МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science Pro»

Слушатель Ишутенко Анна Владимировна

Москва, 2024

**Содержание**

1. Введение
2. Аналитическая часть
   1. Постановка задачи
   2. Описание используемых методов
   3. Разведочный анализ данных
3. Практическая часть
   1. Предобработка данных
   2. Разработка и обучение модели
   3. Тестирование модели
   4. Создание нейронной сети для рекомендации соотношения матрица/наполнитель
   5. Разработка приложения
   6. Создание удаленного репозитория   
      и загрузка результатов работы на него.
4. Заключение
5. Список использованной литературы
6. Приложения
7. **Введение**

Композиционный материал (композит) - многокомпонентный материал, состоящий, как правило, из пластичной основы (матрицы), армированной наполнителями, обладающими высокой прочностью, жёсткостью и т.д. Сочетание разнородных веществ приводит к созданию нового материала, свойства которого количественно и качественно отличаются от свойств каждого из его составляющих. Варьируя состав матрицы и наполнителя, их соотношение, ориентацию наполнителя, можно получить широкий спектр материалов с требуемым набором свойств. Многие композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении её механических характеристик, а также такие материалы имеют высокие значения прочности, жаропрочности, жесткости, сопротивления, усталости при меньшей их плотности.

У композиционных материалов есть также и недостатки, среди которых одним из важнейших является сложность точного определения конечных свойств нового материала на основе характеристик соединяемых компонентов. Подобная задача требует проведения многочисленных экспериментов или создания современных способов прогнозирования конечных свойств. Разработка моделей для точного прогноза особенно актуальна в настоящем времени, так как позволит снизить количество необходимых экспериментов, а значит, сэкономить ресурсы и ускорить процесс разработки новейших материалов.

**2. Аналитическая часть**

**2.1. Постановка задачи**

Композиционные материалы обладают улучшенными свойствами, которые не наблюдаются у их составляющих компонентов по отдельности. Это делает их крайне востребованными в различных отраслевых сферах, включая авиакосмическую промышленность, автомобилестроение, строительную индустрию и другие. Однако определение конечных свойств композитов на основе свойств их исходных компонентов остается сложной задачей. Физические испытания таких материалов зачастую затратны по времени и ресурсам. В этой связи актуальным является использование методов машинного обучения для прогнозирования свойств композитов, что позволяет значительно сократить количество необходимых экспериментальных исследований.

Задачей работы является прогнозирование ключевых механических свойств композитных материалов (таких как модуль упругости при растяжении и прочность при растяжении) на основе характеристик исходных компонентов композита. Помимо этого, необходимо разработать методику автоматизированного подбора оптимального соотношения компонентов (матрица-наполнитель) для получения материалов с заданными свойствами.

Исходные данные представляют собой набор характеристик композитных материалов в двух датасетах формата MS Excel, включающий следующие параметры:

Входные параметры:

* Угол нашивки (градусы)
* Шаг нашивки
* Плотность нашивки
* Плотность материала (кг/м³)
* Модуль упругости (ГПа)
* Количество отвердителя (массовые %)
* Содержание эпоксидных групп (%)
* Температура вспышки (°C)
* Поверхностная плотность (г/м²)
* Потребление смолы (г/м²)

Выходные параметры:

* Модуль упругости при растяжении (ГПа)
* Прочность при растяжении (МПа)
* Соотношение матрица-наполнитель

Датасеты содержат информацию о значениях указанных параметров для множества образцов композитов, что позволяет использовать его для обучения и тестирования различных моделей машинного обучения. При анализе данных были обнаружены определенные особенности, такие как наличие выбросов и значительные вариации в распределении некоторых переменных, что требует проведения предварительной обработки данных.

Датасет X\_nup содержит: число параметров – 3, строк 1040. Датасет X\_bp содержит: число параметров – 11, строк 1023. Потребуется удаление столбцов без названия (содержащие индексы), и объединение датасетов в один для дальнейшей разведки и обработки данных.

Разведочный анализ также включает в себя: нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек. Для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; выполнить предобработку данных: удалить выбросы, сделать нормализацию и стандартизацию. Обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель. Разработать приложение с графическим интерфейсом, которое будет выдавать прогноз соотношения «матрица-наполнитель». Оценить точность модели на тренировочном и тестовом датасете. Создать репозиторий в GitHub и разместить код исследования. Оформить файл README.

**2.2. Описание используемых методов**

Для решения поставленной задачи предполагается использовать следующие методы.

1) Линейная регрессия:

* Описание: линейная регрессия — это метод прогнозирования значений зависимой переменной на основе одной или нескольких независимых переменных, предполагающий линейную связь между ними.
* Достоинства: простота реализации и интерпретации результатов; минимальные вычислительные затраты.
* Недостатки: ограниченная возможность моделирования нелинейных зависимостей; чувствительность к выбросам.
* Область применения: используется для моделирования линейных взаимосвязей между переменными, когда зависимость можно описать прямой линией.

2) Случайный лес (Random Forest):

* Описание: случайный лес — это ансамблевый метод обучения, основанный на построении множества решающих деревьев, каждое из которых обучается на случайной подвыборке данных.
* Достоинства: высокая точность; устойчивость к переобучению; способность моделировать сложные нелинейные зависимости.
* Недостатки: высокая вычислительная сложность; возможна сложность интерпретации результата.
* Область применения: подходит для решения задач регрессии и классификации в условиях сложных взаимосвязей между переменными.

3) К-ближайших соседей (K-Nearest Neighbors/k-NN):

* Описание: К-ближайших соседей — алгоритм классификации и регрессии, основанный на гипотезе компактности, которая предполагает, что расположенные близко друг к другу объекты в пространстве признаков имеют схожие значения целевой переменной или принадлежат к одному классу.
* Достоинства: прост в реализации и понимании полученных результатов; имеет низкую чувствительность к выбросам; не требует построения модели; допускает настройку нескольких параметров; позволяет делать дополнительные допущения; универсален; находит лучшее решение из возможных; решает задачи небольшой размерности.
* Недостатки: замедляется с ростом объёма данных; не создаёт правил; не обобщает предыдущий опыт; основывается на всем массиве доступных исторических данных; невозможно сказать, на каком основании строятся ответы; сложно выбрать близость метрики; имеет высокую зависимость результатов классификации от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоёмкость.
* Область применения: алгоритмы классификации, задачи регрессии.

4)Метод опорных векторов (Support Vector Regression)

Описание: этот бинарный линейный классификатор был выбран, потому что он хорошо работает на небольших датасетах. Данный алгоритм – это алгоритм обучения с учителем, использующихся для задач классификации и регрессионного анализа, это контролируемое обучение моделей с использование схожих алгоритмов для анализа данных и распознавания шаблонов. Учитывая обучающую выборку, где алгоритм помечает каждый объект, как принадлежащий к одной из двух категорий, строит модель, которая определяет новые наблюдения в одну из категорий.

Модель метода опорных векторов – отображение данных точками в пространстве, так что между наблюдениями отдельных категорий имеется разрыв, и он максимален. Каждый объект данных представляется как вектор (точка) в p-мерном пространстве. Он создаёт линию или гиперплоскость, которая разделяет данные на классы.

* Достоинства метода: для классификации достаточно небольшого набора данных. При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных. Эффективен при большом количестве гиперпараметров. Способен обрабатывать случаи, когда гиперпараметров больше, чем количество наблюдений. Существует возможность гибко настраивать разделяющую функцию.  Алгоритм максимизирует разделяющую полосу, которая, как подушка безопасности, позволяет уменьшить количество ошибок классификации.
* Недостатки метода: неустойчивость к шуму; для больших наборов данных требуется долгое время обучения; достаточно сложно подбирать полезные преобразования данных; параметры модели сложно интерпретировать.

5)Градиентный бустинг (Gradient Boosting) — это ансамбль деревьев решений, обученный с использованием градиентного бустинга. В основе данного алгоритма лежит итеративное обучение деревьев решений с целью минимизировать функцию потерь. Основная идея градиентного бустинга: строятся последовательно несколько базовых классификаторов, каждый из которых как можно лучше компенсирует недостатки предыдущих. Финальный классификатор является линейной композицией этих базовых классификаторов.

Достоинства метода: новые алгоритмы учатся на ошибках предыдущих; требуется меньше итераций, чтобы приблизиться к фактическим прогнозам; наблюдения выбираются на основе ошибки; прост в настройке темпа обучения и применения; легко интерпретируем.

Недостатки метода: необходимо тщательно выбирать критерии остановки, иначе это может привести к переобучению; наблюдения с наибольшей ошибкой появляются чаще; слабее и менее гибок чем нейронные сети.

Таблица 1 — Сравнительная таблица методов

|  |  |  |  |
| --- | --- | --- | --- |
| Метод | Достоинство | Недостаток | Область применения |
| Линейная регрессия | Простота, низкая вычислительная сложность | Ограничение по линейности, чувствительность к выбросам | Моделирование линейных зависимостей |
| Случайный лес | Высокая точность, устойчивость к переобучению | Высокая вычислительная сложность, сложность интерпретации | Сложные взаимосвязи, нелинейные зависимости |
| К-ближайших соседей | Простота, универсальность, низкая чувствительность к выбросам | Вычислительная трудоемкость, не обобщает предыдущий опыт | Алгоритмы классификации, задачи регрессии |
| Метод опорных векторов | Достаточно небольшого набора данных; эффективен при большом количестве гиперпараметров | Неустойчивость к шуму, сложность интерпретации | Задачи классификации и регрессии |
| Градиентный бустинг (Gradient Boosting) | Прост в настройке темпа обучения и применения; легко интерпретируем | Наблюдения с наибольшей ошибкой появляются чаще; слабее и менее гибок чем нейронные сети | Алгоритмы классификации, задачи регрессии |

**2.3 Разведочный анализ данных**

Первоначально два датасета были изучены на количество строк и столбцов, после чего из них были удалены столбцы Unnamed, а затем датасеты были объединены по типу INNER. Лишние строки были отброшены, так как не имели соответствующих в соседней таблице.

Для каждой колонки в объединенном датасете были построены гистограммы, которые позволили выявить основные закономерности и особенности распределения данных – гистограммы представлены на рисунке 1. Например, распределение угла нашивки показало наличие двух четких пиков, что может указывать на использование двух различных технологических подходов при производстве композитов. Распределение других переменных, таких как модуль упругости или прочность при растяжении, имеет более нормальное распределение с незначительными отклонениями, что подтверждает корректность проведенных измерений в соответствии с рисунком 2.

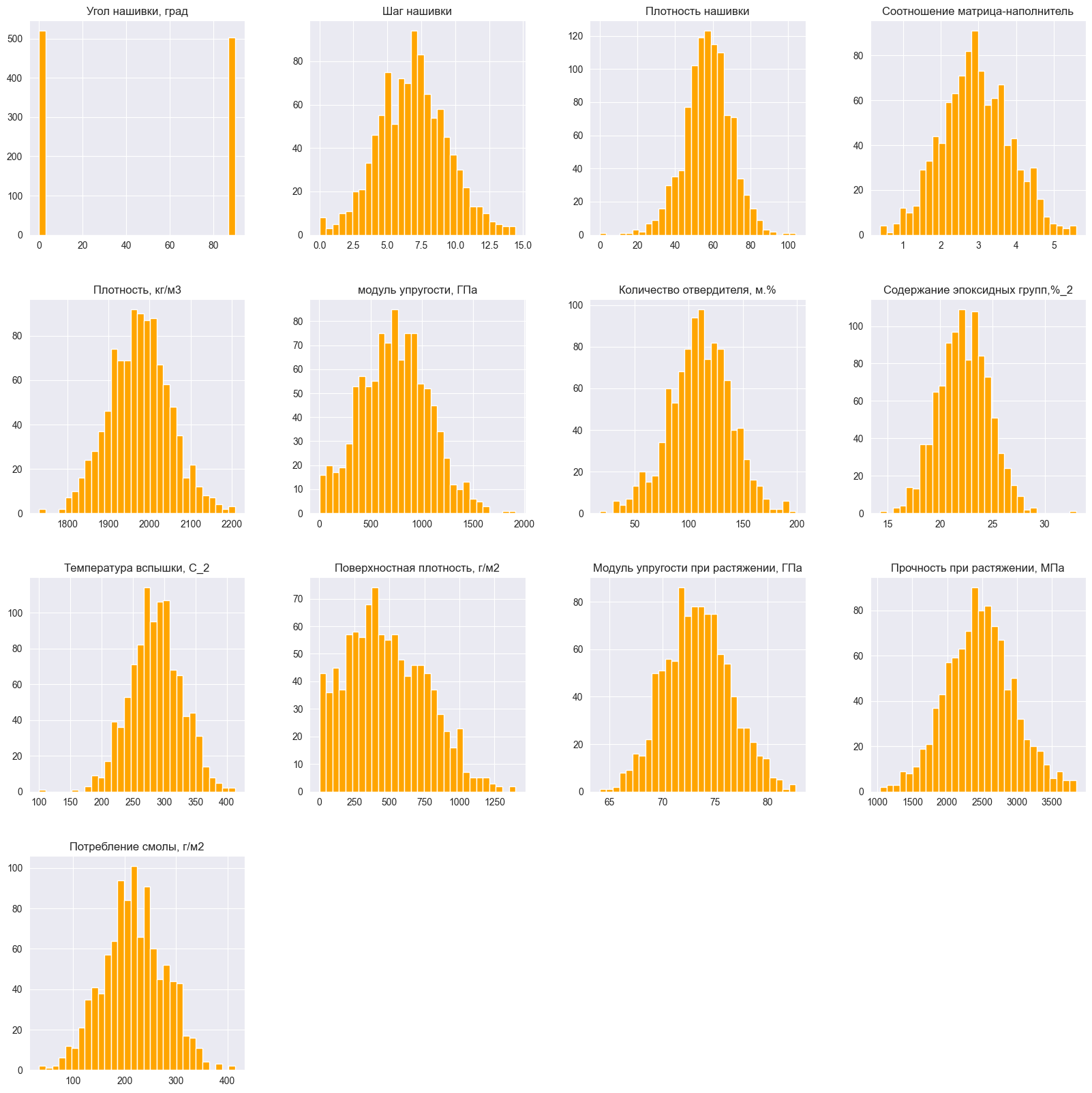


Рисунок 1 — Гистограммы распределения переменных

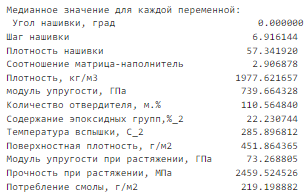
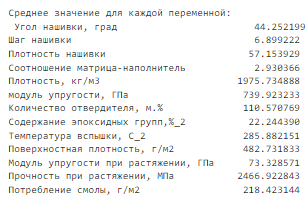


Рисунок 2 — Вывод среднего и медианного значений

для каждой переменной

Было проанализировано количество пропусков в каждом столбце датасета. Пропусков обнаружено не было, что дало возможность не заполнять пустые ячейки средними значениями. Результат поиска пропусков представлен на рисунке 3.

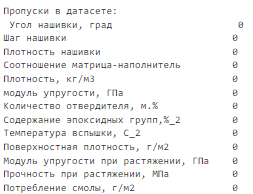


Рисунок 3 — Поиск пропусков в каждой колонке датасета

Выбросы были выявлены при анализе диаграмм ящика с усами. Некоторые переменные, такие как плотность нашивки и модуль упругости, содержат значительное количество выбросов, которые могут существенно повлиять на результаты моделирования. Результат представлен на рисунке 4.

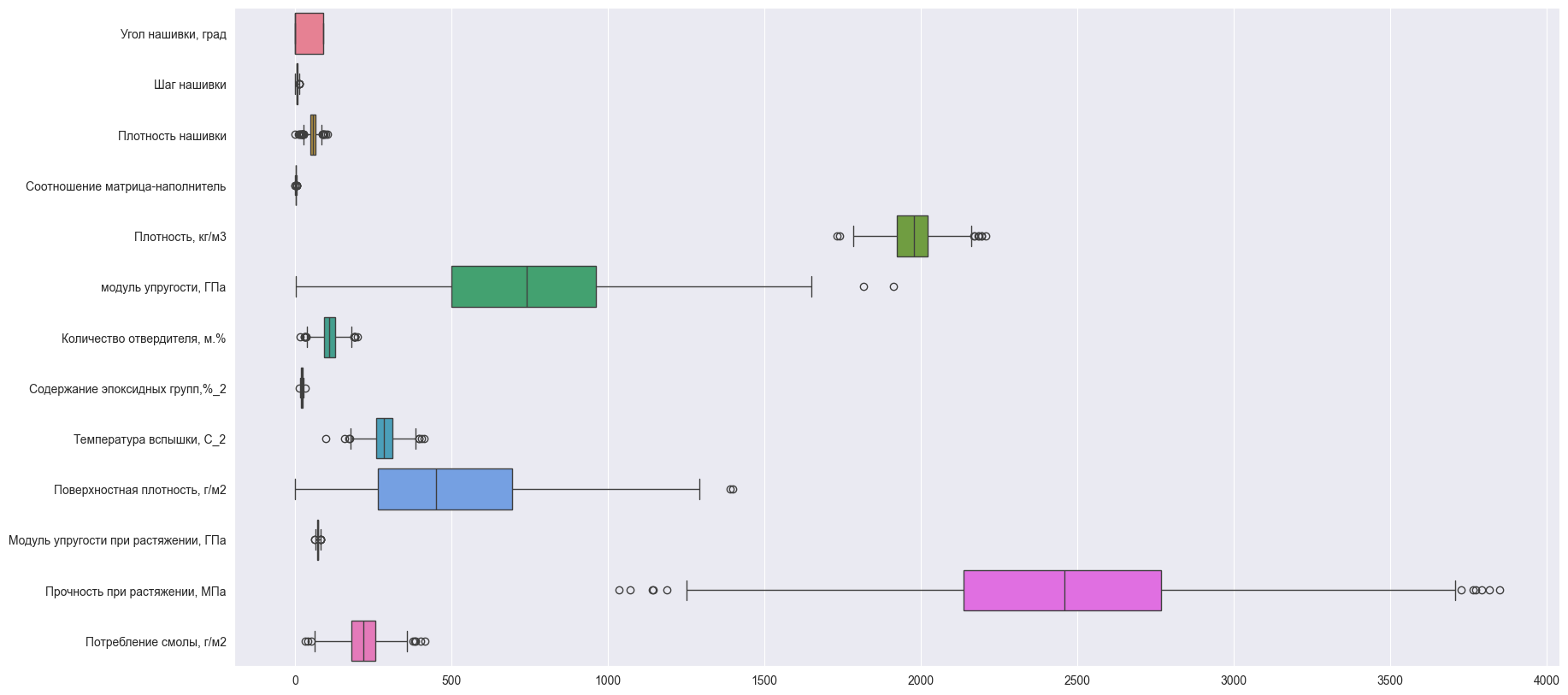


Рисунок 4 — Диаграмма "Ящик с усами"

Для выявления взаимосвязей между переменными были построены попарные графики рассеяния. Эти графики показали наличие как линейных, так и нелинейных зависимостей между некоторыми параметрами. Например, плотность материала и прочность при растяжении показали сильную положительную корреляцию, что соответствует теоретическим ожиданиям и наглядно отображено в соответствии с рисунком 5.

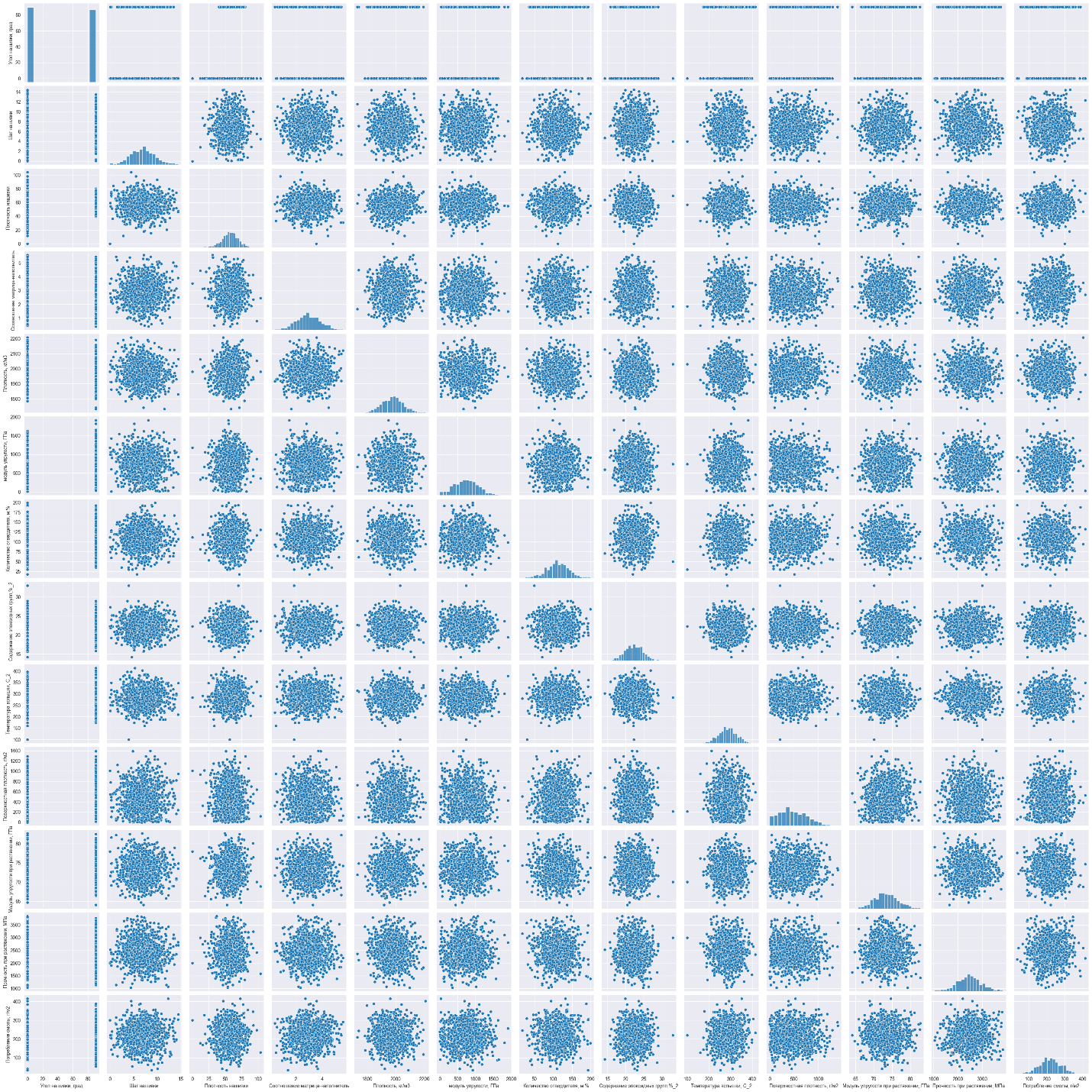


Рисунок 5 — Попарные графики рассеяния

Были проверены типы данных и количество ненулевых значений. Повторяющихся значений практически нет.

Построена тепловая карта корреляционной зависимости переменных. Результат представлен на рисунке 6.

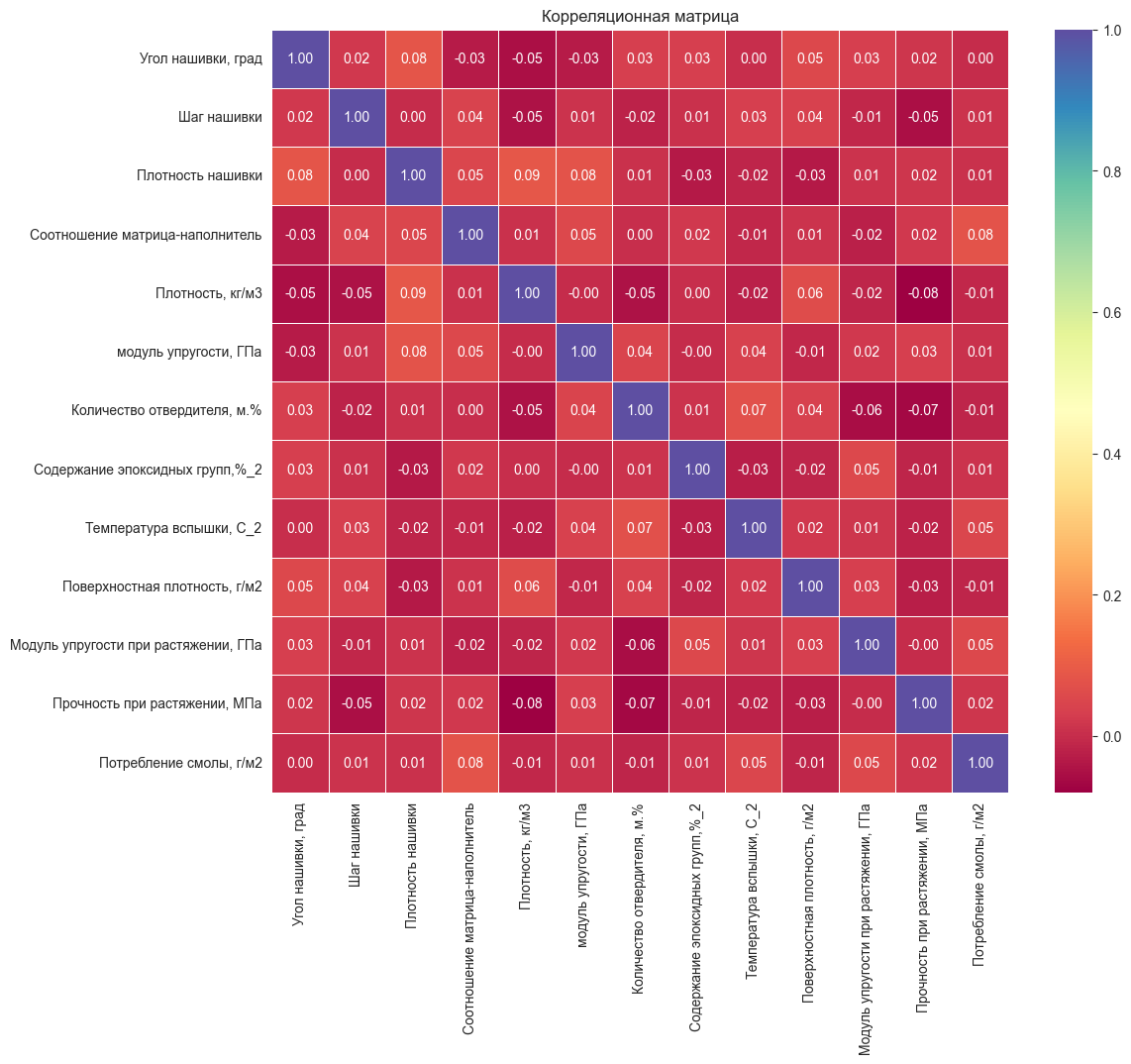


Рисунок 6 – Тепловая карта корреляции данных

На основание гистограмм, диаграммы попарных графиков рассеяния и тепловой карты корреляции данных был сделан вывод о том, что максимальная корреляция между плотностью нашивки и углом нашивки 0.11, значит нет зависимости между этими данными. Корреляция между всеми параметрами очень близка к 0, корреляционные связи между переменными не наблюдаются.

**3. Практическая часть**

**3.1. Предобработка данных**

Выбросы в данных могут существенно влиять на качество моделей машинного обучения, так как они приводят к искажению распределений и могут вызвать переобучение модели. В данной работе для выявления и удаления выбросов был использован метод межквартильного размаха (IQR). Этот метод позволяет определить выбросы как значения, выходящие за пределы интервала , где Q1 и Q3 — это первый и третий квартили соответственно, а IQR — межквартильный размах. Все обнаруженные выбросы были удалены из датасета, что привело к значительному улучшению качества данных.

После удаления выбросов количество строк в датасете уменьшилось с 1023 до 922. Удаление выбросов необходимо выполнить до 3 раз – проверка на наличие выбросов была выполнена через построение диаграммы «Ящик с усами» - рисунок 7. Это указывает на то, что часть данных содержала аномальные значения, что потребовало их исключения для корректного обучения моделей.

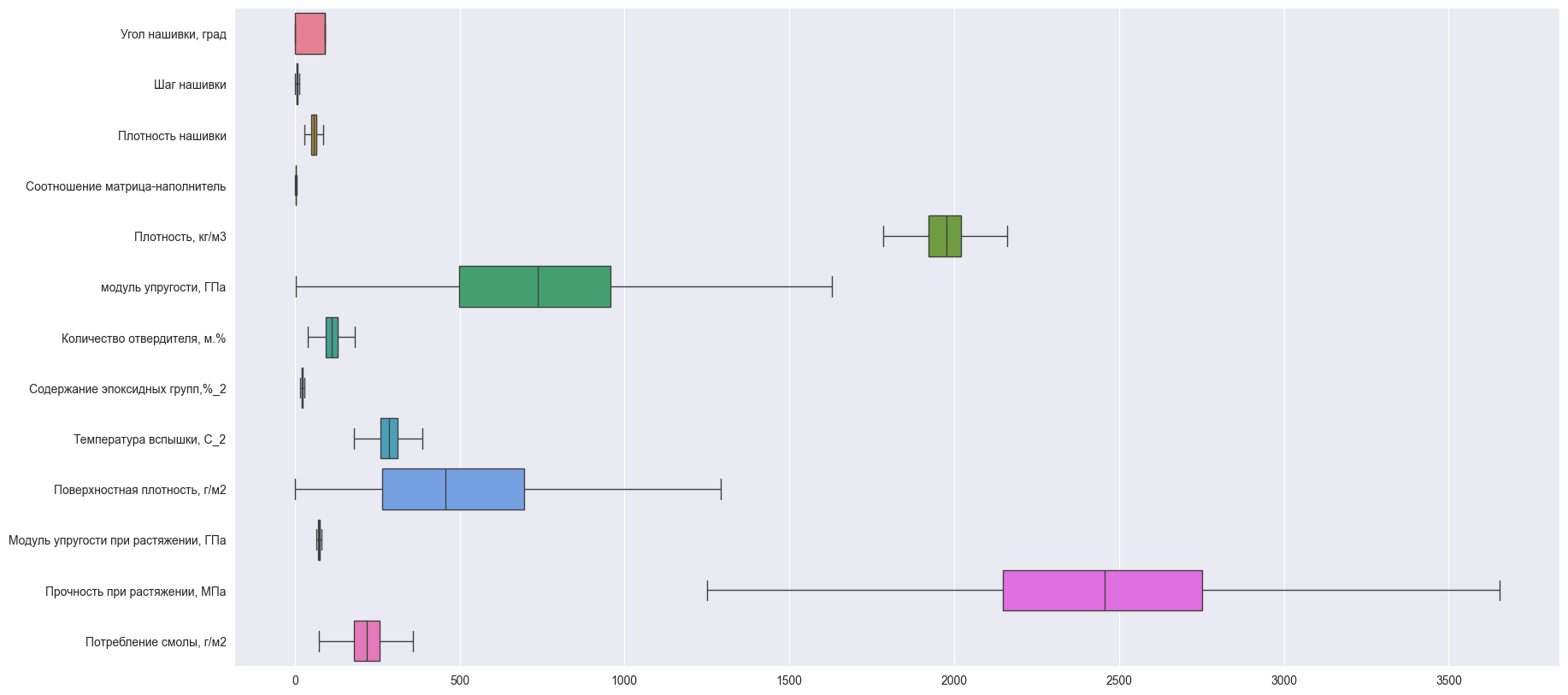


Рисунок 7 — Диаграмма "Ящик с усами" после удаления выбросов

Нормализация данных является важным этапом предобработки, так как она позволяет привести данные к одному масштабу, что особенно важно для методов машинного обучения, чувствительных к масштабу входных признаков, таких как линейная регрессия и нейронные сети. В данной работе использовался метод стандартизации, при котором каждая переменная приводится к распределению с нулевым средним и единичным стандартным отклонением. Для нормализации был применен алгоритм StandardScaler из библиотеки scikit-learn.

Нормализация позволила существенно улучшить распределение данных, что было подтверждено анализом гистограмм до и после нормализации. Визуализация этих изменений показала, что после нормализации все переменные приобрели более компактное и симметричное распределение, что является благоприятным фактором для успешного обучения моделей.

Для каждого признака были построены гистограммы распределения после нормализации. Результат представлен на рисунке 8. Графики показали, что после нормализации значения всех признаков были приведены к одинаковому масштабу, и распределение данных стало более подходящим для обучения моделей. Это особенно важно для методов, которые могут некорректно работать с данными разного масштаба. Такие изменения обеспечивают более стабильное и точное обучение моделей на последующих этапах работы.

Была проведена стандартизация с использованием метода StandardScaler. После стандартизации появились отрицательные значения, увеличилось количество выбросов. Было решено в дальнейшем для машинного обучения использовать нормализованный датасет.

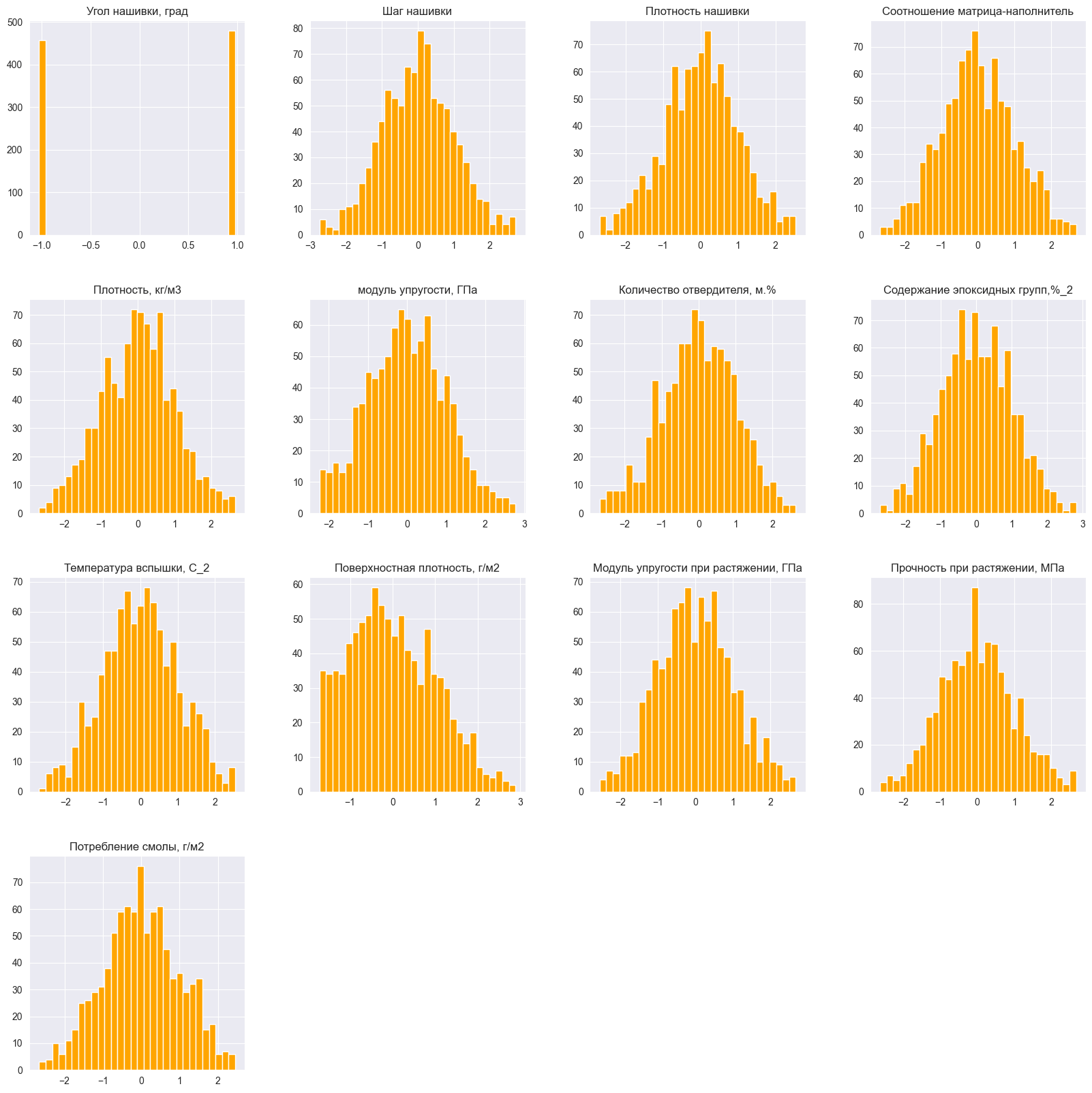


Рисунок 8 — Гистограммы распределения переменных после нормализации и стандартизации

**3.2. Разработка и обучение модели**

Для решения задачи прогнозирования модуля упругости при растяжении и прочности при растяжении были выбраны следующие модели машинного обучения:

* Линейная регрессия
* Случайный лес (Random Forest)
* К-ближайших соседей (K-Nearest Neighbors/k-NN)
* Метод опорных векторов (SVR)
* Градиентный бустинг (Gradient Boosting).

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

Для обучения моделей данные были разделены на тренировочную и тестовую выборки в пропорции 70% и 30% соответственно. Это разделение позволяет модели обучаться на большей части данных, сохраняя при этом достаточное количество данных для объективного тестирования модели.

Результаты тестовых и прогнозных значений каждой из модели представлены на рисунках 9-13 (оранжевый – прогнозные значения, синий – тестовые).

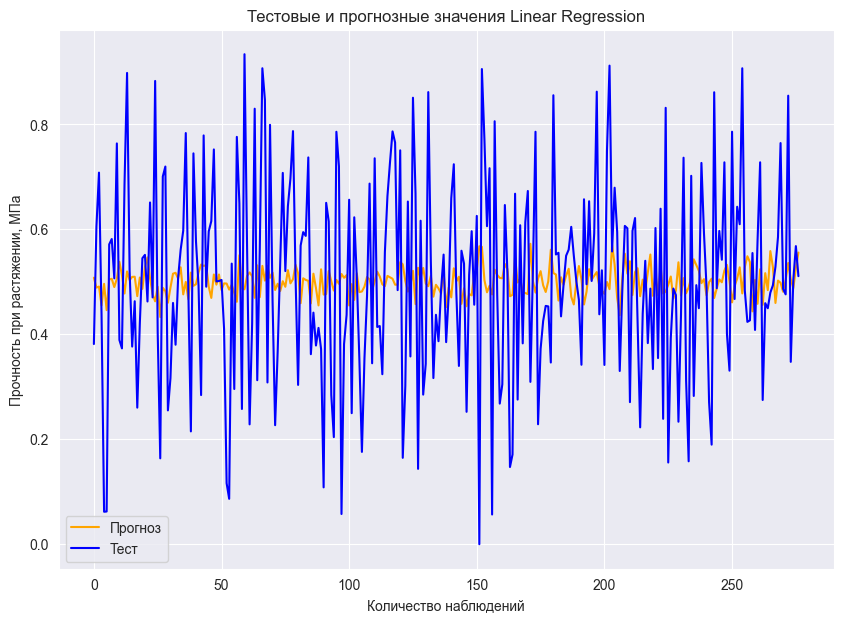


Рисунок 9 — Тестовые и прогнозные значения, метод Линейной регрессии

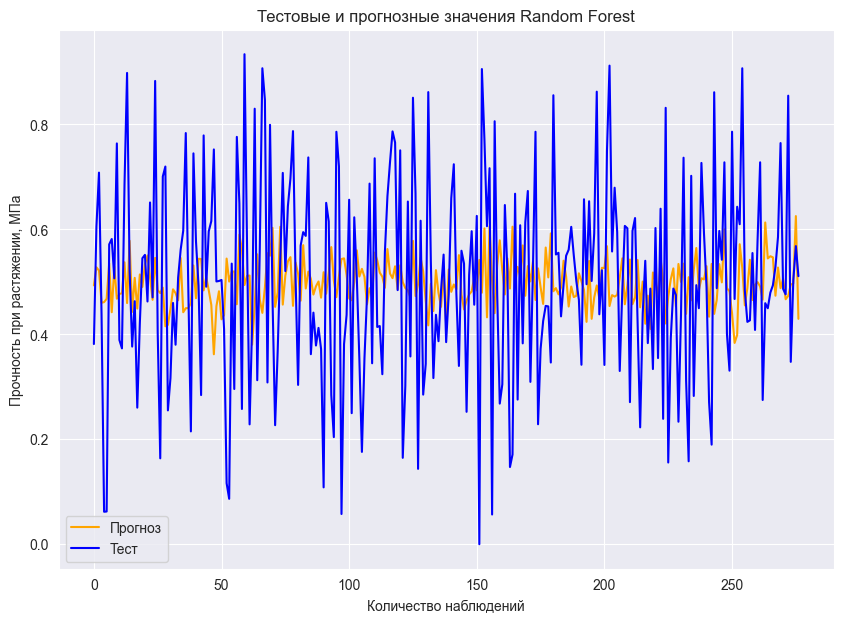


Рисунок 10 — Тестовые и прогнозные значения, метод Случайный лес

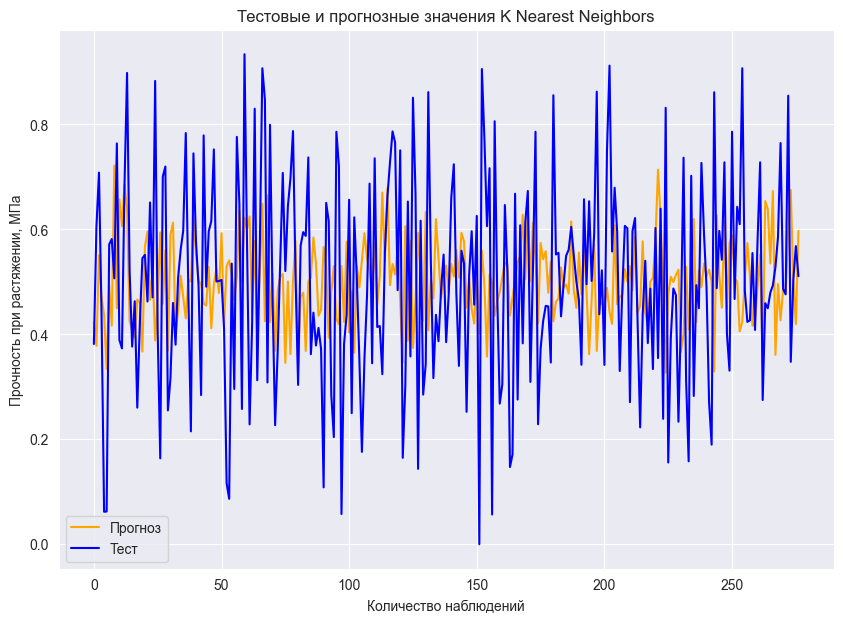


Рисунок 11 — Тестовые и прогнозные значения, метод К-ближайших соседей

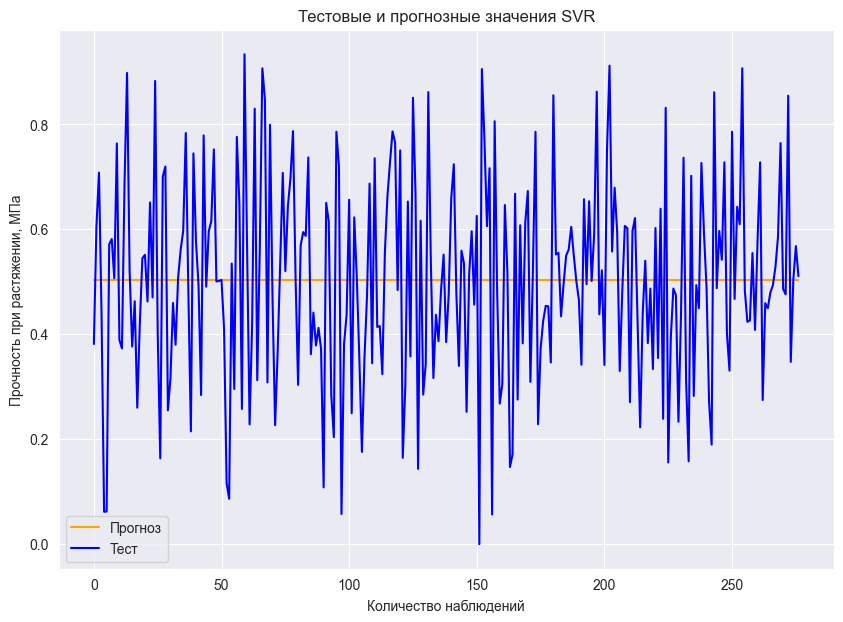


Рисунок 12 — Тестовые и прогнозные значения, метод опорных векторов

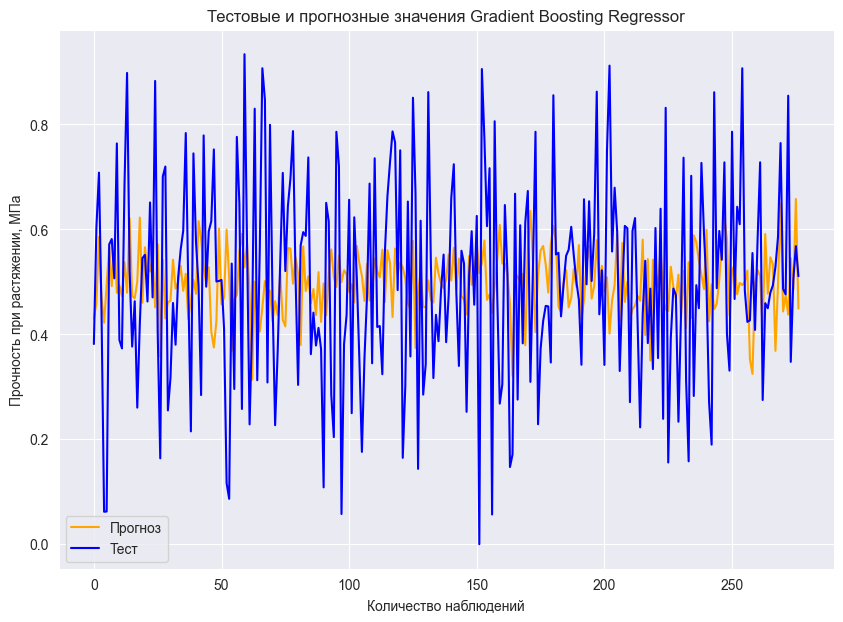


Рисунок 13 — Тестовые и прогнозные значения, метод Градиентного бустинга

Каждая из выбранных моделей была обучена на подготовленных данных. В процессе обучения использовалась техника перекрестной проверки (cross-validation), которая позволяет более объективно оценить качество моделей и избежать переобучения. Было выбрано 10-кратное кросс-валидирование, что позволило получить стабильные и надежные результаты.

Для каждой модели был проведен подбор гиперпараметров с использованием метода поиска по сетке (GridSearchCV). Этот метод позволяет найти оптимальные значения гиперпараметров, которые обеспечивают наилучшее качество модели на тренировочных данных.

Также для каждой модели был рассчитан коэффициент МАЕ для сравнения каждой модели. Результаты представлены на рисунке 14.

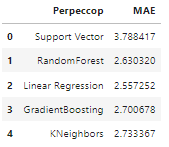


Рисунок 14 – сводная таблица результатов МАЕ по каждой модели обучения

На основании проделанной работы был сделан следующий вывод: параметры эффективности каждой модели неудовлетворительные. Расчет коэффициента детерминации (R2) оказался много ниже 2, все использованные модели не справились с поставленной задачей.

**3.3. Тестирование модели**

После обучения моделей была проведена оценка точности модели К ближайших соседних на обучающей и тестовых выборках. В качестве параметра оценки модели использовалась средняя квадратическая ошибка (MSE). Результат неудовлетворительный. При полученных результатах целесообразнее применять среднее значение переменной в качестве прогнозного.

Таблица 2 – сравнение моделей по коэффициентам эффективности

|  |  |  |  |
| --- | --- | --- | --- |
| Переменная | Модель | MSE | R2 |
| Прочность при растяжении | Линейная регрессия | 0.02397 | -0.018674 |
| Модуль упругости при растяжении | Линейная регрессия | 0.0247 | -0.014876 |
| Прочность при растяжении | Случайный лес (Random Forest) | 0.02765 | -0.005672 |
| Модуль упругости при растяжении | Случайный лес (Random Forest) | 0.02348 | -0.01456 |
| Прочность при растяжении | К-ближайших соседей | 0.028116 | -0.013587 |
| Модуль упругости при растяжении | К-ближайших соседей | 0.0256 | -0.021467 |
| Прочность при растяжении | Метод опорных векторов (SVR) | 0.02772 | -0.018689 |
| Модуль упругости при растяжении | Метод опорных векторов (SVR) | 0.02544 | -0.018035 |
| Прочность при растяжении | Градиентный бустинг | 0.01987 | -0.014384 |
| Модуль упругости при растяжении | Градиентный бустинг | 0.02781 | -0.011906 |

**3.4. Создание нейронной сети для рекомендации соотношения матрица/наполнитель**

Для задачи рекомендации соотношения матрица/наполнитель была разработана нейронная сеть на основе полносвязной архитектуры. Нейронная сеть состоит из следующих слоев:

* Входной слой: Размерность входного слоя соответствует числу признаков в данных.
* Скрытые слои: три скрытых слоя с 128, 64 и 32 нейронами соответственно, каждый из которых использует функцию активации ReLU. Для предотвращения переобучения в каждом слое используется механизм Dropout с коэффициентом 0.4 для первых двух слоев и 0.3 для последнего.
* Выходной слой: Один нейрон с линейной функцией активации, предназначенный для регрессии и предсказания соотношения матрица-наполнитель.

Коэффициент МАЕ для построенной нейронной сети составил 0.7942. Также были построены:

1) визуализация тестовых и прогнозных значений нейронной сети – представлена на рисунке 15;

2) график предсказательных и настоящих значений – представлен на рисунке 16.



Рисунок 15 - Тестовые и прогнозные значения построенной нейронной сети

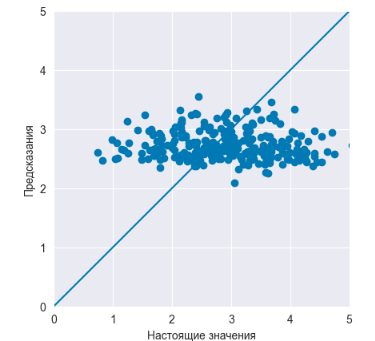


Рисунок 16 – График предсказательных и настоящих значений

**3.5. Разработка приложения**

Для взаимодействия с обученной нейронной сетью было разработано консольное приложение, которое позволяет пользователю вводить значения различных параметров композитного материала и получать прогнозируемое соотношение матрица-наполнитель. Приложение реализовано на языке Python с использованием библиотек TensorFlow и scikit-learn для работы с моделью и нормализацией данных соответственно.

Приложение состоит из следующих основных частей:

Приложение загружает заранее обученную модель нейронной сети из файла "my\_model.keras".

Функция "predict\_matrix\_ratio" отвечает за предсказание соотношения матрица-наполнитель на основе введенных пользователем значений признаков. Входные данные преобразуются в массив NumPy, нормализуются с использованием загруженного normalizer, после чего модель выполняет предсказание и возвращает результат.

Основной функционал приложения реализован в виде консольного интерфейса, где пользователю предлагается ввести значения для ряда признаков композитного материала, таких как угол нашивки, шаг нашивки, плотность материала и другие. Пользователь может вводить значения по каждому признаку, а в случае отсутствия данных можно оставить поле пустым, и значение будет установлено по умолчанию (0.0). После ввода всех данных приложение выполняет предсказание и выводит результат в консоль.

Приложение запускается путем выполнения Python-скрипта в консоли. Для этого необходимо запустить команду "python app.py" в терминале. После запуска приложения пользователю будет предложено последовательно ввести значения для следующих параметров:

* Угол нашивки (градусы)
* Шаг нашивки
* Плотность нашивки
* Плотность материала (кг/м³)
* Модуль упругости (ГПа)
* Количество отвердителя (массовые %)
* Содержание эпоксидных групп (%)
* Температура вспышки (°C)
* Поверхностная плотность (г/м²)
* Модуль упругости при растяжении (ГПа)
* Прочность при растяжении (МПа)
* Потребление смолы (г/м²)

Если для какого-либо параметра у пользователя нет данных, можно оставить поле пустым, и значение по умолчанию будет установлено в 0.0.

После ввода всех данных приложение выполнит предсказание соотношения матрица-наполнитель и выведет результат в консоль в виде числового значения.

Приложение в дальнейшем может быть расширено и адаптировано для работы с новыми моделями и функциями. Например, можно добавить интерфейс для предсказания других параметров композитов или реализовать графический интерфейс для улучшения взаимодействия с пользователем.

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

На Github был создан репозиторий: <https://github.com/anna-ish/VKR_Bauman>. В репозиторий загружены:

1) Исходный датасет (два файла формата MS Ecel);

2) Презентация;

3) Пояснительная записка;

4) Код на Jupyter Notebook;

5) Файл нейронной сети – my\_model.keras.

**4. Заключение**

В заключении проведенной работы выделены основные моменты:

1) В исходном датасете корреляционные признаки стремятся к нулю. Не было обнаружено никакой взаимосвязи между парами признаков, что существенно усложняет работу.

2) Использованные модели машинного обучения не дали удовлетворительного результата – коэффициенты эффективности показали низкое значение.

3) Из свойств материала нет возможности определить соотношение Матрица/наполнитель. Есть вероятность, что база исходных данных недостаточна, либо были неверно подобраны методы предобработки данных или методы машинного обучения.

Работа с композиционными материалами требует в дальнейшем более глубокой проработки темы – общение с экспертом в данной области (консультация с химиком-технологом производства/отдела по разработке), а также применения более широкого спектра методов машинного обучения.

**5. Список использованной литературы**

1) Композиционные материалы : учебное пособие для вузов / Д. А. Иванов, А. И. Ситников, С. Д. Шляпин ; под редакцией А. А. Ильина. — Москва : Издательство Юрайт, 2024. — 253 с.

2) Джоэл Грас. Data science. Наука о данных с нуля: Пер. с англ. – 2-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2024. – 416 с.: ил.

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

4) Аналитическая культура. От сбора данных до бизнес-результатов / Карл Андерсон; пер. с англ. Юлии Константиновой; [науч. ред. Руслан Салахиев]. — М.: Манн, Иванов и Фербер, 2017.

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

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

7) Буценко Е. В. Оптимизация управления проектами: монография / Е. В. Буценко. – Екатеринбург: Изд-во Урал. гос. экон. ун-та, 2023. – 247 с.

8) Гифт Н. Прагматичный ИИ. Машинное обучение и облачные технологии / Н. Гифт. – Спб: Питер, 2019. – 300 с.

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

10) Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с.

11) Тимофеев А. Г., Лебединская О. Г. Статистический и интеллектуальный анализ при построении прогноза цен и финансовых показателей с использованием ИИ // Экономика и предпринимательство. – 2022. – № 4. – С. 1118-1126.

12) Уланов А. А. Оценка потенциала использования ИИ в управлении организацией // Экономика и предпринимательство. – 2023. – № 5. – С. 45-57.

13) Хорватов А. И. Роль ИИ в трансформации промышленности // Технополис. – 2021. – № 3. – С. 62-68.