МГТУ им. Н. Э. Баумана, кафедра ИУ5

курс “Технологии машинного обучения”

Лабораторная работа №4

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

ВЫПОЛНИЛ:

Ерохин И.А.

Группа: ИУ5-61Б

ПРОВЕРИЛ:

Гапанюк Ю.Е.

Москва 2020

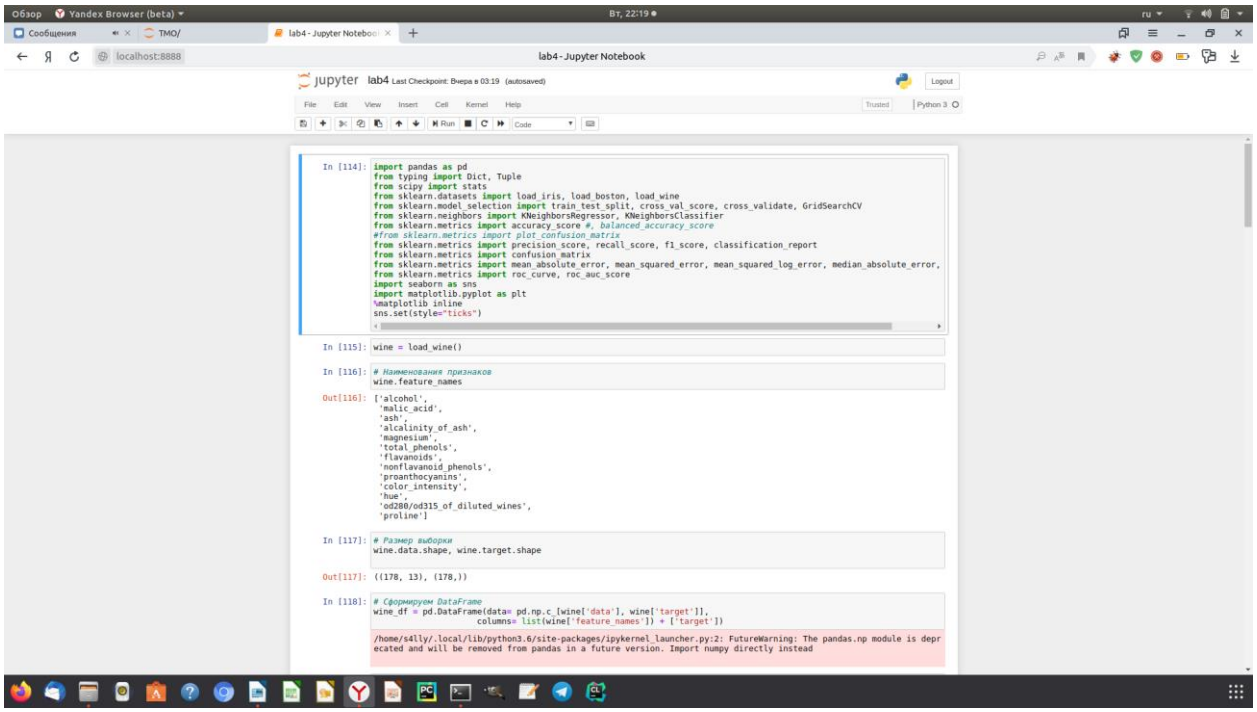
Цель лабораторной работы: изучение сложных способов подготовки выборки и подбора гиперпараметров на примере метода ближайших соседей.

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.

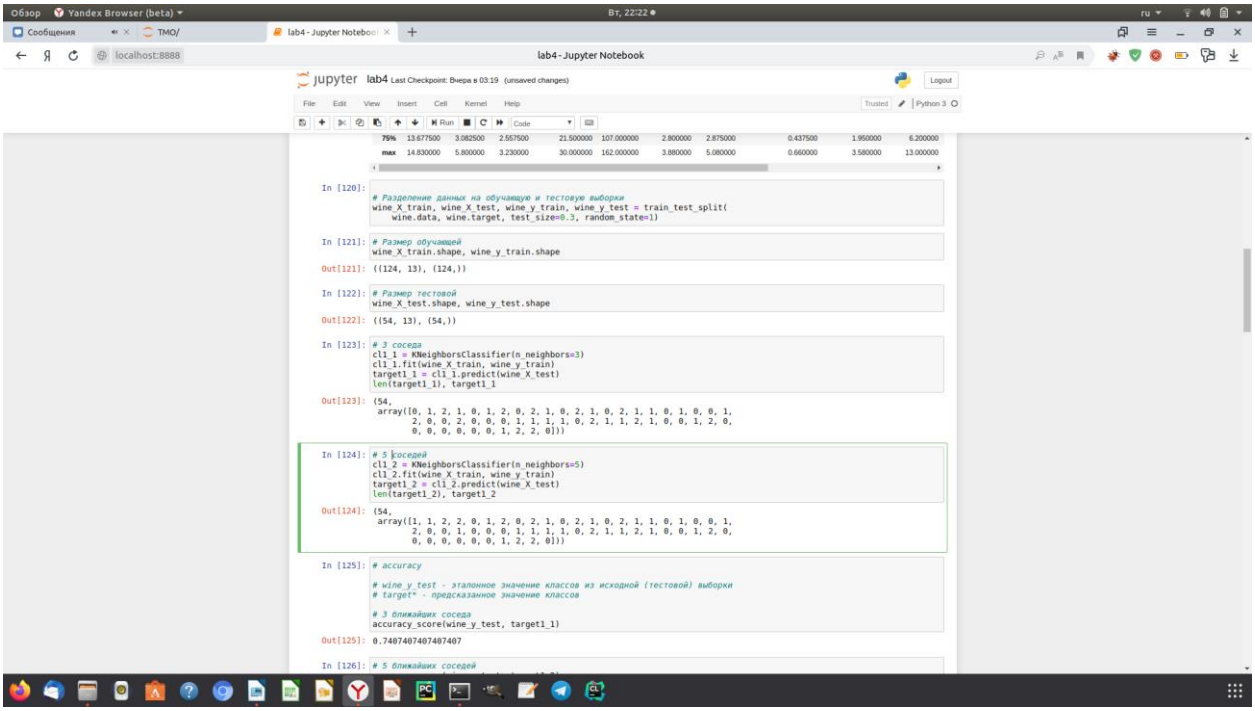
Выполненная работа:

Загрузка и первичный анализ данных

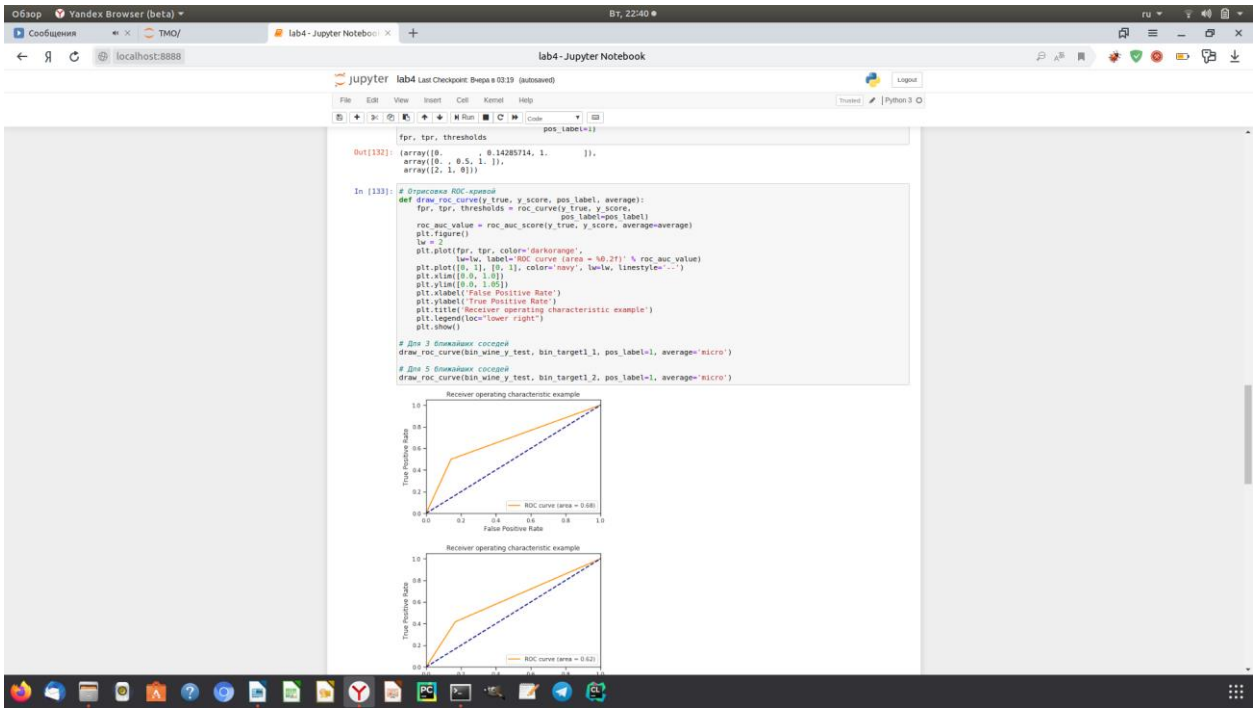


Разделение данных на обучающую и тестовую выборки и построение базовых моделей на основе метода ближайших соседей и подсчет метрик качества классификации

1) Accuracy Precision, recall и F-мера

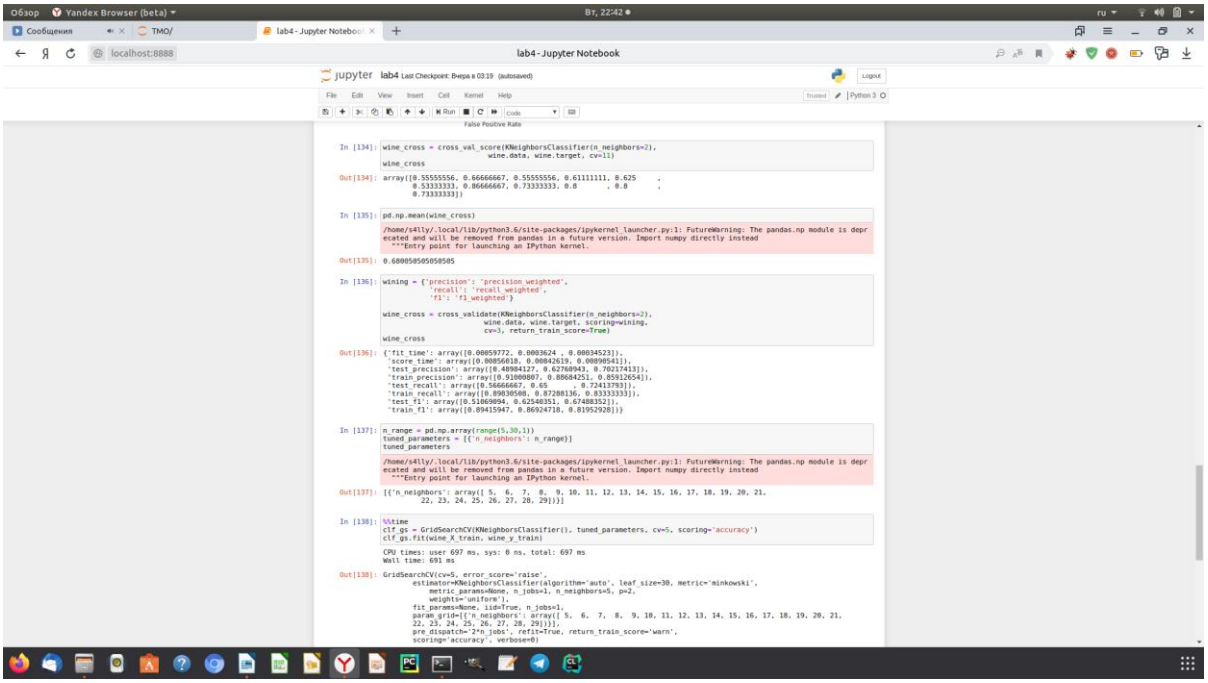


2) ROC-кривая и ROC AUC

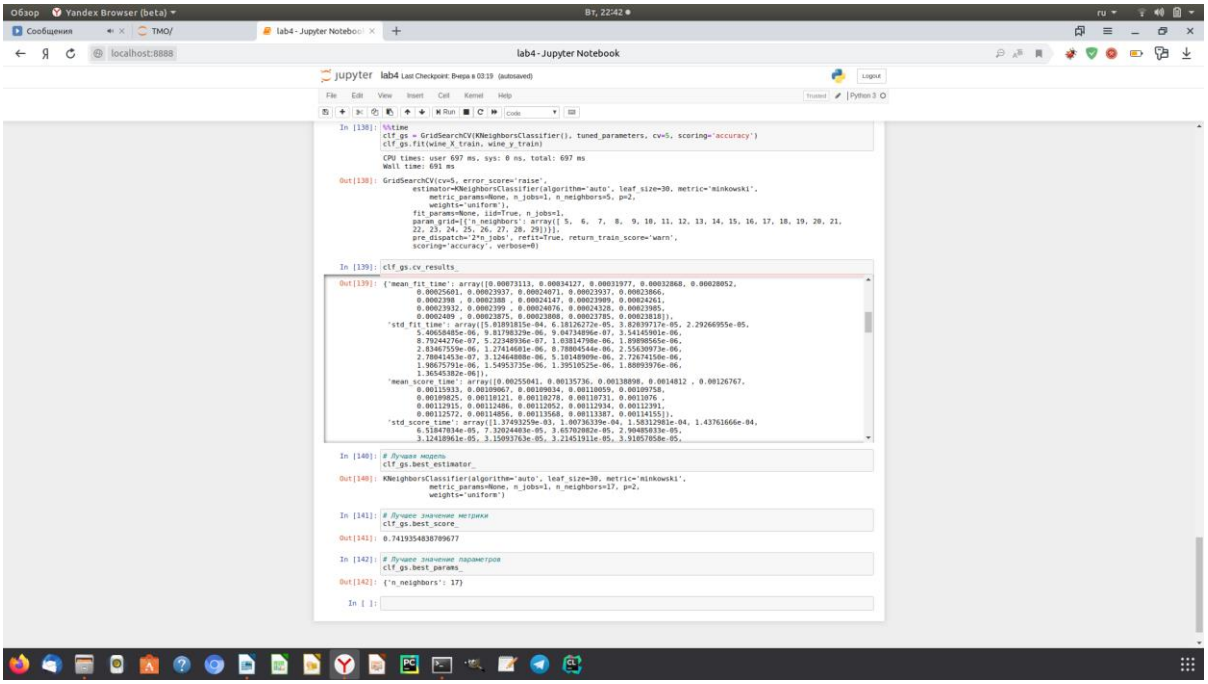


Проанализировав результаты полученных метрик качества классификации, можно судить о среднем качестве классификиции.

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



Нахождение наилучшего гиперпараметра K с использованием GridSearchCV и кросс-валидации



Как видно, лучшее найденное значение гиперпараметра = 17. При этом гиперпараметре получено наилучшее значение метрики = 0.74