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 Pro»

Слушатель Тимбаева Инна Анатольевна\_

(ФИО)

Москва, 2024

Пояснительная записка.

1. Аналитическая часть

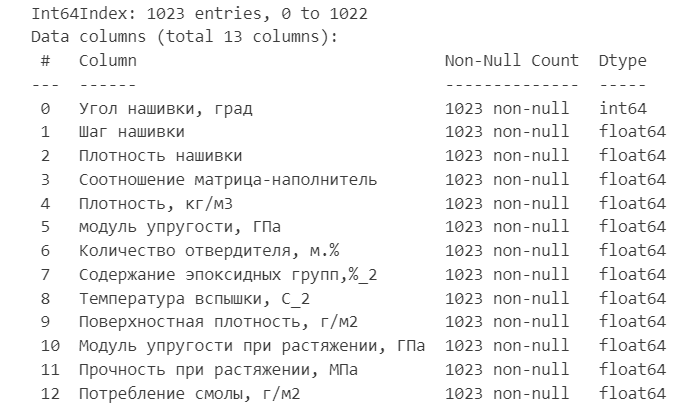
*1.1. Постановка задачи.*

Разработка модели, которая будет предсказывать значения модуля упругости и прочности при растяжении композитов на основе их состава и процесса производства.

Композиты имеют такие структурные элементы, как матрица и нашивка. Матрица - материал, внутри которого располагается набивка (нашивка). Нашивка - это те компоненты, которые призваны усовершенствовать свойства материала (матрицы). Прогнозные модели помогают сократить количество проводимых испытаний, так как процесс производства композитов может быть очень сложным, учитывая что конечные свойства материала зависят от многих факторов, таких как температура, состав материала и многие другие.

Для построения модели использованы, предоставленные в формате excel, экспериментальные наблюдения по 13 показателям: 'Угол нашивки, град', 'Шаг нашивки', 'Плотность нашивки', 'Соотношение матрица-наполнитель', 'Плотность, кг/м3', 'модуль упругости, ГПа', 'Количество отвердителя, м.%', 'Содержание эпоксидных групп,%\_2', 'Температура вспышки, С\_2', 'Поверхностная плотность, г/м2', 'Модуль упругости при растяжении, ГПа', 'Прочность при растяжении, МПа', 'Потребление смолы, г/м2'.Общее число наблюдений - 1023. Практическая часть анализа, обработки данных и построения моделей выполнена при помощи языка Python.

Данные имеют следующую структуру:





*1.2. Описание используемых методов.*

Для предсказания значения модуля упругости и прочности при растяжении использовались следующие регрессионные модели:

1. **Линейная регрессия** — один из самых простых и широко используемых методов машинного обучения.

Регрессионный анализ осуществляется путем оценки коэффициента линейного уравнения. При этом может быть одна или несколько независимых переменных, которые коррелируют и лучше всех подходят для прогнозирования значения зависимой переменной. Можно считать, что регрессионный анализ выполняется путем подгонки прямой линии к данным, которые стремятся к уменьшению расхождения между фактическими и прогнозируемыми значениями зависимой переменной. С работой любой линейной регрессионной модели принято связывать четыре основных допущения.

* **Линейность.** Между средним значением зависимой переменной и независимыми переменными должна существовать линейная связь. Эта связь измеряется путем выявления изменений зависимой переменной в связи с изменениями независимых переменных.
* **Гомоскедастичность.** В линейной регрессии гомоскедастичность имеет важное значение, поскольку представляет собой степень подгонки модели под данные. Она определяет дисперсию по величине погрешности или остатков: если дисперсия увеличивается, значит модель подогнана плохо.
* **Независимость.** Собранные точки данных должны быть независимы друг от друга.
* **Нормальность.** Должно быть нормальное распределение для любого из фиксированных значений зависимой и независимой переменных.

2. **Деревья решений** - принцип работы заключается в разделении данных на подмножества на основе правил принятия решений, выведенных из входных признаков. Каждое разбиение делается так, чтобы минимизировать ошибку в предсказании целевой переменной.

Каждый узел в дереве представляет собой вопрос или условие, а ветви дерева — возможные ответы или исходы этого вопроса:

* Корень дерева — это начальный узел, с которого начинается процесс принятия решения.
* Внутренние узлы представляют собой условия, которые необходимо проверить.
* Листовые узлы — это конечные узлы, которые представляют собой возможные результаты (предсказания).

Деревья решений используют как для задач классификации, так и для задач регрессии. Они просты для интерпретации, но могут быть подвержены [переобучению](https://ru.wikipedia.org/wiki/%D0%9F%D0%B5%D1%80%D0%B5%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5), особенно если дерево становится слишком глубоким и сложным.

3. **Случайный лес** — это алгоритм машинного обучения, основанный на объединении множества деревьев решений для улучшения точности и устойчивости модели.

Идея состоит в том, чтобы создавать несколько деревьев решений на основе разных подвыборок данных (этот процесс называется [бутстрэппингом](https://ru.wikipedia.org/wiki/%D0%91%D1%83%D1%82%D1%81%D1%82%D1%80%D1%8D%D0%BF_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0))) и затем комбинировать их результаты. Чтобы снизить вероятность переобучения и повысить устойчивость модели, метод случайного леса использует [метаалгоритм бэггинга](https://ru.wikipedia.org/wiki/%D0%91%D1%8D%D0%B3%D0%B3%D0%B8%D0%BD%D0%B3):

* В случае классификации, модель выбирает наиболее часто встречающееся предсказание среди всех деревьев (мода).
* В случае регрессии, модель выбирает среднее значение предсказаний всех деревьев.

Случайный лес показывает более высокую точность, чем одно дерево решений, и хорошо работает с большими наборами данных.

Оценка любой модели машинного обучения  —  важнейшая задача, сопутствующая моделированию данных. Оценка моделей производилась при помощи следующих оценок:

**1. MAE (mean absolute error, средняя абсолютная ошибка)**  —  это универсальная метрика, которая позволяет узнать разницу между фактическими и прогнозируемыми значениями.

**2. MSE (mean squared error, средняя квадратичная ошибка)** можно рассматривать как уточненную MAE, поскольку она помогает находить ошибки с помощью квадратичной разницы между фактическими и прогнозируемыми значениями.

**3.R² (R-squared, R-квадрат)** также считается универсальной метрикой, применяемой для оценки эффективности регрессионной модели. Показатель показывает долю дисперсии зависимой переменной, объясняемую рассматриваемой **моделью** зависимости, то есть объясняющими переменными.

*1.3. Разведочный анализ данных.*

Разведочный анализ данных включает нахождение пропусков и дубликатов, определение уникальных значений, анализ распределения данных, анализ наличие выбросов и их удаление, наличие зависимостей (корреляции между признаками).

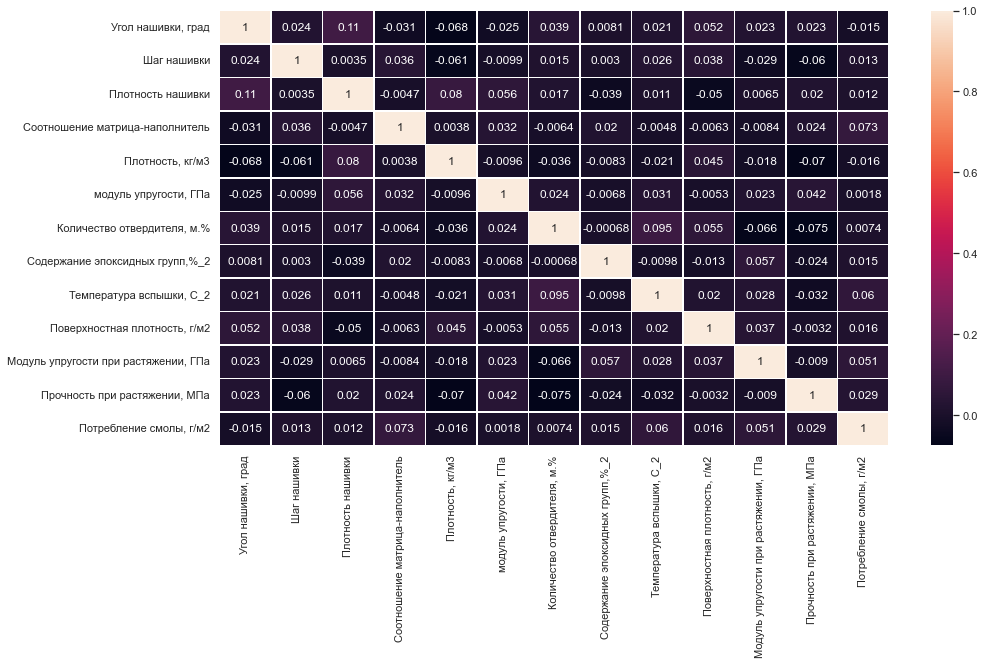
Пропуски и дубликаты в данных отсутствуют.

Количество уникальных значений в каждом столбце:



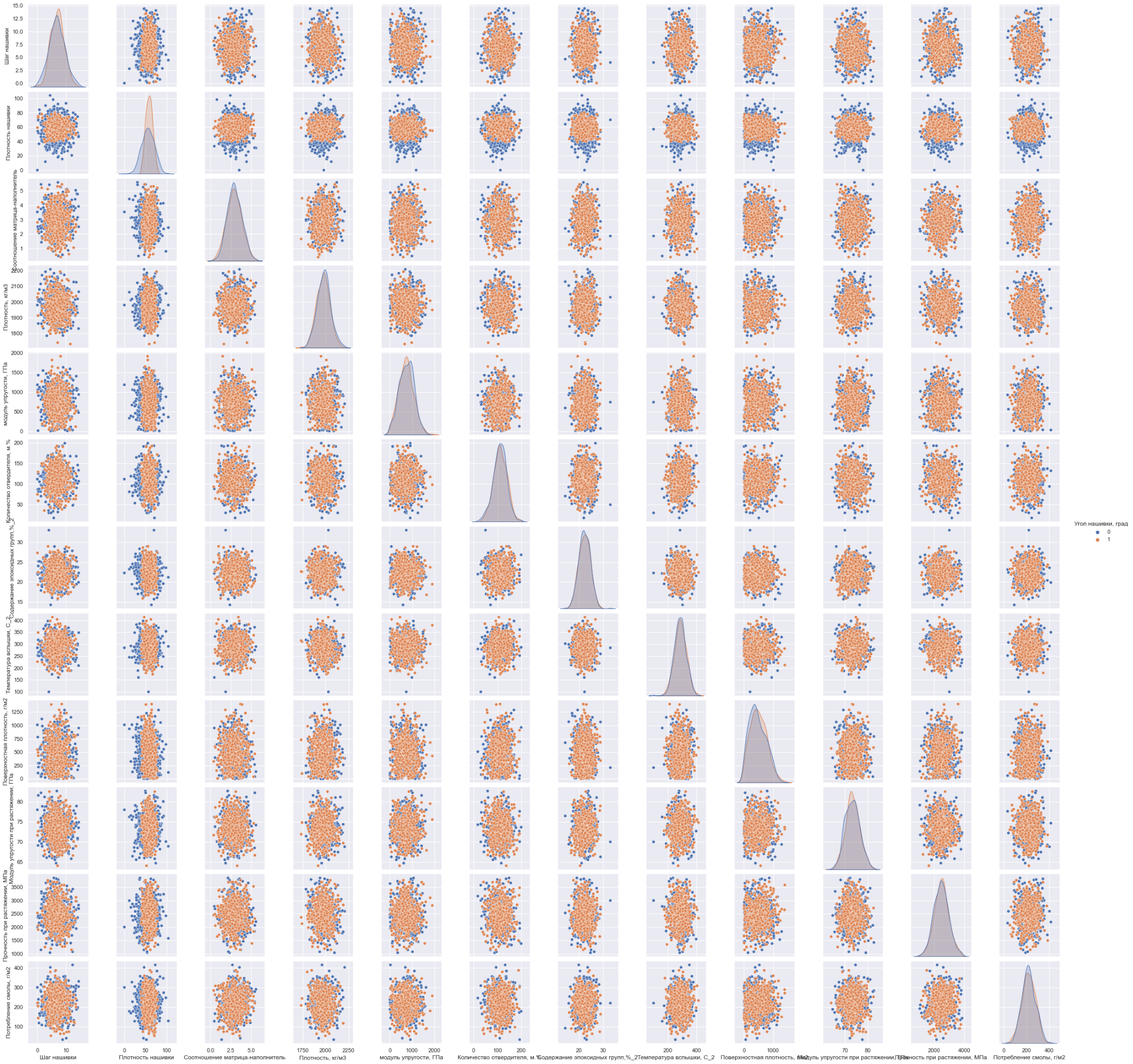
По отдельным показателям количество уникальных значений от 989-1020. По показателю Угол нашивки - 2 уникальных значения (0 и 90 градусов). Данный показатель был отнесен к категориальными со значениями 0 и 1.

Зависимость (корреляция) между данными слабая, все данные были включены в модели:

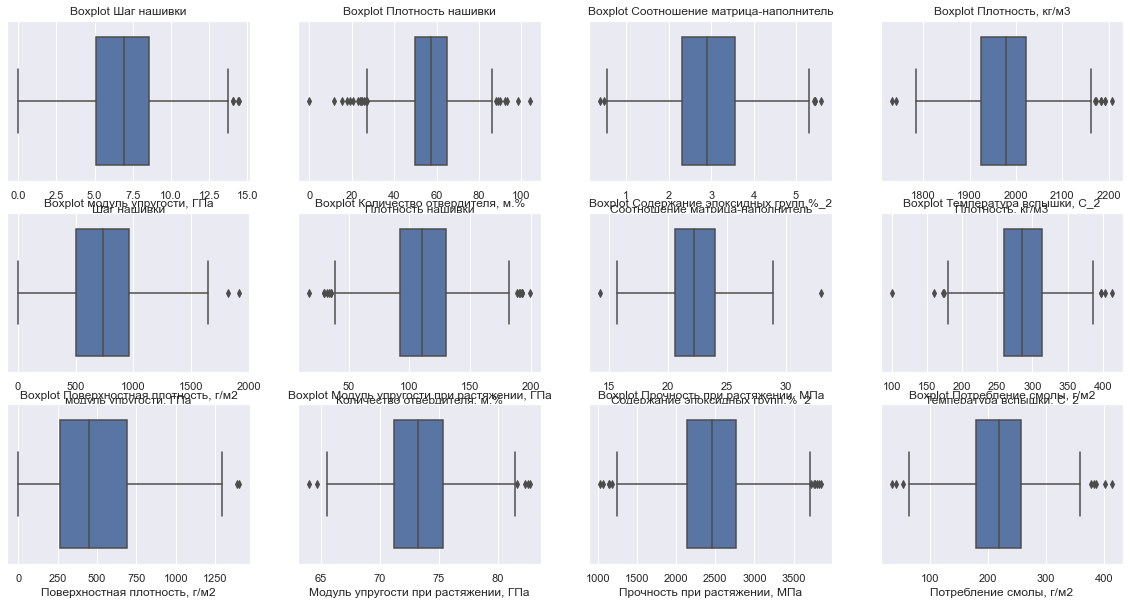


На основании визуального анализа графиков распределения и pairplot, а также boxplot, выявлены признаки с выбросами (точки находящиеся на значительном удалении от основной массы значений), которые в дальнейшем были оценены с помощью z-scores и Изолирующего леса.

***Pairplot***



***Boxplot***

******

2. Практическая часть

*2.1. Предобработка данных.*

**(1) Визуальный анализ:** Признаки с большим количеством относительно компактно расположенных точек за пределами усов boxplot (Q1 - 1.5 \* IQR и Q3 + 1.5 \* IQR), не рассматривались как выбросы - принимались как возможный вариант нормы.

В качестве выбросов рассматривались отдельные точки, находящиеся существенно дальше усов/основной массы точек.

Для дальнейшего анализа были выбраны следующие 8 признаков: 'Плотность нашивки', 'Плотность, кг/м3', 'модуль упругости, ГПа','Температура вспышки, С\_2', 'Поверхностная плотность, г/м2', 'Содержание эпоксидных групп,%\_2', 'Количество отвердителя, м.%', 'Потребление смолы, г/м2'.

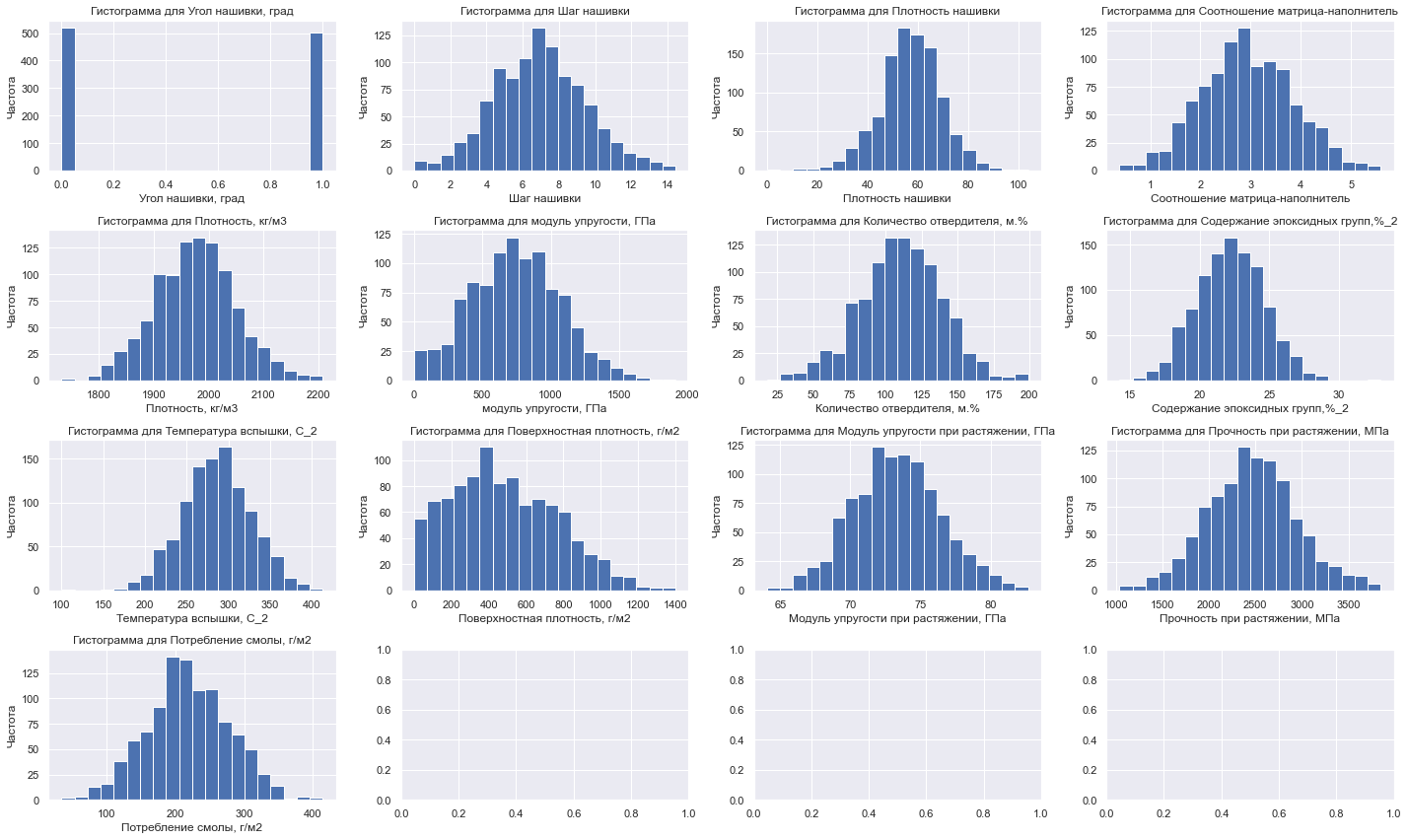
**(2) z-scores:** Далее выбранные признаки оценивались при помощи z-score (данные в Dataframe были стандартизированы функцией StandardScaler и применена функция z-scores, показывающая отклонение значений от среднего) и значения за пределами 3 сигм от среднего были отмечены как выбросы).

**(3) Изолирующий лес:** Алгоритм работает путём случайного разбиения и изоляции точек данных в наборе данных. Для этого он создаёт группы узлов и изолирует их на основе их атрибутов. Данные предварительно были нормализованы при помощи функции MinMaxScaler.

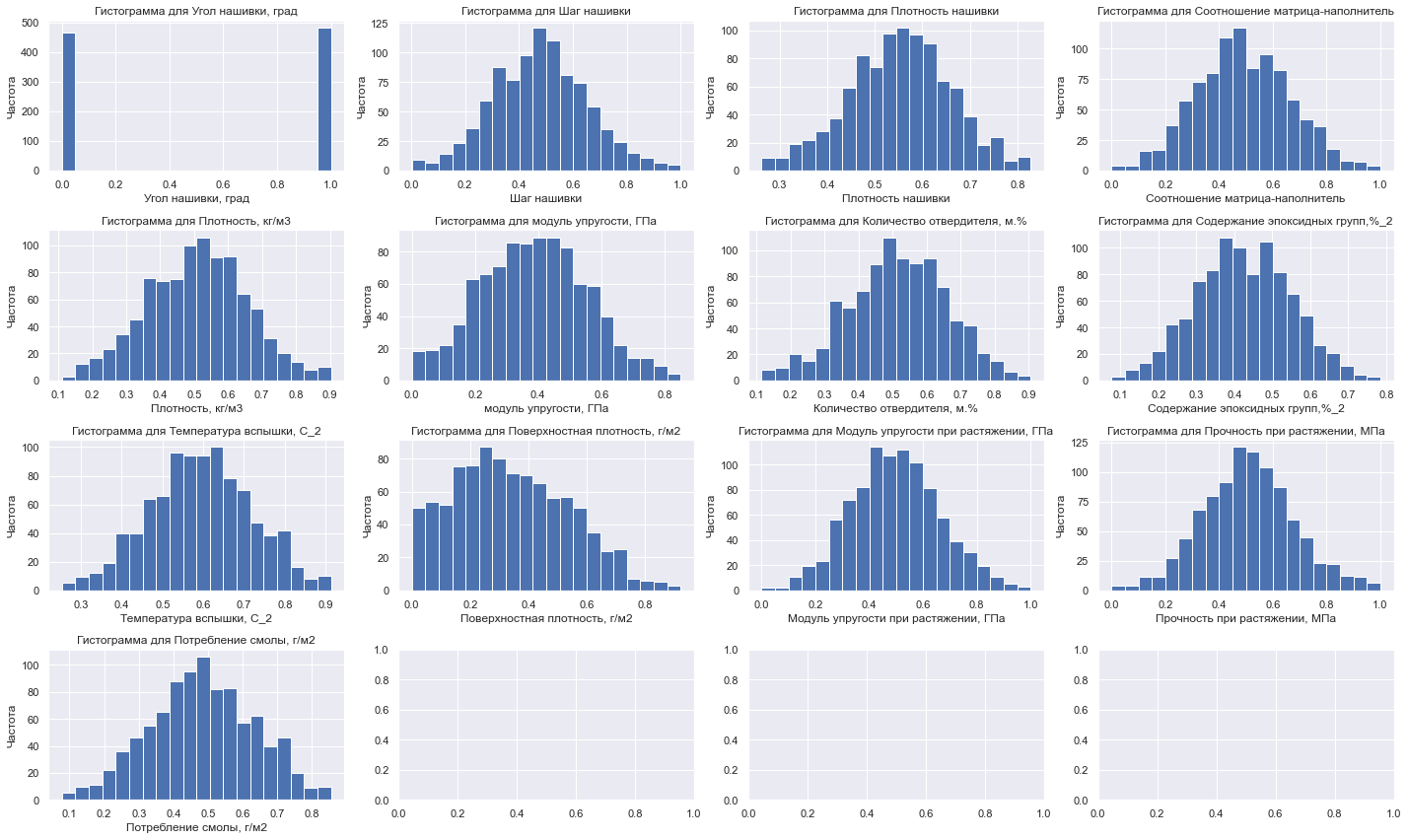
По итогам этапов (2) и (3) были удалены все выявленные выбросы из Набора данных.

Общий объем выбросов в Наборе данных несущественный (по boxplot - не более 2,05% по каждому признаку, по z-score - не превышает 1% по каждому признаку, по Изолирующему лесу - 2,05%.

***Графики распределения до обработки***

**

***Графики распределения после нормализации и удаления выбросов***



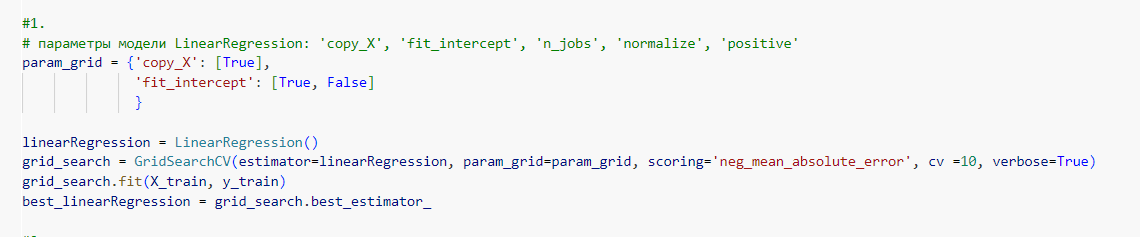
Набор данных после предобработки:

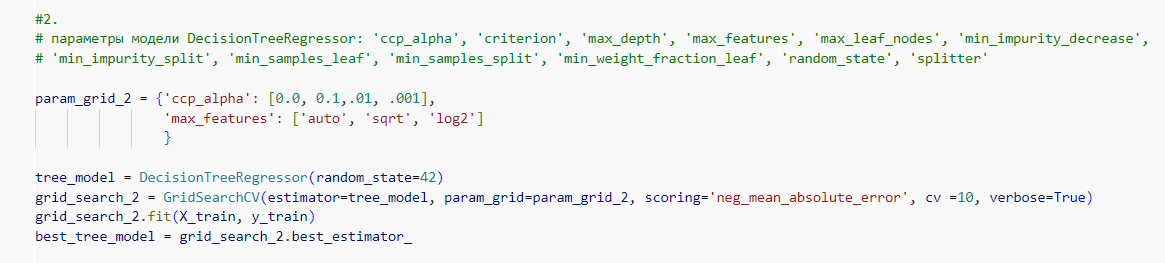
**

Предобработка данных важный этап в подготовке к дальнейшему использованию данных в модели, который определяет точность и качество будущей модели.

*2.2. Разработка и обучение модели.*

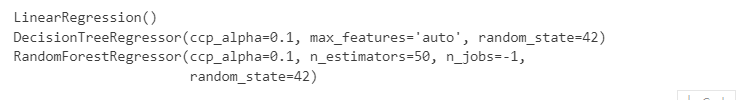
В рамках практической части работы подбор параметров выше рассмотренных моделей осуществлялся при помощи функции GridSearch:





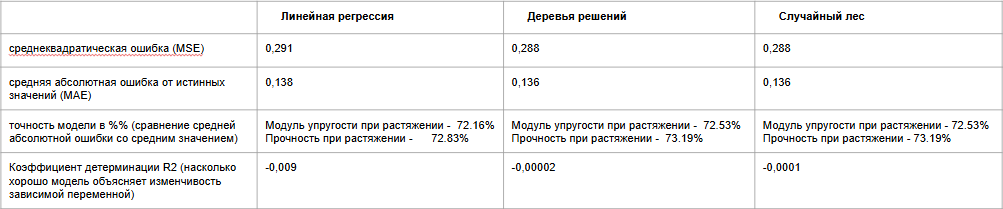


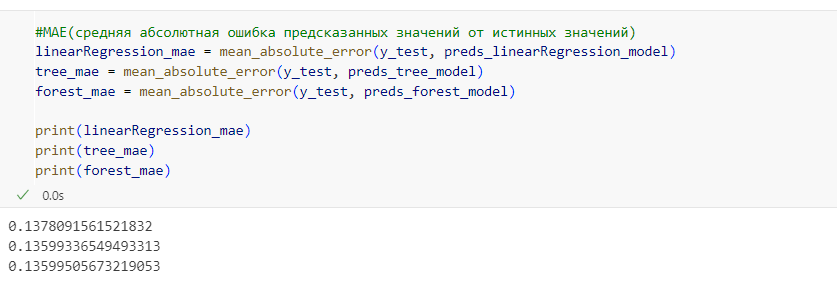
Наилучшие рассчитанные параметры моделей:

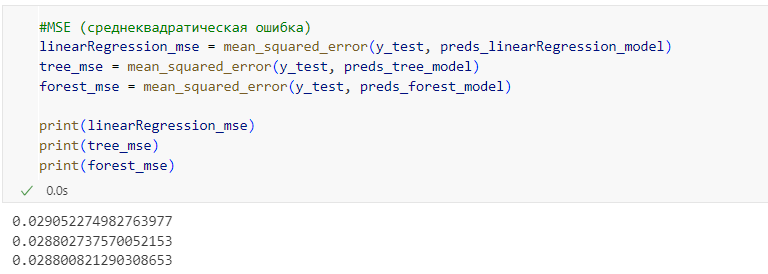


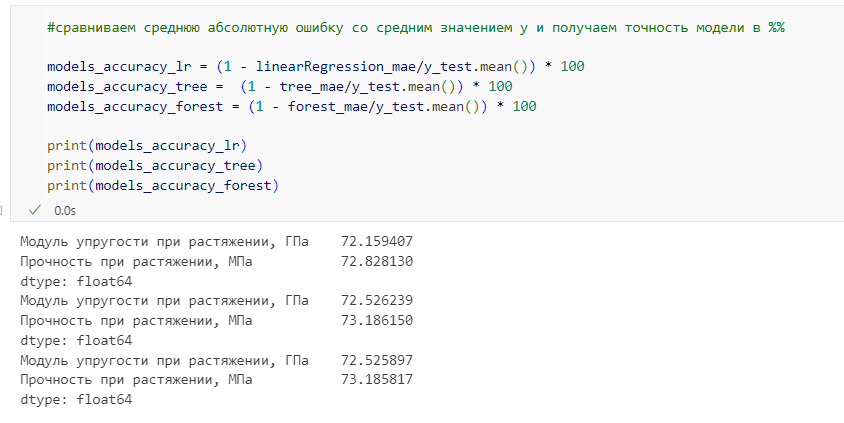
*2.3. Тестирование модели.*

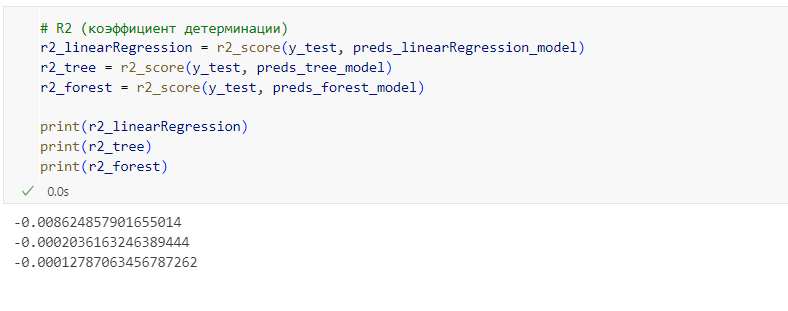
Все рассмотренные модели показали примерно одинаковый уровень точности (72-73%):

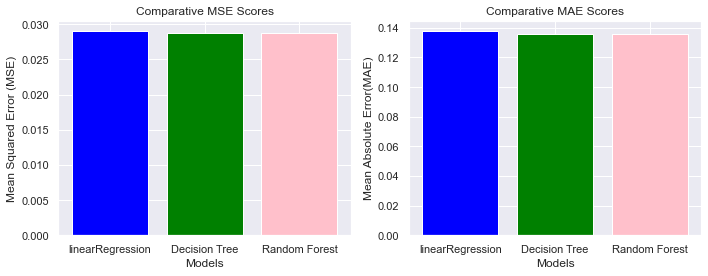
**





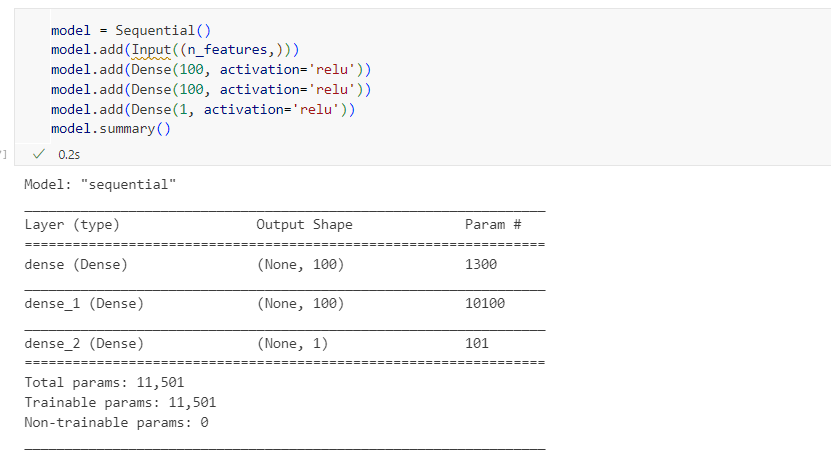
**



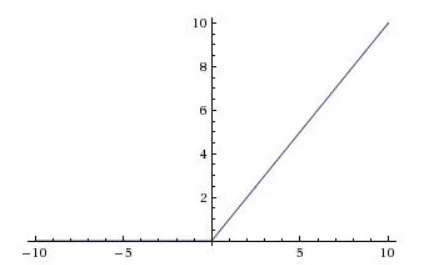
**

*2.4. Написать нейронную сеть, которая будет рекомендовать соотношение матрицы.*

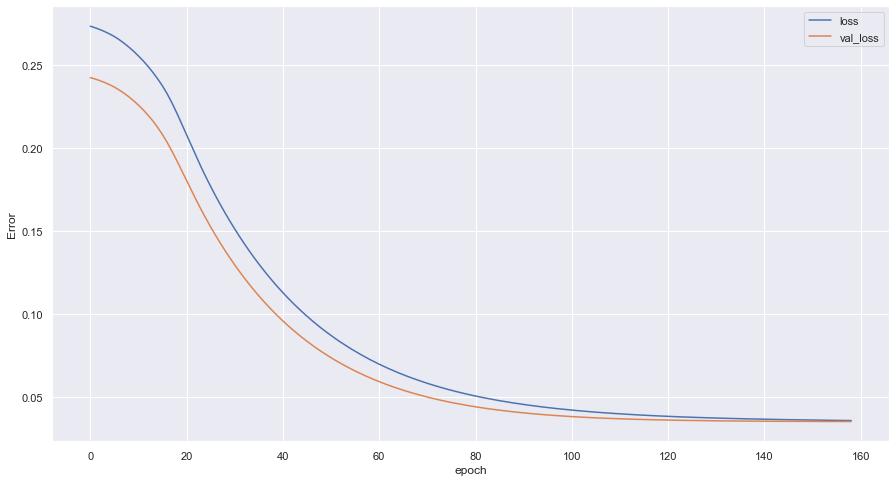
Для рекомендации соотношения матрица-наполнитель была выбрана следующая структура нейронной сети:

**

Входящий слой, два полносвязных слоя с количеством 100 нейронов в каждом и функцией активации Relu. Функция активации определяет выходное значение нейрона в зависимости от результата взвешенной суммы входов и порогового значения. ReLu возвращает значение х, если х положительно, и 0 в противном случае:



Точность модели порядка 68%.



*2.5. Разработка приложения.*

Было разработано flask-приложение, позволяющее в дружественном интерфейсе параметры композиционных материалов и возможных предполагаемых условий их взаимодействия, которые благодаря внедренной в приложение модели простой линейной регрессии будут обрабатываться приложением и выдавать прогноз модуля упругости и и прочности при растяжении композитов при заданных пользователем параметрах.

Для запуска приложения необходимо запустить файл app.py в среде Python и в браузе перейти по ссылке <https://127.0.0.1:5000/>.

*2.6. Создание удаленного репозитория и загрузка результатов работы на него.*

https://github.com/Khilbi/Composits-Bauman-

Список использованной литературы:

1.Интернет-ресурсы: <https://habr.com/>, <https://education.yandex.ru/>, https://blog.skillfactory.ru/

2. Материалы лекций МГТУ им.Баумана