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: Цыганков Павел Игоревич

Москва, 2023

**Содержание**

**Содержание............................................................................................................ 2** **Введение................................................................................................................. 3**

1. **Аналитическая часть…................................................................................ 5**
   1. **Постановка задачи........................................................................... 5**
   2. **Описание используемых методов.................................................. 8**
   3. **Разведочный анализ данных........................................................ 13**
2. **Практическая часть…................................................................................. 17**
   1. **Предобработка данных.................................................................. 17**
   2. **Разработка и обучение модели..................................................... 19**
   3. **Тестирование модели..................................................................... 22**
   4. **Написать нейронную сеть, которая будет рекомендовать соотношение**

**«матрица-наполнитель» ………………………………………... 23**

* 1. **Разработка приложения ............................................................... 27**
  2. **Создание удалённого репозитория и загрузка.......................... 29**
  3. **Заключение...................................................................................... 29**
  4. **Список используемой литературы и веб ресурсы................... 31**

**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 на фреймворке Flask.

# Аналитическая часть

## **1.1. Постановка задачи**

Для исследовательской работы были даны 2 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). Часть кода начала обработки первого датасета представлена ниже на Рисунке 1.



Рисунок 1 – первый датасет и его поверхностный обзор в коде.

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. Ниже на Рисунке 2 представлена часть кода с объединенным датасетом.

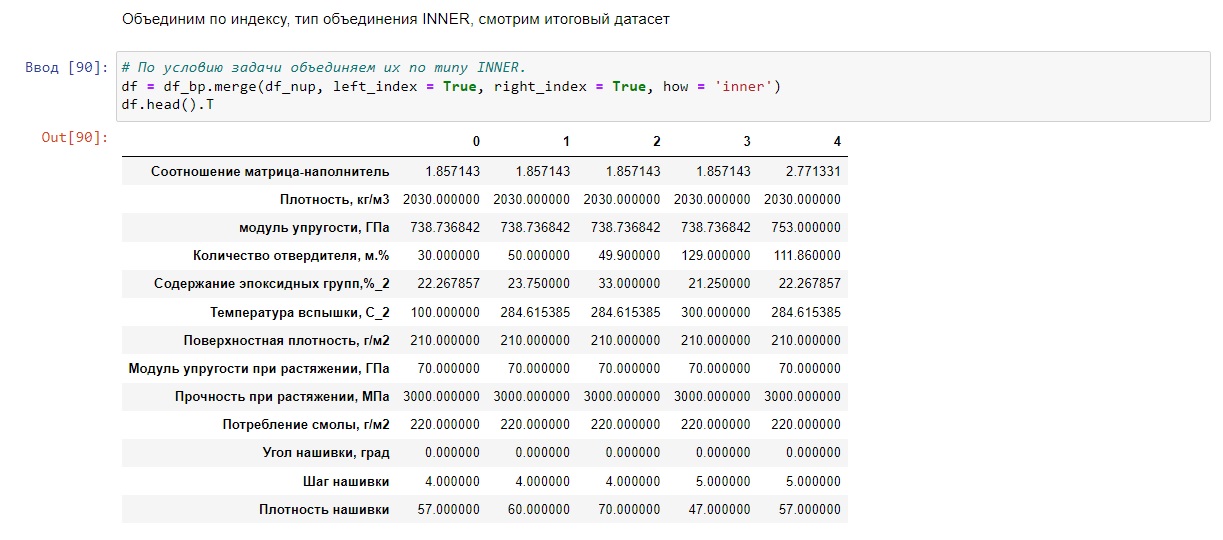


Рисунок 2 – объединенный необработанный датасет

Затем провести разведочный анализ данных, нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек. Пример первого визуального исследования представлен на рисунке 3 ниже.

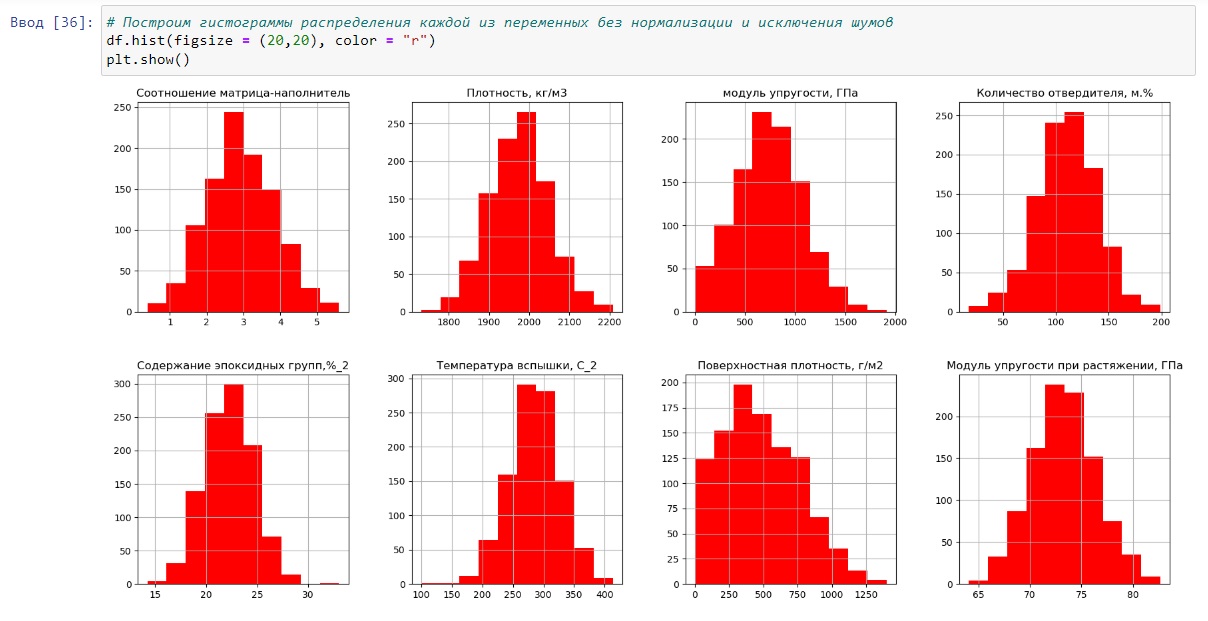


Рисунок 3 - гистограммы распределения переменных

Для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков (часть кода и графического анализа пропусков представлена на рисунке 4 ниже); предобработать 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». Создать профиь в GitHub и разместить код исследования.

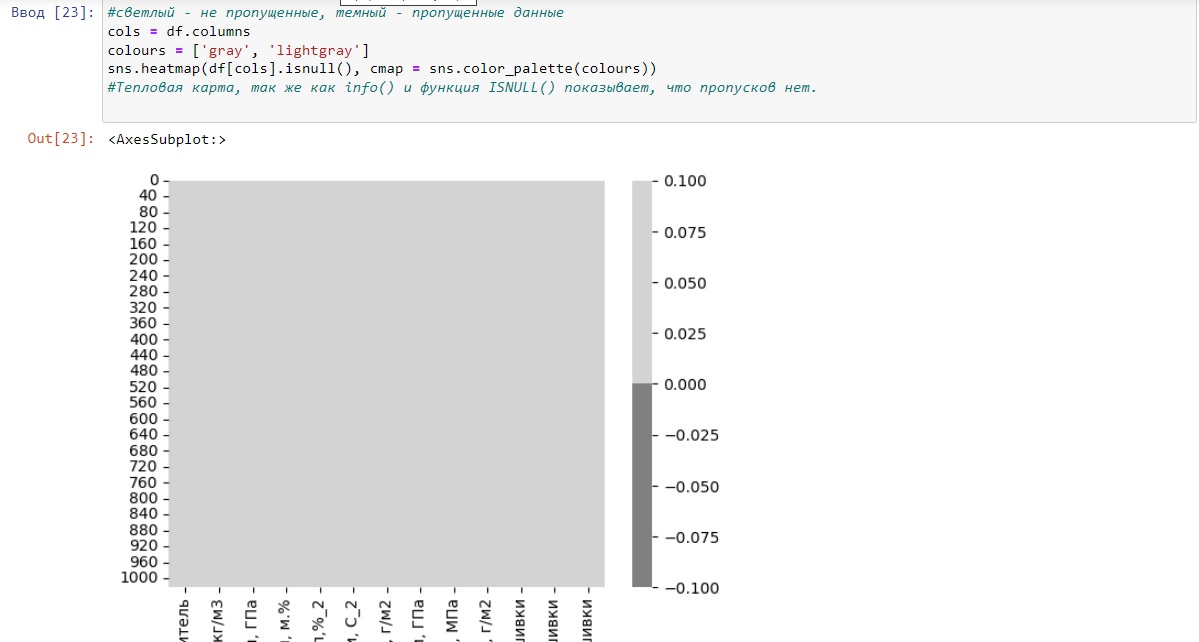


Рисунок 4 – тепловая карта пропусков итогового датасета

## **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; (рисунок 5)

− линейная регрессия; (рисунок 6)

− К-ближайших соседей; (рисунок 7)

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 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.

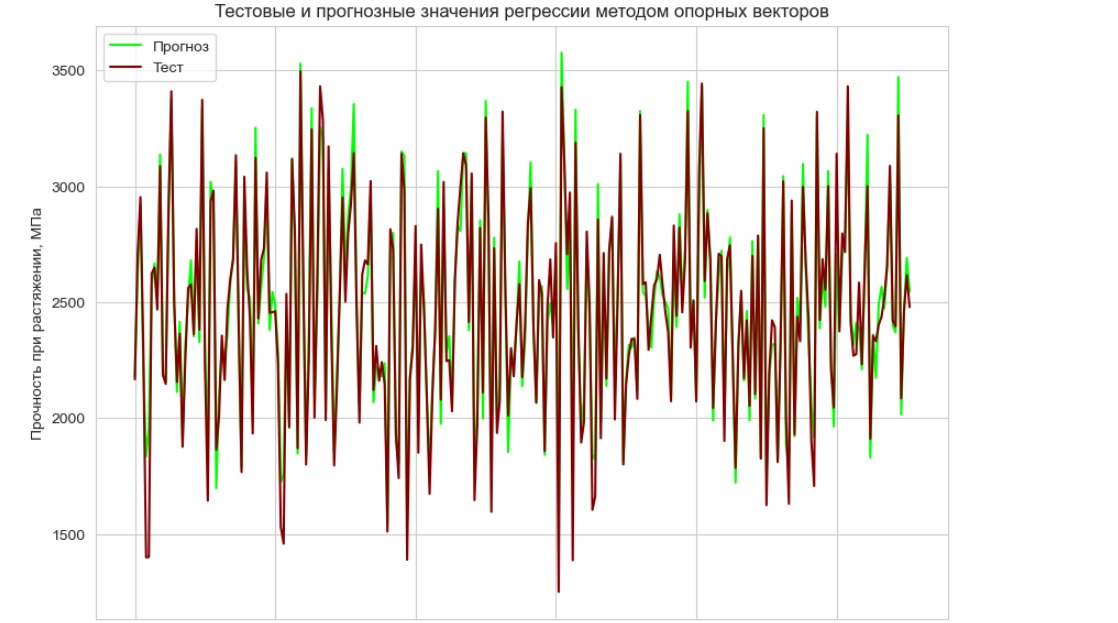


Рисунок 5

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

Метод ближайших соседей - К-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 числительную трудоёмкость.

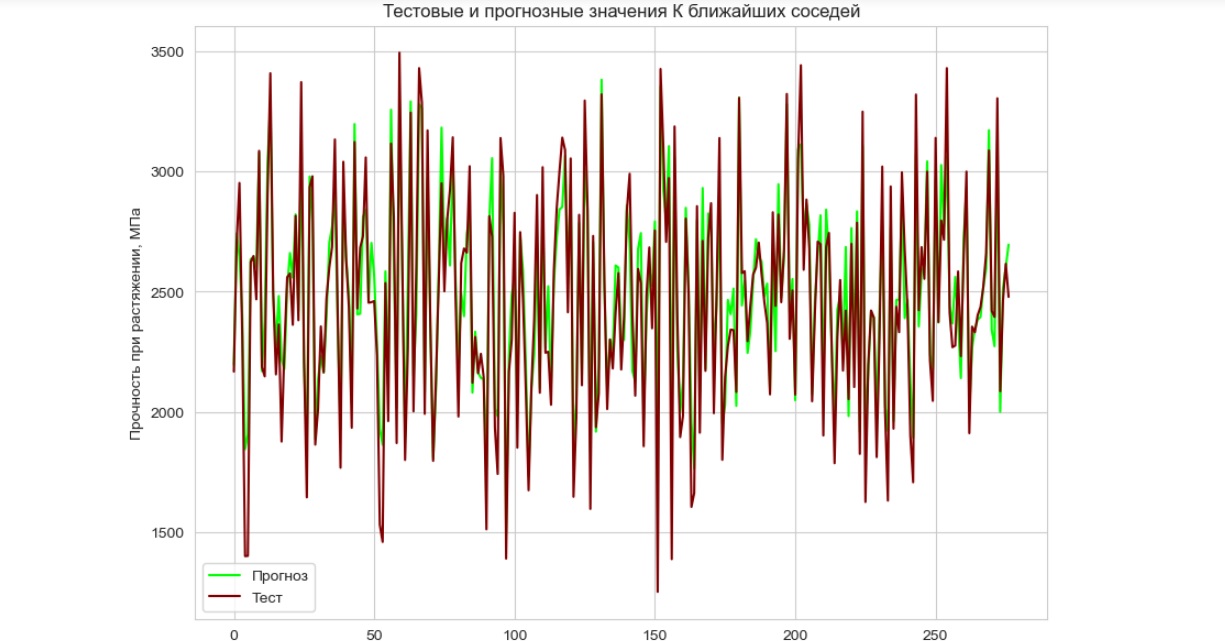


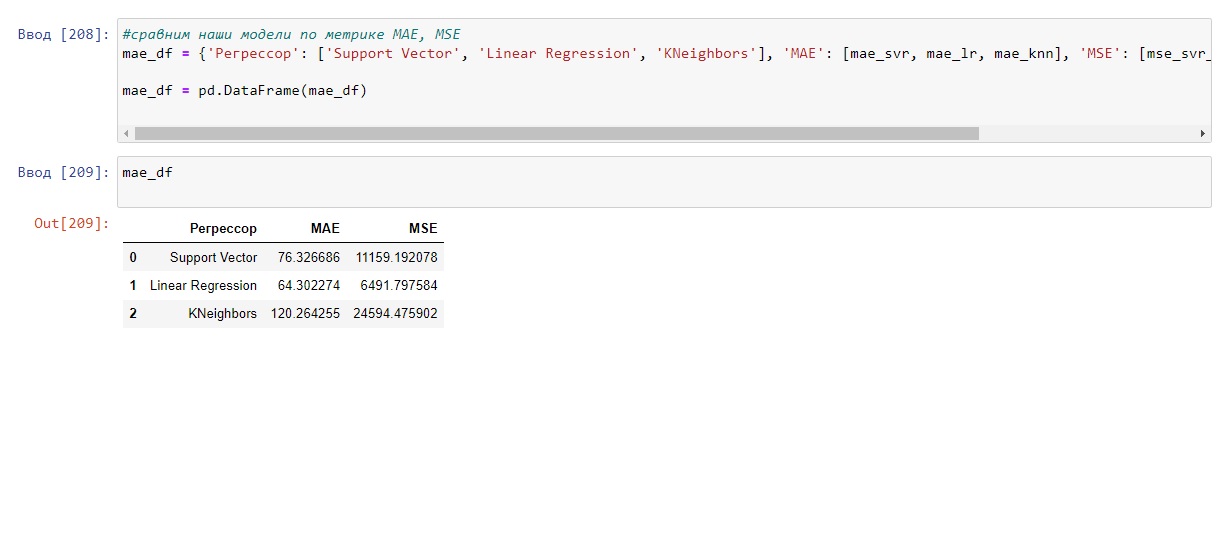
Рисунок 7

Немного расскажем об используемых метриках качества моделей:

MAE (Mean Absolute Error) метрика измеряет среднюю сумму абсолютной разницы между фактическим значением и прогнозируемым значением. Т.е. MAE рассчитывается как среднее абсолютных разностей между наблюдаемым и предсказанным значениями. В отличие от MSE и RMSE она является линейной оценкой, а это значит, что все ошибки в среднем взвешены одинаково. Например, разница между 0 и 10 будет вдвое больше разницы между 0 и 5. Для MSE и RMSE, как отмечено выше, это не так.

Поэтому MAE широко используется, например, в финансовой сфере, где ошибка в 10 долларов должна интерпретироваться как в два раза худшая, чем ошибка в 5 долларов.

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.

Рисунок 8 – сравнение всех наших моделей по метрикам MAE и MSE

**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. В качестве инструментов разведочного анализа используется: оценка статистических характеристик датасета (Рисунок 9); гистограммы распределения каждой из переменной (несколько различных вариантов) (рисунок 10); диаграммы ящика с усами (несколько интерактивных вариантов) (рисунок 11); попарные графики рассеяния точек (несколько вариантов) (Рисунок 12); график «квантиль-квантиль»; тепловая карта (несколько вариантов); описательная статистика для каждой переменной; анализ и полное исключение выбросов (5 повторных итераций); проверка наличия пропусков и дубликатов; ранговая корреляция Кендалла , Пирсона и Спирмана.

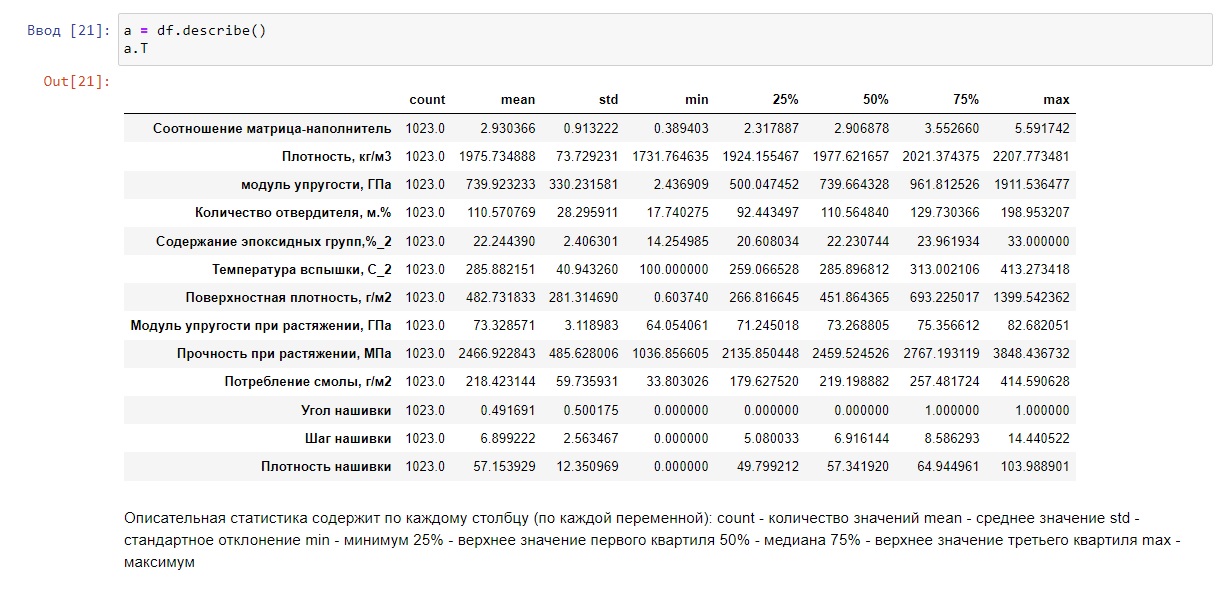


Рисунок 9 - count - количество значений mean - среднее значение std - стандартное отклонение min - минимум 25% - верхнее значение первого квартиля 50% - медиана 75% - верхнее значение третьего квартиля max – максимум

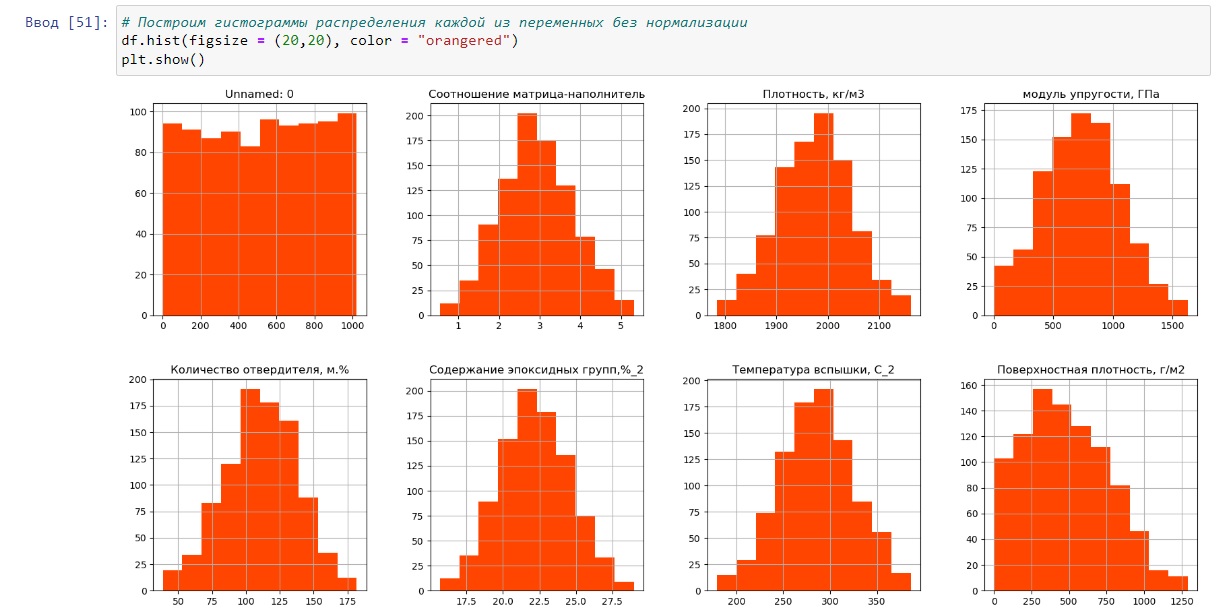


Рисунок 10 – гистограммы распределения переменных

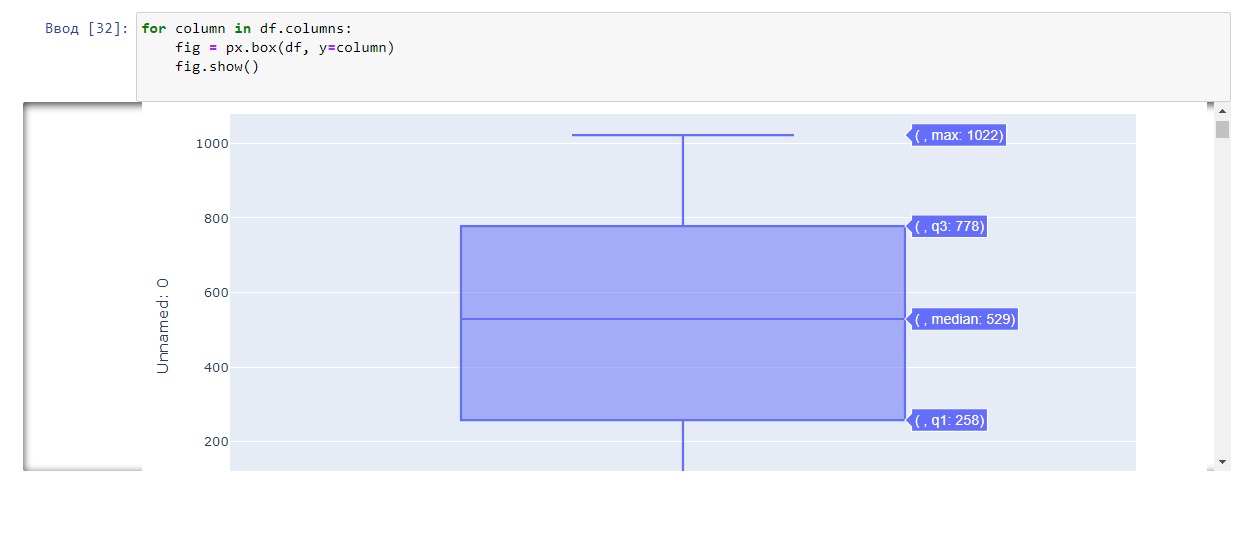


Рисунок 11 – ящик с усами с указанием значений квартилей

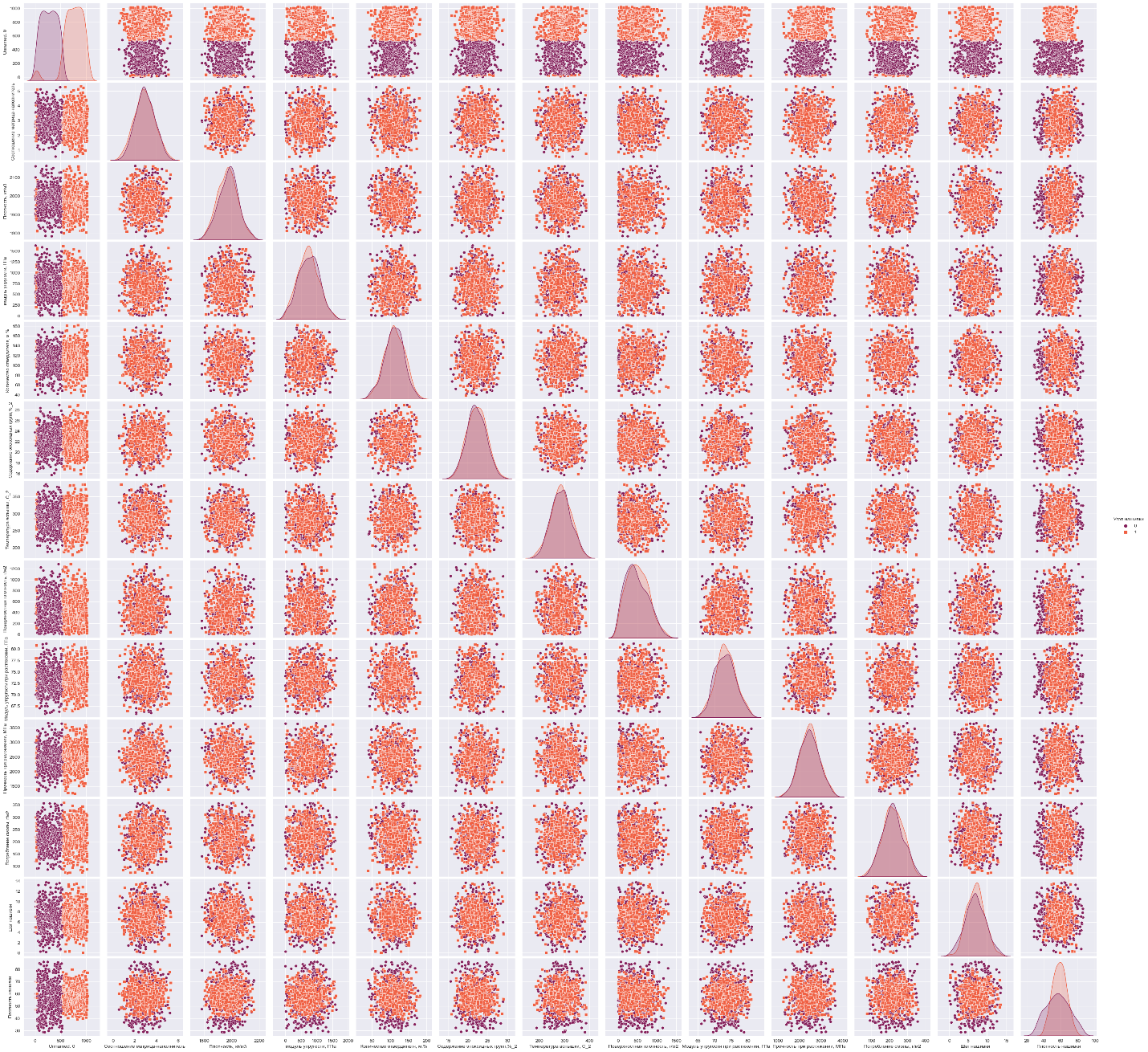


Рисунок 12 – попарные графики рассеяния точек

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 13)

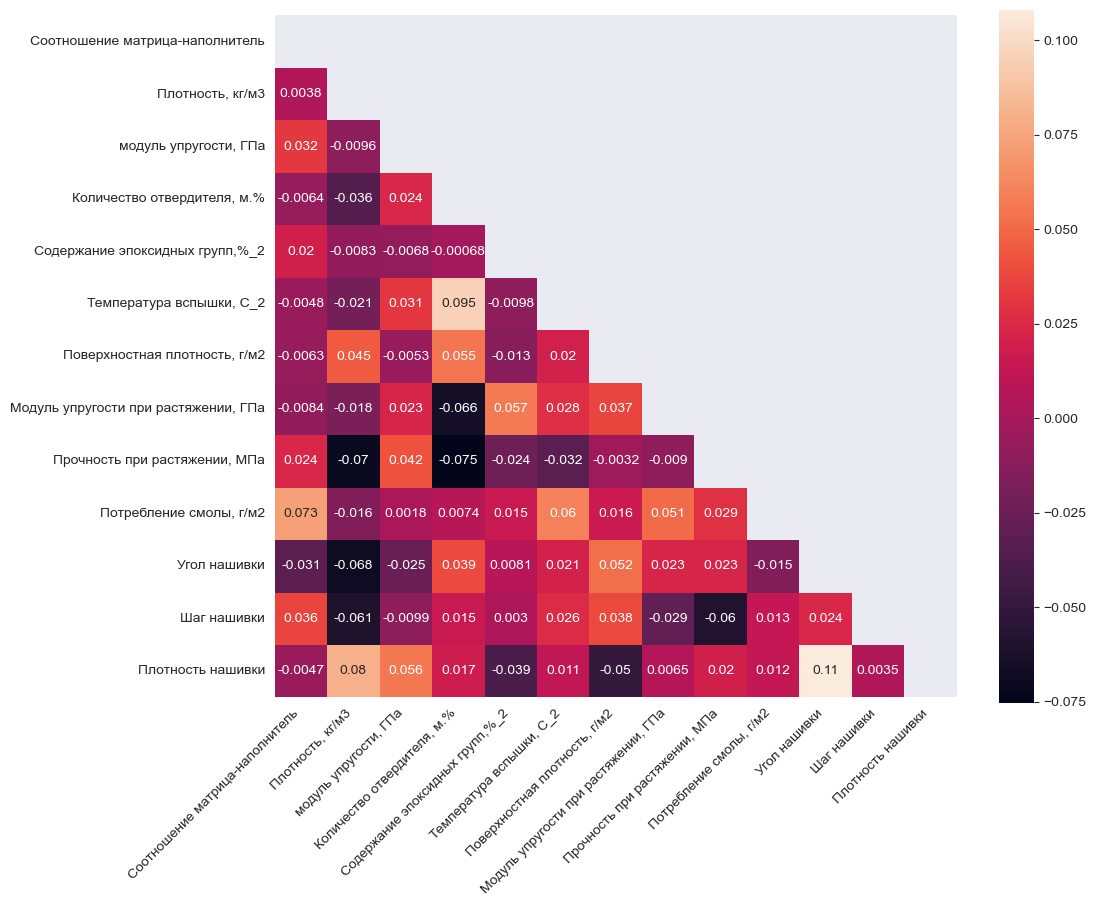


Рисунок 13 – тепловая карта с корреляцией данных окончательного датасета

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.

# 2. Практическая часть

## **2.1. Предобработка данных**

В ходе проведённого анализа принимаем решение столбец "Угол нашивки" привести к виду «0» и «1». (рисунок 14)

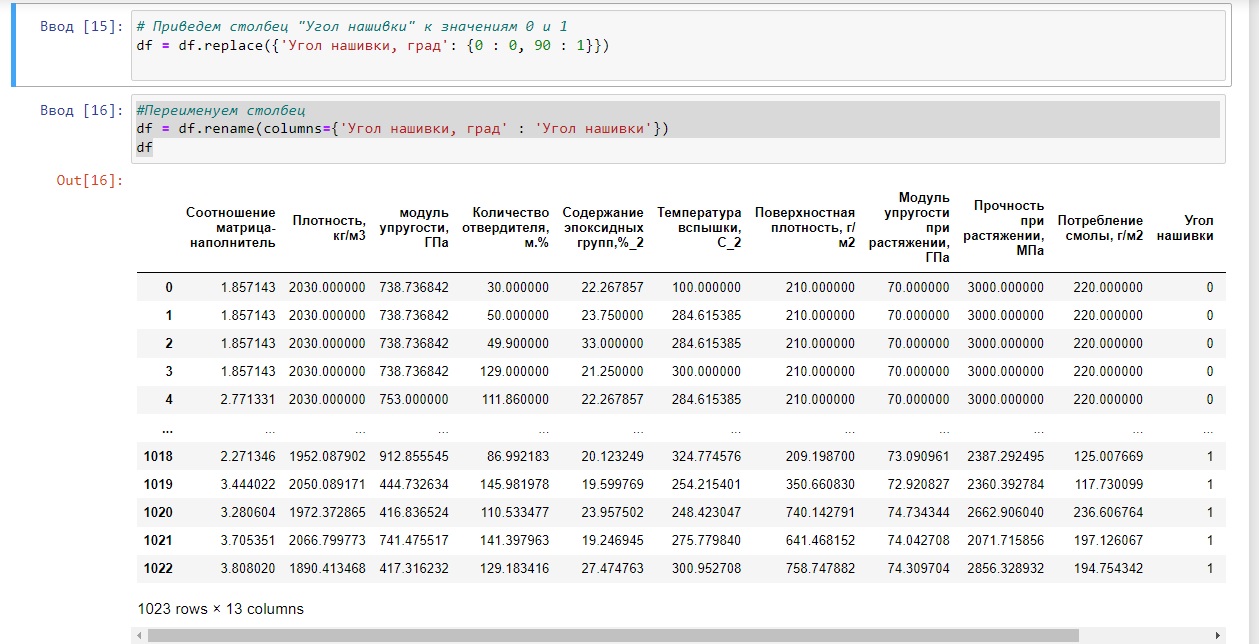
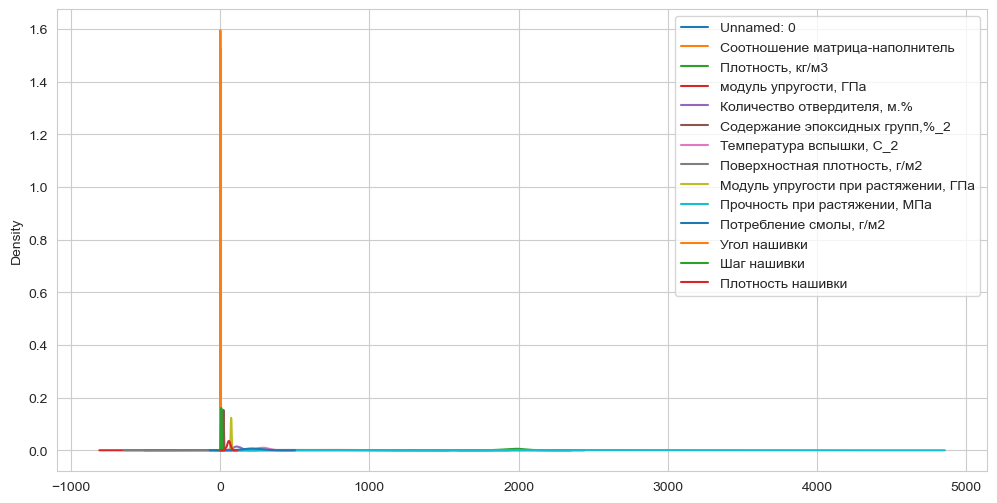


Рисунок 14 – преобразование значений угла нашивки

По условиям задания нормализуем значения. Для этого применим MinMaxScaler(), затем применим Normalizer(). Второе даёт нам больше выбросов.



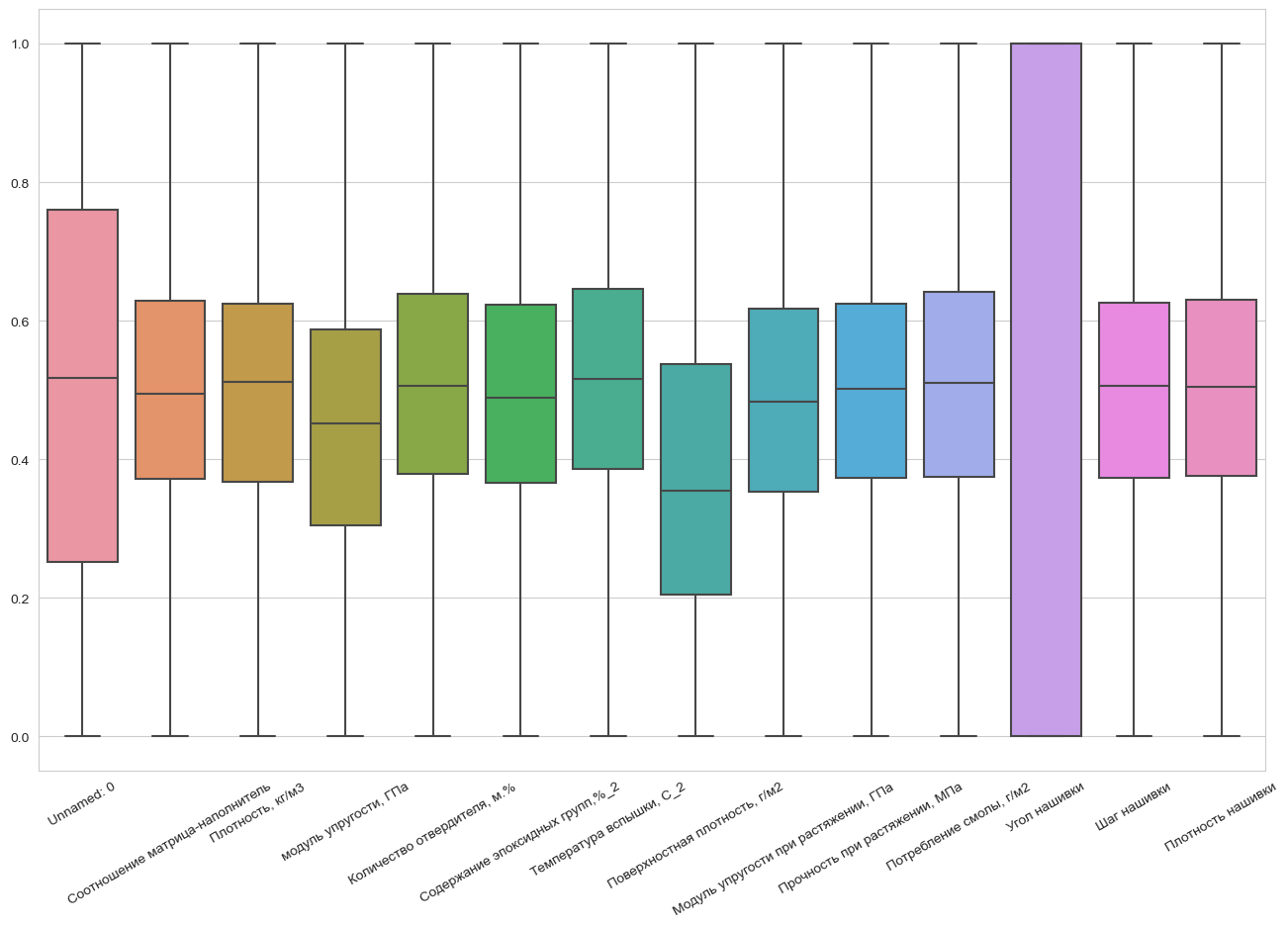
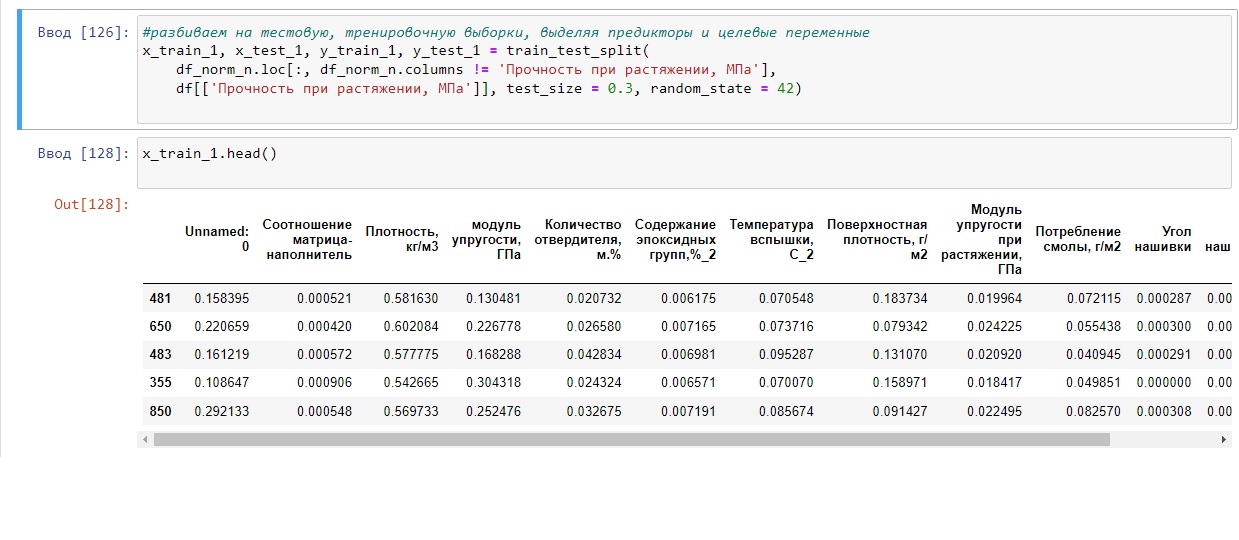


Рисунок 15 – результат нормализации данных с помощью MinMaxScaler

## **2.2. 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.

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%);



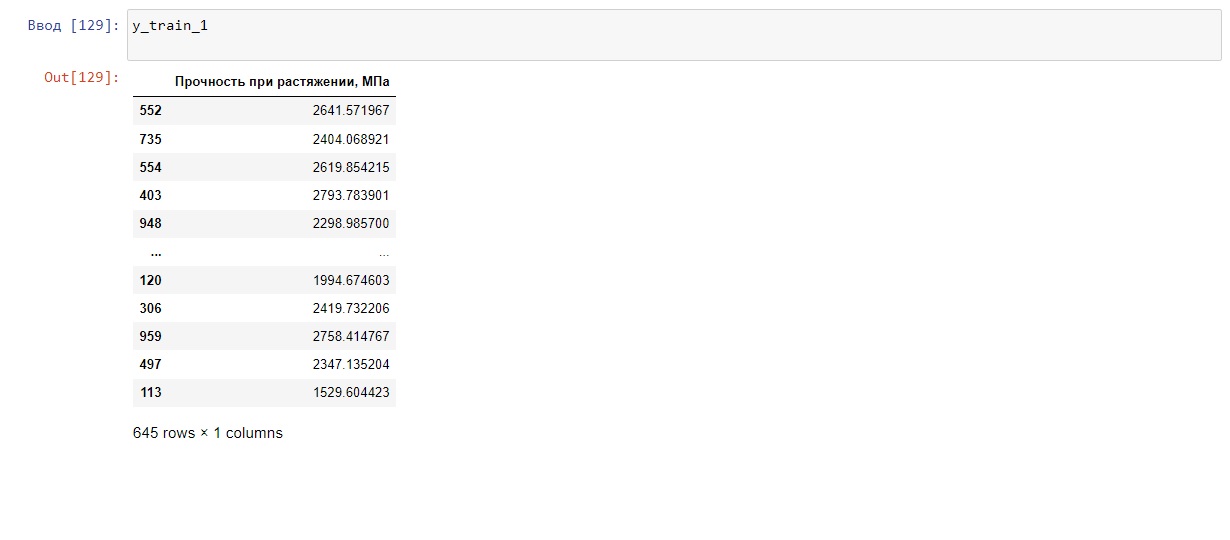


Рисунок 16 – разбиение на обучающую и тестовую выборки

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.

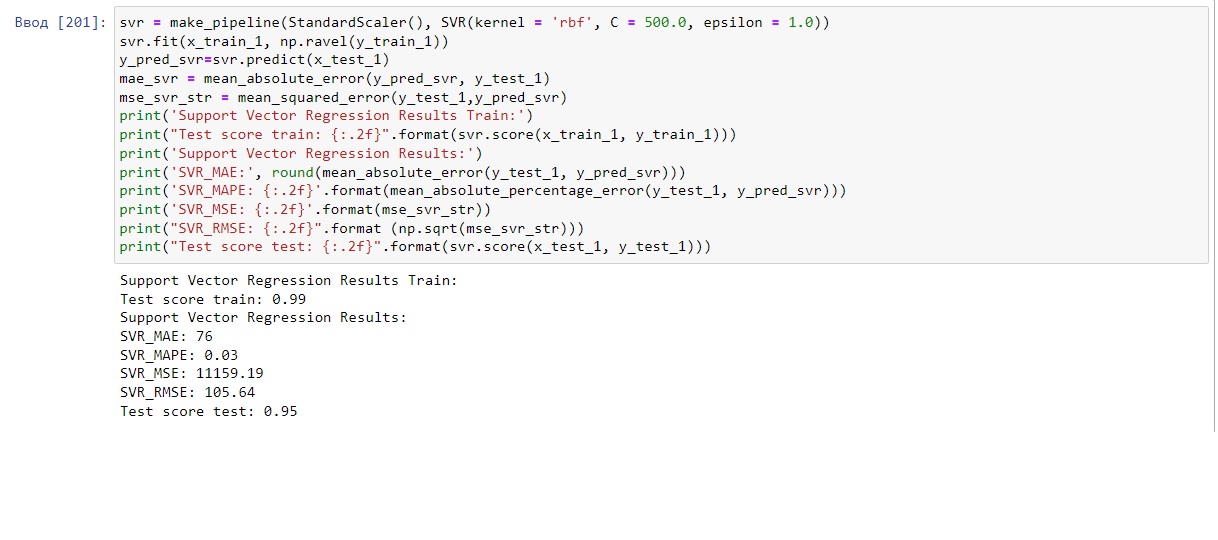


Рисунок 17 – результат модели метода опорных векторов

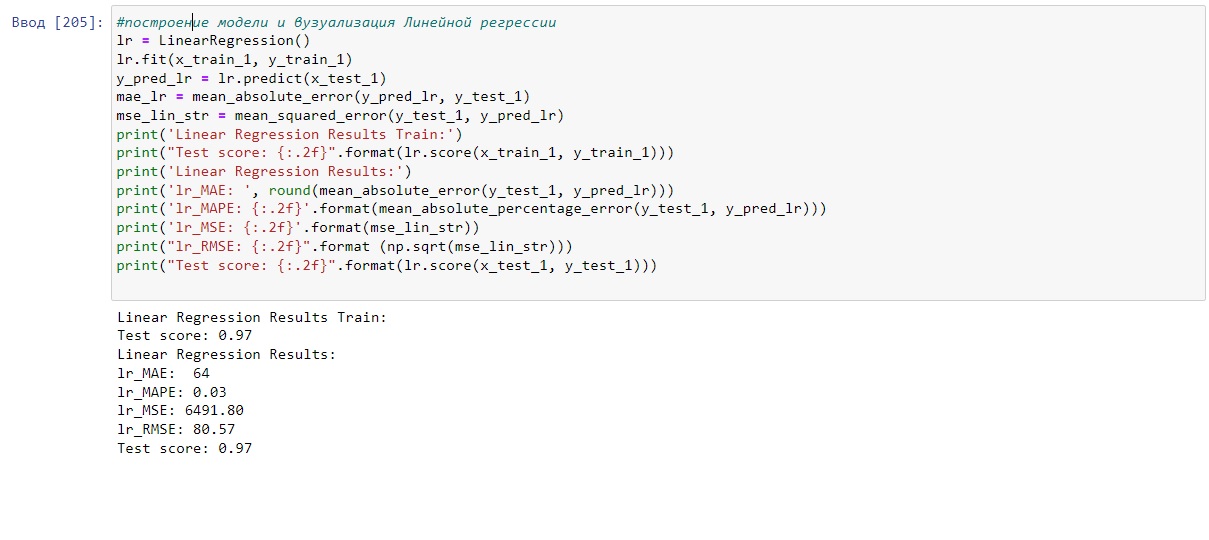
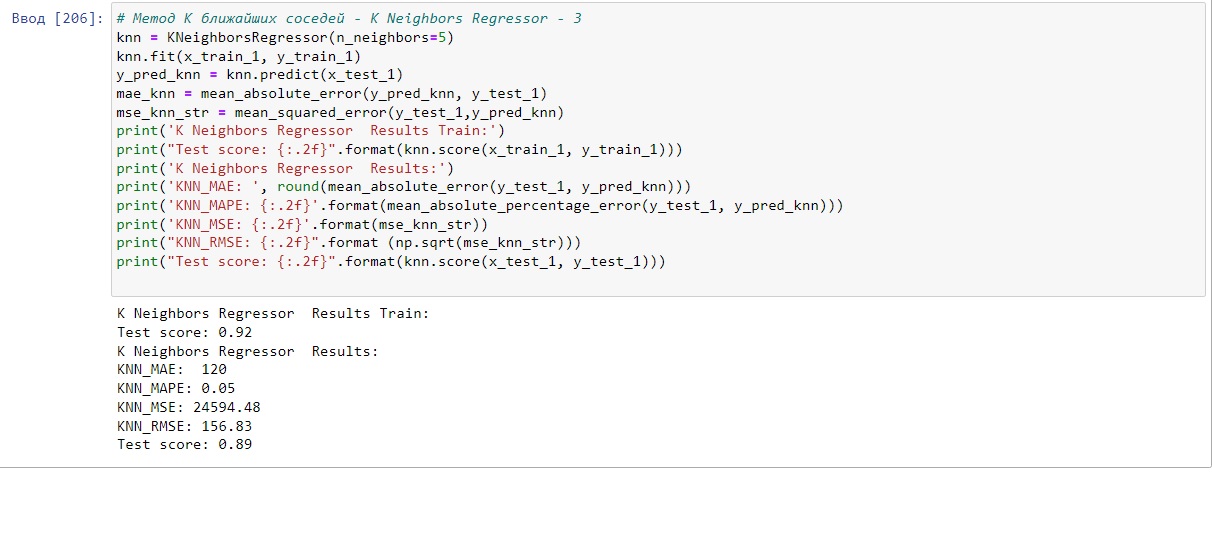


Рисунок 18 – результат модели метода линейной регрессии

Рисунок 19 – результат модели метода К ближайших соседей

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.

## **2.3. Тестирование модели**

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.

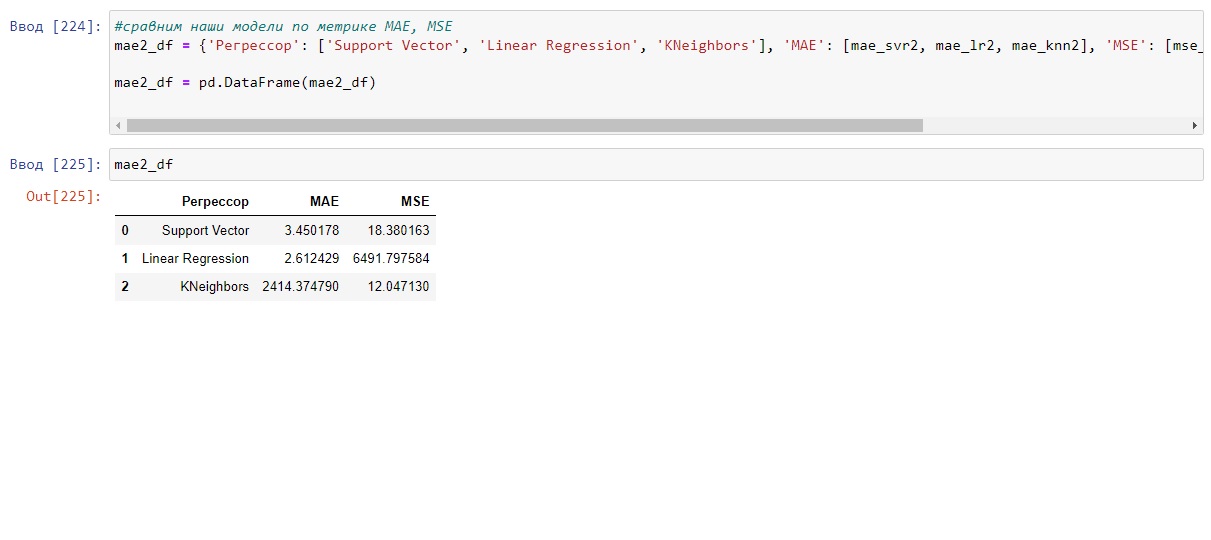


Рисунок 20 – сравнение результатов моделей по метрикам

**2.4. 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 стоить нейронную сеть с помощью класса keras.Sequential.



Рисунок 21 - 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.

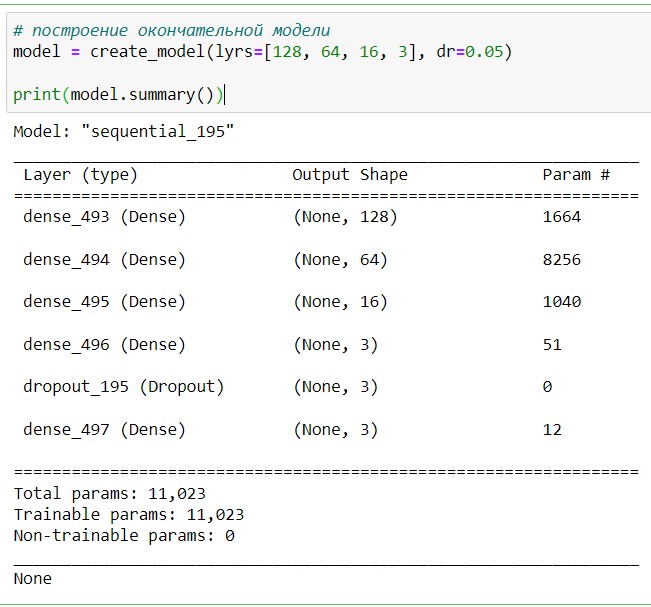


Рисунок 22 - построение первой нейросети

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

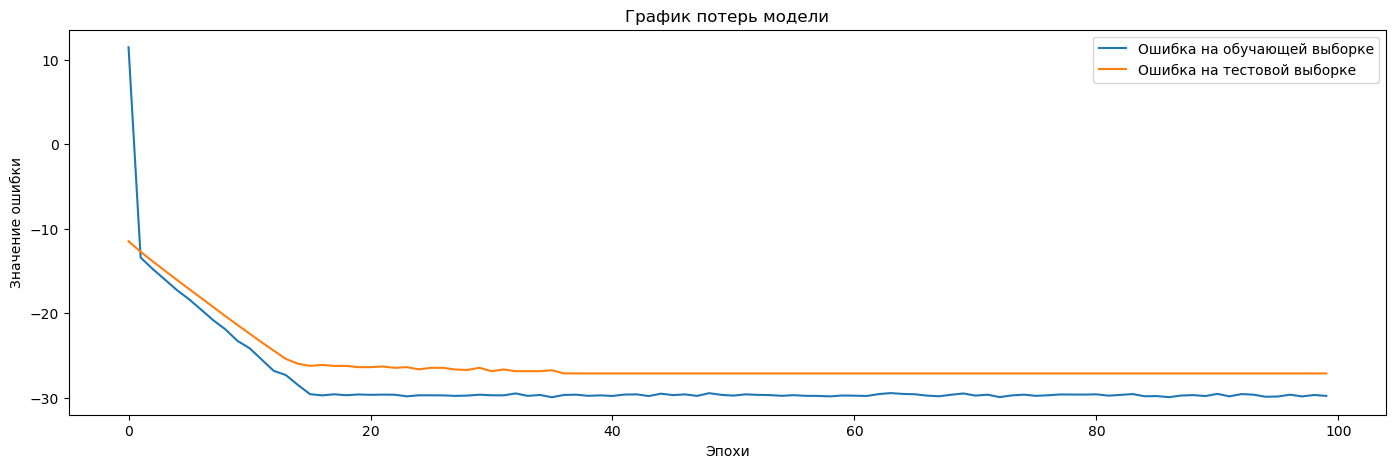


Рисунок 23 – график потерь модели 1

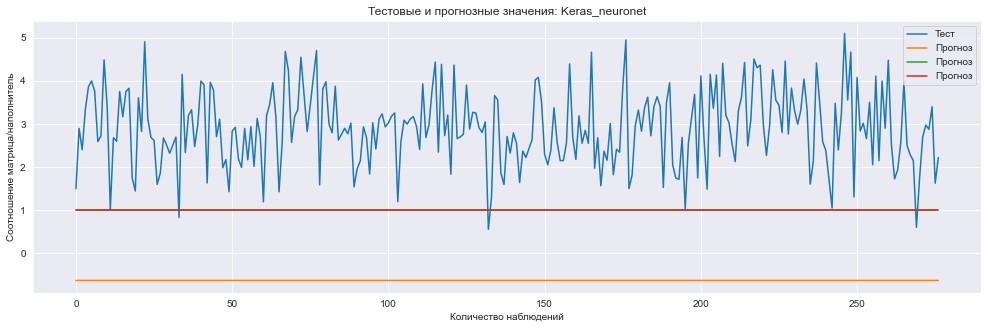


Рисунок 24 - тестовые и прогнозные значения модели 1

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

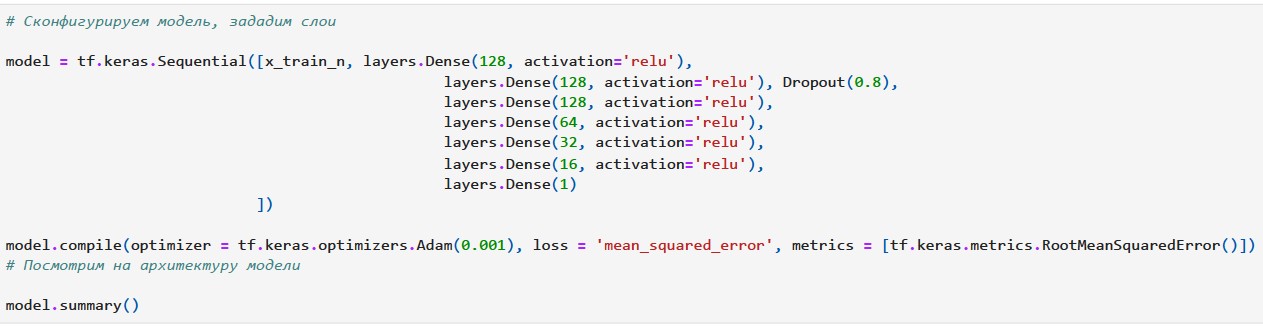


Рисунок 27 – создание слоев модели 2

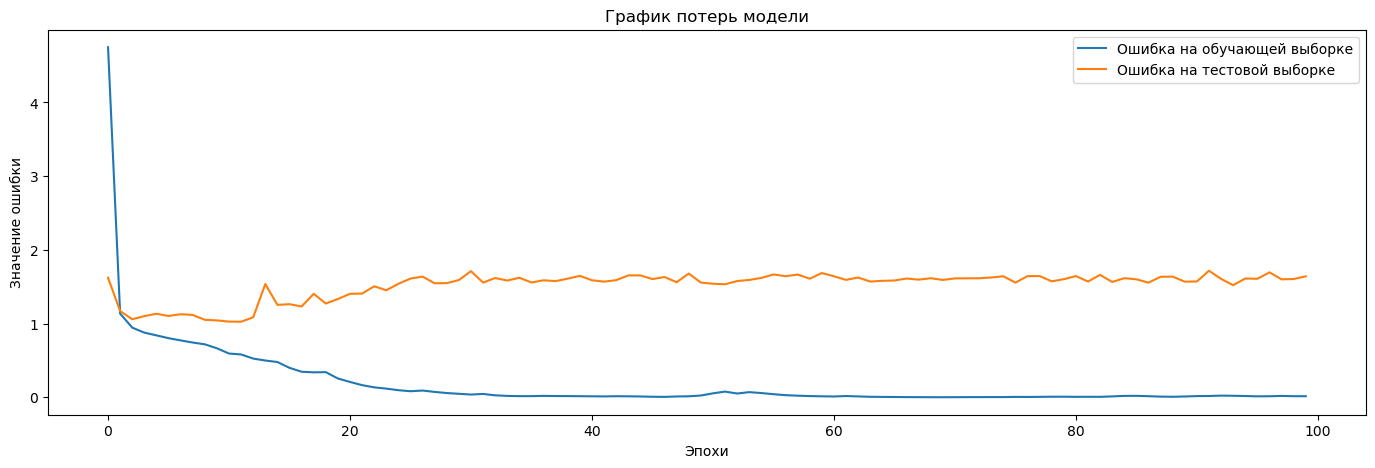


Рисунок 26 – график потерь модели 2

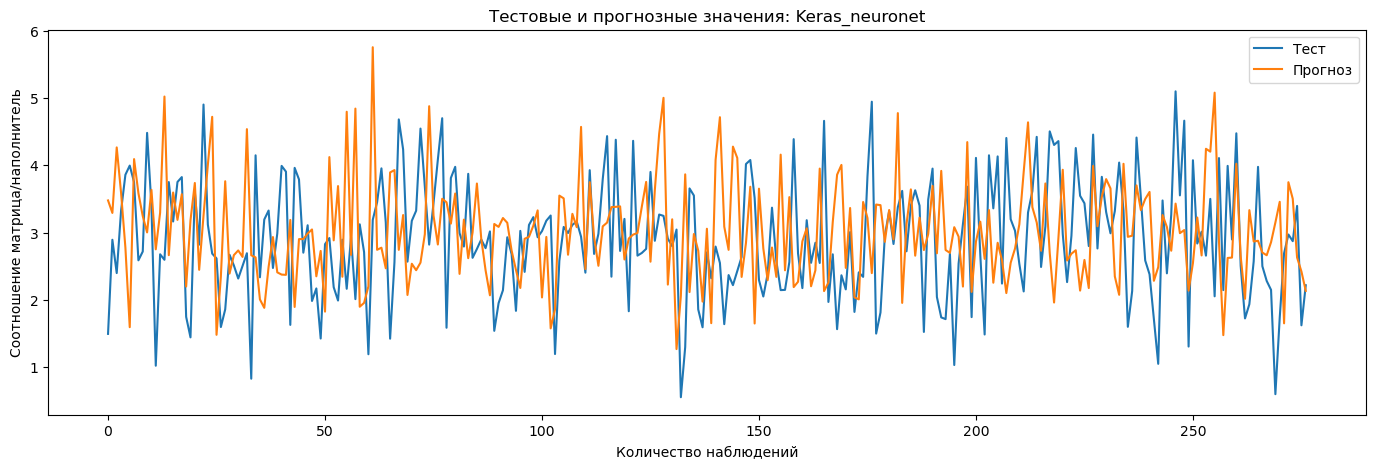


Рисунок 27 - тестовые и прогнозные значения модели 2

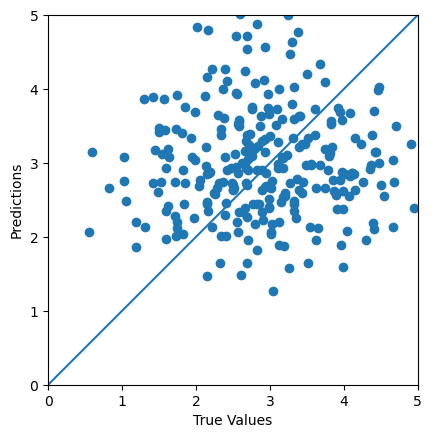


Рисунок 28 – график прогнозных и настоящих значений

В ходе работы был задействован дата-сет с реальными данными, произведена его подробная опись и сопутствующий анализ; построено множество разнообразных графиков; осуществлено разбиение данных на обучающую и тестовую выборки с использованием модулей из библиотеки SkLearn. В рамках машинного обучения и поиска гиперпараметров были задействованы несколько алгоритмов: линейная регрессия, К ближайших соседей, а также опорные вектора. Поиск гиперпараметров осуществлялся при помощи таких методов, как «GridSearch». Было представлено сравнение результатов оценок работы алгоритмов, а также различные графики и диаграммы, позволяющие наглядно оценить итоги проведенного обучения. Обучена нейронная сеть, предсказываюая вероятный прогноз по заданным параметрам. Результаты предсказаний регрессии и нейросети получили крайне неперспективные. Предполагаемо, из-за особенностей датасета. Требуется более детальный анализ его данных, обработка, возможно, разбиение на подсеты или(и) глубокие знания материаловедения в области композитных материалов.

## **2.5 Разработка приложения**

Приложение должно показывать результат прогноза для соотношения «матрица – наполнитель». Предположительно работающий код приложения и шаблон страницы ввода и получения данных был создан, но рабочего варианта приложения не удалось добиться. Проблема, предположительно в настройках веб-сервера, не удается запустить службу Apache (я так понял, она не дает открыть локал-хост), что не позволяет открыть адрес <http://127.0.0.1:5000/>, устанавливаемый Flask по умолчанию для работы приложения.

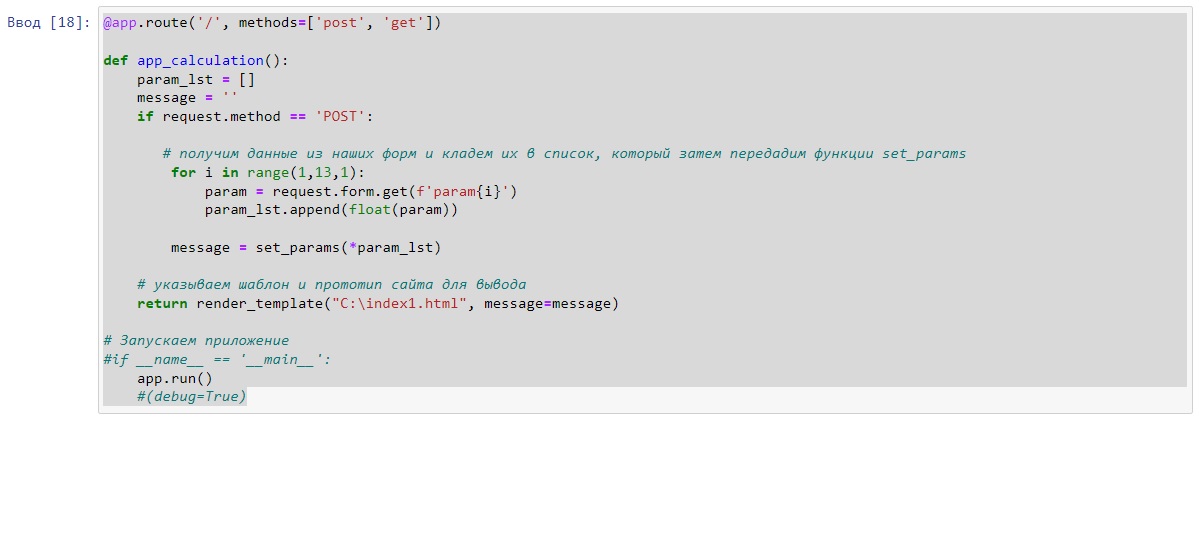
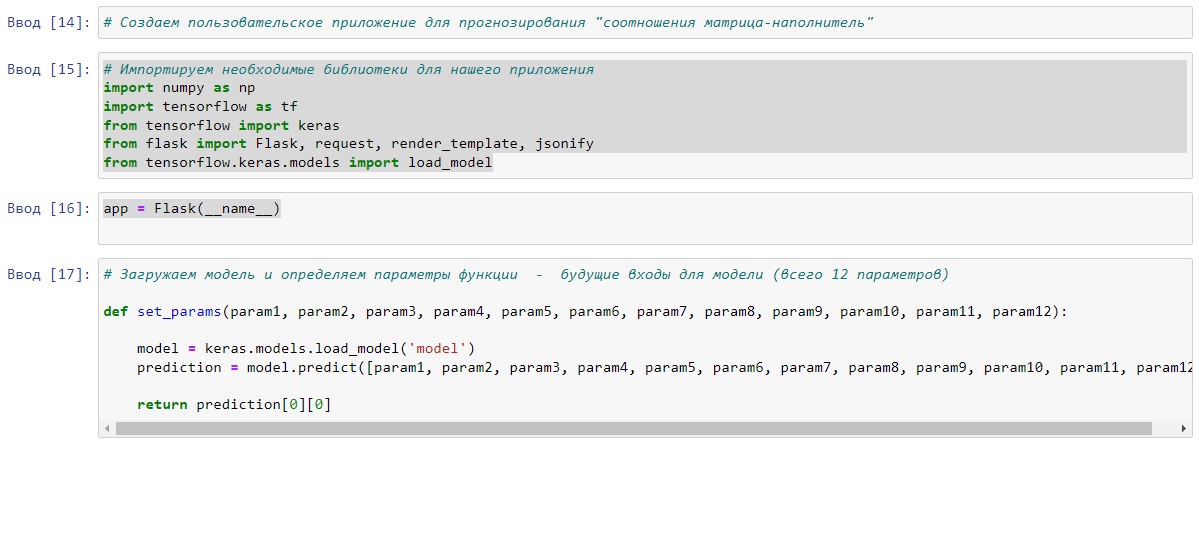


Рисунок 29 – код приложения Flask



Рисунок 30 - HTML страница – шаблон для ввода данных и получения прогноза в приложении.

**2.6. Создание удалённого репозитория и загрузка**

Репозиторий был создан на github.com по адресу:

<https://github.com/Pavel-Tsygankov/VKR>

## **2.7. 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 - лассо-регрессия.

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 учёных.

В целом прогнозирование конечных свойств/характеристик композитных материалов без изучения материаловедения, погружения в вопрос экспериментального анализа характеристик композитных материалов не демонстрирует скольконибудь 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.

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

1. Материалы курса Data Science Центра дополнительного образования МГТУ им. Н.Э. Баумана
2. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour): – Режим доступа: [https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-](https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19)

[2021-07-19.](https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19) (дата обращения: 07.06.2022)

1. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать: – Режим доступа: https://habr.com/ru/company/vk/blog/513842/(дата обращения: 01.06.2022).
2. Документация по библиотеке keras: – Режим доступа: [https://keras.io/api/.](https://keras.io/api/)(дата обращения: 08.06.2022).
3. Документация по библиотеке matplotlib: – Режим доступа: [https://matplotlib.org/stable/users/index.html.](https://matplotlib.org/stable/users/index.html) (дата обращения: 10.06.2022)
4. Документация по библиотеке numpy: – Режим

доступа: [https://numpy.org/doc/1.22/user/index.html#user.](https://numpy.org/doc/1.22/user/index.html#user) (дата обращения:

03.06.2022).

1. Документация по библиотеке pandas: – Режим доступа: [https://pandas.pydata.org/docs/user\_guide/index.html#user-guide.](https://pandas.pydata.org/docs/user_guide/index.html#user-guide) (дата обращения: 04.06.2022).
2. Документация по библиотеке scikit-learn: – Режим доступа: [https://scikit-learn.org/stable/user\_guide.html.](https://scikit-learn.org/stable/user_guide.html) (дата обращения: 05.06.2022).
3. Документация по библиотеке seaborn: – Режим доступа: [https://seaborn.pydata.org/tutorial.html.](https://seaborn.pydata.org/tutorial.html) (дата обращения: 06.06.2022).
4. Документация по библиотеке Tensorflow: – Режим доступа:

<https://www.tensorflow.org/overview>(дата обращения: 10.06.2022).

1. Документация по языку программирования python: – Режим доступа: [https://docs.python.org/3.8/index.html.](https://docs.python.org/3.8/index.html) (дата обращения: 02.06.2022).
2. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) – Режим доступа: <https://habr.com/ru/post/428503/>(дата обращения 07.06.2022)
3. Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.:

Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил.

1. Руководство по быстрому старту в flask: – Режим доступа: [https://flaskrussian-docs.readthedocs.io/ru/latest/quickstart.html.](https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html)