Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

*Дисциплина «Технологии машинного обучения»*

**Отчёт**

по лабораторной работе №4

«Подготовка обучающей и тестовой выборки, кросс-валидация и подбор гиперпараметров на примере метода ближайших соседей»

*Вариант 6*

Студент:

Дыньков А.Д.

Группа ИУ5-62Б

Преподаватель:

Гапанюк Ю.Е.

Москва, 2020 г.

**Цель лабораторной работы:**

Изучение сложных способов подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.

**Задание:**

1. Выберите набор данных (датасет) для решения задачи классификации или регрессии.
2. С использованием метода train\_test\_split разделите выборку на обучающую и тестовую.
3. Обучите модель ближайших соседей для произвольно заданного гиперпараметра K. Оцените качество модели с помощью подходящих для задачи метрик.
4. Постройте модель и оцените качество модели с использованием кросс-валидации.
5. Произведите подбор гиперпараметра K с использованием GridSearchCV и кросс-валидации.

**Выполнение работы:**

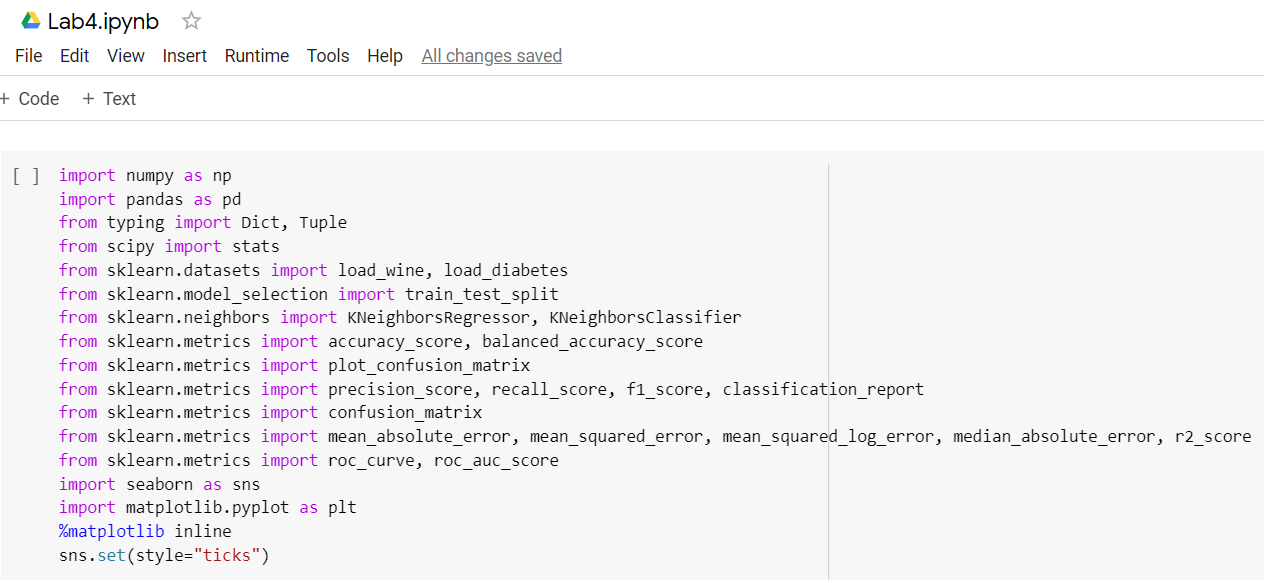
Выбранный датасет:

Решение задачи классификации

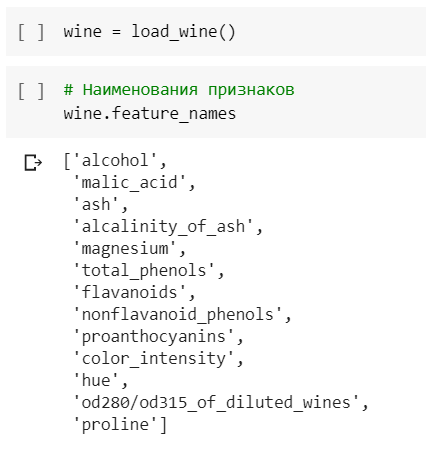


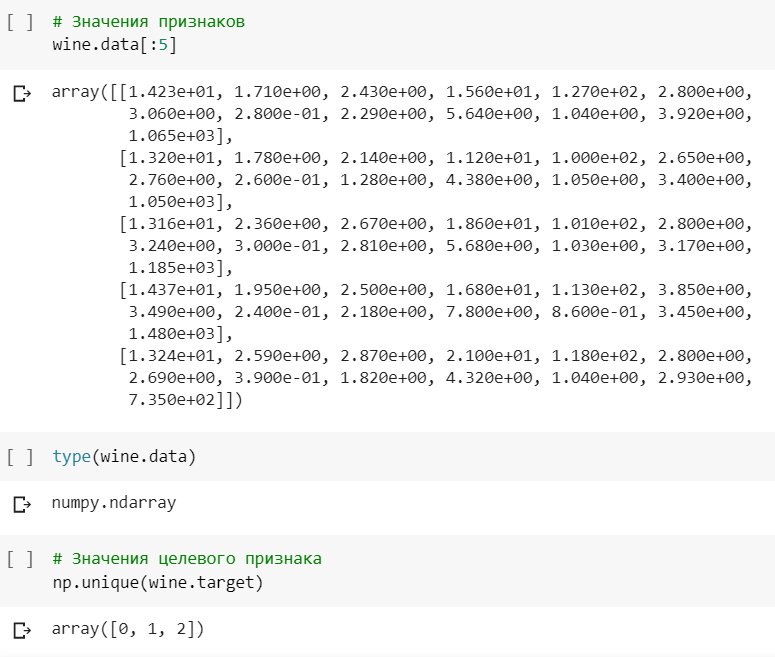
[https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\_wine.html#sklearn.datasets.load\_wine](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_wine.html" \l "sklearn.datasets.load_wine)

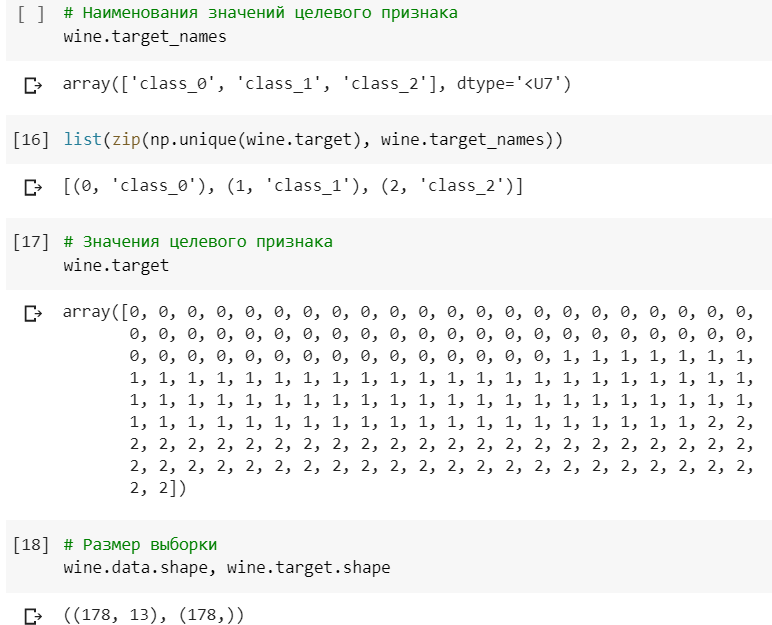
1. **Изучение качества классификации**



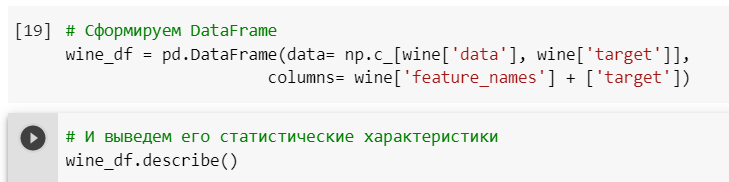
* 1. Подготовка данных и построение базовых моделей для оценки качества

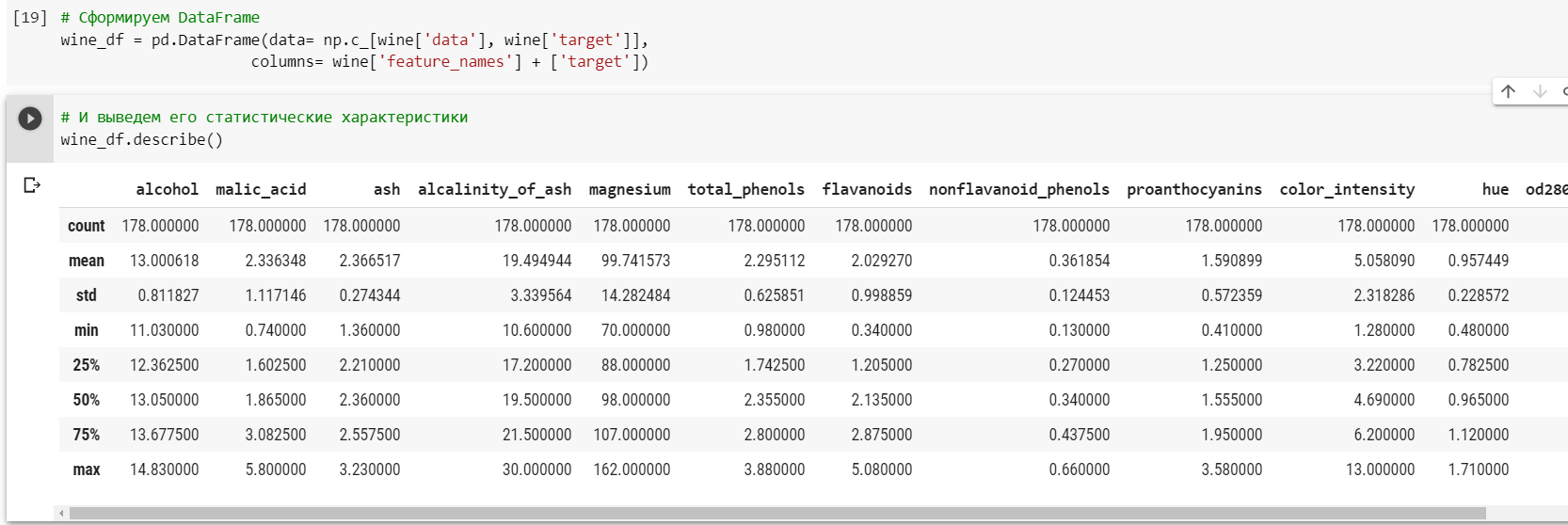




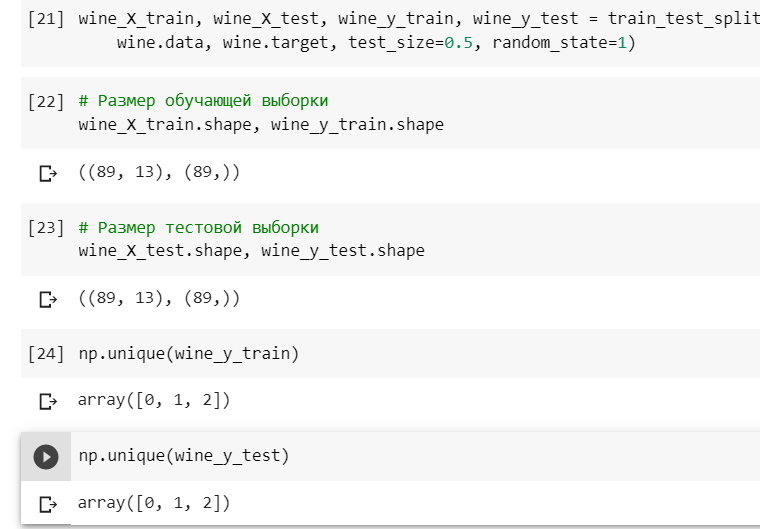


Формируем DataFrame:

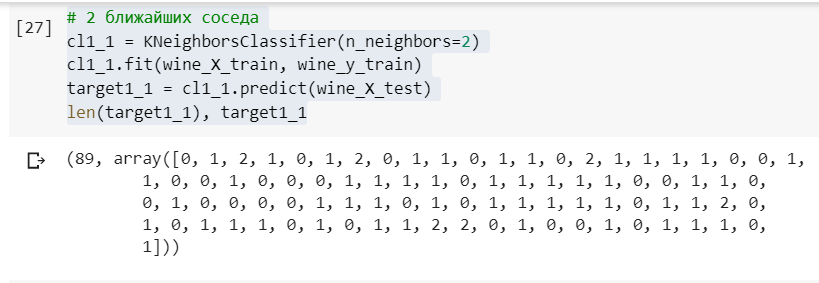


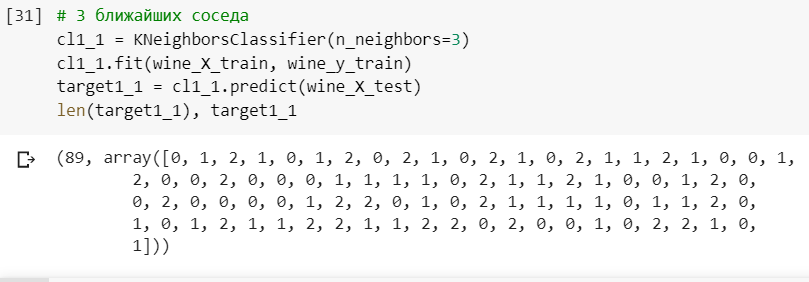


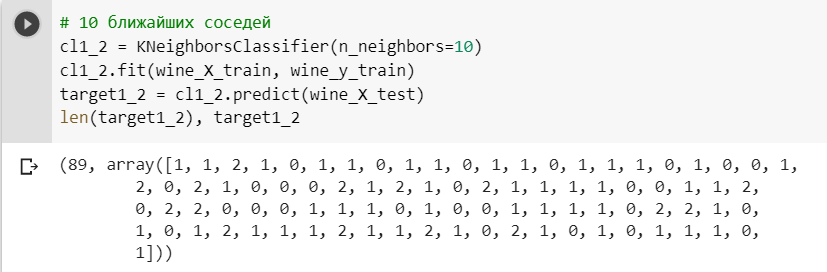
* 1. Разделение выборки на обучающую и тестовую



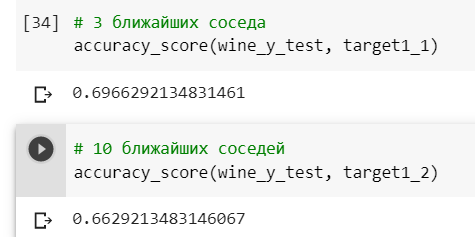
* 1. Построим базовые модели на основе метода ближайших соседей



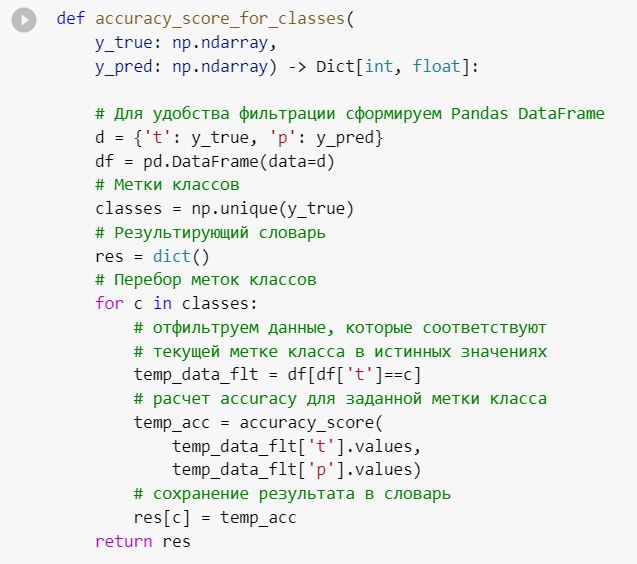
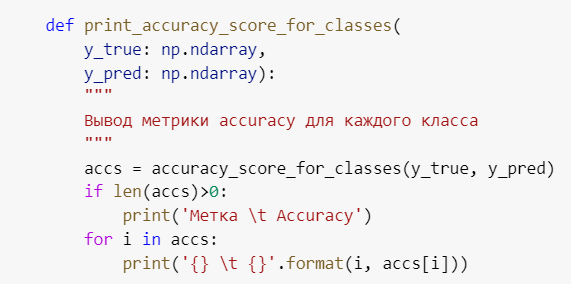


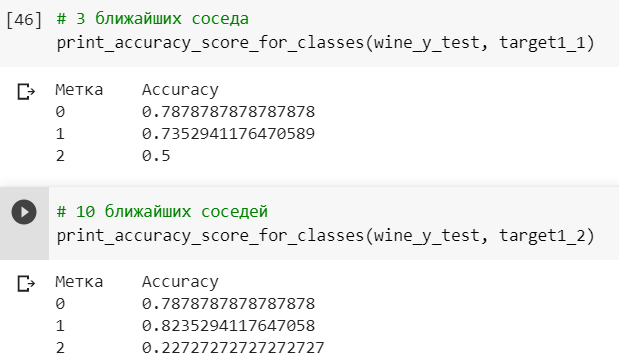


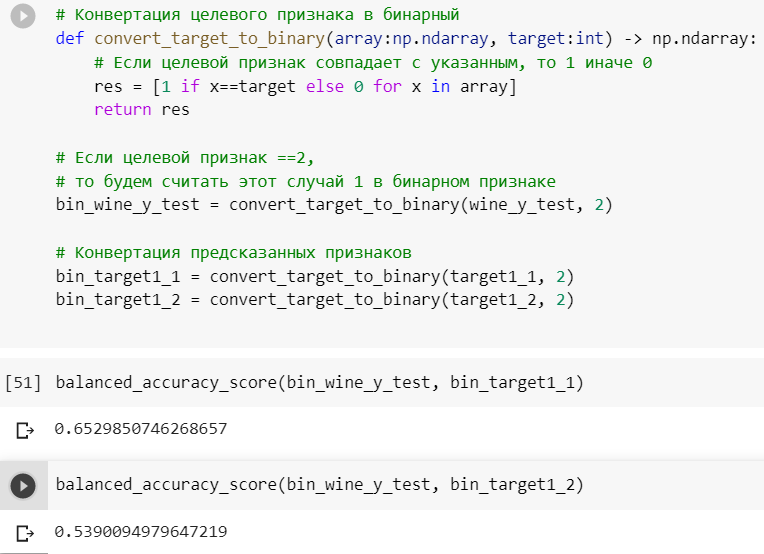
1. **Метрики качества классификации**
   1. Accuracy



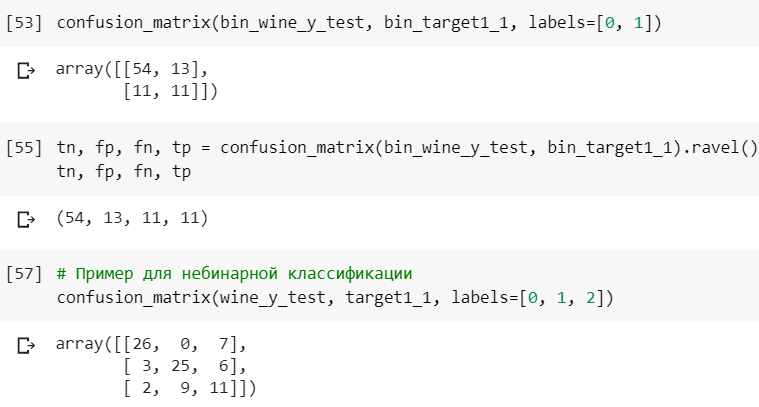
Точность в случае 10 ближайших соседей составляет более 66%, а точность в случае 2 ближайших соседей составляет более 69%.

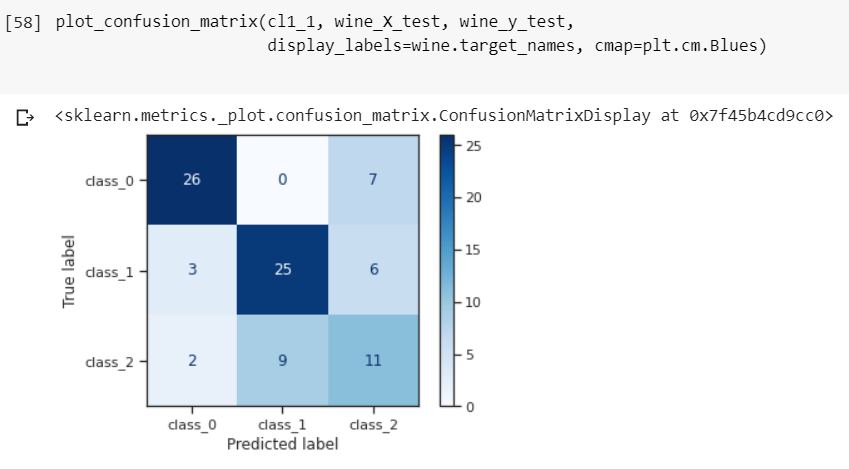
 

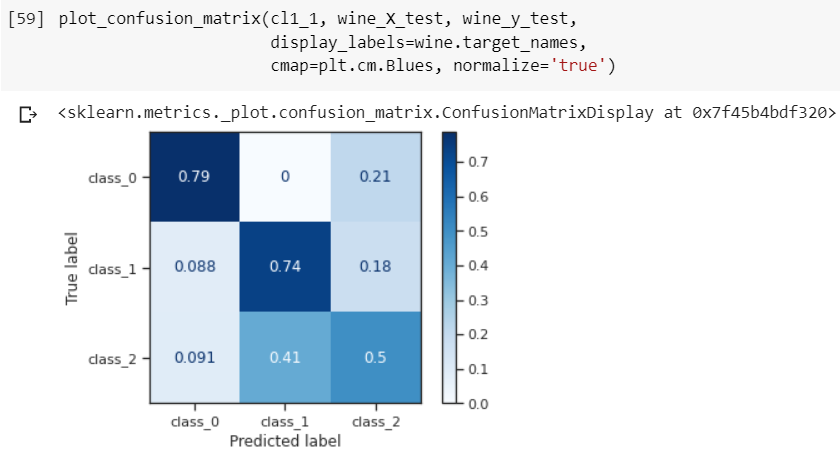


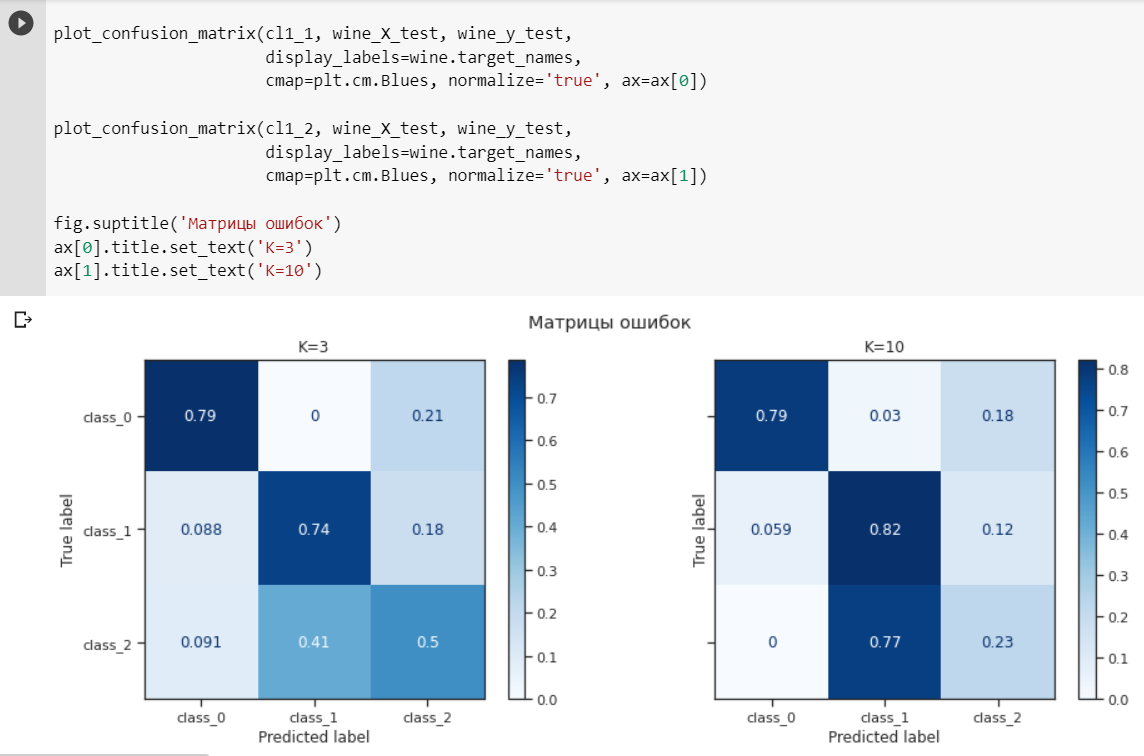


* 1. Матрица ошибок или ConfusionMatrix

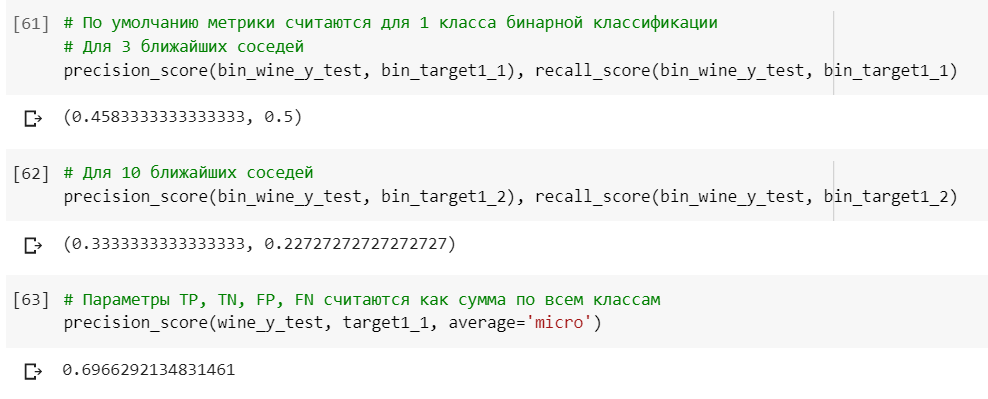


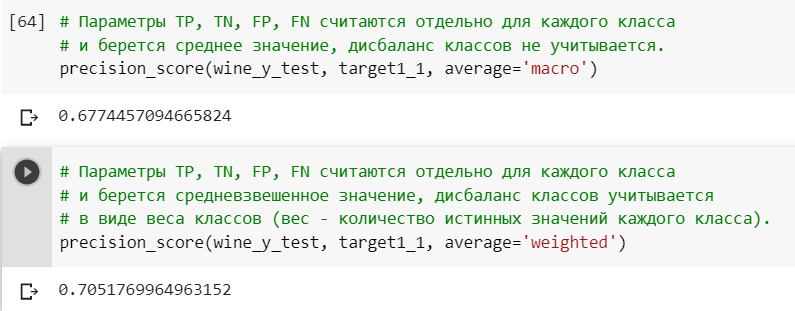


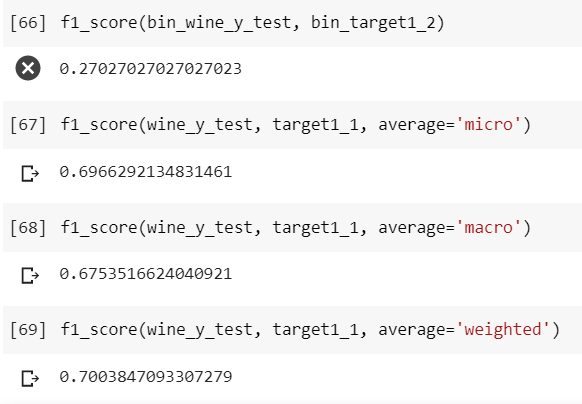


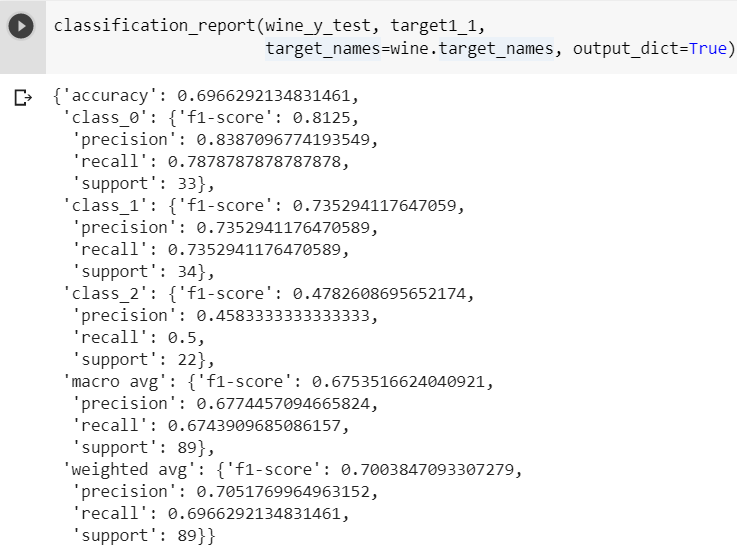


* 1. Precision, recall и F-мера

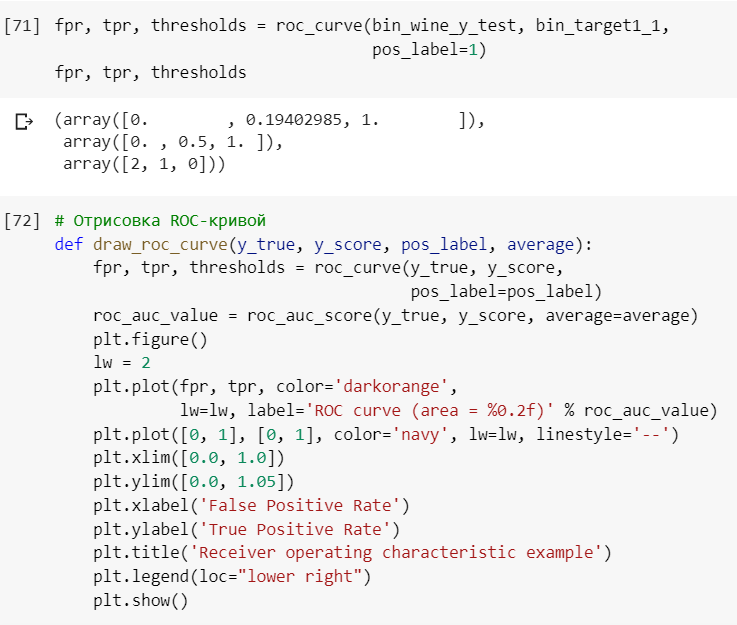


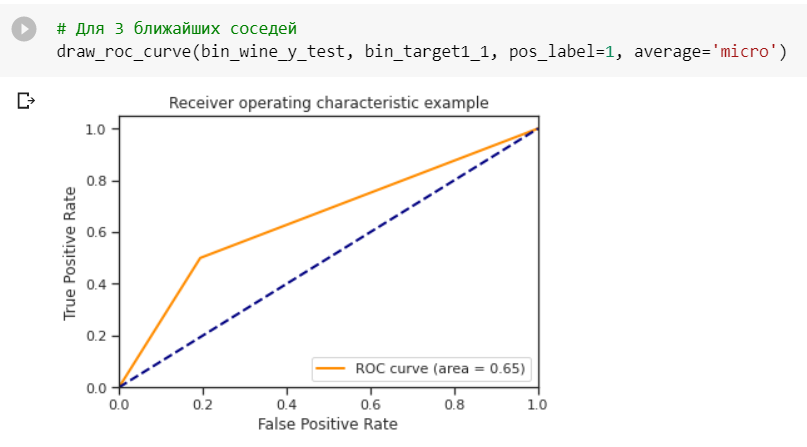


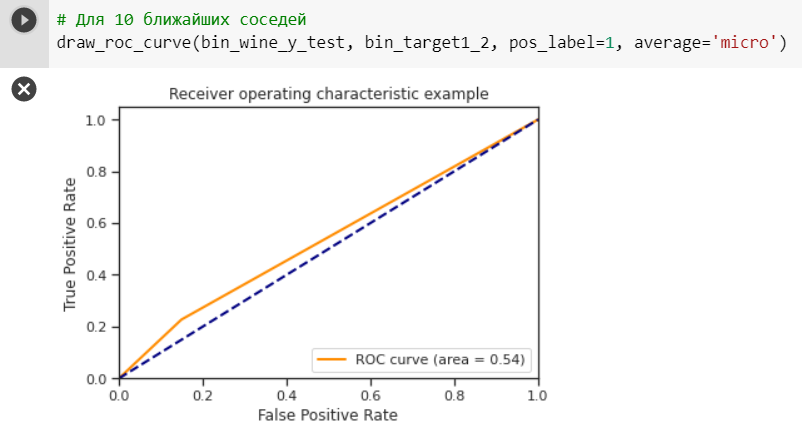




* 1. ROC-кривая и ROC AUC

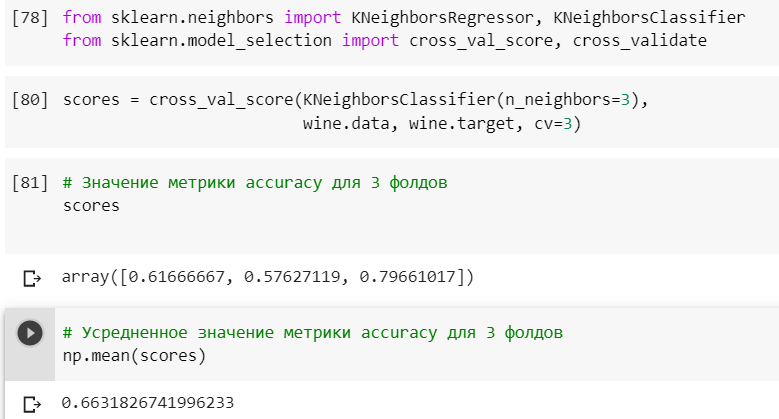




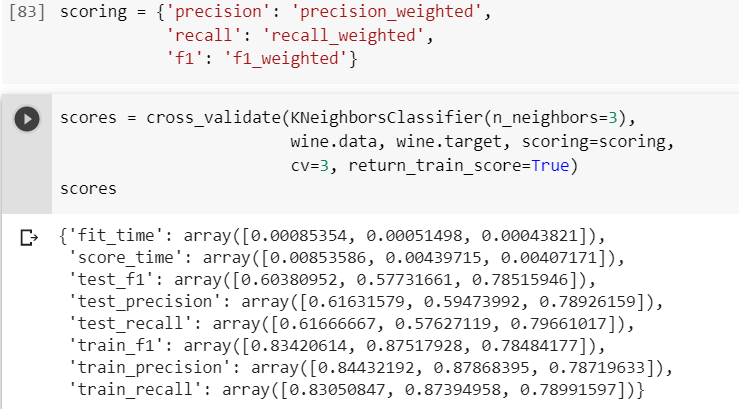


1. **Разбиение выборки на k частей с помощью кросс-валидации.**

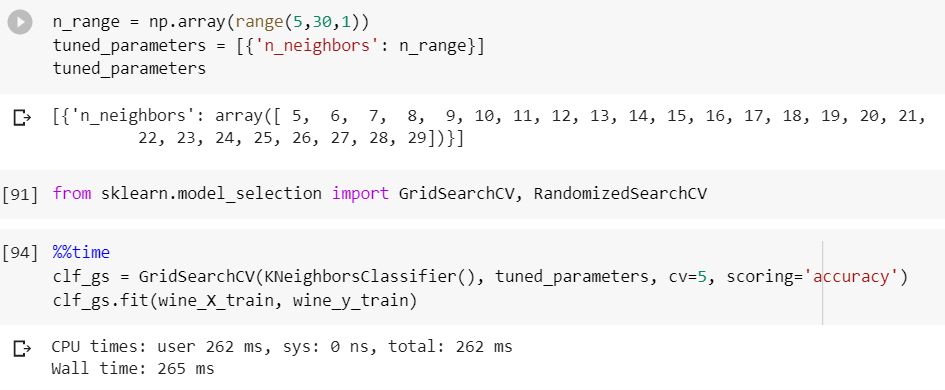
Наиболее простым способом кросс-валидации является вызов функции cross\_val\_score. В этом случае стратегия кросс-валидации определяется автоматически.

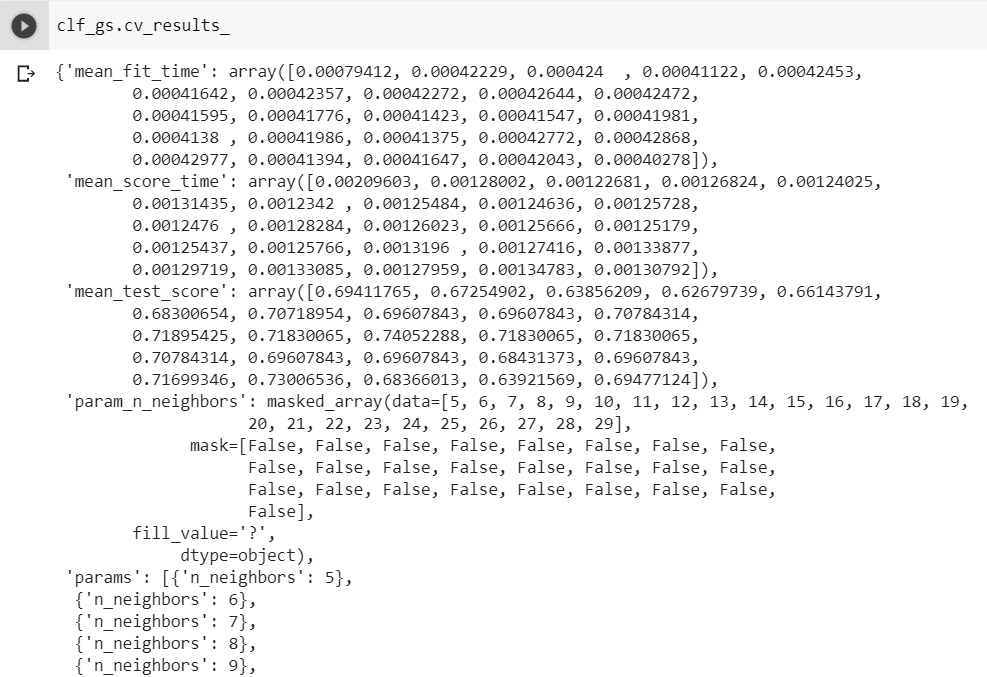


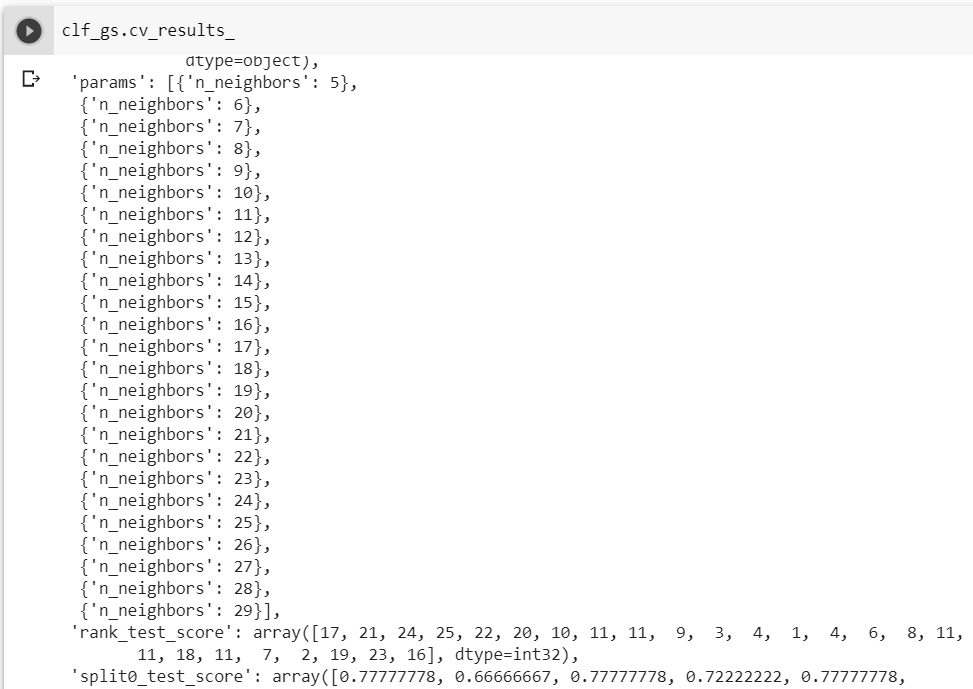
В отличие от функции cross\_val\_score, функция cross\_validate позволяет использовать для оценки несколько метрик и возращает более детальную информацию.

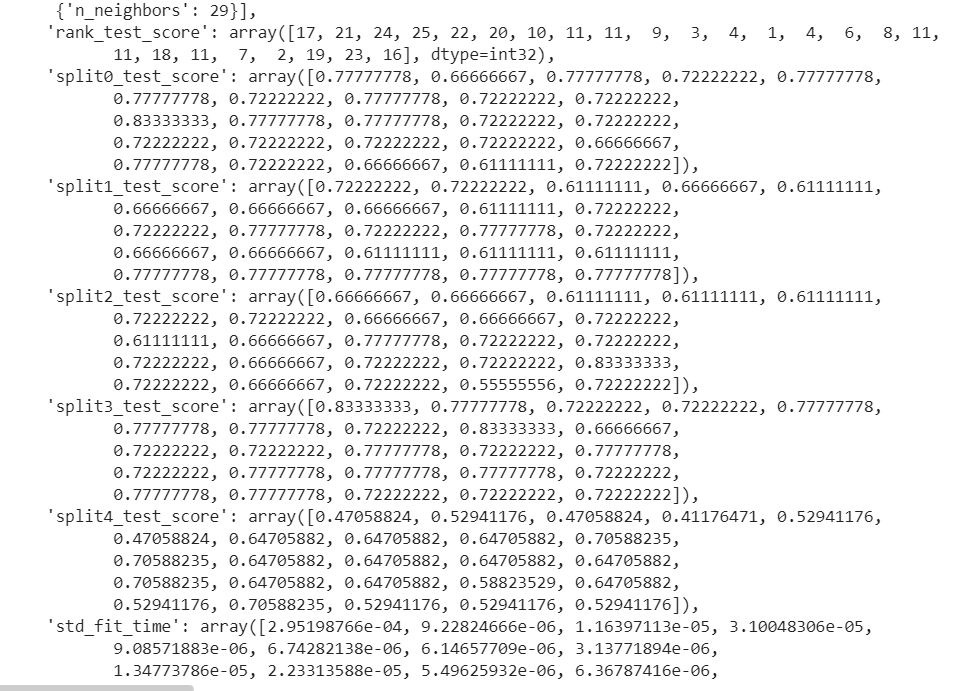


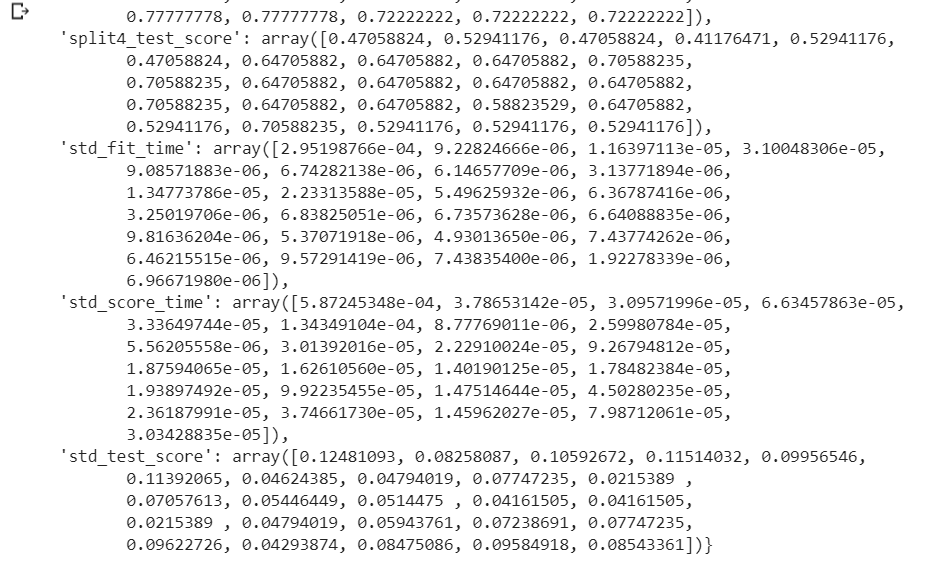
1. **Нахождение наилучшего гиперпараметра K с использованием GridSearchCV и кросс-валидации**

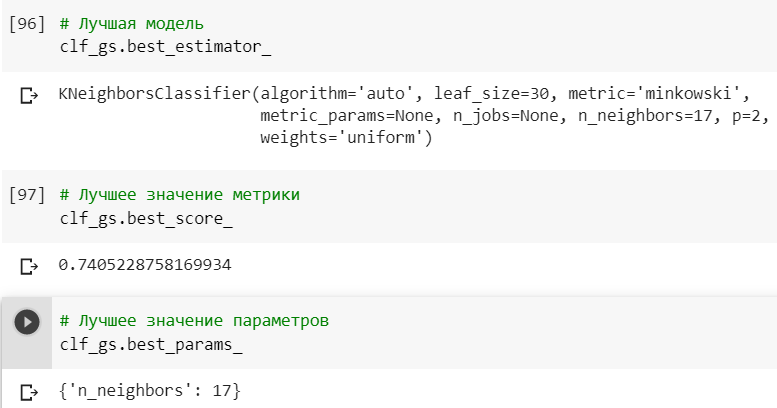












Лучшее значение гиперпараметра =17

Лучшее значение метрики при этом гиперпараметре = 0.7405228758169934

