# 

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 Деженин Е.Ю. (ФИО)

Москва, 2022

Содержание

[Введение 3](#_Toc106370165)

[1. Аналитическая часть 4](#_Toc106370166)

[1.1. Постановка задачи. 4](#_Toc106370167)

[1.2. Описание используемых методов. 6](#_Toc106370168)

[1.2.1 Линейная регрессия 7](#_Toc106370169)

[1.2.2. К-ближайших соседей 8](#_Toc106370170)

[1.2.3. Случайный лес 9](#_Toc106370171)

[1.2.4. Нейронная сеть 9](#_Toc106370172)

[1.2.5. Метрики оценки качества моделей 11](#_Toc106370173)

[2. Практическая часть 12](#_Toc106370174)

[2.1 Разведочный анализ данных 12](#_Toc106370175)

[2.2. Обнаружение и обработка значений выбросов 16](#_Toc106370176)

[2.3 Нормализация данных 16](#_Toc106370177)

[2.4 Разработка и тест моделей 18](#_Toc106370178)

[2.5 Результат работы моделей 16](#_Toc106370179)

[2.6. Разработка нейронной сети для прогнозирования соотношения матрица-наполнитель 16](#_Toc106370180)

[2.7 Разаработка приложения 23](#_Toc106370179)

[Заключение 25](#_Toc106370183)

[Список использованных источников 26](#_Toc106370184)

Ведение

Тема выпускной квалификационной работы по курсу «Data Science»: **«**Прогнозирование конечных свойств новых материалов (композиционных материалов)».

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).

В ходе решения данной задачи применялись методы, изученные на курсе «Data Science».

# 1. Аналитическая часть

## 1.1. Постановка задачи.

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.

Датасет со свойствами композитов. Объединение делать по индексу тип объединения INNER

Требуется:

1. Изучить теоретические основы и методы решения поставленной задачи.
2. Провести разведочный анализ предложенных данных. Необходимо нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек. Необходимо также для каждой колонке получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков.
3. Провести предобработку данных (удаление шумов, нормализация и т.д.).
4. Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении. При построении модели необходимо 30% данных оставить на тестирование модели, на остальных происходит обучение моделей. При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.
5. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.
6. Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в задании 4 или 5 (один или два прогноза, на выбор учащегося).
7. Оценить точность модели на тренировочном и тестовом датасете.
8. Создать репозиторий в GitHub / GitLab и разместить там код исследования. Оформить файл README.

Исходные датасеты о свойствах композиционных материалов получены структурным подразделением МГТУ им. Н.Э. Баумана – Центр НТИ «Цифровое материаловедение: новые материалы и вещества» и основан на реальных производственных задачах.

Датасет состоит из двух файлов X\_bp.xlsx с данными о параметрах базальтопластика, X\_nup.xlsx с данными о нашивках из углепластика.

На входе имеются данные о начальных свойствах компонентов композиционных материалов, состоящие из 13 факторов (переменных):

1. соотношение матрица-наполнитель;
2. плотность, кг/м3;
3. модуль упругости, Гпа;
4. количество отвердителя, м.%;
5. содержание эпоксидных групп,%\_2;
6. температура вспышки, С\_2;
7. поверхностная плотность, г/м2;
8. модуль упругости при растяжении, ГПа;
9. прочность при растяжении, Мпа;
10. потребление смолы, г/м2;
11. угол нашивки, град;
12. шаг нашивки;
13. плотность нашивки.

Датасет со свойствами композитов, полученный после объединения данных по типу INNER, состоит из 1 023 строк (наблюдений).

## 1.2. Описание используемых методов.

В данной работе требуется произвести регрессионный анализ данных.

Регрессионный анализ ­ статистический аналитический метод, позволяющий вычислить предполагаемые отношения между зависимой переменной одной или несколькими независимыми переменными.

Создание регрессионной модели представляет собой итерационный процесс, направленный на поиск эффективных независимых переменных, чтобы объяснить зависимые переменные, которые мы пытаемся смоделировать или понять, запуская инструмент регрессии, чтобы определить, какие величины являются эффективными предсказателями. Затем пошаговое удаление и/или добавление переменных до тех пор, пока вы не найдете наилучшим образом подходящую регрессионную модель. Т.к. процесс создания модели часто исследовательский, он никогда не должен становиться простым "подгоном" данных. Процесс построения регрессионной модели должен учитывать теоретические аспекты, мнение экспертов в этой области и здравый смысл.

## 1.2.1 Линейная регрессия

Линейная регрессия (Linear regression) - это алгоритм машинного обучения, основанный на контролируемом обучении. Он выполняет задачу регрессии. Регрессионные модели представляют собой целевое значение прогноза, основанное на независимых переменных. В основном используется для выяснения взаимосвязи между переменными и прогнозирования.

Линейная регрессия сводится к нахождению уравнения вида:

*Y = а + bx ;*

Уравнение вида Y = а + bx позволяет по заданным значениям фактора x иметь теоретические значения результативного признака, подставляя в него фактические значения фактора X.

Модель линейной регрессии является часто используемой и наиболее изученной в [эконометрике](https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0). А именно изучены свойства оценок параметров, получаемых различными методами при предположениях о вероятностных характеристиках факторов, и случайных ошибок модели. Предельные (асимптотические) свойства оценок нелинейных моделей также выводятся исходя из аппроксимации последних линейными моделями. С эконометрической точки зрения более важное значение имеет [линейность по параметрам](https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%BE%D1%81%D1%82%D1%8C_%D0%BF%D0%BE_%D0%BF%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%B0%D0%BC), чем линейность по факторам модели.

**Достоинства линейной регрессии:**

* скорость и простота получения модели;
* интерпретируемость модели;
* линейная модель является прозрачной и понятной для аналитика;
* по полученным коэффициентам регрессии можно судить о том, как тот или иной фактор влияет на результат, сделать на этой основе дополнительные полезные выводы;
* широкая применимость.

**Недостатки:**

* не может применяться к задачам нелинейной классификации;
* правильный выбор функций не требуется;
* ожидается хорошее соотношение сигнал / шум;
* колинеарность и выбросы влияют на точность модели LR.

1.2.2. Метод ближайших соседей

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 вычислительную трудоёмкость

## 1.2.3. Случайный лес

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

Формула итогового решателя — это усреднение предсказаний отдельных деревьев.

Где N – количество деревьев; i – счетчик для деревьев; b – решающее дерево; x – сгенерированная нами на основе данных выборка.

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

## 1.2.4. Нейронная сеть

Нейро́нная сеть (также искусственная нейронная сеть, ИНС) — [математическая модель](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8C), а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей — сетей [нервных клеток](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD) живого организма. Это понятие возникло при изучении процессов, протекающих в [мозге](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D0%B7%D0%B3), и при попытке [смоделировать](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D0%B4%D0%B5%D0%BB%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5) эти процессы. После разработки алгоритмов обучения получаемые модели стали использовать в практических целях: в [задачах прогнозирования](https://ru.wikipedia.org/wiki/%D0%97%D0%B0%D0%B4%D0%B0%D1%87%D0%B8_%D0%BF%D1%80%D0%BE%D0%B3%D0%BD%D0%BE%D0%B7%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F), для [распознавания образов](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D0%BE%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2), в задачах [управления](https://ru.wikipedia.org/wiki/%D0%90%D0%B4%D0%B0%D0%BF%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B5_%D1%83%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5) и др.

ИНС представляет собой [систему](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0) соединённых и взаимодействующих между собой простых процессоров ([искусственных нейронов](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%BD%D0%B5%D0%B9%D1%80%D0%BE%D0%BD)). Такие процессоры обычно довольно просты (особенно в сравнении с процессорами, используемыми в персональных компьютерах). Каждый процессор подобной сети имеет дело только с [сигналами](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%B3%D0%BD%D0%B0%D0%BB), которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам. И, тем не менее, будучи соединёнными в достаточно большую сеть с управляемым взаимодействием, такие по отдельности простые процессоры вместе способны выполнять довольно сложные задачи.

Типовая ИНС состоит из трех компонентов:

* **входной слой**, нейроны которого принимают исходный вектор значений, кодирующий входной сигнал и передают его на последующий слой, усилив или ослабив;
* **скрытые (промежуточные) слои**, которые выполняют основные вычислительные операции;
* **выходной слой**, нейроны которого представляют собой выходы сети и иногда также могут производить какие-либо вычисления.



Рисунок 2. Схема нейронной сети

**Достоинства** **нейронных** **сетей**:

* Возможность моделирования и прогнозирования нелинейных процесов;
* Способность работать с зашумленными данными;
* Быстрое обучение и гибкость адаптации к изменениями внешней среды

**Недостатки**

* Сравнительная дороговизна аппаратной реализации;
* Трудность тиражирования накопленных знаний;
* Для больших сетей невозможность заранее даже приблизительно оценить время обучения сети.

## 1.2.5. Метрики оценки качества моделей

Качеством модели регрессии называется адекватность построенной модели исходным (наблюдаемым) данным. Для оценки качества модели регрессии наиболее типичными мерами качества в задачах регрессии являются:

R2 или коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то качество прогноза идентично средней величине целевой переменной (т.е. очень низкое). Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

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.

2. Практическая часть

2.1. Разведочный анализ данных

С помощью метода df.isnull().sum().sum() пропусков в данных не выявлено. Для создания описательной статистики для кадра данных использовался метод df.**describe** () рис 1

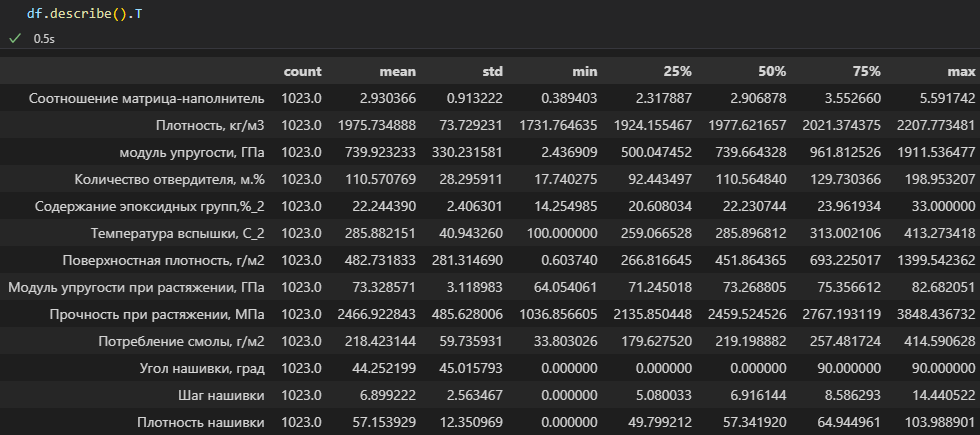
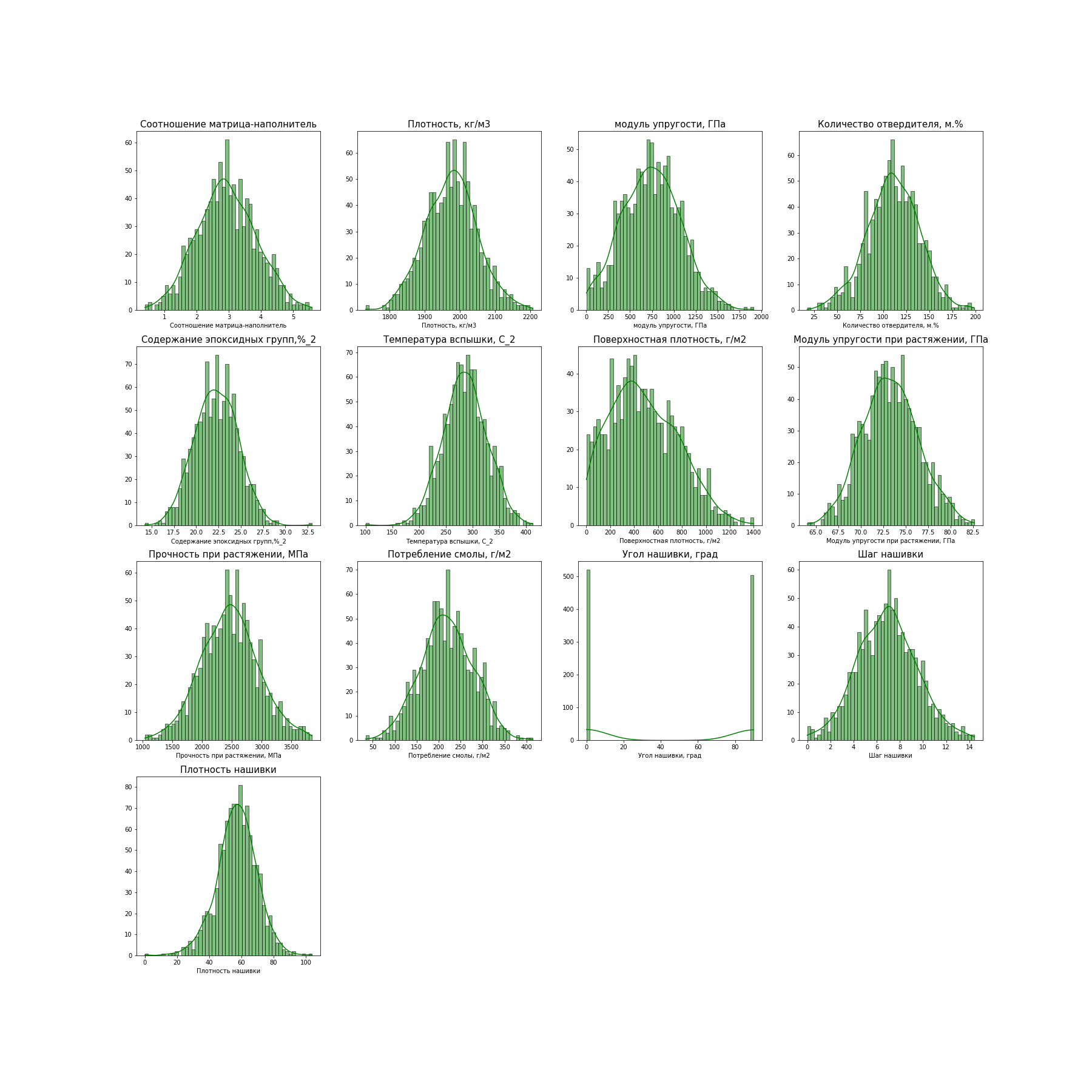


Рисунок 1 – Описательная статистика датасета

Рисунок 2. - Гистограммы распределения для каждой из переменной.

Из полученных гистограмм рис-3 видно, что наши признаки имеют тип распределения близкий к нормальному, кроме признаков «Угол нашивки» и «Поверхностная плотность». «Угол нашивки» имеет всего два значения 0 и 90, «поверхностная плотность» имеет смещение в лево.

Для визуализации коэффициентов корреляции и определения того, между какими переменными установлена более тесная взаимосвязь, была построена тепловая карта коэффициентов корреляции и попарные графики рассеяния точек (рис. 3-4):

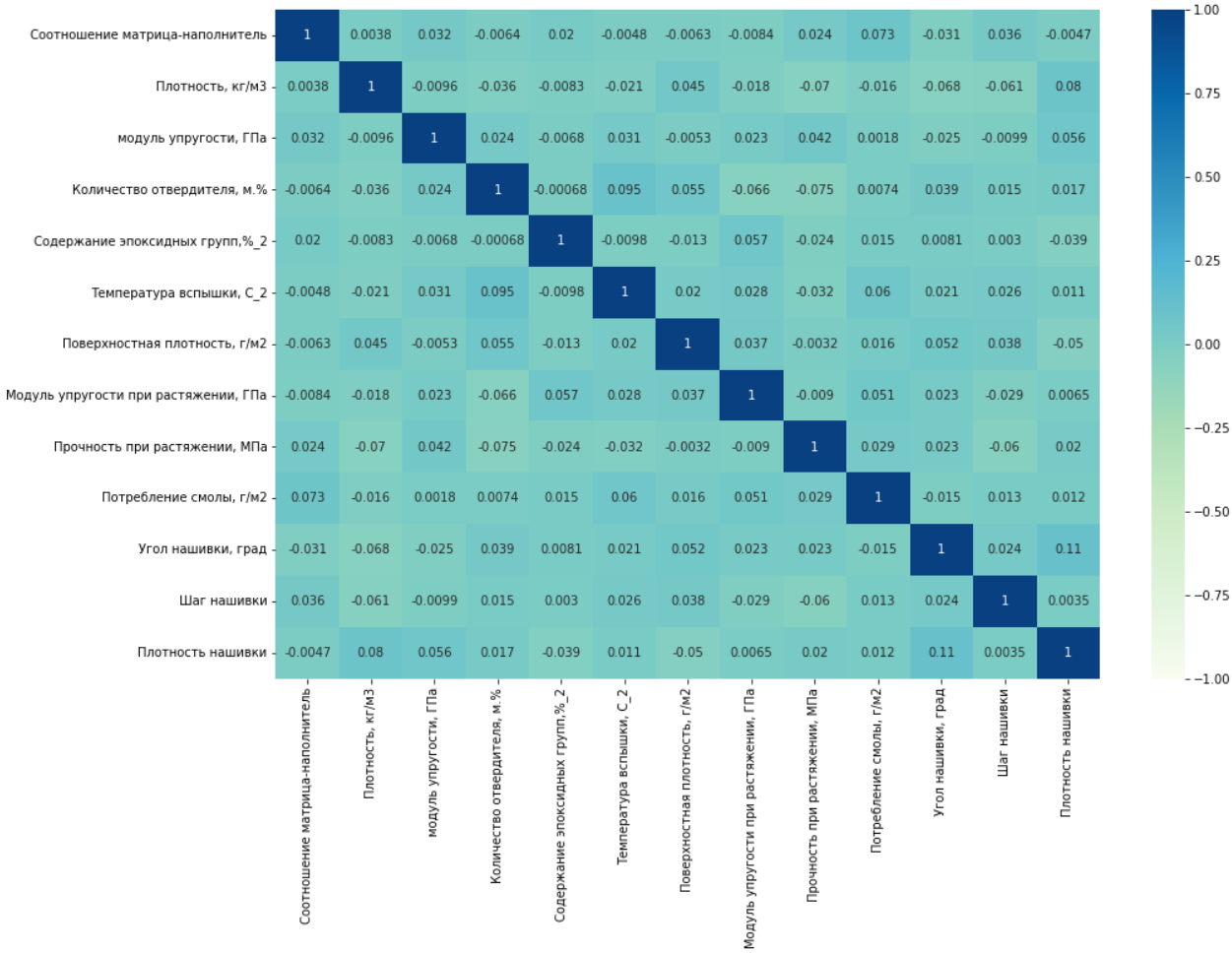


Рисунок 3. Тепловая карта

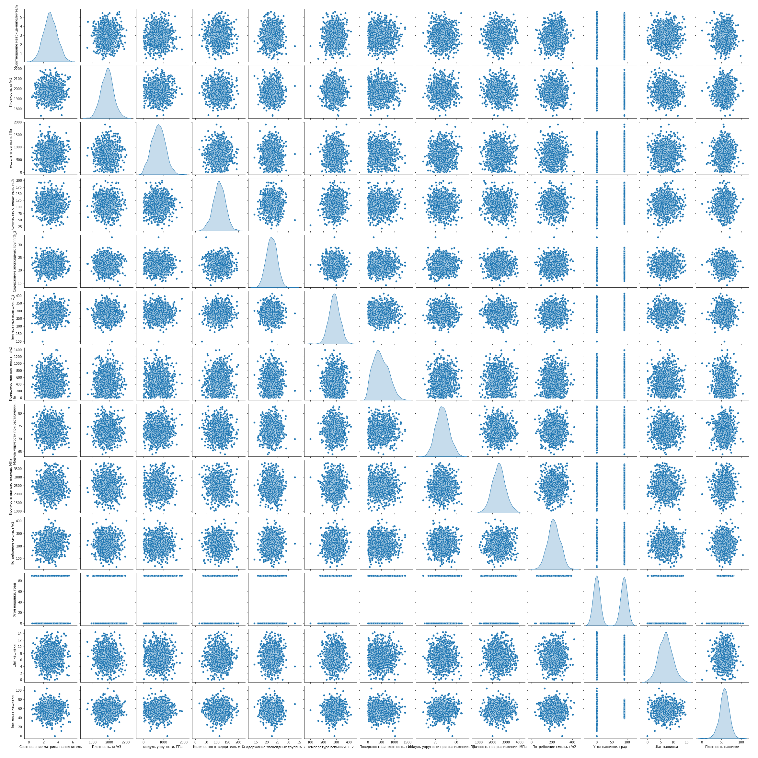


Рисунок 4. попарные графики рассеяния точек

Тепловая карта и попарные диаграммы рассеяния точек показывают, что корреляция практически отсутствует, возможно, наши признаки связаны какой-то другой, нелинейной зависимостью.

Диаграмма "ящик с усами" показывает распределение данных по квартилям, выделяя их выбросы.

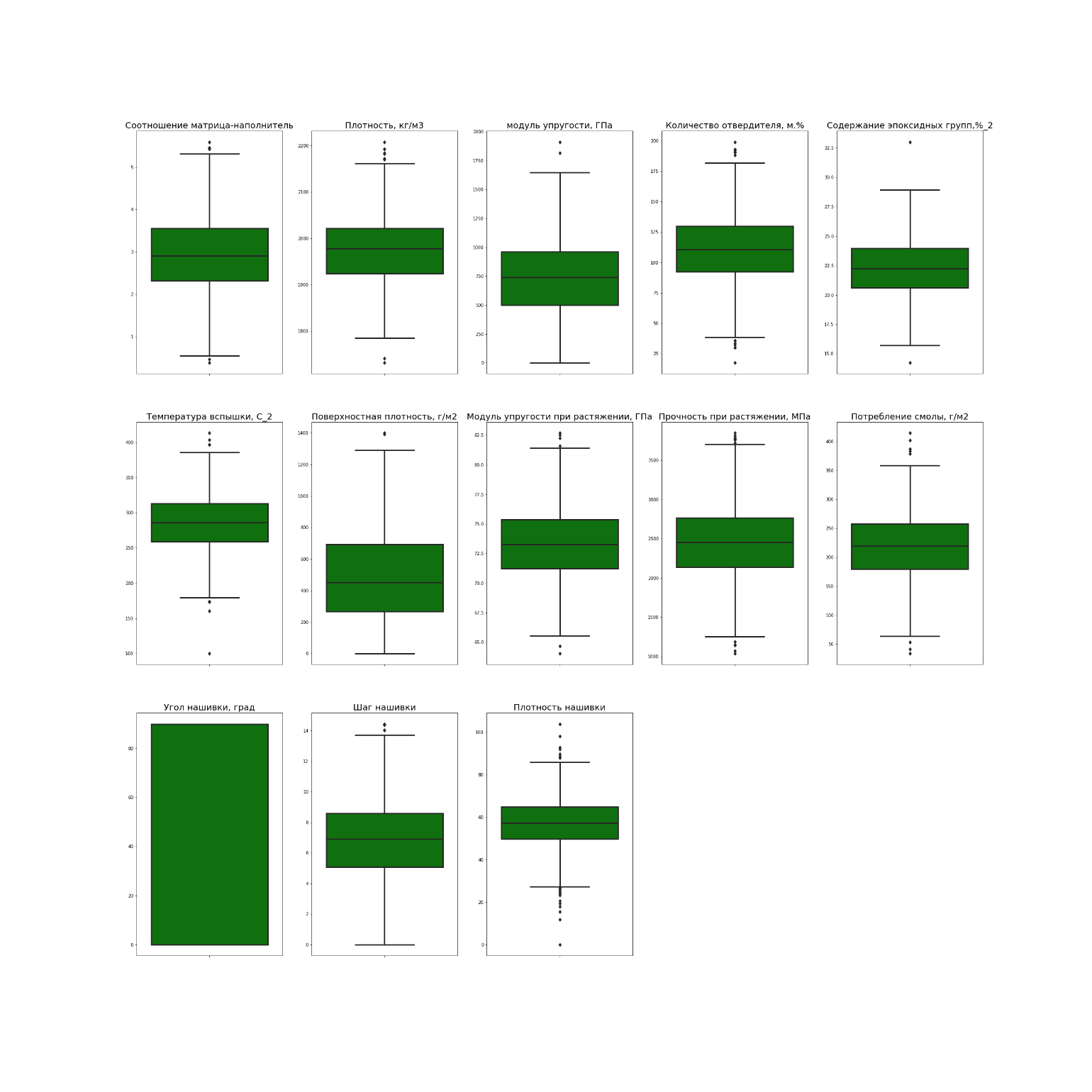


Рисунок 5. Ящик с усами

2.2. Очистка от выбросов

Как показывают графики рис 5 наша выборка имеет некоторые выбросы.

Сравним количество выбросов методом 3-х сигм и межквартильных расстояний:

Метод 3-х сигм, выбросов: - 24

Метод межквартильных расстояний, выбросов: - 93

Т.к датасет уже подготовлен и очищен от явного шума, применю метод 3-х сигм, чтобы потерять меньше значимой информации.

2.3. Нормализация данных

Нормализация — это метод, который часто применяется как часть подготовки данных для машинного обучения. Цель нормализации — изменить значения числовых столбцов в наборе данных для использования общей шкалы без искажения различий в диапазонах значений или потери информации. Нормализация также требуется для некоторых алгоритмов для правильного моделирования данных. На рис 6 видно, что данные находятся в разных диапазонах.

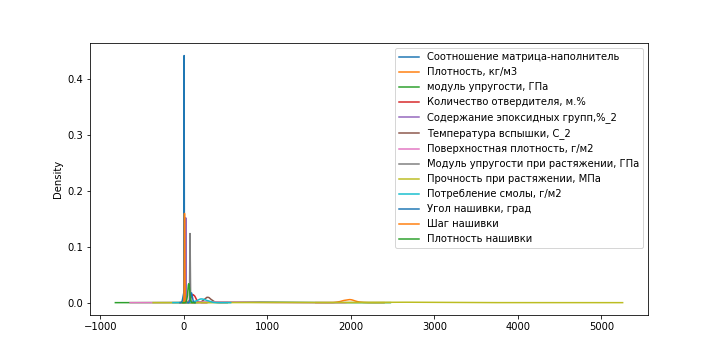


рис 6.

Для нормализации датасета я применил MinMaxScaler() MinMaxScaler – библиотека sklearn.preprocessing, позволяющая произвести нормализацию данных перед использованием в модели машинного обучения , то есть приведение числовых переменных к диапазону от 0 до 1 без изменения формы распределения рис 7,8,9 Таким образом, после предобработки данных получаем нормализованный дата сет размером 999 строк

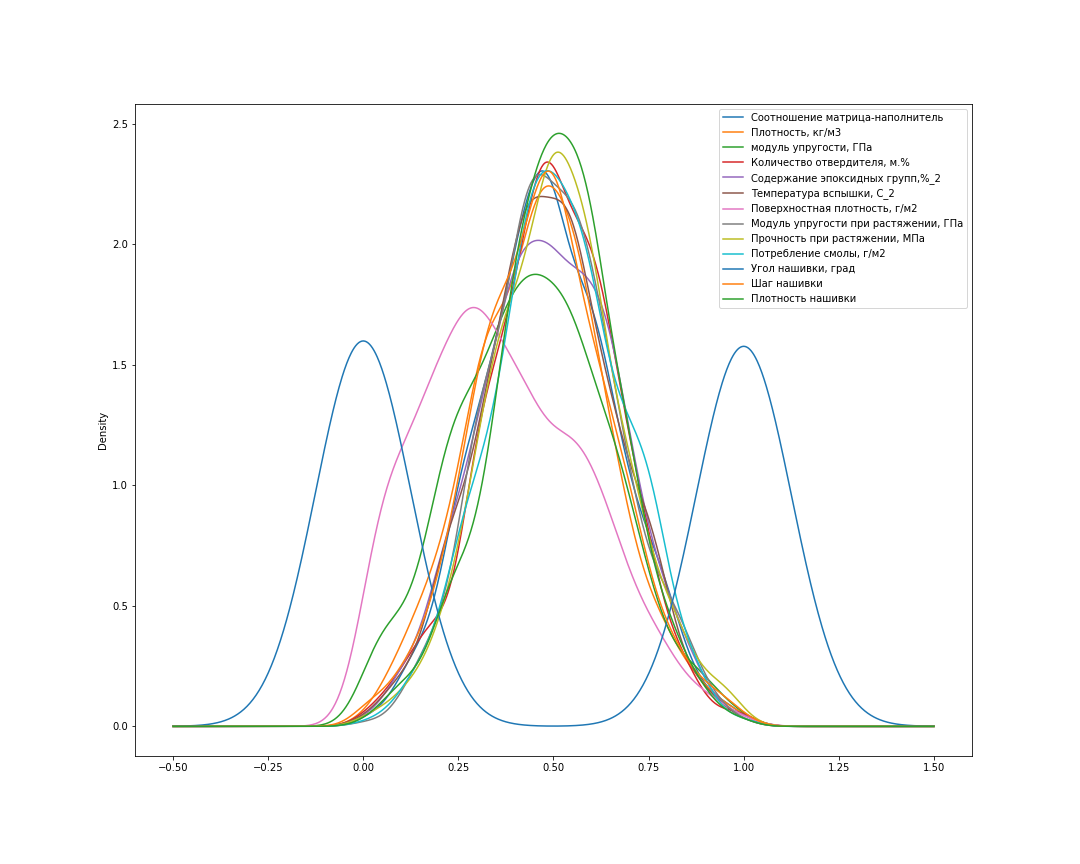


Рис 7.

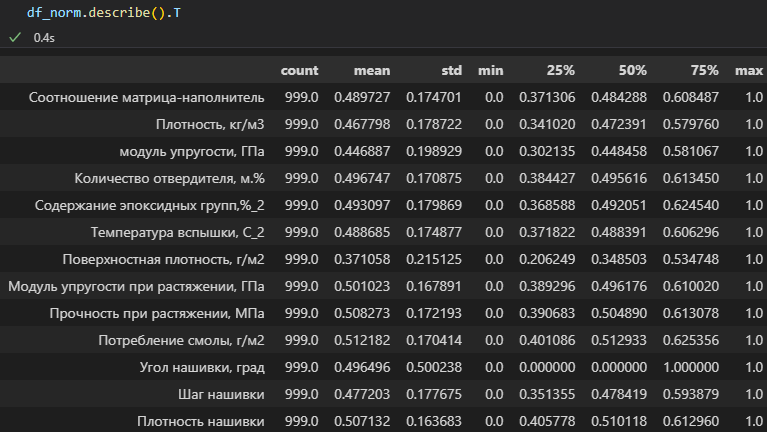


Рисунок 8. Описательная статистика нормализованного датасета

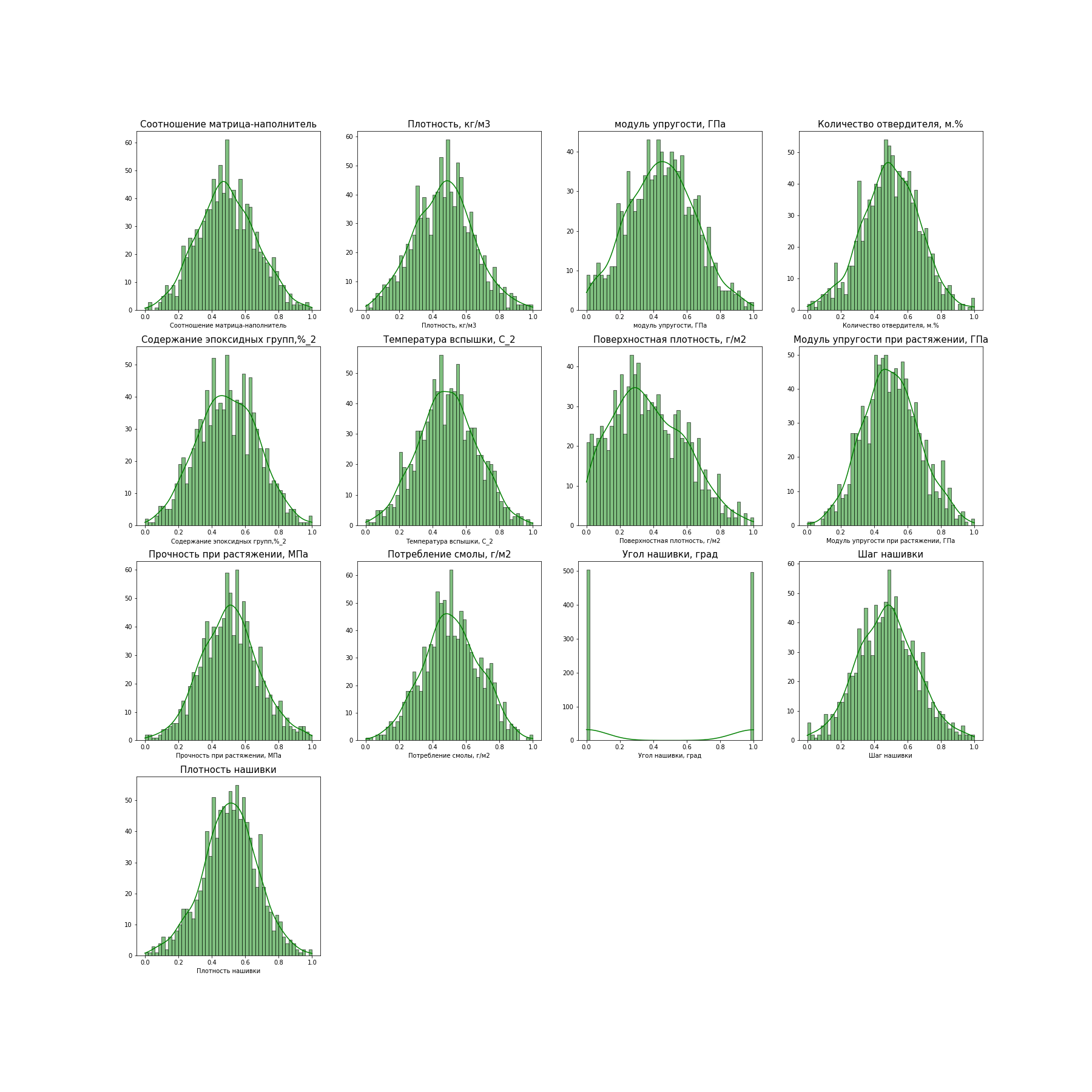


Рисунок 9 Гистограмма распределения после нормализации

Данные предобработаны и нормализованы теперь приступим к регрессионному анализу признаков.

2.4. Разработка и тестирование моделей

В соответствии с поставленной задачей, нужно осуществить разработку и обучение моделей машинного обучения для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении». Для каждого признака построение моделей осуществляется раздельно. Разделение нормализованных данных на обучающую и тестовую выборки (в соотношении 70 на 30%, согласно поставленной задаче).

Для признака «Модуль упругости при растяжении» были разработаны и обучены следующие модели:

− модель случайный лес (метод RandomForestRegressor);

− модель на основе линейной регрессии (метод LinearRegression);

− модель k ближайших соседей (метод KNeighborsRegressor);

## 2.5. Результаты работы моделей

Ниже представлены результаты работы моделей для «Модуль упругости при растяжении» рис 10. а также «Прочность при растяжении» рис. 11

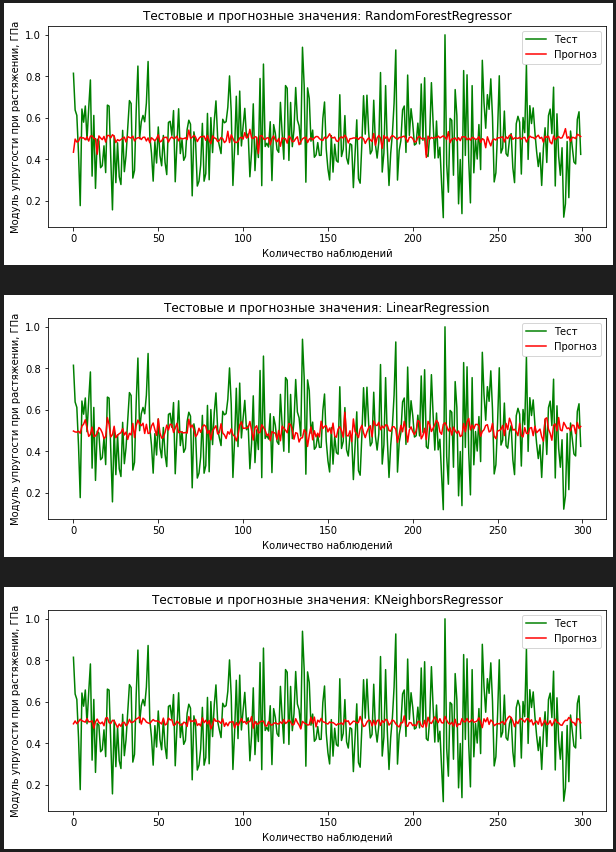


Рис. 10 Модуль упругости при растяжении

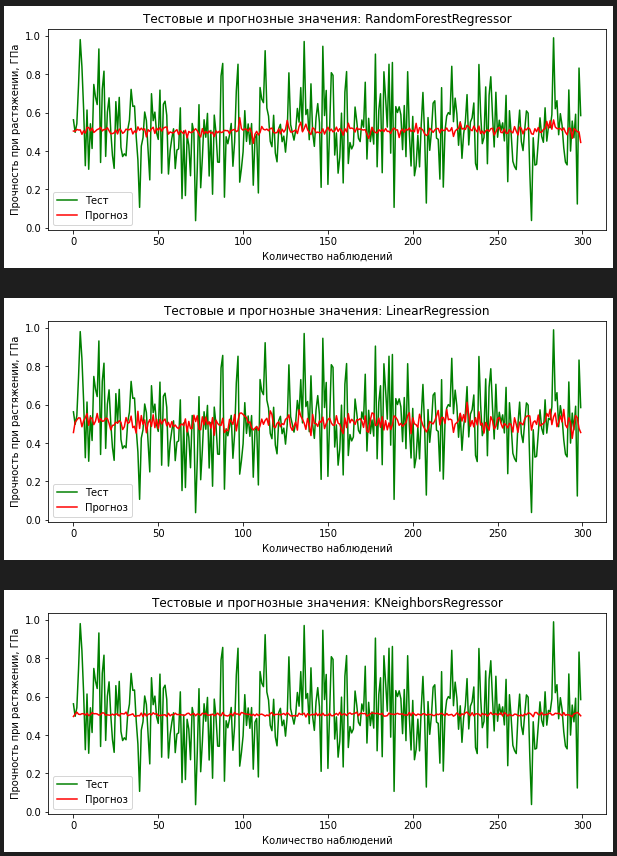


Рис. 11 Прочность при растяжении

Таблица 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Model** | **MSE** | **R2 score** |
| Модуль упругости при растяжении | RandomForestRegressor | 0.026863 | -0.026035 |
| Модуль упругости при растяжении | LinearRegression | 0.026806 | -0.023851 |
| Модуль упругости при растяжении | KNeighborsRegressor | 0.026136 | 0.001729 |
| Прочность при растяжении | RandomForestRegressor | 0.030332 | 0.011061 |
| Прочность при растяжении | LinearRegression | 0.030845 | -0.005656 |
| Прочность при растяжении | KNeighborsRegressor | 0.030678 | -0.000216 |

Выводы: все модели показали неудовлетворительный результат, ни одна из моделей не справилась с задачей.

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

По заданию для соотношения матрица-наполнитель необходимо построить нейросеть модуля Keras библиотеки TensorFlow. Строю нейронную сеть с помощью класса keras.Sequential со следующими параметрами:

* входной слой нормализации 12 признаков;
* выходной слой для 1 признака;
* скрытых слоев: 4;
* нейронов в скрытом слое: 128, 128, 64, 32;
* активационная функция скрытых слоев: relu;
* оптимизатор: Adamax;
* loss-функция: MSE.

Архитектура нейросети приведена на рисунке 12.

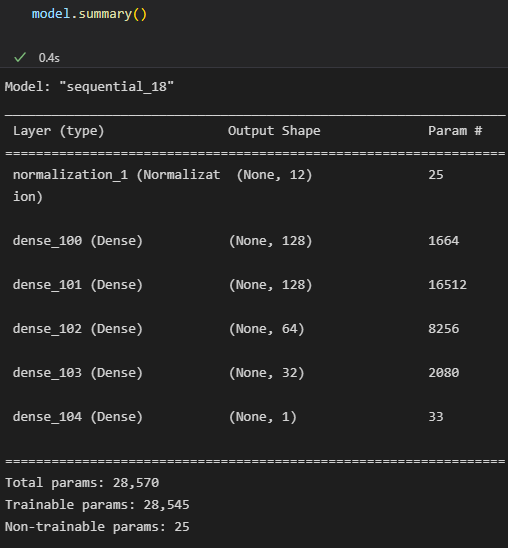
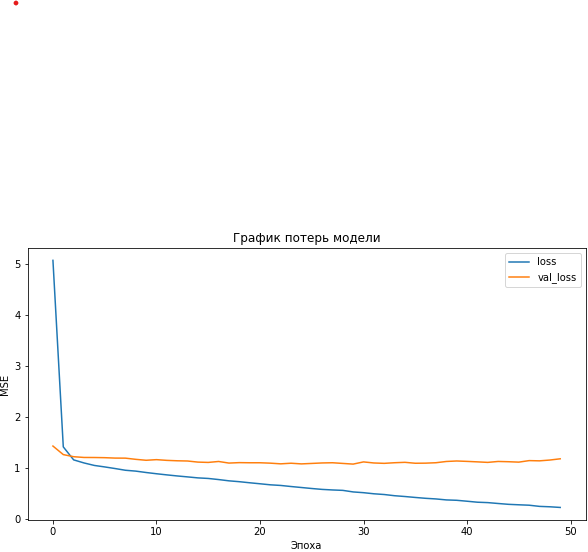


Рисунок 12 архитектура нейронной сети

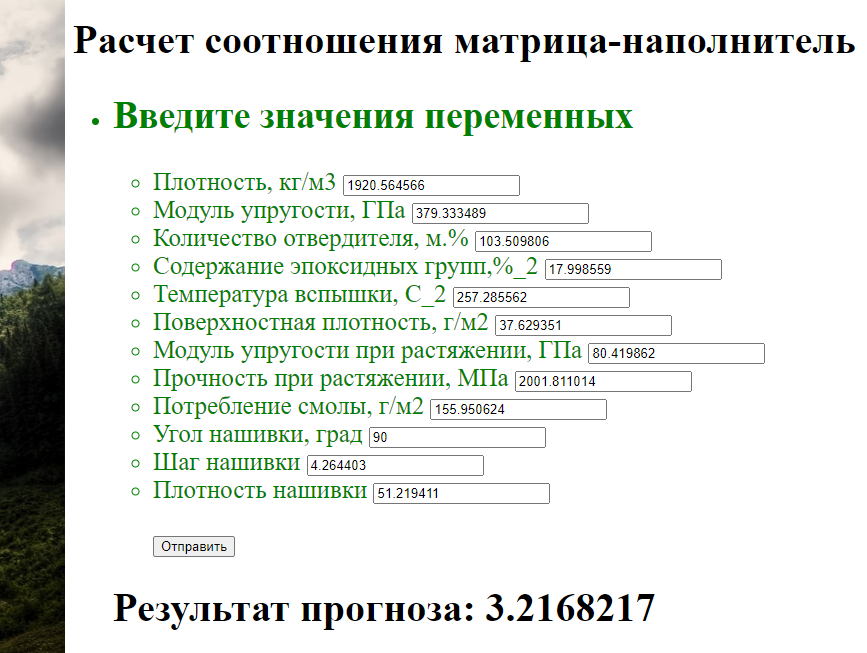




По графику делаем вывод что для обучения спроектированной нейронной сети достаточно 28 эпох. Ошибки модели, следующие: MSE= 1.203631, R^2 = -0.492071. Результаты неудовлетворительны.

## 2.5. Разработка приложения

Несмотря на то, что пригодных к внедрению моделей получить не удалось, можно разработать тестовое веб-приложение на фреймворке Flask. Приложение разработано в среде разработки Visual Studio Code.



Создан репозиторий в GitHub, где размещён код исследования. Оформлен файл README.

Страница слушателя: https://github.com/DezheninEvgeniy/VKR

**Заключение**

Полученные в ходе решения задачи модели не приносят положительного результата, так как имеют высокий уровень ошибки в предсказаниях.

Список использованных источников

1Билл Любанович. Простой Python. Современный стиль программирования. — СПб.: Питер, 2016. — 480 с.: ил. — (Серия «Бестселлеры O’Reilly»).

2 Бринк Х. Машинное обучение. / Х. Бринк, Дж. Ричардс, М. Феверолф. – СПб.: Питер,2017. 336 с.

3 Вандер Плас Дж. Python для сложных задач: наука о данных и машинное обучение. - СПб.: Питер, 2018. - 576 с.

4 Гиздатуллин А.Р., Хозин В.Г., Куклин А.Н., Хуснутдинов А.М. «Особенности испытаний и характер разрушения полимеркомпозитной арматуры».

5 Горбунов П.М., Мацкевич Ю.А., Чубарь А.В. Машинное обучение. Автоматизация подбора модели машинного обучения // Робототехника и искусственный интеллект. 2021. С. 155-160.

6 Миркес Е. М., Нейроинформатика. Учебное пособие с программами для выполнения лабораторных работ. 2003. ISBN 5-7636-0477-6

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

Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.

8 Справочник по композиционным материалам: в 2 - х кн. Кн. 2 / Под ред. Дж. Любина; Пер. с англ. Ф. Б. Геллера, M. М. Гельмонта; Под ред. Б. Э. Геллера - М.: Машиностроение, 1988. - 488 с. : ил;

9 Траск Эндрю. Грокаем глубокое обучение. – СПб.: Питер, 2019. – 352 с