**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 Перфильева Анна Владимировна

Санкт-Петербург, 2022

**Содержание**

Содержание.............................................................................................................................. 2

Введение ...................................................................................................................................3

Анализ данных………..............................................................................................................5

Предобработка данных……………………………..............................................................11

Построение регрессионных моделей....................................................................................13

Сравнение результатов моделей...........................................................................................14

Построение нейронной сети .................................................................................................16

Создание приложения……………………………………....................................................18

Создание удалённого репозитория GitHub и загрузка документов..............................................................................................................................22

Выводы………………………………………………………………………………………23

**Введение**

Композиционный материал или композитный материал, сокращённо композит - многокомпонентный материал, изготовленный (человеком или природой) из двух или более компонентов с существенно различными физическими и (или) химическими свойствами, которые, в сочетании, приводят к появлению нового материала с характеристиками, отличными от характеристик отдельных компонентов и не являющимися простой их суперпозицией.

Таким образом, при известных характеристиках исходных компонентов определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы существует два пути: физические испытания образцов материалов (что достаточно финансово затратно), или прогнозирование характеристик.

Учитывая тот факт, что Россия вошла в четверку стран-лидеров в области композитов, наряду с США и Японией, при этом обеспечив 100%-е импортозамещение по всей линейке продукции, развитие отрасли, в том числе и изучение свойств композитов – одна из актуальных проблем отечественной промышленности. Отрасль растет на 10% в год, в стране реализуется более 150 проектов с объемом инвестиций более 150 млрд. рублей, но есть куда стремиться: по сути России необходимо повторить один из этапов индустриализации 20 века, только в развитии не металлургии как в 30-е годы, а композитостроения, которое по прогнозам в ближайшие годы превзойдет черную металлургию, тогда к 2030 году мы не просто сохраним свое место в большой четверке, но и станем ее лидерами.

В данном исследовании в качестве известных параметров имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов: Модуль упругости при растяжении, Прочность при растяжении и Соотношение матрица-наполнитель. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

Цель данного исследования – получение алгоритмов машинного обучения, которые будут при заданных характеристиках композиционных материалов (всего 13 параметров), определять значения: «Модуль упругости при растяжении» и «Прочность при растяжении», написание нейронной сети, которая будет предлагать значение параметра «Соотношение матрица – наполнитель», а также создание пользовательского веб-приложения с на фреймворке Flask для вывода предсказания этой нейронной сети .

**Анализ данных**

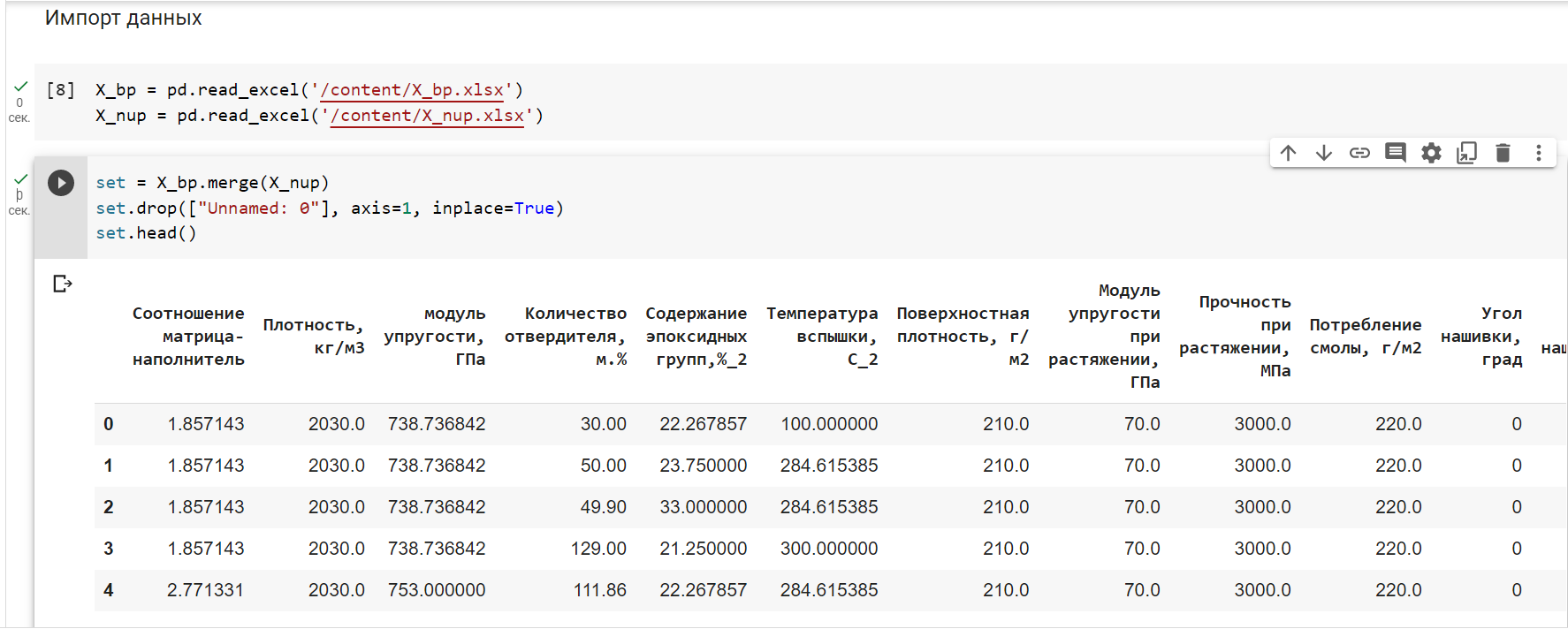
1. **Предобработка данных**

После импорта необходимых библиотек

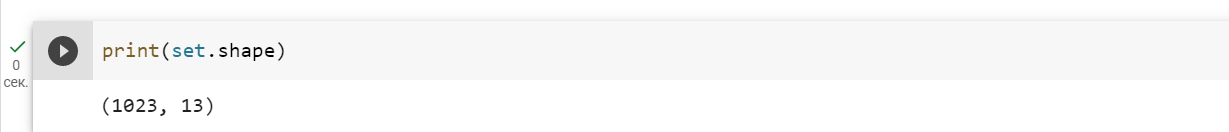


приступаем к изучению данных: дано 2 файла: X\_bp.xlsx и X\_nup.xlsx содержащие данные с характеристиками композитов и различающиеся по количеству строк.

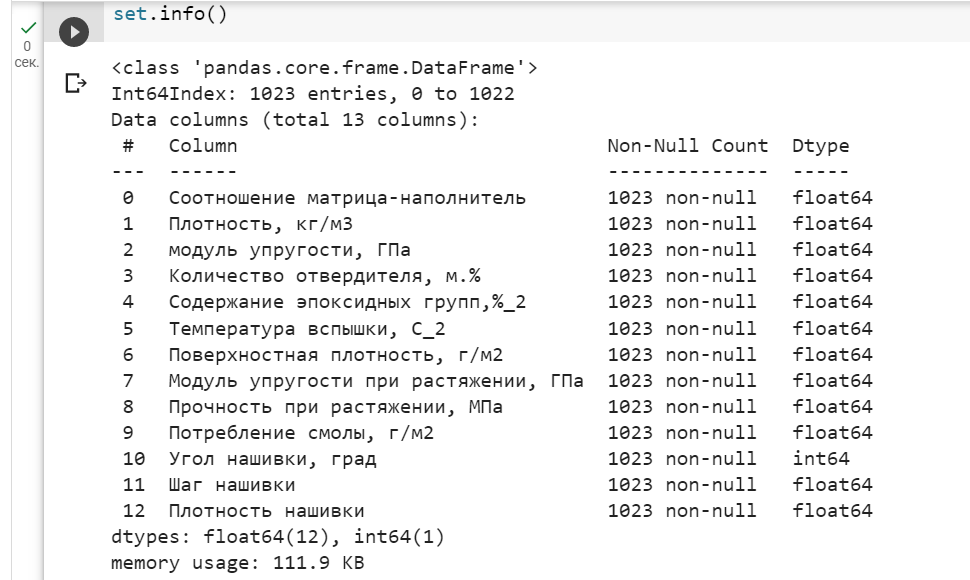
Необходимо объединить файлы методом Inner Join.



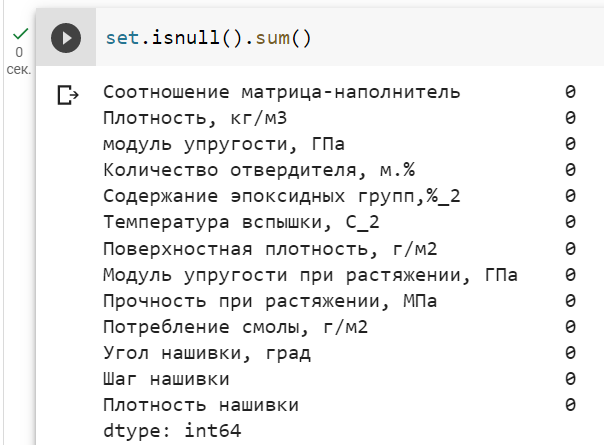
В результате получаем таблицу с 13 столбцами (параметрами) и 1023 строками.



Все данные вещественные, в том числе и «Угол нашивки, град». По моему мнению, это физическая величина и должна рассматриваться как вещественный признак (несмотря на то что в данном датасете она принимает только два значения – 0 и 90 градусов), т.к. предусмотрены и другие значения данного признака.

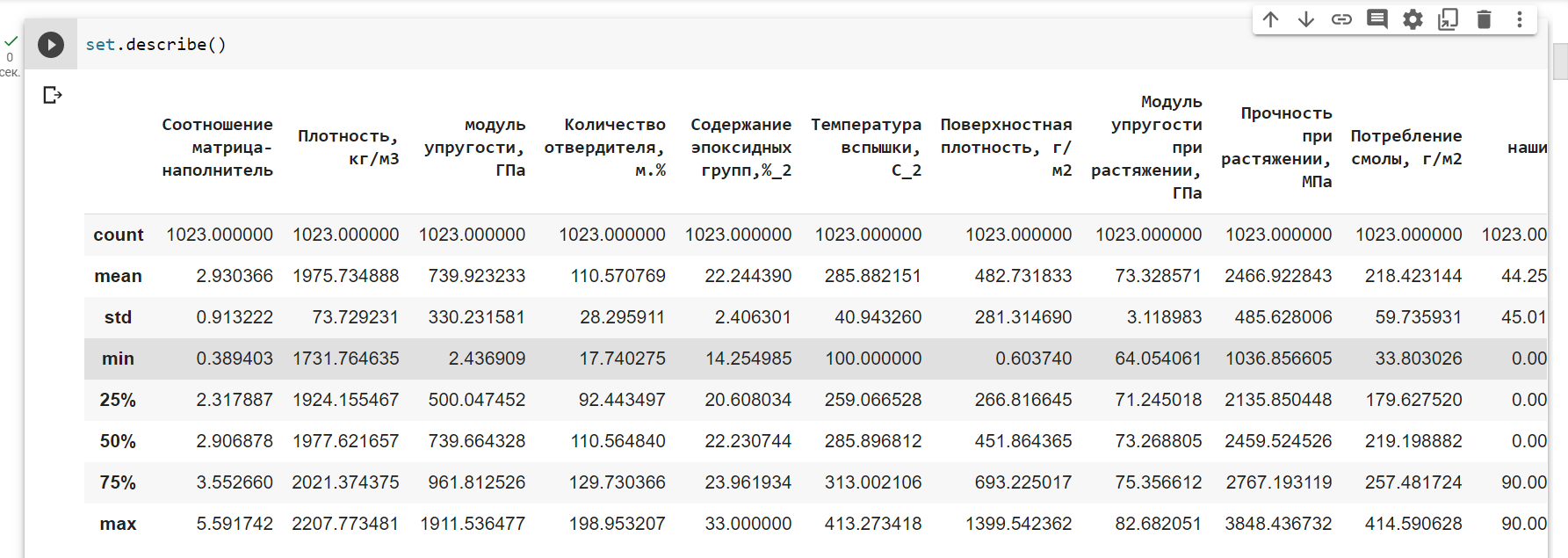


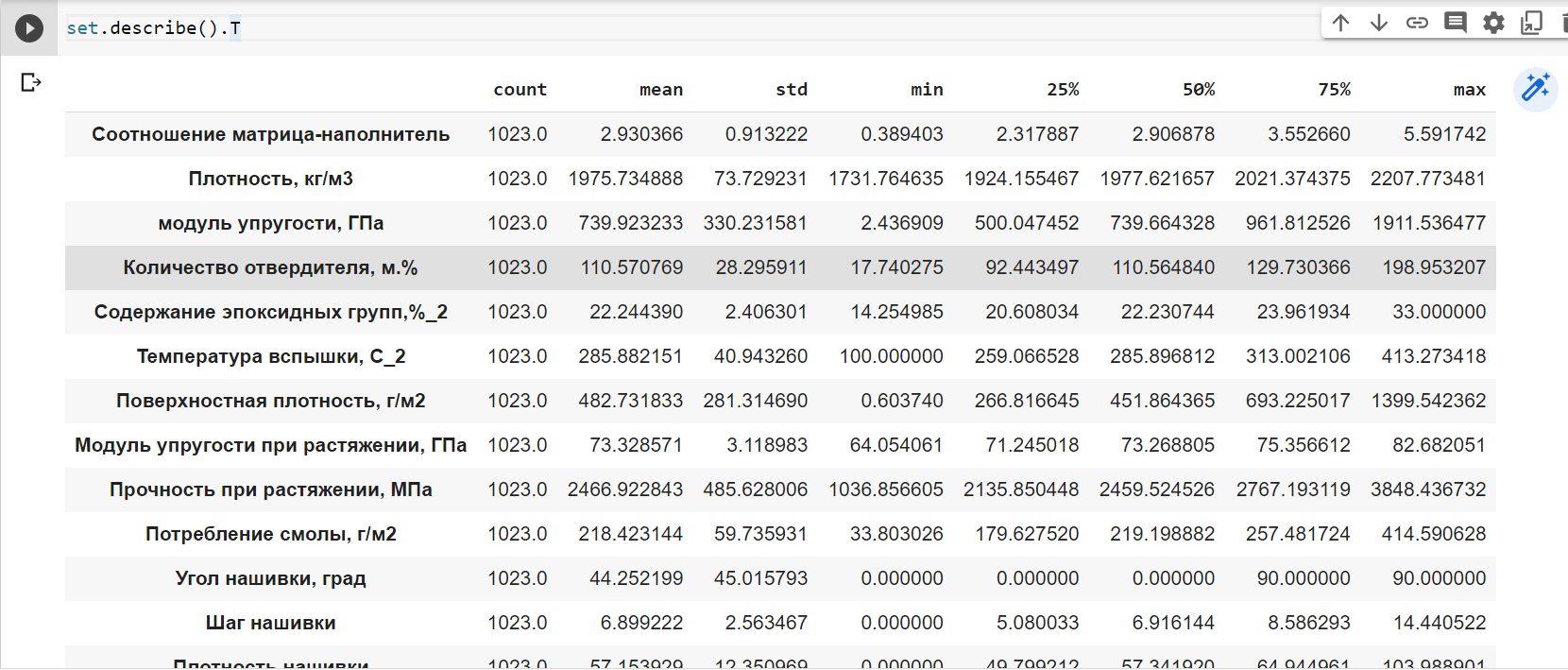
Пропуски отсутствуют:



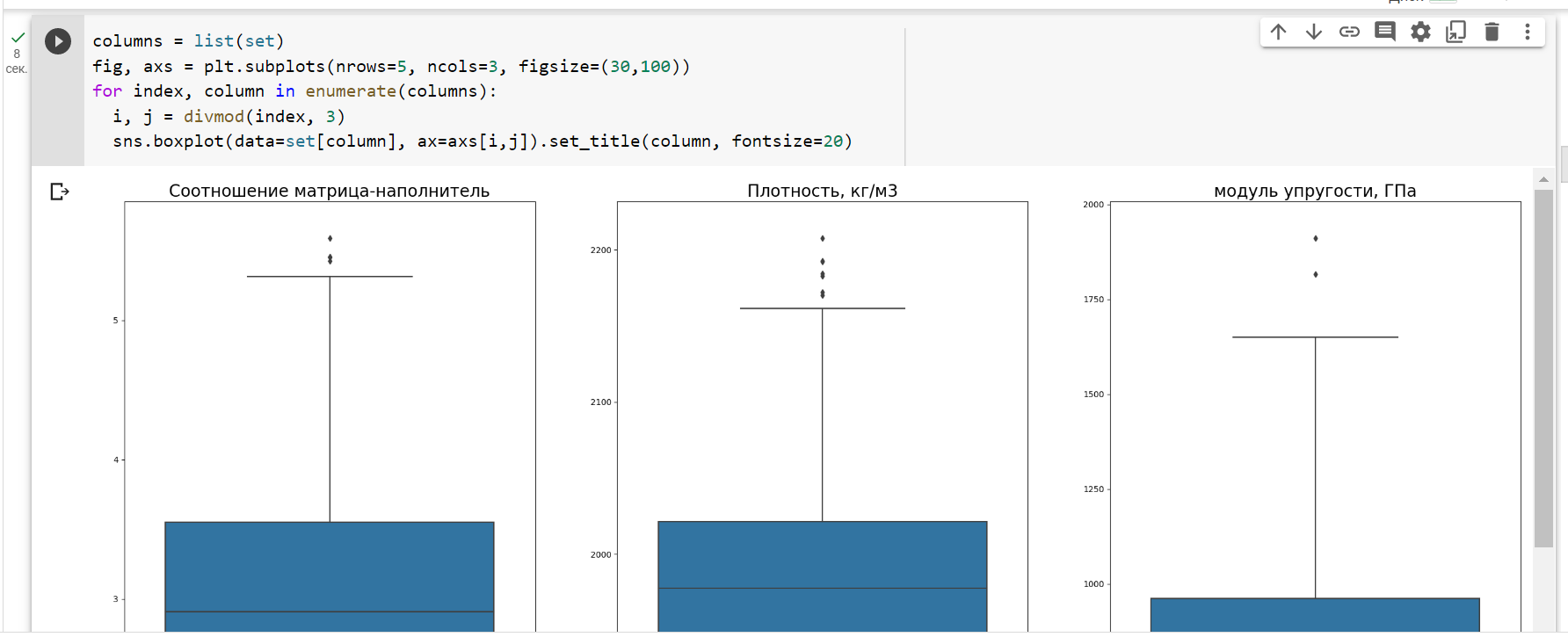
Описательная статистика: проверка распределения данных - количество строк в столбце - count, среднее значение столбца - mean, столбец стандартное отклонение - std, минимальные (min) и максимальные (max) значения, а также границу каждого квартиля - 25%, 50% и 75%, в натуральном и транспонированном виде.

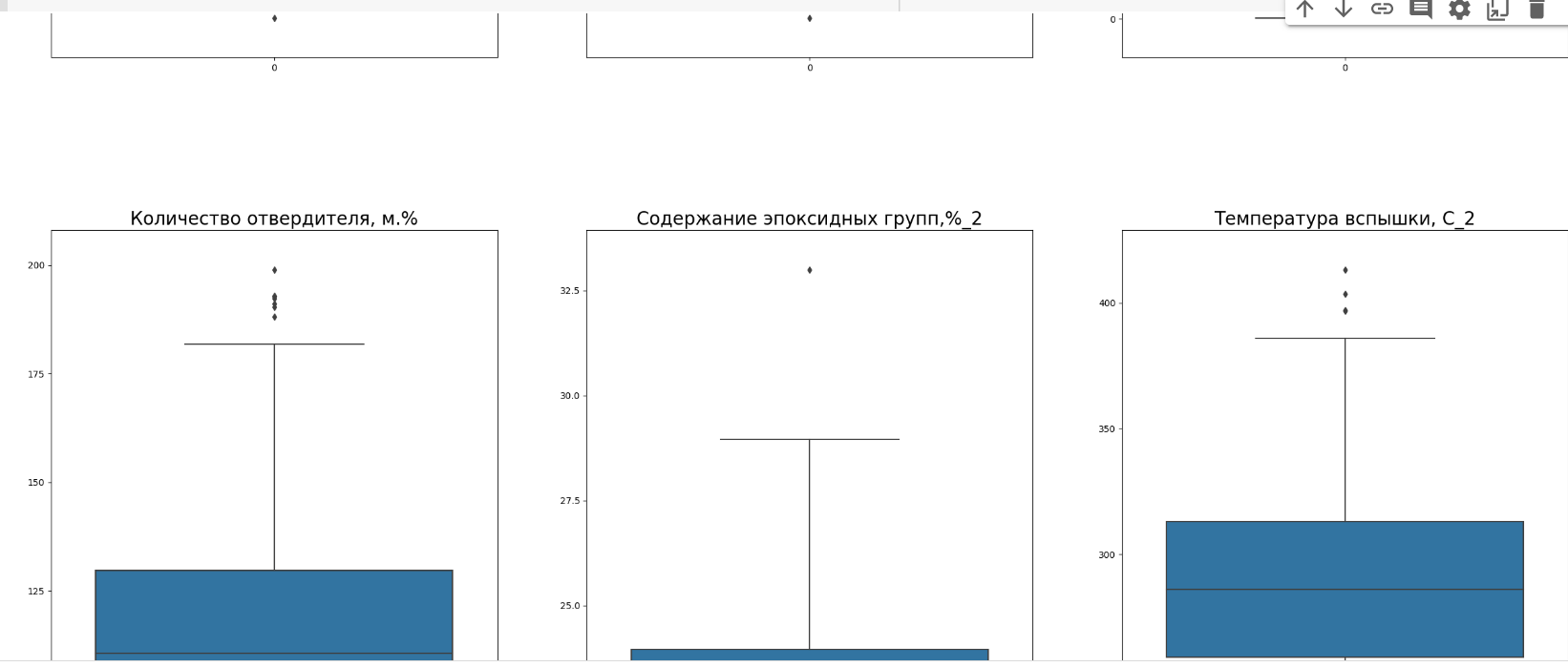
Очевидно, что данные необходимо масштабировать для корректного обучения моделей.

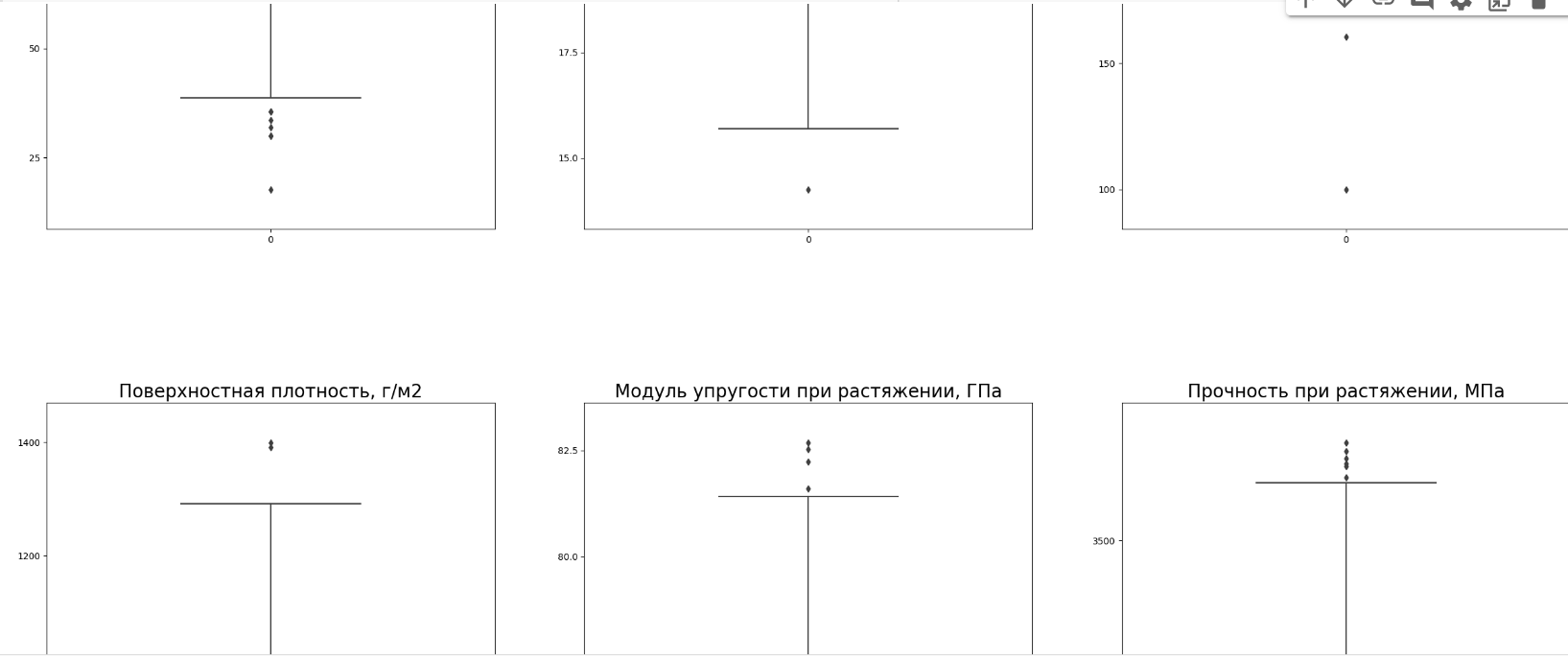




Анализ на выбросы (т.н. «Ящик с усами») свидетельствует о наличии выбросов в датасете, но мной принято решение не устранять выбросы и аномалии, т.к., во-первых, размер данного датасета небольшой, и не хотелось бы терять часть данных, во-вторых, не уверена, что эти выбросы – ошибки, допущенные при дорогостоящем тестировании выходных материалов, предполагаю, что эти данные – действительно физические свойства композитов, которые довольно непредсказуемы и слабо коррелируются со свойствами исходных материалов, в чем мы убедимся далее.







Тепловая карта а также таблицы значения коэффициента корреляции целевых признаков («Модуль упругости при растяжении», «Прочность при растяжении», «Соотношение матрица-наполнитель») говорит об очень низкой связи параметров.

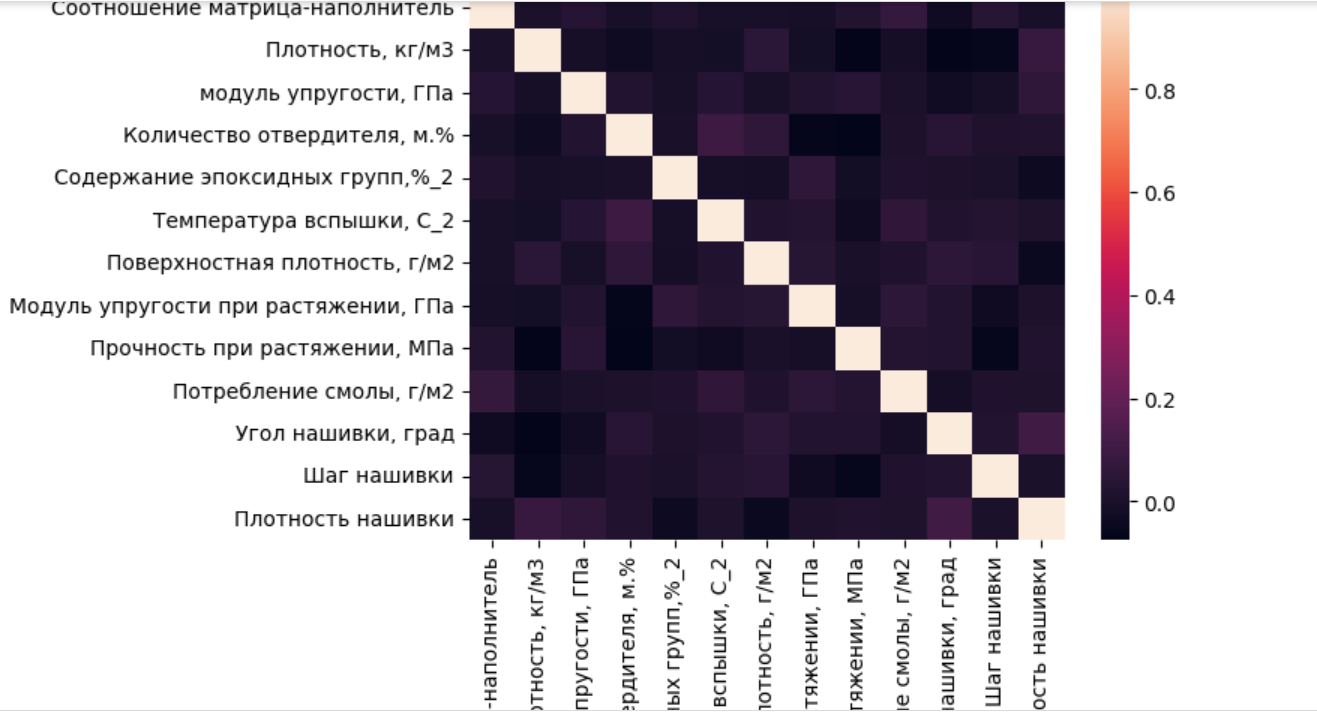
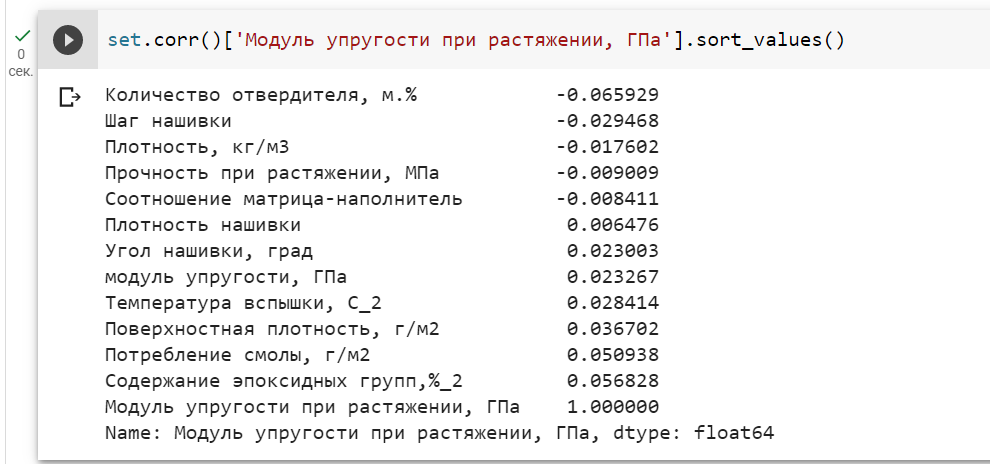
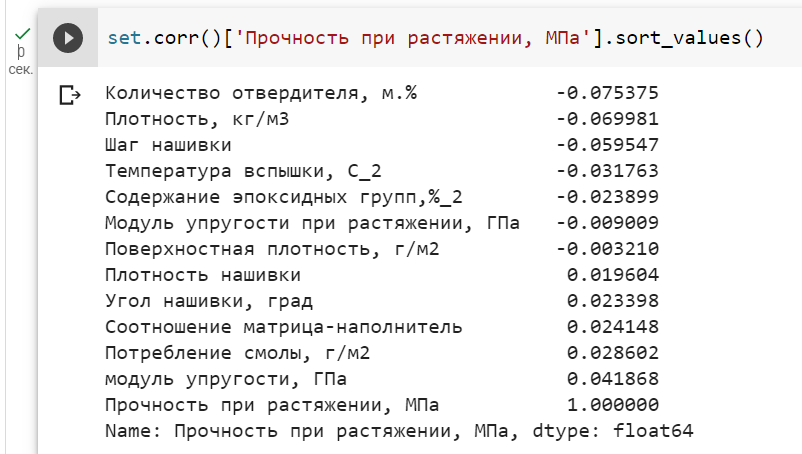
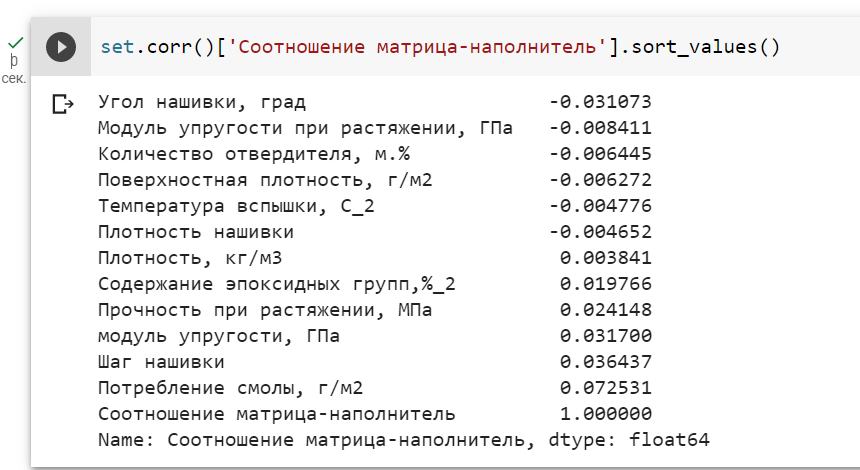
  Коэффициент корреляции Пирсона, который является мерой линейной связи между переменными и всегда принимает значение от **-1 до 1**, где

**-1** указывает на совершенно отрицательную линейную корреляцию,

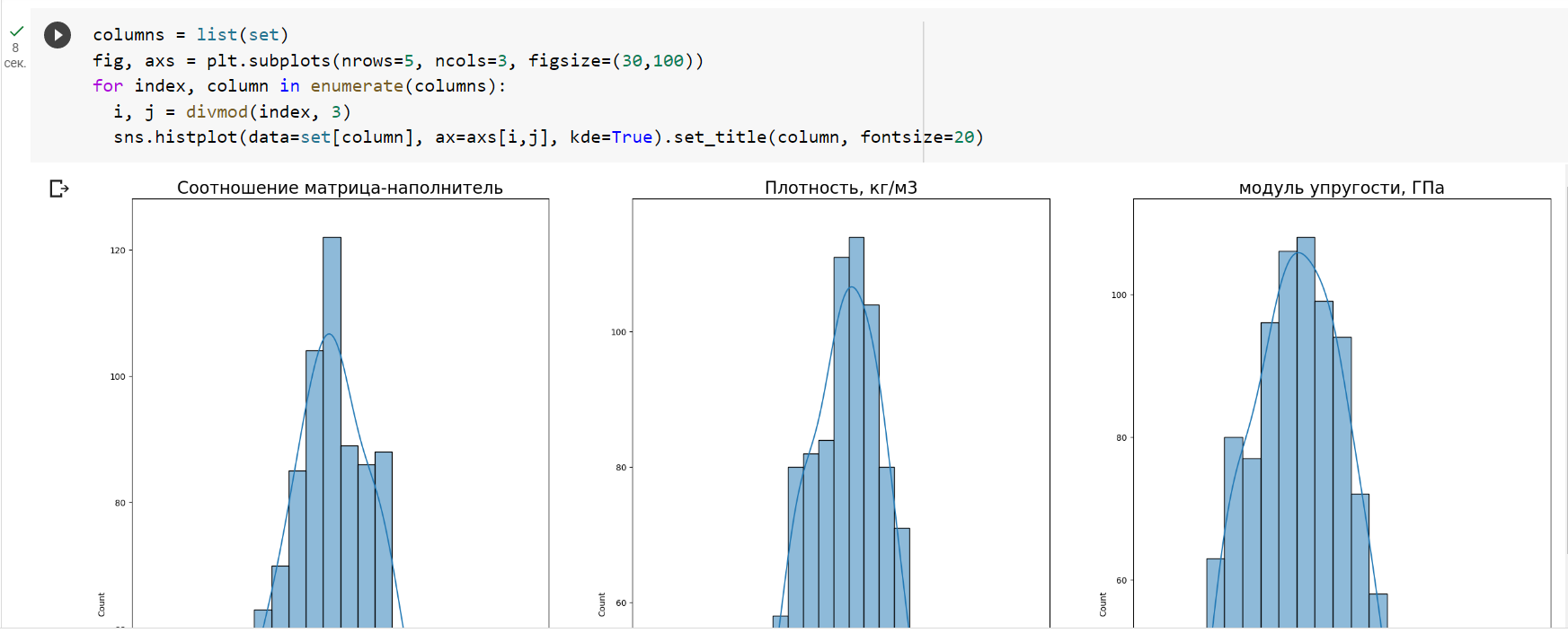
1. указывает на отсутствие линейной корреляции,

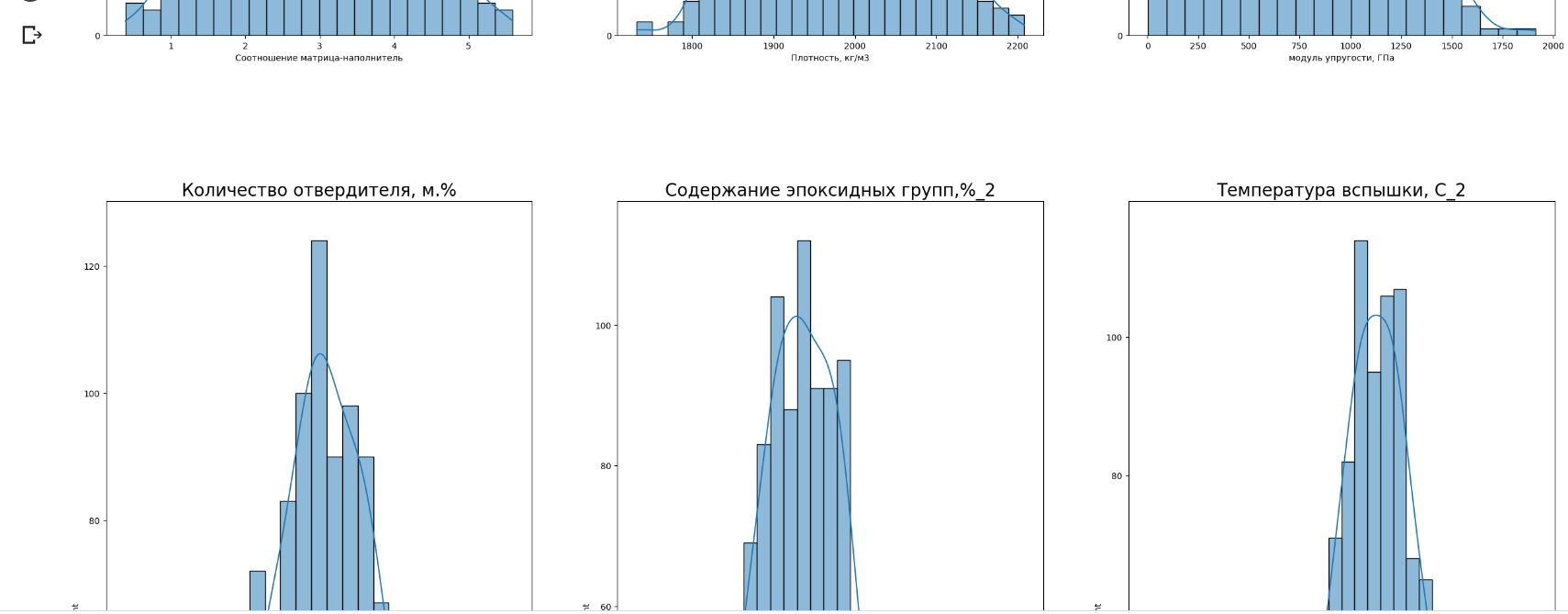
а **1** указывает на совершенно положительную линейную корреляцию между двумя переменными

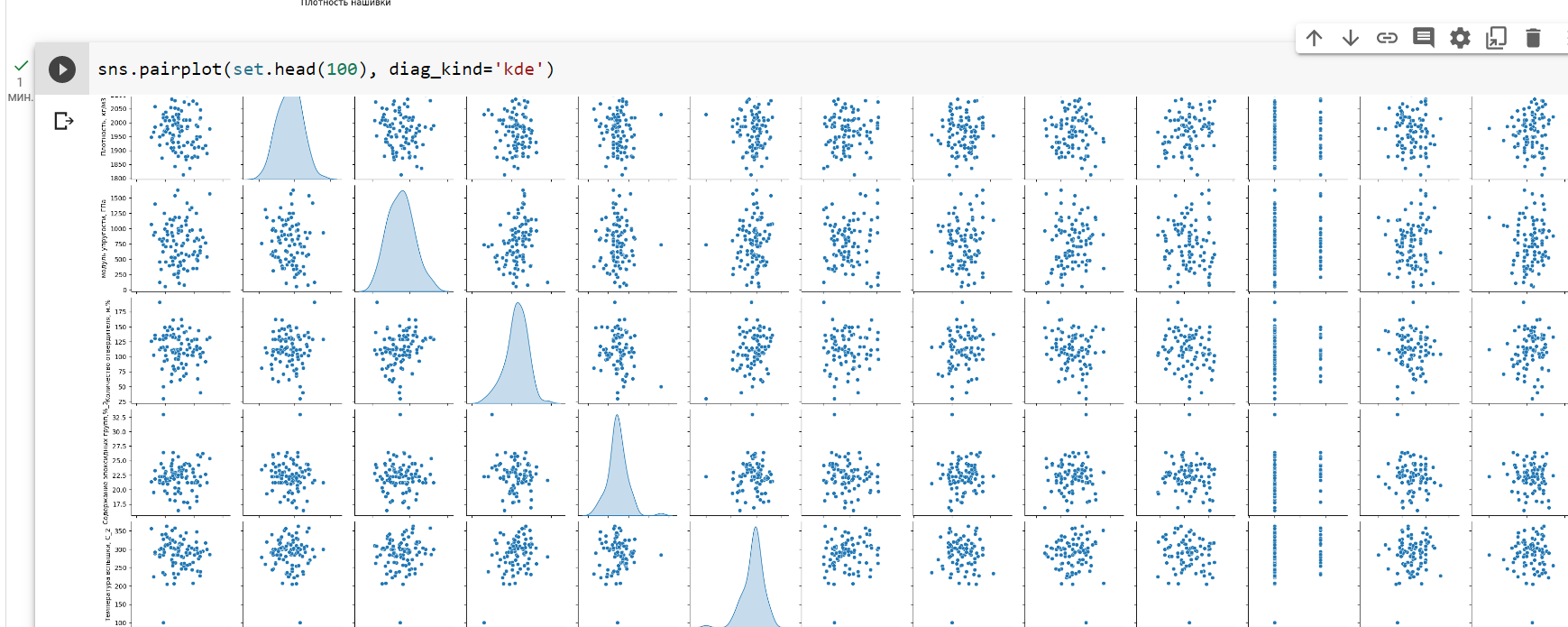
в нашем случае коэффициент корреляции практически везде близок к 0, т. е. данные не имеют четко выраженной зависимости.

Далее приводятся гистограммы распределения каждой из переменной, попарные графики рассеяния точек, по которым можно заключить, что распределение данных близко к нормальным.





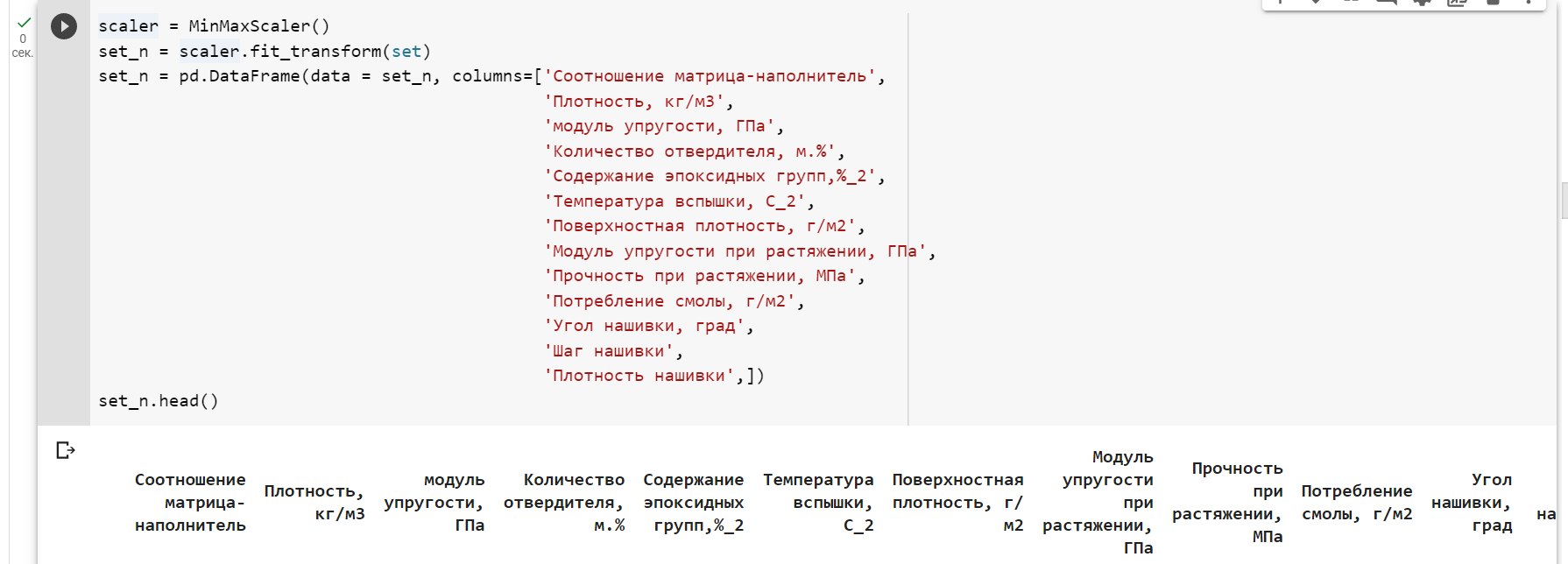


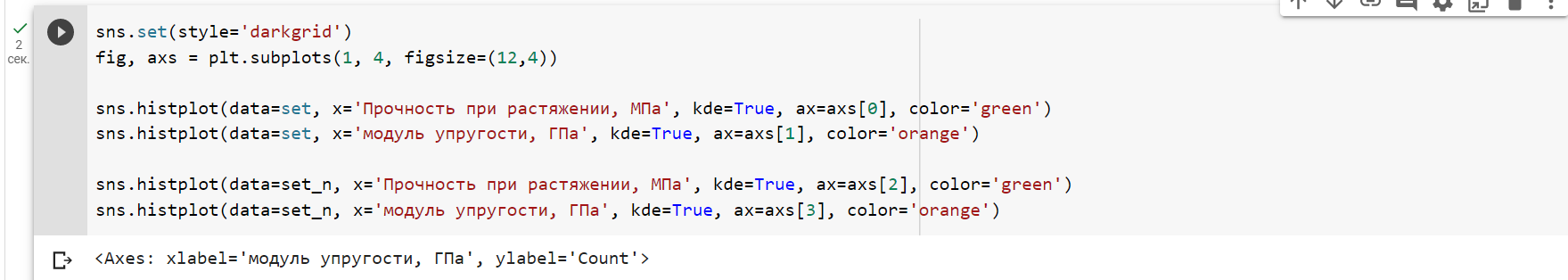
**Предобработка данных**

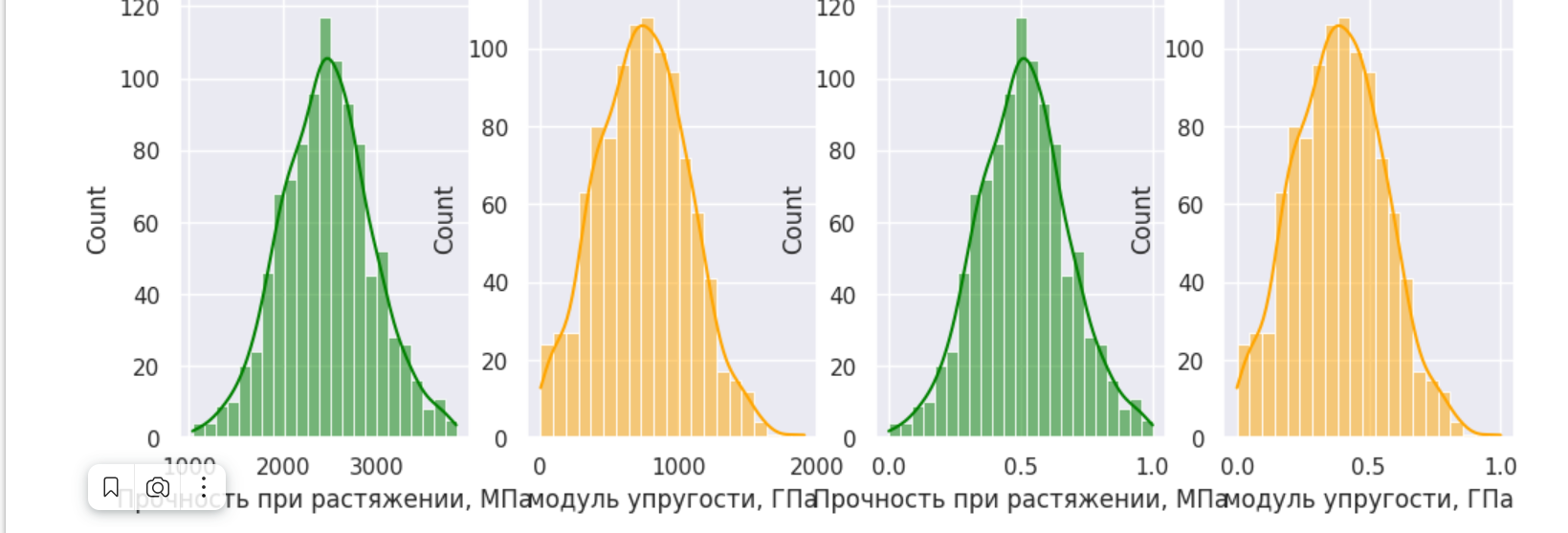
При обучении моделей для того, чтобы они как можно скорее сходились, как правило производится предварительная обработка данных. Для масштабирования данных в датасетах, используют стандартизацию или нормализацию.

Стандартизация заключается в пропорциональном масштабировании данных для размещения их в интервале от -1 до 1, т. е. стандартизованные данные могут быть положительными или отрицательными. Нормализация является способом стандартизации, которая отображает данные в интервале от 0 до 1.

Учитывая то, что наши данные – физические показатели, я посчитала уместным использовать именно нормализацию:

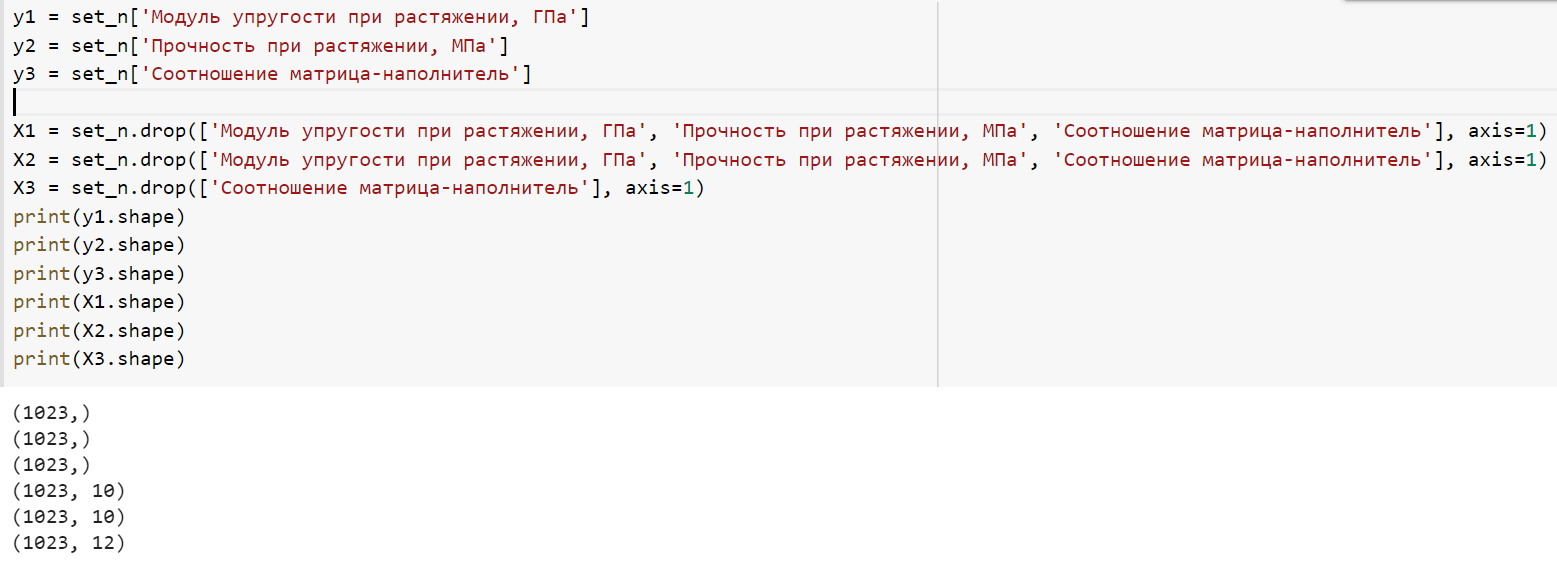




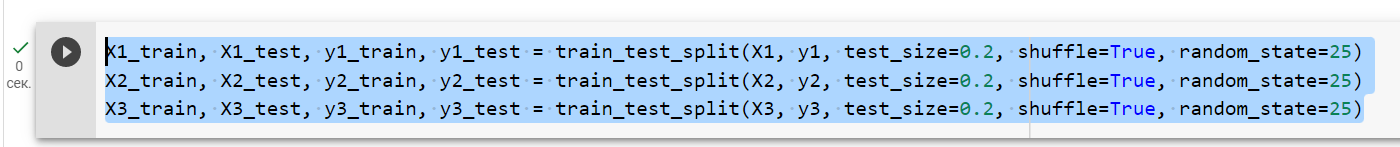


Затем следует выделение целевых переменных:

* для построения регрессионных моделей для определения значений параметров «Модуль упругости при растяжении, Гпа» и «Прочность при растяжении, Мпа» в качестве обучающих переменных был выделен начальный датафрейм с исключением трех столбцов - «Модуль упругости при растяжении, Гпа», «Прочность при растяжении, Мпа» и «Соотношение матрица-наполнитель».
* для построения нейросети, которая должна рекомендовать параметр «Соотношение матрица-наполнитель», в качестве обучающих переменных был выделен начальный датафрейм после исключения одного столбца с целевой переменной «Соотношение матрица-наполнитель».



Далее - разделение обучающих датасетов на тренировочные и тестовые выборки.



**Построение регрессионных моделей**

Разработка и обучение моделей машинного обучения для целей регрессии осуществлялась для предсказания значений параметров: «Прочность при растяжении» и «Модуль упругости при растяжении». Цель обучения — определить функцию потерь (ошибку предсказания модели) и минимизировать её.

В процессе исследования применялись следующие методы:

1. **линейная регрессия** — 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;
2. **случайный лес -** универсальный алгоритм машинного обучения, суть которого состоит в использовании ансамбля решающих деревьев, из-за большого количества которых результат значительно улучшается;
3. **дерево решений** - непараметрический контролируемый метод обучения, цель которого состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной, изучая простые правила принятия решений, выведенные из характеристик данных;
4. **метод опорных векторов** – метод бинарной классификации, основная идея которого заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом;
5. **метод К-ближайших соседей** - метрический алгоритм для автоматической [классификации](https://ru.wikipedia.org/wiki/%D0%97%D0%B0%D0%B4%D0%B0%D1%87%D0%B0_%D0%BA%D0%BB%D0%B0%D1%81%D1%81%D0%B8%D1%84%D0%B8%D0%BA%D0%B0%D1%86%D0%B8%D0%B8) объектов или [регрессии](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F_(%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0)). В случае использования метода для [регрессии](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F_(%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0)), объекту присваивается среднее значение по ближайшим к нему объектам, значения которых уже известны.
6. **метод Лассо** — это вариация линейной регрессии, адаптированная для данных, которые демонстрируют сильную мультиколлинеарность и использует сжатие коэффициентов, то есть процесс, в котором значения данных приближаются к центральной точке (например, среднему значению).

**Сравнение результатов моделей**

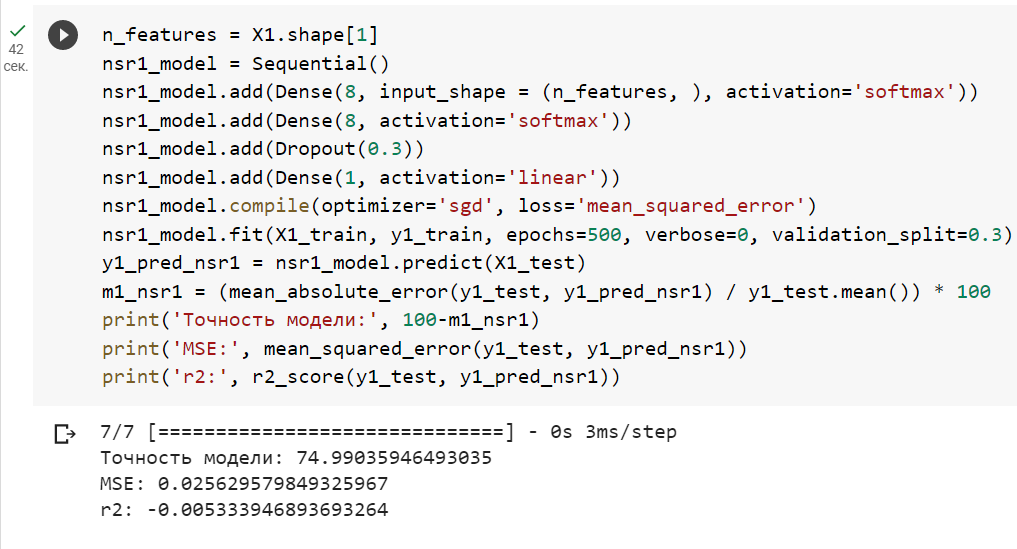
После обучения моделей была проведена оценка их точности. В качестве параметра оценки модели использовались показатели:

* **средняя** **абсолютная** **ошибка** (MAE) – это степень несоответствия между фактическими и прогнозируемыми значениями;
* **средняя квадратичная ошибка модели (MSE)** – это среднее расстояние между прогнозируемыми значениями модели и фактическими значениями данных, чем ближе к нулю MSE, тем лучше работают предсказательные качества модели;
* **коэффициент детерминации (r2)** - показатель, отражающий объясняющую способность [регрессии](https://wiki.loginom.ru/articles/regression.html), это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. 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.
* **точность модели** – доля объектов, верно предсказанная алгоритмом, доля правильных ответов.



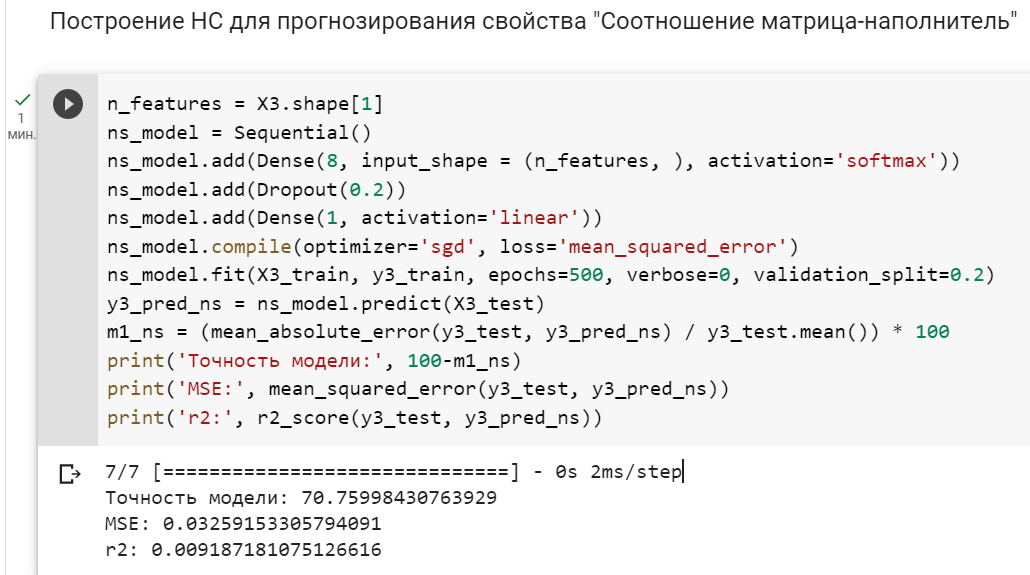
В результате сравнительного анализа шести моделей регрессии можно резюмировать, что лучший результат по обоим показателям («Прочность при растяжении» и «Модуль упругости при растяжении») – у модели Лассо.

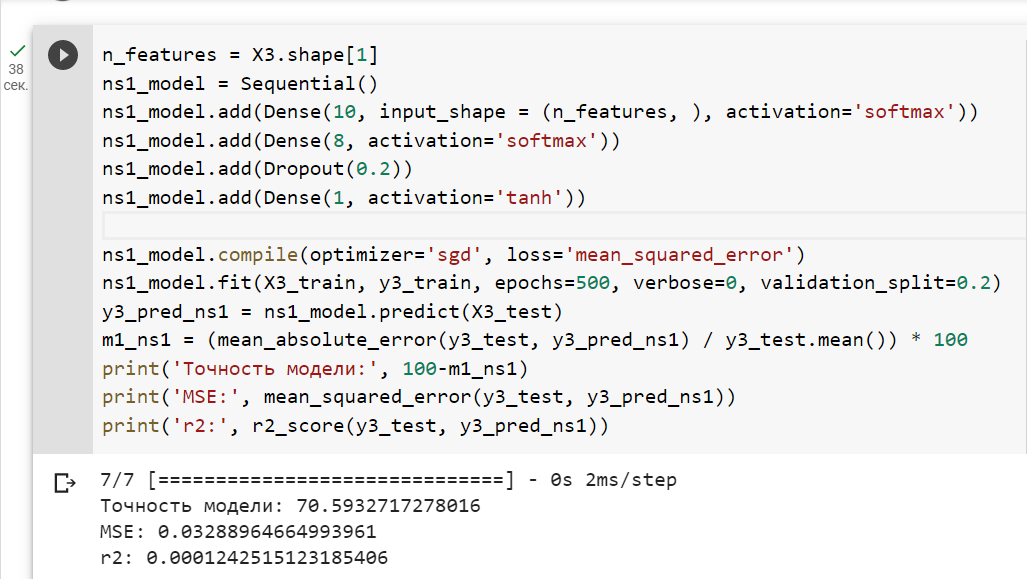
В качестве эксперимента была построена нейронная сеть с двумя скрытыми слоями для предсказания показателя «Прочность при растяжении», которая показала примерно такой же результат, как и предыдущие регрессионные модели.



**Построение нейронной сети**

В данном исследовании мной предложено две полносвязных нейросети с одним и двумя скрытыми слоями соответственно и одним суммирующим слоем для прогнозирования показателя «Соотношение матрица-наполнитель». Количество слоев далее не стала повышать, т. к. для любой численной регрессии гладкой функции достаточно двух скрытых слоев при достаточном числе нейронов.



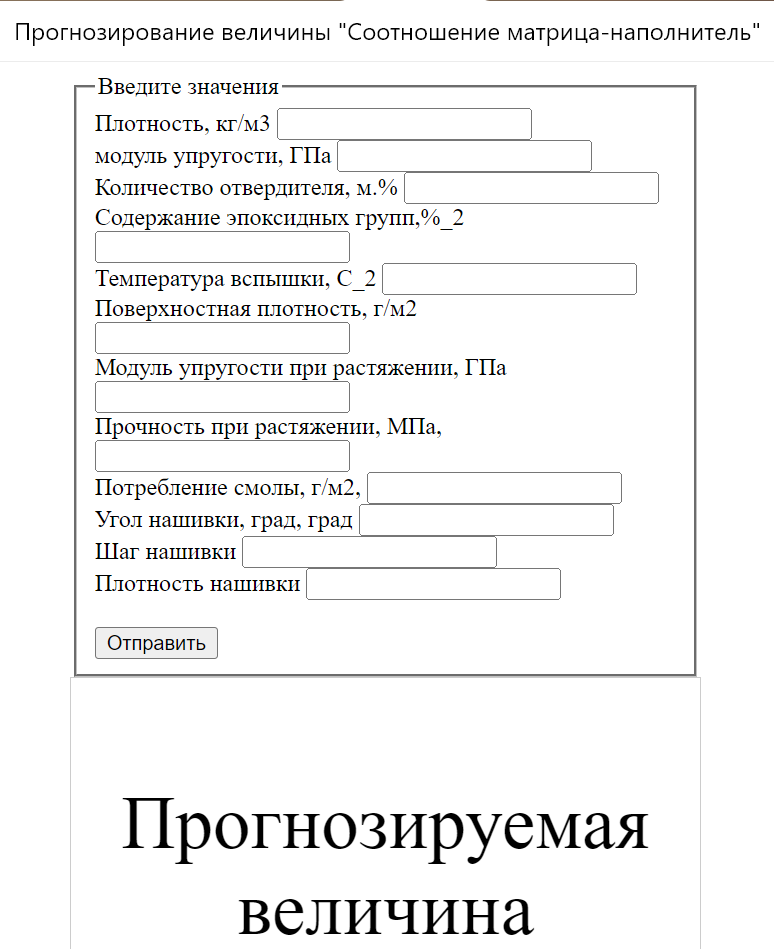
****

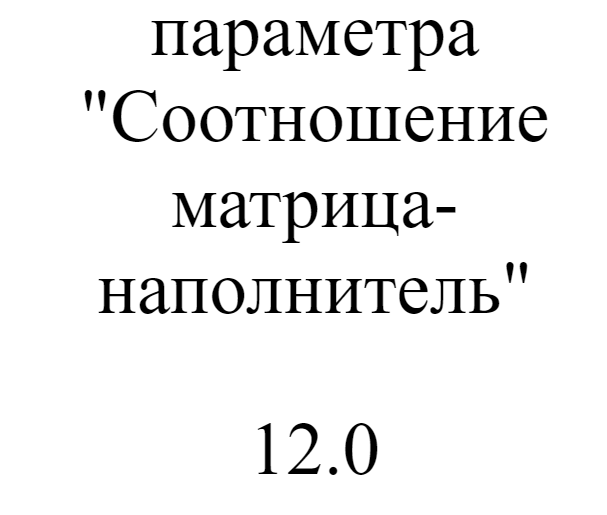
По показателям точности модели можно резюмировать, что результат очень близок к результатам, полученным при использовании моделей регрессии. Коэффициент детерминации также близок к нулю, но уже положительный, что говорит о том, что нейросеть предсказывает целевые значения чуть лучше, чем наивная линейная регрессия.

**Создание приложения**

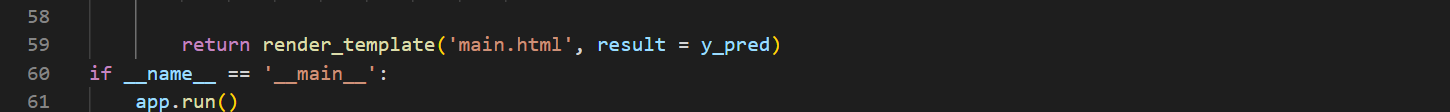
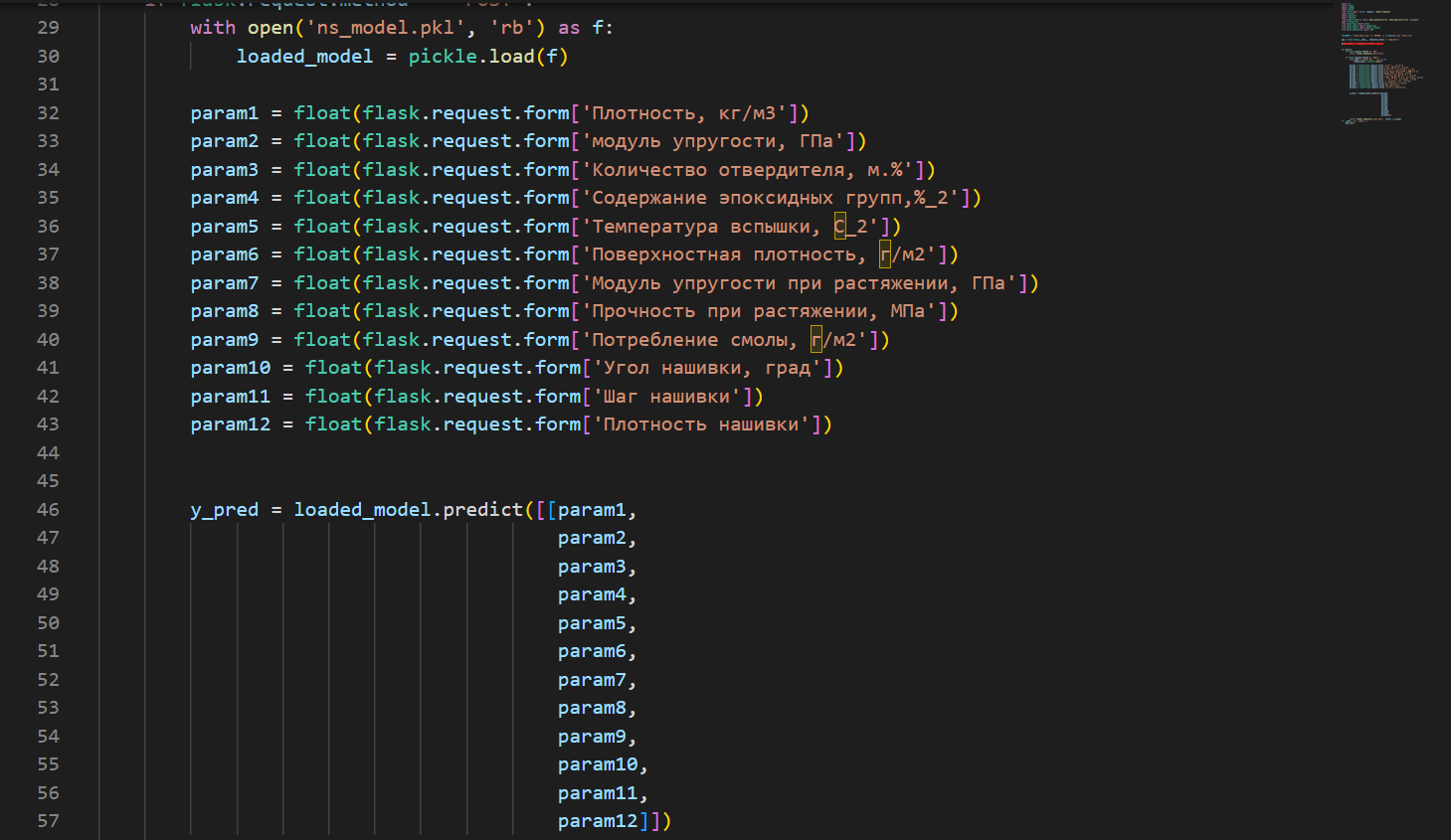
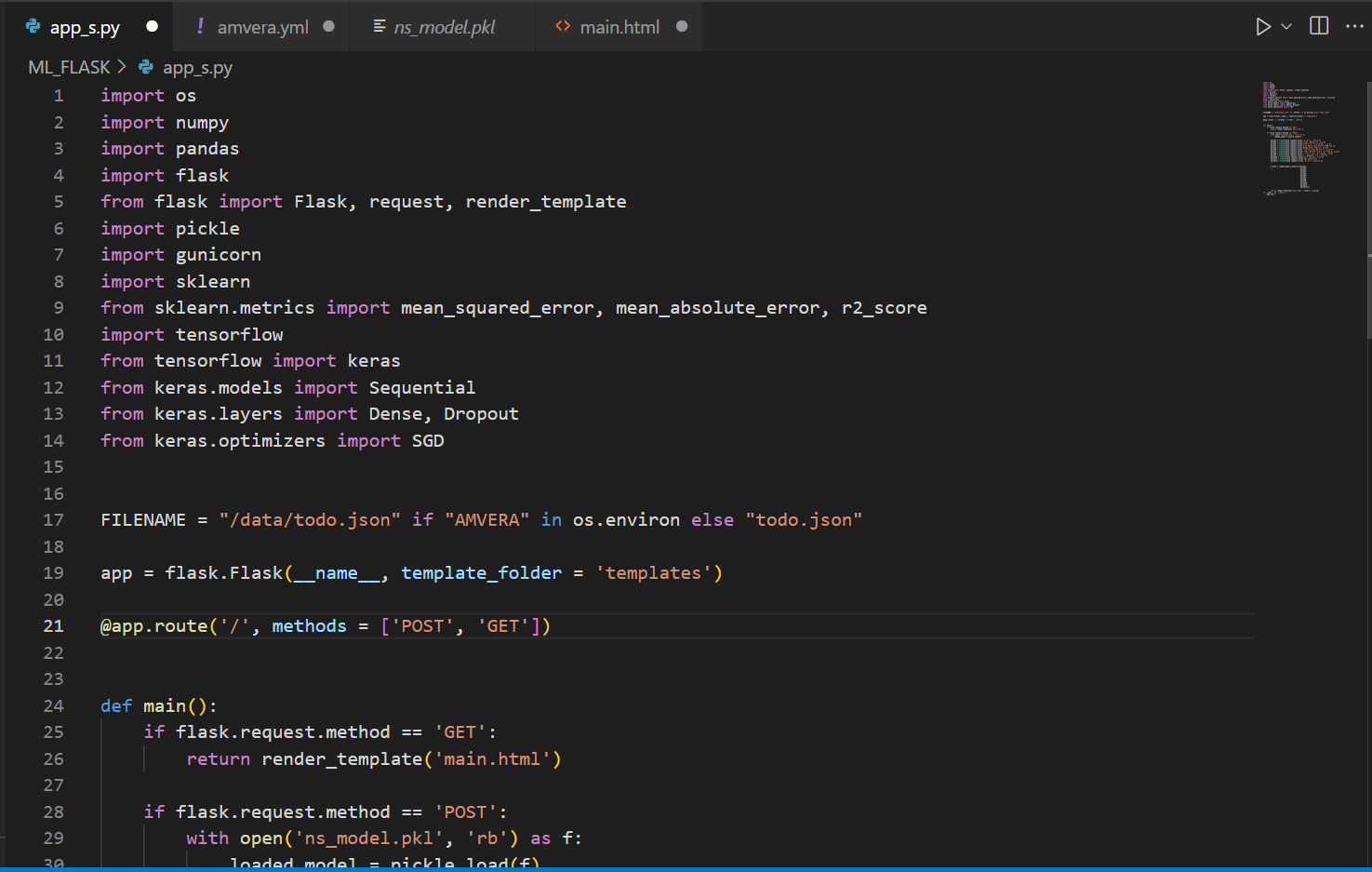
Создано пользовательское веб-приложение с на фреймворке Flask для вывода предсказанного нейросетью параметра **«Соотношение матрица-наполнитель».**

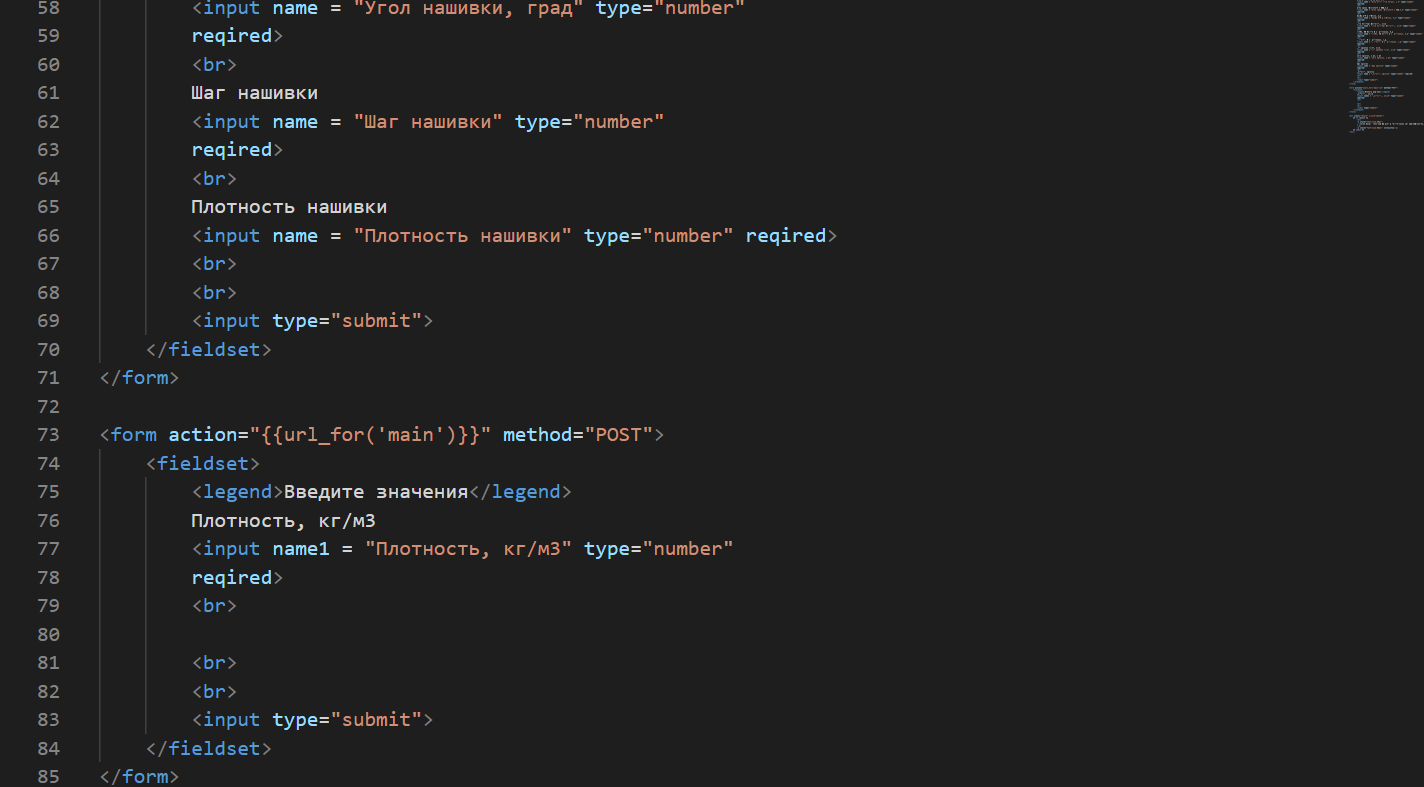
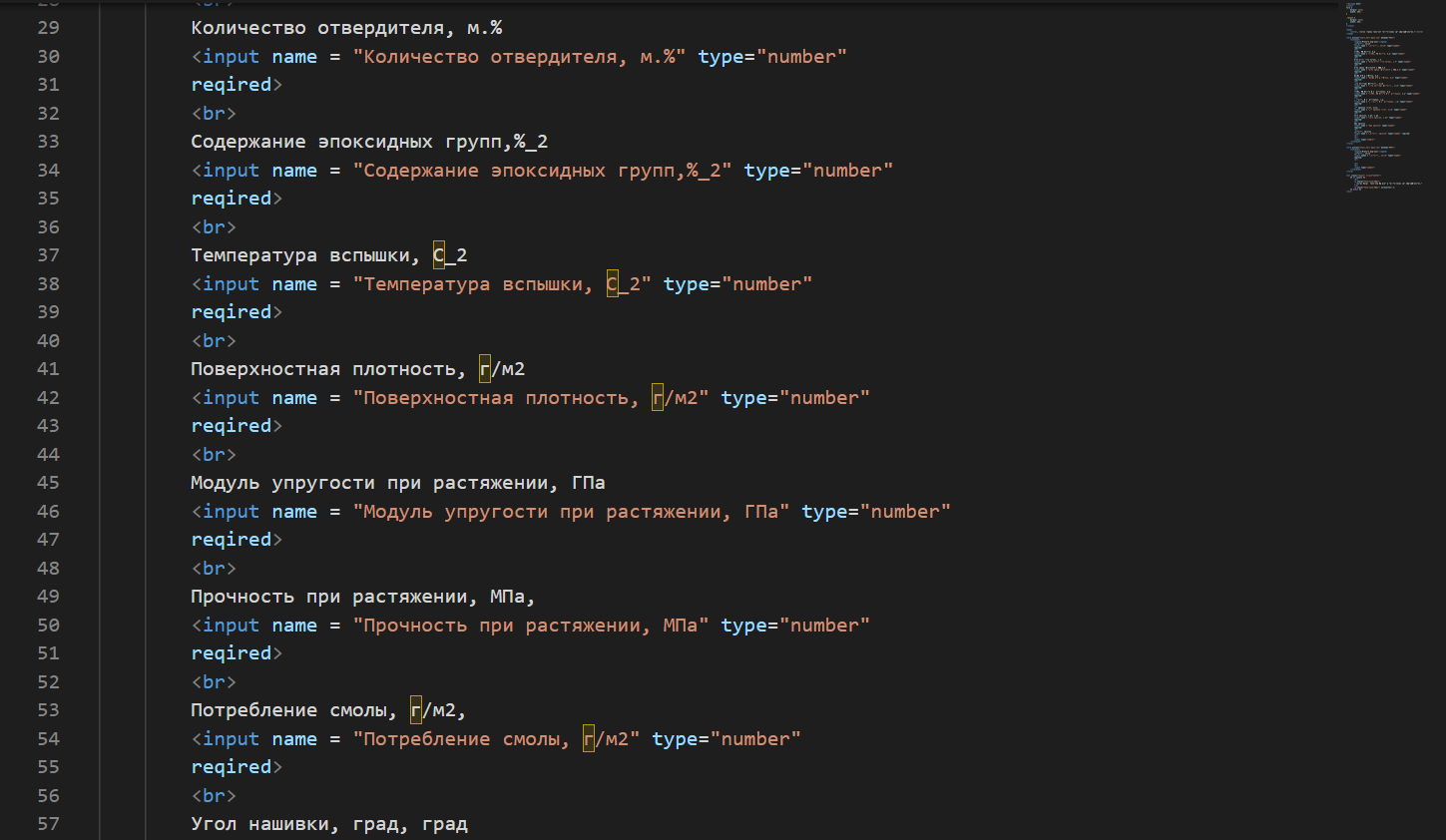
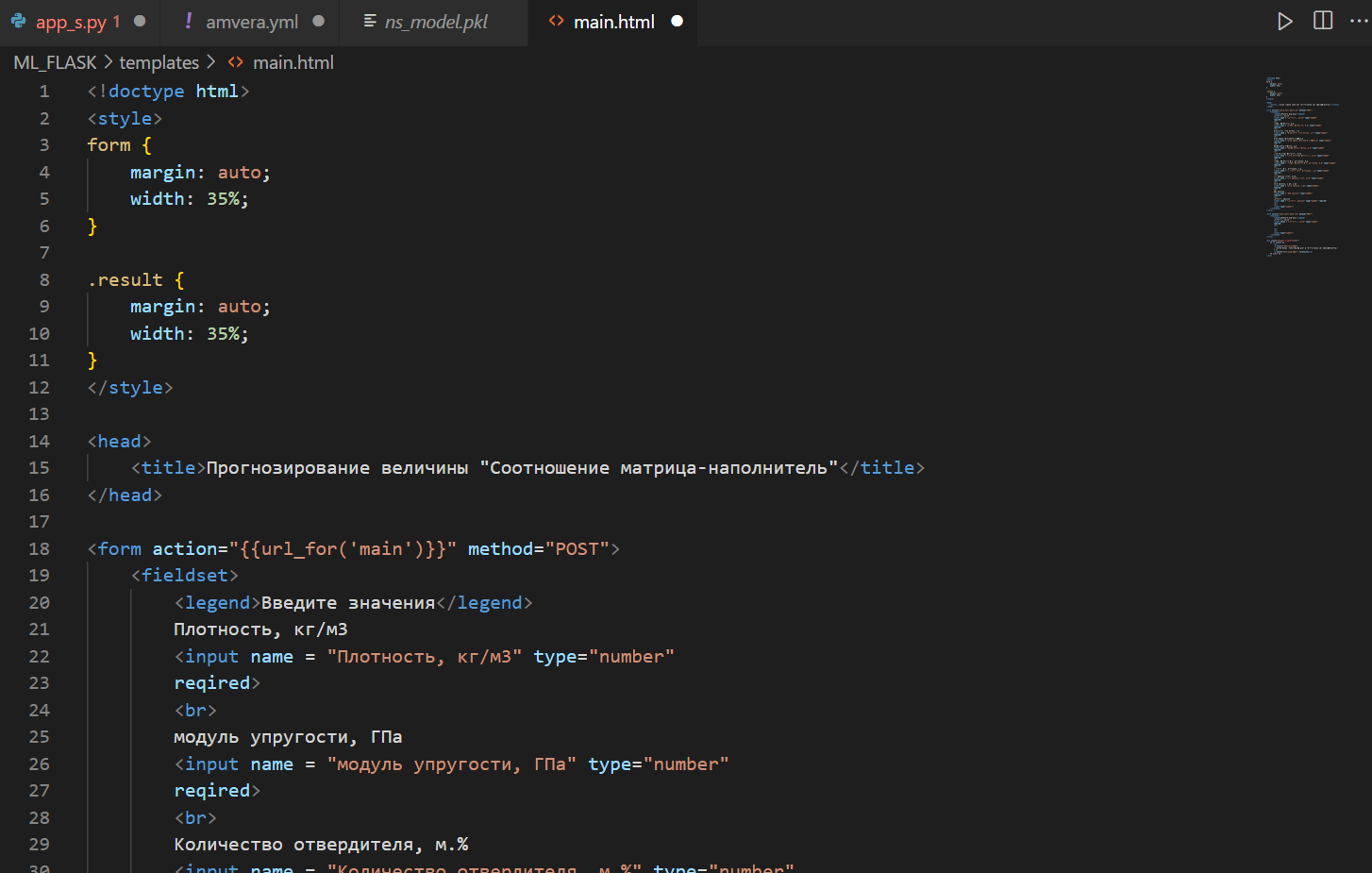
[**http://127.0.0.1:5000/index**](http://127.0.0.1:5000/index)





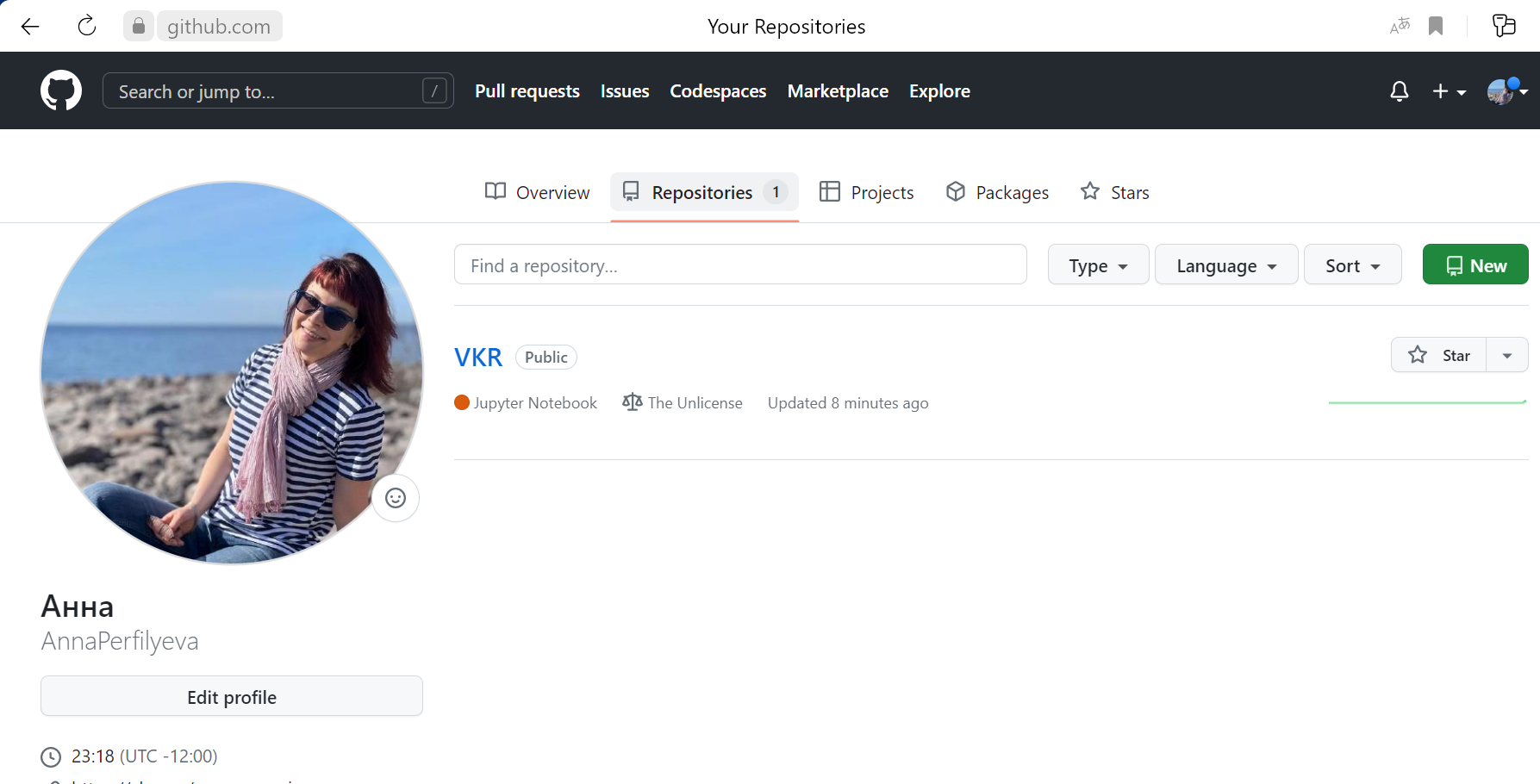
Данное приложение — это файл Flask, собранный из сохраненной модели нейросети, 

непосредственно файла с приложением app.py

файла-шаблона main.html.

**Создание удалённого репозитория GitHub и загрузка документов**

Репозиторий был создан на github.com по адресу: <https://github.com/AnnaPerfilyeva/VKR>



**Заключение**

Данное исследование проводилось последовательно через этапы:

1. Загрузка и обработка входящих данных
2. Объединение датасетов по методу INNER
3. Проведение разведочного анализа данных
4. Нормализация данных
5. Разделение данных на тестовую и обучающую выборки
6. Построение и сравнение регрессионных моделей
7. Построение нейронной сети
8. Сохранение модели для приложения
9. Разработка приложения
10. Сохранение репозитория
11. Выводы.

Разработанные модели регрессии и нейронная сеть не позволили получить прогнозов желаемой достоверности: все модели показывают либо отрицательные, либо близкие к нулю значения коэффициента детерминации. Для более точного прогноза необходимо дальнейшее, более углубленное изучение связи между признаками, получение более полной информации о физических свойствах материалов, консультация с экспертами в данной области, а также желательна работа с более подробными датасетами, расширенными по признакам и количеству данных.