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 Маслова Ольга Владимировна

Москва, 2022

Содержание

[Введение 3](#_Toc106369284)

[1 Аналитическая часть 4](#_Toc106369285)

[1.1. Постановка задачи 4](#_Toc106369286)

[1.2. Описание используемых методов 6](#_Toc106369287)

[1.3. Разведочный анализ данных 9](#_Toc106369288)

[2 Практическая часть 15](#_Toc106369289)

[2.1. Предобработка данных 16](#_Toc106369290)

[2.2 Прогнозирование модуля прочности при растяжении 17](#_Toc106369291)

[2.3 Прогнозирование модуля упругости при растяжении 21](#_Toc106369292)

[2.4 Разработка нейронной сети для прогнозирования соотношения матрица-наполнитель 25](#_Toc106369293)

[2.5 Создание удаленного репозитория 27](#_Toc106369294)

[Заключение 28](#_Toc106369295)

[Список использованной литературы 29](#_Toc106369296)

## Введение

Пояснительная записка подготовлена в рамках выпускной квалификационной работы по курсу «Data Science» на тему «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.

Современные композиты изготавливаются из разных материалов: полимеры, керамика, стеклянные и углеродные волокна, но основной 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).

## Аналитическая часть

## 1.1. Постановка задачи

Целью работы являетсядемонстрирование некоторых изученных методов работы анализа данных дляпрогнозирования конечных свойств новых 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. 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).

**На входе** имеются данные о начальных свойствах компонентов композиционных материалов:

* 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
* Шаг нашивки
* Плотность нашивки
* Угол нашивки, град

На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов:

* Соотношение матрица-наполнитель
* Модуль упругости при растяжении, Гпа
* Прочность при растяжении, МПа

Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

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

В данной выборке после объединения файлов получилось 1023 строки, 13 столбцов, проверка показала отсутствие пропусков, тип всех данных int, значит, ошибочного ввода значений также нет.

## 1.2. Описание используемых методов

Предсказание значений вещественной, непрерывной переменной — это задача регрессии. Эта зависимая переменная должна иметь связь с одной или несколькими независимыми переменными, называемых также предикторами или регрессорами. Регрессионный анализ помогает понять, как «типичное» значение зависимой переменной изменяется при изменении независимых переменных.

В настоящее время разработано много методов регрессионного анализа. Например, простая и множественная линейная регрессия. Эти модели являются параметрическими в том смысле, что функция регрессии определяется конечным числом неизвестных параметров, которые оцениваются на основе данных.

**Линейная регрессия**

Простая линейная регрессия имеет место, если рассматривается зависимость между одной входной и одной выходной переменными. Для этого определяется уравнение регрессии (1) и строится соответствующая прямая, известная как линия регрессии.

(1)

Коэффициенты a и b, называемые также параметрами модели, определяются таким образом, чтобы сумма квадратов отклонений точек, соответствующих реальным наблюдениям данных, от линии регрессии была бы минимальной. Коэффициенты обычно оцениваются методом наименьших квадратов.

Если ищется зависимость между несколькими входными и одной выходной переменными, то имеет место множественная линейная регрессия. Соответствующее уравнение имеет вид (2).

(2)

где n - число входных переменных.

Очевидно, что в данном случае модель будет описываться не прямой, а гиперплоскостью. Коэффициенты уравнения множественной линейной регрессии подбираются так, чтобы минимизировать сумму квадратов отклонения реальных точек данных от этой гиперплоскости.

Линейная регрессия — первый тщательно изученный метод регрессионного анализа. Его главное достоинство — простота. Такую модель можно построить и рассчитать даже без мощных вычислительных средств.

Простота является и главным недостатком этого метода. Тем не менее, именно с линейной регрессии целесообразно начать подбор подходящей модели.

На языке python линейная регрессия реализована в sklearn.linear\_model.LinearRegression.

**Метод k-ближайших соседей**

Еще один метод классификации, который адаптирован для регрессии - метод k-ближайших соседей (k Nearest Neighbors). На интуитивном уровне суть метода проста: посмотри на соседей вокруг, какие из них преобладают, таковым ты и являешься.

В случае использования метода для регрессии, объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны.

Для реализации метода необходима метрика расстояния между объектами. Используется, например, эвклидово расстояние для количественных признаков или расстояние Хэмминга для категориальных. Этот метод — пример непараметрической регрессии.

Метод реализован в sklearn.neighbors.KneighborsRegressor.

**Случайный лес**

Случайный лес (RandomForest) — представитель ансамблевых методов.

Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив. Формула итогового решателя (3) — это усреднение предсказаний отдельных деревьев.

 (3),

где N – количество деревьев;

i – счетчик для деревьев;

b – решающее дерево;

x – сгенерированная нами на основе данных выборка.

Для определения входных данных каждому дереву используется метод случайных подпространств. Базовые алгоритмы обучаются на различных подмножествах признаков, которые выделяются случайным образом.

Преимущества случайного леса в его высокой точности предсказаний; также этот метод редко переобучается, практически не чувствителен к выбросам в данных, одинаково хорошо обрабатывает как непрерывные, так и дискретные признаки, данные с большим числом признаков. У данного метода высокая параллелизуемость и масштабируемость.

Из недостатков можно отметить, что его построение занимает больше времени. Так же теряется интерпретируемость.

Метод реализован в sklearn.ensemble.RandomForestRegressor.

**Нейронная сеть**

Нейронная сеть — это последовательность нейронов, соединенных между собой связями. Структура нейронной сети пришла в мир программирования из биологии. Вычислительная единица нейронной сети — нейрон или персептрон.

У каждого нейрона есть определённое количество входов, куда поступают сигналы, которые суммируются с учётом значимости (веса) каждого входа. Смещение – это дополнительный вход для нейрона, который всегда равен 1 и, следовательно, имеет собственный вес соединения.

Так же у нейрона есть функция активации, которая определяет выходное значение нейрона. Она используется для того, чтобы ввести нелинейность в нейронную сеть. Примеры активационных функций: relu, сигмоида.

У полносвязной нейросети выход каждого нейрона подается на вход всем нейронам следующего слоя. У нейросети имеется входной слой — его размер соответствует входным параметрам; скрытые слои — их количество и размерность определяем специалист; выходной слой — его размер соответствует выходным параметрам.

Прямое распространение – это процесс передачи входных значений в нейронную сеть и получения выходных данных, которые называются прогнозируемым значением.

Прогнозируемое значение сравниваем с фактическим с помощью функции потери. В методе обратного распространения ошибки градиенты (производные значений ошибок) вычисляются по значениям весов в направлении, обратном прямому распространению сигналов. Значение градиента вычитают из значения веса, чтобы уменьшить значение ошибки. Таким образом происходит процесс обучения. Обновляются веса каждого соединения, чтобы функция потерь минимизировалась.

Для обновления весов в модели используются различные оптимизаторы.

Количество эпох показывает, сколько раз выполнялся проход для всех примеров обучения.

Нейронные сети применяются для решения задач регрессии, классификации, распознавания образов и речи, компьютерного зрения и других. На настоящий момент это самый мощный, гибкий и широко применяемый инструмент в машинном обучении.

## 1.3. Разведочный анализ данных

**Описательная статистика**

Описательная статистика – это краткие описательные коэффициенты, которые суммируют данный набор данных, который может быть либо представлением всей совокупности, либо выборкой совокупности. Описательная статистика подразделяется на меры центральной тенденции и меры изменчивости (распространения). Меры центральной тенденции включают среднее, медиану и моду, в то время как меры изменчивости включают стандартное отклонение, дисперсию, минимальные и максимальные переменные, эксцесс и асимметрию.

Описательная статистика датасета представлена в таблице 1. Она в численном виде отражает то, что мы увидим ниже на гистограммах, демонстрирует среднее, медианное значения для каждого параметра, показывает распределение данных по квартилям.

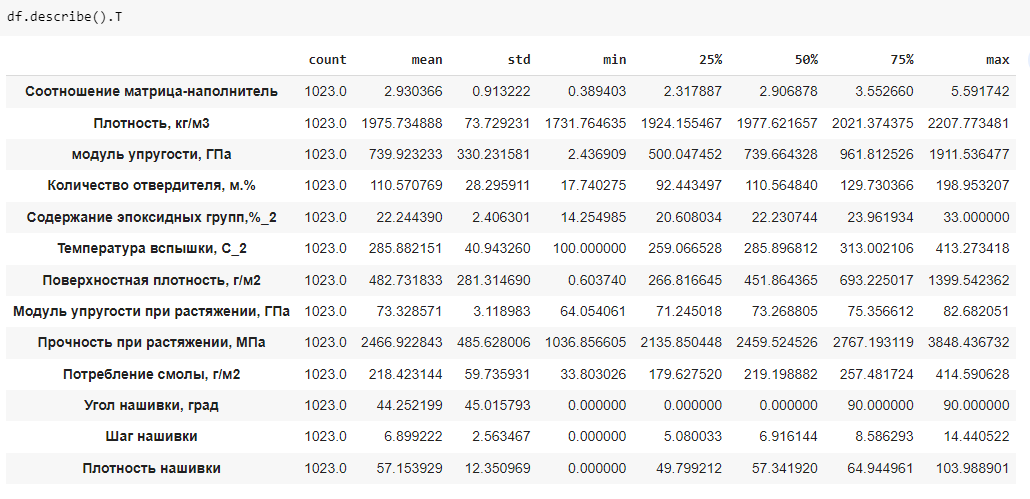


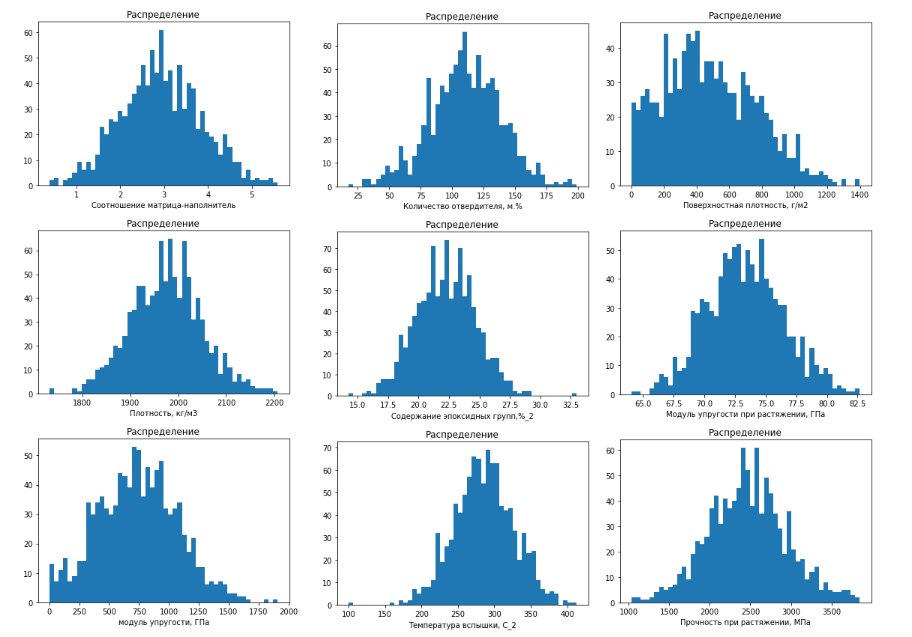
Табл. 1 Описательная статистика датасета

**Гистограмма плотности распределения**

Гистограмма плотности распределения — это столбиковая диаграмма, которая показывает, как данные распределяются по группам значений. Собранные данные представляют в виде ряда прямоугольников, одинаковых по ширине и различающихся по высоте. Анализ характера изменения высот позволяет оценить динамику процесса.

Гистограмму плотности распределения используют, чтобы наглядно показать, в каком интервале располагаются наиболее часто встречающиеся значения и как вообще распределяются данные. Кроме того, эта гистограмма позволяет определить наилучшие результаты процесса. Такое графическое изображение динамики процесса дает возможность наметить приоритетные задачи по его улучшению.

Гистограммы распределения переменных приведены на рисунке 1. По ним видно, что все признаки, кроме «Угол нашивки», имеют распределение, близкое к нормальному, и принимают неотрицательные значения. «Угол нашивки» принимает значения 0, 90.



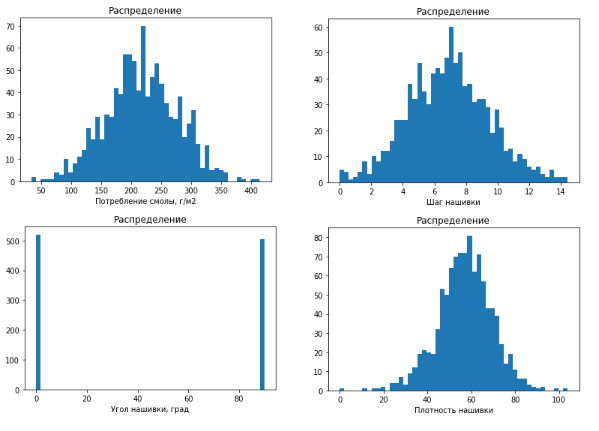


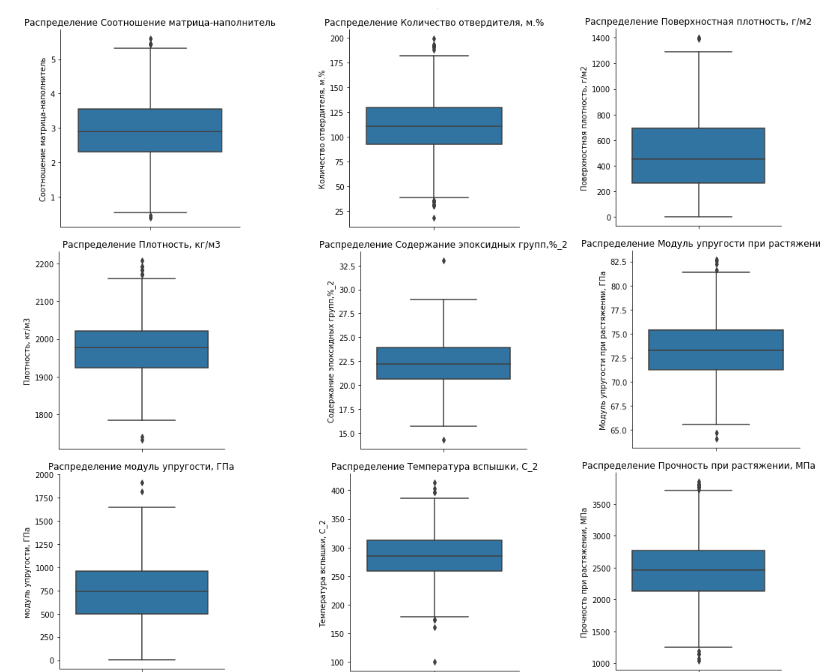
Рис. 1 Гистограммы распределения переменных

**Диаграммы размаха**

«Ящик с усами», диаграмма размаха — график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.

Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы. Несколько таких ящиков можно нарисовать бок о бок, чтобы визуально сравнивать одно распределение с другим; их можно располагать как горизонтально, так и вертикально. Расстояния между различными частями ящика позволяют определить степень разброса (дисперсии) и асимметрии данных и выявить выбросы.

Диаграммы «ящик с усами» приведены на рисунке 2. Данные диаграммы демонстрируют некоторое наличие выбросов по переменным.



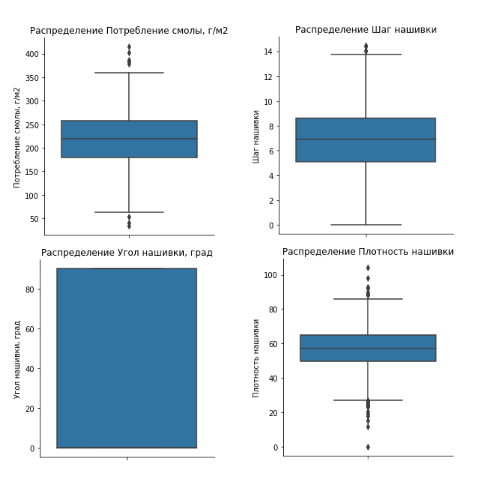


Рис. 2 Диаграммы размаха

**Тепловая карта корреляций**

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.

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

Тепловая карта корреляций приведена на рисунке 3. По ней видно, что в нашем наборе данных нет очевидных зависимостей или они очень слабо выражены.

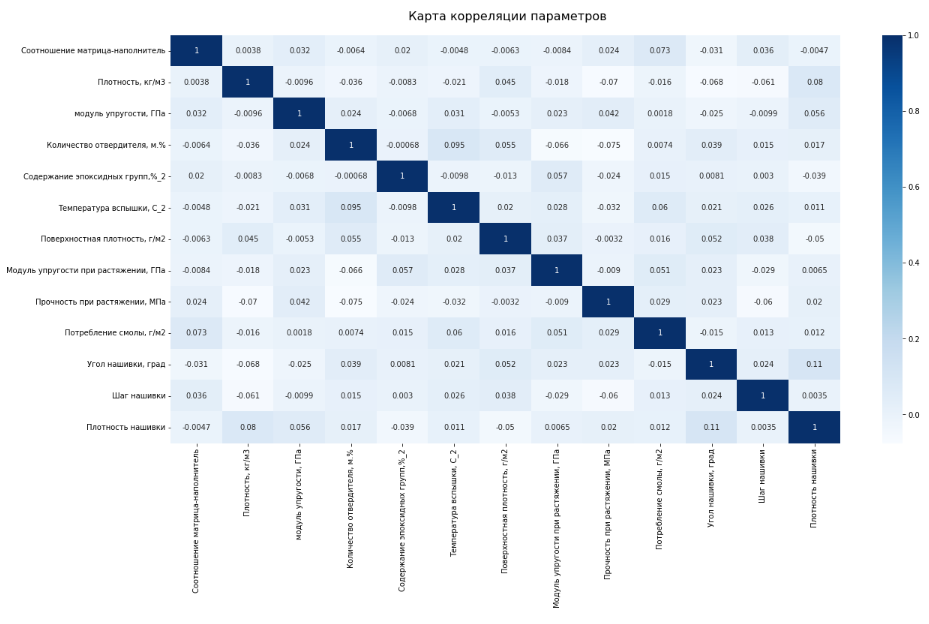
**

Рис. 3 Тепловая карта корреляций

**Попарные диаграммы рассеивания**

Диаграмма рассеивания позволяет оценить зависимости между парами соответствующих элементов. Этим методом можно без математической обработки данных установить зависимости по графическому представлению.

Попарные графики рассеяния точек приведены на рисунке 4. По графикам рассеяния мы видим, что некоторые точки отстоят далеко от общего облака. Так визуально выглядят выбросы — аномальные, некорректные значения данных, выходящие за пределы допустимых значений признака.

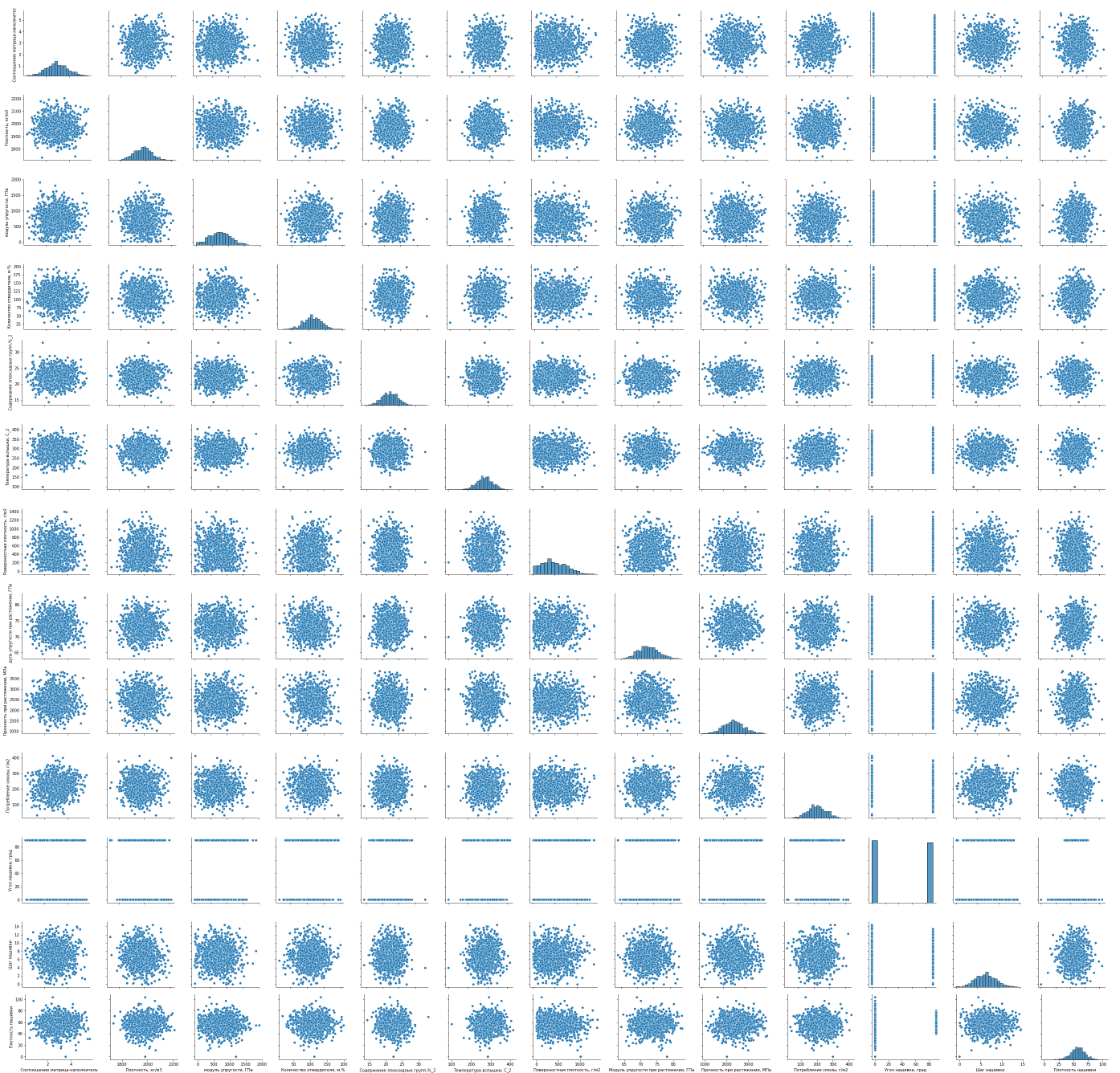


Рис. 4 Попарные диаграммы рассеивания

## 2 Практическая часть

Существует множество различных метрик качества, применимых для анализа моделей. В этой работе были использованы следующие методы.

или коэффициент детерминации – измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то прогнозы сопоставимы по качеству с константным предсказанием.

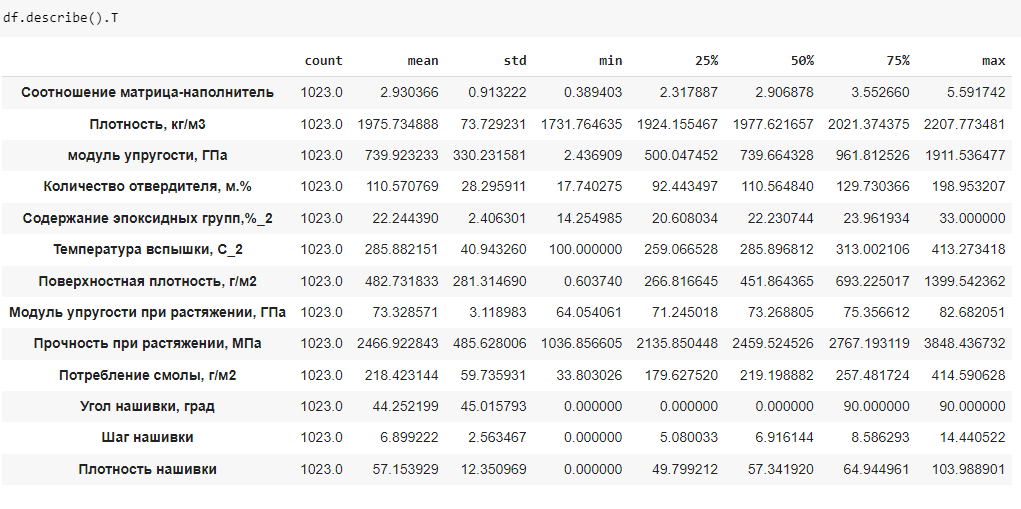
MAE (Mean Absolute Error) - средняя абсолютная ошибка так же принимает значения в тех же единицах, что и целевая переменная.

MSE – это средний квадрат ошибки определения какой-либо величины, то есть среднеквадратичная ошибка.

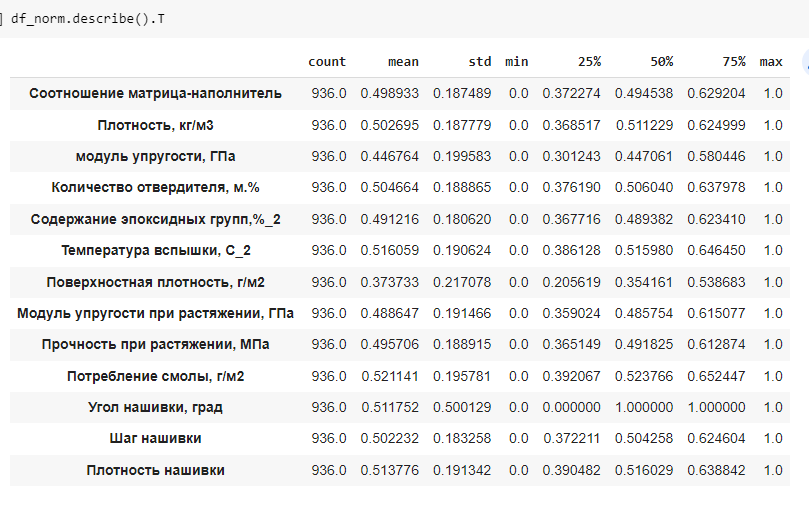
## 2.1. Предобработка данных

Предоставленные заказчиком данные для выпускной квалификационной работы были предварительно предобработаны. Но так как исходные значения признаков могут изменяться в очень большом диапазоне и отличаться друг от друга на несколько порядков, а данные, будучи разными по физическому смыслу, сильно различаются между собой по абсолютным величинам, проведем нормализацию, а также удалим выбросы по правилу трех сигм. После нормализации все числовые значения входных признаков будут приведены к одинаковой области их изменения – некоторому узкому диапазону.

Датасет до нормализации (транспонированный) и удаления выбросов:



Датасет после нормализации (транспонированный) и удаления выбросов:

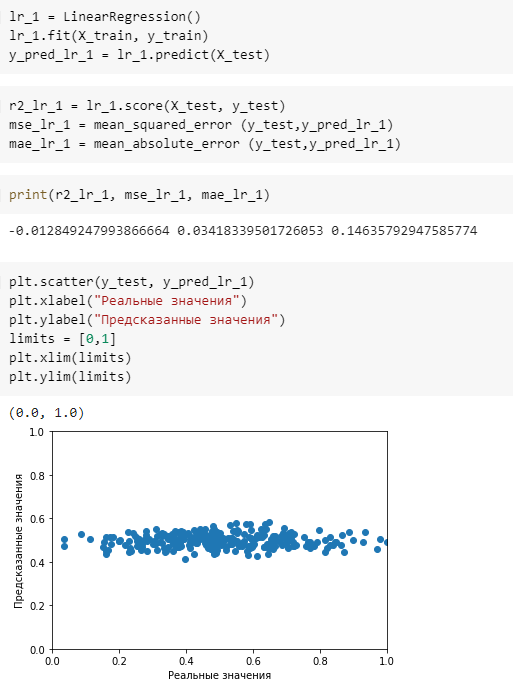


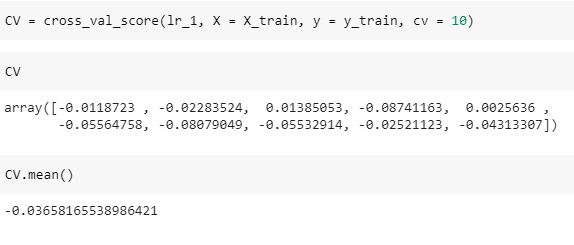
## 2.2 Прогнозирование модели прочности при растяжении

Для прогноза модели прочности при растяжении были использованы следующие модели.

**Модель 1. Линейная регрессия**

На вход подается нормализованный датасет.





Оценка линейная регрессия: -0,0128.

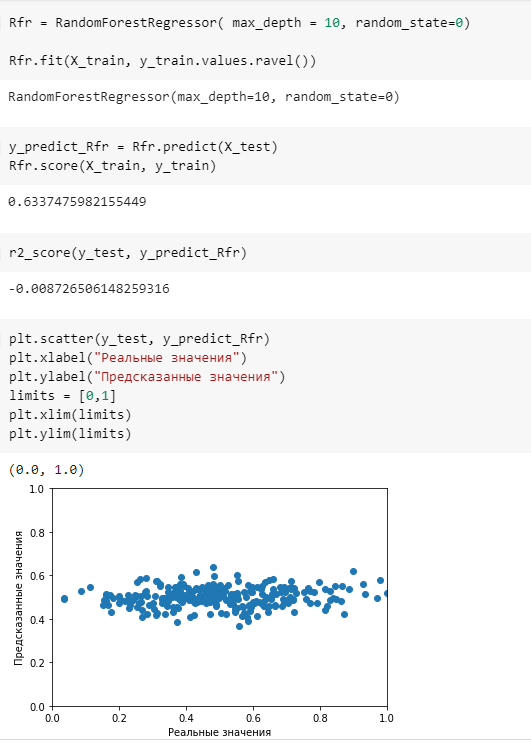
Средняя квадратическая ошибка MSE: 0,0341.

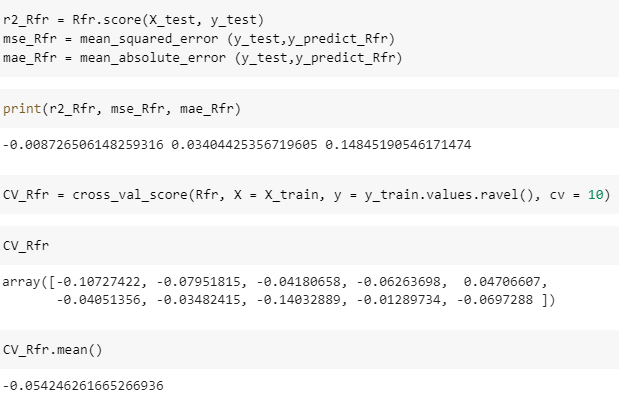
Средняя абсолютная ошибка MAE: 0,1463.

Вывод: линейная регрессия не показала зависимость.

**Модель 2. Случайный лес**

На вход подается нормализованный датасет.





Оценка линейная регрессия: -0,0087.

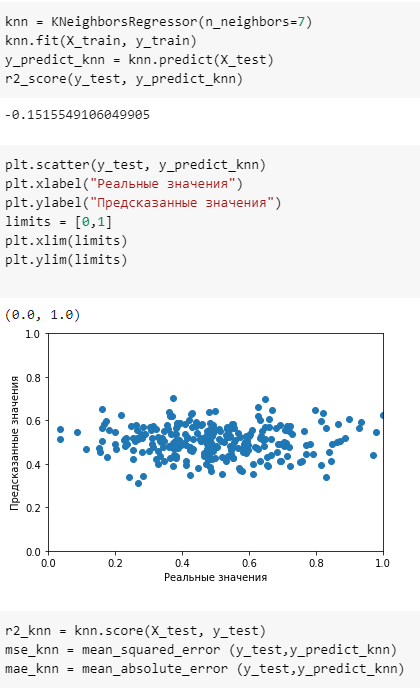
Средняя квадратическая ошибка MSE: 0,0340.

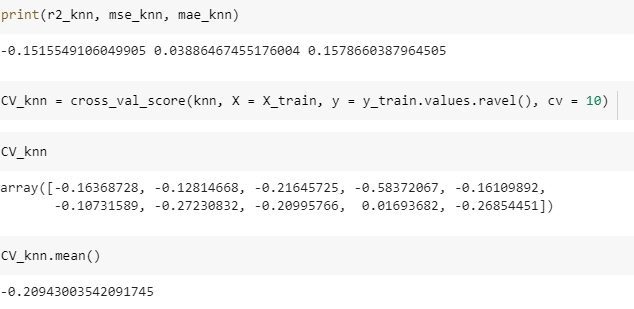
Средняя абсолютная ошибка MAE: 0,1484.

Вывод: метод случайного леса не показал зависимость.

**Модель 3. K-ближайших соседей**

На вход подается нормализованный датасет.





Оценка линейная регрессия: -0,1515.

Средняя квадратическая ошибка MSE: 0,0388.

Средняя абсолютная ошибка MAE: 0,1578.

Вывод: метод k-ближайших соседей не выявил зависимость.

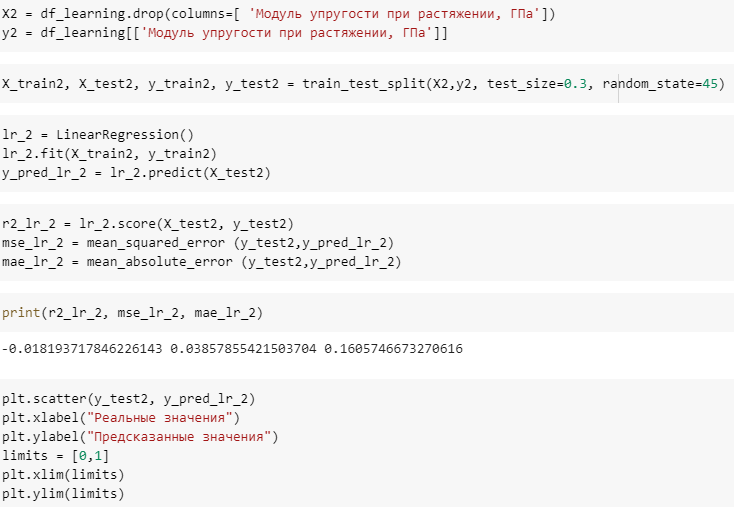
Общий вывод: модели показывают, что линейная зависимость отсутствует, значения средней абсолютной ошибки MAE, средней квадратичной ошибки MSE и коэффициент детерминации R2 низкие, что говорит о плохой предсказательности моделей, соответственно связь между входными переменными и выходной слабая.

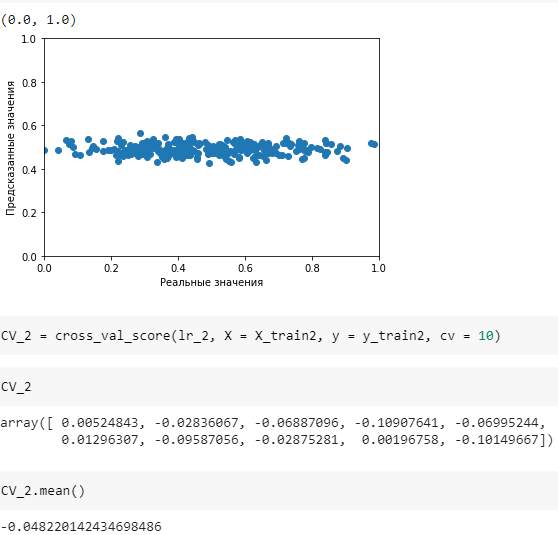
## 2.3 Прогнозирование модели упругости при растяжении

Для прогноза модели упругости при растяжении были использованы следующие модели.

**Модель 1. Линейная регрессия**

На вход подается нормализованный датасет.





Оценка линейная регрессия: -0,0181.

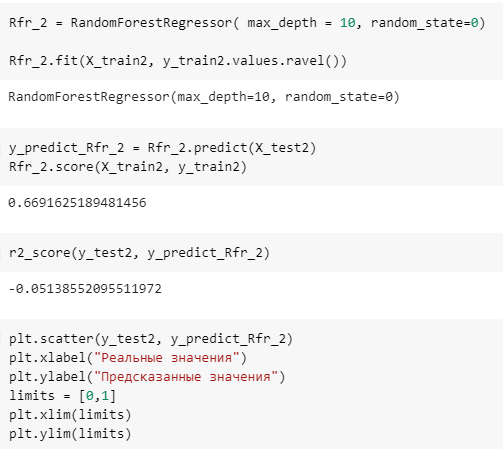
Средняя квадратическая ошибка MSE: 0,0385.

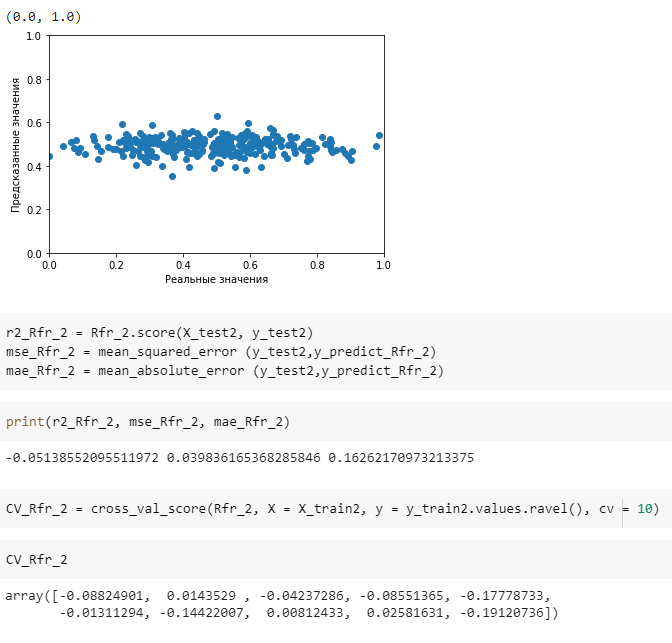
Средняя абсолютная ошибка MAE: 0,1605.

Вывод: линейная регрессия не показала зависимость.

**Модель 2. Случайный лес**

На вход подается нормализованный датасет.





Оценка линейная регрессия: -0,0513.

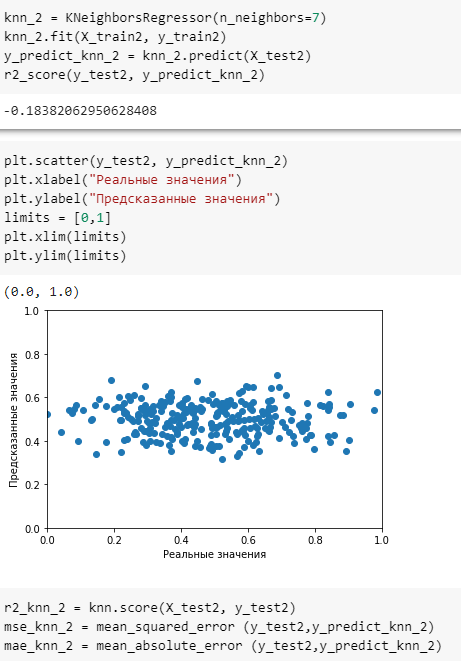
Средняя квадратическая ошибка MSE: 0,0398.

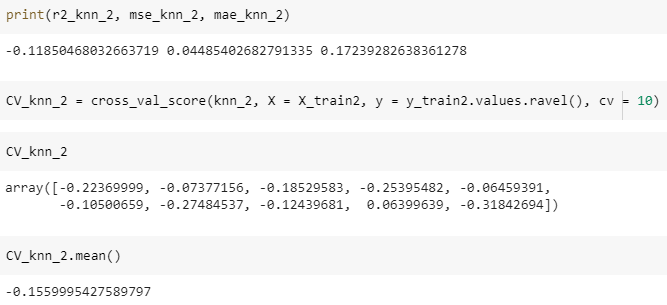
Средняя абсолютная ошибка MAE: 0,1626.

Вывод: метод случайного леса не выявил зависимость.

**Модель 3. K-ближайших соседей**

На вход подается нормализованный датасет.





Оценка линейная регрессия: -0,1185.

Средняя квадратическая ошибка MSE: 0,0448.

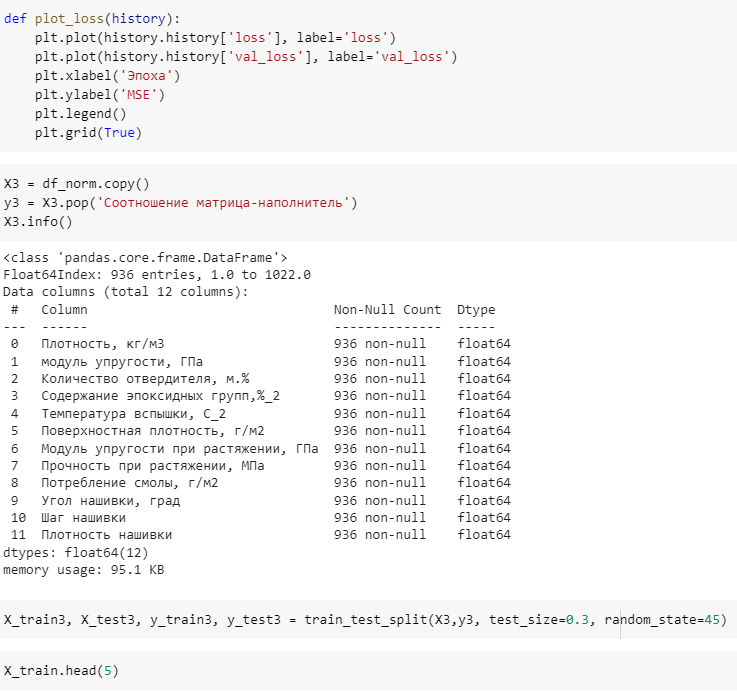
Средняя абсолютная ошибка MAE: 0,1723.

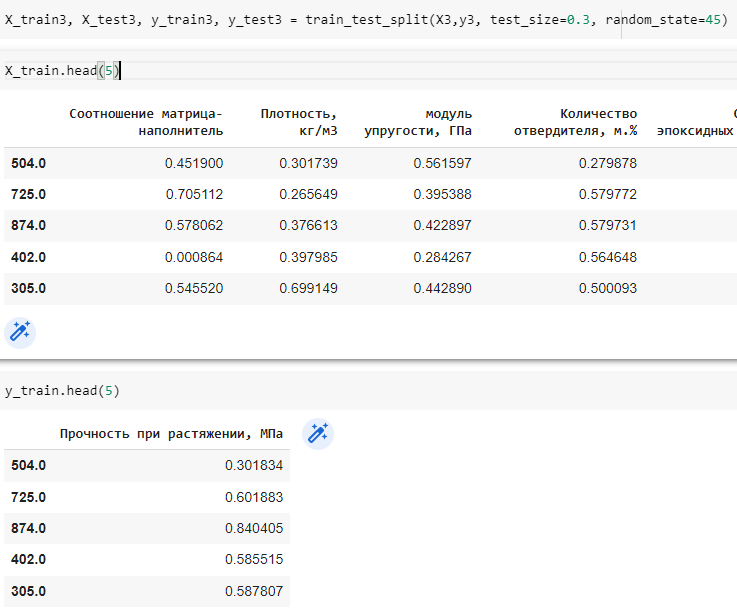
Вывод: метод k-ближайших соседей не выявил зависимость.

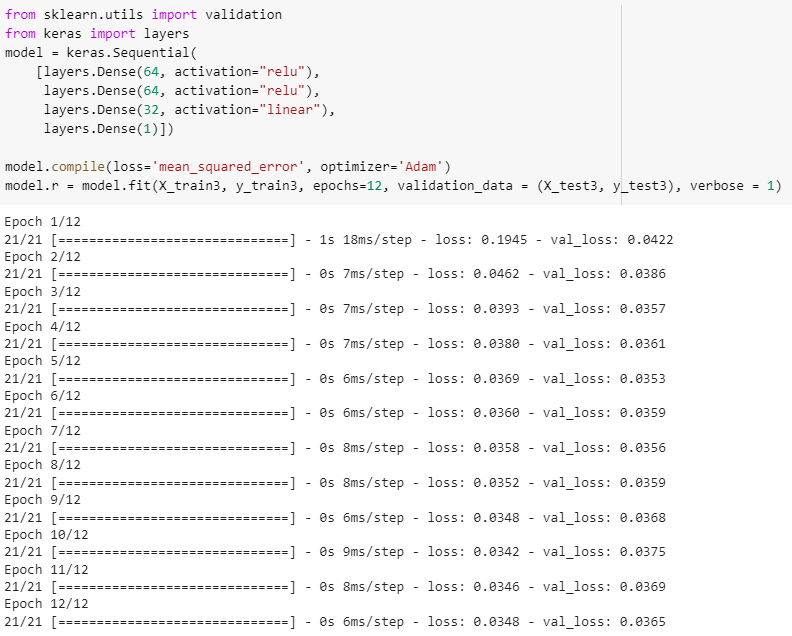
Общий вывод: модели показывают, что линейная зависимость отсутствует, значения средней абсолютной ошибки MAE, средней квадратичной ошибки MSE и коэффициент детерминации R2 низкие, что говорит о плохой предсказательности моделей, соответственно связь между входными переменными и выходной слабая.

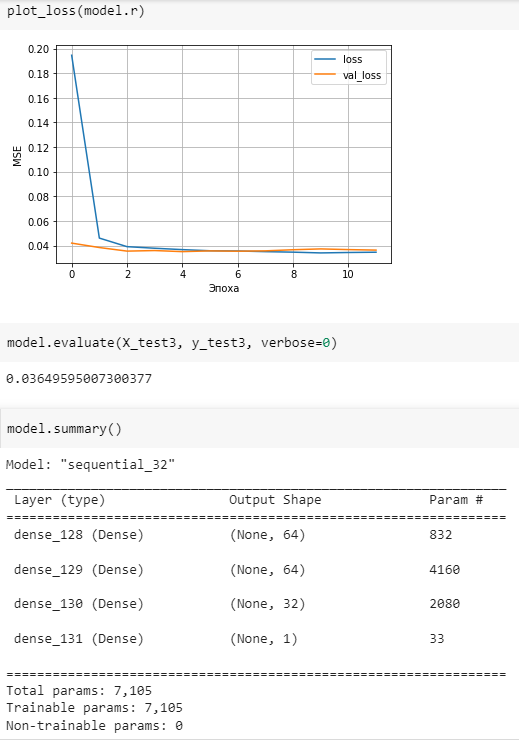
## 2.4 Разработка нейронной сети для прогнозирования соотношения матрица-наполнитель

Приведем описание выбранной архитектуры нейронной сети с 1 нейроном на выходе и ее результаты.









Точность нейронной сети составила 0,0364.

Полученные результаты говорят о том, что либо в нашем наборе данных нет зависимостей, либо наша модель их не смогла определить.

## 2.5 Создание удаленного репозитория

Создан репозиторий в GitHub, где размещен код исследования, оформлен файл README.

Страница слушателя: <https://github.com/OVMaslova/-_-> .

## Заключение

В данной работе продемонстрированы некоторые методы обработки данных и построения моделей зависимости одних параметров от других.

Первичный статистический анализ данных не выявил очевидных зависимостей между переменными. Построенные модели линейной регрессии, случайного леса, k-ближайших соседей, нейронной сети также не дали удовлетворительных результатов по выявлению зависимости модуля упругости при растяжении, прочности при растяжении, соотношения матрица-наполнитель от входных переменных.

Можно сделать вывод, что используемым в работе набором алгоритмов задачу прогнозирования конечных свойств новых материалов (композиционных материалов) решить не удалось. Значит, либо зависимостей в наборе данных нет, либо они могут быть изучены при применении других, отличных от указанных, алгоритмов или параметров исследования.

## Список использованной литературы

1. Композиционные материалы : учебное пособие для вузов / Д. А. Иванов, А. И. Ситников, С. Д. Шляпин ; под редакцией А. А. Ильина. — Москва : Издательство Юрайт, 2019 — 253 с. — (Высшее образование). — Текст : непосредственный.

2. Документация по языку программирования python: – Режим доступа: <https://docs.python.org/3.8/index.html>.

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

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

5. Документация по библиотеке matplotlib: – Режим доступа: <https://matplotlib.org/stable/users/index.html>.

6. Документация по библиотеке seaborn: – Режим доступа: <https://seaborn.pydata.org/tutorial.html>.

7. Документация по библиотеке sklearn: – Режим доступа: <https://scikit-learn.org/stable/user_guide.html>.

8. Документация по библиотеке keras: – Режим доступа: <https://keras.io/api/>.