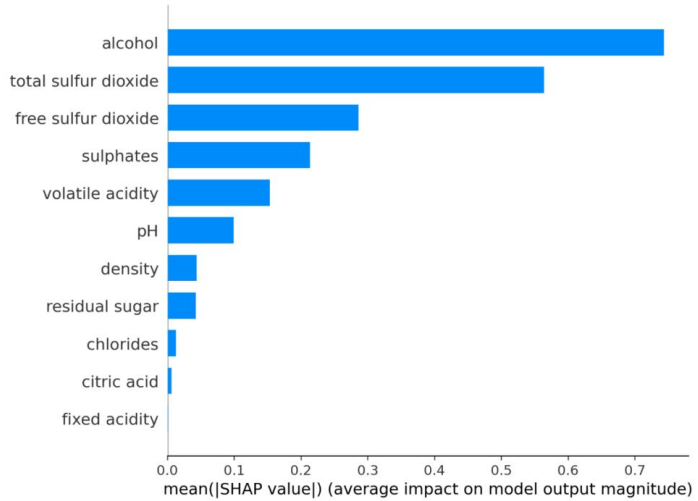
Прогнозировать

Прогнозируемое качество вина: 6.61

Выберите тип визуализации:

Bar Plot

Bar Plot:



Спасибо
за внимание!

ITMO more than a UNIVERSITY