

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»

Тема:

Прогнозирование конечных свойств новых материалов
(композиционных материалов).

Слушатель: Семиврагов Сергей Александрович

Москва, 2023

Содержание

Содержание.....	2
Введение.....	3
1 Аналитическая часть.....	4
1.1 Постановка задачи	4
1.2 Описание используемых методов	5
1.3 Разведочный анализ данных	2
2 Практическая часть	10
2.1 Предобработка данных.....	10
2.2 Разработка и обучение модели	14
2.3 Тестирование модели.....	14
2.4 Написать нейронную сеть, рекомендуя соотношение матрица	18
2.5 Разработка приложения.....	7
2.6 Создание удаленного репозитория и загрузка работы на него	9
Заключение	10
Библиографический список	11
Приложение А Все что не влезло в основную часть. Ошибка! Закладка не определена.	

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 обладает высокими свойствами сопротивления 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, при сохранении принципа формирования композиционных материалов. При таком подходе существуют определенные недостатки: даже если знать 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.

На специальном оборудовании проводятся различные испытания образцов (растяжение, изгиб, крутимость, твердость, ударная вязкость и др.)

Прогнозирование характеристик проводится путем симуляции представительного элемента объема композита на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

1 Аналитическая часть

1.1 Постановка задачи

Изучить теоретические основы и методы решения поставленной задачи.

Цель работы: разработать модели для прогноза модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель».

На входе имеются данные о начальных свойствах компонентов композиционных материалов, содержащихся в двух файлах:

1. X_br.xlsx, состоящий из 1024 строки и 11 столбцов (1 строка содержит – названия столбцов):
 - a) Соотношение матрица-наполнитель
 - b) Плотность, кг/м³
 - c) модуль упругости, ГПа
 - d) Количество отвердителя, м.%
 - e) Содержание эпоксидных групп, %₂
 - f) Температура вспышки, С₂
 - g) Поверхностная плотность, г/м²
 - h) Модуль упругости при растяжении, Гпа
 - i) Прочность при растяжении, Мпа
 - j) Потребление смолы, г/м²
2. X_nup.xlsx, состоящий из 1040 строки и 3 столбцов
 - k) Шаг нашивки
 - l) Плотность нашивки
 - m) Угол нашивки

На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов:

- 1) Модуль упругости при растяжении, ГПа;
- 2) Прочность при растяжении, Мпа.

Также необходимо написать нейронную сеть, которая будет рекомендовать, который будет определять значения «соотношение матрица-наполнитель».

Для этого нужно объединить 2 файла. Объединение делается по индексу тип объединения INNER. Часть информации (17 строк таблицы X_br.xlsx способов компоновки композитов) не имеет соответствующих строк в таблице

X_nur.xlsx (соотношений и свойств используемых компонентов композитов), поэтому данные строки были удалены.

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

Для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; 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».

Оценить точность модели на тренировочном и тестовом датасете. Создать репозиторий в GitHub и разместить код исследования. Оформить файл README.

1.2 Описание используемых методов

Целью поставленной задачи является предсказание значения зависимых переменных (модуль упругости при растяжении, прочность при растяжении) с помощью независимых переменных, поэтому решать ее будем методами регрессионного анализа.

Регрессионный анализ определяется как набор статистических процессов для оценки взаимосвязи между зависимой переменной и одной или несколькими независимыми переменными.

Метод ближайших соседей (kNN - k Nearest Neighbours) - метод решения задач классификации и задач регрессии, основанный на поиске ближайших объектов с известными значениями целевой переменной.

Метод основан на предположении о том, что близким объектам в признаковом пространстве соответствуют похожие метки. Метод предполагает найти для нового объекта ближайшие к нему объекты и построить прогноз по их меткам.

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).

Стохастический градиентный спуск (SGD) – это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под выпуклые функции потерь.

Линейная регрессия (Linear regression) – это алгоритм машинного обучения, основанный на контролируемом обучении, рассматривающий зависимость между одной входной и выходными переменными. Линейная регрессия – один из самых простых и эффективных инструментов статистического моделирования. Она определяет зависимость переменных с помощью линии наилучшего соответствия.

Метод 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 метод случайного леса – один из немногих алгоритмов, который можно использовать в абсолютном большинстве задач.

Выбирается подвыборка обучающей выборки размера `samplesize` (м.б. с возвращением), по ней строится дерево (для каждого дерева – своя подвыборка).

Для построения каждого расщепления в дереве просматриваем `max_features` случайных признаков (для каждого нового расщепления – свои случайные признаки).

Выбираем наилучший признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Таблица 1 – Используемые методы

№	Метод	Достоинства	Недостатки	Априорные предпосылки к работоспособности
1.	К-ближайших соседей	Прост в реализации и понимании полученных результатов; имеет низкую чувствительность к выбросам; допускает настройку нескольких параметров; позволяет делать дополнительные допущения; универсален; находит лучшее решение из возможных; решает задачи небольшой размерности	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 трудоемкость	Прост в реализации и понимании.
2.	Стохастический градиентный спуск	Эффективен; прост в реализации; имеет множество возможностей для настройки кода; способен обучаться на избыточно больших выборках	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.	Простой, но очень эффективный подход к подгонке линейных регрессоров.
3.	Линейная регрессия	Быстр и прост в реализации; легко интерпретируем, имеет меньшую сложность по сравнению с другими алгоритмами	Моделирует только прямые линейные зависимости; требует прямую связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние, а границы линейны.	Прост с алгоритмической точки зрения.
4.	Случайный лес (Random forest)	Не переобучается; не требует предобработки входных данных; эффективно обрабатывает пропущенные данные, данные с большим числом классов и признаков; имеет высокую точность предсказания и внутреннюю оценку обобщающей способности модели, а также высокую параллелизуемость и масштабируемость	Построение занимает много времени; сложно интерпретируемый; не обладает возможностью экстраполяции; может недообучаться; трудоёмко прогнозируемый; иногда работает хуже, чем линейные методы	Универсальный метод.

1.3 Разведочный анализ данных

Разведочный анализ данных (Exploratory Data Analysis) – предварительное исследование датасета с целью определения его основных характеристик, взаимосвязей между признаками.

Предварительная обработка и очистка данных должны проводиться до того, как набор данных будет использоваться для обучения модели. Необработанные данные зачастую искажены и ненадежны, и в них могут быть пропущены значения. Использование таких данных при моделировании может приводить к неверным результатам. Эти задачи являются частью процесса обработки и анализа данных группы и обычно подразумевают первоначальное изучение набора данных, используемого для определения и планирования необходимой предварительной обработки.

В данном разделе приводится краткое описание методов разведочного анализа данных, которые используются для первоначального анализа.

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.

При разведочном анализе данных датасета используются следующие методы:

- 1) анализ статистических характеристик;
- 2) проверка наличия пропусков и дубликатов;
- 3) гистограммы распределения каждой из переменной;
- 4) попарные графики рассеяния точек;
- 5) корреляция Пирсона и диаграмма тепловая карта;
- 6) диаграммы boxplot (ящик с усами).

Анализ статистических характеристик датасета.

Объединенный датасет имеет 13 столбцов и 1023 строки.

17 строк из таблицы X_pur были отброшены в процессе объединения данных.

Просмотрим информацию о датасете, проверим тип данных в каждом столбце (типы признаков):

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1023 entries, 0 to 1022
Data columns (total 13 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Соотношение матрица-наполнитель          1023 non-null   float64
1   Плотность, кг/м3                          1023 non-null   float64
2   модуль упругости, ГПа                     1023 non-null   float64
3   Количество отвердителя, м.%               1023 non-null   float64
4   Содержание эпоксидных групп,%_2          1023 non-null   float64
5   Температура вспышки, C_2                 1023 non-null   float64
6   Поверхностная плотность, г/м2            1023 non-null   float64
7   Модуль упругости при растяжении, ГПа     1023 non-null   float64
8   Прочность при растяжении, МПа            1023 non-null   float64
9   Потребление смолы, г/м2                  1023 non-null   float64
10  Угол нашивки, град                        1023 non-null   int64
11  Шаг нашивки                              1023 non-null   float64
12  Плотность нашивки                         1023 non-null   float64
dtypes: float64(12), int64(1)
memory usage: 111.9 KB
```

Все переменные, кроме "Угол нашивки, град" содержат значения float64. Переменные "Угол нашивки, град" содержат значения int64. Качественные характеристики отсутствуют. Пропусков не имеется. Ни одна из записей не является NaN. Таким образом, очистка данных не требуется. Дубликатов строк в датасете не выявлено.

Статистические характеристики датасета содержат: количество значений, среднее значение, стандартное отклонение, минимум и максимум, верхние значения первого и третьего квартиля и медиану по каждому столбцу.

Таблица 2 – Статистические характеристики датасета

index	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023	2,930366	0,913222	0,389403	2,317887	2,906878	3,55266	5,591742
Плотность, кг/м3	1023	1975,735	73,72923	1731,765	1924,155	1977,622	2021,374	2207,773
модуль упругости, ГПа	1023	739,9232	330,2316	2,436909	500,0475	739,6643	961,8125	1911,536
Количество отвердителя, м. %	1023	110,5708	28,29591	17,74027	92,4435	110,5648	129,7304	198,9532
Содержание эпоксидных групп, % 2	1023	22,24439	2,406301	14,25499	20,60803	22,23074	23,96193	33
Температура вспышки, С 2	1023	285,8822	40,94326	100	259,0665	285,8968	313,0021	413,2734
Поверхностная плотность, г/м2	1023	482,7318	281,3147	0,60374	266,8166	451,8644	693,225	1399,542
Модуль упругости при растяжении, ГПа	1023	73,32857	3,118983	64,05406	71,24502	73,2688	75,35661	82,68205
Прочность при растяжении, МПа	1023	2466,923	485,628	1036,857	2135,85	2459,525	2767,193	3848,437
Потребление смолы, г/м2	1023	218,4231	59,73593	33,80303	179,6275	219,1989	257,4817	414,5906
Угол нашивки, град	1023	44,2522	45,01579	0	0	0	90	90
Шаг нашивки	1023	6,899222	2,563467	0	5,080033	6,916144	8,586293	14,44052
Плотность нашивки	1023	57,15393	12,35097	0	49,79921	57,34192	64,94496	103,9889

Все значения датасета положительные.

Гистограммы распределения всех переменных датасета представлены на Рисунке 1 (Гистограммы распределения переменных датафрейма).

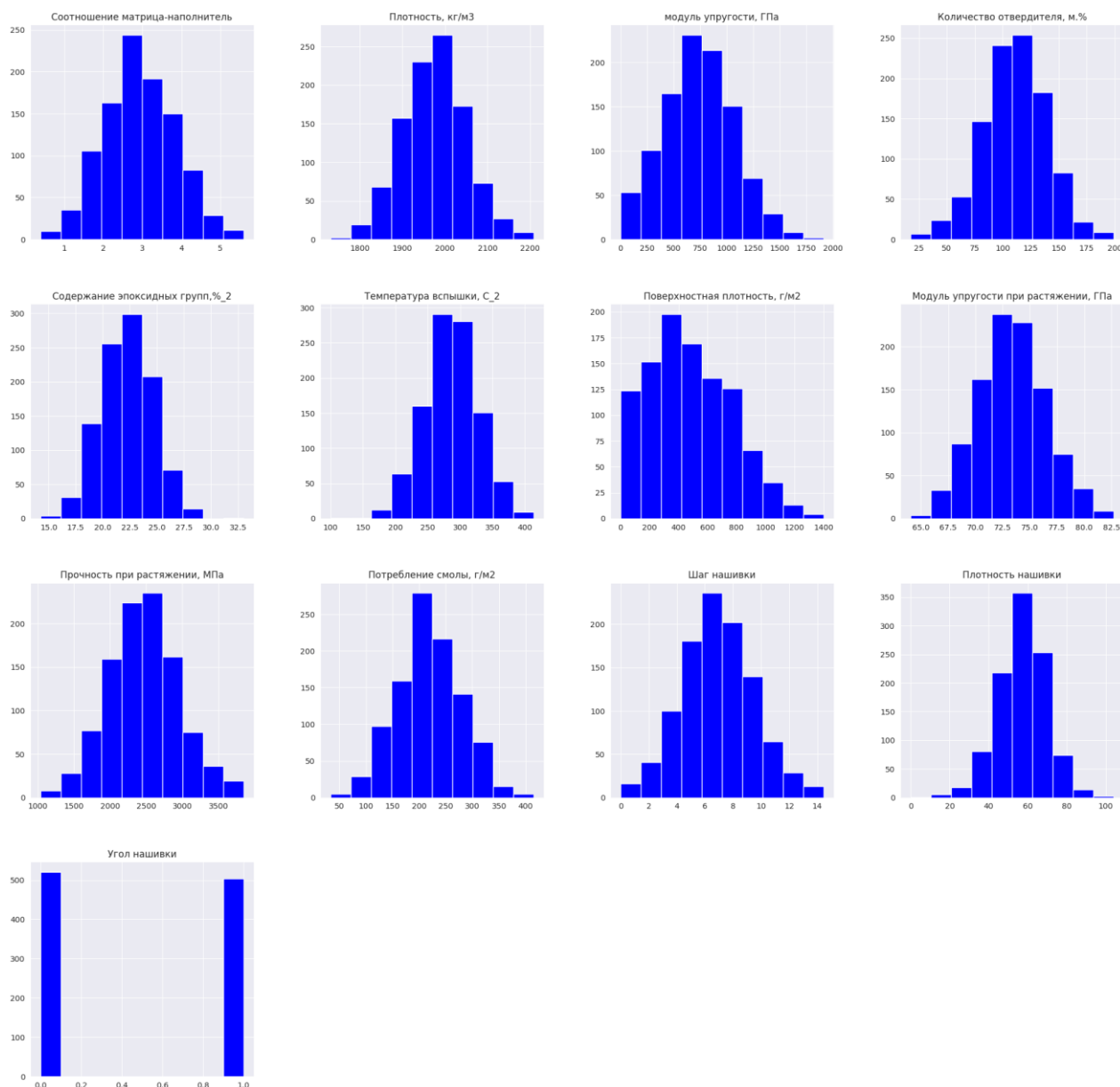


Рисунок 1 – Гистограммы распределения переменных датафрейма

Визуально гистограммы показывают нормальное распределение, за исключением распределения признаков «Потребление смолы, г/м²» – график имеет выраженное смещение влево, а также признака «Угол нашивки», который имеет всего два значения 0 и 90 градусов.

Диаграммы рассеяния, показывают положительные линейные отношения (когда x увеличивается, увеличивается y), либо отрицательные (когда x увеличивается, y уменьшается). Гистограммы в диагональных прямоугольниках, показывают распределение конкретных признаков.

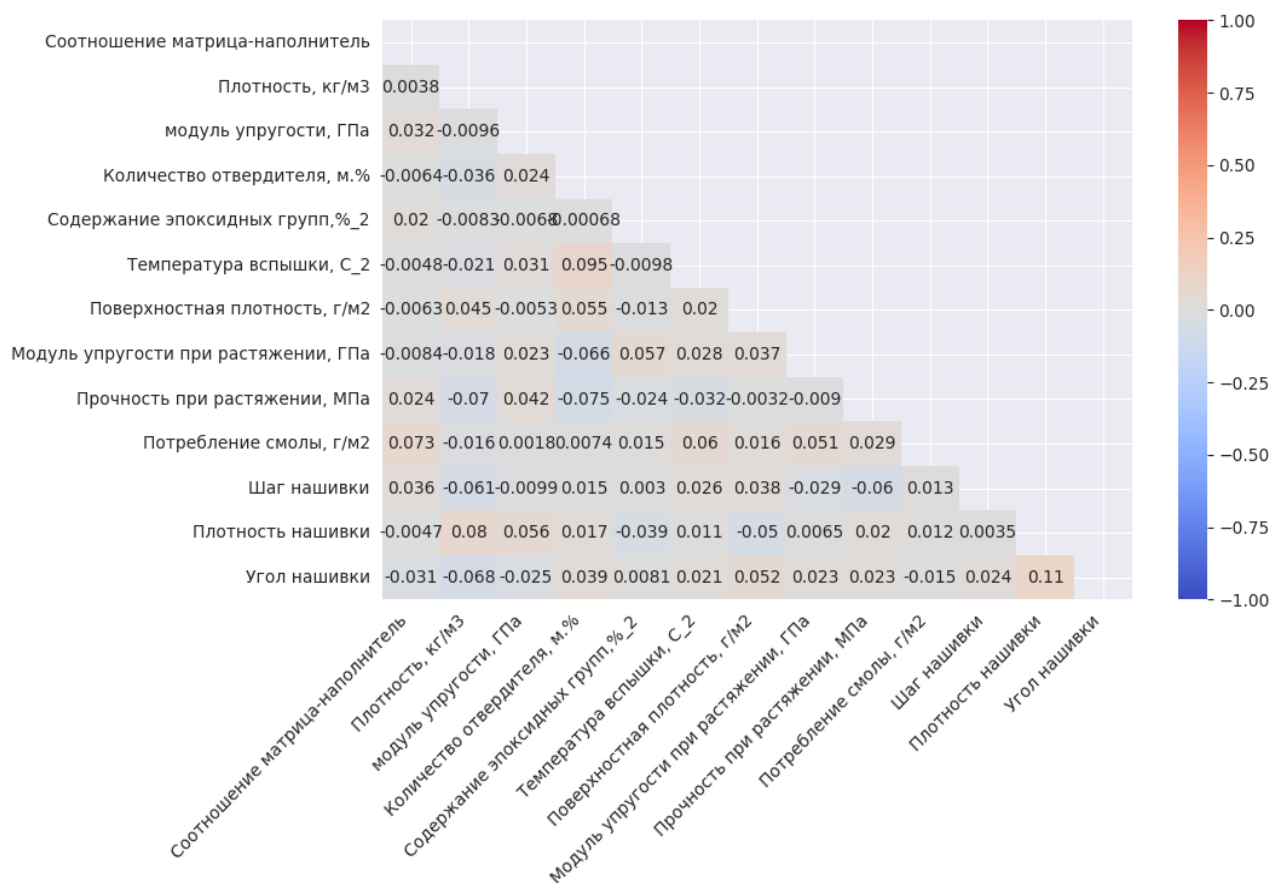


Рисунок 3 – Тепловая карта со значениями корреляции данных

На парных диаграммах визуально корреляции между признаками не наблюдается. Тепловая карта показывает практически отсутствие корреляции между признаками и целевыми переменными.

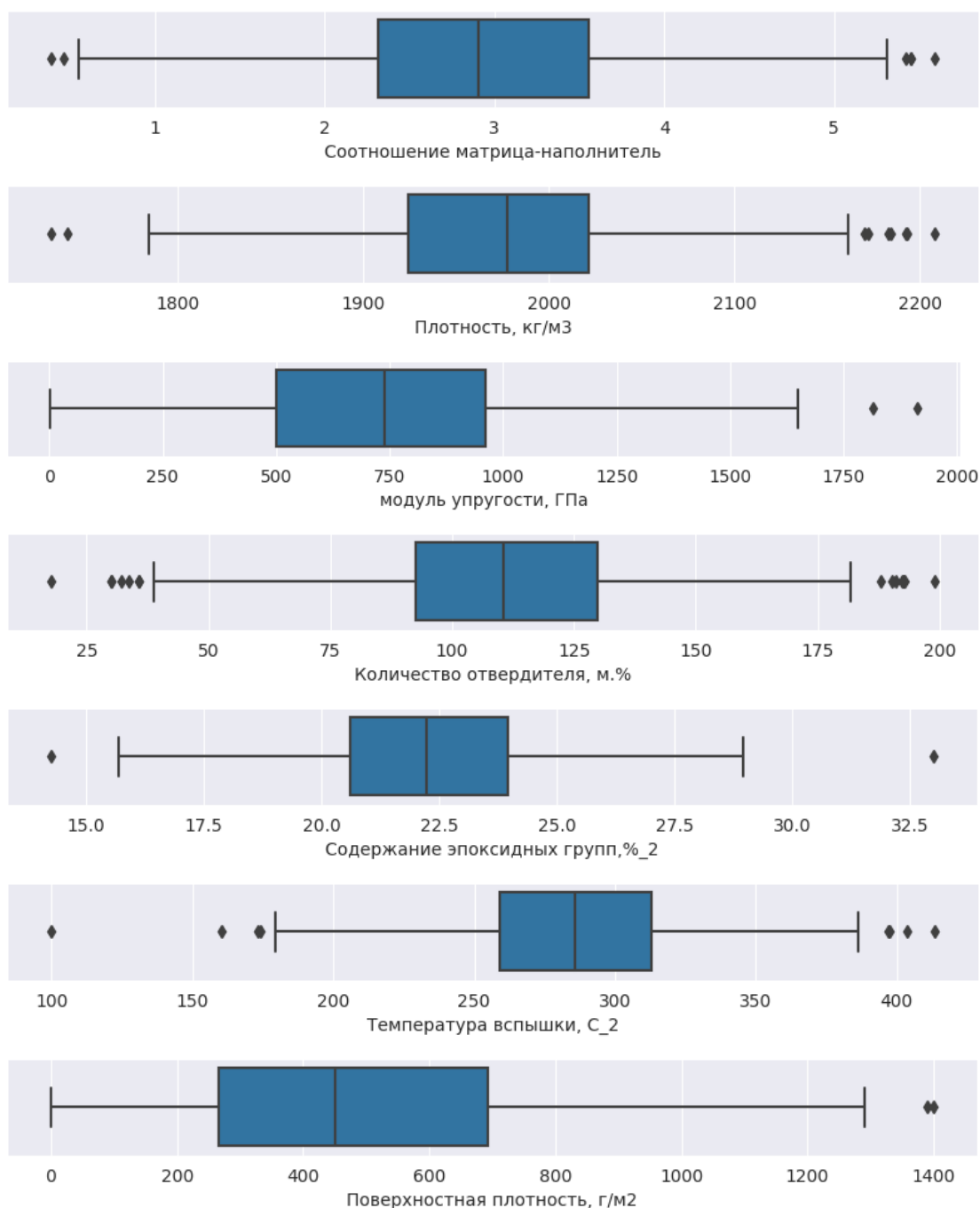
Для оценки силы связи в теории корреляции применяется шкала Чеддока:

- 1) слабая — от 0,1 до 0,3;
- 2) умеренная — от 0,3 до 0,5;
- 3) заметная — от 0,5 до 0,7;
- 4) высокая — от 0,7 до 0,9;
- 5) весьма высокая (сильная) — от 0,9 до 1,0.

В соответствии с матрицей корреляции максимальная корреляция наблюдается между плотностью нашивки и углом нашивки 0.11. В соответствии с таблицей Чедока это свидетельствует о слабых силах связи между значениями переменных.

В остальных случаях корреляция между всеми параметрами очень близка к нулю, следовательно, корреляционные связи между переменными не наблюдаются.

Boxplot – это диаграмма, которая показывает, как распределяются значения переменной. Диаграмма известна как ящик с усами, и она дает информацию об изменчивости и дисперсии данных с использованием сводки из пяти чисел. К ним относятся минимум, первый квартиль (Q1), медиана, третий квартиль (Q3) и максимум. Используются для обнаружения выбросов.



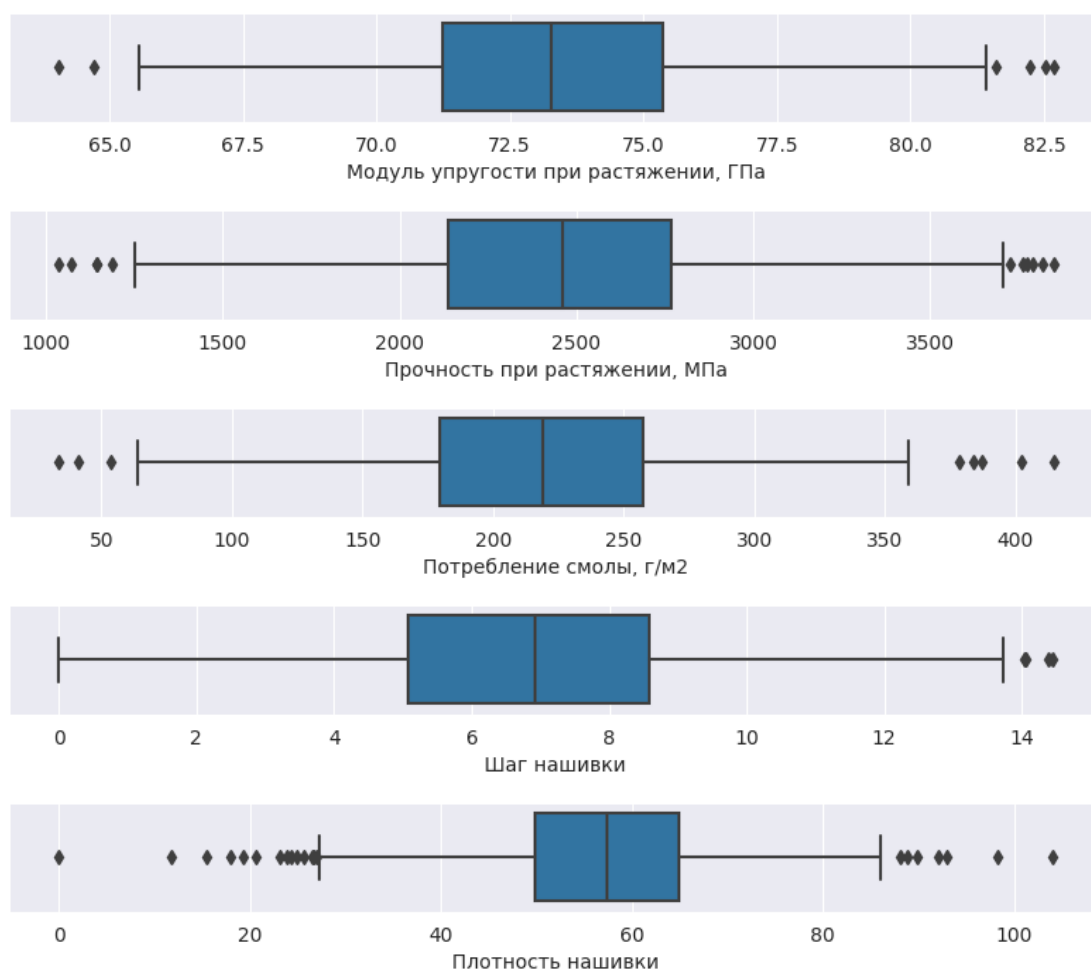


Рисунок 4 – Диаграммы boxplot на начальном датасете

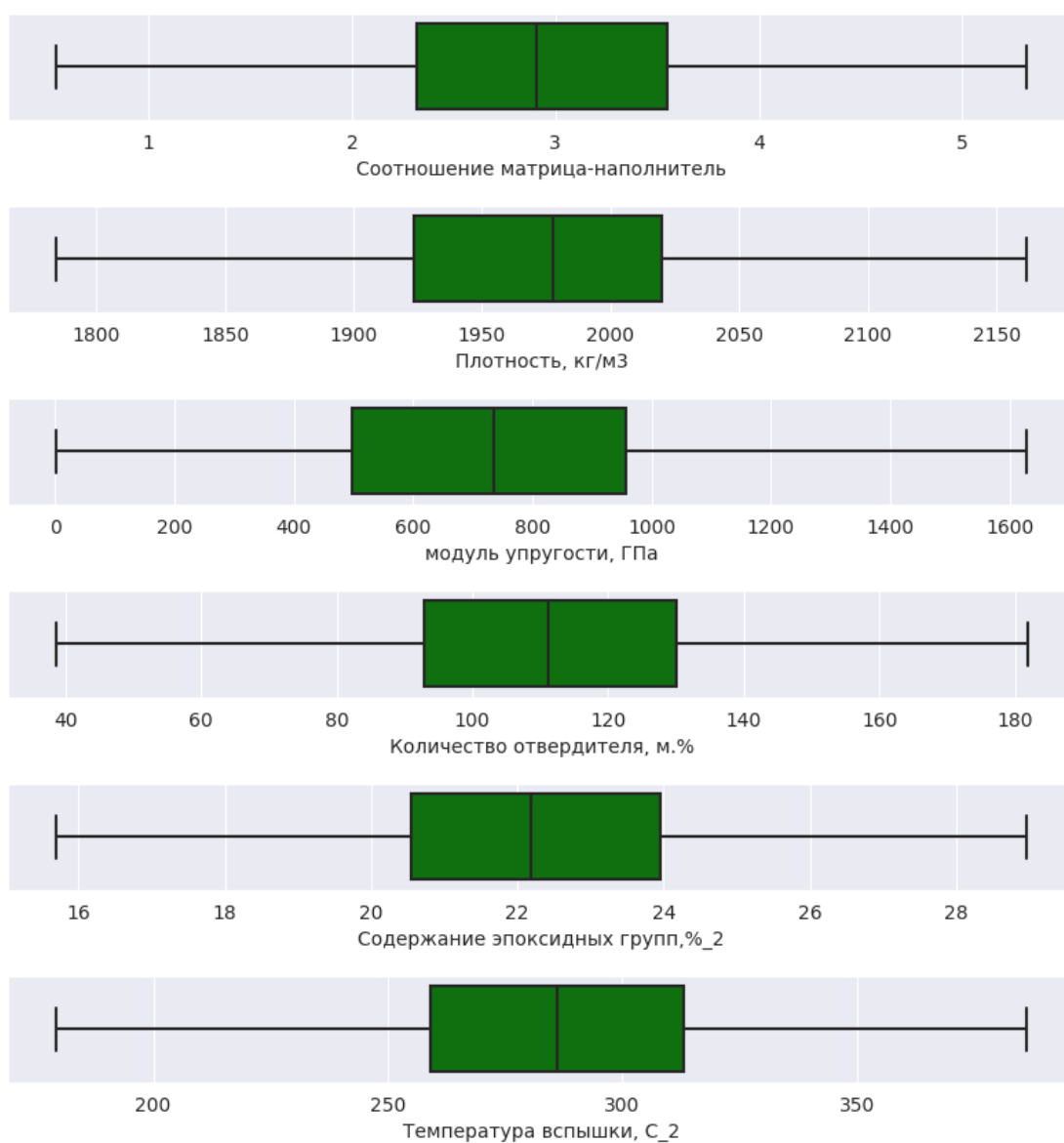
Диаграммы boxplot показывают наличие выбросов, особенно в признаке “Плотность нашивки”.

2 Практическая часть

2.1 Предобработка данных

Так как количество уникальных значений в колонке "Угол нашивки, град" равно 2, приведем данные в этой колонке к значениям 0 и 1 с помощью метода LabelEncoder.

Для удаления выбросов используются методы трех сигм и межквартильного расстояния. В нашем случае применим способ межквартильного расстояния для максимальной чистоты, так как используем методы чувствительные к выбросам.



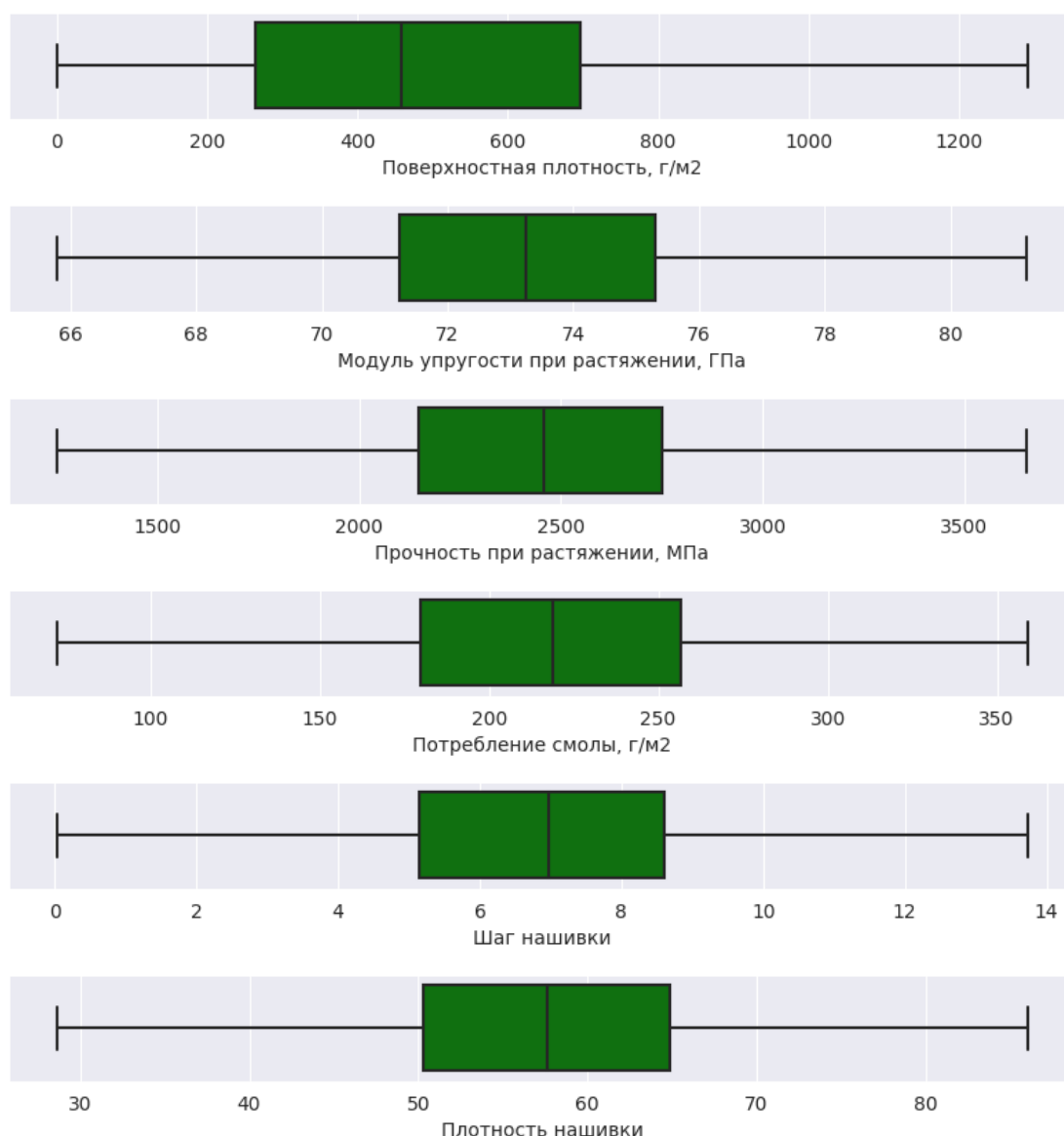


Рисунок 5 – Диаграммы boxplot после удаления выбросов

В результате проведенной очистки выбросов в датасете осталось 922 строки:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 922 entries, 1 to 1022
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Соотношение матрица-наполнитель      922 non-null    float64
1   Плотность, кг/м3                      922 non-null    float64
2   модуль упругости, ГПа                 922 non-null    float64
3   Количество отвердителя, м.%           922 non-null    float64
4   Содержание эпоксидных групп,%_2       922 non-null    float64
5   Температура вспышки, С_2              922 non-null    float64
6   Поверхностная плотность, г/м2         922 non-null    float64
7   Модуль упругости при растяжении, ГПа  922 non-null    float64
8   Прочность при растяжении, МПа         922 non-null    float64
```

9	Потребление смолы, г/м2	922 non-null	float64
10	Шаг нашивки	922 non-null	float64
11	Плотность нашивки	922 non-null	float64

dtypes: float64(12)
memory usage: 93.6 KB

Отобразим распределение данных в датасете на одном графике.

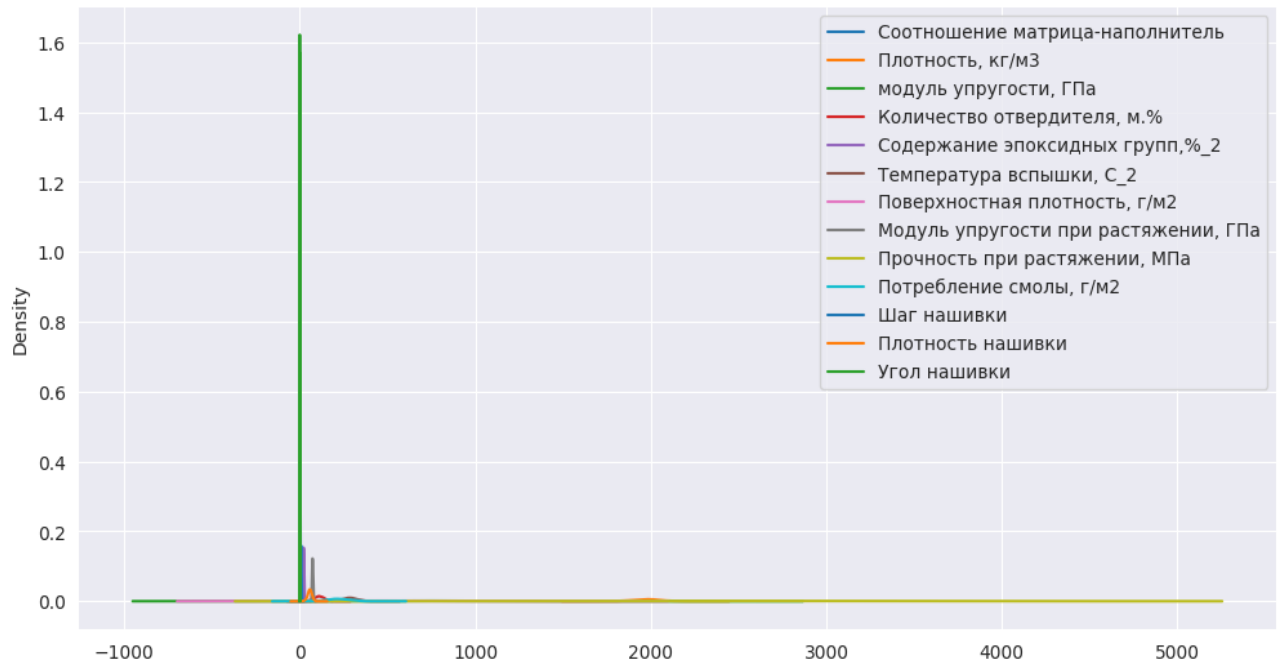


Рисунок 6 – Распределение данных в датасете

Оценка плотности ядра данных датасета показывает, что данные находятся в разных диапазонах, их необходимо нормализовать.

Нормализация датасета – это преобразование данных к неким безразмерным единицам, в рамках заданного диапазона, например, $[0 \dots 1]$.

Нормализуем значения с помощью метода `MinMaxScaler`, содержащегося в библиотеке `Scikit-learn`, и выведем описательную характеристику получившегося датасета.

Таблица 3 – Описательная статистика нормализованного датасета

index	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	922	0,499411883	0,187858	0	0,371909	0,495189	0,629774	1
Плотность, кг/м3	922	0,502903905	0,188395	0	0,368184	0,511396	0,624719	1
модуль упругости, ГПа	922	0,451340878	0,201534	0	0,305188	0,451377	0,587193	1
Количество отвердителя, м.%	922	0,506199998	0,186876	0	0,378514	0,506382	0,638735	1
Содержание эпоксидных групп,%_2	922	0,490578277	0,180548	0	0,366571	0,488852	0,623046	1
Температура вспышки, С_2	922	0,516739445	0,190721	0	0,386228	0,516931	0,646553	1
Поверхностная плотность, г/м2	922	0,373294919	0,217269	0	0,204335	0,354161	0,538397	1
Модуль упругости при растяжении, ГПа	922	0,48734326	0,196366	0	0,353512	0,483718	0,617568	1
Прочность при растяжении, МПа	922	0,503776031	0,188668	0	0,373447	0,501481	0,624299	1
Потребление смолы, г/м2	922	0,507875548	0,199418	0	0,374647	0,510143	0,642511	1
Шаг нашивки	922	0,503425946	0,183587	0	0,372844	0,506414	0,626112	1
Плотность нашивки	922	0,503938174	0,193933	0	0,376869	0,50431	0,630842	1
Угол нашивки	922	0,510845987	0,500154	0	0	1	1	1

В соответствии с таблицей 5 видно, что все данные датасета нормализованы и находятся в диапазоне [0...1].

Визуализируем распределение нормализованных данных.

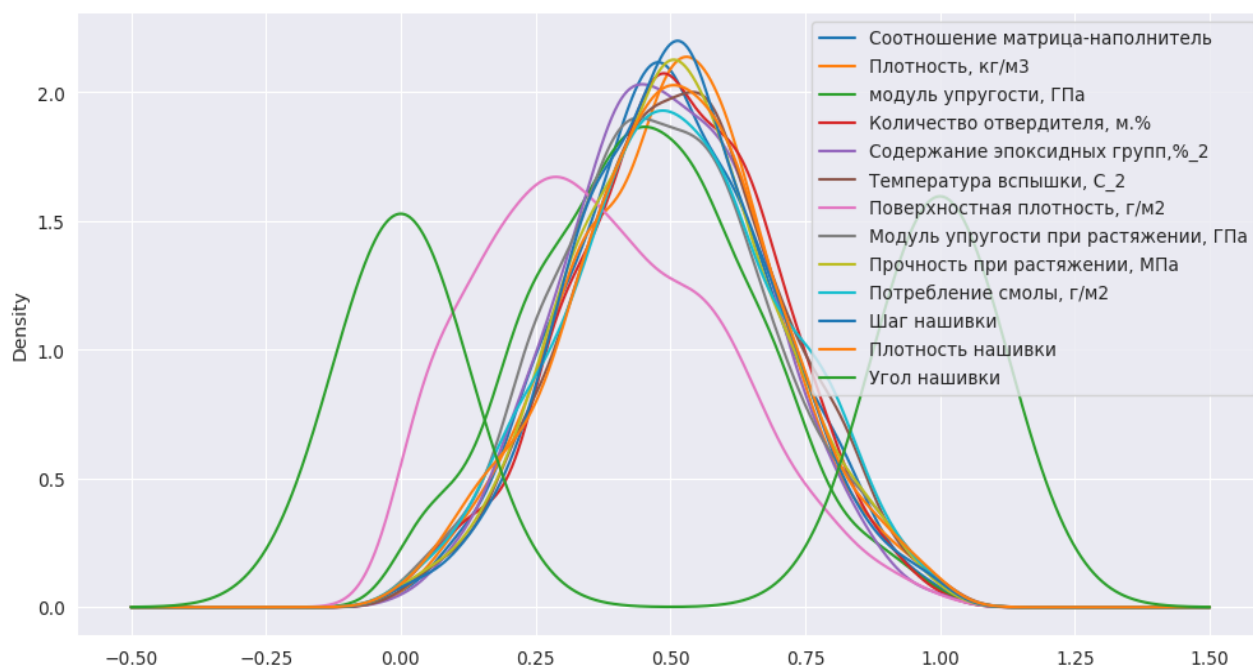


Рисунок 7 – Распределение нормализованных данных в датасете.

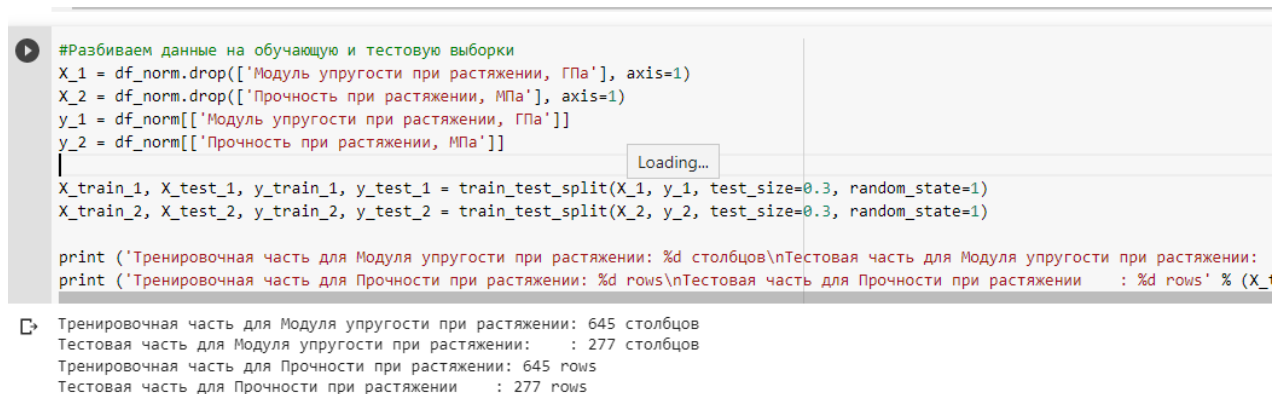
2.2 Разработка и обучение модели

Разработка и обучение моделей машинного обучения осуществлялась для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении» отдельно.

Для решения применим следующие модели:

- 1) k-ближайших соседей;
- 2) стохастический градиентный спуск;
- 3) линейная регрессия;
- 4) случайный лес (Random forest);

Данные датасета разбиваются на тестовую часть (30%) – содержит 277 строки и тренировочная часть (70%) – 645 строк.



```
#Разбиваем данные на обучающую и тестовую выборки
X_1 = df_norm.drop(['Модуль упругости при растяжении, ГПа'], axis=1)
X_2 = df_norm.drop(['Прочность при растяжении, МПа'], axis=1)
y_1 = df_norm[['Модуль упругости при растяжении, ГПа']]
y_2 = df_norm[['Прочность при растяжении, МПа']]

X_train_1, X_test_1, y_train_1, y_test_1 = train_test_split(X_1, y_1, test_size=0.3, random_state=1)
X_train_2, X_test_2, y_train_2, y_test_2 = train_test_split(X_2, y_2, test_size=0.3, random_state=1)

print ('Тренировочная часть для Модуля упругости при растяжении: %d столбцов\nТестовая часть для Модуля упругости при растяжении: %d столбцов' % (X_train_1.shape[1], X_test_1.shape[1]))
print ('Тренировочная часть для Прочности при растяжении: %d rows\nТестовая часть для Прочности при растяжении : %d rows' % (X_train_2.shape[0], X_test_2.shape[0]))
```

Тренировочная часть для Модуля упругости при растяжении: 645 столбцов
Тестовая часть для Модуля упругости при растяжении: : 277 столбцов
Тренировочная часть для Прочности при растяжении: 645 rows
Тестовая часть для Прочности при растяжении : 277 rows

Рисунок 8 – Разбиение данных на тестовую и тренировочную часть

2.3 Тестирование модели

После обучения моделей была проведена оценка точности этих моделей на обучающей и тестовых выборках.

Для оценки качества применяемых моделей использовались следующие метрики:

1. Средняя квадратическая ошибка (MSE);
2. Средняя абсолютная ошибка (MAE);
3. Коэффициент детерминации R-квадрат.

Средняя квадратическая ошибка (MSE) – находит среднеквадратичную ошибку между прогнозируемыми и фактическими значениями и определяется по формуле:



Применяется в случаях, когда требуется подчеркнуть большие ошибки и выбрать модель, которая дает меньше именно больших ошибок.

Средняя абсолютная ошибка (MAE)

Рассчитывается как среднее абсолютных разностей между наблюдаемым и предсказанным значениями.



В отличие от MSE она является линейной оценкой, а это значит, что все ошибки в среднем взвешены одинаково.

Коэффициент детерминации R-квадрат показывает долю дисперсии зависимой переменной, объяснённой с помощью регрессионной модели.

$$R^2 = 1 - \frac{\sum (y_j - \hat{y}_j)^2}{\sum (y_j - \bar{y})^2}$$

Когда коэффициент R2 принимает отрицательные значения (обычно небольшие). Это произойдёт, если ошибка модели среднего становится меньше ошибки модели с переменной. В этом случае оказывается, что добавление в модель с константой некоторой переменной только ухудшает её (т.е. регрессионная модель с переменной работает хуже, чем предсказание с помощью простой средней).

На практике используют следующую шкалу оценок:

- если $R^2 > 0.8$, то модель рассматривается как очень хорошая.
- модель, для которой $R^2 > 0.5$, является удовлетворительной.
- значения, меньшие 0.5 говорят о том, что модель плохая.

Результаты оценки моделей представлены отсортированные по MAE и R2 score в таблице 3.

Таблица 4 – Результаты оценки моделей

№	Модель	Искомое значение	MSE	MAE	R2 score
1.	LinearRegression_pr	Прочность при растяжении	0.034523	0.149937	0.009
2.	RandomForestRegressor_pr	Прочность при растяжении	0.035424	0.151399	-0.017
3.	SGDRegressor_pr	Прочность при растяжении	0.035107	0.151721	-0.008
4.	SGDRegressor_upr	Модуль упругости при растяжении	0.038705	0.159921	-0.011
5.	LinearRegression_upr	Модуль упругости при растяжении	0.039564	0.161937	-0.034
6.	RandomForestRegressor_upr	Модуль упругости при растяжении	0.040606	0.164424	-0.061
7.	KNeighborsRegressor_pr	Прочность при растяжении	0.044102	0.165831	-0.266
8.	KNeighborsRegressor_upr	Модуль упругости при растяжении	0.049561	0.181919	-0.295

Результат всех примененных моделей неудовлетворительный, что может говорить о «лишних» данных в изначальном датасете.

Получена визуализация разброса предсказанных и тестовых данных в зависимости от примененного метода.

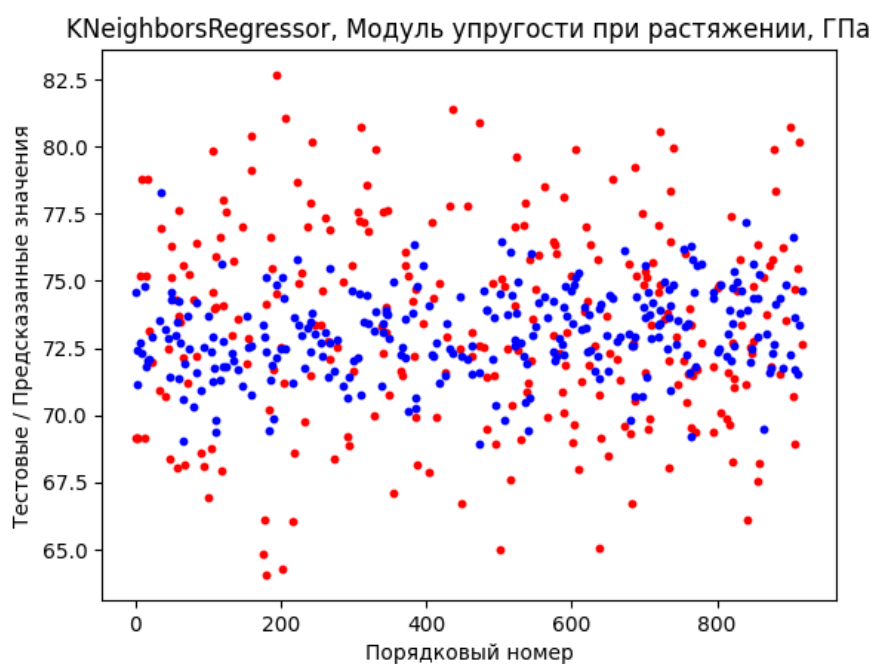


Рисунок 9 – Разброс предсказанных и тестовых данных модели KNeighborsRegressor



Рисунок 10 – Разброс предсказанных и тестовых данных модели SGDRegressor

2.4 Нейронная сеть, которая будет рекомендовать соотношение «матрица – наполнитель»

Загружаем объединенный, очищенный и нормализованный датасет. Разделяем датасет на зависимые и независимые переменные. Зависимые и независимые переменные разделяем на тренировочную и тестирующую выборки:

- тренировочная часть – 645 строк;
- тестовая часть – 277 строк.

Создаем архитектуру нейронной сети и запускаем обучение. Оценивая результаты меняем параметры нейросети: количество нейронов, функции активации, количество слоев, добавление слоя Dropout для решения проблемы переобучения.

	MAE	MSE	R2
['Нейросеть 1 ',	0.15617427623257507,	0.037204866305288825,	-0.001247942991688289],
['Нейросеть 2 ',	0.1577630915770674,	0.037468339572840154,	-0.008338468865492388],
['Нейросеть 3 ',	0.15488497272542256,	0.037296911398320226,	-0.0037250372326391723],
['Нейросеть 4 ',	0.16182539498664764,	0.04181927417240977,	-0.1254297192993412],
['Нейросеть 5 ',	0.15555790582130452,	0.03715067493023215,	0.0002104415744100807]

Рисунок 11 – Оценка нейросетей с различной архитектурой

Model: "sequential_27"

Layer (type)	Output Shape	Param #
dense_109 (Dense)	(None, 32)	416
dropout_36 (Dropout)	(None, 32)	0
dense_110 (Dense)	(None, 64)	2112
dropout_37 (Dropout)	(None, 64)	0
dense_111 (Dense)	(None, 32)	2080
dense_112 (Dense)	(None, 16)	528
dense_113 (Dense)	(None, 1)	17

=====
Total params: 5,153
Trainable params: 5,153
Non-trainable params: 0

Рисунок 12 – Структура нейросети с наименьшей ошибкой

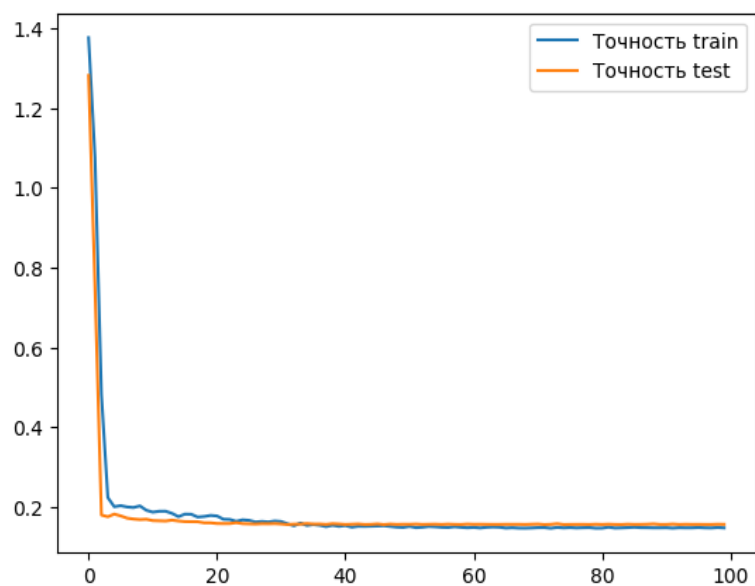


Рисунок 13 – Визуализация работы нейросети с наименьшей ошибкой

Все нейросети показали схожий результат. Коэффициент детерминации меньше 0.5.

2.7 Проверка датасета на вброс сгенерированных данных

В процессе проведения работы, не получив какого-либо предсказания и закономерности возникло предположение, что не весь датасет был получен при проведении испытания композиционных материалов, часть данных была сгенерирована.

Для проведения исследования на добавление сгенерированных данных проведена проверка на повторяющиеся значения в каждом столбце и сформирована визуализация повторяющихся данных.

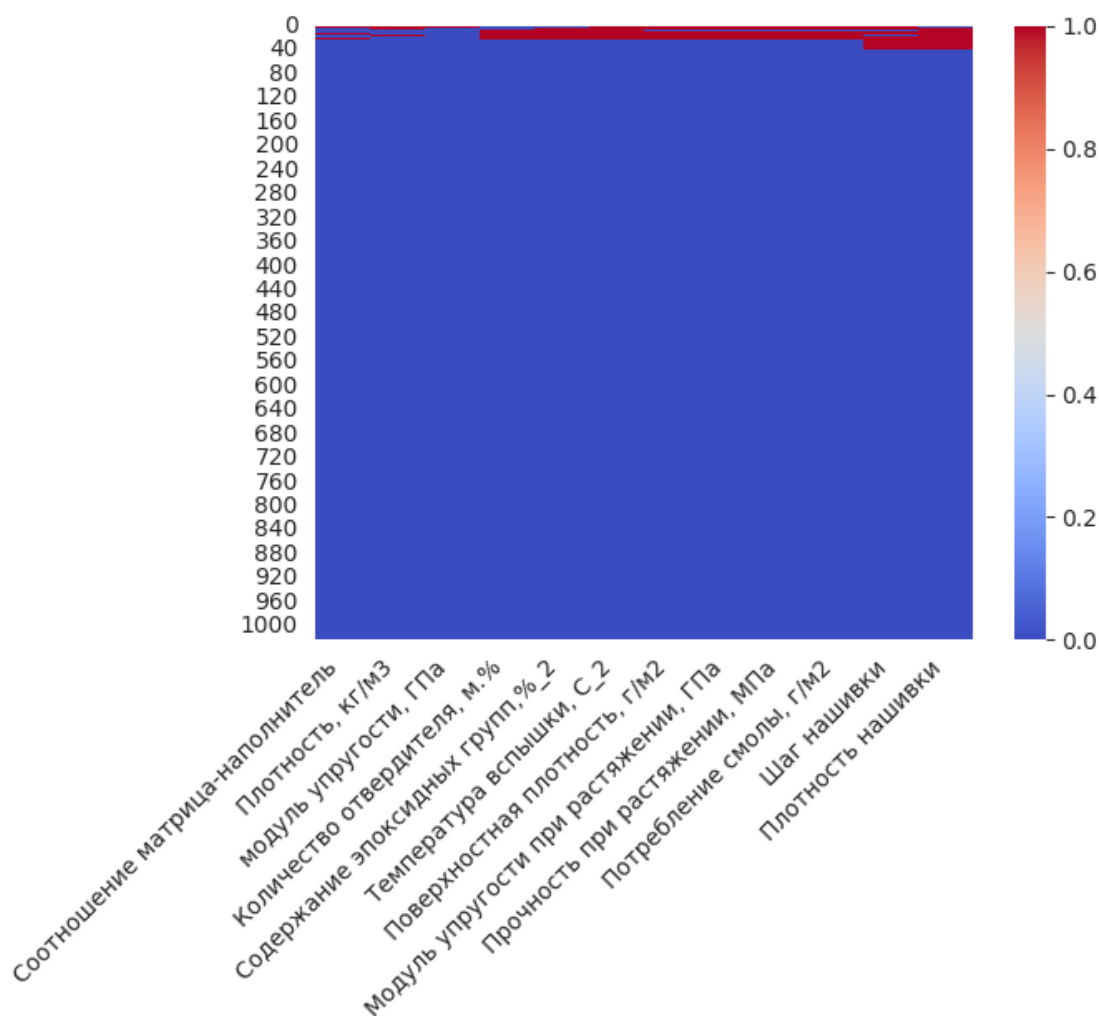


Рисунок 14 – Тепловая карта повторяющихся значений

Тепловая карта показывает, что некоторые данные повторяются в промежутке где-то до 50-й строчки. Составим диаграмму «Тепловая карта» на данных до 50-й строчки.

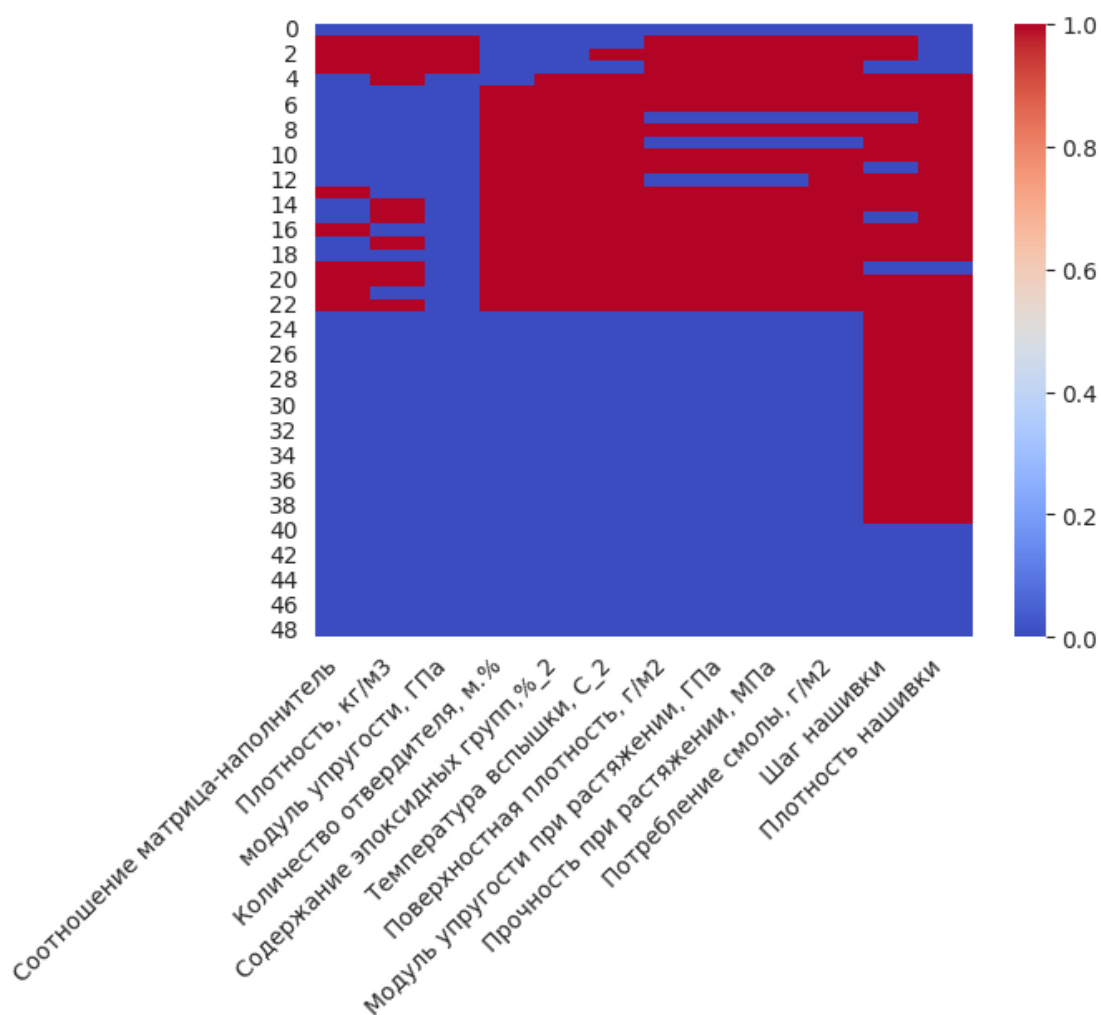


Рисунок 15 – Тепловая карта повторяющихся значений – первых 50 строк

На рисунке видно, что основная часть значений датасета повторяется в 10 первых столбцах и 22 первых строчках, что свидетельствует о добавлении в изначальные экспериментальные данные сгенерированных данных – ровно 1000 строк.

Это косвенно подтверждает характер данных, если их изобразить в таблице с 0 по 49 строку. Сгенерированные данные выделены желтым цветом.

Таблица 5 – Данные объединенного датасета с 0 по 49 строку

	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	Шаг нашивки	Плотность нашивки	Угол нашивки
0	1,857142857	2030	738,7368421	30	22,26785714	100	210	70	3000	220	4	57	0
1	1,857142857	2030	738,7368421	50	23,75	284,6153846	210	70	3000	220	4	60	0
2	1,857142857	2030	738,7368421	49,9	33	284,6153846	210	70	3000	220	4	70	0
3	1,857142857	2030	738,7368421	129	21,25	300	210	70	3000	220	5	47	0
4	2,771331058	2030	753	111,86	22,26785714	284,6153846	210	70	3000	220	5	57	0
5	2,767918089	2000	748	111,86	22,26785714	284,6153846	210	70	3000	220	5	60	0
6	2,569620253	1910	807	111,86	22,26785714	284,6153846	210	70	3000	220	5	70	0
7	2,56147541	1900	535	111,86	22,26785714	284,6153846	380	75	1800	120	7	47	0
8	3,557017544	1930	889	129	21,25	300	380	75	1800	120	7	57	0
9	3,532338308	2100	1421	129	21,25	300	1010	78	2000	300	7	60	0
10	2,919677836	2160	933	129	21,25	300	1010	78	2000	300	7	70	0
11	2,877358491	1990	1628	129	21,25	300	1010	78	2000	300	9	47	0
12	1,598173516	1950	827	129	21,25	300	470	73,33333333	2455,555556	220	9	57	0
13	2,919677836	1980	568	129	21,25	300	470	73,33333333	2455,555556	220	9	60	0
14	4,029126214	1910	800	129	21,25	300	470	73,33333333	2455,555556	220	9	70	0
15	2,934782609	2030	302	129	21,25	300	210	70	3000	220	10	47	0
16	3,557017544	1880	313	129	21,25	300	210	70	3000	220	10	57	0
17	4,193548387	1950	506	129	21,25	300	380	75	1800	120	10	60	0
18	4,897959184	1890	540	129	21,25	300	380	75	1800	120	10	70	0
19	3,532338308	1980	1183	111,86	22,26785714	284,6153846	1010	78	2000	300	0	0	0
20	2,877358491	2000	205	111,86	22,26785714	284,6153846	1010	78	2000	300	4	47	90
21	1,598173516	1920	456	111,86	22,26785714	284,6153846	470	73,33333333	2455,555556	220	4	57	90
22	4,029126214	1880	622	111,86	22,26785714	284,6153846	470	73,33333333	2455,555556	220	4	60	90
23	2,587347643	1953,274926	1136,596135	137,6274196	22,34453357	234,716883	555,8934533	80,80322176	2587,342983	246,6131165	4	70	90
24	2,499917928	1942,595777	901,5199467	146,2522078	23,08175748	351,231874	864,7254838	76,17807508	3705,672523	226,2227604	5	47	90
25	2,046471464	2037,631811	707,570887	101,6172513	23,14639281	312,3072052	547,6012186	73,81706662	2624,026407	178,1985559	5	57	90
26	1,85647617	2018,220332	836,2943816	135,4016966	26,4355146	327,5103767	150,9614485	77,21076158	2473,187195	123,3445614	5	60	90
27	3,305535422	1917,907506	478,2862473	105,7869296	17,87409991	328,1545795	526,6921594	72,34570879	3059,032991	275,5758795	5	70	90
28	2,709554095	1892,071124	641,0525494	96,56329319	22,98929056	262,956722	804,5926208	74,51135922	2288,967377	126,8163389	7	47	90
29	2,282825314	2008,357592	393,9673255	149,3728324	21,66175068	330,498641	535,3714591	72,24492408	2704,445081	261,0770716	7	57	90
30	1,978140173	1973,629097	991,7240946	149,3721279	19,75057789	232,0581913	485,4537781	75,66570056	2448,943079	162,4936936	7	60	90
31	1,771436393	1872,49156	801,0338825	79,79454787	22,29630372	340,7368984	864,9291837	70,94759156	2796,785402	123,3562643	7	70	90
32	3,277086987	2010,047012	339,5504228	67,49899306	24,28060902	254,9490837	117,5352342	67,47870663	2462,605386	207,0185813	9	47	90
33	2,984362226	1912,315437	1183,091845	133,5490007	23,26379657	314,9961255	377,3890094	75,29045222	2303,770656	200,5802494	9	57	90
34	2,916149621	1879,969846	1003,270178	109,2395305	25,68275948	294,0485366	408,3542393	71,70085562	3086,546196	192,1911621	9	60	90
35	3,247617211	1813,2346	757,874479	81,37987084	23,42246524	279,0801575	575,0628571	69,34113288	3188,136358	252,8705688	9	70	90
36	2,423875673	1908,940601	530,2286864	58,26241428	24,07354923	325,138688	456,9080467	74,24435417	1890,505807	222,6994873	10	47	90
37	5,09899309	1977,339047	1572,096042	132,3430598	25,39700098	286,5564309	690,3648357	72,34163973	1386,578973	271,9013937	10	57	90
38	2,444176986	2085,495837	931,3106361	110,5648399	23,48713976	270,2867651	278,2300203	71,47906047	2740,229631	187,8613727	10	60	90
39	2,667696929	2078,894676	1542,168458	132,1474033	22,6501092	357,9728962	787,299217	76,47178847	2559,643047	163,902778	10	70	90
40	3,034399483	1968,401388	455,8710188	61,42129652	23,49072291	316,4145721	637,3768927	75,09037174	2848,490078	311,0523979	7,856166547	64,30196385	0
41	2,465204971	1936,099137	1056,554985	71,29405822	24,5233807	271,9756783	129,0771629	66,42079436	2868,586527	227,0225573	7,401542567	19,25053314	0
42	2,664388649	1996,159145	525,0577741	77,50688254	18,12610706	223,4086854	228,65810234	69,48977348	2220,587445	314,7766687	6,675780339	78,62329934	0
43	1,193529582	1965,929227	899,603701	102,9590686	19,56671624	225,8102229	871,0889548	73,45469452	2335,541792	91,04764633	7,52639832	38,17697532	0
44	2,914333275	2049,373404	382,2633585	81,35204737	16,39159473	233,2960627	561,9921308	69,81461516	2262,784366	303,0754524	8,32569922	46,04542763	0
45	4,315665782	1913,379677	822,9187355	143,5769371	24,2755881	274,9887944	260,8593411	75,95732857	1639,912525	248,2443299	7,656210875	33,57102356	0
46	2,338424288	1963,35156	1155,160504	150,0158298	18,29942138	315,9041781	644,3631421	71,30398977	3407,713581	304,4232741	10,30294472	39,23427979	0
47	1,298167246	1984,511373	1405,786822	130,9427979	21,8292399	288,9520988	161,0077185	74,68091326	2526,814256	228,8677196	8,946891121	72,08459409	0
48	2,134446144	1986,349053	809,2938035	95,08902184	19,36492572	205,499761	196,3576427	76,34020725	2459,524526	289,9571416	3,746624977	57,99777218	0
49	4,147966432	1991,789789	1250,198275	116,8564622	21,57326495	320,7401725	755,5005552	70,35546403	1795,719359	189,8833073	9,094363677	44,80160057	0

Таким образом, часть предоставленных данных (около 1000 строк) о начальных свойствах компонентов композиционных материалов была сгенерирована.

Сгенерированные данные в количестве 1000 строк были удалены, и оставлены только данные полученные экспериментальным путем: первые 23 строки объединённого датасета.

Выведем гистограммы распределений переменных из 23 строчек.

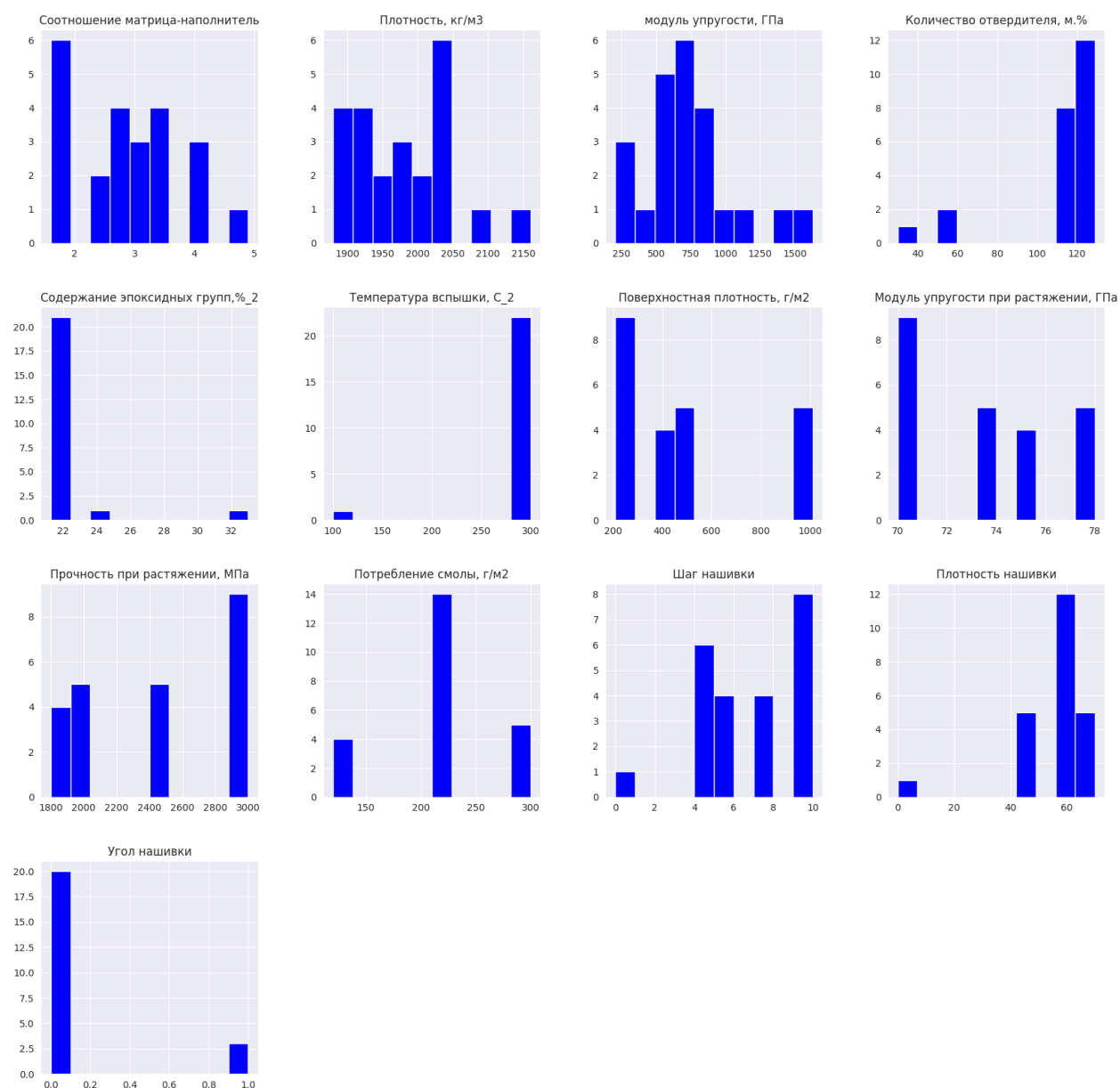


Рисунок 16 – Гистограммы распределений

Данные не подчинены закону нормального распределения.

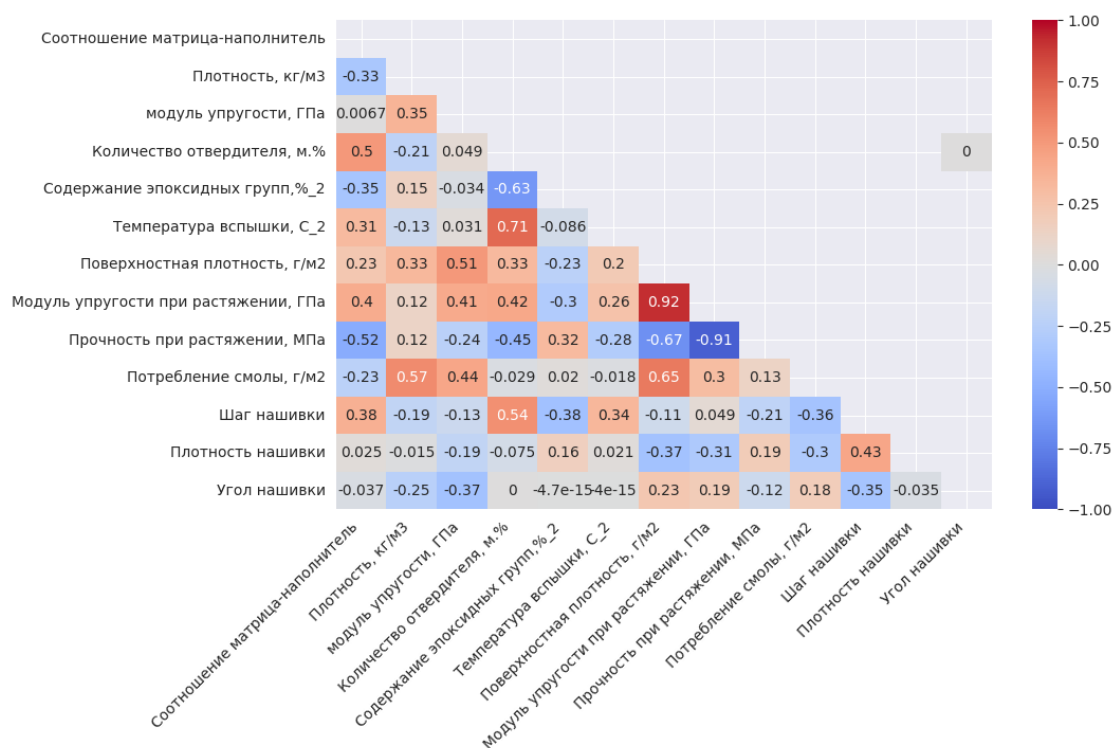


Рисунок 18 – Тепловая карта со значениями корреляции данных

Коэффициент корреляции показывает сильную связь (> 0.7) между:

—значениями "Модуль упругости при растяжении, ГПа" и "Поверхностная плотность, г/м2";

—значениями "Прочность при растяжении, МПа" и "Модуль упругости при растяжении, ГПа"

—значениями "Температура вспышки, С_2" и "Количество отвердителя, м.%".

Датасет с 23 строками начального датасета был нормализован, проведено повторное тестирование моделей.

Таблица 6 – Результаты оценки моделей

Искомое значение	Model	MSE	MAE	R2 score
Модуль упругости при растяжении	LinearRegression_n_upr	2.954099e-29	3.141931e-15	1.000
Прочность при растяжении	LinearRegression_n_pr	3.351072e-29	3.569367e-15	1.000
Прочность при растяжении	SGDRegressor_pr	7.001937e-03	7.352911e-02	0.946

Искомое значение	Model	MSE	MAE	R2 score
Модуль упругости при растяжении	SGDRegressor_upr	1.283970e-02	9.310538e-02	0.910
Модуль упругости при растяжении	RandomForestRegressor_upr	2.781571e-02	1.120357e-01	0.805
Прочность при растяжении	RandomForestRegressor_pr	3.123474e-02	1.457407e-01	0.759
Модуль упругости при растяжении	KNeighborsRegressor_upr	6.233333e-02	2.266667e-01	0.563
Прочность при растяжении	KNeighborsRegressor_pr	6.243004e-02	2.314815e-01	0.518

На основании модели линейной регрессии получен коэффициент детерминации равный 1.

2.5 Разработка приложения

Разработано приложение с графическим интерфейсом, которое будет определять значение: Соотношение матрица-наполнитель.

Рекомендации соотношения матрица-наполнитель для композиционных материалов

Введите параметры

<input type="text"/>	Плотность, кг/м3
<input type="text"/>	Модуль упругости, ГПа
<input type="text"/>	Количество отвердителя, м.%
<input type="text"/>	Содержание эпоксидных групп, %_2
<input type="text"/>	Температура вспышки, С_2
<input type="text"/>	Поверхностная плотность, г/м2
<input type="text"/>	Модуль упругости при растяжении, ГПа
<input type="text"/>	Прочность при растяжении, МПа
<input type="text"/>	Потребление смолы, г/м2
<input type="text"/>	Угол нашивки, град
<input type="text"/>	Шаг нашивки
<input type="text"/>	Плотность нашивки

Рисунок 19 – Графический интерфейс приложения

```
7 https://colab.research.google.com/drive/11a_Hn13KtwxKhk8-TBvH1XjJSi9urVPF
8
9 Создаем приложение с графическим интерфейсом для прогнозирования "соотношения матрица-наполнитель"
10 """
11
12 from flask import Flask, request, render_template
13
14 import tensorflow as tf
15
16 app = Flask(__name__)
17
18 def prediction(params):
19     model = tf.keras.models.load_model('models/mn_model_nn')
20     pred = model.predict([params])
21     return pred
22
23 @app.route('/', methods=['POST', 'GET'])
24 def predict():
25     message = ''
26     if request.method == 'POST':
27         param_list = ('Плотность, кг/м3', 'модуль упругости, ГПа', 'Количество отвердителя, м.%',
28                     'Содержание эпоксидных групп,%_2', 'Температура вспышки, C_2', 'Поверхностная плотность, г/м2',
29                     'Модуль упругости при растяжении, ГПа', 'Прочность при растяжении, МПа', 'Потребление смолы, г/м2',
30                     'Угол нашивки, град', 'Шаг нашивки', 'Плотность нашивки')
31         params = []
32         for i in param_list:
33             param = request.form.get(i)
34             params.append(param)
35             params = [float(i.replace(',', '.')) for i in params]
36
37         message = f'Соотношение матрица-наполнитель: {prediction(params)}'
38         return render_template('mainn.html', message=message)
39
40 if __name__ == '__main__':
41     app.run()
```

[Give feedback](#)

Рисунок 20 – Код приложения

Краткая инструкция использования приложения:

Для получения прогноза необходимо:

- а) запустить app.py,
- б) совершить запуск всех ячеек,
- в) в появившейся строке (* Running on <http://127.0.0.1:5000/> (Press CTRL+C to quit)) - нажать на ссылку: <http://127.0.0.1:5000/>.
- г) В новом открывшемся окне (сайте) ввести 12 входных параметров и нажать "Рассчитать".
- д) появится результат в виде числа с плавающей точкой.

2.6 Создание удаленного репозитория и загрузка работы на него

Репозиторий был создан на github.com по адресу:

<https://github.com/Sergey7OV/CompoZit>

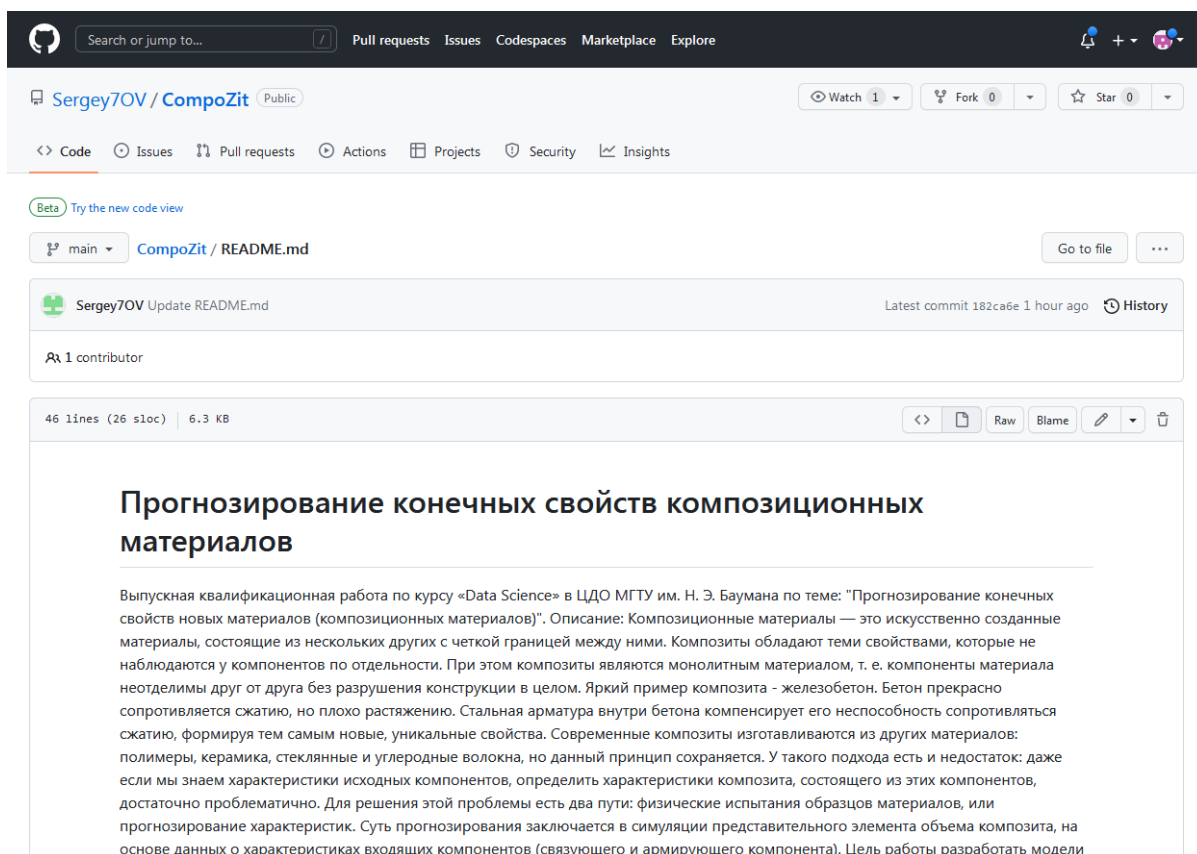


Рисунок 21 – Файл README в созданном репозитории

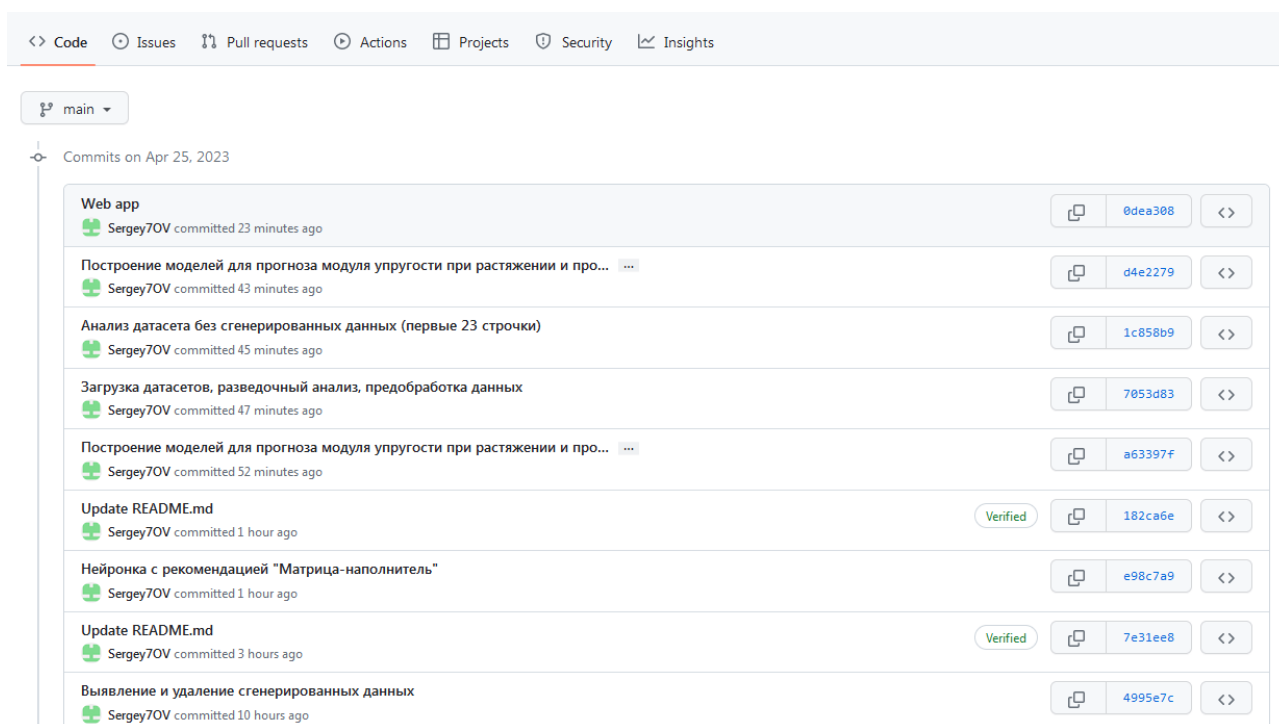


Рисунок 22 – Коммиты в репозитории.

Заключение

В процессе исследовательской работы были изучены теоретические основы и методы решения задачи регрессии с целью разработки модели для прогноза модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель».

Проведен разведочный анализ данных объединенного датасета, на основании которого сделан вывод, что распределение полученных данных близко к нормальному, корреляция между парами признаков близка к нулю.

Примененные методы решения задачи регрессии для объединенного и нормализованного датасета не привели к желаемому результату, коэффициент детерминации всех моделей меньше 0,5.

Исследовательская работа позволила сделать вывод, что часть предоставленных данных о начальных свойствах компонентов композиционных материалов была сгенерирована. Проведен анализ повторяемости данных в датасете, составлена тепловая карта и выявлена «зона» сгенерированных данных.

Сгенерированные данные в количестве 1000 строк были удалены и оставлены только данные полученные экспериментальным путем, что позволило на их основе провести повторное тестирование моделей. На основании модели линейной регрессии получен коэффициент детерминации равный 1.

Однако, очищенный от сгенерированных данных датасет содержит всего лишь 23 строки и его значения не подтверждены нормальному распределению. При этом выявлена высокая корреляция между значениями "Модуль упругости при растяжении, ГПа" и "Поверхностная плотность, г/м²"; значениями "Прочность при растяжении, МПа" и "Модуль упругости при растяжении, ГПа", а также значениями "Температура вспышки, С₂" и "Количество отвердителя, м.%".

В дальнейшем, для получения нормального распределения значений в датасете необходимы исследования для получения дополнительных экспериментальных значений начальных свойств компонентов композиционных материалов.

Библиографический список

- 1 Джоши, Прадик. Д42 Искусственный интеллект с примерами на Python. : Пер. с англ. - СПб. : ООО "Диалектика", 2019. - 448 с. - Парал. тит. англ.
- 2 Луис Педро Коэльо, Вилли Ричарт Построение систем машинного обучения на языке Python. 2-е издание / пер. с англ. Слинкин А. А. - М.: ДМК Пресс, 2016. - 302 с.: пл.
- 3 Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А. А. Слинкина. - М.: ДМК Пресс, 2015. - 400 с.: ил.
- 4 Андреас Мюллер, Сара Свило Введение в машинное обучение с помощью Python Руководство для специалистов по работе с данными
- 5 Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. - М.: ДМК Пресс, 2017.-418 с.: ил.
- 6 Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn.Mayavi. - devpractice.ru. 2020. - 412 с.: ил.
- 7 Абросимов Н.А Методика построения разрешающей системы уравнений динамического деформирования композитных элементов конструкций (Учебно-методическое пособие), ННГУ, 2010
- 8 Абу-Хасан Махмуд, Масленникова Л. Л. Прогнозирование свойств композиционных материалов с учётом наноразмера частиц и акцепторных свойств катионов твёрдых фаз, статья 2006 год
- 9 Гафаров, Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения: учеб. пособие /Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Издательство Казанского университета, 2018. – 121 с.
- 10 Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2021. - 416 с.: ил.9.
- 11 Документация по библиотеке keras: <https://keras.io/api/>.
- 12 Документация по библиотеке matplotlib: <https://matplotlib.org/stable/users/index.html>.

- 13 Документация по библиотеке numpy:
<https://numpy.org/doc/1.22/user/index.html#user>.
- 14 Документация по библиотеке pandas:
https://pandas.pydata.org/docs/user_guide/index.html#user-guide.
- 15 Документация по библиотеке scikit-learn: https://scikit-learn.org/stable/user_guide.html.
- 16 Документация по библиотеке seaborn:
<https://seaborn.pydata.org/tutorial.html>.
- 17 Документация по библиотеке Tensorflow:
<https://www.tensorflow.org/overview>
- 18 Документация по языку программирования python:
<https://docs.python.org/3.8/index.html>.
- 19 Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.
- 20 Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) : <https://habr.com/ru/post/428503/>
- 21 Ларин А.А., Способы оценки работоспособности изделий из композиционных материалов методом компьютерной томографии, Москва, 2013, 148 с. научно-техническая конференция «Полимерные композиционные материалы и производственные технологии нового поколения», 19 ноября 2021 г.
- 22 Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с
- 23 Реутов Ю.А.: Прогнозирование свойств полимерных композиционных материалов и оценка надёжности изделий из них, Диссертация на соискание учёной степени кандидата физико-математических наук, Томск 2016.
- 24 Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.: Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил
- 25 Линейная регрессия. Часть 1. <https://www.dmitrymakarov.ru/opt/mlr-04/?ysclid=lgvuwldfuj858351122>

