МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

НИЖЕГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ

УНИВЕРСИТЕТ им. Р.Е.АЛЕКСЕЕВА



ОТЧЁТ

По лабораторной работе №3

по дисциплине

«методы Data Mining»

Руководитель:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Суркова А.С.

Студент:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Докукин Д.В.

Группа:

20-ПО

Работа защищена «\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_

С оценкой \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Цель работы**

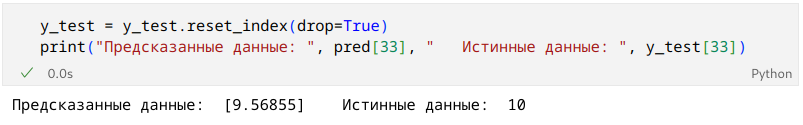
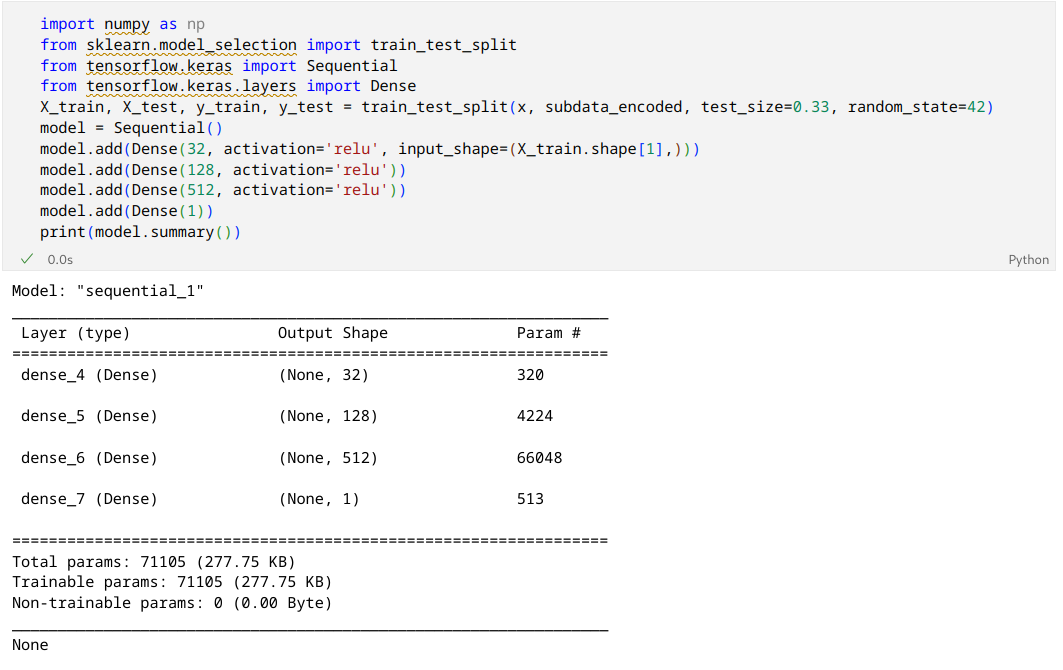
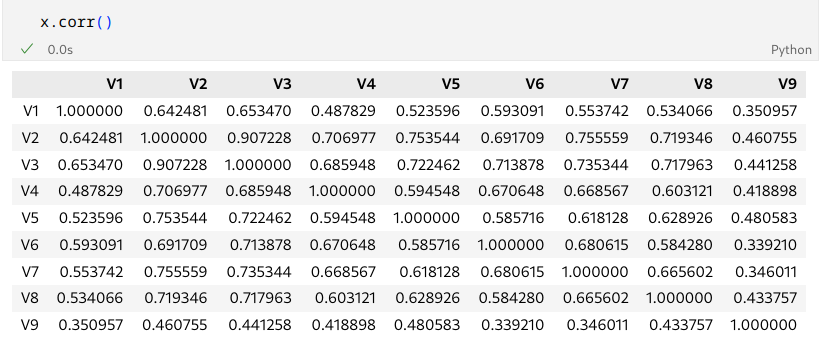
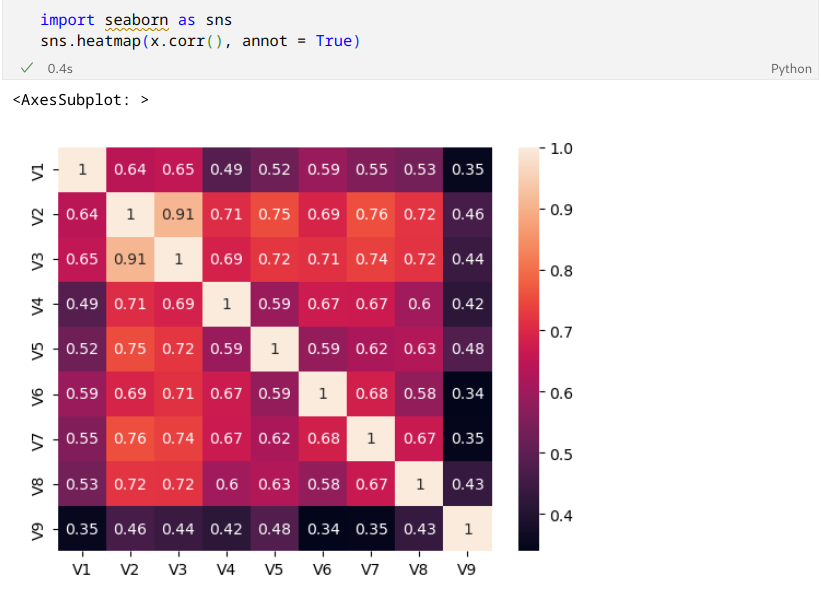
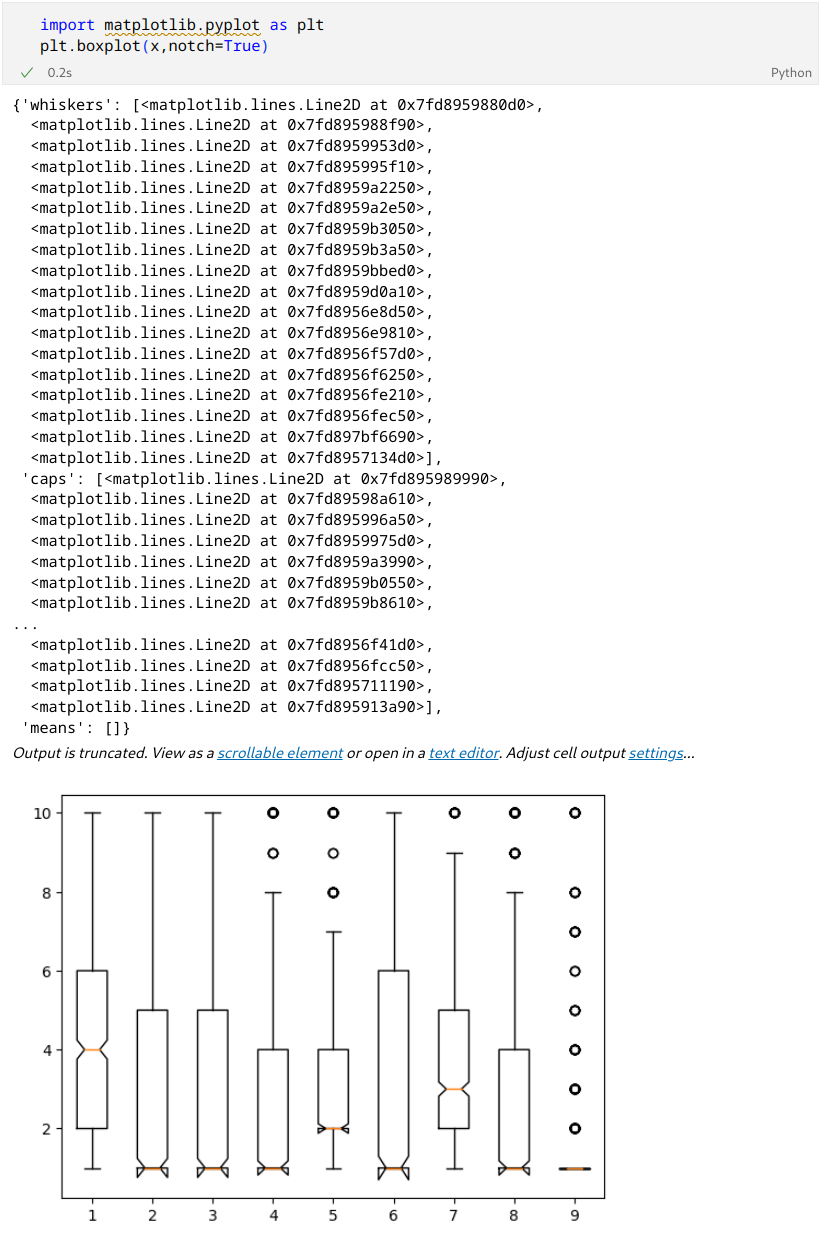
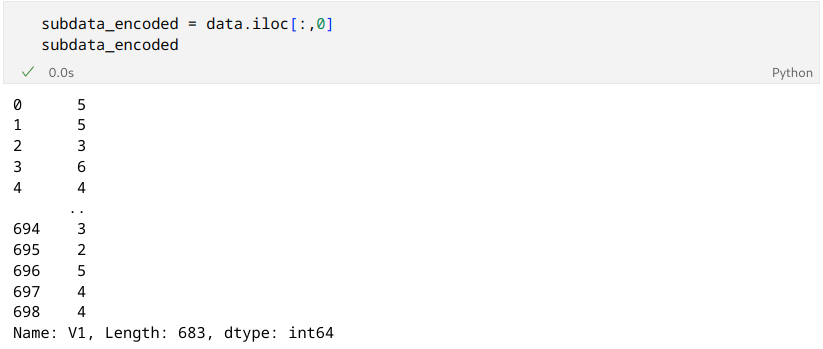
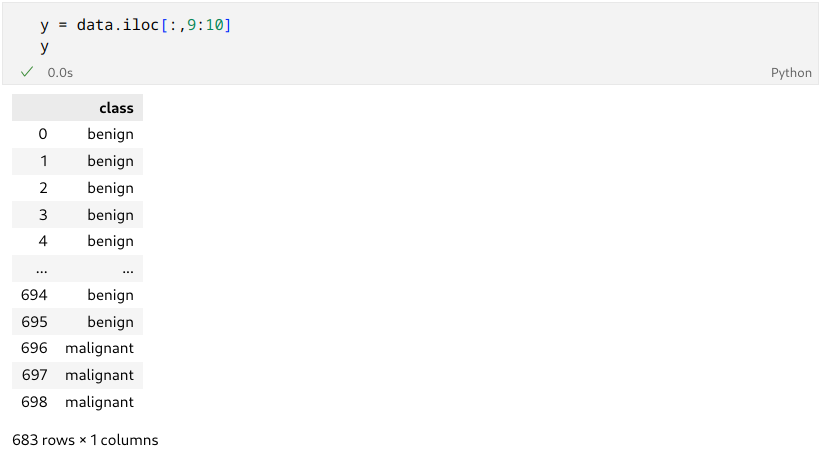
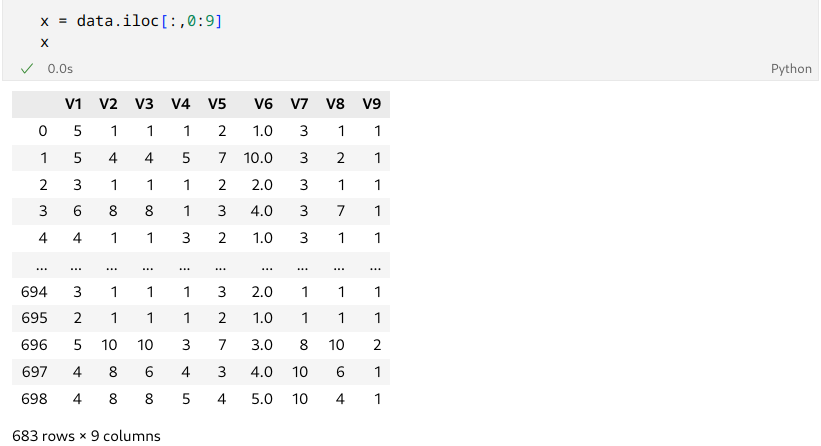
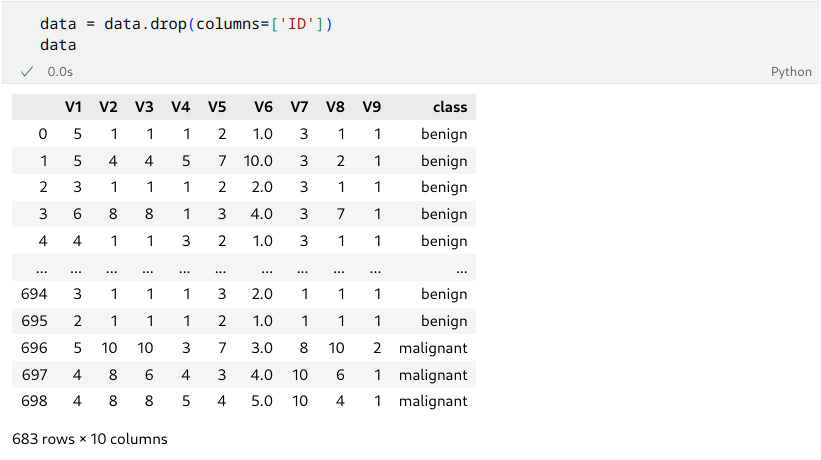
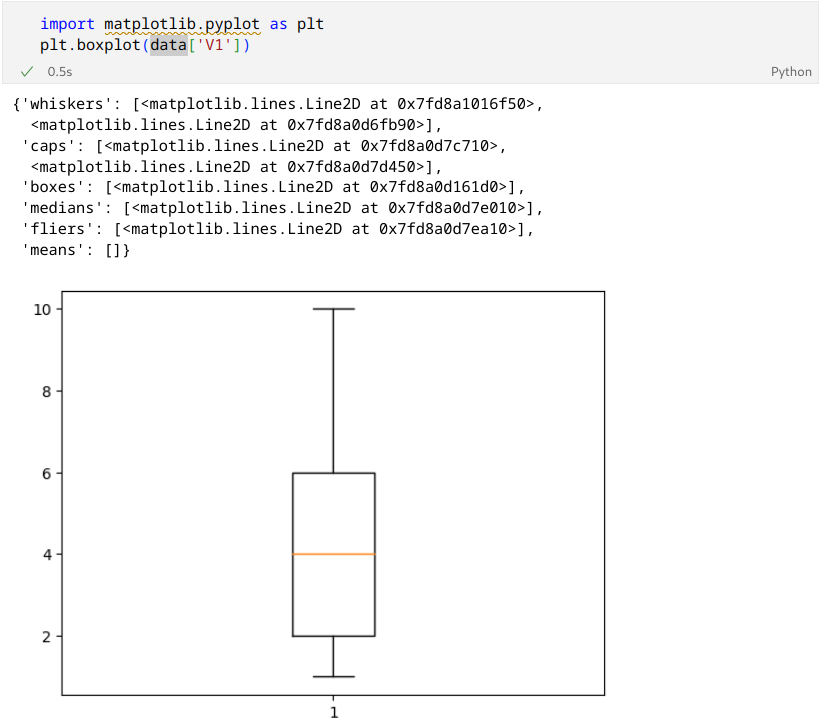
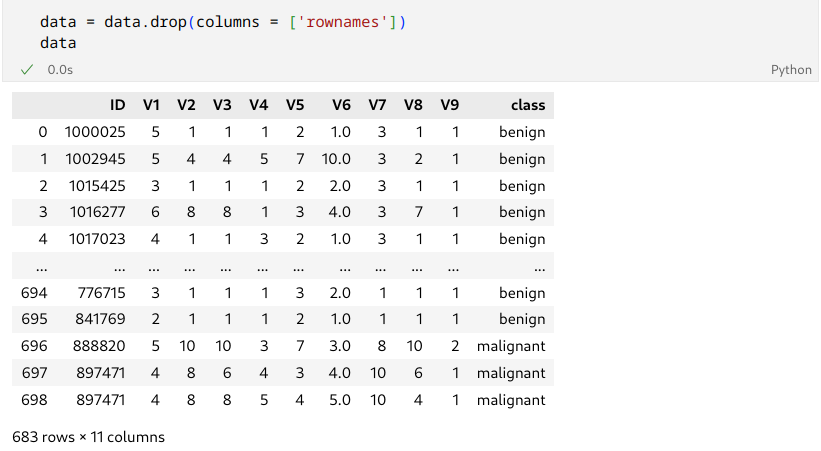
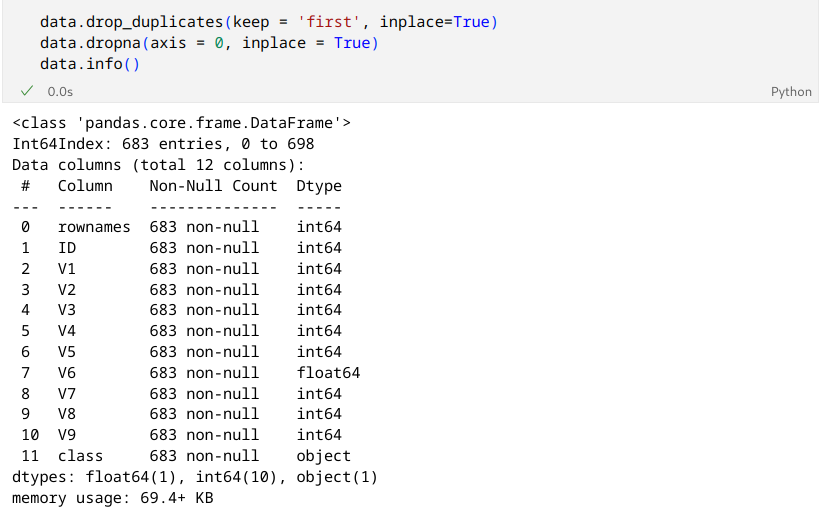
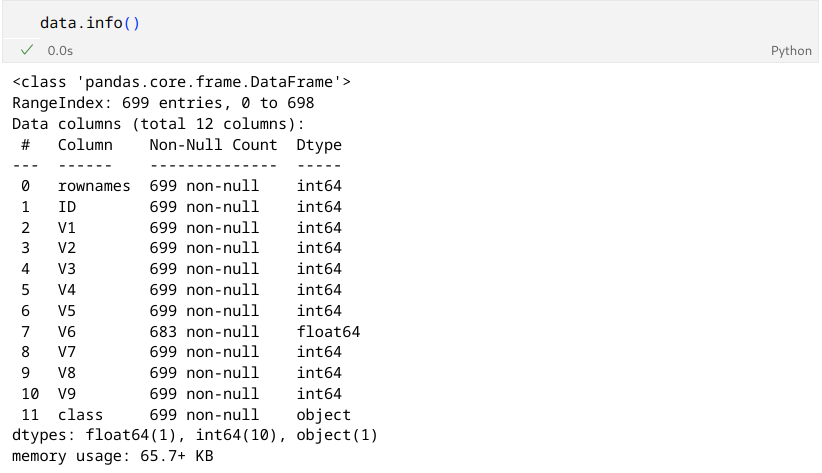
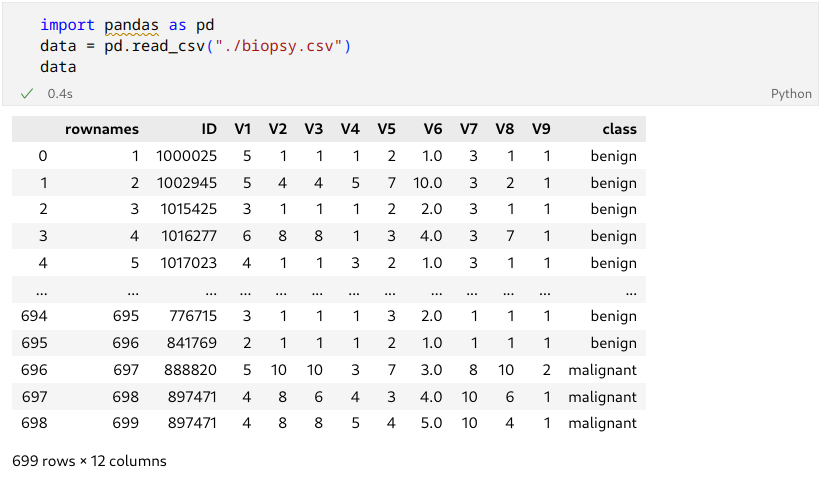
Закрепление знаний по теме «Регрессия» и получение навыков анализа данных.

Ход работы

Задание 1

Произвести анализ данного датасета: углубиться в данные, произвести предобработку, подготовить данные к построению регрессионных моделей, построить графики, описывающие данные, интерпретировать их, посмотреть на взаимосвязь признаков с помощью матрицы корреляции. Исходный датасет разделить на обучающуюся и тестовую выборки. Применить методы снижения размерности, если необходимо. Выбрать и обучить регрессионную модель. Проверить качество модели.

|  |  |  |
| --- | --- | --- |
| **Вариант** | **Краткое вступление** | **Dataset** |
| 7 | Рак молочной железы | data(biopsy, package = 'MASS') |



Задание 2

Исследовать взаимосвязь в данных (Второй столбец – зависимая переменная).

1. Построить регрессионную модель y(второй столб.) = b0 + b1 \* x (первый столб.). Визуализировать полученную модель.
2. Рассчитать коэффициент детерминации.
3. Сформулировать нулевую гипотезу, произвести анализ с помощью t-критерия. (H0 – никакой взаимосвязи в данных нет).
4. Произвести анализ остатков: построить scatter-plot и qq-plot остатков.
5. Предсказать значение по регрессионной модели.

Вариант 7:

Исследовать взаимосвязь процента людей, проживающих в столице и процент домохозяек. Предсказать процент домохозяек, основываясь на проценте людей, проживающих в столице = 100.

|  |  |
| --- | --- |
| **Столица (%)** | **Домохоз. (%)** |
| 55,4 | 14,2 |
| 65,6 | 10,8 |
| 88,2 | 11,1 |
| 52,5 | 12,1 |
| 94,4 | 12,6 |
| 84,5 | 9,6 |
| 87,7 | 12,1 |
| 80,1 | 13,1 |
| 100 | 18,9 |
| 89,3 | 12 |
| 71,6 | 14,5 |
| 91,5 | 12,6 |

1.

Средне: Xср = 80,066 Yср = 12,8

sx = 14,754 sy = 2,254

rxy = 0,188

b1 = 0,004387

b0 = 12,44871

y = 12,44 + 0,0043\*x

2.

R2 = 0,009969

3.

t(b1=0) = 0,0912

Нулевая гипотеза: «связи в данных нет»

P = 0.1

Число степеней свободы: 10

t табл = 1,372

tрасч < tтабл следовательно гипотеза принимается

4. Анализ остатков

|  |
| --- |
| 2,274738 |
| 3,750142 |
| 3,012601 |
| 0,335302 |
| 0,069109 |
| 10,36486 |
| 0,538009 |
| 0,089912 |
| 36,15068 |
| 0,706458 |
| 3,01768 |
| 0,062582 |

Вывод:

Подавляющая часть наблюдений лежит на прямой линии, что говорит о нормальном распределении.

Применение регрессивного анализа было обоснованно. Модели можно доверять, значит можно предсказать.