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 Бурчакова Анна Александровна

Москва, 2023

**Оглавление**



**Список иллюстраций**

**Список таблиц**

**Введение**

Передовые технологические разработки являются основой научного и промышленного прогресса человечества. Они отвечают за перспективы экономического развития государства, определяют его место на международных рынках, гарантируют индустриальный суверенитет, который, в конченом счете, отражается в политической независимости.

В соответствии с государственными программами Российской Федерации «Развитие промышленности и повышение ее конкурентоспособности» и «Научно-технологическое развитие Российской Федерации» Правительством Российской Федерации ставятся цели по:

- формированию в гражданских отраслях промышленности Российской Федерации глобально конкурентоспособного сектора с высоким экспортным потенциалом, обеспечивающего достижение национальных целей развития[1]; а также

- технологическому обновлению научной, научно-технической и инновационной (высокотехнологичной) деятельности[2].

Этим целям соответствует деятельность Центра НТИ «Цифровое материаловедение: новые материалы и вещества» – структурного подразделения МГТУ им. Н.Э. Баумана, созданного 28 декабря 2020 года для реализации цифрового подхода к «быстрому» и «сквозному» проектированию, разработке, испытанию и применению новых материалов и веществ. Центр НТИ формирует национальный банк данных и знаний по материалам и их «цифровым двойникам»[3].

[1] «Развитие промышленности и повышение ее конкурентоспособности» [Электронный ресурс] / Портал Госпрограмм РФ. – https://programs.gov.ru/Portal/program/16/passport / Дата обращения 25.04.2023

[2] «Научно-технологическое развитие Российской Федерации» [Электронный ресурс] / Портал Госпрограмм РФ. – https://programs.gov.ru/Portal/programs/passport/47. Дата обращения 25.04.2023

[3] Официальный сайт Центра НТИ «Цифровое материаловедение: новые материалы и вещества» [Электронный ресурс] . – https://nti.emtc.ru/ Дата обращения 25.04.2023

Целью настоящего исследования является решение актуальной производственной задачи Центра НТИ «Цифровое материаловедение: новые материалы и вещества»: прогнозирование свойств получаемых композиционных материалов.

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними.

Создание композиционных материалов – дорогостоящий, трудоемкий процесс, требующий участия высококвалифицированных кадров. В этой связи, актуальность настоящего исследования обусловлена тем, что наличие прогнозных моделей позволяет сократить количество реально проводимых испытаний, а также дополнить базу данных материалов их новыми характеристиками и цифровыми двойниками новых композитов.

В исследовании использованы данные о начальных свойствах компонентов композиционных материалов (поверхностная плотность, г/м2; потребление смолы, г/м2 и т.д.).

Основные задачи исследования:

1. изучение теоретических основ и методов прогнозирования целевых переменных, в частности свойств получаемых композиционных материалов;
2. проведение разведочного анализа данных;
3. обучение моделей для прогноза целевых признаков: модуля упругости при растяжении и прочности при растяжении;
4. написание нейронной сети, которая будет рекомендовать соотношение матрица-наполнитель;
5. разработка приложения, которое будет выдавать прогноз, соотношения матрица-наполнитель.

Результаты исследования опубликованы в созданном на веб-портале GitHub репозитории и находятся в открытом доступе[4].

[4] Веб-портал GitHub [Электронный ресурс]. – https://github.com/aburchakova/kompozitus

# Глава 1. Аналитическая часть

## 1.1. Постановка задачи

Реализация цифрового подхода к проектированию, разработке, испытанию и применению новых материалов и веществ является важным элементом создания национального банка данных и знаний по материалам.

В настоящей работе исследуются данные о начальных свойствах компонентов композиционных материалов. Таким образом, **объектом** исследования являются композиционные материалы, которые означают искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними.

Основные **задачи** исследования включают:

1. изучение теоретических основ и методов прогнозирования целевых переменных, в частности свойств получаемых композиционных материалов;
2. проведение разведочного анализа данных;
3. обучение моделей для прогноза целевых признаков: модуля упругости при растяжении и прочности при растяжении;
4. написание нейронной сети, которая будет рекомендовать соотношение матрица-наполнитель;
5. разработать приложение, которое будет выдавать прогноз, соотношения матрица-наполнитель.

**Актуальность** поставленных задач обусловлена возможностью использования цифрового подхода к процессу создания композиционных материалов для снижения количества реально проводимых испытаний.

Для настоящего исследования использованы данные по тринадцати начальным свойствам компонентов композиционных материалов, хранящимся в двух таблицах: x_bp и x_nup[5].

[5] https://drive.google.com/file/d/1B1s5gBlvgU81H9GGolLQVw_SOi-vyNf2/view?usp=sharing

Перечислим имеющиеся данные о свойствах композитов из таблицы x_bp (далее – датасет x_bp):

1. соотношение матрица-наполнитель;
2. плотность, кг/м$^3$;
3. модуль упругости, Гпа;
4. количество отвердителя, м.%;
5. содержание эпоксидных групп, %_2;
6. температура вспышки, С_2;
7. поверхностная плотность, г/м$^2$;
8. модуль упругости при растяжении, Гпа;
9. прочность при растяжении, Мпа;
10. потребление смолы, г/м$^2$.

Перечислим имеющиеся данные о свойствах композитов из таблицы x_ nup (далее – датасет x_ nup):

1. угол нашивки, град;
2. шаг нашивки;
3. плотность нашивки.

**Предметом** исследования являются прогнозные данные трех свойств композитов: соотношение матрица-наполнитель, модуль упругости при растяжении, прочность при растяжении.

В датасете x_bp, представляющей собой класс DataFrame, содержатся 1023 строки и 10 столбцов. Отсутствуют пропущенные значения, тип данных – float64 (рисунок 1).

Числа с десятичной точкой (тип данных float) называются числами с плавающей точкой или вещественными числами[6].

[6] Свейгарт Эл. Автоматизация рутинных задач с помощью Python [Текст] / ООО «Дилектика». – г. Санкт-Петербург. – 2021 г. – 672 с.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1023 entries, 0 to 1022
Data columns (total 10 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, С_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
dtypes: float64(10)
memory usage: 87.9 KB
```

Рисунок 1 – Информация по датасету x_bp, выведенная с помощью метода info()

Получаем описание датасета x_bp с использованием статистических параметров (рисунок 2). Узнаем, что:

– во всех столбцах количество значений соответствует длине столбцов и составляет 1023;

– средние значения между столбцами несопоставимы;

– видим стандартное квадратическое отклонение, пороговые значения для 25, 50 и 75% квантилей, минимальные и максимальные значения данных по столбцам датасета.

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 1023.000000 | 1023.000000 | 1023.000000 | 1023.000000 | 1023.000000 | 1023.000000 | 1023.000000 | 1023.000000 | 1023.000000 | 1023.000000 |
| mean | 2.930366 | 1975.734888 | 739.923233 | 110.570769 | 22.244390 | 285.882151 | 482.731833 | 73.328571 | 2466.922843 | 218.423144 |
| std | 0.913222 | 73.729231 | 330.231581 | 28.295911 | 2.406301 | 40.943260 | 281.314690 | 3.118983 | 485.628006 | 59.735931 |
| min | 0.389403 | 1731.764635 | 2.436909 | 17.740275 | 14.254985 | 100.000000 | 0.603740 | 64.054061 | 1036.856605 | 33.803026 |
| 25% | 2.317887 | 1924.155467 | 500.047452 | 92.443497 | 20.608034 | 259.066528 | 266.816645 | 71.245018 | 2135.850448 | 179.627520 |
| 50% | 2.906878 | 1977.621657 | 739.664328 | 110.564840 | 22.230744 | 285.896812 | 451.864365 | 73.268805 | 2459.524526 | 219.198882 |
| 75% | 3.552660 | 2021.374375 | 961.812526 | 129.730366 | 23.961934 | 313.002106 | 693.225017 | 75.356612 | 2767.193119 | 257.481724 |
| max | 5.591742 | 2207.773481 | 1911.536477 | 198.953207 | 33.000000 | 413.273418 | 1399.542362 | 82.682051 | 3848.436732 | 414.590628 |

Рисунок 2 – Статистические характеристики датасета x_bp, выведенные с помощью метода describe()

В таблице x_nup, представляющей собой класс DataFrame, содержатся 1040 строк и 3 столбца. Отсутствуют пропущенные значения, присутствуют два типа данных: int64 и float64 (рисунок 3).

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1040 entries, 0 to 1039
Data columns (total 3 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Угол нашивки, град  1040 non-null   int64  
 1   Шаг нашивки         1040 non-null   float64
 2   Плотность нашивки   1040 non-null   float64
dtypes: float64(2), int64(1)
memory usage: 32.5 KB
```

Рисунок 3 – Информация по датасету x_nup, выведенная с помощью метода info()

Посмотрим, какие уникальные значения хранятся в столбце «Угол нашивки, град», который содержит данные типа int64. Для этого воспользуемся методом unique(). Результатом будет массив array([ 0, 90], dtype=int64). Таким образом, угол нашивки принимает два значения: 0 и 90 градусов.

Получим описание датасета x_ nup с использованием статистических параметров (рисунок 4). Узнаем, что:

- во всех столбцах количество значений соответствует длине столбцов и составляет 1040;
- средние значения между столбцами несопоставимы;
- видим стандартное квадратическое отклонение, пороговые значения для 25, 50 и 75% квантилей, минимальные и максимальные значения данных по столбцам датасета.

| | Угол нашивки, град | Шаг нашивки | Плотность нашивки |
|---|---|---|---|
| count | 1040.00000 | 1040.000000 | 1040.000000 |
| mean | 45.00000 | 6.911385 | 57.248399 |
| std | 45.02165 | 2.555181 | 12.332438 |
| min | 0.00000 | 0.000000 | 0.000000 |
| 25% | 0.00000 | 5.102256 | 49.970740 |
| 50% | 45.00000 | 6.938000 | 57.413594 |
| 75% | 90.00000 | 8.587662 | 65.107235 |
| max | 90.00000 | 14.440522 | 103.988901 |

Рисунок 4 – Информация по таблице x_ nup, выведенная с помощью метода describe ()

Для целей настоящего исследования объединим два датасета и сформируем единый датасет. В исходных датасетах содержится разное количество строк, разница в 17 строк соответствует 1,6% данных большего датасета x_ nup. Исключим из дальнейшего исследования эти данные и сформируем новый датасет joinned_dataset с помощью метода inner join ().

Выводим информацию по joined_dataset (риснок 5). Узнаем, что:

- класс таблицы – DataFrame;
- в таблице 1023 строки и 13 столбцов;
- в таблице нет пропущенных значений;
- в таблице 2 типа данных: int64, float64.

```
<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, С_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
```

Рисунок 5 – Информация по датасету joined_dataset, выведенная с помощью метода info()

Таким образом, на первом этапе настоящего исследования составлен датасет joined_dataset с данными о свойствах композиционных материалов, представленный классом DataFrame, хранящий 1023 строки и 13 столбцов. В датасете нет пропущенных значений, содержатся 2 типа данных: int64 и float64.

Датасет joined_dataset содержит как целевые переменные, так и признаки, используемые для прогнозирования значений целевых переменных.

Для целевых переменных «модуль упругости при растяжении» и «прочность при растяжении» входными данными будут следующие одиннадцать 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. потребление смолы, г/м2;
9. угол нашивки, град;
10. шаг нашивки;
11. плотность нашивки

Для целевой переменной «соотношение матрица-наполнитель «входными данными будут следующие двенадцать переменных:

1. модуль упругости при растяжении;
2. прочность при растяжении;
3. плотность, кг/м3;
4. модуль упругости, Гпа;
5. количество отвердителя, м.%;

6. содержание эпоксидных групп, %_2;
7. температура вспышки, С_2;
8. поверхностная плотность, г/м2;
9. потребление смолы, г/м2;
10. угол нашивки, град;
11. шаг нашивки;
12. плотность нашивки.

## 1.2. Описание используемых методов

В соответствии с поставленными задачами по прогнозированию свойств композитов (модуля упругости при растяжении, прочности при растяжении, соотношение матрица-наполнитель) в настоящем исследовании решается задача регрессии.

Воспользуемся следующим определением регрессионного анализа: «Регрессионный анализ заключается в определении аналитического выражения связи, в котором изменение одной величины (называемой зависимой или результативным признаком) обусловлено влиянием одной или нескольких независимых величин (факторов), а множество всех прочих факторов, также оказывающих влияние на зависимую величину, принимается за постоянные и средние значения»[7].

Будем использовать метод множественной (многофакторной) регрессии, который определяется как «изучение связи между тремя и более связанными между собой признаками»[8].

Ниже приводится описание достоинств, недостатков и области применения методов регрессионного анализа, использованных в настоящем исследовании: линейная, метод лассо, дерево решений, метод ближайших соседей.

**Линейная регрессия** выражается уравнениями прямой (линейной функцией), в случае для парной регрессии и задается следующей формулой:

$$Y = b_0 + b_1 X. \quad (1)$$

Модель множественной линейной регрессии позволяет находить и изучать зависимости переменной Y от нескольких объясняющих переменных $X_1$, $X_2$, $X_3$, $X_4$, $X_5$. Множественная линейная регрессия представлена следующим образом:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + e. \quad (2)$$

[7] Теория статистики. Учебник. – г. Москва. – «Финансы и статистика». – 1998 г.
[8] Там же.

В данном уравнении Y – зависимая переменная, описывающая процесс, который планируется предсказать. X – независимые переменные, используемые для моделирования или прогнозирования значений зависимых переменных.

Независимые переменные в регрессионных моделях называют регрессорами[9].

b – коэффициенты уравнения линейной множественной регрессии, которые рассчитываются в результате регрессионного анализа для каждой независимой переменной X, они представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

e – случайные ошибки.

Построение уравнения линейной регрессии сводится к оценке ее параметров с использованием, например, метода наименьших квадратов (МНК).

МНК позволяет получить такие коэффициенты уравнения линейной множественной регрессии, при которых сумма квадратов отклонений фактических значений зависимой переменной Y от теоретических минимальна.

Метод линейной регрессии – наиболее простой метод регрессионного анализа, характеризуется простотой вычислительных алгоритмов и наглядностью и интерпретируемостью результатов.

Недостатком линейной регрессии является то, что данный метод упрощает реальные задачи, предполагая линейную взаимосвязь между переменными.

**Метод регрессии Лассо** (LASSO, Least Absolute Shrinkage and Selection Operator) – вариация линейной регрессии. Лассо-регрессия

[9] Бабешко Л.О. Основы эконометрического моделирования: Учебное пособие. [Текст] / КомКнига . – г. Москва. – 2006 г. – 432 с.

минимизирует сумму квадратов остатков (RSS, Residual Sum of Squares) вместе с некоторым штрафным ограничением.

Область применения Лассо – датасеты с высокой коррелированностью независимых переменных (мультиколлинеарность).

Лассо также называют методом регуляризации, заключающемся в наложении дополнительных ограничений на искомые параметры, которые могут предотвратить излишнюю сложность модели. Метод заключается во введении дополнительного слагаемого регуляризации в функционал оптимизации модели, что часто позволяет получать более устойчивое решение.

Недостаток регрессии Лассо заключается в том, что становится трудно интерпретировать коэффициенты в окончательной модели, поскольку они сжимаются до нуля.

**Деревья решений** относятся к древовидным моделям классификации и регрессии (CART, classification and regression trees).

Древовидная модель – это набор правил импликации вида «если-то-иначе». В модели деревьев решений используется древесная структура для представления ряда возможных путей принятия решения и результата для каждого пути[10].

В задачах регрессии деревья решений работают путем определения соответствующего выходного значения в соответствии с вектором признаков.

Преимущество модели заключается в наличии визуального инструмента обследования данных для представления о том, какие переменные важны и как они друг с другом связаны. Кроме того, древовидные модели обеспечивают набор правил и могут быть эффективно использованы неспециалистами.

[10]Грас Д. Data Science. Наука о данных с нуля [Текст] / БХВ-Петербург. – г. Санкт-Петербург. – 2021 г. – 416 с

**Метод К ближайших соседей** (KNN, k-nearest neighbors) состоит, во-первых, в поиске K записей, которые имеют схожие значения независимых переменных, во-вторых, в поиске среди этих схожих записей среднего и предсказывания этого среднего для новой записи.

Сосед (neighbor) – запись, чьи предикторные значения схожи с другой записью[11].

Преимуществом метода К ближайших соседей является простота модели и ее интуитивная понятность.

Метрические показатели расстояния (distance metrics) – метрические показатели, которые обобщают в одном числе, насколько далеко одна запись находится от другой[12]. Самым популярным метрическим показателем расстояния между двумя векторами является евклидово расстояние.

K – число соседей, учитываемых при вычислении алгоритма ближайших соседей. Определяется K тем, насколько хорошую результативность алгоритм показывает на тренировочных данных с использованием разных значений K.

**Нейросетевая регрессия** создает модели регрессии с помощью алгоритма нейронной сети.

Искусственные нейронные сети (artificial neural networks, ANN) – упрощенные модели биологических нейронных сетей мозга человека[13]. Упрощенная математическая модель нейрона состоит из следующих элементов (рисунок 6):

- входные параметры $x_1$, $x_2$, $x_3$, …, $x_n$, имеющие свои веса – $w_1$, $w_2$, $w_3$, …, $w_n$;

[11] Брюс П., Брюс Э. Практическая статистика для специалистов Data Science [Текст] / «БХВ-Петербург» – г. Санкт-Петербург. – 2020 г. – 304 с.

[12] Там же.

[13] Постолит, А. Основы искусственного интеллекта в примерах на Python [Текст] / «БХВ-Петербург» – г. Санкт-Петербург. – 2022 г. – 448 с.

– сумматор, где каждый входной коэффициент умножается на некоторый действительный весовой коэффициент и формируется итоговая сумма;
– функция активации – нелинейное преобразование, поэлементно применяющееся к пришедшим на вход данным;
– на выходе осуществляется проверка значения функции активации.

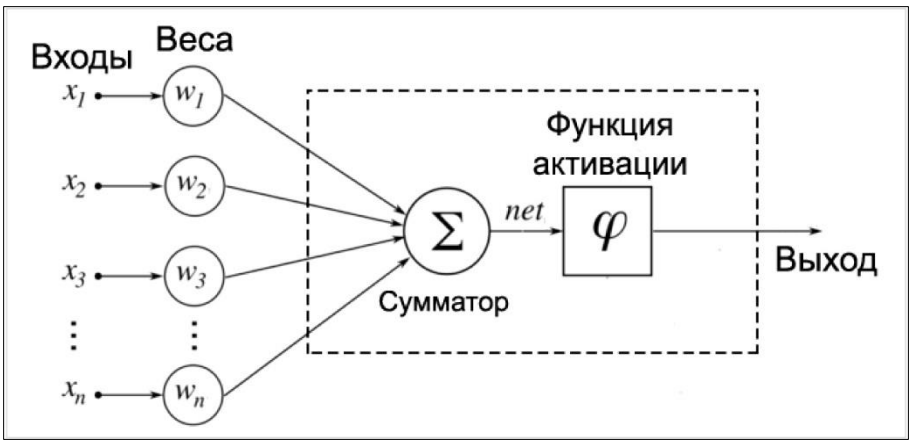


Рисунок 6 – Математическая модель искусственного нейрона

Источник: Основы искусственного интеллекта в примерах на Python. – А.Постолит. – г. Санкт-Петербург. – «БХВ-Петербург». – 2022 г.

Обучение нейронной сети – это поиск такого набора весовых коэффициентов, при котором входной сигнал после прохода по сети преобразуется в нужный выходной.

Обучающая выборка (training set) – это набор входных сигналов вместе с правильными выходными сигналами, по которым происходит обучение сети.

Тестовые выборка (testing set) – это набор входных сигналов вместе с правильными выходными сигналами, по которым происходит оценка качества работы сети после обучения на обучающей выборке.

Таким образом, для обучения моделей прогноза модуля упругости при растяжении и прочности при растяжении будут использованы

рассмотренные выше методы регрессионного анализа: линейная, метод лассо, дерево решений, метод ближайших соседей. Также будет написана нейронная сеть, которая будет рекомендовать соотношение матрица-наполнитель.

# Глава 2. Практическая часть

## 2.1. Предобработка данных

Предварительная обработка данных является важнейшим этапом исследования, позволяющим подготовить датасет, который будет в дальнейшем использоваться при построении моделей прогнозирования значения переменных.

Как уже было определено в *Разделе 1.1. Постановка задачи* решение задач исследования базируется на работе с датасетом, содержащим данные о свойствах композиционных материалов со следующими характеристиками:

- класс DataFrame;
- содержит 1023 строки и 13 столбцов;
- отсутствуют пропущенные значения;
- содержатся 2 типа данных: int64 и float64.

Датасет (joined_dataset) содержит как целевые переменные, так и признаки, используемые для прогнозирования значений целевых переменных.

### 2.1.1. Гистограммы, диаграммы ящика с усами, попарные графики рассеяния точек

В первую очередь построим графики рассеяния исследуемого датасета (рисунок 7). На графиках отображена взаимосвязь между переменными в исследуемом наборе данных. По диагонали на пересечении одного и того же признака отображена его гистограмма, в местах пересечения разных признаков отображена точечная диаграмма.

На графике ниже видно, что видимая попарная зависимость признаков отсутствует.

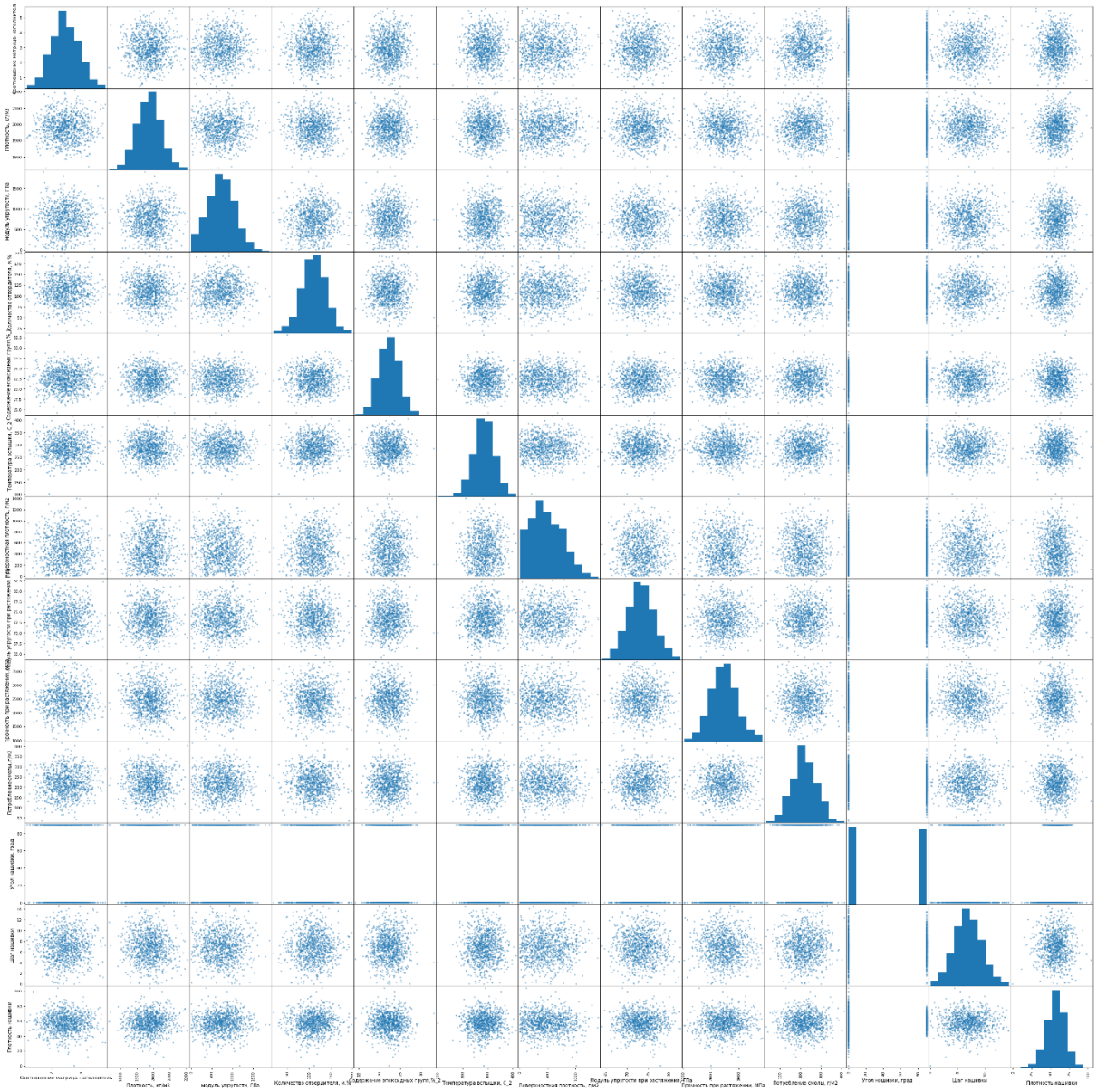

Рисунок 7 – Попарные графики рассеяния точек датасета joined_dataset

Далее построим диаграмму "ящик с усами", чтобы выявить наличие в выборке выбросов – данных, не подпадающих под общее распределение.

На рисунках 8-9 видно, что выбросы есть у всех параметров, кроме угла нашивки, который принимает только два значения.

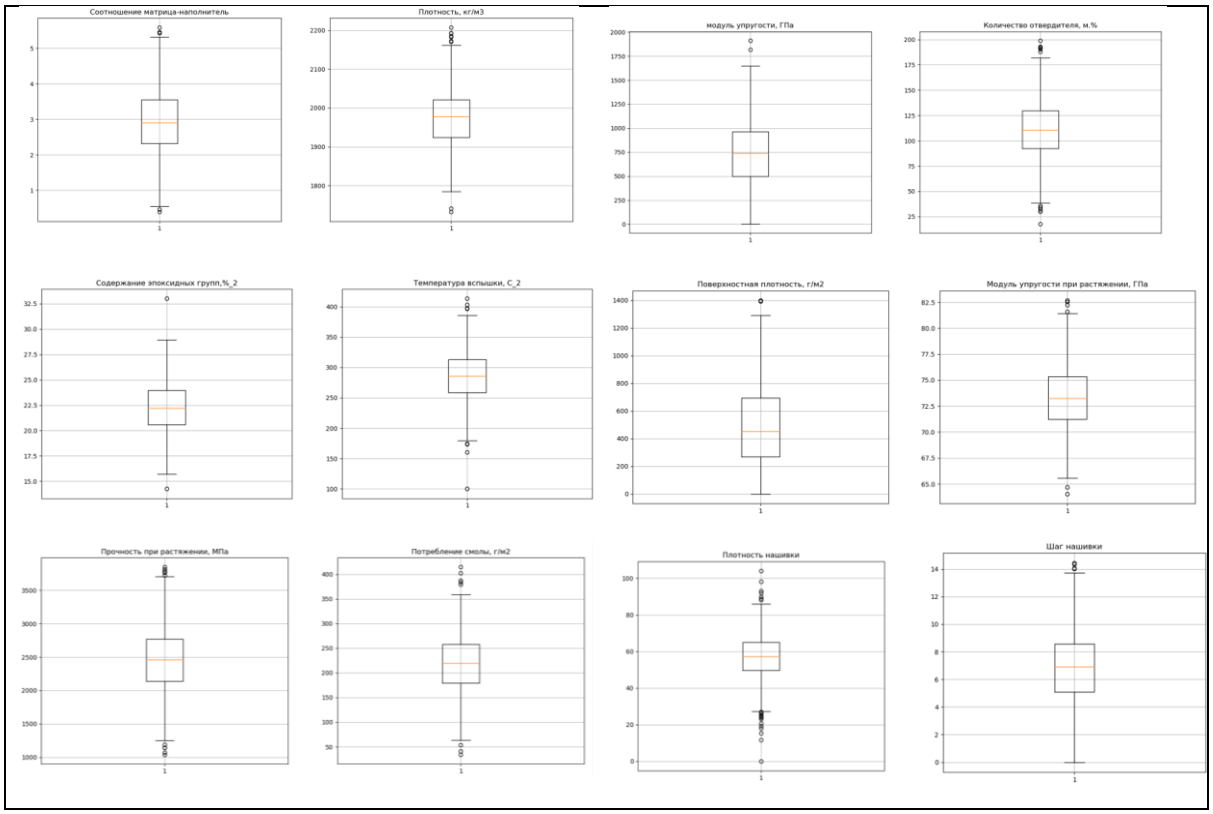


Рисунок 8 – Диаграммы «ящик с усами» признаков исходного датасета joined_dataset (выбросы обнаружены)

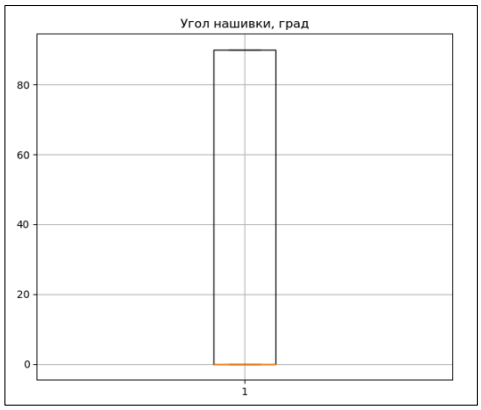


Рисунок 9 – Диаграмма «ящик с усами» признаков исходного датасета joined_dataset (выбросы не обнаружены)

Далее построим гистограммы распределения признаков датасета joined_dataset (рисунки 10-11).

Гистограммой называется ступенчатая фигура, состоящая из прямоугольников, основанием которых служат частичные интервалы длиной h, а высоты равны $w_j/h$[14].

На графиках видно, что распределение признаков имеет колоколообразный вид, что свидетельствует о нормальном распределении данных. Признак «поверхностная плотность, г/м2» смещенное влево распределение, говорит о несимметричности распределения. Количество значений слева от медианы больше, чем справа от медианы.

Признак «угол нашивки» имеет отличную от других признаков гистограмму, поскольку принимает только два значения: 0 и 90 градусов.

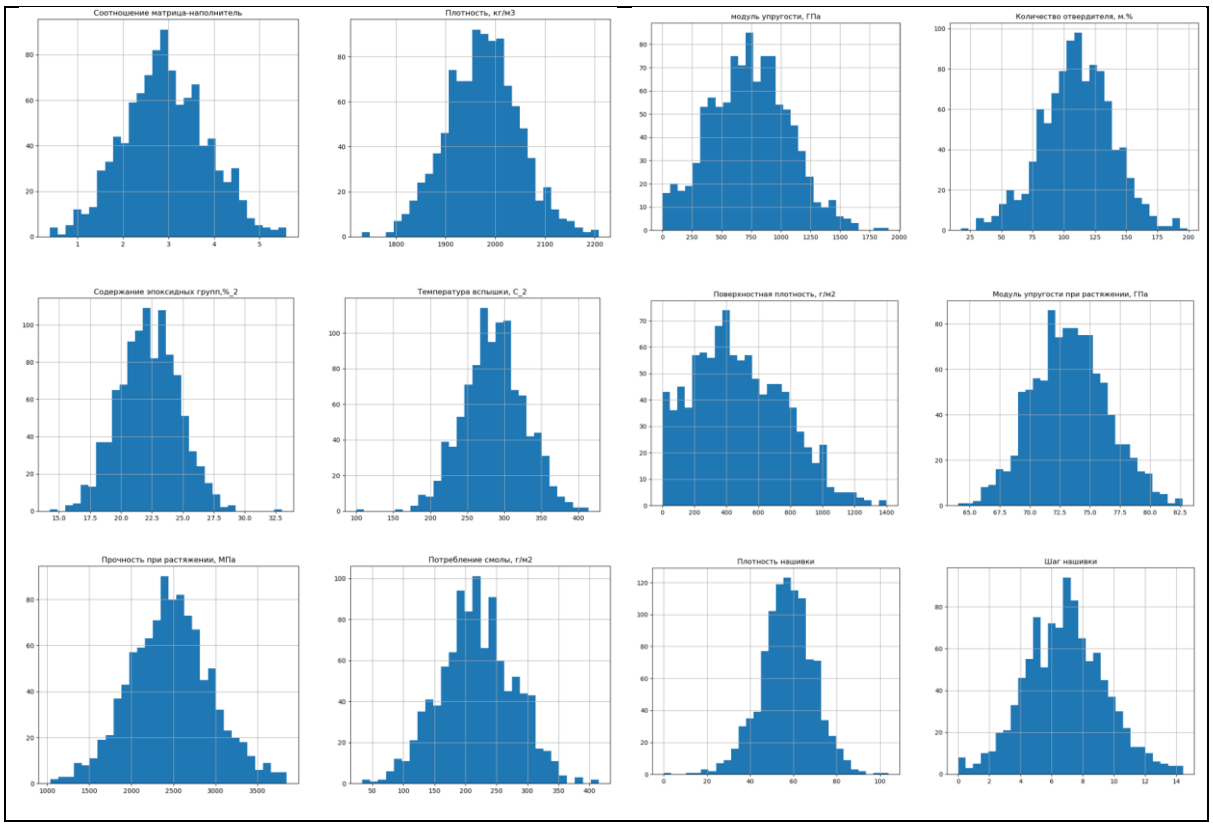


Рисунок 10 – Гистограммы признаков исходного датасета joined_dataset

[14] Фадеева Л.Н., Жуков Ю.В., Лебедев А.В. Математика для экономистов: Теория вероятностей и математическая статистика [Текст] / Эксмо. – г. Москва . – 2006 г. – 336 с.

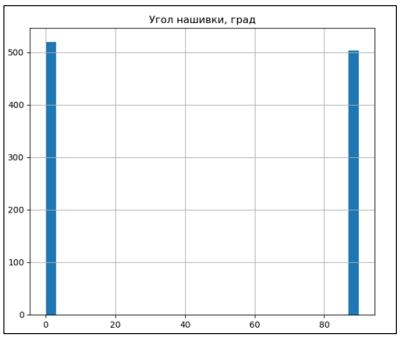


Рисунок 11 – Гистрограмма дискретного признака «угол нашивки» исходного датасета joined_dataset

Рассмотрим подробнее статистические характеристики датасета. Для этого воспользуемся методом describe(), который для каждой переменной датасета покажет количество значений (count), среднее значение (mean), стандартное (среднеквадратичное) отклонение (std), максимальные и минимальные значения (max, max), пороговые значения для 25, 50 и 75% квантилей (таблица 1).

Таблица 1 – Статистические характеристики исходного датасета joined_dataset, выведенные с помощью метода describe ()

| Свойства композитов | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Соотношение матрица-наполнитель | 1023 | 2,9 | 0,9 | 0,4 | 2,3 | 2,9 | 3,6 | 5,6 |
| Плотность, кг/м3 | 1023 | 1 975,7 | 73,7 | 1 731,8 | 1 924,2 | 1 977,6 | 2 021,4 | 2 207,8 |
| Модуль упругости, ГПа | 1023 | 739,9 | 330,2 | 2,4 | 500,0 | 739,7 | 961,8 | 1 911,5 |
| Количество отвердителя, м.% | 1023 | 110,6 | 28,3 | 17,7 | 92,4 | 110,6 | 129,7 | 199,0 |
| Содержание эпоксидных групп, %_2 | 1023 | 22,2 | 2,4 | 14,3 | 20,6 | 22,2 | 24,0 | 33,0 |
| Температура вспышки, С_2 | 1023 | 285,9 | 40,9 | 100,0 | 259,1 | 285,9 | 313,0 | 413,3 |
| Поверхностная плотность, г/м2 | 1023 | 482,7 | 281,3 | 0,6 | 266,8 | 451,9 | 693,2 | 1 399,5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Модуль упругости при растяжении, ГПа | 1023 | 73,3 | 3,1 | 64,1 | 71,2 | 73,3 | 75,4 | 82,7 |
| Прочность при растяжении, МПа | 1023 | 2 466,9 | 485,6 | 1 036,9 | 2 135,9 | 2 459,5 | 2 767,2 | 3 848,4 |
| Потребление смолы, г/м2 | 1023 | 218,4 | 59,7 | 33,8 | 179,6 | 219,2 | 257,5 | 414,6 |
| Угол нашивки, град | 1023 | 44,3 | 45,0 | 0,0 | 0,0 | 0,0 | 90,0 | 90,0 |
| Шаг нашивки | 1023 | 6,9 | 2,6 | 0,0 | 5,1 | 6,9 | 8,6 | 14,4 |
| Плотность нашивки | 1023 | 57,2 | 12,4 | 0,0 | 49,8 | 57,3 | 64,9 | 104,0 |

Исследуем зависимость (корреляцию) свойств композитов. Для этого построим матрицу корреляций, содержащую коэффициенты корреляции между всеми парами переменных, используемых в анализе.

Таблица 2 – Матрица корреляции свойств композитов исходного датасета joined_dataset, выведенная с помощью метода corr ()

| | 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 | Угол нашивки, град | Шаг нашивки | Плотность нашивки |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Соотношение матрица-наполнитель | 1 | 0,004 | 0,032 | -0,006 | 0,020 | -0,005 | -0,006 | -0,008 | 0,024 | 0,073 | -0,031 | 0,036 | -0,005 |
| Плотность, кг/м3 | 0,004 | 1 | -0,010 | -0,036 | -0,008 | -0,021 | 0,045 | -0,018 | -0,070 | -0,016 | -0,068 | -0,061 | 0,080 |
| Модуль упругости, ГПа | 0,032 | -0,010 | 1 | 0,024 | -0,007 | 0,031 | -0,005 | 0,023 | 0,042 | 0,002 | -0,025 | -0,010 | 0,056 |
| Количество отвердителя, м.% | -0,006 | -0,036 | 0,024 | 1 | -0,001 | 0,095 | 0,055 | -0,066 | -0,075 | 0,007 | 0,039 | 0,015 | 0,017 |
| Содержание эпоксидных групп,%_2 | 0,020 | -0,008 | -0,007 | -0,001 | 1 | -0,010 | -0,013 | 0,057 | -0,024 | 0,015 | 0,008 | 0,003 | -0,039 |
| Температура вспышки, С_2 | -0,005 | -0,021 | 0,031 | 0,095 | -0,010 | 1 | 0,020 | 0,028 | -0,032 | 0,060 | 0,021 | 0,026 | 0,011 |
| Поверхностная плотность, г/м2 | -0,006 | 0,045 | -0,005 | 0,055 | -0,013 | 0,020 | 1 | 0,037 | -0,003 | 0,016 | 0,052 | 0,038 | -0,050 |
| Модуль упругости при растяжении, ГПа | -0,008 | -0,018 | 0,023 | -0,066 | 0,057 | 0,028 | 0,037 | 1 | -0,009 | 0,051 | 0,023 | -0,029 | 0,006 |
| Прочность при растяжении, МПа | 0,024 | -0,070 | 0,042 | -0,075 | -0,024 | -0,032 | -0,003 | -0,009 | 1 | 0,029 | 0,023 | -0,060 | 0,020 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Потребление смолы, г/м2 | 0,073 | -0,016 | 0,002 | 0,007 | 0,015 | 0,060 | 0,016 | 0,051 | 0,029 | 1 | -0,015 | 0,013 | 0,012 |
| Угол нашивки, град | -0,031 | -0,068 | -0,025 | 0,039 | 0,008 | 0,021 | 0,052 | 0,023 | 0,023 | -0,015 | 1 | 0,024 | 0,108 |
| Шаг нашивки | 0,036 | -0,061 | -0,010 | 0,015 | 0,003 | 0,026 | 0,038 | -0,029 | -0,060 | 0,013 | 0,024 | 1 | 0,003 |
| Плотность нашивки | -0,005 | 0,080 | 0,056 | 0,017 | -0,039 | 0,011 | -0,050 | 0,006 | 0,020 | 0,012 | 0,108 | 0,003 | 1 |

Дополнительно проиллюстрируем корреляцию признаков с помощью «тепловой карты корреляции» (рисунок 12). На рисунке видно, что зависимость между признаками минимальна.

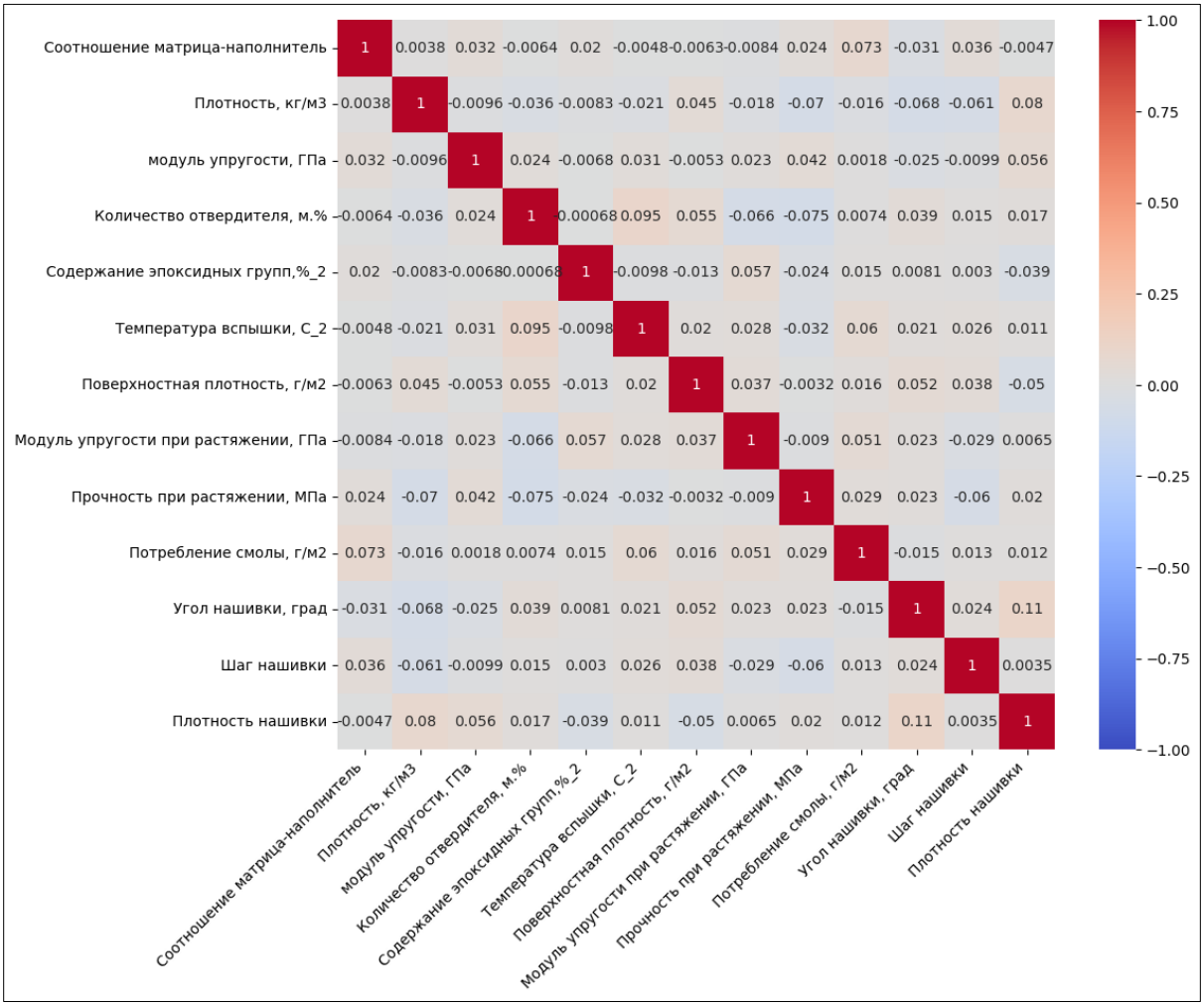


Рисунок 12 – Корреляционная тепловая карта исходного датасета joined_dataset

### 2.1.2. Удаление выбросов

В настоящем исследовании при подготовке датасета для построения моделей удаление выбросов проводилось несколько раз. Окончательный

выбор был сделан в пользу выборки, для которой был вычислен наибольший коэффициент детерминации (R2).

Модели создавались для следующих датасетов:

- датасет, в котором выбросы удалены по 10%-м нижним и верхним квартилям;
- датасет, в котором выбросы удалены вручную для трех признаков;
- датасет, в котором удалены первые 40 строк исходного датасета;
- датасет, в котором удалены первые 40 строк исходного датасета, а также вручную удалены выбросы для трех признаков.

Выведем таблицу с пороговыми значениями 10%-х квантилей для переменных датасета joined_dataset (таблица 3).

Таблица 3 – Пороговые значения десятипроцентных квантилей для переменных датасета joined_dataset

| | 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 | Угол нашивки, град | Шаг нашивки | Плотность нашивки |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,4 | 1731,8 | 2,4 | 17,7 | 14,3 | 100,0 | 0,6 | 64,1 | 1036,9 | 33,8 | 0,0 | 0,0 | 0,0 |
| 0,1 | 1,7 | 1881,4 | 321,1 | 75,9 | 19,1 | 233,2 | 117,7 | 69,3 | 1859,9 | 140,5 | 0,0 | 3,7 | 41,4 |
| 0,2 | 2,2 | 1912,9 | 442,6 | 87,2 | 20,2 | 251,5 | 219,7 | 70,6 | 2055,6 | 169,8 | 0,0 | 4,8 | 47,6 |
| 0,3 | 2,5 | 1936,1 | 563,7 | 96,8 | 21,0 | 265,6 | 311,1 | 71,7 | 2203,6 | 188,2 | 0,0 | 5,5 | 51,3 |
| 0,4 | 2,7 | 1959,6 | 652,5 | 104,4 | 21,6 | 275,4 | 379,6 | 72,4 | 2354,8 | 202,3 | 0,0 | 6,3 | 54,6 |
| 0,5 | 2,9 | 1977,6 | 739,7 | 110,6 | 22,2 | 285,9 | 451,9 | 73,3 | 2459,5 | 219,2 | 0,0 | 6,9 | 57,3 |
| 0,6 | 3,1 | 1993,8 | 830,3 | 118,2 | 22,9 | 295,8 | 536,1 | 74,1 | 2578,8 | 233,6 | 90,0 | 7,5 | 60,1 |
| 0,7 | 3,4 | 2013,0 | 918,4 | 125,9 | 23,6 | 305,7 | 637,6 | 74,9 | 2707,7 | 248,3 | 90,0 | 8,1 | 63,5 |
| 0,8 | 3,7 | 2035,8 | 1022,0 | 133,9 | 24,3 | 320,3 | 734,6 | 76,0 | 2852,0 | 271,1 | 90,0 | 9,0 | 66,9 |
| 0,9 | 4,1 | 2067,1 | 1157,0 | 146,2 | 25,3 | 340,7 | 865,1 | 77,3 | 3088,8 | 299,2 | 90,0 | 10,2 | 72,1 |
| 1 | 5,6 | 2207,8 | 1911,5 | 199,0 | 33,0 | 413,3 | 1399,5 | 82,7 | 3848,4 | 414,6 | 90,0 | 14,4 | 104,0 |

Исходная выборка является недостаточно многочисленной, состоит из 1023 строк и 13 колонок, не содержит пустых значений. Подробное изучение данных позволяет сделать выод о наличии в выборке синтетических данных, которые выделяются из всего датасета отутствием

значений после запятой (рисунок 13). Проделанная автором работа по созданию моделей прогнозирования переменных позволяет сделать выод о необходимости удаления первых сорока строк датасета для формирования датасета, для которого в дальнейшем будут получены наилучшие характеристики достоверности моделей прогноза.

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 1,598173516 | 1950 | 827 | 129 | 21,25 | 300 | 470 | 73,33333333 | 2455,555556 | 220 | 0 | 9 | 57 |
| 13 | 2,919677836 | 1980 | 568 | 129 | 21,25 | 300 | 470 | 73,33333333 | 2455,555556 | 220 | 0 | 9 | 60 |
| 14 | 4,029126214 | 1910 | 800 | 129 | 21,25 | 300 | 470 | 73,33333333 | 2455,555556 | 220 | 0 | 9 | 70 |
| 15 | 2,934782609 | 2030 | 302 | 129 | 21,25 | 300 | 210 | 70 | 3000 | 220 | 0 | 10 | 47 |
| 16 | 3,557017544 | 1880 | 313 | 129 | 21,25 | 300 | 210 | 70 | 3000 | 220 | 0 | 10 | 57 |
| 17 | 4,193548387 | 1950 | 506 | 129 | 21,25 | 300 | 380 | 75 | 1800 | 120 | 0 | 10 | 60 |
| 18 | 4,897959184 | 1890 | 540 | 129 | 21,25 | 300 | 380 | 75 | 1800 | 120 | 0 | 10 | 70 |
| 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 | 90 | 4 | 47 |
| 21 | 1,598173516 | 1920 | 456 | 111,86 | 22,26785714 | 284,6153846 | 470 | 73,33333333 | 2455,555556 | 220 | 90 | 4 | 57 |
| 22 | 4,029126214 | 1880 | 622 | 111,86 | 22,26785714 | 284,6153846 | 470 | 73,33333333 | 2455,555556 | 220 | 90 | 4 | 60 |
| 23 | 2,587347643 | 1953,274926 | 1136,596135 | 137,6274196 | 22,34453357 | 234,716883 | 555,8934533 | 80,80322176 | 2587,342983 | 246,6131165 | 90 | 4 | 70 |
| 24 | 2,499917928 | 1942,595777 | 901,5199467 | 146,2522078 | 23,08175748 | 351,231874 | 864,7254838 | 76,17807508 | 3705,672523 | 226,2227604 | 90 | 5 | 47 |
| 25 | 2,046471464 | 2037,631811 | 707,570887 | 101,6172513 | 23,14639281 | 312,3072052 | 547,6012186 | 73,81706662 | 2624,026407 | 178,1985559 | 90 | 5 | 57 |
| 26 | 1,85647617 | 2018,220332 | 836,2943816 | 135,4016966 | 26,4355146 | 327,5103767 | 150,9614485 | 77,21076158 | 2473,187195 | 123,3445614 | 90 | 5 | 60 |
| 27 | 3,305535422 | 1917,907506 | 478,2862473 | 105,7869296 | 17,87409991 | 328,1545795 | 526,6921594 | 72,34570879 | 3059,032991 | 275,5758795 | 90 | 5 | 70 |
| 28 | 2,709554095 | 1892,071124 | 641,0525494 | 96,56329319 | 22,98929056 | 262,956722 | 804,5926208 | 74,51135922 | 2288,967377 | 126,8163389 | 90 | 7 | 47 |
| 29 | 2,282825314 | 2008,357592 | 393,9673255 | 149,3728324 | 21,66175068 | 330,498641 | 535,3714591 | 72,24492408 | 2704,445081 | 261,0770716 | 90 | 7 | 57 |
| 30 | 1,978140173 | 1973,629097 | 991,7240946 | 149,3721279 | 19,75057789 | 232,0581913 | 485,4537781 | 75,66570056 | 2448,943079 | 162,4936936 | 90 | 7 | 60 |
| 31 | 1,771436393 | 1872,49156 | 801,0338825 | 79,79454787 | 22,29630372 | 340,7368984 | 864,9291837 | 70,94759156 | 2796,785402 | 123,3562643 | 90 | 7 | 70 |
| 32 | 3,277086987 | 2010,047012 | 339,5504228 | 67,49899306 | 24,28060902 | 254,9490837 | 117,5352342 | 67,47870663 | 2462,605386 | 207,0185813 | 90 | 9 | 47 |
| 33 | 2,984362226 | 1912,315437 | 1183,091845 | 133,5490007 | 23,26379657 | 314,9961255 | 377,3890094 | 75,29045222 | 2303,770656 | 200,5802494 | 90 | 9 | 57 |
| 34 | 2,916149621 | 1879,969846 | 1003,270178 | 109,2395305 | 25,68275948 | 294,0485366 | 408,3542393 | 71,70085562 | 3086,546196 | 192,1911621 | 90 | 9 | 60 |
| 35 | 3,247617211 | 1813,2346 | 757,874479 | 81,37987084 | 23,42246524 | 279,0801575 | 575,0628571 | 69,34113288 | 3188,136358 | 252,8705688 | 90 | 9 | 70 |
| 36 | 2,423875673 | 1908,940601 | 530,2286864 | 58,26241428 | 24,07354923 | 325,138688 | 456,9080467 | 74,24435417 | 1890,505807 | 222,6994873 | 90 | 10 | 47 |
| 37 | 5,09899309 | 1977,339047 | 1572,096042 | 132,3430598 | 25,39700098 | 286,5564309 | 690,3648357 | 72,34163973 | 1386,578973 | 271,9013937 | 90 | 10 | 57 |
| 38 | 2,444176986 | 2085,495837 | 931,3106361 | 110,5648399 | 23,48713976 | 270,2867651 | 278,2300203 | 71,47906047 | 2740,229631 | 187,8613727 | 90 | 10 | 60 |
| 39 | 2,667696929 | 2078,894678 | 1542,168458 | 132,1474033 | 22,6501092 | 357,9728962 | 787,299217 | 76,47178847 | 2559,643047 | 163,902778 | 90 | 10 | 70 |
| 40 | 3,034399483 | 1968,401388 | 455,8710188 | 61,42129652 | 23,49072291 | 316,4145721 | 637,3768927 | 75,09037174 | 2848,490078 | 311,0523979 | 0 | 7,856166547 | 64,30196385 |
| 41 | 2,465204971 | 1936,099137 | 1056,554985 | 71,29405822 | 24,5233807 | 271,9756783 | 129,0771629 | 66,42079436 | 2868,586527 | 227,0225573 | 0 | 7,401542567 | 19,25053314 |
| 42 | 2,664388649 | 1996,159145 | 525,0577741 | 77,50688254 | 18,12610706 | 223,4086854 | 28,65810234 | 69,48977348 | 2220,587445 | 314,7766687 | 0 | 6,675780339 | 78,62329934 |
| 43 | 1,193529582 | 1965,929227 | 899,603701 | 102,9590686 | 19,56671624 | 225,8102229 | 871,0889548 | 73,45469452 | 2335,541792 | 91,04764633 | 0 | 7,52639832 | 38,17697532 |
| 44 | 2,914333275 | 2049,373404 | 382,2633585 | 81,35204737 | 16,39159473 | 233,2960627 | 561,9921308 | 69,81461516 | 2262,784366 | 303,0754524 | 0 | 8,32569922 | 46,04542763 |
| 45 | 4,315665782 | 1913,379677 | 822,9187355 | 143,5769371 | 24,2755681 | 274,9887944 | 260,8583411 | 75,95732857 | 1639,912525 | 248,2443299 | 0 | 7,656210875 | 33,57102356 |
| 46 | 2,338424286 | 1963,35156 | 1155,160504 | 150,0158298 | 18,29942138 | 315,9041781 | 644,3631421 | 71,30396977 | 3407,713581 | 304,4232741 | 0 | 10,30294472 | 39,23427979 |
| 47 | 1,298167246 | 1984,511373 | 1405,786822 | 130,9427979 | 21,8292399 | 288,9520988 | 161,0077185 | 74,68091326 | 2526,814256 | 228,8677196 | 0 | 8,946891121 | 72,08459409 |
| 48 | 2,134446144 | 1988,349053 | 809,2938035 | 95,08902184 | 19,36492572 | 205,499761 | 196,3576427 | 76,34020725 | 2459,524526 | 289,9571416 | 0 | 3,746624977 | 57,99777218 |
| 49 | 4,147966432 | 1991,789789 | 1250,198275 | 116,8564622 | 21,57326495 | 320,7401725 | 755,5005552 | 70,35546403 | 1795,719359 | 189,8833073 | 0 | 9,094363677 | 44,80160057 |
| 50 | 3,057830147 | 1987,259859 | 403,3952302 | 112,719826 | 19,39051107 | 336,2453837 | 352,8319835 | 73,01648889 | 2016,503637 | 206,6933789 | 0 | 6,303772846 | 72,15201873 |
| 51 | 4,182128584 | 1940,739237 | 347,9917504 | 111,9697729 | 18,80028336 | 304,1904883 | 229,9283369 | 72,76947322 | 2097,669452 | 226,4226453 | 0 | 7,257961831 | 50,21963553 |

Рисунок 13 – Демонстрация специфики данных исходного датасета joined_dataset

Таким образом, с помощью функии iloc[] из датасета были удалены все строки до сорокой. Для дальнейшего удаления выбросов был создан датасет raw_dataset, содержащий 983 строки и 13 колонок. Тип данных в датасете: float64 и int64.

Для raw_dataset далее были построены диаграммы "ящик с усами" для определения и последующего удаления выбросов. Принято решение об удалении выбросов для трех признаков:

- температура вспышки, С_2 (более 390 и менее 170);
- содержание эпоксидных групп,%_2 (более 30 и менее 16);
- плотность нашивки (более 95, менее 15).

Такой выбор обсуловлен тем, что данные выбросы наиболее выделяются из основной выборки. Отбор выбросов осуществлялся исходя из цели по максимальному сохранению численности выборки, так как для построения моделей необходимо сохранить достаточное количество параметров.

Таким образом, сформирован новый датасет dataset_filtered, в котором содержаться 972 строки и 13 столбцов, пустые значения отсутствуют (рисунок 14).

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 972 entries, 40 to 1022
Data columns (total 13 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Соотношение матрица-наполнитель       972 non-null    float64
 1   Плотность, кг/м3                      972 non-null    float64
 2   модуль упругости, ГПа                 972 non-null    float64
 3   Количество отвердителя, м.%           972 non-null    float64
 4   Содержание эпоксидных групп,%_2       972 non-null    float64
 5   Температура вспышки, С_2              972 non-null    float64
 6   Поверхностная плотность, г/м2         972 non-null    float64
 7   Модуль упругости при растяжении, ГПа  972 non-null    float64
 8   Прочность при растяжении, МПа         972 non-null    float64
 9   Потребление смолы, г/м2               972 non-null    float64
 10  Угол нашивки, град                    972 non-null    int64  
 11  Шаг нашивки                           972 non-null    float64
 12  Плотность нашивки                     972 non-null    float64
dtypes: float64(12), int64(1)
memory usage: 106.3 KB
```

Рисунок 14 – Информация по датасету dataset_filtered, выведенная с помощью метода info()

На таблице 4 отображены статистические характеристики датасета, содержащие количество значений (count), среднее значение (mean), стандартное (среднеквадратичное) отклонение (std), максимальные и минимальные значения (max, max), пороговые значения для 25, 50 и 75% квартилей.

Датасет dataset_filtered содержит данные в исходном виде, которые трудно сопоставить между собой. Об этом свидетельствут средние значения признаков. В этом связи следующим шагом исследования будет проведение процедуры нормализации датасета.

Таблица 4 – Статистические характеристики датасета filtered_dataset, выведенные с помощью метода describe ()

| Свойства композитов | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Соотношение матрица-наполнитель | 972 | 2,94 | 0,91 | 0,39 | 2,33 | 2,92 | 3,56 | 5,59 |
| Плотность, кг/м3 | 972 | 1 975,50 | 73,63 | 1 731,76 | 1 924,70 | 1 977,60 | 2 020,99 | 2 207,77 |
| модуль упругости, ГПа | 972 | 738,37 | 329,62 | 2,44 | 498,44 | 738,28 | 962,17 | 1 911,54 |
| Количество отвердителя, м.% | 972 | 110,49 | 28,29 | 17,74 | 92,17 | 110,16 | 129,85 | 198,95 |
| Содержание эпоксидных групп,%_2 | 972 | 22,24 | 2,39 | 16,05 | 20,55 | 22,21 | 23,98 | 28,96 |
| Температура вспышки, С_2 | 972 | 285,18 | 40,15 | 173,48 | 258,35 | 285,20 | 312,84 | 386,07 |
| Поверхностная плотность, г/м2 | 972 | 480,55 | 281,28 | 0,60 | 266,21 | 450,41 | 692,80 | 1 399,54 |
| Модуль упругости при растяжении, ГПа | 972 | 73,32 | 3,12 | 64,05 | 71,30 | 73,23 | 75,34 | 82,68 |
| Прочность при растяжении, МПа | 972 | 2 464,87 | 485,27 | 1 036,86 | 2 135,89 | 2 455,30 | 2 758,75 | 3 848,44 |
| Потребление смолы, г/м2 | 972 | 219,01 | 59,88 | 33,80 | 179,91 | 218,25 | 258,81 | 414,59 |
| Угол нашивки, град | 972 | 44,35 | 45,02 | 0 | 0 | 0 | 90 | 90 |
| Шаг нашивки | 972 | 6,92 | 2,57 | 0,04 | 5,14 | 6,91 | 8,58 | 14,44 |
| Плотность нашивки | 972 | 57,14 | 12,12 | 15,42 | 49,89 | 57,36 | 64,93 | 92,96 |

### 2.1.3. Нормализация

Перед процедурой нормализации исключаем целевые значения модуля упругости при растяжении и прочности при растяжении. Далее проводим процедуру нормализации для датасета с десятью переменными.

На таблице 5 отображены статистические характеристики нормализованного датасета, средние значения признаков которого теперь расположены в диапазоне от 0 до 1 и являются сопоставимыми.

Таблица 5 – Статистические характеристики нормализованного датасета normalized_dataset, выведенные с помощью метода describe ()

| Свойства композитов | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Соотношение матрица-наполнитель | 972 | 0,0013 | 0,0004 | 0,0002 | 0,0010 | 0,0013 | 0,0016 | 0,0025 |
| Плотность, кг/м3 | 972 | 0,8857 | 0,0507 | 0,6949 | 0,8527 | 0,8890 | 0,9253 | 0,9850 |
| Модуль упругости, ГПа | 972 | 0,3246 | 0,1313 | 0,0011 | 0,2338 | 0,3320 | 0,4208 | 0,6816 |
| Количество отвердителя, м.% | 972 | 0,0496 | 0,0131 | 0,0079 | 0,0412 | 0,0493 | 0,0586 | 0,0935 |
| Содержание эпоксидных групп,%_2 | 972 | 0,0100 | 0,0013 | 0,0063 | 0,0091 | 0,0099 | 0,0109 | 0,0142 |
| Температура вспышки, С_2 | 972 | 0,1279 | 0,0194 | 0,0749 | 0,1146 | 0,1276 | 0,1409 | 0,1883 |
| Поверхностная плотность, г/м2 | 972 | 0,2120 | 0,1185 | 0,0003 | 0,1218 | 0,2033 | 0,3023 | 0,5599 |
| Потребление смолы, г/м2 | 972 | 0,0983 | 0,0276 | 0,0147 | 0,0803 | 0,0970 | 0,1167 | 0,1968 |
| Угол нашивки, град | 972 | 0,0199 | 0,0203 | 0,0000 | 0,0000 | 0,0000 | 0,0406 | 0,0486 |
| Шаг нашивки | 972 | 0,0031 | 0,0012 | 0,0000 | 0,0023 | 0,0031 | 0,0038 | 0,0074 |
| Плотность нашивки | 972 | 0,0256 | 0,0056 | 0,0073 | 0,0220 | 0,0256 | 0,0295 | 0,0443 |

Таким образом, на этапе предварительной обработки и разведочного анализа данных подготовлен датасет, не содержащий пустых значений, очищенный от явных выбросов, над которым произведена процедура нормализации.

Полученный датасет будет использован для создания моделей прогноза значений двух признаков: модуля упругости при растяжении и прочности при растяжении.

## 2.2. Разработка и обучение модели

В предыдущих разделах в результате предварительной обработки и разведочного анализа данных подготовлен датасет для создания моделей прогноза значений двух признаков: модуля упругости при растяжении и прочности при растяжении. Датасет не содержит пустых значений, очищен от явных выбросов, нормализован.

В первую очередь, поместим целевые переменные 'Модуль упругости при растяжении, ГПа' и 'Прочность при растяжении, МПа' в переменные y_upr и y_prochn. Статистические характеристики целевых переменных, выведенные с помощью метода describe () проиллюстрированы на рисунке 15.

| | Модуль упругости при растяжении, ГПа |
|---|---|
| count | 972.000000 |
| mean | 73.323252 |
| std | 3.122066 |
| min | 64.054061 |
| 25% | 71.301753 |
| 50% | 73.230375 |
| 75% | 75.337585 |
| max | 82.682051 |

| | Прочность при растяжении, МПа |
|---|---|
| count | 972.000000 |
| mean | 2464.874731 |
| std | 485.269329 |
| min | 1036.856605 |
| 25% | 2135.886086 |
| 50% | 2455.297778 |
| 75% | 2758.749272 |
| max | 3848.436732 |

Рисунок 15 – Статистические характеристики целевых переменных, выведенные с помощью метода describe ()

Оставшиеся одиннадцать признаков поместим в переменные x_upr и x_prochn (таблицы 6-7).

Таблица 6 – Статистические характеристики независимых переменных x_upr, выведенные с помощью метода describe ()

| Свойства композитов | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Соотношение матрица-наполнитель | 972 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| Плотность, кг/м3 | 972 | 0,89 | 0,05 | 0,69 | 0,85 | 0,89 | 0,93 | 0,98 |
| Модуль упругости, ГПа | 972 | 0,32 | 0,13 | 0,00 | 0,23 | 0,33 | 0,42 | 0,68 |
| Количество отвердителя, м.% | 972 | 0,05 | 0,01 | 0,01 | 0,04 | 0,05 | 0,06 | 0,09 |
| Содержание эпоксидных групп,%_2 | 972 | 0,01 | 0,00 | 0,01 | 0,01 | 0,01 | 0,01 | 0,01 |
| Температура вспышки, С_2 | 972 | 0,13 | 0,02 | 0,07 | 0,11 | 0,13 | 0,14 | 0,19 |
| Поверхностная плотность, г/м2 | 972 | 0,21 | 0,12 | 0,00 | 0,12 | 0,20 | 0,30 | 0,56 |
| Потребление смолы, г/м2 | 972 | 0,10 | 0,03 | 0,01 | 0,08 | 0,10 | 0,12 | 0,20 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Угол нашивки, град | 972 | 0,02 | 0,02 | 0,00 | 0,00 | 0,00 | 0,04 | 0,05 |
| Шаг нашивки | 972 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,01 |
| Плотность нашивки | 972 | 0,03 | 0,01 | 0,01 | 0,02 | 0,03 | 0,03 | 0,04 |

Таблица 7 – Статистические характеристики независимых переменных x_prochn, выведенные с помощью метода describe ()

| Свойства композитов | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Соотношение матрица-наполнитель | 972 | 0,001 | 0,000 | 0,000 | 0,001 | 0,001 | 0,002 | 0,003 |
| Плотность, кг/м3 | 972 | 0,886 | 0,051 | 0,695 | 0,853 | 0,889 | 0,925 | 0,985 |
| модуль упругости, ГПа | 972 | 0,325 | 0,131 | 0,001 | 0,234 | 0,332 | 0,421 | 0,682 |
| Количество отвердителя, м.% | 972 | 0,050 | 0,013 | 0,008 | 0,041 | 0,049 | 0,059 | 0,093 |
| Содержание эпоксидных групп,%_2 | 972 | 0,010 | 0,001 | 0,006 | 0,009 | 0,010 | 0,011 | 0,014 |
| Температура вспышки, С_2 | 972 | 0,128 | 0,019 | 0,075 | 0,115 | 0,128 | 0,141 | 0,188 |
| Поверхностная плотность, г/м2 | 972 | 0,212 | 0,118 | 0,000 | 0,122 | 0,203 | 0,302 | 0,560 |
| Потребление смолы, г/м2 | 972 | 0,098 | 0,028 | 0,015 | 0,080 | 0,097 | 0,117 | 0,197 |
| Угол нашивки, град | 972 | 0,020 | 0,020 | 0,000 | 0,000 | 0,000 | 0,041 | 0,049 |
| Шаг нашивки | 972 | 0,003 | 0,001 | 0,000 | 0,002 | 0,003 | 0,004 | 0,007 |
| Плотность нашивки | 972 | 0,026 | 0,006 | 0,007 | 0,022 | 0,026 | 0,029 | 0,044 |

Разобьем данные на обучающую (X_train_upr, y_train_upr, X_train_prochn, y_train_prochn) и тестовую (X_test_upr, y_test_upr, X_test_prochn, y_test_prochn) выборки. В соответствии с заданными условиями 30% данных оставим на тестирование модели, на остальных произведем обучение моделей, значение случайного числа (random_state) определим равным 42.

Посмотрим на размерности обучающих и тестовых выборок модуля упругости при растяжении и прочности при растяжении (таблица 8).

Таблица 8 – Размерности обучающих и тестовых выборок модуля упругости при растяжении и прочности при растяжении

| X_train_upr | y_train_upr | X_test_upr | y_test_upr |
|---|---|---|---|
| (680, 11) | (680, 1) | (292, 11) | (292, 1) |
| X_train_prochn | y_train_prochn | X_test_ prochn | y_test_ prochn |
| (680, 11) | (680, 1) | (292, 11) | (292, 1) |

Далее проведено обучение моделей для прогноза значений модуля упругости при растяжении и прочности при растяжении, рассчитаны значения средних квадратических ошибок (MSE), корней из средней квадратической ошибки (RMSE), а также коэффициентов детерминации ($R^2$).

В настоящем исследовании использованы следующие методы обучения моделей: линейная регрессия, метод лассо, дерево решений, метод ближайших соседей.

Таблица 9 – Оценка качества моделей для показателя модуль упругости при растяжении

| Метрики | RMSE | MSE | R2_score |
|---|---|---|---|
| Лассо упругость | 3,22 | 10,36 | 0,00 |
| Дерево решений упругость | 3,23 | 10,45 | -0,01 |
| К-соседей упругость | 3,33 | 11,09 | -0,07 |

Таблица 10 – Оценка качества моделей для показателя прочности при растяжении

| Метрики | RMSE | MSE | R2_score |
|---|---|---|---|
| Лассо прочность | 502,02 | 252027,1 | -0,02 |
| Дерево решений прочность | 2500,32 | 6251620 | -24,35 |
| К-соседей прочность | 500,33 | 250329,4 | -0,01 |

Представление выше метрики свидетельствует о достаточно низких способностях моделей к прогнозированию. Наилучшие значения показывает метод регрессии Лассо, что тем не менее не говорит об эффективности модели.

## 2.3. Нейронная сеть для прогнозирования соотношения матрица наполнитель

В первую очередь, подготовим датасет для прогнозирования значений соотношения матрица-наполнитель. Для этого из исходного датасета dataset_index, прошедшего процесс нормализации и удаления выбросов, выведем целевую переменную и сохраним ее в target_smn.

Далее, из исходного датасет dataset_index удалим целевую переменную с помощью метода drop(). Полученный датасет с независимыми переменными df_drop_target_smn содержит 12 столбцов.

Таким образом, получаем важную информацию для разработки нейронной сети: значение input_dim равное 12, для выходного слоя зададим такое количество нейронов, которое соответствует количеству прогнозируемых классов, т.е. 1.

Далее нормализуем датасет аналогично процедуре, проведенной в разделе 2.1: с помощью метода Normalizer. На таблице 11 отображены статистические характеристики датасета normalized_smndataset, средние значения признаков которого теперь расположены в диапазоне от 0 до 1.

Таблица 11 – Статистические характеристики нормализованного датасета normalized_smndataset, выведенные с помощью метода describe ()

| Свойства композитов | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Плотность, кг/м3 | 972 | 0,597 | 0,069 | 0,420 | 0,551 | 0,592 | 0,642 | 0,874 |
| модуль упругости, ГПа | 972 | 0,221 | 0,096 | 0,001 | 0,150 | 0,222 | 0,291 | 0,481 |
| Количество отвердителя, м.% | 972 | 0,033 | 0,010 | 0,005 | 0,027 | 0,033 | 0,039 | 0,073 |
| Содержание эпоксидных групп,%_2 | 972 | 0,007 | 0,001 | 0,004 | 0,006 | 0,007 | 0,007 | 0,011 |
| Температура вспышки, С_2 | 972 | 0,086 | 0,016 | 0,048 | 0,075 | 0,085 | 0,097 | 0,148 |
| Поверхностная плотность, г/м2 | 972 | 0,144 | 0,084 | 0,000 | 0,081 | 0,136 | 0,203 | 0,414 |
| Модуль упругости при растяжении, ГПа | 972 | 0,022 | 0,003 | 0,015 | 0,020 | 0,022 | 0,024 | 0,033 |
| Прочность при растяжении, МПа | 972 | 0,730 | 0,073 | 0,430 | 0,687 | 0,740 | 0,781 | 0,882 |
| Потребление смолы, г/м2 | 972 | 0,066 | 0,019 | 0,011 | 0,053 | 0,066 | 0,079 | 0,145 |
| Угол нашивки, град | 972 | 0,013 | 0,014 | 0,000 | 0,000 | 0,000 | 0,027 | 0,038 |
| Шаг нашивки | 972 | 0,002 | 0,001 | 0,000 | 0,002 | 0,002 | 0,003 | 0,005 |
| Плотность нашивки | 972 | 0,017 | 0,004 | 0,004 | 0,015 | 0,017 | 0,020 | 0,031 |

Далее разобьем данные на обучающую (X_train_smn, y_train _smn) и тестовую (X_test _smn, y_test _smn) выборки. В соответствии с заданными условиями, 30% данных оставим на тестирование модели, на остальных проведем обучение моделей, значение случайного числа (random_state) определим равным 42.

Посмотрим на размерности обучающих и тестовых выборок соотношения матрица-наполнитель (таблица 12).

Таблица 12 – Размерности обучающих и тестовых выборок соотношения матрица-наполнитель

| X_train_smn | y_train_smn |
|---|---|
| (680, 12) | (680, 1) |
| X_test_smn | y_test.smn |
| (292, 12) | (292, 1) |

Разработаем и обучим нейронную сеть для прогнозирования значений соотнрошения матрица-наполнитель. Создадим четыре DENSE-слоя: три скрытых слоя, состоящих из 128 нейронов каждый, и 1 выходной слой из 1 нейрона.

Активационные функции на скрытых слоях relu, на выходном – elu. Входному слою передаем input_dim = 12, а также функцию активации relu. Воспользуемся методом Dropout, чтобы избежать переобучения сети, зададим значение 0.2. Для выходного слоя зададим такое количество нейронов, которое соответствует количеству классов, т.е. 1, а также воспользуемся функцией elu.

Скомпелируем модель с помощью метода compile(), которому зададим функцию потерь mean_absolute_error, оптимизатор обучения Adam.

Запустим обучение нейронной сети с помощью метода fit(). Передадим методу данные x_train_smn и y_train_smn, размер батча

(batch_size) зададим равным 128, количество циклов обучения (epochs) – 100, долю обучающих данных, используемую для проверки нейросети (validation_split) 0,2.

Для нейронной сети рассчитаны значения средней квадратической ошибки (MSE), корня из средней квадратичной ошибки (RMSE), а также коэффициента детерминации ($R^2$) (таблица 13).

Таблица 13 – Оценка качества нейронной сети для соотношения матрица-наполнитель

| Метрики | |
|---|---|
| RMSE | 0,88 |
| MSE | 0,77 |
| R2 | -0,02 |

Представление выше метрики свидетельствует о достаточно низких способностях модели к прогнозированию. Отрицательный коэффициент детерминации говорит о низком качестве модели.

Таким образом, модель для прогнозирования свойства «соотношение матрица-наполнитель» недостаточно эффективна.

## 2.4. Разработка приложения

Приложение создано для прогнозирования признака соотношение-матрица наполнитель (рисунок 16). Для получения прогнозного значения необходимо ввести значения двенадцати переменных:

1. модуль упругости при растяжении;
2. прочность при растяжении;
3. плотность, кг/м3;
4. модуль упругости, Гпа;
5. количество отвердителя, м.%;
6. содержание эпоксидных групп, %_2;
7. температура вспышки, С_2;
8. поверхностная плотность, г/м2;
9. потребление смолы, г/м2;
10. угол нашивки, град;
11. шаг нашивки;
12. плотность нашивки.

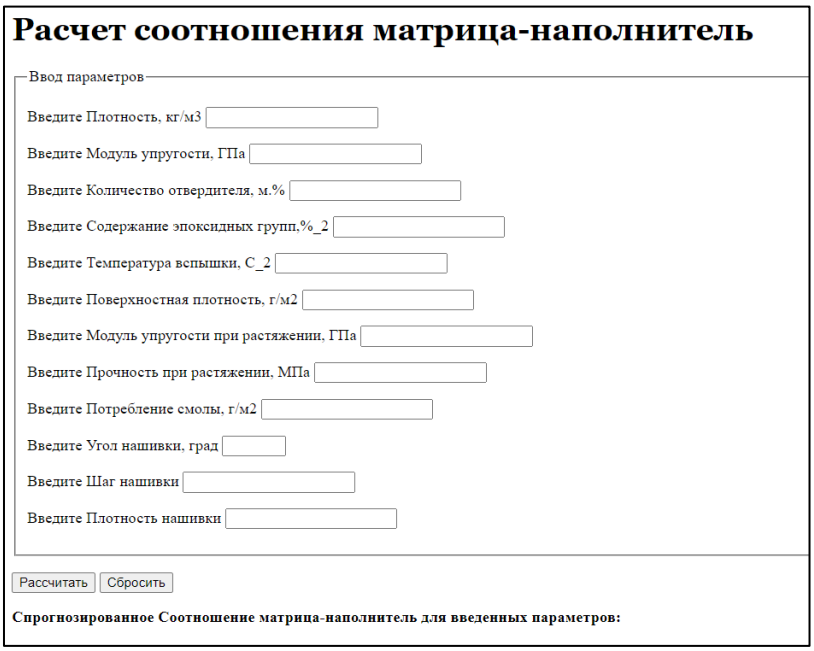


Рисунок 16 – Окно ввода данных для прогнозирования признака соотношение-матрица наполнитель

Далее необходимо заполнить ячейки соответствующими значениями независимых признаков (рисунок 17).

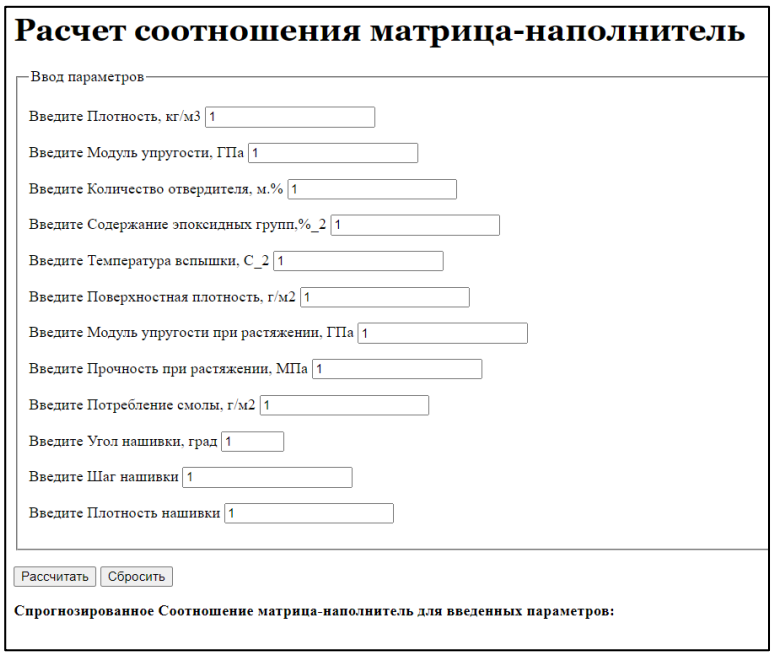


Рисунок 17 – Окно ввода данных с заполненными значениями независимых переменных

После проставления вводных данных необходимо нажать на кнопку «рассчитать». Прогнозное значение признака соотношение-матрица наполнитель отразится в нижней строке (рисунок 18).

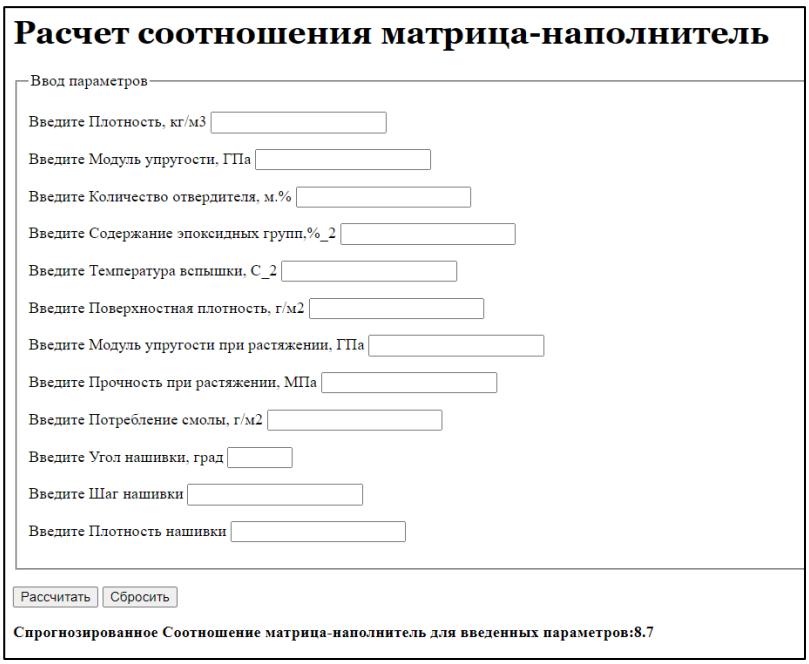


Рисунок 18 – Окно ввода данных с выведенным прогнозным значением признака соотношение-матрица наполнитель

## 2.5. Создание удаленного репозитория

Автором создана личная страница на веб-сервисе GitHub (aburchakova). На страницу добавлен репозиторий «kompozitus», который находится по адресу: https://github.com/aburchakova/kompozitus.

## Заключение

Прогнозирование свойств композиционных материалов является актуальной производственной задачей, позволяющей снизить количество реально проводимых испытаний.

Объектом исследования выступили композиционные материалы, которые представляют собой искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними.

Предметом исследования стали прогнозные данные трех свойств композитов: соотношение матрица-наполнитель, модуль упругости при растяжении, прочность при растяжении.

В настоящем исследовании последовательно решались следующие задачи исследования:

1. изучение теоретических основ и методов прогнозирования целевых переменных, в частности свойств получаемых композиционных материалов;
2. проведение разведочного анализа данных;
3. обучение моделей для прогноза целевых признаков: модуля упругости при растяжении и прочности при растяжении;
4. написание нейронной сети, которая будет рекомендовать соотношение матрица-наполнитель.

На первом этапе исследования сформирован датасет с данными о свойствах композиционных материалов, представленный классом DataFrame, хранящий 1023 строки и 13 столбцов. Проведено исследование данных с использованием статистических характеристик.

На втором этапе исследования проведена обработка данных: построены попарные графики рассеяния точек, диаграммы ящика с усами, гистограммы, исследованы выбросы датасета, изучена корреляция между признаками, а также проведена нормализация независимых переменных. В

результате подготовлен датасет, использованный при построении моделей прогнозирования значений переменных.

Кроме того, осуществлено изучение теоретических основ и методов прогнозирования целевых переменных, в частности линейной регрессии, метода лассо, дерева решений, метода ближайших соседей, исследована специфика нейросетевой регрессии.

На третьем этапе исследования разработаны и обучены модели прогноза значений модуля упругости при растяжении, прочности при растяжении, создана нейронная сеть для прогнозирования соотношения матрица наполнитель.

Оценка метрик моделей показала наличие слабой способности моделей к прогнозированию целевых значений. В этой связи для дальнейшего исследования определены следующие задачи: применение б*о*льшего количества моделей машинного обучения для поиска модели с более высокой способностью к прогнозированию, обработка исходных данных датасета с участием экспертов-составителей датасета и формирование нового датасета.

**Библиографический список**

1. Бабешко Л.О. Основы эконометрического моделирования: Учебное пособие. [Текст] / КомКнига . – г. Москва. – 2006 г. – 432 с.

2. Брюс П., Брюс Э. Практическая статистика для специалистов Data Science [Текст] / «БХВ-Петербург» – г. Санкт-Петербург. – 2020 г. – 304 с.

3. Грас Д. Data Science. Наука о данных с нуля [Текст] / БХВ-Петербург. – г. Санкт-Петербург. – 2021 г. – 416 с.

4. Постолит, А. Основы искусственного интеллекта в примерах на Python [Текст] / «БХВ-Петербург» – г. Санкт-Петербург. – 2022 г. – 448 с.

5. Свейгарт Эл. Автоматизация рутинных задач с помощью Python [Текст] / ООО «Дилектика». – г. Санкт-Петербург. – 2021 г. – 672 с.

6. Теория статистики. Учебник. – г. Москва. – «Финансы и статистика». – 1998 г.

7. Фадеева Л.Н., Жуков Ю.В., Лебедев А.В. Математика для экономистов: Теория вероятностей и математическая статистика [Текст] / Эксмо. – г. Москва. – 2006 г. – 336 с.

8. «Развитие промышленности и повышение ее конкурентоспособности» [Электронный ресурс] / Портал Госпрограмм РФ. – https://programs.gov.ru/Portal/program/16/passport / Дата обращения 25.04.2023

9. «Научно-технологическое развитие Российской Федерации» [Электронный ресурс] / Портал Госпрограмм РФ. – https://programs.gov.ru/Portal/programs/passport/47. Дата обращения 25.04.2023