МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«Дальневосточный федеральный университет»

(ДВФУ)

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

(ШКОЛА)

Департамент информационных и компьютерных систем

ОТЧЕТ

по дисциплине «системы искусственного интеллекта»

Выполнил студенты группы Б9122-09.03.03пикд

Зверев Р. И.

Проверил преподаватель

Бочарова В. В.

зачтено/не зачтено

г. Владивосток

2025 г

Оглавление

Цель работы

Целью работы является применение метода главных компонент (PCA) для задачи снижения числа признаков датасета с минимальными потерями.

Постановка задачи

В данной работе рассматриваются задачи снижения размерности датасета на двух примерах: датасет GiveMeSomeCredit и датасет Turkiye Student Evaluation.

Необходимо реализовать следующие этапы и функции:

- Предполагаемое снижение размерности датасета GiveMeSomeCreadit и подтвердить выводы из методички;
- Выбрать предмет из датасета опроса и произвести по нему PCA;
- Выбрать два предмета одного преподавателя из датасета опроса и произвести по ним PCA, сравнив с предыдущим пунктом;
- Произвести PCA для всего датасета и сравнить результаты;
- Произвести PCA для пунктов, но без стандартизации.

Введение

Целью работы является применение метода главных компонент (PCA) для снижения размерности наборов данных с минимальными потерями информации. На практике это позволит упростить модели, ускорить обучение и выявить основные направления вариативности признаков.

В работе используются два набора данных: GiveMeSomeCredit (финансовые данные кредитного скоринга) и Turkiye Student Evaluation (опросные данные по оценке преподавателей/предметов). В рамках задания необходимо реализовать несколько сценариев применения PCA: предположить и обосновать возможное снижение размерности для GiveMeSomeCredit; провести PCA для выбранного предмета опроса; выполнить PCA для двух предметов одного преподавателя и сравнить результаты с предыдущим случаем; применить PCA ко всему датасету и сопоставить выводы; а также повторить анализ для отдельных пунктов без предварительной стандартизации признаков, чтобы оценить влияние масштабирования на результаты.

В качестве инструментов используются Python-библиотеки pandas, numpy, scikit-learn (PCA, StandardScaler), а также средства визуализации (matplotlib) для иллюстрации доли объяснённой дисперсии и интерпретации компонент.

Воспроизведение вычислений лабораторной

Описание набора данных GiveMeSomeCredit.

Для проведения анализа был выбран набор данных Give Me Some Credit, который содержит информацию о клиентах банка. Датасет включает 251 тыс. записей (клиентов) и 10 признаков (столбцов).

Основные столбцы, представленные в датасете, и их описание:

- **RevolvingUtilizationOfUnsecuredLines**: Общий баланс по кредитным картам и личным кредитным линиям, за исключением долга по недвижимости и без рассрочки;
- **Age**: Возрастзаемщика в годах;
- **NumberOfTime30-59DaysPastDueNotWorse**: Количество просроченных платежей заемщика на 30-59 дней, но не больше чем за последние 2 года;
- **DebtRatio**: Ежемесячные выплаты по долгу, алименты, расходы на жизнь, разделенные на ежемесячный валовой доход;
- **MonthlyIncome**: Ежемесячный доход;
- **NumberOfOpenCreditLinesAndLoans**: Количество открытых займов (рассрочка, например, автокредит или ипотека) и кредитных линий (например, кредитные карты);
- **NumberOfTimes90DaysLate**: Количество просроченных платежей заемщика на 90 дней или более;
- **NumberRealEstateLoansOrLines**: Количество ипотечных кредитов и ссуд на недвижимость, включая кредитные линии под залог собственного капитала;
- **NumberOfTime60-89DaysPastDueNotWorse**: Количество раз, когда заемщик просрочил платеж на 60-89 дней, но не больше чем за последние 2 года;
- **NumberOfDependents**: Количество иждивенцев в семье, исключая их самих (супруга, дети и т.д.).

После удаления пустых значений из данного датасета остается 200 тыс. записей.

После применения стандартизации и PCA получается следующий график:

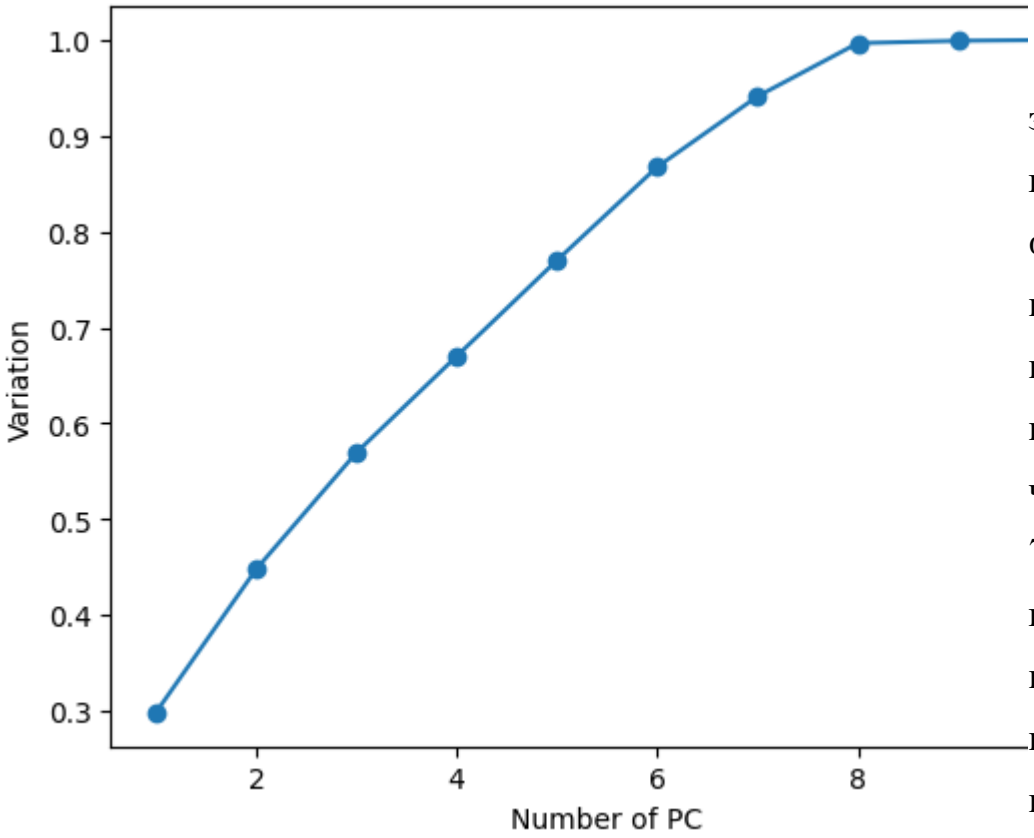

Рисунок 1. Влияние главных компонент

Из графика видно, что 7 главных компонент описывают ~95% дисперсии данных, а 8 компонент уже ~99%. Можно с минимальными потерями снизить количество признаков до 7 - 8 штук. Это полностью совпадает с рассчетами и выводами в лабораторной в методичке.

Исследование датасета опроса

Исходный набор данных Turkiye Student Evaluation содержит ответы студентов на вопросы о качестве преподавания и содержит 5280 записей и 33 признака. Полное описание:

- **Instr**: идентификатор инструктора, значения взяты из {1,2,3};
- **Class**: код курса (дескриптор), значения взяты из {1-13};
- **Nb.repeat**: сколько раз студент проходил этот курс, значения взяты из {0,1,2,3, …};
- **Attendance**: код уровня посещаемости, значения из {0, 1, 2, 3, 4};
- **Difficulty**: уровень сложности курса, который воспринимается студентом; значения взяты из {1,2,3,4,5};
- **Q1**: содержание семестрового курса, метод обучения и система оценивания были предоставлены в начале;
- **Q2**: цели и задачи курса были четко сформулированы в начале периода;
- **Q3**: курс стоил присвоенной ему суммы кредита;
- **Q4**: курс преподавался в соответствии с программой, объявленной в первый день занятий;
- **Q5**: обсуждения в классе, домашние задания, приложения и исследования были удовлетворительными;
- **Q6**: учебники и другие ресурсы курсов были достаточными и актуальными;
- **Q7**: курс допускал полевые работы, приложения, лабораторные, обсуждения и другие исследования;
- **Q8**: тесты, задания, проекты и экзамены способствовали обучению;
- **Q9**: мне очень понравился урок, и я очень хотел активно участвовать во время лекций;
- **Q10**: мои первоначальные ожидания относительно курса оправдались в конце периода или года;

- **Q11**: курс был актуален и полезен для моего профессионального развития;
- **Q12**: курс помог мне взглянуть на жизнь и мир с новой точки зрения;
- **Q13**: знания инструктора были актуальными и актуальными;
- **Q14**: инструктор прибыл подготовленным к занятиям;
- **Q15**: инструктор преподавал в соответствии с объявленным планом урока;
- **Q16**: инструктор был привержен курсу и был понятен;
- **Q17**: инструктор прибыл вовремя на занятия;
- **Q18**: инструктор легко и четко произносит речь;
- **Q19**: инструктор эффективно использовал часы занятий;
- **Q20**: преподаватель объяснил курс и очень хотел помочь студентам;
- **Q21**: преподаватель продемонстрировал положительный подход к студентам;
- **Q22**: преподаватель был открыт и уважительно относился к мнению студентов о курсе
- **Q23**: инструктор поощрял участие в курсе
- **Q24**: преподаватель давал соответствующие домашние задания проекты и помогал руководил студентами
- **Q25**: инструктор ответил на вопросы о курсе внутри и вне курса;
- **Q26**: система оценки преподавателя (промежуточные и заключительные вопросы, проекты, задания и т. Д.) Эффективно измеряла цели курса;
- **Q27**: преподаватель предоставил решения к экзаменам и обсудил их со студентами;
- **Q28**: преподаватель относился ко всем студентам правильно и объективно;

- Q1 — Q28 относятся к типу Лайкерта, что означает, что значения взяты из {1,2,3,4,5}.

В качестве исследуемого предмета был взят предмет №11 и вот его график главных компонент:

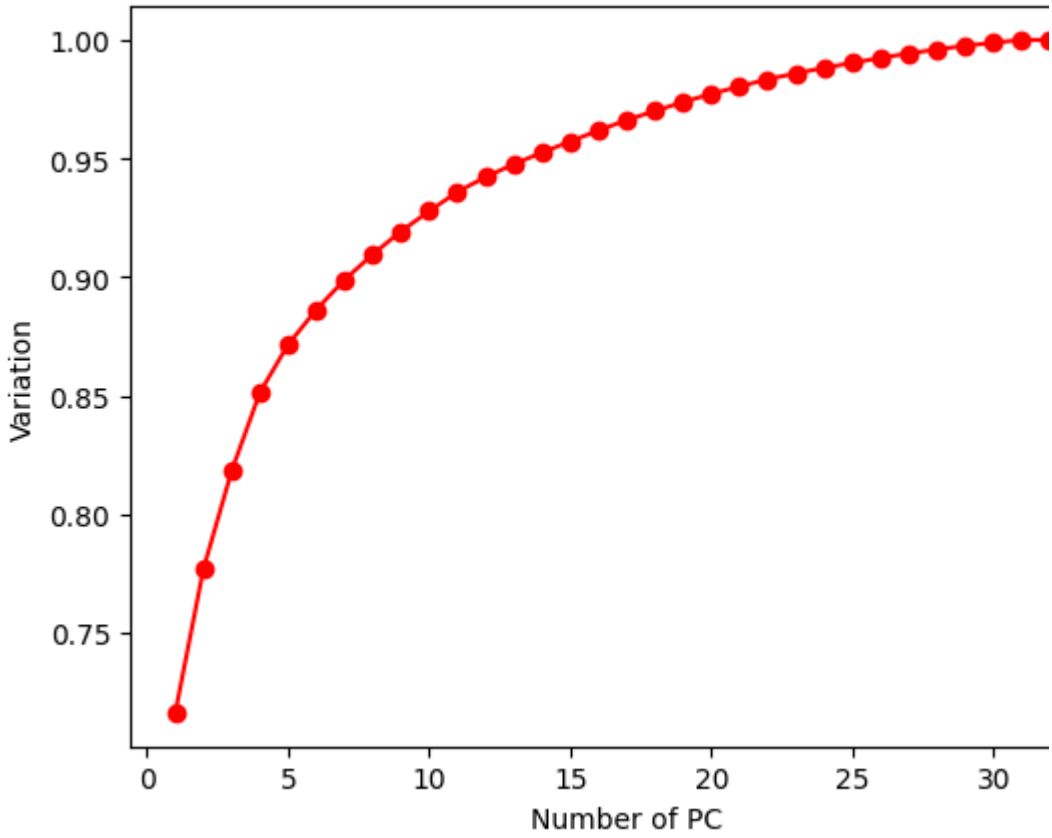

Рисунок 2. PCA для предмета №11

Видно, что 25 главных компонент описывают ~99% дисперсии.

В качестве двух предметов одного преподавателя были выбраны предметы №3 и №9 преподавателя 3. Вот их анализ главных компонент:

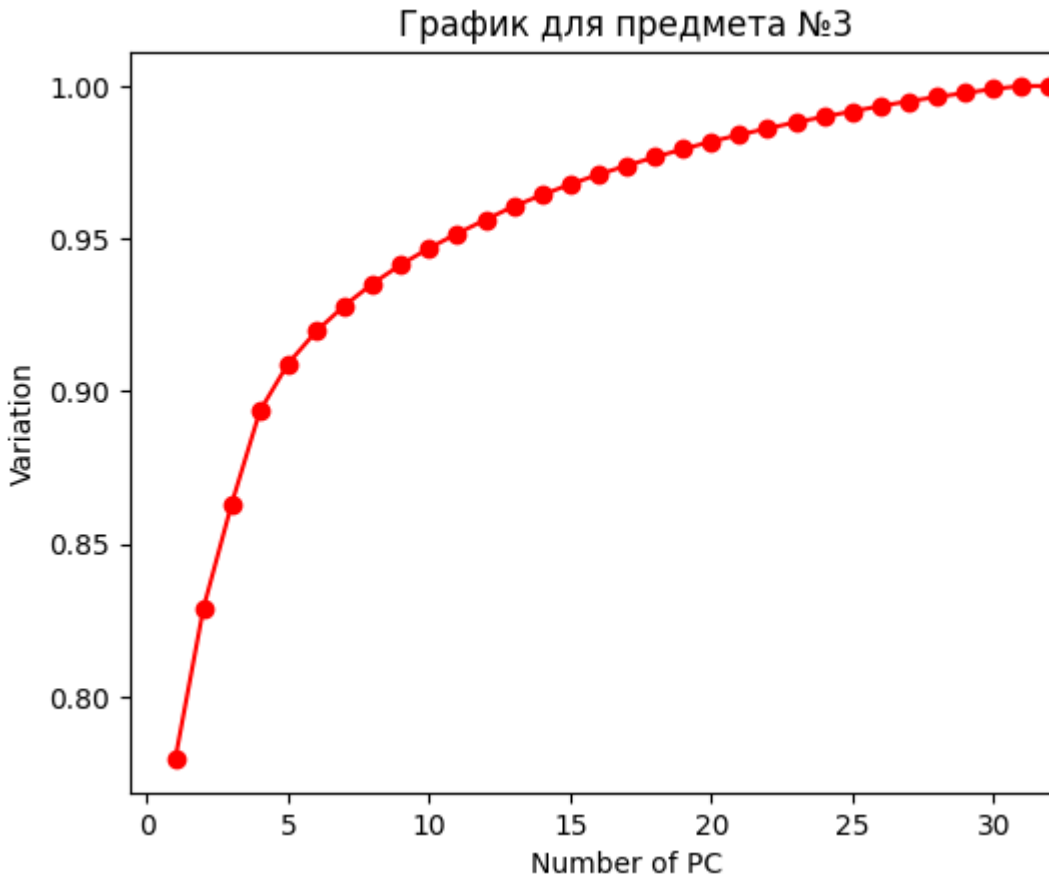

Рисунок 3. PCA для предмета №3

Для предмета №3 25 главных компонент описывают 99% дисперсии.

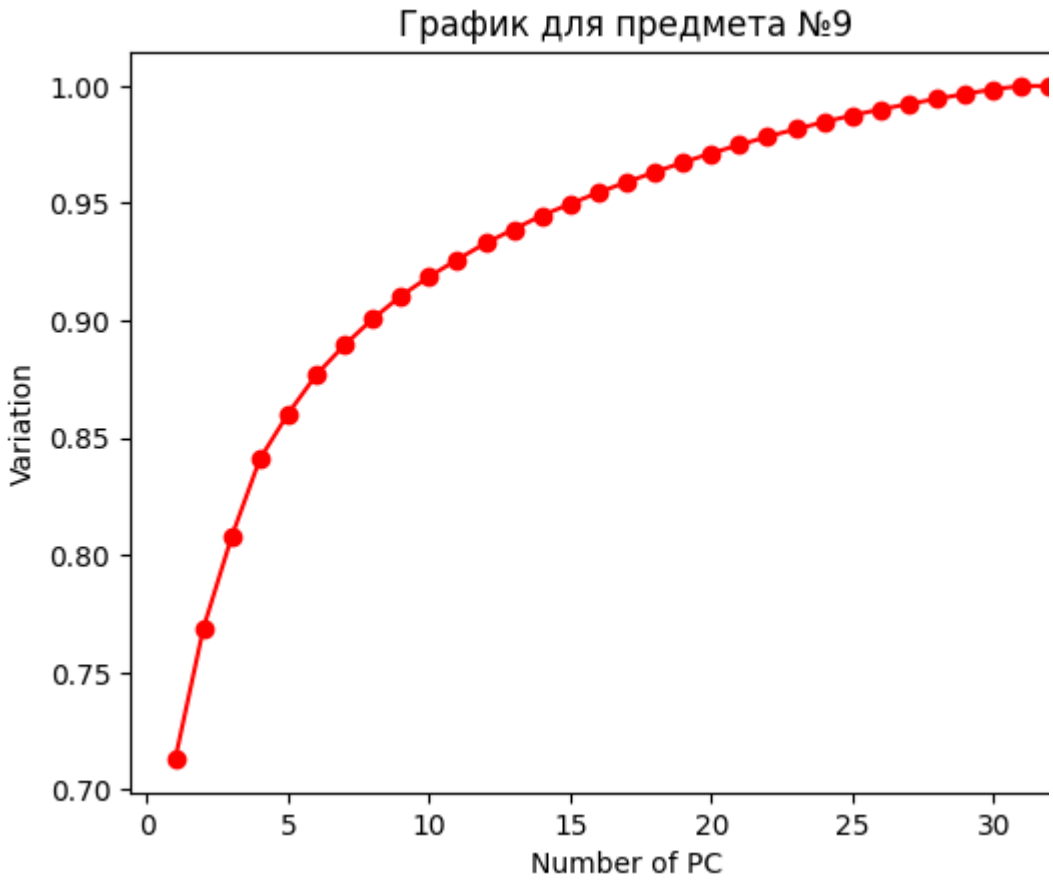

Рисунок 4. PCA для предмета №9

Для предмета №9 26 главных компонент описывают 99% дисперсии, что на одну компоненту хуже, чем у предмета №3. Такая группировка показывает, что оценки учеников для предметов №3 и №9 более разнообразные и требуют больше компонент для описания.

Для всего датасета график выглядит так:

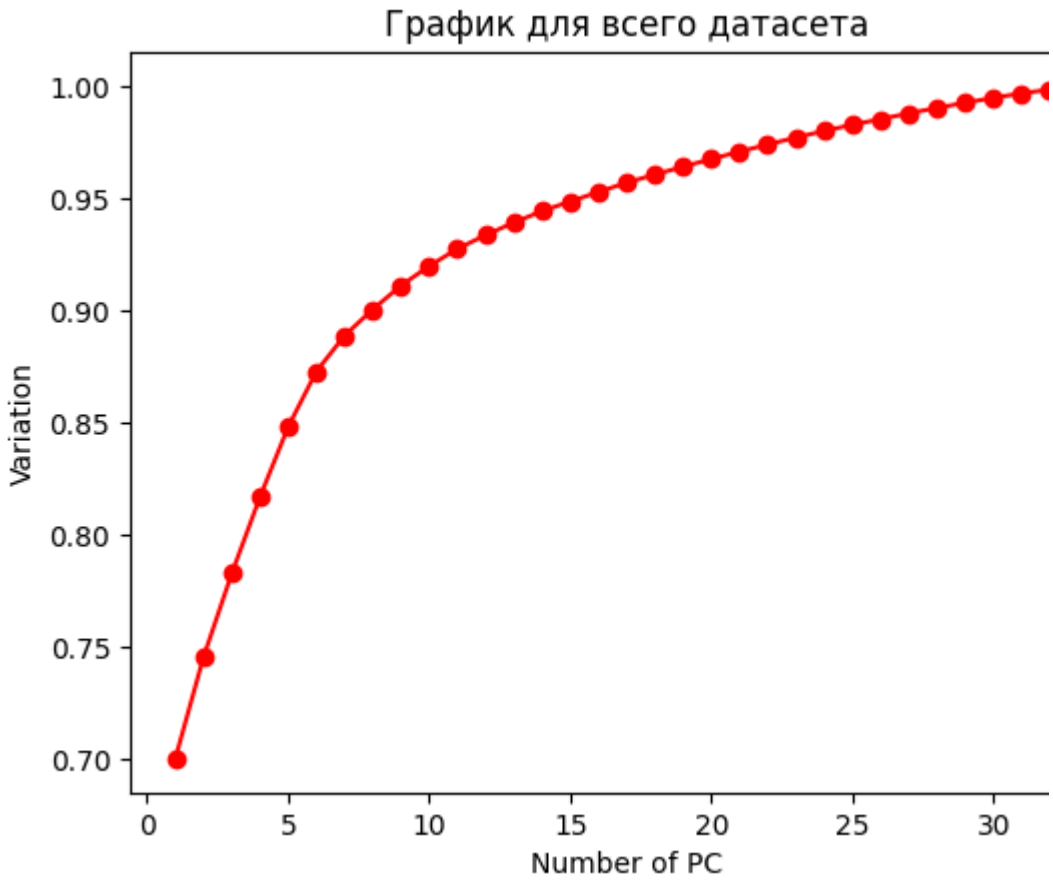

Рисунок 5. PCA для всего датасета

Видно, что 28 главных компонент описывают ~99% дисперсии.

Исходя из пунктов выше, можно сказать что вариативность между предметами и преподавателями добавляет сложности, которую PCA должен учесть.

Также, был произведене PCA для нестандартизированных данных тех же выборок:

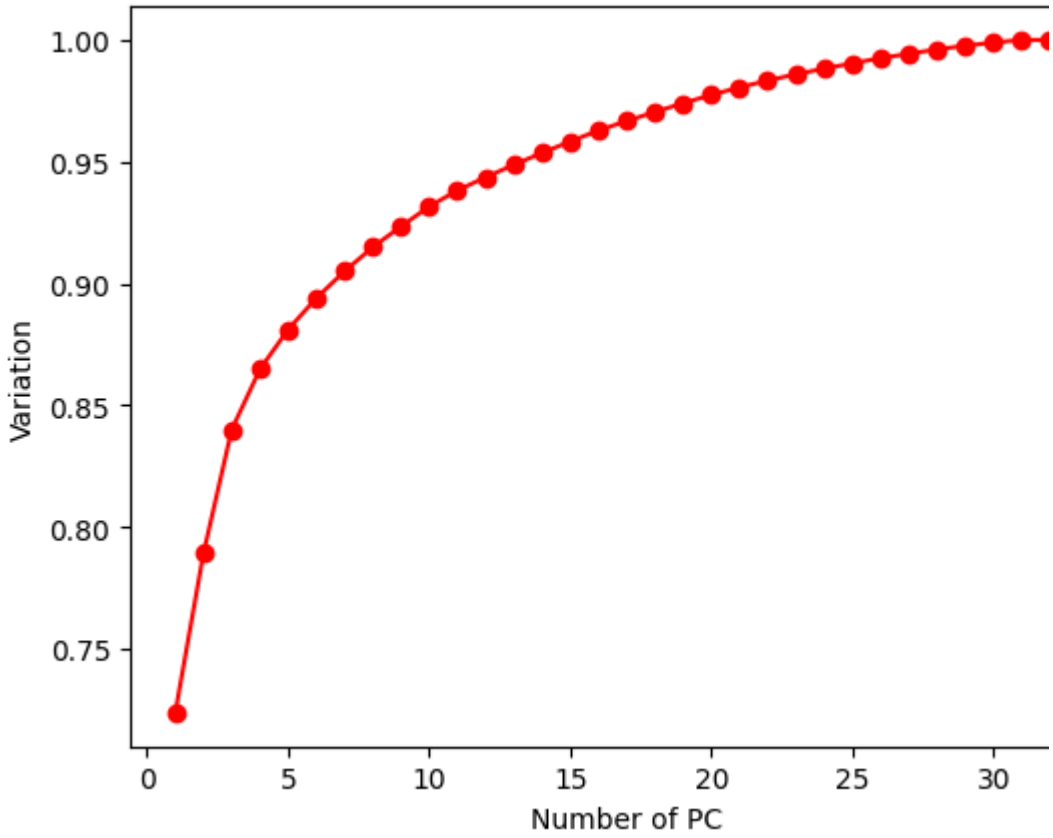

Рисунок 6. PCA для нестандартизированных данных предмета №11

В таком виде требуется 25 компонент, что не отличается от стандартизированного варианта.

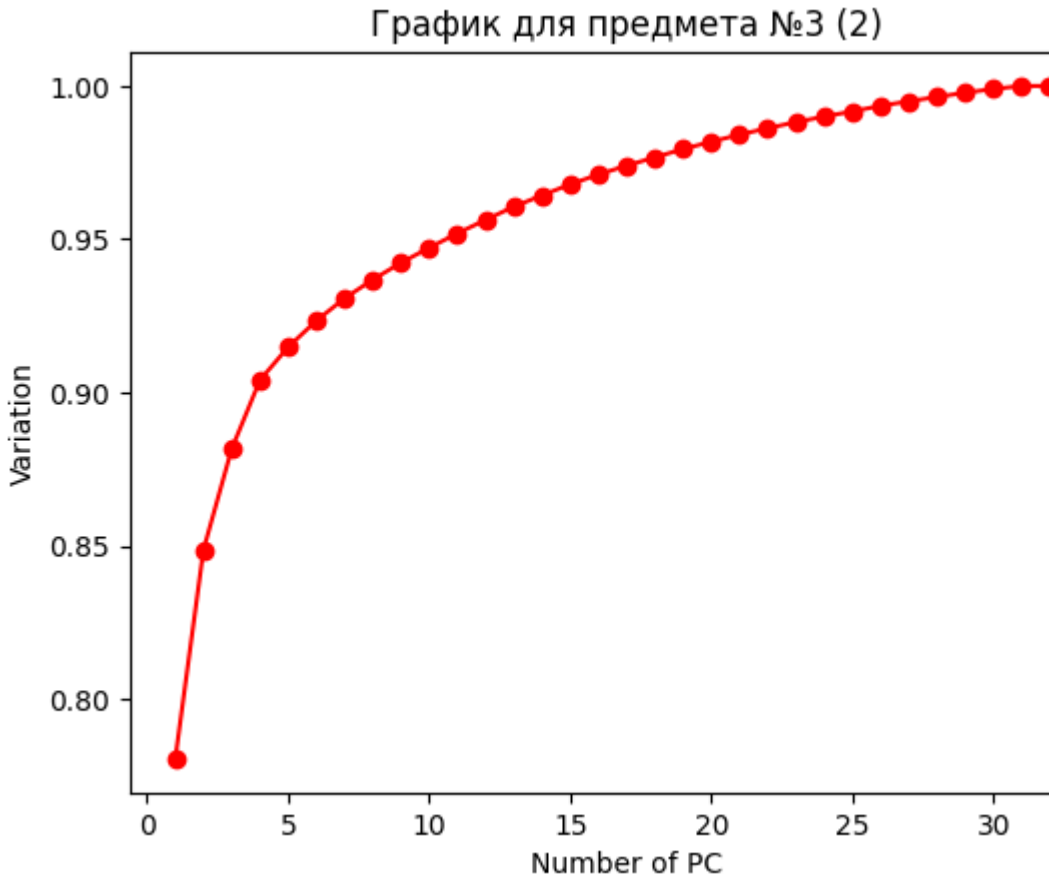

Рисунок 7. PCA для нестандартизированных данных предмета №3

В нестандартизированном виде требуется 24 компоненты для описания 99% дисперсии предмета №3, что быстрее на одну компоненту, чем в стандартизированном варианте.

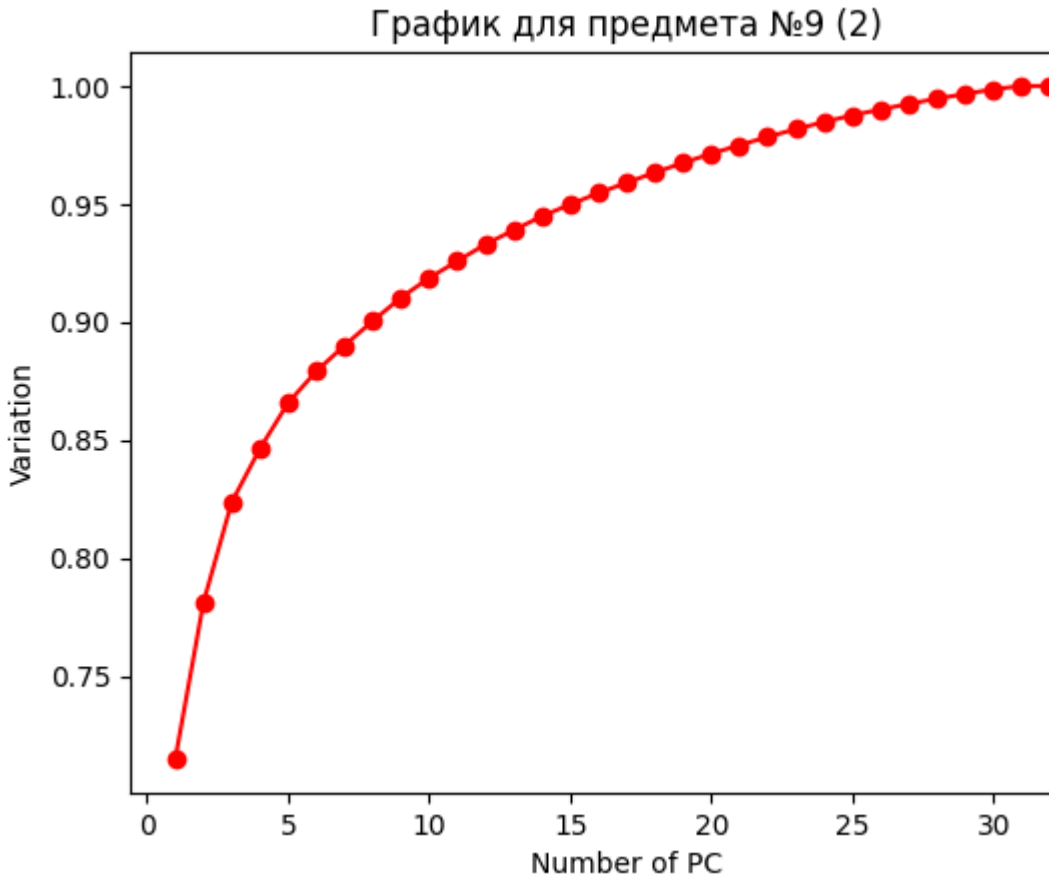

Рисунок 8. PCA для нестандартизированных данных предмета №9

Требуется так же 26 компонент, что и в стандартизированном варианте.

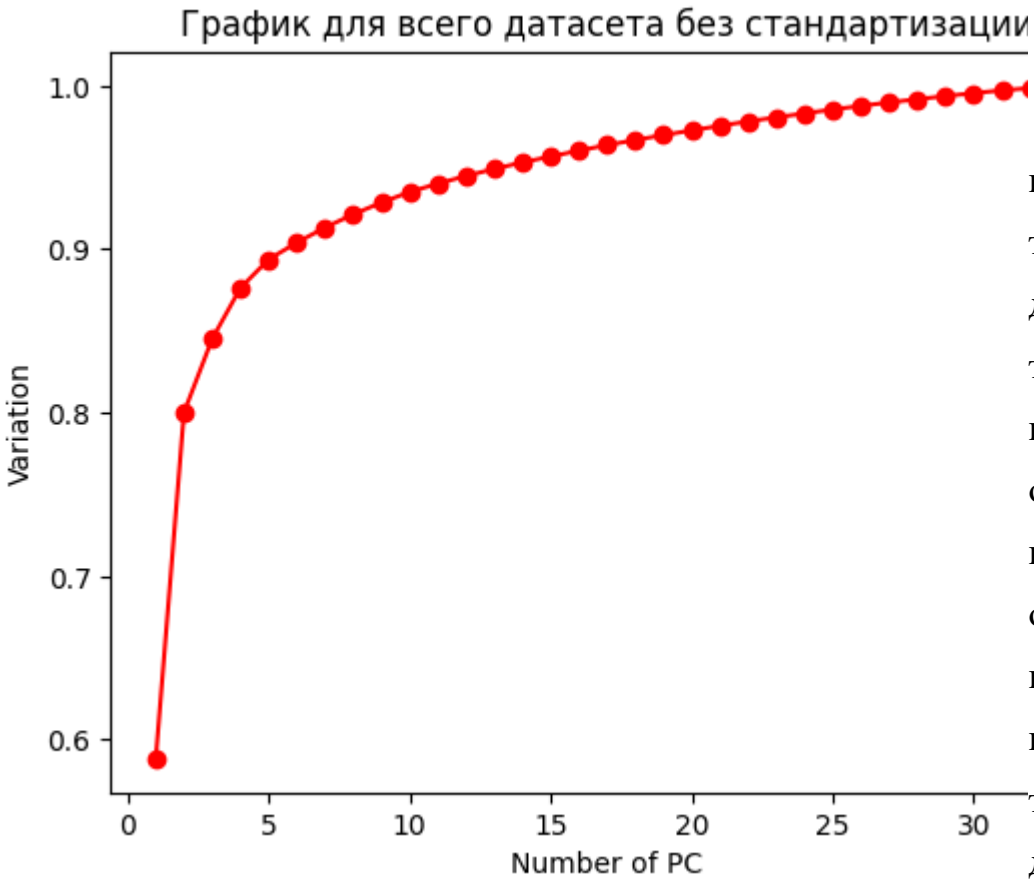

Рисунок 9. PCA для нестандартизированного датасета

В нестандартизированном варианте для всего датасет требуется столько же компонент, что и в стандартизированном.

Заключение

В ходе выполнения данной работы были успешно получены навыки в построении и интерпритации метода главных компонент для снижения размерности данных.

На наглядных примерах показано, что в ряде случаев можно снизить размерность, убрав несколько признаков, сохранив при этом основную часть дисперсии данных.

Также, было продемонтсрированно сравнение стандартизированных и нестандартизированных данных в PCA. Местами нестандартизированные данные даже лучше, так как им требуется меньше компонент для описания 99% дисперсии, но это, скорее, ошибка датасета, так как в нем нет выбросов. PCA чувствителен к выбросам и лучше всегда стандартизировать данные.

Список литературы

1. GitHub: исходный код лабораторной работы. – URL: Лабораторная работа №3.3 (дата обращения: [09.10.2025]). – Текст: электронный.