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 Махров Андрей Сергеевич

Москва, 2022

**Содержание**

Введение………………………………………………………….……………3

1. Аналитическая часть…………………………………………….……………4
   1. Постановка задачи……………………………………………….……………4
   2. Описание используемых методов……………………………….…………...6
   3. Разведочный анализ данных…………………………………...……………10
2. Практическая часть………………………………………………….……….14
   1. Предобработка данных…………………………………………….…...…....14
   2. Разработка и обучение модели………………………………….………......16
   3. Тестирование модели………………………………………………….….....17
   4. Написание модели нейроной сети ………………………………………….23
   5. Разработка приложения……………………………………….……………..27
   6. Создание репозитория…………………………………………………….….27

Заключение…………………………………………………………………....28

Библиографический список………………………………………………….29

**Введение**

Тема выпускной квалификационной работы по курсу «Data Science»: **«**Прогнозирование конечных свойств новых материалов (композиционных материалов)».

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.

Цель данной работы прогнозирование конечных свойств новых композиционных материалов, используя данные о начальных свойствах компонентов композиционных материалов.

Актуальность темы заключается в том, что созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

Предмет исследования – методы используемые в Data Science для выявления закономерностей в наборах данных.

Объект исследования – свойства композитных материалов.

**1. Аналитическая часть**

**1.1 Постановка задачи**

Для выполнения выпускной квалификационной работы предоставлено два датасета:

1) X\_bp.xlsx;

2) X\_nup.xlsx.

Файл X\_bp.xlsx содержит данные по 1023 наблюдениям и содержит следующие параметры:

1. индекс наблюдения
2. соотношение матрица-наполнитель;
3. плотность, кг/м3;
4. модуль упругости, Гпа;
5. количество отвердителя, м.%;
6. содержание эпоксидных групп,%\_2;
7. температура вспышки, С\_2;
8. поверхностная плотность, г/м2;
9. модуль упругости при растяжении, ГПа;
10. прочность при растяжении, Мпа;
11. потребление смолы, г/м2.

Файл X\_nup.xlsx содержит данные по 1040 наблюдениям и содержит следующие параметры:

1. индекс наблюдения;
2. угол нашивки, град;
3. шаг нашивки;
4. плотность нашивки.

Количество наблюдений в датасетах разное. При объединении данных по индексу (тип INNER) не все данные из файла X\_nup.xlsx попадут в итоговый датасет. Итоговый датасет содержит информацию о 13 параметрах по 1023 наблюдениям. Пропуски в данных отсутствуют. Тип данных одинаковый.

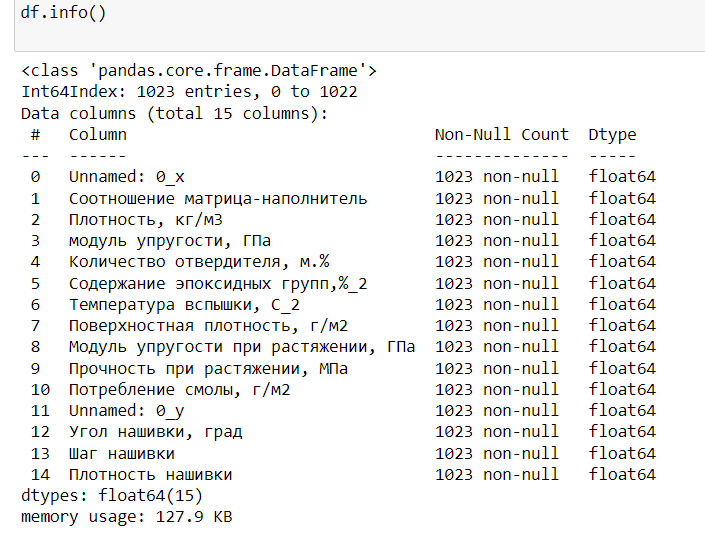


Рисунок 1 - Характеристики параметров наблюдений

Используя исходные данные необходимо обучить алгоритм машинного обучения, который будет определять значения модуль упругости при растяжении, прочность при растяжении и написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.

**1.2 Описание используемых методов**

При решении задачи применялись методы машинного обучения. Машинное обучение — это класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

В данном исследовании требуется получить прогноз на основе выборки объектов с различными признаками, соответственно необходимо использовать методы для решения задач регрессии.

При анализе данных были использованы следующие методы:

1. линейная регрессия;
2. метод k-ближайших соседей;
3. регрессия дерева решений;
4. метод опорных векторов;
5. регрессия нейронной сети.

Линейная регрессия – это метод, используемый для моделирования отношений между одной независимой входной переменной (переменной функции) и выходной зависимой переменной. Модель линейная.

Более общий случай – множественная линейная регрессия, где создаётся модель взаимосвязи между несколькими входными переменными и выходной зависимой переменной. Модель остаётся линейной, поскольку выходное значение представляет собой линейную комбинацию входных значений.

Преимущества:

Быстрое моделирование. В особенности, моделирование можно назвать простым, если отсутствует большой объём данных.

Линейную регрессию легко понять. Она может быть ценна для различных бизнес-решений.

Недостатки:

В случае нелинейных данных полиномиальную регрессию трудно спроектировать. Необходимо иметь информацию о структуре данных и взаимосвязи между переменными.

Основываясь на изложенных выше фактах, линейная регрессия неэффективна, когда речь идёт об очень сложных данных и больших объёмах.

Метод k-ближайших соседей. В случае использования метода для регрессии, объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны. Алгоритм может быть применим к выборкам с большим количеством атрибутов (многомерным). Преимуществом метода является его хорошая математическая обоснованность, недостатком — низкая объясняющая способность.

Регрессия дерева решений. Деревья принятия решений — это непараметрические модели, выполняющие последовательность простых тестов для каждого экземпляра, выполняя обход древовидной структуры двоичных данных до достижения конечного узла (решения).

Деревья принятия решений имеют следующие преимущества:

- они эффективны с точки зрения вычисления и использования памяти во время обучения и прогнозирования;

- они могут представлять границы нелинейного принятия решений;

- они выполняют выбор признаков и классификацию и являются устойчивыми при наличии шумовых признаков.

Эта модель регрессии состоит из совокупности деревьев принятия решений. Каждое дерево в регрессионном лесу решений выводит распределение по Гауссу в виде прогноза. По совокупностям деревьев выполняется агрегирование с целью найти распределение по Гауссу, ближайшее к объединенному распределению для всех деревьев модели.

Метод опорных векторов. Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как метод классификатора с максимальным зазором.

Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с наибольшим зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, создающая наибольшее расстояние до двух параллельных гиперплоскостей. Алгоритм основан на допущении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Регрессия нейронной сети.Несмотря на то, что нейронные сети широко используются для углубленного обучения и моделирования сложных задач, таких как распознавание изображений, они легко адаптируются к задачам регрессии. Любой класс статистических моделей можно назвать нейронной сетью, если эти модели используют адаптивные весовые коэффициенты и могут использоваться для аппроксимации нелинейных функций входных данных. Таким образом, регрессия нейронной сети подходит для задач, которые нельзя решить с помощью более традиционных моделей.

Нейронная сеть выдаст прогнозируемое значение переменной, зависимое от множества входных параметров.

Перед тем, как производить прогноз, алгоритм обучается на тренировочном наборе данных — обучающей выборке. Каждая строка такой выборки содержит:

- в полях, обозначенных как входные — множество входных параметров;

- в поле, обозначенном как выходное — соответствующее входным параметрам значение зависимой переменной.

Технически обучение заключается в нахождении весов — коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными параметрами и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения нейронная сеть способна выдать верный результат на основании данных, которые отсутствовали в обучающей выборке, а также на неполных и/или «зашумленных», частично искажённых данных.

**1.3 Разведочный анализ данных**

Количество наблюдений совпадает с количеством значений каждого параметра, что говорит об отсутствии пропусков. Так же отсутствие пропусков было проверено в главе 1.1.

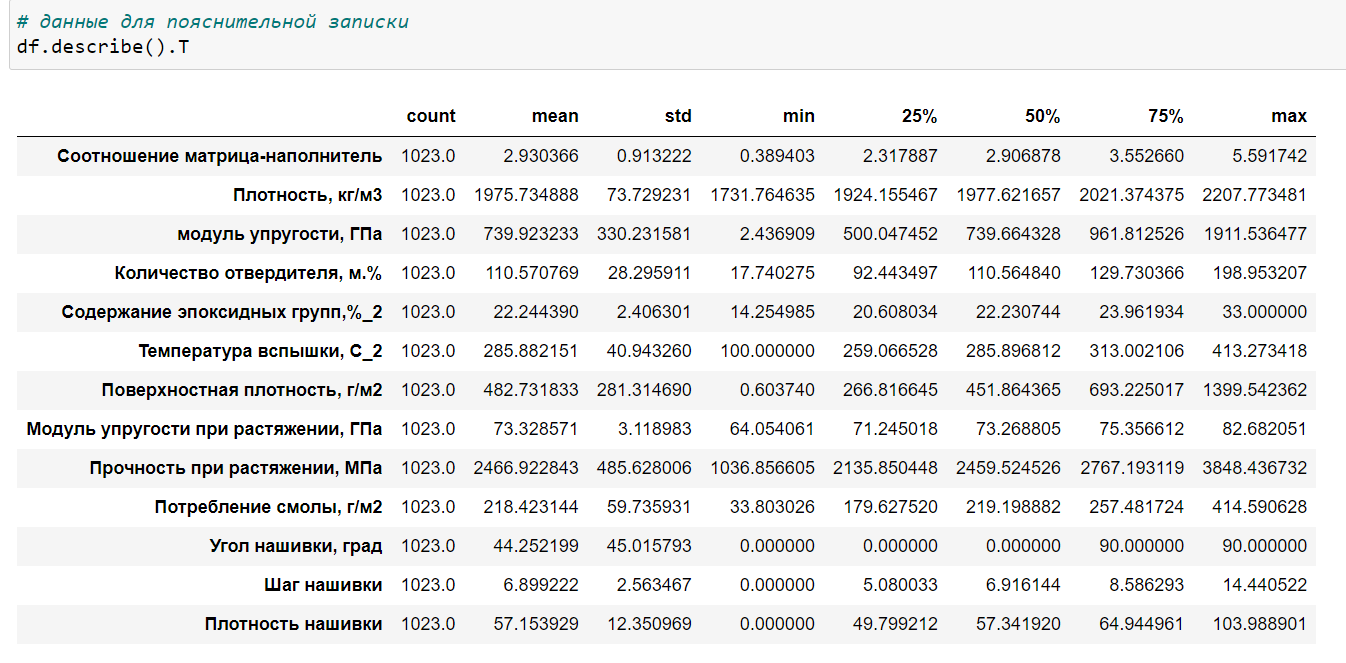


Рисунок 2 – Описательная статистика датасета

На рисунке 3 показано количество уникальных значений параметров. Угол нашивки содержит только 2 значения. Поскольку значение не текстовое к нему можно не применять LabelEncoder. Также перед обучением моделей данные датасета планируется нормализовать, соответственно значения данного параметра после нормализации будут 0 и 1.

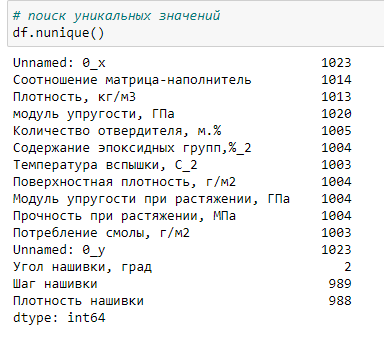


Рисунок 3 – Количество уникальных значений параметров

На рисунке 4 представлены попарные графики значений параметров. Визуально взаимосвязи не наблюдается.

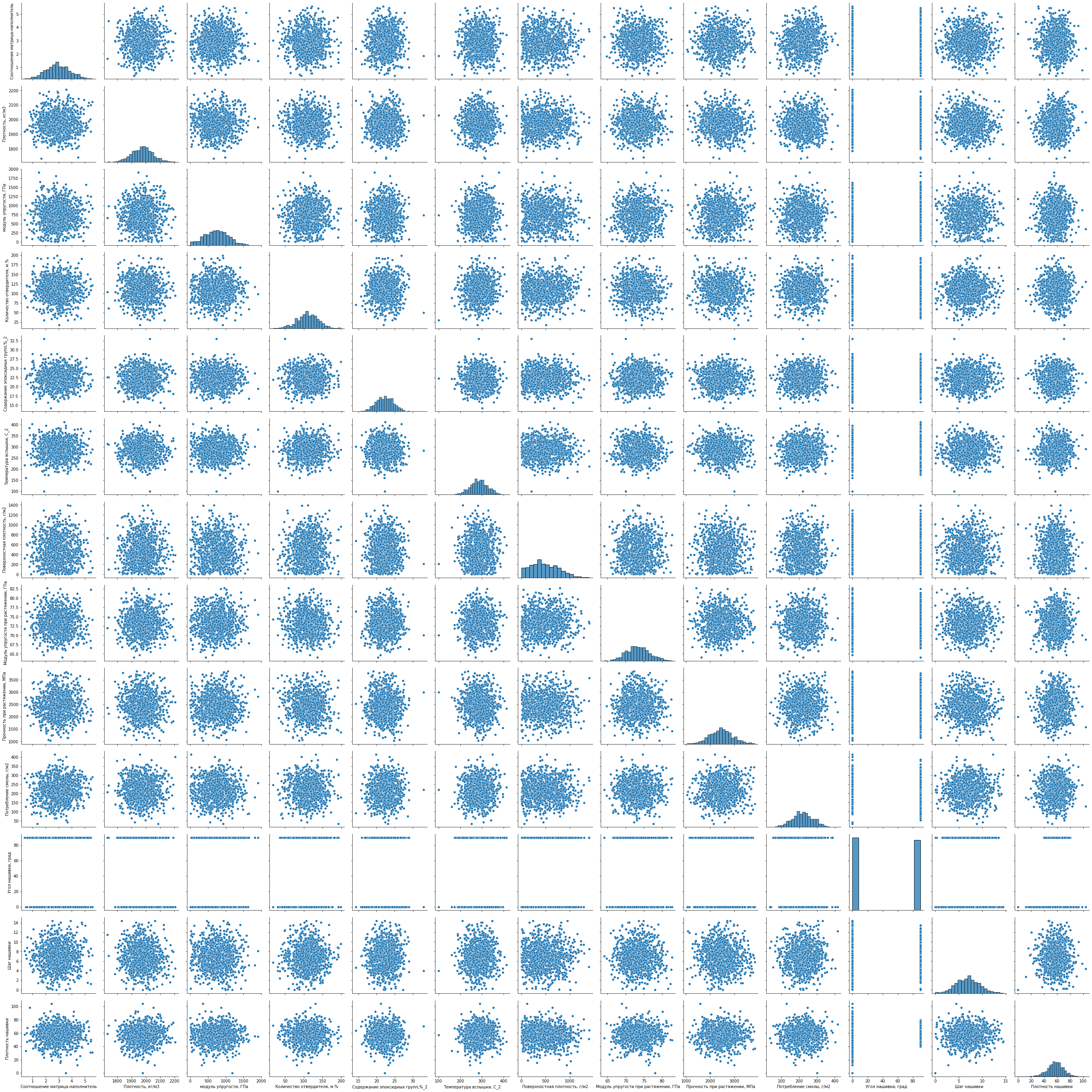


Рисунок 4 – Попарные графики значений параметров

Данные корреляционной матрицы (рисунок 5) так -же не показывают четкой зависимости между данными.

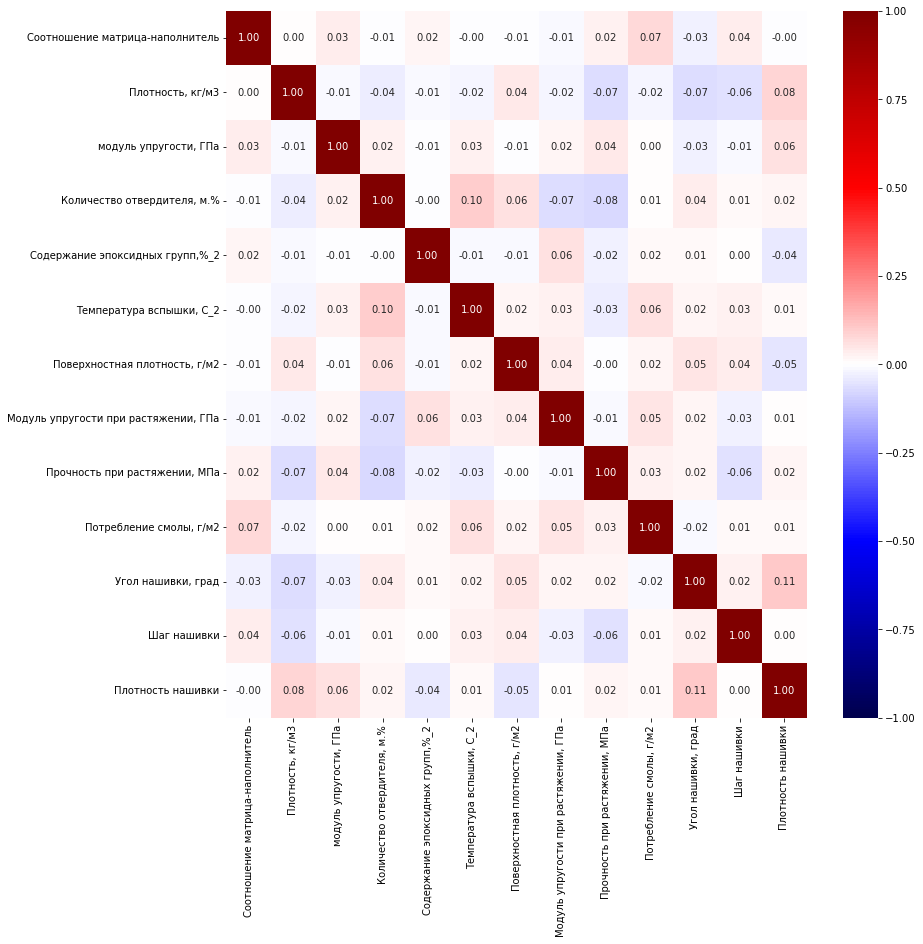


Рисунок 5 – Корреляционная матрица

Графики распределения данных (рисунок 6) показывает, что почти все параметры (кроме угла нашивки) имеют нормальное распределение. Графики «ящик с усами» показывает небольшое кол-во выбросов.

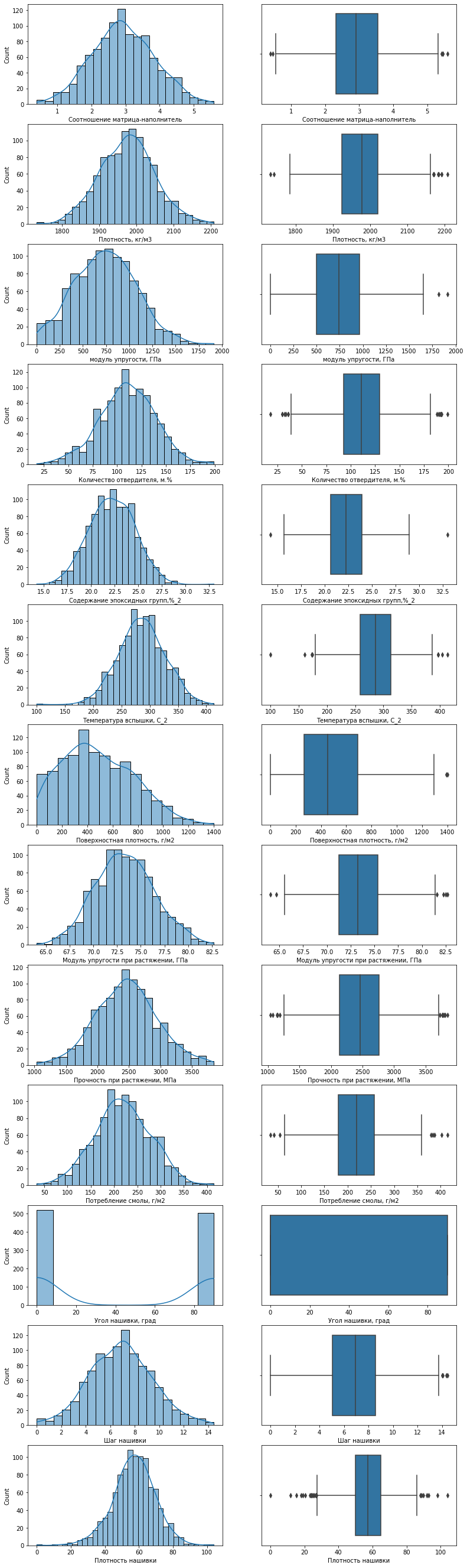


Рисунок 6 – Графики распределения данных

**2. Практическая часть**

**2.1 Предобработка данных**

Как показано на рисунке 6 почти все данные имеют распределение близкое к нормальному и незначительные выбросы. Можно попробовать удалить выбросы использовав правило трех сигм. Количество параметров изменилось не значительно 999 против 1023.

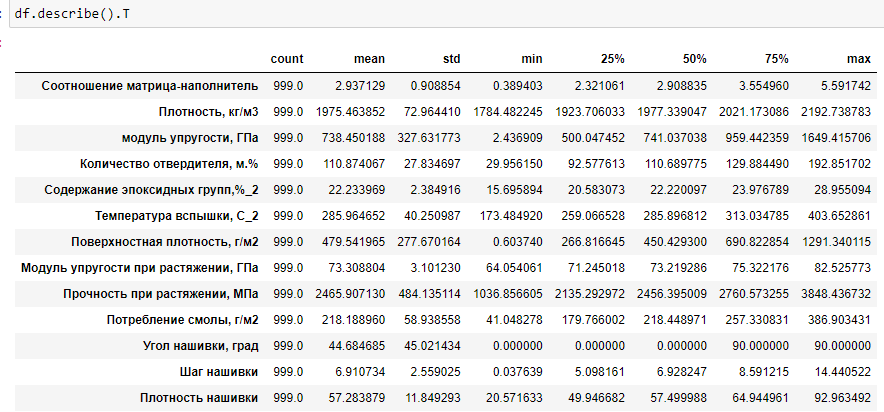


Рисунок 7 – Описательная статистика датасета после удаления выбросов

Корреляция между параметрами (Рисунок 8) также не увеличилась.

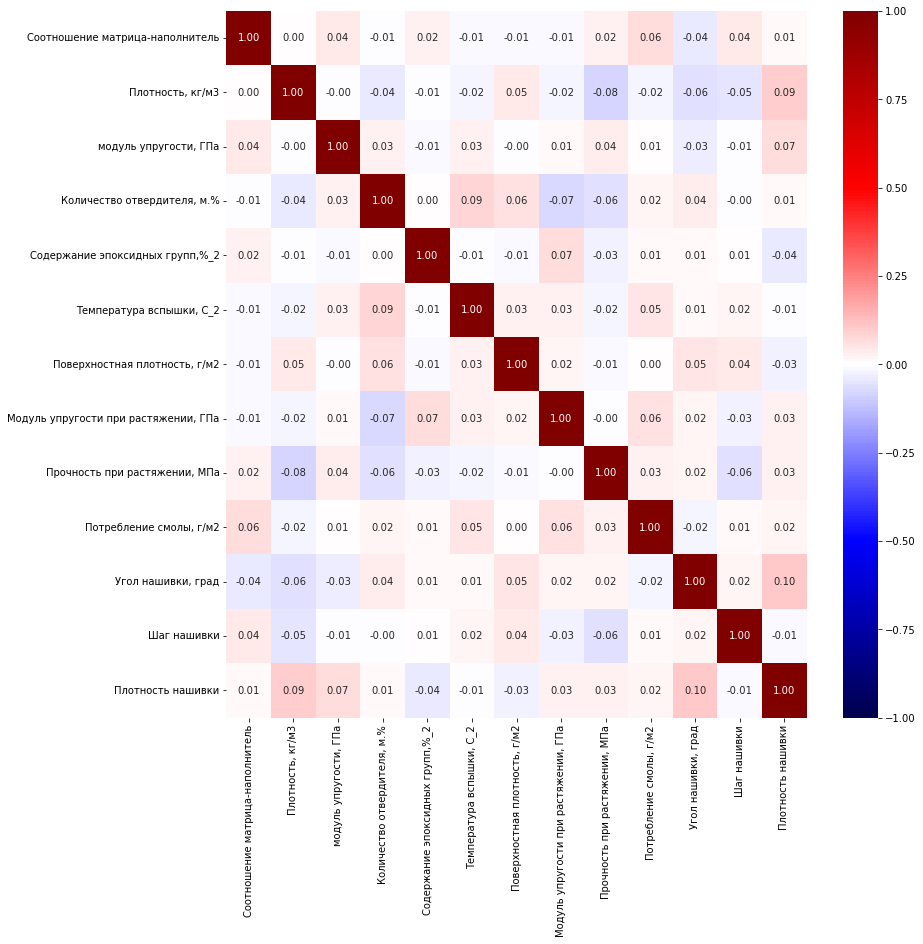


Рисунок 8 – Корреляционная матрица после удаления выбросов

Данные сильно различаются между собой по абсолютным величинам. Работа моделей машинного обучения с такими показателями окажется некорректной из-за увеличения влияния значений, которые имеют большее абсолютное значение. После нормализации все числовые значения входных признаков будут приведены к одинаковой области их изменения. В ВКР для нормализации данных используется метод MinMaxScaler.

После обучения моделей и получения прогнозных данных данные нужно вернуть в исходные значения. Для корректной работы Inverse\_transform датасет первоначально делится на тестовый и тренировочный и только после этого производится нормализация данных.

**2.2 Разработка и обучение модели**

На начальном этапе необходимо разделить выборку на тестовую (30%) и на которой будет происходить обучение. Поскольку предполагается построить несколько моделей для прогнозирования модуля упругости и плотности при растяжении может получится что одна модель хорошо предсказывает одно значение, а вторая другое. Данные переменные предлагается анализировать отдельно:

- df\_y\_elasticity - Модуль упругости при растяжении, ГПа

- df\_y\_strength - Прочность при растяжении, МПа

Далее предполагается настроить и обучить 4 модели регрессии: линейная регрессия, метод k-ближайших соседей, регрессия дерева решений, метод опорных векторов.

После обучения моделей предполагается оценить их по трем параметрам:

### - Средняя абсолютная ошибка (MAE)

### - Средняя квадратичная ошибка (MSE)

### - Корень из средней квадратичной ошибки (RMSE).

Модели, которые имеют гиперпараметры предлагается оптимизировать, используя GridSearchCV.

После получения лучших прогнозных данных их следует преобразовать в исходный вид.

**2.3 Тестирование модели**

Первая модель, которую предлагается протестировать модель линейной регрессии. Полученные данные об ошибках представлены в таблице 1. Значения получили довольно большие (нормализированные данные).

Таблица 1 – Оценка качества модели линейной регрессии

|  |  |
| --- | --- |
| Параметр | Значение |
| MAE для упругости | 0,133 |
| MSE для упругости | 0,028 |
| RMSE для упругости | 0,166 |
| MAE для прочности | 0,136 |
| MSE для прочности | 0,029 |
| RMSE для прочности | 0,171 |

В таблице 2 представлены значения для данных, которые переведены из нормализованного вида.

Таблица 2 - Оценка качества модели линейной регрессии

|  |  |
| --- | --- |
| Параметр | Значение |
| MAE для упругости | 2,46 |
| MSE для упругости | 9,41 |
| RMSE для упругости | 3,07 |
| MAE для прочности | 383,27 |
| MSE для прочности | 230 034,09 |
| RMSE для прочности | 479,62 |

Модуль упругости при растяжении имеет среднее значение 73,3 (см. рисунок 7). Средняя абсолютная ошибка может показаться не большой 2,46. Но при выводе на график прогнозных данных видно, что модель плохо предсказывает значения. Аналогичный график есть для значений плотности при растяжении, представлен в тетраде jypiter.

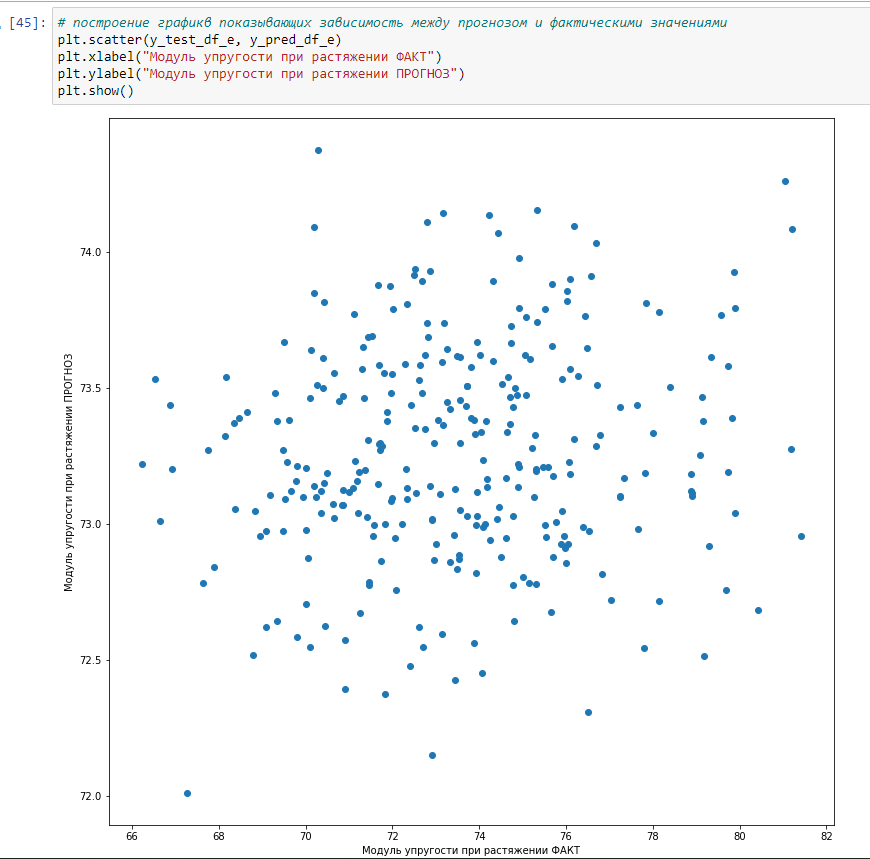


Рисунок 8 – Прогнозные и фактические данные модели линейной регресси

В методе k – ближайших соседей настраивались гиперпараметры через GridSearchCV. Чем больше задавалось количество k – соседей, тем лучше получался результат. Поэтому для подбора оптимальных гиперпараметров было решено использовать метод градиентного спуска. На рисунке 9 показано что качество модели значительно не улучается при количестве соседей больше 20. Поэтому это количество и используется в модели.

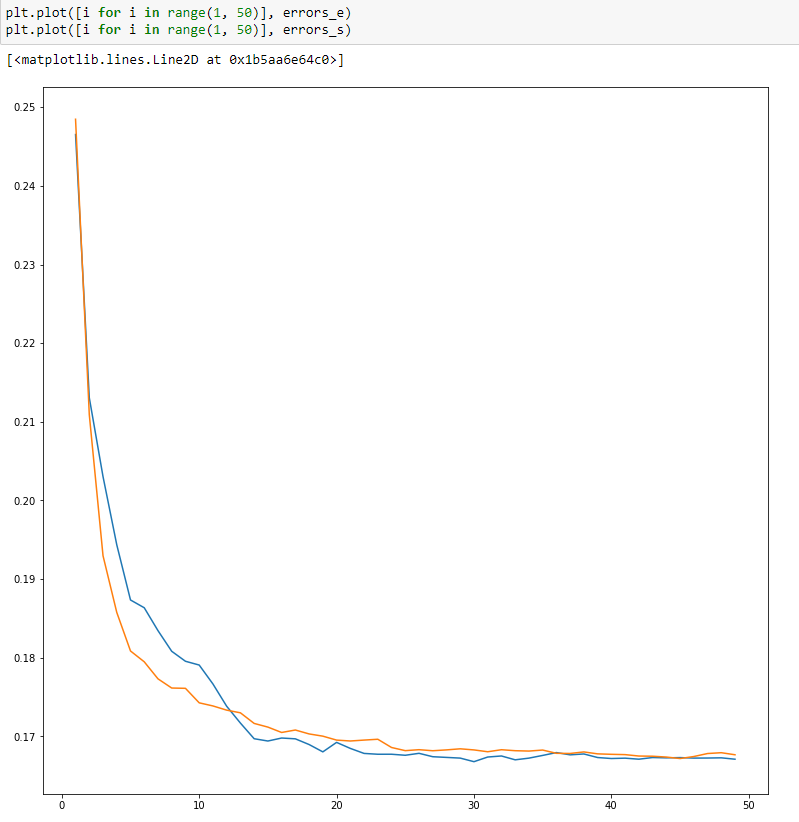


Рисунок 9 – Метод knn – подбор количества ближайших соседей

В таблице 2 представлены данные оценки качества регрессии методом knn с оптимальными гиперпараметрами. Данные получились ненамного лучше, чем в варианте с линейной регрессией.

Таблица 2 – Оценка качества модели регрессии (метод knn)

|  |  |
| --- | --- |
| Параметр | Значение |
| MAE для упругости | 0,136 |
| MSE для упругости | 0,029 |
| RMSE для упругости | 0,169 |
| MAE для прочности | 0,134 |
| MSE для прочности | 0,029 |
| RMSE для прочности | 0,170 |

Метод дерево решений дал похожие результаты. Результаты представлены в таблице 3.

Таблица 3 – Оценка качества модели регрессии (метод дерево решений)

|  |  |
| --- | --- |
| Параметр | Значение |
| MAE для упругости | 0,134 |
| MSE для упругости | 0,028 |
| RMSE для упругости | 0,167 |
| MAE для прочности | 0,132 |
| MSE для прочности | 0,028 |
| RMSE для прочности | 0,168 |

Метод опорных векторов дал аналогичные результаты как и метод дерева решений. Данные представлены в таблице 4.

Таблица 4 – Оценка качества модели регрессии (метод опорных векторов)

|  |  |
| --- | --- |
| Параметр | Значение |
| MAE для упругости | 0,134 |
| MSE для упругости | 0,028 |
| RMSE для упругости | 0,167 |
| MAE для прочности | 0,132 |
| MSE для прочности | 0,028 |
| RMSE для прочности | 0,168 |

Итоговые данные по оценке качества моделей представлены в таблице 5.

Таблица 5 - Итоговые данные по оценке качества моделей

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Параметр | Линейная регрессия | K - ближайшие соседи | Дерево решений | Метод опорных векторов |
| MAE для упругости | 0,133 | 0,136 | 0,134 | 0,134 |
| MSE для упругости | 0,028 | 0,029 | 0,028 | 0,028 |
| RMSE для упругости | 0,166 | 0,169 | 0,167 | 0,167 |
| MAE для прочности | 0,136 | 0,134 | 0,132 | 0,132 |
| MSE для прочности | 0,029 | 0,029 | 0,028 | 0,028 |
| RMSE для прочности | 0,171 | 0,170 | 0,168 | 0,168 |

Поскольку метод дерева решений и метод опорных векторов показали одинаковые результаты, при этом ошибки в прогнозных данных были минимальные можно выделить их как наилучшие. Для примера можно рассмотреть не нормализированные графики, которые показывают взаимосвязь между прогнозными и фактическими значениями. На рисунке 10 видно, что модель спрогнозировала только 2 возможных значения. Аналогичная ситуация с данными для прочности.

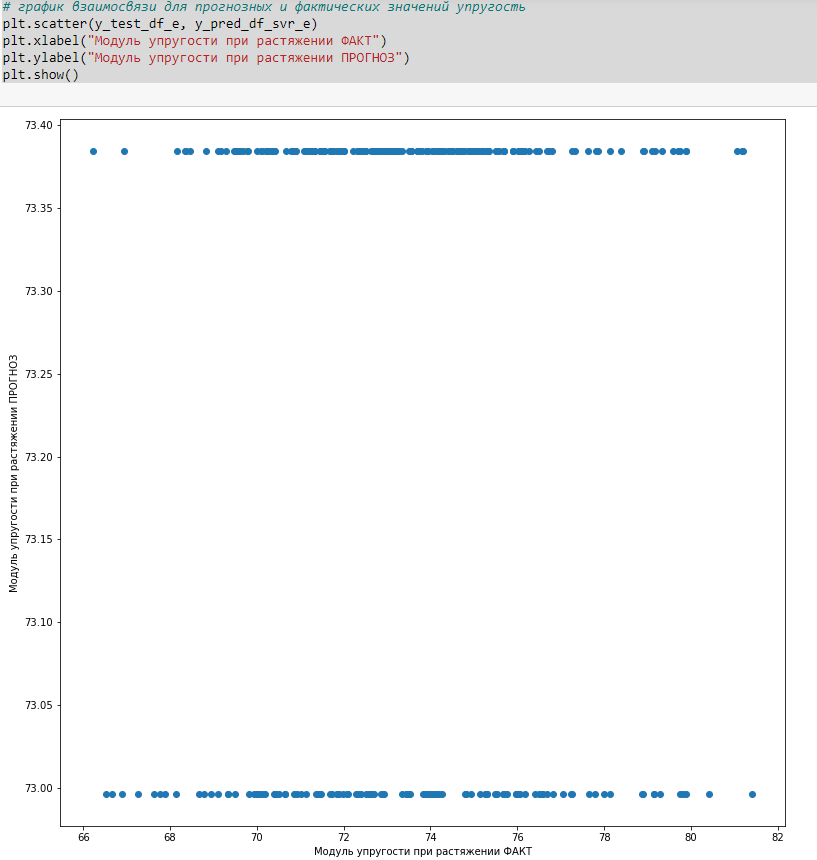


Рисунок 10 – Метод опорных векторов прогнозные и фактические значения для упругости

Из графиков видно, что ни одна из моделей не справилась с задачей хорошо. При этом хуже всего показывает себя метод дерева решений и опорных векторов (прогнозируют только 2 значения). Метод линейной регрессии и k-nn дают лучшие результаты, но также не позволяют спрогнозировать конечные свойства материалов.

**2.4 Написание нейронной сети**

Для прогнозирования соотношения матрица-наполнитель написаны 2 модели, включающая 2 и 4 слоя. В этих целях была задействована библиотека keras. Нормализация данных для нейронной сети производилась предварительно по аналогии с моделью регрессии. Первый слой включает 128 нейронов, выходной 1, т.к. нейронная сеть должна выдавать одно число, показывающее соотношение матрица-наполнитель. Вторая модель имеет 12 входных параметров и 2 промежуточных слоя и один выходной.

В качестве функции ошибки, которая используется для оптимизации в методе обратного распространения ошибки используется среднеквадратичная ошибка.

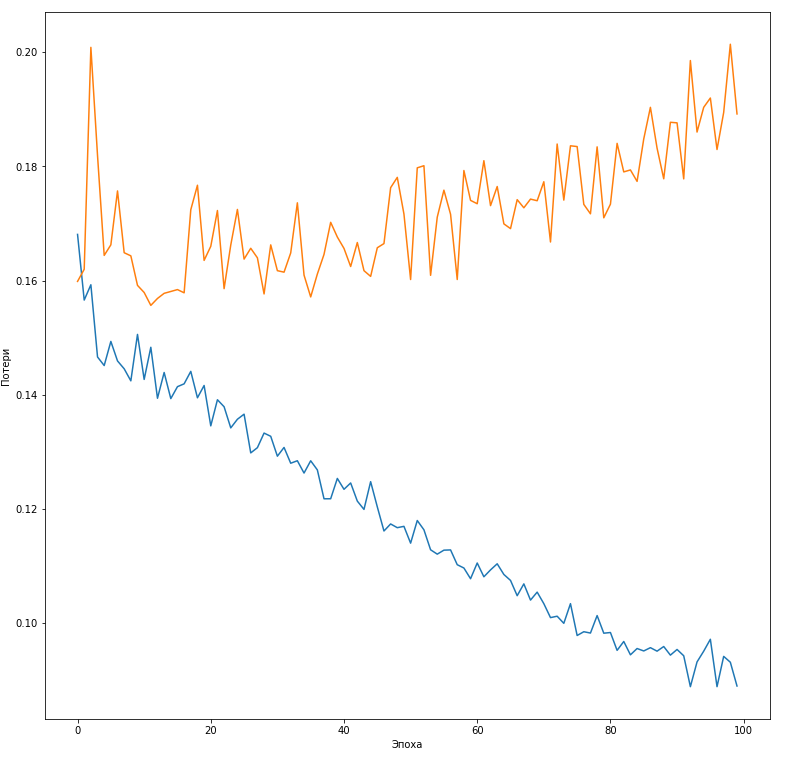


Рисунок 11 – график обучения нейронной сети (mae, val\_mae)

После обучения модели и тестировании на тестовой выборке, получились следующие значения – таблица 6.

Таблица 6 – Оценка качества модели регрессии нейронной сети (2 слоя)

|  |  |
| --- | --- |
| Параметр | Значение |
| MAE матрица-наполнитель | 0,115327 |
| MSE матрица-наполнитель | 0,020508 |
| RMSE матрица-наполнитель | 0,143205 |

График соотношения прогнозных и фактических значений выглядит следующим образом (Рисунок 12).

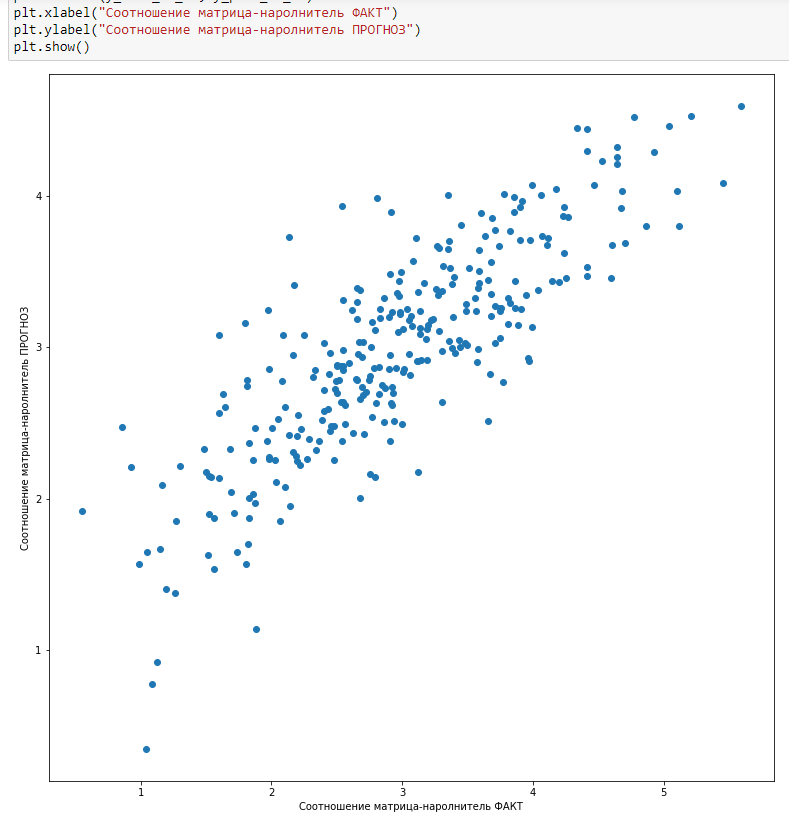


Рисунок 12 - Прогнозные и фактические данные модели регрессии нейронной сети (2 слоя)

Вторая модель нейронной сети имеет немного лучшие показатели качества (таблица 7).

Таблица 7 – Оценка качества модели регрессии нейронной сети (4 слоя)

|  |  |
| --- | --- |
| Параметр | Значение |
| MAE матрица-наполнитель | 0,0708 |
| MSE матрица-наполнитель | 0,0138 |
| RMSE матрица-наполнитель | 0,1176 |

На рисунке 12 видно, что средняя ошибка уменьшается на тренировочных данных в процессе обучения, но увеличивается на тестовых.

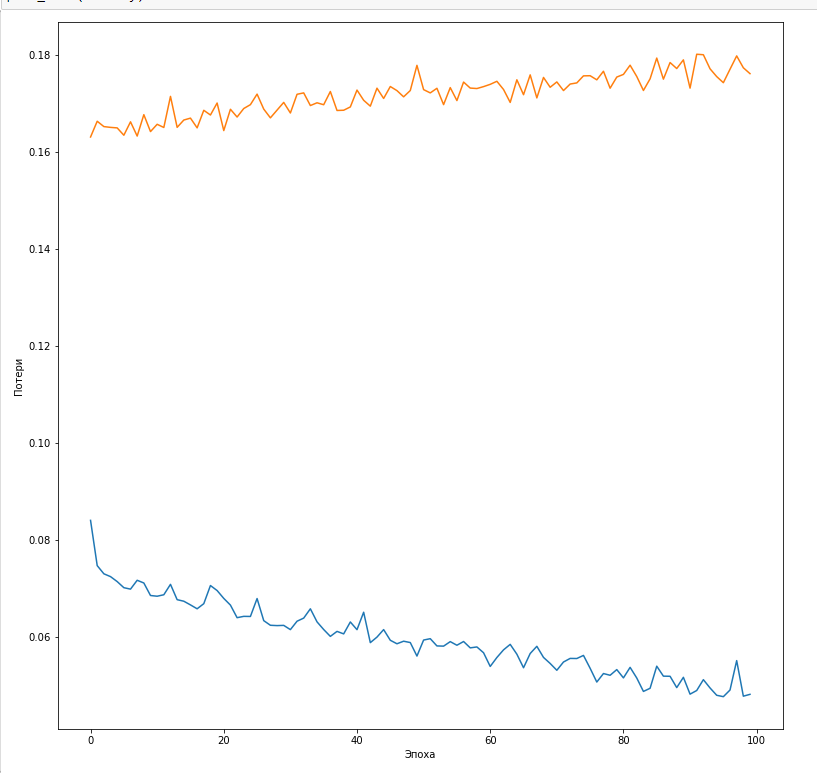


Рисунок 13 – график обучения нейронной сети (mae, val\_mae)

На рисунке 14 показана взаимосвязь между прогнозным и фактическим значением. Прогнозные значения оказались смещены по оси Y.

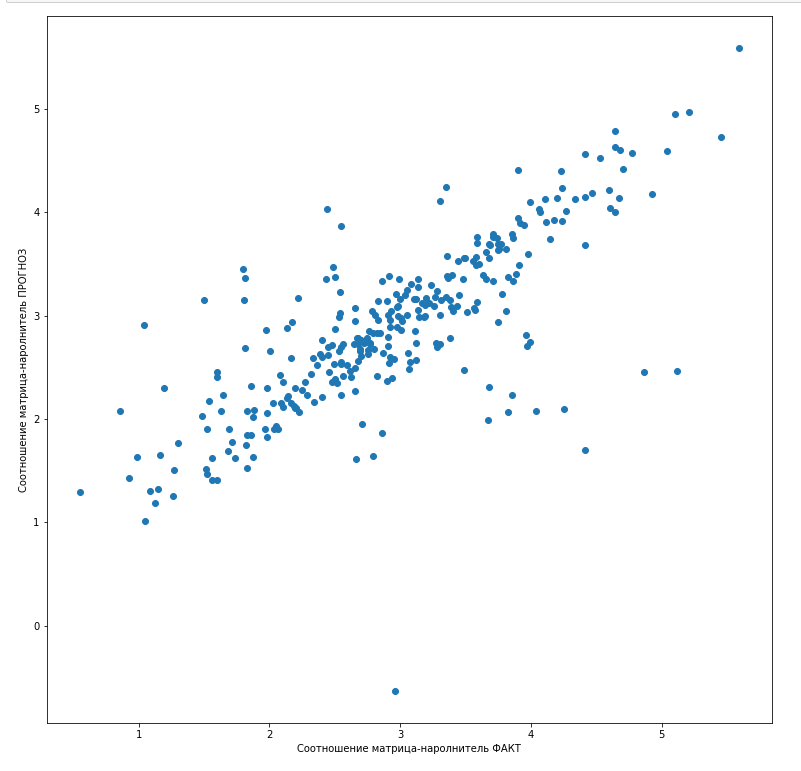


Рисунок 14 - Прогнозные и фактические данные модели регрессии нейронной сети (4 слоя)

Несмотря на большое значение средней абсолютной ошибки график зависимости прогнозных и фактических данных показывает наличие взаимосвязи.

**2.5 Разработка приложения**

В процессе выполнения ВКР было разработано Flask приложение для прогноза модуля упругости при растяжении на основании линейной регрессионной модели. Интерфейс приложения показан на рисунке 15.

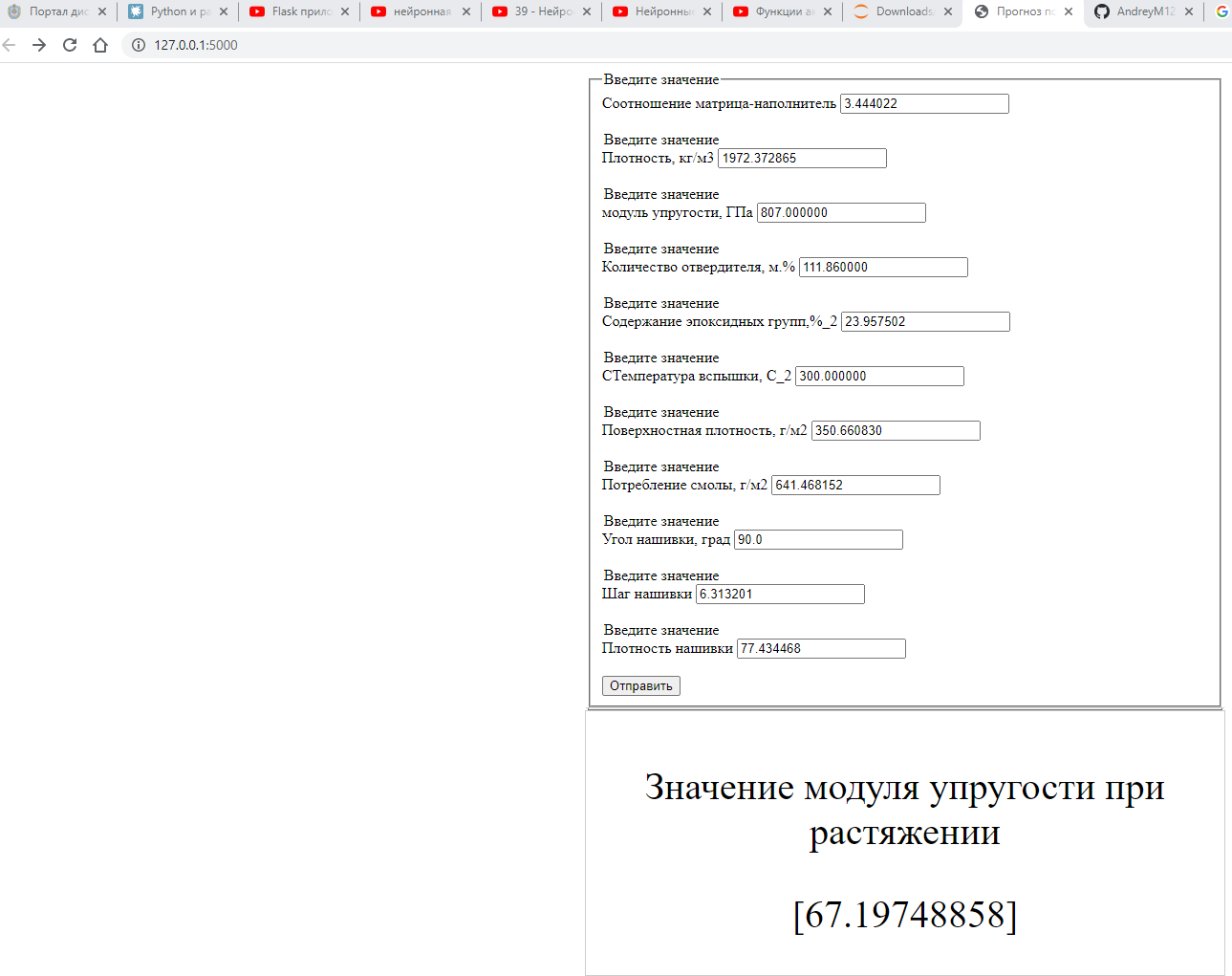


Рисунок 15 – Интерфейс веб-приложения

**2.6 Создание репозитория**

По итогам работы все материалы, включающие исследование в формате jupyter notebook, пояснительная записка, презентация, Flask приложение были размещены в репозитории на GitHub. (<https://github.com/AndreyM123/vkr>

**Заключение**

В ходе решения задачи прогнозирования конечных свойств новых материалов были изучены основные теоретические и практические методы машинного обучения. Проведен предварительный анализ данных и их предобработка. Изучены основные алгоритмы машинного обучения и проведен сравнительный анализ полученных результатов. В моделях были настроены гиперпараметры.

После выполнения исследования разработано веб-приложение, данные загружены в репозиторий.

В ходе выполнения ВКР не удалось разработать модель, которая предсказывала бы значения с приемлемой точностью. Модель нейронной сети показала взаимосвязь, но все предсказанные данные были смещены.

**Библиографический список**

1. Аллен Б. Дауни – Основы Python. Научитесь думать как программист / Аллен Б. Дауни ; пер. с англ. С. Черникова ; [науч. ред. А. Родионов]. — Москва: Манн, Иванов и Фербер, 2021. — 298 с.

2. Билл Любанович. Простой Python. Современный стиль программирования. — СПб.: Питер, 2016. — 328 с.: ил. — (Серия «Бестселлеры O’Reilly»).

3. Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем.Пер. с англ. - СпБ.: ООО Альфа-книга: 2018. - 712 с.

4. Язык программирования Python- Режим доступа: https://www.python.org/. (дата обращения 31.10.2022)

5. Библиотека Pandas- Режим доступа: https://pandas.pydata.org/. (дата обращения 31.10.2022)

6. Библиотека Matplotlib- Режим доступа: https://matplotlib.org/. (дата обращения 31.10.2022)

7. Библиотека Seaborn- Режим доступа: https://seaborn.pydata.org/. (дата обращения 31.10.2022)

8. Библиотека Sklearn- Режим доступа: https://scikit-learn.org/stable/. (дата обращения 31.10.2022)

9. Библиотека Tensorflow: Режим доступа: https://www.tensorflow.org/. (дата обращения 31.10.2022)

10. Машинное обучение [Электронный ресурс]: – Режим доступа: https://ru.wikipedia.org/wiki//Машинное\_обучение (дата обращения: 13.06.2022).