**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 Виноградова Анжела Геннадьевна

Москва 2024

СОДЕРЖАНИЕ

Введение…………………………………………………………………………3

1. Аналитическая часть………………………………………………………...... 5
   1. Постановка задачи……………………………………………………. 5
   2. Изучение предметной области
   3. Описание используемых методов…………………………… 7
   4. Разведочный анализ данных…………………………………… 13
2. Практическая часть…………………………………………………………….. 20
   1. Предобработка данных……………………………………………. 20
   2. Разработка и обучение модели……………………………….. 21
   3. Тестирование модели……………………………………………… .27
   4. Сравнение показателей…………………………
   5. Разработка приложения…………………………………………… 30
   6. Создание репозитория……………………………………………… 31

Заключение……………………………………………………… 32

Список используемой литературы………………………………………………..33

**1 Аналитическая часть**

**1.1 Постановка задачи**

**Тема:**

Прогнозирование конечных свойств новых материалов (композиционных материалов).

**Описание:**

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).

**Актуальность:**

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

Задание:

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

* Модуль упругости при растяжении, ГПа;
* Прочность при растяжении, МПа.

1. Написать нейронную сеть, которая будет рекомендовать:

* Соотношение матрица-наполнитель.

1. Написать приложение, которое будет выдавать прогноз, полученный в задании 1 или 2 (один или два прогноза, на выбор учащегося).
2. Создать профиль на github.com.
3. Сделать commit приложения на github.com.
4. Сделать commit на веб-хостинг (по желанию учащегося).
5. Написать пояснительную записку к проекту, которая включает блок-схему и описание процесса подготовки, обучения моделей и инструкцию по установке и запуску приложения.

**1.2 Изучение предметной области**

Для композиционных конструкционных материалов характерны следующие признаки:

-состав и форма компонентов материала определены заранее;

-компоненты присутствуют в количествах, обеспечивающих заданные свойства материала;

-материал является неоднородными (компоненты различаются по свойствам, между ними существует явная граница раздела).

Один из компонентов, обладающий непрерывностью по всему объему материала, называют матрицей или связующим веществом. Другой компонент или компоненты, разделенные по объему композиции, называют усиливающими или армирующими наполнителями. Матричными материалами могут быть металлы и их сплавы, органические и неорганические полимеры, керамика и другие вещества.

Положительный эффект применения наполнителей выражается в увеличении прочности и жесткости материалов, улучшении теплопроводности и теплостойкости, повышении износостойкости и ударной вязкости, уменьшении коэффициента температурного расширения и пористости, улучшении качества поверхности, а в отдельных случаях -и в удешевлении материалов.

Композиционные материалы-это материалы настоящего и будущего. С развитием современных дисперсно-упрочненных композиционных материалов связаны широкие перспективы использования принципиально новых наноматериалов и нанотехнологий.

Особенность конструкций из композитных материалов состоит в том, что разработчики должны проектировать и сами материалы, обеспечивая расчётные, механические, тепловые и другие параметры и характеристики, необходимые для эффективной реализации инженерных разработок. Это связано с обоснованным выбором исходных материалов (армирующих волокон, связующих наполнителей), структуры композиционных материалов, технологии их изготовления и переработки в готовые изделия.

В задании у нас предложено:

Обучить алгоритм машинного обучения, который будет определять значения:

-Модуль упругости при растяжении, ГПа;

-Прочность при растяжении, МПа.

Написать нейронную сеть, которая будет рекомендовать:

-Соотношение матрица-наполнитель.

Узнаём, что это за параметры.

Прочность на растяжение- это максимальная нагрузка, которую образец может выдержать до того, как он сломается. Обычно выражается в единицах силы на единицу площади, например, в фунтах на квадратный дюйм(psi) или мегапаскалях(Мпа). Прочность материала на растяжение зависит от его состава, структуры и истории обработки.

Модуль упругости зависит от времени действия нагрузки. При мгновенном нагружении конструкции деформации пропорциональны величине внешних сил.

Физический смысл модуля упругости сводится к тому, что он характеризует сопротивление материала упругой деформации. Модуль упругости практически не зависит от структуры и определяется силами межатомной связи.

Соотношение матрица + наполнитель это сам композиционный материал.

Была выдвинута гипотеза, что показатели «Модуль упругости при растяжении» и «Прочность при растяжении» можно спрогнозировать по нашим физическим свойствам и другим данным с помощью машинного обучения. Так же попытаемся построить нейронную сеть, которая будет рекомендовать соотношение показателя «Матрица + наполнитель»

**1.3 Описание используемых методов**

У нас задача регрессии. Регрессия пытается предсказать одну зависимую переменную (обычно обозначаемую У) и серию других изменяющихся переменных (Известных как независимые переменные, обычно обозначаемые Х).

Метод опорных векторов.

Это алгоритм классификации, предназначенный для поиска хороших «решающих границ», разделяющих два класса.

Он выполняется в два этапа.

-Данные отображаются в новое пространство более высокой размерности, где граница может быть представлена как гиперплоскость.

-Хорошая решающая граница (разделяющая гиперплоскость) вычисляется путем максимизации расстояния от гиперплоскости до ближайших точек каждого класса. Это называется максимизацией зазора. Для метода регрессии так же можно использовать этот алгоритм применяя «зазоры».

Деревья решений.

Это иерархические структуры, которые позволяют классифицировать входные данные или предсказать выходные значения по заданным исходным значениям.

Однако, в ряде случаев, их применяют и для задач регрессии, когда алгоритм на выходе формирует одно вещественное значение (или несколько значений) для каждого входного вектора Х. Если говорить простыми словами, то дерево решений представляет собой задачу с несколькими вариантами действий. На карте прорабатываются возможные результаты каждого шага, а так же следующие на них реакции.Эффиктивный инструмент интеллектуального анализа данных и предсказательной аналитики. Оно помогает в решении задач по классификации и регрессии.

Лес деревьев.

Надёжный и практичный подход к обучению на основе деревьев решений. Он создаёт большое количество специализированных деревьев решений и в последующем объединяет выдаваемые ими результаты. Случайные леса применимы к широкому кругу проблем-можно сказать, они почти всегда являются оптимальным алгоритмом для любых задач машинного обучения.

Метод ближайших соседей KNN.

Алгоритм KNN работает путем нахождения К ближайших соседей к данной точке данных на основе метрики расстояния.Затем класс или значение точки данных определяется большинством голосов или средним значением К соседей. Такой подход позволяет алгоритму адаптироваться к различным шаблонам и делать прогнозы на основе локальной структуры данных. Алгоритм К-ближайших соседей работает по принципу подобия , когда он предсказвает метку или значение новой точки данных путем рассмотрения меток или значений ее К-ближайших соседей в обучающем наборе данных.

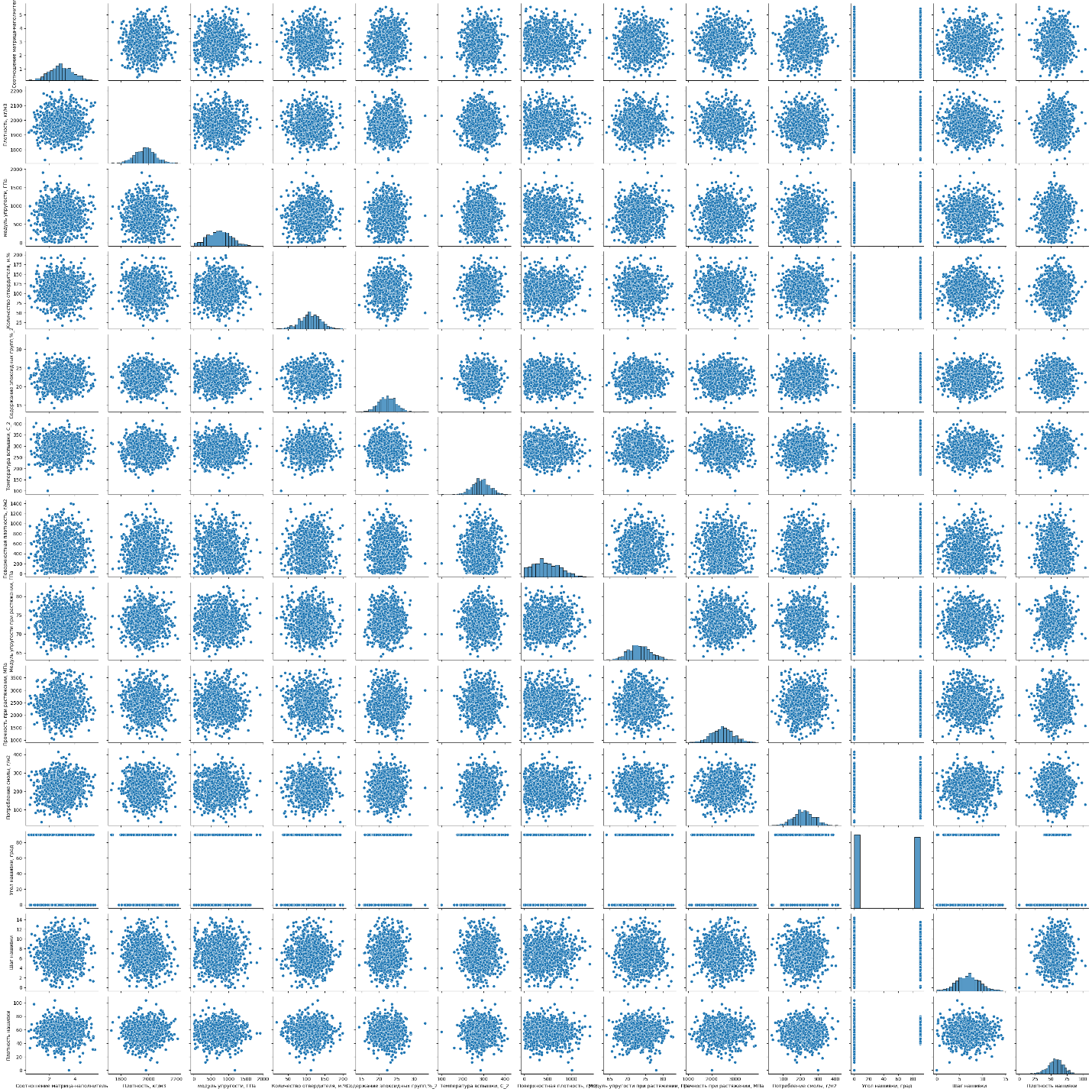
CatBoots.

CatBoost – это алгоритм для повышения градиента на деревьях решений. Она разработана исследователями и инженерами Яндекса и используется для поиска, систем рекомендаций, самоуправляемых автомобилей, прогнозирования погоды и многих других задач.Метод градиентного бустинга ( во многом напоминающий случайный лес) -это приём машинного обучения , основанный на объединении слабых моделей прогнозирования , обычно деревьев решений. Приминительно к деревьям решений настоящий приём позволяет получить модели, которые в большинстве случаев превосходят случайные леса , сохраняя анологичные свойства.

1.4 Разведоточный анализ данных.

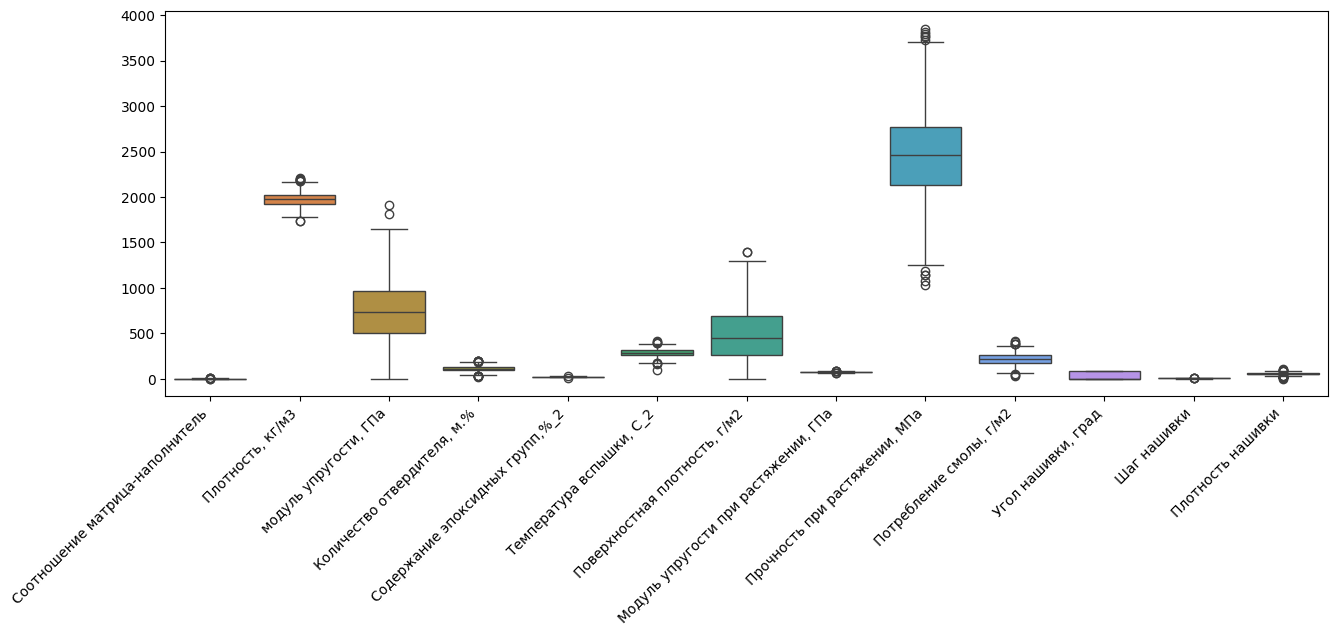
Загрузим данные. По заданию нам нужно объединить две таблицы данных в одну. Для начала проверила размер таблицы, пропуски в данных. Пропуски не обнаружены. Набор данных для машинного обучения и построения нейронной сети является не большим. Все показатели являются числовыми.При выведении описательной статистики мы видим, что данные имеют разные величины измерения и разный диапозон.Значит для задач машинного обучения и построения нейронной сети их нужно будет масштабировать. Будем использовать метод StandardScaler. Этот метод удалеет среднее и масштабирует каждую характеристику/переменную до единичной дисперсии. Эта операция выполняется функционально и независимо. На StandrdScaler могут влиять выбросы. Визуализируем матрицу рассеяния.Смотрим рисунок 1.

Рисунок 1- Матрица рассеяния данных.

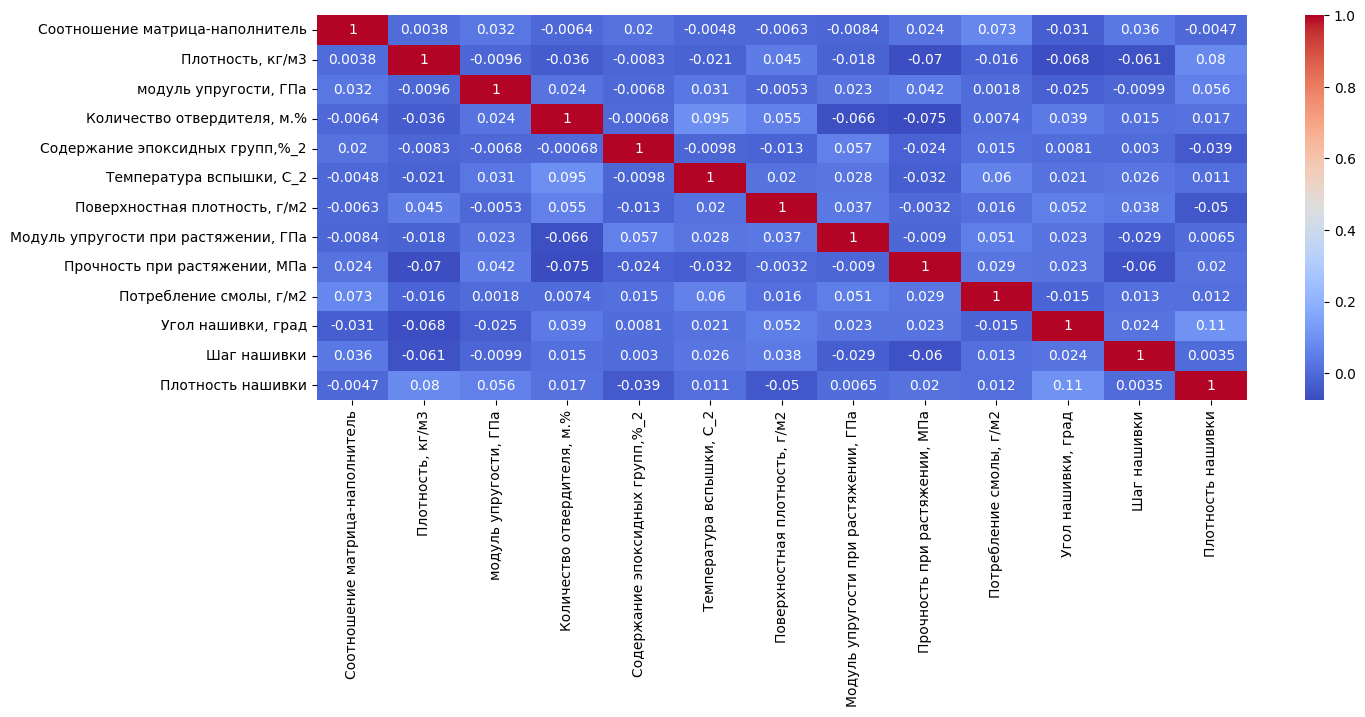


На графике мы видим, что данные между собой не имеют каких -то зависимостей. Кроме того можно заметить, что есть точки которые удалены от общего облака, а значит это скорее всего выбросы. Есть один столбец ‘Угол нашивки’, который принимает два значения. Если проверить уникальные значения этого столбца, то увидим что значения в нем распределены примерно одинаково (перевеса нет). Давайте посмотрим на график «Ящик с усами» ,где основная часть «прямоуголник» показывает межквартильный диапозон, который является средней частью данных.   Так же можно будет увидеть, есть ли данные которые выходят за рамки средней части данных. Такие значения будем считать выбросами. Посмотрим на рисунок 2.

Рисунок2 - Диаграмма «Ящик с усами» для всех данных.



Проанализировав данную диаграмму можно заметить, что есть «выбросы».

Рисунок 3- Матрица корреляций по всем параметрам

На рисунке 3 я построила матрицу корреляций по всем параметрам. Корреляционная матрица – это матрица которая показывает степень линейной зависимости между парами случайных переменных.Она используется для измерения силы и направления связи между переменными в наборе данных. Коэффициент корреляции принимает значения от -1 до 1. В наших данных мы видим, что линейных зависимостей нет. Проверим корреляцию наших целевых показателей с другими данными (таблица 1 и таблица 2).

Таблица 1-Корреляция «Модуль упругости при растяжении»

|  |  |
| --- | --- |
| Плотность, кг /м3 | -0.017602 |
| Модуль упругости, ГПа | 0.023267 |
| Количество отвердителя ,м.% | -0.065929 |
| Содержание эпоксидных групп, %\_2 | 0.056828 |
| Температура вспышки, С\_2 | 0.028414 |
| Поверхностная площадь,г/м2 | 0.036702 |
| Модуль упругости при растяжении , ГПа | 1.000000 |
| Прочность при растяжении, ГПа | -0.009009 |
| Потребление смолы,г/м2 | 0.050938 |
| Угол нашивки, град | 0.023003 |
| Шаг нашивки | -0.029468 |
| Плотность нашивки | 0.006476 |

Таблица 2 -Корреляция «Прочность при растяжении»

|  |  |
| --- | --- |
| Плотность, кг /м3 | -0.069981 |
| Модуль упругости, ГПа | 0.041868 |
| Количество отвердителя ,м.% | -0.075375 |
| Содержание эпоксидных групп, %\_2 | -0.031763 |
| Температура вспышки, С\_2 | -0.003210 |
| Поверхностная площадь,г/м2 | 0.036702 |
| Модуль упругости при растяжении , ГПа | -0.009009 |
| Прочность при растяжении, ГПа | 1.000000 |
| Потребление смолы,г/м2 | 0.028602 |
| Угол нашивки, град | 0.023003 |
| Шаг нашивки | -0.059547 |
| Плотность нашивки | 0.019604 |

Здесь можно более детально увидеть, что зависимости между целевыми переменными и другими показателями нет. Посмотрим на гистограмме распределение наших целевых переменных. Рисунок 4 и рисунок 5.

Рисунок 4-Гистограмма распределения «Модуля упругости при растяжении»

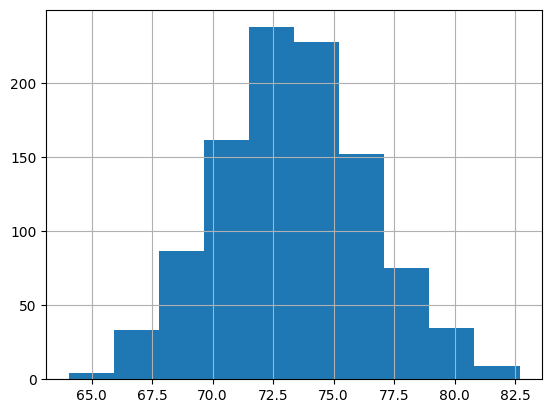
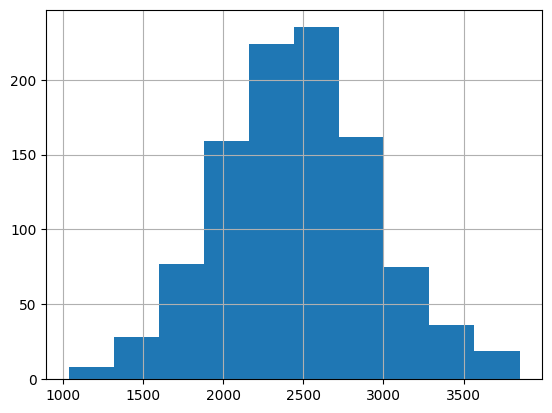


Рисунок 5- Гистограмма распределения «Прочность при растяжении»



Визуально мы посмотрели в каком диапазоне наши целевые переменные.

Посмотрим описательную статистику наших целевых переменных. Таблица 3 и таблица 4.

Таблица 3 -Описательная статистика «Модуль упругости при растяжении»

|  |  |
| --- | --- |
| count | 1023.000000 |
| mean | 73.328571 |
| std | 3.118983 |
| min | 64.054061 |
| 25% | 71.248018 |
| 50% | 73.268805 |
| 75% | 75.356612 |
| max | 82.682051 |

Таблица 4 -Описательная статистика «Прочность при растяжении»

|  |  |
| --- | --- |
| count | 1023.000000 |
| mean | 2466.922843 |
| std | 485.628006 |
| min | 1036.856605 |
| 25% | 2135.850448 |
| 50% | 2459.524526 |
| 75% | 2767.193119 |
| max | 3848.436732 |

Давайте посмотрим матрицу корреляций фик. **Phik** (Φk) — это новый и практичный коэффициент **корреляции**, который последовательно работает с категориальными, порядковыми и интервальными переменными. ... особенно полезен в современном **анализе** при изучении зависимостей между набором переменных со смешанными типами, где часто некоторые переменные являются категориальными. Рисунок 6.

В отличии от обычной корреляции здесь визуально кажется, что есть некоторые небольшие зависимости. Посмотрим также эти показатели в табличных данных с нашими целевыми переменными. Таблица 5 и таблица 6.

Таблица 5 -Корреляция «Модуль упругости при растяжении»

|  |  |
| --- | --- |
| Модуль упругости при растяжении | 1.000000 |
| Угол нашивки, град | 0.134983 |
| Поверхностная плотность | 0.127889 |
| Потребление смолы | 0.118782 |
| Количество отвердителя | 0.108480 |
| Температура вспышки | 0.099609 |
| Содержание эпоксидных групп | 0.097075 |
| Плотность нашивки | 0.027650 |
| Прочность при растяжении | 0.015080 |
| Соотношение матрица-наполнитель | 0.000000 |
| Плотность кг/м3 | 0.000000 |
| Модуль упругости | 0.000000 |
| Шаг нашивки | 0.000000 |

Таблица 6 -Корреляция «Прочность при растяжении»

|  |  |
| --- | --- |
| Прочность при растяжении | 1.000000 |
| Количество отвердителя | 0.221462 |
| Шаг нашивки | 0.167572 |
| Модуль упругости при растяжении | 0.160940 |
| Потребление смолы | 0.127832 |
| Поверхностная плотность | 0.089266 |
| Соотношение матрица-наполнитель | 0.086151 |
| Плотность кг/м3 | 0.076788 |
| Модуль упругости | 0.000000 |
| Содержание эпоксидных групп | 0.000000 |
| Температура вспышки | 0.000000 |
| Угол нашивки | 0.000000 |

Посмотрим диаграммы «Ящик с усами» для каждого показателя.

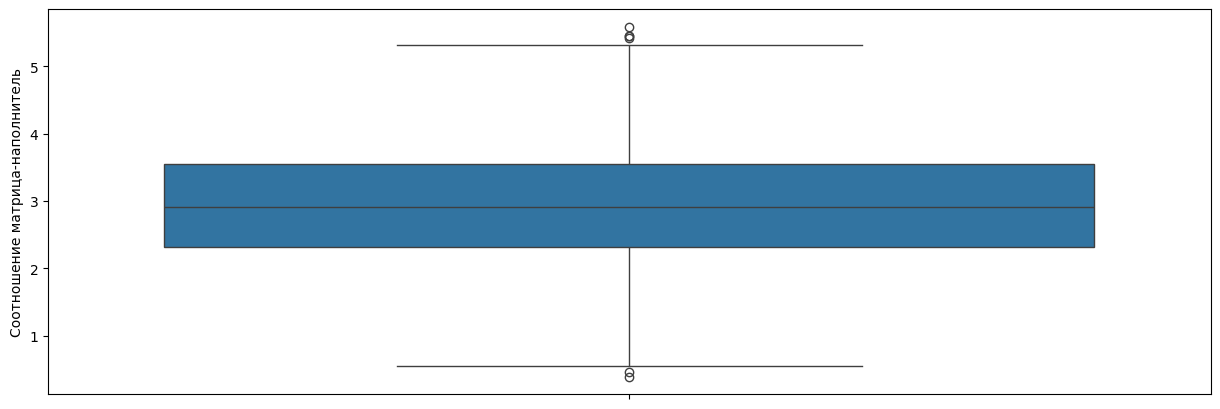


Рисунок 6-«Соотношение матрица+наполнитель

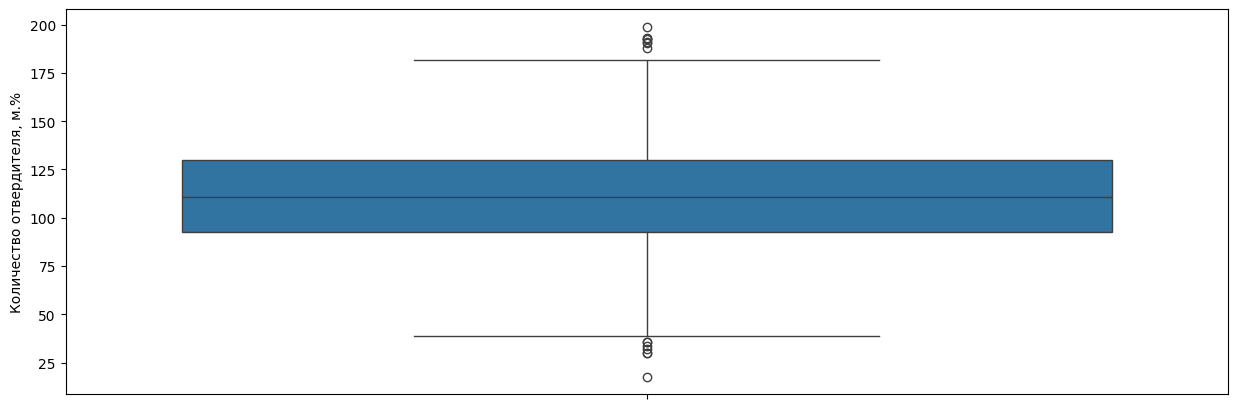


Рисунок 7 -«Количесто отвердителя»

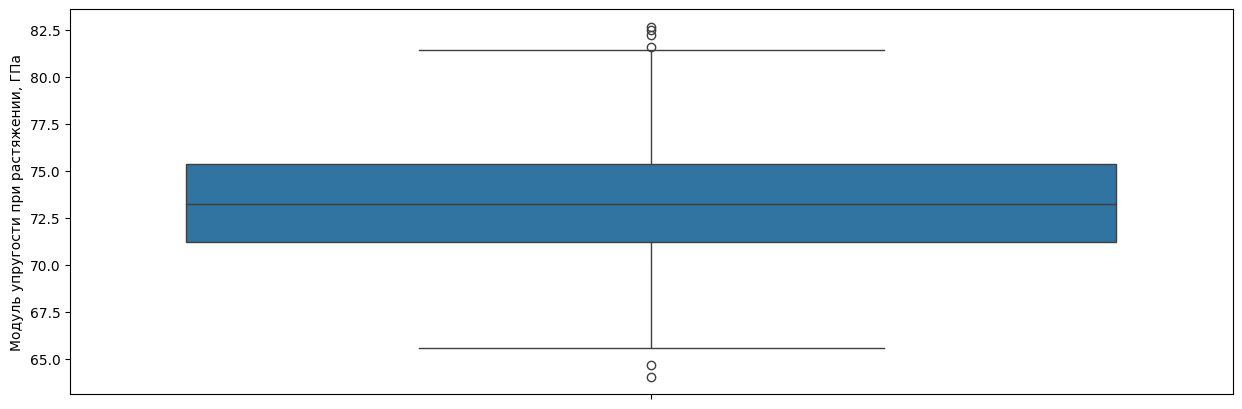


Рисунок 8-«Модуль упругости при растяжении»

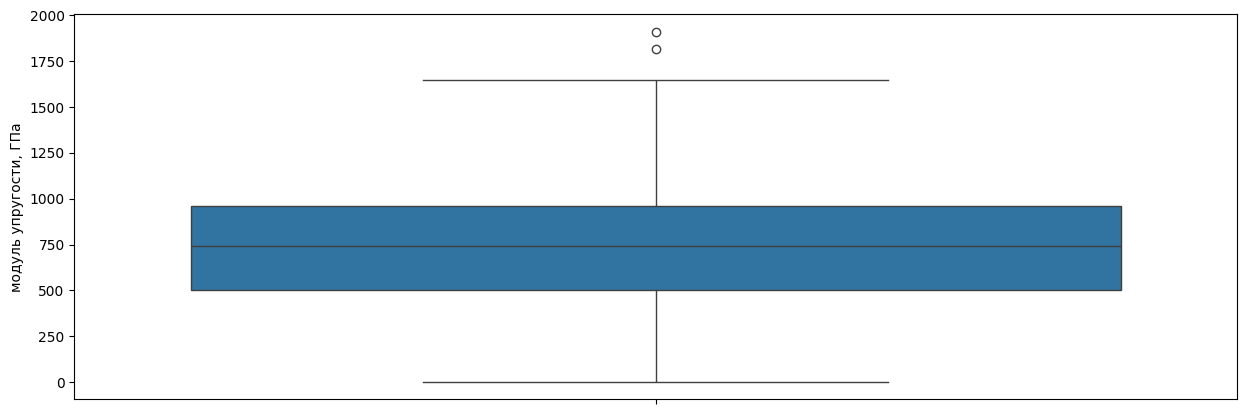


Рисунок 9 -«Модуль упругости»

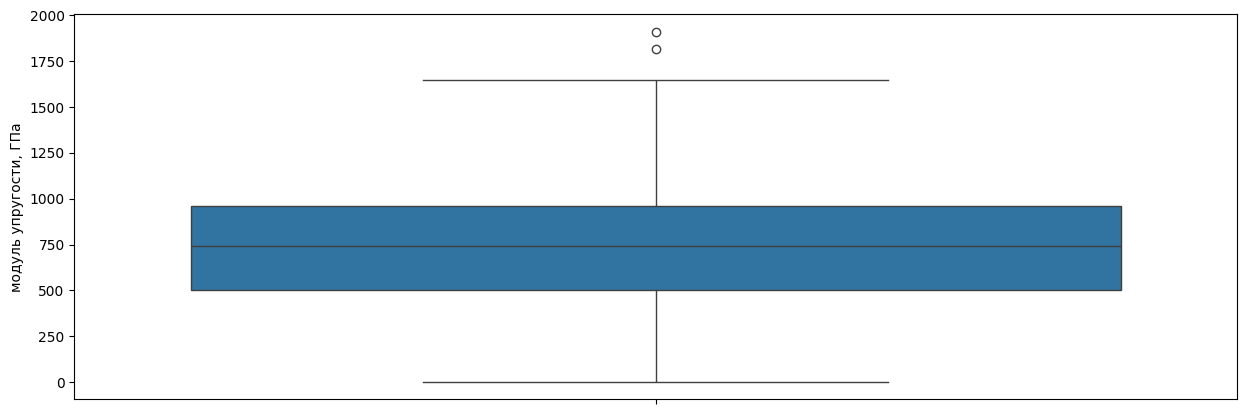


Рисунок 10 -«Плотность»

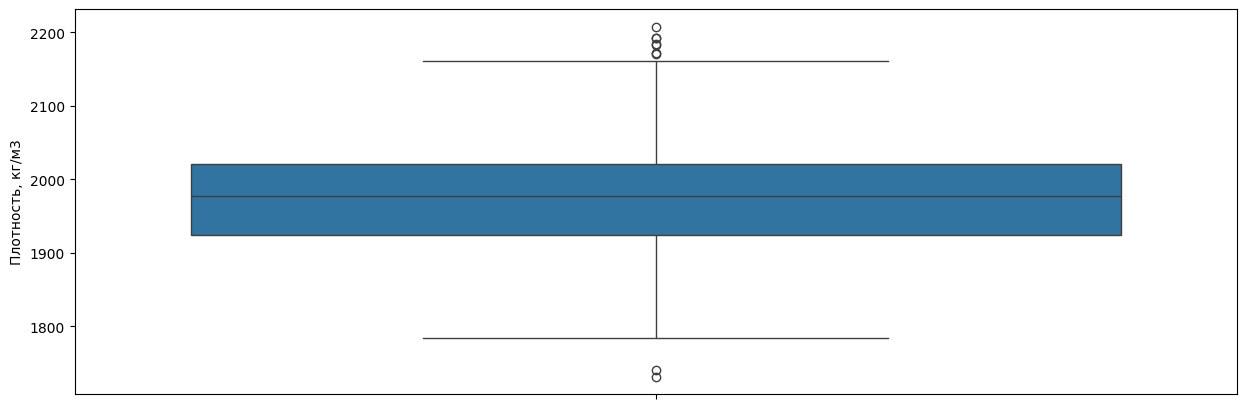


Рисунок 11- «Плотность нашивки»

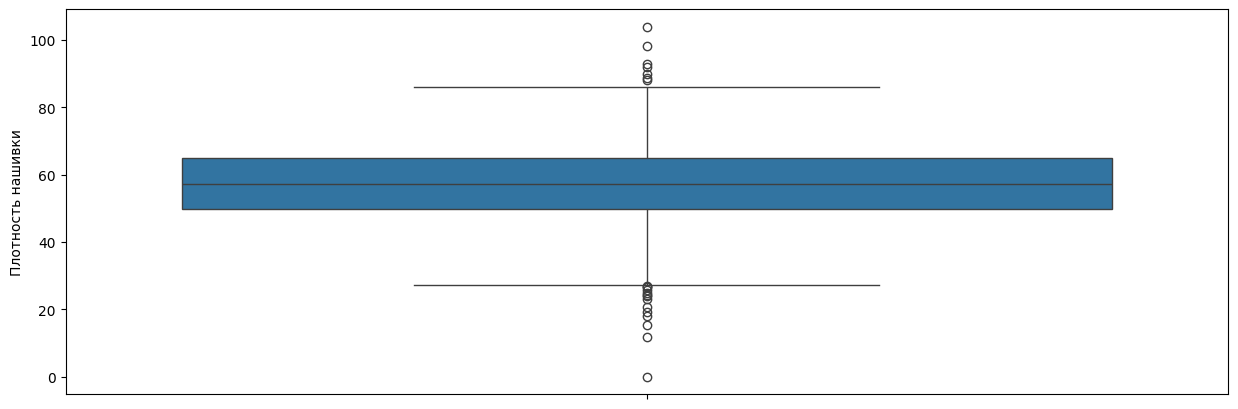


Рисунок 11 -«Поверхностная площадь»

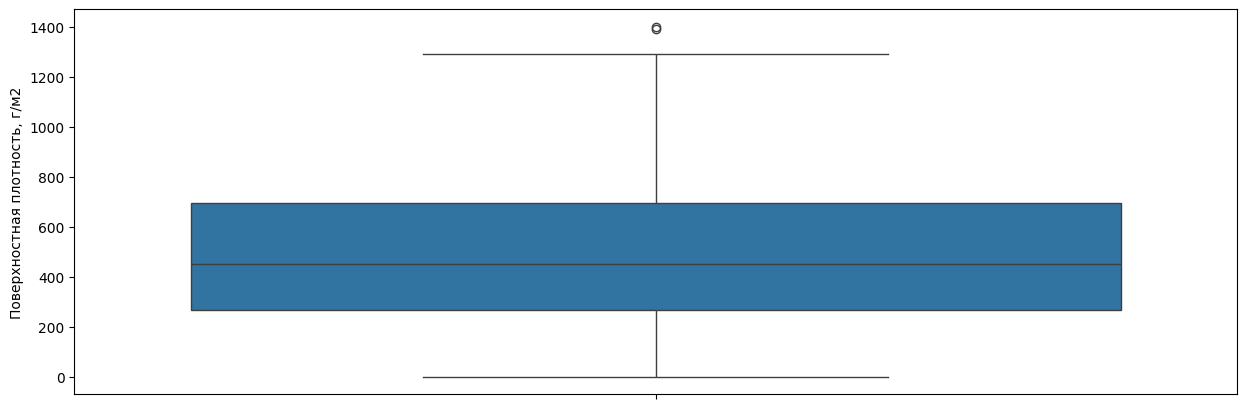


Рисунок 12- «Потребление смолы»

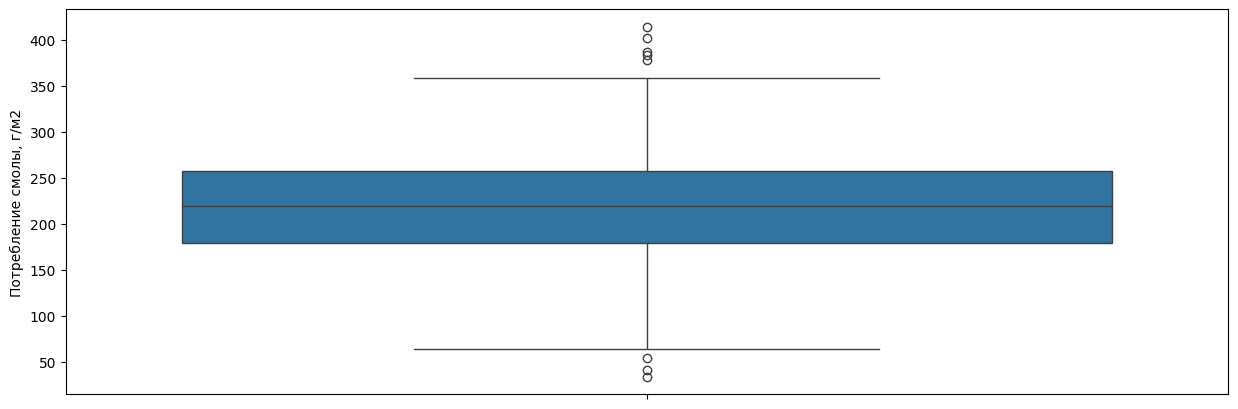


Рисунок 13 -«Прочность при растяжении»

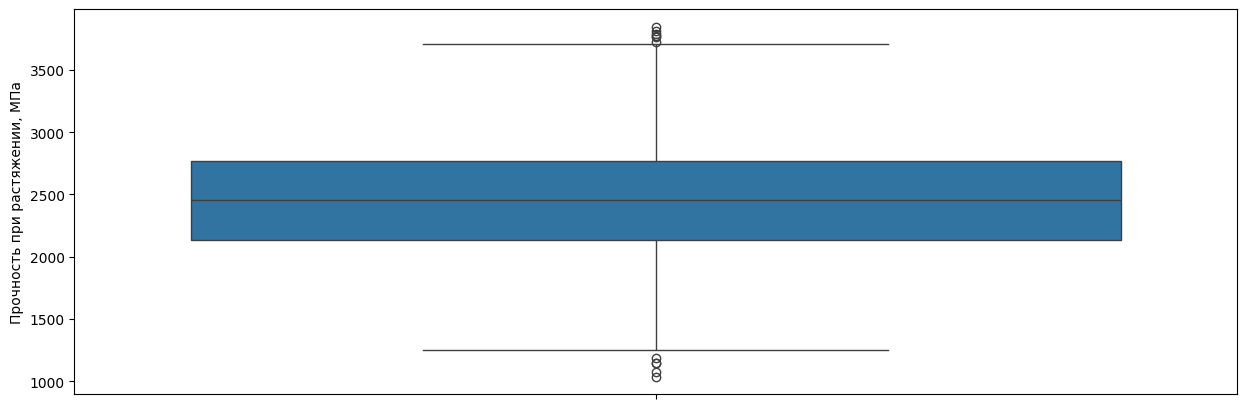


Рисунок 14 -«Содержание эпоксидных групп»

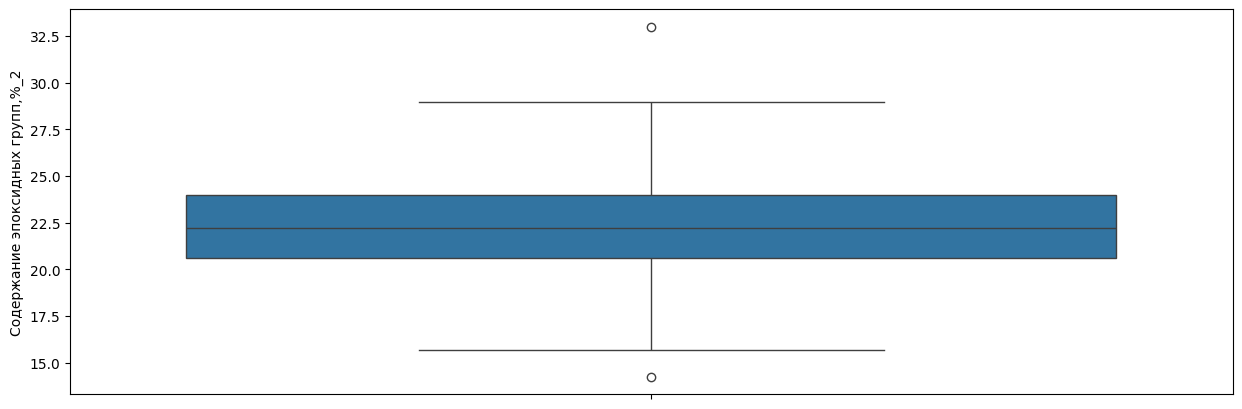


Рисунок 15- «Температура вспышки»

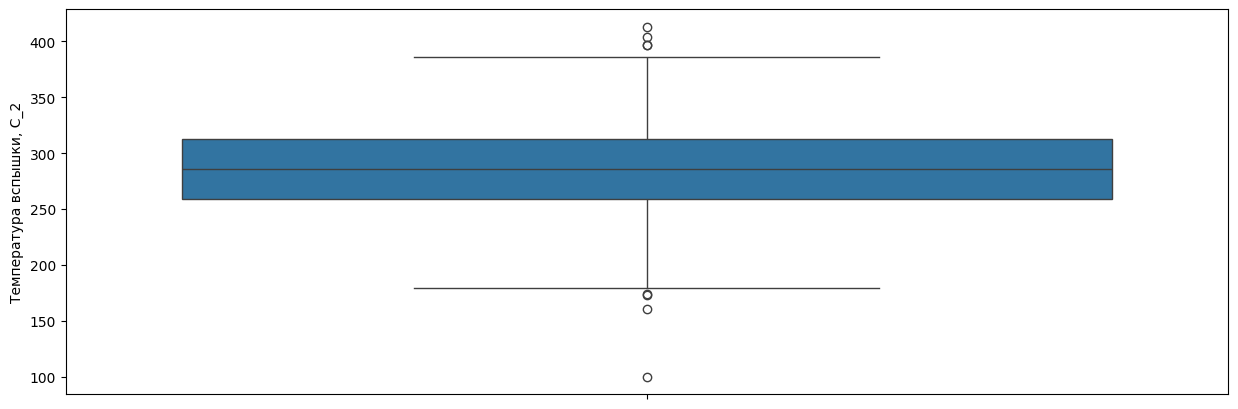
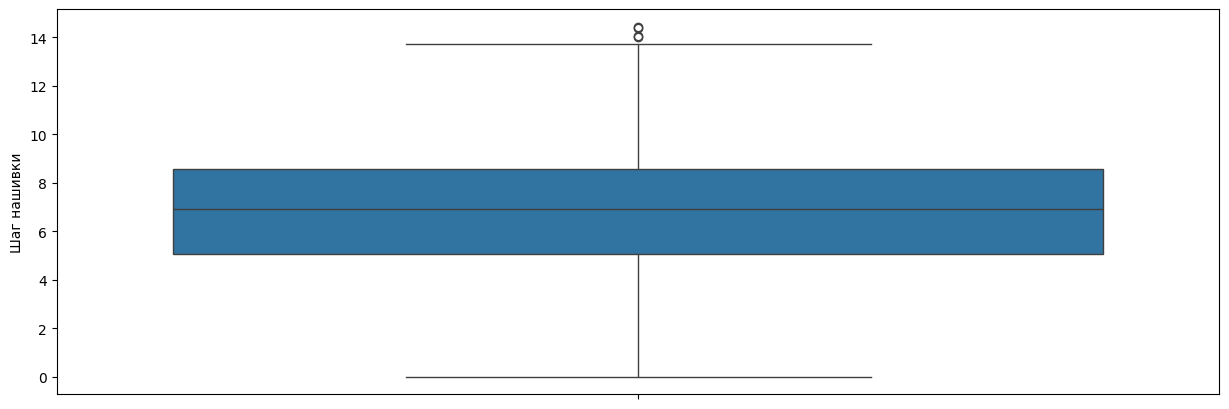


Рисунок 16 -«Шаг нашивки»



**2 Практическая часть**

**2.1 Предобработка данных.**

**Выбросы** **в** данных — это значения, которые сильно отличаются от большинства других значений в наборе данных. Они могут возникнуть из-за ошибок ввода, манипуляций с данными или аномальных процессов. **Выбросы** могут вносить существенные искажения в результаты анализа, поэтому важно определить и обработать их перед анализом данных. Производительность любой модели машинного обучения зависит от данных, на которых она обучается, и на нее можно легко повлиять, изменив распределение или добавив некоторые выбросы во входные данные. Выбросы могут привести к снижению точности моделей машинного обучения и увеличению времени обучения. Для нас становится важным обработать все выбросы, прежде чем предоставлять данные для обучения. Было решено удалить выбросы следующим методом.

IQR или **межквартильный диапазон** - это измерение изменчивости, основанное на разделении набора данных на разные квантили.

Квантили делятся на Q1, Q2 и Q3, где Q1 - среднее значение первой половины набора данных. Q2 - это медианное значение, а Q3 - это среднее значение второй половины набора данных.

IQR равен Q3 минус Q1.

Q1 = df.column.quantile (0,25)  
Q3 = df.column.quantile (0,75)

IQR = Q3 – Q1

После вычисления IQR мы вычисляем нижний и верхний предел, а затем просто отбрасываем все значения, которые меньше или выше предела.

В итоге был удален 91 выброс, для нашего количества данных это не критично. Но могу оговориться, что я создавала модели для данных,которые не выходили сильно за пределы максимума и минимума показателей. Для того, чтобы посчитать с данными которые выходят за границы средних показателей нужно иметь больше таких данных. Я удалила, чтобы повысить точность предсказаний моих моделей. Посмотрим на наши данные после удаления выбросов. Рисунок 17.

Рисунок 17- Коробчатая диаграмма набора данных после удаления выбросов

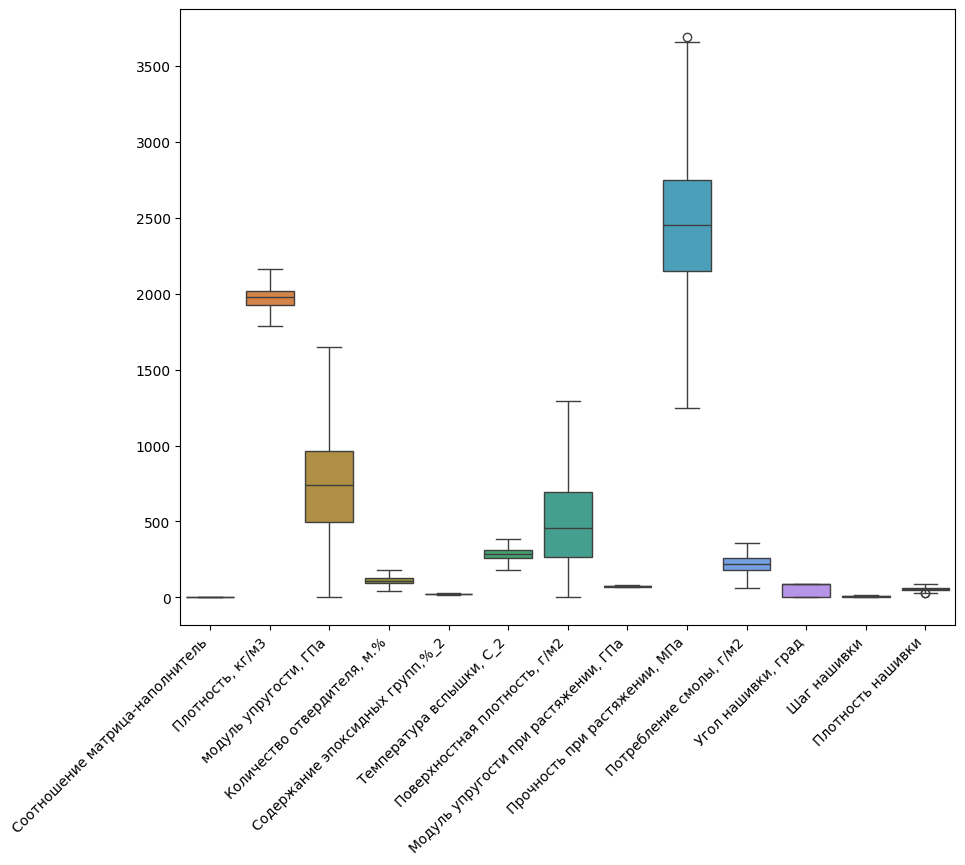
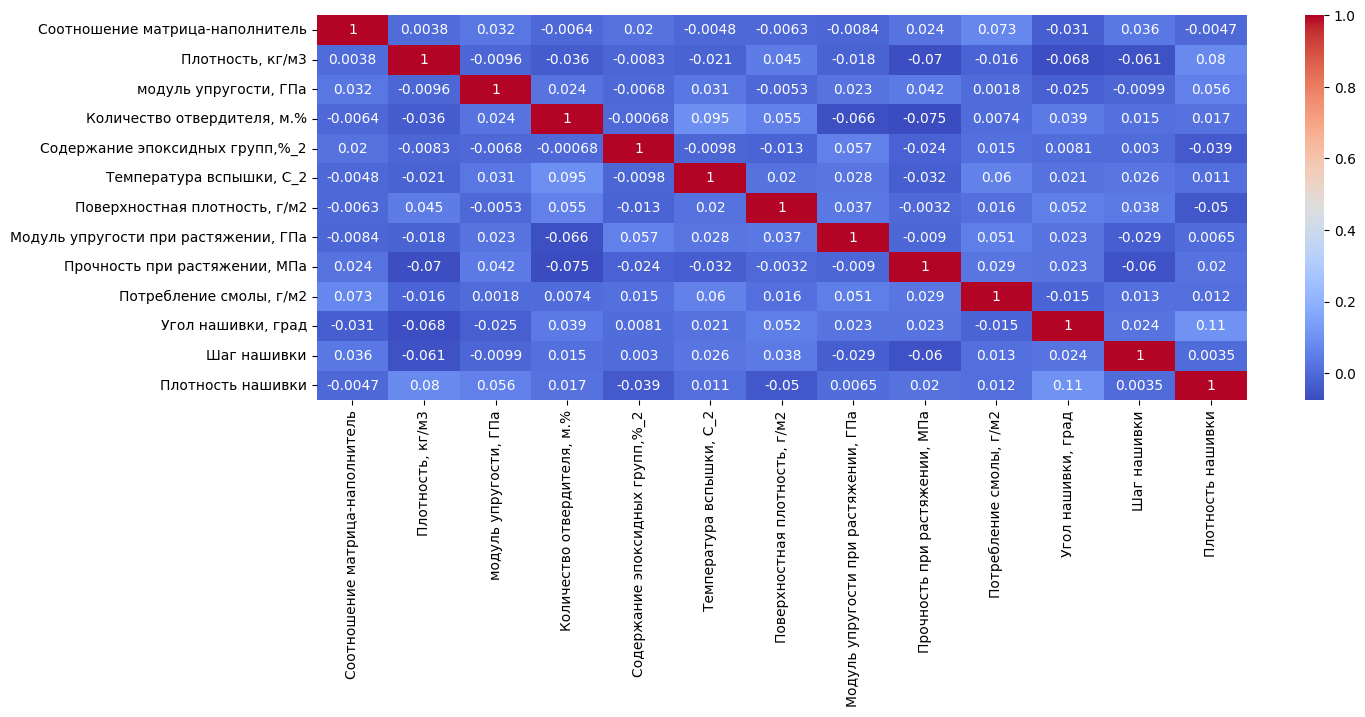


Рисунок 18 -Теплова карта корреляций после удаления выбросов



Посмотрим как изменились распределения данных наших целевых переменных после удаления выбросов. Таблица 7 и таблица 8. На рисунке 19 и 20 посмотрим гистограммы.

Таблица 7- Распределение данных «Модуль упругости при растяжении»

|  |  |
| --- | --- |
| count | 932.000000 |
| mean | 73.301871 |
| std | 3.042231 |
| min | 65.553336 |
| 25% | 71.233343 |
| 50% | 73.259230 |
| 75% | 75.356612 |
| max | 81.417126 |
|  |  |

Рисунок 19 -«Распределение данных «Модуль упругости при растяжении»

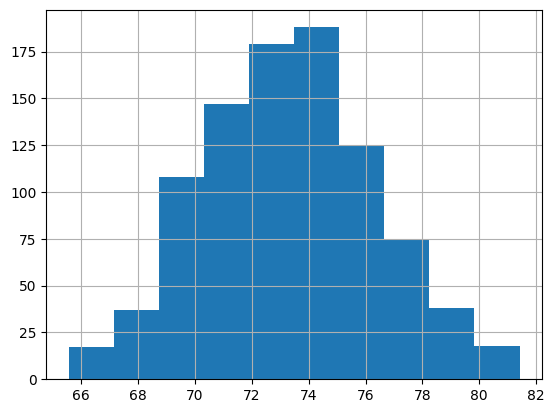


Рисунок 20 -«Распределение данных «Прочность при растяжении»

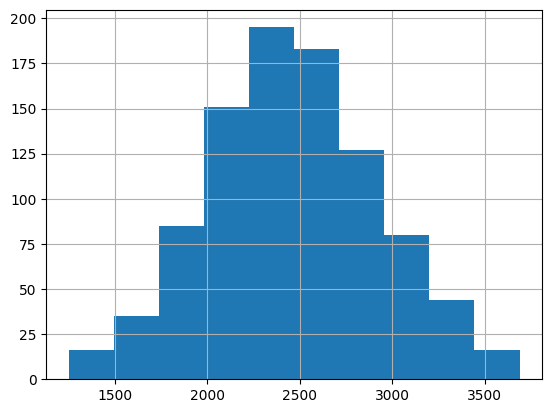
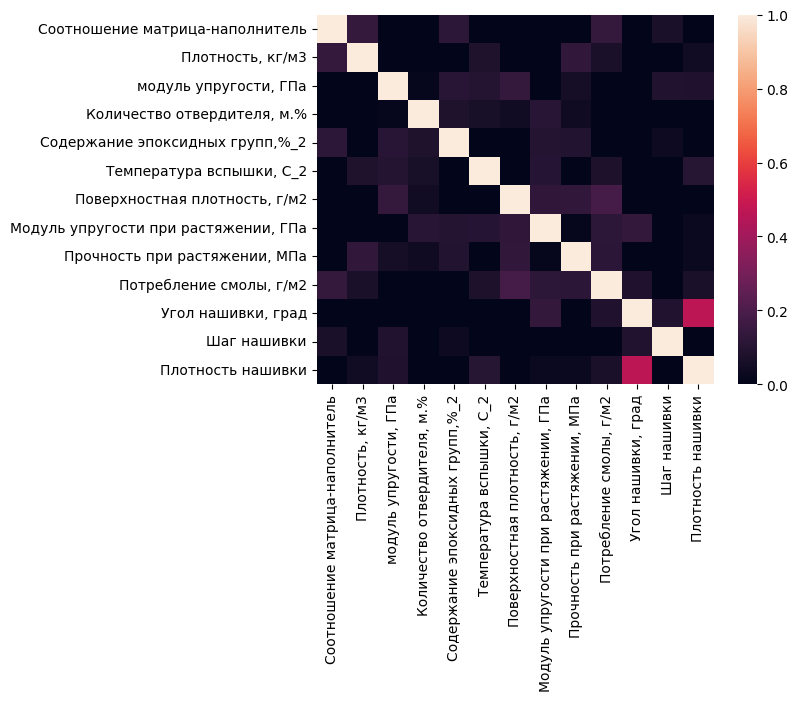


Таблица 8- Распределение данных «Прочность при растяжении»

|  |  |
| --- | --- |
| count | 932.000000 |
| mean | 2464.096062 |
| std | 459.217110 |
| min | 1250.392802 |
| 25% | 2146.936034 |
| 50% | 2456.394188 |
| 75% | 2752.348908 |
| max | 3689.223681 |
|  |  |

Рисунок 21- “Корреляция после удаления выбросов»



Можно заметить, что на тепловой карте пропали зависимости, которые были до удаления выбросов (не большие).

**2.2 Разработка, обучение и тестирование моделей.**

 В машинном обучении наша цель - улучшить модели и показатели точности. Таким образом, для улучшения результатов и хорошей модели прогнозирования используется стандартизация.  Цель стандартизации - изменить значения числовых столбцов в наборе данных для использования общей шкалы без искажения различий в диапазонах значений или потери информациии .Я буду строить и анализировать модели для каждой целевой переменной отдельно. Начну с показателя « Модуль упругости при растяжении».

Проанализируем метод опорных векторов. Сперва построим модель без подбора гиперпараметров. Потом попробуем подобрать гиперпараметры и сделать перекрестную проверку по 10 блокам. Результаты сравним в таблицах.(9,10, 11)

Таблица 9- Оценки без подбора гиперпараметров

|  |  |
| --- | --- |
| MAE | 2.516143 |
| MSE | 9.244201 |
| RMSE | 3.040427 |

Таблица 10- Оценка с подборкой гиперпараметров 10 блоков

|  |  |
| --- | --- |
| MAE | 2.396089 |
| MSE | 8.537738 |
| RMSE | 2.921940 |

Таблица 11- Оценка с подборкой гиперпараметров 5 блоков

|  |  |
| --- | --- |
| MAE | 2.393574 |
| MSE | 8.555448 |
| RMSE | 2.924976 |

Мы знаем распределение наших данных, если на них посмотреть то мы увидим что модель ошибается около 20% я считаю, что это плохой результат.

Дерево решений. Посчитаем на этой модели приблизительную ошибку. Результаты также сравним в таблицах(12,13,14)

Таблица 12- Дерево решений без подборки гиперпараметров

|  |  |
| --- | --- |
| MAE | 3.656764 |
| MSE | 20.898317 |
| RMSE | 4.571467 |

Таблица 13 -Дерево решений с подборкой гиперпараметров и поиском по сетке 5 блоков

|  |  |
| --- | --- |
| MAE | 2.89655 |
| MSE | 12.79844 |
| RMSE | 3.577490 |

Таблица -14 Дерево решений с подборкой гиперпараметров и поиском по сетке 10 блоков

|  |  |
| --- | --- |
| MAE | 3.168672 |
| MSE | 15.042101 |
| RMSE | 3.878414 |

Можно заметить, что благодаря подборке гиперпараметров результаты стали чуть лучше. Но эта модель отработала хуже предыдущей ее ошибка уже больше чем на 22%. Что тоже не удовлетворительный результат.

Посмотрим как отработает модель Лес деревьев с подбором гиперпараметров и без в таблицах(15,16,17).

Таблица 15 -Лес деревьев

|  |  |
| --- | --- |
| MAE | 2.458395 |
| MSE | 8.86086 |
| RMSE | 2.97672 |

Таблица 16- лес деревьев 5 блоков

|  |  |
| --- | --- |
| MAE | 2.413937 |
| MSE | 8.608371 |
| RMSE | 2.93400 |

Таблица 17 -Лес деревьев с подборкой гиперпараметров 10 блоков

|  |  |
| --- | --- |
| MAE | 2.440313 |
| MSE | 8.685633 |
| RMSE | 2.947139 |

Видим так же что ошибка большая. Но при подборе гиперпараметров были небольшие улучшения работы модели.

Давайте посмотрим на метод ближайших соседей и оценим результат работы. Данные приведем в таблице 18 и 19.

Таблица-18 Метод ближайших соседей

|  |  |
| --- | --- |
| MAE | 3.361 |
| MSE | 18.005 |
| RMSE | 4.243 |

Таблица 19- Метод ближайших соседей с нипепараметрами

|  |  |
| --- | --- |
| MAE | 2.425 |
| MSE | 8.661 |
| RMSE | 2.943 |

Конечно заметно, что с подборкой гиперпараметров модель работает значительно лучше, но такой результат более 20% нас тоже не устраивает.

Так же было решено проверить модель воспользовавшись продуктом CatBoost и посмотреть какую ошибка покажет она. Посмотрим результаты в таблице 20 и 21.

Таблица 20 – CatBoost на валидационной выборке

|  |  |
| --- | --- |
| MAE | 2.38 |
| MSE | 8.99 |
| RMSE | 2.99 |

Таблица 21 -CatBoost на тестовой выборке

|  |  |
| --- | --- |
| MAE | 2.38 |
| MSE | 9,00 |
| RMSE | 3.00 |

В результате видим, что модель показала плохие результаты оценок. На тестовых данных результат стал чуть меньше, так как модель эти данные не видела.

Перейдем к параметру «Прочность при растяжении» и постоим модели машинного обучения для нее. Начнем с метода опорных векторов и посмотрим результаты в таблицах 22 и 23.

Таблица 22- Метод опорных векторов

|  |  |
| --- | --- |
| MAE | 372.37 |
| MSE | 212790.75 |
| RMSE | 461.29 |

Таблица 33- Метод опорных векторов с гиперпараметрами

|  |  |
| --- | --- |
| MAE | 372.022 |
| MSE | 212663.00 |
| RMSE | 461.15 |

Видим что результат ошибки более 20% , но сама модель при подборке гиперпараметров почти не улучшила результат.

Перейдем к алгоритму дерево решений. Результаты в таблицах 34,35,36.

Таблица 34 – Дерево решений.

|  |  |
| --- | --- |
| MAE | 533.05 |
| MSE | 429120.40 |
| RMSE | 655.07 |

Таблица 35-Дерево решений с гиперпараметрами количество блоков 5

|  |  |
| --- | --- |
| MAE | 441.15 |
| MSE | 300.793 |
| RMSE | 548.446 |

Таблица 36 -Дерево решений 10 блоков

|  |  |
| --- | --- |
| MAE | 430.16 |
| MSE | 292377.85 |
| RMSE | 540.720 |

Модель дерево решений показала лучше всего себя с применением гиперпараметров, но результат ошибки все равно более 20 %.

Проверим алгоритм машинного обучения лес решений и посмотрим результаты в таблицах 37,38 и 39

Таблица 37- Лес решений

|  |  |
| --- | --- |
| MAE | 370.806 |
| MSE | 206156.25 |
| RMSE | 454.044 |

Таблица 38- Лес решений с типерпараметрами

|  |  |
| --- | --- |
| MAE | 370.003 |
| MSE | 206245.40 |
| RMSE | 454.142 |

Таблица 39- Лес решений с проверкой по 10 блокам

|  |  |
| --- | --- |
| MAE | 371.55 |
| MSE | 207876.97 |
| RMSE | 455.93 |

Лес решений показал результаты лучше всех моделей около 18% ошибки, но для нашего исследования это много.

Давайте посмотрим на последний метод -Метод ближайших соседей и сравним результаты его работы.(рисунок 40 и 41)

Рисунок 40- Метод ближайших соседей

|  |  |
| --- | --- |
| MAE | 492.64 |
| MSE | 391622.00 |
| RMSE | 625.79 |

Рисунок 41- Метод ближайших соседей с подбором гиперпараметров.

|  |  |
| --- | --- |
| MAE | 383.19 |
| MSE | 222984.59 |
| RMSE | 472.212 |

Как и с предыдущими моделями, этот алгоритм тоже показал не удовлетворительные результаты.

В рамках этой работы, я решила провести экспиремент и посмтореть, как будут работать алгоритмы машинного обучения на данных, которые мы не будем обрабатывать и удалять выбросы. Посмотрим на два показатели и для каждого сделаем по две модели. Начнем, как и во всей работе с показателя « Модуль упругости при растяжении». Первый метод будет лес деревьев. Рисунок 42.

Рисунок 42 -Лес деревьев с гиперпараметрами.

|  |  |
| --- | --- |
| MAE | 2.59 |
| MSE | 10.44 |
| RMSE | 3.23 |

На рисунке 43 результат работы метода ближайших соседей

|  |  |
| --- | --- |
| MAE | 2.56 |
| MSE | 10.30 |
| RMSE | 3.21 |

Наглядно видно, что результаты ухудшились.

Давайте посмотрим как покажет себя показатель «Прочность при растяжении» и первая модель будет дерево решений. Рисунок 44

Рисунок 44- Дерево решений с гиперпараметрами

|  |  |
| --- | --- |
| MAE | 440.05 |
| MSE | 299578.78 |
| RMSE | 547.33 |

Видим, что ошибка более 20%. Так посмотри на результат работы модели ближайших соседей и результат ее работы в таблице 45.

Таблица-45 Метод ближайших соседей с гиперпараметрами.

|  |  |
| --- | --- |
| MAE | 378.542 |
| MSE | 223821.50 |
| RMSE | 473.09 |

Результат конечно лучше, чем при применение модели дерево решений, но все равно не удачный. Мы посмотрели как работают модели на данных в которых не удалялись выбросы. Можно сделать определенный вывод, что модели машинного обучения в данном случае показали лучшие результаты на обработанных данных.

**2.4 Построение нейронной сети.**

Нам нужно написать нейронную сеть, которая будет рекомендовать: Соотношение матрица + наполнитель.

Я загрузила данные. Разделила выборку на тестовую и обучающую. Было бы проблематично передать в нейронную сеть значения, имеющие разные диапазоны. Поэтому я сделала нормализацию. Так как данных у нас не много я решила взять маленькую сеть с двумя 64 мерными слоями. Модель закончу без функции активации. Я использую функцию потерь mse. Также для взяла метрику mae это абсолютное значение разности между предсказанными и целевыми значениями. Чтобы оценить качество модели (таких как количество эпох) буду применять перекрёстную проверку по К блокам.Суть в разделении данных на К блоков.По полученным оценкам вычисляется среднее значение , которое принимается как оценка модели.

Для начала я взяла 4 блока и проход 50 эпох. 2 скрытых слоя и выходной слой без функции активации. Оптимизатор использовала «rmsprop» посмотрим на результат работы сети.

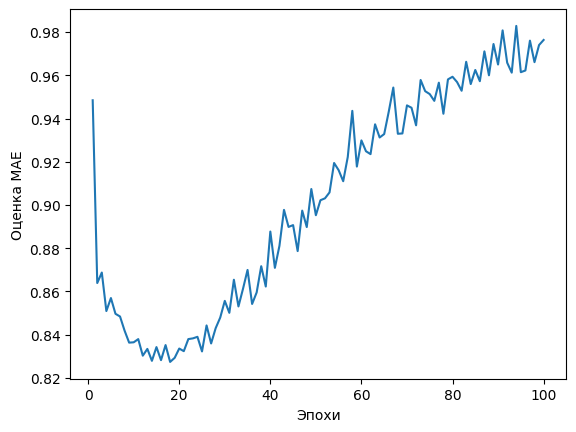
50 эпох – np.mean(all\_score) = 0.914

100эпох - np.mean(all\_score)= 0.914

Ошибка на всех данных test\_mae\_score = 0.9104.

Посмотрим на график с оценками проверок. Рисунок 22.

Рисунок 22 -Оценки MAE по эпохам.



Было построено несколько моделей. Настраивалось количество слоев, менялись функции в выходных слоях. Изменяла «размер пакета» от 64 до 128. Выбирала разное количесто эпох. Разделяла данные на разное количесто блоков для перекрестной проверки. Хорошего результата достигнуто не было. Самая удачная модель оказалась 64-разрядная, с двумя скрытыми слоями, выходной слой без функции активации. Оптимизатор Adam. Количество эпох 10.

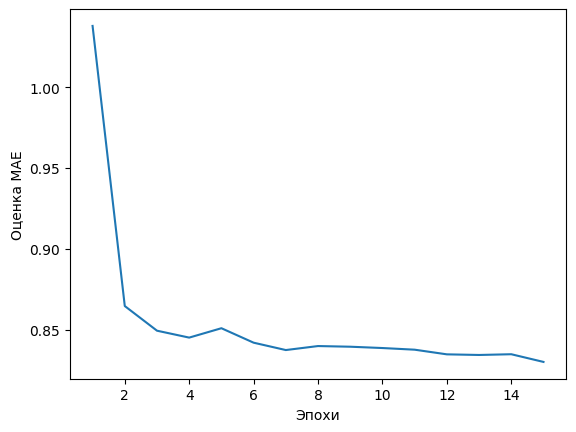
Ошибка np.mean(all\_score) = 0.87

Test\_mae\_score = 0.864

График с оценками проверок рисунок 23

Мною сделан вывод, что при небольшом объеме данных лучше использовать небольшое количество промежуточных слоёв.

Рисунок 23 -Оценки MAE по эпохам.



Благодаря построению нейронной сети для того чтобы предсказать соотношений «Матрица+наполнитель» мне не удалось получить результат для этих данных.