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 Яманкин Андрей Геннадьевич

Москва, 2023

Содержание

Введение

Структура выпускной квалификационной работы представлена введением, двумя главами, заключением и списком используемой литературы.

Композиционные материалы — это материалы, состоящие из двух или более компонентов, нерастворимых друг с другом, с чётко обозначенной границей раздела и сильным взаимодействием по всей зоне контакта. Одним из компонентов композитных материалов является непрерывная фаза, он называется матрица, в которой нерастворимые материалы помещаются в другую природу, называемую арматурой или наполнителем.

Внедрение композиционных материалов обусловлено стремлением использовать их преимущества по сравнению с традиционно используемыми металлами и сплавами. Примеры композита – железобетон (сочетание стали арматуры и камня бетона), древесноволокнистая плита ДВП (сочетание древесной основы – щепы и полимерного связующего).

Базальт - магматическая вулканическая порода. Это самая распространённая порода на поверхности Земли и на других планетах Солнечной системы. Базальты образуются путём затвердевания силикатного магматического расплава. Большая часть базальтов образуется на срединно-океанических хребтах и образует океаническую кору. Активно развивается использование композитных материалов на основе базальта.

Базальтопластик - современный композитный материал на основе базальтовых волокон и органического связующего вещества. В настоящее время базальтопластик успешно конкурирует с металлическими изделиями, превосходя их по коррозионной, щелочной, кислотоустойчивости и некоторым другим свойствам. Целью данной работы является прогнозирование конечных свойств новых материалов на основе базальтопластика (композиционных материалов).

Расширение разнообразия материалов, используемых при проектировании нового композиционного материала, увеличивает необходимость определения свойств нового композита при минимальных финансовых затратах. Для решения этой проблемы обычно используются два способа: физические тесты образцов материалов или оценка свойств, в том числе на основе физико-математических моделей. Традиционно разработка композитных материалов является долгосрочным процессом, так как из свойств отдельных компонентов невозможно рассчитать конечные свойства композита. Для достижения определенных характеристик требуется большое количество различных комбинированных тестов, что делает насущной задачу прогнозирования успешного решения, снижающего затраты на разработку новых материалов и затраты на рабочую силу. Суть прогнозирования заключается в моделировании репрезентативного элемента композитного объёма на основе данных о свойствах входящих компонентов (связующего и армирующего компонента).

Актуальность выпускной работы состоит в том, что созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

Объектом исследования является процесс прогнозирования конечных свойств новых материалов.

Предметом исследования – автоматизация процесса прогнозирования конечных свойств новых материалов.

Цели данной выпускной квалификационной работы:

1. Обучить алгоритм машинного обучения, который будет определять значения:

- Модуль упругости при растяжении, ГПа
- Прочность при растяжении, МПа

2. Написать нейронную сеть, которая будет рекомендовать:

- Соотношение матрица-наполнитель

3. Написать приложение, которое будет выдавать прогнозное значение параметра «Соотношение матрица-наполнитель».

В процессе исследовательской работы были разработаны несколько моделей, способные с высокой вероятностью прогнозировать модули упругости при растяжении и прочности при растяжении, а также были созданы 2 нейронных сети, которые предлагают соотношение «матрицы - наполнитель». На основе одной из нейронных сетей было создано пользовательское веб - приложение на фреймворке Flask.

1. Аналитическая часть

1.1. Постановка задачи

В связи с широким применением композиционных материалов во многих сферах экономической деятельности появилась необходимость разработки автоматизированной информационной системы прогнозирования свойств исследуемых материалов на основе регрессионного анализа, которая позволит снизить трудоемкость экспериментальных исследований, повысить качество прогнозирования физико-механических и технологических свойств материалов, снизить себестоимость изготавливаемых изделий.

В настоящее время при производстве композитов свойства конечных изделий определяются по контрольным образцам, полученным из соответствующей серии партии деталей. Практически не существует эффективных методик, позволяющих прогнозировать свойства конечных изделий на основе информации о компонентах. Конечная концентрация компонентов в композитах зависит от исходных компонентов и параметров технологического процесса их изготовления.

Анализ композиционных материалов, методов обработки экспериментальных данных и программных средств, применяемых для исследования свойств материалов, показал возможность использования корреляционно-регрессионного анализа для формирования прогнозных моделей с помощью прикладных программ.

Для получения прогнозных моделей свойств композитных материалов, а именно модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель», необходимо выполнить следующие шаги:

- провести разведочный анализ предложенных данных, построить гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек;
- провести предобработку данных (удаление шумов, нормализация и стандартизация данных и т.д.);
- обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении;
- написать нейронную сеть, которая будет рекомендовать соотношение «матрица-наполнитель»;
- разработать приложение, которое будет выдавать прогноз соотношения «матрица-наполнитель»;
- провести оценку точности модели на тренировочном и тестовом датасете.

Для выполнения исследования использованы производственные данные Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана). Информация представлена в виде двух файлов формата excel: X_bp.xlsx (характеристики базальтопластика) и X_nup.xlsx (характеристики нашивки из углепластика). Датасет X_bp. Датасет из файла X_bp.xlsx содержит 1023 строки и 11 столбцов, из файла X_nup.xlsx – 1040 строк и 4 столбца (рисунки 1-2).

Ввод [4]:
```
# Загрузка датасета из файла "X_bp.xlsx" и вывод 5 первых строк
dataset_bp = pd.read_excel('C:/Users/AYAmankin/Desktop/Курс Data science МГТУ/ВКР/ВКР/Datasets/X_bp.xlsx')
dataset_bp.head()
```

Out[4]:

	Unnamed: 0	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
0	0	1.857143	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.0	3000.0	220.0
1	1	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.0	3000.0	220.0
2	2	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.0	3000.0	220.0
3	3	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.0	3000.0	220.0
4	4	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0

Ввод [5]:
```
# Вывод размерности первого датасета
dataset_bp.shape
```

Out[5]:
```
(1023, 11)
```

Рисунок 1 - Первые 5 строк и размерность датасета из файла X_bp

Ввод [8]:
```
#Загрузка датасета из файла "X_nup.xlsx", вывод 5 первых строк
dataset_nup = pd.read_excel('C:/Users/AYAmankin/Desktop/Курс Data science МГТУ/ВКР/ВКР/Datasets/X_nup.xlsx')
dataset_nup.head()
```

Out[8]:

	Unnamed: 0	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	0	0	4.0	57.0
1	1	0	4.0	60.0
2	2	0	4.0	70.0
3	3	0	5.0	47.0
4	4	0	5.0	57.0

Ввод [9]:
```
# вывод размерности второго датасета
dataset_nup.shape
```

Out[9]: (1040, 4)

Рисунок 2 - Первые 5 строк и размерность датасета из файла X_nup

Для дальнейшей работы указанные выше файлы были объединены в один при помощи функции merge с типом объединения INNER. Полученный датасет содержит 1023 строки и 13 столбцов (рисунок 3). Это означает, что часть данных (а именно 17 строк из датасета dataset_nup) была удалена из таблицы и исключена из дальнейшего исследования.

Ввод [11]:
```
# Объединим датасеты с помощью метода merge(), тип объединения INNER и выведем первые 5 строк
dataset = pd.merge(dataset_bp, dataset_nup,
                   left_index=True,
                   right_index=True,
                   how = "inner")
dataset.head()
```

Out[11]:

	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	наш
0	1.857143	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.0	3000.0	220.0	0	
1	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.0	3000.0	220.0	0	
2	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.0	3000.0	220.0	0	
3	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.0	3000.0	220.0	0	
4	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0	0	

Ввод [12]:
```
# выведем размерность объединенного датасета
dataset.shape
```

Out[12]: (1023, 13)

Рисунок 3 - Первые 5 строк и размерность объединенного датасета

Объединенный датасет был сохранен под новым именем X_bp_nup для проведения последующего детального анализа.

1.2. Описание используемых методов

При решении конкретной задачи машинного обучения нельзя сказать заранее, какой вид модели будет наиболее эффективен. Проблему выбора модели чаще всего решают перебором – нужно попробовать разные типы моделей и выбрать те, которые показывают наибольшую эффективность. Для задачи прогнозирования модуля упругости при растяжении и прочности при растяжении рассмотрим следующие модели:

1. Метод К-ближайших соседей
2. Метод опорных векторов
3. Линейная регрессия
4. Дерево решений
5. AdaBoost
6. Градиентный бустинг
7. XGBoost
8. Случайный лес
9. Стохастический градиентный спуск
10. Метод регрессии «Lasso».

1. Метод К-ближайших соседей (k-Nearest Neighbors, или kNN). Суть метода достаточно проста: посмотри на соседей вокруг, какие из них преобладают, таковым ты и являешься. В случае использования метода для классификации объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны. В случае использования метода для регрессии, объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны. Регрессия на основе соседей может

использоваться в случаях, когда метки данных являются непрерывными, а не дискретными переменными.

Преимущества:

- Легкая и простая модель машинного обучения.
- Легко добавить больше данных во множество данных.
- Модель принимает только 2 параметра: К и метрика расстояния (обычно это Евклидово расстояние).

Недостатки:

- К следует выбирать мудро.
- Большие вычислительные затраты во время выполнения, если размер выборки велик.
- Работает не так хорошо с категорическими параметрами.

2. Метод опорных векторов (Support Vector Machines, или SVM) – это один из наиболее широко используемых алгоритмов машинного обучения, применяемый для решения задач классификации, регрессии и обнаружения выбросов. Алгоритм для решения задач классификации строит гиперплоскость в n-мерном пространстве для разделения объектов двух или более классов. Гиперплоскость выбирается таким образом, чтобы максимизировать расстояние между гиперплоскостью и ближайшими объектами разных классов (зазор). Объекты, которые расположены ближе всего к гиперплоскости, называются опорными векторами.

Метод классификации опорных векторов может быть расширен для решения задач регрессии. Этот метод называется регрессией опорных векторов (Support Vector Regression, или SVR). В SVR идентифицируется гиперплоскость с максимальным запасом, так что максимальное количество точек данных находится в пределах этого поля.

Преимущества:

- Является одним из наиболее точных алгоритмов машинного обучения, которые могут обучаться на больших наборах данных.
- Хорошо работает с данными, которые имеют большое количество признаков.
- Работает с небольшими выборками данных.
- Малое количество гиперпараметров: имеет только несколько гиперпараметров, что делает его относительно простым для настройки.

Недостатки:

- Чувствительность к шуму. Шум в данных может привести к тому, что SVM строит границу принятия решений, которая не обобщается хорошо на новые данные.
- Вычислительная сложность для обучения на больших наборах данных.
- Выбор правильной функции ядра может быть сложной задачей. Некоторые ядра работают лучше на определенных типах данных, и выбор неправильного ядра может привести к плохим результатам.

3. Линейная регрессия (Linear regression) – это метод машинного обучения с учителем, который используется для предсказания непрерывной целевой переменной от одного или нескольких независимых признаков. В основе метода лежит предположение предполагает о том, что существует линейная связь между признаками и целевой переменной. Эта связь моделируется с помощью линейной функции. Модель линейной регрессии пытается найти лучшую прямую, которая может описывать зависимость между независимыми признаками и зависимой переменной. Это делается с помощью поиска оптимальных коэффициентов, которые могут быть использованы для описания линейной функции. Эта модель может быть

использована как для предсказания, так и для анализа влияния признаков на целевую переменную.

Преимущества линейной регрессии:

- Простота и удобство в использовании.
- Эффективность при линейных зависимостях.
- Интерпретируемость: в линейной регрессии каждый коэффициент регрессии может быть использован для определения влияния каждой независимой переменной на зависимую переменную.

Недостатки линейной регрессии:

- Ограниченная эффективность при нелинейных зависимостях: если между независимыми и зависимыми переменными существует нелинейная зависимость, линейная регрессия может давать неточные предсказания.
- Необходимость проведения предварительной подготовки данных: линейная регрессия чувствительна к выбросам и мультиколлинеарности, так что необходимо выполнить предварительную подготовку данных.

4. Дерево принятия решений (Decision Tree) – это непараметрический контролируемый метод обучения, используемый для классификации и регрессии. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной, изучая простые правила принятия решений, выведенные из характеристик данных. Дерево можно рассматривать как кусочно-постоянное приближение.

Некоторые преимущества деревьев решений:

- Просто понять и интерпретировать. Деревья можно визуализировать.
- Требуется небольшая подготовка данных.
- Стоимость использования дерева является логарифмической по количеству точек данных, используемых для обучения дерева.

- Может обрабатывать как числовые, так и категориальные данные.
- Способен обрабатывать проблемы с несколькими выходами.
- Возможна проверка модели с помощью статистических тестов. Это позволяет учитывать надежность модели.

К недостаткам деревьев решений можно отнести:

- Обучающиеся дереву решений могут создавать слишком сложные деревья, которые плохо обобщают данные. Это называется переобучением.
- Деревья решений могут быть нестабильными, поскольку небольшие изменения в данных могут привести к созданию совершенно другого дерева.
- Предсказания деревьев решений не являются ни гладкими, ни непрерывными, а являются кусочно-постоянными приближениями. Следовательно, они не годятся для экстраполяции.
- Ученики дерева решений создают предвзятые деревья, если некоторые классы доминируют. Поэтому рекомендуется сбалансировать набор данных перед подгонкой к дереву решений.

Одно решающее дерево имеет достаточно грубые границы между листьями, и при этом либо существенно недообучается (при малом количестве листьев), либо сильно переобучается (при большом количестве листьев). Эту проблему можно исправить, обучая сразу много разных решающих деревьев.

В целом ансамблированием называется комбинация нескольких моделей машинного обучения в одну модель. Ансамблирование решающих деревьев как правило осуществляется «одноуровнево», то есть все деревья работают параллельно и независимо выдают ответ, а затем их предсказания складываются или усредняются. Процесс обучения при этом может выполняться параллельно (бэггинг) или последовательно (бустинг).

Наиболее распространенные алгоритмы ансамблирования решающих деревьев: Random forest, Adaboost, Gradient Boosting, XGBoost, LightGBM, CatBoost.

Бустинг (Boosting) – это способ построения ансамбля, в котором обучается много копий более слабой модели («weak learner»), то есть такой модели, которая не может достичь высокой точности на обучающем датасете, переобучившись на нем. Как правило такой моделью является решающее дерево небольшой глубины. На каждом шаге новый weak learner концентрируется на исправлении ошибок, допущенных предыдущими weak learner'ами. В итоге предсказания всех weak learner'ов суммируются с определенными весами. Бустинг чем-то похож на бэггинг, но в бэггинге модели обучаются совершенно независимо и параллельно, а в бустинге последовательно, с оглядкой на предыдущие.

5. AdaBoost (AdaBoostRegressor) – алгоритм машинного обучения, в котором каждый следующий weak learner фокусировал внимание на тех примерах, на которых предыдущие weak learner'ы дали неверные ответы. При этом он не знал, какие именно ответы даны предыдущими weak learner'ами - было лишь известно, что ответы неверны или неточны. Задачей нового weak learner'a было дать верные ответы преимущественно на этих примерах. Заметим, что при этом не используется никакого валидационного датасета. Используется только обучающий датасет, на нем же оценивается точность предыдущих weak learner'ов. Это означает, что если очередной weak learner после обучения дал верные ответы на все примеры, то бустинг продолжить будет невозможно. Например, если в качестве weak learner'a мы используем решающее дерево неограниченной глубины, то так и произойдет. Нужно использовать решающие деревья небольшой глубины: weak learner должен быть действительно "слабым", не переобучаясь слишком сильно.

Преимущества алгоритма AdaBoost:

- Можно легко, быстро и просто запрограммировать.

- Достаточно гибкий, чтобы комбинировать его с любым алгоритмом машинного обучения без настройки параметров.
- Расширяем до задач обучения сложнее, чем двоичная классификация, и достаточно универсален, поскольку его можно использовать с числовыми или текстовыми данными.

Недостатки:

- Этот алгоритм доказывается эмпирически и очень уязвим к равномерно распределенному шуму.
- Слабые классификаторы в случае, если они слишком слабые, могут привести к плохим результатам и переобучению.

6. Градиентный бустинг (GradientBoostingRegressor) – это продвинутый алгоритм машинного обучения для решения задач классификации и регрессии. Он строит предсказание в виде ансамбля слабых предсказывающих моделей, которыми в основном являются деревья решений. Из нескольких слабых моделей в итоге собирается одна, но уже эффективная. В градиентном бустинге целевыми данными для следующего weak learner'a является градиент (со знаком минус) функции потерь по предсказаниям предыдущих алгоритмов. Таким образом следующий weak learner корректирует предсказания предыдущих. Общая идея алгоритма – последовательное применение предиктора (предсказателя) таким образом, что каждая последующая модель сводит ошибку предыдущей к минимуму.

Преимущества:

- Алгоритм работает с любыми функциями потерь.
- Предсказания в среднем лучше, чем у других алгоритмов.
- Самостоятельно справляется с пропущенными данными.

Недостатки:

- Алгоритм крайне чувствителен к выбросам и при их наличии будет тратить огромное количество ресурсов на эти моменты.
- Модель будет склонна к переобучению при слишком большом количестве деревьев. Данная проблема присутствует в любом алгоритме, связанном с деревьями.
- Вычисления могут занять много времени.

7. XGBoost (XGBRegressor) – является вычислительно эффективной реализацией градиентного бустинга над решающими деревьями. Помимо оптимизированного программного кода, авторы предлагают различные улучшения алгоритма. Основной ценностью библиотеки XGBoost является эффективная программная реализация. За счет разных оптимизаций, таких как эффективная работа с пропущенными значениями, поиск порога только среди персентилей, оптимизация работа с кэшем и распределенное обучение, достигается выигрыш в десятки или даже сотни раз по сравнению с наивной реализацией.

8. Случайный лес (Random forest) – это метод использующий ансамбль деревьев решений, созданных на случайно разделенном датасете. Набор таких деревьев-классификаторов образует лес. Каждое отдельное дерево решений генерируется с использованием метрик отбора показателей, таких как критерий прироста информации, отношение прироста и индекс Джини для каждого признака.

Любое такое дерево создается на основе независимой случайной выборки. В задаче классификации каждое дерево голосует, и в качестве окончательного результата выбирается самый популярный класс. В случае регрессии конечным результатом считается среднее значение всех выходных данных ансамбля. Метод случайного леса является более простым и эффективным по сравнению с другими алгоритмами нелинейной классификации.

Преимущества:

- Имеет высокую точность предсказания.
- Не требует тщательной настройки параметров.
- Практически не чувствителен к выбросам в данных из-за случайного семплирования (random sample).
- Не чувствителен к масштабированию и к другим монотонным преобразованиям значений признаков.
- Редко переобучается. На практике добавление деревьев только улучшает композицию.
- В случае наличия проблемы переобучения, она преодолевается путем усреднения или объединения результатов различных деревьев решений.
- Способен эффективно обрабатывать данные с большим числом признаков и классов.
- Хорошо работает с пропущенными данными – сохраняет хорошую точность даже при их наличии.
- Одинаково хорошо обрабатывает как непрерывные, так и дискретные признаки.
- Высокая параллелизуемость и масштабируемость.

Недостатки:

- Для реализации алгоритма случайного дерева требуется значительный объем вычислительных ресурсов.
- Большой размер моделей.
- Построение случайного леса отнимает больше времени, чем деревья решений или линейные алгоритмы.
- Алгоритм склонен к переобучению на зашумленных данных.
- Нет формальных выводов, таких как p-values, которые используются для оценки важности переменных.

- В отличие от более простых алгоритмов, результаты случайного леса сложнее интерпретировать.
- Когда в выборке очень много разреженных признаков, таких как тексты или наборы слов (bag of words), алгоритм работает хуже, чем линейные методы.
- В отличие от линейной регрессии, Random Forest не обладает возможностью экстраполяции. Это можно считать и плюсом, так как в случае выбросов не будет экстремальных значений.
- Если данные содержат группы признаков с корреляцией, которые имеют схожую значимость для меток, то предпочтение отдается небольшим группам перед большими, что ведет к недообучению.
- Процесс прогнозирования с использованием случайных лесов очень трудоемкий по сравнению с другими алгоритмами.

9. Стохастический градиентный спуск (Stochastic Gradient Descent, или SGD) – это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под выпуклые Функции потерь (Loss Function), такие как Метод опорных векторов (SVM) и Логистическая регрессия (Logistic Regression).

Преимущества:

- Эффективность.
- Простота реализации (множество возможностей для настройки кода).

К недостаткам можно отнести:

- SGD требует ряда гиперпараметров, таких как параметр регуляризации и количество итераций.
- SGD чувствителен к масштабированию функций.

10. Регрессия по методу «лассо» (LASSO, Least Absolute Shrinkage and Selection Operator). Регрессия по методу наименьших квадратов часто может стать неустойчивой, то есть сильно зависящей от обучающих данных, что обычно является

проявлением тенденции к переобучению. Избежать такого переобучения помогает регуляризация – общий метод, заключающийся в наложении дополнительных ограничений на искомые параметры, которые могут предотвратить излишнюю сложность модели. Смысл процедуры заключается в «стягивании» в ходе настройки вектора коэффициентов таким образом, чтобы они в среднем оказались несколько меньше по абсолютной величине, чем это было бы при оптимизации по методу наименьших квадратов.

Метод регрессии «Лассо» заключается во введении дополнительного слагаемого регуляризации в функционал оптимизации модели, что часто позволяет получать более устойчивое решение.

Преимущества:

- Более точные и стабильные оценки истинных параметров.
- Уменьшение ошибок выборки и отсутствия выборки.

Метрики эффективности для регрессии оценивают отклонение (расстояние) между предсказанными значениями и реальными. Обычно метрики сравнивают данную модель с тривиальной – моделью, которая всегда предсказывает среднее реальное значение целевой переменной. Модели могут быть точны на 100%, но плохи они могут быть без ограничений.

Коэффициент детерминации (R2) показывает силу связи между двумя случайными величинами. Если модель всегда предсказывает точно, метрика равна 1. Для тривиальной модели – 0. Значение метрики может быть отрицательно, если модель предсказывает хуже, чем тривиальная. Это одна из немногих несимметричных метрик эффективности.

Именно коэффициент детерминации чаще всего используется как метрика по умолчанию, которую можно посмотреть при помощи метода score() у модели

регрессии. Этот метод принимает на вход саму обучающую выборку. Но более универсально будет использовать эту метрику независимо от модели. Данная метрика называется r2_score. При использовании этой функции ей надо передавать два вектора целевой переменной – сначала эмпирический, а вторым аргументом - теоретический.

Средняя абсолютная ошибка (mean absolute error, MAE) показывает среднее абсолютное отклонение предсказанных значений от реальных. Чем выше значение MAE, тем модель хуже. У идеальной модели MAE = 0. MAE очень легко интерпретировать – на сколько в среднем ошибается модель.

Средний квадрат ошибки (mean squared error, MSE) показывает средний квадрат отклонений предсказанных значений от реальных. Чем выше значение MSE, тем модель хуже. У идеальной модели MSE = 0. MSE больше учитывает сильные отклонения, но хуже интерпретируется, чем MAE.

Среднеквадратичная ошибка (root mean squared error, RMSE) – это, по сути, корень из MSE. Выражается в тех же единицах, что и целевая переменная. Чаще применяется при статистическом анализа данных. Данная метрика очень чувствительна к аномалиям и выбросам.

Средняя абсолютная процентная ошибка (mean absolute percentage error, MAPE). В ней каждое отклонение оценивается в процентах от истинного значения целевой переменной. Идеальная модель имеет MAPE = 0. Верхний предел – не ограничен. Эта метрика имеет одно критическое преимущество над остальными – с ее помощью можно сравнивать эффективность моделей на разных обучающих выборках. Ведь если мы возьмем классические метрики (например, MAE), то размер отклонений будет очевидно зависеть от самих данных. А в двух разных выборках и средняя величина скорее всего будет разная. Поэтому метрики MAE, MSE, RMSE не сопоставимы при сравнении предсказаний, сделанных на разных выборках.

Анализируя описанные выше метрики, можно сделать следующие выводы. R2 значение очень интуитивно понятно. Но исследования показывают, что R2 действителен только для линейной регрессии. Однако большинство моделей регрессии, такие, например, как дерево решений или KNN, являются нелинейными моделями. Для нелинейных моделей не следует полностью доверять $R2$. Предпочтительно всегда использовать $R2$ вместе с другими показателями, такими как MAE и RMSE. Когда необходимо уменьшить влияние выбросов, лучше использовать MAE, когда выбросы нельзя игнорировать, лучше использовать RMSE.

Для сравнения моделей будем использовать коэффициент детерминации R2 и средняя абсолютная ошибка MAE, которые показывают, насколько модель улавливает изменение объясняемой переменной и среднее абсолютное отклонение предсказанных значений от реальных.

1.3 Разведочный анализ данных

Разведочный анализ данных (Exploratory Data Analysis) играет важнейшую роль после получения набора данных и ставит своей целью выяснить, как с ним работать и получить требуемый результат.

Разведочный анализ данных – предварительное исследование датасета с целью определения его основных характеристик, взаимосвязей между признаками, а также сужения набора методов, используемых для создания модели машинного обучения.

Вычислительные методы разведочного анализа данных включают основные статистические методы, а также более сложные, специально разработанные методы многомерного анализа, предназначенные для отыскания закономерностей в многомерных данных. К основным методам разведочного статистического анализа

относится процедура анализа распределений переменных, просмотр корреляционных матриц с целью поиска коэффициентов, превосходящих по величине определенные пороговые значения, или анализ многовходовых таблиц частот.

Исследование предоставленных данных по композитам целесообразно начать с анализа статистических характеристик датасета (рисунок 4): количества элементов, среднего арифметическое, медианы, стандартного отклонения, минимального и максимального значения, перцентилей.

Ввод [18]:
```
# Статистические характеристики датасета
df.describe()
```

Out[18]:

	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	У нашив г
count	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000000	1023.000
mean	2.930366	1975.734888	739.923233	110.570769	22.244390	285.882151	482.731833	73.328571	2466.922843	218.423144	0.491
std	0.913222	73.729231	330.231581	28.295911	2.406301	40.943260	281.314690	3.118983	485.628006	59.735931	0.500
min	0.389403	1731.764635	2.436909	17.740275	14.254985	100.000000	0.603740	64.054061	1036.856605	33.803026	0.000
25%	2.317887	1924.155467	500.047452	92.443497	20.608034	259.066528	266.816645	71.245018	2135.850448	179.627520	0.000
50%	2.906878	1977.621657	739.664328	110.564840	22.230744	285.896812	451.864365	73.268805	2459.524526	219.198882	0.000
75%	3.552660	2021.374375	961.812526	129.730366	23.961934	313.002106	693.225017	75.356612	2767.193119	257.481724	1.000
max	5.591742	2207.773481	1911.536477	198.953207	33.000000	413.273418	1399.542362	82.682051	3848.436732	414.590628	1.000

Рисунок 4 - Статистические характеристики объединенного датасета

Следующий шаг – проверка датасета на пропущенные значения. Для этого посмотрим информацию по кол-ву данных в столбцах датасета. По результатам убеждаемся, что пропусков в данных нет.

```
Ввод [14]: # проверка на полноту данных по столбцам
           df.info()

<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 - Проверка на полноту данных

Важным способом описания переменной является форма ее распределения, которая показывает, с какой частотой значения переменной попадают в определенные интервалы. Обычно исследователя интересует, насколько точно распределение можно аппроксимировать нормальным.

Гистограмма – это классический инструмент визуализации, позволяющий качественно оценить различные характеристики распределения, в том числе оценить нормальность эмпирического распределения. На гистограмму также накладывается кривая распределения (рисунок 6).

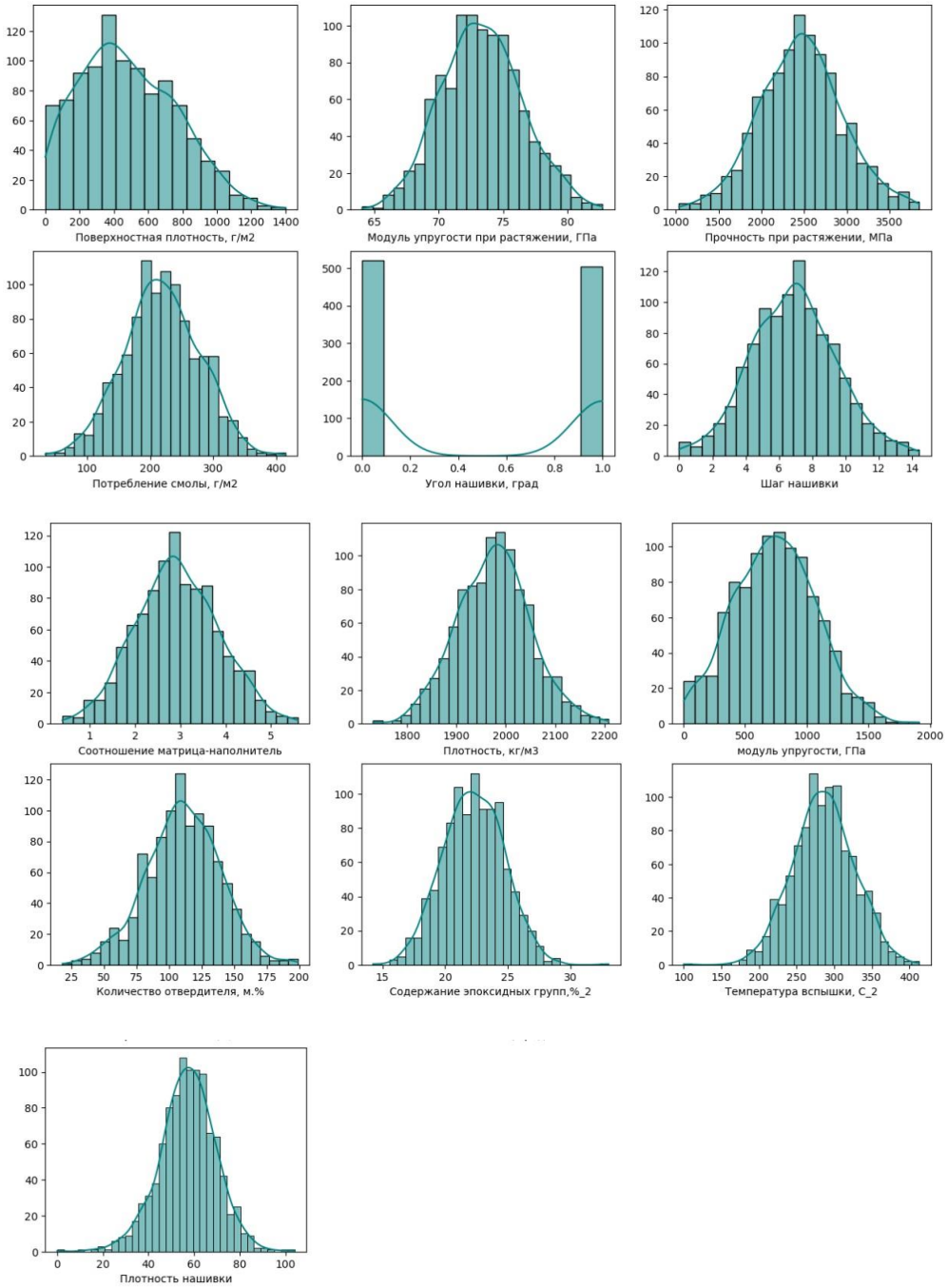

Рисунок 6 - Гистограммы и графики плотности

Анализ графиков показывает, что распределения всех параметров (кроме «Угла нашивки» и «Поверхностной плотности, г/м2») стремятся к нормальному. Параметр «Угол нашивки» имеет два значения: 0 и 1, что соответствует величинам

0 и 90 градусов. Т.к. другие значения данного признака в датасете не встречаются, будем считать его бинарным и категориальным.

Диаграмма «ящик с усами» – график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей. Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квантили, минимальное и максимальное значение выборки и выбросы.

Выброс (в статистике) – это измерительная точка данных, которая значительно выделяется из общей выборки. Выбросы могут быть вызваны вариативностью измерений или указывать на экспериментальную ошибку; в последнем случае они иногда исключаются из набора данных. Выброс может вызвать серьезные проблемы при статистическом анализе.

В boxplot можно считать нижний и верхний усики как о границы распределения данных. Любые точки данных, которые показывают выше или ниже усов, могут считаться выбросами или аномальными.

Из диаграмм «ящик с усами» (рисунок 7) видно, что выбросы имеют все характеристики, кроме параметра «Угол нашивки».

Влияние выбросов на модель:

- данные оказались в искаженном формате;

- изменяет общее статистическое распределение данных с точки зрения среднего значения, дисперсии и т.д;

- приводит к искажению уровня точности модели.

Из-за вышеуказанных причин необходимо обнаружить и избавиться от выбросов до моделирования набора данных.

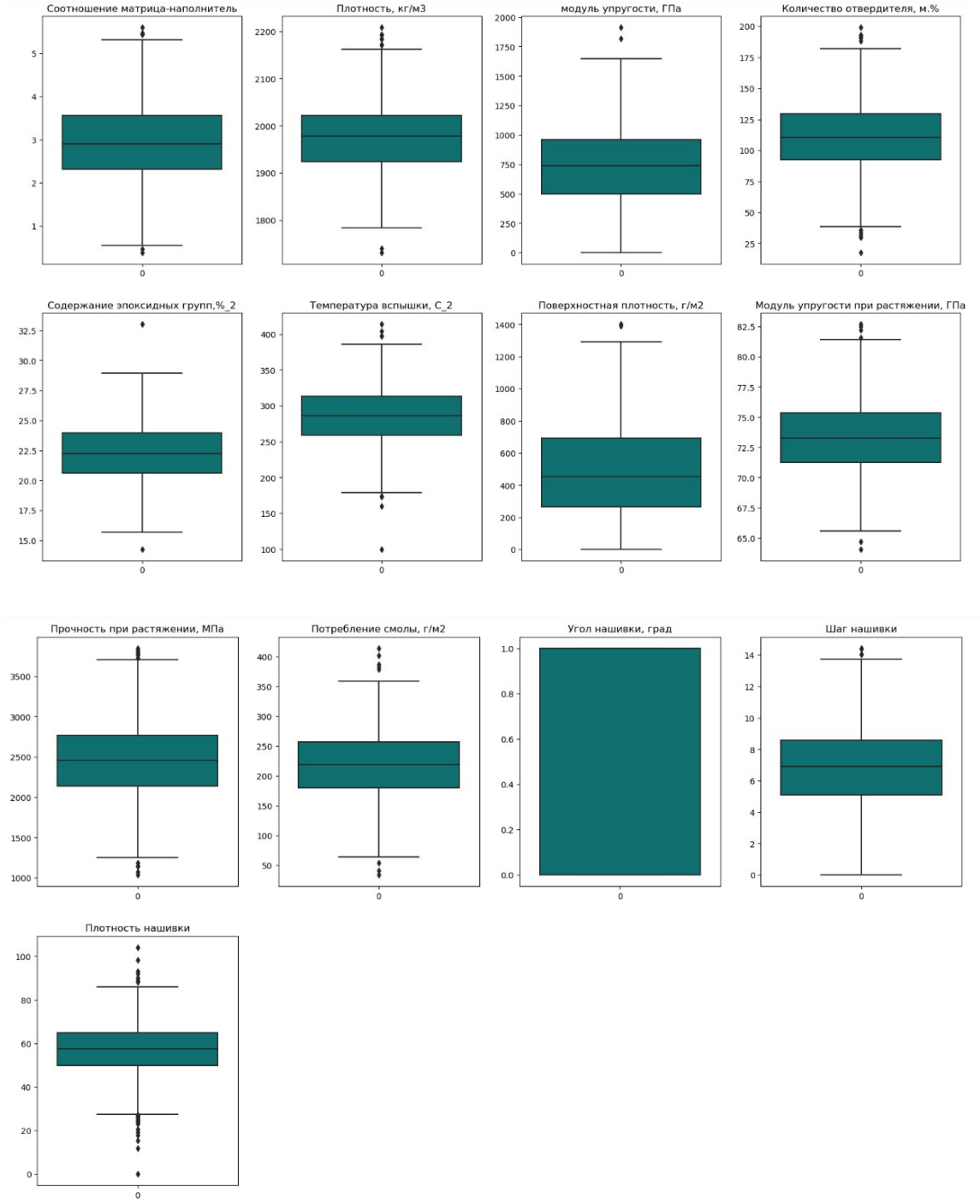

Рисунок 7 - Диаграммы «Ящик с усами»

Матричная диаграмма рассеяния (рисунок 8) представляет собой все возможные попарные диаграммы рассеяния, представленные в виде большой квадратной матрицы. Диагональные элементы матрицы являются графиками ядерной оценки

плотности распределения вероятности каждой из переменных. А остальные элементы – это диаграммы рассеяния переменных относительно друг друга.

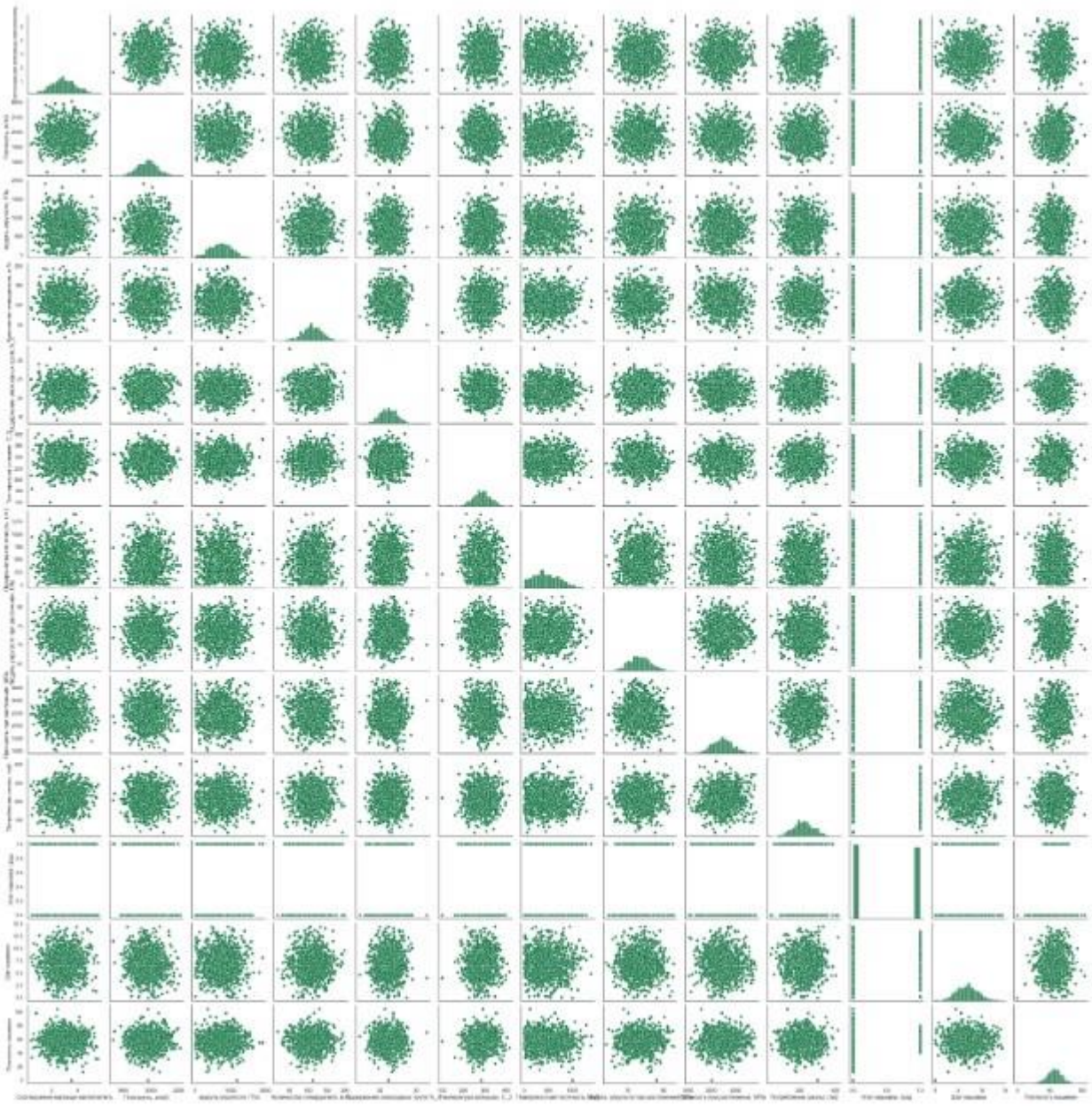

Рисунок 8 -Попарные графики рассеяния

Приведенные выше графики свидетельствуют об отсутствии линейной зависимости между характеристиками композитных материалов.

Далее необходимо установить взаимосвязи (корреляции) между переменными.

Корреляционная зависимость – это согласованные изменения двух (парная корреляционная связь) или большего количества признаков (множественная корреляционная связь). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой(-их) переменной(-ых).

Коэффициент корреляции – двумерная описательная статистика, количественная мера взаимосвязи (совместной изменчивости) двух переменных.

Значение коэффициента корреляции может варьироваться от -1 до 1, где -1 указывает на полную отрицательную связь, 0 указывает на отсутствие связи и 1 указывает на полную положительную связь.

В зависимости от того в каких шкалах измерены переменные, для установления наличия корреляционной связи используют следующие коэффициенты корреляции:

1. Коэффициент корреляции Пирсона (рисунок 9): используется, когда переменные измерены в шкале интервалов или отношений, эмпирические данные подчиняются нормальному закону распределения, число данных по каждому признаку одинаково.

   Недостатки коэффициента Пирсона:

   - для распределений, отличных от нормального, перестаёт быть эффективной оценкой коэффициента корреляции;
   - служит мерой только линейной взаимосвязи;
   - чувствителен к выбросам.

Ввод [20]:
```
# 1. Коэффициент корреляции Пирсона
df.corr(method = 'pearson')
```

Out[20]:

	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
Соотношение матрица-наполнитель	1.000000	0.003841	0.031700	-0.006445	0.019766	-0.004776	-0.006272	-0.008411	0.024148	0.072531
Плотность, кг/м3	0.003841	1.000000	-0.009647	-0.035911	-0.008278	-0.020695	0.044930	-0.017602	-0.069981	-0.015937
модуль упругости, ГПа	0.031700	-0.009647	1.000000	0.024049	-0.006804	0.031174	-0.005306	0.023267	0.041868	0.001840
Количество отвердителя, м.%	-0.006445	-0.035911	0.024049	1.000000	-0.000684	0.095193	0.055198	-0.065929	-0.075375	0.007446
Содержание эпоксидных групп,%_2	0.019766	-0.008278	-0.006804	-0.000684	1.000000	-0.009769	-0.012940	0.056828	-0.023899	0.015165
Температура вспышки, С_2	-0.004776	-0.020695	0.031174	0.095193	-0.009769	1.000000	0.020121	0.028414	-0.031763	0.059954
Поверхностная плотность, г/м2	-0.006272	0.044930	-0.005306	0.055198	-0.012940	0.020121	1.000000	0.036702	-0.003210	0.015692
Модуль упругости при растяжении, ГПа	-0.008411	-0.017602	0.023267	-0.065929	0.056828	0.028414	0.036702	1.000000	-0.009009	0.050938
Прочность при растяжении, МПа	0.024148	-0.069981	0.041868	-0.075375	-0.023899	-0.031763	-0.003210	-0.009009	1.000000	0.028602
Потребление смолы, г/м2	0.072531	-0.015937	0.001840	0.007446	0.015165	0.059954	0.015692	0.050938	0.028602	1.000000
Угол нашивки, град	-0.031073	-0.068474	-0.025417	0.038570	0.008052	0.020695	0.052299	0.023003	0.023398	-0.015334
Шаг нашивки	0.036437	-0.061015	-0.009875	0.014887	0.003022	0.025795	0.038332	-0.029468	-0.059547	0.013394
Плотность нашивки	-0.004652	0.080304	0.056346	0.017248	-0.039073	0.011391	-0.049923	0.006476	0.019604	0.012239

Рисунок 9 - Коэффициент корреляции Пирсона

2. Коэффициент корреляции рангов Спирмена: используется для измерения взаимозависимости между признаками, значения которых могут быть упорядочены или проранжированы по степени убывания (или возрастания).
3. Коэффициент корреляции Кендалла: применяется, когда переменные измерены в ранговой шкале, но размер выборки мал. Число данных по каждому признаку должно быть одинаковым, не допускается использование равных рангов.

Коэффициенты корреляции Спирмена и Кендалла используются как меры взаимозависимости между рядами рангов, а не как меры связи между самими переменными.

Коэффициенты Спирмена и Кендалла обладают примерно одинаковыми свойствами, но коэффициент корреляции Кендалла в случае многих рангов, а также при введении дополнительных объектов в ходе исследования имеет определенные вычислительные преимущества.

Для визуализации матриц корреляции можно воспользоваться тепловой картой (рисунок 10).

```
Ввод [26]: # Визуализация матрицы корреляции в виде тепловой карты
f, ax = plt.subplots(figsize=(12,10))
sns.heatmap(df.corr(), annot=True, ax=ax, square = True, cmap='RdPu')
plt.show()
```

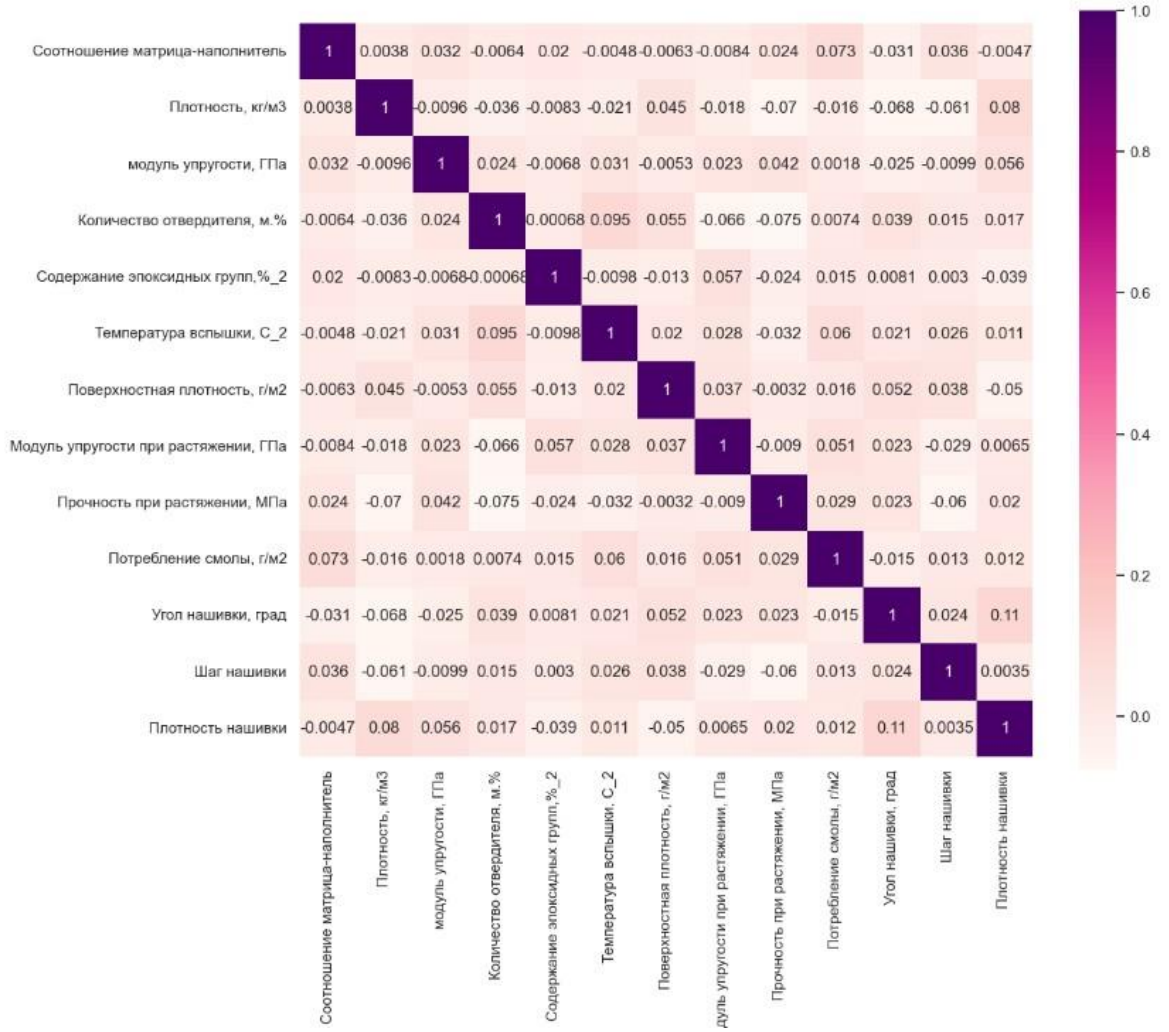

Рисунок 10 - Тепловая карта

Из приведенных выше данных видно, что наибольшая корреляция наблюдается между «Плотностью нашивки» и «Углом нашивки» и равняется 0,11. Другие коэффициенты корреляции близки к 0 и, соответственно, корреляционная зависимость между переменными крайне слабая.

2. Практическая часть

2.1 Предобработка данных

В области науки о данных и машинного обучения предварительная обработка значений данных является важным шагом. Первым шагом предварительной обработкой рассмотрим удаление всех выбросов из данных до моделирования.

Для поиска выбросов воспользуемся двумя методами:

- стандартное отклонение (правило трех сигм);
- метод межквартильного диапазона (IQR - Inter Quartile Range).

Метод 1. Стандартное отклонение (правило трех сигм) (рисунок 13): когда данные подчиняются нормальному распределению, 99,7% значений должны находиться в пределах 3 стандартных отклонений от среднего; когда значение превышает это расстояние, это можно считать выбросом.

Для реализации этого метода применяют z-оценку. Z-оценка показывает количество стандартных отклонений данного значения от среднего. Формула (1) для расчета z-показателя:

$$z = (X - \mu) / \sigma \quad (1)$$

где: X – это одно необработанное значение данных;

μ – среднее значение населения;

σ – стандартное отклонение населения.

Наблюдение можно идентифицировать как выброс (формула (2)), если его z-оценка меньше -3 или больше 3.

Выбросы = наблюдения с z-показателями > 3 или < -3 (2)

Метод 2. Метод межквартильного диапазона (IQR - Inter Quartile Range) (рисунок 14): Межквартильный размах (IQRб) – это разница между 75-м процентилем и 25-м процентилем в наборе данных (формула (3)). Он измеряет разброс средних 50% значений.

Другими словами, IQR – это первый квартиль (Q1), вычтенный из третьего квартиля (Q3); эти квартили можно четко увидеть на диаграмме «ящик с усами».

$$IQR = Q3 - Q1 \quad (3)$$

Наблюдение можно рассматривать как выброс (формула (4), если оно в 1,5 раза превышает межквартильный размах, превышающий третий квартиль (Q3), или в 1,5 раза превышает межквартильный размах, меньше первого квартиля (Q1).

$$\text{Выбросы} = \text{наблюдения} > Q3 + 1{,}5IQR \text{ или } Q1 - 1{,}5IQR \quad (4)$$

Ввод [27]:
```
# Метод межквартильного диапазона, предварительная оценка кол-ва выбросов
def detect_outliers_IQR(data):
    outliers = []
    Q1 = data.quantile(0.25)
    Q3 = data.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    for i in data:
        if (i <= lower_bound) | (i >= upper_bound):
            outliers.append(i)
    return outliers

sum_of_outliers = 0
for col in df.columns:
    df_outliers_IQR = detect_outliers_IQR(df[col])
    sum_of_outliers += len(df_outliers_IQR)
    print("Выбросов в столбце ", df[col].name, ": ", len(df_outliers_IQR))
print("Итого выбросов: ", sum_of_outliers)
```

```
Выбросов в столбце  Соотношение матрица-наполнитель :  6
Выбросов в столбце  Плотность, кг/м3 :  9
Выбросов в столбце  модуль упругости, ГПа :  2
Выбросов в столбце  Количество отвердителя, м.% :  14
Выбросов в столбце  Содержание эпоксидных групп,%_2 :  2
Выбросов в столбце  Температура вспышки, С_2 :  8
Выбросов в столбце  Поверхностная плотность, г/м2 :  2
Выбросов в столбце  Модуль упругости при растяжении, ГПа :  6
Выбросов в столбце  Прочность при растяжении, МПа :  11
Выбросов в столбце  Потребление смолы, г/м2 :  8
Выбросов в столбце  Угол нашивки, град :  0
Выбросов в столбце  Шаг нашивки :  4
Выбросов в столбце  Плотность нашивки :  21
Итого выбросов:  93
```

Рисунок 11- Оценка кол-ва выбросов при помощи метода межквартильного диапазона

При использовании метода трех сигм было выявлено 25 выбросов, при использовании метода межквартильного диапазона - 93 выброса.

Подход к поиску выбросов в интервале между квартилями является наиболее часто используемым и наиболее надежным подходом, применяемым в области исследований. Воспользуемся результатами второго метода.

Выбросы из данных удалены после трех этапов очистки (рисунок 12). Размерность датасета стала (922, 13).

```
Ввод [35]: sum_of_outliers_4 = 0
           for col in df_clean.columns:
               df_outliers_IQR = detect_outliers_IQR(df_clean[col])
               sum_of_outliers_4 += len(df_outliers_IQR)
               print("Выбросов в столбце ", df_clean[col].name, ": ", len(df_outliers_IQR))
           print("Итого выбросов: ", sum_of_outliers_4)

           Выбросов в столбце  Соотношение матрица-наполнитель :  0
           Выбросов в столбце  Плотность, кг/м3 :  0
           Выбросов в столбце  модуль упругости, ГПа :  0
           Выбросов в столбце  Количество отвердителя, м.% :  0
           Выбросов в столбце  Содержание эпоксидных групп,%_2 :  0
           Выбросов в столбце  Температура вспышки, С_2 :  0
           Выбросов в столбце  Поверхностная плотность, г/м2 :  0
           Выбросов в столбце  Модуль упругости при растяжении, ГПа :  0
           Выбросов в столбце  Прочность при растяжении, МПа :  0
           Выбросов в столбце  Потребление смолы, г/м2 :  0
           Выбросов в столбце  Угол нашивки, град :  0
           Выбросов в столбце  Шаг нашивки :  0
           Выбросов в столбце  Плотность нашивки :  0
           Итого выбросов:  0
```

Рисунок 12 - Результаты очистки датасета от выбросов при помощи метода IQR

Тестирование данных на нормальность часто является первым этапом их анализа, так как большое количество статистических методов исходит из предположения нормальности распределения изучаемых данных.

Многие алгоритмы машинного обучения построены на предположении о Гауссовом (нормальном) распределении входных данных. Поэтому для качественной работы таких моделей, обязательна проверка данных на нормальность, а при необходимости – приведение их к распределению, близкому к нормальному.

Проверку выборки на нормальность можно производить несколькими способами:

- построение гистограммы признака;

- построение QQ-графика;

- проведение теста на нормальность.

Тест Шапиро-Уилка (рисунок 13) используется для определения того, соответствует ли выборка нормальному распределению. Если p-значение ниже определенного уровня значимости (0,05), то у нас есть достаточно доказательств, чтобы сказать, что данные выборки не получены из нормального распределения.

Ввод [41]:
```
# Проведем тест Шапиро-Уилка на нормальность
for col in df_clean.columns:
    print(df_clean[col].name, shapiro(df_clean[col]))
```

```
Соотношение матрица-наполнитель ShapiroResult(statistic=0.9971880316734314, pvalue=0.10904344916343689)
Плотность, кг/м3 ShapiroResult(statistic=0.997011661529541, pvalue=0.08352091163396835)
модуль упругости, ГПа ShapiroResult(statistic=0.995254397392273, pvalue=0.0058130305260419846)
Количество отвердителя, м.% ShapiroResult(statistic=0.9966756105422974, pvalue=0.05000615119934082)
Содержание эпоксидных групп,%_2 ShapiroResult(statistic=0.9977133870124817, pvalue=0.23541033267974854)
Температура вспышки, С_2 ShapiroResult(statistic=0.9971266984939575, pvalue=0.09941922873258591)
Поверхностная плотность, г/м2 ShapiroResult(statistic=0.9776217937469482, pvalue=1.0688074730813568e-10)
Модуль упругости при растяжении, ГПа ShapiroResult(statistic=0.9955782890319824, pvalue=0.009416559711098671)
Прочность при растяжении, МПа ShapiroResult(statistic=0.9973730444908142, pvalue=0.14375117421150208)
Потребление смолы, г/м2 ShapiroResult(statistic=0.9955164790153503, pvalue=0.008584197610616684)
Угол нашивки, град ShapiroResult(statistic=0.6364129781723022, pvalue=2.8589291269154918e-40)
Шаг нашивки ShapiroResult(statistic=0.9980173110961914, pvalue=0.3559962809085846)
Плотность нашивки ShapiroResult(statistic=0.9967383742332458, pvalue=0.05505112186074257)
```

Рисунок 13 - Результаты теста Шапиро-Уилка до нормализации

В данном случае нулевая гипотеза отвергается (p-value < 0.05) в случаях:

- модуль упругости, ГПа;

- поверхностная плотность, г/м2;

- модуль упругости при растяжении, ГПа;

- потребление смолы, г/м2;

- угол нашивки;

Однако, проверка нормальности статистическими тестами является очень строгой, т.к. идет сравнение с идеальным распределением. Поэтому несмотря на то, что статистический тест говорит о ненормальности распределения, стоит посмотреть на гистограмму распределения (рисунок 6), либо QQ тест (рисунок 14).

График QQ, сокращение от графика «квантиль-квантиль», используется для оценки того, потенциально ли набор данных получен из некоторого теоретического распределения.

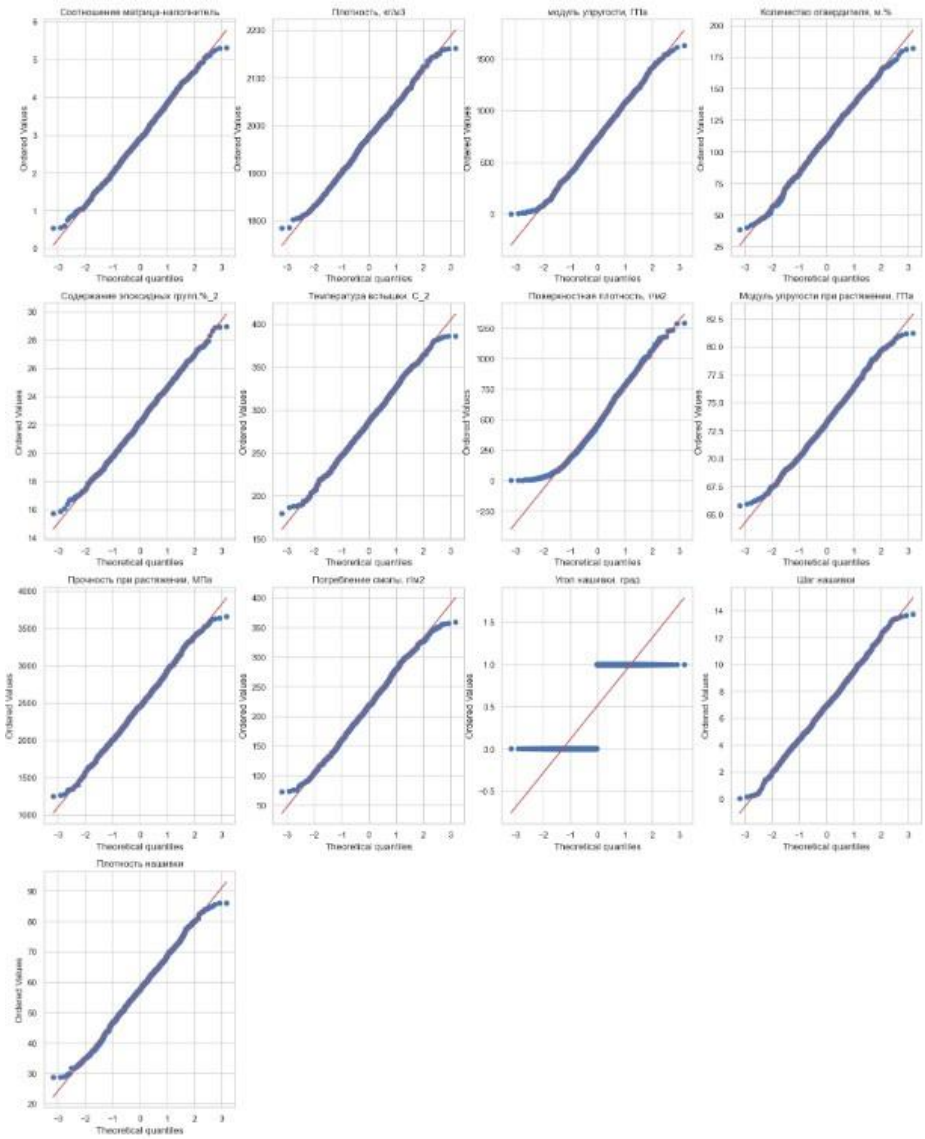

Рисунок 14 - QQ графики до нормализации

В большинстве случаев этот тип графика используется для определения того, соответствует ли набор данных нормальному распределению. Если данные распределены нормально, точки на графике QQ будут лежать на прямой диагональной линии. И наоборот, чем больше точки на графике значительно отклоняются от прямой диагональной линии, тем менее вероятно, что набор данных следует нормальному распределению.

Из графиков видно, что большинство данных близки к нормальному распределению, которое на данном графике представляет красная линия.

Алгоритмы машинного обучения, как правило, работают лучше или сходятся быстрее, когда различные переменные примерно одинаковый масштаб и близки к нормальному распределению. Поэтому общепринятой практикой является нормализация и стандартизация данных перед обучением на них моделей машинного обучения. Кроме того, в данном датасете присутствуют численные признаки разных масштабов, поэтому необходимо отмасштабировать признаки (рисунок 15).

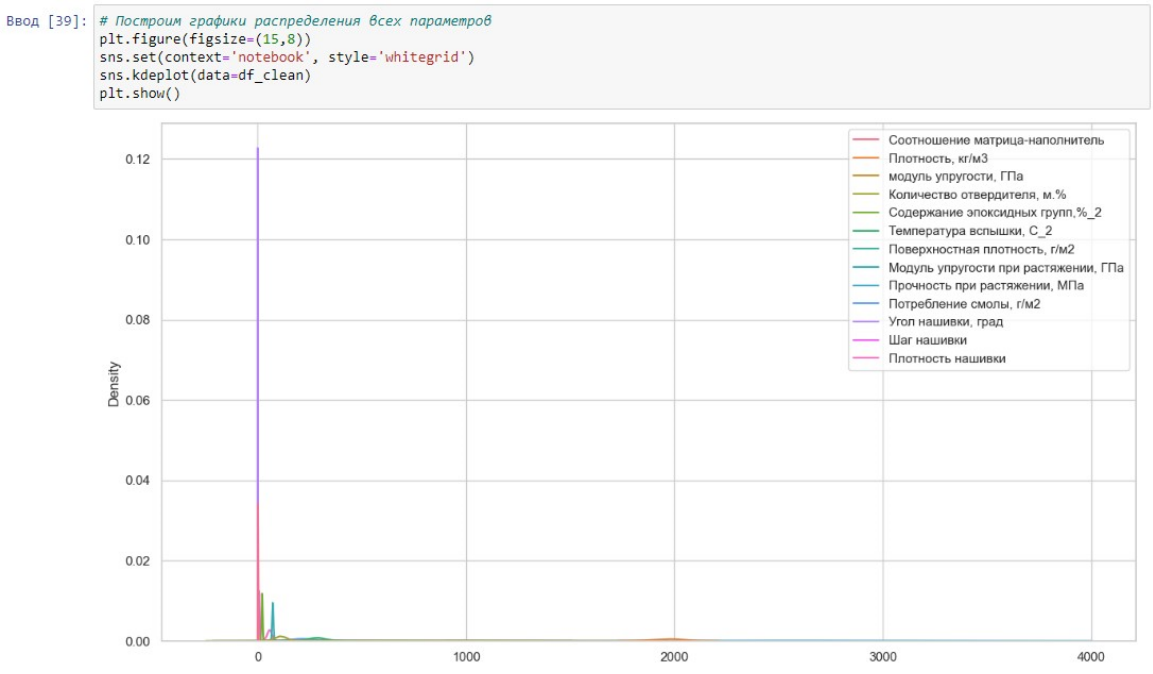

Рисунок 15 - Оценка масштаба распределения данных исходного датасета

Нормализация – процедура предобработки входной информации (обучающих, тестовых и валидационных выборок, а также реальных данных), при которой значения признаков во входном векторе приводятся к некоторому заданному диапазону, например, [0...1] или [-1...1].

Ключевая цель нормализации – приведение данных в самых разных единицах измерения и диапазонных значениях к единому виду, который позволит сравнить данные между собой и использовать для расчета схожести объектов в выборке. До начала обучения модели необходимо привести все признаки к равному влиянию друг на друга. В Python-библиотеке Scikit-learn есть для этого классы MinMaxScaler и RobustScaler.

Стандартизация – приведения данных к определенному формату и представлению, которые обеспечивают их корректное применение в многомерном анализе. Стандартизация позволяет устранить возможное влияние отклонений по каждому признаку и приводит все исходные значения в датасете к набору значений к нормальному распределению с математическим ожиданием равным 0 и стандартным отклонением равным 1. В результате получается стандартизированная шкала, которая определяет место каждого значения в наборе данных и измеряет его отклонение от среднего в единицах стандартного отклонения.

Недостатком стандартизации является возможность присутствия в этих шкалах отрицательных значений, что может привести к потере логики анализа данных. Отрицательные значения могут исключаться путем дополнительных преобразований.

Для стандартизации данных в Scikit-learn есть класс StandardScaler, который применяет к каждому из атрибутов следующее: вычитает из значений среднее и делит полученную разность на стандартное отклонение.

Для работы с датасетом был выбран метод PowerTransformer. Преобразование удаляет сдвиг из распределения данных, чтобы сделать распределение более нормальным (гауссовским). Популярными примерами являются лог-преобразование (положительные значения) или обобщенные версии, такие как преобразование Бокса-Кокса (положительные значения) или преобразование Йео-Джонсона (положительные и отрицательные значения).

После форматирования данных с помощью функции MinMaxScaler все параметры имеют одинаковый относительный масштаб (рисунки 20-21). Относительные пробелы между значениями каждого объекта были сохранены. Максимальное значение признака равно 1, минимальное - 0.

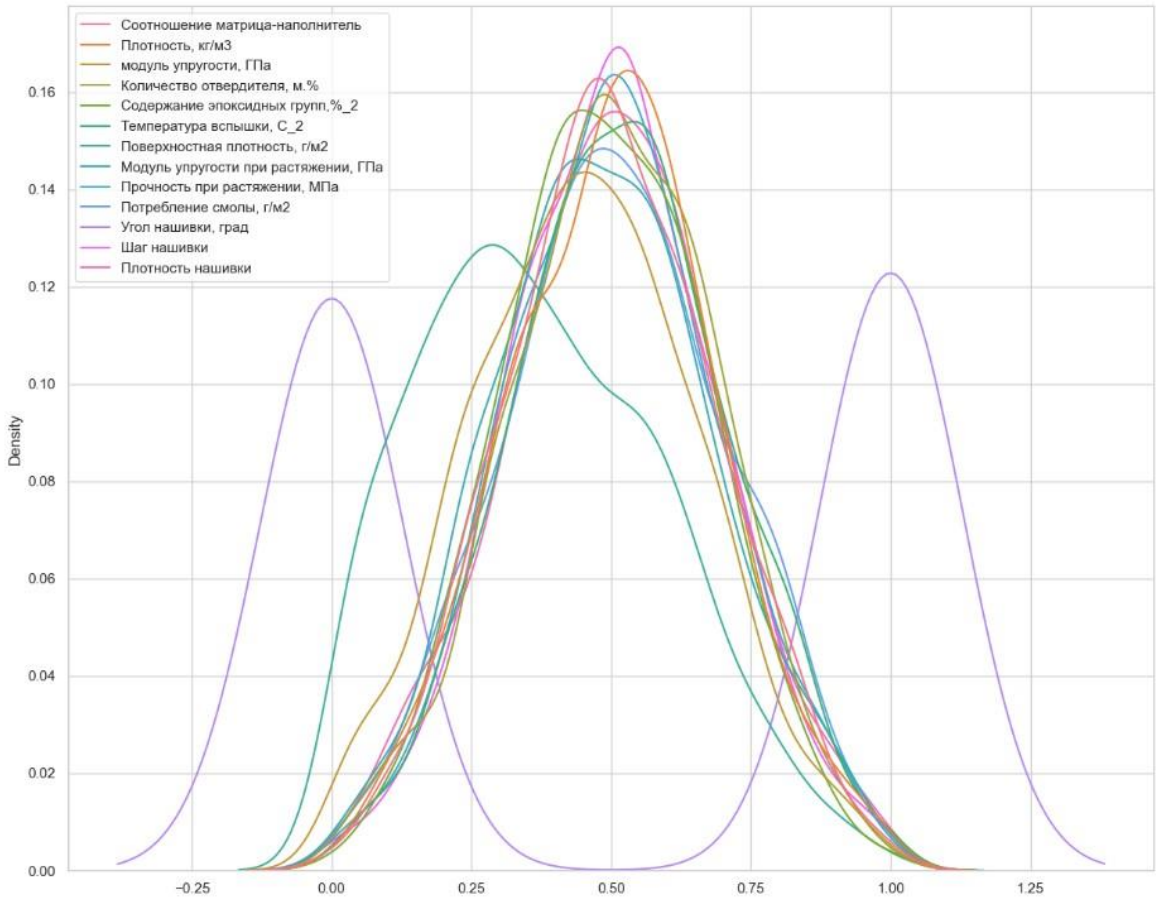

Рисунок 16 - Оценка распределения данных после использования метода MinMaxScaler

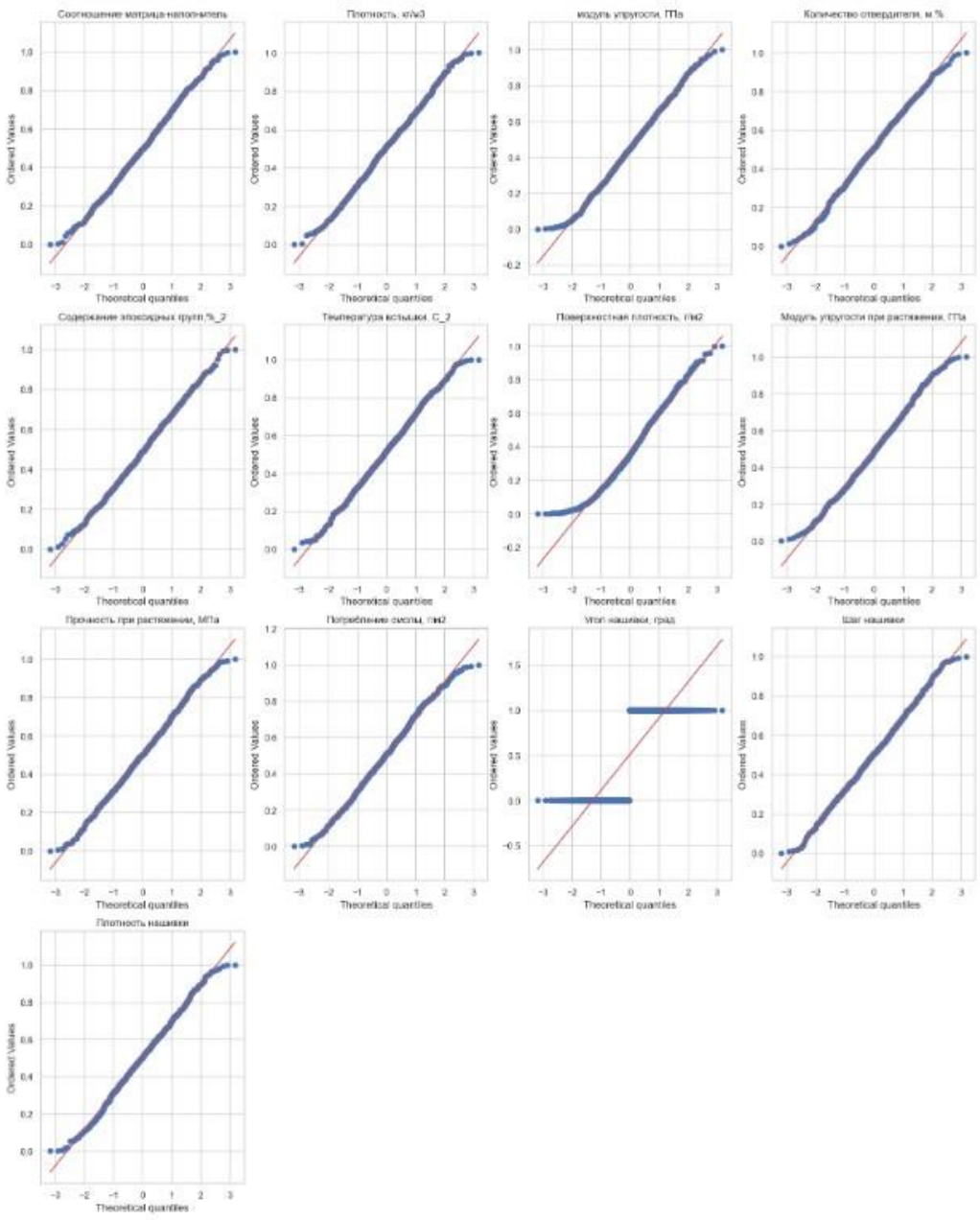

Рисунок 17 - QQ графики после масштабирования при помощи метода MinMaxScaler()

2.2 Разработка и обучение модели

Разработка модели машинного обучения для прогнозирования таких характеристик композитных материалов, как модуль упругости при растяжении и прочность при растяжении, включает следующие этапы:

1. Разделение нормализованных данных на обучающую и тестовую выборку: 30% данных оставлено на тестирование моделей, на остальных происходит обучение моделей.

2. Анализ работы различных моделей на стандартных параметрах. Для задачи прогнозирования модуля упругости при растяжении и прочности при растяжении используем описанные ранее модели (рисунки 18 - 27).

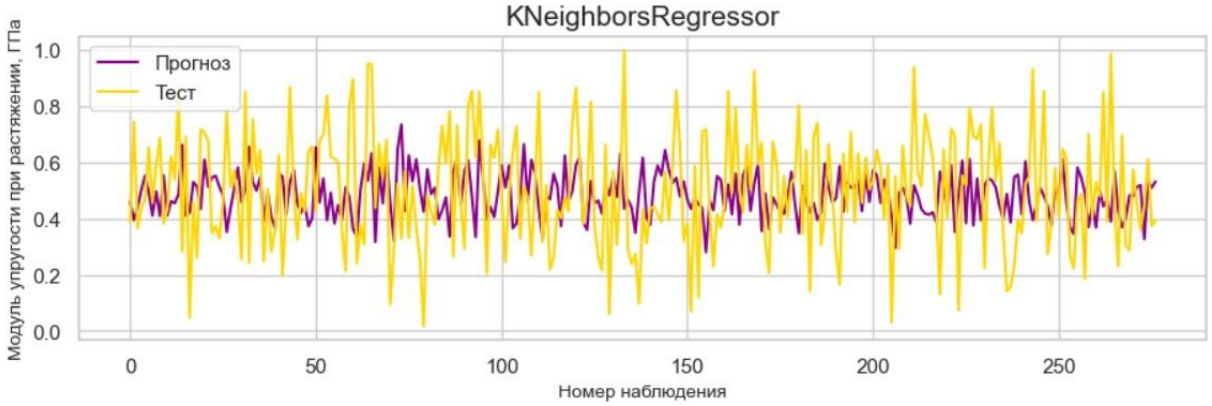

Рисунок 18 - Прогнозные и тестовые результаты: метод К-ближайших соседей

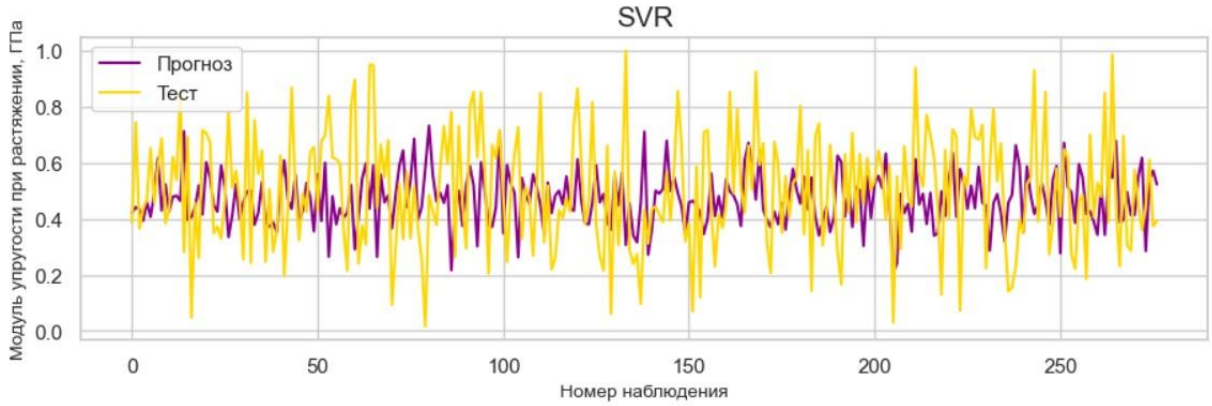

Рисунок 19 - Прогнозные и тестовые результаты: метод опорных векторов

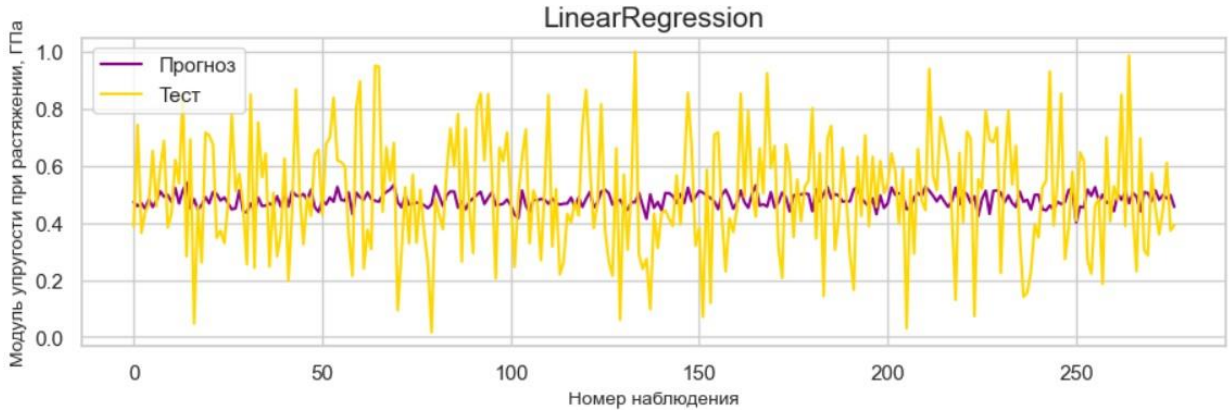

Рисунок 20 - Прогнозные и тестовые результаты: линейная регрессия

Рисунок 21 - Прогнозные и тестовые результаты: дерево решений

Рисунок 22 - Прогнозные и тестовые результаты: AdaBoost

Рисунок 23 - Прогнозные и тестовые результаты: градиентный бустинг

Рисунок 24 - Прогнозные и тестовые результаты: XG Boost

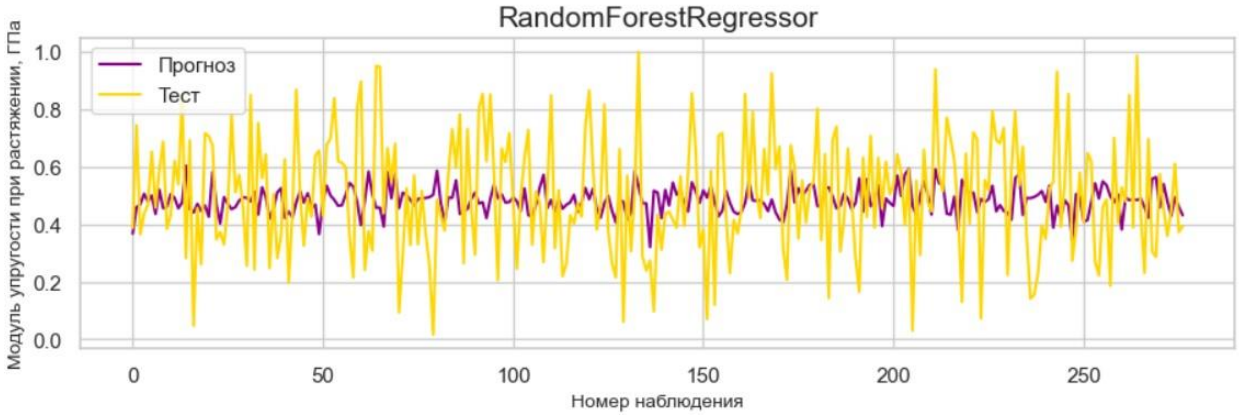

Рисунок 25 - Прогнозные и тестовые результаты: случайный лес

Рисунок 26 - Прогнозные и тестовые результаты: стохастический градиентный спуск

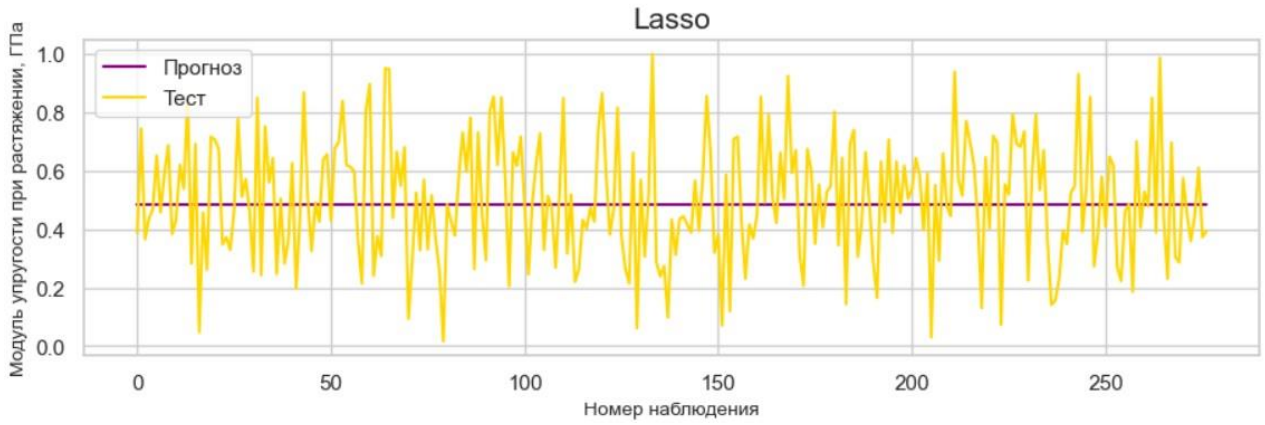

Рисунок 27 - Прогнозные и тестовые результаты: метод «Лассо»

3. Сравнение моделей по следующим метрикам: коэффициент детерминации R2, средняя абсолютная ошибка (MAE), средний квадрат ошибки (MSE) и средне-квадратичная ошибка (RMSE) (рисунок 28).

Ввод [66]:
```
# Рассчитаем метрики для следующих моделей:
KNR = KNeighborsRegressor()
SVReg = SVR()
LR = LinearRegression()
DTR = DecisionTreeRegressor()
Abr = AdaBoostRegressor(base_estimator=DecisionTreeRegressor())
Gbr = GradientBoostingRegressor()
XgbR = XGBRegressor()
RFR = RandomForestRegressor() #n_estimators = 20, max_depth = 10, random_state = 42)
SGDR = SGDRegressor()
LassoR = Lasso(alpha=0.1)

Model_Comparision_Train_Test([KNR, SVReg, LR, DTR, Abr, Gbr, XgbR, RFR, SGDR, LassoR], X_train_elastic, np.ravel(y_train_elastic)
```

Out[66]:

Model	R2_score	MAE	MSE	RMSE	MAPE
KNeighborsRegressor	-0.1435 (0.197)	0.18 (0.14)	0.05 (0.03)	0.22 (0.17)	0.65 (2088968184903.12)
SVR	-0.2554 (0.4649)	0.18 (0.11)	0.05 (0.02)	0.23 (0.14)	0.66 (1409875462373.38)
LinearRegression	-0.0245 (0.0172)	0.17 (0.15)	0.04 (0.04)	0.21 (0.19)	0.65 (3530183592566.66)
DecisionTreeRegressor	-1.0137 (1.0)	0.23 (0.0)	0.08 (0.0)	0.29 (0.0)	0.78 (0.0)
AdaBoostRegressor	-0.082 (0.9999)	0.17 (0.0)	0.05 (0.0)	0.21 (0.0)	0.64 (0.0)
GradientBoostingRegressor	-0.1124 (0.4548)	0.17 (0.11)	0.05 (0.02)	0.22 (0.14)	0.68 (1828505402718.78)
XGBRegressor	-0.1728 (0.9989)	0.18 (0.0)	0.05 (0.0)	0.22 (0.01)	0.7 (28797473498.31)
RandomForestRegressor	-0.025 (0.8494)	0.17 (0.06)	0.04 (0.01)	0.21 (0.07)	0.65 (1334496858305.19)
SGDRegressor	-0.0669 (-0.0241)	0.17 (0.16)	0.04 (0.04)	0.21 (0.19)	0.64 (3515840230841.68)
Lasso	-0.0089 (0.0)	0.17 (0.16)	0.04 (0.04)	0.21 (0.19)	0.65 (3362407323544.2)

Рисунок 28 - Метрики для тестовых и тренировочных данных (в скобках)

Отрицательные значение коэффициента детерминации означают слабую обобщающую способность моделей. Если R2 отрицательна, то модель работает хуже, чем простой подсчет среднего.

Лучшие показатели R2 и MAE на тестовой выборке у алгоритма регрессии «Lasso», на тренировочной выборке – у алгоритмов DecisionTreeRegressor и AdaBoostRegressor.

4. Поиск гиперпараметров моделей с помощью поиска по сетке с перекрестной проверкой.

Задачу выбора модели усложняет тот факт, что у многих типов моделей существуют разные вариации, особые параметры. Гиперпараметр модели – это численное значение, которое влияет на работу модели, но не подбирается в процессе обучения. Они задаются при определении модели и должны оставаться неизменными до схождения алгоритма обучения.

Поиск по сетке – полный перебор всех комбинаций значений гиперпараметров для поиска оптимальных значений. Для его организации надо задать список гиперпараметров и их конкретных значений (рисунок 29). Поиск по сетке имеет экспоненциальную сложность. Чем больше параметров и значений задать, тем лучше получится модель, но дольше поиск. Рекомендуется использовать кросс-валидацию. По умолчанию используется оценка модели, встроенная в сам объект модели через метод score, то есть точность (accuracy) для классификации и коэффициент детерминации (R2) для регрессии (рисунок 30).

```
# Создаем словарь с наборами гиперпараметров всех моделей
all_params = {'kneighborsregressor':       {'kneighborsregressor__n_neighbors': [i for i in range(1, 201, 2)],
                                            'kneighborsregressor__weights':     ['uniform', 'distance'],
                                            'kneighborsregressor__algorithm':   ['auto', 'ball_tree', 'kd_tree', 'brute']
                                           },
              'svr':                       {'svr__kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
                                            'svr__C':      [0.01, 0.1, 1],
                                            'svr__gamma':  [0.01, 0.1, 1]
                                           },
              'linearregression':          {'linearregression__fit_intercept': [True, False]},
              'decisiontreeregressor':     {'decisiontreeregressor__max_depth':         [3, 5, 7, 9, 11, 13, 15],
                                            'decisiontreeregressor__min_samples_leaf':  [1, 2, 5, 10, 20, 50, 100, 150, 200],
                                            'decisiontreeregressor__min_samples_split': [200, 250, 300],
                                            'decisiontreeregressor__max_features':      ['auto', 'sqrt', 'log2']
                                           },
              'adaboostregressor':         {'adaboostregressor__base_estimator__max_depth':        [i for i in range(2, 11,
                                            'adaboostregressor__base_estimator__min_samples_leaf': [5,  10],
                                            'adaboostregressor__n_estimators':                     [10, 50, 100, 250, 1000],
                                            'adaboostregressor__learning_rate':                    [0.01, 0.05, 0.1, 0.5]
                                           },
              'gradientboostingregressor': {'gradientboostingregressor__learning_rate': [0.01,  0.02, 0.03, 0.04],
                                            'gradientboostingregressor__subsample'    : [0.9, 0.5, 0.2, 0.1],
                                            'gradientboostingregressor__n_estimators' : [100, 500, 1000, 1500],
                                            'gradientboostingregressor__max_depth'    : [4, 6, 8, 10]
                                           },
              'xgbregressor':              {'xgbregressor__learning_rate':    [0.05, 0.10, 0.15],
                                            'xgbregressor__max_depth':        [ 3, 4, 5, 6, 8],
                                            'xgbregressor__min_child_weight': [ 1, 3, 5, 7],
                                            'xgbregressor__gamma':            [ 0.0, 0.1, 0.2],
                                            'xgbregressor__colsample_bytree': [ 0.3, 0.4]
                                           },
              'randomforestregressor':     {'randomforestregressor__n_estimators':     [30, 100, 200, 300, 500],
                                            'randomforestregressor__max_depth':        [1, 2, 3, 4, 5, 6, 7, 8],
                                            'randomforestregressor__min_samples_leaf': [1, 2],
                                            'randomforestregressor__max_features':     ['auto', 'sqrt', 'log2']
                                           },
              'sgdregressor':              {'sgdregressor__penalty': ['l2', 'l1', 'elasticnet', None],
                                            'sgdregressor__alpha':   [0.0001, 0.001, 0.01, 0.1]
                                           },
              'lasso':                     {'lasso__alpha': [0.01, 0.02, 0.1, 0.2, 0.03, 0.3, 0.05, 0.5, 0.07, 0.7, 1]}
```

Рисунок 29 - Набор гиперпараметров для всех моделей

```
Ввод [69]: Model_Selection(X_train_elastic, np.ravel(y_train_elastic), X_test_elastic, np.ravel(y_test_elastic))

KNeighborsRegressor() Лучшее значение R2 на тренировочной выборке: 0.0091
KNeighborsRegressor() Лучшее значение R2 на тестовой выборке -0.0105
KNeighborsRegressor() Лучшее значение R2 на перекрестной проверке: -0.0021
KNeighborsRegressor() Лучшие параметры модели: {'kneighborsregressor__algorithm': 'auto', 'kneighborsregressor__n_neighbors':
199, 'kneighborsregressor__weights': 'uniform'}

_____________________________________________________________________________________________

SVR() Лучшее значение R2 на тренировочной выборке: 0.0326
SVR() Лучшее значение R2 на тестовой выборке -0.0146
SVR() Лучшее значение R2 на перекрестной проверке: -0.0031
SVR() Лучшие параметры модели: {'svr__C': 0.01, 'svr__gamma': 1, 'svr__kernel': 'rbf'}

_____________________________________________________________________________________________

LinearRegression() Лучшее значение R2 на тренировочной выборке: 0.0172
LinearRegression() Лучшее значение R2 на тестовой выборке -0.0245
LinearRegression() Лучшее значение R2 на перекрестной проверке: -0.0307
LinearRegression() Лучшие параметры модели: {'linearregression__fit_intercept': True}
```

Рисунок 30 - Пример вывода лучших гиперпараметров для моделей

Лучшим алгоритмом для прогнозирования модуля упругости при растяжении при использовании функции GridSearchCV() выбран регрессор AdaBoostRegressor со значением R2 = -0.0025 на тестовой выборке (рисунок 31).

```
_____________________________________________________________________________________________

Регрессор с лучшим значением R2 = -0.0025 на тестовой выборке: AdaBoostRegressor(base_estimator=DecisionTreeRegressor())

Лучший алгоритм:
Pipeline(steps=[('adaboostregressor',
                 AdaBoostRegressor(base_estimator=DecisionTreeRegressor(max_depth=2,
                                                                        min_samples_leaf=10),
                                   learning_rate=0.5, n_estimators=10))])
```

Рисунок 31 - Вывод лучшего алгоритма прогнозирования параметров

2.3 Тестирование модели

После определения лучших параметров для каждой модели произведено тестирование моделей на тренировочном и тестовом наборе данных.

Для сравнения моделей будем использовать коэффициент детерминации R2 и средняя абсолютная ошибка MAE, которые показывают, насколько модель

улавливает изменение объясняемой переменной и среднее абсолютное отклонение предсказанных значений от реальных.

Коэффициенты детерминации (R2) отрицательные у всех моделей прогнозирования модуля упругости при растяжении (рисунок 32). Это означает, что прогнозы моделей хуже, чем предсказание среднего значения. Модели плохо обучаются на тренировочной выборке и, соответственно, плохо предсказывают значения для тестовой выборки. Коэффициент MAE одинаков у всех моделей.

Ввод [70]:
```
# Рассчитаем метрики для моделей с их лучшими параметрами:
KNR_best = KNeighborsRegressor(algorithm='auto', n_neighbors=199, weights='uniform')
SVReg_best = SVR(C=0.01, gamma=1, kernel='rbf')
LR_best = LinearRegression(fit_intercept=True)
DTR_best = DecisionTreeRegressor(max_depth=9, max_features='sqrt', min_samples_leaf=200, min_samples_split=250)
Abr_best = AdaBoostRegressor(base_estimator=DecisionTreeRegressor(max_depth=2, min_samples_leaf=10), learning_rate=0.01, n_estima
Gbr_best = GradientBoostingRegressor(learning_rate=0.01, max_depth=4, n_estimators=100, subsample=0.5)
XgbR_best = XGBRegressor(colsample_bytree=0.3, gamma=0.2, learning_rate=0.05, max_depth=3, min_child_weight=5)
RFR_best = RandomForestRegressor(max_depth=1, max_features='log2', min_samples_leaf=2, n_estimators=100)
SGDR_best = SGDRegressor(alpha=0.1, penalty='l1')
LassoR_best = Lasso(alpha=0.01)

Model_Comparision_Train_Test([KNR_best, SVReg_best, LR_best, DTR_best, Abr_best, Gbr_best, XgbR_best, RFR_best, SGDR_best, LassoR
```

Out[70]:

Model	R2_score	MAE	MSE	RMSE	MAPE
KNeighborsRegressor	-0.0105 (0.0091)	0.17 (0.15)	0.04 (0.04)	0.21 (0.19)	0.65 (3324567063626.71)
SVR	-0.0146 (0.0326)	0.17 (0.15)	0.04 (0.04)	0.21 (0.19)	0.65 (3293971898525.6)
LinearRegression	-0.0245 (0.0172)	0.17 (0.15)	0.04 (0.04)	0.21 (0.19)	0.65 (3530183592566.66)
DecisionTreeRegressor	0.006 (0.0061)	0.17 (0.15)	0.04 (0.04)	0.2 (0.19)	0.65 (3479792181714.39)
AdaBoostRegressor	-0.0066 (0.0228)	0.17 (0.15)	0.04 (0.04)	0.2 (0.19)	0.65 (3231476794929.06)
GradientBoostingRegressor	-0.017 (0.1559)	0.17 (0.14)	0.04 (0.03)	0.21 (0.18)	0.65 (2959719650427.48)
XGBRegressor	-0.0218 (0.0918)	0.17 (0.15)	0.04 (0.03)	0.21 (0.18)	0.66 (3274354675248.97)
RandomForestRegressor	-0.0106 (0.0171)	0.17 (0.15)	0.04 (0.04)	0.21 (0.19)	0.66 (3366865472577.18)
SGDRegressor	-0.0092 (-0.0)	0.17 (0.16)	0.04 (0.04)	0.21 (0.19)	0.65 (3359840987530.74)
Lasso	-0.0089 (0.0)	0.17 (0.16)	0.04 (0.04)	0.21 (0.19)	0.65 (3362407323544.2)

Рисунок 32 - Метрики для тестовых и тренировочных данных после подбора гиперпараметров для прогнозирования модуля упругости при растяжении

Все использованные модели плохо справились с поставленной задачей прогнозирования модуля упругости при растяжении. Получен неудовлетворительный результат.

При разработке модели машинного обучения для прогнозирования прочности при растяжении также получены модели со слабой обобщающей способностью (рисунок 33). Полученный результат не решает поставленную задачу.

Ввод [79]:
```
# Рассчитаем метрики для моделей с их лучшими параметрами:
KNR_best = KNeighborsRegressor(algorithm='auto', n_neighbors=159, weights='distance')
SVReg_best = SVR(C=0.01, gamma=0.1, kernel='poly')
LR_best = LinearRegression(fit_intercept=True)
DTR_best = DecisionTreeRegressor(max_depth=15, max_features='sqrt', min_samples_leaf=100, min_samples_split=250)
Abr_best = AdaBoostRegressor(base_estimator=DecisionTreeRegressor(max_depth=2, min_samples_leaf=10), learning_rate=0.01, n_estima
Gbr_best = GradientBoostingRegressor(learning_rate=0.01, max_depth=8, n_estimators=100, subsample=0.1)
XgbR_best = XGBRegressor(colsample_bytree=0.3, gamma=0.2, learning_rate=0.05, max_depth=3, min_child_weight=5)
RFR_best = RandomForestRegressor(max_depth=1, max_features='log2', min_samples_leaf=2, n_estimators=30)
SGDR_best = SGDRegressor(alpha=0.1, penalty='l1')
LassoR_best = Lasso(alpha=0.01)

Model_Comparision_Train_Test([KNR_best, SVReg_best, LR_best, DTR_best, Abr_best, Gbr_best, XgbR_best, RFR_best, SGDR_best, Lassof
```

Out[79]:

Model	R2_score	MAE	MSE	RMSE	MAPE
KNeighborsRegressor	-0.002 (1.0)	0.15 (0.0)	0.04 (0.0)	0.19 (0.0)	8176160847752.77 (0.0)
SVR	-0.0012 (0.0012)	0.15 (0.15)	0.04 (0.03)	0.19 (0.19)	8151410332325.56 (0.69)
LinearRegression	-0.0035 (0.017)	0.15 (0.15)	0.04 (0.03)	0.19 (0.19)	9069348467609.18 (0.68)
DecisionTreeRegressor	-0.0116 (0.0284)	0.15 (0.15)	0.04 (0.03)	0.19 (0.18)	8145511233200.83 (0.68)
AdaBoostRegressor	-0.0072 (0.0501)	0.15 (0.15)	0.04 (0.03)	0.19 (0.18)	8684037540203.58 (0.67)
GradientBoostingRegressor	-0.0226 (0.161)	0.16 (0.14)	0.04 (0.03)	0.19 (0.17)	8432488024396.78 (0.63)
XGBRegressor	0.0005 (0.1155)	0.15 (0.14)	0.04 (0.03)	0.19 (0.18)	8709751775321.21 (0.65)
RandomForestRegressor	-0.0011 (0.0231)	0.15 (0.15)	0.04 (0.03)	0.19 (0.18)	8410926210985.66 (0.68)
SGDRegressor	-0.0002 (-0.0)	0.15 (0.15)	0.04 (0.03)	0.19 (0.19)	8168139964879.41 (0.69)
Lasso	-0.0001 (0.0)	0.15 (0.15)	0.04 (0.03)	0.19 (0.19)	8183534682026.43 (0.69)

Рисунок 33 - Метрики для тестовых и тренировочных данных после подбора гиперпараметров для прогнозирования прочности при растяжении.

Лучшим алгоритмом для прогнозирования прочности при растяжении при использовании функции GridSearchCV() выбран регрессор XGBRegressor со значением R2 = 0.0005 на тестовой выборке .

С учетом полученных неудовлетворительных результатов в качестве прогноза для модуля упругости при растяжении и прочности при растяжении можно использовать среднее значение признака.

2.4 Нейронная сеть для рекомендации соотношения «матрица – наполнитель»

Нейронная сеть – это метод в искусственном интеллекте, который учит компьютеры обрабатывать данные таким же способом, как и человеческий мозг. Это тип процесса машинного обучения, называемый глубоким обучением, который использует взаимосвязанные узлы или нейроны в слоистой структуре, напоминающей человеческий мозг. Он создает адаптивную систему, с помощью которой компьютеры учатся на своих ошибках и постоянно совершенствуются. Таким образом, искусственные нейронные сети пытаются решать сложные задачи с более высокой точностью.

Часто архитектуры нейронных сетей строят в виде последовательности слоев, начиная с входного и заканчивая выходным. Теоретически, число скрытых слоев может быть сколь угодно большим. Для описания такой модели, как раз применяется класс Sequential, который используется в нейронной сети для рекомендации соотношения «матрица – наполнитель».

Объект оболочки Keras для использования в качестве регрессионной оценки называется KerasRegressor. Создадим экземпляр и передадим ему как имя функции для создания модели нейронной сети, так и некоторые параметры для дальнейшей их передачи в функции компиляции и обучения модели. Далее с помощью функции GridSearchCV() произведем сравнение моделей нейронной сети (рисунок 34).

Ввод [83]:
```
# Создадим функцию для генерации слоев нейронной сети
def create_NN_model(layers, activation, drop, opt):
        model= Sequential()
        for i, neurons in enumerate(layers):
            if i==0:
                model.add(Dense(neurons, input_dim=X_train_matrix.shape[1], activation = activation))
            else:
                model.add(Dense(neurons, activation))
                model.add(Dropout(drop))
            model.add(Dense(1))

            model.compile(loss = 'mse', optimizer = opt, metrics = ['mae'])

            return model
```

Ввод [85]:
```
# Построим нейронную сети с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 5 (cv = 5)
# Воспользуемся методом GridSearchCV

reg = KerasRegressor(model = create_NN_model, layers = [128], activation = 'relu', drop = 0.1, opt = 'Adam', verbose = 2)

# Зададим параметры для модели
param_grid = {'activation': ['relu', 'softmax', 'sigmoid'],
                  'layers': [[128, 64, 16], [128, 128, 64, 32], [128, 128, 64, 16]],
                    'opt': ['Adam', 'SGD'],
                   'drop': [0.0, 0.1, 0.2],
              'batch_size': [10, 20, 40],
                  'epochs': [10, 50, 100]
                 }

# Произведем поиск лучших параметров
grid = GridSearchCV(estimator = reg,
                    param_grid = param_grid,
                    cv = 5,
                    verbose = 0,
                    n_jobs = -1)

grid_result = grid.fit(X_train_matrix, np.ravel(y_train_matrix))
```

Рисунок 34 - Построение нейронной сети с помощью поиска по сетке с перекрестной проверкой

Выбор модели осуществляется по лучшему значению коэффициента KerasRegressor.score(), который возвращает коэффициент детерминации прогноза (также известный как оценка R2) (рисунок 35).

Ввод [86]:
```
print('Лучший коэффициент R2: {:.4f} при использовании модели с параметрами {} \n'.format(grid_result.best_score_, grid_result.be
```

```
Лучший коэффициент R2: -0.0025 при использовании модели с параметрами {'activation': 'softmax', 'batch_size': 10, 'drop': 0.2,
'epochs': 50, 'layers': [128, 128, 64, 32], 'opt': 'SGD'}
```

Ввод [87]:
```
# Создадим модель с полученными значениями
best_model = Sequential()
best_model.add(Dense(128, input_dim = X_train_matrix.shape[1], activation = 'softmax'))      # входной слой
best_model.add(Dense(128, activation = 'softmax'))                                           # добавляем полносвязный слой
best_model.add(Dropout(0.0))                                                                 # исключаем переобучения
best_model.add(Dense(64, activation = 'softmax'))                                            # добавляем полносвязный слой
best_model.add(Dropout(0.0))                                                                 # исключаем переобучения
best_model.add(Dense(32, activation = 'softmax'))                                            # добавляем полносвязный слой
best_model.add(Dropout(0.0))                                                                 # исключаем переобучения
best_model.add(Dense(1))                                                                     # выходной слой

# Компиляция модели: определяем метрики и алгоритм оптимизации
best_model.compile(loss = 'mse',
                   optimizer = 'SGD',
                   metrics = ['mae'])

# Обучение модели
best_history = best_model.fit(X_train_matrix, np.ravel(y_train_matrix),
                              epochs=10,
                              batch_size=10,
                              verbose=1,
                              validation_split=0.2)
```

Рисунок 35 - Выбор и построение лучшей модели

Структура нейронной сети, выбранной с помощью функции KerasRegressor(), приведена на рисунке 36.

```
Ввод [89]: # Структура нейронной сети
best_model.summary()

Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_2 (Dense)             (None, 128)               1664

 dense_3 (Dense)             (None, 128)               16512

 dropout (Dropout)           (None, 128)               0

 dense_4 (Dense)             (None, 64)                8256

 dropout_1 (Dropout)         (None, 64)                0

 dense_5 (Dense)             (None, 32)                2080

 dropout_2 (Dropout)         (None, 32)                0

 dense_6 (Dense)             (None, 1)                 33

=================================================================
Total params: 28,545
Trainable params: 28,545
Non-trainable params: 0
_________________________________________________________________
```

Рисунок 36 - Структура нейронной сети

После выбора модели произведем ее обучение на тренировочном датасете. На рисунке 37 представлен график потерь на тренировочной и тестовой выборках, на рисунке 38 – визуализация прогнозных результатов для модели.

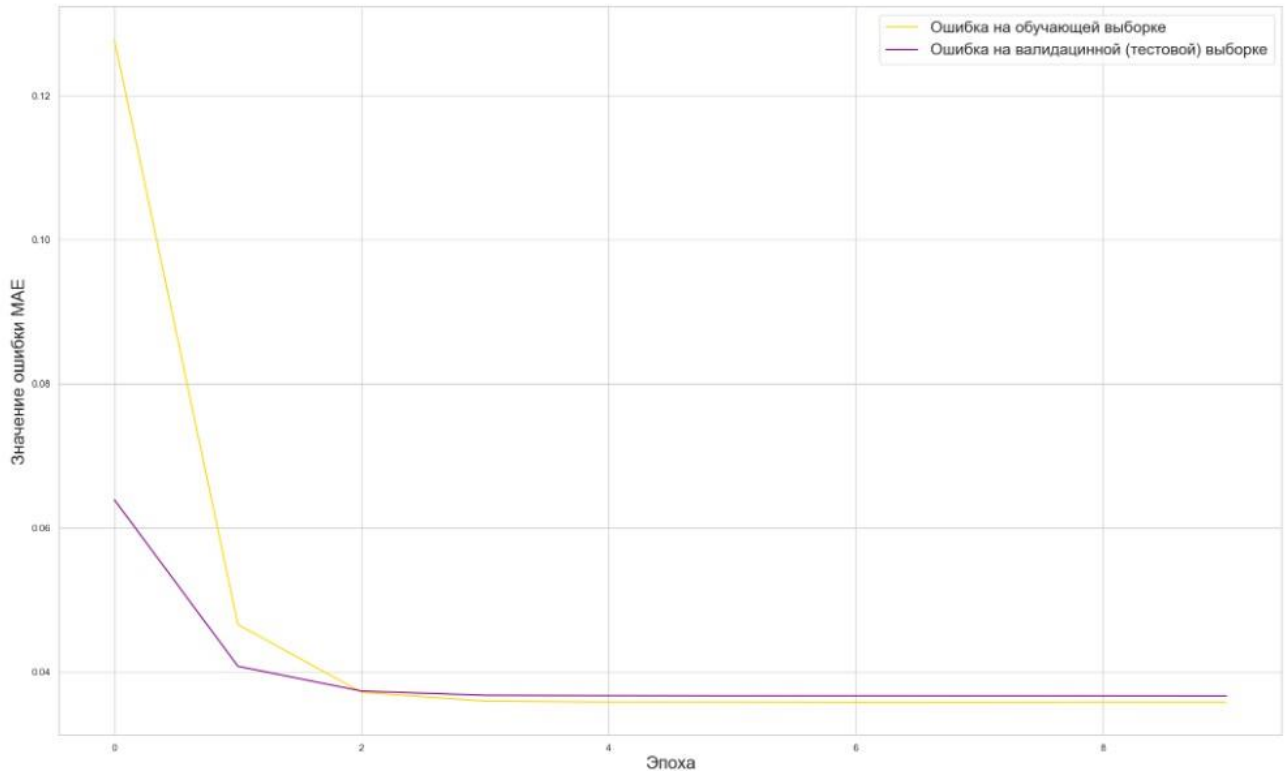

Рисунок 37 - График потерь на тренировочной и тестовой выборках

Рисунок 38 - Визуализация прогнозных результатов для модели

Результат прогноза нейронной сети неудовлетворительный. Значение функции потерь – среднего квадрата ошибки (R2) – составило 0.0336, а средней абсолютной ошибки (MAE) – 0.1490 (рисунок 39). Полученная модель нейронной сети плохо справились с поставленной задачей прогнозирования соотношения

«матрица-наполнитель».

```
Ввод [88]: # Оценка получившейся модели
           best_model.evaluate(X_test_matrix, np.ravel(y_test_matrix), verbose = 1)

           9/9 [==============================] - 0s 2ms/step - loss: 0.0336 - mae: 0.1490

Out[88]: [0.033644143491983414, 0.149032860994339]
```

Рисунок 39 - Оценка работы модели

2.5 Разработка приложения для прогнозирования соотношения «матрица – наполнитель»

Разработка веб-приложения в фреймворке Flask включает следующие этапы:

1. Инициализация приложения Flask и загрузка модели машинного обучения, а также необходимых масштабаторов (т.к. при обучении модели были использованы нормализованные данные) (рисунок 40).

```
Ввод [101]: # Инициализируем приложение Flask
            app = Flask(__name__)

Ввод [103]: # Загружаем модель и масштабаторы
            nn_model = load_model('C:/Users/AYAmankin/Desktop/Курс Data science МГТУ/ВКР/ВКР/Application//model_matrix/')
            scaler_x = load('C:/Users/AYAmankin/Desktop/Курс Data science МГТУ/ВКР/ВКР/Application//minmax_scl_x.pkl')
            scaler_y = load('C:/Users/AYAmankin/Desktop/Курс Data science МГТУ/ВКР/ВКР/Application//minmax_scl_y.pkl')
```

Рисунок 40 - Создание flask приложения

2. Определение маршрута приложения для страницы веб-приложения по умолчанию: маршруты относятся к шаблонам URL-адресов приложения. @app.route('/') – это декоратор Python, который Flask предоставляет для простого назначения URL-адресов в создаваемом приложении функциям. Декоратор сообщает нашему @app, что всякий раз, когда пользователь посещает домен приложения (localhost: 5000 для локальных серверов) с заданным .route(), выполнять функцию home() (рисунок 44). Flask использует библиотеку шаблонов Jinja для

визуализации шаблонов. В создаваемом приложении используются шаблоны для рендеринга HTML, который будет отображаться в браузере.

3. Перенаправление API для прогнозирования соотношения «матрица-наполнитель». Создатся новый маршрут приложения («/predict»), который считывает ввод из формы «main.html» и при нажатии кнопки «Рассчитать» выводит результат с помощью render_template (рисунок 40).

4. Запуск сервера Flask вызывается app.run(), и веб-приложение размещается локально на [localhost: 5000]. «Debug = True» и «Use_reloader = False» гарантирует, что не нужно загружать и запускать созданное приложение каждый раз, когда будут внесены изменения, и дает возможность просто обновить нашу веб-страницу, чтобы увидеть изменения, пока сервер все еще работает (рисунок 41).

Ввод [104]:
```
# Определяем маршрут приложения для страницы веб-приложения по умолчанию
@app.route('/')
def home():
    return render_template('main.html')
```

Ввод [105]:
```
# Создаем новый маршрут приложения, который считывает ввод из формы «main.html»
# и при нажатии кнопки "Рассчитать" выводит результат
@app.route('/predict', methods = ['POST'])
def predict():
    int_features = [float(x) for x in request.form.values()]
    X = scaler_x.transform(np.array(int_features).reshape(1,-1))
    prediction = nn_model.predict(X)
    output = scaler_y.inverse_transform(prediction)
    return render_template('main.html',
                           prediction_text = 'Прогнозное значение соотношения "матрица - наполнитель": {}'.format(output[0][0]))
```

Ввод [*]:
```
# Запуск сервера Flask
if __name__ == "__main__":
    app.run(debug=True, use_reloader=False)
```

```
 * Serving Flask app "__main__" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
INFO:werkzeug: * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Рисунок 41 - Запуск приложения flask

Проект сохраняется в папке с именем «Application» включает следующее:

- Папка «model_matrix» с моделью нейронной сети;

- Папка «templates» с файлом main.html;

- minmax_scl_x.pkl, minmax_scl_y.pkl – сохранные нормализаторы MinMaxScaler();

При запуске приложения открывается локальный сервер на порту 5000 (рисунок 42). Рекомендуется сначала запустить приложение на локальном сервере и проверить его функциональность, прежде чем размещать его в интернете на облачной платформе.

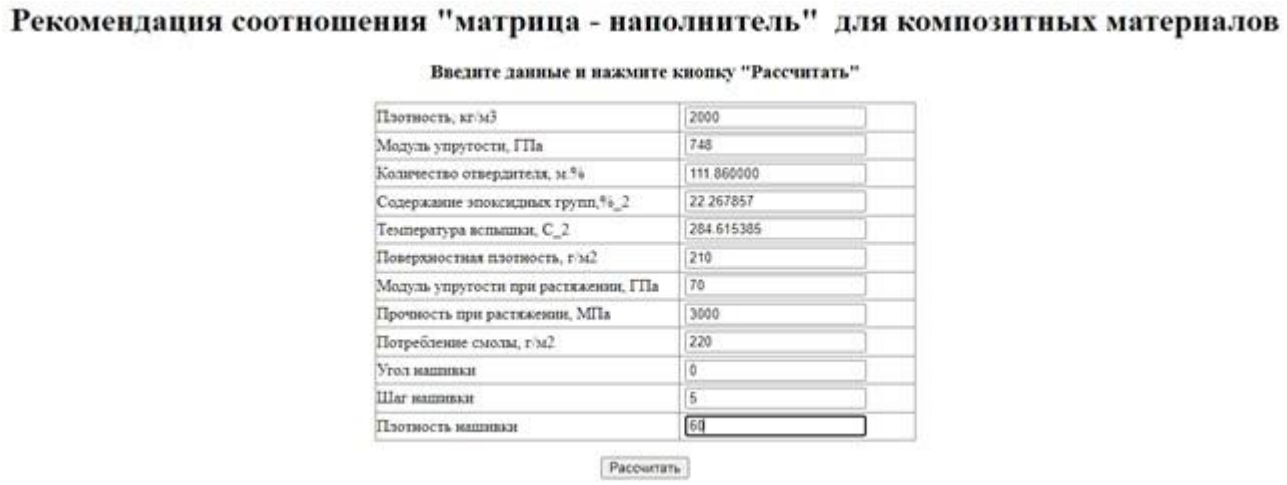

Рисунок 42 – приложение flask

На открывшейся html-странице необходимо ввести данные для прогноза соотношения «матрица-наполнитель» и нажать кнопку «Рассчитать».

2.6 Создание удаленного репозитория и загрузка результатов работы на него

Репозиторий с материалами, разработанными в ходе выполнения выпускной квалификационной работы размещен по адресу https://github.com/Andrey-Yamankin/Composite. Структура данных в репозитории:

- файл VKR Yamankin AG.ipynb – ноутбук, содержащий все произведенные вычисления;
- пояснительная записка к ВКР;
- презентация к ВКР;
- требования к ВКР;

- папка application с сохраненными масштабаторами и моделью;
- папка images с изображениями, использованными в работе;
- папка datasets с исходными и преобразованными данными;
- файл README.

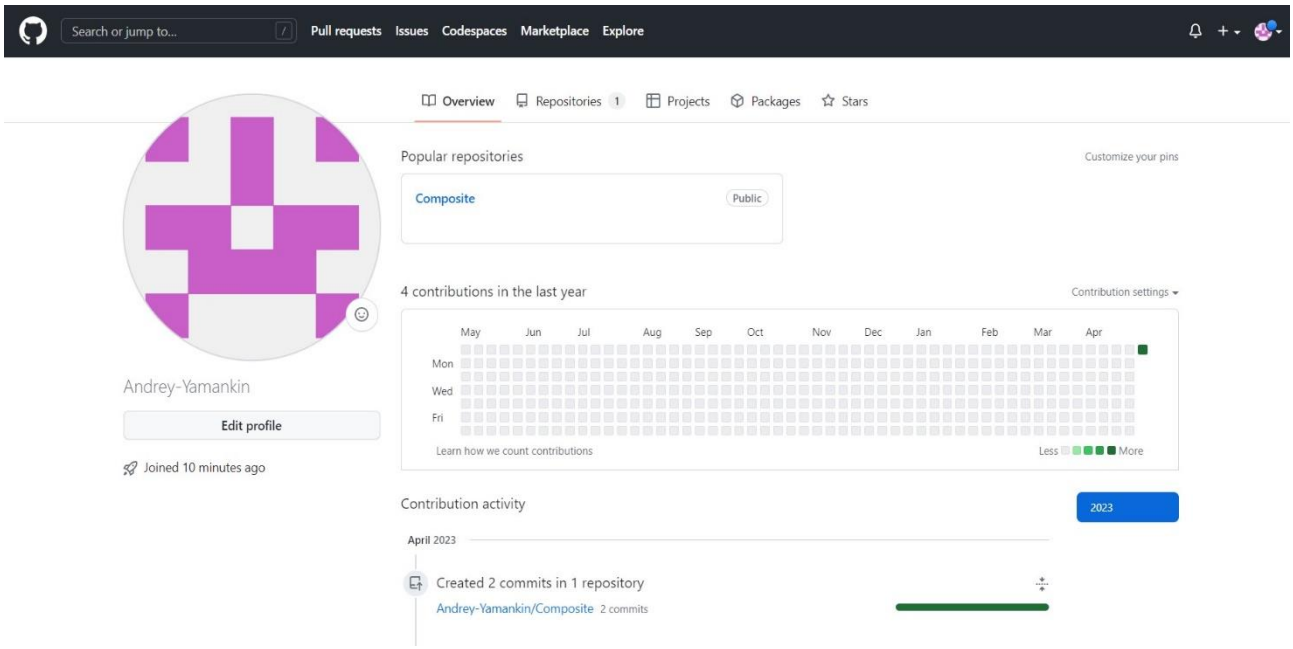

Рисунок 43 – профиль пользователя на GitHub

Заключение

В результате выполнения выпускной квалификационной работы, цель которой – изучение способов прогнозирования конечных свойств новых композиционных материалов, были проанализированы характеристики композитных материалов, а также разработаны модели машинного обучения для выполнения прогнозов этих характеристик.

С использованием разработанных алгоритмов была проведена обработка экспериментальных данных модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель» с использованием языка программирования python.

Как показал анализ исходных данных, корреляционная зависимость между характеристиками композитов крайне слабая и стремится к нулю. Этот факт непосредственно повлиял на результат работы регрессионных моделей. Все использованные модели показали низкую прогнозирующую способность. Лучшим алгоритмом для прогноза модуля упругости при растяжении выбран AdaBoostRegressor, для прогнозирования прочности при растяжении – XGBRegressor.

Созданная для рекомендации соотношения «матрица-наполнитель» нейронная сеть также плохо справилась с поставленной задачей прогноза. Такие низкие показатели работы алгоритмов машинного обучения говорят о том, что прогнозирование свойств композиционных материалов – достаточно сложный процесс, требующий как знаний в области композиционных материалов, так и опыта в построении и использовании алгоритмов машинного обучения.

Полученный неудовлетворительный результат может также свидетельствовать о недостатках и ошибках в наборе исходных данных, недостаточно

глубокой и детальной обработке данных, неточностях в выборе алгоритмов машинного обучения и их параметров.

Таким образом, для успешного решения задачи, поставленной в выпускной квалификационной работе, необходимы более глубокие знания в области материаловедения и технологии конструкционных материалов, математического анализа и статистики, а также в области решения задач машинного обучения и обработки данных. Более детальное изучение данных вопросов и консультация квалифицированных специалистов из указанных областей определенно положительно повлияют на уточнение подходов и оптимизацию алгоритмов для решения задачи прогнозирования конечных свойств композиционных материалов.

Список литературы

1. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour): – Режим доступа: https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19.

2. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать: – Режим доступа: https://habr.com/ru/company/vk/blog/513842/(дата обращения: 22.02.2023).

3. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.

4. Абросимов Н.А.: Методика построения разрешающей системы уравнений динамического деформирования композитных элементов конструкций (Учебно-методическое пособие), ННГУ, 2010

5. Бизли Д. Python. Подробный справочник: учебное пособие. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.

6. 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.

7. 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.

8. Документация по библиотеке keras: – Режим доступа: https://keras.io/api/.

9. Документация по библиотеке matplotlib: – Режим доступа: https://matplotlib.org/stable/users/index.html

10. Документация по библиотеке numpy: – Режим доступа: https://numpy.org/doc/1.22/user/index.html#user.

11. Документация по библиотеке pandas: – Режим

доступа: https://pandas.pydata.org/docs/user_guide/index.html#user-guide.

12. Документация по библиотеке scikit-learn: – Режим доступа: https://scikit-learn.org/stable/user_guide.html.

13. Документация по библиотеке seaborn: – Режим доступа: https://seaborn.pydata.org/tutorial.html.

14. Документация по библиотеке Tensorflow: – Режим доступа: https://www.tensorflow.org/overview

15. Документация по языку программирования python: – Режим доступа: https://docs.python.org/3.8/index.html.

16. Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.

17. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) – Режим доступа: https://habr.com/ru/post/428503/ (дата обращения 03.03.2023)

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

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