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**

«DataScience»

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

**(композиционных материалов)»**

Слушатель Зуева Наталия Витальевна

# 

Москва, 2022

# **Содержание**

[Содержание 2](#_Toc106232838)

[Введение 3](#_Toc106232839)

[1. Аналитическая часть 5](#_Toc106232840)

[1.1. Постановка задачи 5](#_Toc106232841)

[1.2. Описание используемых методов 6](#_Toc106232842)

[1.3. Разведочный анализ данных 18](#_Toc106232843)

[2. Практическая часть 23](#_Toc106232844)

[2.1. Предобработка данных 23](#_Toc106232845)

[2.2. Разработка и обучение модели 23](#_Toc106232846)

[2.3. Тестирование модели 26](#_Toc106232847)

[2.4. Написать нейронную сеть, которая будет рекомендовать соотношение «матрица-наполнитель» 27](#_Toc106232848)

[2.5. Разработка приложения 30](#_Toc106232849)

[2.6. Создание удалённого репозитория и загрузка 32](#_Toc106232850)

[2.7. Заключение 33](#_Toc106232851)

[2.8. Список используемой литературы и веб ресурсы. 34](#_Toc106232852)

[2.9. Приложение 1 37](#_Toc106232853)

# **Введение**

Композиционные материалы — это материалы, состоящие из двух или более компонентов, нерастворимых друг с другом, с чётко обозначенной границей раздела и сильным взаимодействием по всей зоне контакта. Одним из компонентов композитных материалов является непрерывная фаза, он называется матрица, в которой нерастворимые материалы помещаются в другую природу, называемую арматурой или наполнителем.

Внедрение композиционных материалов обусловлено стремлением использовать их преимущества по сравнению с традиционно используемыми металлами и сплавами. Примеры композита – железобетон (сочетание стали арматуры и камня бетона), древесноволокнистая плита ДВП (сочетание древесной основы – щепы и полимерного связующего).

Базальт - магматическая вулканическая порода. Это самая распространённая порода на поверхности Земли и на других планетах Солнечной системы. Базальты образуются путём затвердевания силикатного магматического расплава. Большая часть базальтов образуется на срединно-океанических хребтах и образует океаническую кору. Активно развивается использование композитных материалов на основе базальта.

Базальтопластик - современный композитный материал на основе базальтовых волокон и органического связующего вещества. В настоящее время базальтопластик успешно конкурирует с металлическими изделиями, превосходя их по коррозионной, щелочной, кислотоустойчивости и некоторым другим свойствам. Целью данной работы является прогнозирование конечных свойств новых материалов на основе базальтопластика (композиционных материалов).

Расширение разнообразия материалов, используемых при проектировании нового композиционного материала, увеличивает необходимость определения свойств нового композита при минимальных финансовых затратах. Для решения этой проблемы обычно используются два способа: физические тесты образцов материалов или оценка свойств, в том числе на основе физико-математических моделей. Традиционно разработка композитных материалов является долгосрочным процессом, так как из свойств отдельных компонентов невозможно рассчитать конечные свойства композита. Для достижения определенных характеристик требуется большое количество различных комбинированных тестов, что делает насущной задачу прогнозирования успешного решения, снижающего затраты на разработку новых материалов и затраты на рабочую силу. Суть прогнозирования заключается в моделировании репрезентативного элемента композитного объёма на основе данных о свойствах входящих компонентов (связующего и армирующего компонента). В процессе исследовательской работы были разработаны несколько моделей, способные с высокой вероятностью прогнозировать модули упругости при растяжении и прочности при растяжении, а также была создана нейронная сеть, которая предлагает соотношение «матрицы - наполнитель». На основе нейронной сети было создано пользовательское веб - приложение на фреймворке Flask.

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

Для исследовательской работы были даны 2 файла: X\_bp.xlsx (с данными о параметрах базальтопластика, состоящий из 1024 строки и 11 столбцов) и X\_nup.xlsx (данными нашивок углепластика, состоящий из 1041 строки и 4 столбцов).

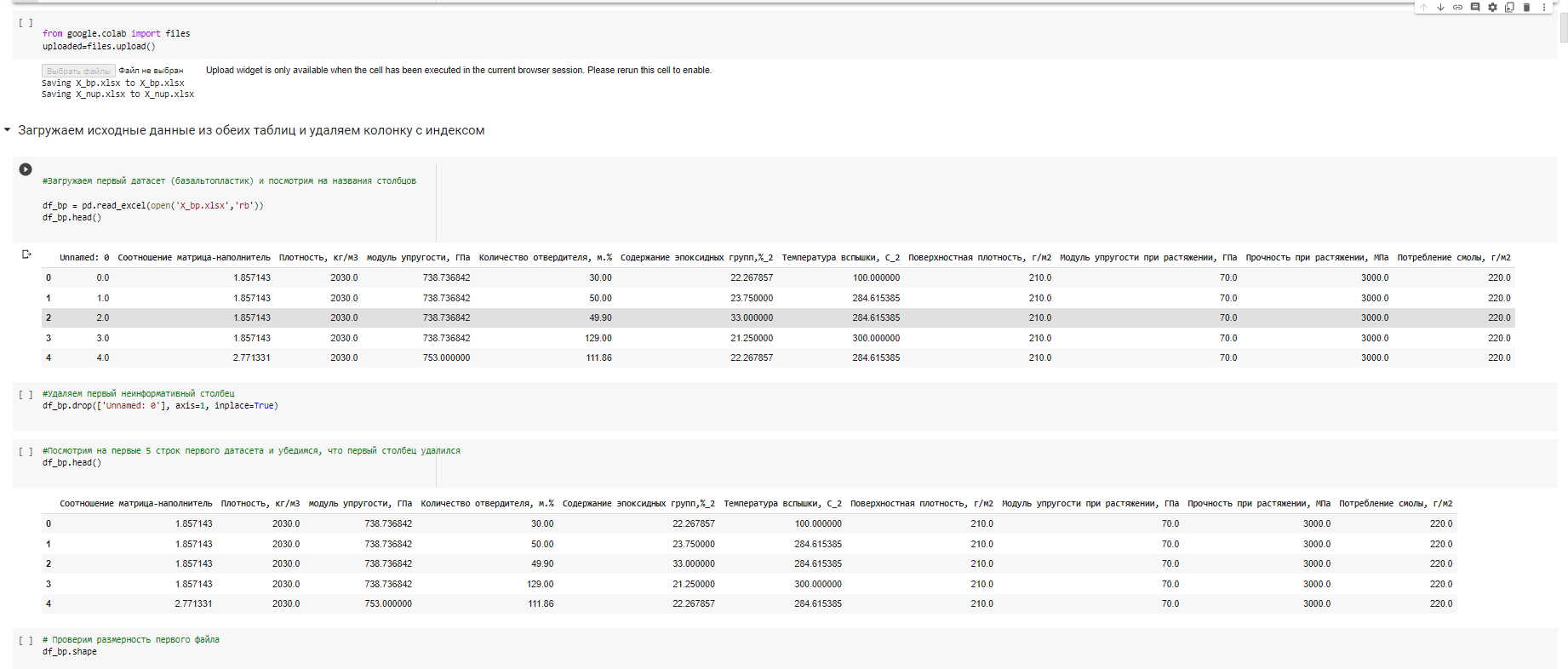


Рисунок 1– пример начала работы с файлом X\_bp.xlsx

Цель работы разработать модели для прогноза модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель». Для этого нужно объединить 2 файла. Часть информации (17 строк таблицы способов компоновки композитов) не имеют соответствующих строк в таблице соотношений и свойств используемых компонентов композитов, поэтому были удалены.

Затем провести разведочный анализ данных, нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек.

Для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; предобработать 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**

Данная задача в рамках классификации категорий машинного обучения относится к машинному обучению с учителем и традиционно это задача 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-ближайшихсоседей;
* дерево решений;
* стохастический градиентный спуск;
* многослойный перцептрон;
* Лассо;

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

Модель метода опорных векторов – отображение данных точками в пространстве, так что между наблюдениями отдельных категорий имеется разрыв, и он максимален.

Каждый объект данных представляется как вектор (точка) в p-мерном пространстве. Он создаёт линию или гиперплоскость, которая разделяет данные на классы.

Достоинства метода: для классификации достаточно небольшого набора данных. При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных. Эффективен при большом количестве гиперпараметров. Способен обрабатывать случаи, когда гиперпараметров больше, чем количество наблюдений. Существует возможность гибко настраивать разделяющую функцию.  Алгоритм максимизирует разделяющую полосу, которая, как подушка безопасности, позволяет уменьшить количество ошибок классификации.

Недостатки метода: неустойчивость к шуму, поэтому в работе была проведена тщательнейшая работа с выбросами, иначе в обучающих данных шумы становятся опорными объектами-нарушителями и напрямую влияют 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.

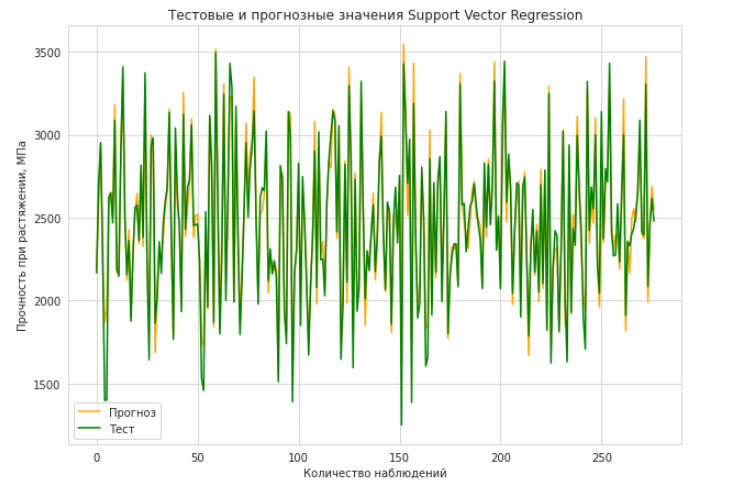


Рисунок 2- график метода опорных векторов для прочности при растяжении, МПа

Случайный лес(RandomForest)— это множество решающих деревьев. Универсальный алгоритм машинного обучения с учителем, представитель ансамблевых методов.  Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив.

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 параллелизуемость и масштабируемость.

Недостатки метода: построение занимает много времени; сложно интерпретируемый; не обладает возможностью экстраполяции; может недо обучаться; трудоёмко прогнозируемый; иногда работает хуже, чем линейные методы.

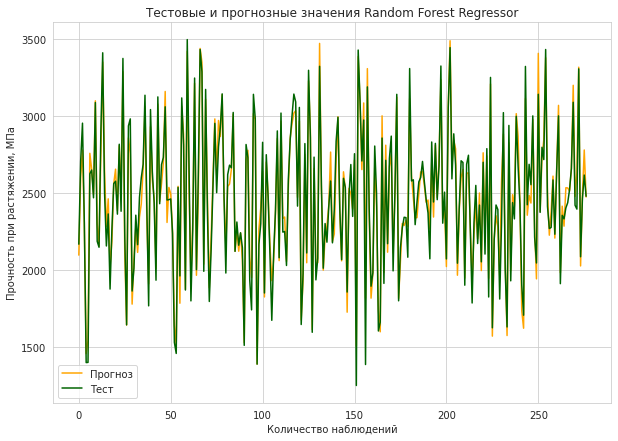


Рисунок 3 - график "случайного леса" для прочности при растяжении, МПа

Линейная регрессия (Linearregression) — 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 R-квадрат равен 1, это значит, что модель описывает все данные. Если же R-квадрат 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.

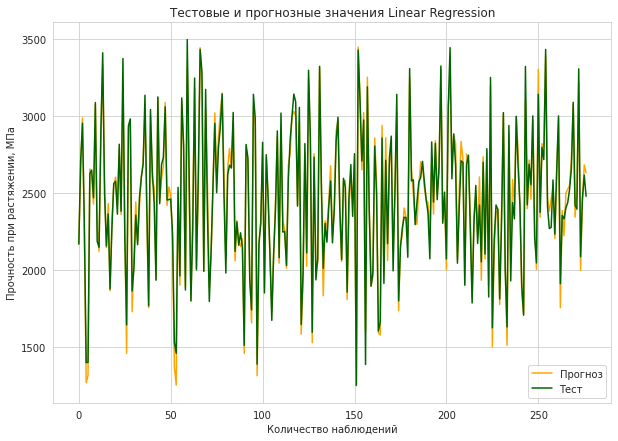


Рисунок 4 график линейной регрессии для прочности при растяжении, МПа

Градиентный бустинг (GradientBoosting)— 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.

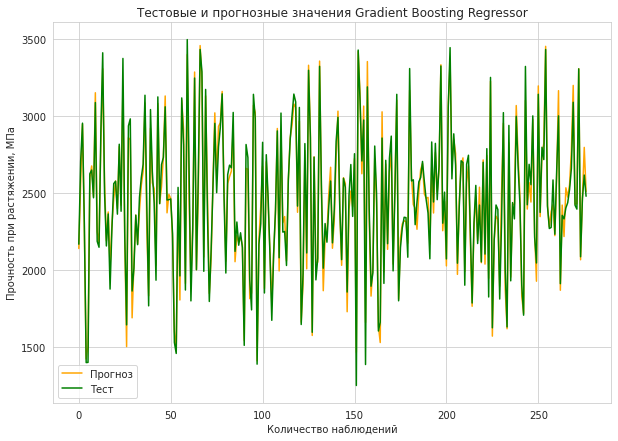


Рисунок 5 - график градиентного бустинга для прочности при растяжении, МП

Метод ближайших соседей - К-ближайших соседей (kNN - kNearestNeighbours) 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 реализациии 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 создаёт правил; не обобщает предыдущий опыт; основывается на всем массиве доступных исторических данных; невозможно сказать, на каком основании строятся ответы; сложно выбрать близость метрики; имеет высокую зависимость результатов классификации от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоёмкость.

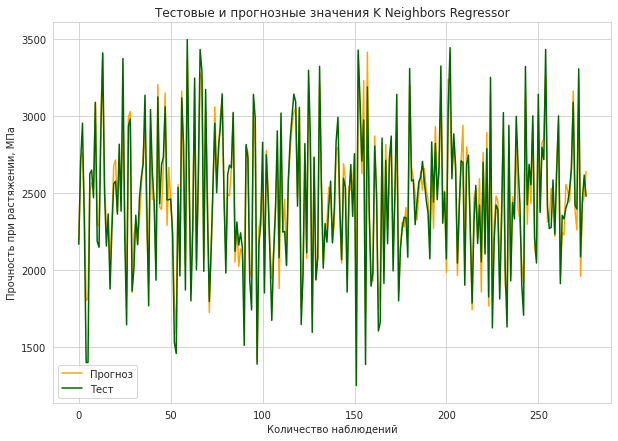


Рисунок 6 - график К-ближайших соседей для прочности при растяжении, Мпа

Дерево принятия 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 принятия решений- эффективный инструмент интеллектуального анализа данных и предсказательной аналитики. Алгоритм дерева решений подпадает под категорию контролируемых алгоритмов обучения. Он работает как для непрерывных, так и для категориальных выходных переменных. Правила генерируются за счёт обобщения множества отдельных наблюдений (обучающих примеров), описывающих предметную область.  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: помогают визуализировать процесс принятия решения и сделать правильный выбор в ситуациях, когда результаты одного решения влияют на результаты следующих решений; создаются по понятным правилам; просты в применении и интерпретации; заполняют пропуски в данных наиболее вероятным решением; работают с разными переменными; 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.

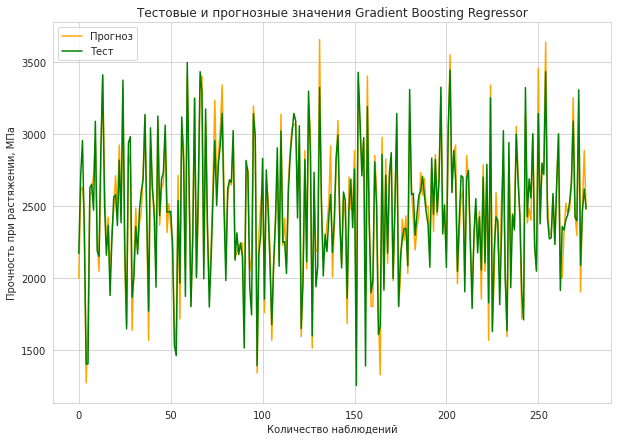


Рисунок 7- график дерева принятия решений для прочности при растяжении

Стохастический градиентный спуск (SGDRegressor) — это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под выпуклые функции потерь. Этот подход подразумевает корректировку весов нейронной сети, используя аппроксимацию градиента функционала, вычисленную только на одном случайном обучающем примере из выборки.

Достоинства метода: эффективен; прост в реализации; имеет множество возможностей для настройки кода; способен обучаться на избыточно больших выборках.

Недостатки метода: требует ряд гиперпараметров; чувствителен к масштабированию функций; может не сходиться или сходиться слишком медленно; функционал многоэкстремален; процесс может "застрять" в одном из локальных минимумов; возможно переобучение.

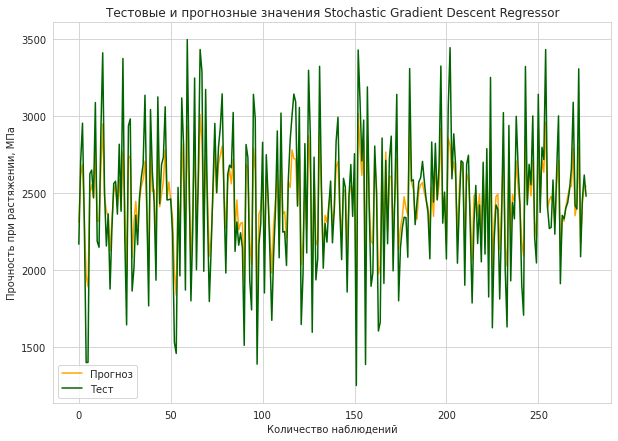


Рисунок 8-график стохастического градиентного спуска для модуля упругости

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: имеет невыпуклую функцию потерь; разные инициализации случайных весов могут привести к разной точности проверки; требует настройки ряда гиперпараметров; чувствителен к масштабированию функций.

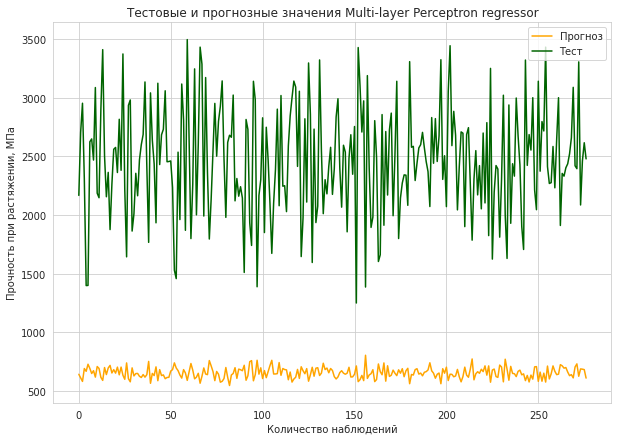


Рисунок 9 -график многослойного персептрона для модуля упругости, ГПа

Лассо регрессия (Lasso)— это линейная модель, которая оценивает разреженные коэффициенты.  Это простой метод, позволяющий уменьшить сложность модели и предотвратить переопределение, которое может возникнуть в результате простой линейной регрессии. Данный метод вводит дополнительное слагаемое регуляризации в оптимизацию модели. Это даёт более устойчивое решение. В регрессии лассо 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, таким как авторегрессии.

Достоинства метода: легко полностью избавляется от шумов в данных; быстро работает; не очень энергоёмко; способно полностью убрать признак из датасета; доступно обнуляет значения коэффициентов.

Недостатки метода: выбор модели не помогает и обычно вредит; часто страдает качество прогнозирования; выдаёт ложное срабатывание результата; случайным образом выбирает одну из коллинеарных переменных; не оценивает правильность формы взаимосвязи между независимой и зависимой переменными; не всегда лучше, чем пошаговая регрессия.

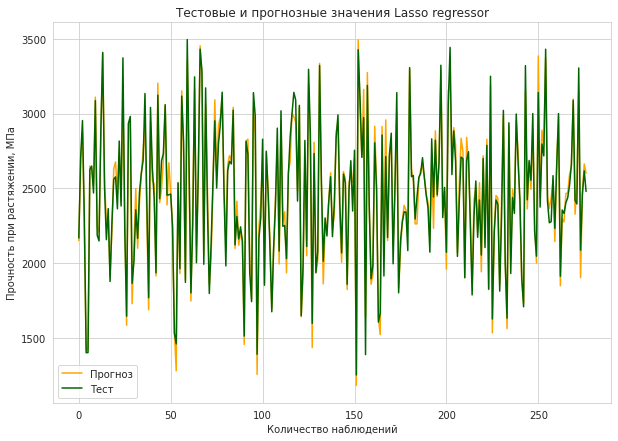


Рисунок 10-график метода Лассо для модуля упругости, Гпа

Немного расскажем об используемых метриках 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.

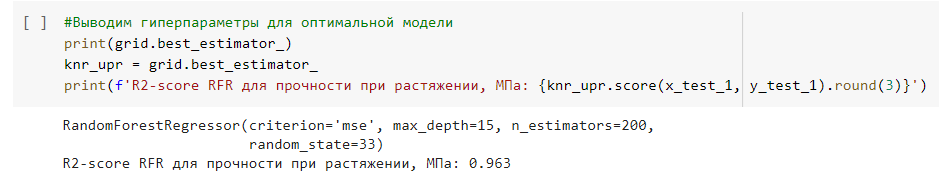


Рисунок 11 – часть кода для результата метрики R2 для метода «Случайный лес»

MAE показывает среднее абсолютное отклонение предсказанных значений от реальных. Чем выше значение MAE, тем модель хуже. У идеальной модели MAE=0

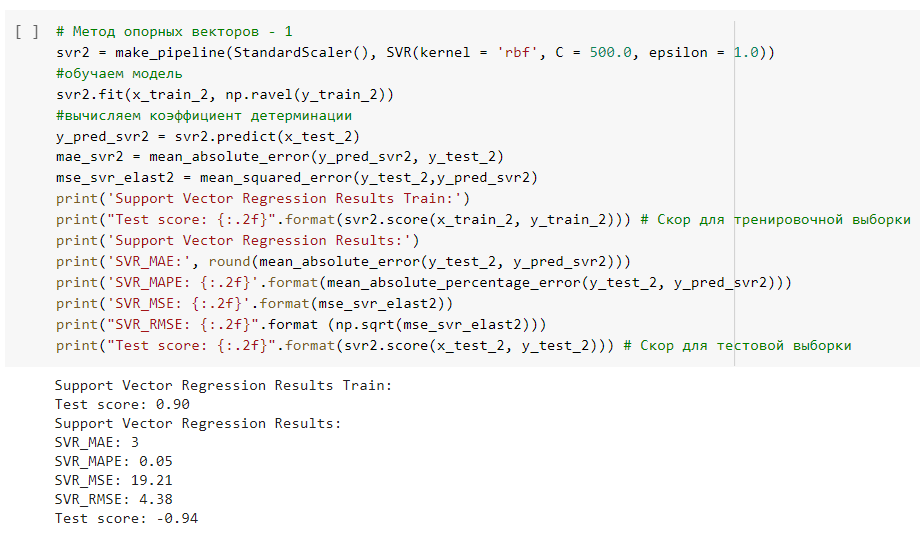


Рисунок - код для вывода различных метрик для метода опорных векторов

* 1. **Разведочный анализ данных**

Прежде чем передать данные в работу моделей машинного обучения, необходимо обработать и очистить их. Очевидно, что «грязные» и необработанные данные могут содержать искажения и пропущенные значения – это ненадёжно, поскольку способно привести к крайне неверным результатам по итогам моделирования. Но безосновательно удалять что-либо тоже неправильно. Именно поэтому сначала набор данных надо изучить.

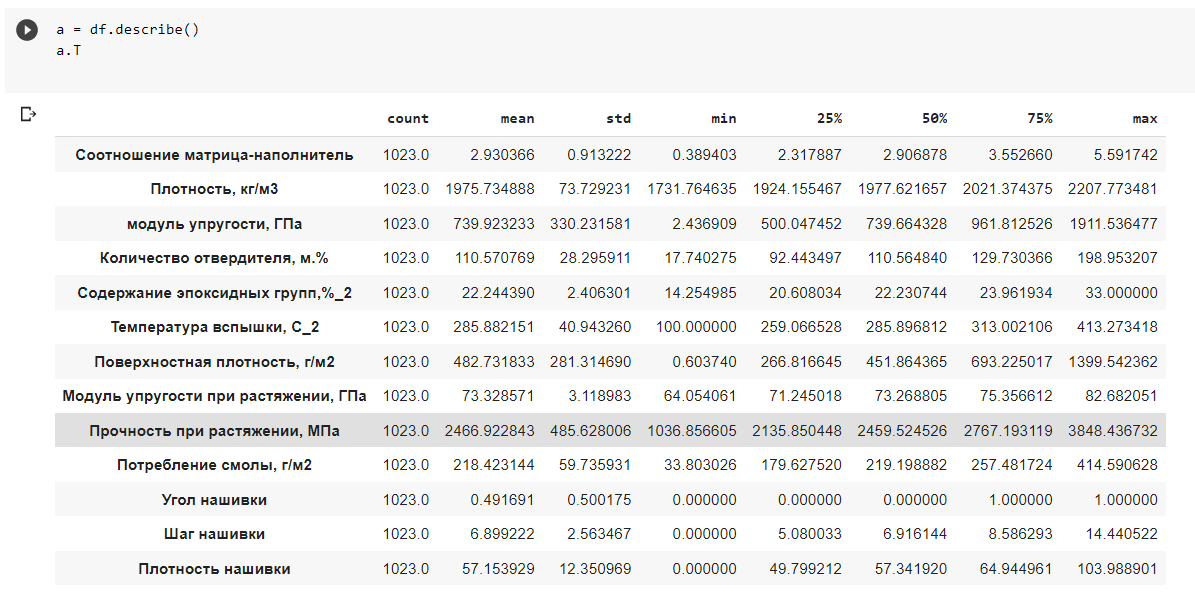


Рисунок 13 - описательная статистика датасета

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

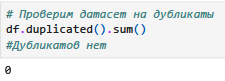


Рисунок 14 - проверка датасета на наличие дубликатов

В качестве инструментов разведочного анализа используется: оценка статистических характеристик датасета; гистограммы распределения каждой из переменной; диаграммы ящика с усами; попарные графики рассеяния точек;тепловая карта; описательная статистика для каждой переменной; анализ и полное исключение выбросов; проверка наличия пропусков и дубликатов.

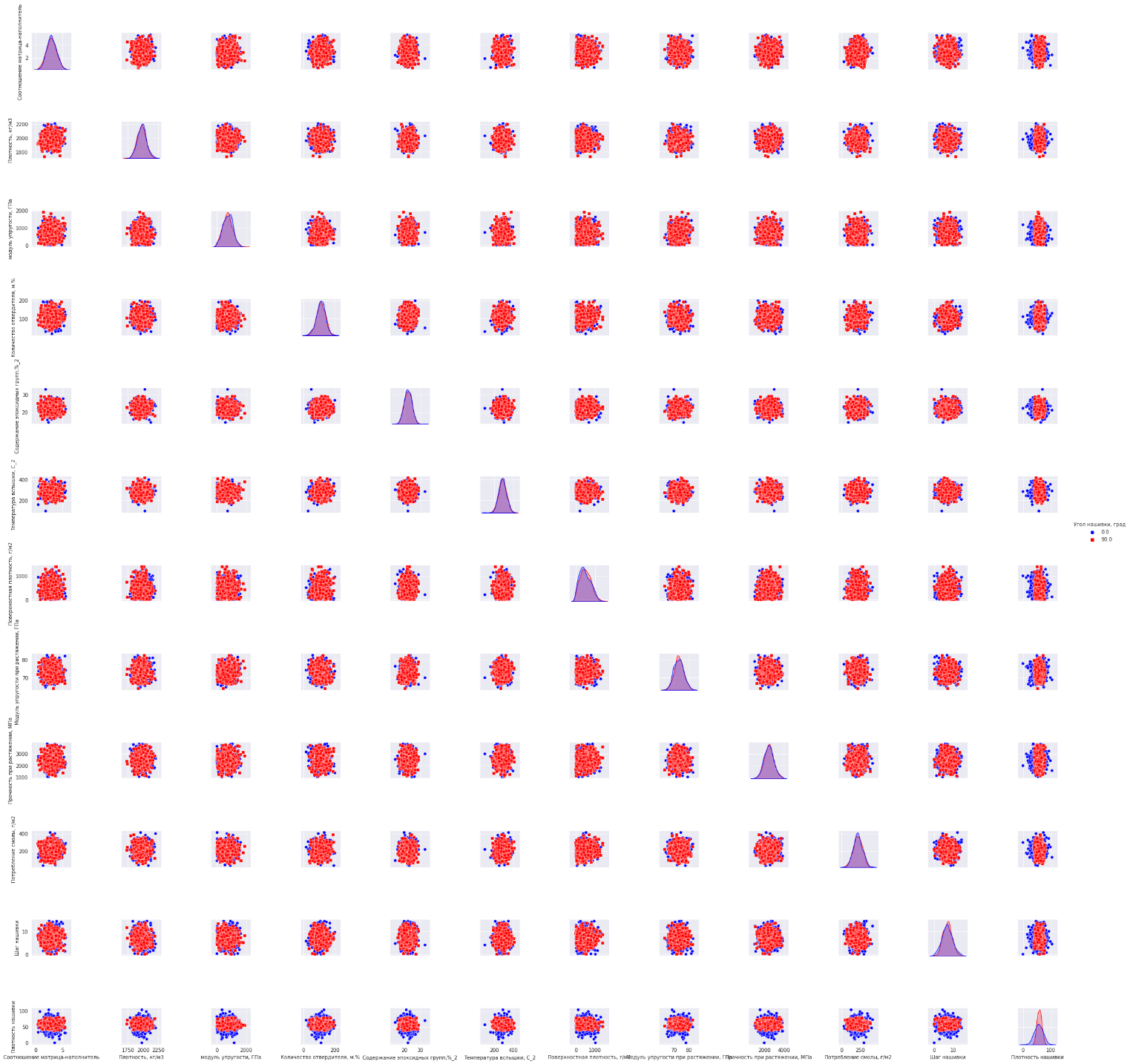


Рисунок 15 - попарные графики рассеяния точек

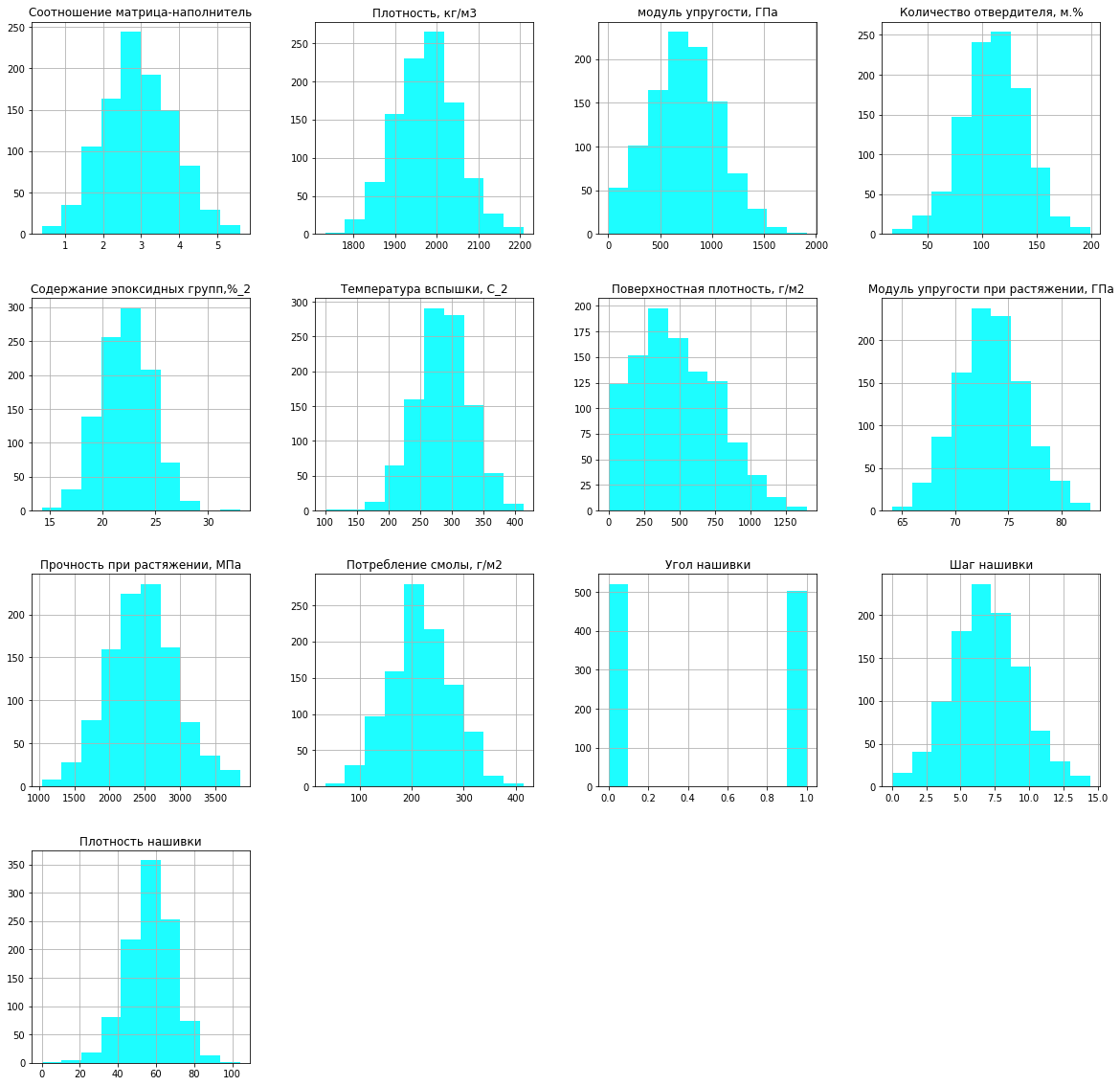


Рисунок 16 - гистограммы распределения

Гистограммы используются для изучения распределений частот значений переменных. Мы видим очень слабую корреляцию между переменными.

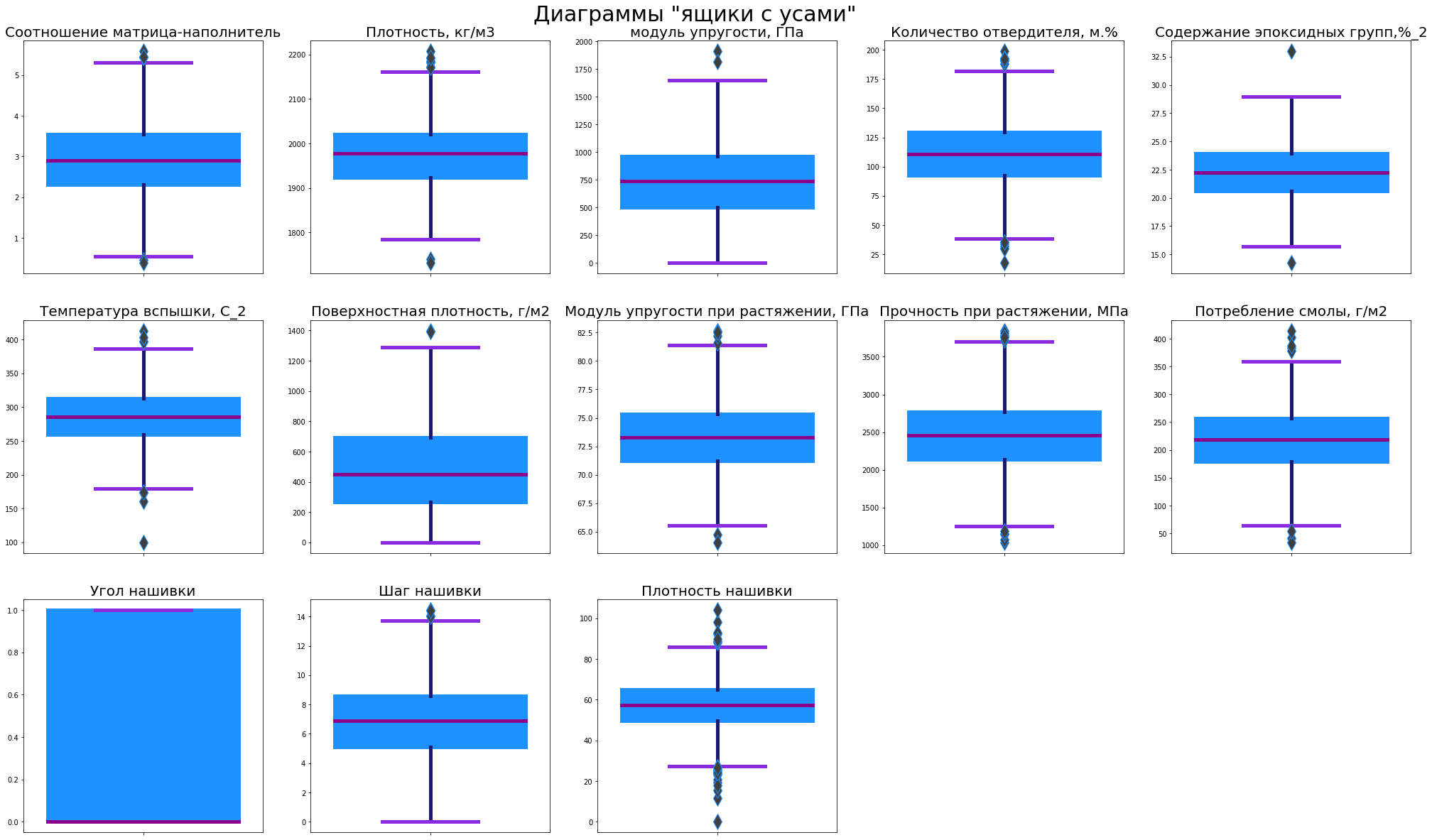


Рисунок 17 - график "ящиков с усами" для всех переменных

После обнаружения выбросов данные, значительно отличающиеся от выборки, будут полностью удалены. Для расчёта этих данных мы будем использовать методы трех сигм и межквартильного расстояния.

Данные объединённого датасета не имеют чётко выраженной зависимости, что подтверждает тепловая карта с матрицей корреляции и матрицы диаграмм рассеяния.

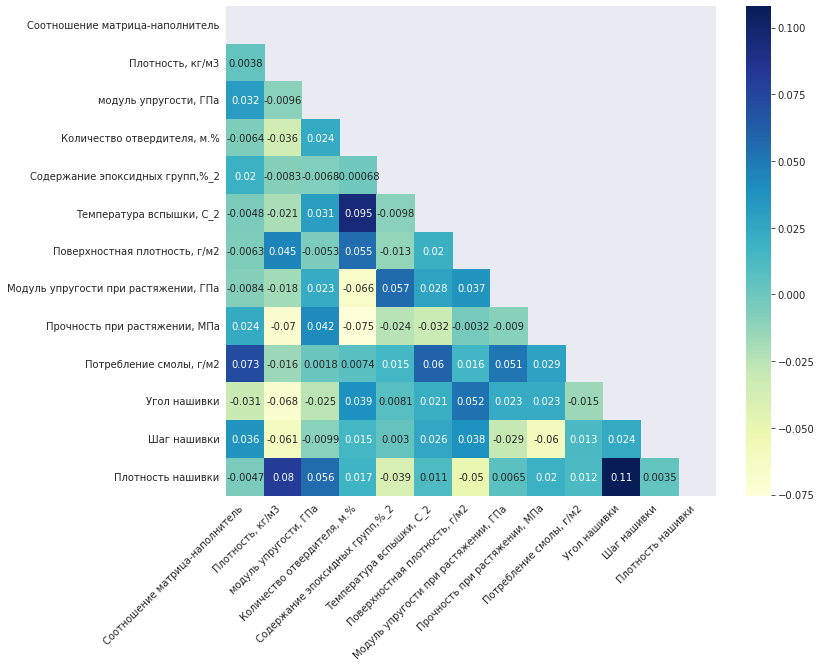


Рисунок 18 - тепловая карта с корреляцией данных

Максимальная корреляция между плотностью нашивки и углом нашивки 0.11, значит нет зависимости между этими данными. Корреляция между всеми параметрами очень близка к 0, корреляционные связи между переменными не наблюдаются.

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

По условиям задания нормализуем значения. Для этого применим MinMaxScaler(), затем применимNormalizer(). Второе даёт нам больше выбросов.

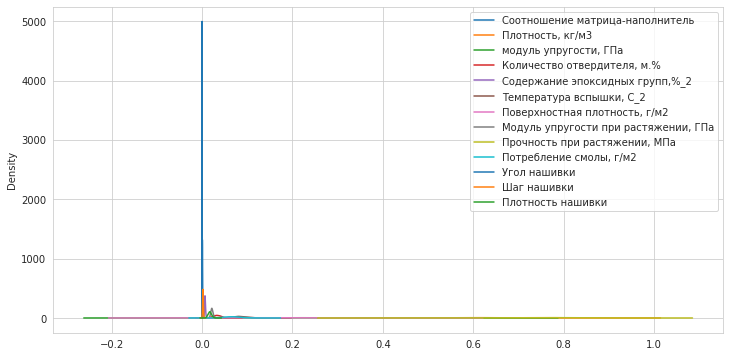
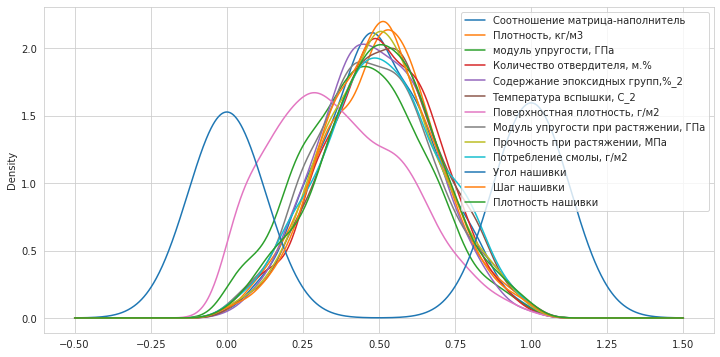


Рисунок 19 - визуализированные данные до и после нормализации

* 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 применим все методы, описанные выше.

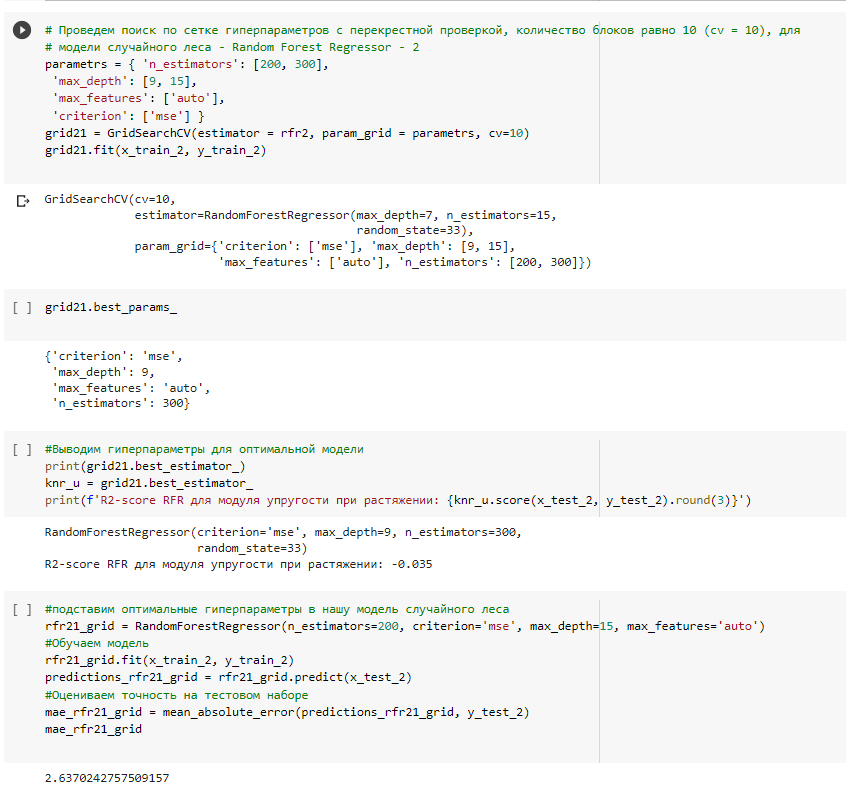


Рисунок 20- поиск гиперпараметров

Порядок разработки модели для каждого параметра и для каждого выбранного метода 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 сетки 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; подстановка оптимальных гиперпараметров в модель и обучение модели на тренировочных данных; оценка полученных данных; сравнение со стандартными значениями.

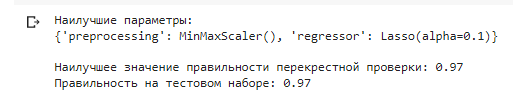


Рисунок 21 - наилучшие гиперпараметры

Модель после настройки гиперпараметров показала результат немного лучше. Однако, ниже, чем базовая модель. Прочность при растяжении и модуль упругости не имеет линейной зависимости. Все использованные модели не справились с задачей. Результат неудовлетворительный. Свойства композитных материалов в первую очередь зависят от используемых материалов.

Таблица 1. Результаты построения и обучения моделей

|  | **Model** | **MAE** | **R2 score** |
| --- | --- | --- | --- |
| **Модуль упругости при растяжении** | KNeighborsRegressor | 2.751513 | -0.018 |
| **Прочность при растяжении** | KNeighborsRegressor | 98.875972 | 0.923 |
| **Модуль упругости при растяжении** | DecisionTree | 2.614842 | -0.009 |
| **Прочность при растяжении** | DecisionTree | 168.618756 | 0.779 |
| **Модуль упругости при растяжении** | RandomForestRegressor | 2.616907 | -0.034 |
| **Прочность при растяжении** | RandomForestRegressor | 67.718326 | 0.964 |

* 1. **Тестирование модели**

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

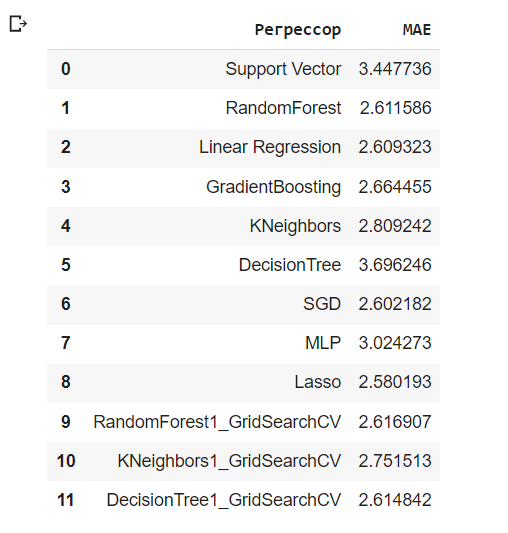


Рисунок 23 - результат оценки точности по MAE для модуля

Хотя в целом при таких результатах можно применять среднее значение переменной в качестве прогнозного.

* 1. **Написать нейронную сеть, которая будет рекомендовать**

**соотношение «матрица – наполнитель».**

Обучение нейронной сети — это такой процесс, при котором происходит подбор оптимальных параметров модели, с точки зрения минимизации функционала ошибки. Создадим простую модель глубокого обучения. Обучим её, посмотрим на потери, оценим MSE, построим график.

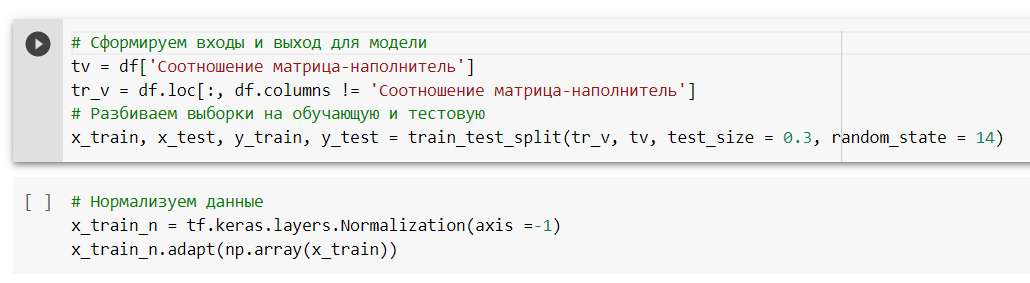




Рисунок 24 - создание нейронной сети

Определим параметры, поищем оптимальные параметры, посмотрим на результаты. Выйдем на наилучшие параметры для нашей нейронной сети и построим окончательную нейросеть.

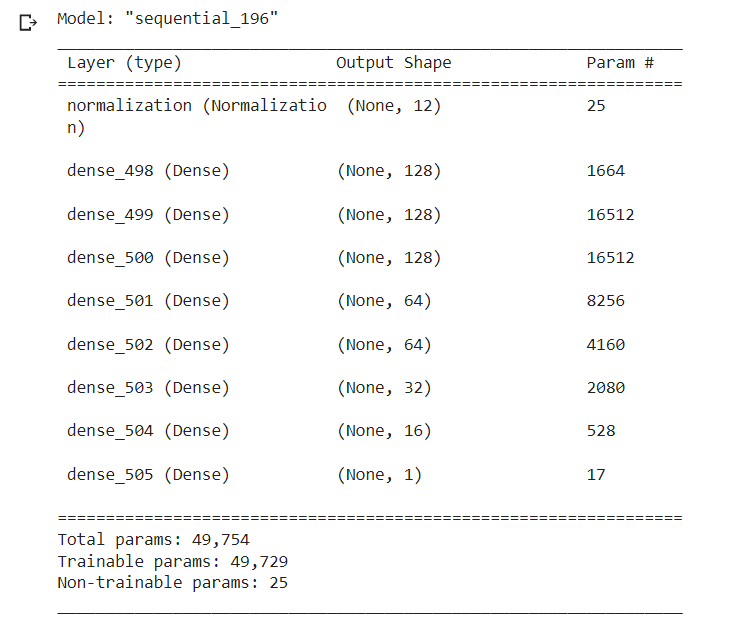


Рисунок 25 - построение нейросети

Обучим и оценим модель, посмотрим на потери, зададим функцию для визуализации факт/прогноз для результатов моделей.

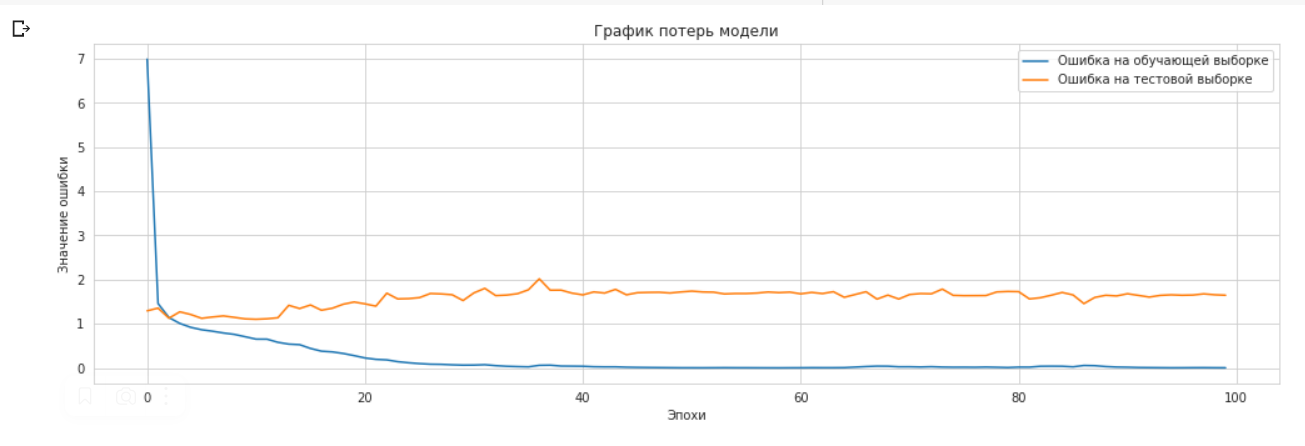


Рисунок 26 - график потерь модели



Рисунок 27 - тестовые и прогнозные значения модели

* 1. **Разработка приложения**

Приложение разработано на основе приложения Visual Studio Code, как прогноз для соотношения «матрица – наполнитель».

Данное приложение — это основной файл Flask, папка templates, с шаблоном html -страницы, папка ВКР c сохранённой моделью для данных.

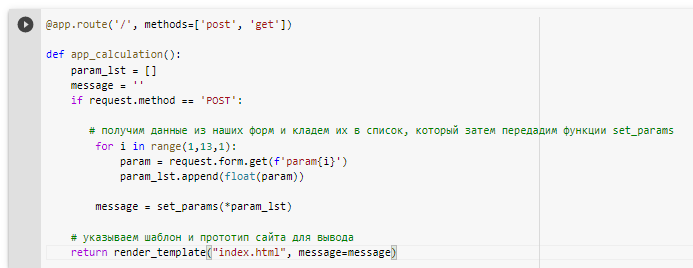


Рисунок 28 - часть кода приложения

При запуске приложения, пользователь переходит на :<http://127.0.0.1:5000/>.

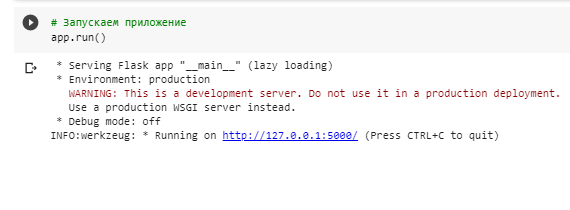


Рисунок 29 - ссылка для открытия html - файла

В открывшемся окне пользователю необходимо ввести в соответствующие ячейки требуемые значения и нажать на кнопку «Готово».

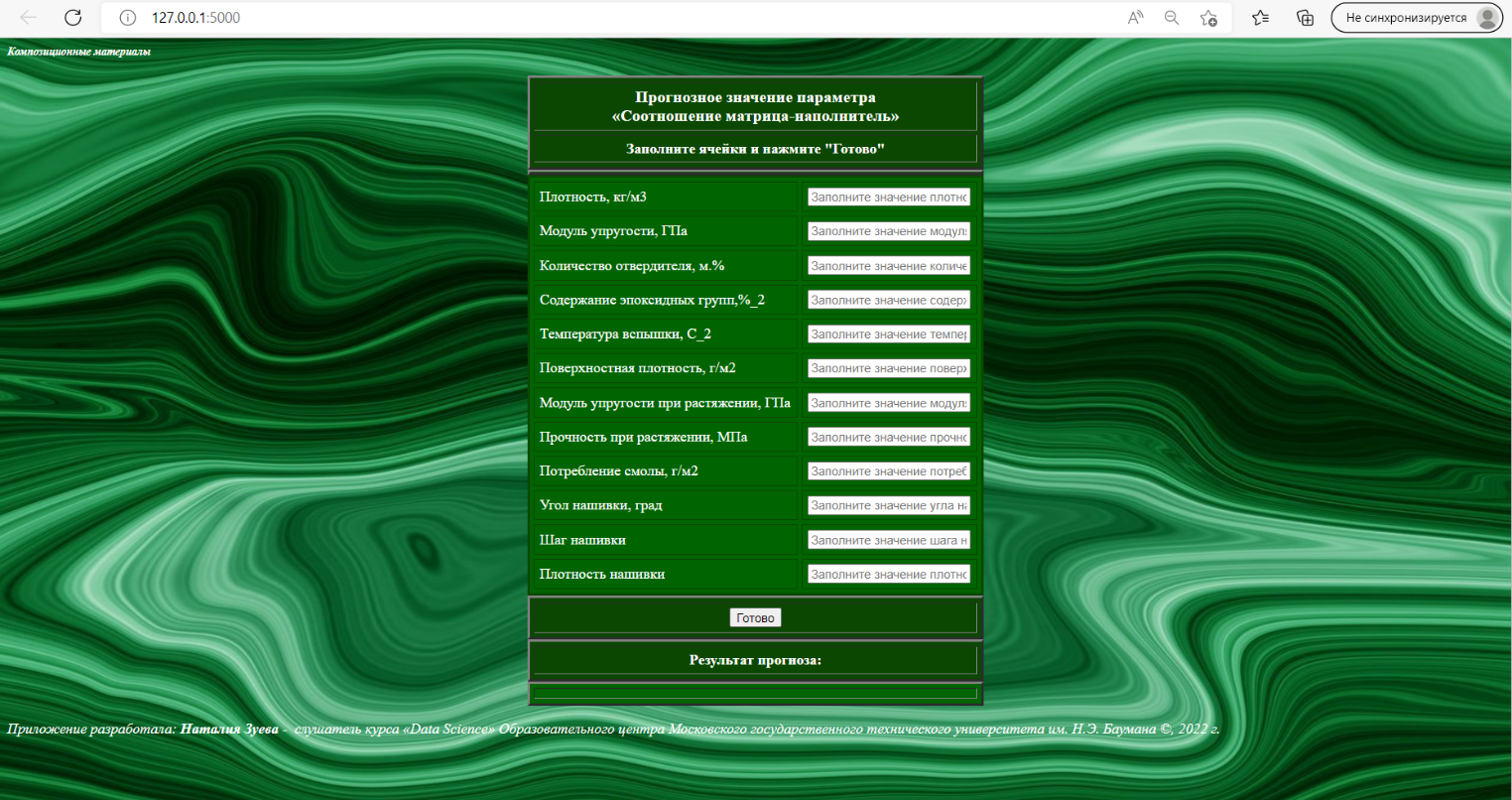


Рисунок 30- скриншот пользовательского приложения

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

* 1. **Создание удалённого репозитория и загрузка**

Репозиторий был создан на github.com по адресу:

<https://github.com/Nata7914>

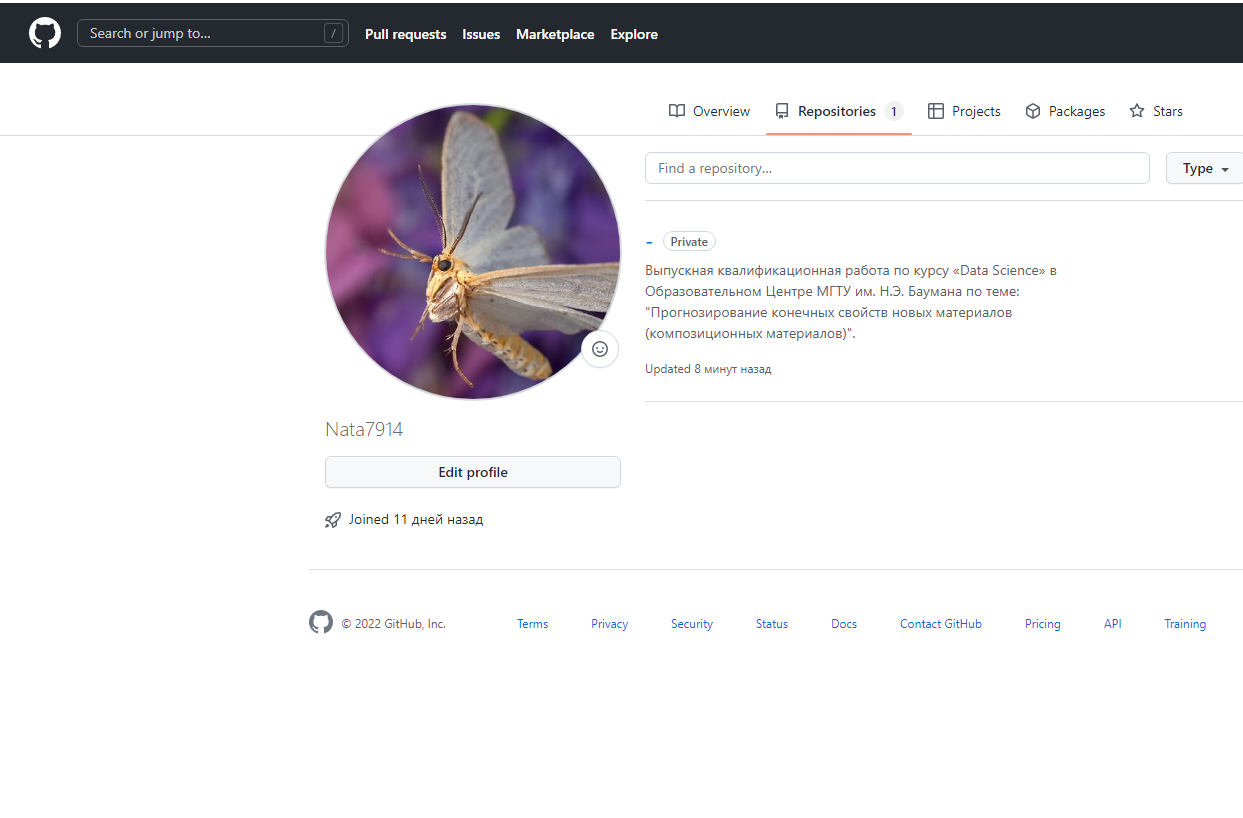


Рисунок 31 - часть страницы на github.com

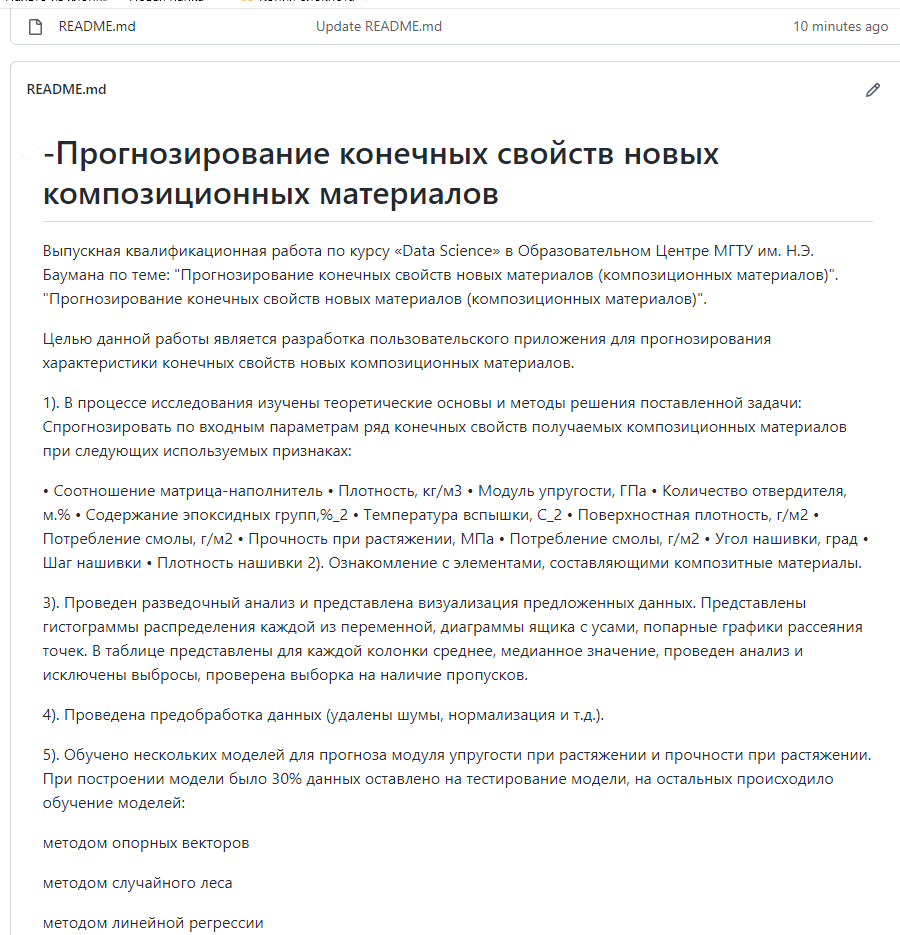


Рисунок 32 - часть созданного файла README

Ноутбук с решением и приложением так же можно найти по адресу: <https://colab.research.google.com/drive/1OT0pDZZO9dTnxxb0EBV5nQbCYet6CYXQ>

* 1. **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, ГПа –метод опорных векторов, для прочности при растяжении, МПа - лассо-регрессия.

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 веб ресурсы**.

1. AlexMaszański. Метод k-ближайших соседей (k-nearestneighbour): – Режим доступа: <https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19>.(дата обращения: 07.06.2022)
2. AndreYe. 5 алгоритмов регрессии в машинном обучении, о которых вам сле-дует знать: – Режим доступа:https://habr.com/ru/company/vk/blog/513842/(дата обращения: 01.06.2022).
3. DevpracticeTeam. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.
4. Абросимов Н.А.: Методика построения разрешающей системы уравнений динамического деформирования композитных элементов конструкций (Учебно-методическое пособие), ННГУ, 2010
5. Абу-Хасан Махмуд, Масленникова Л. Л.: Прогнозирование свойств композиционных материалов с учётом наноразмера частиц и акцепторных свойств катионов твёрдых фаз, статья 2006 год
6. Бизли Д. Python. Подробный справочник: учебное пособие. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.
7. 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.
8. ГрасД. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.
9. Документация по библиотеке keras: – Режим доступа: <https://keras.io/api/>.(дата обращения: 08.06.2022).
10. Документация по библиотеке matplotlib: – Режим доступа: <https://matplotlib.org/stable/users/index.html>.(дата обращения: 10.06.2022)
11. Документация по библиотеке numpy: – Режим доступа: <https://numpy.org/doc/1.22/user/index.html#user>. (дата обращения: 03.06.2022).
12. Документация по библиотеке pandas: – Режим доступа: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>.(дата обращения: 04.06.2022).
13. Документация по библиотеке scikit-learn: – Режим доступа: <https://scikit-learn.org/stable/user_guide.html>. (дата обращения: 05.06.2022).
14. Документация по библиотеке seaborn: – Режим доступа: <https://seaborn.pydata.org/tutorial.html>. (дата обращения: 06.06.2022).
15. Документация по библиотеке Tensorflow: – Режим доступа: <https://www.tensorflow.org/overview>(дата обращения: 10.06.2022).
16. Документация по языку программирования python: – Режим доступа:  <https://docs.python.org/3.8/index.html>.(дата обращения: 02.06.2022).
17. ИвановД.А., СитниковА.И., ШляпинС.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.
18. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) – Режим доступа: <https://habr.com/ru/post/428503/> (дата обращения 07.06.2022)
19. ЛаринА.А., Способы оценки работоспособности изделий из композиционных материалов методом компьютерной томографии, Москва, 2013, 148 с.
20. Материалы конференции: V Всероссийская научно-техническая конференция «Полимерные композиционные материалы и производственные технологии нового поколения», 19 ноября 2021 г.
21. МироновА.А. Машинное обучение часть I ст.9 – Режим доступа:<http://is.ifmo.ru/verification/machine-learning-mironov.pdf>. (дата обращения 08.06.2022)
22. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с.
23. Реутов Ю.А.: Прогнозирование свойств полимерных композиционных материалов и оценка надёжности изделий из них, Диссертация на соискание учёной степени кандидата физико-математических наук, Томск 2016.
24. Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.: Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил.
25. Руководство по быстрому старту в flask: – Режим доступа: <https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html>.(дата обращения: 09.06.2022)
26. Силен Дэви, МейсманАрно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
27. Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.
28. Справочник по композиционным материалам: в 2 - х кн. Кн. 2 / Под ред. Дж. Любина; Пер. с англ. Ф. Б. Геллера, M. М. Гельмонта; Под ред. Б. Э. Геллера - М.: Машиностроение, 1988. - 488 с. : ил;
29. Траск Эндрю. Грокаем глубокое обучение. – СПб.: Питер, 2019. – 352 с.: ил.
30. Чун-Те Чен и Грейс Х. Гу. Машинное обучение для композитных материалов (март 2019г.) – Режим доступа: [https://www.cambridge.org/core/journals/mrs-communications/article/machine- learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15](https://www.cambridge.org/core/journals/mrs-communications/article/machine-%20learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15). (дата обращения 02.06.2022)
    1. **Приложение 1**

Подробный план работы:

1. Загружаем и обрабатываем входящие датасеты
   1. Удаляем неинформативные столбцы
   2. Объединяем датасеты по методу INNER
2. Проводим разведочный анализ данных:
   1. Изучим описательную статистику каждой переменной - среднее, медиана, стандартное отклонение, минимум, максимум.
   2. Проверим датасет на пропуски и дубликаты данных
   3. Получим среднее, медианное значение для каждой колонки (по заданию необходимо получить их отдельно, поэтому продублируем их только отдельно)
3. Визуализируем наш разведочный анализ сырых данных (до выбросов и нормализации)
   1. Построим гистограммы распределения каждой переменной
   2. Построим диаграммы ящиков с усами каждой переменной
   3. Построим гистограмму распределения и диаграмма "ящик с усами" одновременно вместе с данными по каждому столбцу
   4. Построим попарные графики рассеяния точек (матрицы диаграмм рассеяния)
   5. Построим корреляционную матрицу с помощью тепловой карты
4. Проведём предобработку данных (в данном пункте только очистка датасета от выбросов)
   1. Проверим выбросы по 2 методам: 3-х сигм или межквартильных расстояний
   2. Посчитаем распределение выбросов по каждому столбцу (с целью предотвращения удаления особенностей признака или допущения ошибки)
   3. Исключим выбросы методом межквартильного расстояния
   4. Удалим строки c выбросами
   5. Визуализируем датасет без выбросов, и убедимся, что выбросы еще есть.
   6. Для полной очистки датасета от выбросов повторим пункты (4.3 – 4.5) ещё 3 раза.
   7. Сохраняем идеальный, без выбросов датасет
   8. Изучим чистые данные по всем параметрам
   9. Визуализируем «чистый» датасет (без выбросов)
5. Проведём нормализацию и стандартизацию (продолжим предобработку данных)
   1. Визуализируем плотность ядра
   2. Нормализуем данные с помощью MinMaxScaler()
   3. Нормализуем данные с помощью Normalizer()
   4. Сравним с данными до нормализации
   5. Проверим перевод данных из нормализованных в исходные
   6. Рассмотрим несколько вариантов корреляции между параметрами после нормализации
   7. Стандартизируем данные
   8. Визуализируем данные корреляции
   9. Посмотрим на описательную статистику после нормализации и после стандартизации
6. Разработаем и обучим нескольких моделей прогноза прочности при растяжении (с 30% тестовой выборки)
   1. Определим входы и выходы для моделей
   2. Разобьём данные на обучающую и тестовую выборки
   3. Проверим правильность разбивки
   4. Построим модели и найдём лучшие гиперпараметры (задача по заданию):
   5. Построим и визуализируем результат работы метода опорных векторов
   6. Построим и визуализируем результат работы метода случайного леса
   7. Построим и визуализируем результат работы линейной регрессии
   8. Построим и визуализируем результат работы метода градиентного бустинга
   9. Построим и визуализируем результат работы метода К ближайших соседей
   10. Построим и визуализируем результат работы метода деревья решений
   11. Построим и визуализируем результат работы стохастического градиентного спуска
   12. Построим и визуализируем результат работы многослойного перцептрона
   13. Построим и визуализируем результат работы лассо регрессии
   14. Сравним наши модели по метрике МАЕ
   15. Найдём лучшие гиперпараметры для случайного леса
   16. Подставим значения в нашу модель случайного леса
   17. Найдём лучшие гиперпараметры для К ближайших соседей
   18. Подставим значения в нашу модель К ближайших соседей
   19. Найдём лучшие гиперпараметры метода деревья решений
   20. Подставим значения в нашу модель метода деревья решений
   21. Проверим все модели и процессинги и выведем лучшую модель и процессинг
7. Разработаем и обучим нескольких моделей прогноза модуля упругости при растяжении (с 30% тестовой выборки)
   1. Определим входы и выходы для моделей
   2. Разобьём данные на обучающую и тестовую выборки
   3. Проверим правильность разбивки
   4. Построим модели и найдём лучшие гиперпараметры (задача по заданию):
   5. Построим и визуализируем результат работы метода опорных векторов
   6. Построим и визуализируем результат работы метода случайного леса
   7. Построим и визуализируем результат работы линейной регрессии
   8. Построим и визуализируем результат работы метода градиентного бустинга
   9. Построим и визуализируем результат работы метода К ближайших соседей
   10. Построим и визуализируем результат работы метода деревья решений
   11. Построим и визуализируем результат работы стохастического градиентного спуска
   12. Построим и визуализируем результат работы многослойного перцептрона
   13. Построим и визуализируем результат работы лассо регрессии
   14. Сравним наши модели по метрике МАЕ
   15. Найдём лучшие гиперпараметры для случайного леса
   16. Подставим значения в нашу модель случайного леса
   17. Найдём лучшие гиперпараметры для К ближайших соседей
   18. Подставим значения в нашу модель К ближайших соседей
   19. Найдём лучшие гиперпараметры метода деревья решений
   20. Подставим значения в нашу модель метода деревья решений
   21. Проверим все модели и процессинги и выведем лучшую модель и процессинг
8. Нейронная сеть для рекомендации соотношения матрица-наполнитель
   1. Сформируем входы и выход для модели
   2. Нормализуем данные
   3. Сконфигурируем модель, зададим слои
   4. Посмотрим на архитектуру модели
   5. Обучим модель
   6. Посмотрим на потери модели
   7. Посмотрим на график потерь на тренировочной и тестовой выборках
   8. Зададим функцию для визуализации факт/прогноз для результатов моделей
   9. Посмотрим на график результата работы модели
   10. Оценим модель MSE
   11. Сохраняем модель для разработки веб-приложения для прогнозирования соотношения "матрица-наполнитель" в фреймворкеFlask
9. Создаём приложение
   1. Импортируем необходимые бибилиотеки
   2. Загрузим модель и определим параметры функции
   3. Получим данные из наших форм и положим их в список
   4. Укажем шаблон и прототип сайта для вывода
   5. Запустим приложение
   6. Откроем <http://127.0.0.1:5000/>
10. Создание удалённого репозитория и загрузка результатов работы на него.

<https://github.com/Nata7914>

* 1. Создадим README (<https://github.com/Nata7914/-/blob/main/README.md>)
  2. Выгрузим все необходимые файлы в репозиторий