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»

на тему:

Прогнозирование конечных свойств новых материалов

(композиционных материалов)

Слушатель Никитин Иван Сергеевич (ФИО)

Москва, 2023

**Содержание:**

**Введение 3**

**1. Аналитическая часть 4**

**1.1 Постановка задачи 4**

**1.2. Описание используемых методов 6**

**1.3. Разведочный анализ данных 7**

**2. Практическая часть 8**

**2.1. Предобработка данных 8**

**2.2 Разработка и обучение модели 19**

**2.3. Тестирование модели 20**

**2.4. Нейронная сеть для рекомендации соотношения матрица-наполнитель 21**

**2.5 Создание Web приложения 24**

**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.

**Аналитическая часть**

**1.1. Постановка задачи**

Даже если мы знаем характеристики исходных компонентов композитного материала, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Для проведения физических испытаний образцов тратится много денежных средств и времени. Можно попробовать решить задачу определения конечных свойств композитов путем прогнозирования этих свойств с помощью обученных моделей или нейронных сетей. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

Для решения задачи прогнозирования конечных свойств новых материалов даны два файла в формате в формате exel-таблиц. В файлах представлены свойства композитов (Таблица 1, файл X\_bp.xlsx, Таблица 2, файл X\_nup.xlsx).

Таблица1

|  | Соотношение матрица-наполнитель | Плотность, кг/м3 | модуль упругости, ГПа | Количество отвердителя, м.% | Содержание эпоксидных групп,%\_2 | Температура вспышки, С\_2 | Поверхностная плотность, г/м2 | Модуль упругости при растяжении, ГПа | Прочность при растяжении, МПа | Потребление смолы, г/м2 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1,857142857 | 2030 | 738,7368421 | 30 | 22,26785714 | 100 | 210 | 70 | 3000 | 220 |
| 1 | 1,857142857 | 2030 | 738,7368421 | 50 | 23,75 | 284,6153846 | 210 | 70 | 3000 | 220 |
| … | … | … | … | … | … | … | … | … | … | … |

В файле X\_bp.xlsx 1023 элемента, в файле X\_nup.xlsx 1040 элементов. Таблицы были объединены в одну методом INNER по индексу элемента. После объединения таблиц, количество элементов в рассматриваемом датасете составило 1023, а количество параметров равно 13. Далее было проверено наличие пропусков и дубликатов в рассматриваемом датасете (Рисунок 1). Дубликатов и пропусков не обнаружено.

Таблица 2.

|  | Угол нашивки, град | Шаг нашивки | Плотность нашивки |
| --- | --- | --- | --- |
| 0 | 0 | 4 | 57 |
| 1 | 0 | 4 | 60 |
| … | … | … | … |

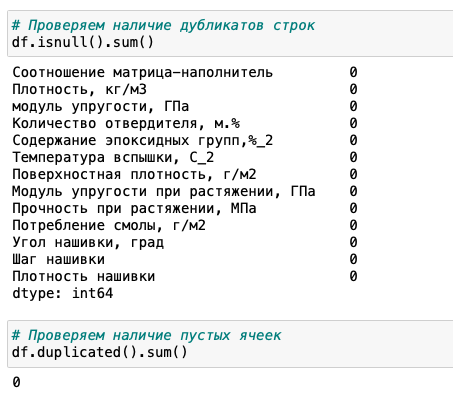


Рисунок 1 - Поиск пропусков и дубликатов в данных

Данные в датасете принадлежат к числовому типу данных, в датасете есть параметр “Угол нашивки”, содержащий 2 значения.

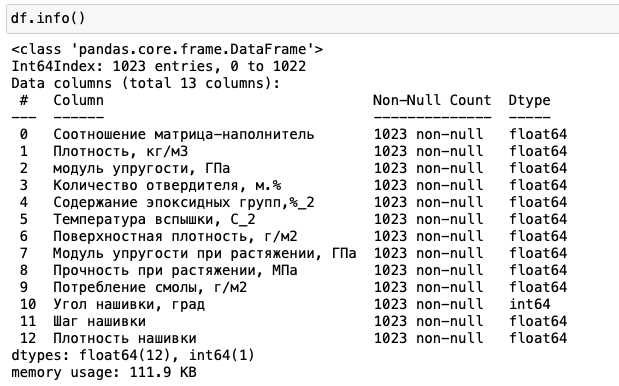


Рисунок 2 - Типы данных в датасете.

**1.2. Описание используемых методов**

1. **Линейная регрессия** Линейная регрессия (англ. Linear regression) — используемая в статистике регрессионная модель зависимости одной (объясняемой, зависимой) переменной y от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) x с линейной функцией зависимости.

2. **Случайный лес**

Random forest (с англ. — «случайный лес») — алгоритм машинного обучения, заключающийся в использовании комитета (ансамбля) решающих деревьев. Алгоритм сочетает в себе две основные идеи: метод бэггинга, и метод случайных подпространств. Алгоритм применяется для задач классификации, регрессии и кластеризации. Основная идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим.

3. **Метод k-ближайших соседей**

Метод k-ближайших соседей (англ. *k-nearest neighbors algorithm*, k- NN) - метрический алгоритм для автоматической классификации объектов или регрессии. В случае использования метода для классификации объект присваивается тому классу, который является наиболее распространённым среди соседей данного элемента, классы которых уже известны. В случае использования метода для регрессии, объекту присваивается среднее значение по ближайшим к нему объектам, значения которых уже известны.

4. **Многослойный перцептрон**

Многослойный персептрон — это класс искусственных нейронных сетей прямого распространения, состоящих как минимум из трех слоёв: входного, скрытого и выходного. За исключением входных, все нейроны использует нелинейную функцию активации. При обучении MLP используется обучение с учителем и алгоритм обратного распространения ошибки.

**1.3. Разведочный анализ данных**

**Разведочный анализ данных** (англ. *Exploratory data analysis, EDA*) — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов визуализации.

В проекте были использованы следующие методы разведочного анализа данных:

1. Визуальный анализ гистограмм
2. Визуальный анализ диаграмм размаха («ящик с усами»)
3. Проверка нормальности распределения по критерию Пирсона
4. Анализ попарных графиков рассеяния переменных
5. Корреляционный анализ c целью поиска коэффициентов

**2. Практическая часть**

**2.1. Предобработка данных**

Для каждого параметра построены гистограммы распределения.

Рисунок 3. Гистограмма: Соотношение матрица-наполнитель.

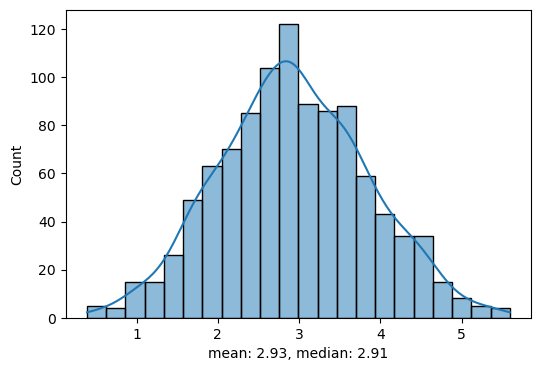
****

Рисунок 4. Гистограмма: Плотность, кг/м3.

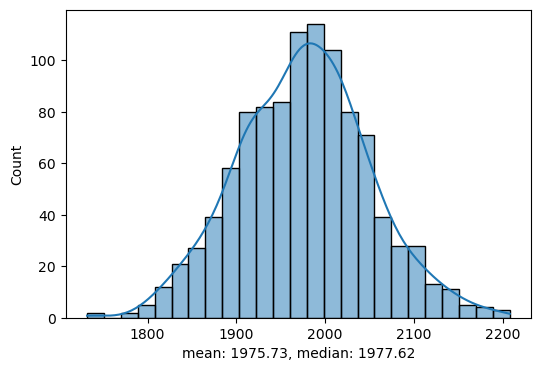


Рисунок 5. Гистограмма: Модуль упругости, Гпа.

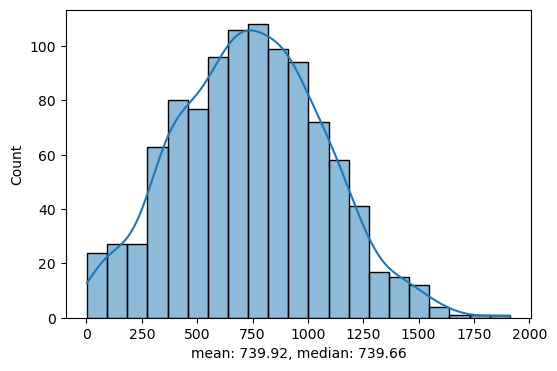


Рисунок 6. Гистограмма: Количество отвердителя, м.%.

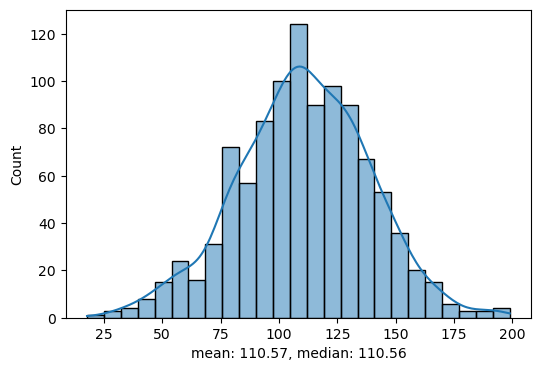


Рисунок 7. Гистограмма: Содержание эпоксидных групп, %\_2.

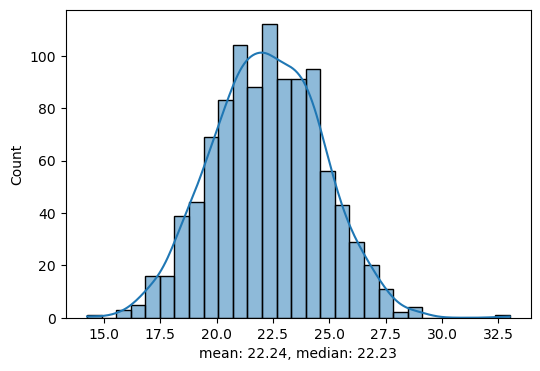


Рисунок 8. Гистограмма: Температура вспышки С\_2.

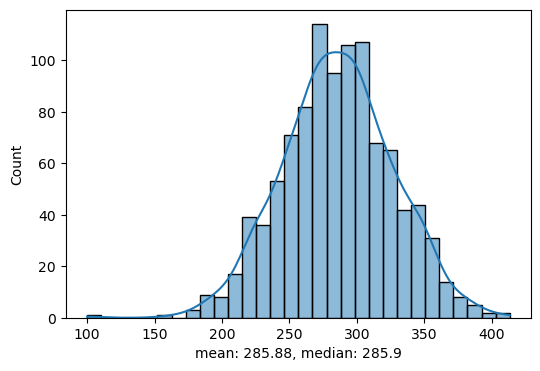


Рисунок 9. Гистограмма: Поверхностная плотность, г/м2.

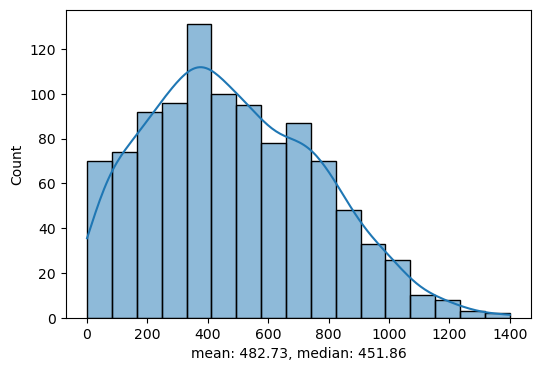


Рисунок 10. Гистограмма: Модуль упругости при растяжении, ГПа.

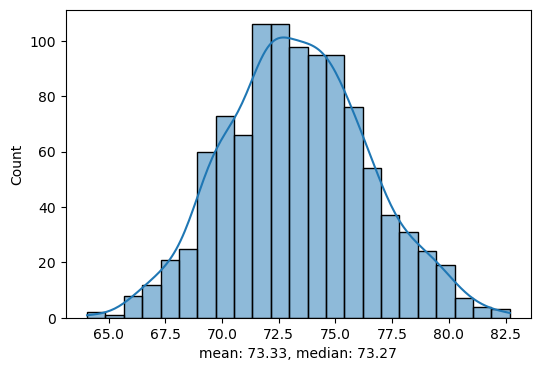


Рисунок 11. Гистограмма: Прочность при растяжении, МПа.

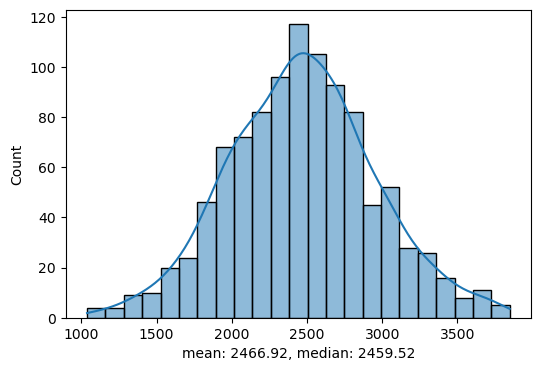


Рисунок 12. Гистограмма: Потребление смолы, г/м2.

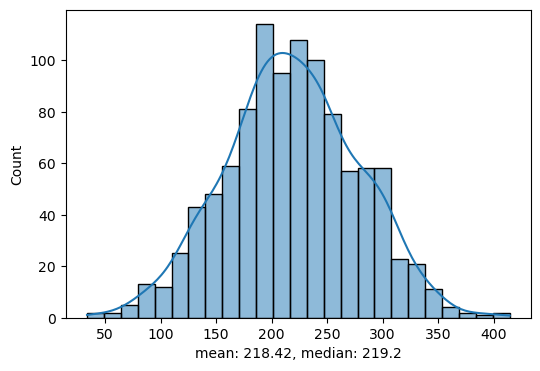


Рисунок 13. Гистограмма: Угол нашивки.

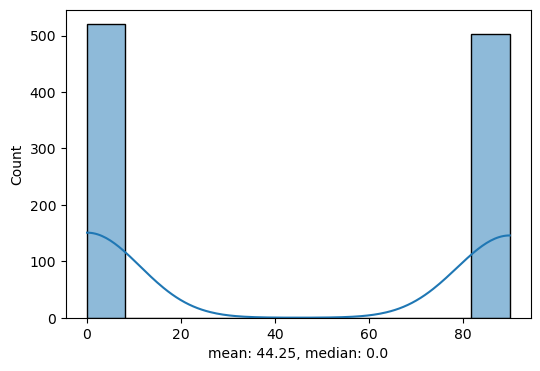


Рисунок 14. Гистограмма: Шаг нашивки.

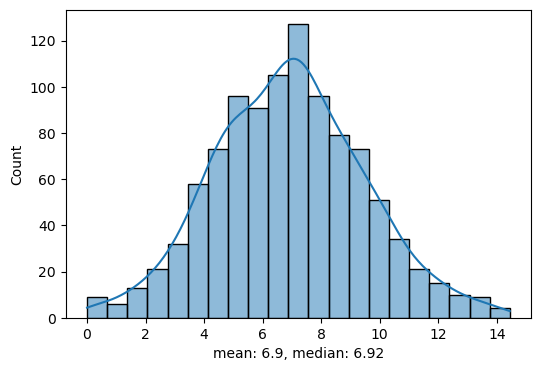
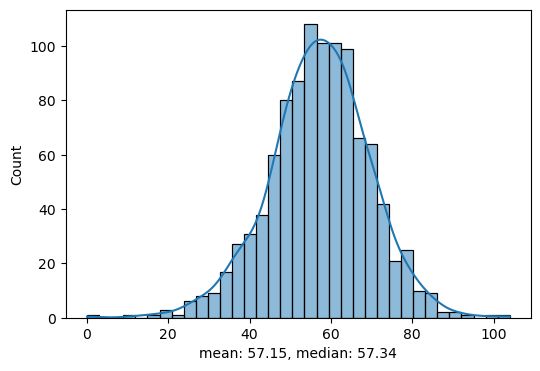
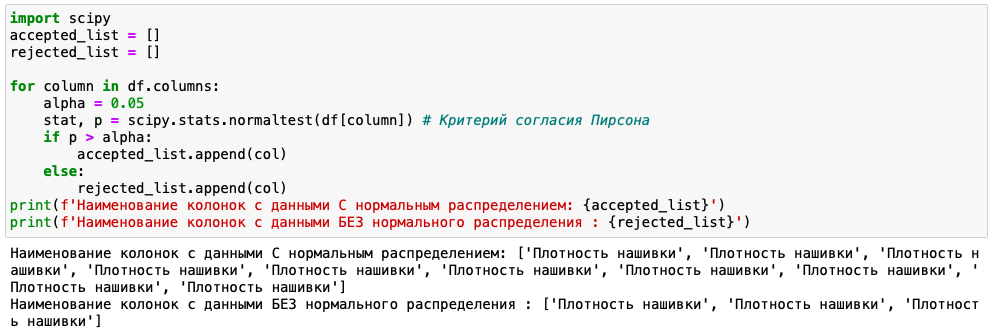


Рисунок 15. Гистограмма: Плотность нашивки.



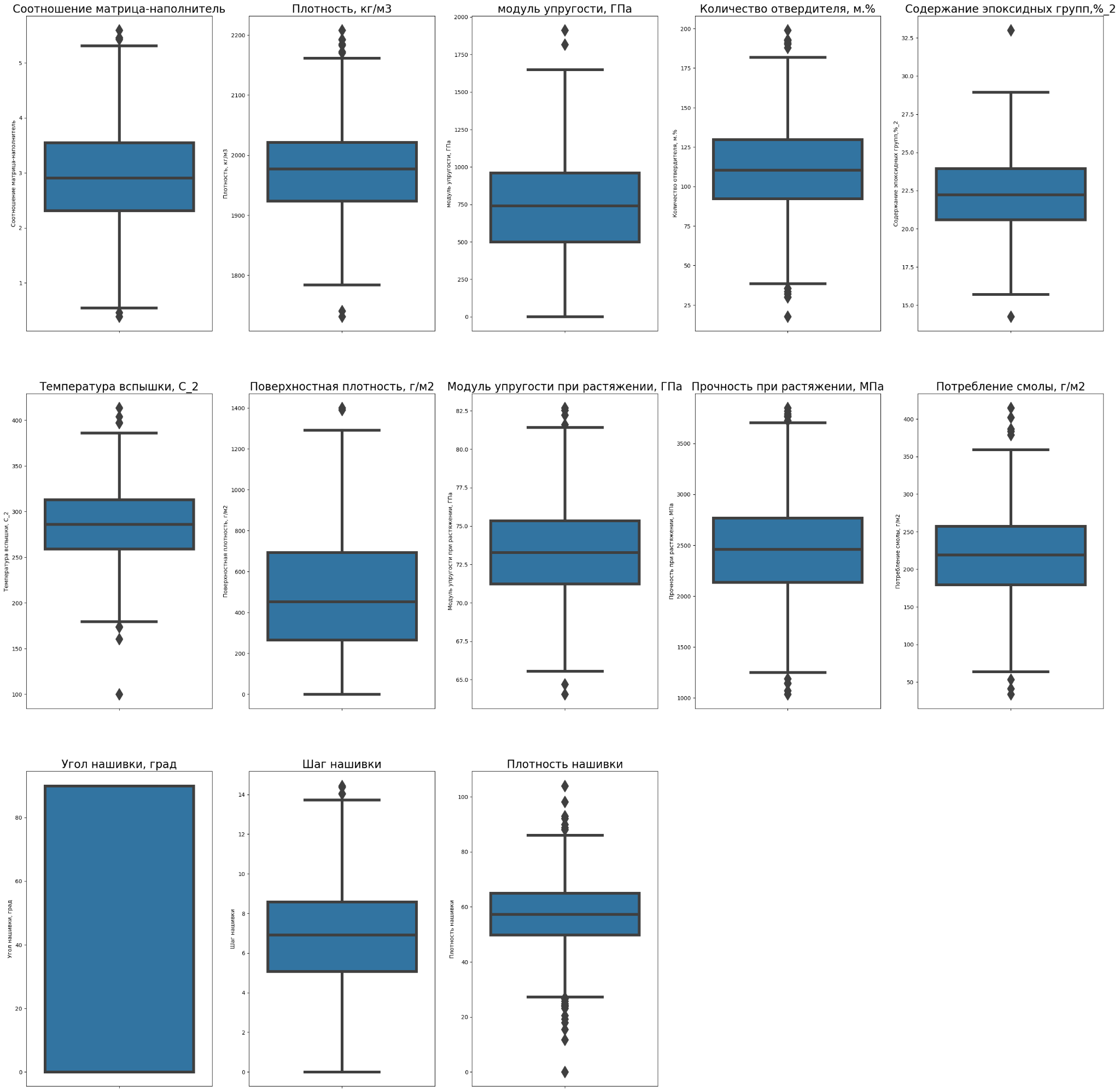
На гистограммах можно увидеть, что распределения параметров являются нормальными или близкими к нормальному.

Распределения переменных были проверены на нормальность по критерию Пирсона (Рисунок 16).

Рисунок 16. Проверка распределения по критерию Пирсона.

Построены диаграммы размаха «ящик с усами» для определения выбросов (Рисунок 17).

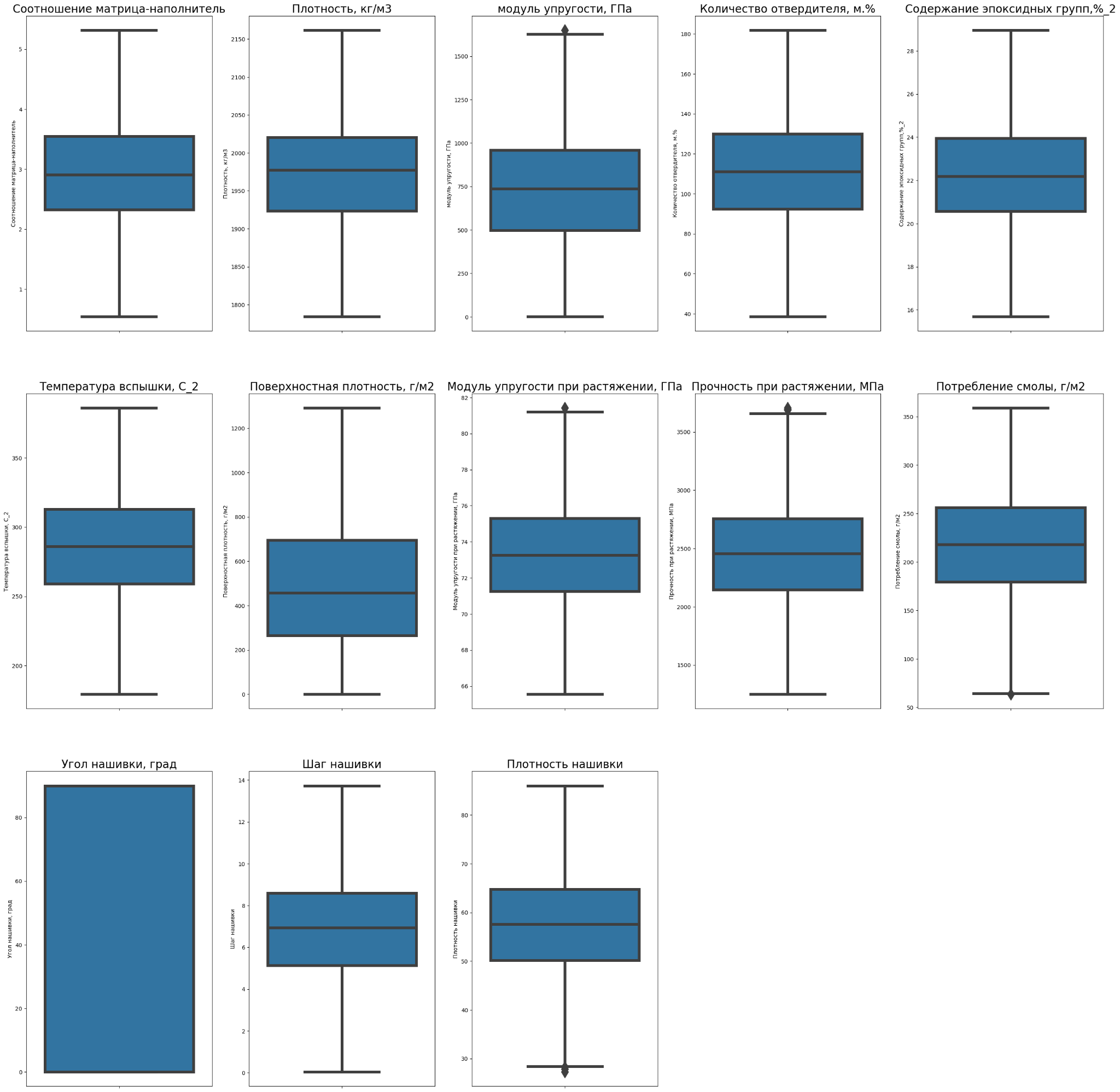
Рисунок 17. Диаграммы размаха переменных до удаления выбросов.



На диаграммах видно наличие выбросов. Выбросами являются точки, превышающие 1,5 межквартильного расстояния. Межквартильное расстояние — это разница между 1-м и 3-м квартилями, т.е. между 25-м и 75-м процентилями.

Данные, выходящие за пределы 1,5 межквартильных расстояния, были заменены на пустые значения и посчитаны. Их кол-во оказалось небольшим, было принято решение удалить строки, содержащие выбросы. (Рисунок 18)

Рисунок 18. Диаграммы размаха переменных после удаления выбросов.



После очистки датасета от выбросов была проведена нормализация данных функцией MinMaxScaler из библиотеки sklearn. Данные были приведены к общей шкале, в которой минимальное значение параметра принимало 0, а максимальное 1. (Рисунок 19)

Рисунок 19. Описательная статистика нормализованных данных.

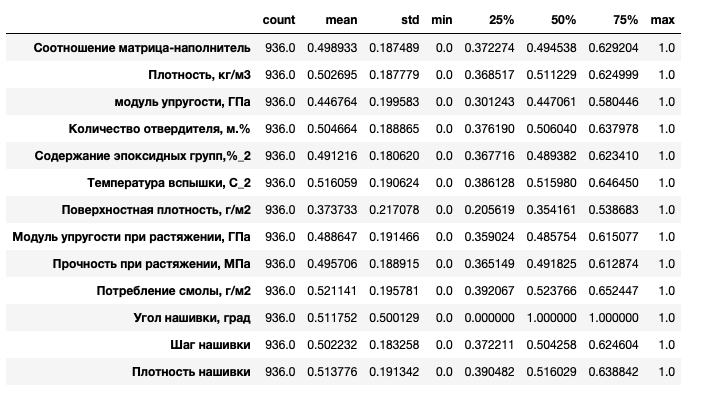


График попарной зависимости переменных (Рисунок 20) был построен для всех параметров. На графике синим выделены точки для угла нашивки равного 0 градусов, а оранжевым для угла нашивки равного 90 градусам. Согласно графику ни для одной пары переменных не прослеживается какая-либо зависимость. Это подтверждается матрицей корреляции (Рисунок 21), в которой максимальное значение корреляции равно 0.093 между параметрами Плотность нашивки и угол нашивки. Корреляция близка к 0, что говорит о том, что переменные не связаны между собой.

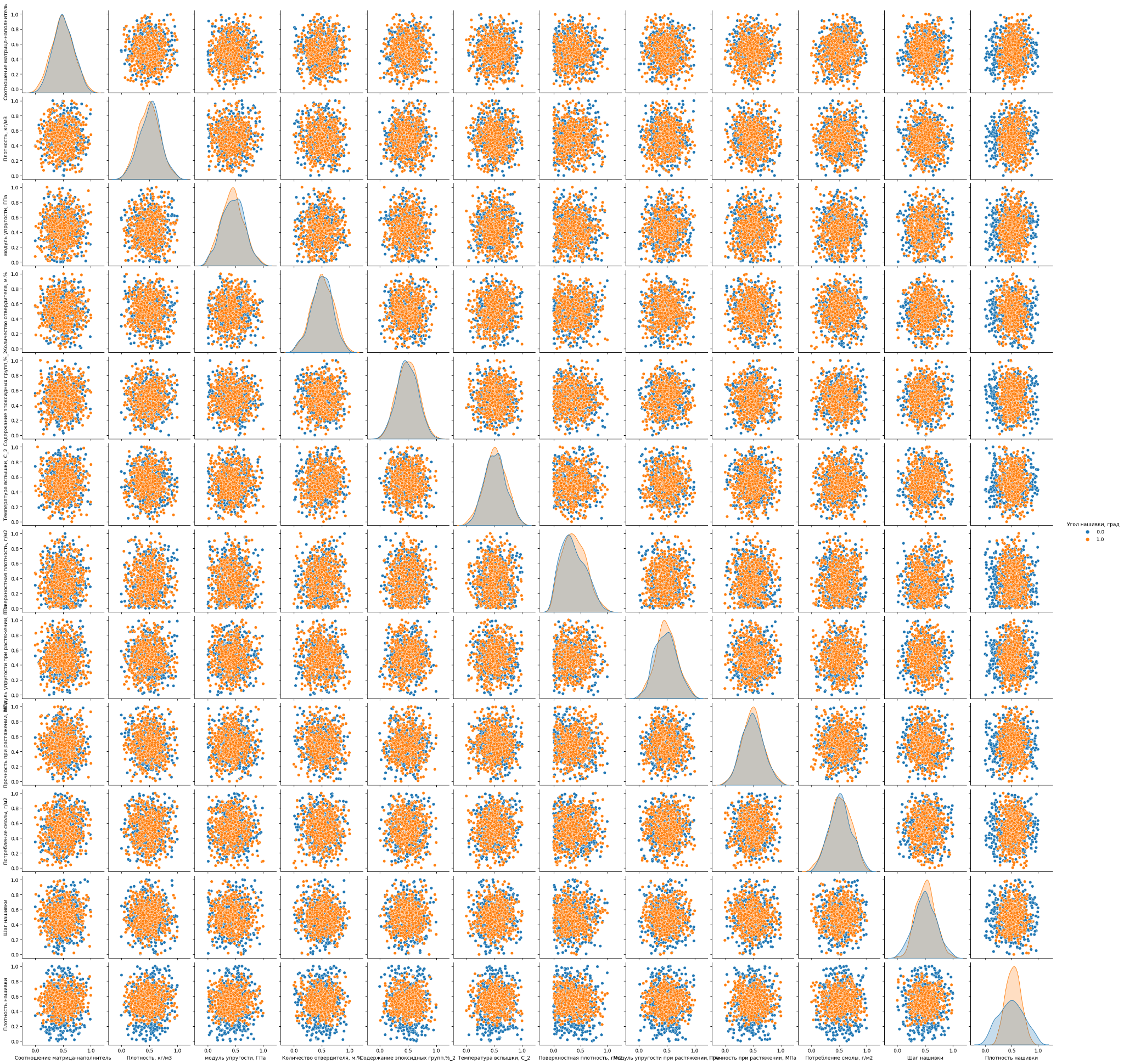
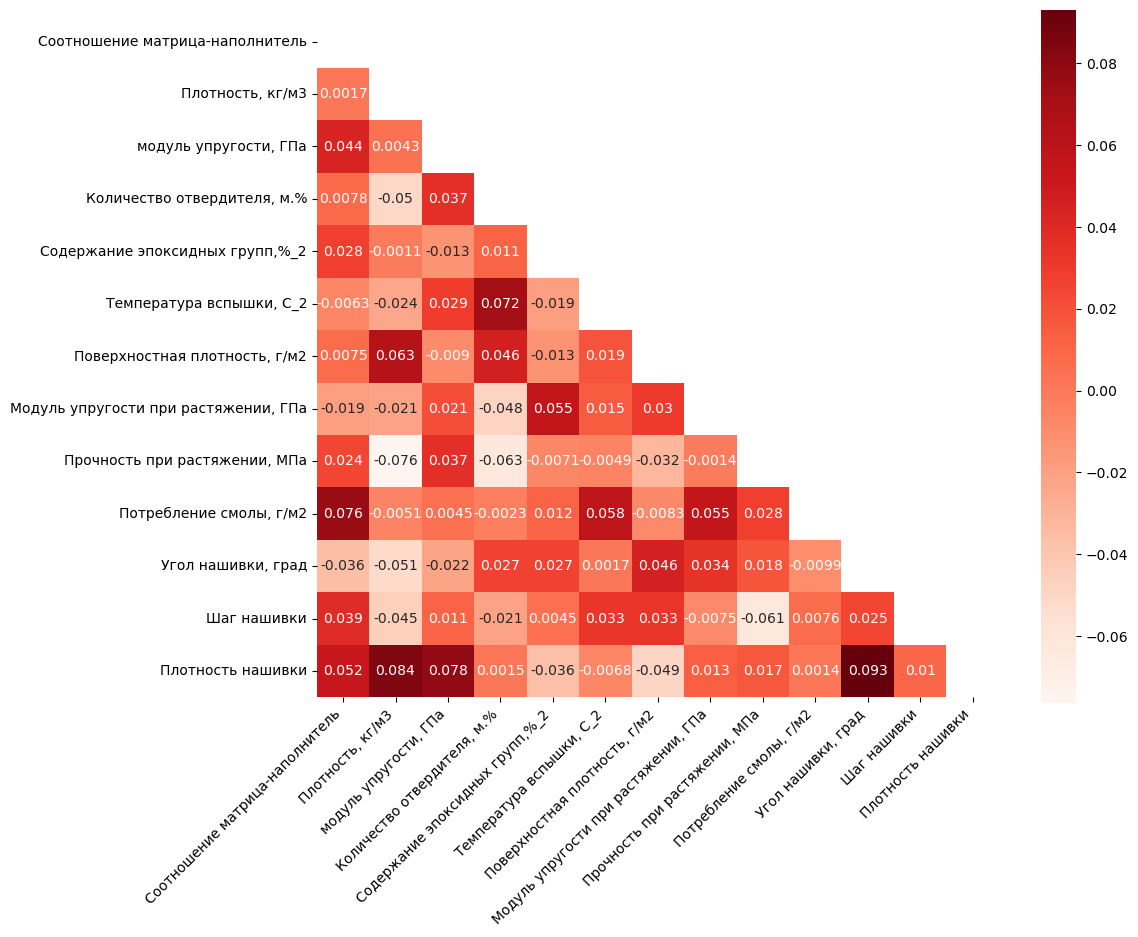
Рисунок 20. График попарной зависимости переменных.

Рисунок 21. Матрица корреляции данных.

**2.2 Разработка и обучение модели**

Согласно поставленной задаче, данные были разделены на обучающую и тестовую выборки в соотношении 70/30. Для каждой модели был создан словарь с гиперпараметрами. С помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10, были найдены лучшие гиперпараметры для каждой модели. Каждая модель с лучшими гиперпараметрами давала свои прогнозы на тестовой выборке. В качестве параметров оценки моделей были выбраны: Средняя абсолютная ошибка (МАЕ) и коэффициент детерминации (R2).

В ходе выполнения проекта в качестве моделей для прогноза модуля упругости при растяжении и прочности при растяжении были использованы:

• Метод k-ближайших соседей

• Стохастический градиентный спуск

• Линейная регрессия

• Случайный лес

**2.3. Тестирование модели**

Создается объект модели. Далее создается сетка гиперпараметров, в данном случае словарь с названием гиперпараметра и его возможными значениями. Следующим шагом объявленная модель и сетка гиперпараметров подаются в поиск по сетке и объявляется кол-во блоков равное 10, т.е. тренировочные данные делятся на 10 равных частей. Далее, для каждого возможного сочетания значений гиперпараметров, метод считает ошибку и в конце выбирает сочетание, при котором ошибка минимальна. С помощью лучшей модели прогнозируются значения на тестовой выборке, результаты средней абсолютной ошибки и коэффициента детерминации заносятся в сводную таблицу (Рисунок 22).

Для прогноза модуля упругости при растяжении лучше всего показала результаты модель стахастического градиентного спуска со средней абсолютной ошибкой = 2.53 Гпа, как и для прогноза прочности при растяжении со средней абсолютной ошибкой = 365.86 Мпа.

При этом, во всех использованных моделях коэффициент детерминации очень близок к 0, это говорит о том, что результат использования моделей не точнее использования для прогноза среднего значения прогнозируемого параметра.

Рисунок 22. Сводная таблица результатов работы моделей.

**2.4. Нейронная сеть для рекомендации соотношения матрица-наполнитель**

Нейронная сеть была написана с помощью библиотеки Keras на языке программирования Python.

Характеристики нейронной сети для прогноза соотношения матрица- наполнитель:

* Последовательная модель (Sequential) нейронной сети
* Модель состоит из 3 скрытых Dense слоев, количество нейронов в которых равно 128, 64, 1 и выходного слоя с одним нейроном. Функция активации слоев – selu (Масштабная экспоненциальная линейная единица).
* Используются слои Batch-Normalization (Пакетная нормализация (англ. batch-normalization) — метод, который позволяет повысить производительность и стабилизировать работу искусственных нейронных сетей. Суть данного метода заключается в том, что некоторым слоям нейронной сети на вход подаются данные, предварительно обработанные и имеющие нулевое математическое ожидание и единичную дисперсию.)
* В качестве оптимизатора нейронной сети используется SGD (стохастический градиентный спуск) с импульсом ускорения = 0.5.
* Размер батча при обучении – 64
* Кол-во эпох обучения – 100
* Размер валидационной выборки – 20% от размера обучающей выборки

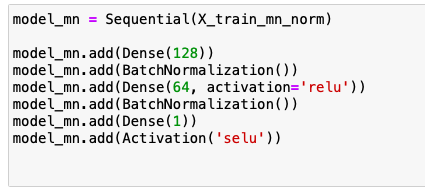
Рисунок 23. Структура нейронной сети.

Рисунок 24. График потерь нейронной сети на валидационной и тестовой выборке.



Рисунок 25. График прогнозных значений нейронной сети.



На графике прогнозных значений нейронной сети (Рисунок 25) можно увидеть, что построенная модель нейронной сети не так хорошо прогнозирует значения на тестовой выборке данных, хотя местами правильно предсказывает направление тренда и величину прогнозного значения.

Рисунок 26. График рассеивания тестовых и прогнозных значений.



На графике рассеяния тестовых и прогнозных значений (Рисунок 26) можно увидеть, что нейронная сеть показывает не самый лучший результат прогноза. В идеальном варианте точки тестовых и прогнозных значений должны выстроиться в прямую линию под углом 45°.

Результат средней абсолютной ошибки нейронной сети на тестовой выборке равен 0.8698.

Результат средней абсолютной ошибки при прогнозе соотношения матрица-наполнитель средним значением этой переменной равен 0.74.

**2.5 Создание Web приложения**