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)»

Москва, 2023

# Содержание

[Содержание 2](#_bookmark0)

[Введение 3](#_bookmark1)

1. [Аналитическая часть 5](#_bookmark2)
   1. [Постановка задачи 5](#_bookmark3)
   2. [Описание используемых методов 7](#_bookmark4)
   3. [Разведочный анализ данных 17](#_bookmark5)
2. [Практическая часть 21](#_bookmark6)
   1. [Предобработка данных 21](#_bookmark7)
   2. [Разработка и обучение модели 22](#_bookmark8)
   3. [Тестирование модели 24](#_bookmark9)
   4. [Написать нейронную сеть, которая будет рекомендовать соотношение](#_bookmark10)

[«матрица-наполнитель» 25](#_bookmark10)

* 1. [Разработка приложения 28](#_bookmark11)
  2. [Создание удалённого репозитория и загрузка 30](#_bookmark12)
  3. [Заключение 31](#_bookmark13)
  4. [Список используемой литературы и веб ресурсы 32](#_bookmark14)
  5. [Приложение 1 35](#_bookmark15)

### Введение

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

Внедрение композиционных материалов обусловлено стремлением использо- вать 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 океани- ческую кору. Активно развивается использование композитных материалов на ос- нове базальта.

Базальтопластик - современный композитный материал на основе базальто- вых волокон и органического связующего вещества. В настоящее время базальто- пластик успешно конкурирует с металлическими изделиями, превосходя их по кор- розионной, щелочной, кислотоустойчивости и некоторым другим свойствам. Це- лью 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 ма- териалов или оценка свойств, в том числе на основе физико-математических моде- лей. Традиционно разработка композитных материалов является долгосрочным процессом, так как из свойств отдельных компонентов невозможно рассчитать ко- нечные свойства композита. Для достижения определенных характеристик требу- ется большое количество различных комбинированных тестов, что делает насущ- ной задачу прогнозирования успешного решения, снижающего затраты на разра- ботку новых материалов и затраты на рабочую силу. Суть прогнозирования заклю- чается в моделировании репрезентативного элемента композитного объёма на ос- нове данных о свойствах входящих компонентов (связующего и армирующего ком- понента). В процессе исследовательской работы были разработаны несколько мо- делей, способные с высокой вероятностью прогнозировать модули упругости при растяжении и прочности при растяжении, а также были созданы 2 нейронных сети, которые предлагают соотношение «матрицы - наполнитель». На основе одной из нейронных сетей было создано дружелюбное и доступное пользовательское веб - приложение с высоким юзабилити на фреймворке Flask.

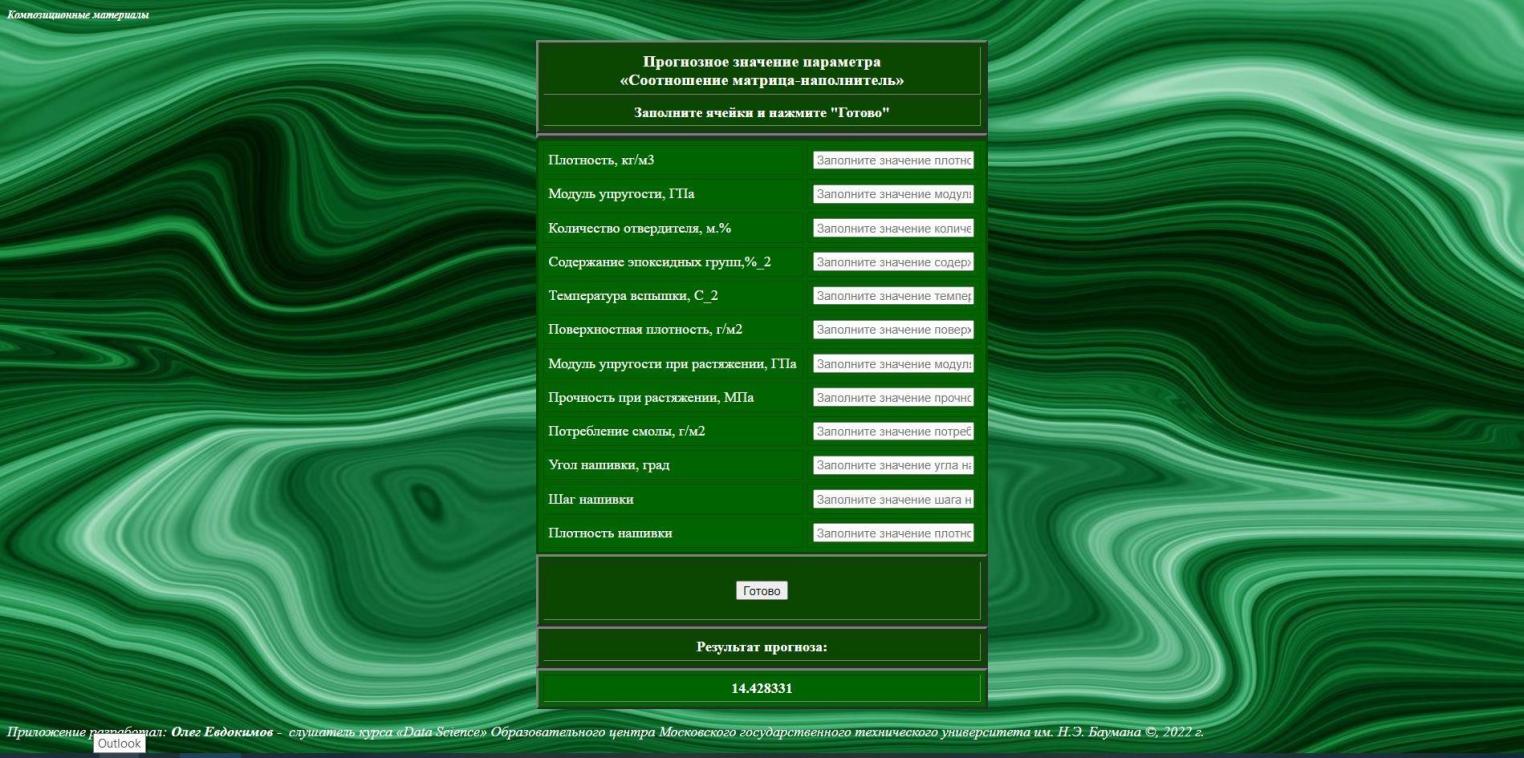


Рисунок 1 - скриншот работающего приложения на фреймворке Flask

## Аналитическая часть

### 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 столб- цов).

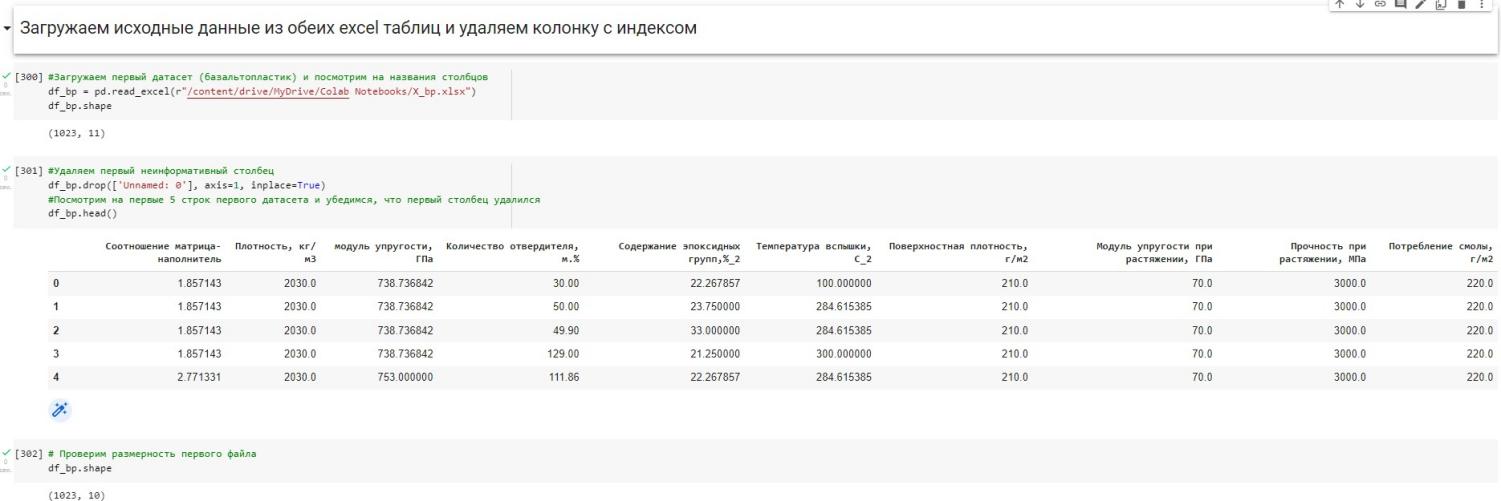


Рисунок 2 – пример начала работы с файлом X\_bp.xlsx

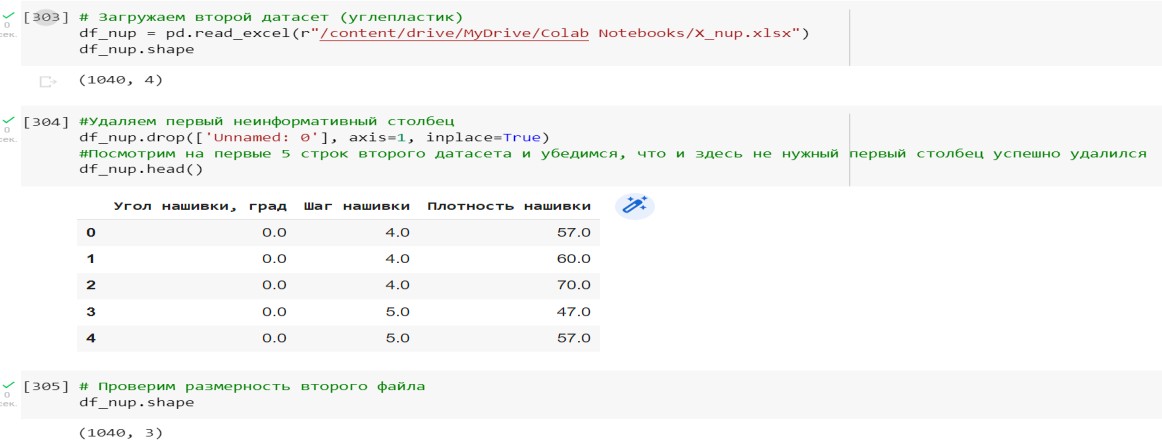
Цель работы разработать модели для прогноза модуля упругости при растя- жении, прочности при растяжении и соотношения «матрица-наполнитель». Для этого нужно объединить 2 файла. Часть информации (17 строк таблицы способов компоновки композитов) не имеют соответствующих строк в таблице соотношений и свойств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

Рисунок 3 - пример начала работы с файлом X\_nup.xlsx

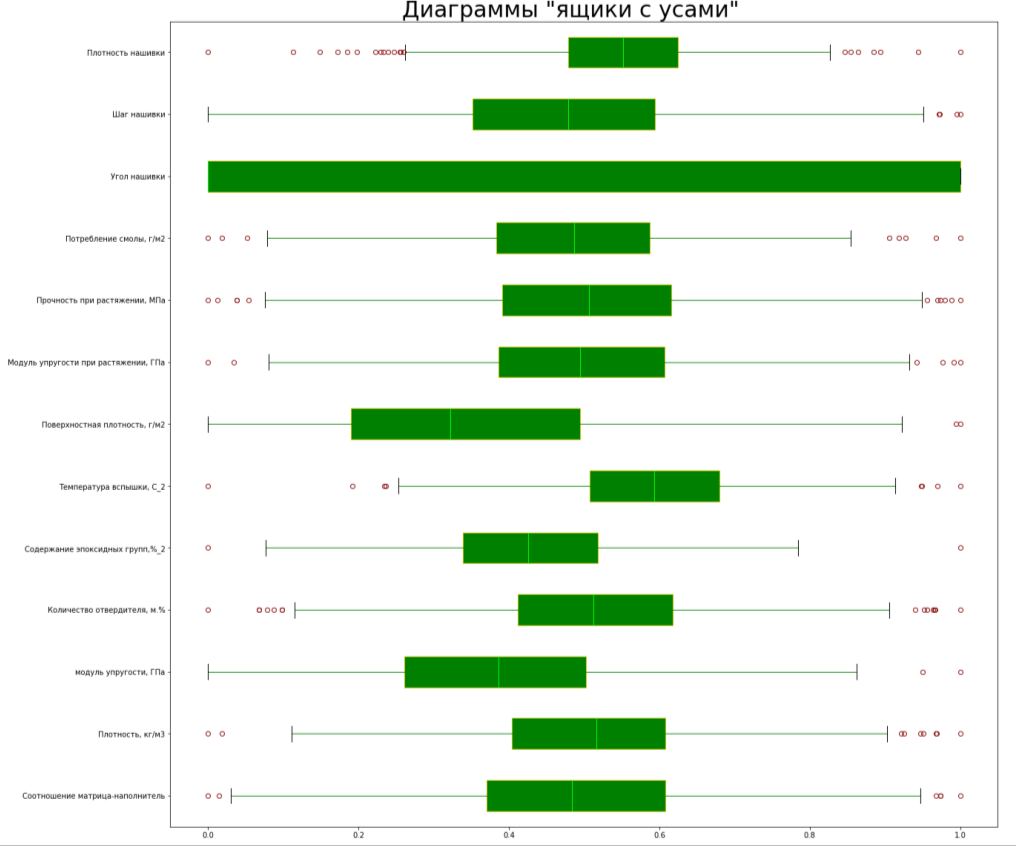
Затем провести разведочный анализ данных, нарисовать гистограммы распре- деления каждой из переменной, диаграммы ящика с усами, попарные графики рас- сеяния точек.

Рисунок 4 - диаграмма "ящик с усами" в объединённом датасете

Для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; пред обработать данные: удалить шумы и выбросы, сделать нормализацию и стандартизацию. Обучить не- сколько 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, кото- рое будет выдавать прогноз соотношения «матрица-наполнитель». Оценить точ- ность модели на тренировочном и тестовом датасете. Создать репозиторий в GitHub и разместить код исследования. Оформить файл README.

### Описание используемых методов

Данная задача в рамках классификации категорий машинного обучения отно- сится к машинному обучению с учителем и традиционно это задача регрессии. Цель любого алгоритма обучения с учителем — определить функцию потерь и миними- зировать её, поэтому для наилучшего решения в процессе исследования были при- менены следующие методы:

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

Метод опорных векторов (Support Vector Regression) – этот бинарный линей- ный классификатор был выбран, потому что он хорошо работает на небольших да- тасетах. 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 простран- стве, так что между наблюдениями отдельных категорий имеется разрыв, и он мак- симален.

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

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

Недостатки метода: неустойчивость к шуму, поэтому в работе была прове- дена тщательнейшая работа с выбросами, иначе в обучающих данных шумы стано- вятся опорными объектами-нарушителями и напрямую влияют на построение раз- деляющей гиперплоскости; для больших наборов данных требуется долгое время обучения; достаточно сложно подбирать полезные преобразования данных; пара- метры модели сложно интерпретировать, поэтому были рассмотрены и другие ме- тоды.



Рисунок 5 - график метода опорных векторов для прочности при растяжении, МПа Случайный лес (RandomForest) — это множество решающих деревьев. Уни- версальный алгоритм машинного обучения с учителем, представитель ансамблевых

методов. Если точность дерева решений оказалось недостаточной, мы можем мно- жество моделей собрать в коллектив.

Достоинства метода: не переобучается; не требует предобработки входных данных; эффективно обрабатывает пропущенные данные, данные с большим чис- лом 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 интерпре- тируемый; не обладает возможностью экстраполяции; может недо обучаться; тру- доёмко прогнозируемый; иногда работает хуже, чем линейные методы.

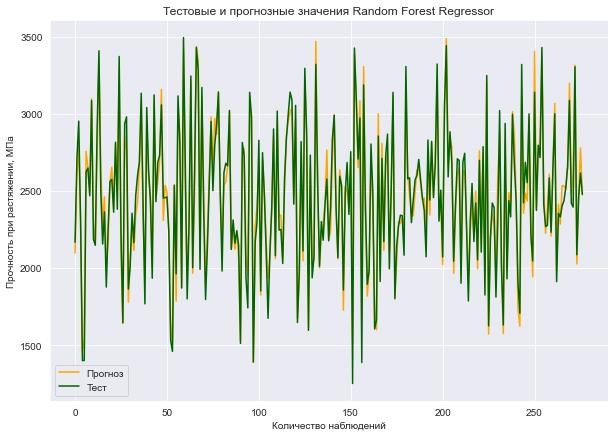


Рисунок 6 - график "случайного леса" для прочности при растяжении, МПа Линейная регрессия (Linear regression) — это алгоритм машинного обучения,

основанный на контролируемом обучении, рассматривающий зависимость между

одной входной и выходными переменными. Это один из самых простых и эффек- тивных инструментов статистического моделирования. Она определяет зависи- мость переменных с помощью линии наилучшего соответствия. Модель регрессии создаёт несколько метрик. R2 , или коэффициент детерминации, позволяет изме- рить, 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, мо- дель 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; тре- бует прямую связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние, а границы линейны.

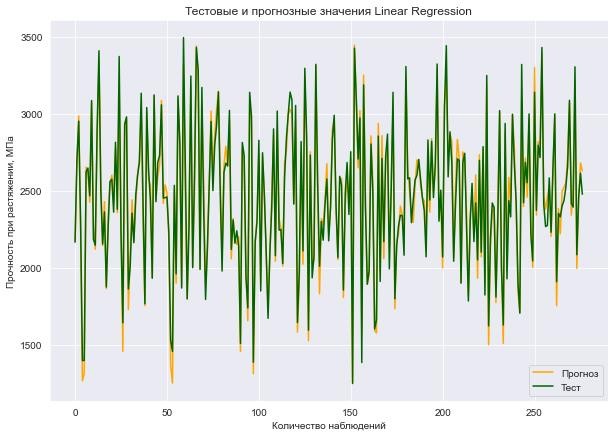


Рисунок 7 график линейной регрессии для прочности при растяжении, МПа Градиентный бустинг (Gradient Boosting) — это ансамбль деревьев решений,

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 не- достатки предыдущих. Финальный классификатор является линейной композицией этих базовых классификаторов.

Достоинства метода: новые алгоритмы учатся на ошибках предыдущих; тре- буется меньше итераций, чтобы приблизиться к фактическим прогнозам; наблюде- ния выбираются на основе ошибки; прост в настройке темпа обучения и примене- ния; легко интерпретируем.

Недостатки метода: необходимо тщательно выбирать критерии остановки, иначе это может привести к переобучению; наблюдения с наибольшей ошибкой по- являются чаще; слабее и менее гибок чем нейронные сети.

Рисунок 8 - график градиентного бустинга для прочности при растяжении, МПа Метод ближайших соседей - К-ближайших соседей (kNN - k Nearest Neigh-

bours) ищет ближайшие объекты с известными значения целевой переменной и ос-

новывается на хранении данных в памяти для сравнения с новыми элементами. Ал- горитм 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 результа- тов; имеет низкую чувствительность к выбросам; не требует построения модели; допускает настройку нескольких параметров; позволяет делать дополнительные до- пущения; универсален; находит лучшее решение из возможных; решает задачи не- большой размерности.

Недостатки метода: замедляется с ростом объёма данных; не создаёт правил; не обобщает предыдущий опыт; основывается на всем массиве доступных истори- ческих данных; невозможно сказать, на каком основании строятся ответы; сложно выбрать близость метрики; имеет высокую зависимость результатов классифика- ции от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоёмкость.

Рисунок 9 - 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 катего- рию контролируемых алгоритмов обучения. Он работает как для непрерывных, так и для категориальных выходных переменных. Правила генерируются за счёт обоб- щения 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 по- лучения значимого непрерывного вывода. Дерево решений один из вариантов ре- шения регрессионной задачи, в случае если зависимость в данных не имеет очевид- ной 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 (измене- ние в одном узле может привести к построению совсем другого дерева); имеет за- тратные вычисления; необходимо обращать внимание на размер; ограниченное число вариантов решения проблемы.

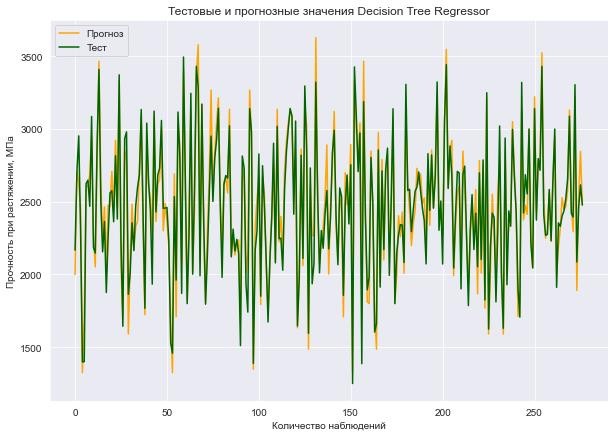


Рисунок 10 - график дерева принятия решений для прочности при растяжении Стохастический градиентный спуск (SGDRegressor) — это простой, но очень

эффективный подход к подгонке линейных классификаторов и регрессоров под вы- пуклые функции потерь. Этот подход подразумевает корректировку весов нейрон- ной сети, используя аппроксимацию градиента функционала, вычисленную только на одном случайном обучающем примере из выборки.

Достоинства метода: эффективен; прост в реализации; имеет множество воз- можностей для настройки кода; способен обучаться на избыточно больших выбор- ках.

Недостатки метода: требует ряд гиперпараметров; чувствителен к масштаби- рованию функций; может не сходиться или сходиться слишком медленно; функци- онал многоэкстремален; процесс может "застрять" в одном из локальных миниму- мов; возможно переобучение.

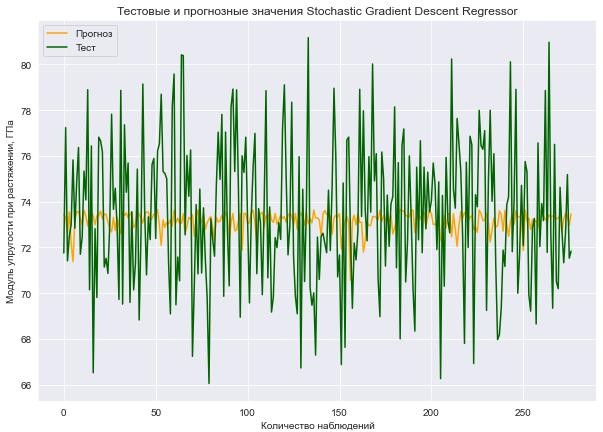


Рисунок 11-график стохастического градиентного спуска для модуля упругости Многослойный персептрон (MLPRegressor) — это алгоритм обучения с учи-

телем, 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 персептро- нов.

Достоинства метода: построение сложных разделяющих поверхностей; воз- можность осуществления любого отображения входных векторов в выходные; легко обобщает входные данные; не требует распределения входных векторов; изу- чает нелинейные модели.

Недостатки метода: имеет невыпуклую функцию потерь; разные инициализа- ции случайных весов могут привести к разной точности проверки; требует настройки ряда гиперпараметров; чувствителен к масштабированию функций.

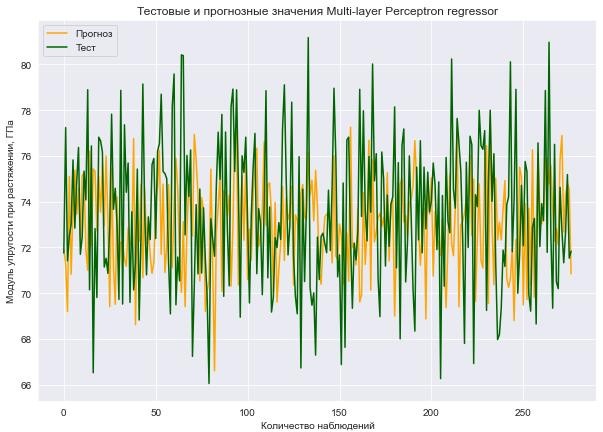


Рисунок 12-график многослойного персептрона для модуля упругости, ГПа

Лассо регрессия (Lasso) — это линейная модель, которая оценивает разрежен- ные коэффициенты. Это простой метод, позволяющий уменьшить сложность мо- дели и предотвратить переопределение, которое может возникнуть в результате простой линейной регрессии. Данный метод вводит дополнительное слагаемое ре- гуляризации в оптимизацию модели. Это даёт более устойчивое решение. В регрес- сии лассо добавляется условие смещения в функцию оптимизации для того, чтобы уменьшить коллинеарность и, следовательно, дисперсию модели. Но вместо квад- ратичного смещения, используется смещение абсолютного значения. Лассо регрес- сия 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 да- тасета; доступно обнуляет значения коэффициентов.

Недостатки метода: выбор модели не помогает и обычно вредит; часто стра- дает качество прогнозирования; выдаёт ложное срабатывание результата; случай- ным образом выбирает одну из коллинеарных переменных; не оценивает правиль- ность формы взаимосвязи между независимой и зависимой переменными; не всегда лучше, чем пошаговая регрессия.

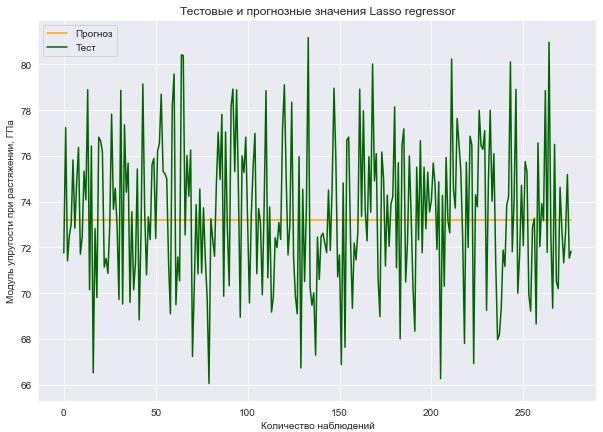


Рисунок 13 - график метода Лассо для модуля упругости, Гпа

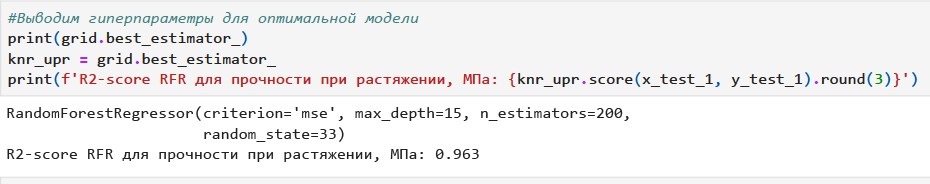
Немного расскажем об используемых метриках качества моделей: R2 или ко- эффициент детерминации измеряет долю дисперсии, объяснённую моделью, в об- щей дисперсии целевой переменной.

Рисунок 14 – часть кода для результата метрики R2 для метода «Случайный лес»

Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то качество прогноза идентично средней величине целевой перемен- ной (т.е. очень низкое). Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

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

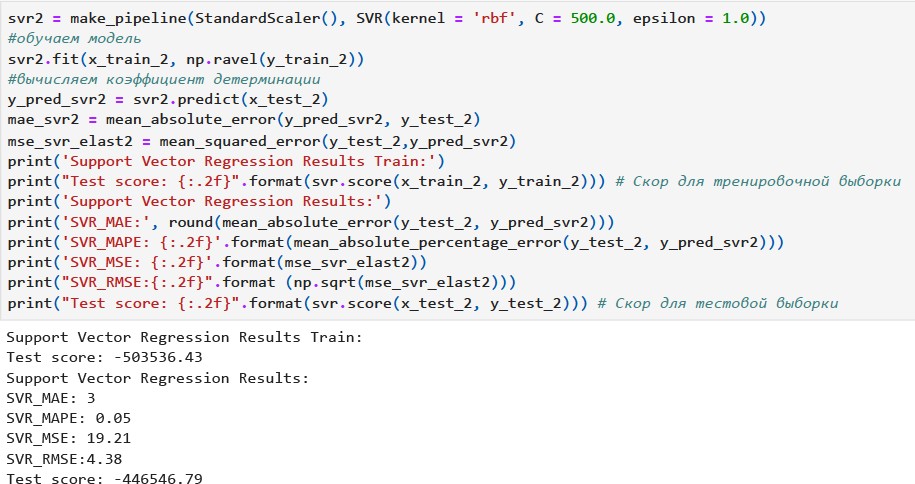


Рисунок 15 - код для вывода различных метрик для метода опорных векторов

### Разведочный анализ данных

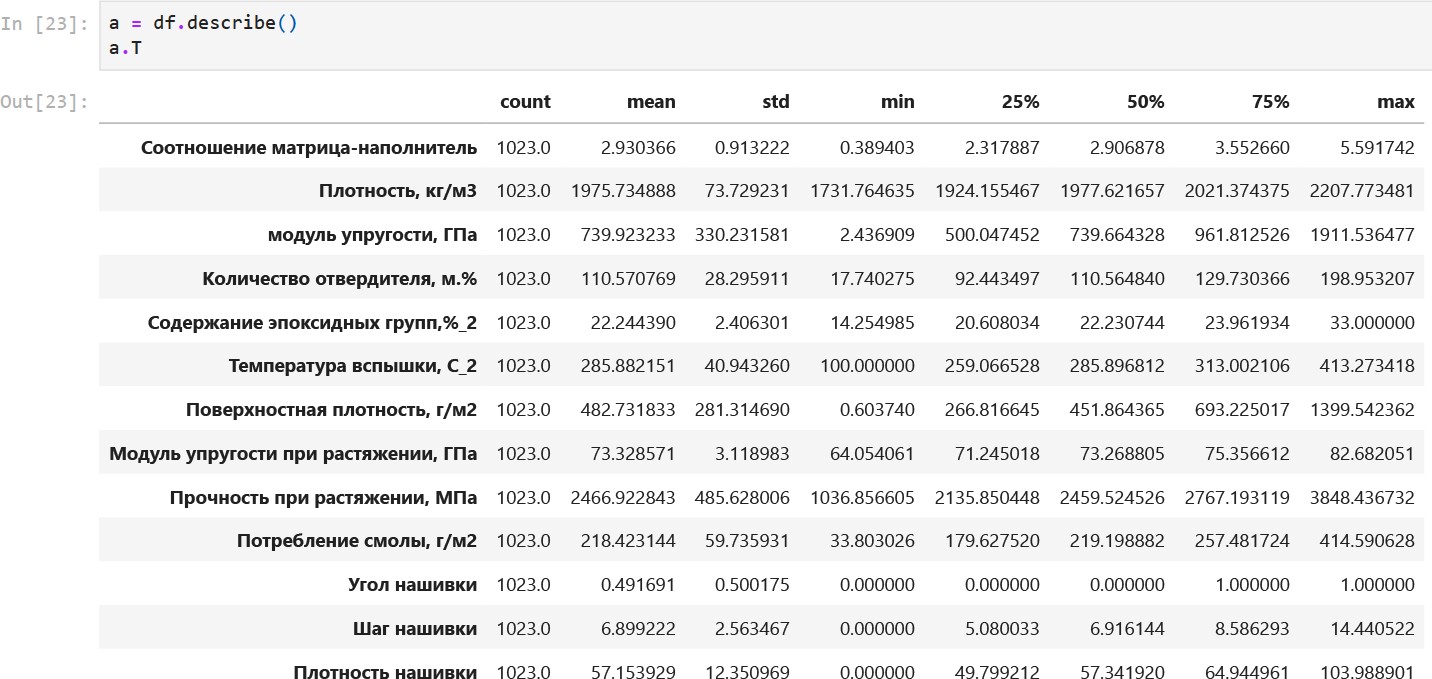
Прежде чем передать данные в работу моделей машинного обучения, необхо- димо обработать и очистить их. Очевидно, что «грязные» и необработанные данные могут содержать искажения и пропущенные значения – это ненадёжно, поскольку способно привести к крайне неверным результатам по итогам моделирования. Но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

Рисунок 16 - описательная статистика датасета

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

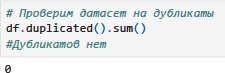


Рисунок 17 - проверка датасета на наличие дубликатов

В качестве инструментов разведочного анализа используется: оценка стати- стических характеристик датасета; гистограммы распределения каждой из перемен- ной (несколько различных вариантов); диаграммы ящика с усами (несколько интер- активных вариантов); попарные графики рассеяния точек (несколько вариантов); график «квантиль-квантиль»; тепловая карта (несколько вариантов); описательная статистика для каждой переменной; анализ и полное исключение выбросов (5 по- вторных итераций); проверка наличия пропусков и дубликатов; ранговая корреля- ция Кендалла и Пирсона.

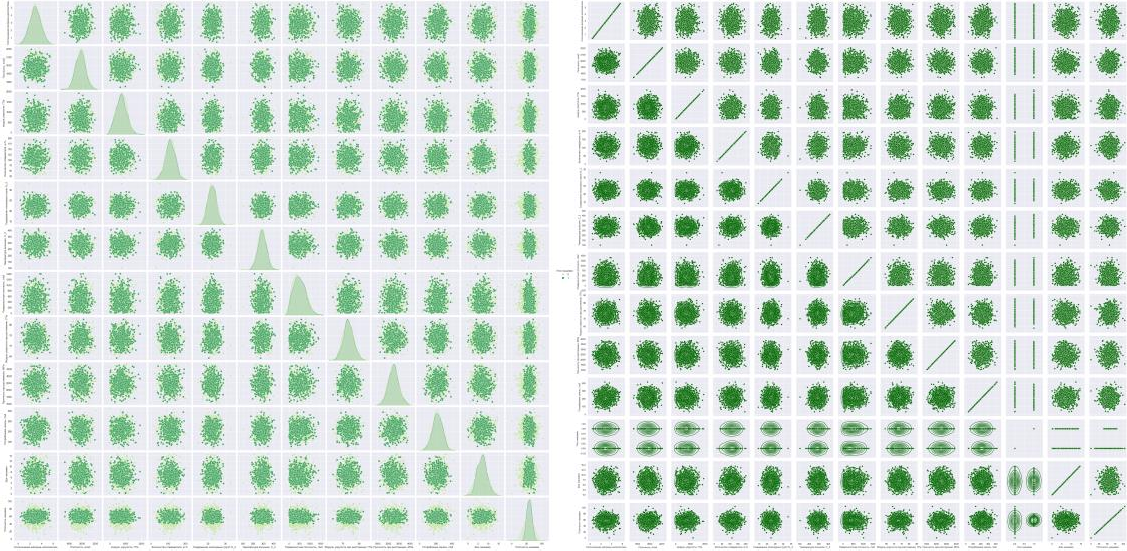


Рисунок 18 - попарные графики рассеяния точек (два разных варианта)

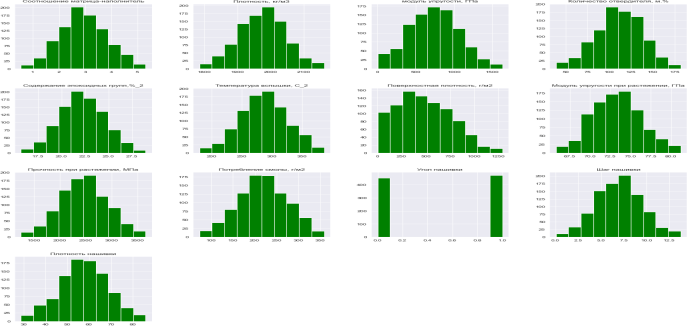
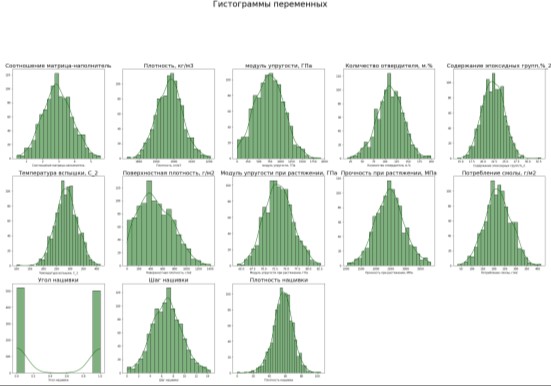
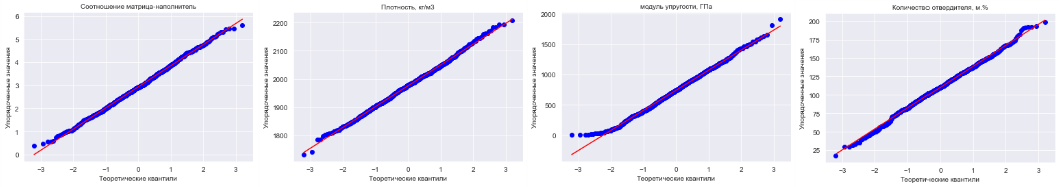


Рисунок 19 - гистограммы распределения (2 разных варианта) Гистограммы используются для изучения распределений частот значений пе-

ременных. Мы видим очень слабую корреляцию между переменными.



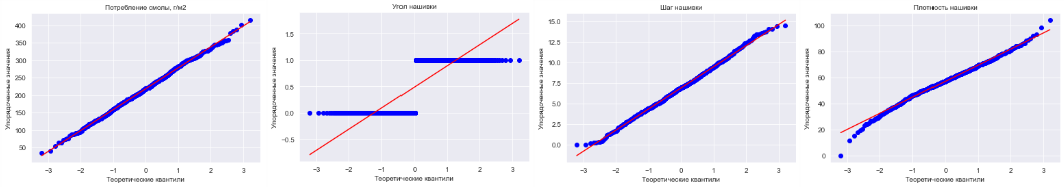


Рисунок 20 - графики «квантиль-квантиль»

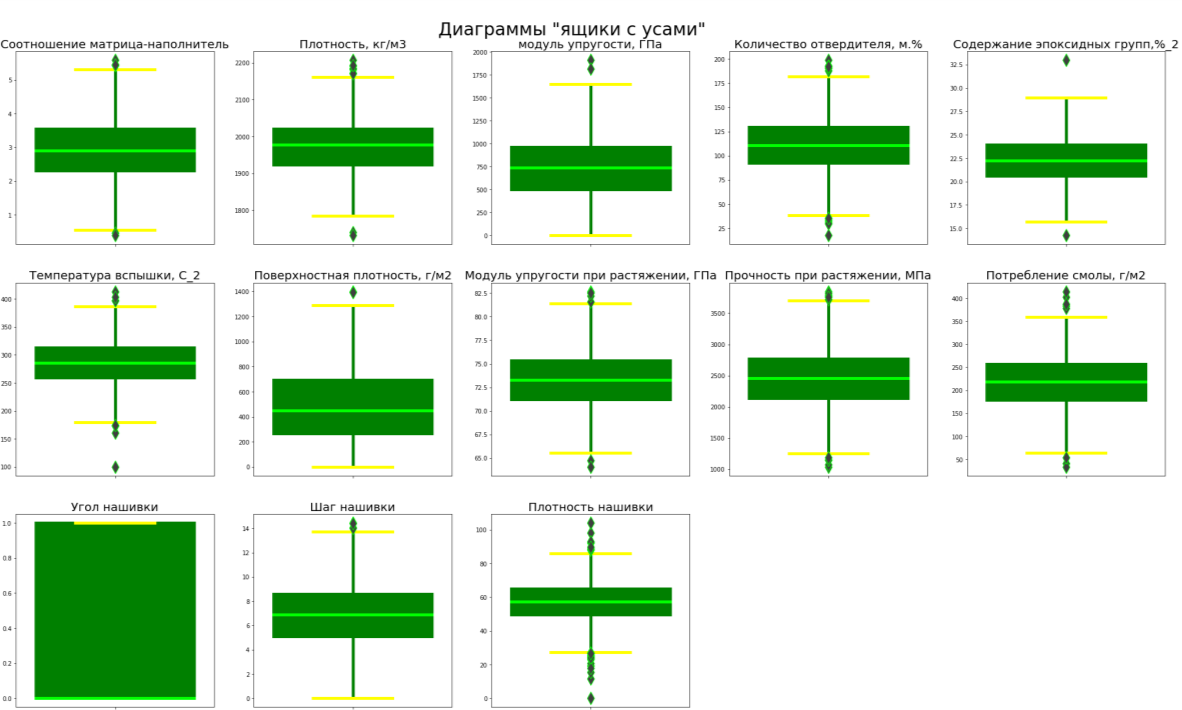


Рисунок 21 - график "ящиков с усами" для всех переменных

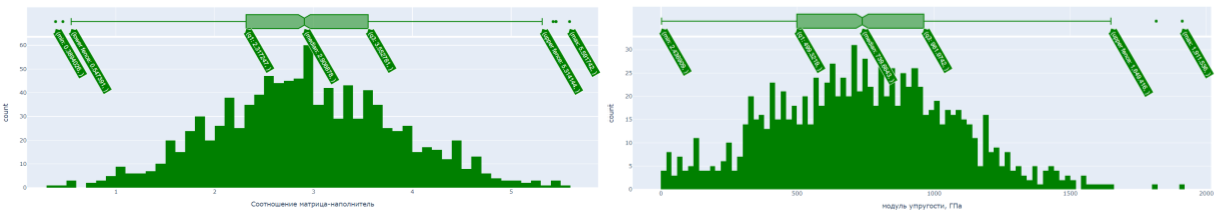


Рисунок 22 - интерактивный график "ящиков с усами" и гистограммы распределе- ния на примере двух переменных

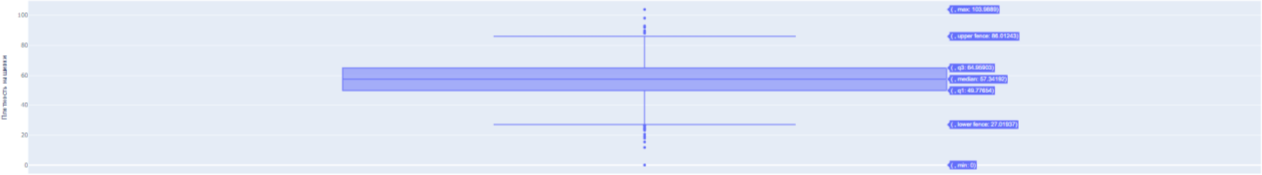


Рисунок 23 - график "ящиков с усами" для одной переменной

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

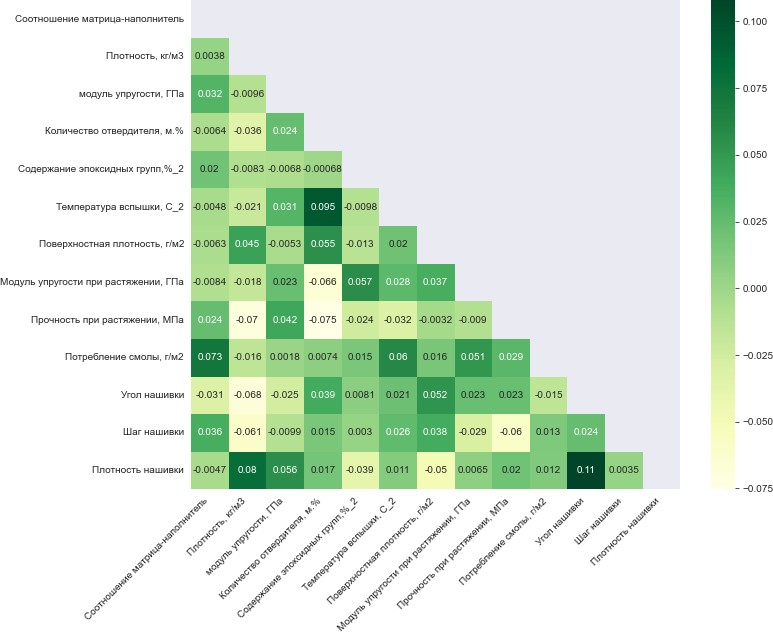
Данные объединённого датасета не имеют чётко выраженной зависимости, что подтверждает тепловая карта с матрицей корреляции и матрицы диаграмм рас- сеяния.

Рисунок 24 - тепловая карта с корреляцией данных Максимальная корреляция между плотностью нашивки и углом нашивки

0.11, значит нет зависимости между этими данными. Корреляция между всеми па- раметрами очень близка к 0, корреляционные связи между переменными не наблю- даются.

## Практическая часть

### Предобработка данных

В ходе проведённого анализа принимаем решение столбец "Угол нашивки" привести к виду «0» и «1».

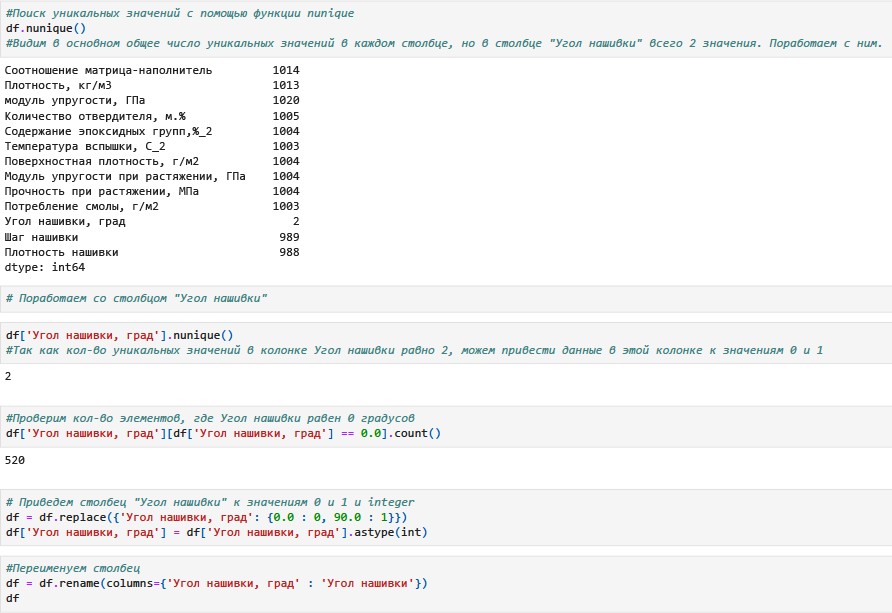


Рисунок 25 - часть кода с преобразованием столбца "Угол нашивки"

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

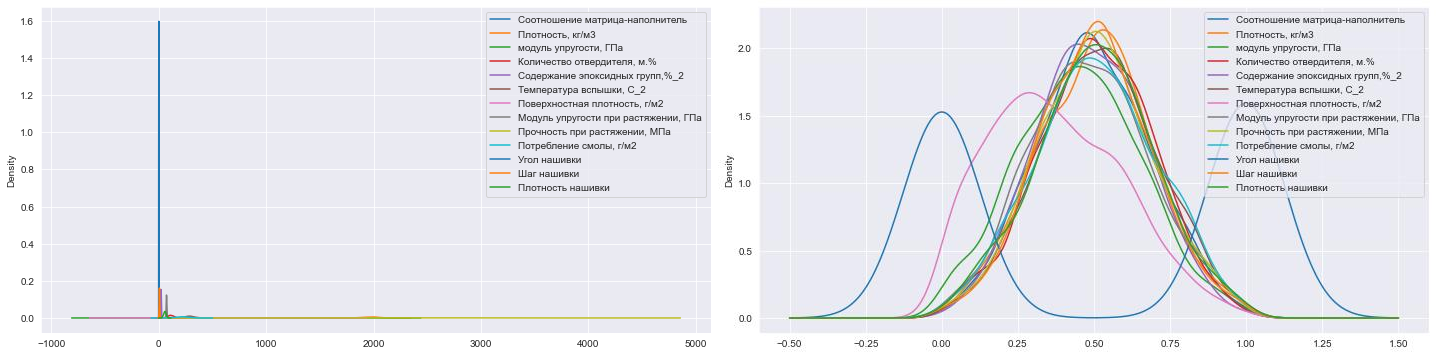


Рисунок 26 - визуализированные данные до и после нормализации

### Разработка и обучение модели

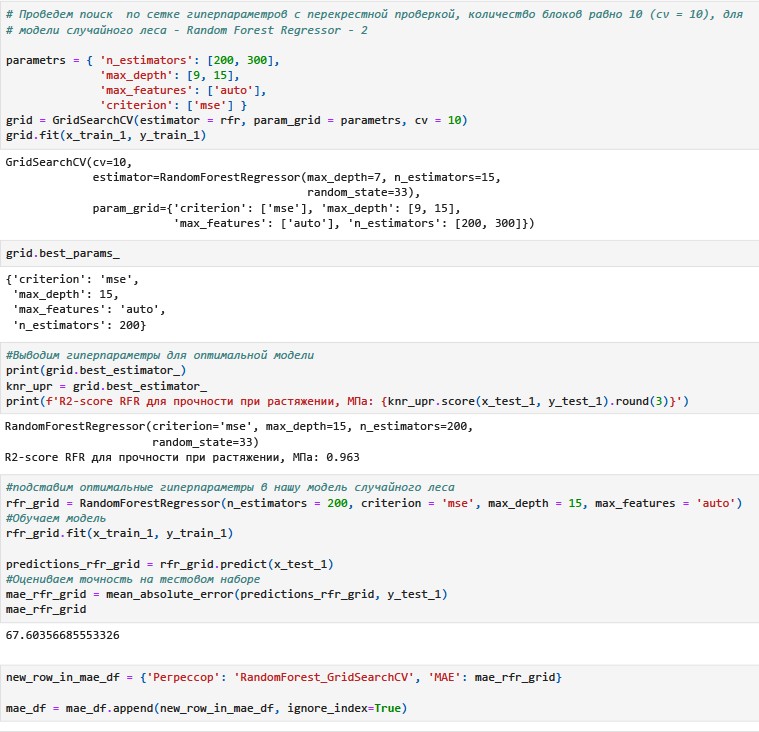
Разработка и обучение моделей машинного обучения осуществлялась для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении» отдельно. Для решения применим все методы, описанные выше.

Рисунок 27- поиск гиперпараметров

Порядок разработки модели для каждого параметра и для каждого выбран- ного 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

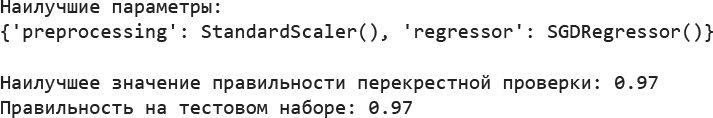
качестве параметра оценки выбран коэффициент детерминации (R2); оптимизация подбора гиперпараметров модели с помощью выбора по сетке и перекрёстной про- верки; подстановка оптимальных гиперпараметров в модель и обучение модели на тренировочных данных; оценка полученных данных; сравнение со стандартными значениями.

Рисунок 28 - 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 справи- лись с задачей. Результат неудовлетворительный. Свойства композитных материа- лов в первую очередь зависят от используемых материалов.

Таблица 1. Результаты построения и обучения моделей

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Model** | **MAE** | **R2 score** |
| **Модуль упругости при растяжении** | KNeighborsRegressor\_upr | 2.517884 | -0.001000 |
| **Прочность при растяжении** | KNeighborsRegressor\_pr | 368.126111 | -0.009000 |
| **Модуль упругости при растяжении** | LinearRegression\_upr | 2.546419 | -0.021000 |
| **Прочность при растяжении** | LinearRegression\_pr | 370.542618 | -0.021000 |
| **Модуль упругости при растяжении** | RandomForestRegressor\_upr | 2.597992 | -0.061000 |
| **Прочность при растяжении** | RandomForestRegressor\_pr | 370.967356 | -0.013000 |
| **Модуль упругости при растяжении** | MLPRegressor\_upr | 2.514920 | -0.000084 |
| **Прочность при растяжении** | MLPRegressor\_pr | 367.611428 | -0.009000 |

### Тестирование модели

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



Рисунок 29 - результат оценки точности по MSE и R2

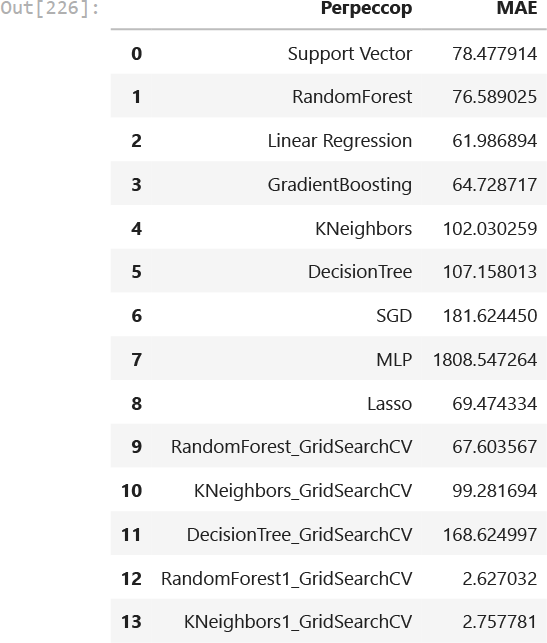


Рисунок 30 - результат оценки точности по MAE

Хотя в целом при таких результатах можно применять среднее значение пе- ременной в качестве прогнозного.

### Написать нейронную сеть, которая будет рекомендовать соотношение «матрица – наполнитель».

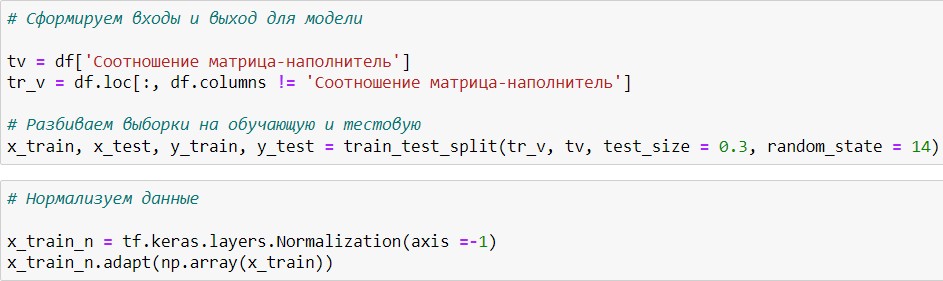
Обучение нейронной сети — это такой процесс, при котором происходит под- бор оптимальных параметров модели, с точки зрения минимизации функционала ошибки. Начнём стоить нейронную сеть с помощью класса keras.Sequential.

Рисунок 31 - создание нейронной сети

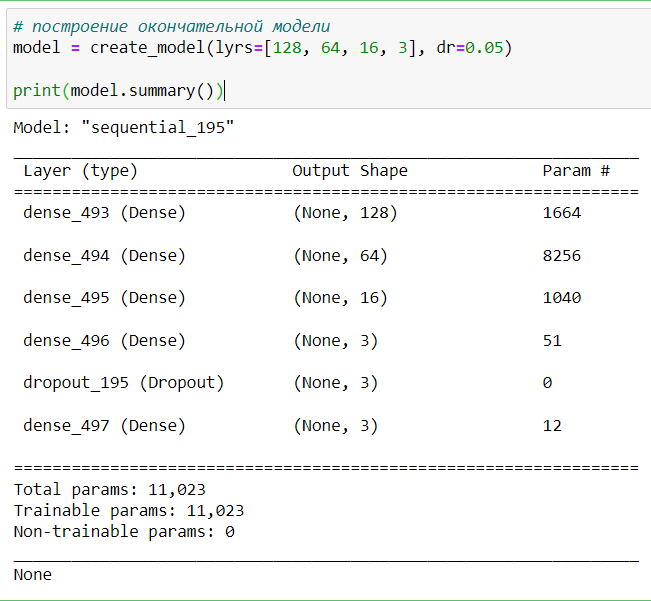
Определим параметры, поищем оптимальные параметры, посмотрим на ре- зультаты. С помощью KerasClassifier выйдем на наилучшие параметры для нашей нейронной сети и построим окончательную нейросеть.

Рисунок 32 - построение первой нейросети

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

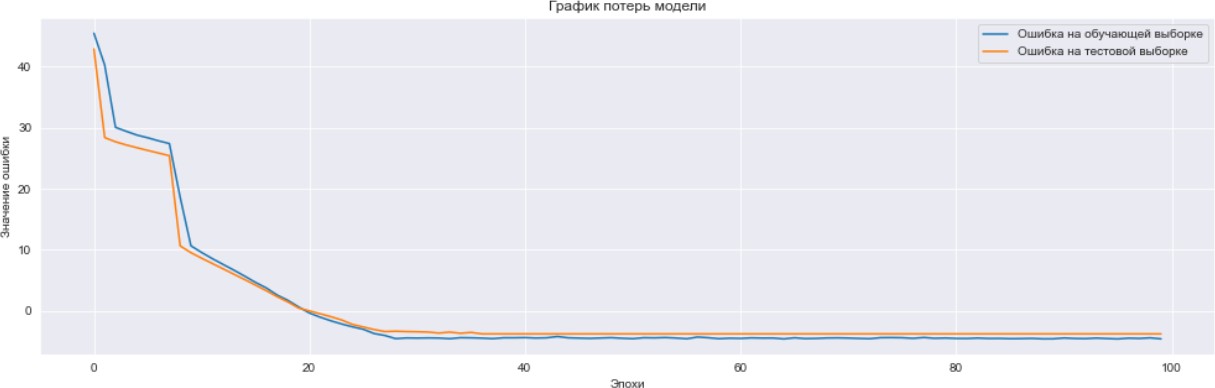


Рисунок 33 - график потерь модели 1

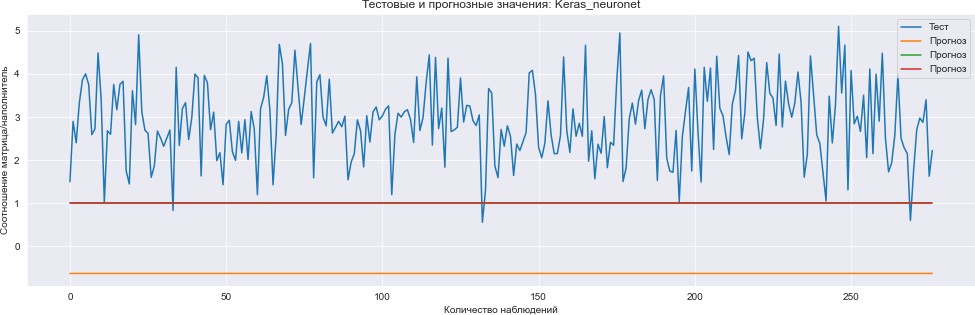


Рисунок 34 - тестовые и прогнозные значения модели 1

Не удовлетворившись таким результатом, создадим другую простую модель глубокого обучения с другой архитектурой. Обучим её, посмотрим на потери, оце- ним MSE, построим график.

Рисунок 35 - создание второй модели

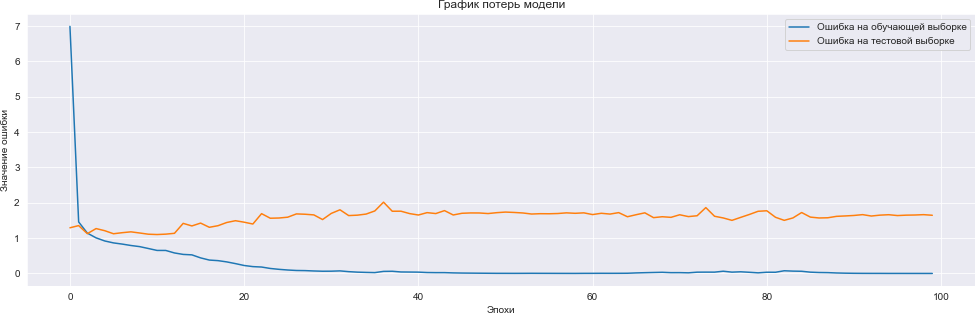


Рисунок 36 - график потерь второй модели

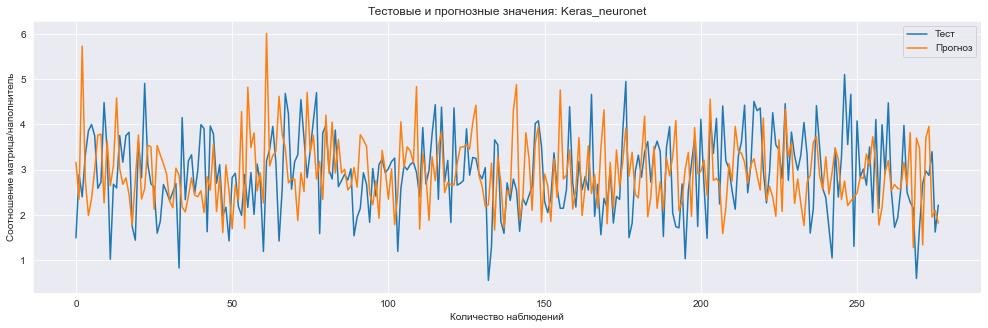


Рисунок 37- тестовые и прогнозные значения второй модели

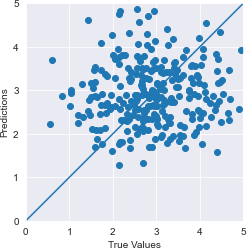


Рисунок 38- график прогнозных и настоящих значений

### Разработка приложения

Приложение успешно работает и показывает результат прогноза для соотно- шения «матрица – наполнитель».



Рисунок 39 - пример результата работы приложения

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

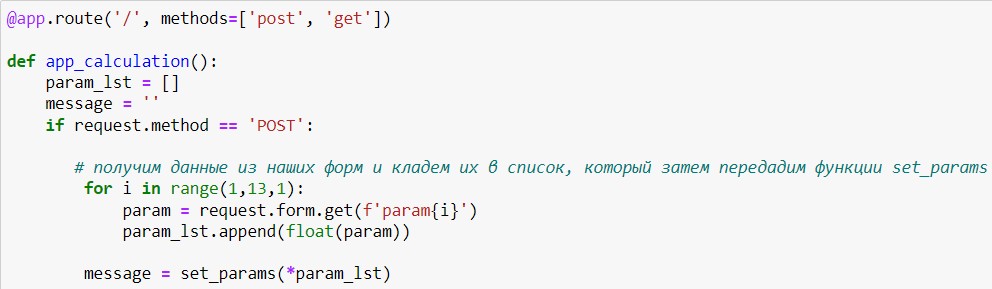


Рисунок 40 - часть кода приложения

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

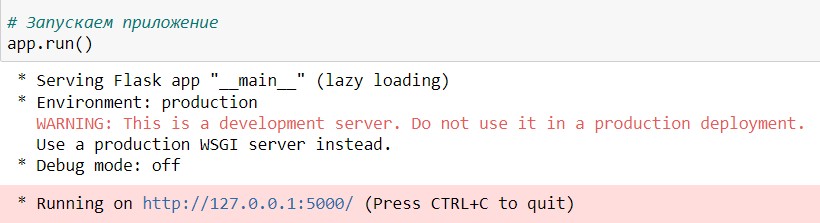


Рисунок 41 - ссылка для открытия html - файла

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

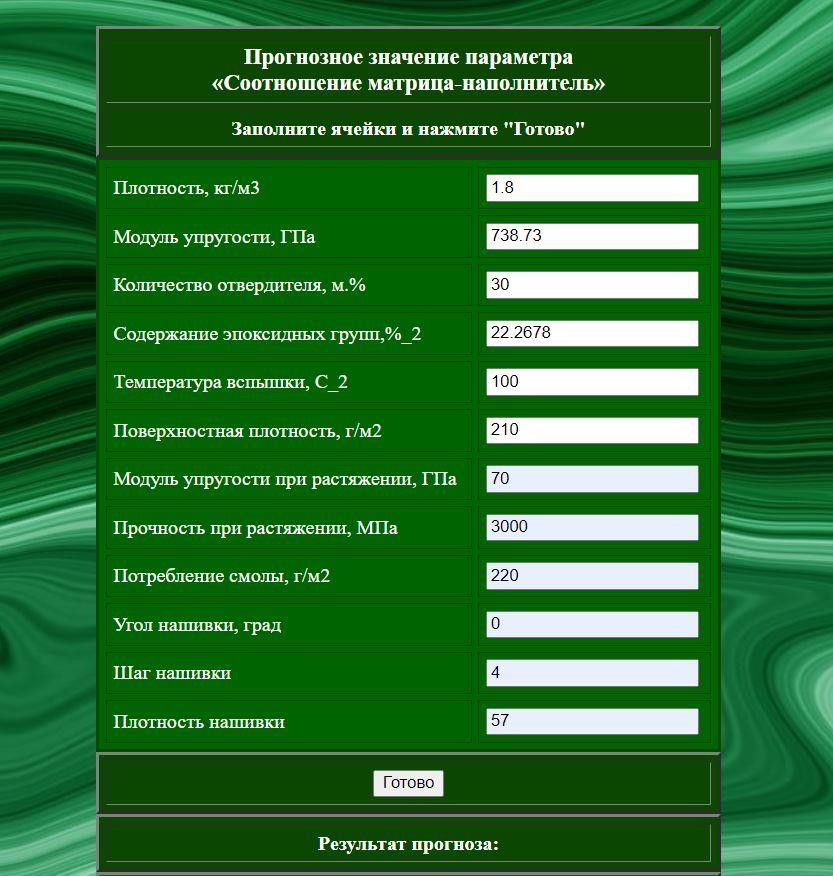


Рисунок 42- скриншот пользовательского приложения

На выходе пользователь получает результат прогноза для значения параметра

«Соотношение «матрица – наполнитель»».

### Создание удалённого репозитория и загрузка

Репозиторий был создан на github.com по адресу:[https://github.com/](https://github.com/Oleg-Evdokimov/KOMPOSIT)OVShishkova

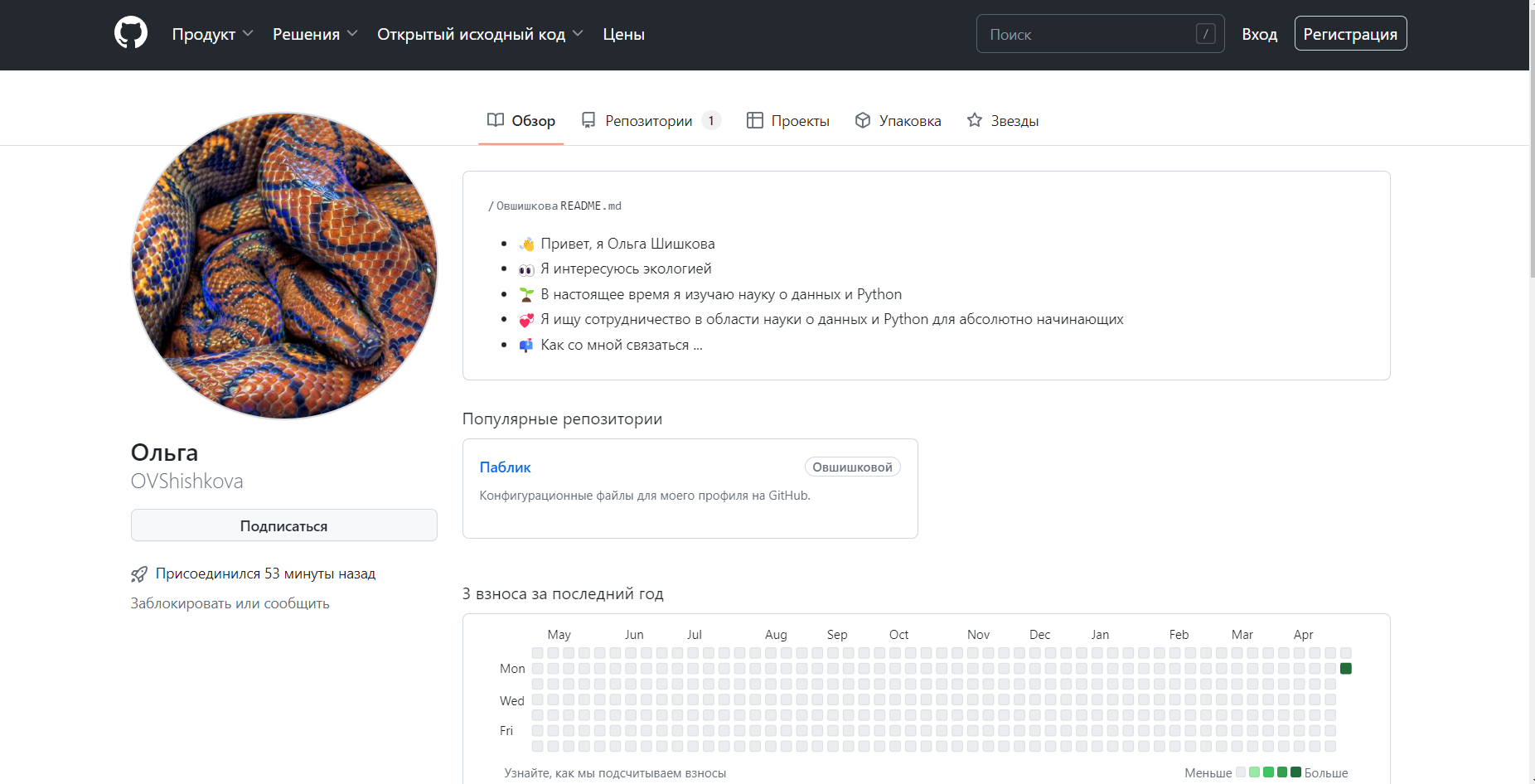


Рисунок 43 - часть страницы на github.com

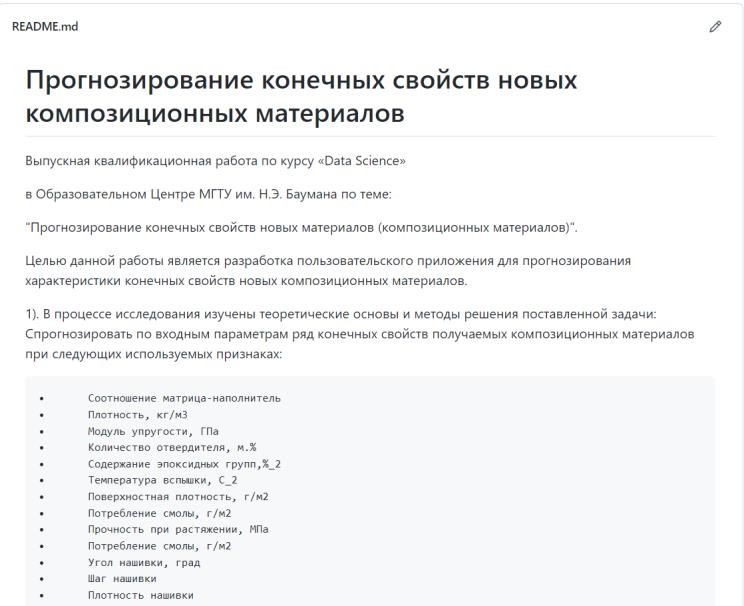


Рисунок 44 - часть созданного файла README

Ноутбук с решением и приложением так же можно найти по адресу: [https://colab.research.google.com/drive/1Pa9Y6gQq-](https://colab.research.google.com/drive/1Pa9Y6gQq-vfE3cyAeAO7EhlQSHq2SO1P?authuser=2)

[vfE3cyAeAO7EhlQSHq2SO1P?authuser=2](https://colab.research.google.com/drive/1Pa9Y6gQq-vfE3cyAeAO7EhlQSHq2SO1P?authuser=2)

### Заключение

Данная исследовательская работа позволяет сделать некоторые основные вы- воды 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 пока- зали высокой эффективности в прогнозировании свойств композитов. Лучшие мет- рики 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 соот- ношение «матрица – наполнитель». Данный факт не указывает на то, что прогнози- рование характеристик композитных материалов на основании предоставленного набора данных невозможно, но может указывать на недостатки базы данных, под- ходов, использованных при прогнозе, необходимости пересмотра инструментов для прогнозирования.

Необходимы дополнительные вводные данные, получение новых результиру- ющих признаков в результате математических преобразований, релевантных до- менной области, консультации экспертов предметной области, новые исследова- ния, работа эффективной команды, состоящей из различных учёных.

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

## Список используемой литературы и веб ресурсы.

1. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour): – Режим доступа: [https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-](https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19) [2021-07-19](https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour-2021-07-19). (дата обращения: 07.04.2023)
2. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам сле-дует знать: – Режим доступа: https://habr.com/ru/company/vk/blog/513842/(дата обращения: 01.03.2023).
3. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.
4. Абросимов Н.А.: Методика построения разрешающей системы уравне- ний 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 при- ложения: учеб. пособие /Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Издательство Казанского университета, 2018. – 121 с.
8. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., пе- рераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.
9. Документация по библиотеке keras: – Режим до- ступа: <https://keras.io/api/>.(дата обращения: 08.04.2023).
10. Документация по библиотеке matplotlib: – Режим до-

ступа: <https://matplotlib.org/stable/users/index.html>. (дата обращения: 10.04.2023)

1. Документация по библиотеке numpy: – Режим

доступа: <https://numpy.org/doc/1.22/user/index.html#user>. (дата обращения: 03.04.2023).

1. Документация по библиотеке pandas: – Режим до-

ступа: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>. (дата обраще- ния: 04.04.2023).

1. Документация по библиотеке scikit-learn: – Режим до-

ступа: <https://scikit-learn.org/stable/user_guide.html>. (дата обращения: 05.04.2023).

1. Документация по библиотеке seaborn: – Режим до-

ступа: <https://seaborn.pydata.org/tutorial.html>. (дата обращения: 06.04.2023).

1. Документация по библиотеке Tensorflow: – Режим доступа: <https://www.tensorflow.org/overview>(дата обращения: 10.04.2023).
2. Документация по языку программирования python: – Режим до- ступа: <https://docs.python.org/3.8/index.html>. (дата обращения: 02.04.2023).
3. Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные матери- алы: учебное пособие для вузов, 2019. 13 с.
4. Краткий обзор алгоритма машинного обучения Метод Опорных Век- торов (SVM) – Режим доступа: <https://habr.com/ru/post/428503/>(дата обращения 07.04.2023)
5. Ларин А. А., Способы оценки работоспособности изделий из компози- ционных материалов методом компьютерной томографии, Москва, 2013, 148 с.
6. Материалы конференции: V Всероссийская научно-техническая конфе- ренция «Полимерные композиционные материалы и производственные технологии нового поколения», 19 ноября 2021 г.
7. Миронов А.А. Машинное обучение часть I ст.9 – Режим доступа: <http://is.ifmo.ru/verification/machine-learning-mironov.pdf>. (дата обращения 08.04.2023)
8. Плас Дж. Вандер, Python для сложных задач: наука о данных и машин-

ное обучение. Санкт-Петербург: Питер, 2018, 576 с.

1. 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.
2. Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.: Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил.
3. Руководство по быстрому старту в flask: – Режим доступа: [https://flask-](https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html) [russian-docs.readthedocs.io/ru/latest/quickstart.html](https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html). (дата обращения: 09.04.2023)
4. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
5. Скиена, Стивен С. С42 Наука о данных: учебный курс.: Пер. с англ. - СПб.: ООО "Диалектика", 2020. - 544 с. : ил.
6. 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;
7. Траск Эндрю. Грокаем глубокое обучение. – СПб.: Питер, 2019. – 352

с.: ил.

1. Чун-Те Чен и Грейс Х. Гу. Машинное обучение для композитных мате-

риалов (март 2019г.) – Режим доступа: [https://www.cambridge.org/core/journals/mrs-](https://www.cambridge.org/core/journals/mrs-communications/article/machine-%20learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15) [communications/article/machine- learning-for-composite-](https://www.cambridge.org/core/journals/mrs-communications/article/machine-%20learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15) [materials/F54F60AC0048291BA47E0B671733ED15](https://www.cambridge.org/core/journals/mrs-communications/article/machine-%20learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15). (дата обращения 02.04.2023)

### Приложение 1

Подробный план работы:

1. Загружаем и обрабатываем входящие датасеты 1.1.Удаляем неинформативные столбцы 1.2.Объединяем датасеты по методу INNER
2. Проводим разведочный анализ данных:
   1. Данные в столбце "Угол нашивки» приведём к 0 и 1
   2. Изучим описательную статистику каждой переменной - среднее, медиана, стандартное отклонение, минимум, максимум, квартили
   3. Проверим датасет на пропуски и дубликаты данных
   4. Получим среднее, медианное значение для каждой колонки (по заданию необходимо получить их отдельно, поэтому продублируем их только от- дельно)
   5. Вычислим коэффициенты ранговой корреляции Кендалла 2.6.Вычислим коэффициенты корреляции Пирсона
3. Визуализируем наш разведочный анализ сырых данных (до выбросов и норма- лизации)
   1. Построим несколько вариантов гистограмм распределения каждой перемен- ной
   2. Построим несколько вариантов диаграмм ящиков с усами каждой перемен- ной
   3. Построим гистограмму распределения и диаграмма "ящик с усами" одновре- менно вместе с данными по каждому столбцу
   4. Построим несколько вариантов попарных графиков рассеяния точек (мат- рицы диаграмм рассеяния)
   5. Построим графики квантиль-квантиль
   6. 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 4.2.Посчитаем распределение выбросов по каждому столбцу (с целью предот-

вращения удаления особенностей признака или допущения ошибки) 4.3.Исключим выбросы методом межквартильного расстояния 4.4.Удалим строки c выбросами

4.5.Визуализируем датасет без выбросов, и убедимся, что выбросы еще есть. 4.6.Для полной очистки датасета от выбросов повторим пункты (4.3 – 4.5) ещё 3

раза.

4.7.Сохраняем идеальный, без выбросов датасет 4.8.Изучим чистые данные по всем параметрам 4.9.Визуализируем «чистый» датасет (без выбросов)

1. Проведём нормализацию и стандартизацию (продолжим предобработку данных) 5.1.Визуализируем плотность ядра

5.2.Нормализуем данные с помощью MinMaxScaler() 5.3.Нормализуем данные с помощью Normalizer() 5.4.Сравним с данными до нормализации

* 1. Проверим перевод данных из нормализованных в исходные
  2. Рассмотрим несколько вариантов корреляции между параметрами после нор- мализации
  3. Стандартизируем данные 5.8.Визуализируем данные корреляции

5.9.Посмотрим на описательную статистику после нормализации и после стан- дартизации

1. 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 6.3.Проверим правильность разбивки

6.4.Построим модели и найдём лучшие гиперпараметры (задача по заданию): 6.5.Построим и визуализируем результат работы метода опорных векторов 6.6.Построим и визуализируем результат работы метода случайного леса 6.7.Построим и визуализируем результат работы линейной регрессии 6.8.Построим и визуализируем результат работы метода градиентного бустинга 6.9.Построим и визуализируем результат работы метода К ближайших соседей

* 1. Построим и визуализируем результат работы метода деревья решений
  2. Построим и визуализируем результат работы стохастического гради- ентного спуска
  3. Построим и визуализируем результат работы многослойного перцеп- трона
  4. Построим и визуализируем результат работы лассо регрессии
  5. Сравним наши модели по метрике МАЕ
  6. Найдём лучшие гиперпараметры для случайного леса
  7. Подставим значения в нашу модель случайного леса
  8. Найдём лучшие гиперпараметры для К ближайших соседей
  9. Подставим значения в нашу модель К ближайших соседей
  10. Найдём лучшие гиперпараметры метода деревья решений
  11. Подставим значения в нашу модель метода деревья решений
  12. Проверим все модели и процессинги и выведем лучшую модель и про- цессинг

1. Разработаем и обучим нескольких моделей прогноза модуля упругости при рас- тяжении (с 30% тестовой выборки)
   1. Определим входы и выходы для моделей 7.2.Разобьём данные на обучающую и тестовую выборки
   2. Проверим правильность разбивки
   3. Построим модели и найдём лучшие гиперпараметры (задача по заданию): 7.5.Построим и визуализируем результат работы метода опорных векторов 7.6.Построим и визуализируем результат работы метода случайного леса 7.7.Построим и визуализируем результат работы линейной регрессии 7.8.Построим и визуализируем результат работы метода градиентного бустинга 7.9.Построим и визуализируем результат работы метода К ближайших соседей
   4. Построим и визуализируем результат работы метода деревья решений
   5. Построим и визуализируем результат работы стохастического гради- ентного спуска
   6. Построим и визуализируем результат работы многослойного перцеп- трона
   7. Построим и визуализируем результат работы лассо регрессии
   8. Сравним наши модели по метрике МАЕ
   9. Найдём лучшие гиперпараметры для случайного леса
   10. Подставим значения в нашу модель случайного леса
   11. Найдём лучшие гиперпараметры для К ближайших соседей
   12. Подставим значения в нашу модель К ближайших соседей
   13. Найдём лучшие гиперпараметры метода деревья решений
   14. Подставим значения в нашу модель метода деревья решений
   15. Проверим все модели и процессинги и выведем лучшую модель и про- цессинг
2. Нейронная сеть для рекомендации соотношения матрица-наполнитель 8.1.Сформируем входы и выход для модели
   1. Нормализуем данные
   2. Построим модель, определим параметры 8.4.Найдем оптимальные параметры для модели 8.5.Посмотрим на результаты
   3. Повторим шаги 8.4 – 8.5 до построения окончательной модели
   4. Обучим нейросеть 80/20 8.8.Оценим модель 8.9.Посмотрим на потери модели
   5. Посмотрим на график результата работы модели
   6. Посмотрим на график потерь на тренировочной и тестовой выборках
   7. Сконфигурируем другую модель, зададим слои
   8. Посмотрим на архитектуру другой модели
   9. Обучим другую модель
   10. Посмотрим на потери другой модели
   11. Посмотрим на график потерь на тренировочной и тестовой выборках
   12. Зададим функцию для визуализации факт/прогноз для результатов мо- делей
   13. Посмотрим на график результата работы модели
   14. Оценим модель MSE
   15. Сохраняем вторую модель для разработки веб-приложения для прогно- зирования соотношения "матрица-наполнитель" в фреймворке Flask
3. Создаём приложение
   1. Импортируем необходимые бибилиотеки 9.2.Загрузим модель и определим параметры функции

9.3.Получим данные из наших форм и положим их в список 9.4.Укажем шаблон и прототип сайта для вывода 9.5.Запустим приложение

9.6.Откроем http://127.0.0.1:5000/

1. Создание удалённого репозитория и загрузка результатов работы на него.

[https://github.com/](https://github.com/Oleg-Evdokimov/KOMPOSIT)OVShishkova

* 1. Создадим README ([https://github.com/](https://github.com/Oleg-Evdokimov/KOMPOSIT#readme)OVShishkova#readme
  2. Выгрузим все необходимые файлы в репозиторий