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»

Слушатель Колесников Анатолий Федорович

Москва, 2023

**Содержание**

[Введение 3](#_gjdgxs)

[**1. Аналитическая часть 5**](#_30j0zll)

[1.1. Постановка задачи 5](#_y77thine5guv)

[1.2. Описание используемых методов 7](#_1fob9te)

[1.2.1 Линейная регрессия 7](#_3znysh7)

[1.2.2 Регрессия k-ближайших соседей 9](#_2et92p0)

[1.2.3 Случайный лес 9](#_tyjcwt)

[1.2.4 Многослойный перцептрон 11](#_3dy6vkm)

[1.2.5 Метрики качества моделей 12](#_1t3h5sf)

[1.3. Разведочный анализ данных 13](#_nnsnrzohmgly)

**2. Практическая часть 16**

[2.1 Предобработка данных 16](#_m9xdibtlyf9f)

[2.2 Разработка и обучение модели 21](#_nc55g514bz9n)

[2.3 Тестирование моделей 21](#_2s8eyo1)

[2.3.1 Линейная регрессия 22](#_jp6dkkn49m3e)

[2.3.2 Метод k-ближайших соседей 23](#_3k2vvwc7h91y)

[2.3.3 Случайный лес 24](#_2jeqxis814yc)

[2.4. Нейронная сеть, которая будет рекомендовать соотношение матрица-наполнитель 25](#_3rdcrjn)

**Заключение 28**

[**Список литературы 29**](#_lnxbz9)

# **Введение**

Благодаря машинному обучению программист не обязан писать инструкции, учитывающие все возможные проблемы и содержащие все решения. Вместо этого в компьютер (или отдельную программу) закладывают алгоритм самостоятельного нахождения решений путём комплексного использования статистических данных, из которых выводятся закономерности и. на основе которых. делаются прогнозы.  
Технология машинного обучения на основе анализа данных берёт начало в 1950 году, когда начали разрабатывать первые программы для игры в шашки. За прошедшие десятилетия общий принцип не изменился. Зато благодаря взрывному росту вычислительных мощностей компьютеров многократно усложнились закономерности и прогнозы, и расширился круг проблем и задач, решаемых с использованием машинного обучения.  
Чтобы запустить процесс машинного обучение, для начала необходимо загрузить в компьютер «датасет» (некоторое количество исходных данных), на которых алгоритм будет учиться обрабатывать запросы. Мы рассмотрим задачу предсказания свойств композиционных материалов на основе стандартного учебного датасета.

Композиционный материа́л или композитный материа́л (КМ), сокращённо компози́т — многокомпонентный материал, изготовленный (человеком или природой) из двух или более компонентов с существенно различными физическими и/или химическими свойствами, которые, в сочетании, приводят к появлению нового материала с характеристиками, отличными от характеристик отдельных компонентов и не являющимися простой их суперпозицией. В составе композита принято выделять матрицу/матрицы и наполнитель/наполнители. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура (наполнитель) внутри бетона (матрица) 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.

# **1. Аналитическая часть**

## **1.1. Постановка задачи**

Композиты - многокомпонентные материалы, состоящие из полимерной, металлической., углеродной, керамической или др. основы (матрицы), армированной наполнителями из волокон, нитевидных кристаллов, тонкодиспeрсных частиц и др. Путем подбора состава и свойств наполнителя и матрицы (связующего), их соотношения, ориентации наполнителя можно получить материалы с требуемым сочетанием эксплуатационных и технологических свойств. Использование в одном материале нескольких матриц (полиматричные композиционные материалы) или наполнителей различной природы (гибридные композиционные материалы) значительно расширяет возможности регулирования свойств композиционных материалов.

Эффективность и работоспособность материала зависят от правильного выбора исходных компонентов и технологии их совмещения, призванной обеспечить прочную связь между компонентами при сохранении их первоначальных характеристик. В результате совмещения армирующих элементов и матрицы образуется комплекс свойств композита, не только отражающий исходные характеристики его компонентов, но и включающий свойства, которыми изолированные компоненты не обладают. В частности, наличие границ раздела между армирующими элементами и матрицей существенно повышает трещиностойкость материала, и в композитах, в отличие от металлов, повышение статической прочности приводит не к снижению, а, как правило, к повышению характеристик вязкости разрушения.

Современные композиты изготавливаются из различных материалов, но даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из известных компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов (затратно по времени, и потребляемым ресурсам) или теоретическое прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

Прогнозирование модели может существенно сократить количество проводимых испытаний, а также пополнить базу данных материалов новыми свойствами материалов, и цифровыми двойниками новых композитов. Актуальность решения задачи обусловлена широким использованием композитных материалов практически во всех областях производства.

Предметом выпускной квалификационной работы является построение моделей для прогнозирования таких характеристик композиционных материалов, как модуль упругости при растяжении, прочность при растяжении и создание нейронной сети для рекомендации соотношения матрица-наполнитель.

Датасет состоит из двух файлов:

1. файл X\_bp.xlsx с данными о параметрах базальтопластика;
2. файл X\_nup.xlsx с данными о нашивках из углепластика.

Для решения задачи проводится объединение двух файлов X\_bp.xlsx и X\_nup.xlsx по индексу, используя тип объединения INNER.

Количество строк в файле X\_bp.xlsx было 1023, столбцов 10. А количество строк в файле X\_nup.xlsx – 1040, столбцов 3.

После объединения таблиц, 17 строк из файла X\_nup.xlsx было отброшено. Дальнейшие исследования проводились с датасетом содержащим 1023 строк и 13 столбцов.

## **1.2. Описание используемых методов**

Предсказание значений вещественной, непрерывной переменной — это задача регрессии. Эта зависимая переменная должна иметь связь с одной или несколькими независимыми переменными, называемых также предикторами или регрессорами. Регрессионный анализ помогает понять, как типичное значение зависимой переменной изменяется при изменении независимых переменных.

### 

### 1.2.1 Линейная регрессия

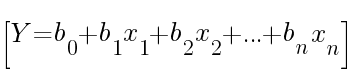
За базовую модель для прогнозирования всех искомых параметров принята линейная регрессия. LinearRegression соответствует линейной модели с коэффициентами w = (w1, …, wp), чтобы минимизировать остаточную сумму квадратов между наблюдаемыми целями в наборе данных и целями, предсказанными линейным приближением.

Простая линейная регрессия имеет место, если рассматривается зависимость между одной входной и одной выходной переменными. Для этого определяется уравнение регрессии (1.1) и строится соответствующая прямая, известная как линия регрессии.

 (1.1)

Коэффициенты a и b, называемые также параметрами модели, определяются таким образом, чтобы сумма квадратов отклонений точек, соответствующих реальным наблюдениям данных, от линии регрессии была бы минимальной. Коэффициенты обычно оцениваются методом наименьших квадратов.

Если ищется зависимость между несколькими входными и одной выходной переменными, то имеет место множественная линейная регрессия. Соответствующее уравнение имеет вид (2).

 (1.2)

где n - число входных переменных.

Очевидно, что в данном случае модель будет описываться не прямой, а гиперплоскостью. Коэффициенты уравнения множественной линейной регрессии подбираются так, чтобы минимизировать сумму квадратов отклонения реальных точек данных от этой гиперплоскости.

Линейная регрессия — первый тщательно изученный метод регрессионного анализа. Его главное достоинство — простота. Такую модель можно построить и рассчитать даже без мощных вычислительных средств. Простота является и главным недостатком этого метода. Тем не менее, именно с линейной регрессии целесообразно начать подбор подходящей модели.

На языке python линейная регрессия реализована в sklearn.linear\_model.LinearRegression.

### 1.2.2 Регрессия k-ближайших соседей

Метод ближайших соседей (kNN - kNearestNeighbours) - метод решения задач классификации и задач регрессии, основанный на поиске ближайших объектов с известными значения целевой переменной.

Для целевой переменной метод предполагает найти ближайшие к нему объекты x1, x2…xk и построить прогноз по их меткам, то есть определить границы классов и выстроить гиперплоскость регрессии. Метка, назначенная целевой переменной, вычисляется на основе среднего значения меток ее ближайших соседей.

Алгоритм находит расстояния между запросом и всеми примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, затем голосует за наиболее часто встречающуюся метку (в случае задачи классификации) или усредняет метки (в случае задачи регрессии).

Преимущества данного метода – простая реализация, низкая чувствительность к выбросам, отсутствие необходимости строить модель, настраивать несколько параметров или делать дополнительные допущения, универсальность.

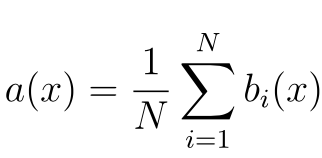
Метод k-ближайших соседей (k-nearest neighbors) – это простой алгоритм машинного обучения с учителем, который можно использовать для решения задач классификации и регрессии. Он прост в реализации и понимании, но имеет существенный недостаток – значительное замедление работы, когда объем данных растет.

### 1.2.3 Случайный лес

Случайный лес (RandomForest) — это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме:

* Выбирается подвыборка обучающей выборки размера samplesize (м.б. с возвращением) – по ней строится дерево (для каждого дерева — своя подвыборка).
* Для построения каждого расщепления в дереве просматриваем max\_features случайных признаков (для каждого нового расщепления — свои случайные признаки).
* Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив. Формула итогового решателя (1.3) — это усреднение предсказаний отдельных деревьев.

 (1.3)

где

N – количество деревьев;

i – счетчик для деревьев;

b – решающее дерево;

x – сгенерированная нами на основе данных выборка.

Для определения входных данных каждому дереву используется метод случайных подпространств. Базовые алгоритмы обучаются на различных подмножествах признаков, которые выделяются случайным образом.

Преимущества случайного леса:

* высокая точность предсказания;
* редко переобучается;
* практически не чувствителен к выбросам в данных;
* одинаково хорошо обрабатывает как непрерывные, так и дискретные признаки, данные с большим числом признаков;

Из недостатков можно отметить, что его построение занимает больше времени. Так же теряется интерпретируемость.

Метод реализован в sklearn.ensemble.RandomForestRegressor.

### 1.2.4 Многослойный перцептрон

Нейронная сеть — это последовательность нейронов, соединенных между собой связями. Нейронные сети применяются для решения задач регрессии, классификации, распознавания образов и речи, компьютерного зрения и других. На настоящий момент это самый мощный, гибкий и широко применяемый инструмент в машинном обучении. Персептроны пытаются имитировать функциональность человеческого мозга для решения задач. Вычислительная единица нейронной сети — нейрон или персептрон.

У каждого нейрона есть определённое количество входов, куда поступают сигналы, которые суммируются с учётом значимости (веса) каждого входа. Смещение – это дополнительный вход для нейрона, который всегда равен 1 и, следовательно, имеет собственный вес соединения. Так же у нейрона есть функция активации, которая определяет выходное значение нейрона. Она используется для того, чтобы ввести нелинейность в нейронную сеть. Примеры активационных функций: relu, сигмоида.

У полносвязной нейросети выход каждого нейрона подается на вход всем нейронам следующего слоя. У нейросети имеется:

* входной слой — его размер соответствует входным параметрам;
* скрытые слои — их количество и размерность определяем специалист;
* выходной слой — его размер соответствует выходным параметрам.

Прямое распространение – это процесс передачи входных значений в нейронную сеть и получения выходных данных, которые называются прогнозируемым значением. Прогнозируемое значение сравниваем с фактическим с помощью функции потери. В методе обратного распространения ошибки градиенты (производные значений ошибок) вычисляются по значениям весов в направлении, обратном прямому распространению сигналов. Значение градиента вычитают из значения веса, чтобы уменьшить значение ошибки. Таким образом происходит процесс обучения. Обновляются веса каждого соединения, чтобы функция потерь минимизировалась. Для обновления весов в модели используются различные оптимизаторы. Количество эпох показывает, сколько раз выполнялся проход для всех примеров обучения.

### 1.2.5 Метрики качества моделей

Существует множество различных метрик качества, применимых для регрессии. В этой работе используются:

R2 или коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то качество прогноза идентично средней величине целевой переменной (т.е. очень низкое). Отрицательные значение коэффициента детерминации означают очень плохую объясняющую способность модели.

MSE (Mean Squared Error) или средняя квадратичная ошибка принимает значениях в тех же единицах, что и целевая переменная. Чем ближе к нулю MSE, тем лучше работают предсказательные качества модели.

## **1.3. Разведочный анализ данных**

Разведочный анализ (Exploratory Data Analysis) — это процесс анализа данных, способствующий выдвижению гипотез, когда отсутствует представление о связи между переменными или оно недостаточно.

Важно отметить, что РА— это не набор готовых методов анализа, а философия того, как следует анализировать данные, что искать и как интерпретировать.

В разведочном анализе часто используется статистическая графика — графическая техника, используемая для визуализации количественных данных.

Результаты разведочного анализа не используются для выработки управленческих решений. Их назначение — помощь в разработке наилучшей стратегии углубленного анализа, выдвижение гипотез, уточнение особенностей применения тех или иных математических методов и моделей. Без разведочного анализа углубленный анализ данных будет производиться практически «вслепую».

Это необходимо для того, чтобы максимизировать понимание набора данных:

* выявить базовую структуру;
* извлечь важные переменные;
* выявить выбросы и аномалии;
* проверить исходные предположения;
* разработать скупые модели;
* определять оптимальные параметры.

Общая информация о датасете

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Column** | **Count** | **Non-Null** | **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 |

Описательная статистика

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Соотношение матрица-наполнитель** | 1023 | 2.930366 | 0.913222 | 0.389403 | 2.317887 | 2.906878 | 3.552660 | 5.591742 |
| **Плотность, кг/м3** | 1023 | 1975.734888 | 73.729231 | 1731.764635 | 1924.155467 | 1977.621657 | 2021.374375 | 2207.773481 |
| **модуль упругости, ГПа** | 1023 | 739.923233 | 330.231581 | 2.436909 | 500.047452 | 739.664328 | 961.812526 | 1911.536477 |
| **Количество отвердителя, м.%** | 1023 | 110.570769 | 28.295911 | 17.740275 | 92.443497 | 110.564840 | 129.730366 | 198.953207 |
| **Содержание эпоксидных групп,%\_2** | 1023 | 22.244390 | 2.406301 | 14.254985 | 20.608034 | 22.230744 | 23.961934 | 33.000000 |
| **Температура вспышки, С\_2** | 1023 | 285.882151 | 40.943260 | 100.000000 | 259.066528 | 285.896812 | 313.002106 | 413.273418 |
| **Поверхн. плотность, г/м2** | 1023 | 482.731833 | 281.314690 | 0.603740 | 266.816645 | 451.864365 | 693.225017 | 1399.542362 |
| **Модуль упругости при растяжении, ГПа** | 1023 | 73.328571 | 3.118983 | 64.054061 | 71.245018 | 73.268805 | 75.356612 | 82.682051 |
| **Прочность при растяжении, МПа** | 1023 | 2466.922843 | 485.628006 | 1036.856605 | 2135.850448 | 2459.524526 | 2767.193119 | 3848.436732 |
| **Потребление смолы, г/м2** | 1023 | 218.423144 | 59.735931 | 33.803026 | 179.627520 | 219.198882 | 257.481724 | 414.590628 |
| **Угол нашивки, град** | 1023 | 44.252199 | 45.015793 | 0.000000 | 0.000000 | 0.000000 | 90.000000 | 90.000000 |
| **Шаг нашивки** | 1023 | 6.899222 | 2.563467 | 0.000000 | 5.080033 | 6.916144 | 8.586293 | 14.440522 |
| **Плотность нашивки** | 1023 | 57.153929 | 12.350969 | 0.000000 | 49.799212 | 57.341920 | 64.944961 | 103.988901 |

Мы видим отсутствие пропусков, большой разброс значений и, следовательно, необходимость нормализации, которая будет проведена в дальнейшем, категориальную характеристику 'Угол нашивки, град', закодируем при помощи LabelEncoder.

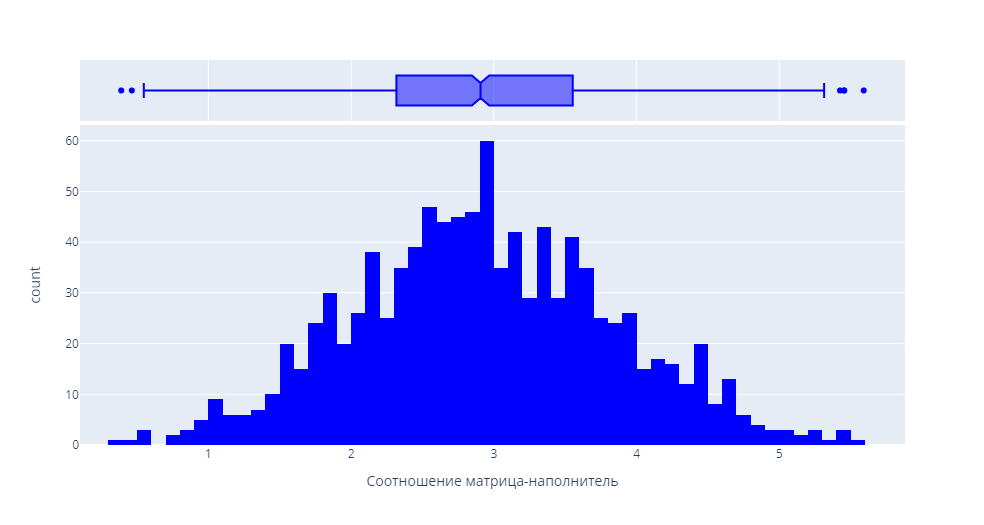
# **2. Практическая часть**

## **2.1 Предобработка данных**

Приступим к выполнению задачи по прогнозированию конечных свойств новых композиционных материалов.

**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.

Гистограммы распределения переменных, объединенные для наглядности с боксплотами (для примера в тексте приведены первые две переменные датасета, остальные находятся в коде) представлены на рис. 1, рис.2. По ним видно, что все признаки, кроме «Угол нашивки», имеют распределение близкое к нормальному и принимают неотрицательные значения. По боксплотам видно наличие выбросов. Далее при предобработке данных выбросы будут удалены.

Рис. 1. Гистограмма распределения переменных и боксплот (соотношение матрица-наполнитель)

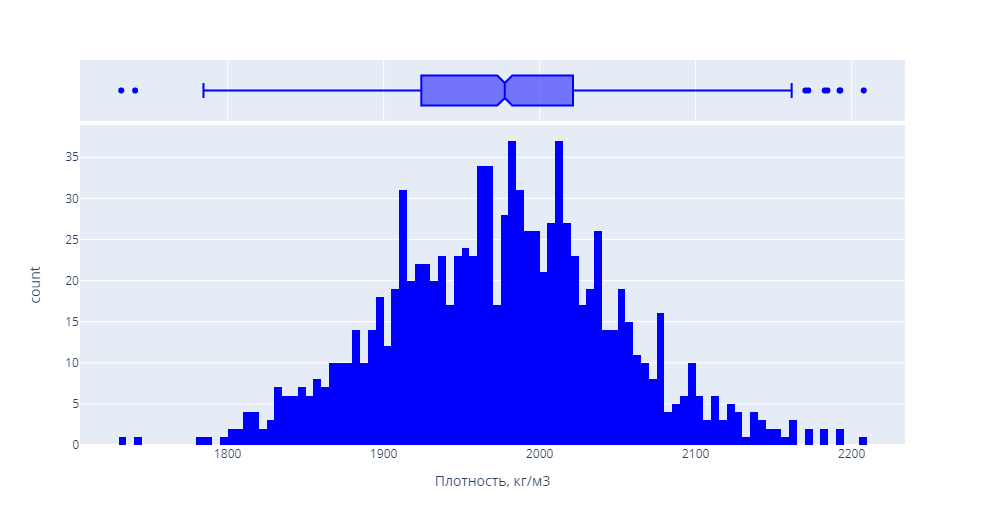


Рис. 2. Гистограмма распределения переменных и боксплот (плотность)

Попробуем выявить связь между характеристиками с помощью матрицы корреляции.

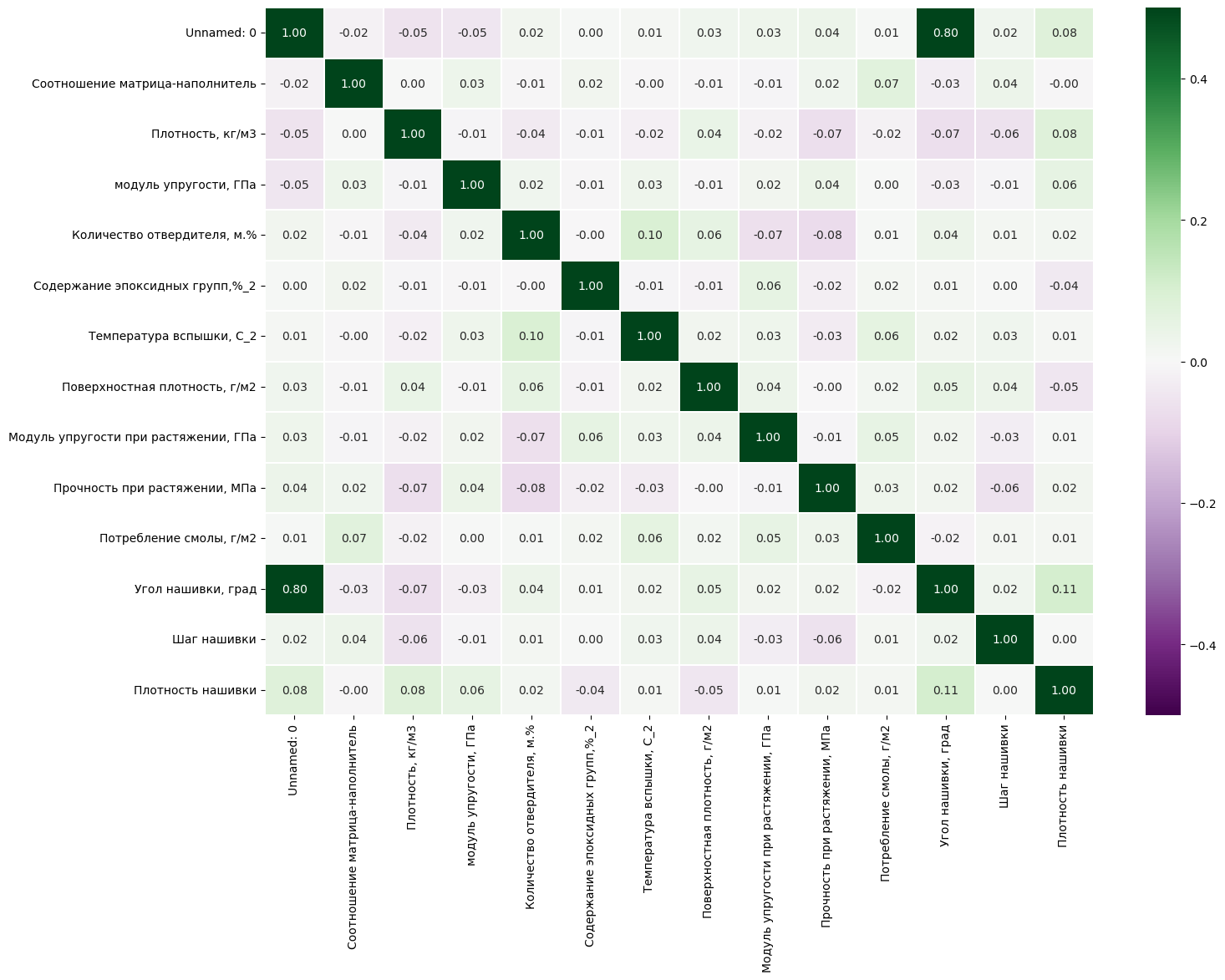


Рис 3. Корреляционная матрица

Визуализация коэффициентов корреляции выполнена с помощью тепловой карты sns.heatmap.

Видно, что все коэффициенты корреляции близки к нулю. Это означает отсутствие линейной зависимости между признаками.

Масштабировать будем с помощью приведения каждого признака к диапазону от 0 до 1 с помощью метода MinMaxScaler.

Описательная статистика после нормализации

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Соотношение матрица-наполнитель | count | mean | std | min | 25% | 50% | 75% | max |
| Плотность, кг/м3 | 936 | 0.498933 | 0.187489 | 0 | 0.372274 | 0.494538 | 0.629204 | 1 |
| модуль упругости, ГПа | 936 | 0.502695 | 0.187779 | 0 | 0.368517 | 0.511229 | 0.624999 | 1 |
| Количество отвердителя, м.% | 936 | 0.446764 | 0.199583 | 0 | 0.301243 | 0.447061 | 0.580446 | 1 |
| Содержание эпоксидных групп,%\_2 | 936 | 0.504664 | 0.188865 | 0 | 0.376190 | 0.506040 | 0.637978 | 1 |
| Температура вспышки, С\_2 | 936 | 0.491216 | 0.180620 | 0 | 0.367716 | 0.489382 | 0.623410 | 1 |
| Поверхностная плотность, г/м2 | 936 | 0.516059 | 0.190624 | 0 | 0.386128 | 0.515980 | 0.646450 | 1 |
| Модуль упругости при растяжении, ГПа | 936 | 0.373733 | 0.217078 | 0 | 0.205619 | 0.354161 | 0.538683 | 1 |
| Прочность при растяжении, МПа | 936 | 0.488647 | 0.191466 | 0 | 0.359024 | 0.485754 | 0.615077 | 1 |
| Потребление смолы, г/м2 | 936 | 0.495706 | 0.188915 | 0 | 0.365149 | 0.491825 | 0.612874 | 1 |
| Угол нашивки, град | 936 | 0.521141 | 0.195781 | 0 | 0.392067 | 0.523766 | 0.652447 | 1 |
| Шаг нашивки | 936 | 0.511752 | 0.500129 | 0 | 0 | 1 | 1 | 1 |
| Плотность нашивки | 936 | 0.502232 | 0.183258 | 0 | 0.372211 | 0.504258 | 0.624604 | 1 |

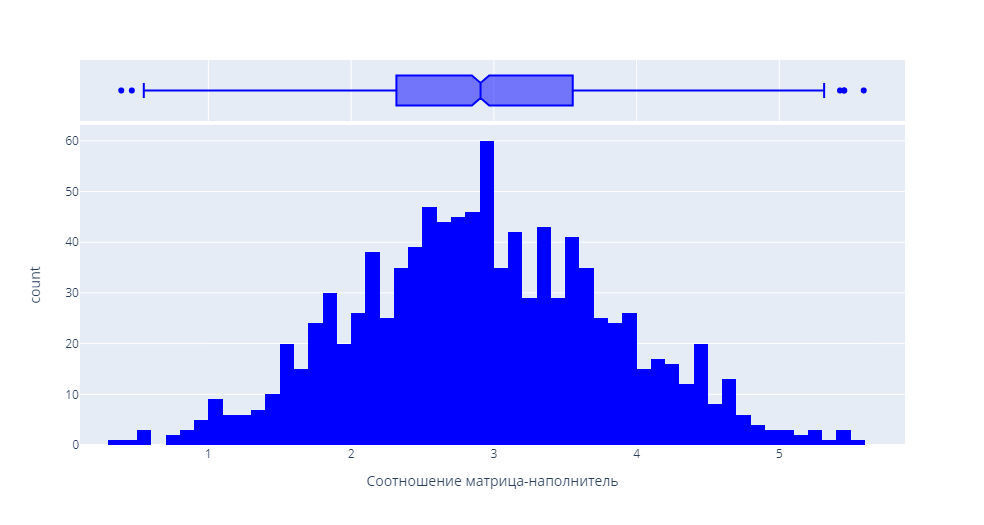
Удаляем выбросы методом межквартильных расстояний.

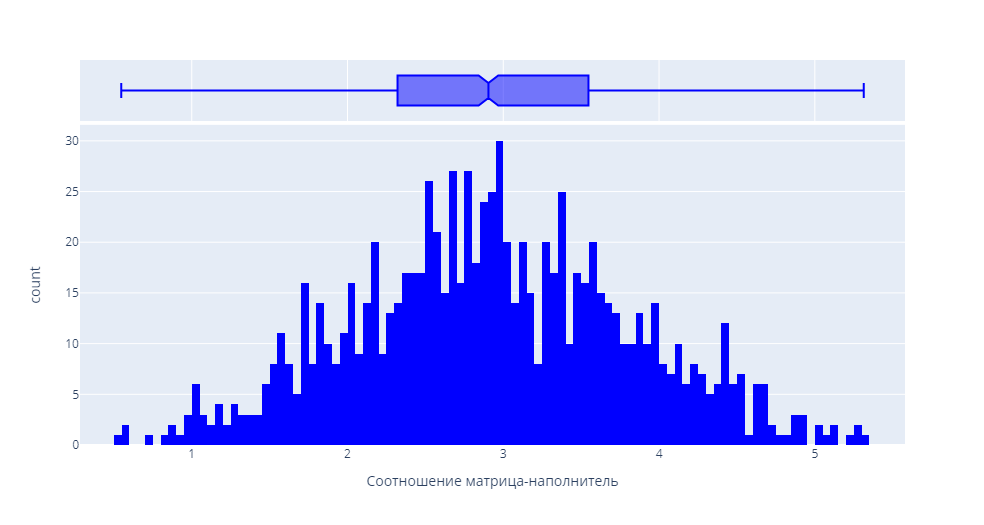
Рис. 4. Распределение и боксплот до удаления выбросов (соотношение матрица-наполнитель)****

Рис. 5. Распределение и боксплот после удаления выбросов (соотношение матрица-наполнитель)

## **2.2 Разработка и обучение модели**

Для прогноза модуля упругости при растяжении и прочности при растяжении использованы модели:

- Линейной регрессии;

- Метод k-ближайших соседей;

- «Случайный лес»;

Для обучения используем 70 % данных, а для тестирования — 30 %.

Зерно генератора случайных чисел зададим постоянным для воспроизводимости результатов обучения.

## **2.3 Тестирование моделей**

Ниже представлены результаты работы моделей для двух целевых переменных в виде соотношения тест/прогноз, а также итоговый датасет ошибок:

|  |  |  |  |
| --- | --- | --- | --- |
| model | target | MSE | R2 |
| Linear Regression | Модуль упругости | 0.035074 | -0.024522 |
| KNeighborsRegressor | Модуль упругости | 0.034788 | -0.016102 |
| Random Forest | Модуль упругости | 0.034912 | -0.019703 |

## 

### 2.3.1 Линейная регрессия

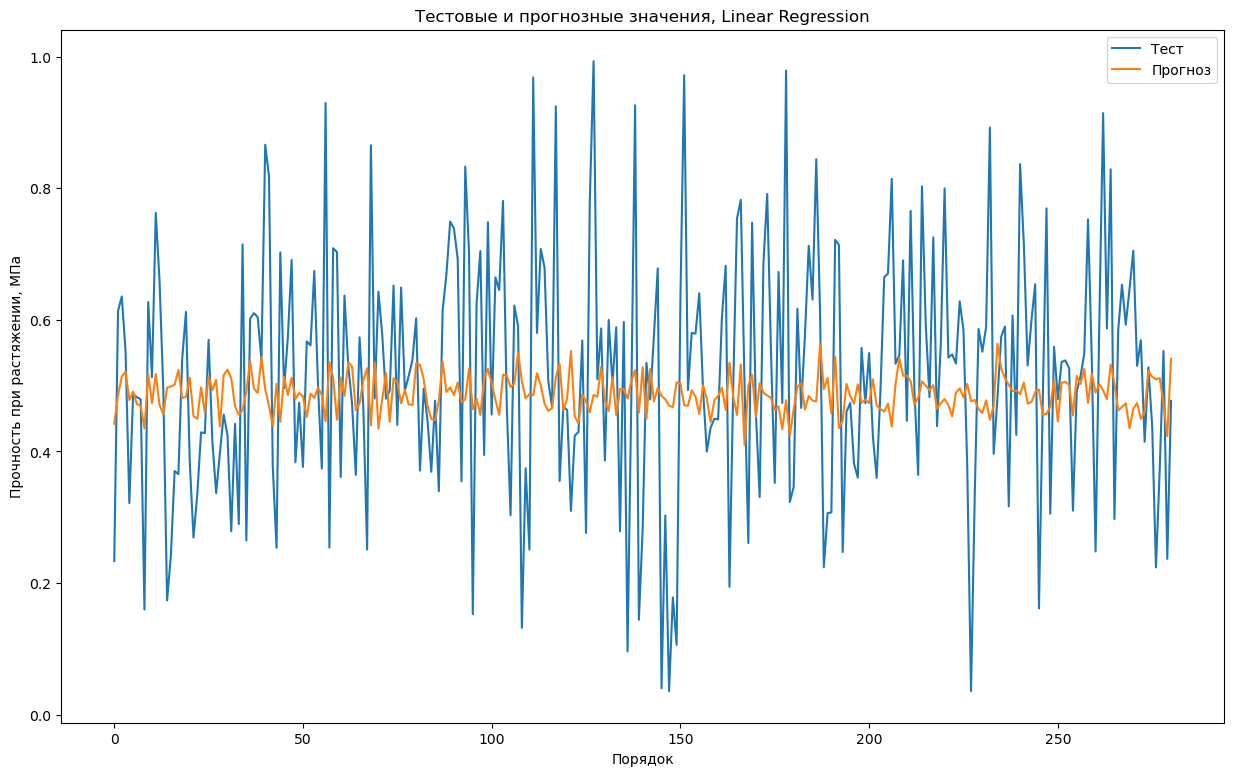
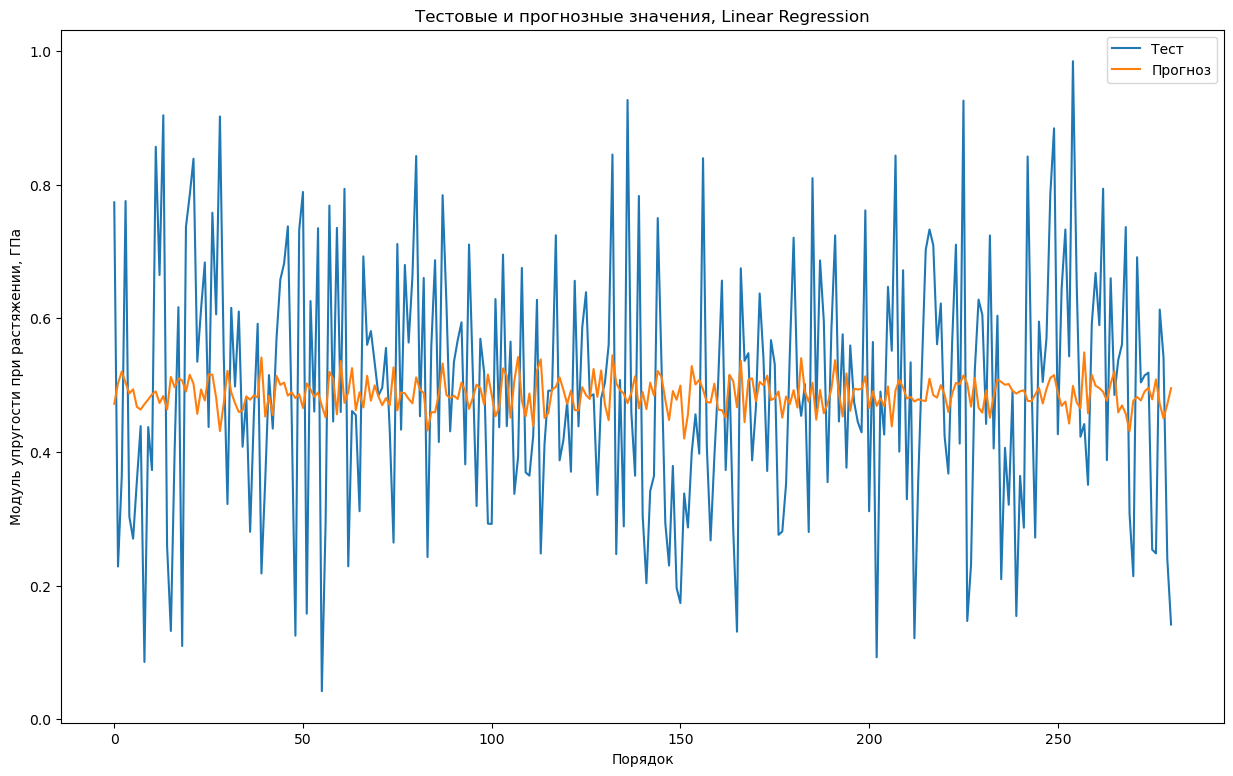


Рис. 6. Визуализация линейной регрессии

### Метод k-ближайших соседей

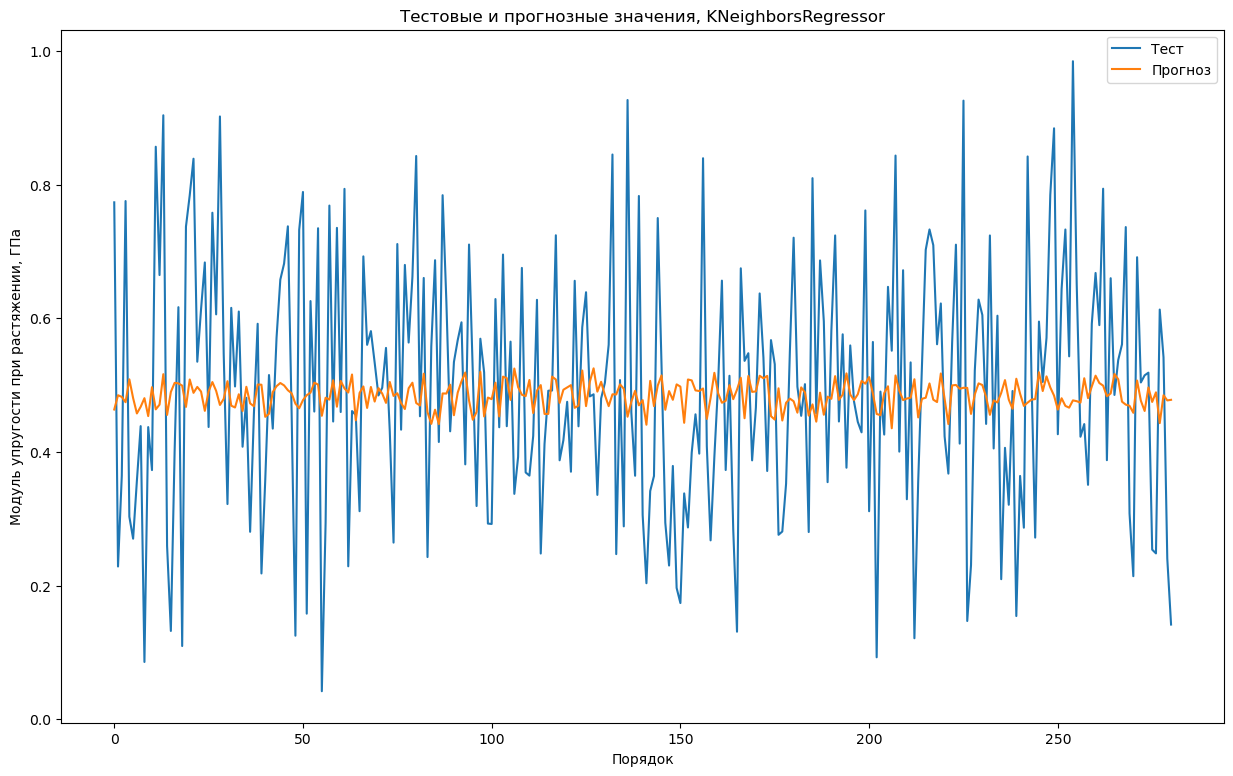
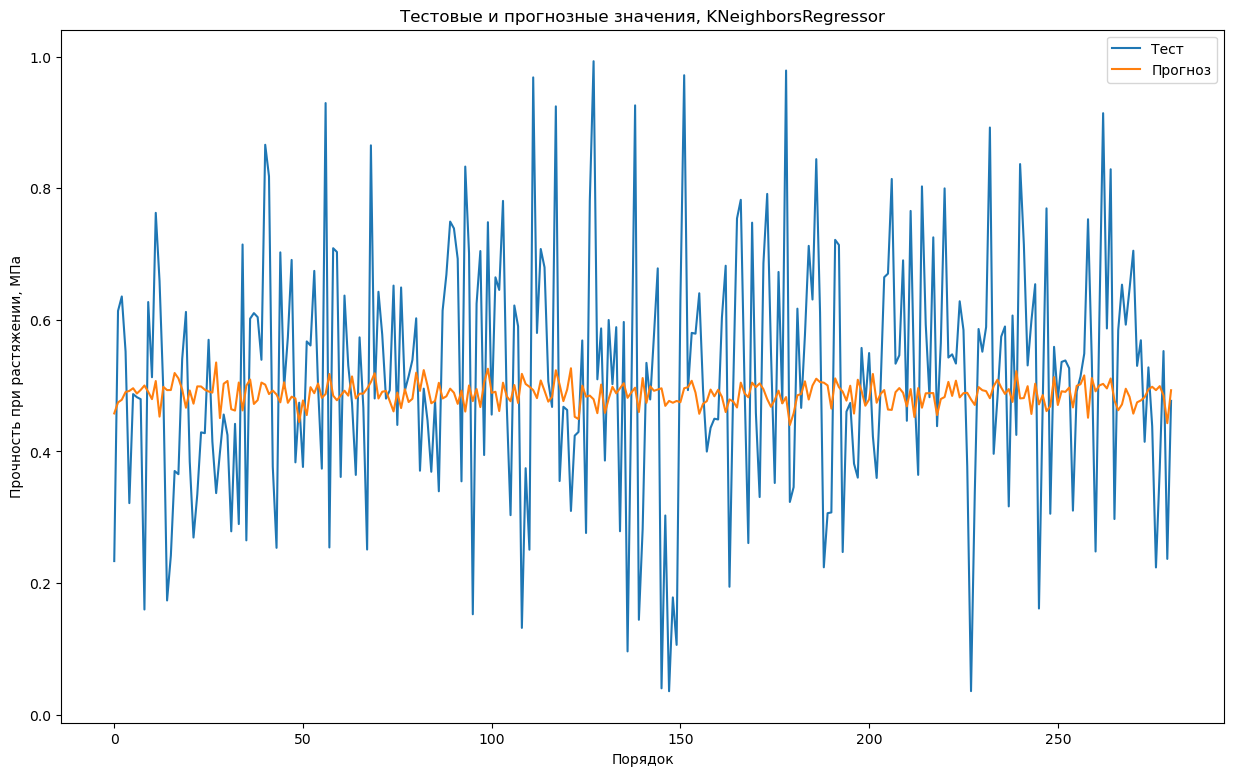


Рис.7. Визуализация k-ближайших соседей

### Случайный лес

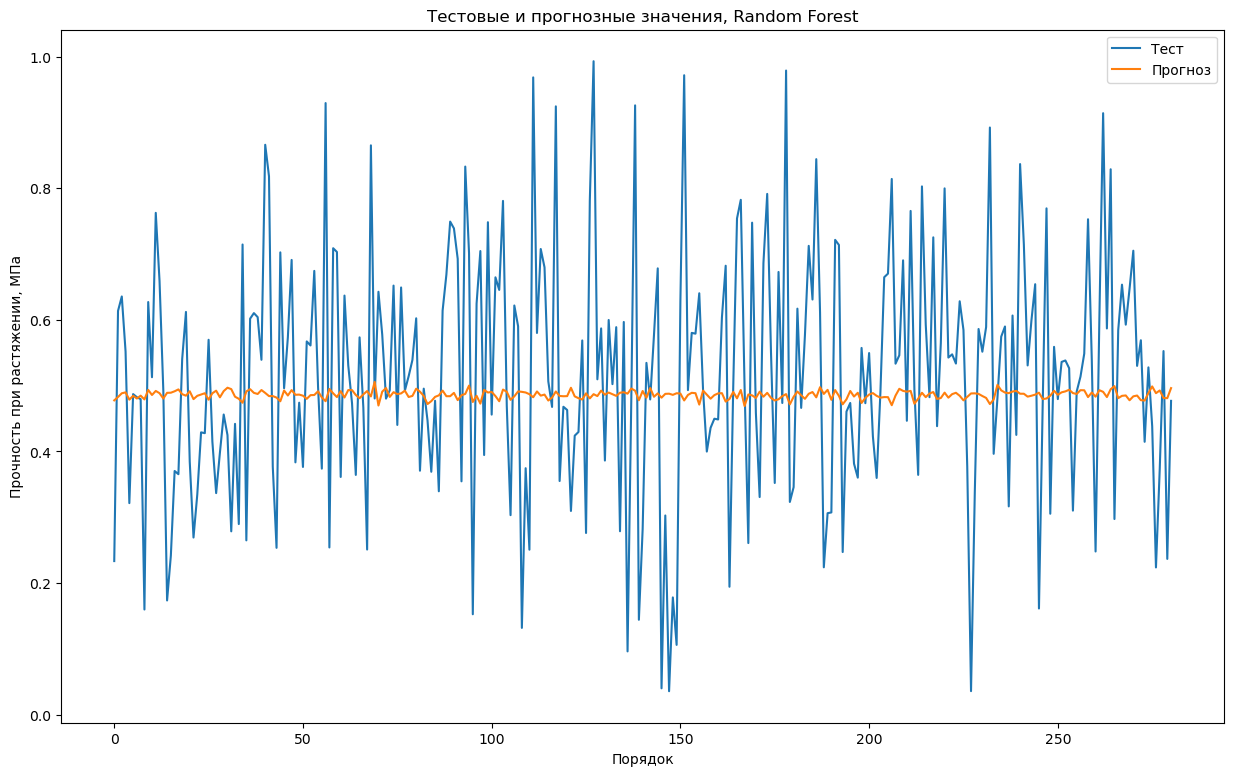
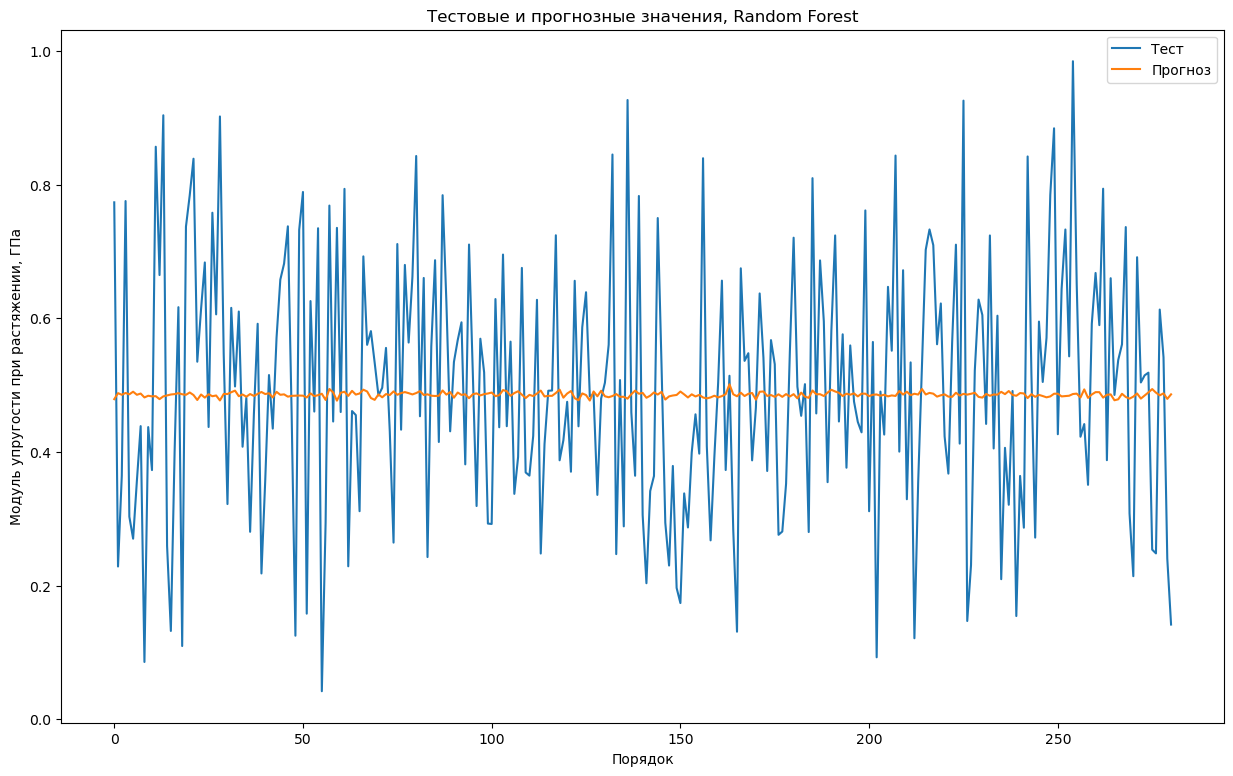


Рис. 8. Визуализация «случайный лес»

## **2.4. Нейронная сеть, которая будет рекомендовать соотношение матрица-наполнитель**

Постоим нейронную сеть с помощью класса keras.Sequential со следующими параметрами:

* входной слой нормализации 12 признаков;
* выходной слой для 1 признака;
* скрытых слоев: 5;
* нейронов в скрытом слое: 50, 128, 19, 64, 32;
* активационная функция скрытых слоев: relu;
* оптимизатор: Adam;
* loss-функция: MeanSquaredError.

Model: "sequential"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

dense (Dense) (None, 50) 650

dropout (Dropout) (None, 50) 0

dense\_1 (Dense) (None, 128) 6528

dropout\_1 (Dropout) (None, 128) 0

dense\_2 (Dense) (None, 19) 2451

dropout\_2 (Dropout) (None, 19) 0

dense\_3 (Dense) (None, 64) 1280

dropout\_3 (Dropout) (None, 64) 0

dense\_4 (Dense) (None, 32) 2080

dense\_5 (Dense) (None, 1) 33

=================================================================

Total params: 13,022

Trainable params: 13,022

Non-trainable params: 0

Рис. 9. Слои и конфигурация нейросети

Параметры нейросети следующие:

* пропорция разбиения данных на тестовые и валидационные: 30%;
* количество эпох: 60.

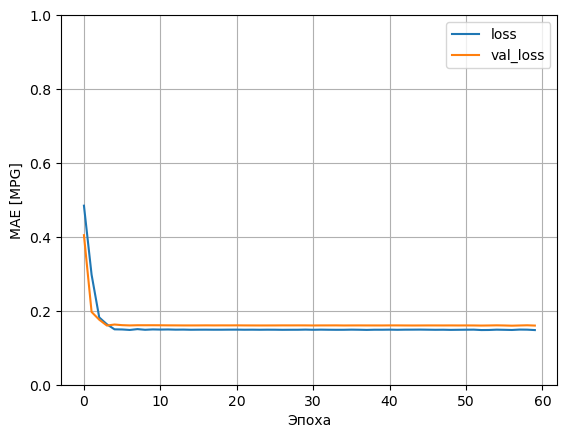


Рис. 10. График потерь

По графику делаем вывод что для обучения спроектированной нейронной сети достаточно 10 эпох.

# C:\Users\User\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\19400FED.tmp

Рис. 11. Визуализация тест/прогноз соотношения матрицы-наполнитель

Ошибки модели:

MSE = 0.03505013680909669

R2 = -0.012632641827147184

Любопытства ради была запущена нейросеть с тремя скрытыми слоями; нейронов в скрытом слое: 50, 19, 32; активационная функция скрытых слоев: tanh;

Отличий невооруженным взглядом обнаружить не удалось.

Результаты неудовлетворительны.

# **Заключение**

В ходе выполнения данной работы было выполнено:

- изучение теоретических методов анализа данных и машинного

обучения;

- разведочный анализ данных;

- предобработка данных;

- построение регрессионных моделей;

- визуализация модели и оценка качества прогноза.

Использованные при разработке моделей подходы не позволили получить достоверных прогнозов.

Работа размещена:

https://github.com/Anatoliy0461/Malon.git

# 

# **Список литературы**

1. Д.А. Иванов А.И., Ситников, С.Д Шляпин – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.
2. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
3. ГрасД. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.
4. 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.
5. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать: – Режим доступа: https://habr.com/ru/company/vk/blog/513842/