МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

**Тема:** «**Прогнозирование конечных свойств новых материалов**

**(композиционных материалов)»**

Слушатель Фесенко Елена Николаевна

Москва, 2022

СОДЕРЖАНИЕ

Введение…………………………………………………………………………3

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

Заключение………………………………………………………………………32

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

ВВЕДЕНИЕ

Композиционный материал - 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](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B8%D0%BD%D1%86%D0%B8%D0%BF_%D1%81%D1%83%D0%BF%D0%B5%D1%80%D0%BF%D0%BE%D0%B7%D0%B8%D1%86%D0%B8%D0%B8" \o "Принцип суперпозиции). В составе композита принято выделять матрицу/матрицы и наполнитель/наполнители, последние выполняют функцию армирования (по аналогии с [арматурой](https://ru.wikipedia.org/wiki/%D0%90%D1%80%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5" \o "Армирование) в таком композиционном строительном материале, как [железобетон](https://ru.wikipedia.org/wiki/%D0%96%D0%B5%D0%BB%D0%B5%D0%B7%D0%BE%D0%B1%D0%B5%D1%82%D0%BE%D0%BD" \o "Железобетон)). В качестве наполнителей композитов как правило выступают углеродные или стеклянные волокна, а роль матрицы играет полимер.

Сочетание разных компонентов позволяет улучшить характеристики материала и делает его одновременно лёгким и прочным. При этом отдельные компоненты остаются таковыми в структуре композитов, что отличает их от смесей и затвердевших растворов. Варьируя состав матрицы и наполнителя, их соотношение, ориентацию наполнителя, получают 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.

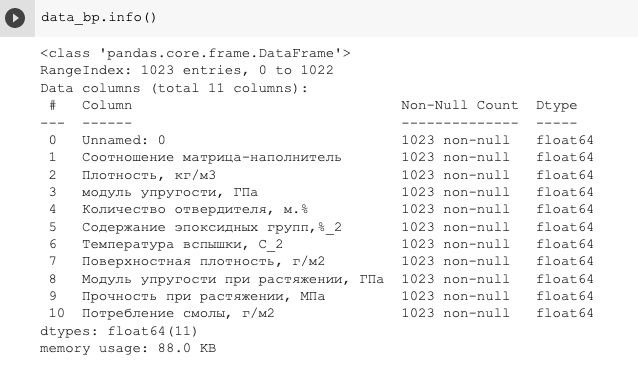
Базальтопластик представляет собой конструкцию композиционного порядка, состоящую из стеклянного наполнителя, а также из полимерного синтетического связующего. Наполнителем, который выполняет скрепляющую и армирующую функции, служат стеклянные волокна, сделанные в виде жгутов и нитей. Это стекловолокно обрабатывается особым связующим материалом, способным обеспечить высокую прочность на сдвиг и увеличить сопротивление разрушающему воздействию агрессивных химических веществ.

Актуальность выбранной темы обусловлена высокой эффективностью конечного материала при правильно выбранных компонентах, их параметрах и характеристиках. Кроме того, адекватно работающая модель может существенно сократить временные, финансовые и другие затраты на проведение испытаний.

В процессе исследовательской деятельности были разработаны несколько моделей, прогнозирующих модули упругости при растяжении и прочности при растяжении, а также были создана нейронная сеть, которая предлагает соотношение «матрицы - наполнитель». На основе нейронной сети было создано доступное пользовательское веб - приложение на фреймворке Flask.

1. АНАЛИТИЧЕСКАЯ ЧАСТЬ
   1. ПОСТАНОВКА ЗАДАЧИ

Для выполнения работы были предоставлены два датасета: X\_bp.xlsx (с данными о параметрах базальтопластика, состоящий из 1023 строки и 11 столбцов) и X\_nup.xlsx (данными нашивок углепластика, состоящий из 1040 строки и 4 столбцов).

 Рисунок 1. Состав датасета с параметрами базальтопластика

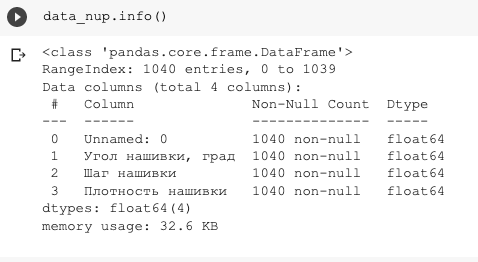


Рисунок 2. Данные о нашивках

Для того, чтобы достичь цели, то есть разработать модели для прогноза прочности при растяжении, модуля упругости при растяжении, а также, нейронную сеть, которая будет рекомендовать соотношение «матрица-наполнитель», необходимо объединить датасеты. Так как в них различное количество строк, используем тип объединения INNER. В результате объединения у нас получается один датасет, в котором 1023 строки и 13 столбцов.



Рисунок 3. Объединение датасетов

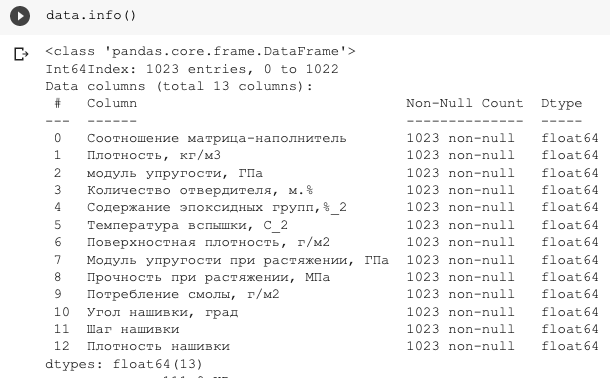


Рисунок 4. Объединенный датасет

Затем необходимо провести разведочный анализ данных, построить гистограммы распределения, попарные графики рассеяния точек для определения зависимостей между параметрами, диаграммы «ящик с усами» (boxplot).

Кроме того, для каждой колонки нужно получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; произвести предварительную обработку данные: удалить шумы и 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.

* 1. ОПИСАНИЕ ИСПОЛЬЗУЕМЫХ МЕТОДОВ

Для решения данной задачи обратимся к принципу обучения с учителем (supervised learning). В основном обучение с учителем применяется для решения двух типов задач: классификации и регрессии. В задачах классификации имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества. В задачах регрессии требуется определить влияние одной или нескольких независимых переменных на зависимую переменную.

Данная задача, очевидно, относится к задачам регрессии. Цель любого алгоритма обучения с учителем — определить функцию потерь и минимизировать её, поэтому для наилучшего решения в процессе исследования были применены следующие методы:

* метод опорных векторов;
* случайный лес;
* линейная регрессия;
* градиентный бустинг;
* К-ближайших соседей;
* дерево решений;
* многослойный перцептрон;
* Лассо.

1. Метод опорных векторов (Support Vector Regression) – этот бинарный линейный классификатор был выбран, потому что он хорошо работает на небольших датасетах, каким и является наш датасет. Данный алгоритм – это алгоритм обучения с учителем, использующийся для задач классификации и регрессионного анализа, это контролируемое обучение моделей с использованием 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. Также говорят, что это метод классификатора с максимальным зазором.

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 тестовой выборке, вполне возможно применение данного метода на реальных данных. Эффективен при большом количестве гиперпараметров. Существует возможность гибко настраивать разделяющую функцию.  Алгоритм максимизирует разделяющую полосу, которая, как подушка безопасности, позволяет уменьшить количество ошибок классификации.

Недостатки метода: неустойчивость к шуму, в обучающих данных шумы могут стать опорными объектами и напрямую повлиять на построение разделяющей гиперплоскости. Для больших наборов данных требуется долгое время обучения; достаточно сложно подбирать полезные преобразования данных; параметры модели сложно интерпретировать, поэтому были рассмотрены и другие методы.

2. Метод случайного леса (Random Forest) - это множество решающих деревьев. Универсальный алгоритм машинного обучения с учителем, представитель ансамблевых методов.

Достоинства метода: не переобучается, не требует предобработки входных данных, эффективно обрабатывает пропущенные данные, данные с большим числом классов и признаков, имеет высокую точность предсказания и внутреннюю оценку обобщающей способности модели, а также высокую масштабируемость.

Недостатки метода: построение занимает много времени, сложно интерпретируемый, не обладает возможностью экстраполяции, может недообучаться, трудоёмко прогнозируемый.

3. Линейная регрессия (Linear regression) — это алгоритм машинного обучения, основанный на контролируемом обучении, рассматривающий зависимость между одной входной и выходными переменными. Это один из самых простых и эффективных инструментов статистического моделирования. Она определяет зависимость переменных с помощью линии наилучшего соответствия. Модель регрессии создаёт несколько метрик. R2 , или коэффициент детерминации, позволяет измерить, насколько модель может объяснить дисперсию данных. Если R2 равен 1, это значит, что модель описывает все данные. Если же R2 равен 0,5, модель объясняет лишь 50 процентов дисперсии данных. Оставшиеся отклонения не имеют объяснения. Чем ближе R2 к единице, тем лучше.

Достоинства метода: быстр и прост в реализации, легко интерпретируем, имеет меньшую сложность по сравнению с другими алгоритмами.

Недостатки метода: моделирует только прямые линейные зависимости, требует прямую связь между зависимыми и независимыми переменными, выбросы оказывают огромное влияние, а границы линейны.

4. 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.

5. Метод К - ближайших соседей (kNN - k - Nearest Neighbours) ищет ближайшие объекты с известными значения целевой переменной и основывается на хранении данных в памяти для сравнения с новыми элементами. 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).

Достоинства метода: 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 метода: замедляется с ростом объёма данных, не обобщает предыдущий опыт, основывается на всем массиве доступных исторических данных, сложно выбрать близость метрики, имеет высокую зависимость результатов классификации от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоёмкость.

6. Дерево решений (DecisionTreeRegressor) – метод автоматического анализа больших массивов данных.  Это инструмент принятия решений, в котором используется древовидная структура, подобная блок-схеме, или модель решений и всех их возможных результатов, включая результаты, затраты и полезность. Дерево решений - эффективный инструмент интеллектуального анализа данных и предсказательной аналитики. Алгоритм дерева решений подпадает под категорию контролируемых алгоритмов обучения. Регрессия дерева решений отслеживает особенности объекта и обучает модель в структуре дерева прогнозированию данных в будущем для получения значимого непрерывного вывода. Дерево решений - один из вариантов решения регрессионной задачи, в случае, если зависимость в данных не имеет очевидной корреляции.

Достоинства метода: создаются по понятным правилам, просты в применении и интерпретации, заполняют пропуски в данных наиболее вероятным решением, выделяют наиболее важные поля для прогнозирования.

Недостатки метода: ошибаются при классификации с большим количеством классов и небольшой обучающей выборкой, имеют нестабильный процесс (изменение в одном узле может привести к построению совсем другого дерева), имеет затратные вычисления, необходимо обращать внимание на размер, ограниченное число вариантов решения проблемы.

7. Многослойный персептрон (MLPRegressor) — это алгоритм обучения с учителем, который изучает функцию f(⋅): Rm→Ro обучением на наборе данных, где m — количество измерений для ввода и o- количество размеров для вывода. Это искусственная нейронная сеть, имеющая 3 или более слоёв персептронов. Эти слои - один входной слой, 1 или более скрытых слоёв и один выходной слой персептронов.

Достоинства метода: построение сложных разделяющих поверхностей, возможность осуществления любого отображения входных векторов в выходные, легко обобщает входные данные, не требует распределения входных векторов: изучает нелинейные модели.

Недостатки метода: разные инициализации случайных весов могут привести к разной точности проверки, чувствителен к масштабированию функций.

8. Лассо регрессия (Lasso) — это линейная модель, которая оценивает разреженные коэффициенты.  Это простой метод, позволяющий уменьшить сложность модели и предотвратить переобучение, которое может возникнуть в результате простой линейной регрессии. Данный метод вводит дополнительное слагаемое регуляризации в оптимизацию модели. Это даёт более устойчивое решение. В лассо регрессии добавляется условие смещения в функцию оптимизации для того, чтобы уменьшить коллинеарность и, следовательно, дисперсию модели. Но вместо квадратичного смещения, используется смещение абсолютного значения. Лассо регрессия хорошо прогнозирует модели временных рядов на основе регрессии (авторегрессии).

Достоинства метода: быстро работает, не очень энергоёмка, доступно обнуляет значения коэффициентов.

Недостатки метода: часто страдает качество прогнозирования, случайным образом выбирает одну из коллинеарных переменных, не оценивает правильность формы взаимосвязи между независимой и зависимой переменными.

В работе использовались следующие метрики качества моделей: R2 или коэффициент детерминации измеряет долю дисперсии, объяснённую 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 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.

MAE (Mean Absolute Error) или средняя абсолютная ошибка - мера ошибок между парными наблюдениями, выражающими одно и то же явление.

1.3. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

Прежде, чем начинать работу по построению моделей машинного обучения, данные необходимо очистить и предобработать. Неочищенные и необработанные данные могу содержать выбросы, шумы, пропуски, искажения. Все это может существенно повлиять на качество модели и ее дальнейшую работу. Кроме того, результаты могут привести нас в заблуждение, соответственно, будут сделаны неверные выводы, что может иметь крайне негативные последствия. Однако, просто так удалять или добавлять какие-либо данные тоже неверно. Поэтому, сначала нужно внимательно изучить первоначальный набор данных. Посмотрим на описательную статистику датасета.

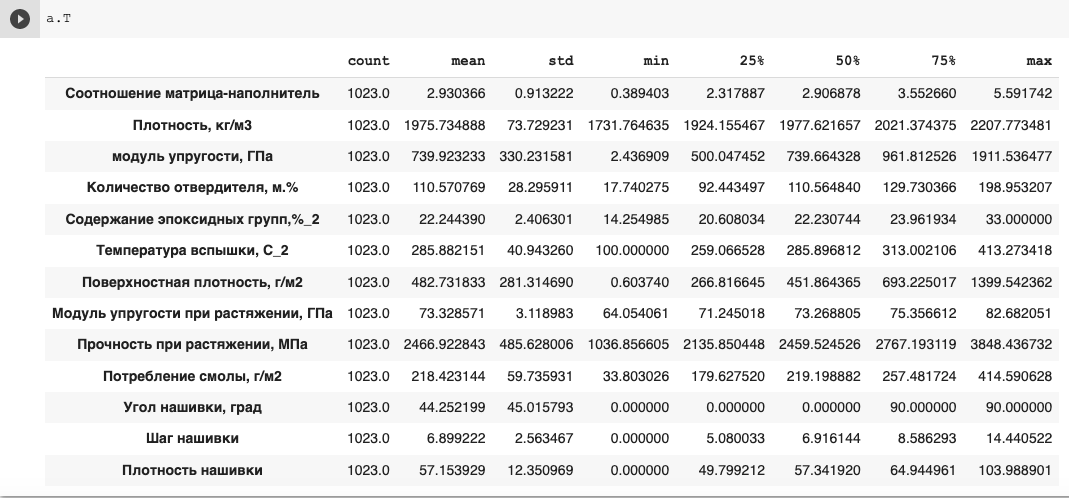


Рисунок 5. Описательная статистика

Цель разведочного анализа - получение первоначальных представлений о характерах распределений переменных исходного набора данных, формирование оценки качества исходных данных (наличие пропусков, выбросов, шумов), выявление характера взаимосвязи между переменными с целью последующего выдвижения гипотез о наиболее подходящих для решения задачи моделях машинного обучения.

Посмотрим количество уникальных значений по каждому столбцу, наличие пропусков – 0, наличие дубликатов – 0. Также, выведем среднее и медианное значение. Затем, построим гистограммы распределения каждой из переменных, диаграммы «ящики с усами» и попарные графики рассеяния точек.

В качестве инструментов разведочного анализа используется: гистограммы распределения каждой из переменных, диаграммы «ящики с усами», попарные графики рассеяния точек, тепловая карта, описательная статистика для каждой переменной, анализ и работа с выбросами, проверка наличия пропусков и дубликатов.

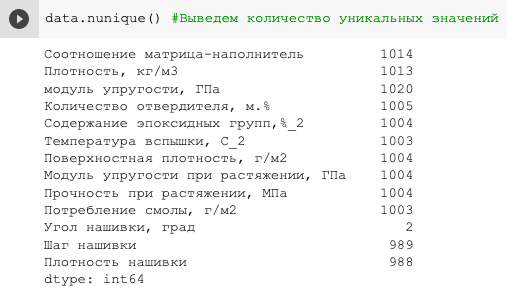


Рисунок 6.Количество уникальных значений

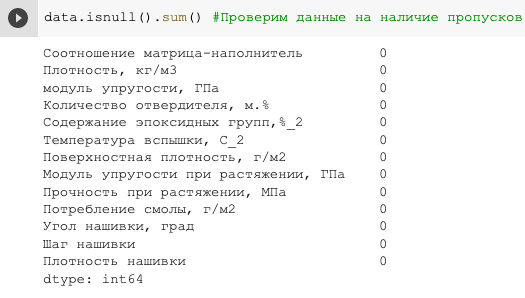


Рисунок 7. Наличие пропусков

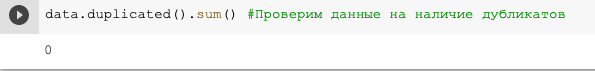


Рисунок 8. Наличие дубликатов

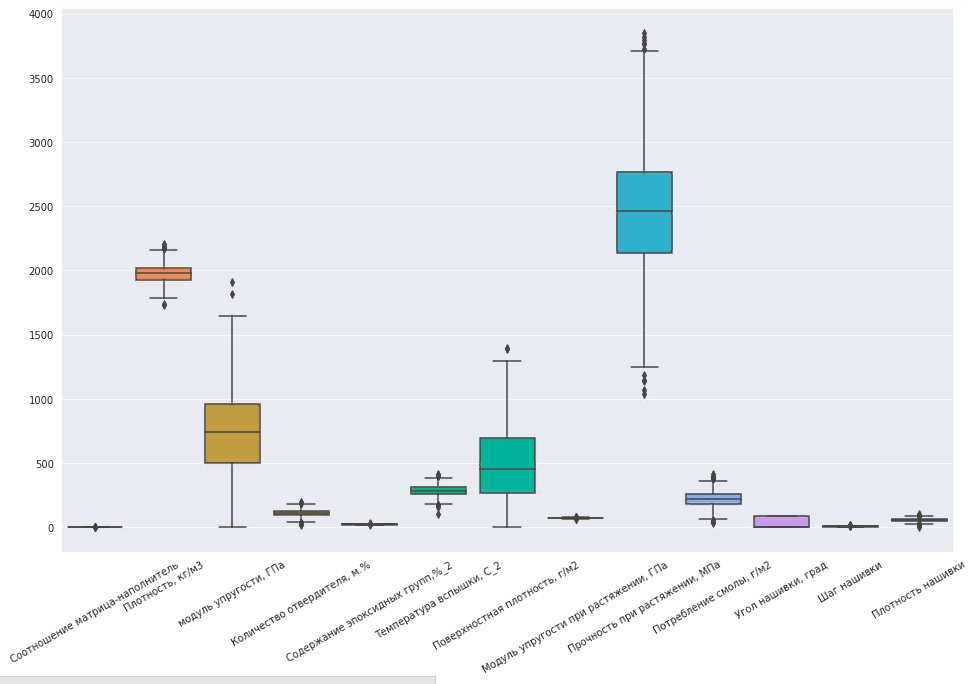


Рисунок 9."Ящики с усами" до удаления выбросов

Основываясь на графиках «ящики с усами»(или boxplot), мы видим, что в наших данных есть некоторое количество выбросов, которые могут повлиять на результаты анализа данных, а также, на дальнейшее построение моделей и их результаты. Поэтому проведем очистку данных от выбросов по z оценке.

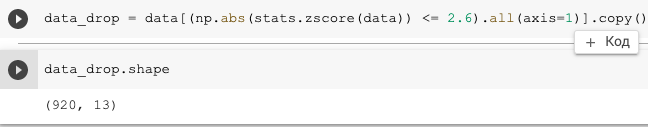


Рисунок 10. Удаление выбросов

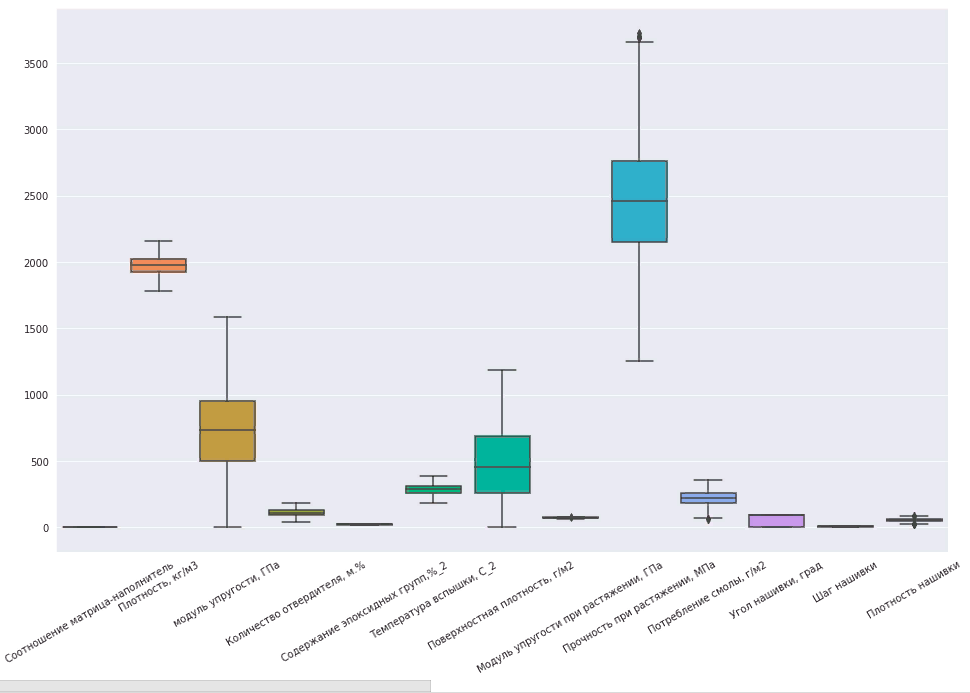


Рисунок 11."Ящики с усами" после удаления выбросов

Удалив около 10% датасета, мы видим, что не все выбросы исчезли. Но удалять больше уже нецелесообразно, так как будут слишком большие потери. Поэтому оставим так.

Посмотрим на среднее и медианное значение параметров очищенного датасета.

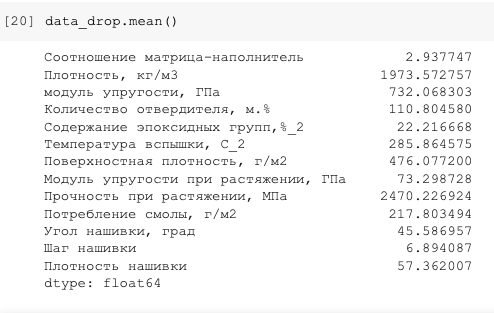


Рисунок 12. Среднее значение после удаления выбросов

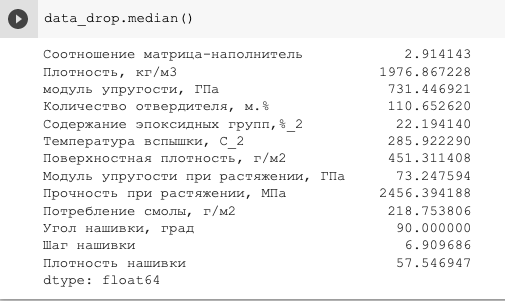


Рисунок 13.Медианное значение после удаления выбросов

Построив и изучив графики попарного рассеяния точек, можем сделать вывод, что какая-либо корреляция между параметрами отсутствует. Тепловая карта также показывает максимальную корреляцию между углом нашивки и плотностью нашивки и между плотностью и плостонстью нашивки равную 0, 10, что не является сколько-то значимым. Между остальными параметрами корреляция стремится к 0, зависимости между параметрами не наблюдаются.

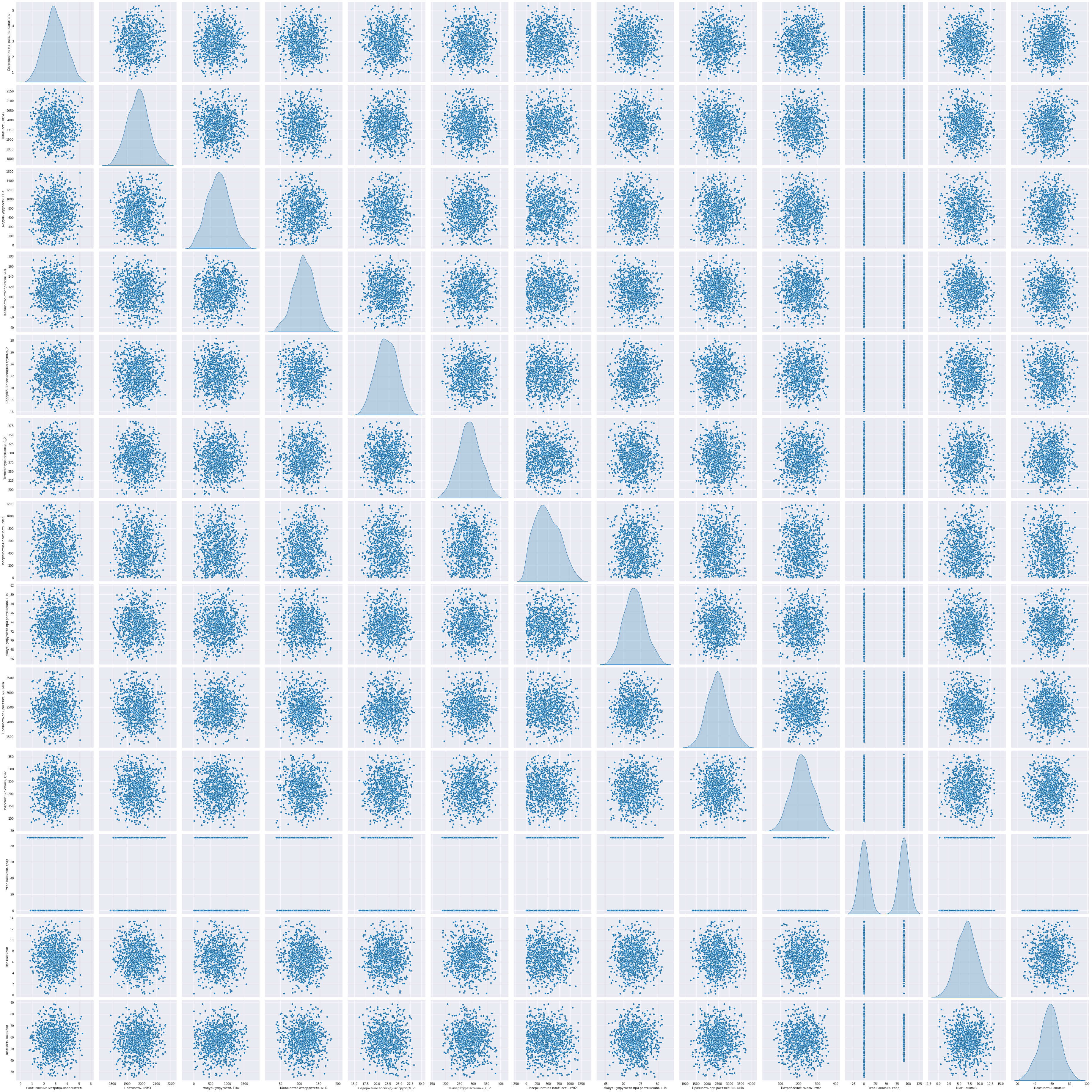


Рисунок 14.Графики попарного рассеяния точек после удаления выбросов

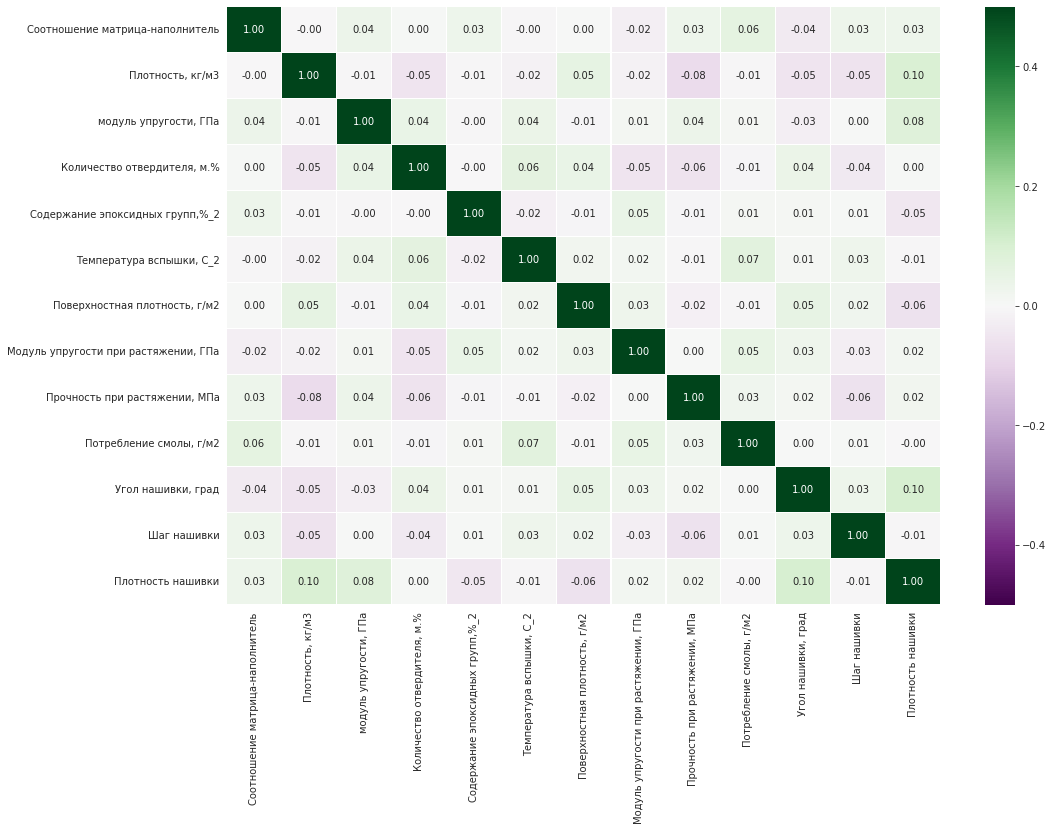


Рисунок 15.Тепловая карта после удаления выбросов

2. ПРАКТИЧЕСКАЯ ЧАСТЬ

* 1. ПРЕДОБРАБОТКА ДАННЫХ

Для того, чтобы привести данные к единому виду, проведем масштабирование данных: нормализацию и стандартизацию.

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

Стандартизация данных - приведения к определённому формату и представлению, которые обеспечивают их корректное применение в многомерном анализе, совместных исследованиях, сложных технологиях аналитической обработки.

Далее мы будем использовать нормализованные данные с помощью метода MinMaxScaler из библиотеки SKLearn.

Посмотрим на гистограммы распределения до и после нормализации.

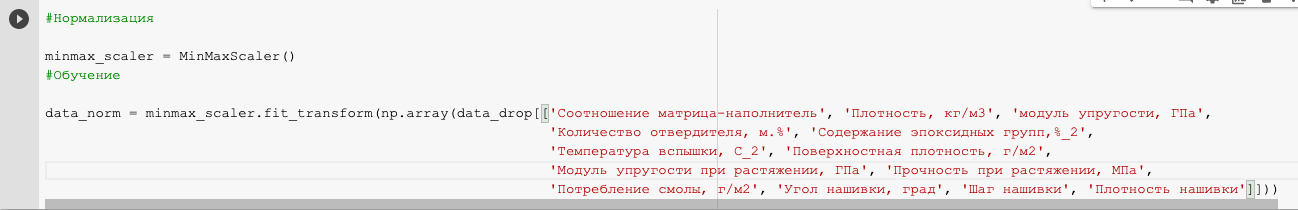
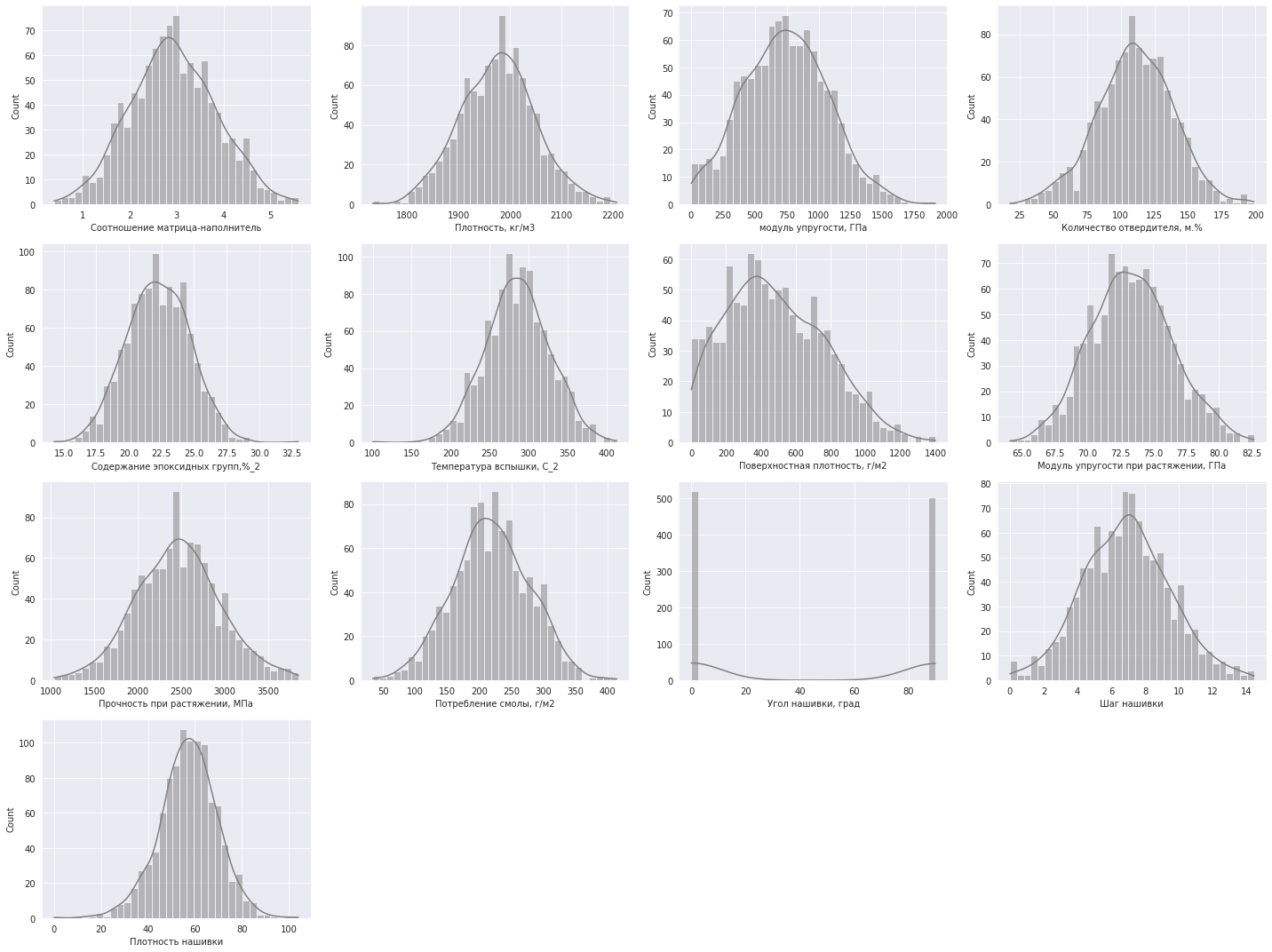
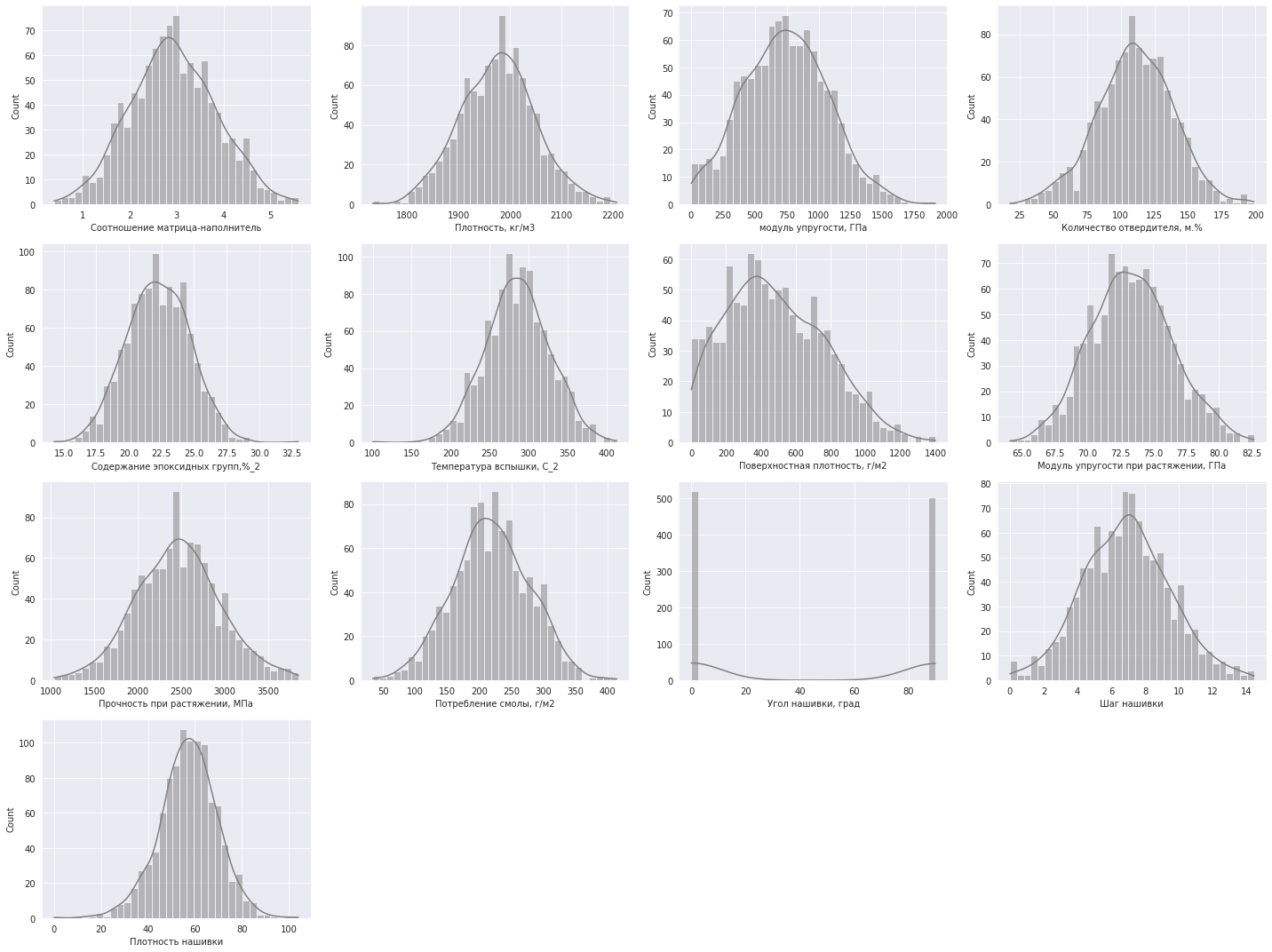


Рисунок 16. Нормализация данных.

 Рисунок 17. Гистограммы распределения до(слева) и после(справа) нормализации.

* 1. РАЗРАБОТКА И ОБУЧЕНИЕ МОДЕЛИ

Разработка и обучение моделей машинного обучения осуществлялась для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении» отдельно. Для решения мы применили все модели, описанные в Главе 1.

Порядок разработки модели можно разделить на следующие этапы: разделение нормализованных данных на обучающую и тестовую выборки (в соотношении 70 на 30%), проверка моделей при стандартных значениях, сравнение с результатами модели, выдающей среднее значение, создание графика, сравнение моделей по метрике МАЕ, поиск гиперпараметров, по которым будет происходить оптимизация модели. В качестве параметра оценки выбран коэффициент детерминации (R2), оптимизация подбора гиперпараметров модели с помощью выбора по сетке и перекрёстной проверки(кросс-валидации), подстановка оптимальных гиперпараметров в модель и обучение модели на тренировочных данных, оценка полученных данных, сравнение со стандартными значениями.

Для начала, разделим датасет на обучающую и тестовую выборки методом train\_test\_split в соотношении 70 на 30 (эту процедуру проведем для прогноза модуля упругости и для прочности при растяжении).

1. Построим модели методом Опорных векторов для обоих параметров, построим графики, выведем метрики (посмотрим на сравнительные таблицы по метрикам ниже).

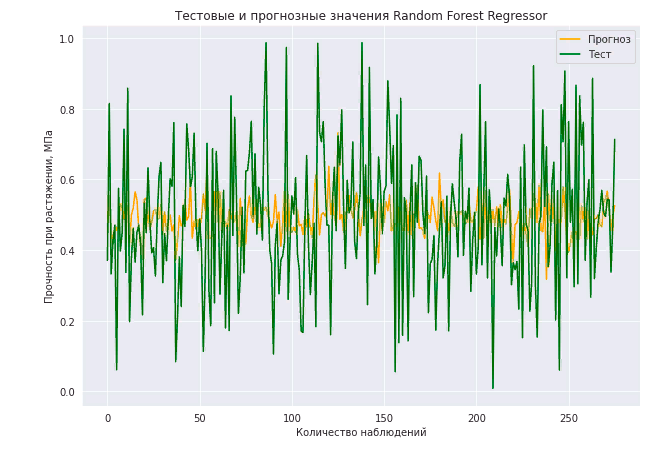
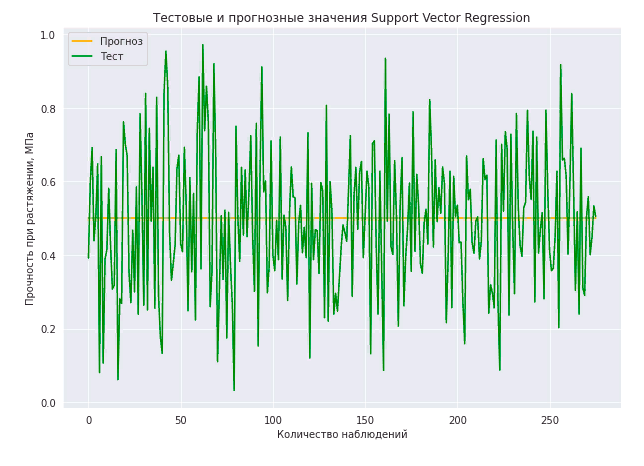


Рисунок 18. Метод опорных векторв для прочност при растяжении

Рисунок 19.Метод опорных векторов для упругости при растяжении

1. Построим модели методом Случайного леса, построим графики.

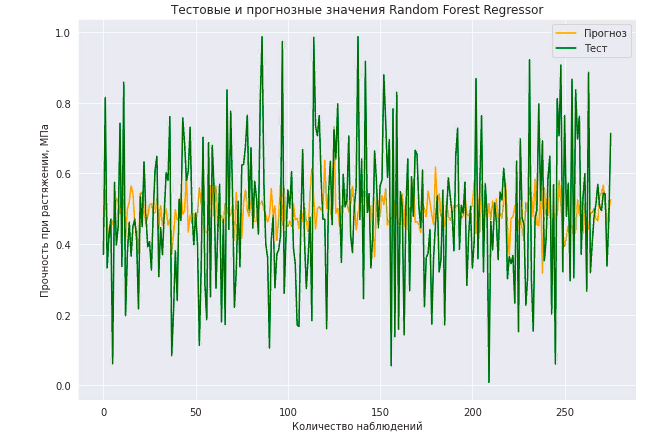


Рисунок 20. Метод Случайного леса для прочности при растяжении

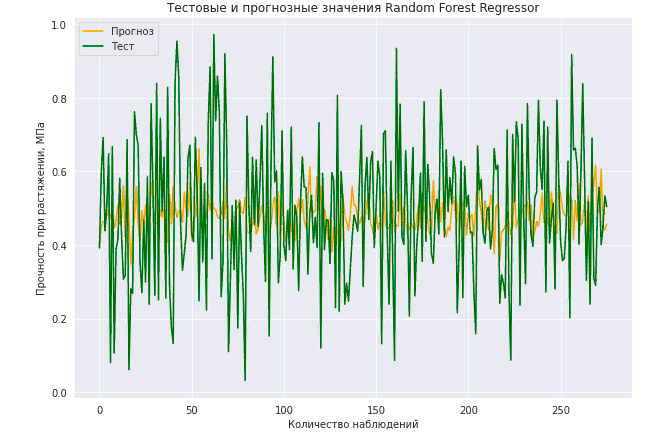


Рисунок 21.Метод Случайного леса для упругости при растяжении

1. Построим модели методом линейной регресии, построим графики.

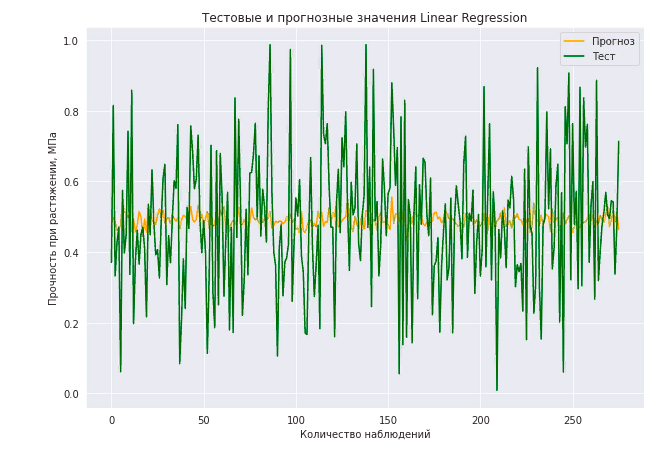


Рисунок 22. Линейная регрессия для прочности при растяжении

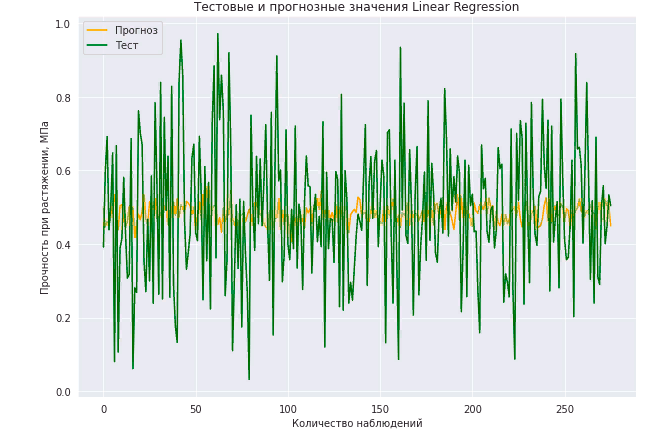


Рисунок 23. Линейная регрессия для упругости при растяжении

1. Построим модели методом k-ближайших соседей, построим графики.

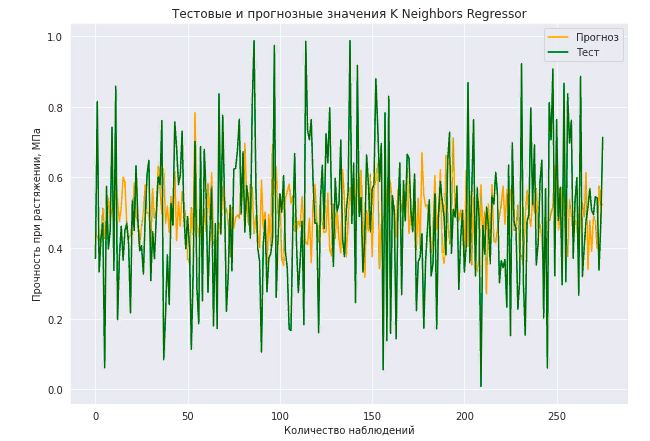


Рисунок 24. Метод k-means для прочности при растяжении

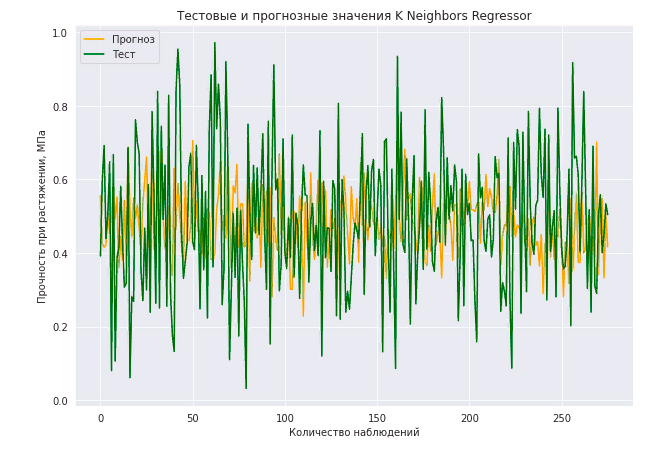


Рисунок 25. Метод k-means для упругости при растяжении

1. Построим модели методом Дерево решений, построим графики.

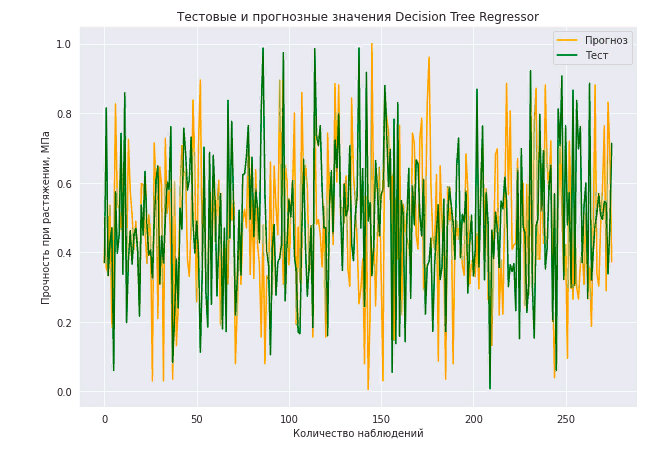


Рисунок 26. Метод Decision Tree для прочности при растяжении

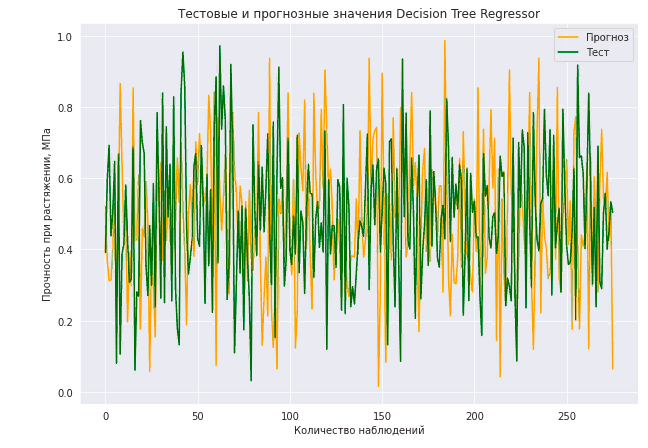


Рисунок 27. Метод Decision Tree для упругости при растяжении

1. Построим модели методом градиентного бустинга, построим график.

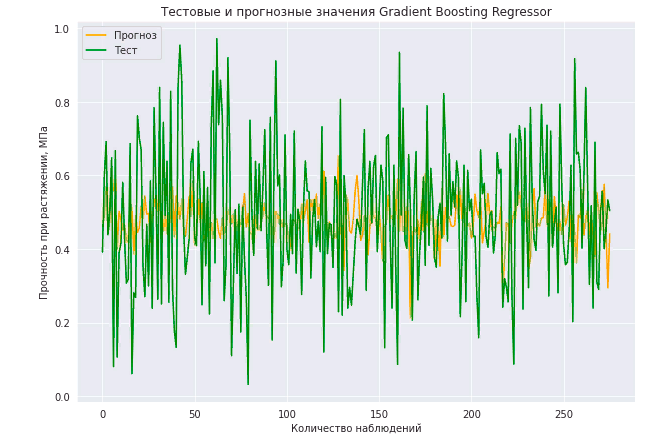
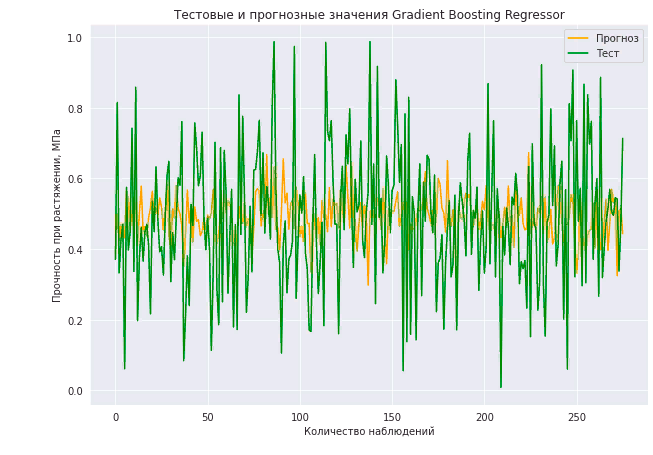
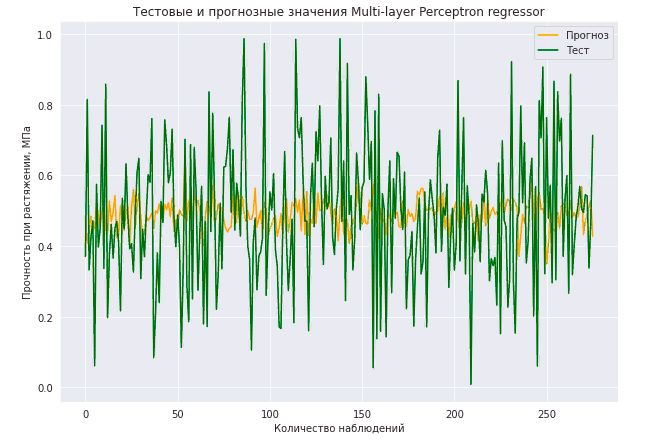
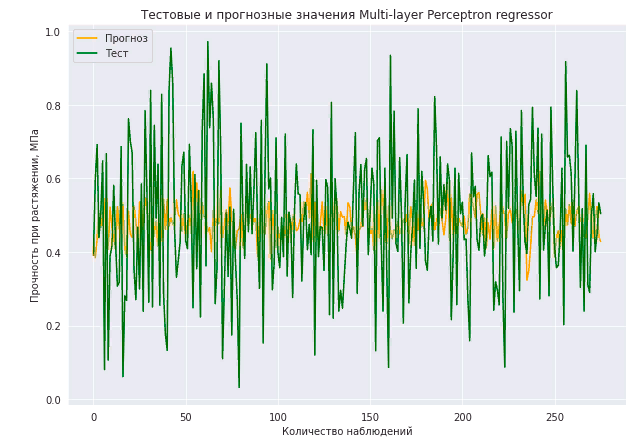


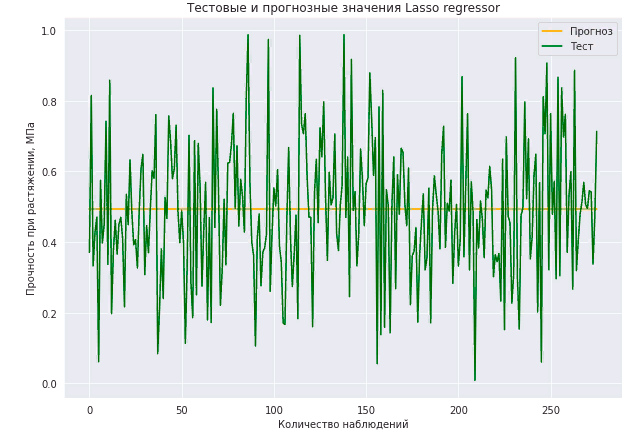
Рисунок 29. Метод градиентного бустинга для упругости при растяжении

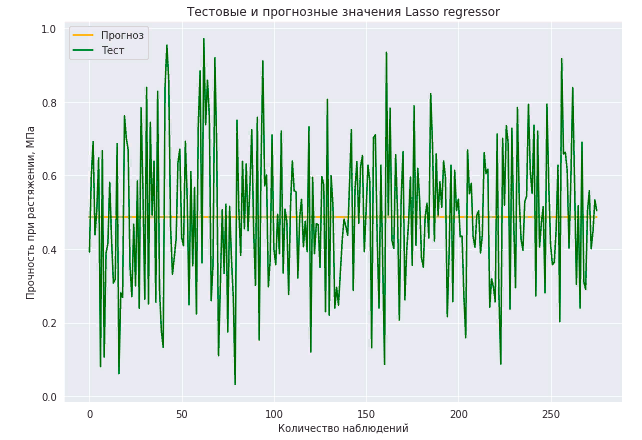
1. Построим модели методом многослойного персептрона, построим графики.

Рисунок 30. Метод многослойного персептрона для прочности при растяжении

Рисунок 31. Метод многослойного персептрона для упругости при растяжении

1. Построим модели методом Лассо регрессии, построим графики.

Рисунок 32. Метод Лассо регрессии для прочности при растяжении

Рисунок 33. Метод Лассо регрессии для упругости при растяжении

* 1. ТЕСТИРОВАНИЕ МОДЕЛИ

После обучения моделей была проведена оценка точности этих моделей на обучающей и тестовой выборках. В качестве параметра оценки модели использовалась средняя абсолютная ошибка (MAE). Результат неудовлетворительный.

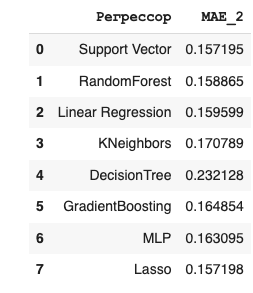
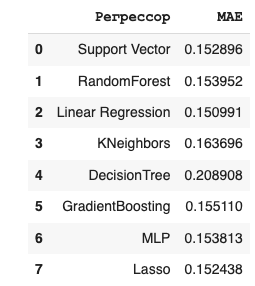


Рисунок 34. Средняя абсолютная ошибка при обучении моделей прогноза прочности при растяжени(слева) и упругости при растяжении(справа)

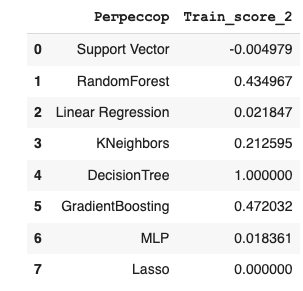
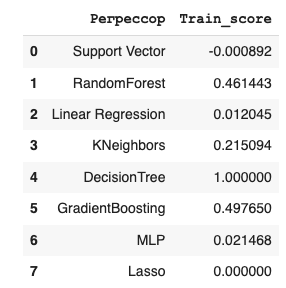


Рисунок 35. Результат работы моделей по прогнозу прочности при растяжении(слева) и упругости при растяжении(справа) на обучающей выборке

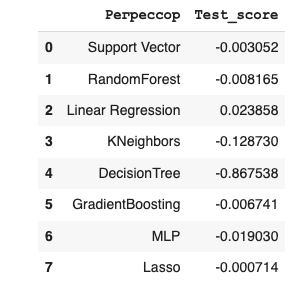
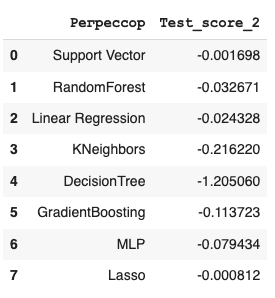
 

Рисунок 36. Результаты работы моделей по прогнозу прочности при растяжении(слева) и упругости при растяжении(справа) на тестовой выборке

При построении следующих моделей был осуществлен поиск по сетке гиперпараметровс перекрестной проверкой: Случайный лес, К-Ближайших соседей, Дерево решений. Поиск гиперпараметров дал положительный результат R2(0.003) для прогноза прочности при растяжении только в модели К-ближайших соседей, лишь незначительно уменьшив параметр МАЕ. Поиск гиперпараметров дал положительный результат R2(0.007) для прогноза упругости при растяжении также только в модели К-ближайших соседей. МАЕ уменьшился несколько больше, чем при прогнозе прочности при растяжении. Стоит отметить, что отрицательный коэффициент детерминации также может давать уменьшение ошибки, как показал пример модели Дерево решений для прогноза упругости при растяжении. При этом на тестовой выборке лучше всего себя показал метод Опорных векторов. Он был выбран лучшей моеделью.

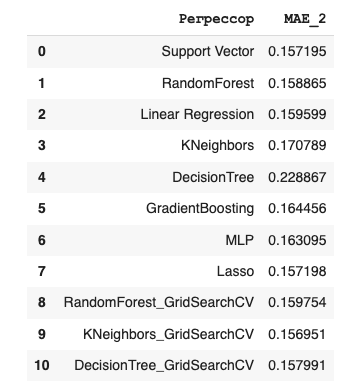
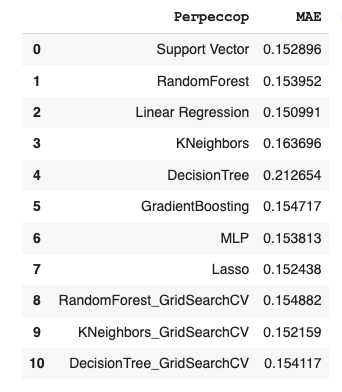


Рисунок 37. Сравнительные таблицы ошибки МАЕ с добавлением гиперпараметров при прогнозе прочности при растяжении(слева) и упругости при растяжении(справа).

* 1. НЕЙРОННАЯ СЕТЬ

Обучение нейронной сети — это такой процесс, при котором происходит подбор оптимальных параметров модели, с точки зрения минимизации функционала ошибки. Начнём стоить нейронную сеть с помощью класса keras.Sequential.

Весь датасет так же разделил на обучающую и тестовую выборки в пропорции 70 на 30, в качестве целевой переменной взяли «Соотношение Матрица-Наполнитель», которое должна рекомендовать наша нейронная сеть.

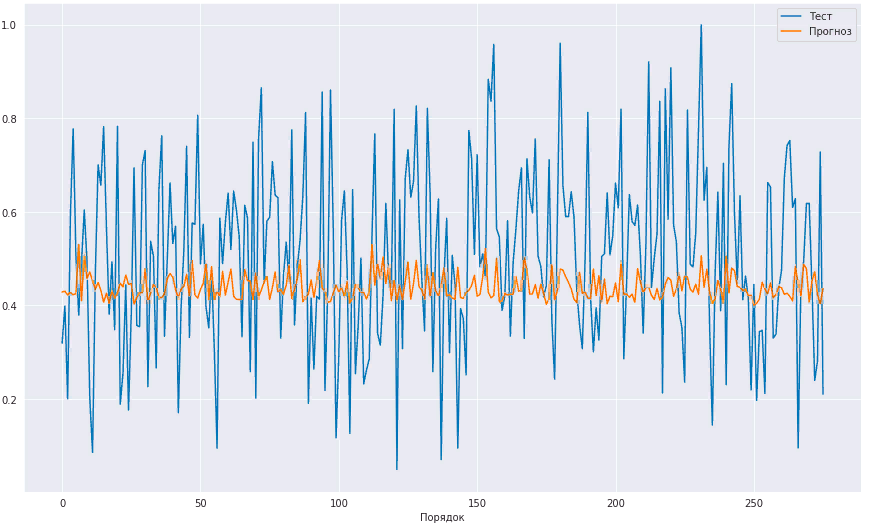


Рисунок 38. Результат работы нейронной сети.

График очень наглядно показывает, что результат работы нейронной сети оказался неудовлетворительный.

* 1. РАЗРАБОТКА ПРИЛОЖЕНИЯ

Приложение успешно работает и показывает результат прогноза для соотношения «матрица – наполнитель».

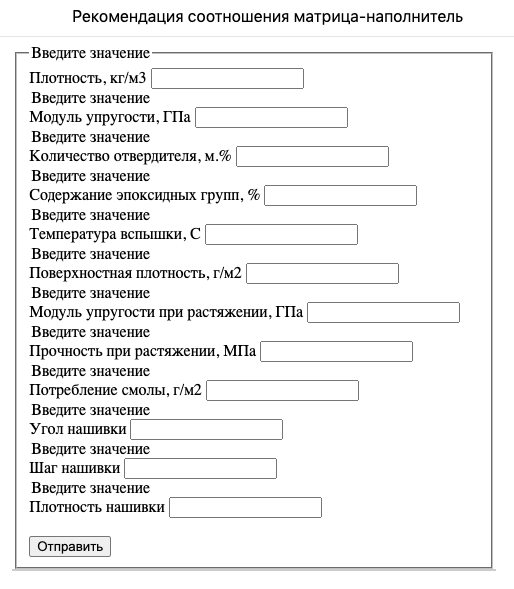
****

Рисунок 39. Приложение

Данное приложение — это файл Flask, папка templates, с шаблоном html - страницы, папка model\_nn.joblib c сохранённой моделью для данных. При запуске приложения пользователь переходит на <http://43a2-34-173-193-155.ngrok.io/>. В открывшемся окне пользователю необходимо ввести в соответствующие ячейки требуемые значения и нажать на кнопку «Отправить».

На выходе пользователь получает результат прогноза для значения параметра «Соотношение «матрица – наполнитель»».

2.6. СОЗДАНИЕ РЕПОЗИТОРИЯ

Репозиторий был создан на платформе github, по адресу: <https://github.com/ElenaFesenko/Final_project_bauman_Fesenko>.

Ноутбук с решением и приложением так же можно найти по адресу:

<https://colab.research.google.com/drive/1d2qWJyZa-IFkdBayIA8PSpTCmDSg5SzI?usp=share_link>.

ЗАКЛЮЧЕНИЕ

Данная исследовательская работа позволяет сделать некоторые основные выводы по теме. Распределение полученных данных в объединённом датасете близко к нормальному, но коэффициенты корреляции между парами признаков стремятся к нулю. Использованные при разработке моделей подходы не позволили получить сколько-нибудь достоверных прогнозов. Применённые модели регрессии не показали высокой эффективности в прогнозировании свойств композитов. Лучшие метрики для модуля упругости при растяжении и для прочности для растяжении– метод опорных векторов.

Был сделан вывод, что невозможно определить из свойств материалов соотношение «матрица – наполнитель». Данный факт не указывает на то, что прогнозирование характеристик композитных материалов на основании предоставленного набора данных невозможно, но может указывать на недостатки базы данных, подходов, использованных при прогнозе, необходимости пересмотра инструментов для прогнозирования.

Необходимы дополнительные вводные данные, получение новых результирующих признаков в результате математических преобразований, релевантных доменной области, консультации экспертов предметной области, новые исследования, работа эффективной команды, состоящей из различных учёных.

В целом прогнозирование конечных свойств/характеристик композитных материалов без изучения материаловедения, погружения в вопрос экспериментального анализа характеристик композитных материалов не демонстрирует сколько-нибудь удовлетворительных результатов. Проработка моделей и построение прогнозов требует внедрения в процесс производных от имеющихся показателей для выявления иного уровня взаимосвязей. Отсюда, также учитывая отсутствие корреляции между признаками, делаем вывод, что текущим набором алгоритмов задача не решается, возможно, решается трудно или не решается совсем.

1. СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ И РЕСУРСЫ
2. Бизли Д. Python. Подробный справочник: учебное пособие. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.
3. 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.
4. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.
5. Документация по библиотеке keras: – Режим доступа: [https://keras.io/api/](https://keras.io/api/" \t "_blank).(дата обращения: 08.06.2022).
6. Документация по библиотеке matplotlib: – Режим доступа: [https://matplotlib.org/stable/users/index.html](https://matplotlib.org/stable/users/index.html" \t "_blank).
7. Документация по библиотеке numpy: – Режим доступа: [https://numpy.org/doc/1.22/user/index.html#user](https://numpy.org/doc/1.22/user/index.html" \l "user" \t "_blank).
8. Документация по библиотеке pandas: – Режим доступа: [https://pandas.pydata.org/docs/user\_guide/index.html#user-guide](https://pandas.pydata.org/docs/user_guide/index.html" \l "user-guide" \t "_blank).
9. Документация по библиотеке scikit-learn: – Режим доступа: [https://scikit-learn.org/stable/user\_guide.html](https://scikit-learn.org/stable/user_guide.html" \t "_blank).
10. Документация по библиотеке Tensorflow: – Режим доступа: <https://www.tensorflow.org/overview>
11. Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.
12. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) – Режим доступа: <https://habr.com/ru/post/428503/>
13. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с.
14. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
15. Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.