МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Слушатель Федоренко Д.В, (ФИО)

Москва, 2023

Оглавление

[Введение 3](#__RefHeading___Toc956_833192896)

[Основная часть 4](#__RefHeading___Toc958_833192896)

[1 Аналитическая часть 4](#__RefHeading___Toc960_833192896)

[1.1 Постановка задачи 4](#__RefHeading___Toc962_833192896)

[1.2 Описание используемых методов 4](#__RefHeading___Toc964_833192896)

[1.3 Разведочный анализ данных 8](#__RefHeading___Toc966_833192896)

[2 Практическая часть 1](#__RefHeading___Toc968_833192896)6

[2.1 Предобработка данных 1](#__RefHeading___Toc970_833192896)6

[2.2 Разработка и обучение моделей 2](#__RefHeading___Toc972_833192896)3

[2.3 Сравнение результатов](#__RefHeading___Toc974_833192896) 24

[2.4 Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель](#__RefHeading___Toc976_833192896) 24

[2.5 Создание удаленного репозитория и загрузка на него результатов работы](#__RefHeading___Toc980_833192896) 25

[Заключение](#__RefHeading___Toc982_833192896) 26

# Введение

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

# Основная часть

## Аналитическая часть

### Постановка задачи

Имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

Цель работы:

1. Обучить несколько алгоритмов машинного обучения, которые будут определять значения:

* Модуль упругости при растяжении, ГПа
* Прочность при растяжении, МПа

1. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель

### Описание используемых методов

#### Линейная регрессия

Линейная регрессия - это алгоритм машинного обучения, основанный на обучении под наблюдением. Он выполняет задачу регрессии. Регрессионные модели представляют собой целевое значение прогноза, основанное на независимых переменных. В основном используется для выяснения взаимосвязи между переменными и прогнозирования.

Несколько ключевых моментов линейной регрессии:

* + Моделирование выполняется быстро и просто, что особенно подходит для ситуаций, когда моделируемые отношения не очень сложны и объем данных невелик.
  + Есть интуитивное понимание и объяснение.
  + Линейная регрессия очень чувствительна к выбросам.

Мы можем смоделировать многомерную линейную регрессию следующим образом:



#### Полиномиальная регрессия

Когда мы хотим создать модель, подходящую для обработки нелинейных разделяемых данных, нам необходимо использовать полиномиальную регрессию. В этом методе регрессии наиболее подходящей линией является не прямая линия, а кривая, которая соответствует точкам данных. Для полиномиальной регрессии выражается следующей формулой:



У некоторых переменных есть индексы, у других - нет. Однако выбор точного индекса для каждой переменной, естественно, требует некоторых предварительных знаний о текущем наборе данных и окончательных результатах.

Несколько ключевых моментов полиномиальной регрессии:

* + Возможность моделирования нелинейных разделяемых данных, линейная регрессия не может этого сделать. Как правило, он более гибкий и может моделировать довольно сложные отношения.
  + Полностью контролировать моделирование переменных функций (для установки индекса переменной).
  + Требует тщательного проектирования. Некоторые предварительные знания данных необходимы, чтобы выбрать лучший индекс.
  + Если индекс выбран неправильно, его легко переобучить.

#### Метод Лассо

Регрессия лассо очень похожа на гребневую регрессию, потому что оба метода имеют одну и ту же предпосылку: они оба добавляют член смещения к функции оптимизации регрессии, чтобы уменьшить эффект коллинеарности, тем самым уменьшая дисперсию модели. Однако вместо использования квадрата отклонения, такого как регрессия гребня, регрессия Лассо использует отклонение абсолютного значения в качестве члена регуляризации:



Между регрессией гребня и регрессией Лассо есть некоторые различия, которые в основном можно отнести к разнице в природе регуляризации L2 и L1:

• Встроенный выбор функций:Это очень полезное свойство нормы L1, тогда как норма L2 не имеет этой характеристики. На самом деле это связано с тем, что норма L1 имеет тенденцию давать разреженные коэффициенты. Например, предположим, что модель имеет 100 коэффициентов, но только 10 из них являются ненулевыми коэффициентами. На самом деле это означает, что «остальные 90 переменных бесполезны для прогнозирования целевого значения». Норма L2 дает не разреженные коэффициенты, поэтому у нее нет этого свойства. Следовательно, можно сказать, что регрессия Лассо представляет собой форму «выбора параметра», и вес невыбранной переменной функции в целом равен 0.

• Редкость: Относится к очень небольшому количеству ненулевых элементов матрицы (или вектора). Норма L1 имеет свойство создавать множество коэффициентов с нулевыми значениями или очень маленькими значениями с несколькими большими коэффициентами.

• Вычислительная эффективность: Норма L1 не имеет аналитического решения, а норма L2 имеет. Это позволяет вычислить решение нормы L2. Однако решение нормы L1 является разреженным, что позволяет использовать его с разреженными алгоритмами, что делает его более эффективным с точки зрения вычислений.

#### Эластичная сеть

ElasticNet – это гибрид методов регрессии лассо и гребневой регрессии. Он использует регуляризацию L1 и L2, а также достигает эффекта двух технологий:



Практическое преимущество компромисса между лассо и регрессией гребня состоит в том, что он позволяет Elastic-Net унаследовать некоторую стабильность регрессии гребня в цикле.

Несколько ключевых моментов регрессии ElasticNet:

* + Он поощряет групповые эффекты в случае сильно коррелированных переменных, а не обнуляет некоторые из них, как это делает Лассо. Эластичные сети очень полезны, когда несколько функций связаны с другой функцией. Лассо обычно выбирает одно из них случайным образом, тогда как эластичные сети предпочитают выбирать два.
  + Нет ограничений на количество выбранных переменных.

#### Метод опорных векторов

В основе метода опорных векторов для задач регрессии или регрессии опорных векторов (SVR) лежит поиск гиперплоскости, при которой риск в многомерном пространстве будет минимальным. По сравнению с традиционной регрессионной моделью SVR оценивает коэффициенты путем минимизации квадратичных потерь. Так, если прогнозное значение попадает в область гиперплоскости, то потери равны нулю. В противном случае разности прогнозного и фактического значений.

Существует неявная взаимосвязь между переменными, в отличие от предыдущих моделей, где взаимосвязь была определена явно с помощью уравнения (коэффициентов достаточно, чтобы сбалансировать масштаб переменных). Поэтому здесь требуется масштабирование функций.

#### Дерево принятия решений

Регрессия дерева решений строит регрессионную модель в виде древовидной структуры. Поскольку набор данных разбивается на более мелкие подмножества, связанное с ним дерево решений строится постепенно. Для точки в тестовом наборе мы прогнозируем значение, используя построенное дерево решений.

#### Бэггинг

Бэггинг (метод случайных подвыборок) — это техника, которая используется для обучения нескольких классификаторов, но не на абсолютно одинаковых выборках. Если мы будете обучать классификаторы на одинаковых данных, они сами станут одинаковыми, но нам нужно разнообразие моделей.

#### Случайный лес

Регрессия случайного леса – в этом мы берем k точек данных из обучающего набора и строим дерево решений. Мы повторяем это для разных наборов k точек. Мы должны определить количество деревьев решений, которые будут построены вышеуказанным образом. Пусть число построенных деревьев равно n. Мы прогнозируем значение, используя все n деревьев, и берем их среднее значение, чтобы получить окончательное прогнозируемое значение для прогнозируемой точки.

Используя случайные наборы характеристик для каждого дерева в случайном лесу, мы декоррелируем деревья и дисперсия полученной модели уменьшается. Эта декорреляция – главное преимущество в использовании случайных лесов в сравнении с деревьями решений

### Разведочный анализ данных

#### Цифровые характеристики датасета

В таблице 1 представлены статистические данные атрибутов учебного датасета: минимальное и максимальное значения, среднее медианное значение, распределение по квартилям.

Таблица Таблица с исходным датасетом



**Таблица** **2 Основные статистические данные атрибутов датасета**



Таблица 3 Перекрестная таблица коэффициентов корреляции атрибутов датасета

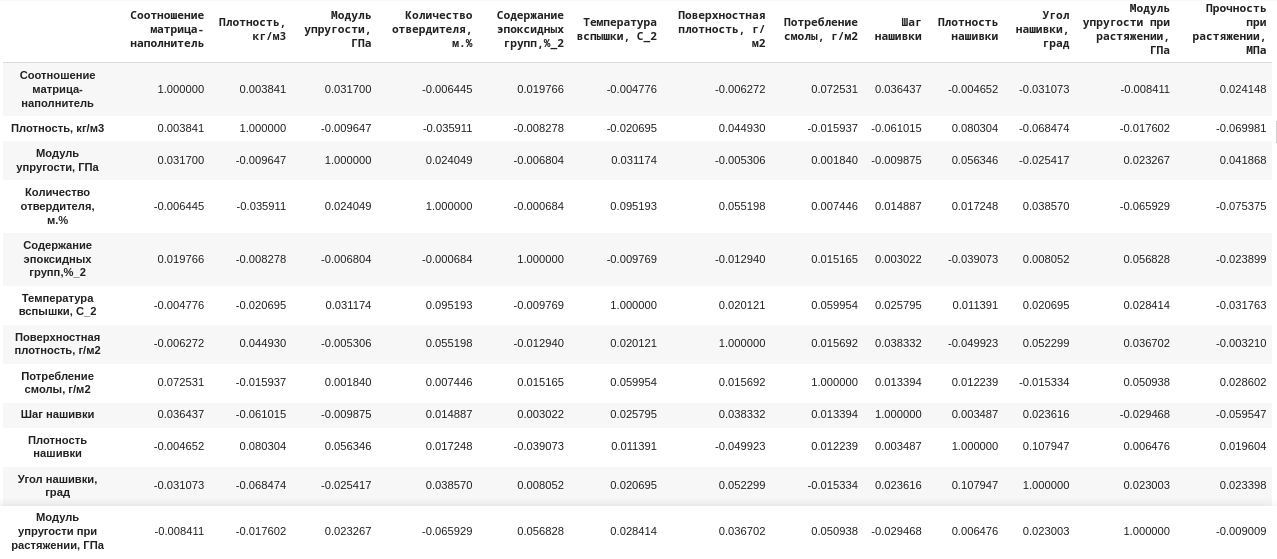
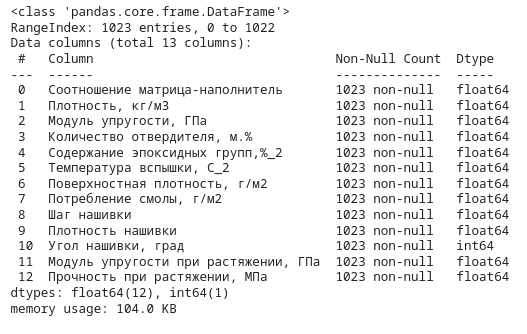


Таблица 4 Характеристики данных по принадлежности к атрибутам датасета



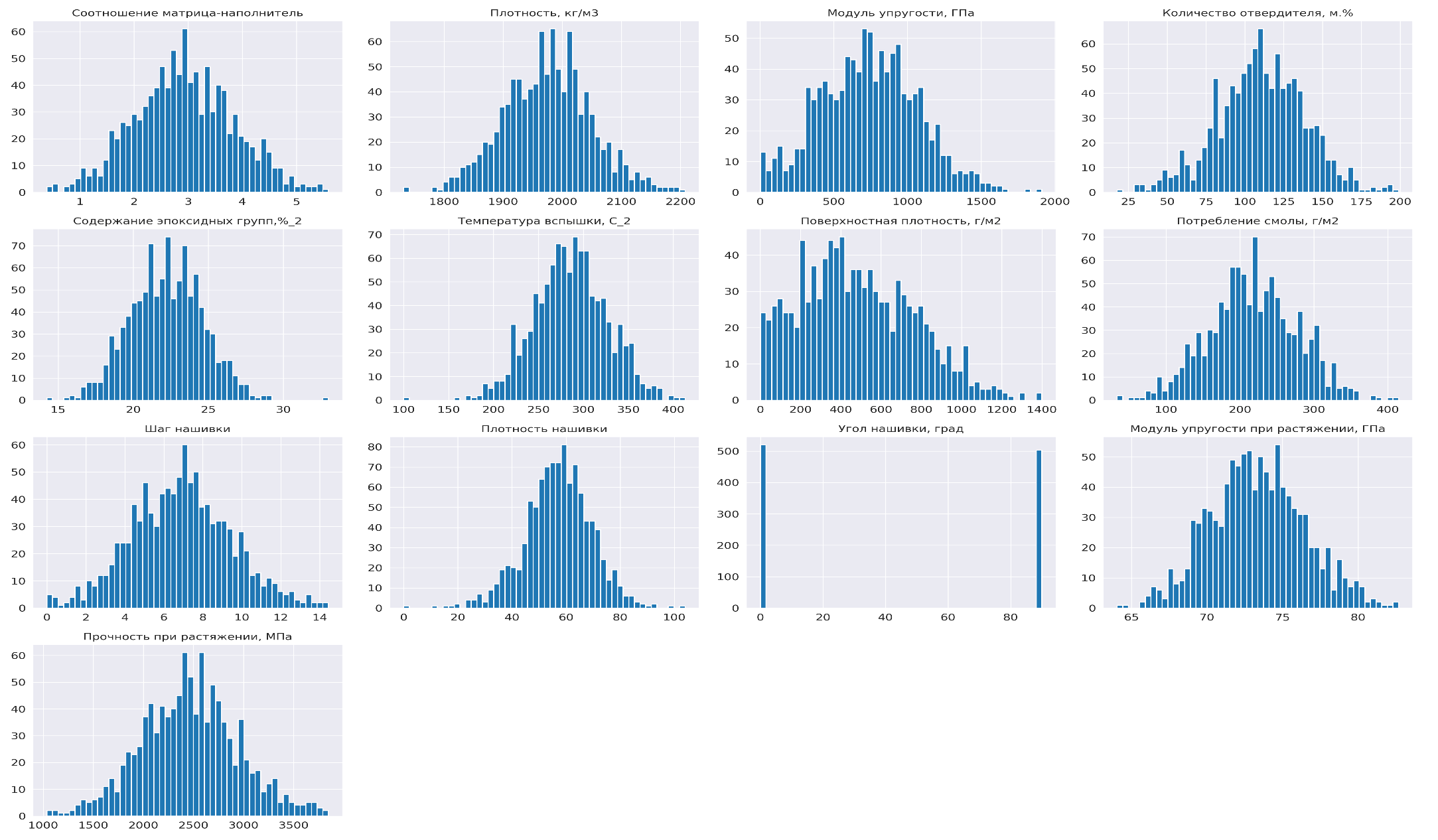


Рисунок 1 Гистограммы распределения по атрибутам датасета

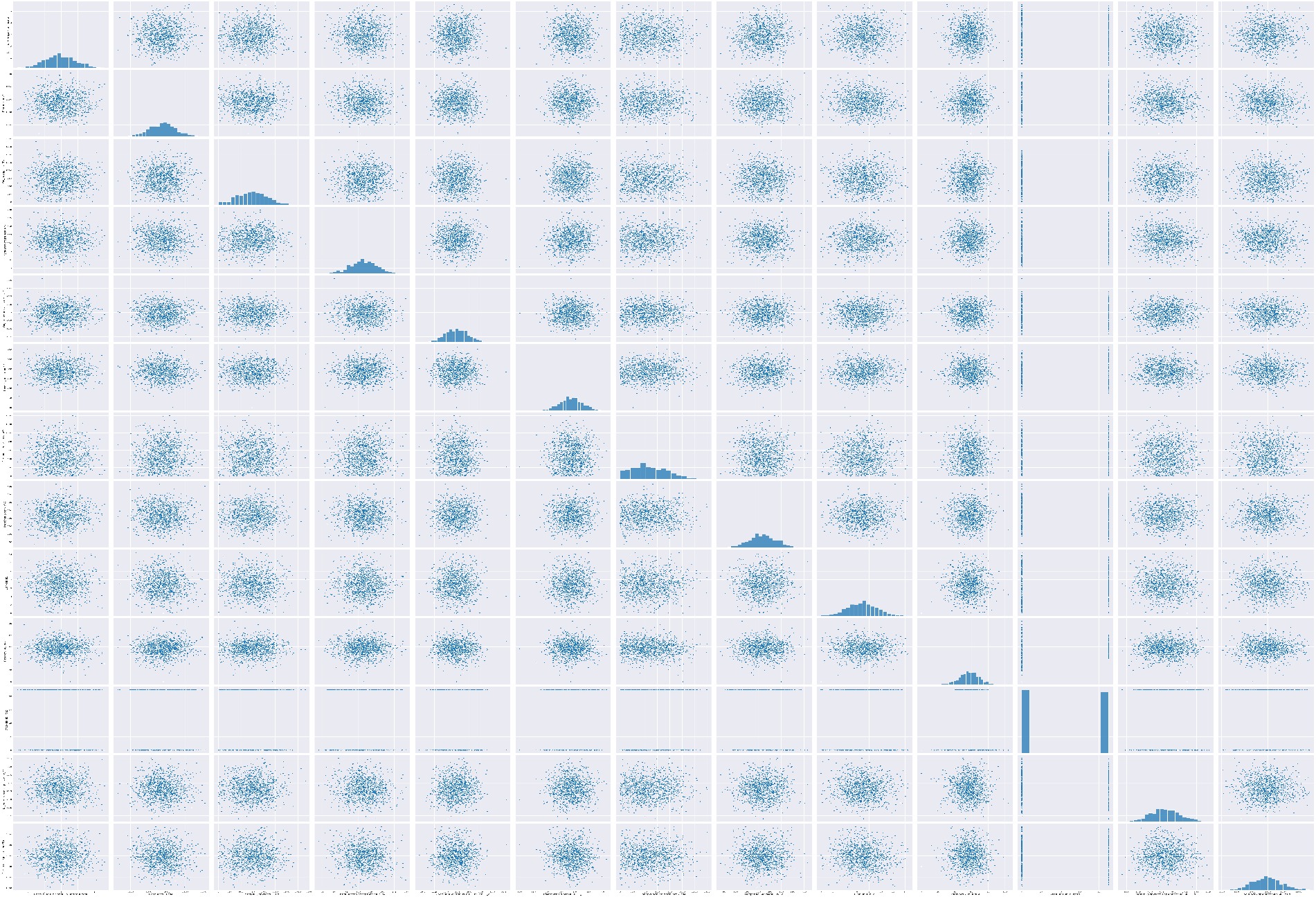


Рисунок 2 Попарные диаграммы рассеивания сведенные в перекрестную таблицу

Рисунок 3 Трехмерная диаграмма распределения

В перекрестной таблице 2 представлены коэффициенты корреляции между атрибутами датасета.

В таблице 3 представлены характеристики данных по принадлежности к атрибутам датасета.

#### Графические характеристики датасета

На рисунках представлены графическая интерпретация данных из таблиц характеристик датасета.

На рисунке 1 представлены гистограммы распределения по атрибутам датасета.

На рисунке 2 представлены попарные диаграммы рассеивания, сведенные в перекрестную таблицу.

На рисунке  3 изображена трехмерная диаграмма рассеивания, которая добавлена в Jupyter notebook работы для укрупнения изображения и позволяет, меняя атрибуты для ее построения рассмотреть заинтересовавшие зависимости.

#### Краткие выводы разведочного анализа.

Из табличных данных и рисунков можно сделать вывод, что линейная зависимость и другие паттерны между атрибутами не просматриваются.

## Практическая часть

### Предобработка данных

Во время предобработки данных, с ними были проведены следующие манипуляции:

1. Значения атрибута «Угол нашивки, град» из значений 0 и 90 были приведены к значениям 0 и 1;
2. Датасет был нормализован по алгоритму MinMaxScaler, с целью приведения значений атрибутов к диапазону от 0 до 1;
3. Датасет был нормализован по алгоритму StandardScaler, с целью приведения среднемедианного значения атрибутов к 0.

Табличные и графические данные после нормализаций представлены далее в таблицах и на рисунках.

Таблица Датасет после нормализации по минимаксу



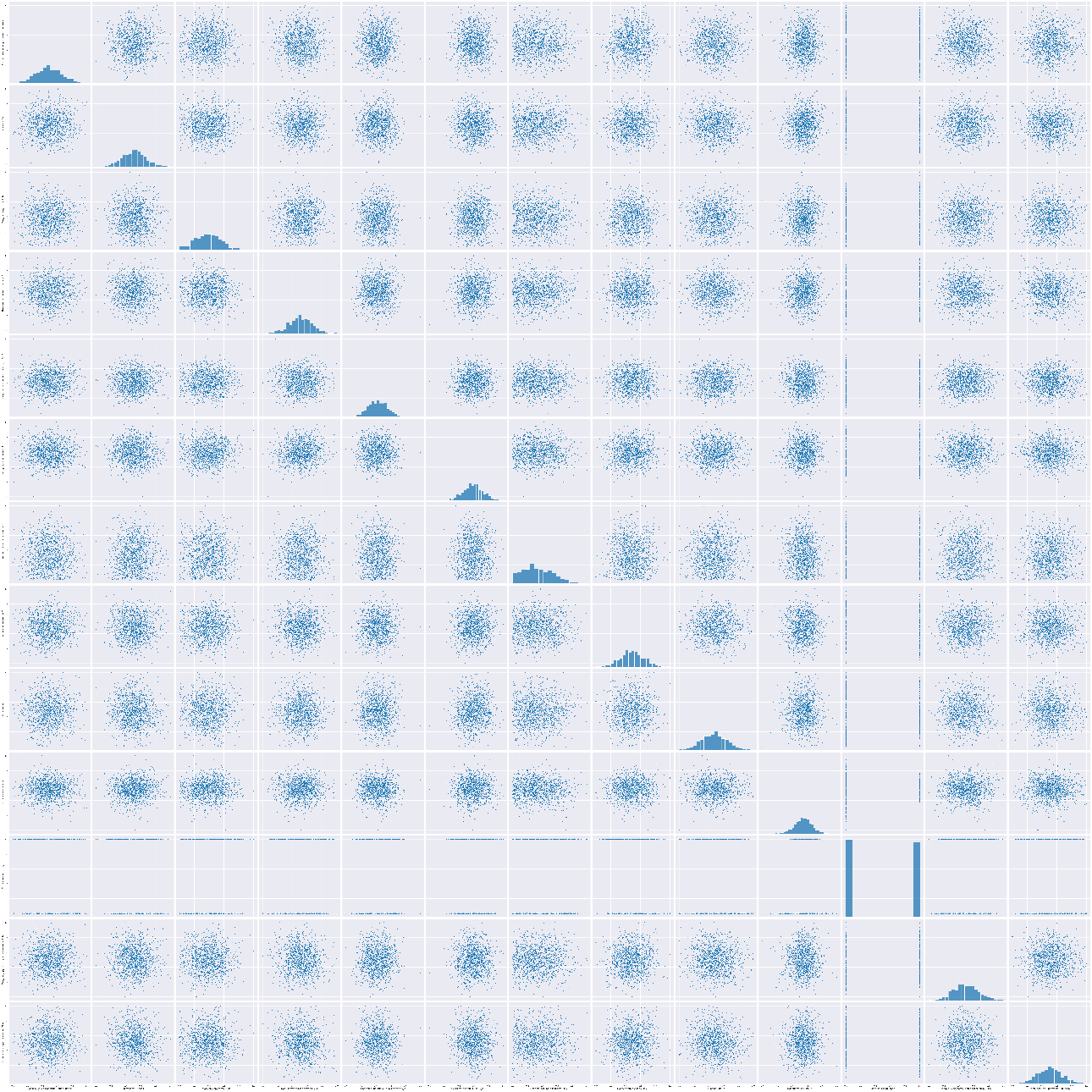


Рисунок Диаграммы рассеивания сведенные в перекрестную таблицу после нормализации по минимаксу

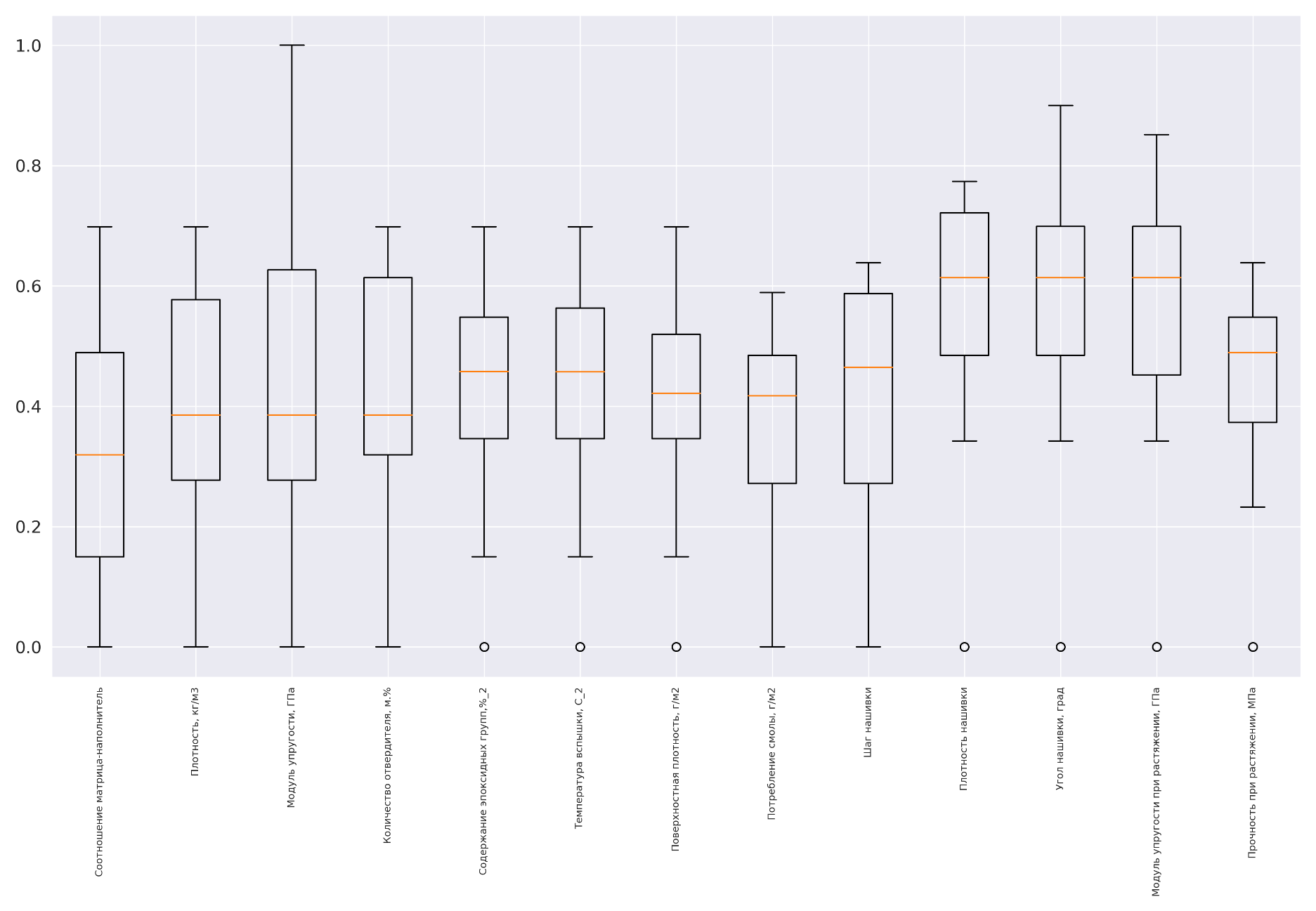


Рисунок Диаграмма ящиков с усами после нормализации по минимаксу

Таблица 8 Датасет после нормализации по алгоритму StandardScaler



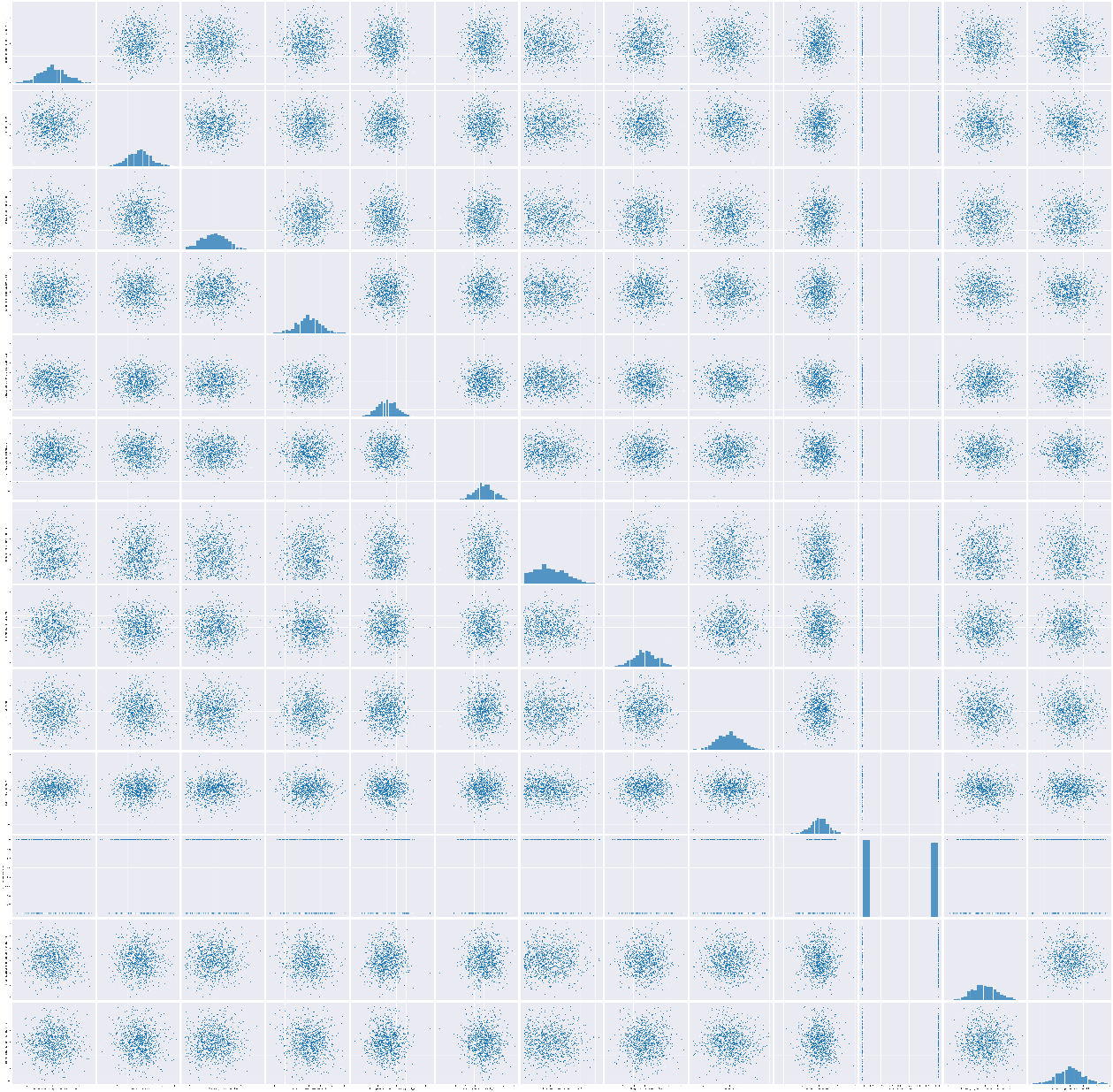


Рисунок 9 Диаграммы рассеивания, сведенные в перекрестную таблицу после нормализации по алгоритму StandardScaler

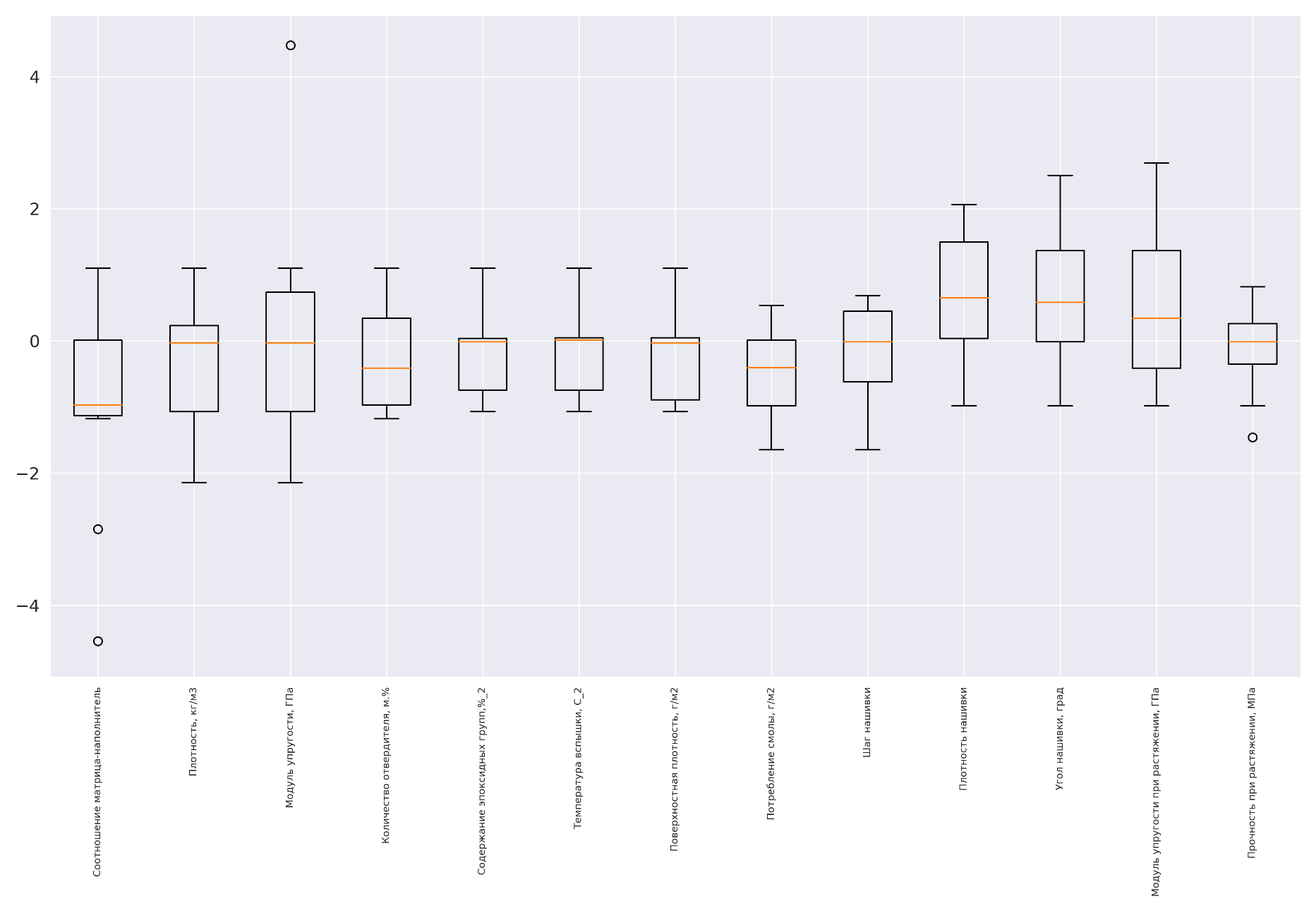


Рисунок 10 Диаграмма ящиков с усами после нормализации по алгоритму StandardScaler

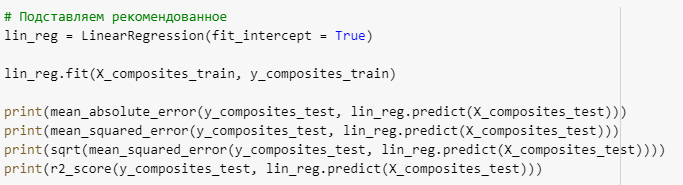
### Разработка и обучение моделей

Используемые для моделей обучения методы регрессии были описаны в подразделе 1.2. А именно:

* + Линейная регрессия;
  + Полиномиальная регрессия;
  + Метод Лассо;
  + Эластичная сеть;
  + Метод опорных векторов;
  + Дерево принятия решений;
  + Бэггинг;
  + Случайный лес.

Для выбора значений глобальных параметров этих моделей использовался поиск по сетке с разбивкой тренировочной выборки на десять подвыборок.

Все модели строились по следующему шаблону кода с использованием библиотеки Scitit-Learn:



с некоторыми изменениями для отдельных моделей в зависимости от реализации их в библиотеке.

После чего создавалась модель с рекомендованными значениями глобальных параметров для обучения модели. Для проверки правильности рекомендованных значений, модели также обучались с другими, отличными от рекомендованных, параметрами.

### Сравнение результатов

Ошибки между прогнозируемыми и заданными значениями для обученных моделей, рассчитанных по различным методикам, были сведены в таблице 4.

Таблица 7 Ошибки для обученных моделей

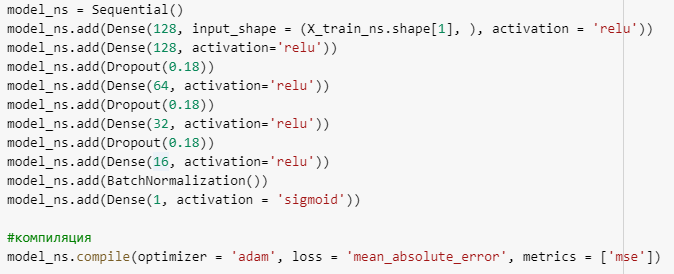
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Метод регрессии** | **Ошибки** | | | |
| **MAE** | **MSE** | **RMSE** | **r2** |
| Линейная регрессия | 0.13610 | 0.02867 | 0.16935 | -0.00698 |
| Полиномиальная регрессия | 0.13610 | 0.02867 | 0.16935 | -0.00698 |
| Метод Лассо | 0.13601 | 0.02871 | 0.16946 | -0.00835 |
| Эластичная сеть | 0.13601 | 0.02871 | 0.16946 | -0.00835 |
| Метод опорных векторов | 0.13551 | 0.02854 | 0.16895 | -0.00229 |
| Дерево принятия решений | 0.13841 | 0.02994 | 0.17303 | -0.05122 |
| Бэггинг | 0.13781 | 0.02933 | 0.17126 | -0.02976 |
| Случайный лес | 0.14047 | 0.03113 | 0.17642 | -0.09275 |

### Нейронная сеть, которая будет рекомендовать соотношение матрица-наполнитель

Для выполнения задания была выбрана нейронная сеть прямого распространения (feed forward neural networks, FF или FFNN) и перцептроны (perceptrons, P), они передают информацию от входа к выходу. FFNN обучается по методу обратного распространения ошибки, в котором сеть получает множества входных и выходных данных.

Принятая архитектура нейронной сети включает в себя входной и выходной слои и 8 скрытых слоев. Из последних 4 слоя Dense, 3 слоя Dropout между ними предпоследнего слоя BatchNormalization. Слои Dropout и BatchNormalization включены в состав сети для предотвращения переобучения (overfiting). Код из блокнота Jupyter notebook проекта представлен на рисунке 4

Рисунок 4 Архитектура используемой нейронной сети



Обучение модели проходит за 100 эпох.

Значение корня из среднеквадратичной ошибки после обучения составляет 0.20136.

### Создание удаленного репозитория и загрузка на него результатов работы

Файлы приложение были размещены на GitHub по адресу:

<https://github.com/SonyaMeloman/course-work/tree/main>

# Заключение

В результате проведенных работ можно сделать следующие выводы:

* + Корреляция между входными и целевыми признаками практически отсутствует,
  + Нейросеть никак не поможет в решении этого вопроса

- ???

- Profit!