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**

**(композитных материалов)»**

Слушатель Ниазбаева Ольга Рашидовна

Москва, 2023

**Содержание**

**Содежание 2**

**Введение 3**

1. **Аналитическая часть 5**
   1. **Постановка задачи 5**
   2. **Описание используемых методов 5**
   3. **Разведочный анализ данных 9**
2. **Практическая часть 14**
   1. **Предобработка данных 14**
   2. **Разработка и обучение моделей 17**
   3. **Тестирование моделей 23**
   4. **Написание нейронной сети, которая рекомендует соотношение «матрица-наполнитель» 24**
   5. **Разработка приложения 26**
   6. **Создание удаленного репозитория и загрузка 27**

**Заключение 28**

**Библиографический список 29**

**Введение**

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

Данная работа посвящена проблеме прогнозирования характеристик композитных материалов: модуля упругости при растяжении, прочности при растяжении, соотношение «матрица-наполнитель».

**Актуальность:** Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

1. **Аналитическая часть**
   1. **Постановка задачи**

Задача данного проекта заключается в создании моделей для прогнозирования свойств композитных материалов на основе данных о характеристиках их компонентов. Для этого необходимо создать и обучить модели или нейронные сети. Для их создания были использованы представленные данные с характеристиками материалов.

Для определения зависимостей между параметрами был проведен разведочный анализ, в котором использовались гистограммы распределения, попарные графики рассеяния точек и диаграммы boxplot («ящики с усами»).

Также, для каждого параметра необходимо вычислить среднее и медианное значение, провести анализ и исключение выбросов, применить нормализацию. Затем обучить модели для прогнозирования модуля упругости при растяжении и прочности при растяжении, и написать нейронную сеть, которая будет рекомендовать соотношение матрица наполнитель и разработать приложение, которое будет выдавать прогноз соотношения матрица-наполнитель.

* 1. **Описание используемых методов**

Для выполнение проекта использовались два файла (рис. 1, 2) с расширением .xlsx: X\_bp.xlsx, который содержит физические характеристики базальтопластика и X\_nup.xlsx, который содержит геометрические характеристики нашивки углепластика.

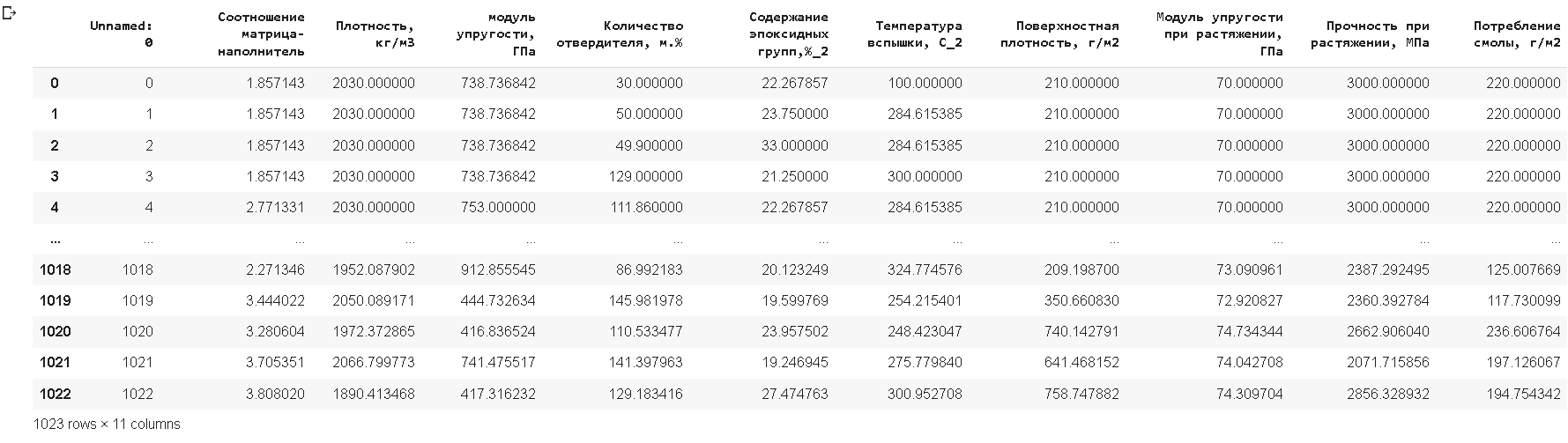


Рисунок 1. Данные физических характеристик базальтопластика.

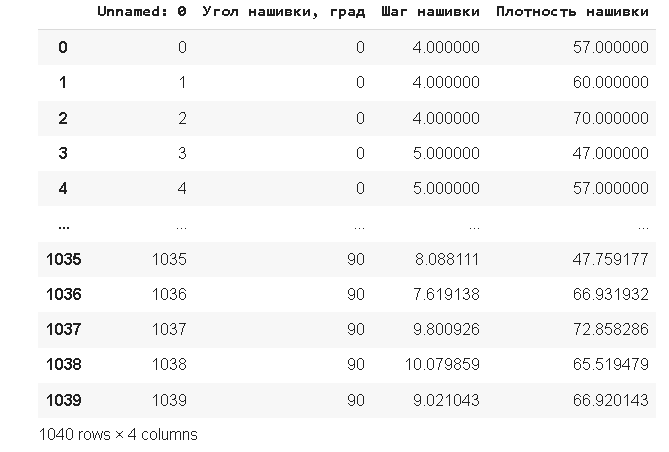


Рисунок 2. Данные геометрических характеристик углепластика.

Таблица на рисунке 1 содержит 1023 строки и 11 параметров. Таблица, представленная на рисунке 2 содержит 1040 строк и 4 параметра. В соответствии с заданием, таблицы с данными были объединены по типу INNER. В итоге был получен датасет, состоящий из 1023 строк и 15 параметров. Однако, два параметра из 15 не информативны, поэтому были удалены.

Далее, при просмотре информации о датасете видим, что все параметры, кроме угла нашивки, типа float64 (рис.3).

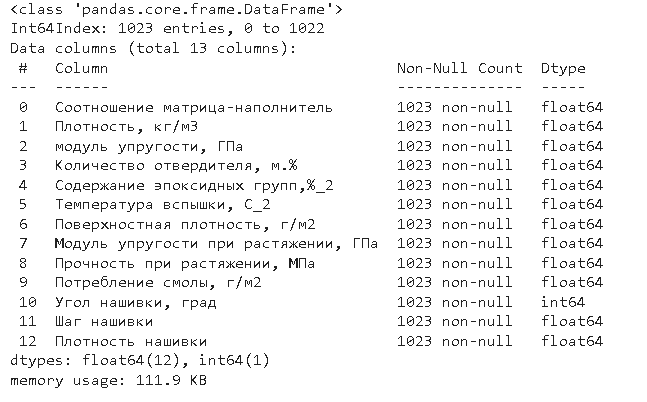


Рисунок 3. Информация о датасете.

Проверка на пропуски показала их отсутствие (рис.4).



Рисунок 4. Результаты проверки на пропуски.

Поиск уникальных значений показал большую уникальность по всем параметрам, кроме угла нашивки (рис.5).

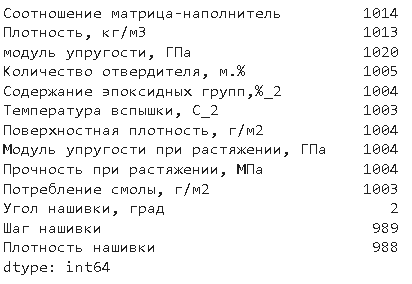


Рисунок 5. Результат поиска уникальных значений.

Описательная статистика показывает среднее и медианное значение по каждому параметру (рис.6).



Рисунок 6. Описательная статистика датасета.

Для определения целевых параметров были использованы следующие модели:

Линейная регрессия, метод анализа данных и статистического моделирования, который позволяет определить линейную зависимость между двумя переменными: зависимой (целевой переменной) и одной или несколькими независимыми переменными (факторами).

Метод k ближайших соседей, основан на идеи, что объекты, близкие друг к другу в пространстве признаков, имеют похожие значение целевой переменной. Основным преимуществом метода k-NN является его простота и интуитивность, а также возможность обработки неструктурированных данных. Он также может быть эффективным в случае, когда зависимость между признаками и целевой переменной нелинейна или сложна.

Метод опорпных векторов, основная идея заключается в том, чтобы находить гиперплоскость, которая максимизирует отступ между различными классами данных. SVM может также использоваться для решения задач регрессии, где вместо классов данных используются числовые значения. В этом случае SVM пытается найти гиперплоскость, которая максимизирует отступ между предсказанными значениями и значениями данных.

Метод случайного леса, основан на комбинации нескольких деревьев решений, каждое из которых обучается на различной подвыборке данных и случайном подмножестве признаков.

Lasso регрессия и Ridge регрессия являются двумя методами регуляризации в линейной регрессии. Оба метода добавляют штрафное условие к минимизируемой функцией потерь, чтобы предотвратить переобучение модели.

* 1. **Разведочный анализ данных**

Разведочный анализ данных является процессом исследования и анализа данных с целью извлечения информации, выявление закономерностей и понимания структуры данных. Основная задача данного метода в том, чтобы максимально использовать имеющуюся информацию для получения полного понимания данных и выделения важных характеристик. Для этого используются различные методы визуализации, статистические методы и алгоритмы машинного обучения.

В данной работе использовались следующие методы разведочного анализа данных:

Визуализация данных с помощью гистограмм может быть полезна для анализа данных и выявления особенностей распределения, таких, как среднее значение, медиана, мода и выбросы. Гистограммы это графики, которые используются для визуализации распределения частоты значений в наборе данных. Она состоит из столбцов, каждый из которых представляет диапазон значений, а высота столбца отображает частоту, с которой значение в этом диапазоне появляются в наборе данных (рис. 7).

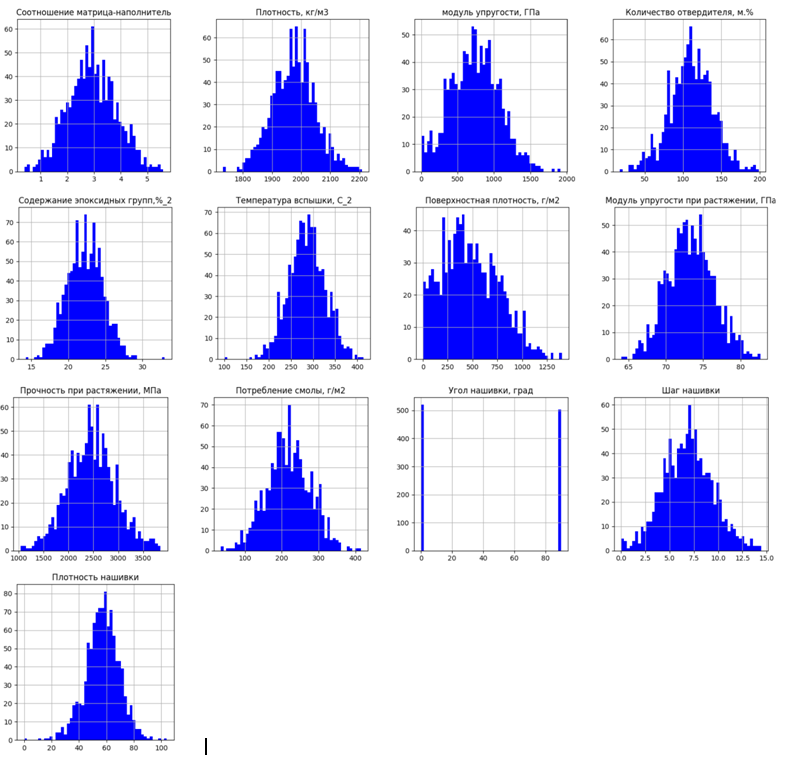


Рисунок 7. Гистограммы параметров данных.

По виду гистограмм можно сделать вывод, что у параметров датасета распределение нормальное или близкое к нормальному, кроме Поверхностной плотности (ассиметричное распределение) и Угла нашивки (имеет два значения, дискретное распределение).

Графики попарного рассеяния не показывают наличие корреляции между параметрами датасета (рис. 9).

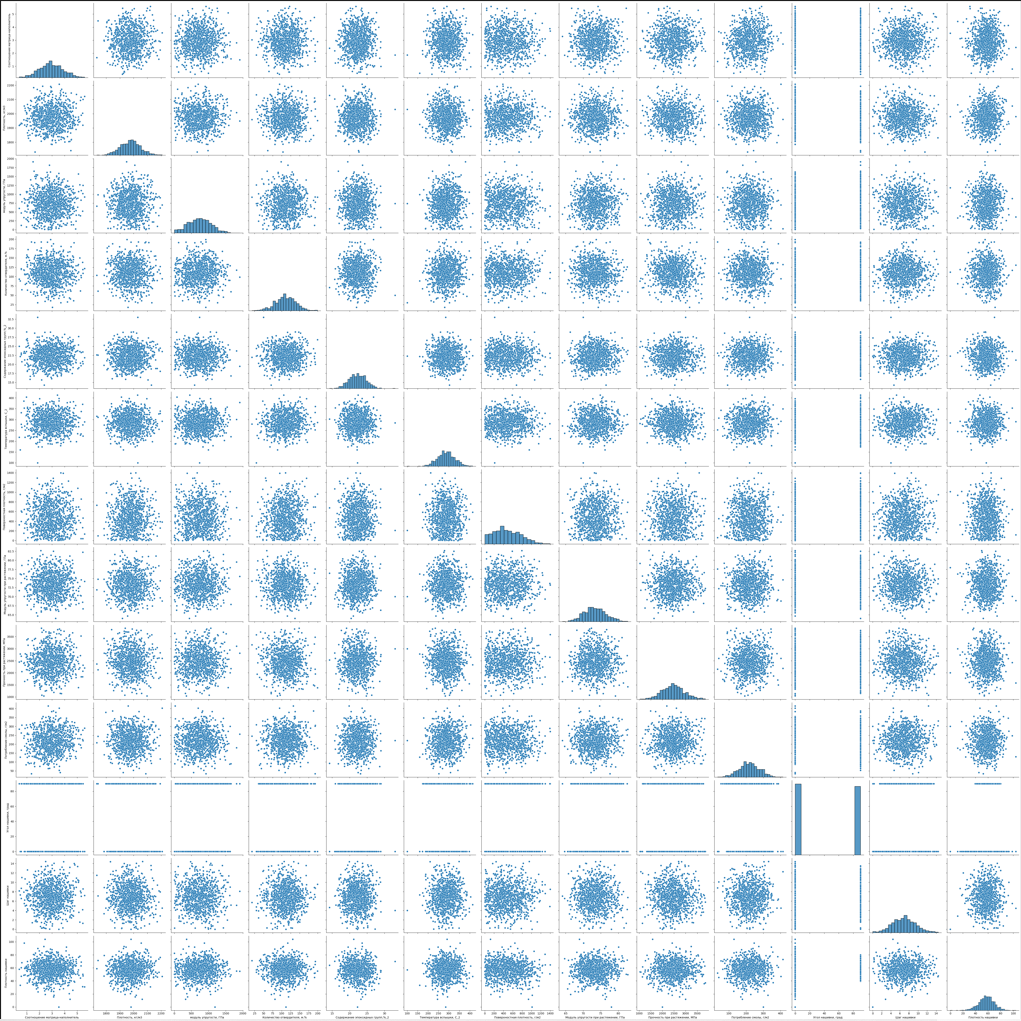


Рисунок 9. График попарного рассеяния.

Визуализация корреляционной матрицы (тепловая карта) также демонстрирует отсутствие корреляции между какими-либо параметрами. Обнаруживается слабая корреляция между Температурой вспышки и Количеством отвердителя (r =0.10), и Плотностью нашивки и Углом нашивки (r=0.11) (рис.10).

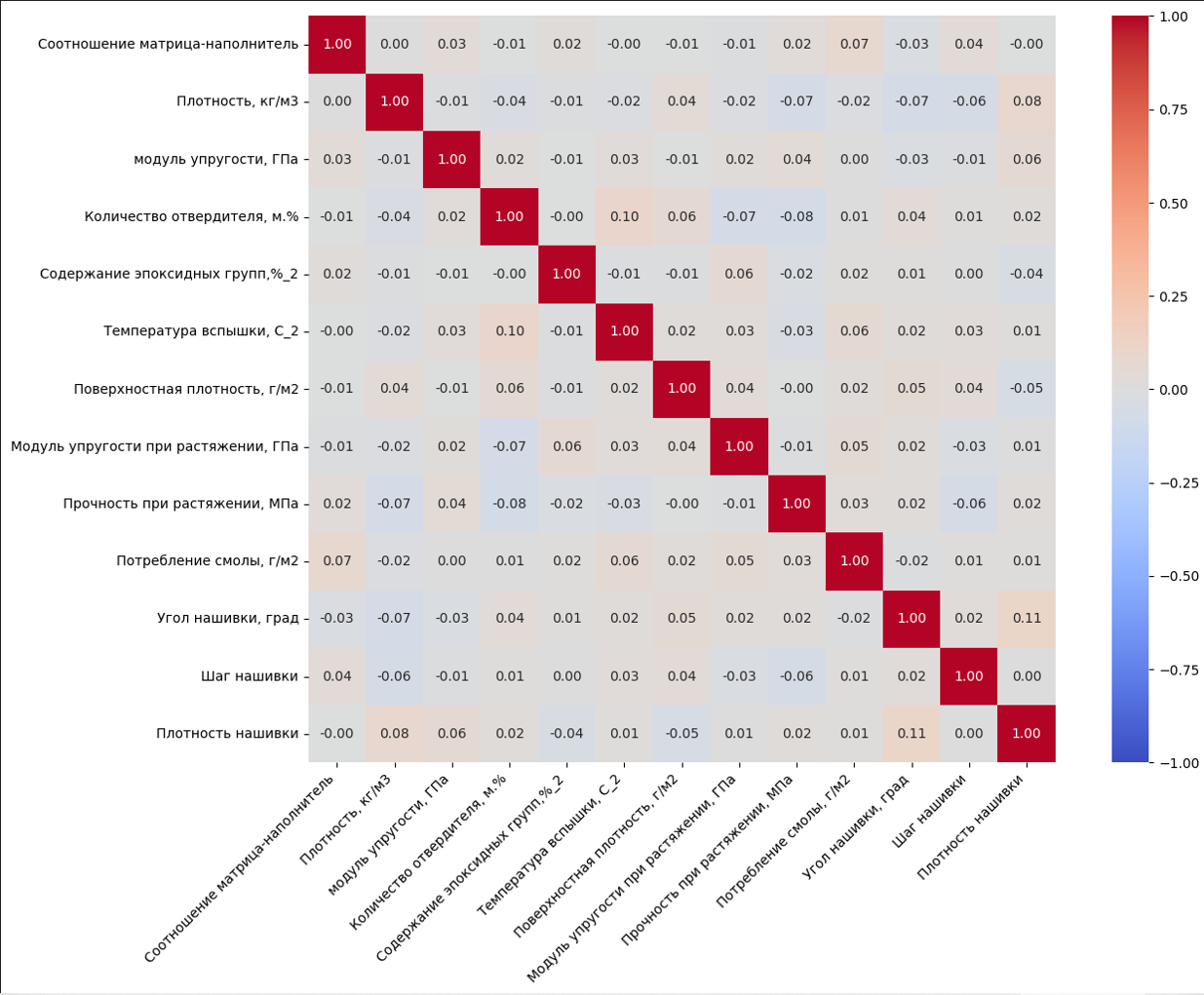


Рисунок 10. Корреляционная матрица датасета.

Визуализация с помощью диаграмм Boxplots (ящики с усами) используется для определения распределения данных и выявления выбросов. Также, позволяет быстро оценить медиану, интерквартильный размах. Ящик с усами представляет собой прямоугольник, который рисуется между первым и третьим квартилями данных, средней линией, которая обозначает медиану, и вертикальными линиями, которые выходят из прямоугольника и обозначают интервалы, содержащие большинство данных. Выбросы, если они есть, отображаются на графике в виде отдельных точек за пределами усов (рис. 8).



Рисунок. 8. Диаграммы Boxplots (ящики с усами) сырых данных.

Данный метод демонстрирует наличие выбросов в датасете, однако выбросы могут содержать важную информацию, и удаление их может исказить результаты анализа.

1. **Практическая часть**
   1. **Предобработка данных**

Удаление выбросов в датасете осуществлялось с помощью метода межквартильных расстояний (рис. 11).

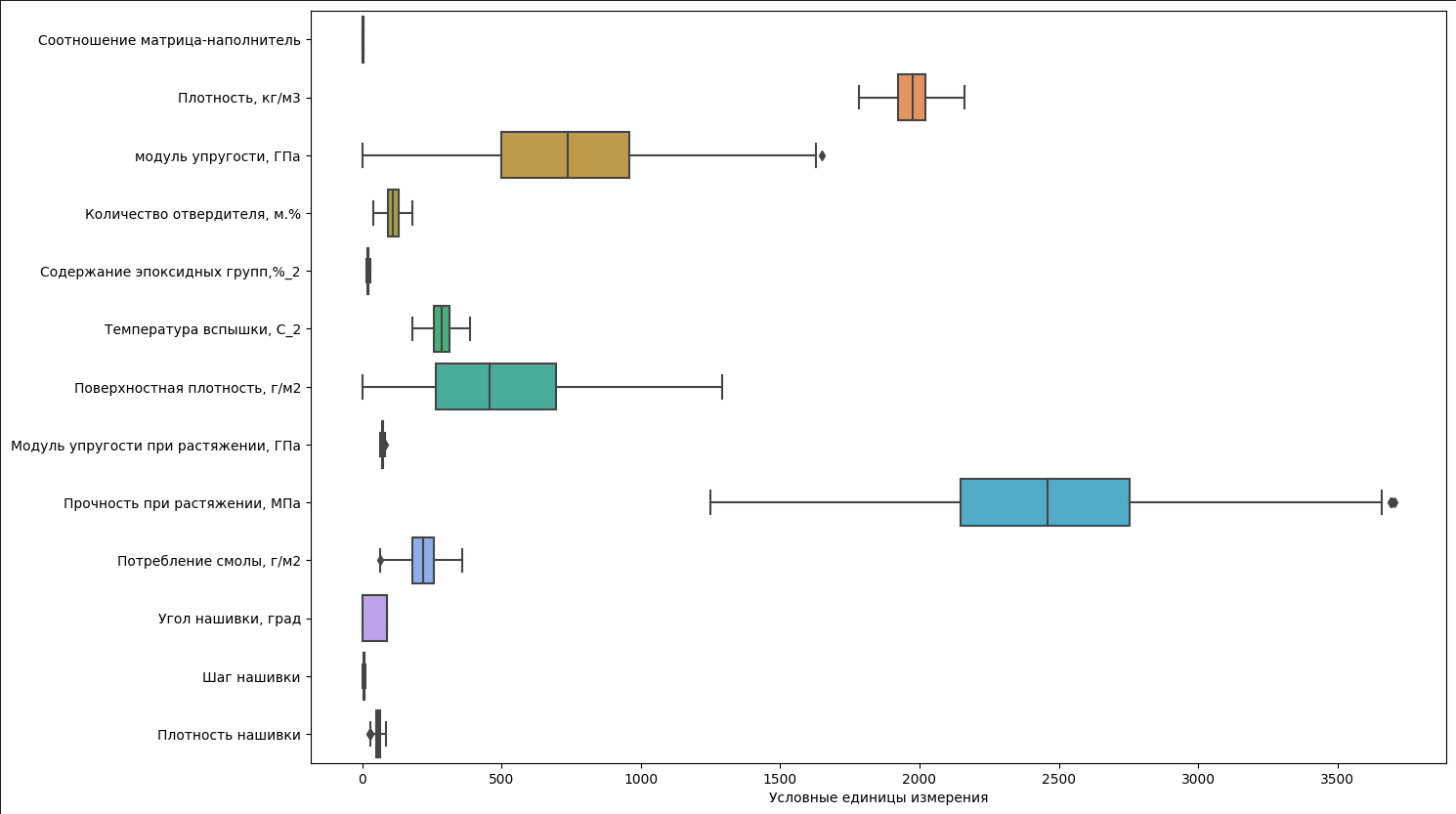


Рисунок 11. Диаграммы Boxplots (ящики с усами) в результате удаления выбросов.

Визуализация плотности ядра (рис. 12) демонстрирует значительный разброс плотности распределения данных.

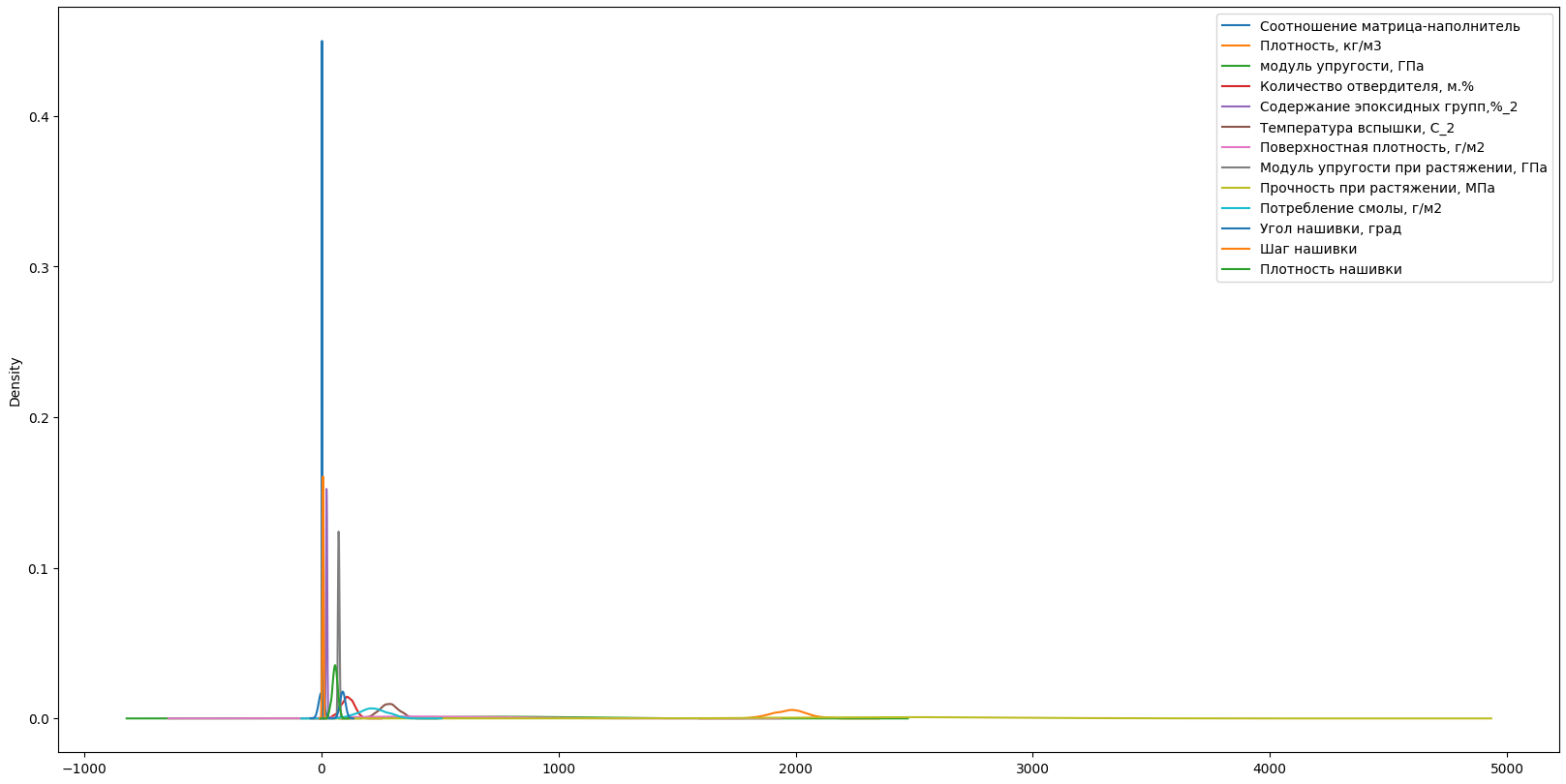


Рисунок 12. График распределения плотности ядра.

Для разработки и обучения моделей необходимо провести масштабирование данных. В данной работе использовался метод MinMaxScaler для нормализации данных (рис. 13). Метод MinMaxScaler масштабирует каждую функцию до заданного диапазона от 0 до 1. Формула для масштабирования функции x выглядит следующим образом:

x\_scaled=(x-x\_min)/(x\_max-x\_min),

где x\_min это минимальное значение функции, а x\_max это максимальное значение функции.

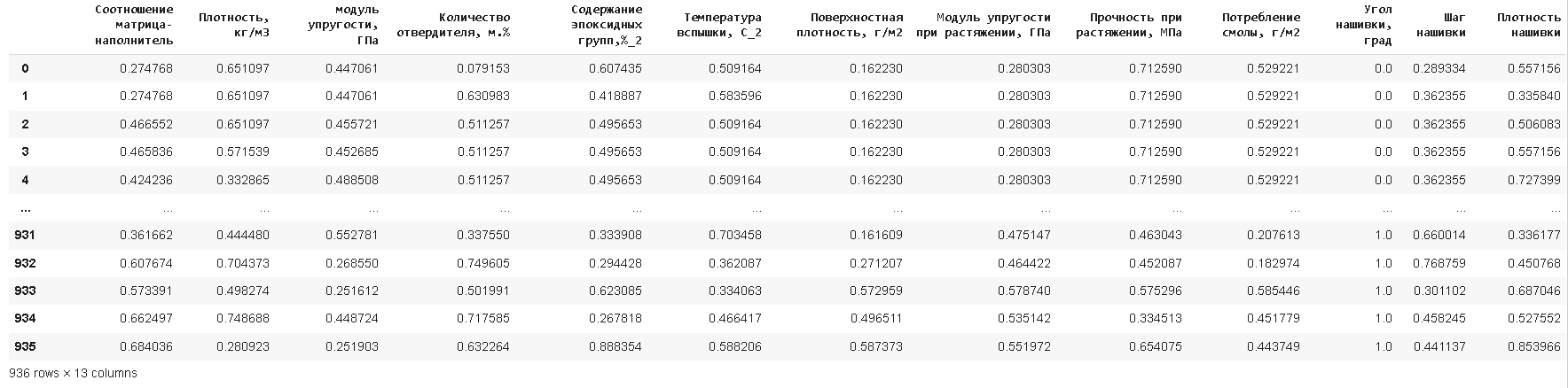


Рисунок 13. Таблица нормализованных данных.

Описательная статистика нормализованных данных демонстрирует диапазон значений в пределах от 0 до 1 (рис. 14).



Рисунок 14. Описательная статистика нормализованных данных.

Оценка плотности ядра нормализованного датасета показывает распределения данных, и позволяет выявить частоту встречаемости различных значений (рис. 15.).

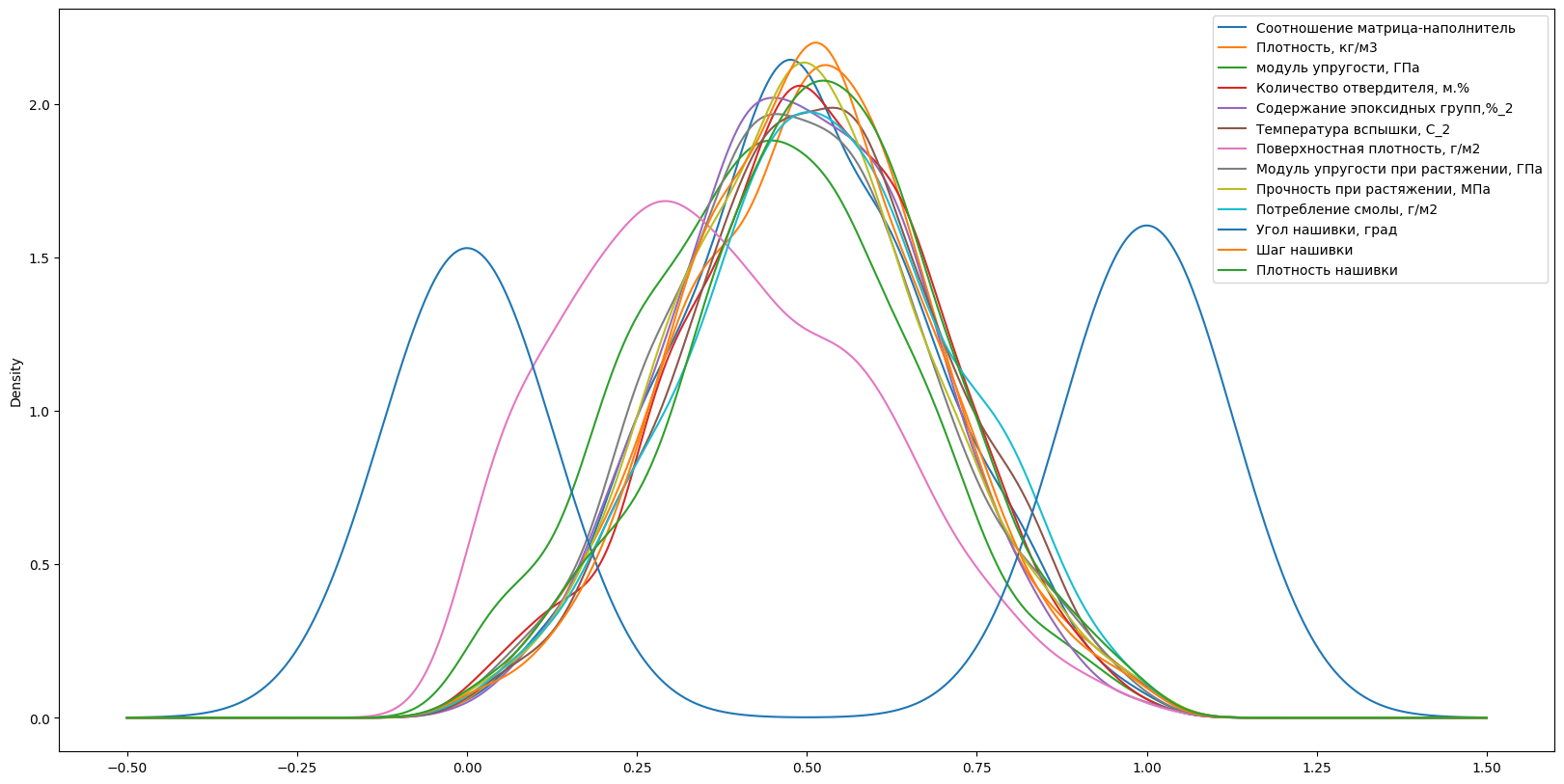


Рисунок 15. График распределения плотности ядра нормализованного датасета.

Ящики с усами после нормализации показывают относительно равномерное распределение данных, их медианы и интерквартильного размаха, кроме Угла нашивки (значительно больший размах интерквартильного размаха) и Поверхностная плотность (значительное смещение медианы) (рис. 16).

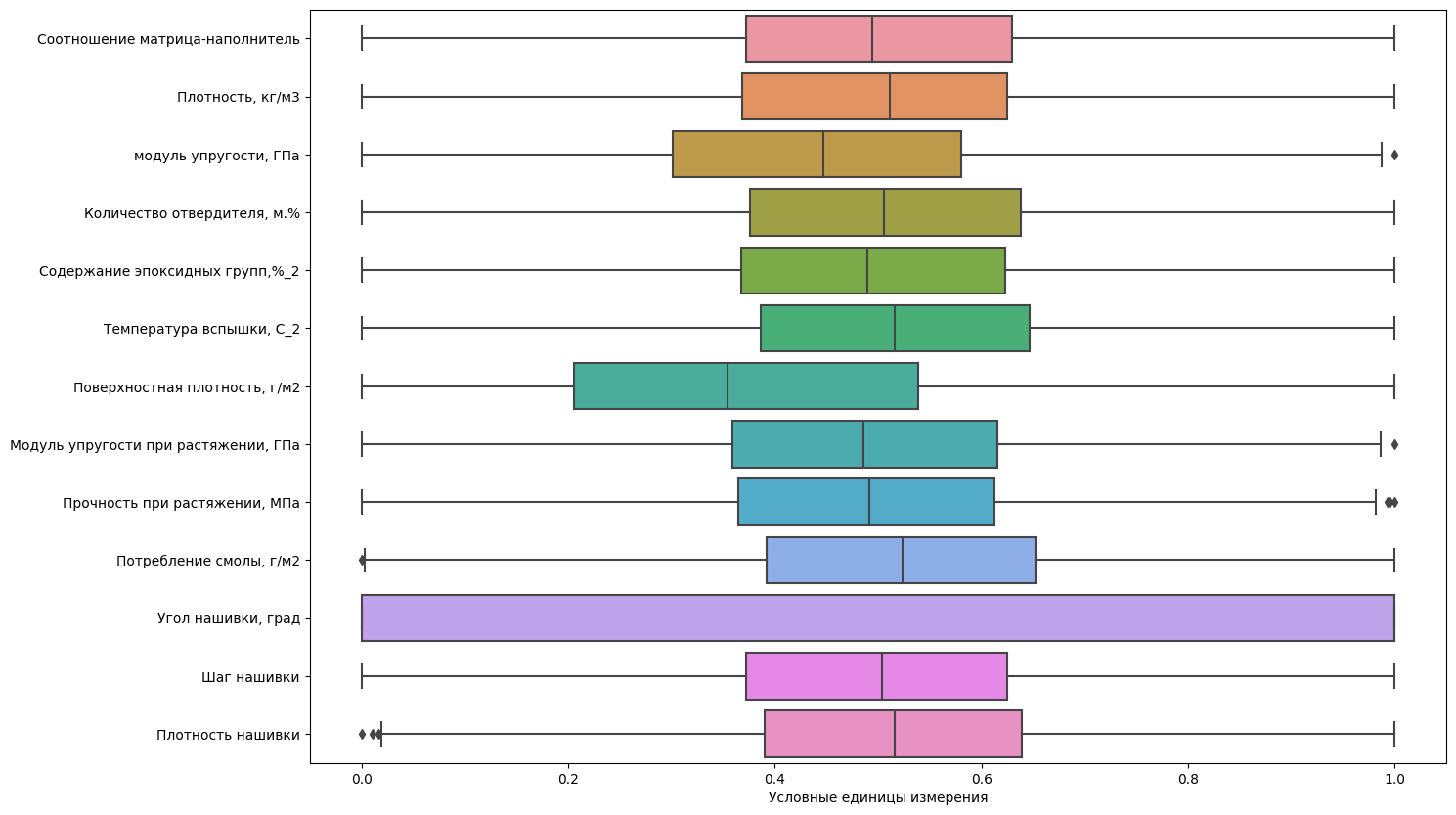


Рисунок 16. Диаграммы ящики с усами нормализованных данных.

Корреляционная матрица показывает отсутствие значимых значений коэффецентов корреляции Пирсона (рис. 17).

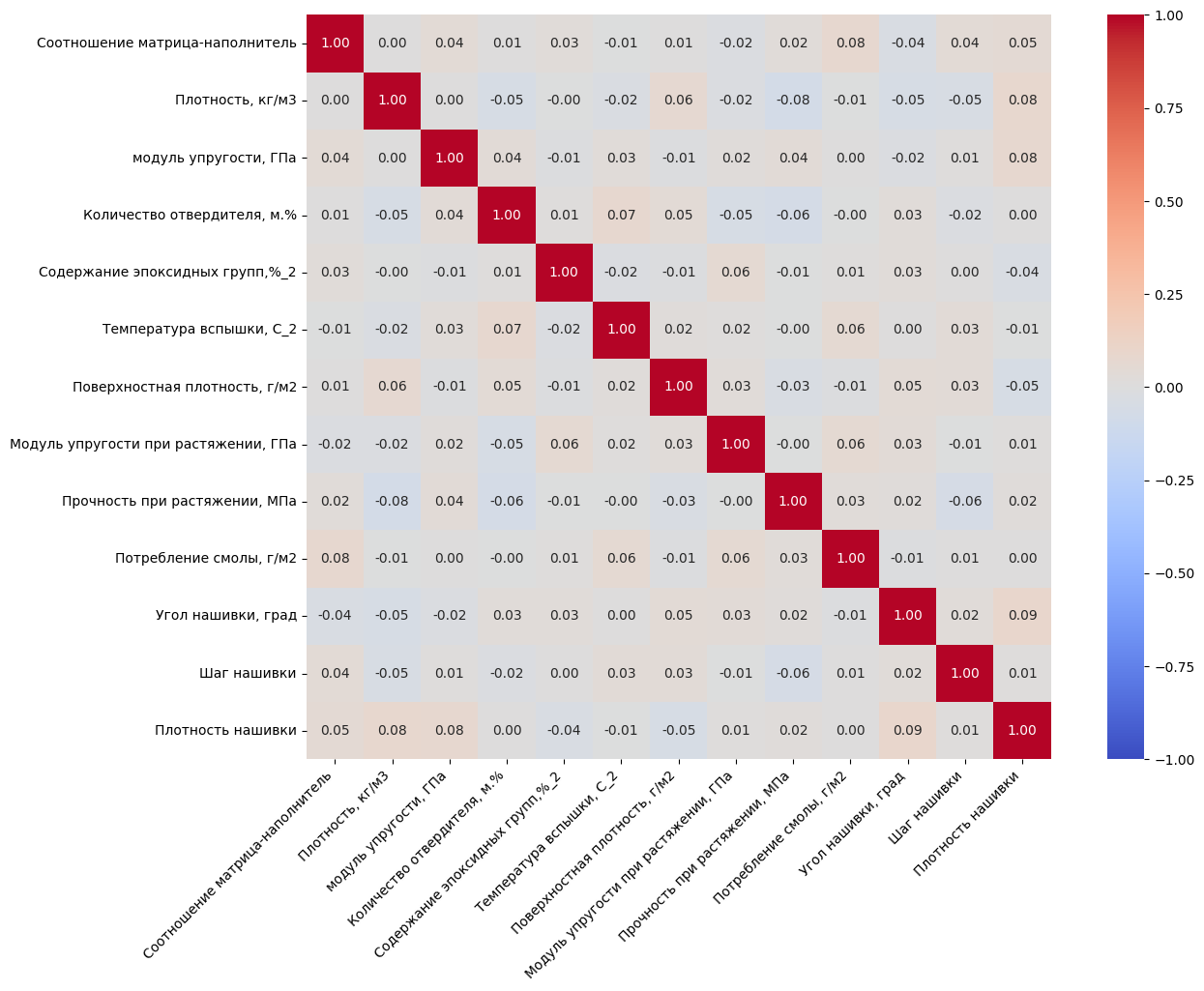


Рисунок 17. Корреляционная матрица нормализованных данных.

* 1. **Разработка и обучение моделей**

Модели машинного обучения разрабатывались для двух параметров: Модуля упругости при растяжении и Прочности при растяжении. Разработка модели состоит из нескольких этапов:

Для визуализации предсказанных значений были разработаны функции для целевых переменных Модуля упругости при растяжении и Прочности при растяжении, и одной целевой переменной Матрица-наполнитель.

Нормализация датасета предполагает сохранение масштабаторов для масштабирования входных и выходных данных (minmax\_y.joblin, minmax\_x.joblin). Количественное (в процентах, %) разделение на обучающую (70%) и тестовую (30%) выборки было разделено в соответствии с условиями поставленной задачи.

Использование базовых моделей со стандартными параметрами.

Проверка моделей на нормализованном датасете.

Оценка результатов базовых моделей по параметрам MSE и R2.

Оптимизация гиперпараметров и поиск лучшей модели.

При подборе лучшей модели для прогнозирования параметров рассматривались следующие методы:

Метод линейной регрессии (рис. 18, 19) предсказывает ценность неизвестных данных с помощью другого связанного и известного значения данных, математически моделируя неизвестную или зависимую переменную и известную или независимую переменную в виде линейного уравнения.

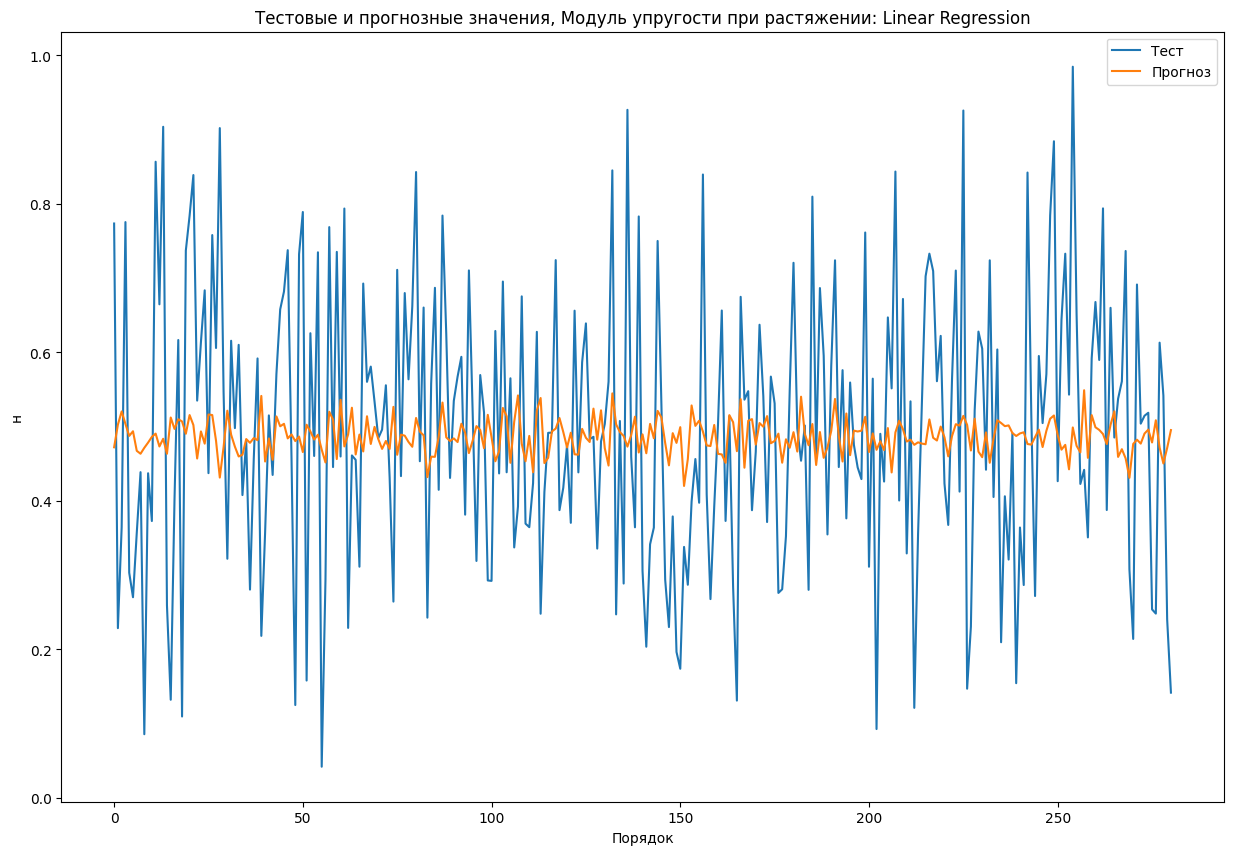


Рисунок 18. Тестовые и прогнозные значения Модуля упругости при растяжении методом линейной регрессии.



Рисунок 19. Тестовые и прогнозные значения прочности при растяжении методом линейной регрессии**.**

Метод k ближайших соседей (рис. 20, 21) простейший метрический классификатор, основанный на оценивании сходства объектов. Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки. Данный метод опирается на одно важное предположение, называемое гипотезой компактности: если мера сходства объектов введена достаточно удачно, то схожие объекты гораздо чаще лежат в одном классе, чем в разных.

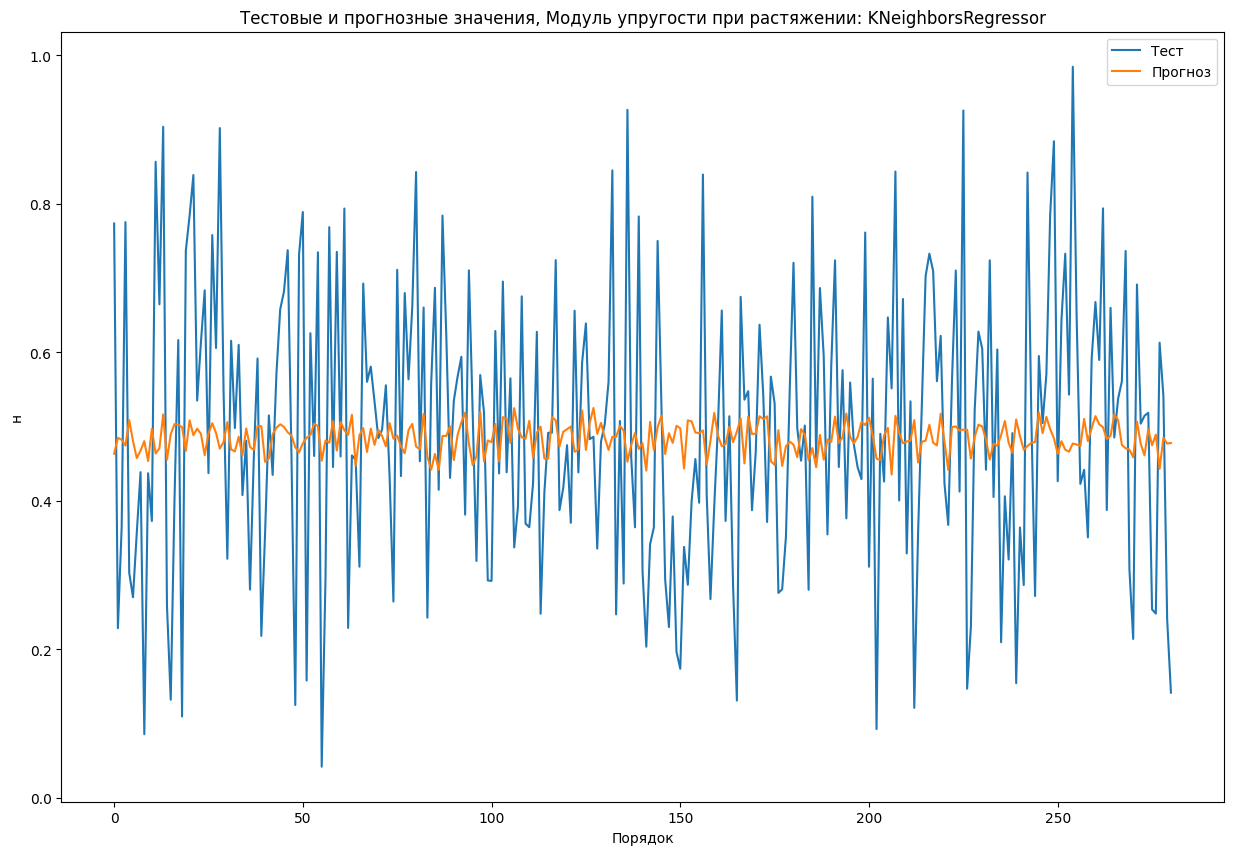


Рисунок 20. Тестовые и прогнозные значения Модуля упругости при растяжении методом k ближайших соседей.

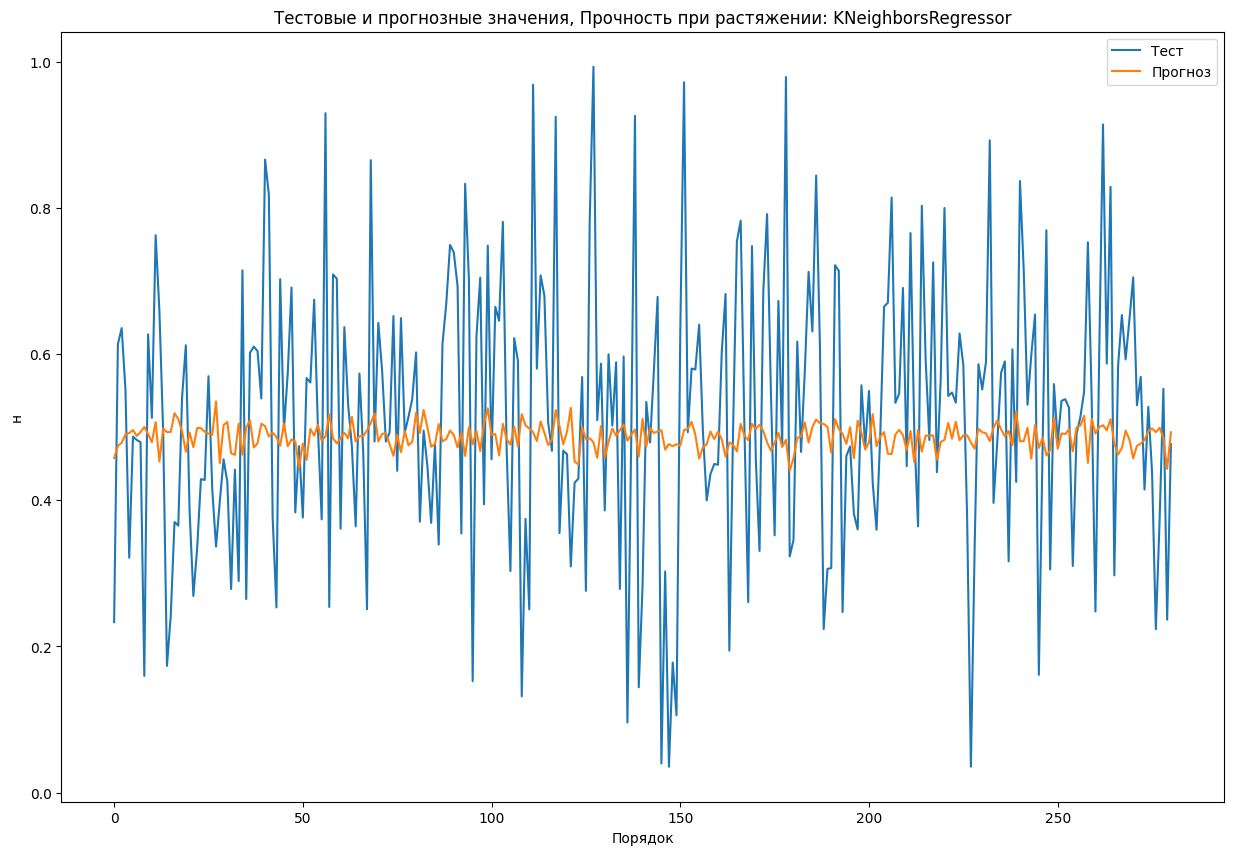
****

Рисунок 21. Тестовые и прогнозные значения прочности при растяжении методом k ближайших соседей**.**

Метод опорных векторов (рис. 22, 23) основан на идеи, что перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с наибольшим зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, создающая наибольшее расстояние до двух параллельных гиперплоскостей. Алгоритм основан на допущении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

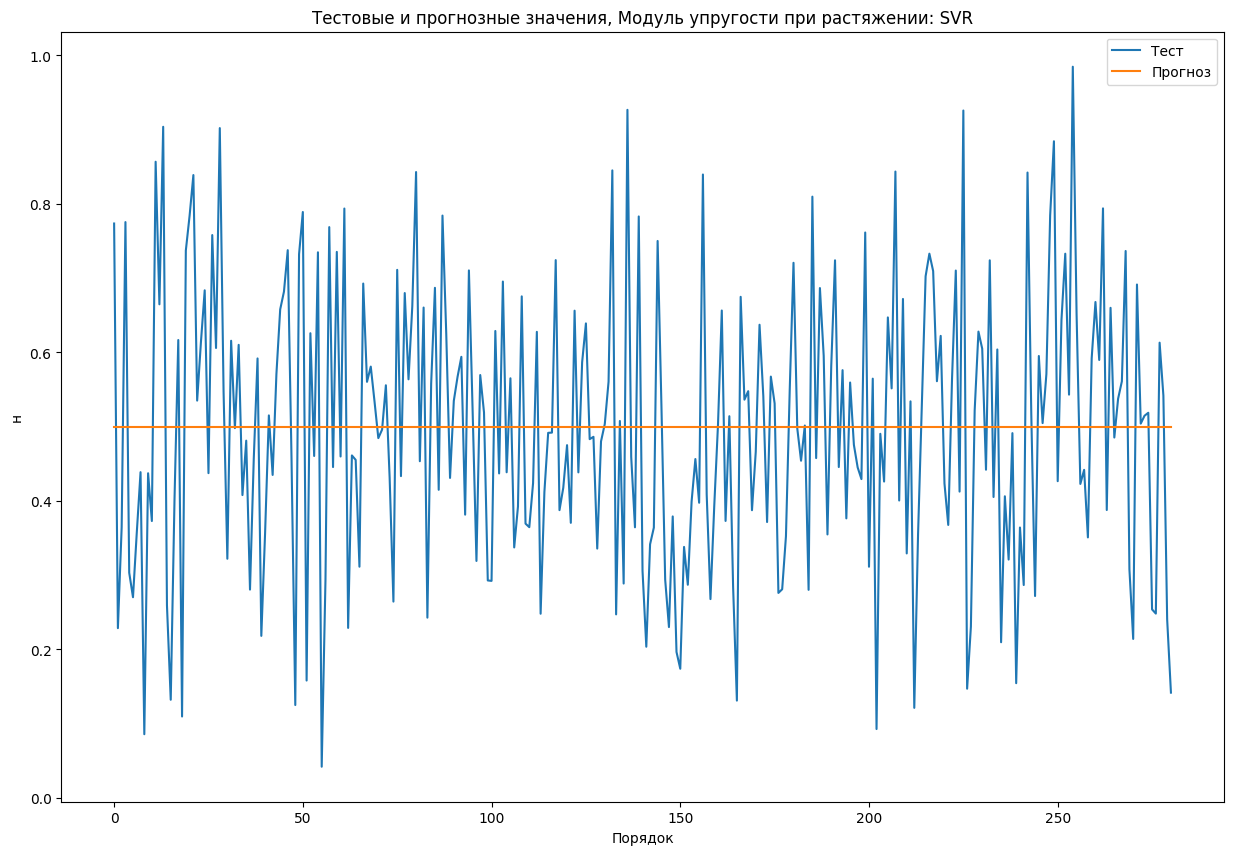


Рисунок 22. Тестовые и прогнозные значения Модуля упругости при растяжении методом опорных векторов.

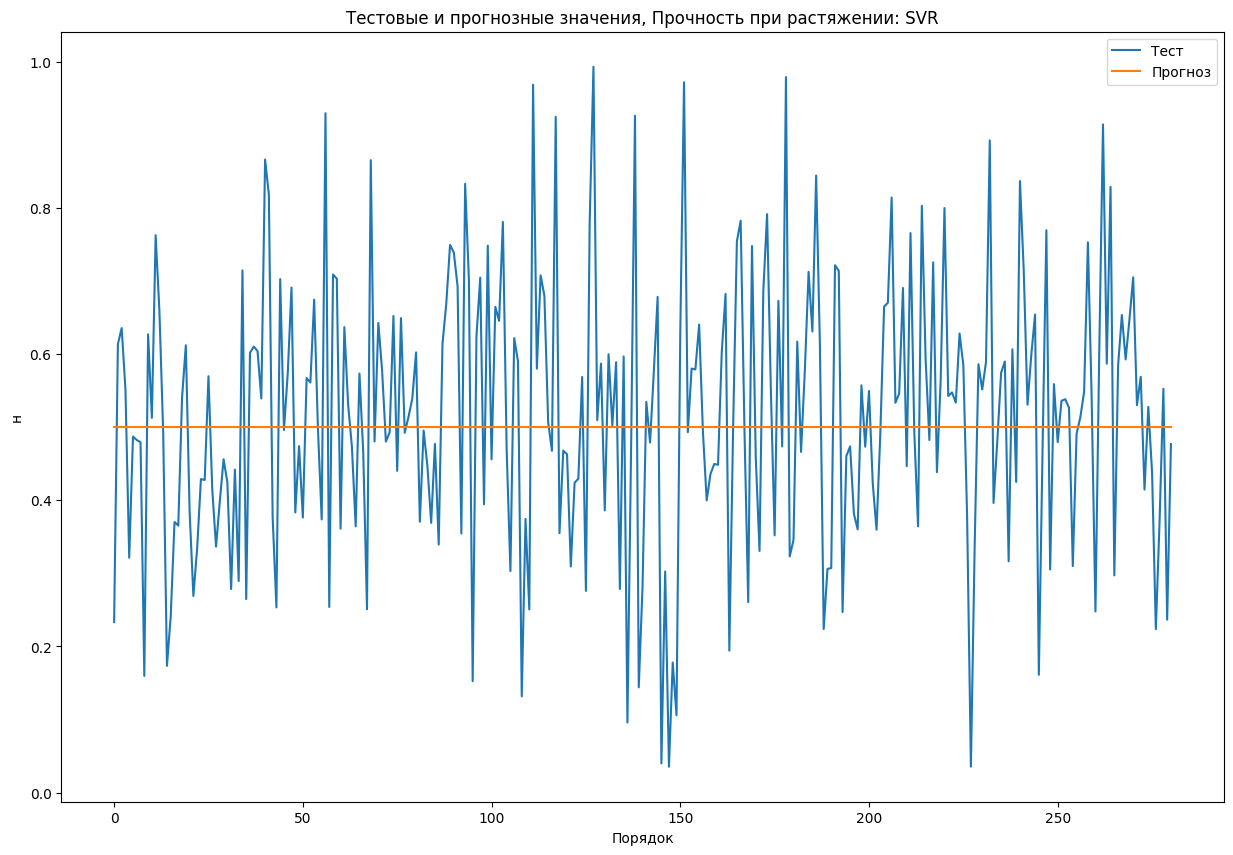


Рисунок 23. Тестовые и прогнозные значения прочности при растяжении методом опорных векторов.

Метод случайного леса (рис. 24, 25) использует ансамбль деревьев решений для задач классификации. Каждое отдельное дерево в таком лесу дает предсказание класса, и набравший наибольшее количество голосов класс, становится предсказанием модели.

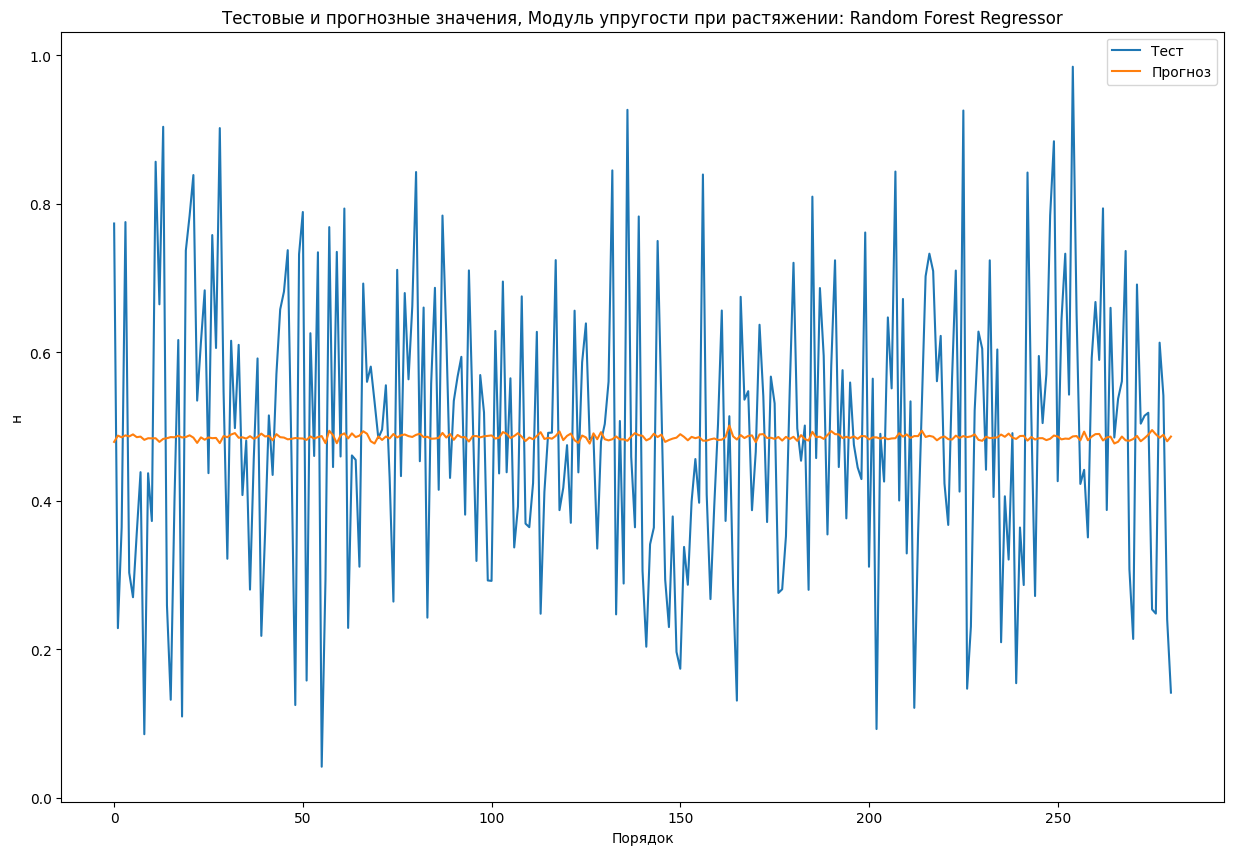


Рисунок 24. Тестовые и прогнозные значения Модуля упругости при растяжении методом случайного леса.

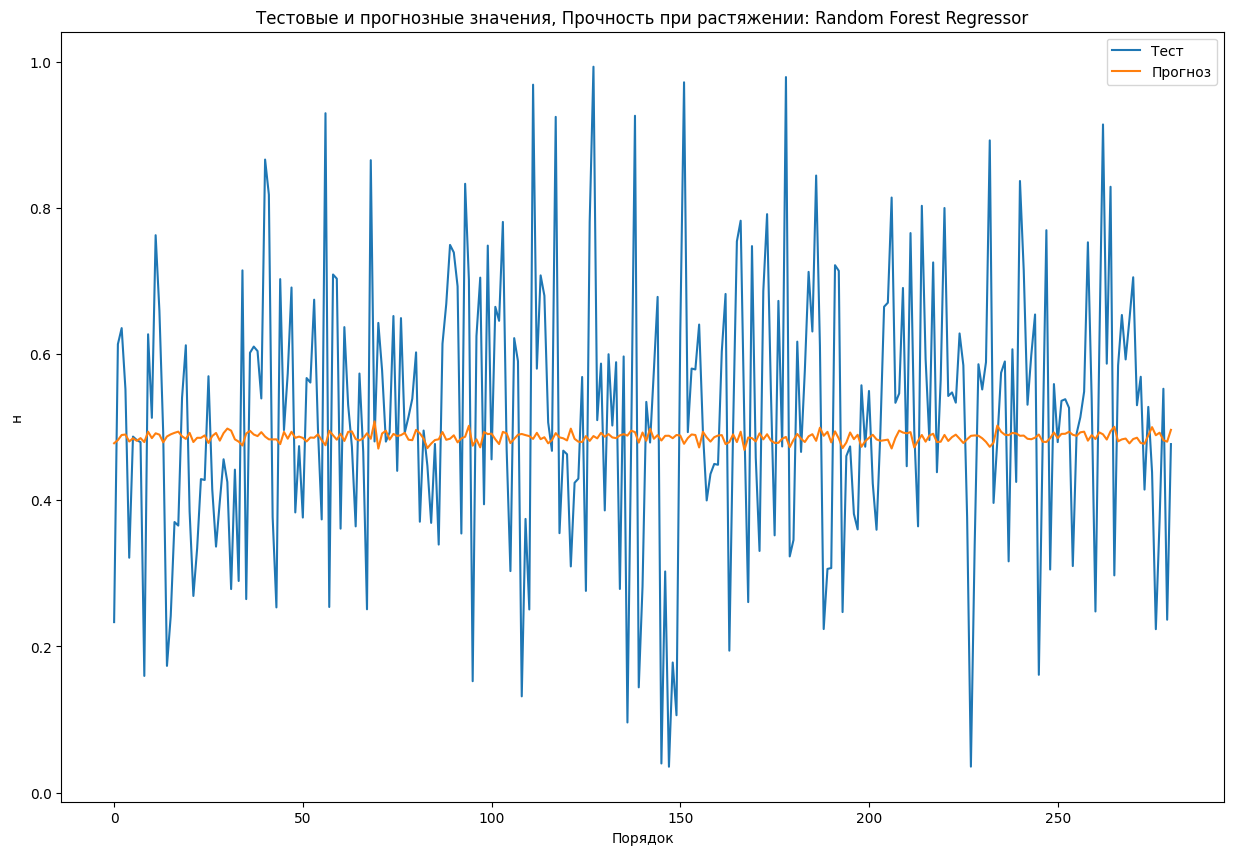


Рисунок 25. Тестовые и прогнозные значения прочности при растяжении методом случайного леса.

Метод Лассо регрессия (рис. 26, 27) заключается во введении дополнительного слагаемого регуляризации в функционал оптимизации модели, что часто позволяет получать более устойчивое решение. Иными словами это регуляризованный алгоритм регрессии, который выполняет регуляризацию L1, которая добавляет штраф, равный абсолютному значению величины коэффициентов.



Рисунок 26. Тестовые и прогнозные значения Модуля упругости при растяжении методом Лассо регрессии.

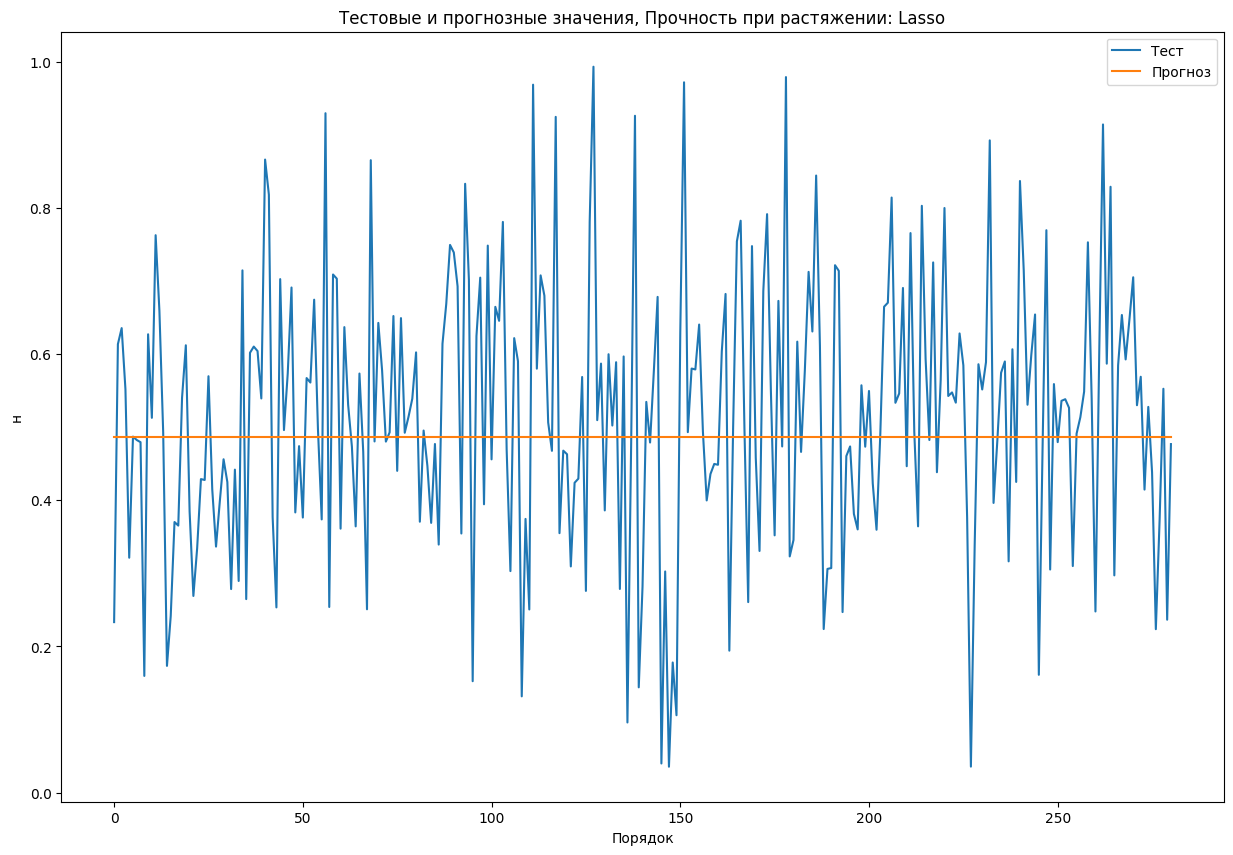


Рисунок 27. Тестовые и прогнозные значения прочности при растяжении методом Лассо регрессии.

* 1. **Тестирование моделей**

Оценка результатов базовых моделей производилась по параметрам среднеквадратичной ошибки (MSE – среднее арифметическое квадратов разностей между предсказанными и реальными значениями) и коэффициенту детерминации (R2 – показывает, что на 74,5% расчетные параметры модели, то есть сама модель, объясняют зависимость и изменения изучаемого параметра -Y от исследуемых факторов - X. Можно сказать что, это показатель качества модели и чем он выше, тем лучше. Он не может быть больше 1 и считается неплохо, когда R2 выше 0,8, а если меньше 0,5, то смысл такой модели можно смело ставить под большой вопрос.) (рис. 28).



Рисунок 28. Метрики эффективности моделей (без индекса общая, \_0 – Модуль упругости при растяжении, \_1 – Прочность при растяжении).

По итогам подбора и тестирования моделей, точность базовых моделей и моделей с подобранными гиперпараметрами не сильно отличаются. В среднем, подбор гиперпараметров позволяет улучшить результаты, но не до такой степени, чтобы можно было получить надежный метод прогнозирования с приемлемыми метриками. Во всех случаях значение коэффициента детерминации отрицательное, то есть ни одна из моделей не способна дать результат лучше среднего значения прогнозируемого параметра. По итогам тестирования наиболее удачной моделью для прогнозирования Модуля упругости при растяжении и Прочности при растяжении решено было выбрать метод опорных векторов с гиперпараметрами MultiOutputRegressor (estimator=SVR(epsilon=0.5)).

* 1. **Написание нейронной сети, которая рекомендует соотношение «матрица-наполнитель»**

Для данной задачи была использована модель нейронной сети Sequential из библиотеке Keras. Модель Sequential представляет собой линейную структуру, в которой слои добавляются один за другим. Каждый слой принимает выходные данные предыдущего слоя в качестве входных данных. Таким образом, в модели Sequential нет возможности объединять неколько входов и выходов.

В качестве подготовки к работе с нейросетью нормализованный датасет был разделен на тестовую (30%) и обучающую (70%) выборки, с целевым параметром Соотношение Матрица-наполнитель.

В модели входной слой Dense() имеет 128 нейронов и активационная функция relu. Далее два скрытых слоя 64 и 8 нейронов с активационной функцией relu. Выходной слой имеет 1 нейрон без активационной функции. Количество эпох 30, размер валидационной выборки 20%. Обученная модель показала следующие результаты (рис. 29, 30, 31).

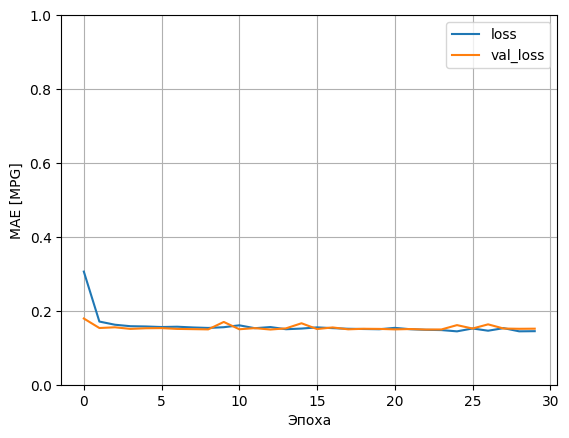


Рисунок 29. График потерь.

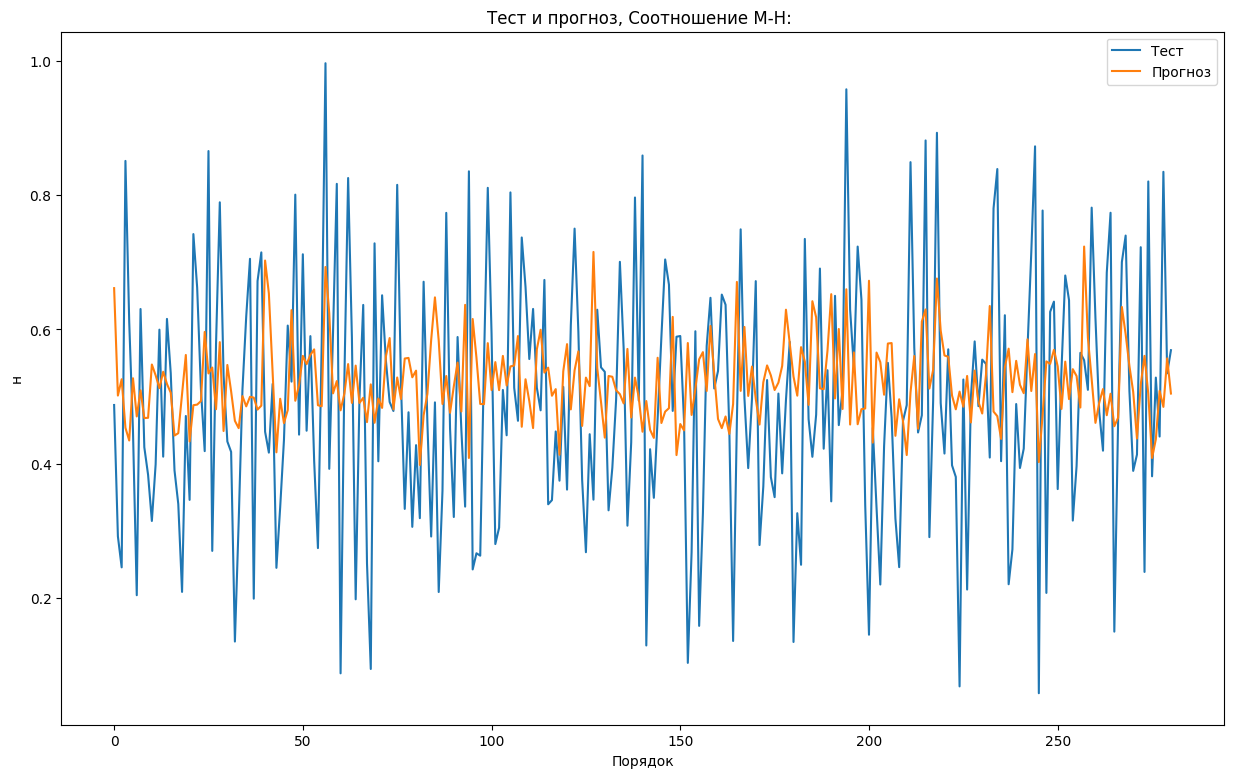
****

Рисунок 30. График тестовых и прогнозных значений параметра Соотношение матрица-наполнитель.

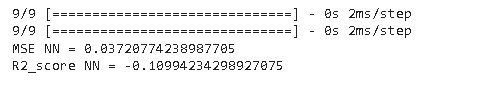


Рисунок 31. Метрики эффективности модели.

Разработанная модель показывает низкую эффективность прогноза.

* 1. **Разработка приложения**

Для данной задачи была использована модель прогноза параметров Модуля упругости при растяжении и Прочности при растяжении. Приложение принимает на вход 11 параметров и для прогноза использует модель, построенную на основе метода опорных векторов (рис.32.).



Рисунок 32. Код приложения.

* 1. **Создание удаленного репозитория и загрузка**

Адрес репозитория на Github: https://github.com/OlgaRash/-\_-

**Заключение**

Целью проекта являлось создание моделей прогнозирования, которые помогут сократить количество проводимых испытаний и пополнить базу данных материалов. В ходе работы был решен ряд задач, связанных с обработкой и анализом данных.

В работе был проведен разведочный анализ данных, который показал, что распределение данных в объединенном датасете близко к нормальному, а коэффициенты корреляций между параметрами очень низкие. Далее была проведена предобработка данных, включающая проверку на пропуски и выбросы, и нормализацию. Далее были разработаны и обучены модели машинного обучения с разными параметрами, протестировано несколько вариантов настроек, а также создано приложение для прогноза параметров Модуля упругости при растяжении и Прочности при растяжении.

Базовые модели, простой подбор гиперпараметров и модель нейросети не показали высокой точности в прогнозировании целевых результатов. Возможно решение лежит через получение дополнительных данных или через глубокое изучение взаимосвязей между параметрами и поиск сложных закономерностей.

**Библиографический список**

1. Любанович Билл, Простой питон. Совеменный стиль программирования. 2-е изд.-СПб.: Питер, 2022.-592с.:ил.
2. Силен Дэв, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных.-СПб.:Питер, 2017.-226с.:ил.
3. <https://matplotlib.org/stable/api/>
4. <https://ru-keras.com/home/>
5. <https://scikit-learn.org/stable/index.html>
6. <https://seaborn.pydata.org/tutorial.html>
7. <https://pandas.pydata.org/docs/user_guide/index.html>
8. <https://joblib.readthedocs.io/en/latest/index.html>