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»

Тема: «Прогнозирование конечных свойств новых материалов

(композиционных материалов)»

Слушатель Москалева Юлия Михайловна

Москва, 2022

Содержание

[**Введение**](#_79c57xalwdgc) **2**

[**1. Аналитическая часть**](#_xl9f2npgkivk) **4**

[1.1 Постановка задачи](#_y4evnvwm3z6q) 4

[1.2 Описание используемых методов](#_67blcgvuu0c8) 4

[1.3 Разведочный анализ данных](#_y8exvm1bfsy1) 5

[**2. Практическая часть**](#_eygcro1ei8qh) **13**

[2.1 Предобработка данных](#_x3yvmdurwxkd) 13

[2.2 Разработка и обучение моделей](#_e10ici55icm5) 17

[2.3 Рекомендательные нейросети для соотношения](#_3ls649jhql1u)

[«матрица – наполнитель»](#_3ls649jhql1u) 22

[**3. Разработка приложения**](#_2zkzgp19w7fv) **27**

[**4. Создание удаленного репозитория и загрузка**](#_6ycvoiu992u2) **28**

[**5. Заключение**](#_xlnt1upl9tzw) **29**

[**6. Список используемой литературы и веб ресурсы**](#_4dgcf6vr6zad) **30**

# Введение

Композиционные материалы — это материалы, состоящие из двух или более компонентов, нерастворимых друг с другом, с чётко обозначенной границей раздела и сильным взаимодействием по всей зоне контакта. Одним из компонентов композитных материалов является непрерывная фаза, он называется матрица, в которой нерастворимые материалы помещаются в другую природу, называемую арматурой или наполнителем.

Внедрение композиционных материалов обусловлено стремлением использовать их преимущества по сравнению с традиционно используемыми металлами и сплавами. Примеры композита – железобетон (сочетание стали арматуры и камня бетона), древесноволокнистая плита ДВП (сочетание древесной основы – щепы и полимерного связующего).

Базальт - магматическая вулканическая порода. Это самая распространённая порода на поверхности Земли и на других планетах Солнечной системы. Базальты образуются путём затвердевания силикатного магматического расплава. Большая часть базальтов образуется на срединно-океанических хребтах и образует океаническую кору. Активно развивается использование композитных материалов на основе базальта.

Базальтопластик - современный композитный материал на основе базальтовых волокон и органического связующего вещества. В настоящее время базальтопластик успешно конкурирует с металлическими изделиями, превосходя их по коррозионной, щелочной, кислотоустойчивости и некоторым другим свойствам. Целью данной работы является прогнозирование конечных свойств новых материалов на основе базальтопластика (композиционных материалов).

Расширение разнообразия материалов, используемых при проектировании нового композиционного материала, увеличивает необходимость определения свойств нового композита при минимальных финансовых затратах. Для решения этой проблемы обычно используются два способа: физические тесты образцов материалов или оценка свойств, в том числе на основе физико-математических моделей. Традиционно разработка композитных материалов является долгосрочным процессом, так как из свойств отдельных компонентов невозможно рассчитать конечные свойства композита. Для достижения определенных характеристик требуется большое количество различных комбинированных тестов, что делает насущной задачу прогнозирования успешного решения, снижающего затраты на разработку новых материалов и затраты на рабочую силу. Суть прогнозирования заключается в моделировании репрезентативного элемента композитного объёма на основе данных о свойствах входящих компонентов (связующего и армирующего компонента). В процессе исследовательской работы были разработаны несколько моделей, способные с высокой вероятностью прогнозировать модули упругости при растяжении и прочности при растяжении, а также были созданы 2 нейронных сети, которые предлагают соотношение «матрицы - наполнитель». На основе одной из нейронных сетей было создано дружелюбное и доступное пользовательское веб-приложение.

# 1. Аналитическая часть

## 1.1 Постановка задачи

Целью настоящей работы является разработка модели для прогноза модуля упругости при растяжении, прочности при растяжении и рекомендации оптимального соотношения «матрица-наполнитель».

Для исследовательской работы были предоставлены два файла: X\_bp.xlsx (с данными о параметрах базальтопластика, состоящий из 1024 строки и 11 столбцов) и X\_nup.xlsx (данными нашивок углепластика, состоящий из 1041 строки и 4 столбцов).

Для каждой колонки необходимо получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; пред обработать данные: удалить шумы и выбросы, сделать нормализацию и стандартизацию. Обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель. Разработать приложение с графическим интерфейсом, которое будет выдавать прогноз соотношения «матрица-наполнитель». Оценить точность модели на тренировочном и тестовом датасете. Создать репозиторий в GitHub и разместить код исследования. Оформить файл README.

## **1.2 Описание используемых методов**

Данная задача в рамках классификации категорий машинного обучения относится к машинному обучению с учителем и традиционно это задача регрессии. Цель любого алгоритма обучения с учителем — определить функцию потерь и минимизировать её, поэтому для наилучшего решения в процессе исследования были применены следующие методы:

метод опорных векторов;

случайный лес;

линейная регрессия;

К-ближайших соседей;

Известные и широкоприменяемые методы, такие как градиентный бустинг,

дерево решений, стохастический градиентный спуск, лассо-регрессия, многослойный перцептрон, стохастический градиентный спуск в данной работе не применялись, поскольку показали свою низкую эффективность на обширном предварительном этапе исследования.

Цель работы разработать модели для прогноза модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель». Для этого нужно объединить 2 файла

Данные объединенного датасета не имеют чётко выраженной зависимости, что подтверждает тепловая карта с матрицей корреляции и матрицы диаграмм рассеяния, приведенные ниже:

<class 'pandas.core.frame.DataFrame'>

Int64Index: 1023 entries, 0 to 1022

Data columns (total 15 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Unnamed: 0\_x 1023 non-null float64

1 Соотношение матрица-наполнитель 1023 non-null float64

2 Плотность, кг/м3 1023 non-null float64

3 модуль упругости, ГПа 1023 non-null float64

4 Количество отвердителя, м.% 1023 non-null float64

5 Содержание эпоксидных групп,%\_2 1023 non-null float64

6 Температура вспышки, С\_2 1023 non-null float64

7 Поверхностная плотность, г/м2 1023 non-null float64

8 Модуль упругости при растяжении, ГПа 1023 non-null float64

9 Прочность при растяжении, МПа 1023 non-null float64

10 Потребление смолы, г/м2 1023 non-null float64

11 Unnamed: 0\_y 1023 non-null float64

12 Угол нашивки, град 1023 non-null float64

13 Шаг нашивки 1023 non-null float64

14 Плотность нашивки 1023 non-null float64

dtypes: float64(15)

memory usage: 127.9 KB

Рисунок 1 - пример из кода с результатами объединения файлов.

## 1.3 Разведочный анализ данных

Прежде чем передать данные в работу моделей машинного обучения, необходимо обработать и очистить их. Очевидно, что «грязные» и необработанные данные могут содержать искажения и пропущенные значения – это ненадёжно, поскольку способно привести к крайне неверным результатам по итогам моделирования. Но безосновательно удалять что-либо тоже неправильно. Именно поэтому сначала набор данных надо изучить.

Цель разведочного анализа - получение первоначальных представлений о характерах распределений переменных исходного набора данных, формирование оценки качества исходных данных (наличие пропусков, выбросов), выявление характера взаимосвязи между переменными с целью последующего выдвижения гипотез о наиболее подходящих для решения задачи моделях машинного обучения.

В качестве инструментов разведочного анализа используется: оценка статистических характеристик датасета; гистограммы распределения каждой из переменной (несколько различных вариантов); диаграммы “Ящики с усами” (несколько интерактивных вариантов); попарные графики рассеивания точек (несколько вариантов); тепловая карта (несколько вариантов); описательная статистика для каждой переменной; анализ и полное исключение выбросов (возможно несколько повторных итераций).

Разведочный анализ данных. Гистограмма.

Библиотека seaborn упрощает создание гистограмм и графиков плотности с помощью метода histplot, который позволяет одновременно строить как гистограмму, так и непрерывную оценку плотности. В качестве примера рассматривается бимодальное распределение, состоящее из двух разных стандартных нормальных распределений. Использован colorbar\_factory из библиотеки matplotlib.

Кривая — график распределения, который, по сути, является сглаженным аналогом гистограммы. По оси y откладывается плотность. Гистограмма по умолчанию нормализована. Поэтому её масштаб по оси y соответствует масштабу графика распределения.

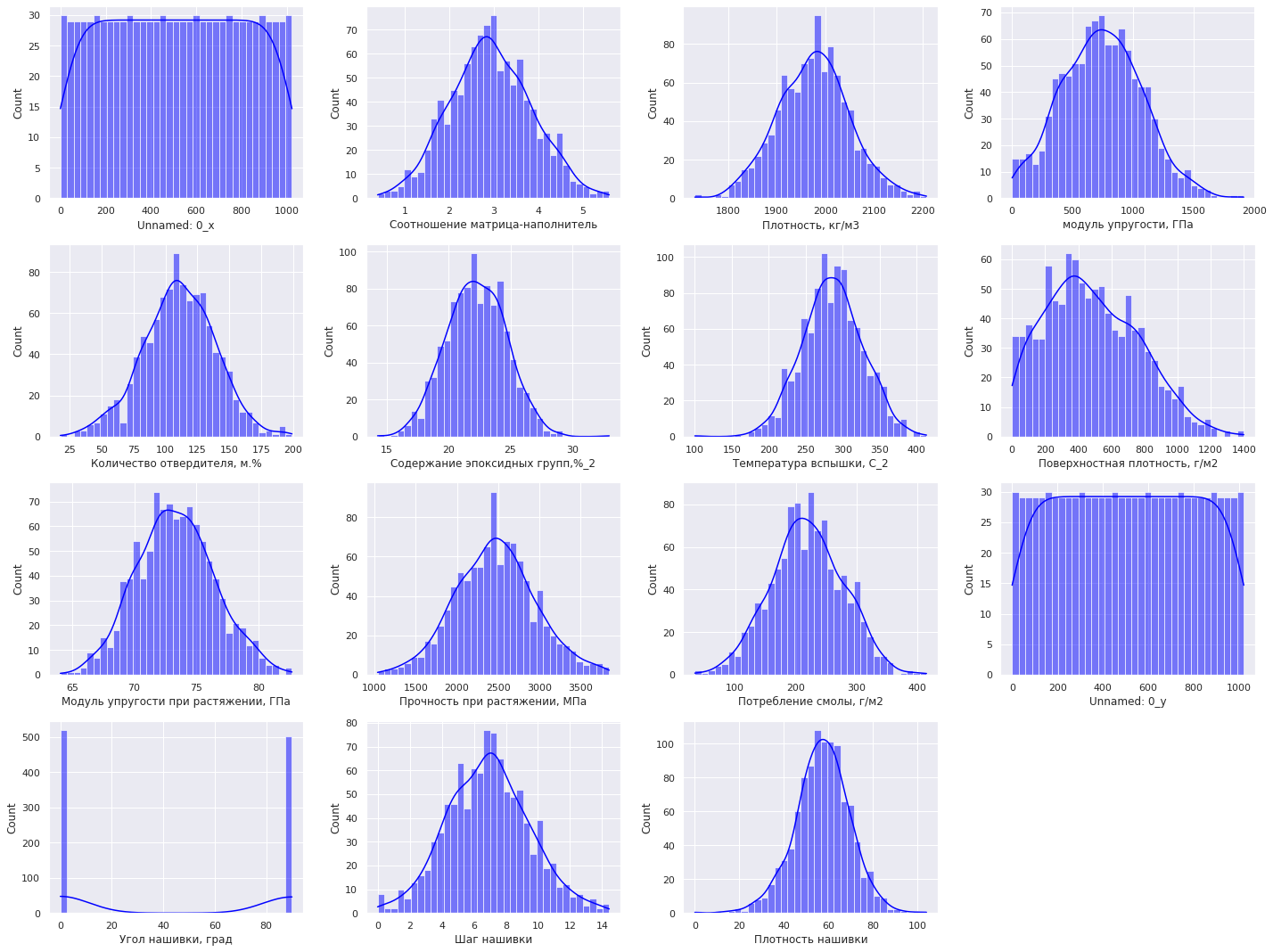


Рисунок 2 - Разведочный анализ данных. Гистограмма.

Дальнейший этап исследовательского анализ данных - построение попарных графиков рассеивания точек. Необходим для выяснения того, что еще данные могут показать и используется, чтобы найти закономерности, отношения или аномалии для информирования последующего анализа. Несмотря на то, что в EDA существует почти подавляющее число методов, одним из наиболее эффективных стартовых инструментов является построение парных пар (также называемое матрицей диаграмм рассеяния). Сюжет пар позволяет нам видеть как распределение отдельных переменных, так и отношения между двумя переменными. Попарные графики являются отличным методом для определения тенденций для последующего анализа и легко реализуются в Python. Стоит отметить, что построение данных графиков (выполнение запущенной ячейки с кодом) занимает сравнительно много времени, тем не менее удобство предоставляемой ими информации имеет большее значение.

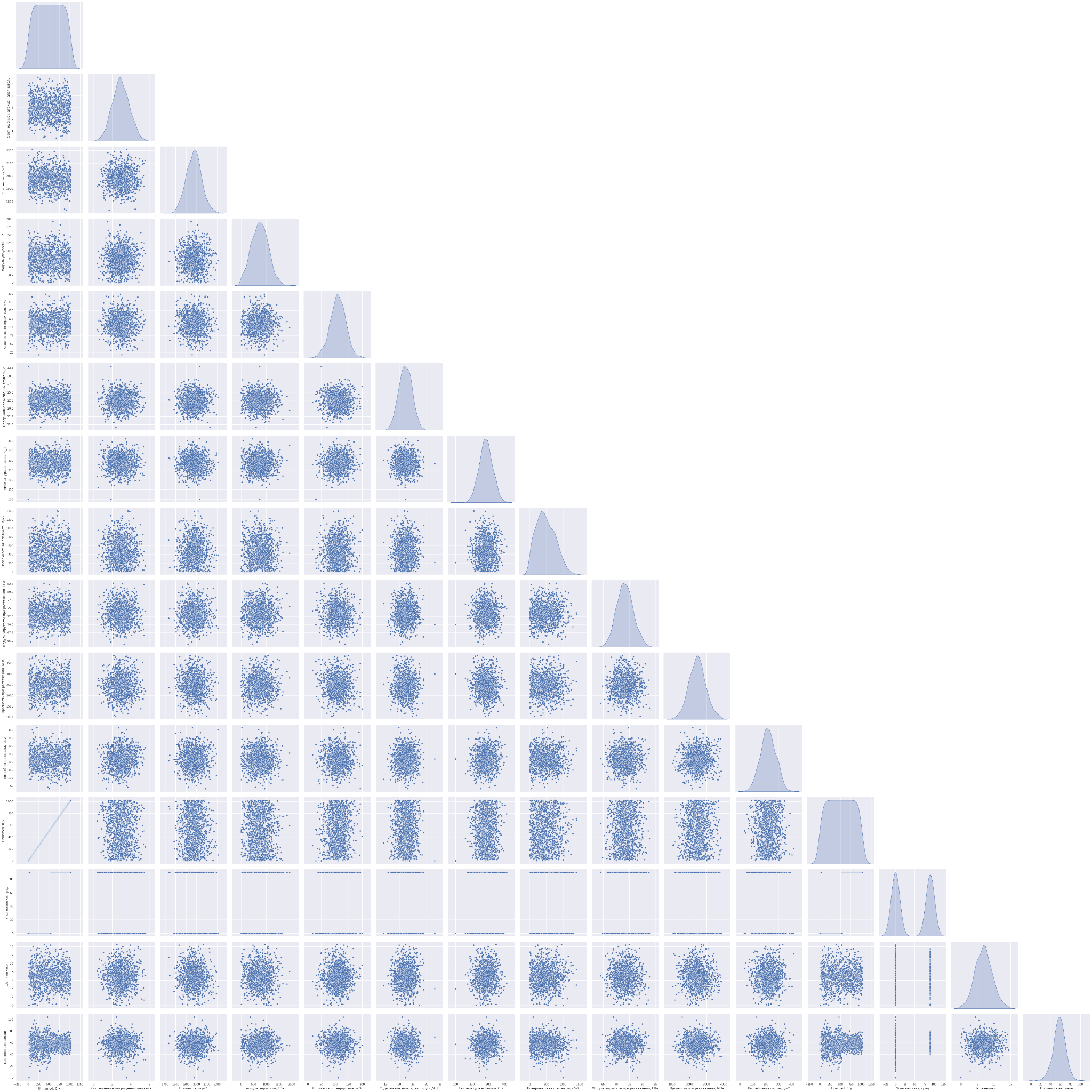


Рисунок 3 - График попарного рассеивания точек.

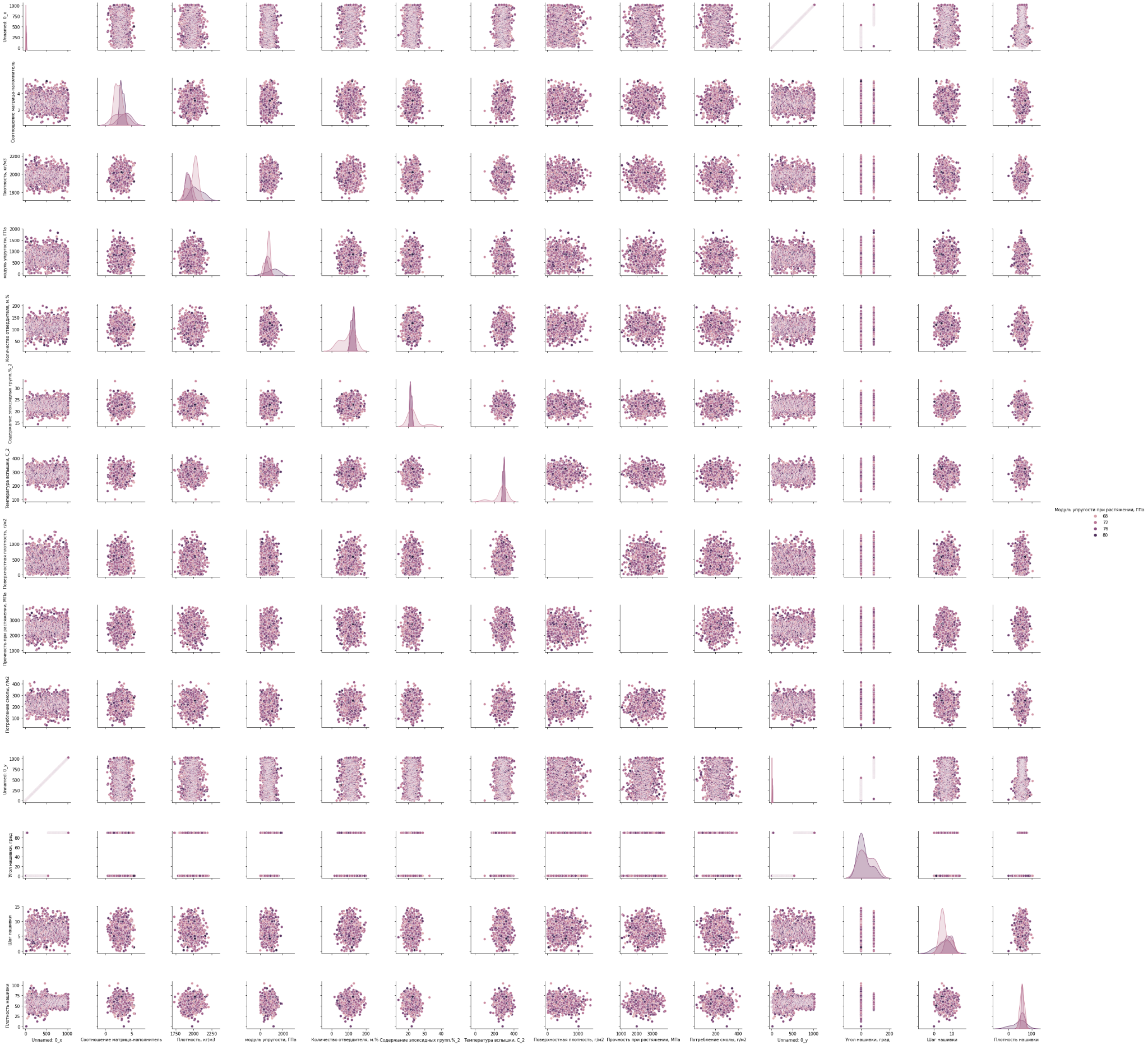


Рисунок 4 - Дополнительный график попарного рассеивания точек в привязке к целевой переменной “Модуль упругости при растяжении, ГПа”

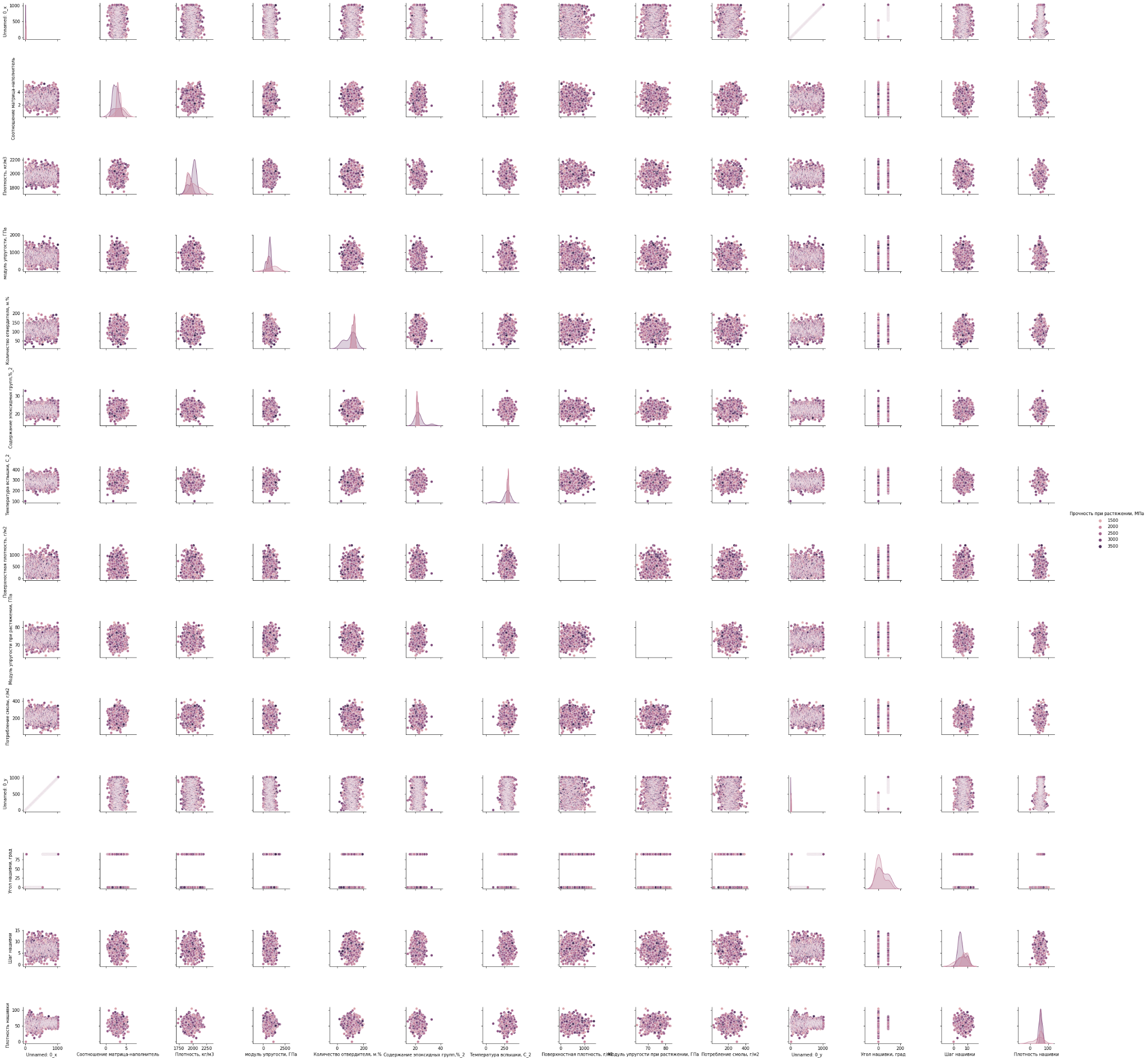


Рисунок 5 - Дополнительный график попарного рассеивания точек в привязке к целевой переменной “Прочность при растяжении, МПа”

Корреляция - важнейший фактор, лежащий в основе анализа данных. Он сообщает, как переменные в наборе данных связаны друг с другом и как они перемещаются относительно друг друга. Значение корреляции колеблется от -1 до +1. 0 Корреляция означает, что две переменные не зависят друг от друга. Положительная корреляция указывает на то, что переменные движутся в одном направлении, а отрицательная корреляция указывает на противоположное.

В данной работе корреляционная матрица строится с помощью модуля seaborn Это помогает легко понять набор данных и очень часто используется для анализа.

Тепловая карта используется для графического представления значений матрицы с разными цветовыми оттенками для разных значений. Он очень четко визуализирует общую матрицу.

Особенность графического представления значений состоит также в том, что если удалить половину данных на одной стороне главной диагонали, то не потеряется никакой важной информации, поскольку она повторяется (благодаря этому, также может создаваться треугольный сюжет).

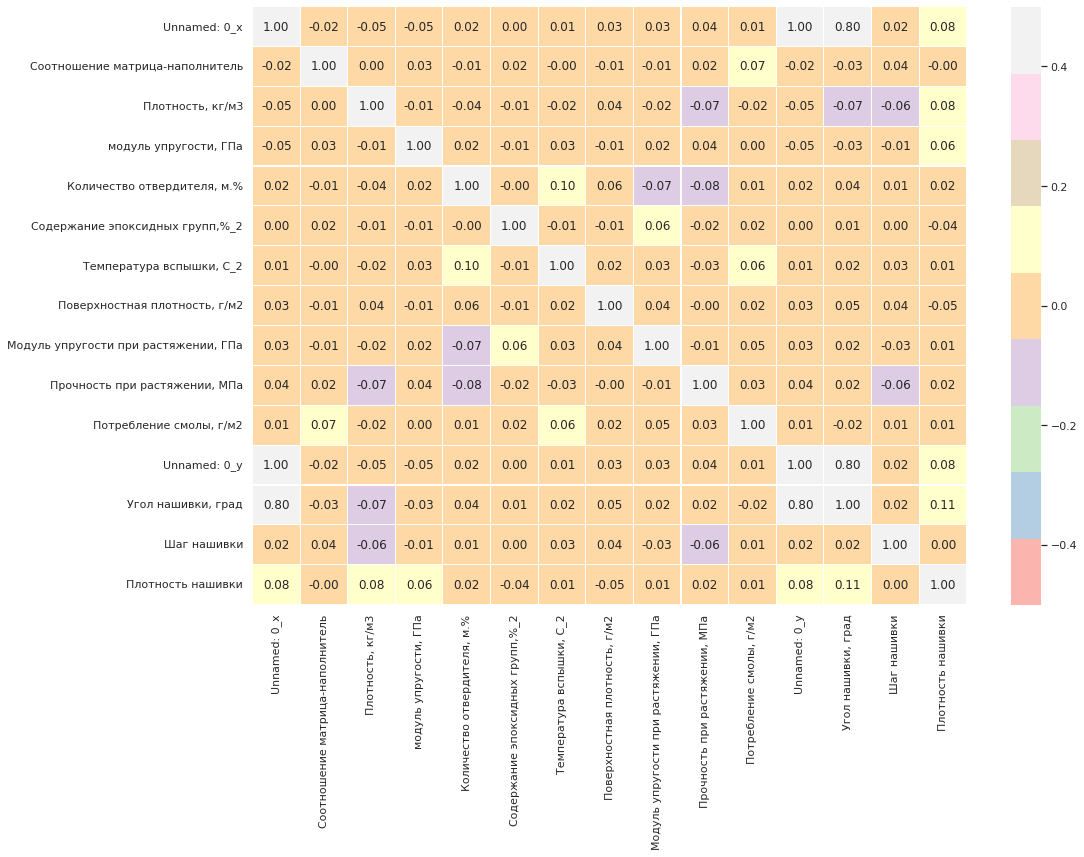


Рисунок 6 - Матрица корреляции данных.

Максимальная корреляция между плотностью нашивки и углом нашивки 0.11, значит практически нет зависимости между этими данными. Корреляция между всеми параметрами очень близка к 0, корреляционные связи между переменными не наблюдаются.

Диаграмма “Ящик с усами” - график, использующийся в описательной статистике, компактно изображающий одномерное [распределение вероятностей](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9). По сути, ящик с усами — это быстрый способ изучения одного или нескольких наборов данных в графическом виде. Этот график может показаться более примитивным, чем, например, [гистограммы](https://ru.wikipedia.org/wiki/%D0%93%D0%B8%D1%81%D1%82%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0), но он имеет некоторые преимущества. Он занимает меньше места и поэтому особенно полезен для сравнения распределений между несколькими группами или наборами данных. Кроме того, ящик с усами в своей первоначальной форме прост для построения. Данный график более компактный и по нему легко можно оценить [медианы](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B4%D0%B8%D0%B0%D0%BD%D0%B0_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)), [квартили](https://ru.wikipedia.org/wiki/%D0%9A%D0%B2%D0%B0%D0%BD%D1%82%D0%B8%D0%BB%D1%8C), [дисперсию](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D0%B8%D1%8F_%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D0%BE%D0%B9_%D0%B2%D0%B5%D0%BB%D0%B8%D1%87%D0%B8%D0%BD%D1%8B) и асимметрию в данных, а также выявить [выбросы](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D1%80%D0%BE%D1%81_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)). Асимметрию данных можно увидеть не только по медиане, смещенной к какому-либо концу ящика, но и по разной длине усов, выходящих из ящика.

В настоящей работе построение графика “Ящики с усами” используются главным образом, для выявления выбросов.

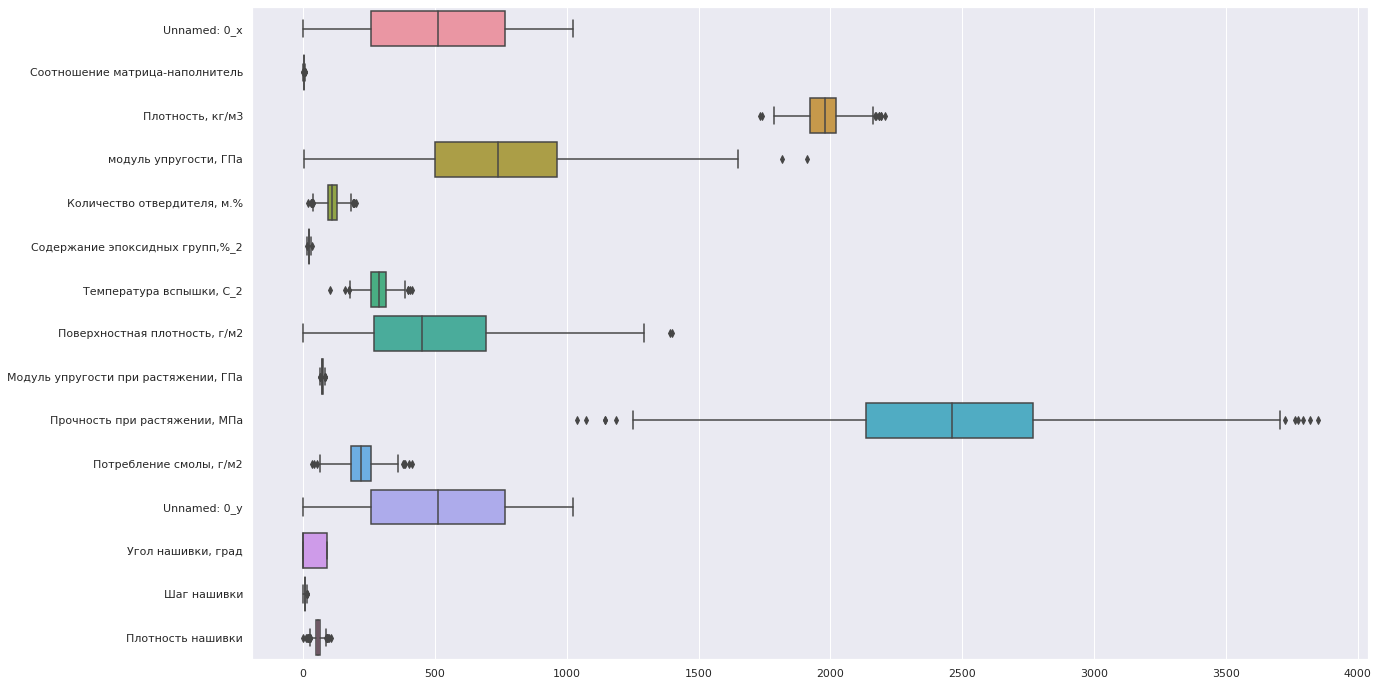


Рисунок 7 - Диаграмма “Ящики с усами” до удаления выбросов.

# 2. Практическая часть

## 2.1 Предобработка данных

Предобработку данных удобно начинать сразу же с удаления выбросов.

Выброс — это наблюдение, которое лежит аномально далеко от других значений в наборе данных. Выбросы могут быть проблематичными, поскольку они могут повлиять на результаты анализа.

Прежде чем удалять выбросы, нужно решить, что считать выбросом. Есть два распространенных способа сделать это: метод межквартильных расстояний и метод трех сигм.

Межквартильный размах (IQR) — это разница между 75-м процентилем (Q3) и 25-м процентилем (Q1) в наборе данных. Он измеряет разброс средних 50% значений.

Определить наблюдение как выброс можно, если оно в 1,5 раза превышает межквартильный размах, превышающий третий квартиль (Q3), или в 1,5 раза превышает межквартильный размах, меньше первого квартиля (Q1).

Выбросы = наблюдения > Q3 + 1,5\*IQR или Q1 – 1,5\*IQR

Если в данных присутствует один или несколько выбросов, необходимо убедиться, что они не являются результатом ошибки ввода данных.

Если выброс оказался результатом ошибки ввода данных, можно присвоить ему новое значение, такое как [среднее значение или медиана](https://www.codecamp.ru/blog/measures-central-tendency/) набора данных.

Если значение является истинным выбросом, можно удалить его, если оно окажет значительное влияние на общий анализ. Об удалении выброса необходимо упомянуть в окончательном отчете или анализе.

В данной работе было произведено удаление выбросов при помощи метода межквартильных расстояний.

for i in df.columns:

q75, q25 = np.percentile(df.loc[:,i], [75,25])

intr\_qr = q75 - q25

max = q75 + (1.5 \* intr\_qr)

min = q25 - (1.5 \* intr\_qr)

df.loc[df[i] < min, i] = np.nan

df.loc[df[i] > max, i] = np.nan

df.dropna(inplace = True)

Оценка результата удаления выбросов произведена путем повторного построения графика “Ящики с усами”.

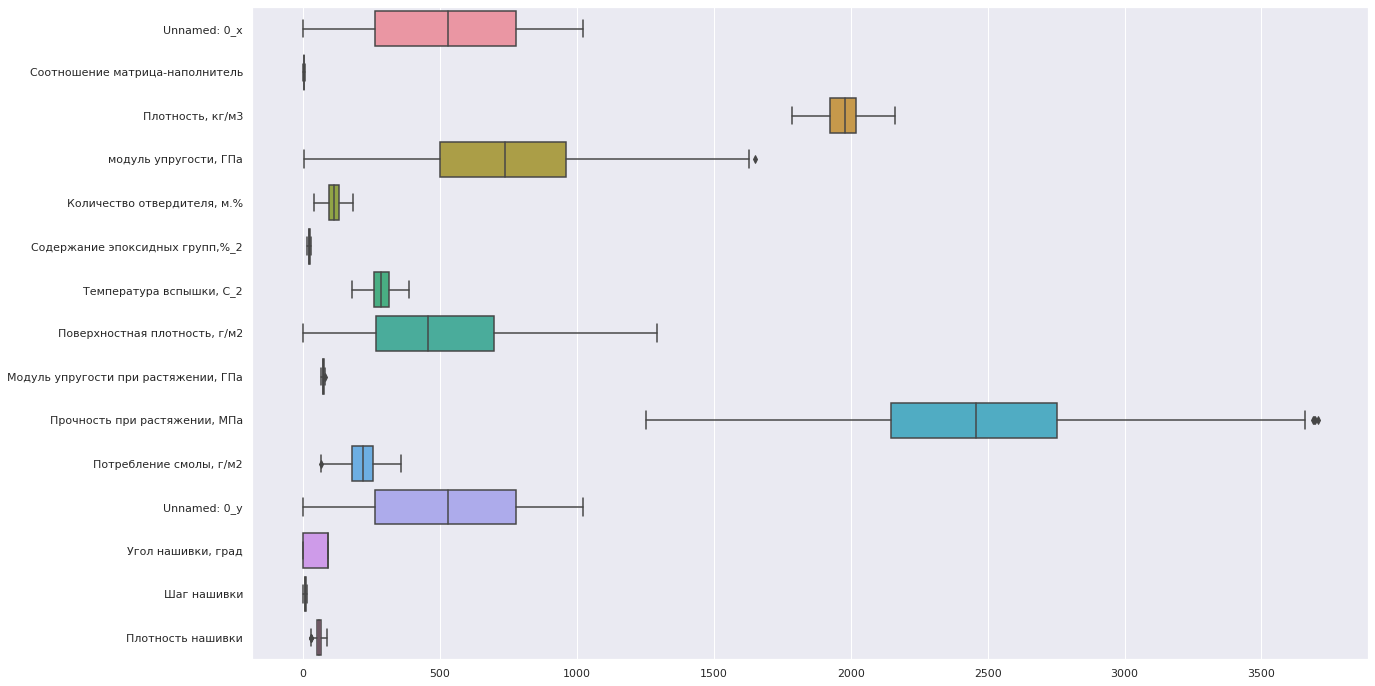


Рисунок 8 - Диаграмма “Ящики с усами” после удаления выбросов.

Таким образом выведение df.shape показывает теперь (936, 15), тогда, как до удаления было (1023, 15).

Самое время оценить плотность ядра. Стоит отметить, что с увеличением полосы пропускания распределение становится более сглаженным. В большинстве случаев, ядро Гаусса по умолчанию и стандартная оценка полосы пропускания работают хорошо.

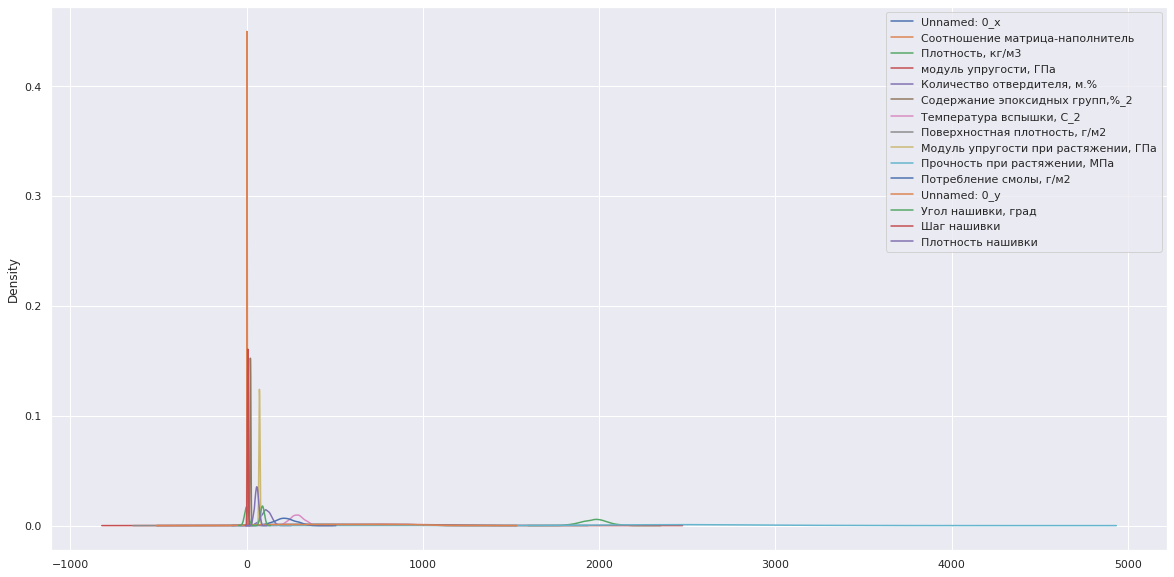


Рисунок 9 - Оценка распределения до нормализации данных.

После проведения разведочного анализа данных производится их нормализация.

Нормализация данных – это метод, который ускоряет получение желаемого результата за счет того, что машине приходится обрабатывать меньший диапазон данных. Нормализация – непростая задача, потому что все результаты зависят от выбора правильного метода нормализации и от типа данных. Зависимость здесь прямая: каждый тип данных имеет свои методы нормализации.

В данной работе применен наиболее популярный метод MinMaxScaler, который предоставляет библиотека sklearn.

Результаты также оцениваются при помощи графиков:

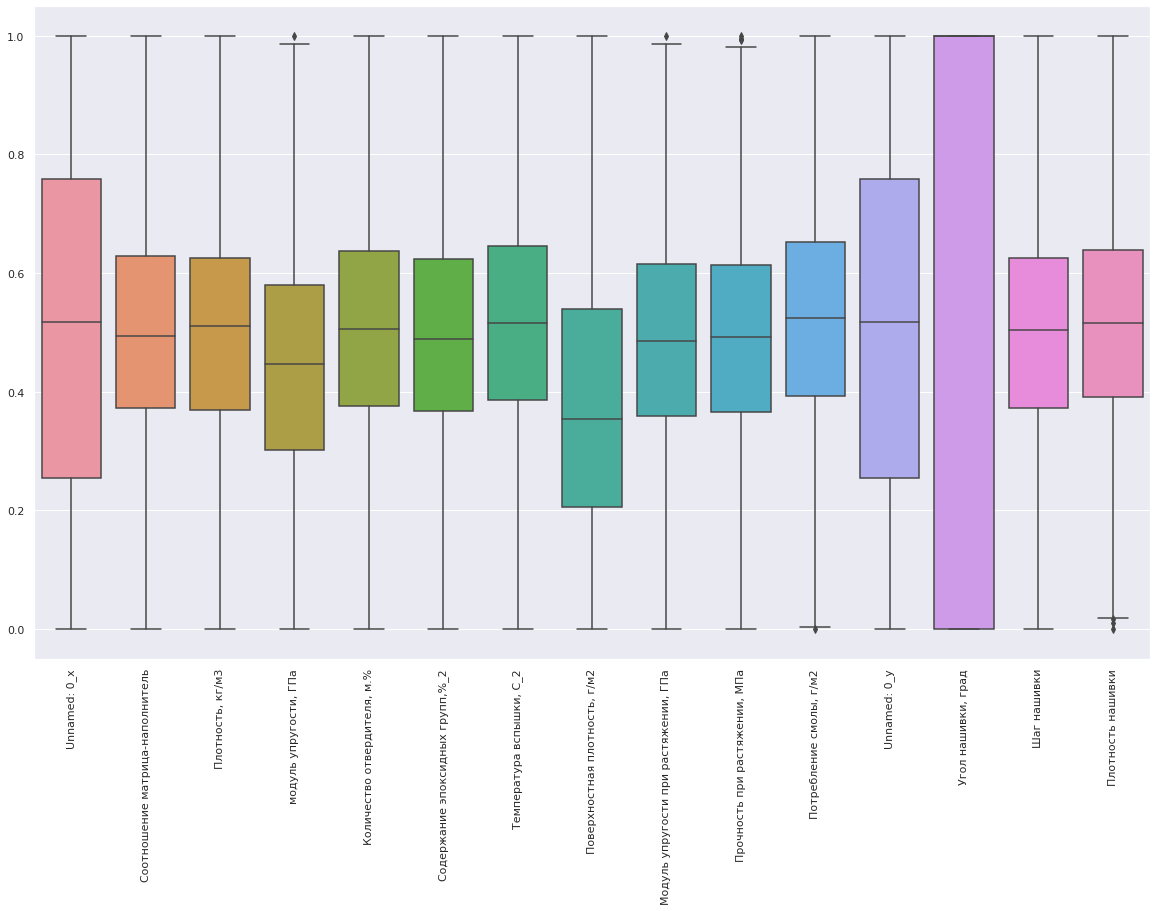


Рисунок 10 - Диаграмма “Ящики с усами” после нормализации.

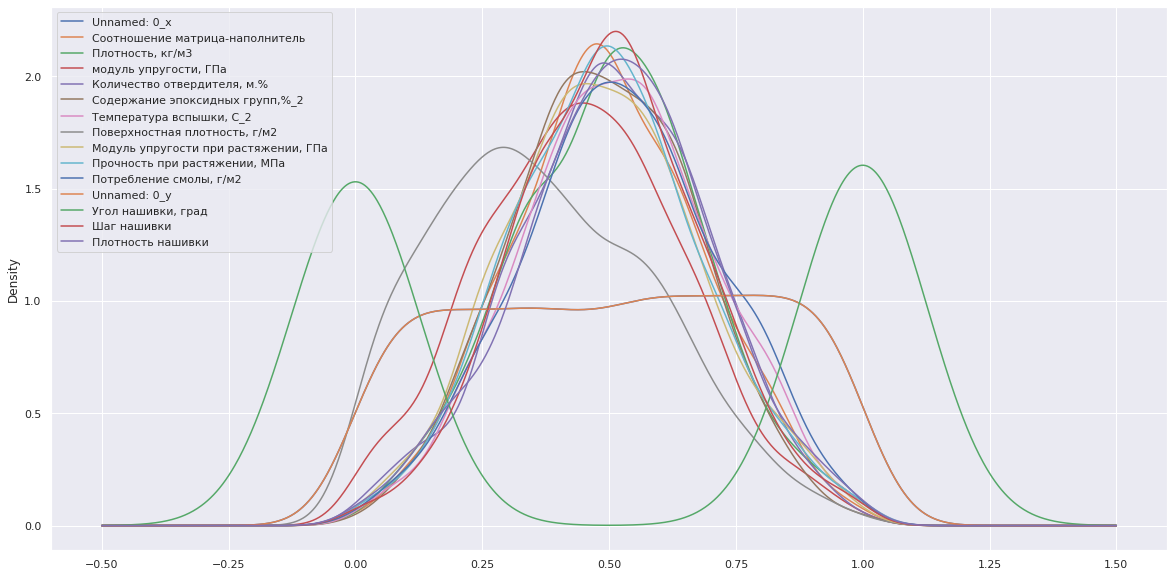


Рисунок 11 - Оценка распределения после нормализации данных.

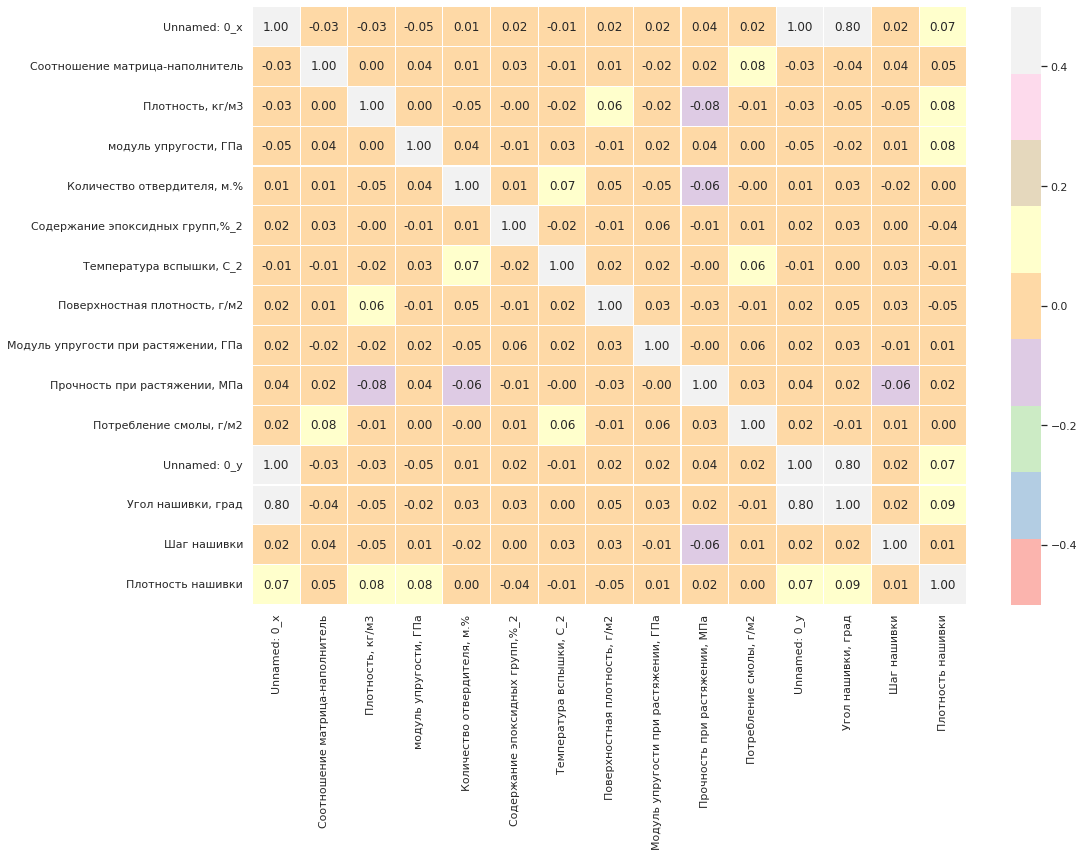


Рисунок 11 - Оценка корреляции после нормализации

## 2.2 Разработка и обучение моделей

Разработка и обучение моделей машинного обучения осуществлялась для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении» отдельно. Для решения применим все методы, описанные выше.

Обозначенная в работе задача в рамках классификации категорий машинного обучения относится к машинному обучению с учителем и является это задачей регрессии. Цель же любого алгоритма обучения с учителем — определение функции потерь и минимизация. Потому для наилучшего решения в процессе исследования были применены следующие методы:

метод опорных векторов;

случайный лес;

линейная регрессия;

k-ближайших соседей;

Линейная регрессия (Linear regression) — это алгоритм машинного обучения, основанный на контролируемом обучении, рассматривающий зависимость между одной входной и выходными переменными. Это один из самых простых и эффективных инструментов статистического моделирования. Она определяет зависимость переменных с помощью линии наилучшего соответствия. Модель регрессии создаёт несколько метрик. R-квадрат, или коэффициент детерминации, позволяет измерить, насколько модель может объяснить дисперсию данных. Если R-квадрат равен 1, это значит, что модель описывает все данные. Если же R-квадрат равен 0,5, модель объясняет лишь 50 процентов дисперсии данных. Оставшиеся отклонения не имеют объяснения. Чем ближе R-квадрат к единице, тем лучше.

Достоинства метода: быстр и прост в реализации; легко интерпретируем; имеет меньшую сложность по сравнению с другими алгоритмами;

Недостатки метода: моделирует только прямые линейные зависимости; требует наличия прямой связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние, а границы линейны.

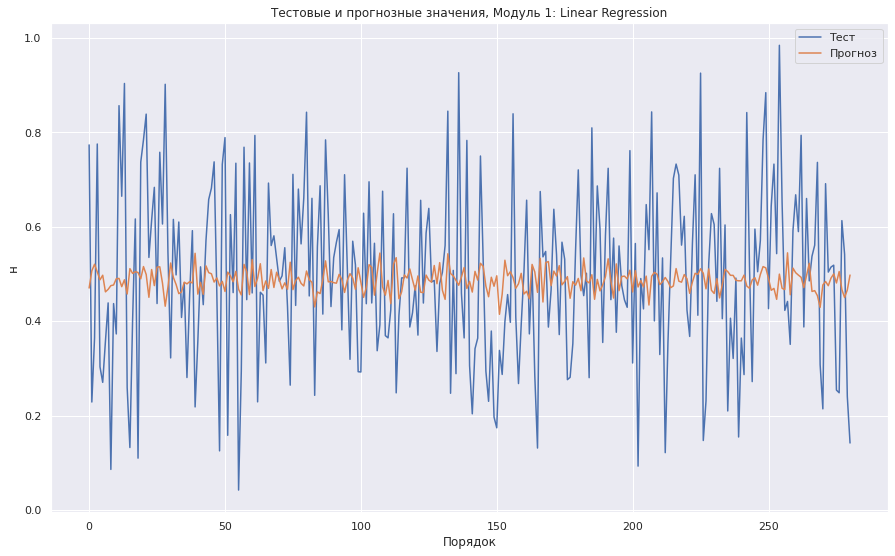


Рисунок 12 - График линейной регрессии для прочности при растяжении, МПа

Метод ближайших соседей - k-ближайших соседей (kNN - k Nearest Neighbours) ищет ближайшие объекты с известными значения целевой переменной и основывается на хранении данных в памяти для сравнения с новыми элементами. Алгоритм находит расстояния между запросом и всеми примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, затем голосует за наиболее часто встречающуюся метку (в случае задачи классификации) или усредняет метки (в случае задачи регрессии).

Достоинства метода: прост в реализации и понимании полученных результатов; имеет низкую чувствительность к выбросам; не требует построения модели; допускает настройку нескольких параметров; позволяет делать дополнительные допущения; универсален; находит лучшее решение из возможных; решает задачи небольшой размерности.

Недостатки метода: замедляется с ростом объёма данных; не создаёт правил; не обобщает предыдущий опыт; основывается на всем массиве доступных исторических данных; невозможно сказать, на каком основании строятся ответы; сложно выбрать близость метрики; имеет высокую зависимость результатов классификации от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоёмкость.



Рисунок 13 - График k-ближайших соседей для прочности при растяжении, Мпа

Метод опорных векторов (Support Vector Regression) – этот бинарный линейный классификатор был выбран, потому что он хорошо работает на небольших датасетах. Данный алгоритм – это алгоритм обучения с учителем, использующихся для задач классификации и регрессионного анализа, это контролируемое обучение моделей с использование схожих алгоритмов для анализа данных и распознавания шаблонов. Учитывая обучающую выборку, где алгоритм помечает каждый объект, как принадлежащий к одной из двух категорий, строит модель, которая определяет новые наблюдения в одну из категорий.

Модель метода опорных векторов – отображение данных точками в пространстве, так что между наблюдениями отдельных категорий имеется разрыв, и он максимален.

Каждый объект данных представляется как вектор (точка) в p-мерном пространстве. Он создаёт линию или гиперплоскость, которая разделяет данные на классы.

Достоинства метода: для классификации достаточно небольшого набора данных. При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных. Эффективен при большом количестве гиперпараметров. Способен обрабатывать случаи, когда гиперпараметров больше, чем количество наблюдений. Существует возможность гибко настраивать разделяющую функцию. Алгоритм максимизирует разделяющую полосу, которая, как подушка безопасности, позволяет уменьшить количество ошибок классификации.

Недостатки метода: неустойчивость к шуму, поэтому в применении требует проведения тщательной работы с выбросами, иначе в обучающих данных шумы становятся опорными объектами-нарушителями и напрямую влияют на построение разделяющей гиперплоскости; для больших наборов данных требуется долгое время обучения; достаточно сложно подбирать полезные преобразования данных; параметры модели сложно интерпретировать, поэтому представляется рациональным использование данного метода только наряду с другими.

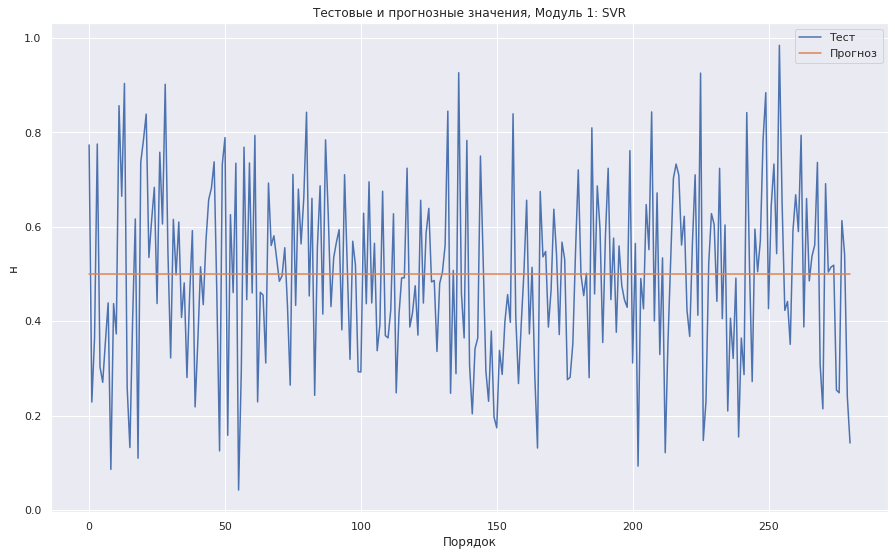


Рисунок 14 - график метода опорных векторов для прочности при растяжении, МПа.

Случайный лес (Random Forest) — это множество решающих деревьев. Универсальный алгоритм машинного обучения с учителем, представитель ансамблевых методов. Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив.

Достоинства метода: не переобучается; не требует предобработки входных данных; эффективно обрабатывает пропущенные данные, данные с большим числом классов и признаков; имеет высокую точность предсказания и внутреннюю оценку обобщающей способности модели, а также высокую параллелизуемость и масштабируемость.

Недостатки метода: построение занимает много времени; сложно интерпретируемый; не обладает возможностью экстраполяции; может недо- обучаться; трудоёмко прогнозируемый; иногда работает хуже, чем линейные методы.

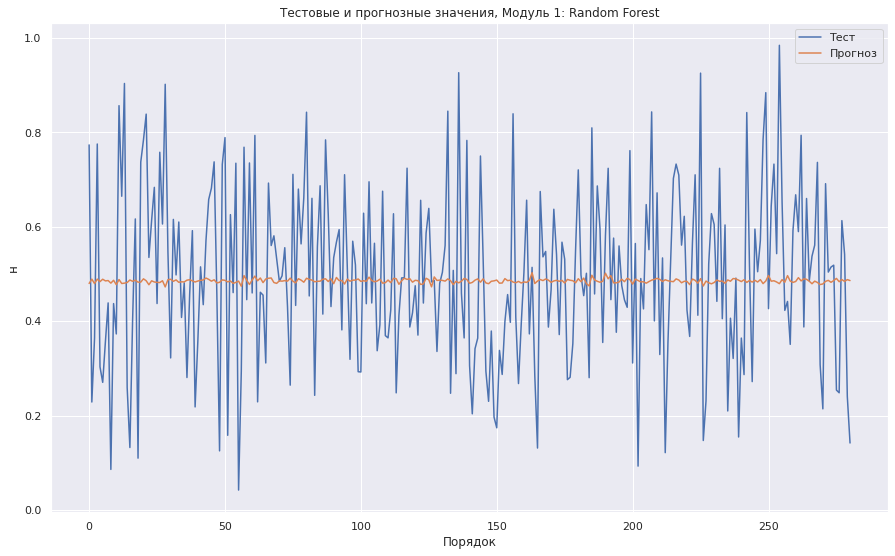


Рисунок 15 - График "Random Forest" для прочности при растяжении, МПа.

Немного об используемых метриках качества моделей: R2 или коэффициент детерминации измеряет долю дисперсии, объяснённую моделью, в общей дисперсии целевой переменной.

| **index** | **model** | **target** | **MSE** | **R2** |
| --- | --- | --- | --- | --- |
| **0** | Linear Regression | Модуль упругости и Прочность | 0.03511176833132294 | -0.02562122717298332 |
| **1** | KNeighborsRegressor | Модуль упругости и Прочность | 0.034938154217849154 | -0.020566741880638506 |
| **2** | SVR | Модуль упругости и Прочность | 0.03440306889238362 | -0.004749605079926722 |
| **3** | Random Forest | Модуль упругости и Прочность | 0.03473766906469722 | -0.01461417831252021 |

Рисунок 16 – Часть кода loss\_df.

Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то качество прогноза идентично средней величине целевой переменной (т.е. очень низкое). Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

MSE (Mean Squared Error) или средняя квадратичная ошибка принимает значениях в тех же единицах, что и целевая переменная. Чем ближе к нулю MSE, тем лучше работают предсказательные качества модели.

## 2.3 Рекомендательные нейросети для соотношения «матрица – наполнитель»

Обучение нейронной сети — это процесс, при котором происходит подбор оптимальных параметров модели, с точки зрения минимизации функционала ошибки. Построение нейросетей производится с помощью класса keras.Sequential.

Порядок разработки модели для каждого параметра и для каждого выбранного метода можно разделить на следующие этапы: разделение нормализованных данных на обучающую и тестовую выборки (в соотношении 80 на 20%); проверка моделей при стандартных значениях; сравнение с результатами модели, выдающей среднее значение; создание графика; сравнение моделей по метрике МАЕ; поиск сетки гиперпараметров, по которым будет происходить оптимизация модели. В качестве параметра оценки выбран коэффициент детерминации (R2); оптимизация подбора гиперпараметров модели с помощью выбора по сетке и перекрёстной проверки; подстановка оптимальных гиперпараметров в модель и обучение модели на тренировочных данных; оценка полученных данных; сравнение со стандартными значениями.

Построение первой нейросети.

Обучение и оценка модели, Задание функции для визуализации факт/прогноз для результатов модели.

model = Sequential()

model.add(layers.Dense(64, input\_dim=X.shape[1], activation='tanh'))

model.add(layers.Dense(64, activation='tanh'))

model.add(layers.Dense(32, activation='sigmoid'))

model.add(layers.Dense(1))

model.summary()

dfmodel = model.compile(optimizer='adam', loss='mae', metrics=['mae'])

history = model.fit(

X\_train,

y\_train,

validation\_split=0.2,

verbose=1, epochs=10)

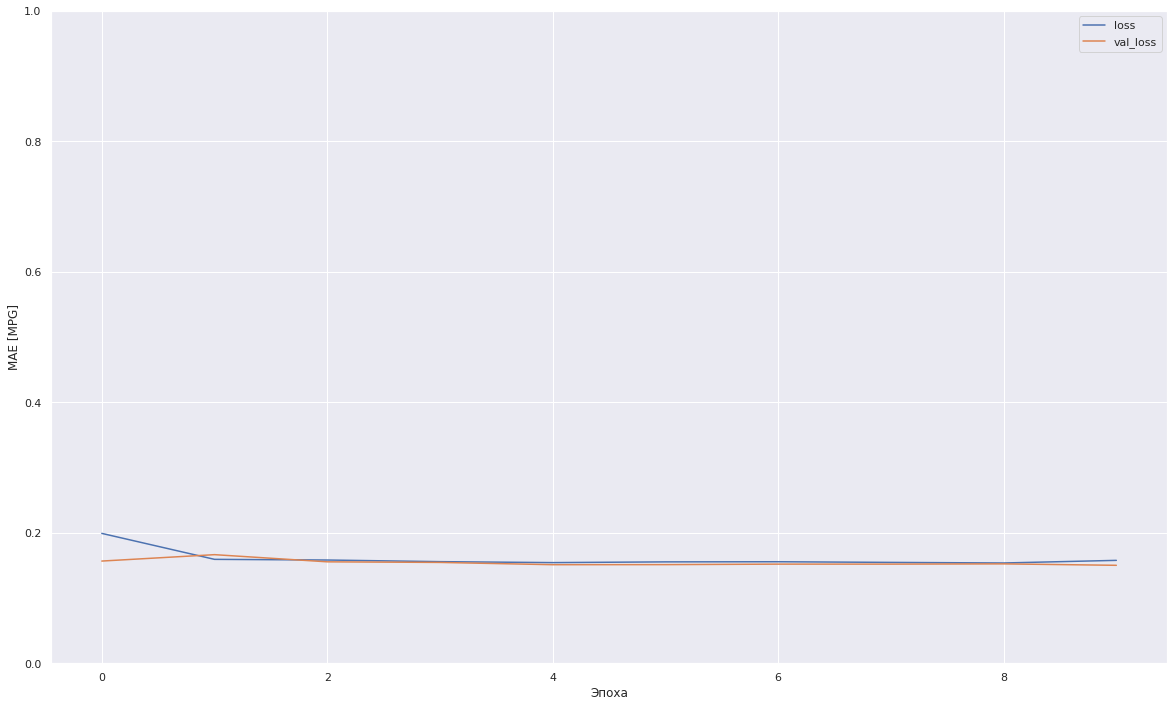


Рисунок 17 – Нейросеть №1, график потерь первой модели

Построение второй нейросети

Обучение и оценка модели, Оценка потерь, задание функции для визуализации факт/прогноз для результатов модели.

model = Sequential()

model.add(layers.Dense(16, input\_dim=X.shape[1], activation='elu'))

model.add(layers.Dense(32, activation='tanh'))

model.add(layers.Dense(8, activation='sigmoid'))

model.add(layers.Dense(1))

model.summary()

dfmodel = model.compile(optimizer='adam', loss='mae', metrics=['mae'])

history = model.fit(

X\_train,

y\_train,

validation\_split=0.2,

verbose=1, epochs=10)

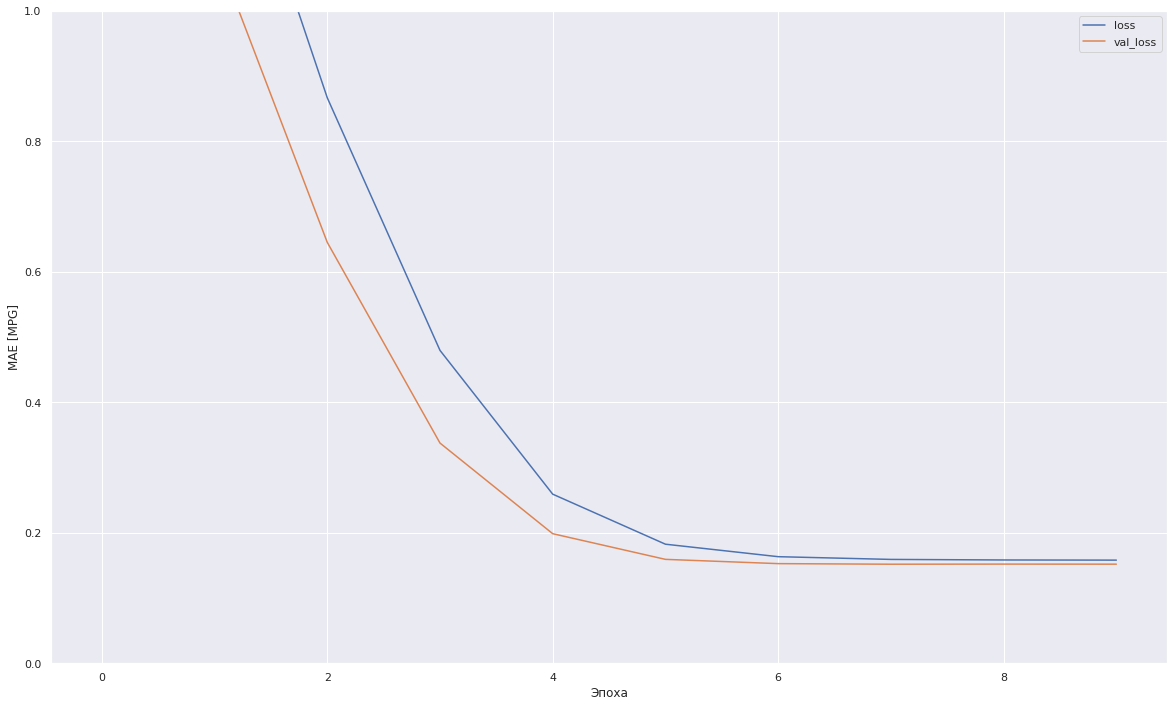


Рисунок 18 – Нейросеть №2, график потерь второй модели

Построение третьей нейросети

Обучение и оценка модели, Задание функции для визуализации факт/прогноз для результатов модели.

model = Sequential()

model.add(layers.Dense(16, input\_dim=X.shape[1], activation='relu'))

model.add(layers.Dense(32, activation='tanh'))

model.add(layers.Dense(8, activation='sigmoid'))

model.add(layers.Dense(1))

model.summary()

dfmodel = model.compile(optimizer='adam', loss='mae', metrics=['mae'])

history = model.fit(

X\_train,

y\_train,

validation\_split=0.2,

verbose=1, epochs=6)

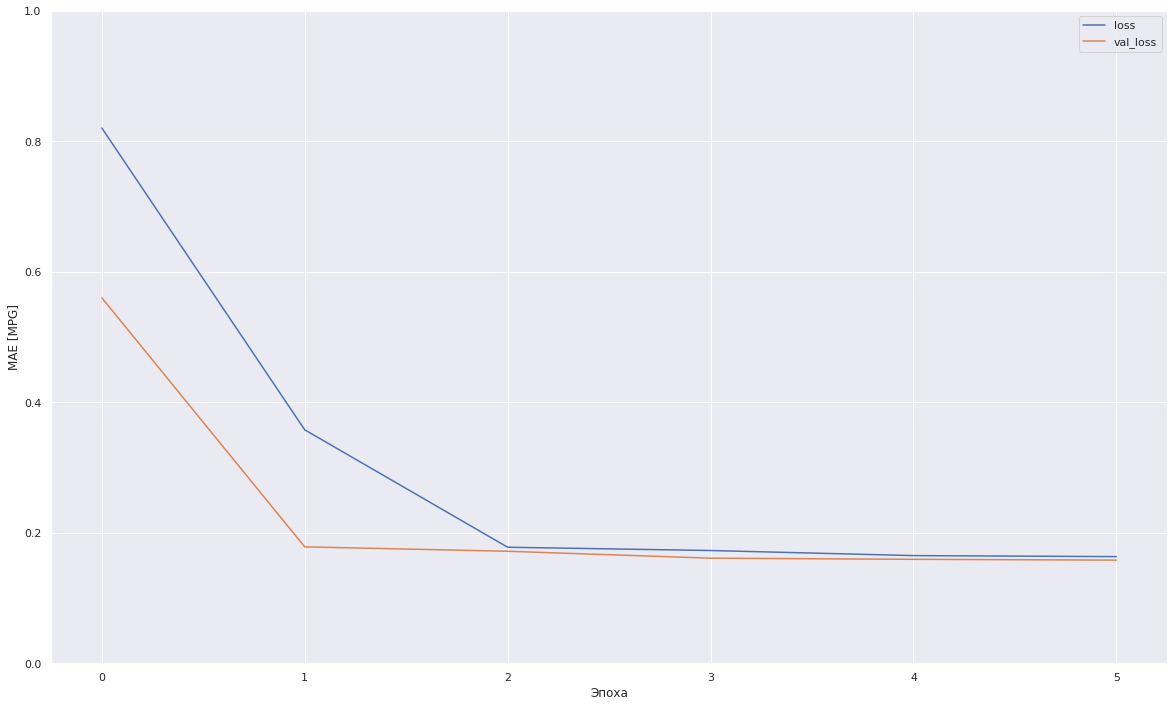


Рисунок 19 – Нейросеть №3, график потерь третьей модели.

Построение четвертой нейросети

Обучение и оценка модели, Оценка потерь, Задание функции для визуализации факт/прогноз для результатов модели.

model = Sequential()

model.add(layers.Dense(16, input\_dim=X.shape[1], activation='tanh'))

model.add(layers.Dense(8, activation='sigmoid'))

model.add(layers.Dense(1))

model.summary()

dfmodel = model.compile(optimizer='adam', loss='mae', metrics=['mae'])

history = model.fit(

X\_train,

y\_train,

validation\_split=0.2,

verbose=1, epochs=10)

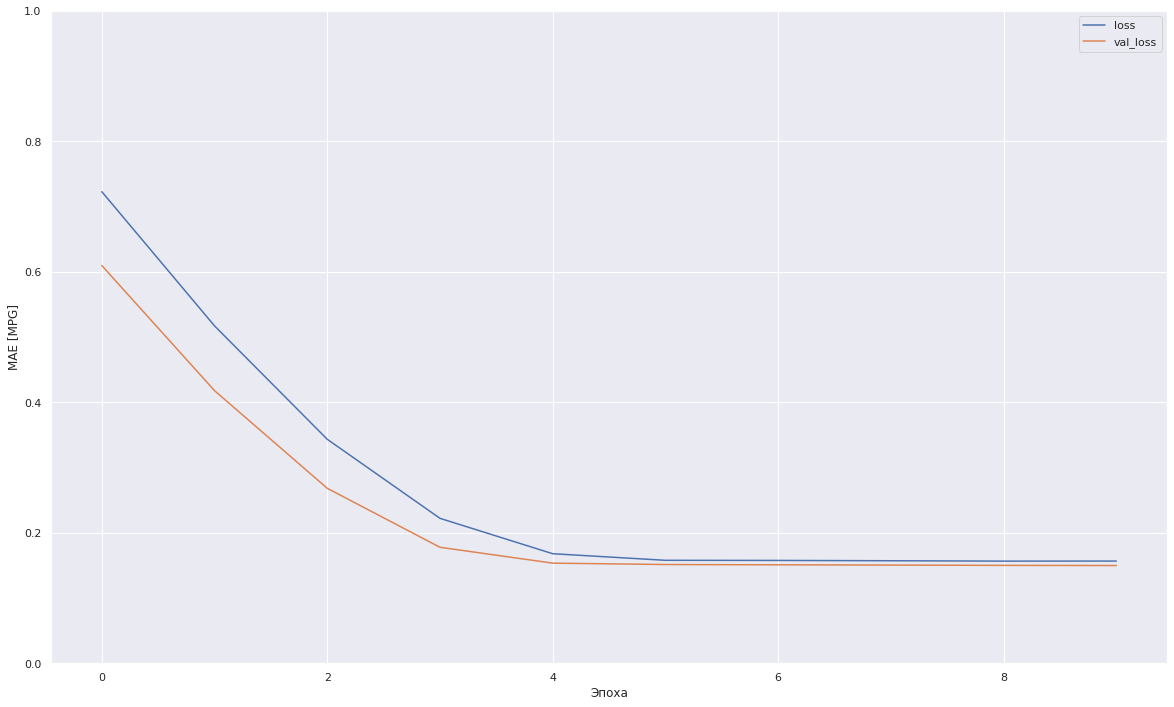


Рисунок 20 – Нейросеть №4, график потерь четвертой модели

Прочность при растяжении и модуль упругости не имеет линейной зависимости. Все использованные модели не справились с задачей. Результат неудовлетворительный. Свойства композитных материалов в первую очередь зависят от используемых материалов.

После обучения моделей была проведена оценка точности этих моделей на обучающей и тестовых выборках. В качестве параметра оценки модели использовалась средняя квадратическая ошибка (MSE). Результат неудовлетворительный.

# 3. Разработка приложения

from joblib import dump, load

def input\_variable():

x1 = float(input('Введите значение переменной Соотношение матрица-наполнитель: '))

x2 = float(input('Введите значение переменной Плотность: '))

x3 = float(input('Введите значение переменной Модуль упругости: '))

x4 = float(input('Введите значение переменной Количество отвердителя: '))

x5 = float(input('Введите значение переменной Содержание эпоксидных групп: '))

x6 = float(input('Введите значение переменной Температура вспышки: '))

x7 = float(input('Введите значение переменной Модуль упругости при растяжении: '))

x8 = float(input('Введите значение переменной Потребление смолы: '))

x9 = float(input('Введите значение переменной Угол нашивки: '))

x10 = float(input('Введите значение переменной Шаг нашивки: '))

x11 = float(input('Введите значение переменной Плотность нашивки: '))

return x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11

def input\_proc(X):

print('вызов модели')

res = model\_l.predict(X)

print(res)

return(res)

def app\_model():

model\_l = load('filename.joblib')

print('Приложение прогнозирует значения модулей упругости и растяжения')

for i in range(110):

try:

print('введите 1 для прогноза, 2 для выхода')

check = input()

if check == '1':

print('Введите данные')

input\_proc(input\_variable())

elif check == '2':

break

else:

print('Повторите выбор')

except:

print('Неверные данные. Повторите операцию')

app\_model()

Рисунок 21. Пользовательское приложение для определения соотношения “матрица - наполнитель”.

# 4. Создание удаленного репозитория и загрузка

Репозиторий был создан на github.com по адресу: <https://github.com/Avelaksom/Baum_VKR_Composite>

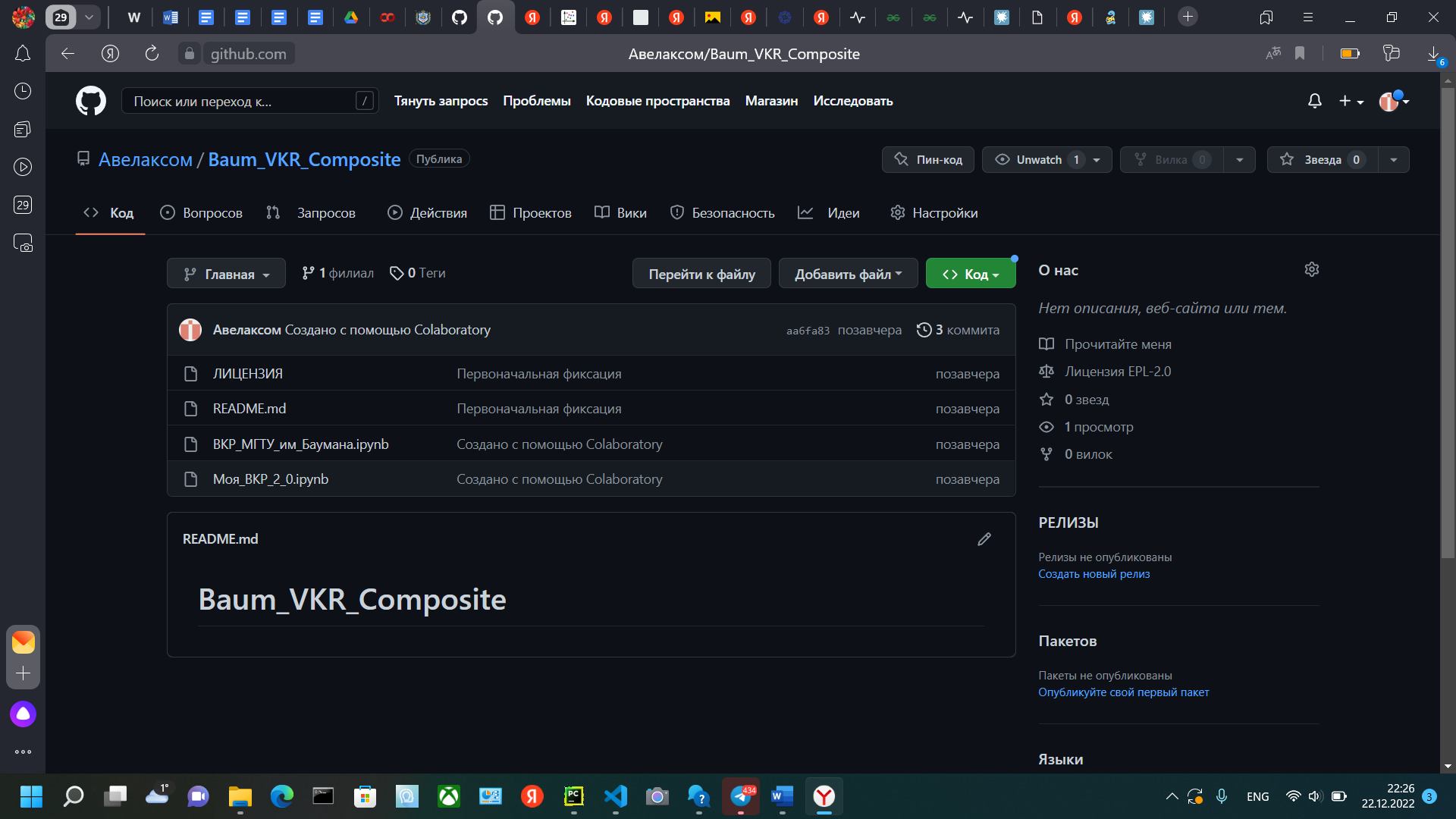


Рисунок 22 - скриншот страницы на github.com

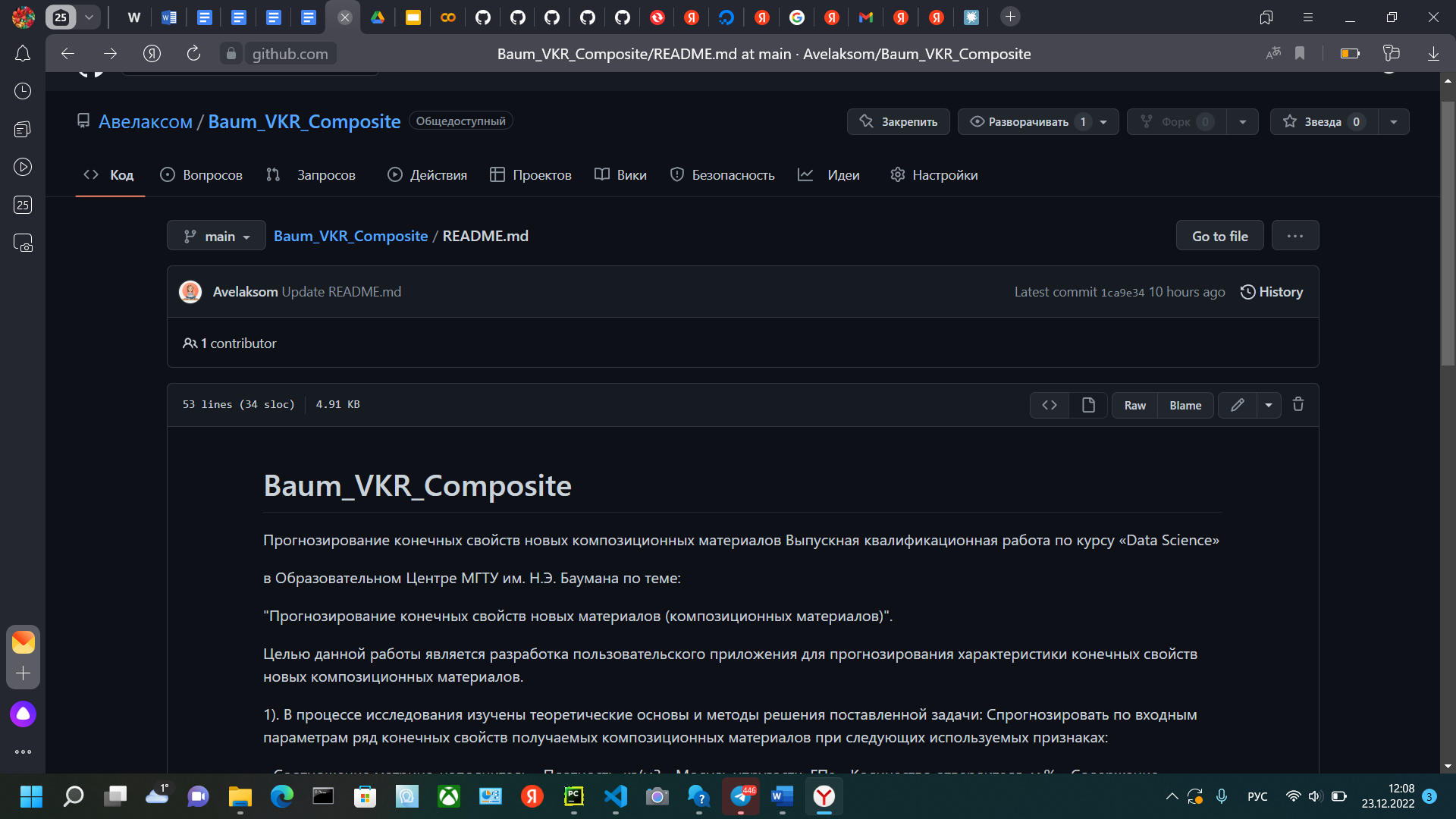


Рисунок 23 - Часть созданного файла README

Ноутбук с решением и приложением так же можно найти по адресу: <https://colab.research.google.com/drive/1hft3Y5tTJJoESA_qACe9p5HVduS_a03X?usp=sharing>

# 5. Заключение

Данная исследовательская работа позволяет сделать некоторые основные выводы по теме. Распределение полученных данных в объединённом датасете близко к нормальному, но коэффициенты корреляции между парами признаков стремятся к нулю. Использованные при разработке моделей подходы не позволили получить сколько-нибудь достоверных прогнозов. Применённые модели регрессии не показали высокой эффективности в прогнозировании свойств композитов.

Был сделан вывод, что невозможно определить из свойств материалов соотношение «матрица – наполнитель». Данный факт не указывает на то, что прогнозирование характеристик композитных материалов на основании предоставленного набора данных невозможно, но может указывать на недостатки базы данных, подходов, использованных при прогнозе, необходимости пересмотра инструментов для прогнозирования.

Необходимы дополнительные вводные данные, получение новых результирующих признаков в результате математических преобразований, релевантных доменной области, консультации экспертов предметной области, новые исследования, работа эффективной команды, состоящей из различных специалистов.

В целом прогнозирование конечных свойств/характеристик композитных материалов без изучения материаловедения, погружения в вопрос экспериментального анализа характеристик композитных материалов не демонстрирует сколько-нибудь удовлетворительных результатов. Проработка моделей и построение прогнозов требует внедрения в процесс производных от имеющихся показателей для выявления иного уровня взаимосвязей. Отсюда, также учитывая отсутствие корреляции между признаками, делаем вывод, что текущим набором алгоритмов задача не решается, возможно, решается трудно или не решается совсем.

# 6. Список используемой литературы и веб ресурсы

1. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour): – Режим доступа: https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19. (дата обращения: 07.12.2022)

2. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам сле-дует знать: – Режим доступа: https://habr.com/ru/company/vk/blog/513842/(дата обращения: 01.12.2022).

3. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.

4. Абросимов Н.А.: Методика построения разрешающей системы уравнений динамического деформирования композитных элементов конструкций (Учебно-методическое пособие), ННГУ, 2010

5. Абу-Хасан Махмуд, Масленникова Л. Л.: Прогнозирование свойств композиционных материалов с учётом наноразмера частиц и акцепторных свойств катионов твёрдых фаз, статья 2006 год

6. Бизли Д. Python. Подробный справочник: учебное пособие. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.

7. Гафаров, Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения: учеб. пособие /Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Издательство Казанского университета, 2018. – 121 с.

8. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.

9. Документация по библиотеке keras: – Режим доступа: https://keras.io/api/.(дата обращения: 08.12.2022).

10. Документация по библиотеке matplotlib: – Режим доступа: https://matplotlib.org/stable/users/index.html. (дата обращения: 10.12.2022)

11. Документация по библиотеке numpy: – Режим доступа: https://numpy.org/doc/1.22/user/index.html#user. (дата обращения: 03.12.2022).

12. Документация по библиотеке pandas: – Режим доступа: https://pandas.pydata.org/docs/user\_guide/index.html#user-guide. (дата обращения: 04.12.2022).

13. Документация по библиотеке scikit-learn: – Режим доступа: https://scikit-learn.org/stable/user\_guide.html. (дата обращения: 05.12.2022).

14. Документация по библиотеке seaborn: – Режим доступа: https://seaborn.pydata.org/tutorial.html. (дата обращения: 06.12.2022).

15. Документация по библиотеке Tensorflow: – Режим доступа: https://www.tensorflow.org/overview (дата обращения: 10.12.2022).

16. Документация по языку программирования python: – Режим доступа: https://docs.python.org/3.8/index.html. (дата обращения: 02.12.2022).

17. Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.

18. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) – Режим доступа: https://habr.com/ru/post/428503/ (дата обращения 07.12.2022)

19. Ларин А. А., Способы оценки работоспособности изделий из композиционных материалов методом компьютерной томографии, Москва, 2013, 148 с.

20. Материалы конференции: V Всероссийская научно-техническая конференция «Полимерные композиционные материалы и производственные технологии нового поколения», 19 ноября 2021 г.

21. Миронов А.А. Машинное обучение часть I ст.9 – Режим доступа: http://is.ifmo.ru/verification/machine-learning-mironov.pdf. (дата обращения 08.12.2022)

22. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с.

23. Реутов Ю.А.: Прогнозирование свойств полимерных композиционных материалов и оценка надёжности изделий из них, Диссертация на соискание учёной степени кандидата физико-математических наук, Томск 2016.

24. Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.: Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил.

25. Руководство по быстрому старту в flask: – Режим доступа: https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html. (дата обращения: 09.06.2022)

26. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.

27. Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.

28. Справочник по композиционным материалам: в 2 - х кн. Кн. 2 / Под ред. Дж. Любина; Пер. с англ. Ф. Б. Геллера, M. М. Гельмонта; Под ред. Б. Э. Геллера - М.: Машиностроение, 1988. - 488 с. : ил;

29. Траск Эндрю. Грокаем глубокое обучение. – СПб.: Питер, 2019. – 352 с.: ил.