МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Московский государственный технический университет

имени Н.Э. Баумана (национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

**Тема:** «**Прогнозирование конечных свойств новых материалов (композиционных материалов)»**

Слушатель Комарова Оксана Александровна

Москва, 2023

# Содержание

|  |  |
| --- | --- |
| Содержание……………………………………………………….... | 2 |
| Введение……………………………………………………………. | 3 |
| 1 Аналитическая часть……………………………………………. | 5 |
| 1.1 Постановка задачи…………………………………………….. | 5 |
| 1.2 Описание используемых методов……………………………. | 6 |
| 1.3 Разведочный анализ данных…………………………………. | 16 |
| 2 Практическая часть…………………………………………….. | 21 |
| 2.1 Предобработка данных………………………………………. | 21 |
| 2.2 Описание созданных моделей машинного обучения……… | 23 |
| 2.3 Создание нейронной сети……………………………………. | 26 |
| 2.4 Экспорт модели для создания приложения………………….. | 27 |
| 2.5 Создание профиля и commit на github.com………………… | 28 |
| Заключение………………………………………………………… | 29 |
| Приложение……………………………………………………….. | 30 |

**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).

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

Расширение разнообразия материалов, используемых при проектировании нового композиционного материала, увеличивает необходимость определения свойств нового композита при минимальных финансовых затратах. 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 затрат 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 были созданы нейронных сети, которые предлагают соотношение «матрицы - наполнитель».

**1 Аналитическая часть**

1.1 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).

Цель работы разработать алгоритмы машинного обучения для прогноза модуля упругости при растяжении, прочности при растяжении и написать нейронную сеть для прогноза 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. В задании сказано объединение делать по индексу тип объединения INNER. В результате такого объединения получился датасет, состоящий из 1023 строк.

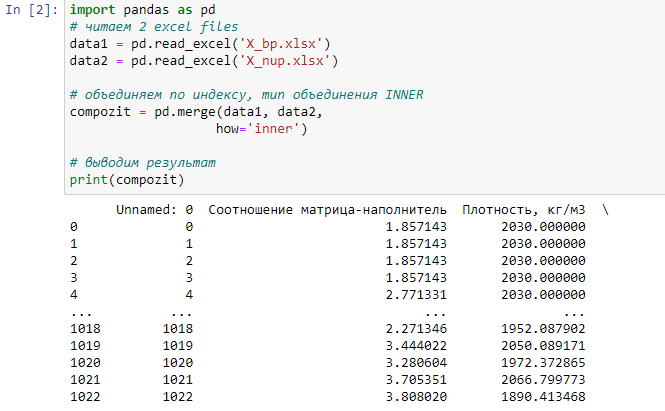


Рисунок 1 - пример начала работы с файлами

1.2 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:

линейная регрессия;

метод опорных векторов;

случайный лес;

градиентный бустинг;

К-ближайших соседей;

дерево решений;

стохастический градиентный спуск.

**Линейная регрессия (Linear regression)** — Используемая в статистике регрессионная модель зависимости одной переменной y от другой или нескольких других переменных x с линейной функцией зависимости. 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. Коэффициент детерминации не может быть отрицательным, данный вывод исходит из свойств коэффициента детерминации. Однако скорректированный коэффициент детерминации вполне может принимать отрицательные значения.

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.

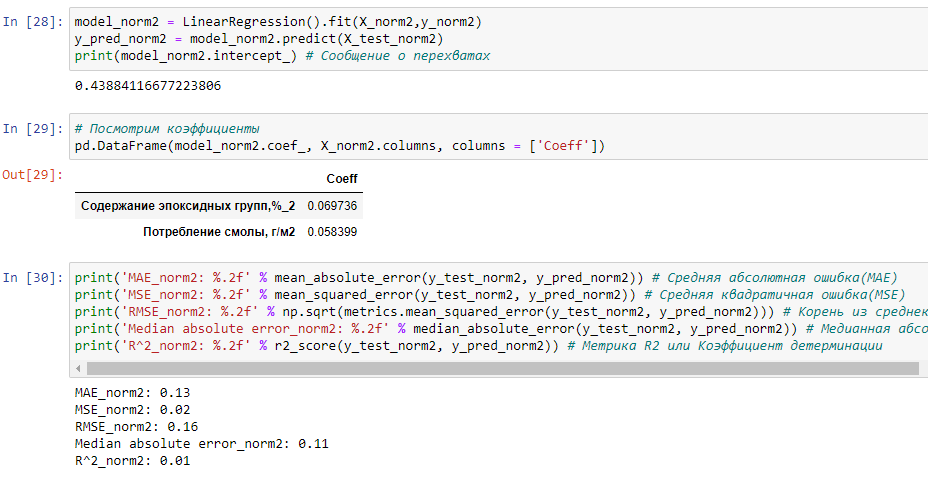


Рисунок 2 – создание модели линейной регрессии

**Метод опорных векторов (Support Vector Regression**) – 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.

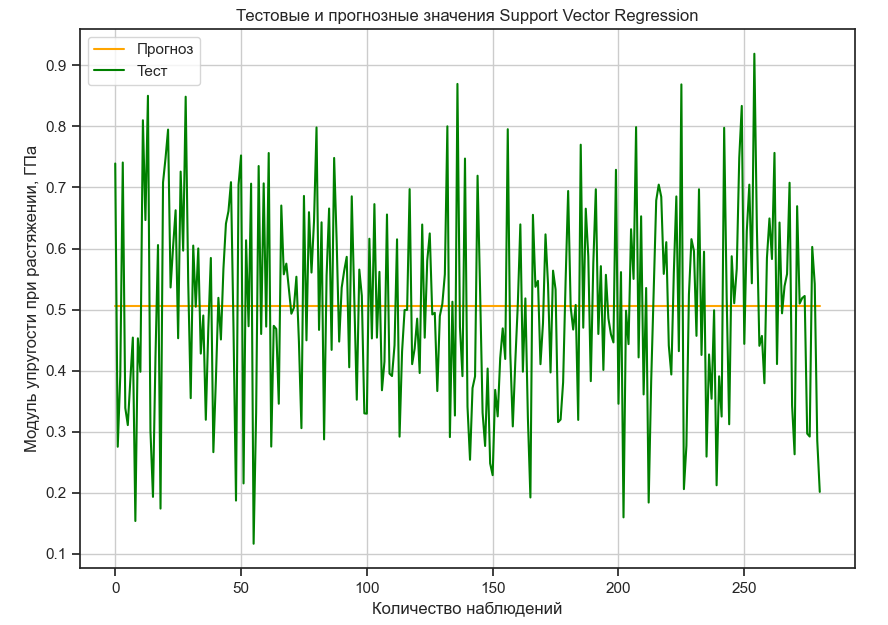


Рисунок 3 – тестовые и прогнозные значения модели Методом опорных векторов для 'Модуль упругости при растяжении, ГПа'

**Случайный лес (Random Forest)** — 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.

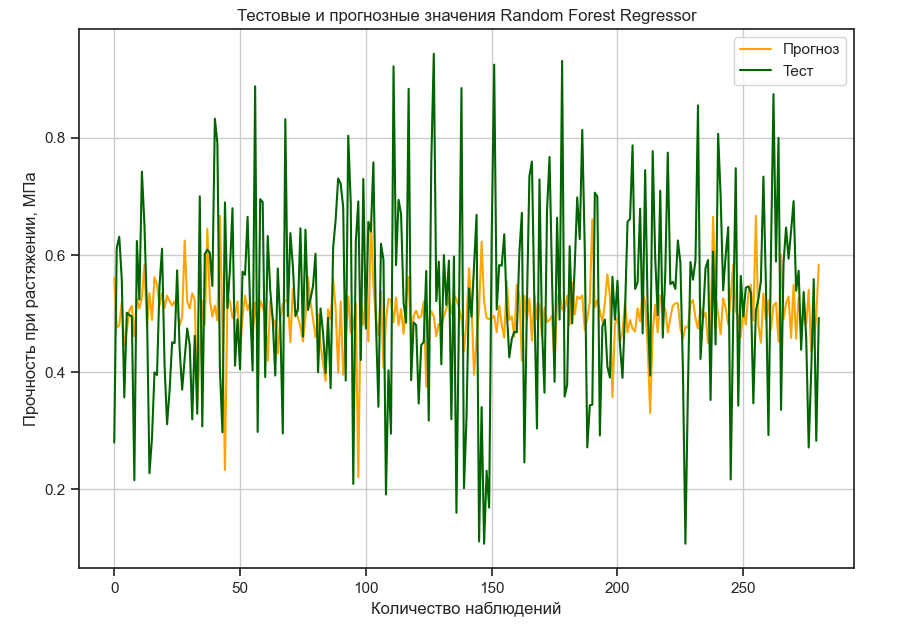


Рисунок 4 – тестовые и прогнозные значения модели Методом случайный лес для 'Прочность при растяжении, МПа '

**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 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.

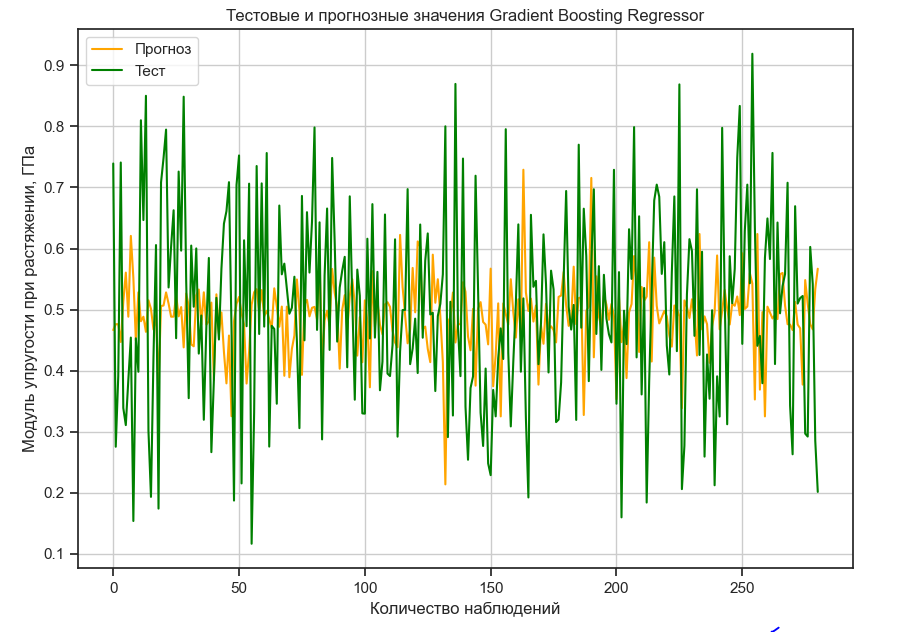


Рисунок 5 – тестовые и прогнозные значения модели Методом градиентного бустинга для 'Модуль упругости при растяжении, ГПа'

**Метод ближайших соседей - К-ближайших соседей (kNN - k Nearest Neighbours)** ищет ближайшие объекты с известными значения целевой переменной и основывается на хранении данных в памяти для сравнения с новыми элементами. Алгоритм находит расстояния между запросом и всеми примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, затем голосует за наиболее часто встречающуюся метку (в случае задачи классификации) или усредняет метки (в случае задачи регрессии).

Достоинства метода: прост в реализации и понимании полученных результатов; имеет низкую чувствительность к выбросам; не требует построения модели; допускает настройку нескольких параметров; позволяет делать дополнительные допущения; универсален; находит лучшее решение из возможных; решает задачи небольшой размерности.

Недостатки метода: замедляется с ростом объёма данных; не 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.

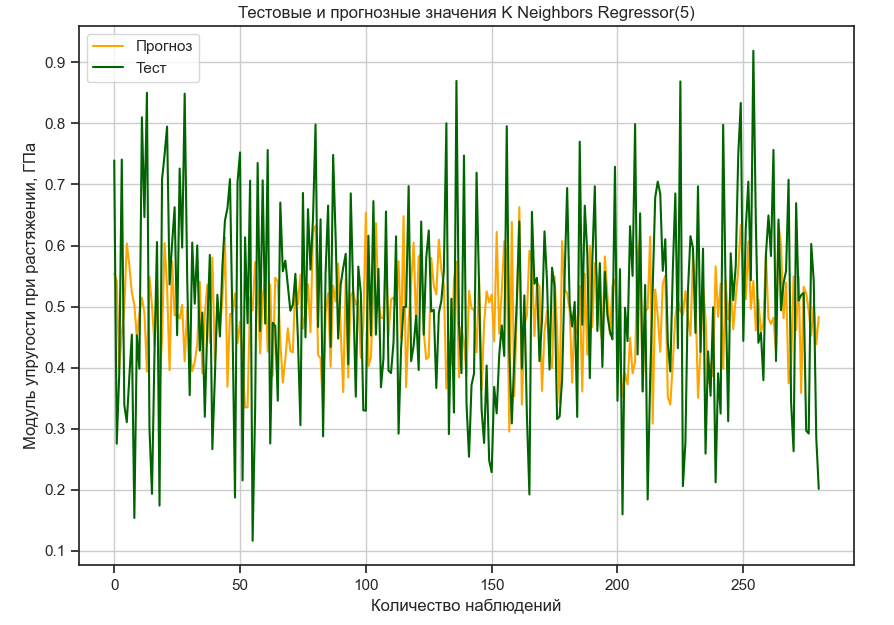


Рисунок 5 – тестовые и прогнозные значения модели Методом К-ближайших соседей для 'Модуль упругости при растяжении, ГПа'

**Дерево принятия решений (Decision Tree Regressor)** – 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: ошибается 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.

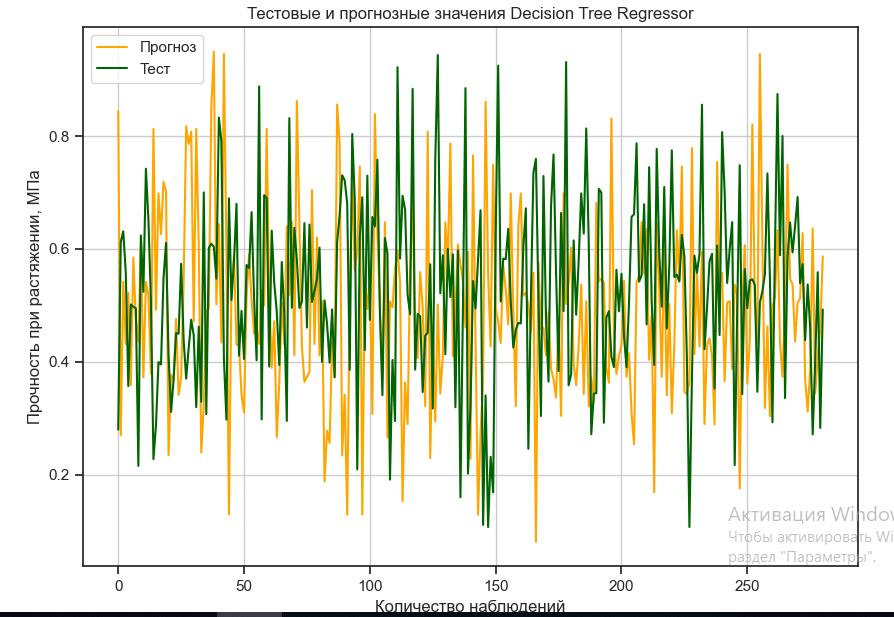


Рисунок 6 - график дерева принятия решений для Прочность при растяжении

**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.

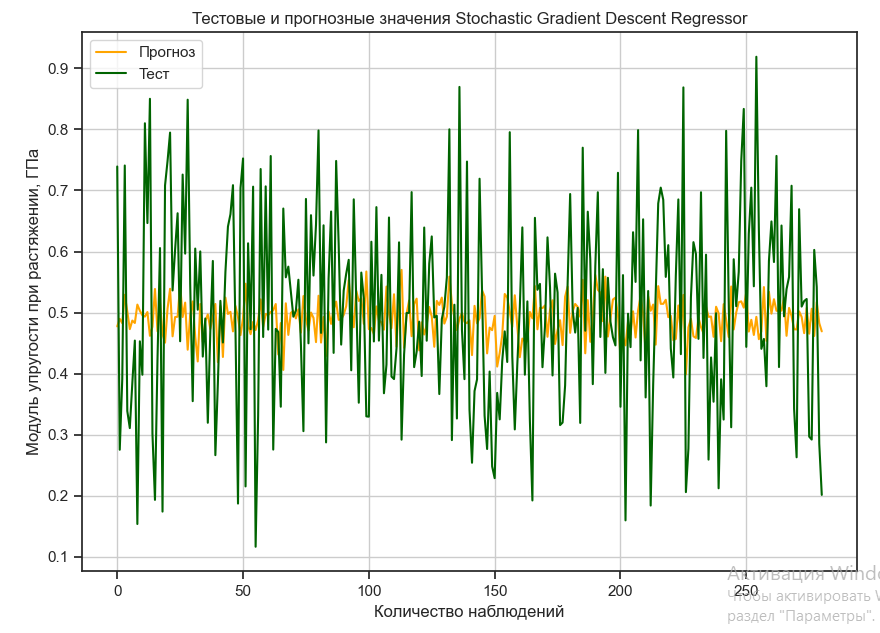


Рисунок 7 – тестовые и прогнозные значения модели Методом Стохастический градиентный спуск для 'Модуль упругости при растяжении'

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.

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

1.3 Разведочный анализ данных

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.

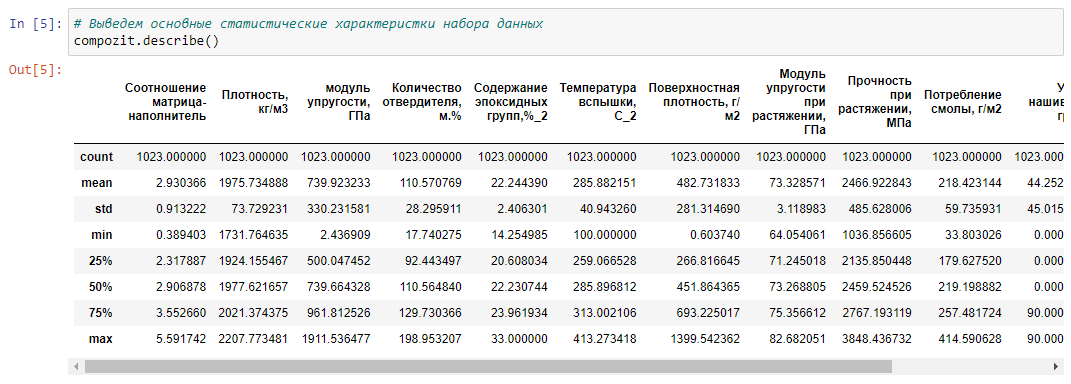


Рисунок 8 – основные статистические характеристики набора данных

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.

В качестве инструментов разведочного анализа используется: оценка статистических характеристик датасета; гистограммы распределения каждой из переменной (несколько различных вариантов); диаграммы ящика с усами (несколько интерактивных вариантов); попарные графики рассеяния точек (несколько вариантов); график «квантиль-квантиль»; тепловая карта (несколько вариантов); описательная статистика для каждой переменной; анализ и полное исключение выбросов; проверка наличия пропусков и дубликатов; ранговая корреляция Кендалла и Пирсона.

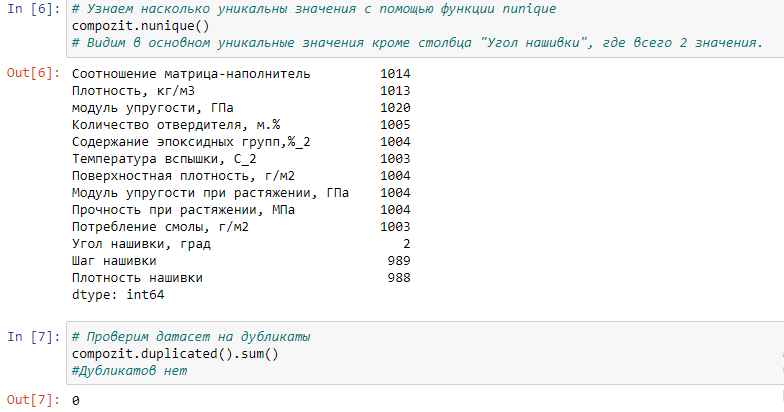


Рисунок 9 – проверка датасета на уникальность значений и дубликаты

Была проведена визуализация сырых данных и их анализ. Построены гистограммы распределения каждой из переменных и боксплоты, диаграммы "ящиков с усами", попарные графики рассеяния точек, графики квантиль-квантиль, тепловая карта без нормализации и исключения шумов.



Рисунок 10 – гистограммы распределения каждой из переменных без нормализации и исключения шумов



Рисунок 11 – график "ящиков с усами" для всех переменных

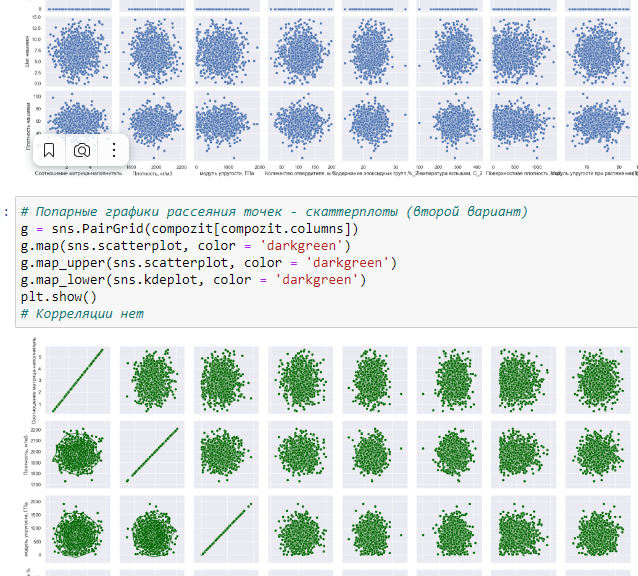


Рисунок 12 – попарные графики рассеяния точек (два разных варианта)

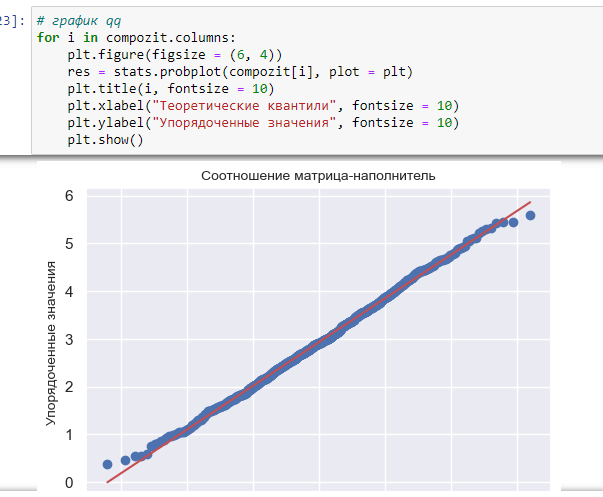


Рисунок 13 - графики «квантиль-квантиль»

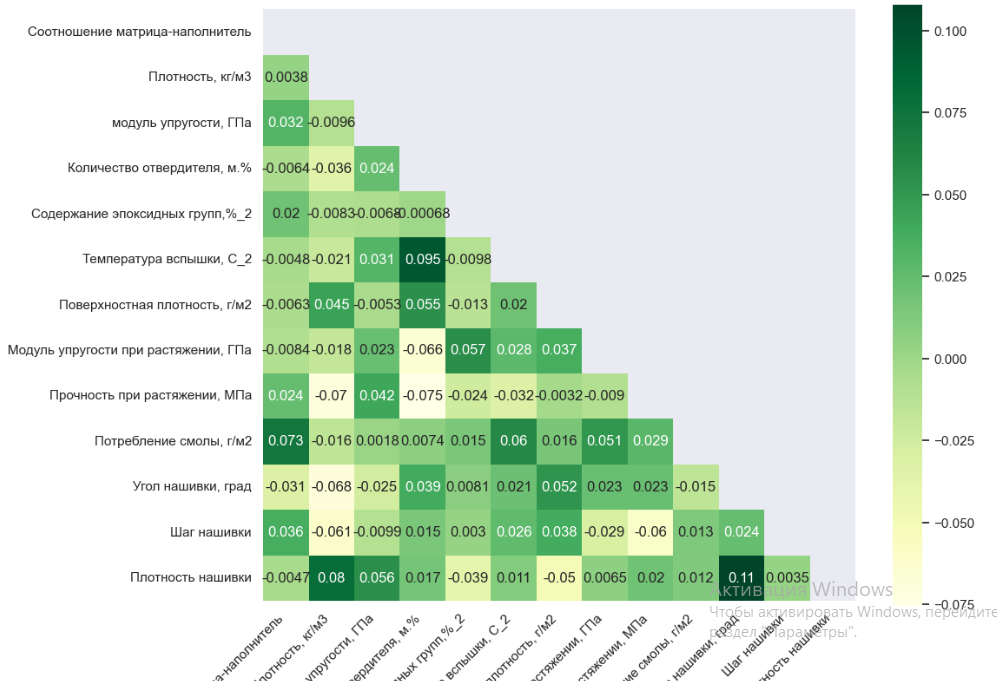


Рисунок 14 - тепловая карта с корреляцией данных

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

Данные объединённого датасета не имеют чётко выраженной зависимости, что подтверждает тепловая карта с матрицей корреляции и матрицы диаграмм рассеяния. Максимальная корреляция между плотностью нашивки и углом нашивки 0.11, значит, нет зависимости между этими данными. Корреляция между всеми параметрами очень близка к 0, то есть корреляционные связи между переменными не наблюдаются.

**2 Практическая часть**

2.1 Предобработка данных

После объединения датасета, я удалила первый неинформативный столбец с индексами.

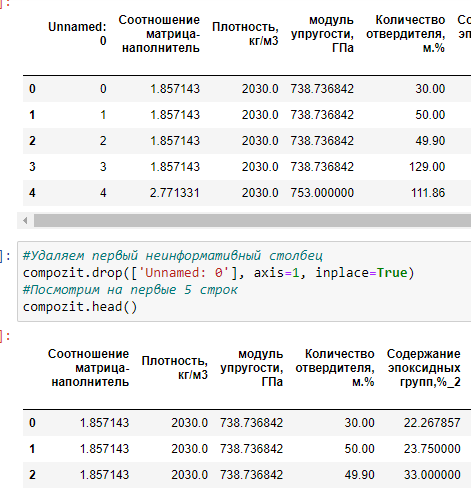


Рисунок 15 – удаление первого столбца

Затем провела с помощью MinMaxScaler нормализацию всех данных.

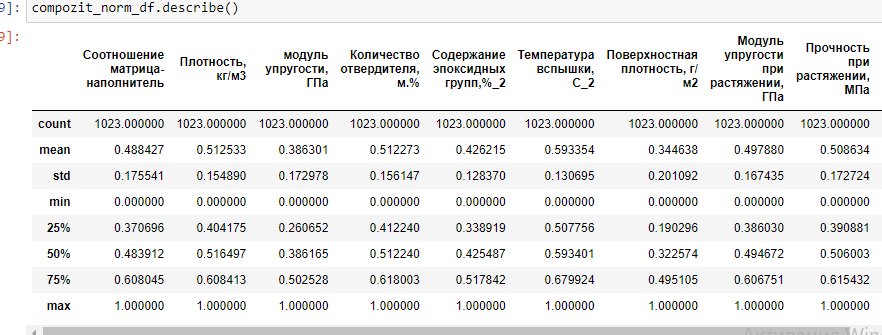


Рисунок 16 – нормализованный датасет

Потом методом межквартильного расстояния удалила выбросы.

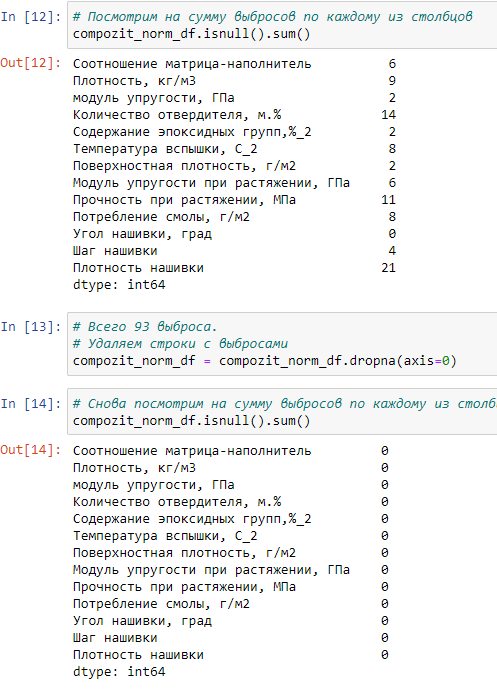


Рисунок 17 – очищенный от выбросов датасет

При работе с нейронными сетями для нормализации данных использовала Normalization.

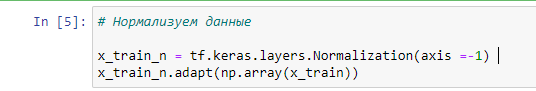


Рисунок 18 – нормализация датасета с помощью Normalization

2.2 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 применила все методы, описанные выше.

Сначала создала 1norm-модель линейной регрессии для признака «Модуль упругости при растяжении, Гпа», используя все остальные признаки

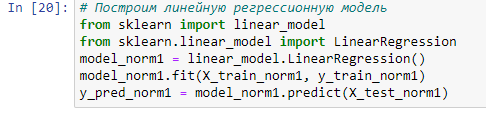


Рисунок 19 – линейная регрессионная модель

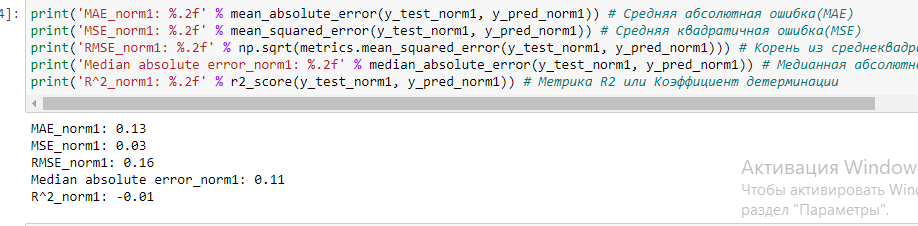


Рисунок 20 – метрики 1-й линейной регрессионной модели

Получившаяся модель плохо выдает прогнозные значения на тестовой выборке. Поэтому построила новую модель model\_norm2, используя только 2 признака 'Содержание эпоксидных групп,%\_2' и 'Потребление смолы, г/м2'.

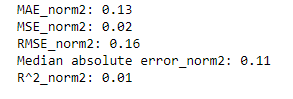


Рисунок 21 – метрики 2-й линейной регрессионной модели

Видно, что коэффициент детерминации стал положительным, хотя и слишком мал (0.01).

Затем попробовала построить модели линейной регрессии на StandardScaler-стандартизированных предварительно очищенных данных, используя для построения model\_std1-модели все признаки, а для model\_std2 - только 2 признака 'Содержание эпоксидных групп,%\_2' и 'Потребление смолы, г/м2'. Но коэффициент детерминации этих моделей не стал положительным.

Поэтому при создании 5-й модели линейной регрессии для определения значения "Прочность при растяжении, МПа", работала на MinMaxScaler-нормализованных данных по одному признаку 'модуль упругости, ГПа'. Так как хороших корреляционных связей между признаками не наблюдается. Коэффициент детерминации этой модели не стал положительным.

Таким образом, не зависимо от того какие и какое количество признаков подаются на вход модели, данные модели линейной регрессии недостаточно справляются с прогнозами целевых признаков.

В 3 части своей работы построила ряд моделей, используя следующие методы: метод опорных векторов; случайный лес; градиентный бустинг; К-ближайших соседей; дерево решений; стохастический градиентный спуск.

Для прогнозирования 'Модуль упругости при растяжении, ГПа' разбила датасет на тестовую и обучающую выборки используя признаки 'Содержание эпоксидных групп,%\_2' и 'Потребление смолы, г/м2'. А для прогнозирования 'Прочность при растяжении, МПа' использовала 'модуль упругости, ГПа'. Для визуализации тестовых и прогнозных значений использовала plt.figure. При создании моделей К-ближайших соседей использовала 5 и 20 соседей. Лучше результаты оказались при меньшем количестве соседей.

Во всех созданных моделях коэффициент детерминации получился отрицательным. Таким образом, ни одна из рассмотренных моделей не годится для прогноза ‘Модуль упругости при растяжении, ГПа’ и 'Прочность при растяжении, МПа' по предоставленному набору данных.

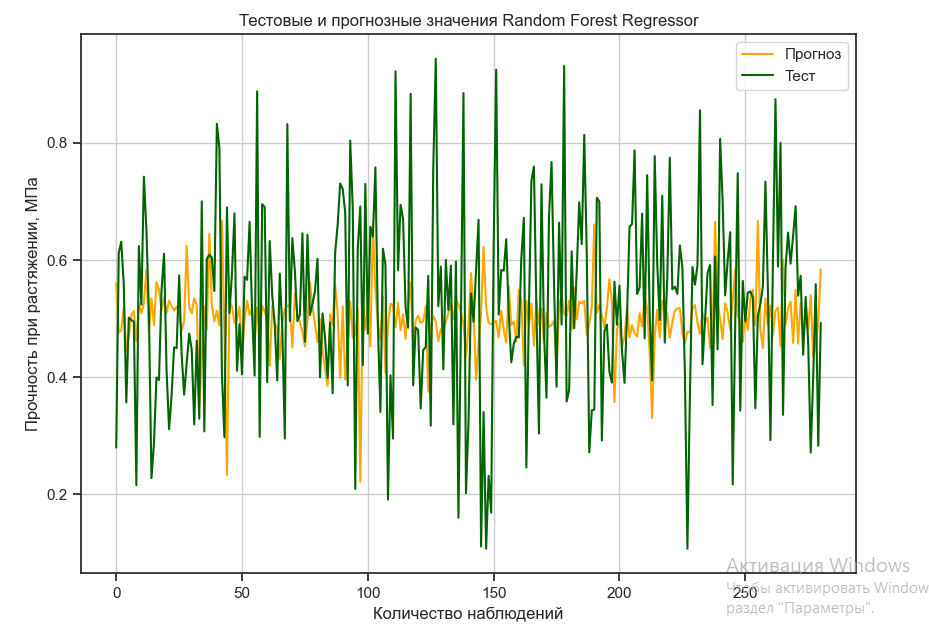


Рисунок 22 – Тестовые и прогнозные значения Прочности при растяжении методом Random Forest Regressor

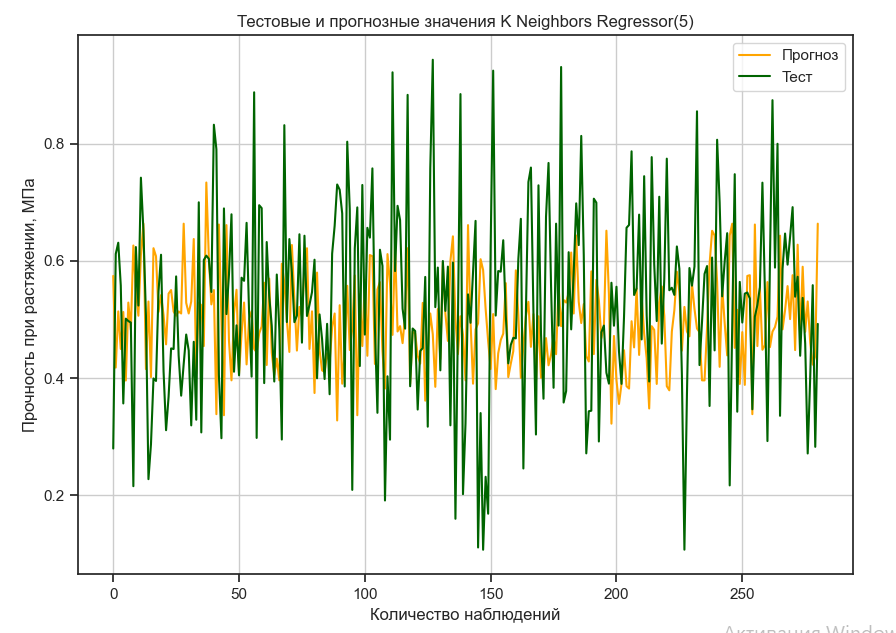


Рисунок 23 – Тестовые и прогнозные значения Прочности при растяжении методом 5-Ближайших соседей

2.3 Создание нейронной сети

При создании моделей нейронной сети использовала Sequential. Функции активации слоев relu, выходного слоя sigmoid.

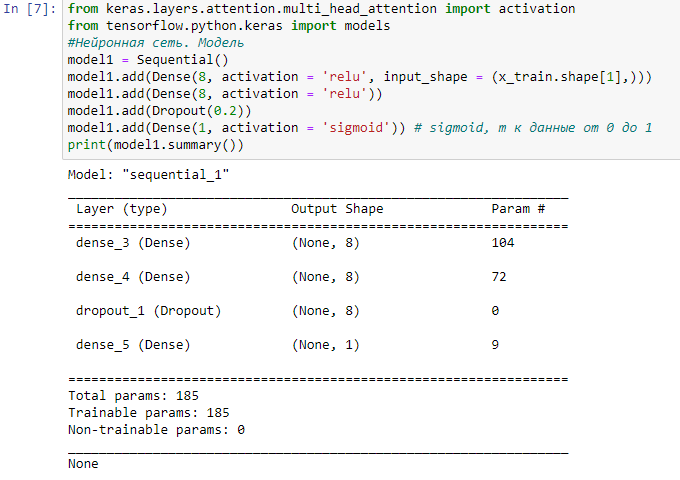


Рисунок 24 – Модель нейронной сети

Экспериментировала с количеством слоев, количеством нейронов, эпох. Коэффициент детерминации везде оказывался отрицательным. И визуализация показывает несостоятельность моделей.

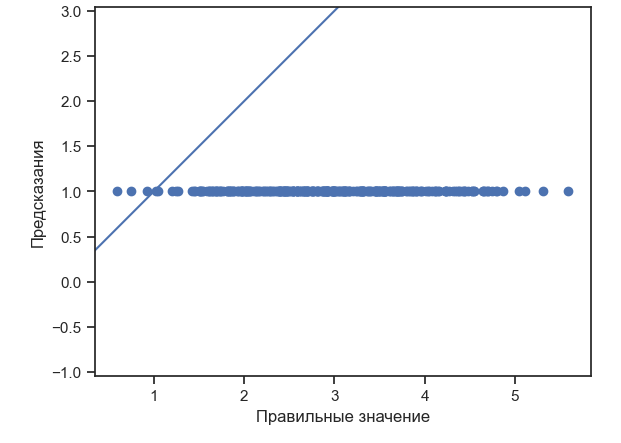


Рисунок 25 – Разброс предсказаний модели нейронной сети

2.4 Экспорт модели для создания приложения

В Google Colab удалось создать модель машинного обучения и экспортировать ее на локальный компьютер.

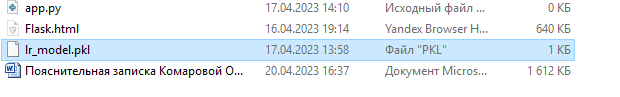


Рисунок 26 –Файл lr\_model.pkl

Дальнейшая работа по созданию интерактивных веб-приложений была отложена в связи с нехваткой времени по изучению Flask и Streamlit. А также возникшими затруднениями при работе с этими приложениями в VS Code.

2.5 Создание профиля и commit на github.com

Репозиторий был создан на github.com по адресу: <https://github.com/Oxanalex?tab=repositories> .

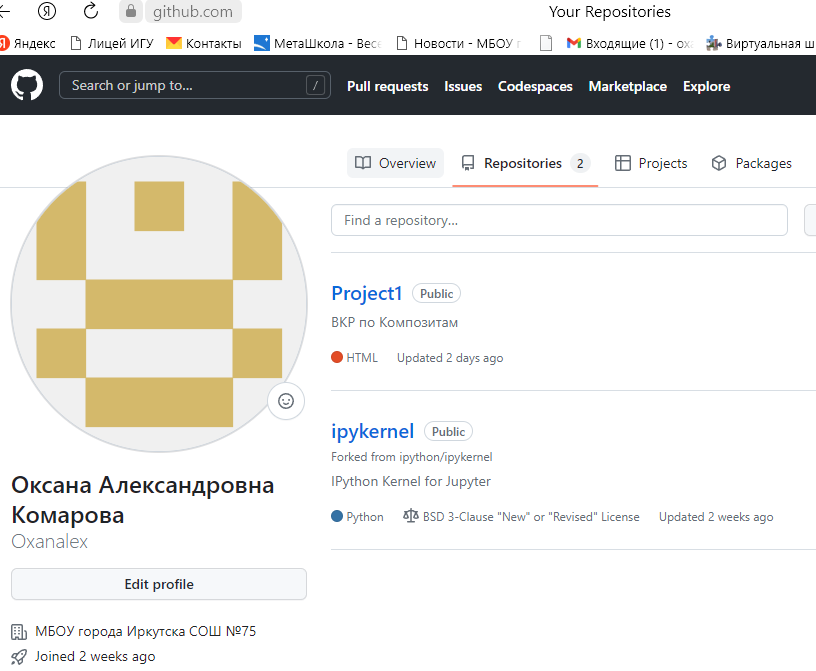


Рисунок 27 - часть страницы на github.com

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.

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 результиру ющих 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.

Приложение

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. Изучим описательную статистику каждой переменной - среднее, медиана, стандартное отклонение, минимум, максимум, квартили
   2. Проверим датасет на пропуски и дубликаты данных
   3. Получим среднее, медианное значение для каждой колонки
   4. Вычислим коэффициенты ранговой корреляции Кендалла
   5. 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 методу межквартильных расстояний
   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.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 модуля упругости при рас- тяжении и прогноза прочности при растяжении (с 30% тестовой выборки)
   1. Определим входы и выходы для моделей

6.2. Разобьём данные на обучающую и тестовую выборки

6.3.Проверим правильность разбивки

6.4.Построим модели

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

6.10.Построим и визуализируем результат работы метода деревья решений

6.11.Построим и визуализируем результат работы стохастического градиентного спуска

1. 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

8.7 Обучим нейросеть

8.8.Оценим модель

8.9.Посмотрим на потери модели

* 1. Посмотрим на график результата работы модели
  2. Посмотрим на график потерь на тренировочной и тестовой выборках
  3. Сконфигурируем другую модель, зададим слои
  4. Посмотрим на архитектуру другой модели
  5. Обучим другую модель
  6. Посмотрим на потери другой модели
  7. Посмотрим на график потерь на тренировочной и тестовой выборках
  8. Зададим функцию для визуализации факт/прогноз для результатов мо делей
  9. Посмотрим на график результата работы модели
  10. Оценим модель MSE

1. Создаём приложение
   1. Импортируем необходимые бибилиотеки
   2. Загрузим модель и определим параметры функции
   3. Получим данные из наших форм и положим их в список
2. Создание удалённого репозитория и загрузка результатов работы на него.
   1. Выгрузим все необходимые файлы в репозиторий

<https://github.com/Oxanalex?tab=repositories> .