МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«Московский государственный технический университет имени Н.Э. Баумана

(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Слушатель Ларионова Вера Юрьевна

Москва, 2023

# Содержание

[Содержание 2](#_heading=h.gjdgxs)

[Введение 3](#_heading=h.30j0zll)

[1 Аналитическая часть 4](#_heading=h.1fob9te)

[1.1 Постановка задачи 4](#_heading=h.3znysh7)

[1.2 Описание используемых методов 6](#_heading=h.2et92p0)

[1.3 Разведочный анализ данных 6](#_heading=h.tyjcwt)

[2 Практическая часть 17](#_heading=h.3dy6vkm)

[2.1 Предобработка данных 17](#_heading=h.1t3h5sf)

[2.2 Разработка и обучение модели 18](#_heading=h.4d34og8)

[2.3 Тестирование модели 18](#_heading=h.2s8eyo1)

[2.4 Нейронная сеть 19](#_heading=h.17dp8vu)

[2.5 Разработка приложения 20](#_heading=h.3rdcrjn)

[3 Создание удаленного репозитория 21](#_heading=h.26in1rg)

[Заключение 22](#_heading=h.lnxbz9)

[Библиографический список 23](#_heading=h.35nkun2)

# Введение

Данная работа выполнена в рамках курса Data Science.

В данной выпускной квалификационной работе будет рассмотрен датасет композитов, который можно использовать для анализа свойств и характеристик различных композитов. Кроме того, данный датасет может быть использован для поиска оптимальных сочетаний материалов и оптимизации процесса создания композитов.

# 1 Аналитическая часть

## 1.1 Постановка задачи

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов).

Достаточно известно определение, согласно которому: композиты – это материалы, созданные путем комбинирования двух или более различных материалов (армирующих элементов и скрепляющей их матрицы) и обладающие свойствами, отличными от суммарных свойств компонентов.

При этом предполагается, что компоненты, входящие в состав композита, должны быть хорошо совместимыми и не растворяться или иным способом поглощать друг друга.

В широком смысле композиционный материал – это любой материал с гетерогенной структурой, т. е. со структурой, состоящей минимум из двух фаз.

Такое определение позволяет отнести к композиционным материалам абсолютное большинство металлических материалов, поскольку они либо намеренно создаются многофазными, либо считаются однофазными, но в них есть неметаллические включения. Полимерные материалы также можно отнести к композитам, поскольку кроме основного компонента (полимера) в них присутствуют различные наполнители, красители и др. Материалы природного происхождения (кости человека и животных, древесина) также можно отнести к композиционным.

Например, древесина представляет собой композицию из пучков целлюлозных волокон трубчатого строения, скрепленных матрицей из органического вещества – лигнина.

Для того чтобы выделить композиционные материалы искусственного происхождения, подчеркнуть их характерные особенности наиболее полным считается определение, согласно которому к композитам относятся материалы, обладающие рядом признаков:

1. состав, форма и распределение компонентов материала «запроектированы заранее»;

2. материал не встречается в природе, а создан человеком;

3. материал состоит из двух или более компонентов, различающихся по химическому составу и разделенных выраженной границей;

4. свойства материала определяются каждым из его компонентов,

которые должны присутствовать в материале в достаточно больших количествах (больше некоторого критического содержания);

5. материал обладает такими свойствами, которых не имеют его

компоненты, взятые в отдельности;

6. материал неоднороден в микромасштабе и однороден в макромасштабе

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если известны характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

В качестве входных данных приняты данные о начальных свойствах компонентов композиционных материалов:

* Соотношение матрица-наполнитель ;
* Плотность;
* Модуль упругости;
* Количество отвердителя;
* Содержание эпоксидных групп;
* Температура вспышки;
* Поверхностная плотность;
* Модуль упругости при растяжении;
* Прочность при растяжении;
* Потребление смолы;
* Угол нашивки;
* Шаг нашивки;
* Плотность нашивки.

Общее количество параметров для анализа – 13.

Для исследовательской работы были даны 2 файла: X\_bp.xlsx (с данными о параметрах, состоящий из 1023 строк и 10 столбцов данных) и X\_nup.xlsx (данными нашивок, состоящий из 1040 строк и 3 столбцов данных).

Для разработки моделей по прогнозу модуля упругости при растяжении, прочности при растяжении и соотношения «матрица-наполнитель нужно объединить 2 файла. Датасеты были объединены по индексу тип объединения INNER, поэтому часть информации (17 строк таблицы X\_nup.xlsx) не имеет соответствующих строк в таблице X\_bp.xlsx и будет удалена.

Итоговый датасет нормализованный, пропуски отсутствуют. Элементы массива соответствуют типу float64.

На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков; сделать предобработку: 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

## 1.2 Описание используемых методов

Для решения поставленной задачи, которая 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 были

исследованы (и некоторые из них применены) следующие методы:

− линейная регрессия (Linear regression);

− лассо регрессия (Lasso);

− гребневая регрессия (Ridge);

− эластичная регрессия (ElasticNet) ;

− градиентный бустинг(GradientBoostingRegressor);

− К-ближайших соседей (KNeighborsRegressor);

− дерево решений (DecisionTreeRegressor);

− случайный лес (RandomForest);

− градиентный бустинг (AdaBoostRegressor);

− стохастический градиентный спуск (SGDRegressor);

− метод опорных векторов (Support Vector Regression);

− многослойный перцептрон.

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

- R2 или коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то прогнозы сопоставимы по качеству с константным предсказанием;

- RMSE (Root Mean Squared Error) или корень из средней квадратичной ошибки принимает значениях в тех же единицах, что и целевая переменная. Метрика использует возведение в квадрат, поэтому хорошо обнаруживает грубые ошибки, но сильно чувствительна к выбросам;

- MAE (Mean Absolute Error) - средняя абсолютная ошибка так же принимает значениях в тех же единицах, что и целевая переменная;

- MAPE (Mean Absolute Percentage Error) или средняя абсолютная процентная ошибка — безразмерный показатель, представляющий собой взвешенную версию MAE;

- max error или максимальная ошибка данной модели в единицах измерения целевой переменной.

RMSE, MAE, MAPE и max error принимают положительные значения. Но отображать я их буду со знаком «-». Так корректно отработает выделение цветом лучших моделей — эти метрики надо минимизировать.

R2 (коэффициент детерминации) измеряет долю дисперсии, объяснённую моделью, вобщей 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 коэффициентадетермина ции означают плохую объясняющую способность модели.

## 1.3 Разведочный анализ данных

Прежде чем передать данные в работу моделей машинного обучения, необходимо обработать и очистить их. Необработанные данные могут содержать искажения и пропущенные значения и способны привести к неверным результатам по итогам моделирования. Но безосновательно удалять что-либо тоже неправильно. Именно поэтому сначала набор данных надо изучить.

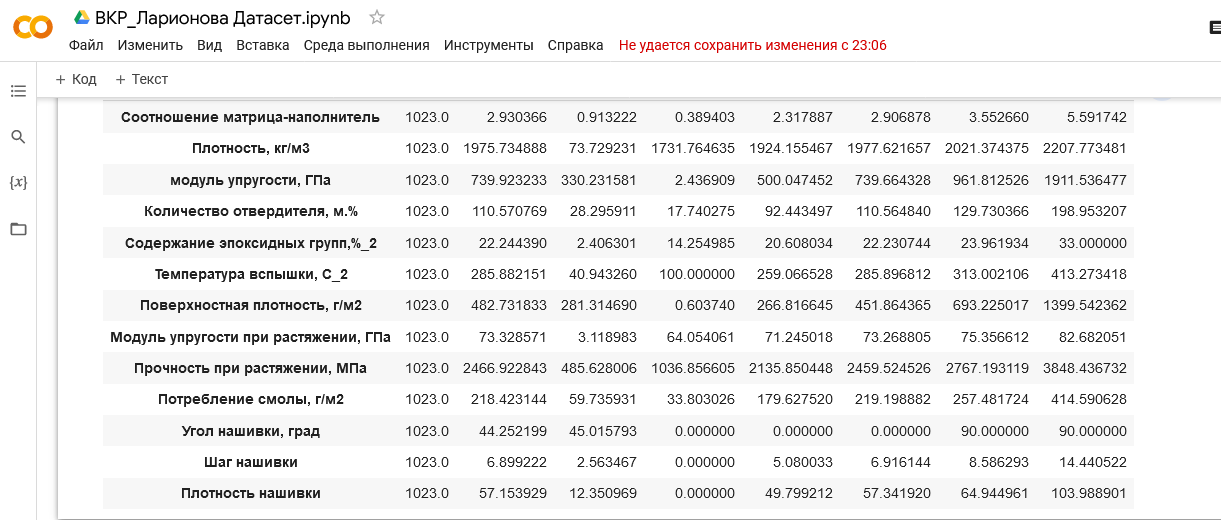


Рисунок 1 - Описательная статистика датасета

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.

С помощью построения гистограмм было выявлено распределение величин, близкое к нормальному, для большей части параметров.

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

С помощью построения гистограмм было выявлено распределение величин, близкое к нормальному, для большей части параметров.

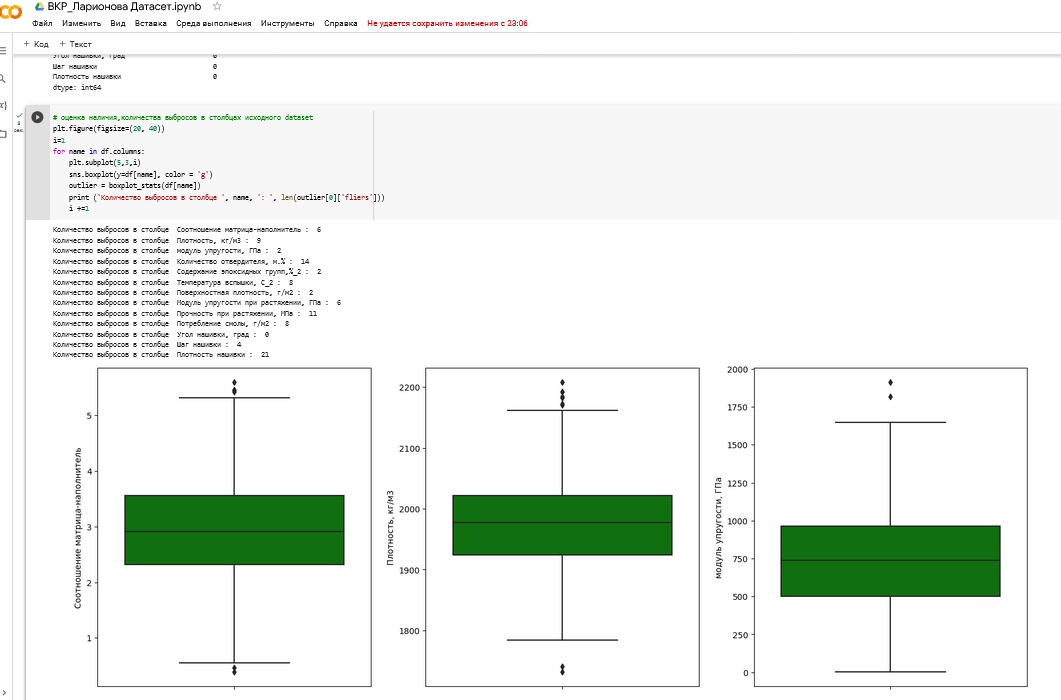


Рисунок 2 – Начальное количество выбросов

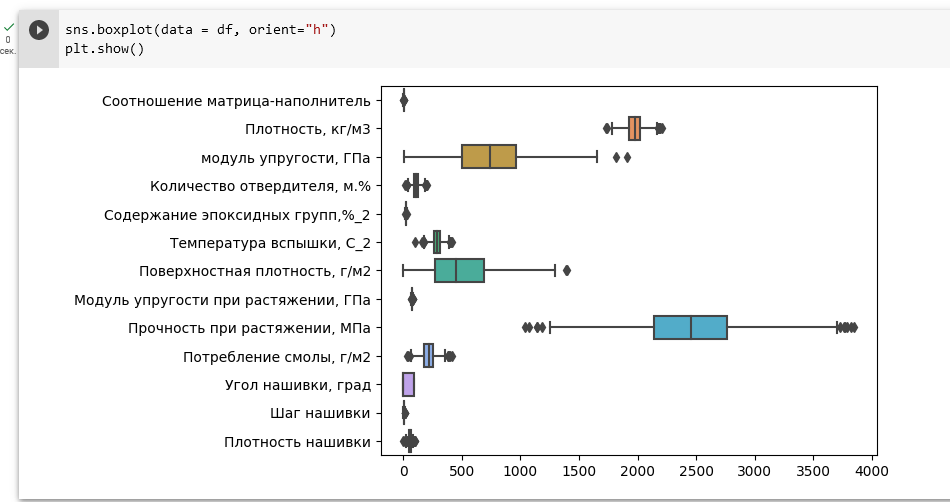


Рисунок 3 – Начальный boxplot

Для удаления выбросов используются методы трех сигм и межквартильного расстояния. В данном случае удалим способом межквартильного расстояния для максимальной чистоты, так как используем методы чувствительные к выбросам.

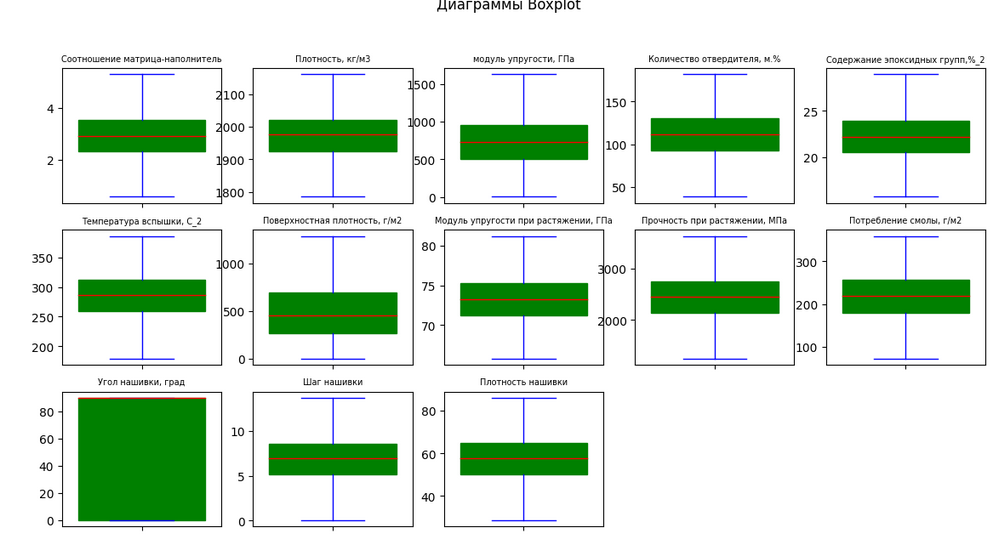


Рисунок 5 – Boxplot после удаления выбросов

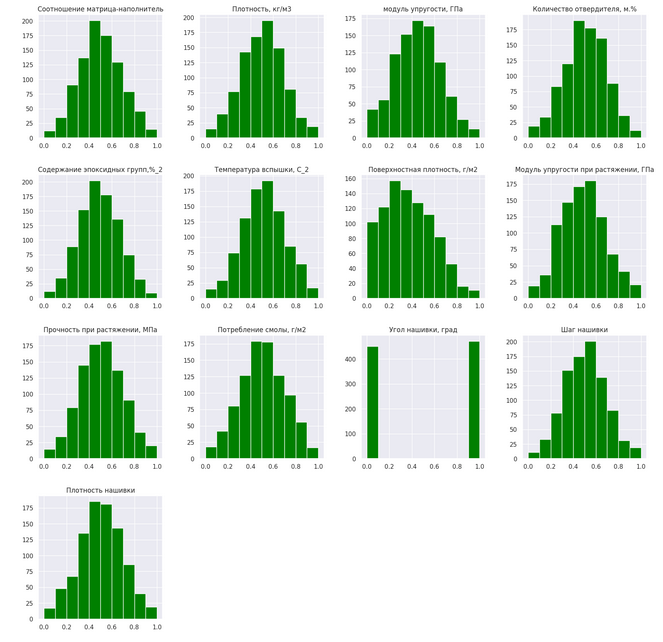


Рисунок 6 – Гистрограммы распределения

С помощью построения гистограмм было выявлено распределение величин, близкое к нормальному, для большей части параметров, за исключением признака. Угол нашивки, который имеет всего два значения 0 и 90 градусов.

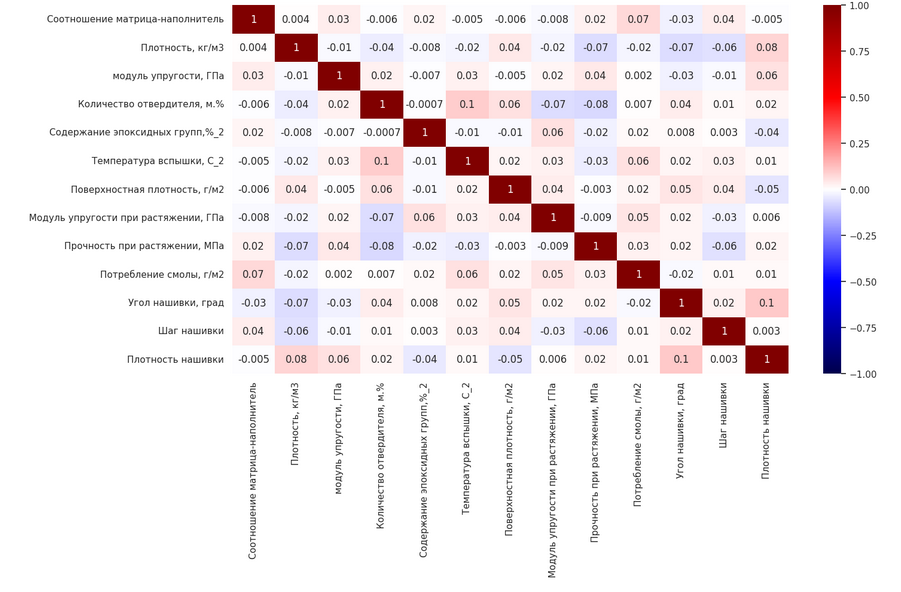


Рисунок 6 – Тепловая карта с корреляцией данных

Тепловая карта показывает практически отсутствие корреляции между признаками и целевыми переменн

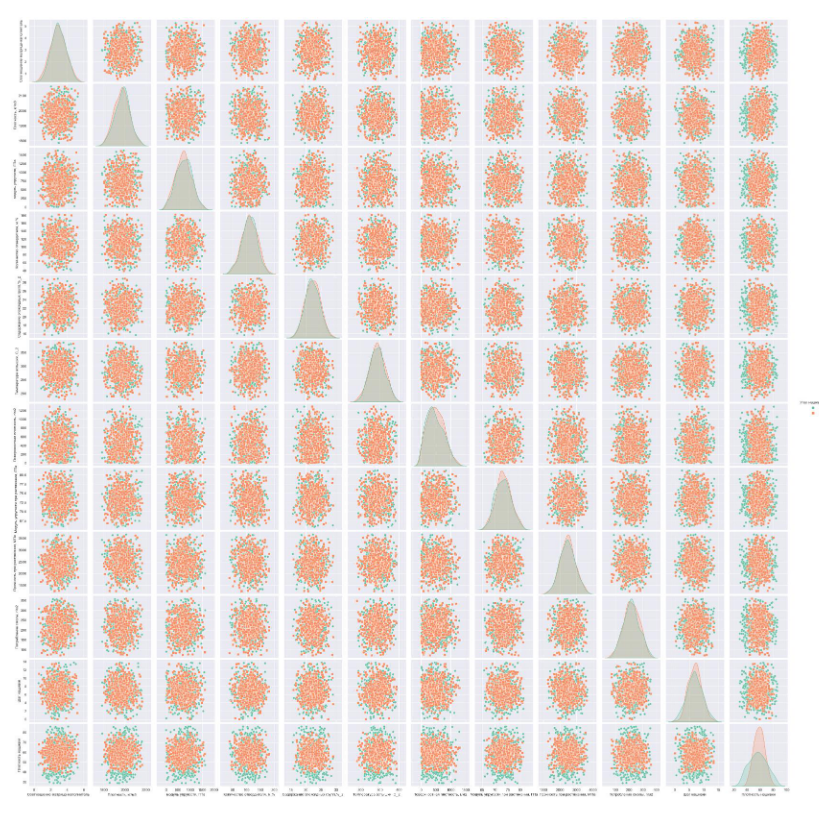


Рисунок 7 - Попарные графики рассеяния точек с выделением значений Угол нашивки

На попарных графиках распределения не видно корреляции между признаками. Единственная зависимость, которую можно отметить – это меньшая дисперсия значений Плотности нашивки при 90 градусов Угла нашивки, по сравнению со значением 0 градусов.

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.

**2 Практическая часть**

## 2.1 Предобработка данных

Предобработка данных осуществлялась на основании разведочного анализа данных, который показал наличие выбросов. Для удаления выбросов был произведен расчет количества выбросов для каждого параметра/ Количество выбросов говорит о их незначительности, соответственно, данные значения можно удалить из датасета.

Размер датасета после удаления выбросов и проверка наличия пропусков:

* Количество строк в очищенном датасете: 918;
* Количество столбцов (переменных) в очищенном датасете: 13;
* Количество пропусков в данных очищенного датасета: 0.

Вызов команды df\_used.describe() для датасета показывает, что количество строк датасета уменьшилось. Датасет очищен от выбросов.

Для выделения наиболее весомых признаков датасета использован факторный анализ.

С помощью метода главных компонент были получены значения влияния 90% факторов Указанные значения позволяют сделать вывод о том, что влияние очень слабое.

С целью определения весов характеристик датасета для первых пяти факторов влияния был проведен анализ, последовательно добавляя по одному фактору. Полученные результаты

* Для двух факторов: выделить наименование и смысл нового фактора сложно. но можно увидеть высокие доли характеристик внутри факторов: модуль упругости, ГПа, Температура вспышки, С\_2 Плотность нашивки;
* Для трех факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки;
* Для четырех факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки Угол нашивки, град;
* Для пяти факторов: выделить наименование и смысл нового фактора сложно. Но можно увидеть высокие доли характеристик внутри факторов: Количество отвердителя, м.% Плотность, кг/м3 Плотность нашивки Угол нашивки, град Потребление смолы, г/м2.

Анализ полученных результатов показывает, что возможно характеристики: Количество отвердителя, Плотность нашивки, Плотность, Угол нашивки, Потребление смолы, являются наиболее существенными для построения будущих моделей.

Для дальнейшей разработки и обучения модели была выполнена нормализация данных с помощью MinMaxScaler.

## 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 применим все методы, описанные выше.



Рисунок 9- Поиск гиперпараметров по сетке

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%); обучение моделей на нормализованных значениях; сравнение моделей по метрике МАЕ; поиск гиперпараметров, по которым будет происходить оптимизация модели, с помощью выбора по сетке и перекрёстной проверки. Оценка полученных результатов работы моделей. В качестве параметра оценки выбран также коэффициент детерминации (R2).



Рисунок 10 - Модель Ridge регрессия

Модели после настройки гиперпараметров показали результат немного лучше. В результате все модели показали примерно одинаковый результат: ошибка MAE примерно равна стандартному отклонению, значения R2 находятся около нуля, то есть все модели предсказывают результат сопоставимый со средним значением. Можно считать, что все примененные модели не справились с задачей, результат неудовлетворительный. Для улучшения работы алгоритмов можно уменьшить количество признаков. Посмотрим коэффициенты вклада признаков в результат в моделях Lasso и Ridge.

Для улучшения работы алгоритмов можно уменьшить количество признаков. Посмотрим коэффициенты вклада признаков в результат в моделях

22

Lasso и Ridge.

Рисунок – Коэффициенты признаков Lasso регрессии

Рисунок – Коэффициенты признаков Ridge регрессии

Как видно, коэффициенты вклада в результат близки для всех признаков к нулю. Поэтому не вижу смысла в уменьшении размерности.

23

Рисунок - Оценка MAE результатов работы моделей

Рисунок - Оценка R2 результатов работы моделей

2.4. Нейронная сеть для рекомендации «Соотношение матрица-наполнитель».

Загружаем очищенный датасет, для X удаляем целевой столбец «Соотношение матрица-наполнитель» и сохраняем его в у. Далее разбиваем на обучающую и тестовую выборку (для вычисления MAE). Затем делаем нормализацию, и снова разбиение данных уже для подачи на вход нейросети.

Рисунок 15 – Разбиение и нормализация данных для нейросети

24

Создаем архитектуру нейронной сети и запускаем обучение. Оценивая результаты меняем параметры нейросети: количество нейронов, функции активации, количество слоев, добавление слоя Dropout.

Рисунок 16 – Ошибка MAE по результатам работы нейросетей с различной архитектурой

Рисунок 17 — Код нейросети с наименьшей ошибкой

25

Рисунок 18 — Визуализация работы нейросети с наименьшей ошибкой

Все нейросети показали схожий результат с ошибкой MAE чуть меньшей, чем среднее отклонение.

Разработка приложения

Создание приложения для расчета параметра «Соотношение матрица-наполнитель». Данное приложение — это основной файл Flask, папка templates, с Шаблоном html - страницы, папка mn\_model\_nn c сохранённой моделью.

Рисунок 19 - Код приложения

26

Рисунок 20 - Форма пользовательского приложения для ввода параметров

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

Рисунок 21 – Результат расчета Соотношение матрица-наполнитель

27

2.5. Создание удалённого репозитория и загрузка Репозиторий был создан на github.com по адресу : https://github.com/VeraLarionova/proverka\_vkr/

Рисунок 23 - часть созданного файла README

28

2.6. Заключение

Данная исследовательская работа позволяет сделать некоторые основные выводы по теме. Распределение полученных данных в объединённом датасете близко к нормальному, корреляция между парами признаков близка к нулю. Использованные при разработке моделей подходы не позволили получить достоверных прогнозов. Применённые модели регрессии не показали эффективности в прогнозировании свойств композитов.

Можно сделать 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, консультации экспертов предметной области, новые исследования. В целом прогнозирование конечных свойств/характеристик композитных материалов без изучения материаловедения, погружения в вопрос экспериментального анализа характеристик композитных материалов не демонстрирует удовлетворительных результатов. Учитывая отсутствие корреляции между признаками, делаем вывод, что текущим набором алгоритмов задача не решается, возможно, решается трудно или не решается совсем.

29

2.7. Список используемой литературы и веб ресурсы.

1. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour) https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour2021-07-19.

2. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам сле-дует знать: https://habr.com/ru/company/vk/blog/513842/

3. Anthony Schams. Смещение, дисперсия и регуляризация в линейной регрессии: лассо, хребет и эластичная сеть - различия и использовании https://machinelearningmastery.ru/bias-variance-and-regularization-in-linear-regression-lasso-ridge-and-elastic-net-8bf81991d0c5/

4. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.

5. Абросимов Н.А Методика построения разрешающей системы уравнений динамического деформирования композитных элементов конструкций (Учебно-методическое пособие), ННГУ, 2010

6. Абу-Хасан Махмуд, Масленникова Л. Л. Прогнозирование свойств композиционных материалов с учётом наноразмера частиц и акцепторных свойств катионов твёрдых фаз, статья 2006 год

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. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.9.

9. Документация по библиотеке keras: https://keras.io/api/.

10. Документация по библиотеке matplotlib: https://matplotlib.org/stable/users/index.html.

11. Документация по библиотеке numpy: https://numpy.org/doc/1.22/user/index.html#user.

12. Документация по библиотеке pandas: https://pandas.pydata.org/docs/user\_guide/index.html#user-guide.

13. Документация по библиотеке scikit-learn: https://scikit-learn.org/stable/user\_guide.html.

30

14. Документация по библиотеке seaborn: https://seaborn.pydata.org/tutorial.html.

15. Документация по библиотеке Tensorflow: https://www.tensorflow.org/overview

16. Документация по языку программирования python: https://docs.python.org/3.8/index.html.

17. Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.

18. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) : https://habr.com/ru/post/428503/

19. Ларин А. А., Способы оценки работоспособности изделий из композиционных материалов методом компьютерной томографии, Москва, 2013, 148 с.

20. Макс Ветков. Градиентный бустинг(AdaBoost) https://your-scorpion.ru/gradient-boosting-adaboost

21. Материалы конференции: V Всероссийская научно-техническая конференция «Полимерные композиционные материалы и производственные технологии нового поколения», 19 ноября 2021 г.

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://flaskrussian-docs.readthedocs.io/ru/latest/quickstart.html. (дата обращения: 09.06.2022)