|  |
| --- |
|  |
| 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](#_Toc4)

[ХОД РАБОТЫ 4](#_Toc5)

[ВЫВОД 8](#_Toc6)

# 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).

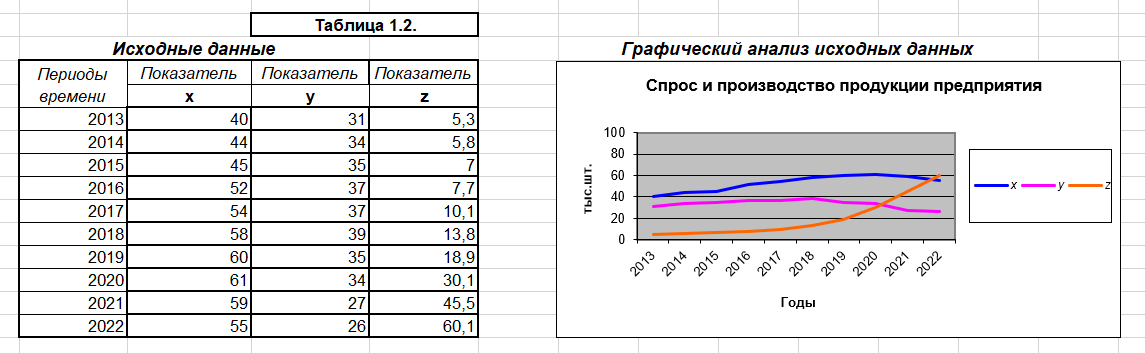


Рисунок 1 – Исходные данные и графический анализ по ним

Основываясь на графике был сформулирован вывод по нему.

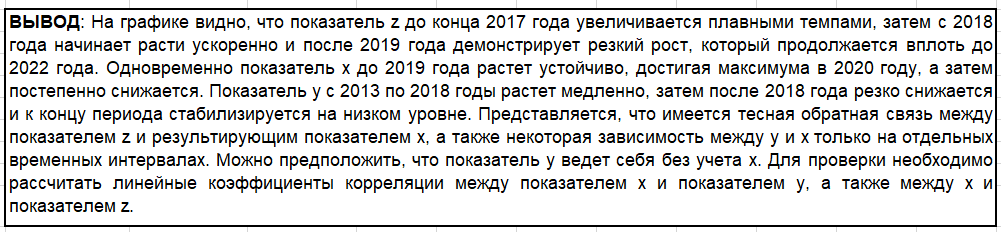


Рисунок 2 – Вывод по графику

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.

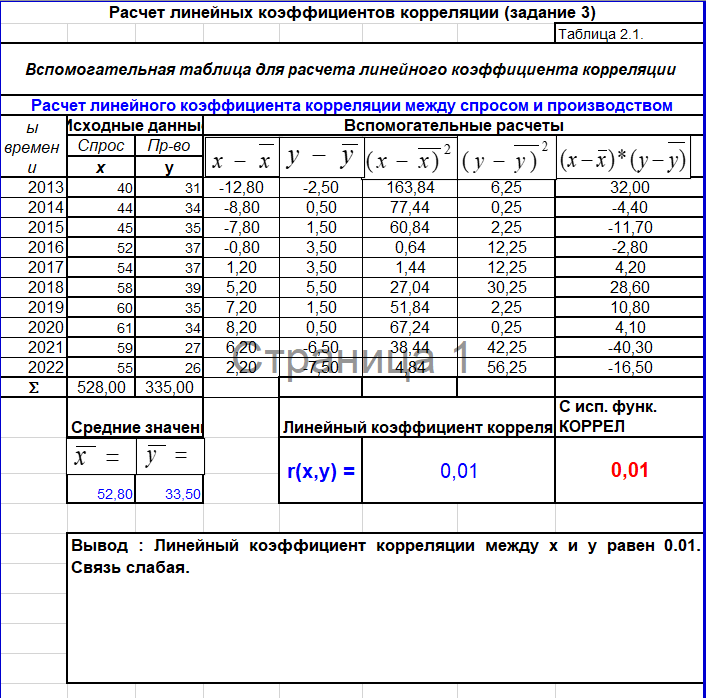


Рисунок 3 – Расчет линейного коэффициента корреляции между х и у

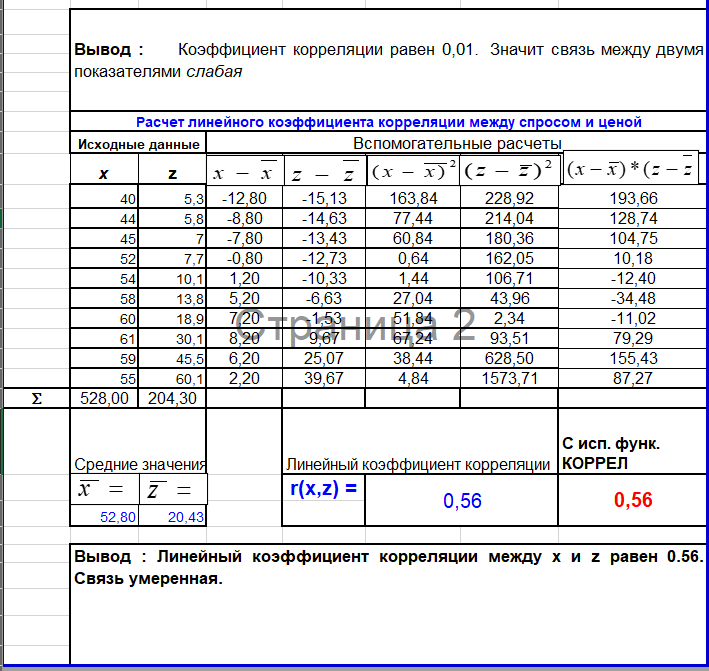


Рисунок 4 – Расчет линейного коэффициента корреляции между x и z

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.

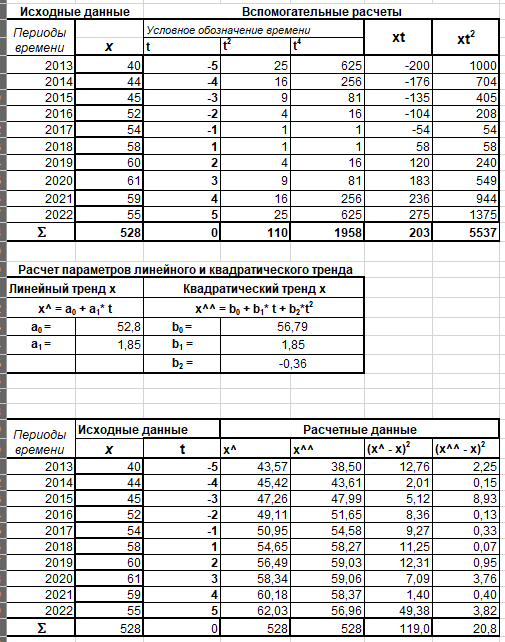


Рисунок 5 – Расчет показателей тредов для Х

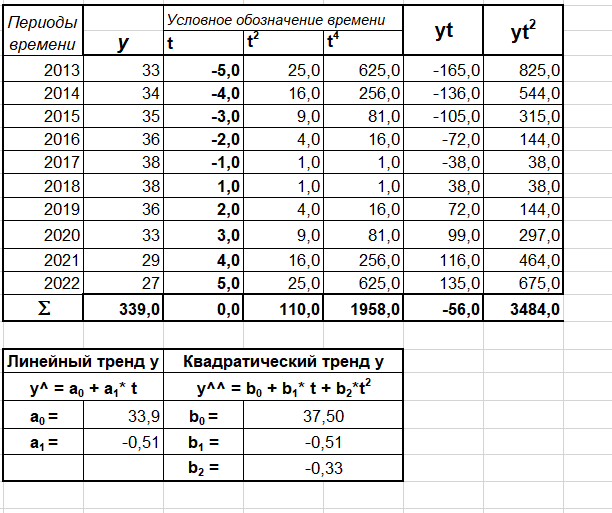


Рисунок 6 – Расчет показателей тредов для Y

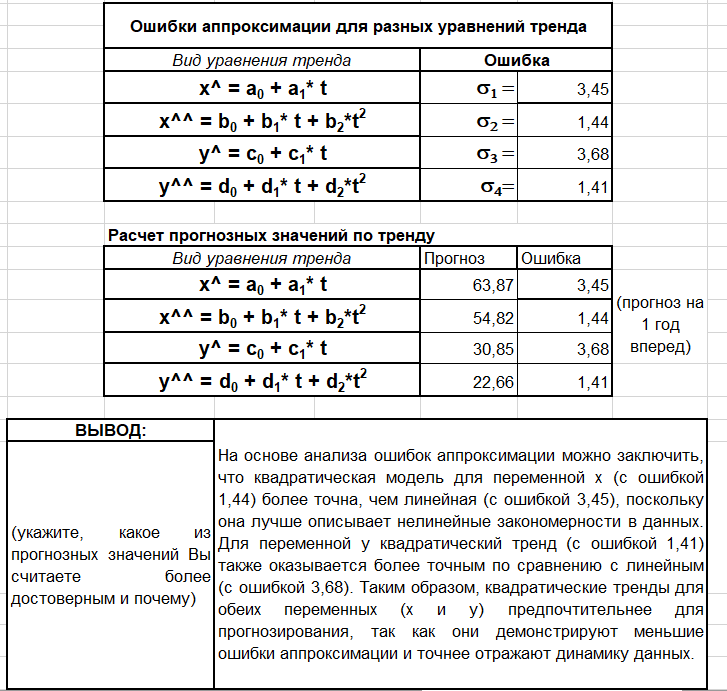


Рисунок 7 – Ошибки аппроксимации и вывод

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).

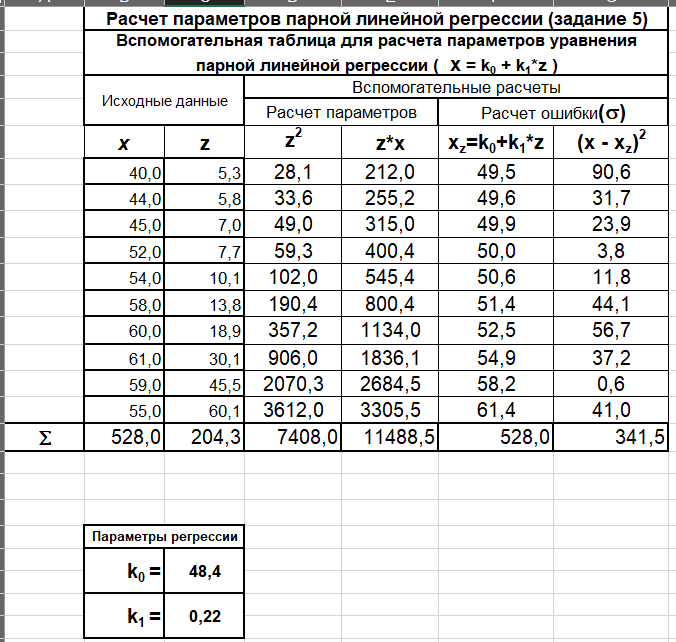


Рисунок 8 – Расчет параметров парной линейной регрессии

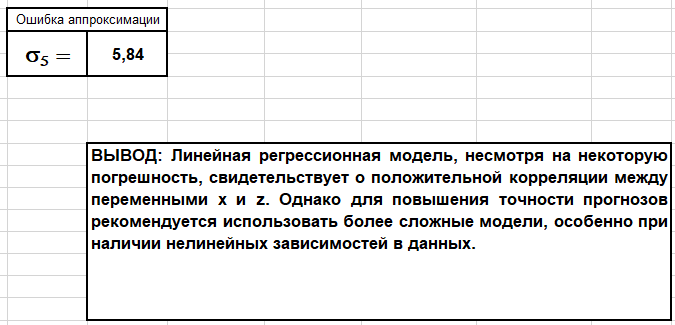


Рисунок 9 – Ошибка аппроксимации, индекс детерминации и вывод

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.

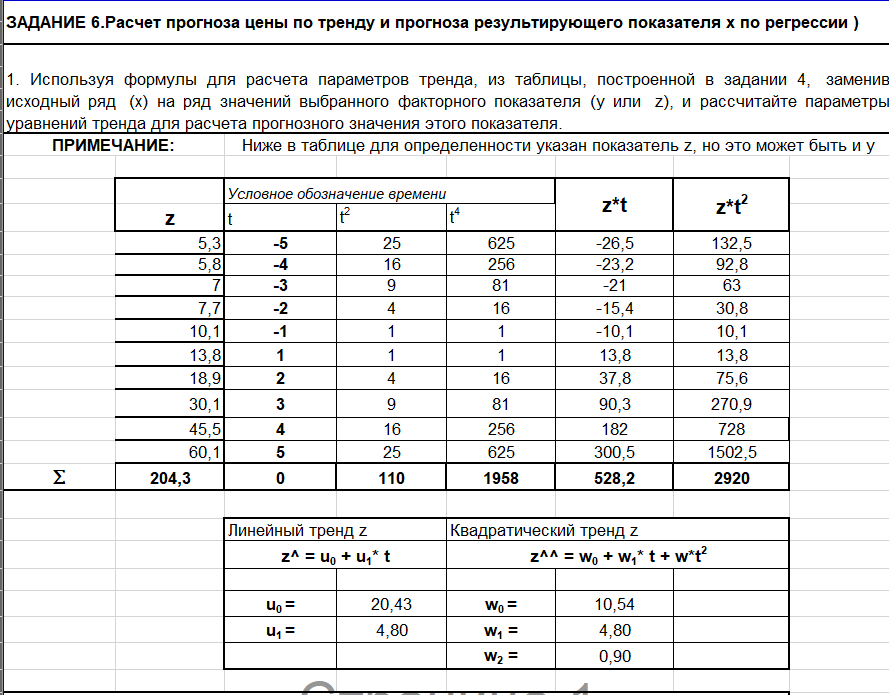


Рисунок 10 – Расчет параметров уравнений трендов

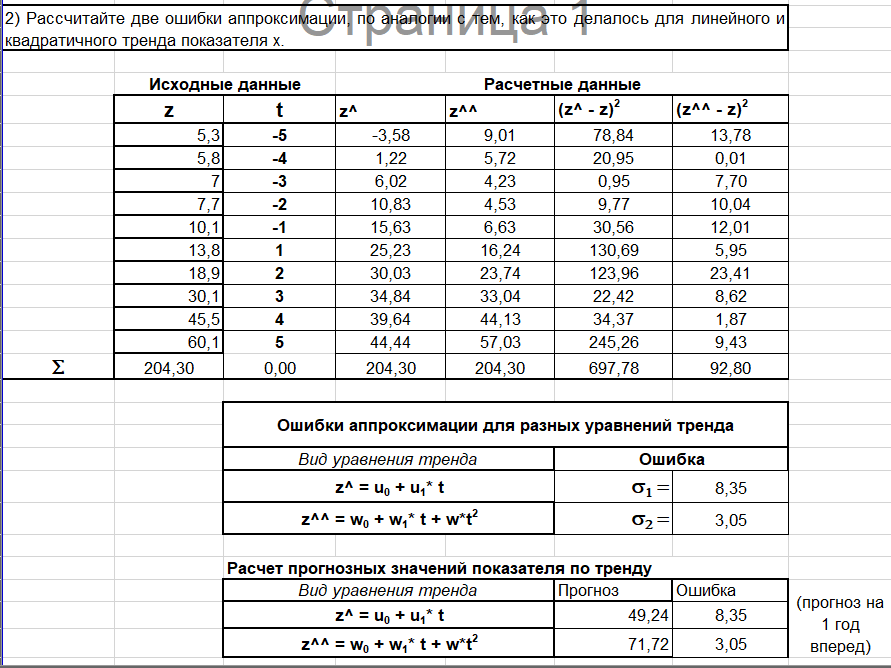


Рисунок 11 – Расчет ошибок аппроксимации

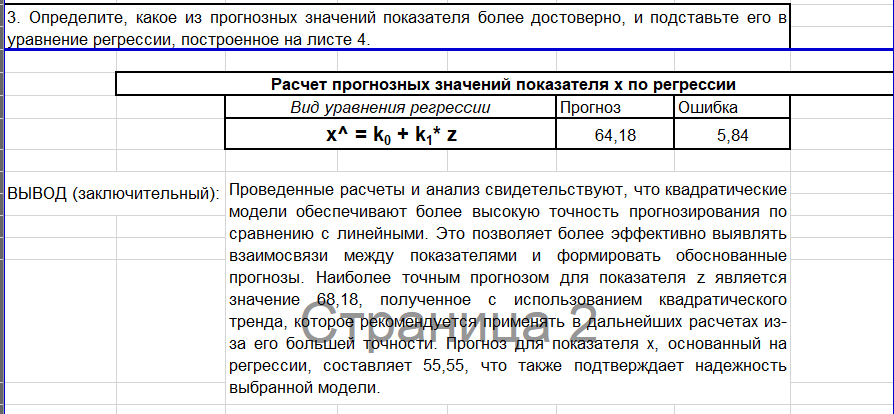


Рисунок 12 – Расчет прогнозных значений показателя x по регрессии и вывод

**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.

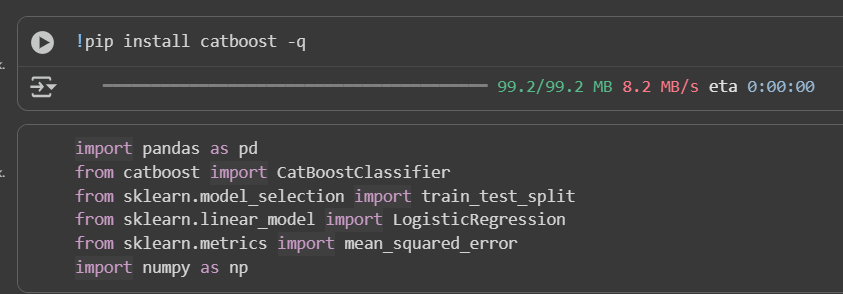


Рисунок 13 – Импорт библиотек

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.

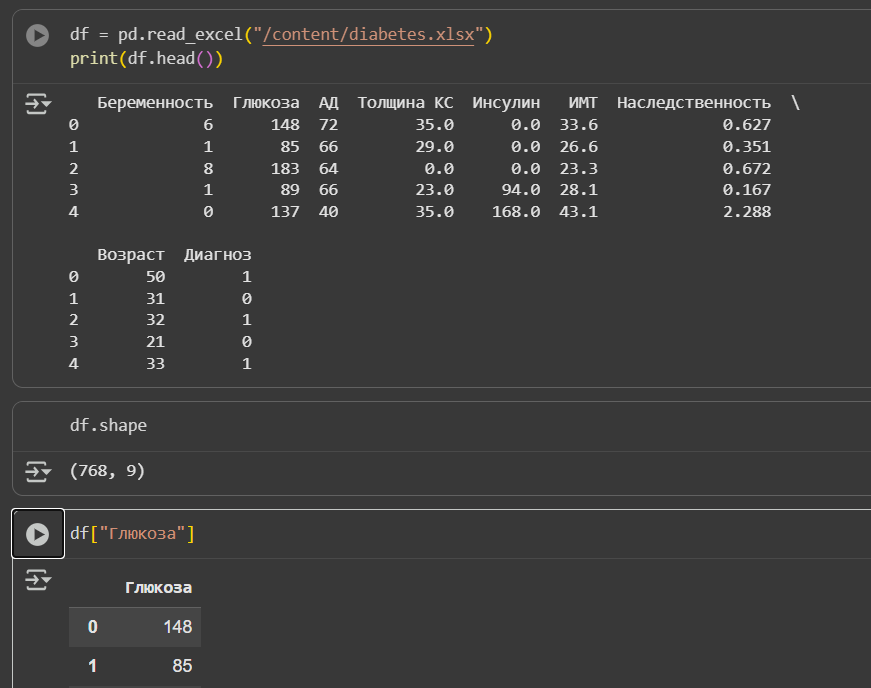


Рисунок 14 – Импорт датасета и проверка данных

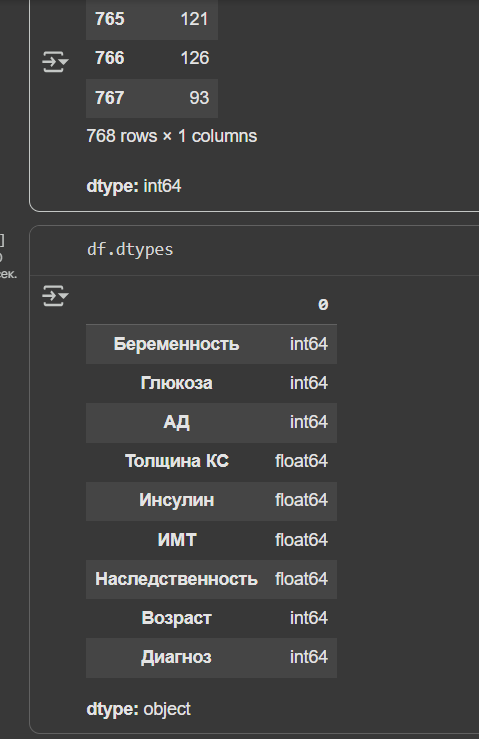


Рисунок 15 – Вывод типов данных датасета

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.

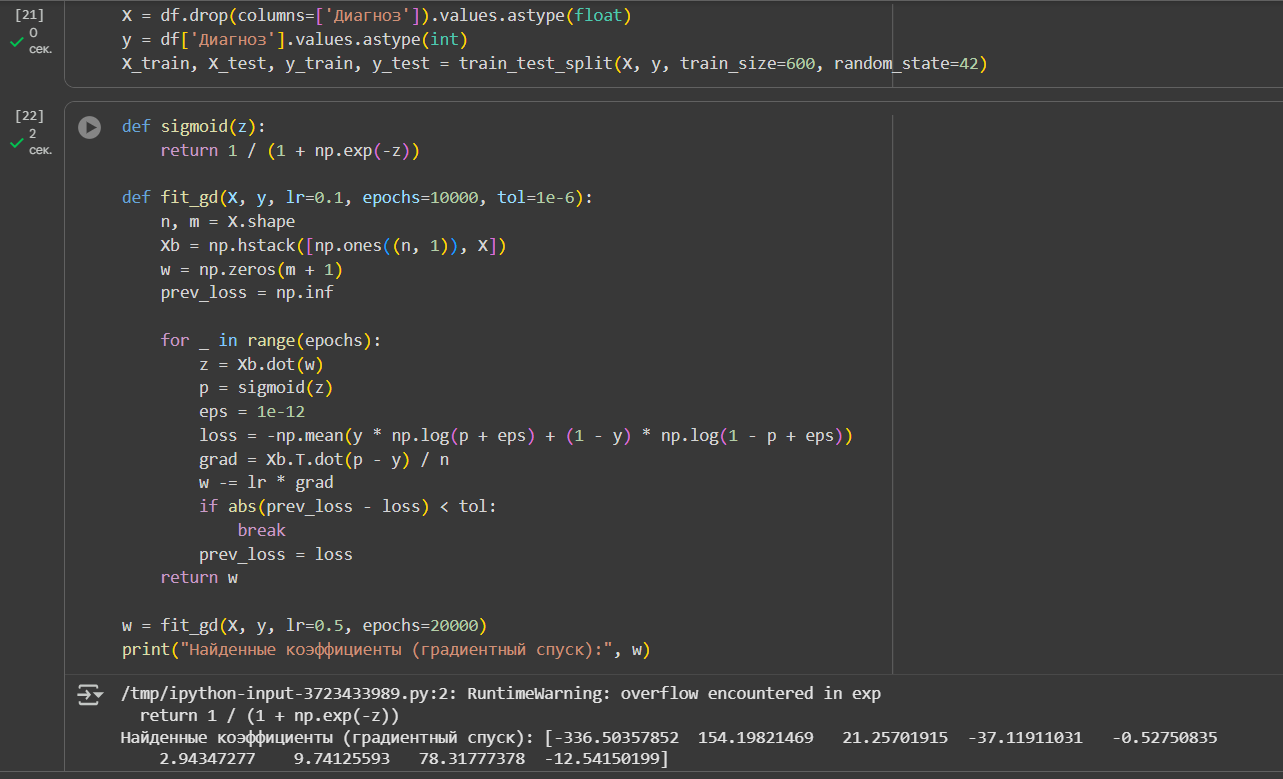


Рисунок 16 – Разбиение данных и подготовка функций

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).

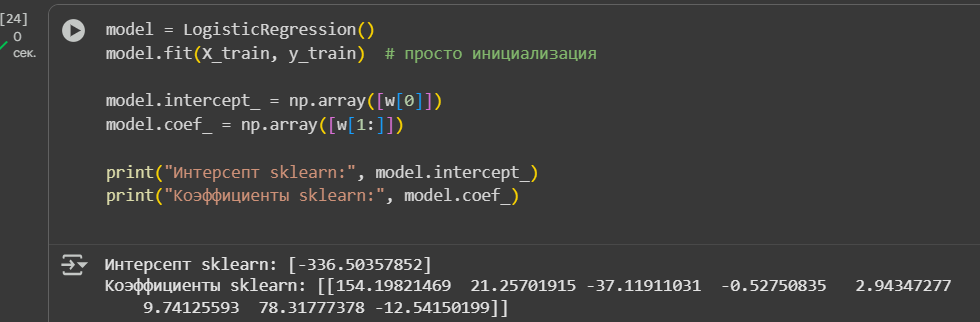


Рисунок 17 – Вывод интерсепта

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).

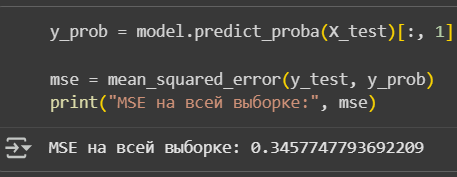


Рисунок 18 – Вычисление MSE

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).

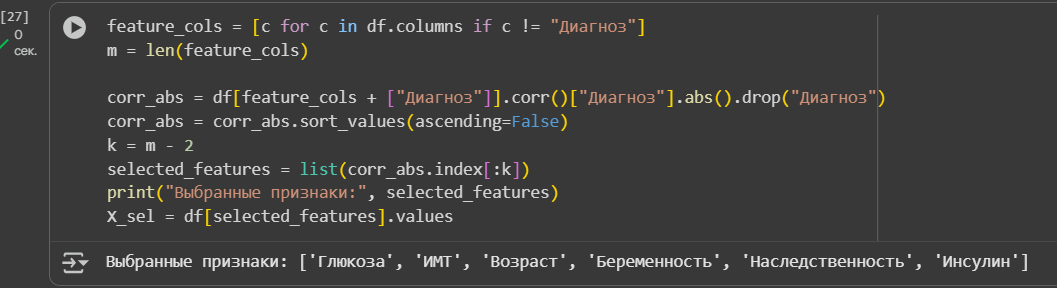


Рисунок 19 – Выбор признаков

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).

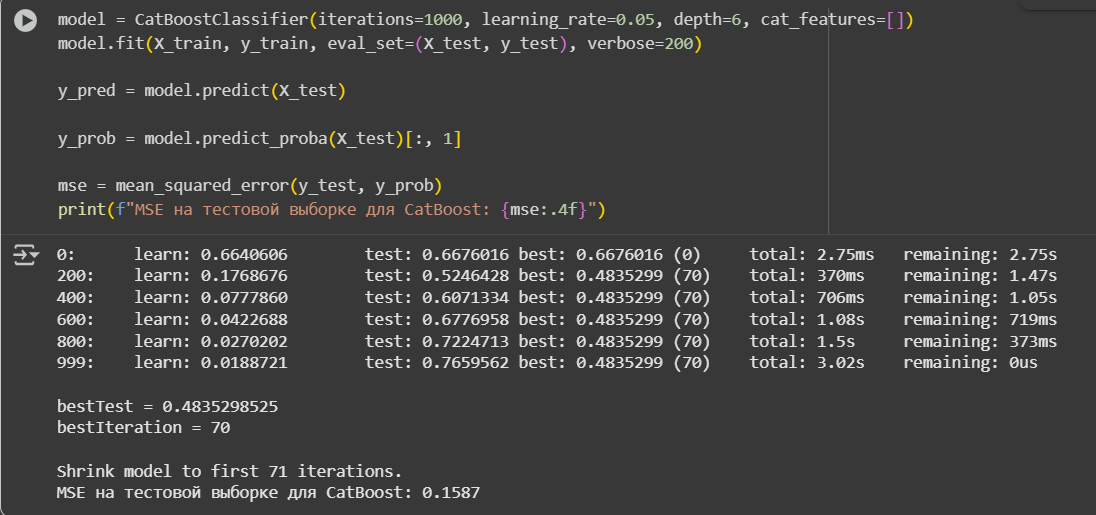


Рисунок 20 – Обчуение CatBoost

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.