ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ассистент |  |  |  | В.В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №4 |
| ПРИМЕНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4917 |  |  |  | В.С. Тихонов |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург

2022

**Цель работы:** изучить алгоритмы и методы классификации на практике.

Вариант 21. Ссылка на репозиторий: https://github.com/Deiwas/a4

1. Получен набор данных от преподавателя для проведения анализа согласно цели выполняемой работы (рис. 1).

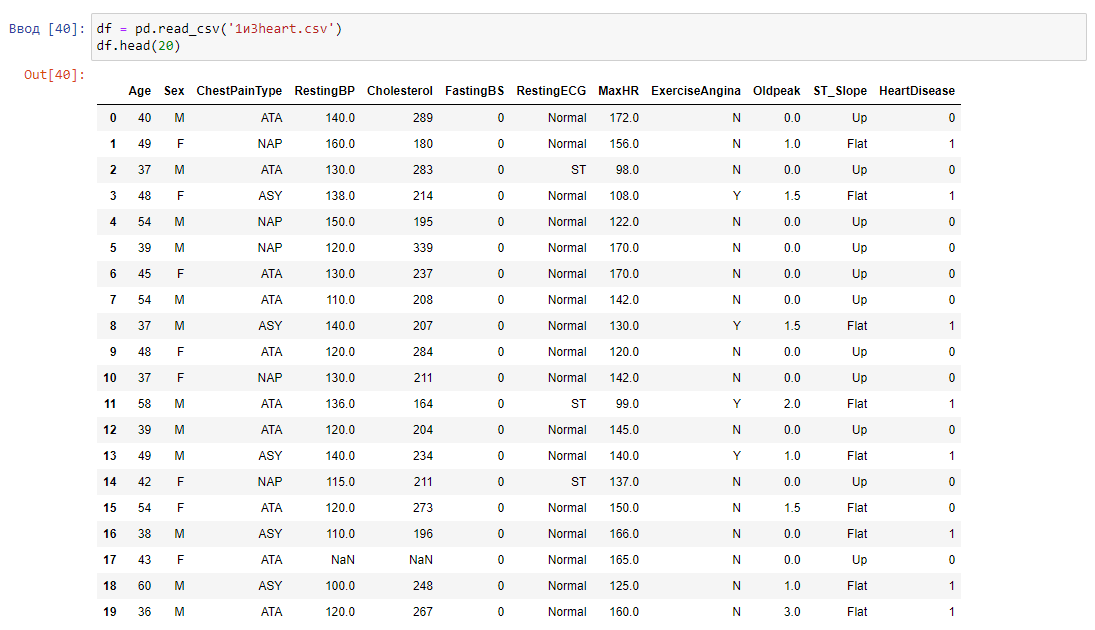


Рисунок 1 – Загруженный датафрейм

1. Проведена предварительная обработка данных. В процессе ее выполнения удалены дубликаты (рис. 2), пустые строки (рис. 3), проведена конвертация данных (рис. 4), а также проведена реиндексация датасета (рис. 5).

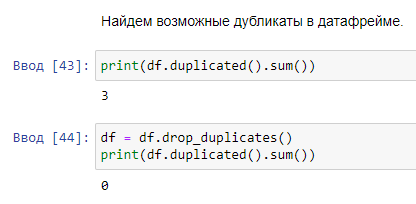


Рисунок 2 – Удаление дубликатов

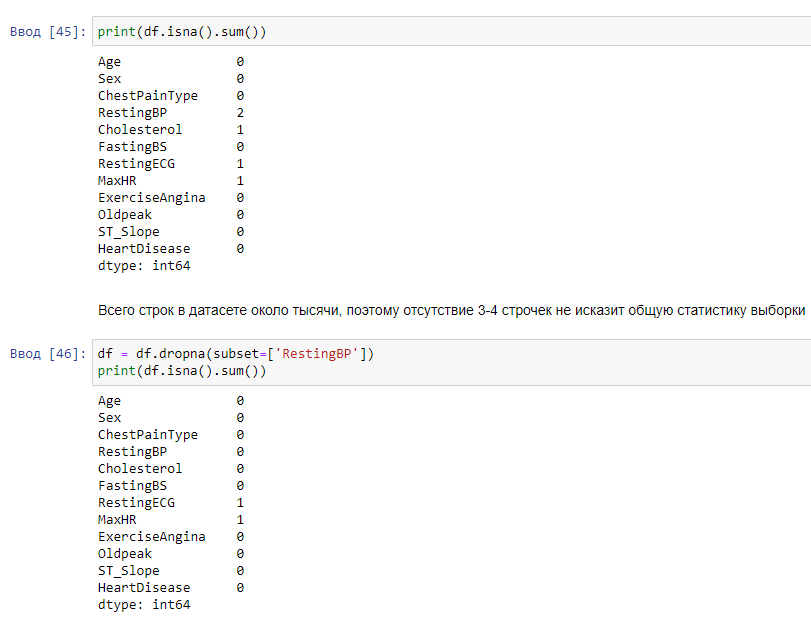


Рисунок 3 – Удаление пустых строк

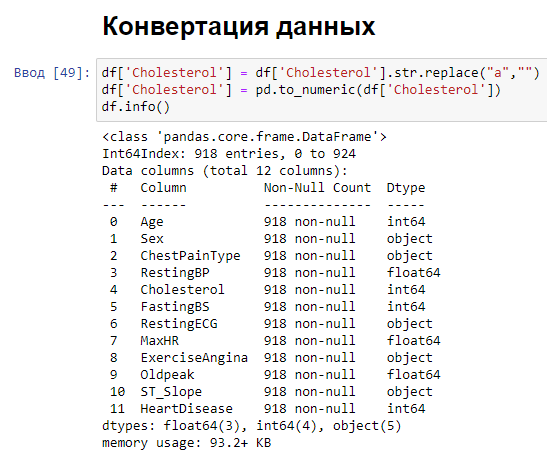


Рисунок 4 – Конвертация столбца «Cholesterol»

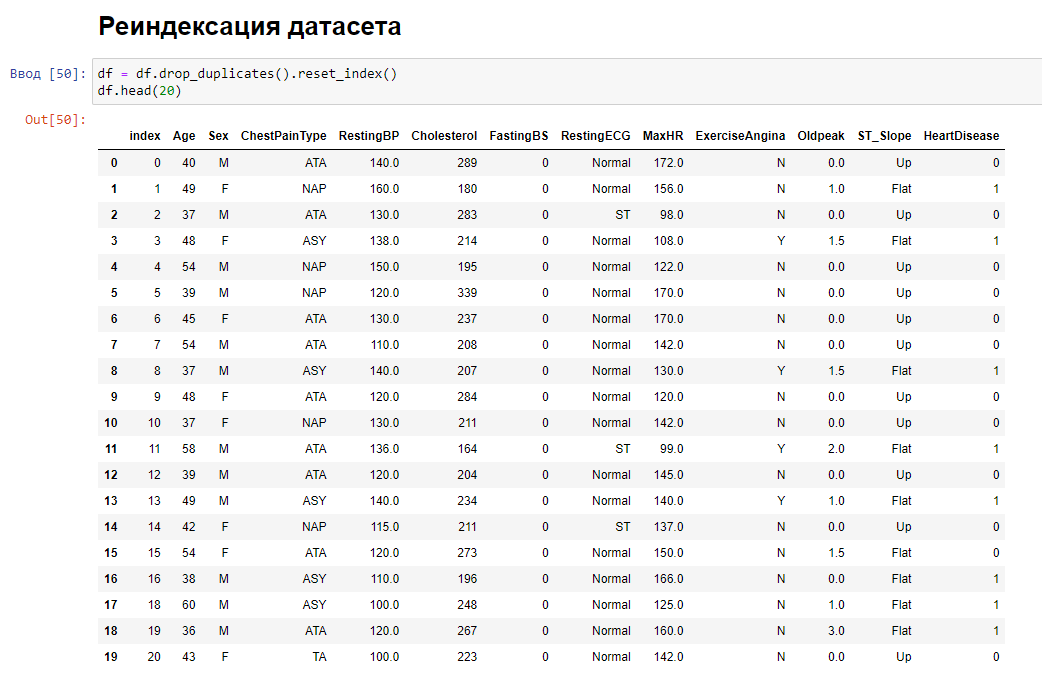


Рисунок 5 – Реиндексация датасета

1. Поиск корреляции столбцов с помощью тепловой карты (рис. 6).

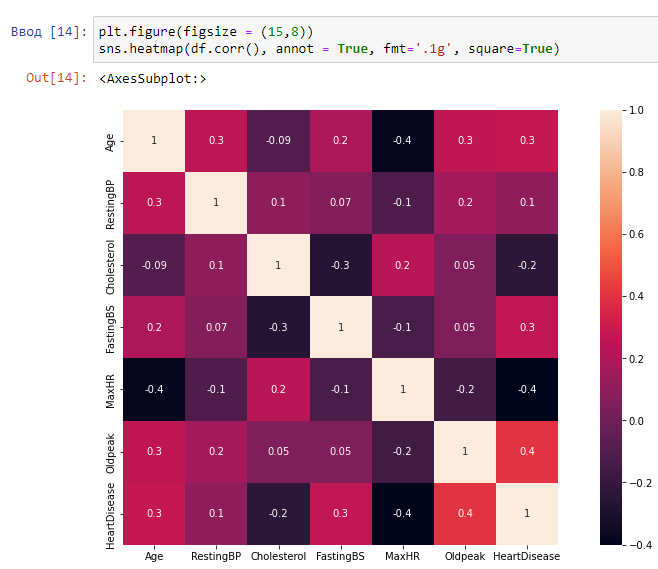


Рисунок 6 – Тепловая карта датафрейма

1. Разбиение набора данных на тренировочный и тестовый датасеты (рис. 7).

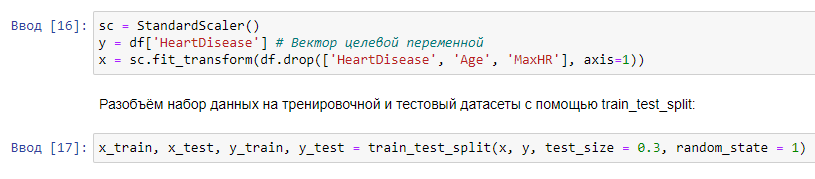


Рисунок 7 – Разбиение набора данных

1. Предсказательная модель методом k-ближайших соседей (рис. 8).

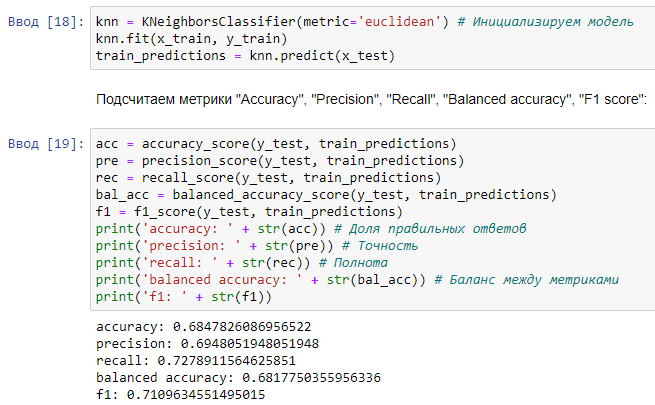


Рисунок 8 – Метод k-ближайших соседей

1. Создание матрицы ошибок (рис. 9).

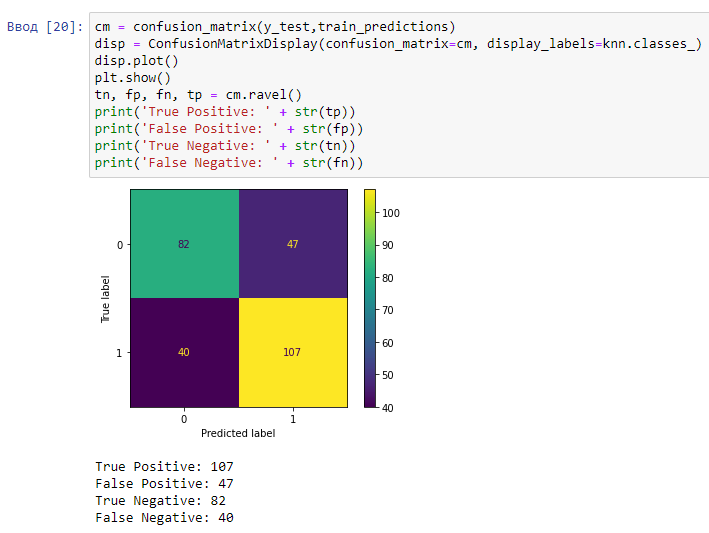


Рисунок 9 – Матрица ошибок

1. Расчет ROC AUC (рис. 10).

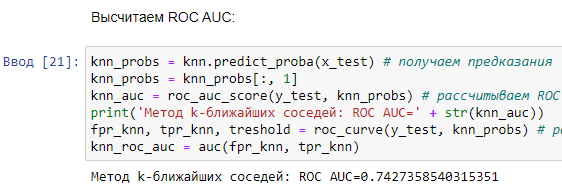


Рисунок 10 – ROC AUC методом k-ближайших соседей

1. Метод логистической регрессии (рис. 11).

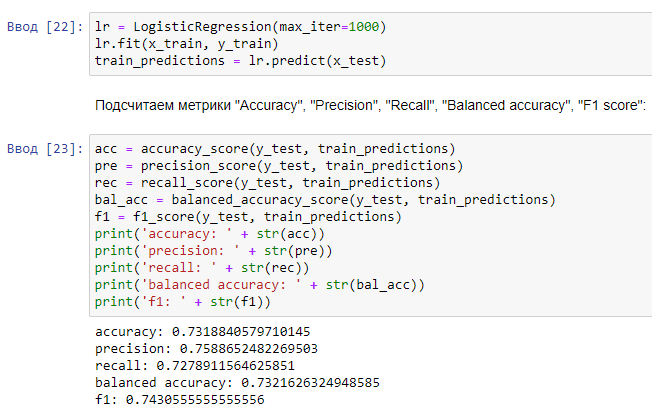


Рисунок 11 – Метод логистической регрессии

1. Создание матрицы ошибок (рис. 12).

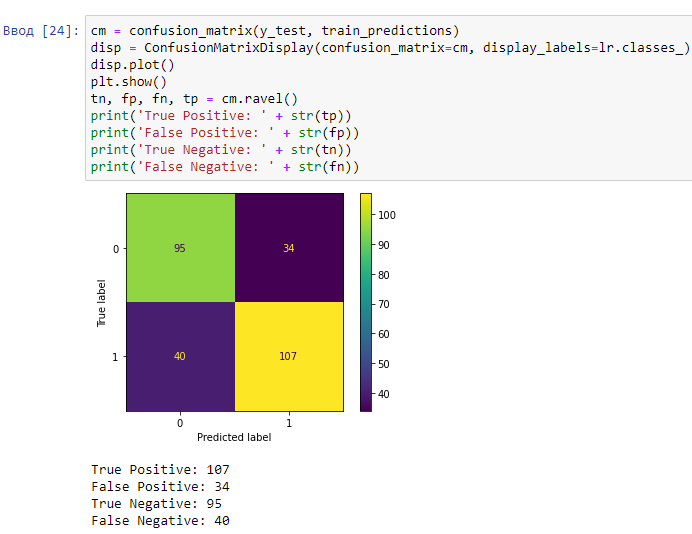


Рисунок 12 – Матрица ошибок метода логистической регрессии

1. Расчет ROC AUC (рис. 13).

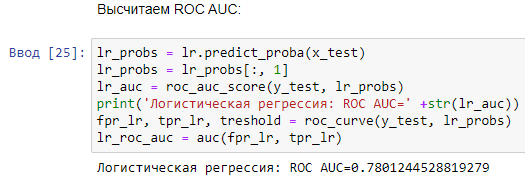


Рисунок 13 – Расчет ROC AUC

1. Разработана предсказательная модель качественного отклика методом случайного леса (рис. 14).

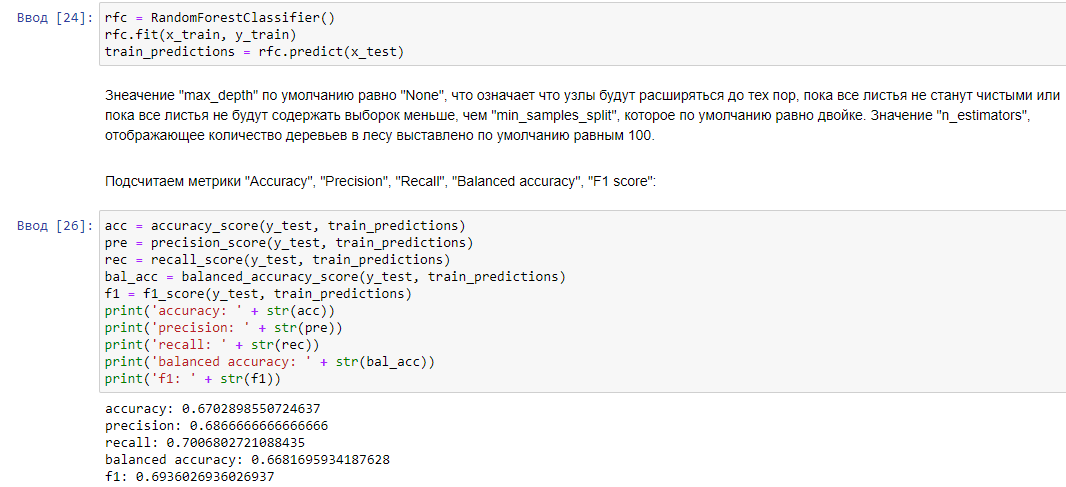


Рисунок 14 – Метод случайного леса

1. Построение матрицы ошибок (рис. 15).

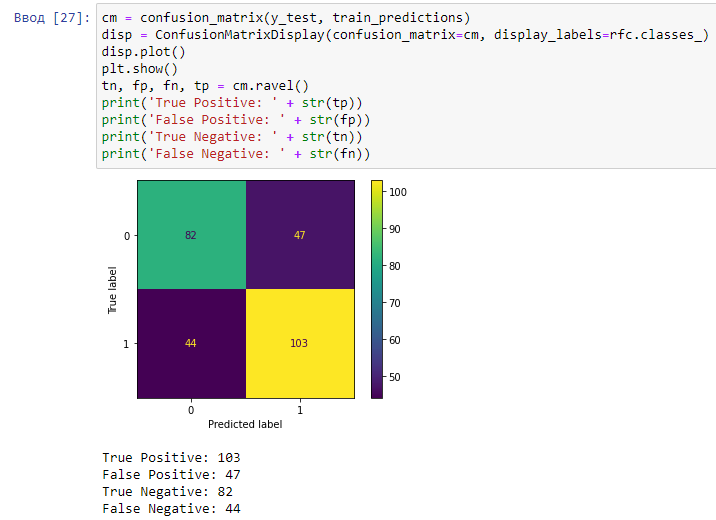


Рисунок 15 – Матрица ошибок метода случайного леса

1. Построим графики ROC-кривых (рис. 16).

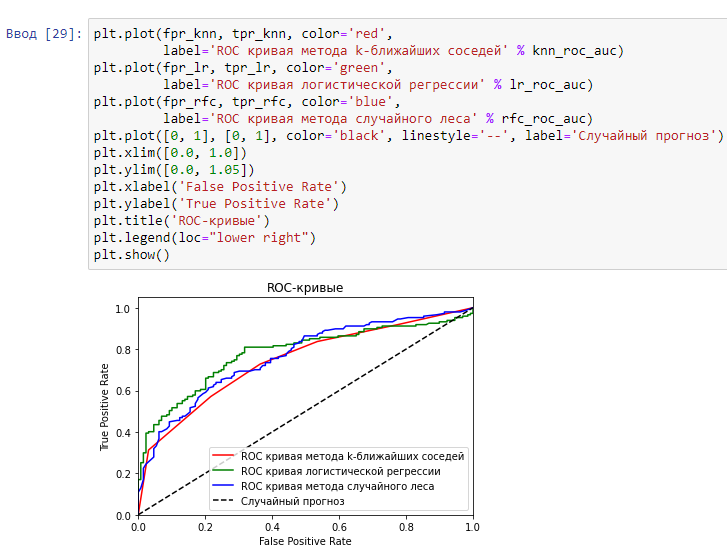


Рисунок 16 – График ROC-кривых

Вывод: в ходе выполнения лабораторной работы изучены алгоритмы и методы классификации данных. Были разработаны предсказательные модели, вычисляющие по набору признаков, на наличие сердечного заболевания на основе методов: k- ближайших соседей, логистической регрессии. Были высчитаны метрики, позволяющие оценить качество работы классификаторов. Лучше всего высчитывает предсказания классификатор, основанный на методе логистической регрессии, а хуже всех себя показал метод k-ближайших соседей. Тем не менее, все три метода классификации данных показали неплохие результаты. Для каждого метода были, также, были построены матрицы неточностей и график ROC-кривых, который подтвердил полученные оценки качества работы методов классификации.

Список источников:

1. Методические указания к лабораторной работе №4: https://pro.guap.ru/get-task/bbccd38503b79ab9d6a1452923da169d