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**

«DataScience»

**Тема:** «**Прогнозирование конечных свойств новых материалов**

**(композиционных материалов)»**

Слушатель Андрюшина М.В.

# 

Москва, 2023

# **Содержание**

[Содержание 2](#_Toc106232838)

[Введение 3](#_Toc106232839)

[1. Аналитическая часть 4](#_Toc106232840)

[1.1. Постановка задачи 4](#_Toc106232841)

[1.2. Описание используемых методов 6](#_Toc106232842)

[1.3. Разведочный анализ данных 19](#_Toc106232843)

[2. Практическая часть 26](#_Toc106232844)

[2.1. Предобработка данных 26](#_Toc106232845)

[2.2. Разработка и обучение модели 27](#_Toc106232846)

[2.3. Тестирование модели 29](#_Toc106232847)

[2.4. Написать нейронную сеть, которая будет рекомендовать соотношение «матрица-наполнитель» 30](#_Toc106232848)

[2.5. Разработка приложения](#_Toc106232849)

[2.6. Создание удалённого репозитория и загрузка 35](#_Toc106232850)

[2.7. Заключение 36](#_Toc106232851)

[2.8. Список используемой литературы и веб ресурсы. 37](#_Toc106232852)

# **Введение**

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. 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 (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. 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.

1. **Аналитическая часть**
   1. **Постановка задачи**

Для исследовательской работы были даны 2 файла: X\_bp.xlsx (с данными о параметрах базальтопластика, состоящий из 1023 строк и 11 столбцов) и X\_nup.xlsx (данными нашивок углепластика, состоящий из 1040 строк и 4 столбцов).

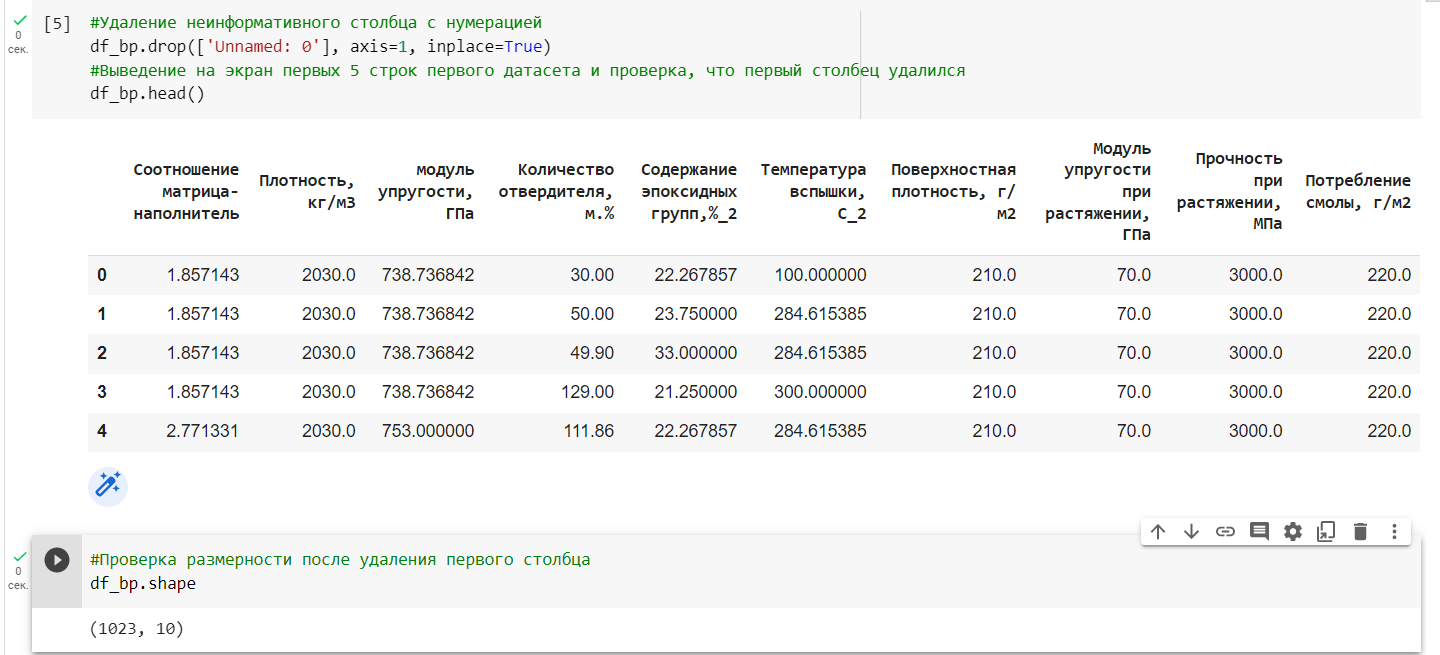


Рисунок 1 –процесс работы с файлом X\_bp.xlsx

Цель работы разработать модели для прогноза модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель». Для этого нужно объединить 2 файла. Часть информации (17 строк таблицы способов компоновки композитов) не имеют соответствующих строк в таблице соотношений и свойств используемых компонентов композитов, поэтому были удалены.

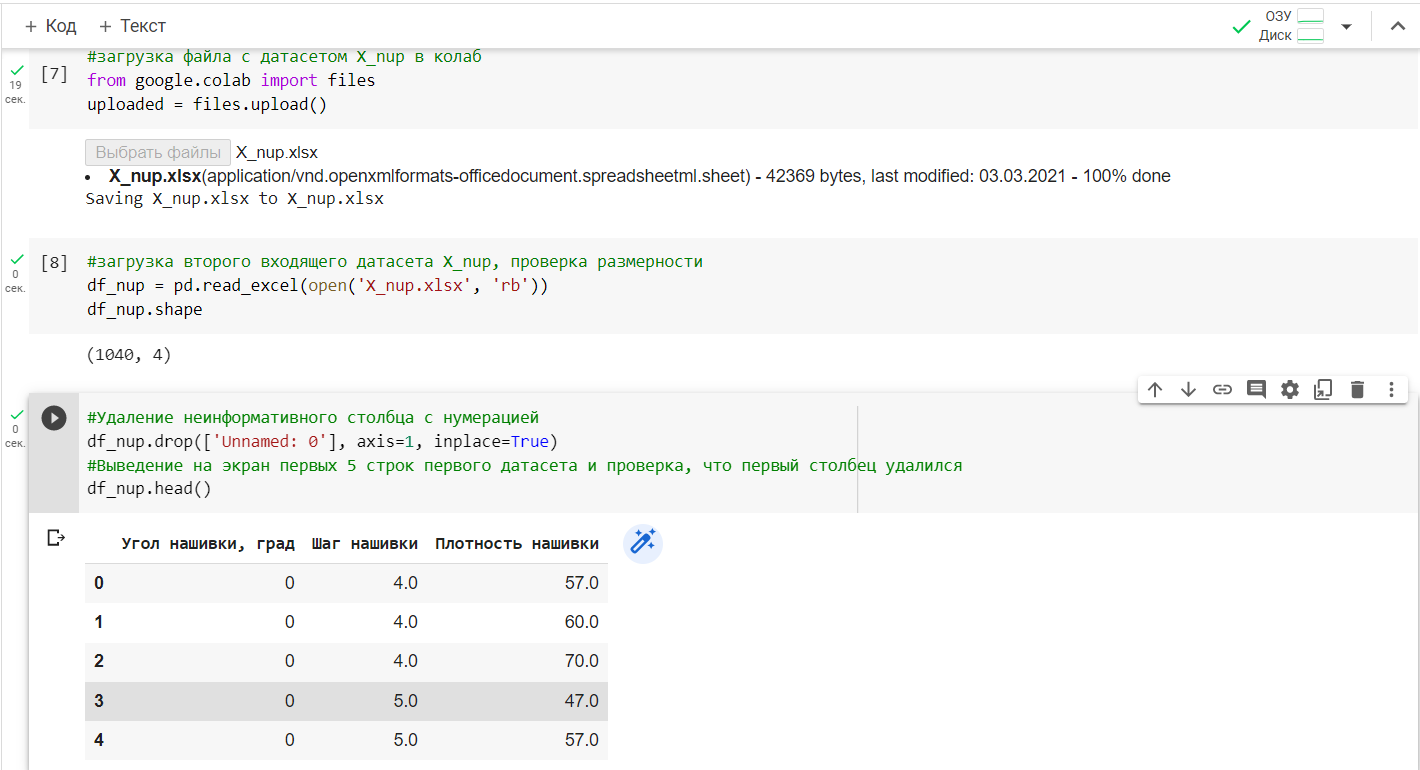


Рисунок 2 - начало работы с файлом X\_nup.xlsx

Затем провести разведочный анализ данных, нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек.



Рисунок 3 - диаграмма "ящик с усами" в объединённом датасете

Для каждой колонки 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 датасетах. Создать репозиторий в GitHub и разместить код исследования. Оформить файл README.

* 1. **Описание используемых методов**

Данная задача в рамках классификации категорий машинного обучения относится к машинному обучению с учителем и 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;
* градиентный бустинг;
* К-ближайших соседей;
* дерево решений;
* стохастический градиентный спуск;
* многослойный персептрон;
* Лассо регрессии;

Метод опорных векторов(Support Vector Regression - SVM) – это очень мощная и универсальная модель машинного обучения, способная выполнять линейную и нелинейную классификацию, регрессию и даже выявление выбросов. Методы SVM особенно хорошо подходят для классификации сложных, но небольших и средних наборов данных.

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 работе была проведена тщательнейшая работа 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 были рассмотрены и другие методы.

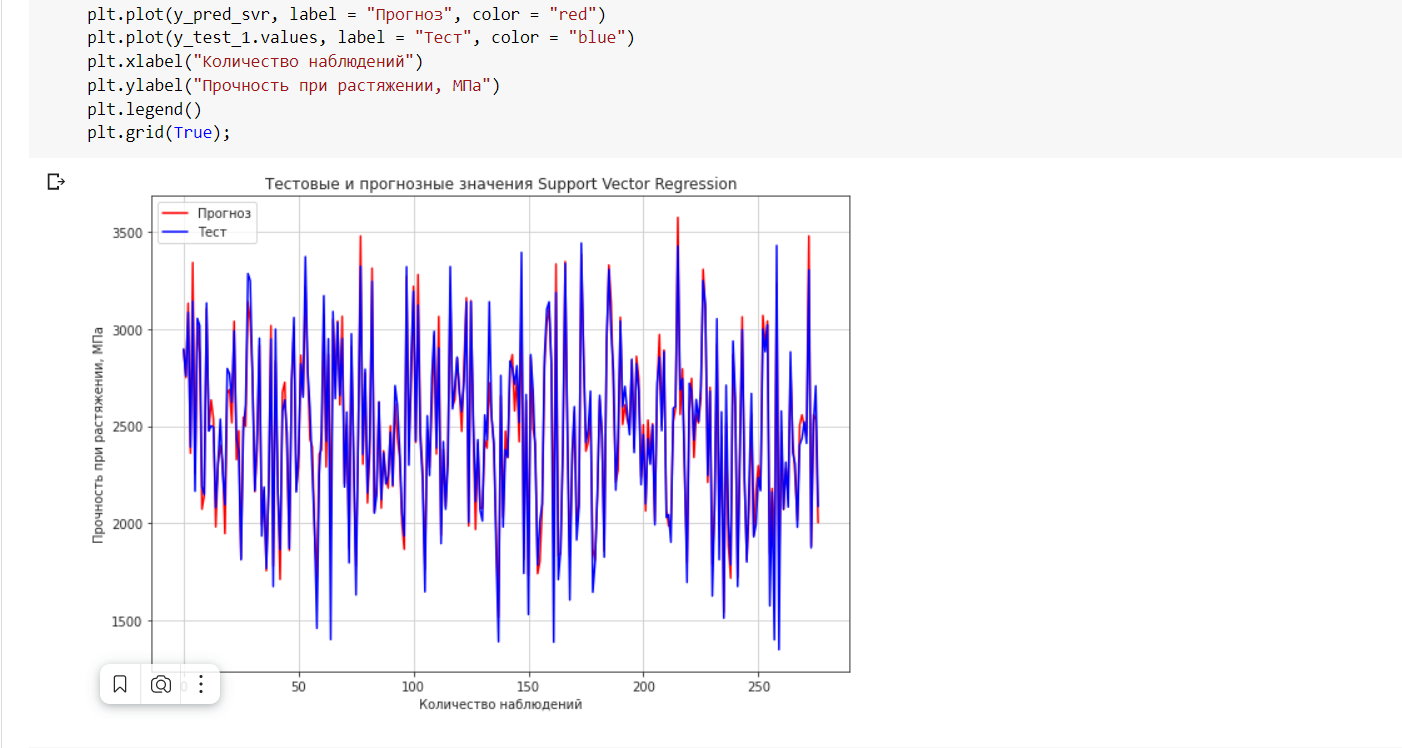


Рисунок 4- график метода опорных векторов для прочности при растяжении, МПа

Случайный лес (Random Forest)— это универсальный алгоритм машинного обучения с учителем, суть которого состоит в использовании ансамбля решающих деревьев. Само по себе решающее дерево представляет крайне невысокое качество классификации, но из-за большого их количества результат значительно улучшается. Также это один их немногих алгоритмов, который можно использовать в абсолютном большинстве задач.

Достоинства метода: имеет высокую точность предсказания, которая сравнима с результатами градиентного бустинга, не требует тщательной настройки параметров, практически не чувствителен к выбросам в данных, не чувствителен к масштабированию и к другим монотонным преобразованиям значений признаков, редко переобучается, в случае наличия проблемы переобучения, она преодолевается путем усреднения или объединения результатов различных деревьев решений, не требует предобработки входных данных, эффективно обрабатывает пропущенные данные, данные с большим числом классов и признако,; 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 недообучаться; трудоёмко прогнозируемый; иногда работает хуже, чем линейные методы.

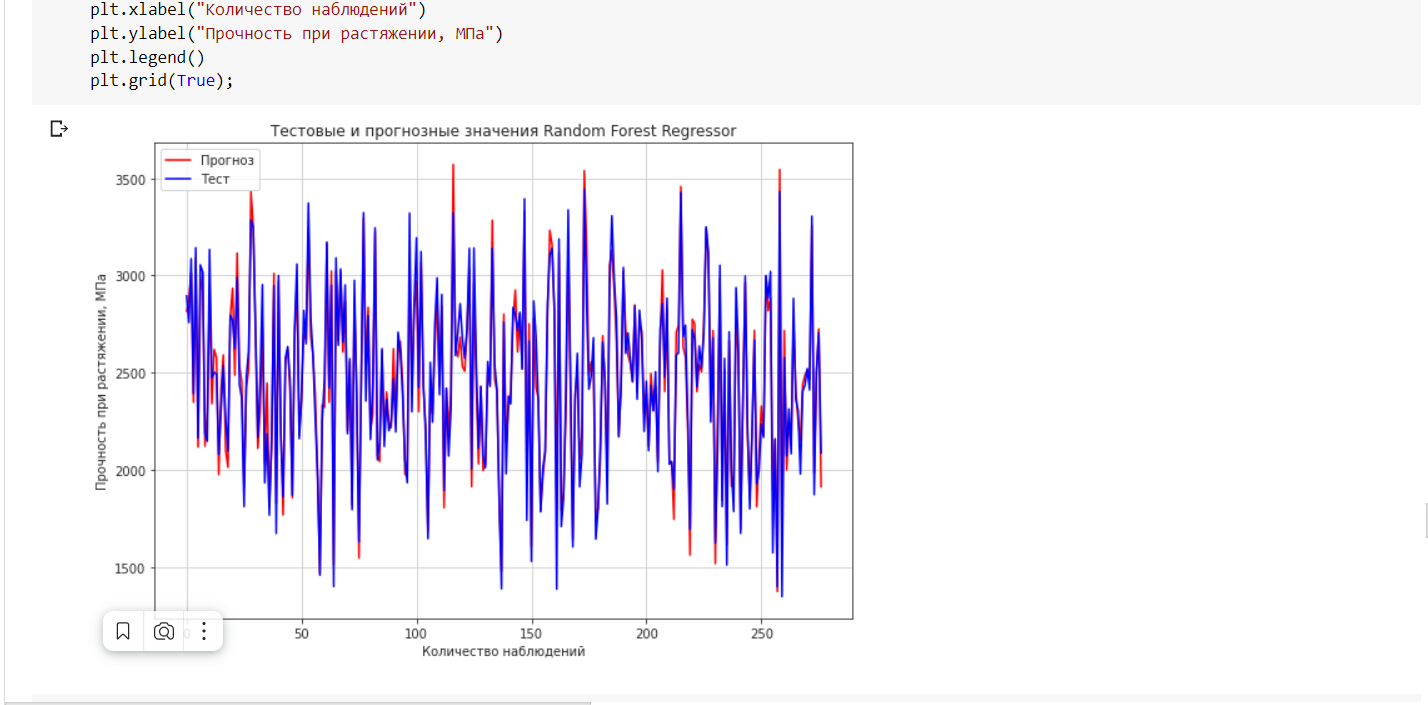


Рисунок 5 - график "случайного леса" для прочности при растяжении, МПа

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. R2 , 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 R2 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 прямую связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние, а границы линейны.

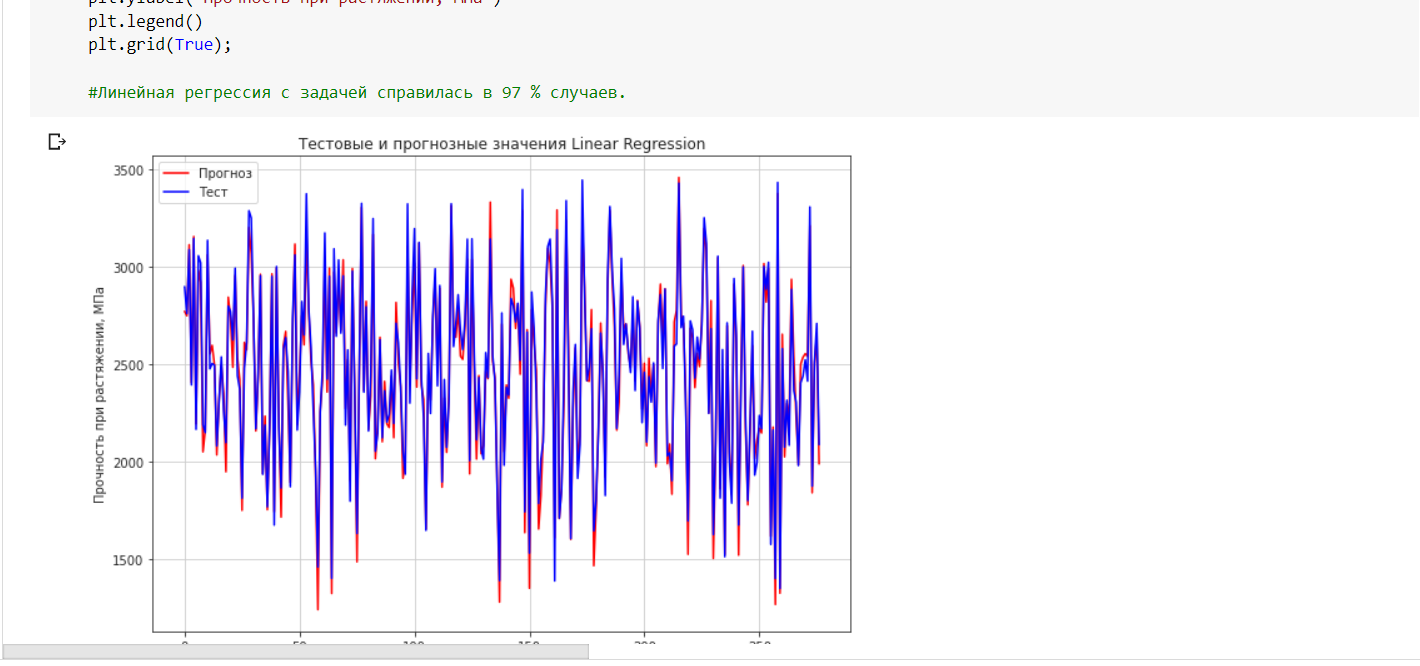


 Рисунок 6 график линейной регрессии для прочности при растяжении, МПа

Градиентный бустинг (GradientBoosting)— это ансамбль деревьев решений, обученный с использованием градиентного бустинга. Подобно AdaBoost градиентный бустинг работает, последовательно добавляя в ансамбль прогнозаторы, каждый из которых корректирует своего предшественника. Тем не менее, вместо подстройки весов образцов на каждой итерации, как делает AdaBoost, этот метод старается подогнать новый прогнозатор к остаточным ошибкам, допущенным предыдущим прогнозатором.

Достоинства метода: новые алгоритмы учатся на ошибках предыдущих; требуется меньше итераций, чтобы приблизиться к фактическим прогнозам; наблюдения выбираются на основе ошибки; прост в настройке темпа обучения и применения; легко интерпретируем.

Недостатки метода: необходимо тщательно выбирать критерии остановки, иначе это может привести к переобучению; наблюдения с наибольшей ошибкой появляются чаще; слабее и менее гибок чем нейронные сети.

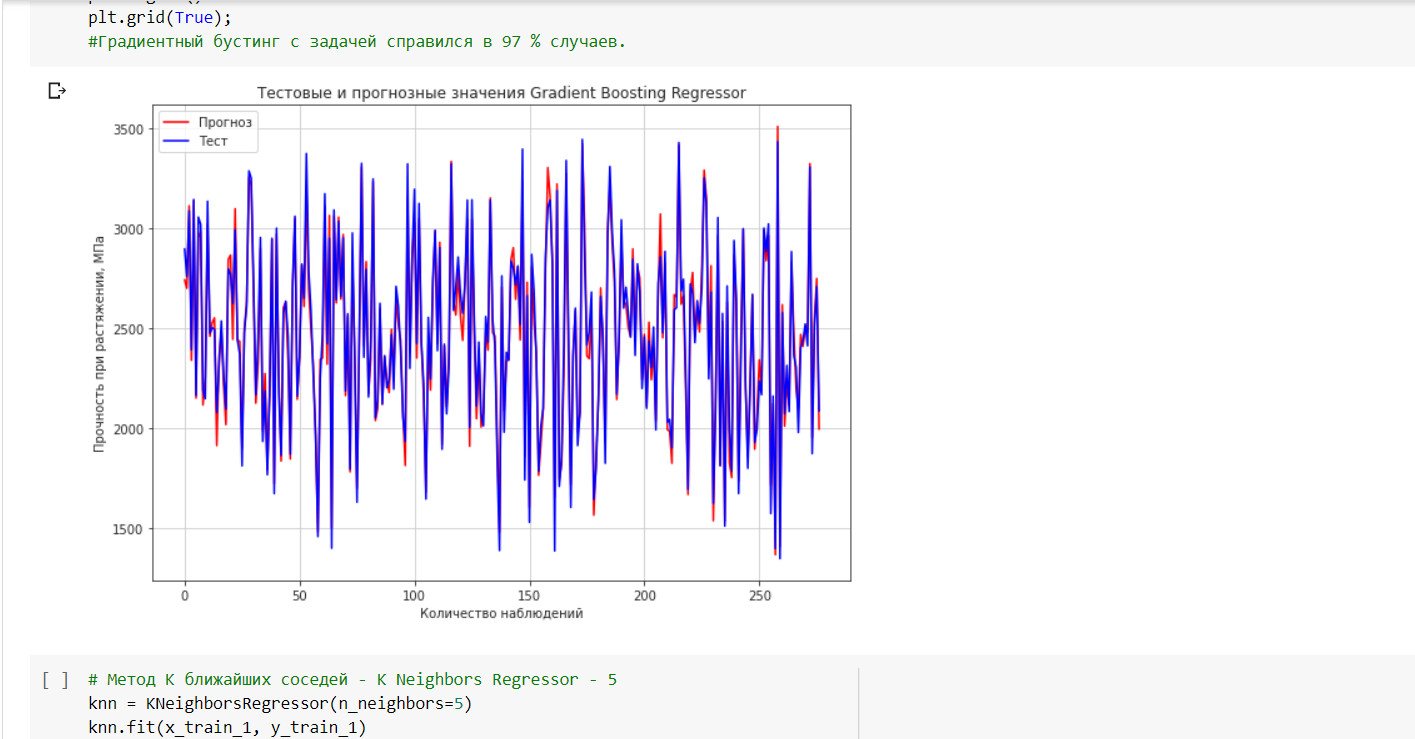


Рисунок 7 - график градиентного бустинга для прочности при растяжении, МПа

Метод ближайших соседей - К-ближайших соседей (kNN - kNearestNeighbours) - простейший метрический классификатор, основанный на оценивании сходства объектов. Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки. 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 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 ростомобъёма 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.

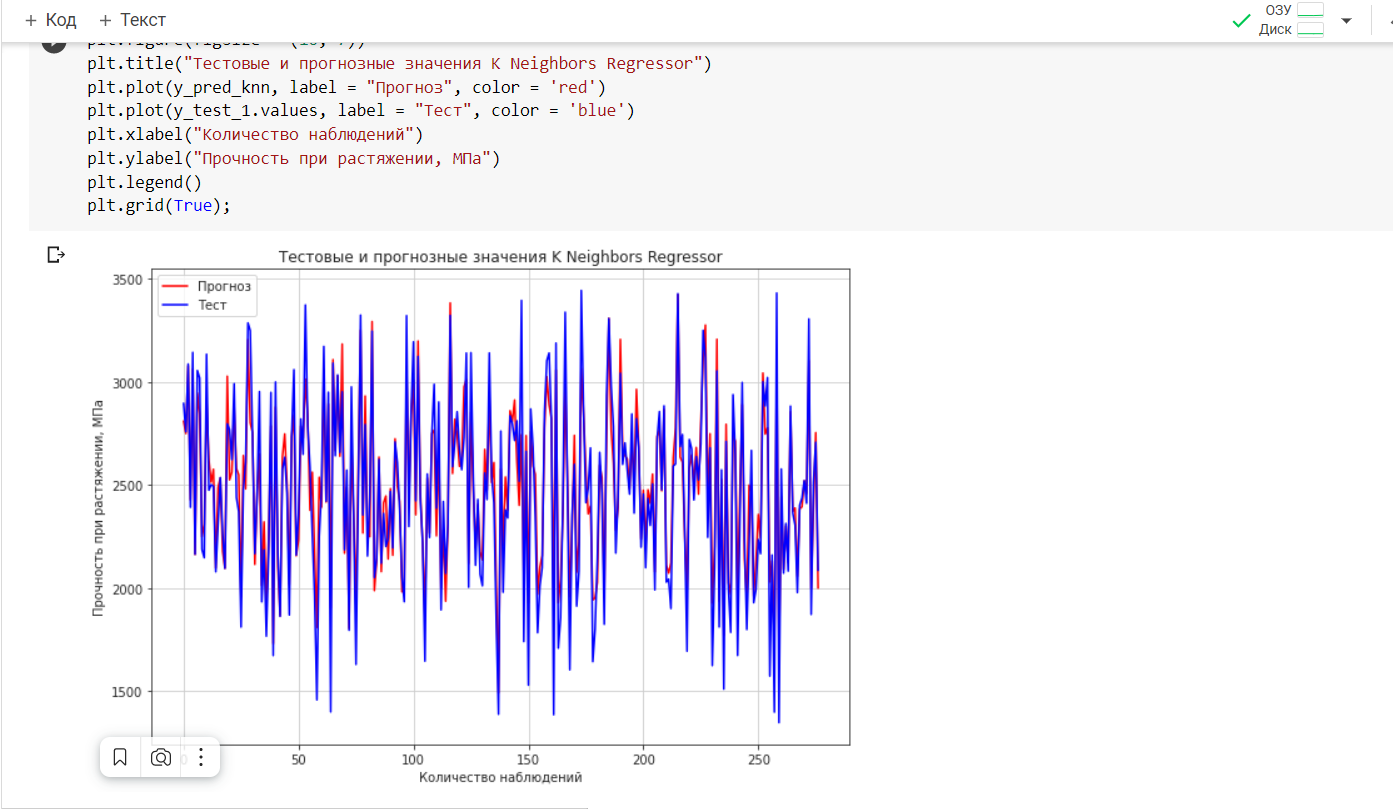


Рисунок 8 - график К-ближайших соседей для прочности при растяжении, Мпа

Дерево принятия решений (DecisionTreeRegressor) – подобно методам опорных векторов деревья принятия решений являются универсальными алгоритмами машинного обучения, которые могут заниматься задачами классификации и регрессии, включая даже многовыходовые задачи. Они представляют собой очень мощные алгоритмы, способные подгоняться к сложным наборам данных. Деревья принятия решений также являются фундаментальными компонентами случайных лесов, которые входят в число самых мощных алгоритмов машинного обучения, доступных на сегодняшний день. Это инструмент принятия решений, в котором используется древовидная структура, подобная блок-схеме, или модель решений и всех их возможных результатов, включая результаты, затраты и полезность. Дерево принятия решений- эффективный инструмент интеллектуального анализа данных и предсказательной аналитики. Алгоритм дерева решений подпадает под категорию контролируемых алгоритмов обучения. Он работает как для непрерывных, так и для категориальных выходных переменных. Правила генерируются за счёт обобщения множества отдельных наблюдений (обучающих примеров), описывающих предметную область.  Регрессия дерева решений отслеживает особенности объекта и обучает модель в структуре дерева прогнозированию данных в будущем для получения значимого непрерывного вывода. Дерево решений - один из вариантов решения регрессионной задачи, в случае если зависимость в данных не имеет очевидной корреляции.

Достоинства метода: помогают визуализировать процесс принятия решения и сделать правильный выбор в ситуациях, когда результаты одного решения влияют на результаты следующих решений; создаются по понятным правилам; просты в применении и интерпретации; заполняют пропуски в данных наиболее вероятным решением; работают с разными переменными; выделяют наиболее важные поля для прогнозирования;

Недостатки метода: ошибаются при классификации с большим количеством классов и небольшой обучающей выборкой; имеют нестабильный процесс (изменение в одном узле может привести к построению совсем другого дерева); имеет затратные вычисления; необходимо обращать внимание на размер; ограниченное число вариантов решения проблемы.

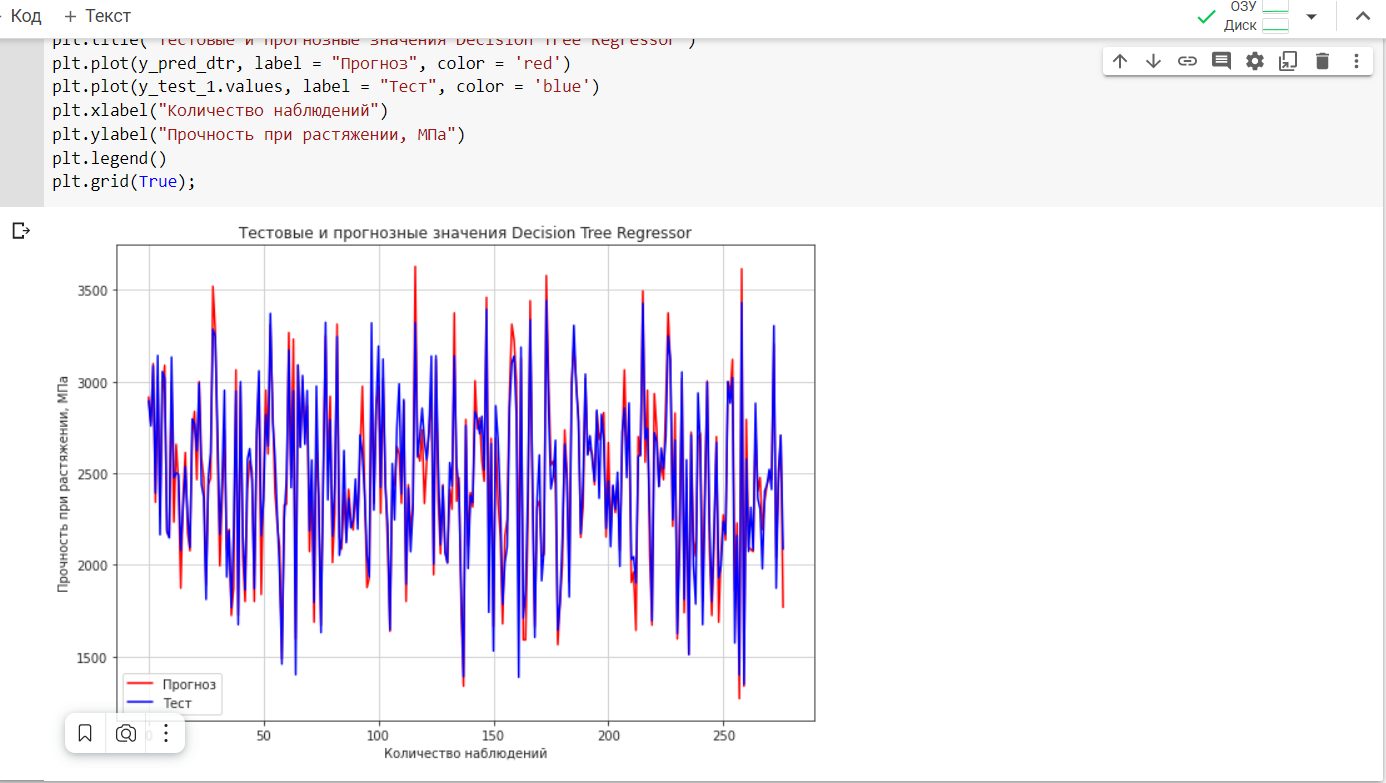


Рисунок 9- график дерева принятия решений для прочности при растяжении

Стохастический градиентный спуск (SGDRegressor) — это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под выпуклые функции потерь. На каждом шаге выбирает из обучающего набора случайный образец и вычисляет градиенты на основе только этого единственного образца. Очевидно, алгоритм становится гораздо быстрее, т.к. на каждой операции ему приходится манипулировать совсем малым объемом данных. Также появляется возможность проводить обучение на гигантских обучающих наборах, потому что на каждой итерации в памяти должен находиться только один образец. С другой стороны, из-за своей стохастической (т.е. случайной) природы этот алгоритм гораздо менее нормален, чем пакетный градиентный спуск: вместо умеренного понижения вплоть до достижения минимума функция издержек будет скачками изменяться вверх и вниз, понижаясь только в среднем. Со временем алгоритм будет очень близок к минимуму, но как только он туда доберется, скачкообразные изменения продолжатся, никогда не угомоняясь. Таким образом, после окончания алгоритма финальные значения параметров оказываются хорошими, но не оптимальными.

Достоинства метода: эффективен; простой в реализации; имеет множество возможностей для настройки кода; способен обучаться на избыточно больших выборках.

Недостатки метода: требует ряд гиперпараметров; чувствителен к масштабированию функций; может не сходиться или сходиться слишком медленно; функционал многоэкстремален; процесс может "застрять" в одном из локальных минимумов; возможно переобучение.

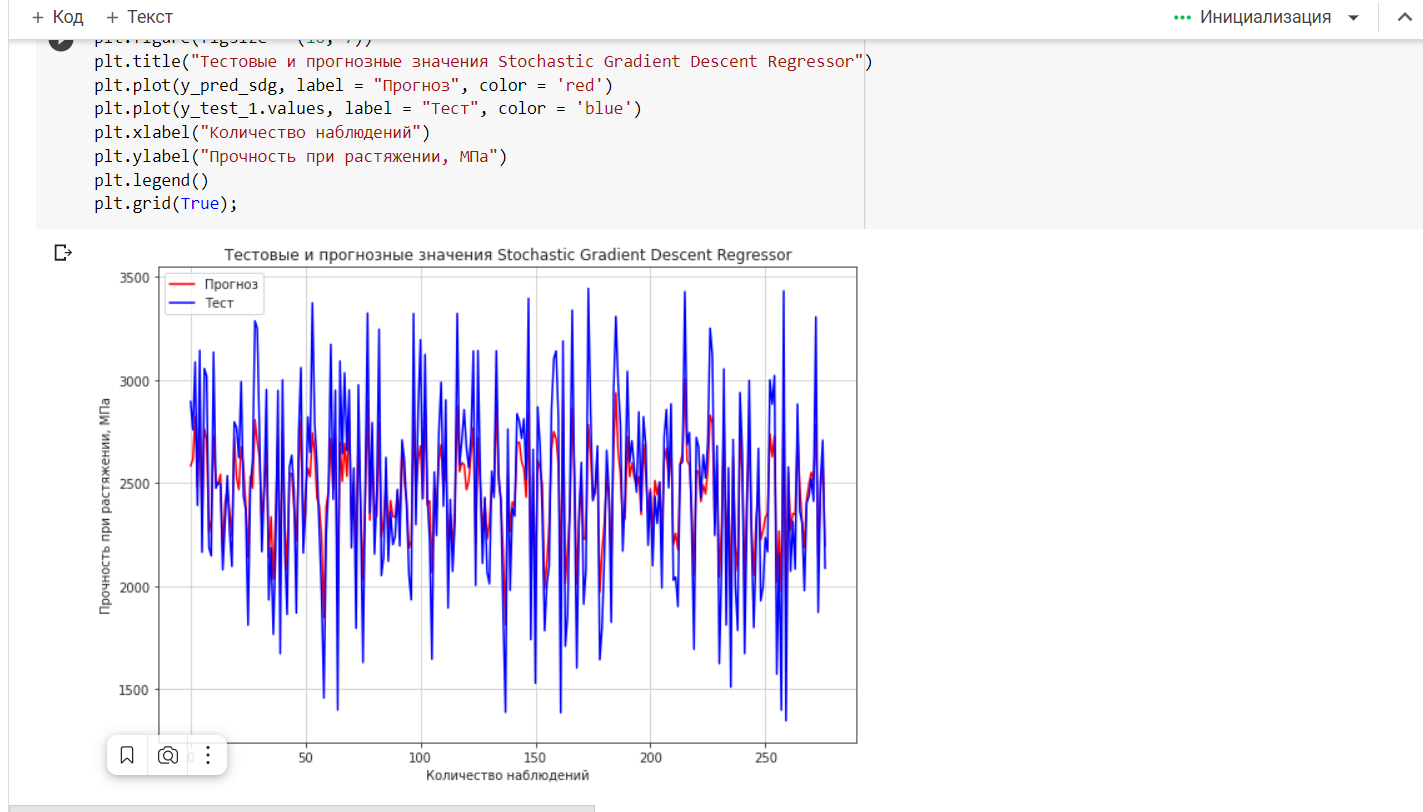


Рисунок 10-график стохастического градиентного спуска для прочности при растяжении

Многослойный персептрон (MLPRegressor) — это алгоритм обучения с учителем, который изучает функцию f(⋅):Rm→Ro обучением на наборе данных, где m — количество измерений для ввода и o- количество размеров для вывода. Это искусственная нейронная сеть, имеющая 3 или более слоёв персептронов. Эти слои - один входной слой, 1 или более скрытых слоёв и один выходной слой персептронов.

Достоинства метода: построение сложных разделяющих поверхностей; возможность осуществления любого отображения входных векторов в выходные; легко обобщает входные данные; не требует распределения входных векторов; изучает нелинейные модели.

Недостатки метода: имеет невыпуклую функцию потерь; разные инициализации случайных весов могут привести к разной точности проверки; требует настройки ряда гиперпараметров; чувствителен к масштабированию функций.

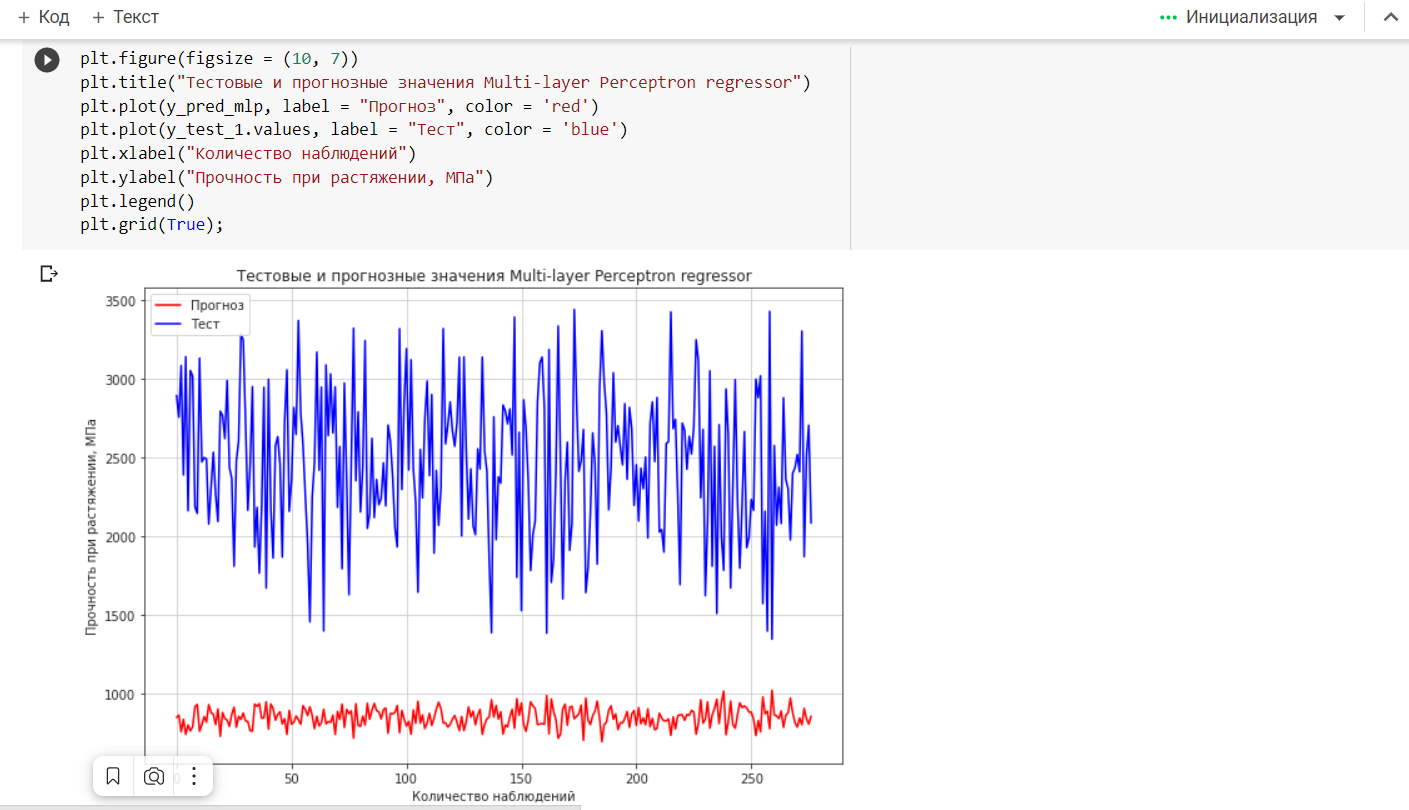


Рисунок 11-график многослойного персептрона для модуля прочности при растяжении

Лассо - регрессия (Lasso)— это регуляризированная версия линейной регрессии, которая оценивает разреженные коэффициенты. Это простой метод, позволяющий уменьшить сложность модели и предотвратить переопределение, которое может возникнуть в результате простой линейной регрессии. Данный метод вводит дополнительное слагаемое регуляризации в оптимизацию модели. Это даёт более устойчивое решение. В регрессии лассо добавляется условие смещения в функцию оптимизации для того, чтобы уменьшить коллинеарность и, следовательно, дисперсию модели. Но вместо квадратичного смещения, используется смещение абсолютного значения. Лассо регрессия хорошо прогнозирует модели временных рядов на основе регрессии, таким как авторегрессии. Важной характеристикой лассо – регрессии является то, что она стремится полностью исключить веса наименее важных признаков (т.е. устанавливает их в ноль).

Достоинства метода: легко полностью избавляется от шумов в данных; быстро работает; не очень энергоёмко; способно полностью убрать признак из датасета; доступно обнуляет значения коэффициентов.

Недостатки метода: выбор модели не помогает и обычно вредит; часто страдает качество прогнозирования; выдаёт ложное срабатывание результата; случайным образом выбирает одну из коллинеарных переменных; не оценивает правильность формы взаимосвязи между независимой и зависимой переменными; не всегда лучше, чем пошаговая регрессия.

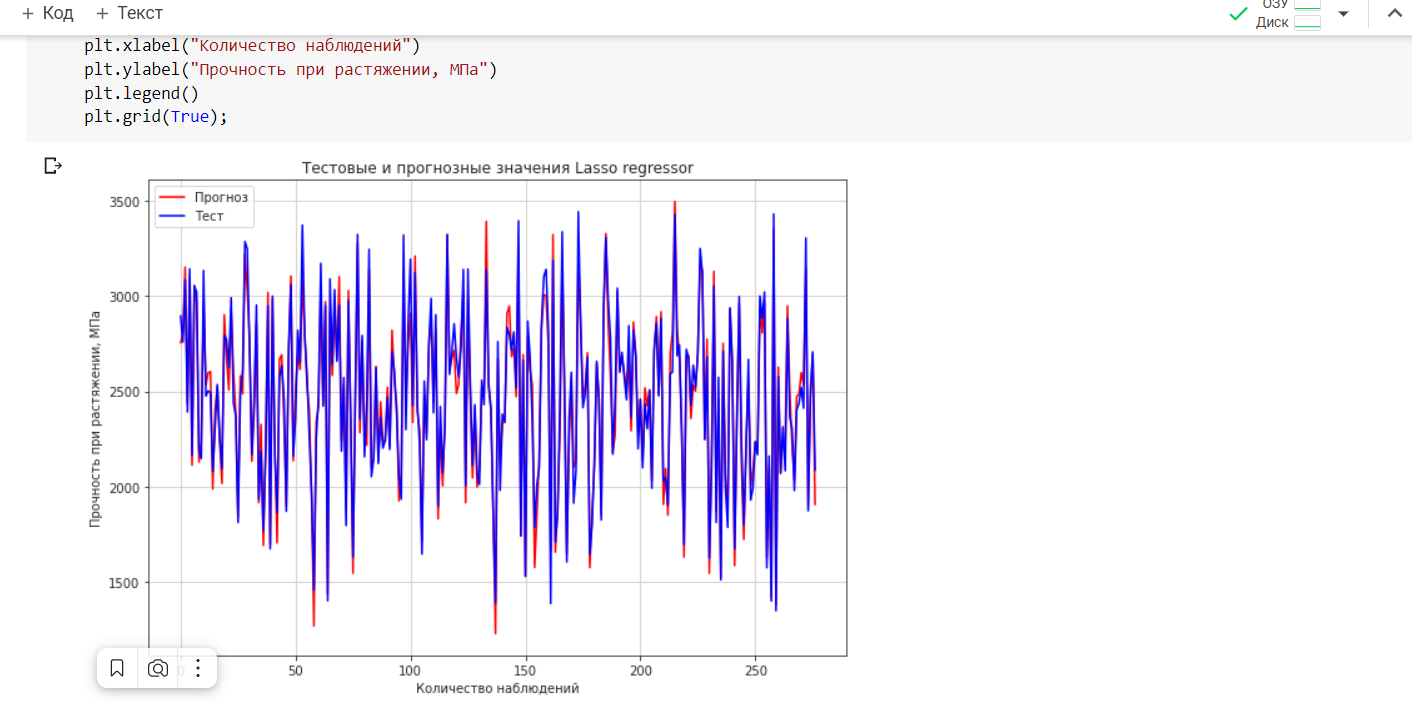


Рисунок 12-график метода Лассо для прочности при растяжении

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

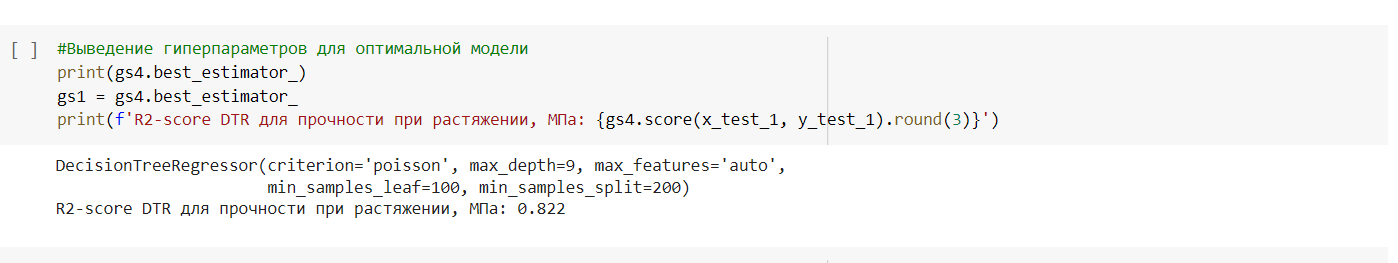


Рисунок 13 – часть кода для результата метрики R2 для метода Дерево принятия решений

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 (MeanSquaredError) 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.

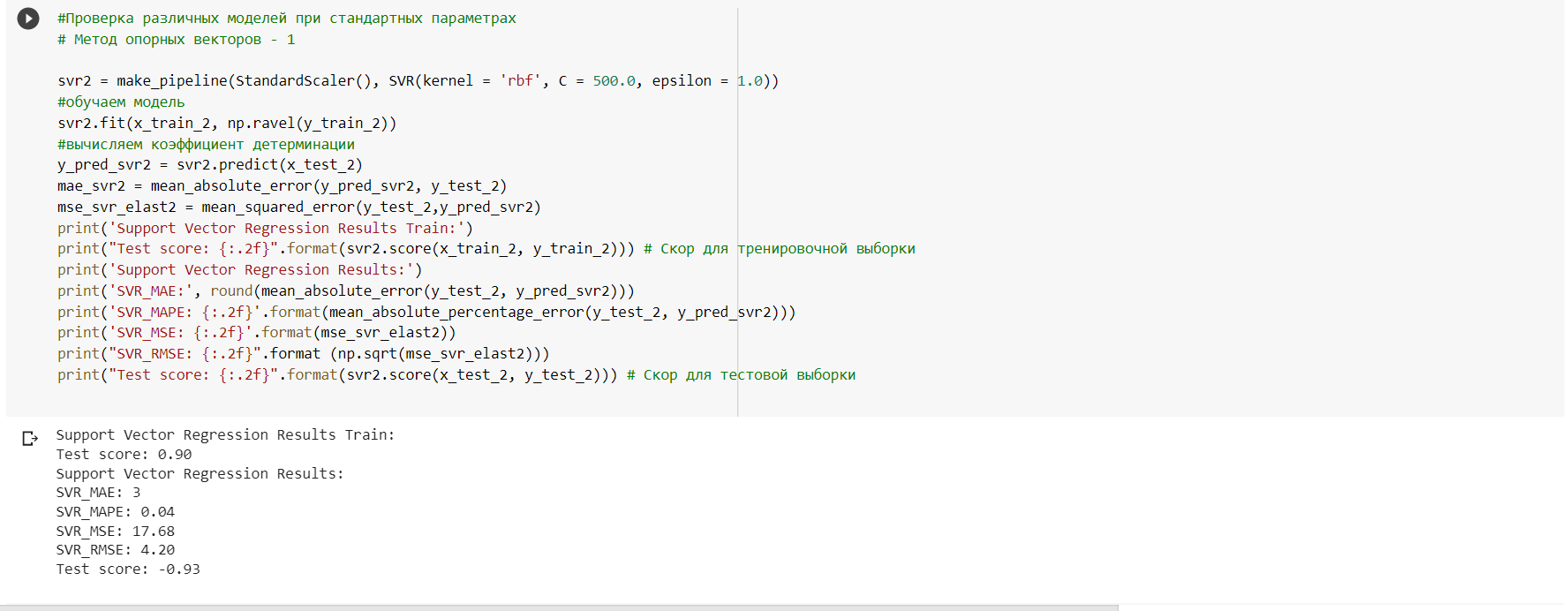


Рисунок 14 - код для вывода различных метрик для метода опорных векторов

* 1. **Разведочный анализ данных**

Разведочный анализ данных (Exploratory Data Analysis) – предварительное исследование датасета с целью определения его основных характеристик, взаимосвязей между признаками, а также сужения набора методов, используемых для создания модели машинного обучения.

Разведочный анализ — это предварительный анализ данных с целью выявления наиболее общих зависимостей, закономерностей и тенденций, характера и свойств анализируемых данных, законов распределения анализируемых величин. Применяется для нахождения связей между переменными в ситуациях, когда отсутствуют (или недостаточны) априорные представления о природе этих связей.

Как правило, при разведочном анализе учитывается и сравнивается большое число признаков, а для поиска закономерностей используются самые разные методы.

Результаты разведочного анализа не используются для выработки управленческих решений. Их назначение — помощь в разработке наилучшей стратегии углубленного анализа, выдвижение гипотез, уточнение особенностей применения тех или иных математических методов и моделей. Без разведочного анализа углубленный анализ данных будет производиться практически «вслепую».

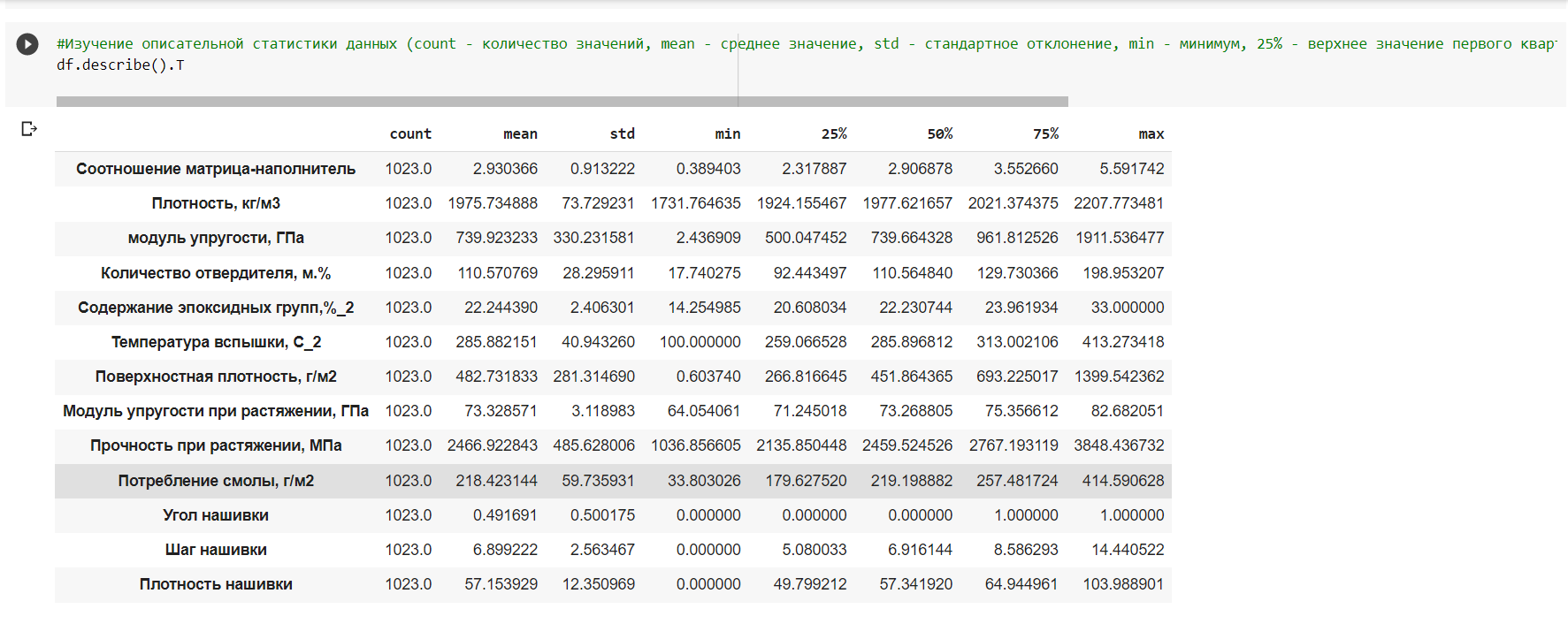


Рисунок 15 - описательная статистика датасета

В первую очередь для анализа данных обычно применяются гистограммы и "ящики с усами". Гистограмма просто разбивает весь диапазон данных на несколько отрезков, и для каждого отрезка выводит количество элементов набора данных, попадающих в этот отрезок. Обычно выводятся гистограммы сразу для нескольких признаков.

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.

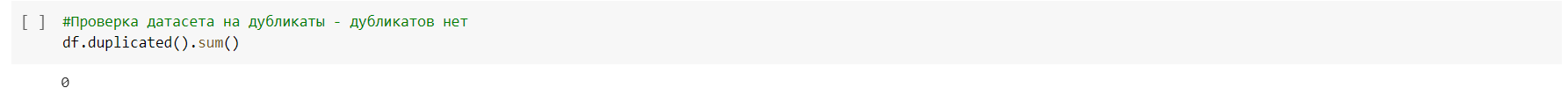


Рисунок 16 - проверка датасета на наличие дубликатов

В качестве инструментов разведочного анализа используется: оценка статистических характеристик датасета; гистограммы распределения каждой из переменной (несколько различных вариантов); диаграммы ящика с усами (несколько интерактивных вариантов); попарные графики рассеяния точек (несколько вариантов);график «квантиль-квантиль»; тепловая карта (несколько вариантов); описательная статистика для каждой переменной; анализ и полное исключение выбросов (несколько повторных итераций); проверка наличия пропусков и дубликатов; ранговая корреляция Пирсона и Спирмена ( от 0,7 до 1 – сильная корреляция).

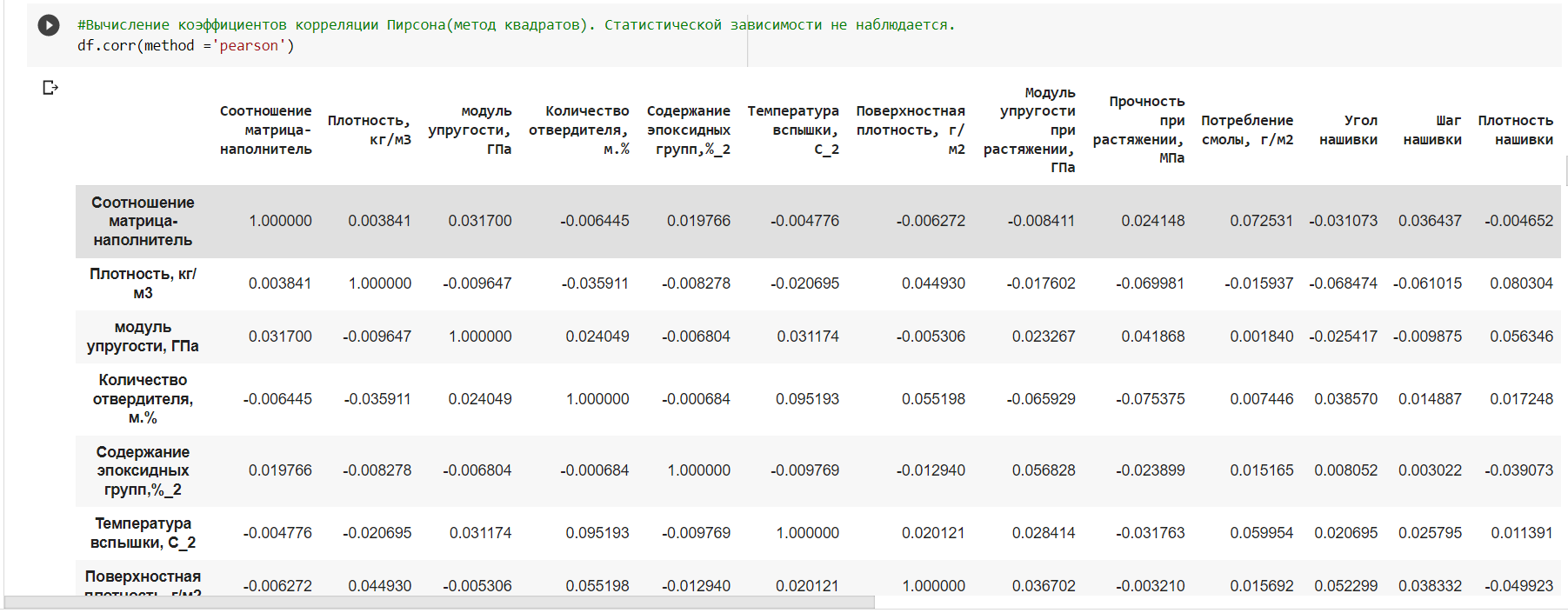


Рисунок 17 – вычисление коэффициентов корреляции Пирсона

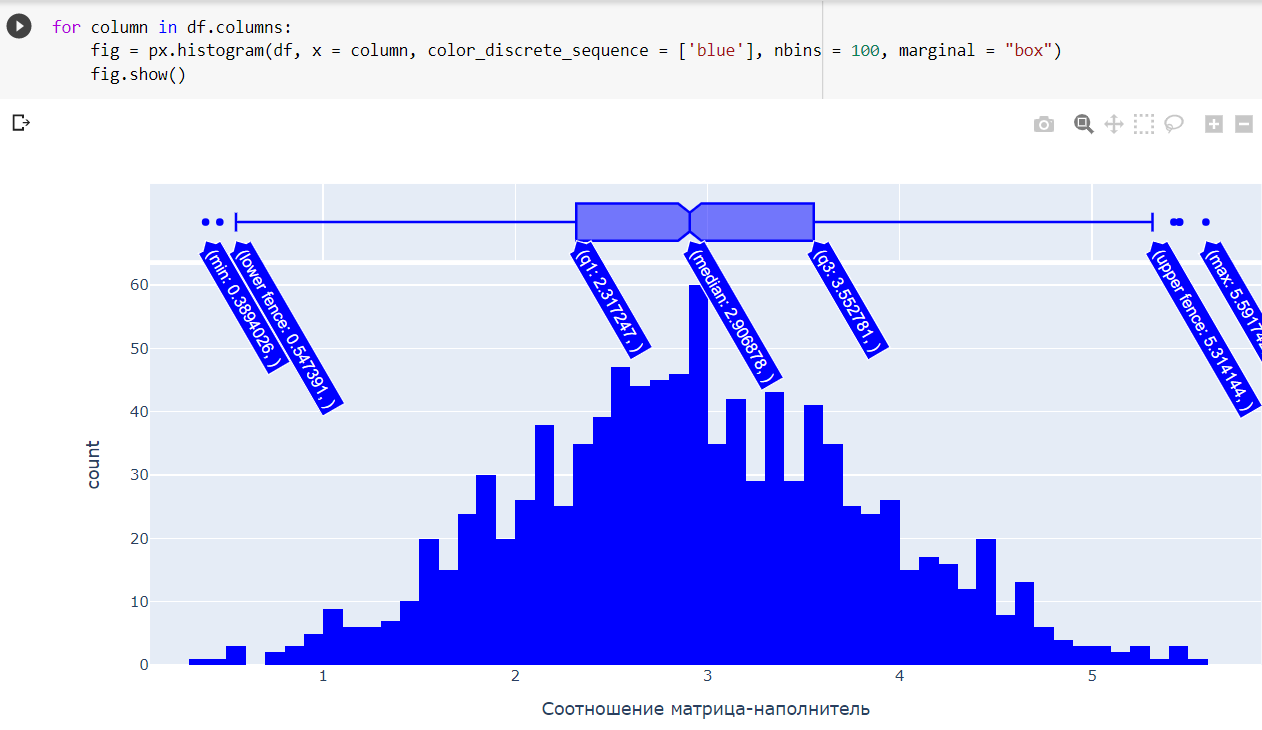


Рисунок – интерактивная гистограмма распределения и коробчатая диаграмма (ящик с усами)

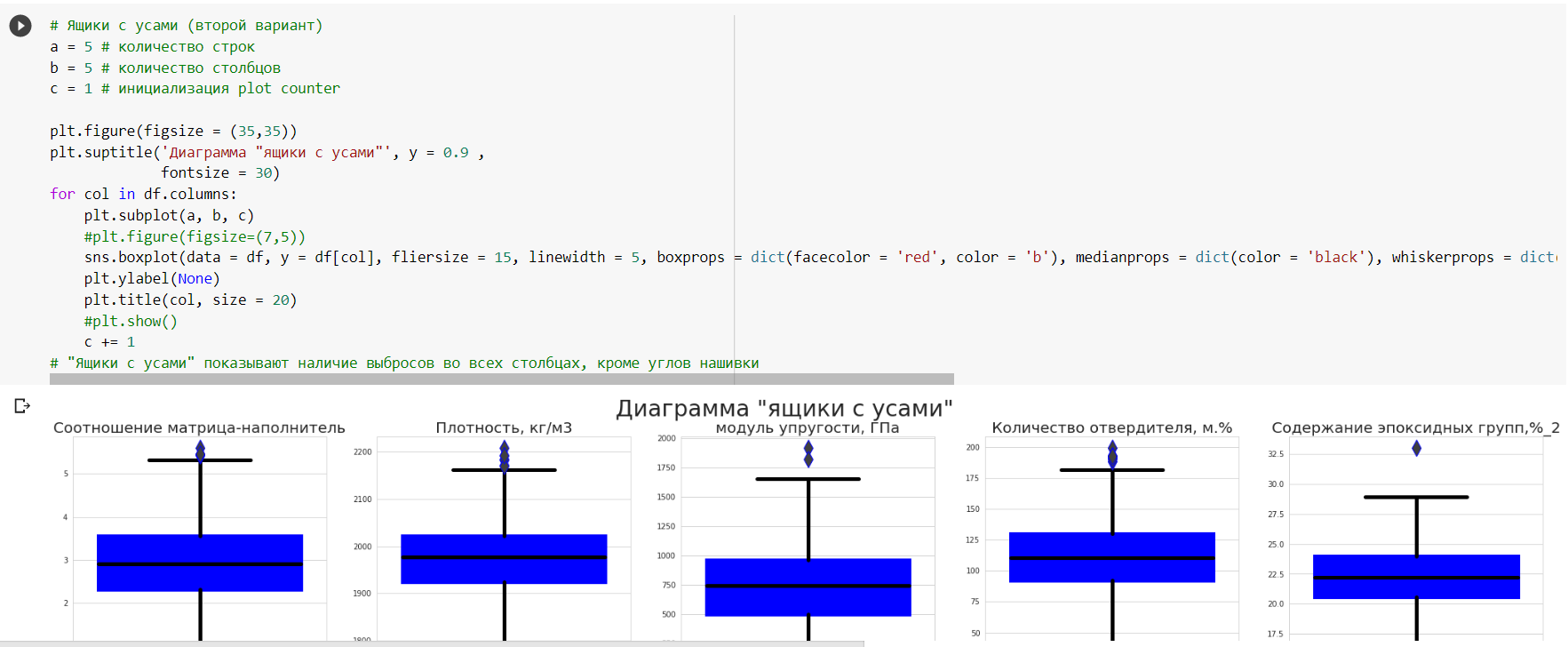


Рисунок 19 –коробчатая диаграмма (ящик с усами)

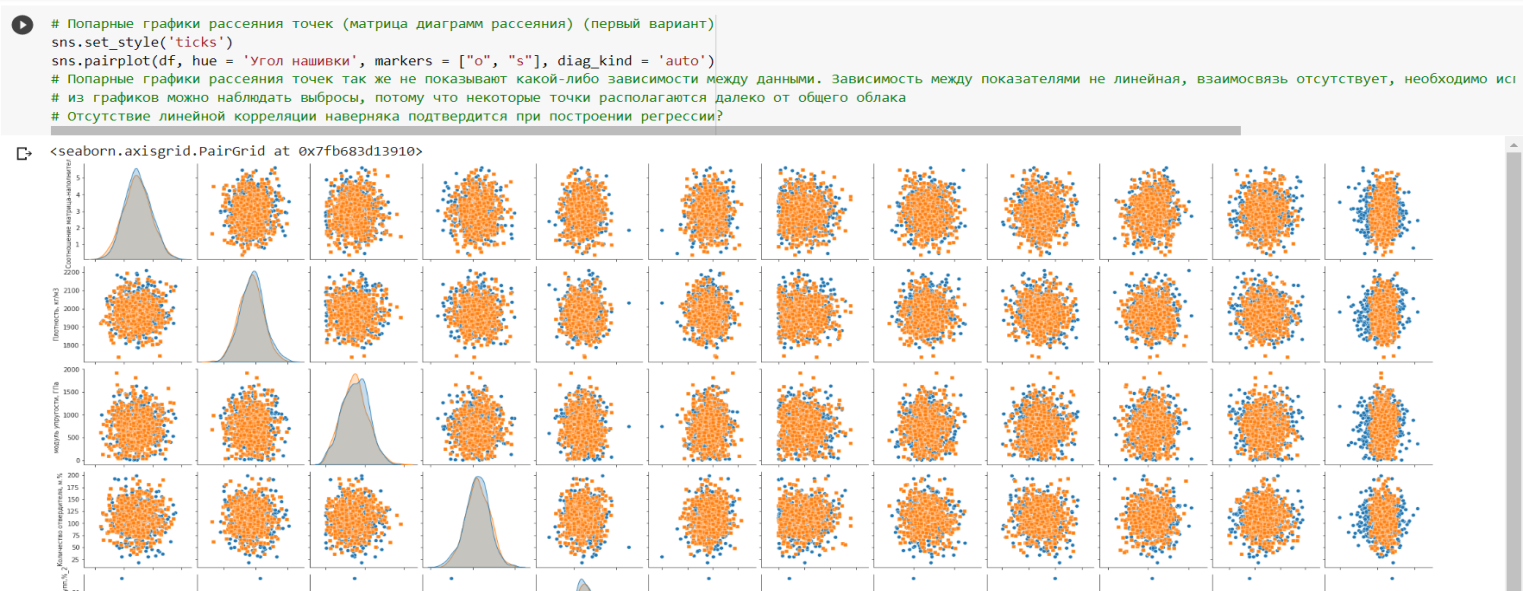


Рисунок 20 - попарные графики рассеяния точек (вариант 1)

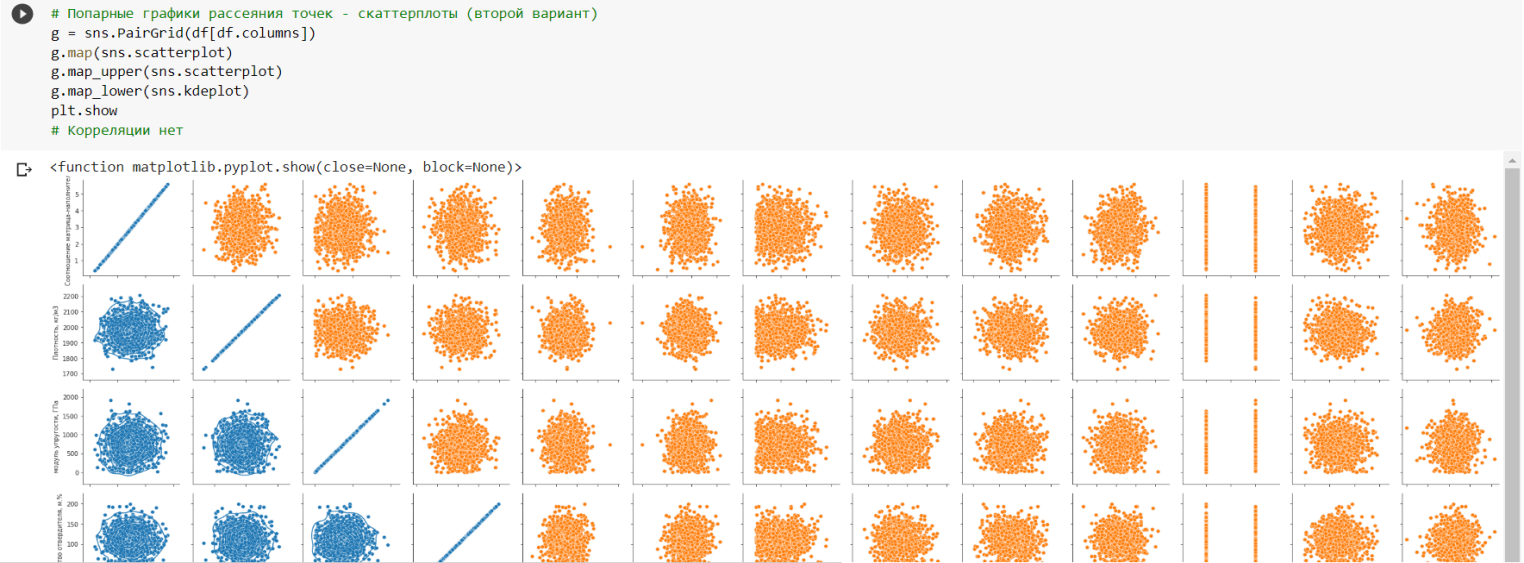


Рисунок 21 - попарные графики рассеяния точек (вариант 2)

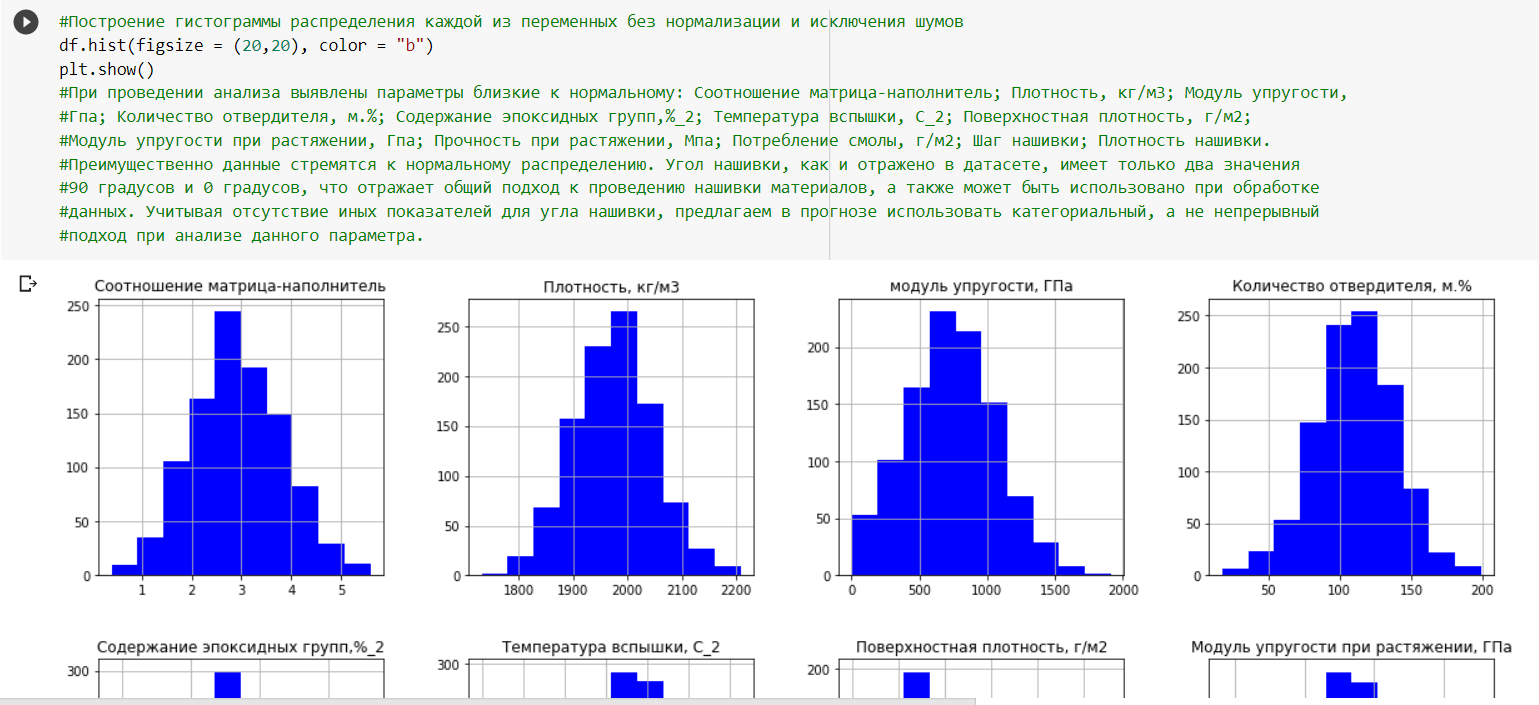


Рисунок - гистограмма распределения (вариант 1)

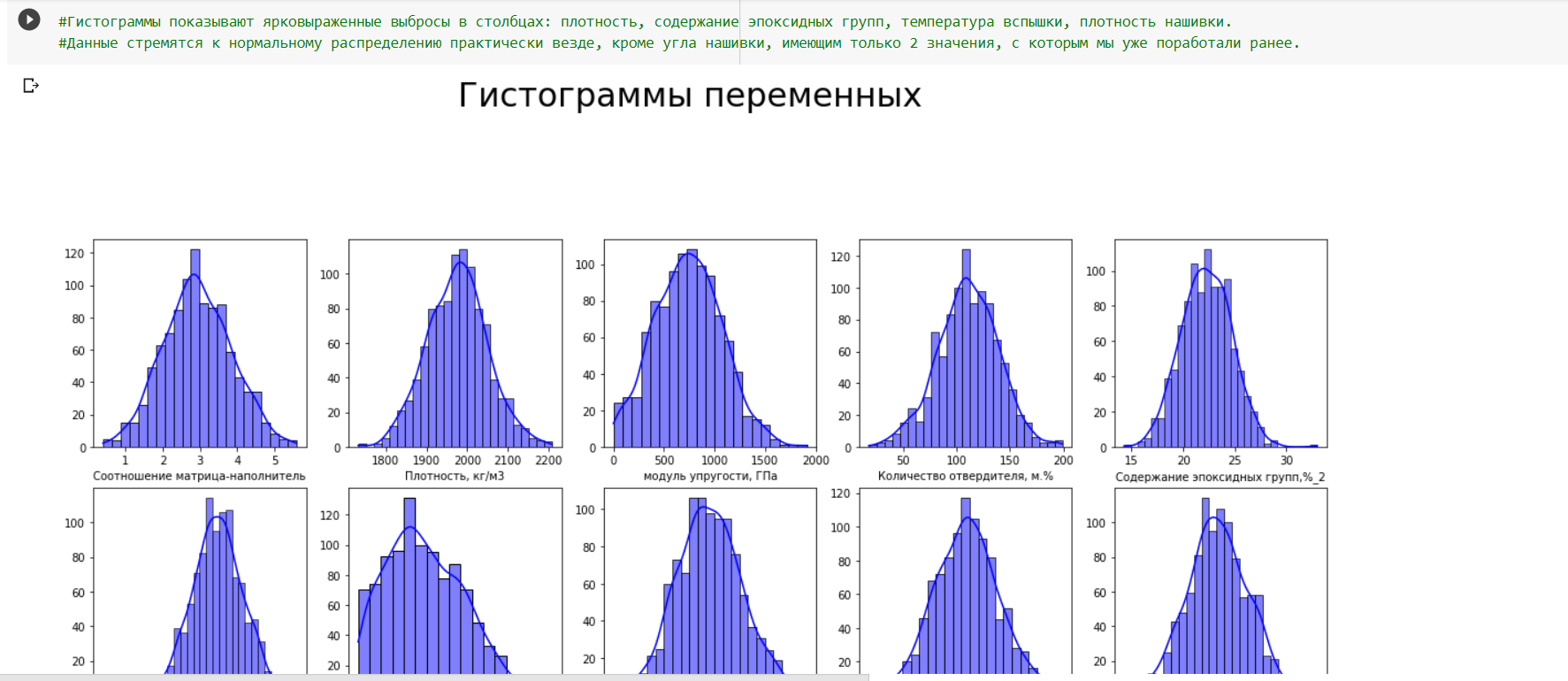
****

Рисунок - гистограмма распределения (вариант 2)

Гистограммы используются для изучения распределений частот значений переменных. Мы видим очень слабую корреляцию между переменными.

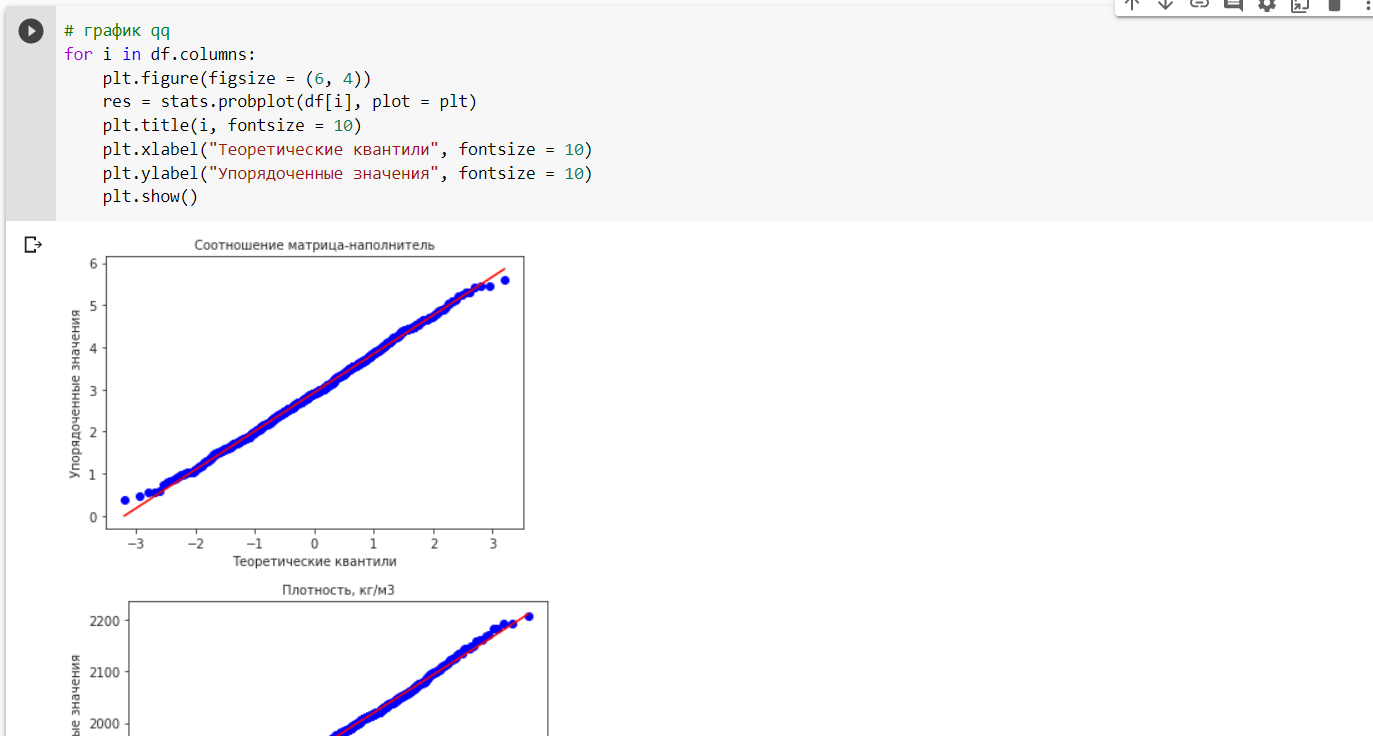


Рисунок 24 - график «квантиль-квантиль»

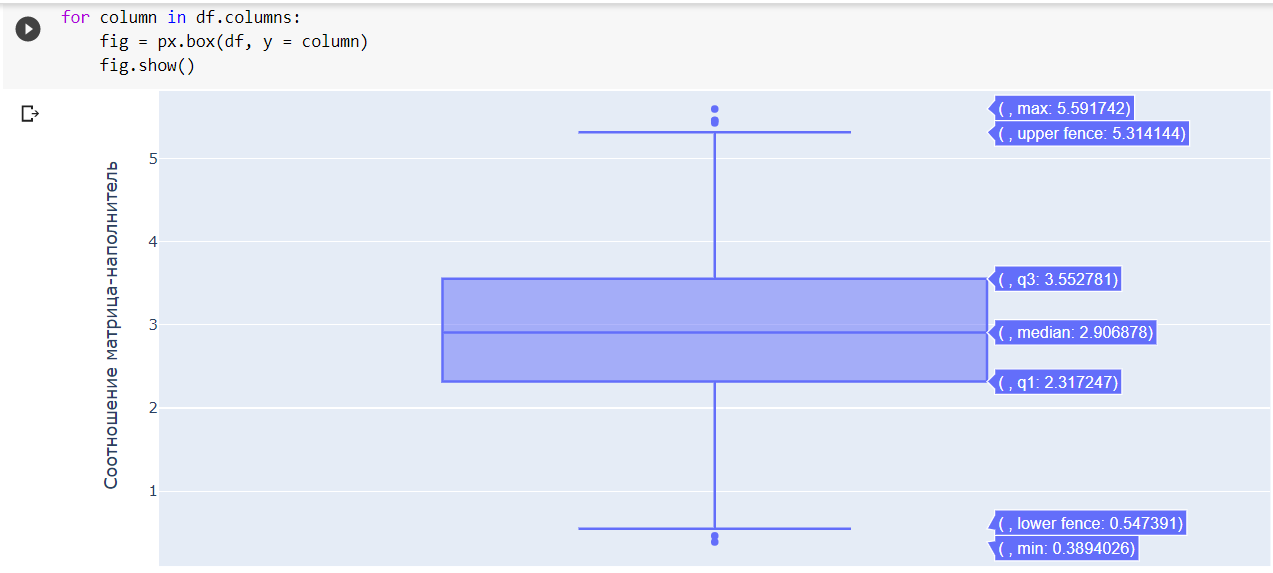


Рисунок 25 - график "ящиков с усами" для одной переменной

После обнаружения выбросов данные, значительно отличающиеся от выборки, будут полностью удалены. Для расчёта этих данных мы будем использовать методы трех сигм и межквартильного расстояния.

Данные объединённого датасета не имеют чётко выраженной зависимости, что подтверждает тепловая карта с матрицей корреляции и матрицы диаграмм рассеяния.

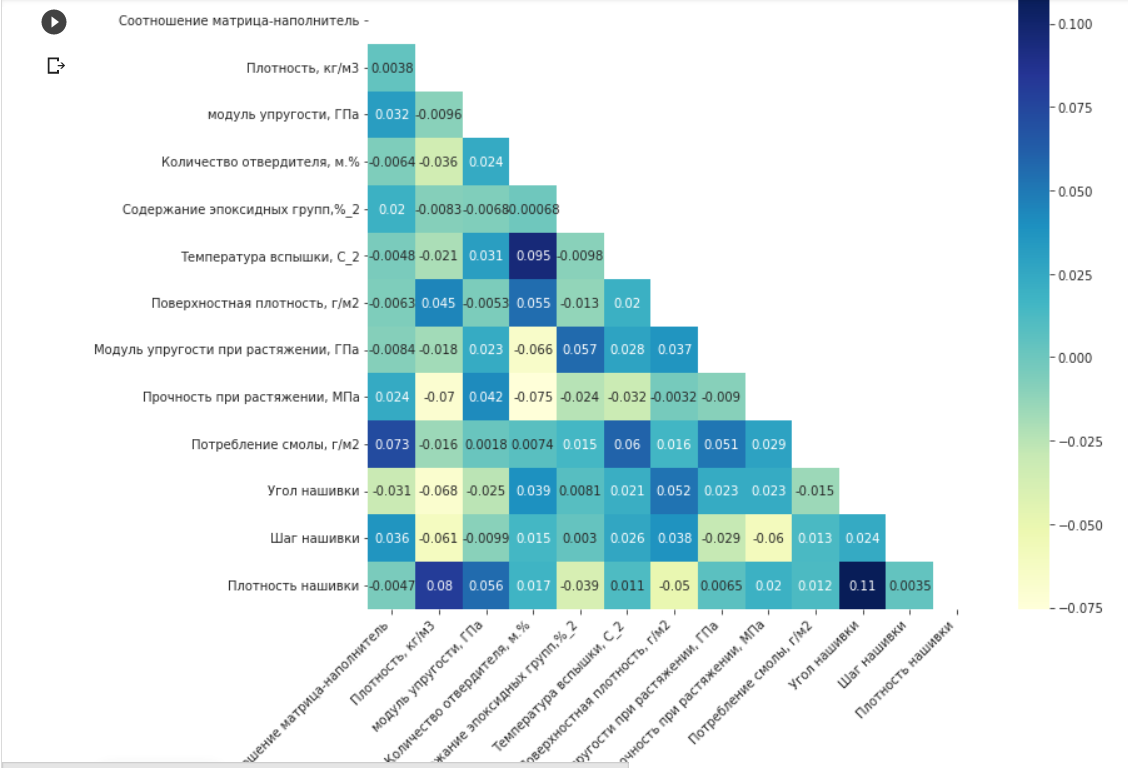


Рисунок 26 - тепловая карта с корреляцией данных

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

1. **Практическая часть**
   1. **Предобработка данных**

Одной из самых важных трансформаций, которые понадобиться применять к данным, является масштабирование признаков. Существует два распостраненных способа обеспечения того же самого масштаба у всех атрибутов: нормализация и стандартизация. По условиям задания нормализуем значения. Для этого применим MinMaxScaler(), затем применимNormalizer(). Второе даёт нам больше выбросов.

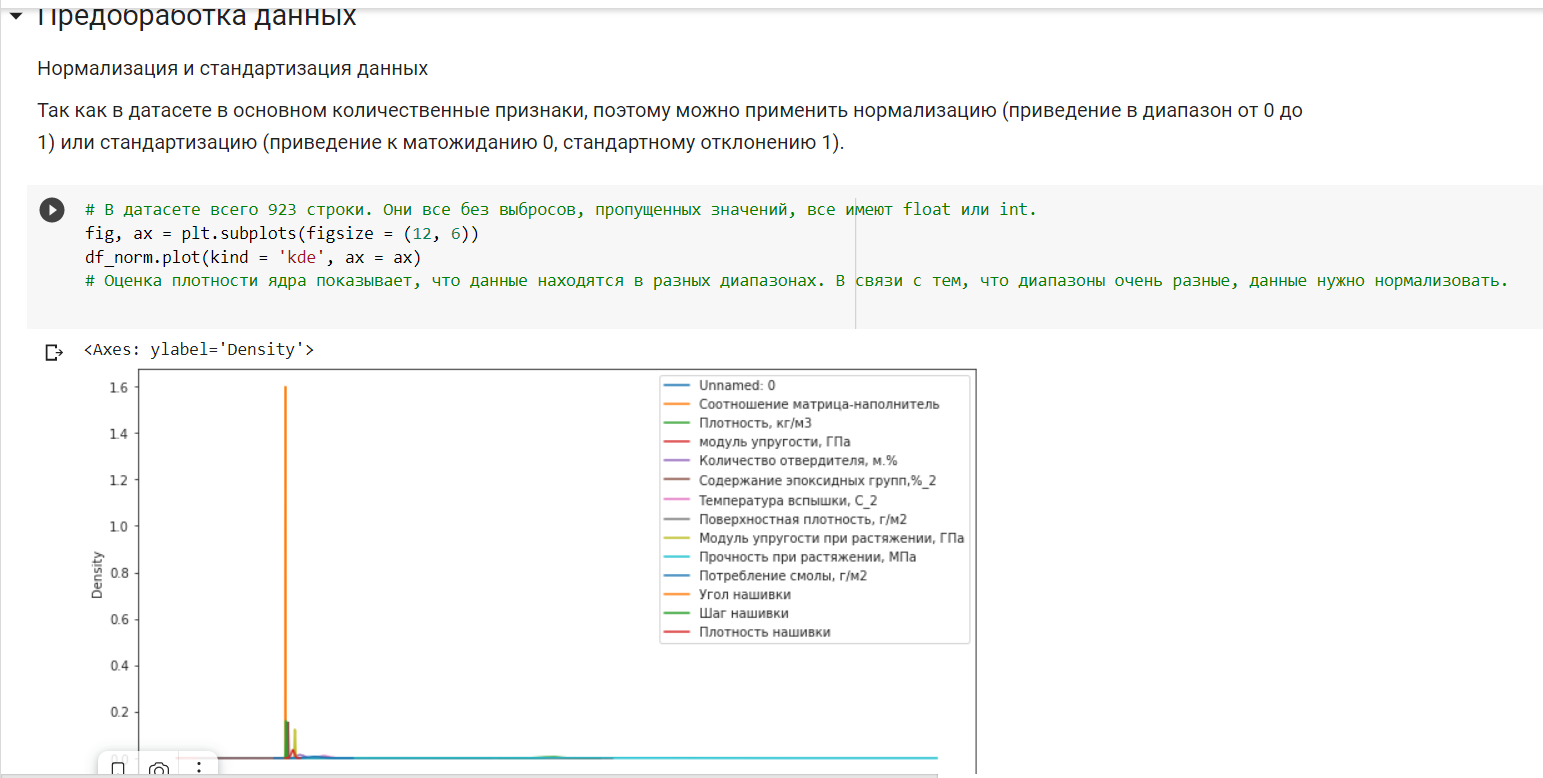


Рисунок 27 – Визуализированные данные до нормализации

Для решения задачи нормализации данных в Scikit-Learn предназначен трансформатор по имени MinMaxScaler.

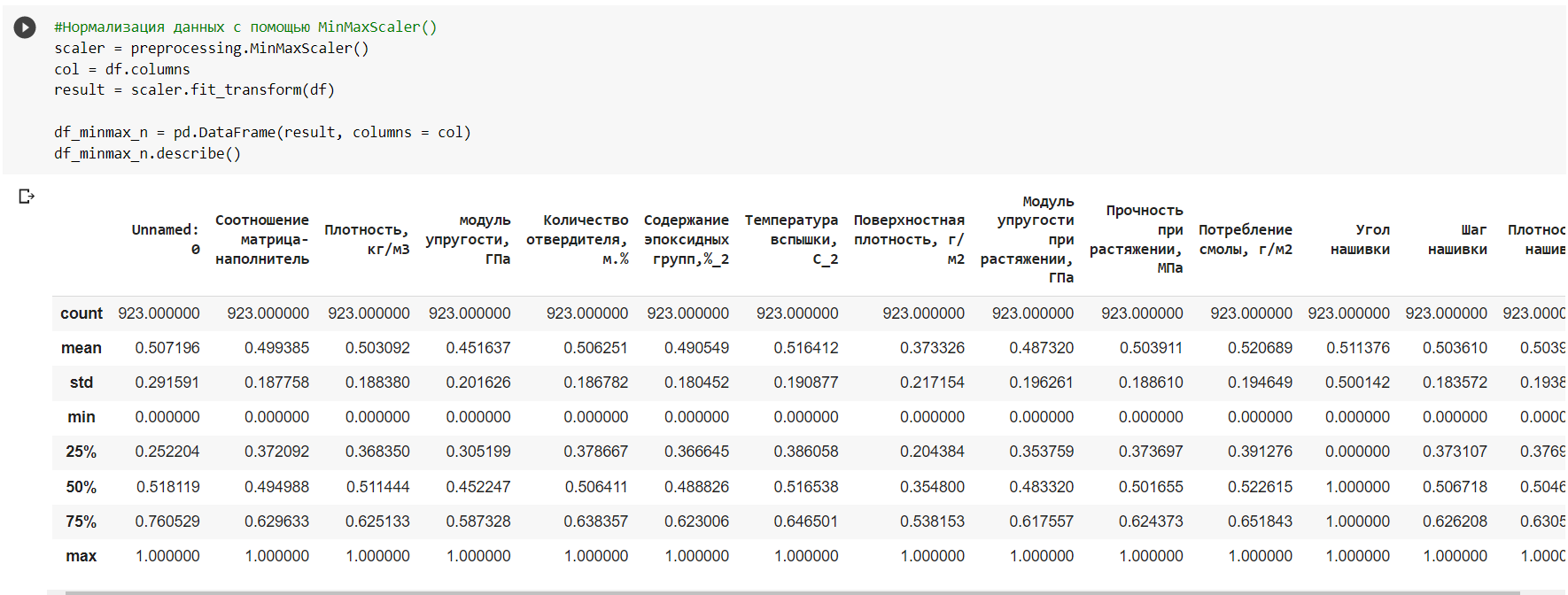


Рисунок 28 – Часть кода с проведенной нормализацией данных

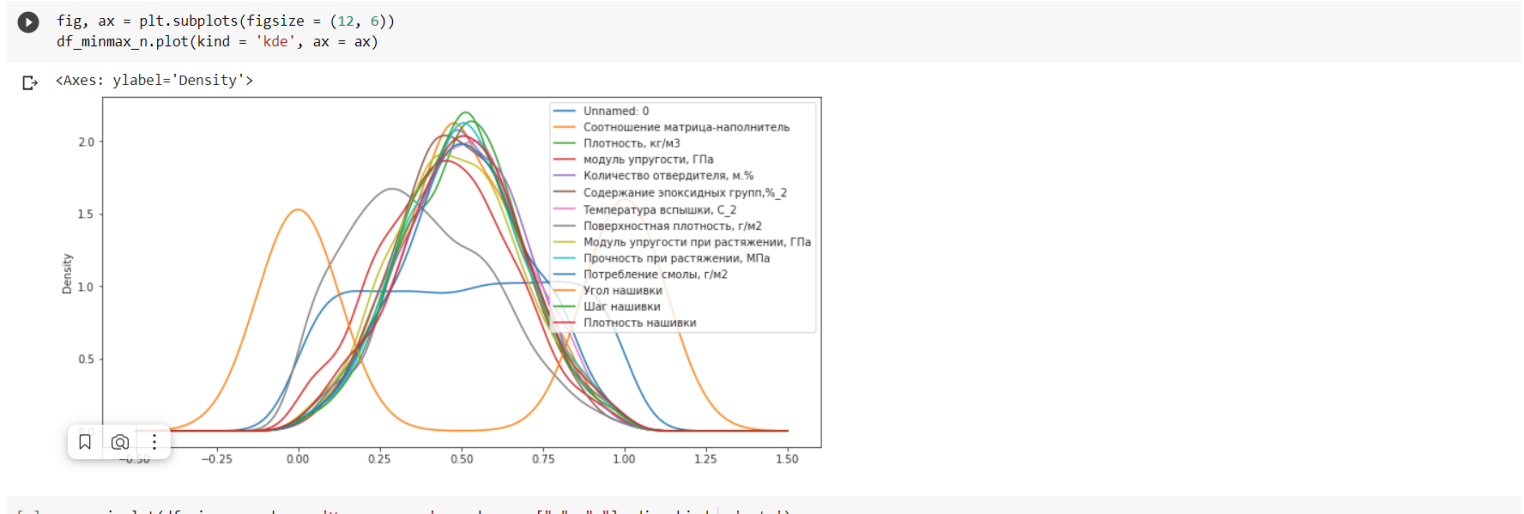


Рисунок 29 - визуализированные данные после нормализации

* 1. **Разработка и обучение модели**

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.



Рисунок 30- поиск гиперпараметров

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 70 на 30%); проверка моделей при стандартных значениях; сравнение с результатами модели, выдающей среднее значение; создание графика; сравнение моделей по метрике МАЕ(средняя абсолютная ошибка); 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.

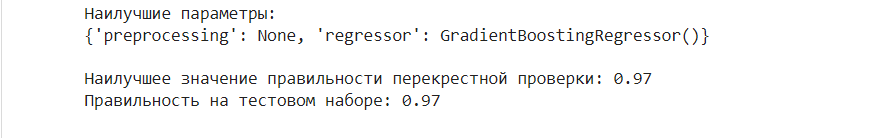


Рисунок 31 - наилучшие гиперпараметры

Модель после настройки гиперпараметров показала результат немного лучше. Однако, ниже, чем базовая модель. 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.

Таблица 1. Результаты построения и обучения моделей (выборочно)

|  | **Model** | **MAE** |  |
| --- | --- | --- | --- |
| **Модуль упругости при растяжении** | KNeighborsRegressor | 2.673409 |  |
| **Прочность при растяжении** | KNeighborsRegressor | 120.551015 |  |
| **Модуль упругости при растяжении** | LinearRegression | 2.535163 |  |
| **Прочность при растяжении** | LinearRegression | 63.502329 |  |
| **Модуль упругости при растяжении** | RandomForestRegressor | 2.648371 |  |
| **Прочность при растяжении** | RandomForestRegressor | 79.766000 |  |
| **Модуль упругости при растяжении** | MLPRegressor | 3.052449 |  |
| **Прочность при растяжении** | MLPRegressor | 1615.681526 |  |

* 1. **Тестирование модели**

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

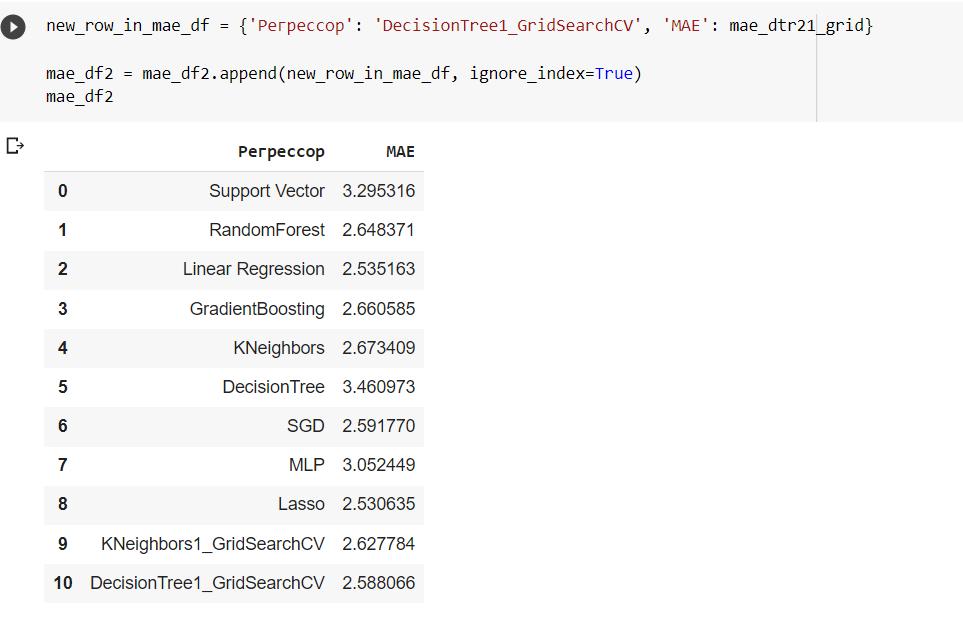


Рисунок 32 - результат оценки точности по MAE (модуль упругости при растяжении)

После обучения моделей была проведена оценка точности этих моделей на обучающей и тестовых выборках. В качестве параметра оценки модели использовалась средняя абсолютная ошибка (MAE). Обе модели даже на тренировочном датасете не смогли обучиться и приблизиться к исходным данным.

* 1. **Написать нейронную сеть, которая будет рекомендовать**

**соотношение «матрица – наполнитель».**

Обучение нейронной сети — это такой процесс, при котором происходит подбор оптимальных параметров модели, с точки зрения минимизации функционала ошибки. Начнём строить нейронную сеть с помощью класса keras.Sequential.

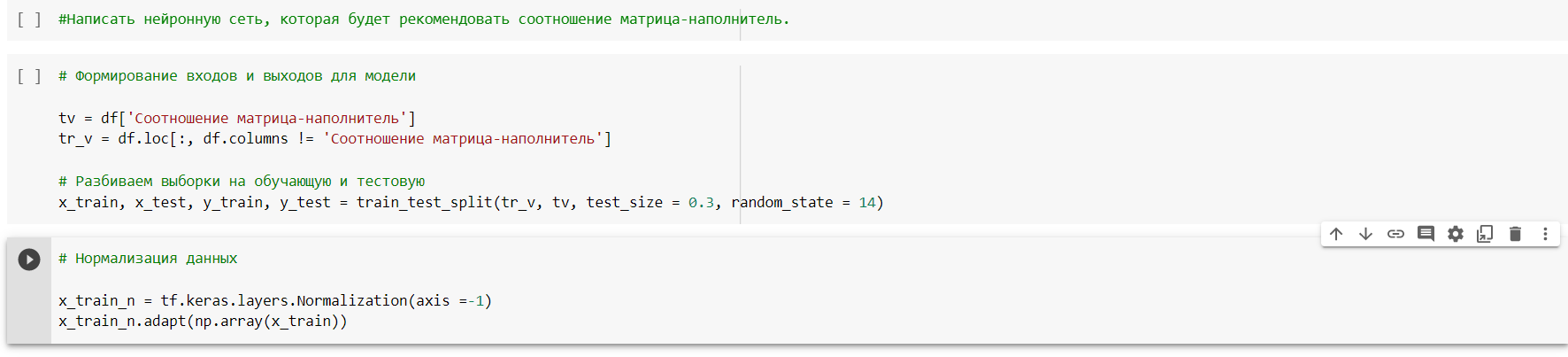


Рисунок 33 - создание нейронной сети

Определим параметры, поищем оптимальные параметры, посмотрим на результаты. С помощью KerasClassifier выйдем на наилучшие параметры для нашей нейронной сети и построим окончательную нейросеть.

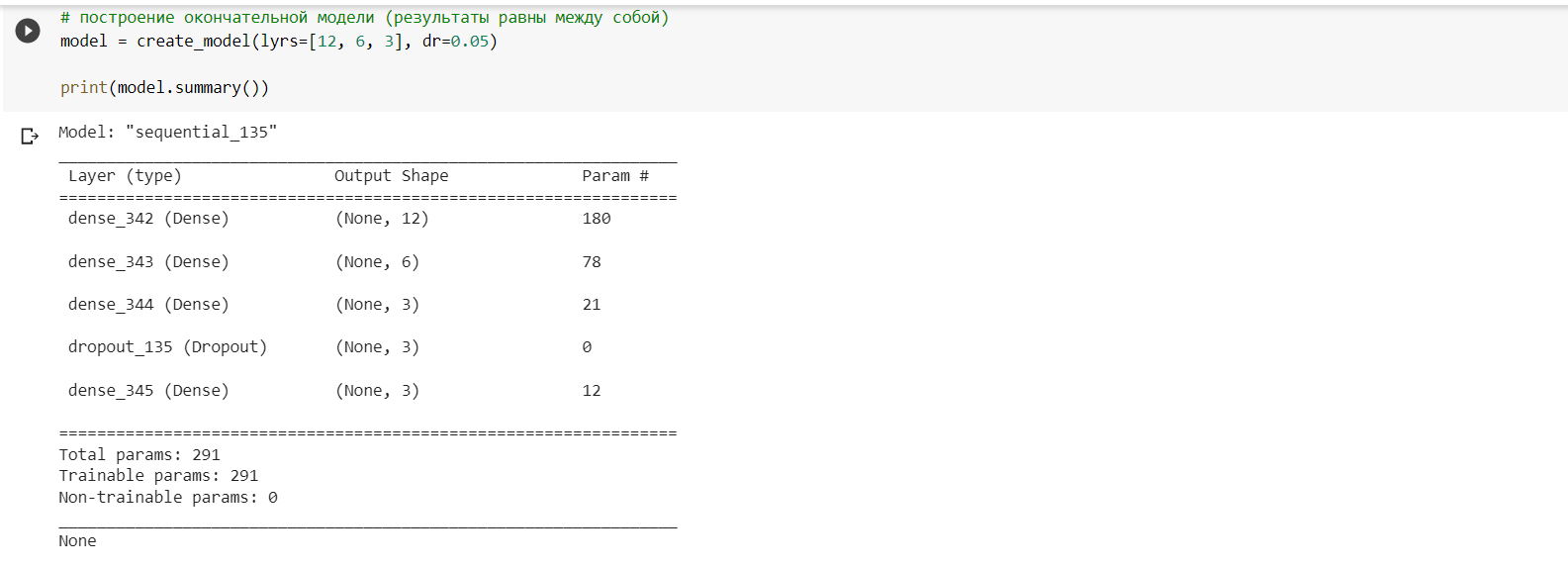


Рисунок 34 - построение первой нейросети

Обучим и оценим модель, посмотрим на потери, зададим функцию для визуализации факт/прогноз для результатов моделей.

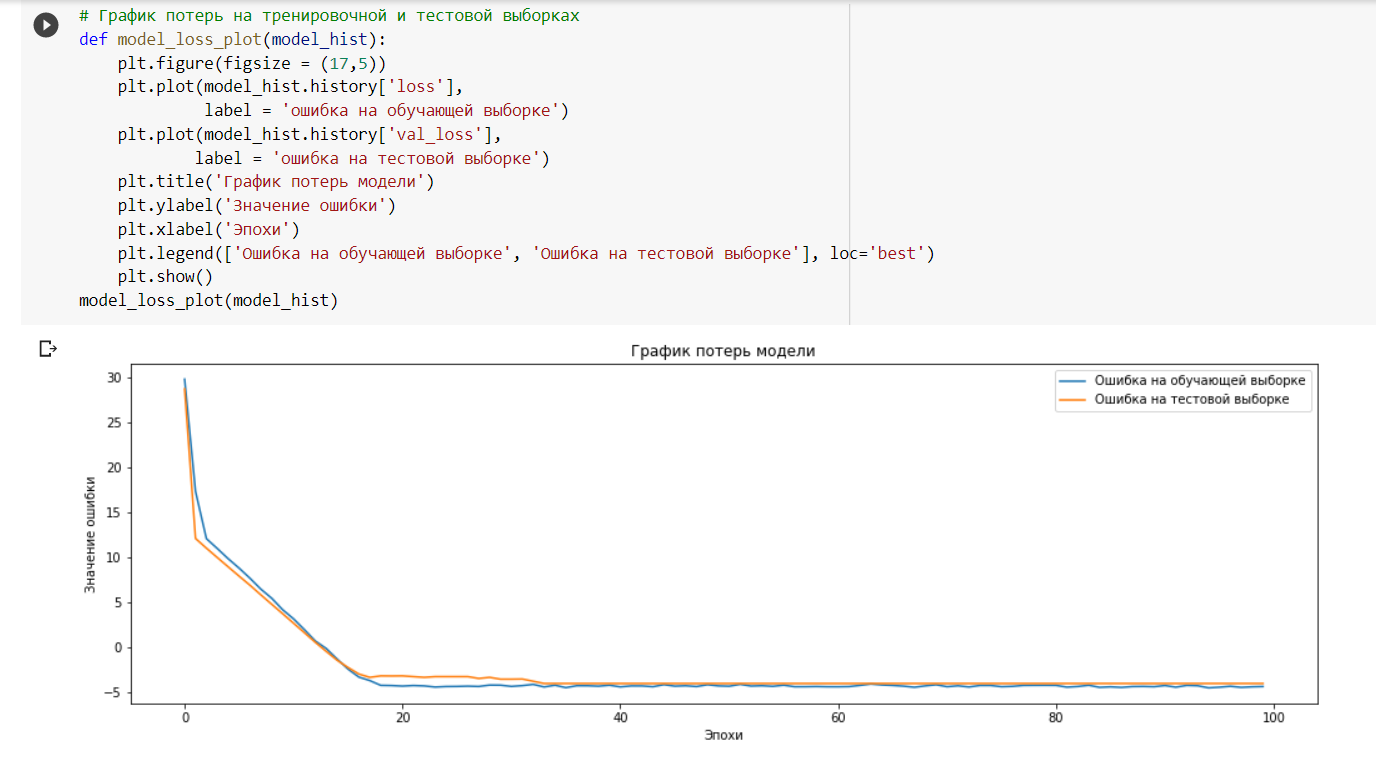


Рисунок 35 - график потерь модели 1

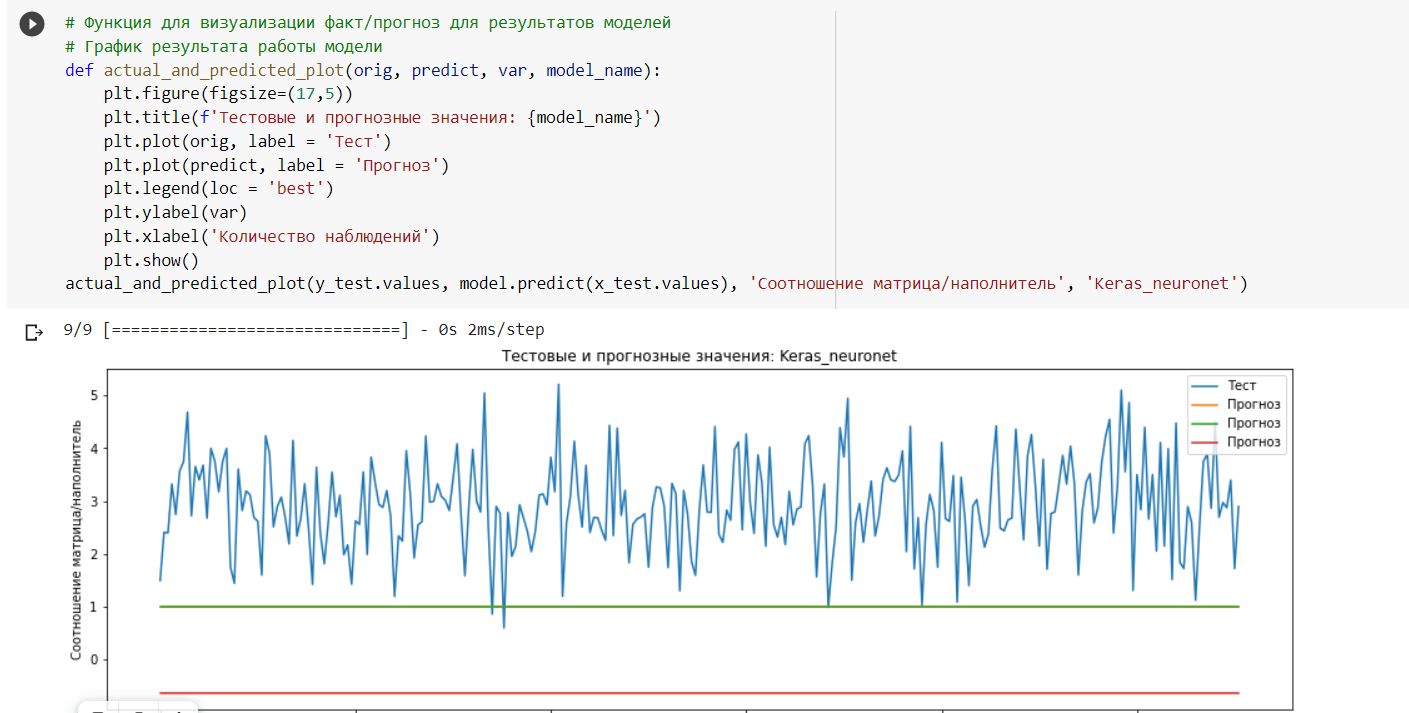


Рисунок 36 - тестовые и прогнозные значения модели 1

Не удовлетворившись таким результатом, создадим другую простую модель глубокого обучения с другой архитектурой. Обучим её, посмотрим на потери, оценим по MSE, построим график.

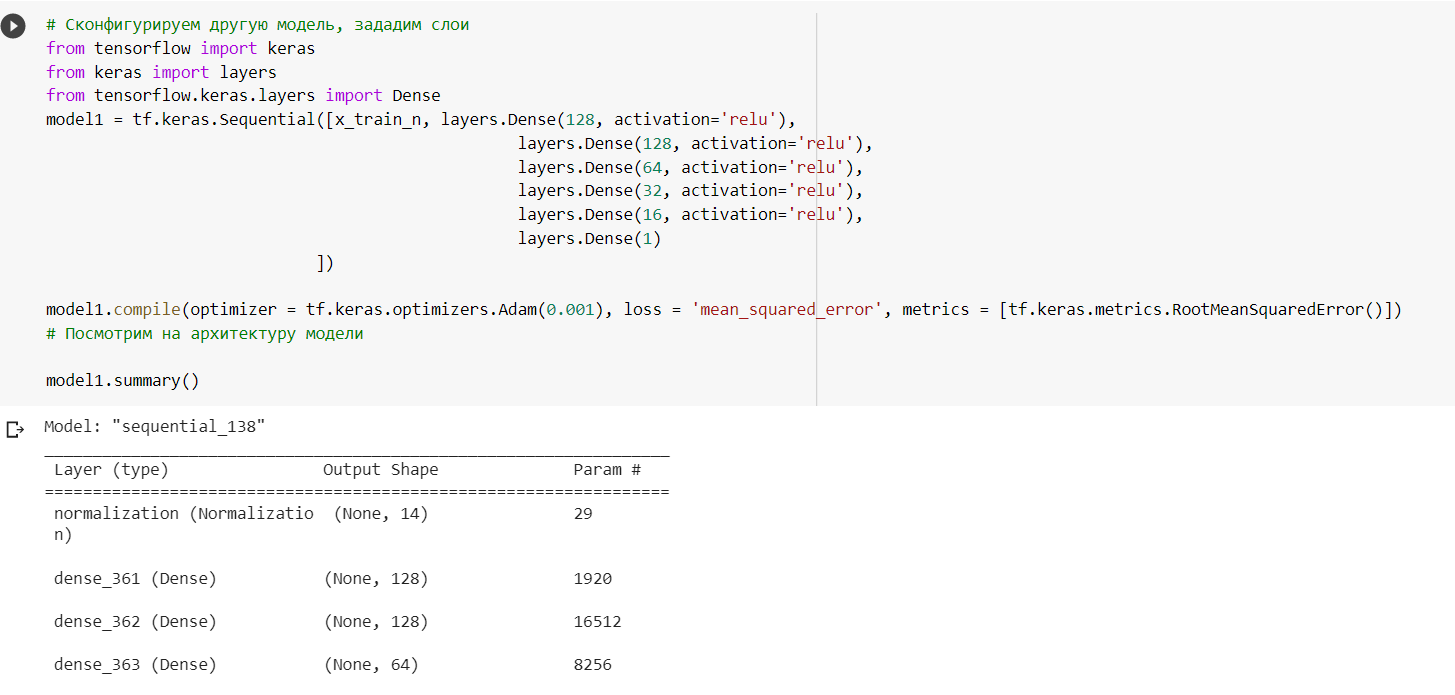


Рисунок 37 - создание второй модели

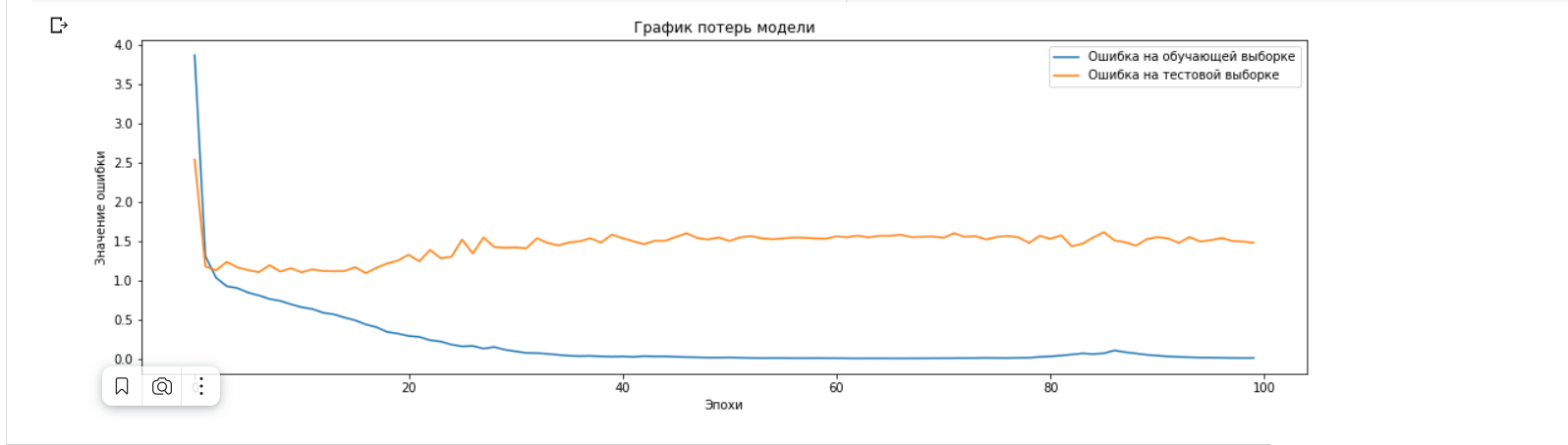


Рисунок 38 - график потерь второй модели

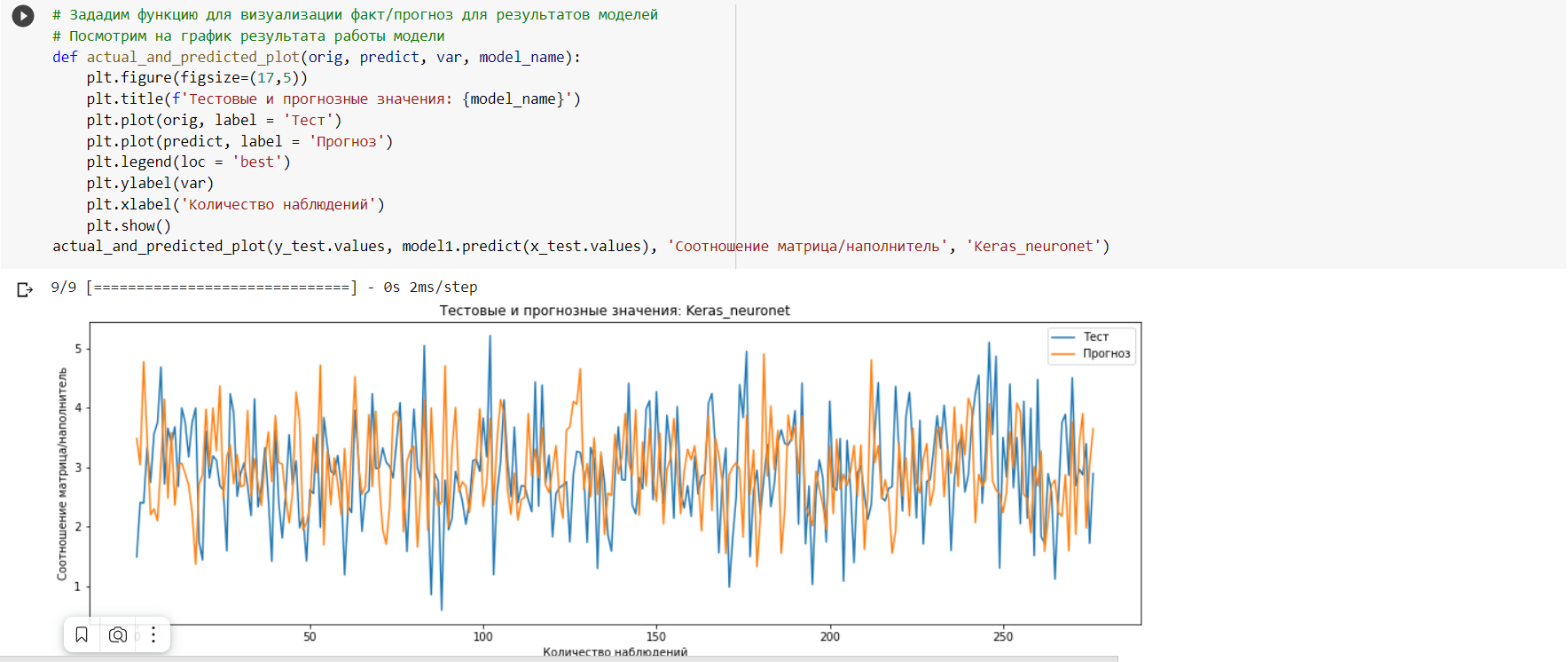


Рисунок 39- тестовые и прогнозные значения второй модели

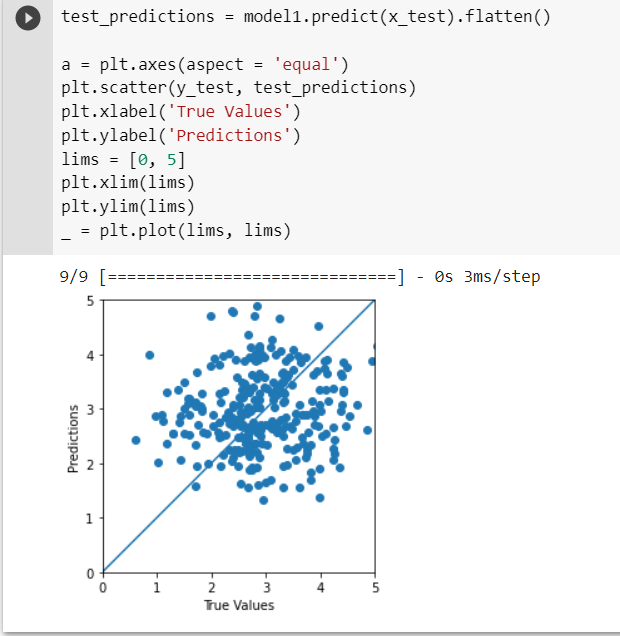


Рисунок 40- график прогнозных и настоящих значений

* 1. **Создание удалённого репозитория и загрузка**

Страница слушателя: <https://github.com/MarinaAndr>, репозиторий был создан на github.com по адресу: <https://github.com/MarinaAndr/->

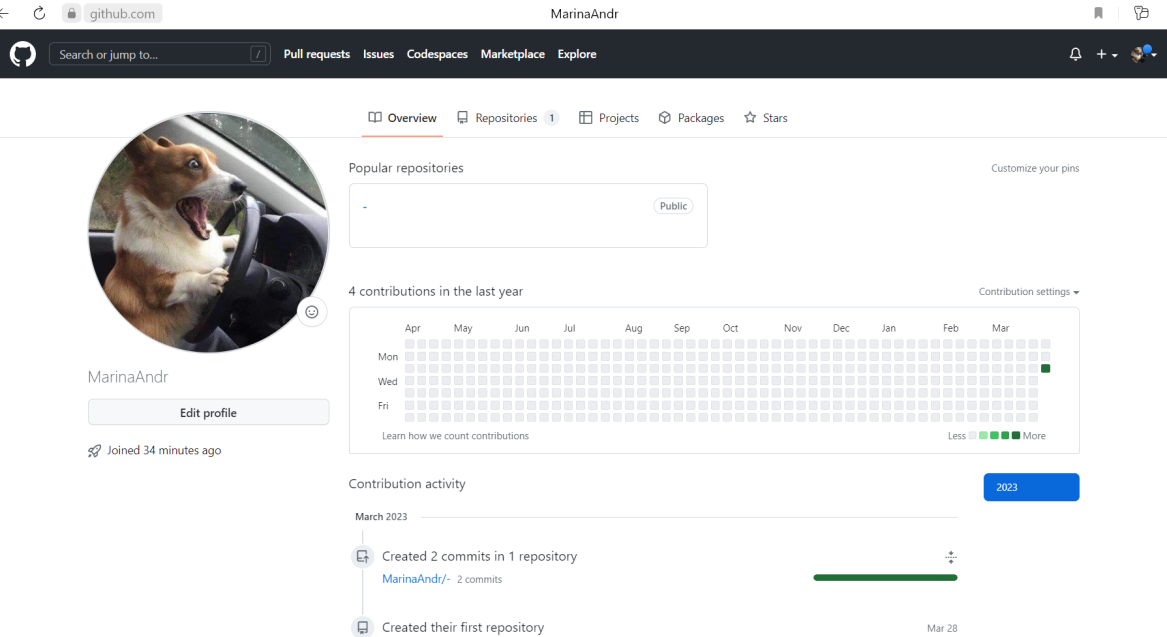


Рисунок 41 - часть страницы на github.com

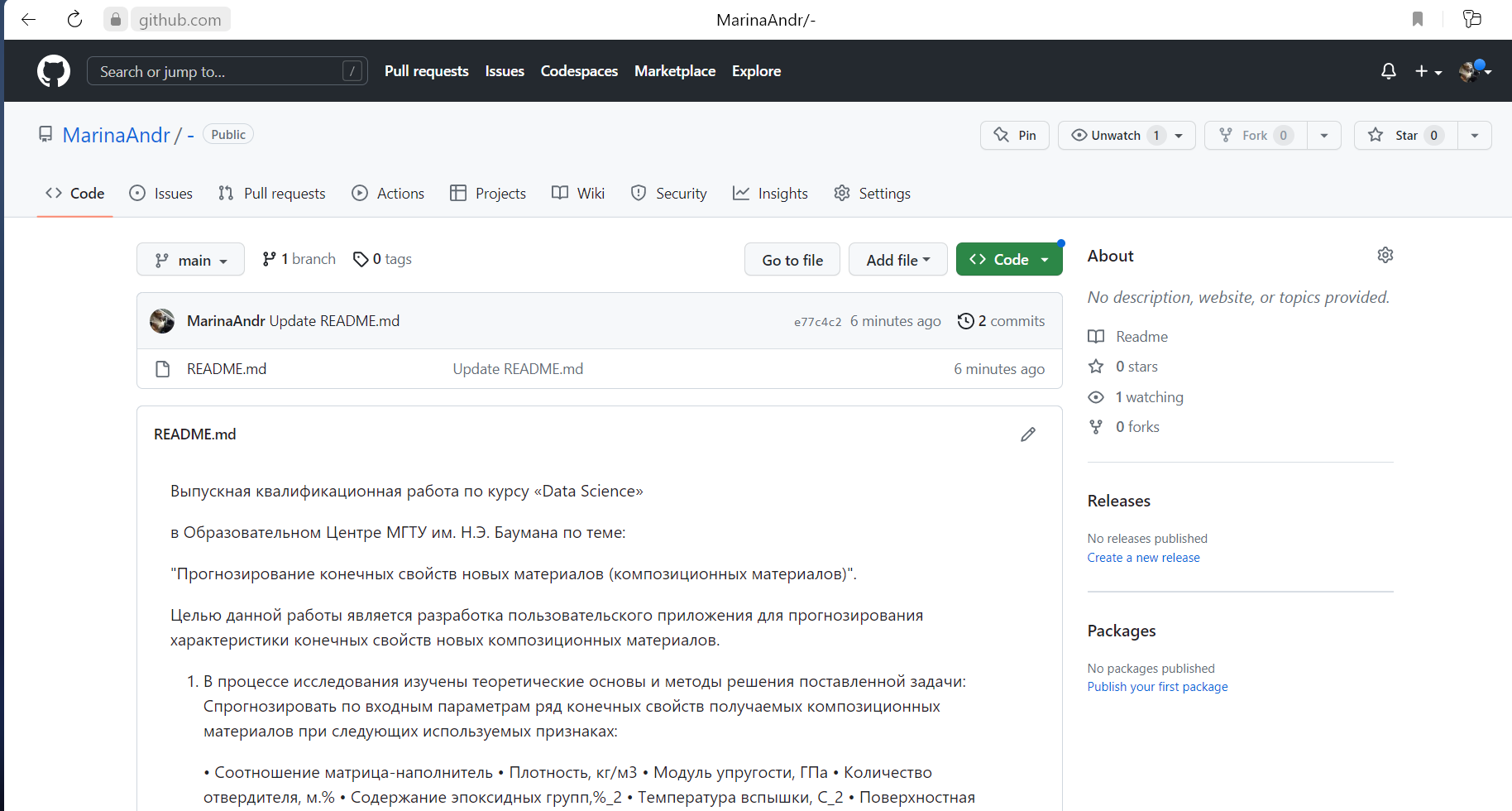


Рисунок 42 - часть созданного файла README

* 1. **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. Лучшие метрики для модуля упругости при растяжении, ГПа – метод опорных векторов, для прочности при растяжении, МПа – метод градиентного бустинга.

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.

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.

* 1. **Список используемой литературы и веб ресурсы**.

1. Орельен Жерон: Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow – пер. с англ. – Москва, Санкт-Петербург: Диалектика, 2020г. – 684с. с ил.
2. Брюс П., Брюс Э., Гедек П.: Практическая статистика для специалистов Data Science – пер. с англ. - Санкт-Петербург: БХВ-Петербург, 2021г. – 346с. с ил.
3. Дауни А.Д. Основы Python. Научитесь думать, как программист – пер. с англ. – Москва.: Манн, Иванов и Фербер, 2021. – 302 с., ил.
4. Джулли А., Пал С. Библиотека Keras – инструмент глубокого обучения – Москва: ДМК, 2018г. – 293 с.
5. 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.
6. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
7. Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.
8. Уилан Ч. Голая Статистика – пер. с англ. - Москва.: Манн, Иванов и Фербер, 2017. – 346 с.
9. Билл Любанович. Простой Python. Современный стиль программирования. — СПб.: Питер, 2016. — 480 с.: ил.
10. В. Ш. Берикашвили, С. П.Оськин Статистическая обработка данных, планирование эксперимента и случайные процессы​ : учебное пособие для вузов​ - 2-е изд., испр. и доп.​ - М.​ : Юрайт, 2021.​ - 163 с.​
11. Дауни Аллен. Рекомендательные системы на практике - пер. с англ. Д. М. Павлова. – М.: ДМК Пресс, 2020. – 448 с.: ил.
12. Траск Эндрю. Грокаем глубокое обучение. – СПб.: Питер, 2019. – 352 с.: ил.
13. Элис Джен. Машинное обучение. Конструирование признаков. - Москва: Бомбора, 2022. – 240с.
14. [Уатт Д](https://www.prodalit.ru/cat?Author=%22%D0%A3%D0%B0%D1%82%D1%82%20%D0%94%D0%B6%D0%B5%D1%80%D0%B5%D0%BC%D0%B8%22&FindMode=Short)., [Борнани Р](https://www.prodalit.ru/cat?Author=%22%D0%91%D0%BE%D1%80%D0%BD%D0%B0%D0%BD%D0%B8%20%D0%A0%D0%B5%D0%B7%D0%B0%22&FindMode=Short)., Катсаггелос А. Машинное обучение: Основы, алгоритмы и практика применения. - СПб.: БХВ-Петербурr, 2022. – 640с.
15. [Берман Кеннеди](https://www.prodalit.ru/cat?Author=%22%D0%91%D0%B5%D1%80%D0%BC%D0%B0%D0%BD%20%D0%9A%D0%B5%D0%BD%D0%BD%D0%B5%D0%B4%D0%B8%22&FindMode=Short). Основы Python для Data Science. - Питер Книга, 2023. – 272с.
16. Документация по библиотеке keras: – Режим доступа: <https://keras.io/api/>.(дата обращения: 19.03.2023).
17. Документация по библиотеке matplotlib: – Режим доступа: <https://matplotlib.org/stable/users/index.html>.(дата обращения: 19.03.2023)
18. Документация по библиотеке numpy: – Режим доступа: <https://numpy.org/doc/1.22/user/index.html#user>. (дата обращения: 19.03.2023)
19. Документация по библиотеке pandas: – Режим доступа: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>.(дата обращения: 20.03.2023).
20. Документация по библиотеке scikit-learn: – Режим доступа: <https://scikit-learn.org/stable/user_guide.html>. (дата обращения: 17.03.2023).
21. Документация по библиотеке seaborn: – Режим доступа: <https://seaborn.pydata.org/tutorial.html>. (дата обращения: 17.03.2023).
22. Документация по библиотеке Tensorflow: – Режим доступа: <https://www.tensorflow.org/overview>(дата обращения: 19.03.2023).
23. Документация по языку программирования python: – Режим доступа:  <https://docs.python.org/3.8/index.html>.(дата обращения: 15.03.2023).
24. SVM. Подробный разбор метода опорных векторов, реализация на Python – Режим доступа: <https://habr.com/ru/company/ods/blog/484148/> (дата обращения 20.03.2023)
25. Различия между искусственным интеллектом, машинным обучением и глубоким обучением – Режим доступа: <https://habr.com/ru/post/401857/> (дата обращения 12.03.2023)
26. Машинное обучение для начинающих: алгоритм случайного леса (Random Forest) – Режим доступа: <https://proglib.io/p/mashinnoe-obuchenie-dlya-nachinayushchih-algoritm-sluchaynogo-lesa-random-forest-2021-08-12> (дата обращения 27.03.2023г.)
27. Разведочный анализ данных (EDA) – Режим доступа: <https://www.helenkapatsa.ru/razvedochnyy-analiz-dannykh-chast-1/> (дата обращения 23.03.2023г.)
28. Виды и применение композитных материалов – Режим доступа: <https://epolymer.ru/article/vidy-i-primenenie-kompozitnyh-materialov> (дата обращения 20.03.2023)
29. Материалы на основе базальтвовых волокон – Режим доступа: <http://basaltm.com/materialy/materials.html> (дата обращения 13.03.2023)
30. Руководство по быстрому старту в flask: – Режим доступа: <https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html>. (дата обращения: 27.03.2023)