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 Долганов Денис Геннадьевич (ФИО)

Москва, 2022

Оглавление

[Введение 3](#_Toc101992618)

[Аналитическая часть 4](#_Toc101992619)

[Постановка задачи. 4](#_Toc101992620)

[Смысловое описание решаемой задачи. 4](#_Toc101992621)

[Характеристика датасета 4](#_Toc101992622)

[Описание используемых методов 10](#_Toc101992623)

[Разведочный анализ данных 11](#_Toc101992624)

[Практическая часть 12](#_Toc101992625)

[Предобработка данных 12](#_Toc101992626)

[Разработка и обучение модели 17](#_Toc101992627)

[Тестирование модели 17](#_Toc101992628)

[Написать нейронную сеть, которая будет рекомендовать соотношение матрица. 19](#_Toc101992629)

[Разработка приложения 20](#_Toc101992630)

[Создание удаленного репозитория и загрузка результатов работы на него 20](#_Toc101992631)

[Заключение 21](#_Toc101992632)

[Список литературы 22](#_Toc101992633)

# 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. В большинстве композитов компоненты можно разделить на матрицу и включённые в неё армирующие элементы (или наполнители). В композитах конструкционного назначения армирующие элементы обычно обеспечивают необходимые механические характеристики материала (прочность, жёсткость и т. д.), а матрица обеспечивает совместную работу армирующих элементов и защиту их от механических повреждений и агрессивной химической среды. При совмещении армирующих элементов и матрицы образуется композиция, обладающая набором свойств, отражающими не только исходные характеристики его компонентов, но и новые свойства, которыми отдельные компоненты не обладают.

Изделия из композиционных материалов применяются в различных отраслях промышленности: машиностроении, строительстве, радиоэлектронике, медицине и др.Ежегодно разрабатываются десятки и сотни новых композиционных материалов. Важнейшие эксплуатационные характеристики этих материалов определяют, в основном, экспериментальными методами, что требует сложного оборудования и значительных временных и материальных затрат. В этой связи теоретический анализ композиционных материалов путем построения моделей на основе методов вычислительной механики и прогнозирование их эффективных характеристик является актуальным.

Сегодня новые материалы разрабатываются уже не в лабораторных условиях, а с помощью цифрового подхода и суперкомпьютеров, обрабатывая большие объемы данных с применением искусственного интеллекта.

Данная выпускная квалификационная работа посвящена итоговому контрольному срезу по курсу Data Science. Целью ВКР является закрепление пройденного материала в части теории и практики. Также в качестве практической задачи было предложено рассмотреть актуальный кейс по созданию прогнозных моделей, которые помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

# Аналитическая часть

## Постановка задачи.

В данном разделе приводится в текстовом виде смысловое описание решаемой задачи анализа данных. Также дается характеристика датасета – число входных и выходных переменных (факторов), объем выборки, характеристика выборки с точки зрения ее особенностей (выбросы, пропуски и т.д.).

## Смысловое описание решаемой задачи.

Имея данные о свойствах отдельно взятых материалов, предсказать свойства композитов, созданных из исходных компонентов. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

## Характеристика датасета

Выборка «x\_bp.df» размерностью (1024, 10)

Выборка «x\_nup.df» размерностью (1041, 3)

После объединения датасетов и удаления строки с нулевыми значениями получили объединенную выборку «x\_bp\_nup.df» размерностью (1022, 13)

**Входные параметры, 10 шт:**

* Плотность, кг/м3
* Модуль упругости, ГПа
* Количество отвердителя, м.%
* Содержание эпоксидных групп,%\_2
* Температура вспышки, С\_2
* Поверхностная плотность, г/м2
* Потребление смолы, г/м2
* Угол нашивки, град 1040
* Шаг нашивки
* Плотность нашивки

**Выходные параметры, 3 шт:**

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

Методом isnull() определили, что пропусков нет.

Методотом duplicated() определили, что дубликатов нет

Для определения наличия выбросов был построена диаграмма «Ящик с усами», который показал наличие выбросов во всех параметрах, кроме параметра «Угол нашивки». Также данная диаграмма показала сильный разброс параметров по диапазону.

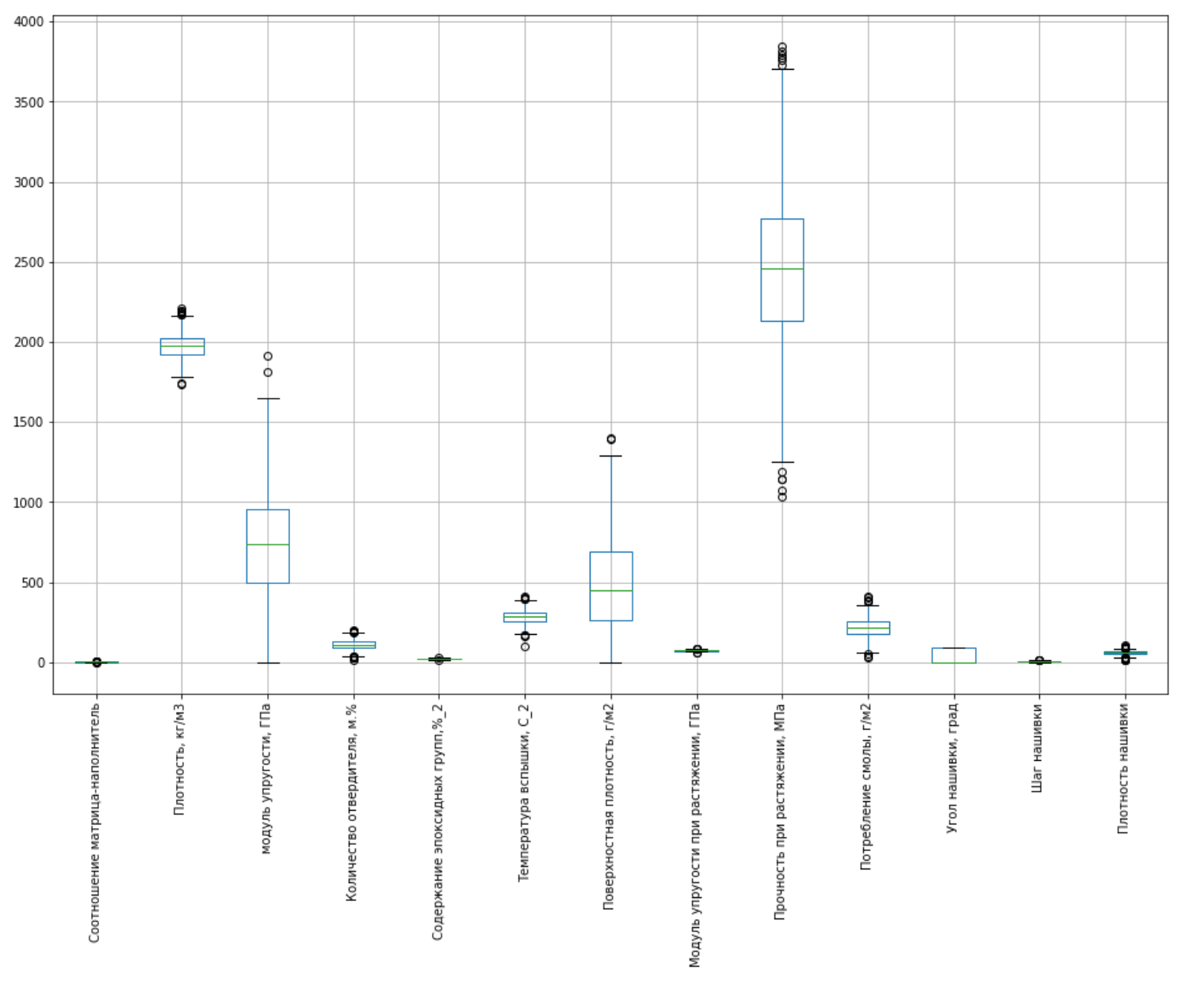


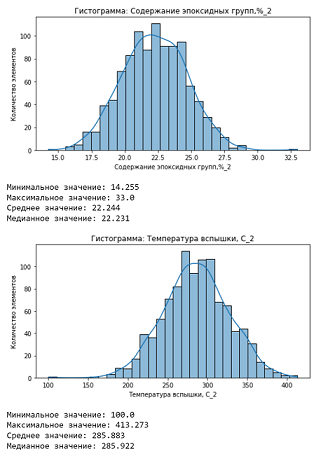
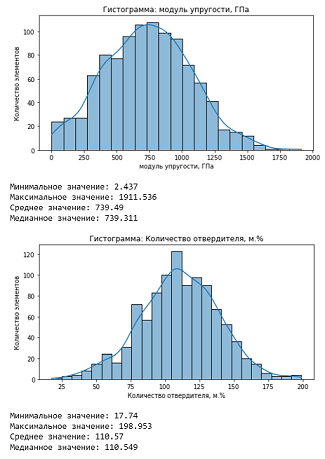
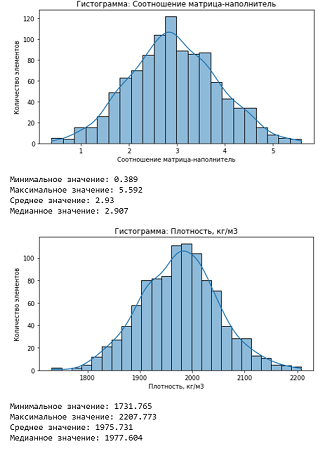
Рисунок 1. Диаграмма «Ящик с усами»

Использовали Метод .describe () для получения сводки числовых значений в нашем наборе данных. Он вычисляет среднее значение, стандартное отклонение, минимальное значение, максимальное значение, 1-й процентиль, 2-й процентиль, 3-й процентиль столбцов с числовыми значениями. Он также подсчитывает количество переменных в наборе данных.



Рисунок 2. Описательная статистика объединенного датасета

Построение гистограмм распределения для каждой переменной выявило нормальное распределение данных, даже излишне нормальное. Это позволило сделать предположение о недостаточном кол-ве реальных наблюдений, о применении метода увеличения выборки путем добавления усредненных данных.



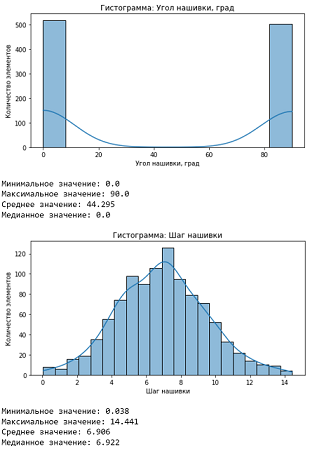
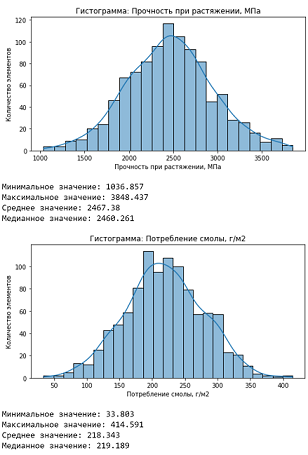
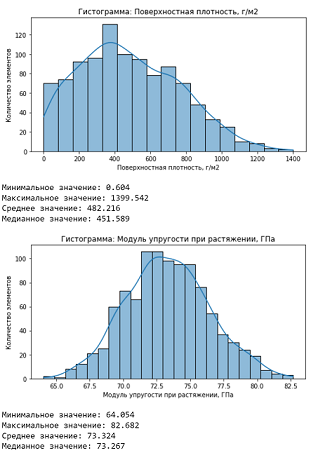




Рисунок 3. Гистограммы распределения признаков

Построение попарных графиков рассеяния показало отсутствие зависимости данных друг от друга.

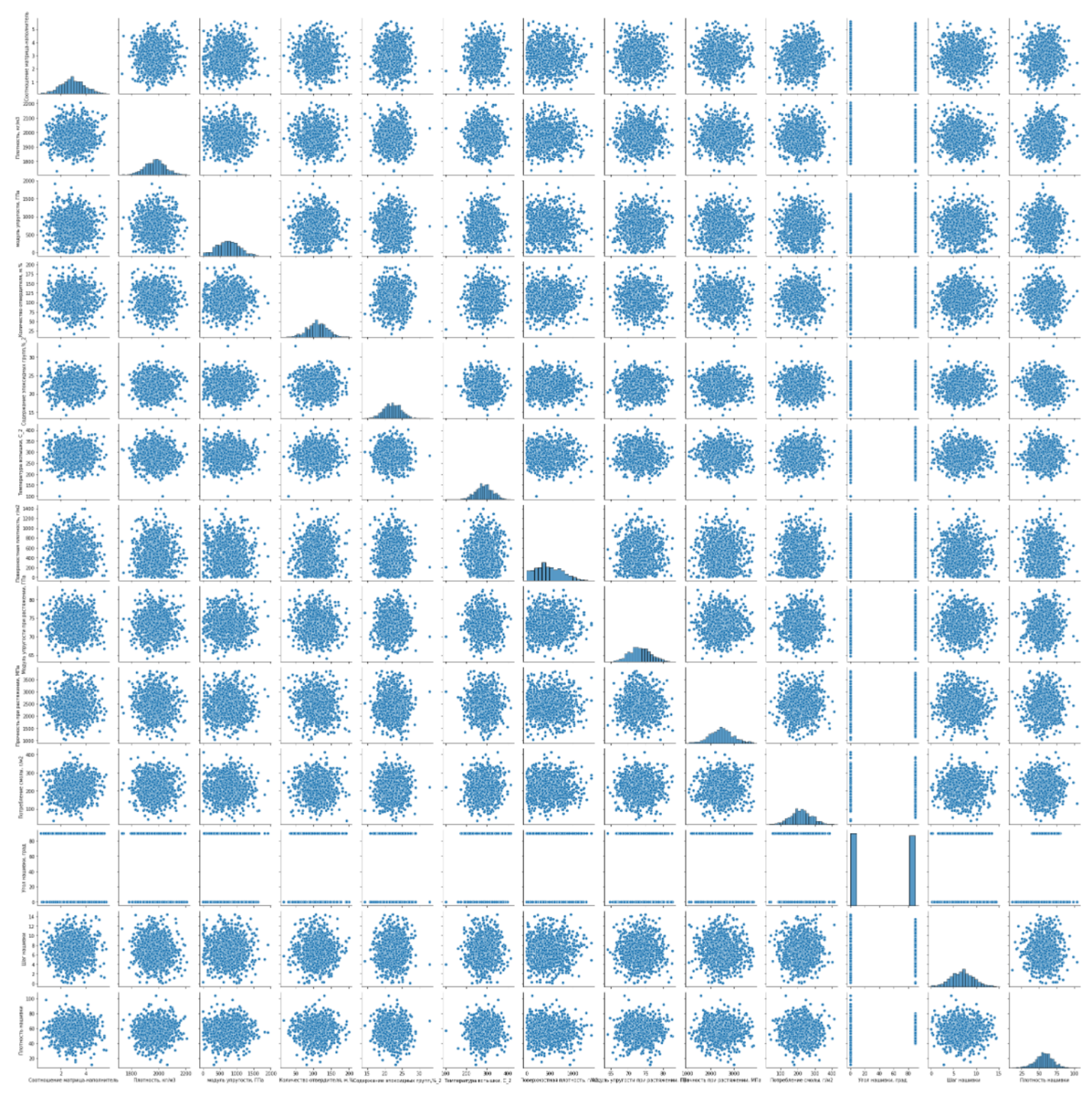


Рисунок 4. Попарные графики рассеяния признаков

Максимальная корреляция выявилась на уровне 0,11 у показателей Угол нашивки и Плотность нашивки. По всем остальным параметрам корреляция близится к нулю, что говорит об отсутствии корреляционных связей между параметрами.

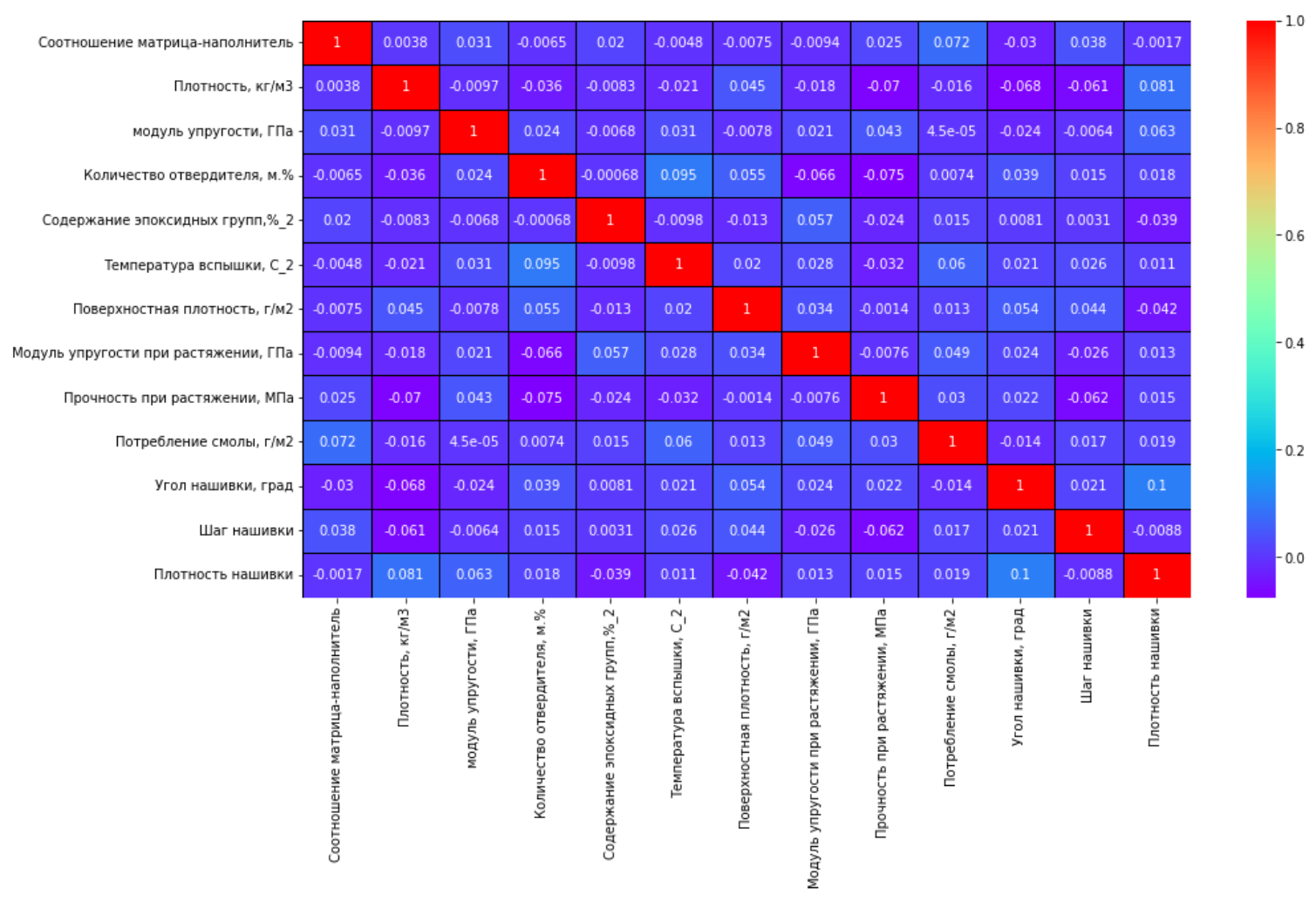


Рисунок 5. Тепловая карта признаков

Таким образом, имеем объединенную выборку размерностью (1022, 13), без пропусков, с нормально распределёнными данными, не зависящими друг от друга, содержащими небольшое количество выбросов, требующими проведения нормализации.

## Описание используемых методов

В ходе предварительного анализа данных были использованы такие методы как: describe, info, head, unique, shape, drop, merge. Все эти методы возвращают информацию об имеющемся датасете, а также позволяют делать такие манипуляции как объединение и отброс данных.

В ходе разведочного анализа данных были использованы такие методы как: isnull (поиск пропусков), duplicated (поиск дублей). Для поиска выбросов была построена диаграмма Ящик с усами (boxplot). Для определения нормального распределения были построены гистограммы (hist). Для оценки показателей корреляции использовались попарные графики рассеяния (pairplot) и тепловые карты (heatmap). Оценка выбросов производилась методом межквартильного расстояний и методом трех сигм. Нормализация производилась методом MinMaxScaler. Выбор моделей регрессоров был рандомным: линейная регрессия, метод К ближайших соседей, случайный лес. Наиболее простой (линейная регрессия) и наиболее универсальные (К соседей и случайный лес). В качестве модели для нейронной сети была выбрана последовательная модель Sequential.

## Разведочный анализ данных

Методы разведочного анализа данных:

Наличие выбросов определялось с помощью диаграммы «Ящик с усами».

Количество выбросов определялось двумя методами: методом Трёх сигм, а также методом Межквартильных расстояний. Далее, для удаления выбросов был выбран метод, определивший наименьшее количество выбросов – метод Трёх сигм (24 выброса, против 92 выбросов в методе Межквартильных расстояний. Такой выбор был сделан, исходя из предположений о недостаточном количестве в смоделированной выборке реальных данных, полученных опытным путем.

Описательная статистика, диаграмма «Ящик с усами», показали, что данные имеют большой разброс по диапазонам. Для визуализации разброса использовали функцию subplots.

Нормализация данных производилась методом MinMaxScaller.

# Практическая часть

## Предобработка данных

В данном разделе приводятся графики распределения для каждого признака до и после нормализации, указаны максимальные и минимальные значения.

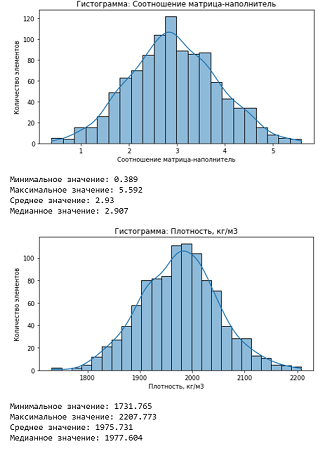
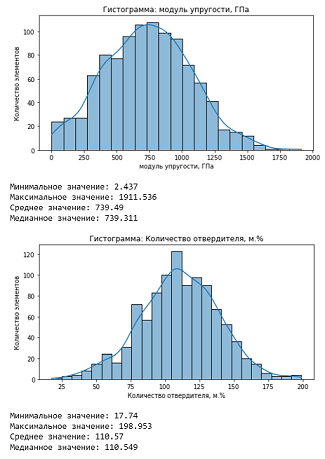
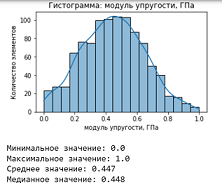
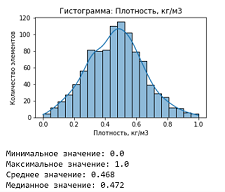
Рисунок 6. Описательная статистика до проведения нормализации

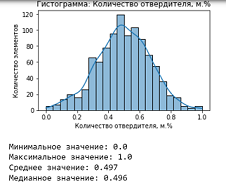


Рисунок 7. Описательная статистика после проведения нормализации

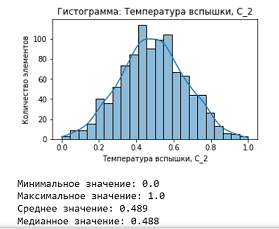
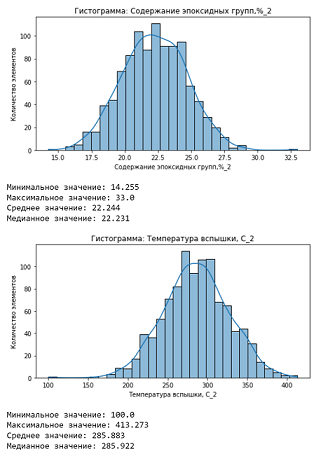
Видим, что после нормализации значения параметров находятся в диапазоне от нуля до единицы.

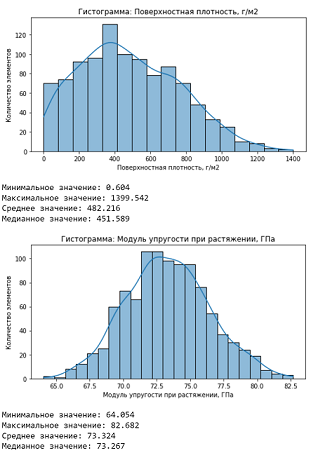
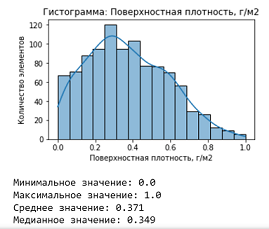
Сравним гистограммы распределения параметров До и После нормализации:





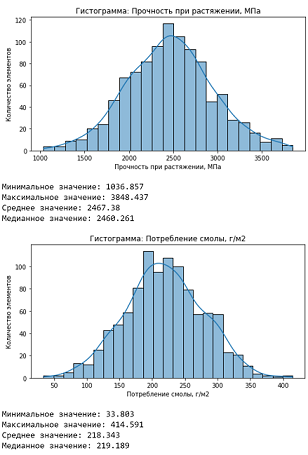


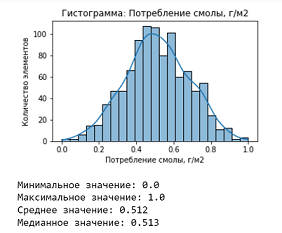












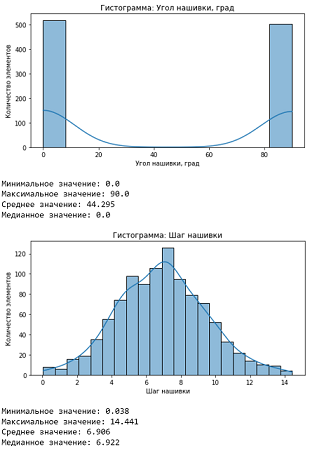
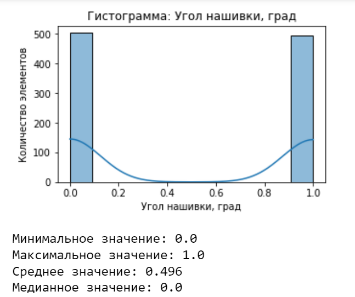






Рисунок 8. Гистограммы распределения параметров До и После нормализации

Видим, что данные имеют нормальное распределение До и После нормализации.

Сравним диаграммы «Ящик с усами» До и После нормализации.

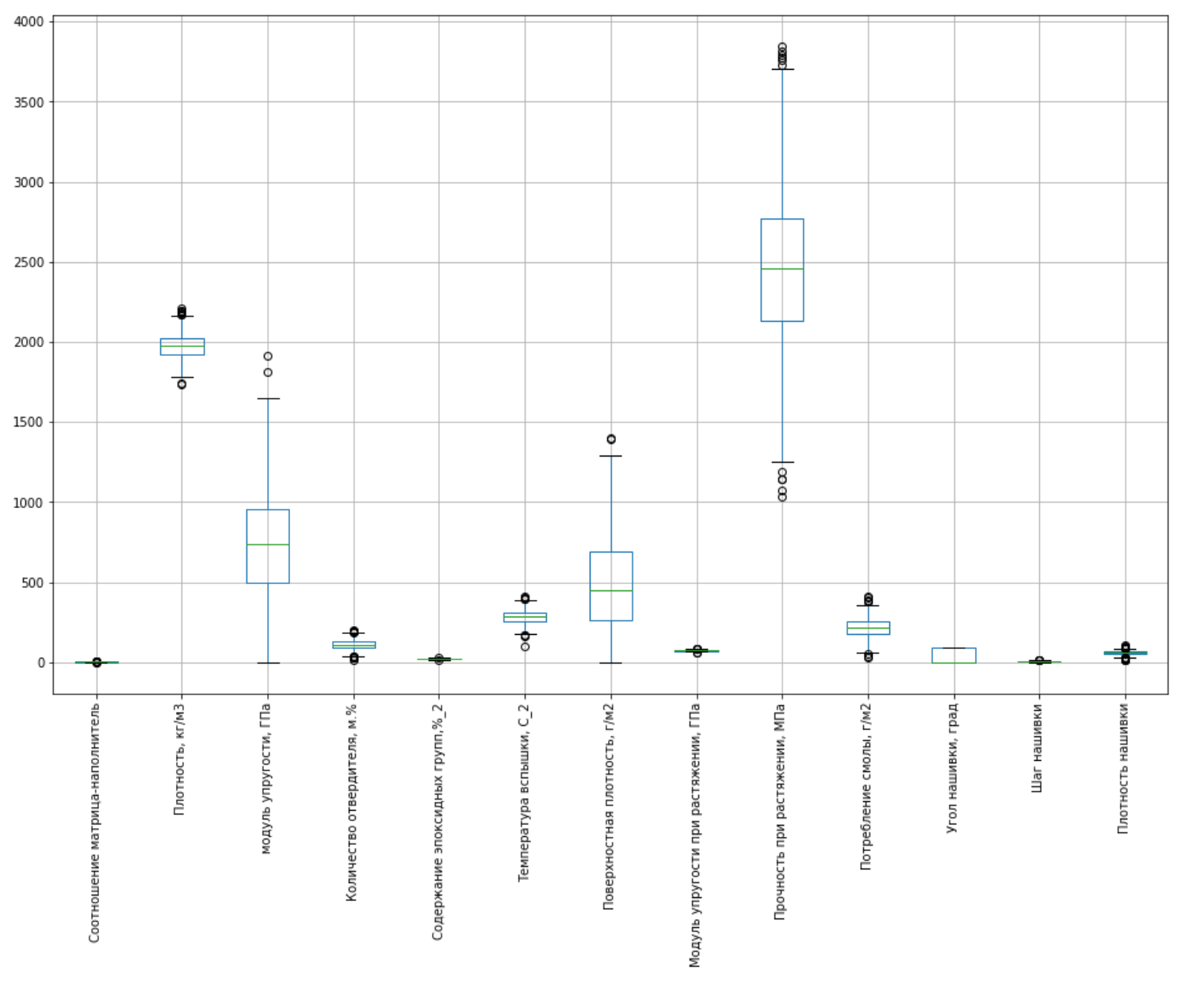


Рисунок 9. Диаграмма «Ящик с усами» До нормализации и До удаления выбросов

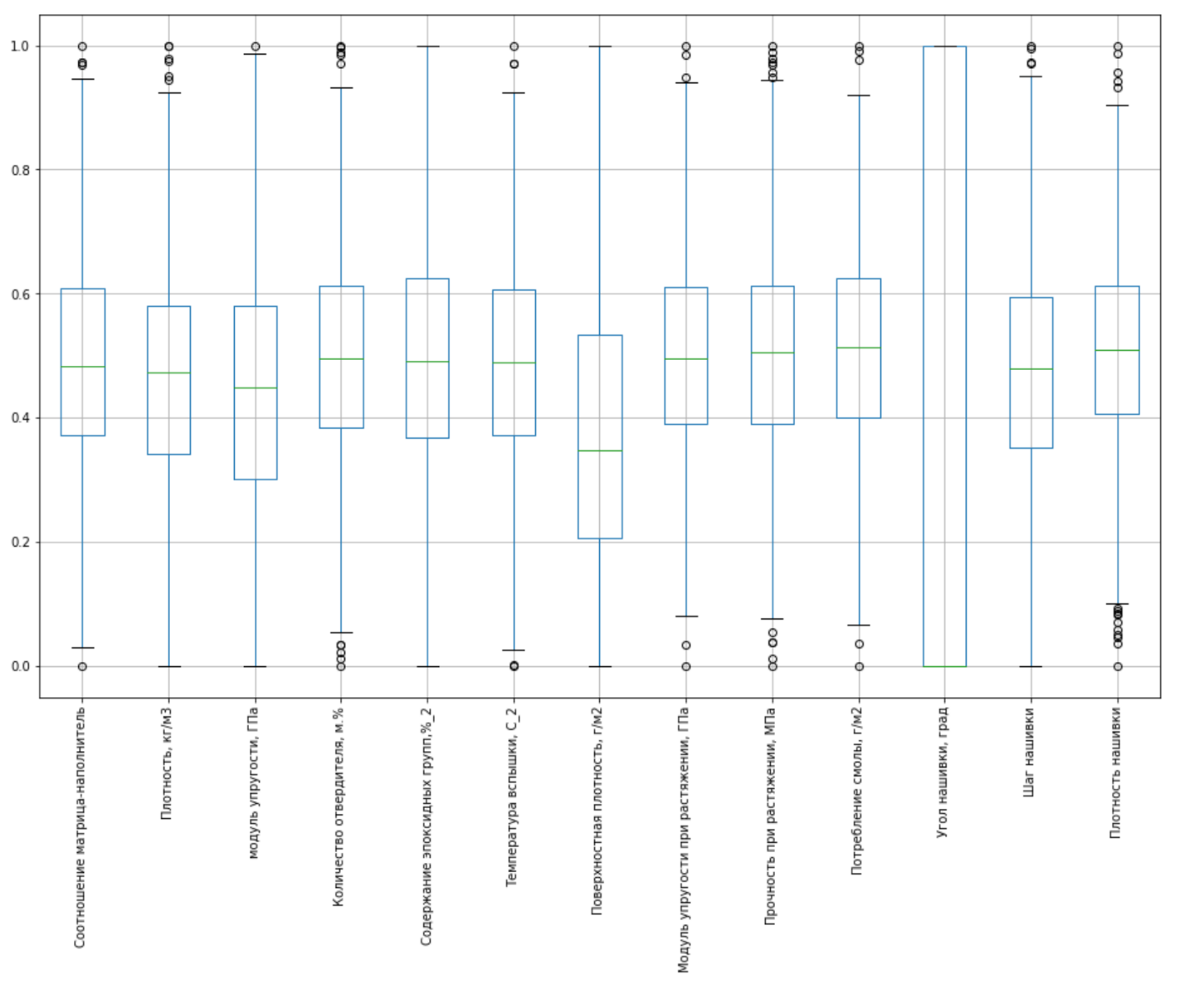


Рисунок 10. Диаграмма «Ящик с усами» после нормализации

Виден результат нормализации: данные принимают значения в диапазоне от нуля до единицы.

## Разработка и обучение модели

В данной части приводится список моделей, которые будут использоваться для прогноза модуля упругости при растяжении и прочности при растяжении.

Для прогноза модуля упругости при растяжении и прочности при растяжении обучались следующие модели:

Линейная регрессия LinearRegression() – простая модель, используется на небольших датасетах, исследует взаимосвязь данных, очень чувствительна к выбросам.

Метод К ближайших соседей KNeighborsRegressor() – модель очень чувствительна к нормализации данных, универсальная модель (подходит как для задач классификации, так и для задач регресссии), нет необходимости настройки нескольких параметров, оптимальным параметром k является 5. Чувствительна к размерам дата сета.

Случайный лес RandomForestRegressor() – универсальная модель, нечувствительная к выбросам.

По заданию необходимо было использовать 10 фолдов.

Выборка делится на обучающую и тестовую в соотношении 70% на 30%.

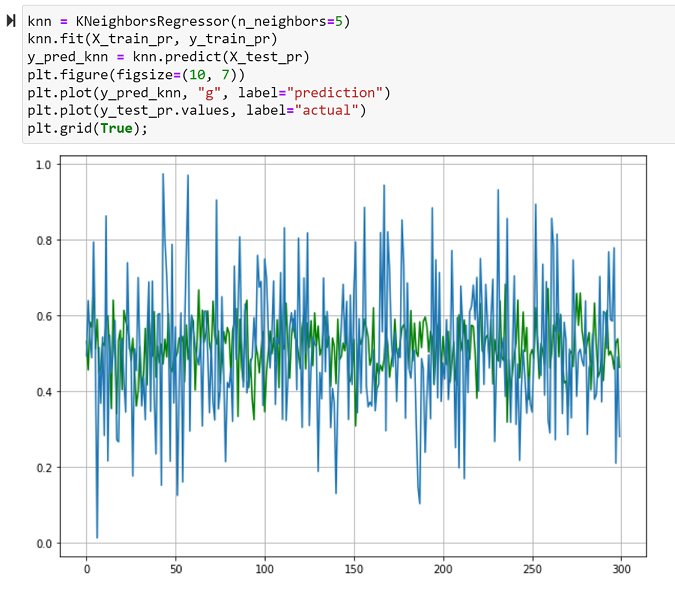
## Тестирование модели

В данном разделе показывается ошибка каждой модели на тренировочной и тестирующей части выборки. Обосновывается выбор модели.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Model** | **MAE** | **R2 score** |
| **Прочность при растяжении** | **KNeighborsRegressor\_pr** | **0.138447** | **-0.0** |
| **Модуль упругости при растяжении** | **KNeighborsRegressor\_upr** | **0.131265** | **0.004** |
| **Прочность при растяжении** | LinearRegression\_pr | 0.141692 | -0.025 |
| **Модуль упругости при растяжении** | LinearRegression\_upr | 0.131349 | -0.0 |
| **Прочность при растяжении** | RandomForestRegressor\_pr | 0.140092 | -0.013 |
| **Модуль упругости при растяжении** | RandomForestRegressor\_upr | 0.131213 | -0.002 |

Наилучшим образом себя показала модель Метод К ближайших соседей.

На втором месте модель Случайный лес.



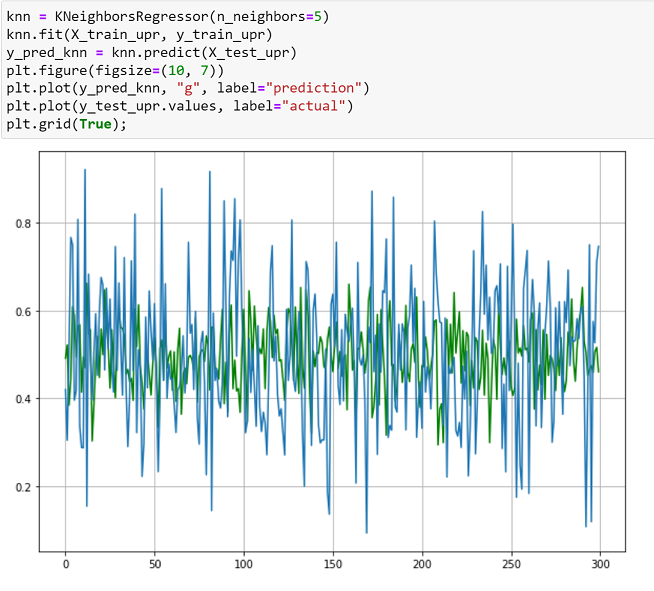


Рисунок 11. График обучения модели К ближайших соседей

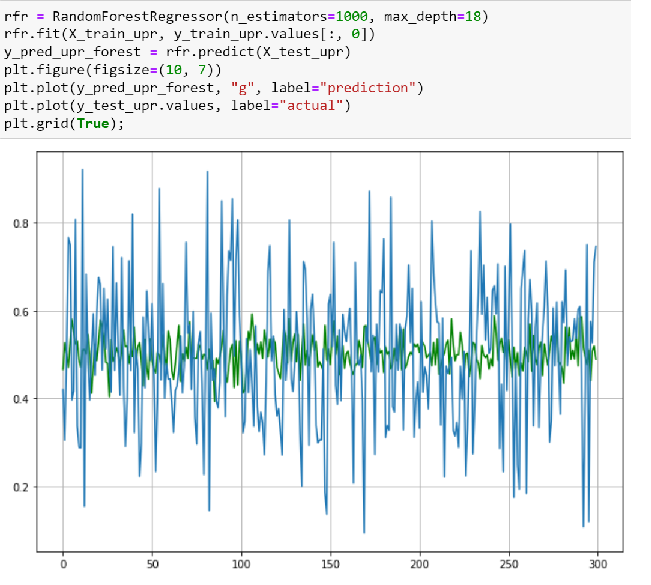
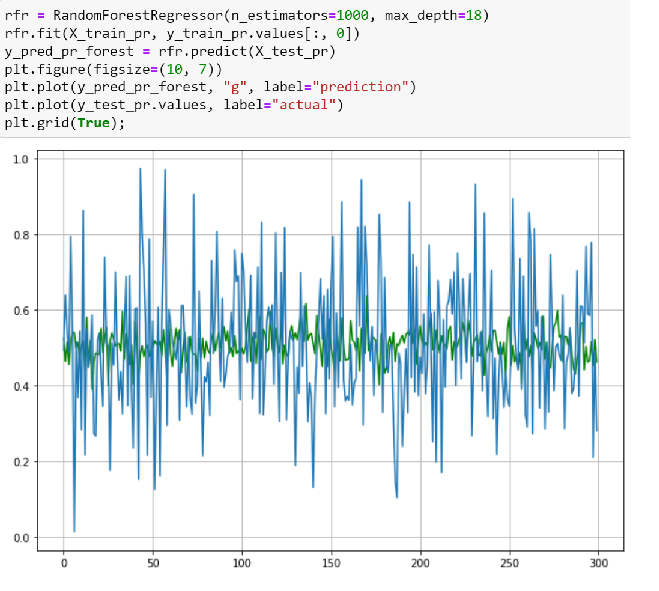


Рисунок 12. График обучения модели Случайный лес

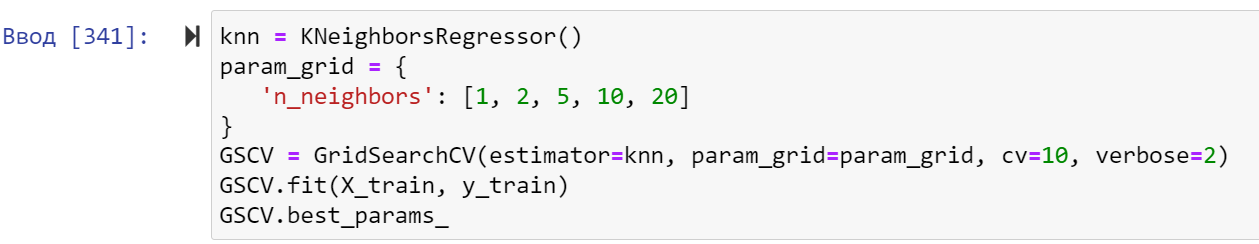
## Написать нейронную сеть, которая будет рекомендовать соотношение матрица.

Описывается выбранная архитектура нейронной сети и ее результаты.

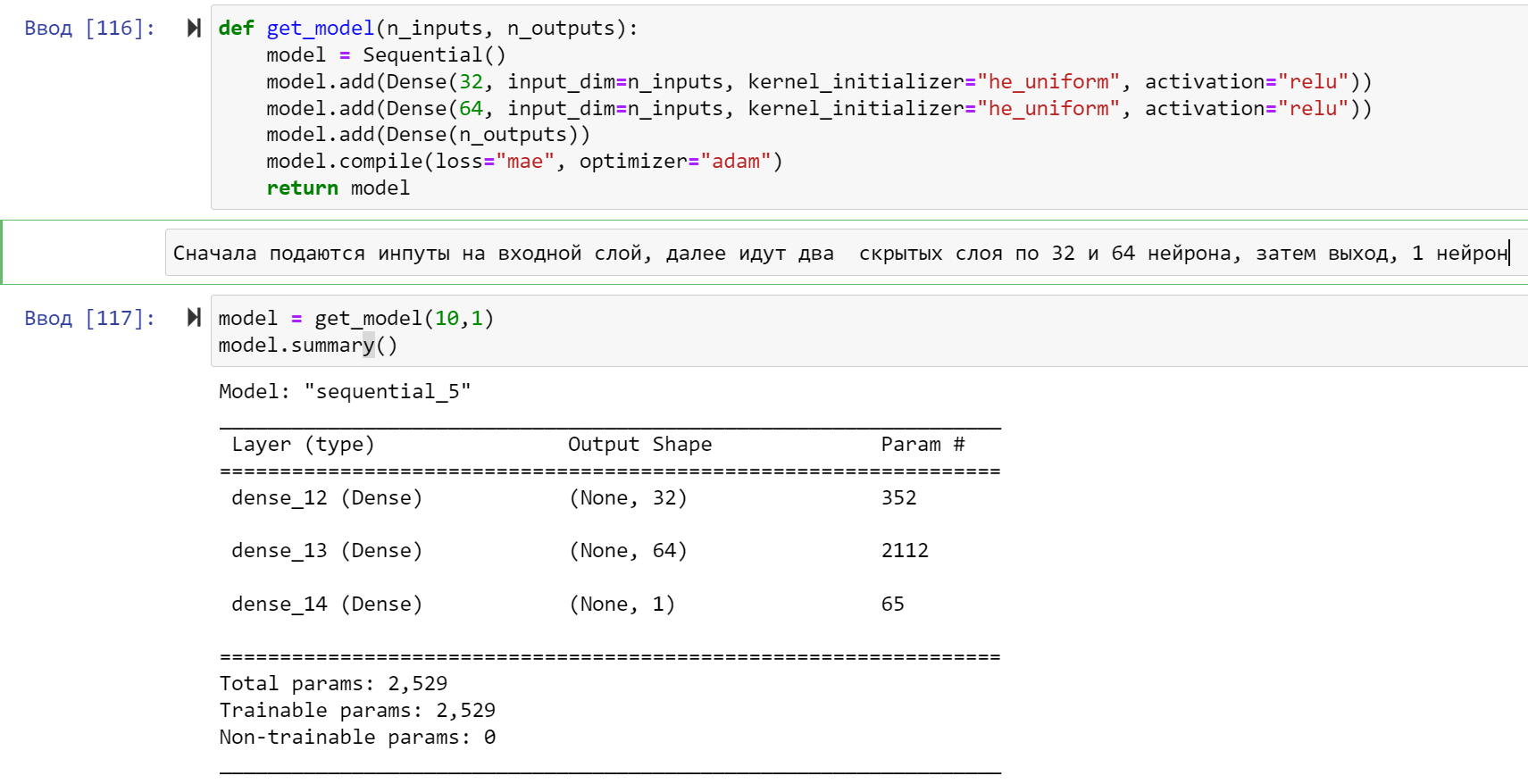
Создаем датасет для НС: на вход подаются 12 параметров, на выходе – параметр Матрица-наполнитель.

Делим выборку на обучающую и тестирующую в соотношении 80% на 20%.

Используем модель НС К ближайших соседей.



Последовательная модель Sequential.



|  |  |
| --- | --- |
| **Соотношение матрица-наполнитель** | |
|  | **R2 score** |
|  |  |
| Sequential | 0.020377 |
| **К ближайших соседей** | **0.037823** |

## **Разработка приложения**

Описать функционал приложения и краткую инструкцию использования.

Данный раздел находится в разработке.

## Создание удаленного репозитория и загрузка результатов работы на него

Указывается страница слушателя, созданный репозиторий, коммиты в репозитории.

https://github.com/DenXM/VKR\_Dolganov/upload/main

# 

# Список литературы

1. Лекции Образовательного центра МГТУ им. Баумана
2. <https://habr.com/ru/all/>