Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

«СЕВЕРО-КАВКАЗСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

|  |  |
| --- | --- |
| Институт | математики и компьютерных наук |
| Кафедра | компьютерной безопасности |

ОТЧЕТ

ЛАБОРАТОРНАЯ РАБОТА 11.

АНАЛИЗ МНОГОМЕРНЫХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ СНИЖЕНИЯ РАЗМЕРНОСТИ

Выполнил: Окунев Николай Александрович,

студент 2 курса

группы КМБ-с-о-23-1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Проверено с оценкой:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2025 г.

Ставрополь, 2025

1. **Цели и задачи**

Цель лабораторной работы: научиться применять обучаемые алгоритмы снижения размерности для визуализации наборов данных сбольшой размерности. Основные задачи: – получение навыков рефакторинга кода в проектах машинного обучения; – получение навыков по использованию обучаемых алгоритмов снимжения размерности; – получение навыков по аргументрированному выбору и анализу алгоритмов снижения размерности.

1. **Теоретическое обоснование**

Необходимо повторить теоретический материал работы 10.

1. **Методика и порядок выполнения работы**
   1. Учебная задача Задание. На основе предоставленного набора данных Fashion-MNIST выполните снижение размерности. Исходный набор данных содержит изображения размером 28x28. Для понимания структуры данных, понимания возможности их использования для решения задач кластеризации, классификации, необходимо выполнить визуализацию данных изображений с отображения расстояния между отдельными экземплярами.

Решение:

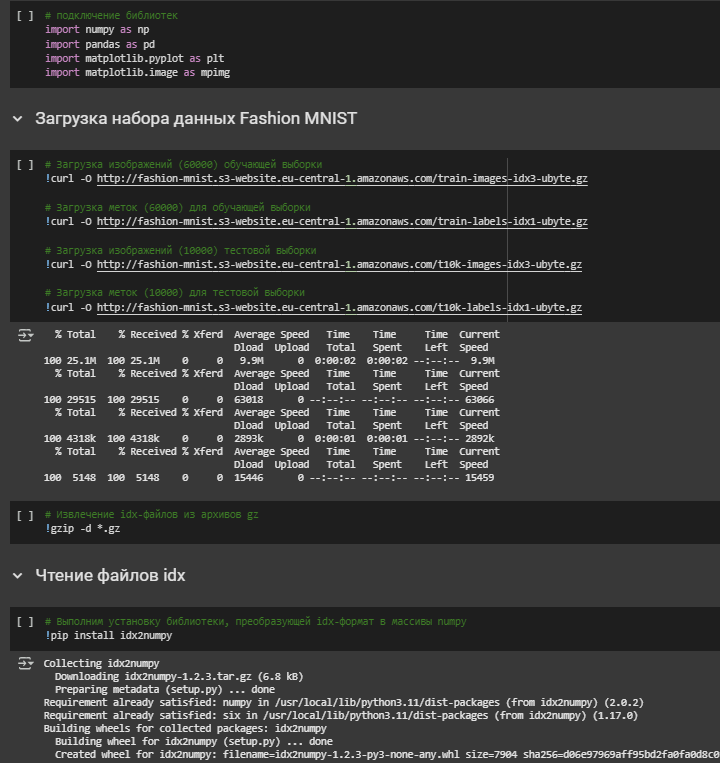


Рисунок 1 - Подключение библиотек и датасета.

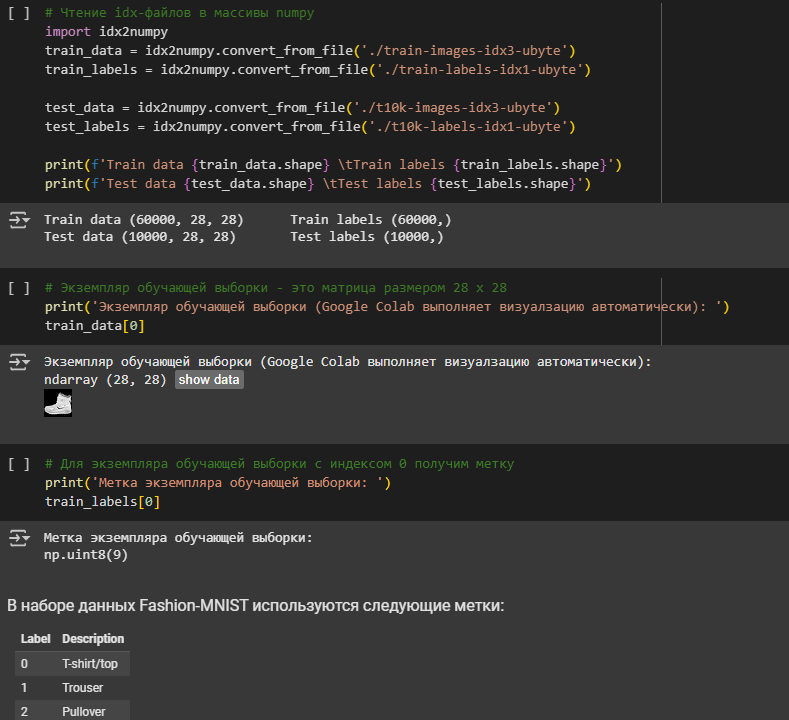


Рисунок 2 - подготовка датасета



Рисунок 3 -вывод содержимого датасета и тэгов

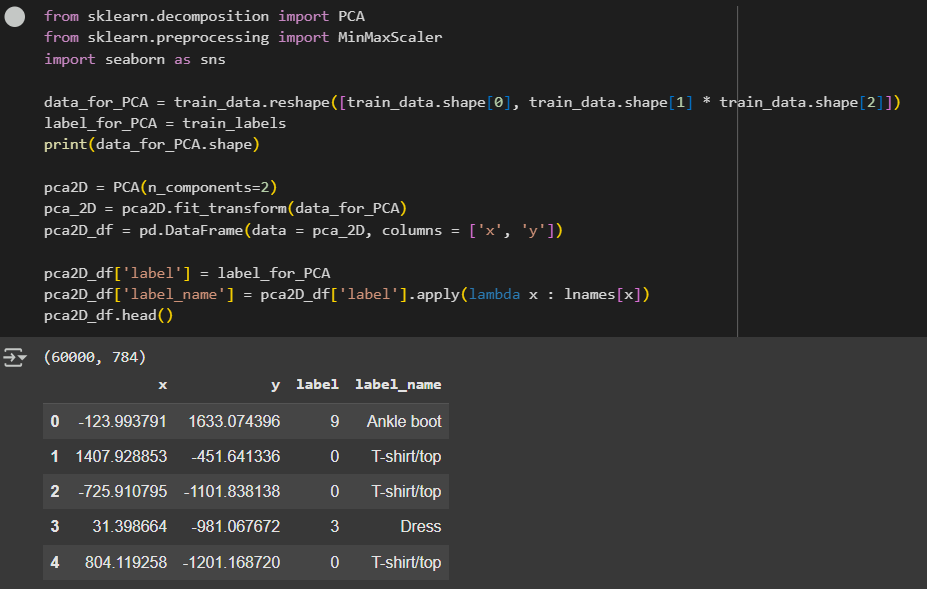


Рисунок 4 - анализируем главные компоненты

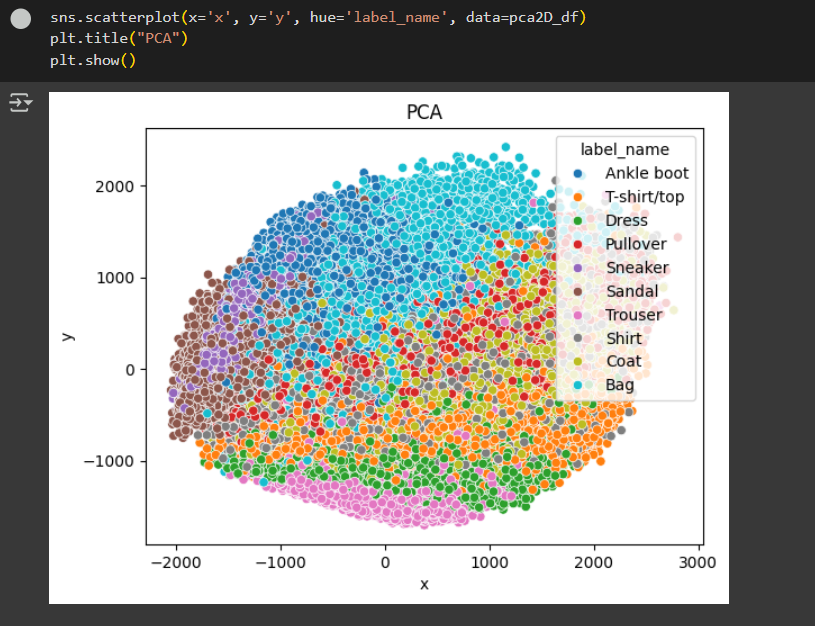


Рисунок 5 - начальный датасет

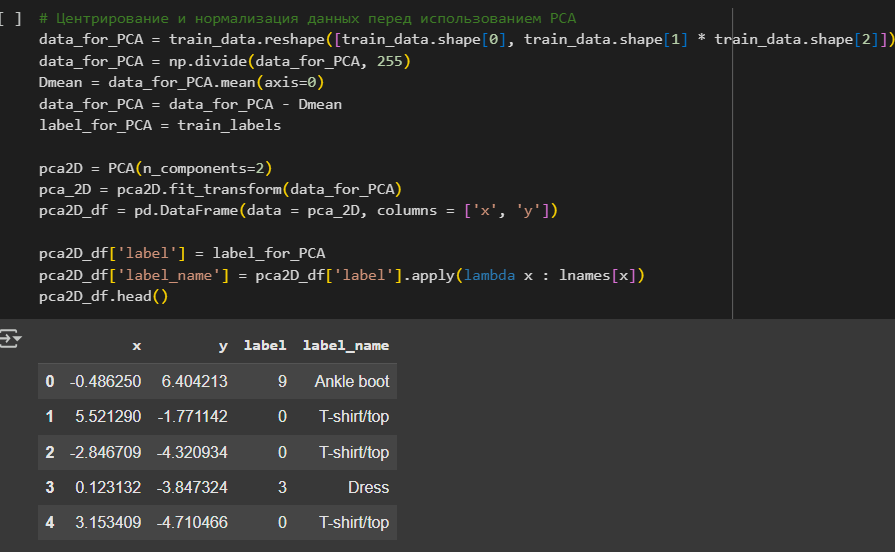


Рисунок 6 - центровка датасета перед сниженем размерности

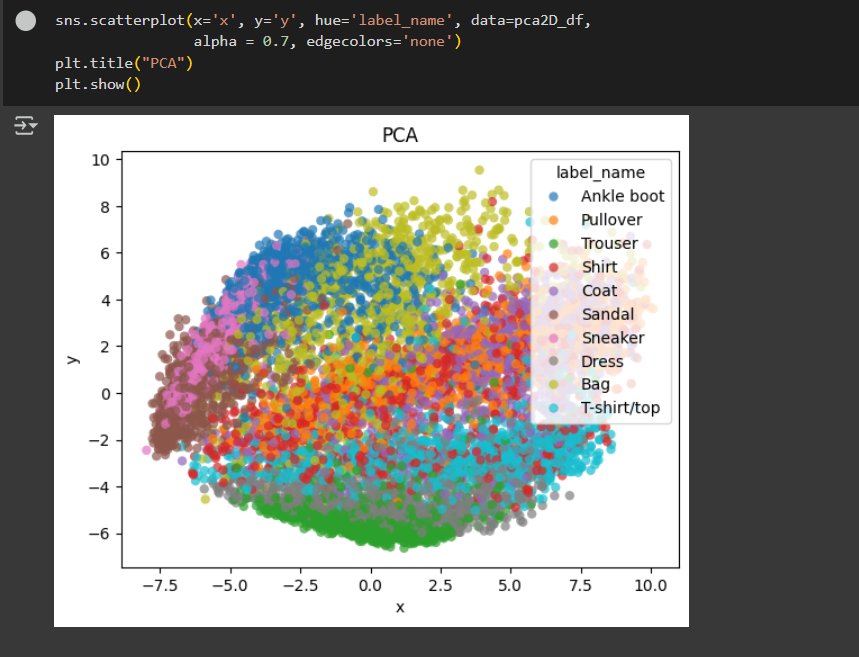


Рисунок 7 - полученный датасет с пониженной размерностью

1. **Контрольные вопросы**

1)Что такое проклятие размерности?

"Проклятие размерности" (Curse of Dimensionality) — это явление, которое возникает при работе с данными высокой размерности. Оно заключается в том, что с увеличением количества признаков (размерности) объем пространства данных экспоненциально увеличивается, что приводит к ряду проблем:

Разреженность данных: Данные становятся более разреженными, то есть расстояния между объектами увеличиваются, и объекты становятся менее похожими друг на друга.

Переобучение: Модели машинного обучения становятся более склонными к переобучению, так как они могут запомнить шум в данных.

Вычислительная сложность: Вычислительная сложность алгоритмов увеличивается экспоненциально с ростом размерности.

Визуализация: Визуализация данных становится невозможной при размерности больше 3.

2)Перечислите основные алгоритмы понижения размерности.

Существует много разных алгоритмов понижения размерности, вот некоторые из основных:

PCA (Principal Component Analysis): Метод главных компонент.

t-SNE (t-distributed Stochastic Neighbor Embedding): t-распределенное стохастическое вложение соседей.

LDA (Linear Discriminant Analysis): Линейный дискриминантный анализ.

NMF (Non-negative Matrix Factorization): Неотрицательная матричная факторизация.

Autoencoders: Автокодировщики (нейронные сети).

ICA (Independent Component Analysis): Независимый компонентный анализ.

3)Опишите алгоритм и основную идею метода PCA.

Основная идея PCA: PCA — это метод, который используется для уменьшения размерности данных путем нахождения главных компонент. Главные компоненты — это новые признаки, которые являются линейными комбинациями исходных признаков и объясняют наибольшую дисперсию в данных. Алгоритм PCA:

Стандартизация данных: Исходные данные стандартизируются, чтобы иметь нулевое среднее и единичную дисперсию.

Вычисление ковариационной матрицы: Вычисляется ковариационная матрица для стандартизированных данных.

Вычисление собственных векторов и собственных значений: Вычисляются собственные векторы и собственные значения ковариационной матрицы.

Сортировка собственных значений: Собственные значения сортируются в порядке убывания.

Выбор главных компонент: Выбираются k собственных векторов, соответствующих k наибольшим собственным значениям. Эти собственные векторы называются главными компонентами.

Проецирование данных на главные компоненты: Исходные данные проецируются на главные компоненты, чтобы получить данные в новом пространстве меньшей размерности.

4)Опишите концепцию метода t-SNE.

Концепция t-SNE: t-SNE — это метод понижения размерности, который особенно хорошо подходит для визуализации данных высокой размерности.

Сохранение локальной структуры: t-SNE пытается сохранить локальную структуру данных, то есть объекты, которые близки друг к другу в исходном пространстве, должны быть близки друг к другу и в пространстве меньшей размерности.

Вероятностное моделирование: t-SNE использует вероятностное моделирование для определения схожести между объектами.

t-распределение: t-SNE использует t-распределение Стьюдента для моделирования расстояний между объектами в пространстве меньшей размерности. Это позволяет уменьшить эффект "скученности" объектов в центре пространства.

В отличие от PCA, t-SNE является нелинейным методом, что позволяет ему обнаруживать более сложные структуры в данных. Однако, t-SNE является вычислительно сложным алгоритмом и может потребовать много времени для обучения на больших наборах данных.

5)Теоретический предел: В теории, предел снижения размерности определяется количеством информации, содержащейся в данных. Если мы уменьшим размерность до такой степени, что потеряем важную информацию, то модель потеряет свою предсказательную способность. Практический предел: На практике, предел снижения размерности зависит от конкретной задачи и данных.

Визуализация: Если цель — визуализация данных, то обычно уменьшают размерность до 2 или 3.

Улучшение производительности: Если цель — улучшение производительности модели, то уменьшают размерность до тех пор, пока не будет достигнут оптимальный баланс между производительностью и точностью.

Сохранение информации: Если цель — сохранить как можно больше информации, то уменьшают размерность до тех пор, пока не будет потеряно значительное количество дисперсии (например, при использовании PCA).

**Вывод:**

Научились применять обучаемые алгоритмы снижения размерности для визуализации наборов данных с большой размерности.