МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА –Российскийтехнологическийуниверситет»

РТУ МИРЭА

Институт кибербезопасности и цифровых технологий

Кафедра КБ-4 «Интеллектуальные системы информационной безопасности»

Дисциплина «Технологии извлечения знаний из больших данных»

Отчет
о проделанной практической работе

Выполнил студент 1 курса
Группы: ББМО-01-25
Мухаметшин Александр Ринатович

Москва
2025

ОГЛАВЛЕНИЕ

ЗАДАНИЕ 3
ХОД РАБОТЫ 4
ВЫВОД 18

ЗАДАНИЕ

Часть 1.

Построить прогностическую модель для набора данных в файле, проверить связь признаков, построить прогностические модели и модели тренда линейного и квадратичного. Оценить погрешность. (Можно использовать язык программирования)

Часть 2.

Разработать прогностическую модель для набора данных диабетических обследований diabetes.txt. Использовать логистическую регрессию, и метод максимального правдоподобия. Коэффициенты логистической регрессии найти с помощью метода градиентного спуска, который необходимо запрограммировать вручную. Разбить выборку на обучающую и тестовую. Вычислить точность классификации.

Применить отбор признаков на основе корреляции: выбрать наилучшее признаковое пространство, имеющее на два измерения меньше исходного. Построить новую модель и вычислить точность классификации. (Используя Pyhton или любой другой язык программирования)

ХОД РАБОТЫ

Часть 1

Для начала работы были сформированы новые исходные данные (рис. 1).

Таблица 1.2.

Исходные данные

Периоды времени	Показатель x	Показатель y	Показатель z
2013	40	31	5,3
2014	44	34	5,8
2015	45	35	7
2016	52	37	7,7
2017	54	37	10,1
2018	58	39	13,8
2019	60	35	18,9
2020	61	34	30,1
2021	59	27	45,5
2022	55	26	60,1

Графический анализ исходных данных

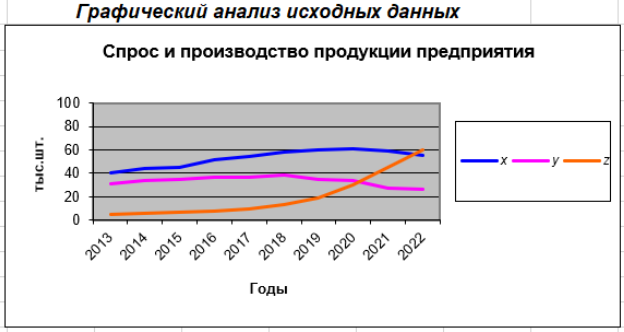


Рисунок 1 – Исходные данные и графический анализ по ним

Основываясь на графике был сформулирован вывод по нему.

ВЫВОД: На графике видно, что показатель z до конца 2017 года увеличивается плавными темпами, затем с 2018 года начинает расти ускоренно и после 2019 года демонстрирует резкий рост, который продолжается вплоть до 2022 года. Одновременно показатель x до 2019 года растет устойчиво, достигая максимума в 2020 году, а затем постепенно снижается. Показатель y с 2013 по 2018 годы растет медленно, затем после 2018 года резко снижается и к концу периода стабилизируется на низком уровне. Представляется, что имеется тесная обратная связь между показателем z и результирующим показателем x, а также некоторая зависимость между y и x только на отдельных временных интервалах. Можно предположить, что показатель y ведет себя без учета x. Для проверки необходимо рассчитать линейные коэффициенты корреляции между показателем x и показателем y, а также между x и показателем z.

Рисунок 2 – Вывод по графику

Были рассчитаны линейные коэффициенты корреляции между показателями x и y (рис. 3), а также между показателями x и z (рис. 4) и сделаны соответствующие выводы по данным расчетам.

Расчет линейных коэффициентов корреляции (задание 3)

Таблица 2.1.

Вспомогательная таблица для расчета линейного коэффициента корреляции

Расчет линейного коэффициента корреляции между спросом и производством

ы времен и	Исходные данны		Вспомогательные расчеты				
	Спрос	Пр-во	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x-\bar{x})*(y-\bar{y})$
	x	y					
2013	40	31	-12,80	-2,50	163,84	6,25	32,00
2014	44	34	-8,80	0,50	77,44	0,25	-4,40
2015	45	35	-7,80	1,50	60,84	2,25	-11,70
2016	52	37	-0,80	3,50	0,64	12,25	-2,80
2017	54	37	1,20	3,50	1,44	12,25	4,20
2018	58	39	5,20	5,50	27,04	30,25	28,60
2019	60	35	7,20	1,50	51,84	2,25	10,80
2020	61	34	8,20	0,50	67,24	0,25	4,10
2021	59	27	6,20	-6,50	38,44	42,25	-40,30
2022	55	26	2,20	-7,50	4,84	56,25	-16,50
Σ	528,00	335,00					

Средние значен		Линейный коэффициент корреля		С исп. функ. КОРРЕЛ
$\bar{x} =$	$\bar{y} =$	r(x,y) =	0,01	0,01
52,80	33,50			

Вывод : Линейный коэффициент корреляции между x и y равен 0.01. Связь слабая.

Рисунок 3 – Расчет линейного коэффициента корреляции между x и y

Вывод : Коэффициент корреляции равен 0,01. Значит связь между двумя показателями *слабая*

Расчет линейного коэффициента корреляции между спросом и ценой

	Исходные данные		Вспомогательные расчеты				
	x	z	$x - \overline{x}$	$z - \overline{z}$	$(x - \overline{x})^2$	$(z - \overline{z})^2$	$(x - \overline{x}) * (z - \overline{z})$
	40	5,3	-12,80	-15,13	163,84	228,92	193,66
	44	5,8	-8,80	-14,63	77,44	214,04	128,74
	45	7	-7,80	-13,43	60,84	180,36	104,75
	52	7,7	-0,80	-12,73	0,64	162,05	10,18
	54	10,1	1,20	-10,33	1,44	106,71	-12,40
	58	13,8	5,20	-6,63	27,04	43,96	-34,48
	60	18,9	7,20	-1,53	51,84	2,34	-11,02
	61	30,1	8,20	9,67	67,24	93,51	79,29
	59	45,5	6,20	25,07	38,44	628,50	155,43
	55	60,1	2,20	39,67	4,84	1573,71	87,27
Σ	528,00	204,30					

Средние значения		Линейный коэффициент корреляции		С исп. функ. КОРРЕЛ
$\overline{x} =$	$\overline{z} =$	r(x,z) =	0,56	0,56
52,80	20,43			

Вывод : Линейный коэффициент корреляции между x и z равен 0.56. Связь умеренная.

Рисунок 4 – Расчет линейного коэффициента корреляции между x и z

Был сделан прогноз показателя x и z по двум вариантам уравнений тренда (рис. 5-6). На рисунке 7 рассчитаны ошибки аппроксимации и сделан вывод о том, какой прогноз более достоверен.

Исходные данные		Вспомогательные расчеты				
Периоды времени		*Условное обозначение времени*			xt	xt^2
	x	t	t^2	t^4		
2013	40	-5	25	625	-200	1000
2014	44	-4	16	256	-176	704
2015	45	-3	9	81	-135	405
2016	52	-2	4	16	-104	208
2017	54	-1	1	1	-54	54
2018	58	1	1	1	58	58
2019	60	2	4	16	120	240
2020	61	3	9	81	183	549
2021	59	4	16	256	236	944
2022	55	5	25	625	275	1375
Σ	528	0	110	1958	203	5537

Расчет параметров линейного и квадратического тренда

Линейный тренд x		Квадратический тренд x	
$x^ = a_0 + a_1 * t$		$x^^ = b_0 + b_1 * t + b_2 * t^2$	
a_0 =	52,8	b_0 =	56,79
a_1 =	1,85	b_1 =	1,85
		b_2 =	-0,36

Периоды времени	Исходные данные		Расчетные данные			
	x	t	x^	x^^	$(x^ - x)^2$	$(x^^ - x)^2$
2013	40	-5	43,57	38,50	12,76	2,25
2014	44	-4	45,42	43,61	2,01	0,15
2015	45	-3	47,26	47,99	5,12	8,93
2016	52	-2	49,11	51,65	8,36	0,13
2017	54	-1	50,95	54,58	9,27	0,33
2018	58	1	54,65	58,27	11,25	0,07
2019	60	2	56,49	59,03	12,31	0,95
2020	61	3	58,34	59,06	7,09	3,76
2021	59	4	60,18	58,37	1,40	0,40
2022	55	5	62,03	56,96	49,38	3,82
Σ	528	0	528	528	119,0	20,8

Рисунок 5 – Расчет показателей тредов для X

Периоды времени		*Условное обозначение времени*			yt	yt^2
	y	**t**	t^2	t^4		
2013	33	**-5,0**	25,0	625,0	-165,0	825,0
2014	34	**-4,0**	16,0	256,0	-136,0	544,0
2015	35	**-3,0**	9,0	81,0	-105,0	315,0
2016	36	**-2,0**	4,0	16,0	-72,0	144,0
2017	38	**-1,0**	1,0	1,0	-38,0	38,0
2018	38	**1,0**	1,0	1,0	38,0	38,0
2019	36	**2,0**	4,0	16,0	72,0	144,0
2020	33	**3,0**	9,0	81,0	99,0	297,0
2021	29	**4,0**	16,0	256,0	116,0	464,0
2022	27	**5,0**	25,0	625,0	135,0	675,0
Σ	**339,0**	**0,0**	**110,0**	**1958,0**	**-56,0**	**3484,0**

Линейный тренд y		**Квадратический тренд y**	
$y^ = a_0 + a_1 * t$		**$y^^ = b_0 + b_1 * t + b_2 * t^2$**	
a_0 =	33,9	**b_0 =**	37,50
a_1 =	-0,51	**b_1 =**	-0,51
		b_2 =	-0,33

Периоды времени	**Исходные данные**		**Расчетные данные**			
	y	**t**	**y^**	**y^^**	**$(y^ - y)^2$**	**$(y^^ - y)^2$**
2013	31	**-5,0**	36,4	31,9	29,7	0,7
2014	34	**-4,0**	35,9	34,3	3,7	0,1
2015	35	**-3,0**	35,4	36,1	0,2	1,2
2016	37	**-2,0**	34,9	37,2	4,3	0,0
2017	37	**-1,0**	34,4	37,7	6,7	0,5
2018	39	**1,0**	33,4	36,7	31,5	5,4
2019	35	**2,0**	32,9	35,2	4,5	0,0
2020	34	**3,0**	32,4	33,0	2,6	0,9
2021	27	**4,0**	31,9	30,2	23,7	10,4
2022	26	**5,0**	31,4	26,8	28,7	0,6
Σ	335,0	0,0	339,0	339,0	135,6	19,9

Рисунок 6 – Расчет показателей тредов для Y

Ошибки аппроксимации для разных уравнений тренда		
Вид уравнения тренда	**Ошибка**	
$x^ = a_0 + a_1 * t$	$\sigma_1 =$	3,45
$x^{\wedge\wedge} = b_0 + b_1 * t + b_2 * t^2$	$\sigma_2 =$	1,44
$y^ = c_0 + c_1 * t$	$\sigma_3 =$	3,68
$y^{\wedge\wedge} = d_0 + d_1 * t + d_2 * t^2$	$\sigma_4 =$	1,41

Расчет прогнозных значений по тренду

Вид уравнения тренда	Прогноз	Ошибка	
$x^ = a_0 + a_1 * t$	63,87	3,45	(прогноз на 1 год вперед)
$x^{\wedge\wedge} = b_0 + b_1 * t + b_2 * t^2$	54,82	1,44	
$y^ = c_0 + c_1 * t$	30,85	3,68	
$y^{\wedge\wedge} = d_0 + d_1 * t + d_2 * t^2$	22,66	1,41	

ВЫВОД:	
(укажите, какое из прогнозных значений Вы считаете более достоверным и почему)	На основе анализа ошибок аппроксимации можно заключить, что квадратическая модель для переменной x (с ошибкой 1,44) более точна, чем линейная (с ошибкой 3,45), поскольку она лучше описывает нелинейные закономерности в данных. Для переменной y квадратический тренд (с ошибкой 1,41) также оказывается более точным по сравнению с линейным (с ошибкой 3,68). Таким образом, квадратические тренды для обеих переменных (x и y) предпочтительнее для прогнозирования, так как они демонстрируют меньшие ошибки аппроксимации и точнее отражают динамику данных.

Рисунок 7 – Ошибки аппроксимации и вывод

Далее были рассчитаны параметры уравнения парной линейной регрессии, выражающей зависимость между показателем x и тем из двух показателей (y или z), с которым связь показателя x более сильная (рис. 8). Рассчитаны ошибка аппроксимации и индекс детерминации, сделан вывод о том, насколько хорошо построенное уравнение отражает существующую зависимость (рис 9).

	Расчет параметров парной линейной регрессии (задание 5)					
	Вспомогательная таблица для расчета параметров уравнения парной линейной регрессии ($x = k_0 + k_1 * z$)					
	Исходные данные		Вспомогательные расчеты			
			Расчет параметров		Расчет ошибки(σ)	
	x	z	z^2	$z*x$	$x_z = k_0 + k_1 * z$	$(x - x_z)^2$
	40,0	5,3	28,1	212,0	49,5	90,6
	44,0	5,8	33,6	255,2	49,6	31,7
	45,0	7,0	49,0	315,0	49,9	23,9
	52,0	7,7	59,3	400,4	50,0	3,8
	54,0	10,1	102,0	545,4	50,6	11,8
	58,0	13,8	190,4	800,4	51,4	44,1
	60,0	18,9	357,2	1134,0	52,5	56,7
	61,0	30,1	906,0	1836,1	54,9	37,2
	59,0	45,5	2070,3	2684,5	58,2	0,6
	55,0	60,1	3612,0	3305,5	61,4	41,0
Σ	528,0	204,3	7408,0	11488,5	528,0	341,5

Параметры регрессии	
$k_0 =$	48,4
$k_1 =$	0,22

Рисунок 8 – Расчет параметров парной линейной регрессии

Ошибка аппроксимации	
$\sigma_5 =$	5,84

ВЫВОД: Линейная регрессионная модель, несмотря на некоторую погрешность, свидетельствует о положительной корреляции между переменными x и z. Однако для повышения точности прогнозов рекомендуется использовать более сложные модели, особенно при наличии нелинейных зависимостей в данных.

Рисунок 9 – Ошибка аппроксимации, индекс детерминации и вывод

В конце был выполнен прогноз показателя, выбранного ранее, по любому из уравнений тренда и рассчитайте прогноз спроса по уравнению регрессии. Расчеты приведены на рисунках 10-12.

ЗАДАНИЕ 6.Расчет прогноза цены по тренду и прогноза результирующего показателя x по регрессии)

1. Используя формулы для расчета параметров тренда, из таблицы, построенной в задании 4, заменив исходный ряд (x) на ряд значений выбранного факторного показателя (y или z), и рассчитайте параметры уравнений тренда для расчета прогнозного значения этого показателя.

ПРИМЕЧАНИЕ: Ниже в таблице для определенности указан показатель z, но это может быть и y

	z	Условное обозначение времени t	t^2	t^4	z*t	$z*t^2$
	5,3	**-5**	25	625	-26,5	132,5
	5,8	**-4**	16	256	-23,2	92,8
	7	**-3**	9	81	-21	63
	7,7	**-2**	4	16	-15,4	30,8
	10,1	**-1**	1	1	-10,1	10,1
	13,8	**1**	1	1	13,8	13,8
	18,9	**2**	4	16	37,8	75,6
	30,1	**3**	9	81	90,3	270,9
	45,5	**4**	16	256	182	728
	60,1	**5**	25	625	300,5	1502,5
Σ	**204,3**	**0**	**110**	**1958**	**528,2**	**2920**

Линейный тренд z		Квадратический тренд z		
$z^\wedge = u_0 + u_1 * t$		$z^{\wedge\wedge} = w_0 + w_1 * t + w*t^2$		
$u_0 =$	20,43	$w_0 =$	10,54	
$u_1 =$	4,80	$w_1 =$	4,80	
		$w_2 =$	0,90	

Рисунок 10 – Расчет параметров уравнений трендов

2) Рассчитайте две ошибки аппроксимации, по аналогии с тем, как это делалось для линейного и квадратичного тренда показателя x.

	Исходные данные		Расчетные данные			
	z	t	$z^\wedge$	$z^{\wedge\wedge}$	$(z^\wedge - z)^2$	$(z^{\wedge\wedge} - z)^2$
	5,3	**-5**	-3,58	9,01	78,84	13,78
	5,8	**-4**	1,22	5,72	20,95	0,01
	7	**-3**	6,02	4,23	0,95	7,70
	7,7	**-2**	10,83	4,53	9,77	10,04
	10,1	**-1**	15,63	6,63	30,56	12,01
	13,8	**1**	25,23	16,24	130,69	5,95
	18,9	**2**	30,03	23,74	123,96	23,41
	30,1	**3**	34,84	33,04	22,42	8,62
	45,5	**4**	39,64	44,13	34,37	1,87
	60,1	**5**	44,44	57,03	245,26	9,43
Σ	204,30	0,00	204,30	204,30	697,78	92,80

Ошибки аппроксимации для разных уравнений тренда		
Вид уравнения тренда	**Ошибка**	
$z^\wedge = u_0 + u_1 * t$	$\sigma_1 =$	8,35
$z^{\wedge\wedge} = w_0 + w_1 * t + w*t^2$	$\sigma_2 =$	3,05

Расчет прогнозных значений показателя по тренду

Вид уравнения тренда	Прогноз	Ошибка
$z^\wedge = u_0 + u_1 * t$	49,24	8,35
$z^{\wedge\wedge} = w_0 + w_1 * t + w*t^2$	71,72	3,05

(прогноз на 1 год вперед)

Рисунок 11 – Расчет ошибок аппроксимации

3. Определите, какое из прогнозных значений показателя более достоверно, и подставьте его в уравнение регрессии, построенное на листе 4.

Расчет прогнозных значений показателя x по регрессии		
Вид уравнения регрессии	Прогноз	Ошибка
$x^{\wedge} = k_0 + k_1 * z$	64,18	5,84

ВЫВОД (заключительный): Проведенные расчеты и анализ свидетельствуют, что квадратические модели обеспечивают более высокую точность прогнозирования по сравнению с линейными. Это позволяет более эффективно выявлять взаимосвязи между показателями и формировать обоснованные прогнозы. Наиболее точным прогнозом для показателя z является значение 68,18, полученное с использованием квадратического тренда, которое рекомендуется применять в дальнейших расчетах из-за его большей точности. Прогноз для показателя x, основанный на регрессии, составляет 55,55, что также подтверждает надежность выбранной модели.

Рисунок 12 – Расчет прогнозных значений показателя x по регрессии и вывод

Часть 2

Сначала импортируем необходимые библиотеки для работы с данными и моделями машинного обучения. Используются pandas для работы с данными, CatBoostClassifier для построения модели на основе градиентного бустинга, sklearn для логистической регрессии, разбиения данных и вычисления метрик, а также numpy для численных операций. Импорт моделей показан на рисунке 13.

```
!pip install catboost -q
```

```
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 99.2/99.2 MB 8.2 MB/s eta 0:00:00
```

```
import pandas as pd
from catboost import CatBoostClassifier
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import mean_squared_error
import numpy as np
```

Рисунок 13 – Импорт библиотек

Загружаем набор данных diabetes.xlsx с помощью pandas.read_excel. Выводим первые пять строк данных с помощью df.head() для ознакомления с их структурой. Также проверяем размеры данных (df.shape) и типы данных (df.dtypes) для анализа. Код и результат работы показан на рисунках 14-15.

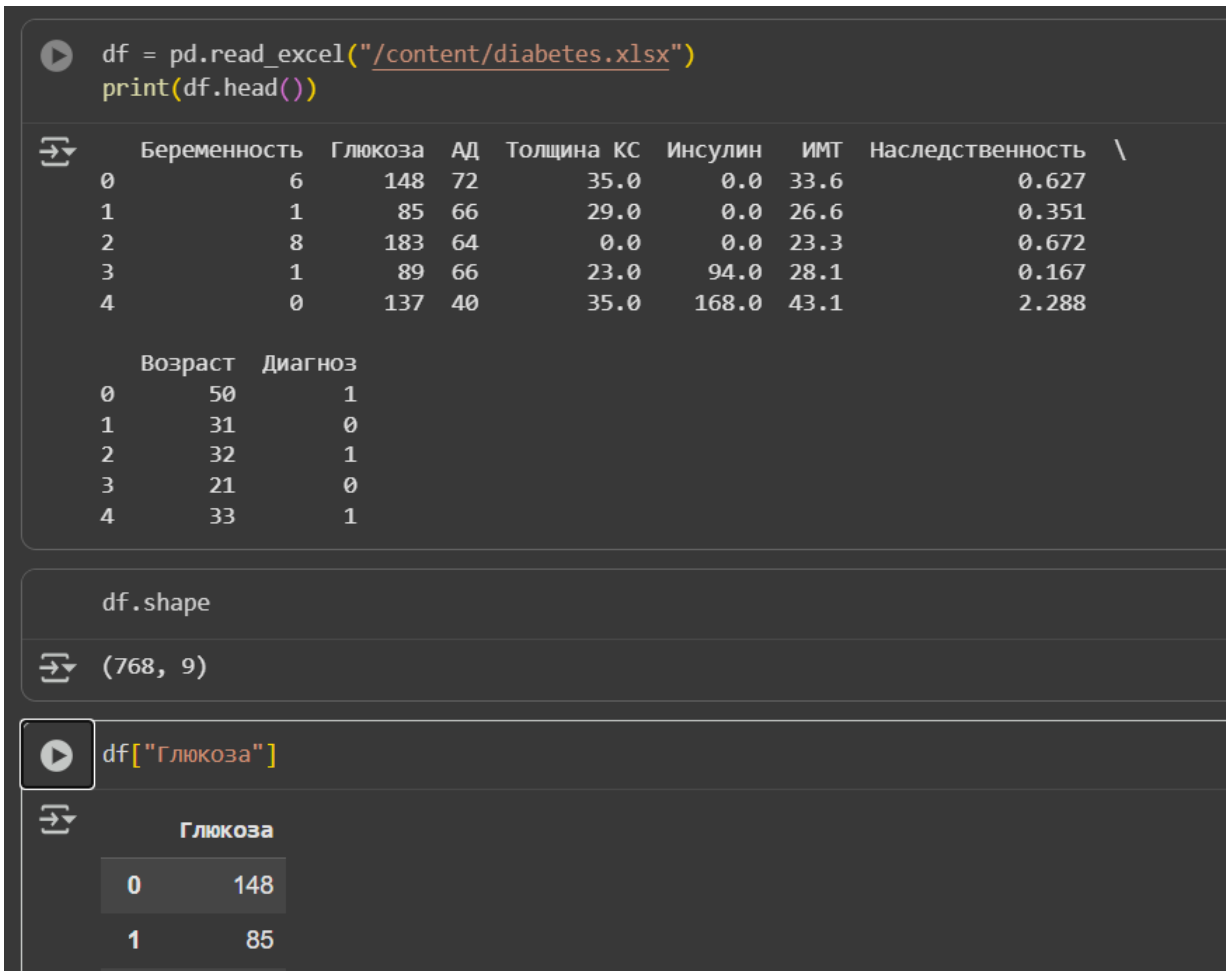


Рисунок 14 – Импорт датасета и проверка данных

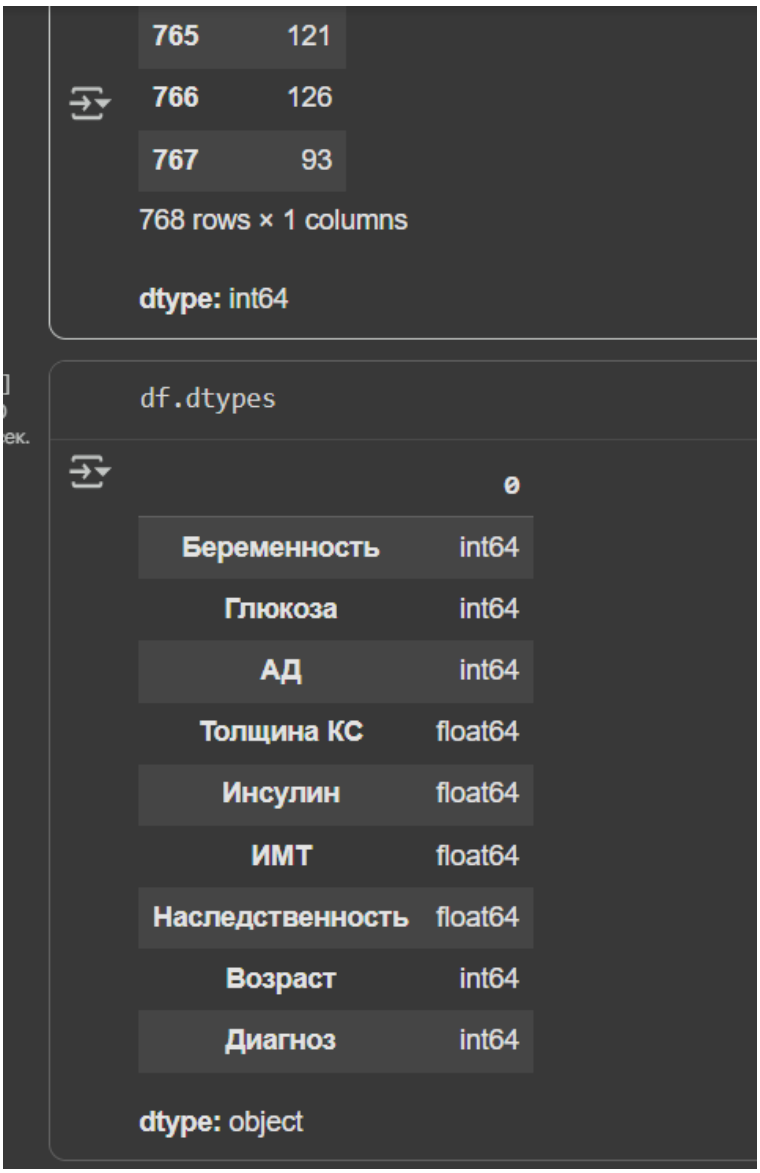


Рисунок 15 – Вывод типов данных датасета

Данные успешно загружены, содержат 768 строк и 9 столбцов (8 признаков и целевая переменная "Диагноз"). Признаки включают "Беременность", "Глюкоза", "АД", "Толщина КС", "Инсулин", "ИМТ", "Наследственность" и "Возраст". Типы данных: int64 для целочисленных признаков и float64 для вещественных. Это подтверждает, что данные готовы для дальнейшей обработки.

Разделяем данные на признаки (X) и целевую переменную (y). Используем train_test_split для разбиения на обучающую (600 строк) и тестовую выборки с фиксированным random_state=42 для воспроизводимости.

Реализуем функцию fit_gd для обучения логистической регрессии

методом градиентного спуска. Функция включает сигмоидную активацию, вычисление логарифмической функции потерь и обновление весов с учетом скорости обучения (lr=0.5) и количества итераций (epochs=20000). Модель обучается на всех данных.

Реализация разбиения и функции показаны на рисунке 16.

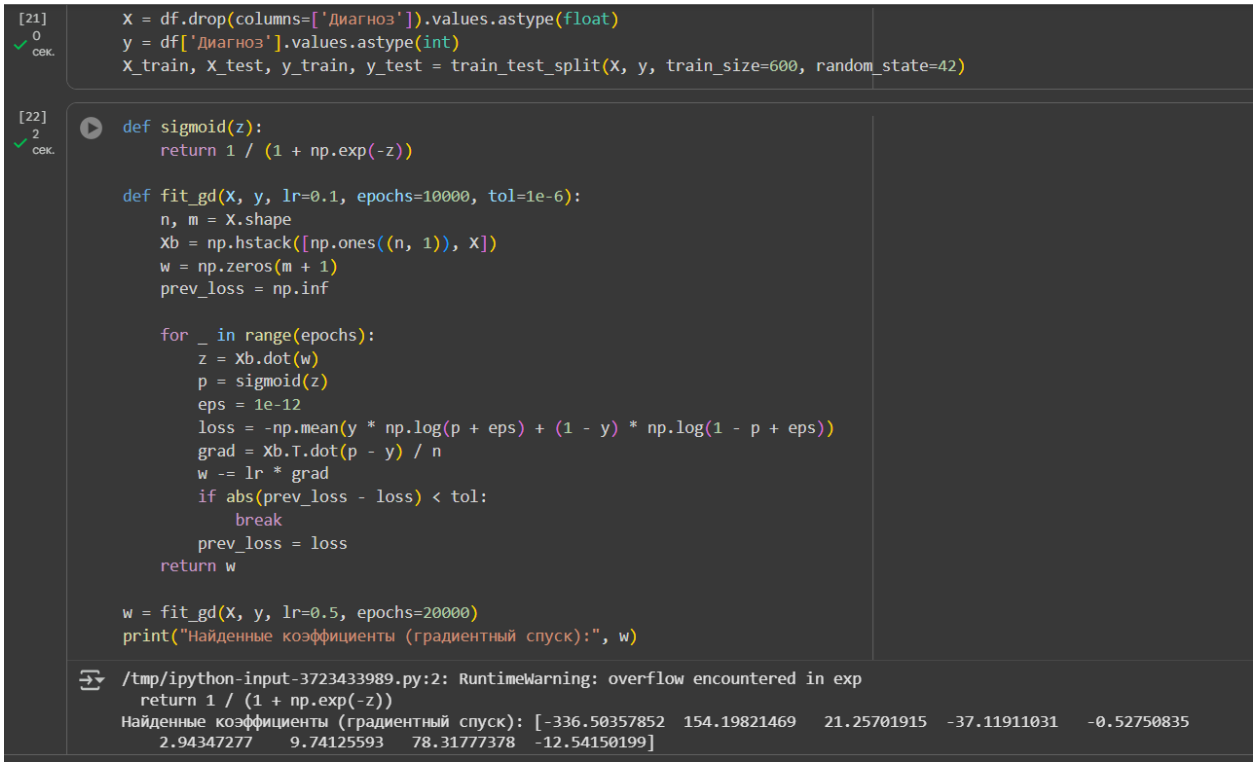


Рисунок 16 – Разбиение данных и подготовка функций

Получены коэффициенты модели: интерсепт -336.50 и веса для признаков.

Инициализируем модель LogisticRegression из sklearn и присваиваем ей коэффициенты, полученные из градиентного спуска. Выводим интерсепт и коэффициенты для проверки (рис. 17).

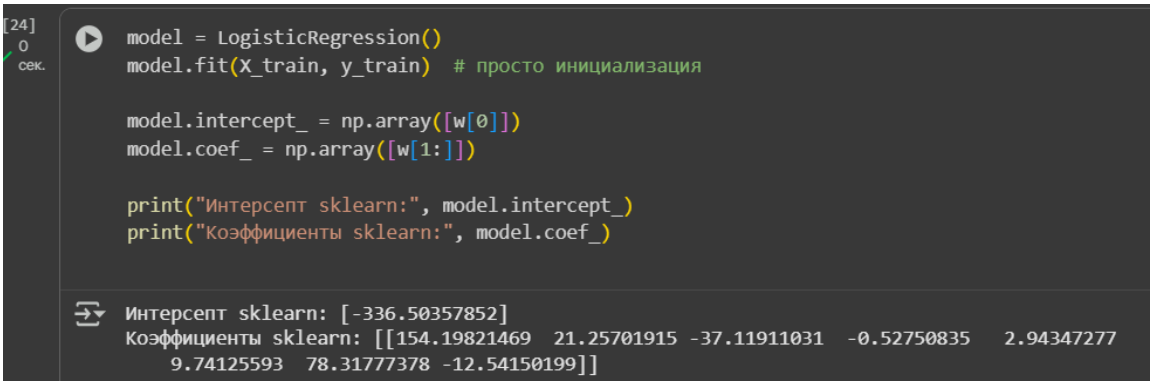


Рисунок 17 – Вывод интерсепта

Коэффициенты из градиентного спуска успешно перенесены в модель sklearn. Интерсепт и веса совпадают с предыдущим блоком.

Вычисляем вероятности класса 1 на тестовой выборке с помощью predict_proba и оцениваем качество модели с помощью среднеквадратичной ошибки (MSE) между вероятностями и истинными метками (рис 18).

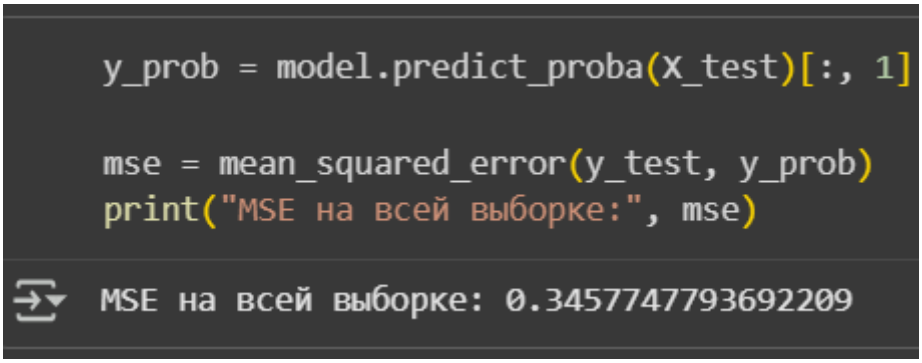


Рисунок 18 – Вычисление MSE

MSE на тестовой выборке составляет 0.345.

Выбираем признаки на основе абсолютных значений корреляции с целевой переменной "Диагноз". Исключаем два признака с наименьшей корреляцией, чтобы получить признаковое пространство на два измерения меньше исходного (рис. 19).

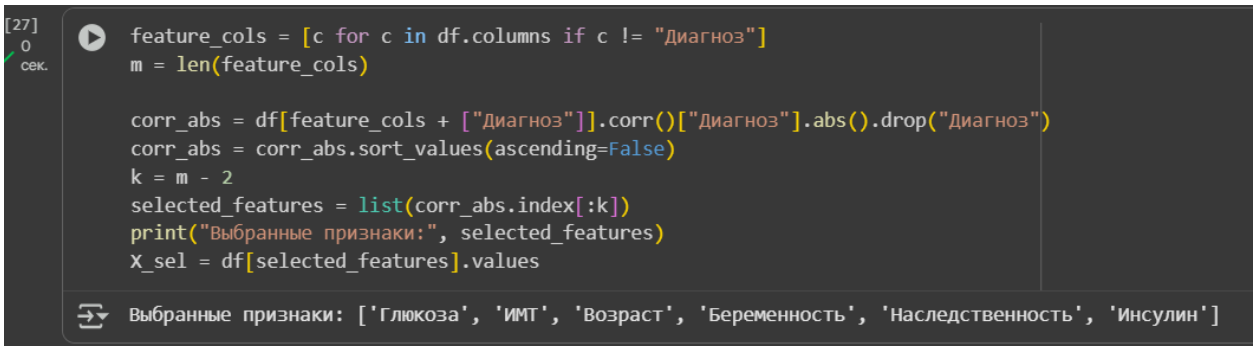
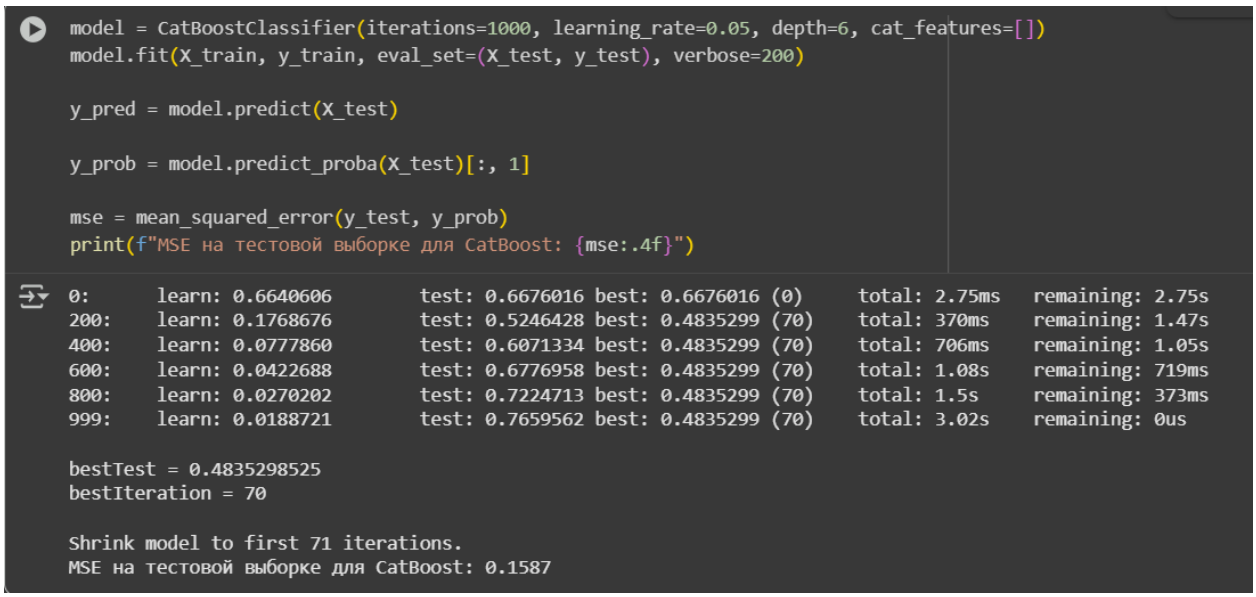


Рисунок 19 – Выбор признаков

Выбраны признаки: "Глюкоза", "ИМТ", "Возраст", "Беременность", "Наследственность", "Инсулин". Признаки "АД" и "Толщина КС" исключены из-за низкой корреляции с целевой переменной.

Обучаем модель CatBoostClassifier с 1000 итерациями, скоростью обучения 0.05 и глубиной дерева 6. Оцениваем модель на тестовой выборке и вычисляем MSE для вероятностей (рис. 20).



```
model = CatBoostClassifier(iterations=1000, learning_rate=0.05, depth=6, cat_features=[])
model.fit(X_train, y_train, eval_set=(X_test, y_test), verbose=200)

y_pred = model.predict(X_test)

y_prob = model.predict_proba(X_test)[:, 1]

mse = mean_squared_error(y_test, y_prob)
print(f"MSE на тестовой выборке для CatBoost: {mse:.4f}")
```

```
0:      learn: 0.6640606        test: 0.6676016 best: 0.6676016 (0)     total: 2.75ms   remaining: 2.75s
200:    learn: 0.1768676        test: 0.5246428 best: 0.4835299 (70)    total: 370ms    remaining: 1.47s
400:    learn: 0.0777860        test: 0.6071334 best: 0.4835299 (70)    total: 706ms    remaining: 1.05s
600:    learn: 0.0422688        test: 0.6776958 best: 0.4835299 (70)    total: 1.08s    remaining: 719ms
800:    learn: 0.0270202        test: 0.7224713 best: 0.4835299 (70)    total: 1.5s     remaining: 373ms
999:    learn: 0.0188721        test: 0.7659562 best: 0.4835299 (70)    total: 3.02s    remaining: 0us

bestTest = 0.4835298525
bestIteration = 70

Shrink model to first 71 iterations.
MSE на тестовой выборке для CatBoost: 0.1587
```

Рисунок 20 – Обчуение CatBoost

Модель CatBoost показала MSE 0.1587 на тестовой выборке, что лучше, чем у логистической регрессии.

ВЫВОД

В результате выполнения практической работы был получен опыт построения прогностических моделей для набора данных диабетических обследований (diabetes.txt) с использованием Python. Реализована логистическая регрессия с методом максимального правдоподобия, коэффициенты которой найдены с помощью вручную запрограммированного градиентного спуска. Проведено разбиение выборки на обучающую и тестовую, вычислена точность классификации. Выполнен отбор признаков на основе корреляции Пирсона, что позволило сократить признаковое пространство на два измерения и построить улучшенную модель с оценкой ее точности. Кроме того, исследованы связи признаков и построены линейные и квадратичные модели тренда с оценкой их погрешности, что обеспечило понимание зависимостей в данных и качества прогнозов.