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

Сотникова Анастасия Андреевна

Москва, 2023

Содержание

[1. Введение 3](#_Toc117941029)

[2. Аналитическая часть 5](#_Toc117941030)

[2.1. Постановка задачи 5](#_Toc117941031)

[2.2. Описание используемых методов 11](#_Toc117941032)

[2.2.1. Линейная регрессия и метод наименьших квадратов (МНК) 12](#_Toc117941033)

[2.2.2. Гребневая регрессия или ридж-регрессия 13](#_Toc117941034)

[2.2.3. Опорная векторная регрессия 13](#_Toc117941035)

[2.2.4. Стохастический градиентный спуск 14](#_Toc117941036)

[2.2.5. Деревья решений 14](#_Toc117941037)

[2.2.6. Градиентный бустинг 15](#_Toc117941038)

[2.2.7. Метод случайного леса 15](#_Toc117941039)

[2.2.8. Нейронная сеть 16](#_Toc117941040)

[2.3. Разведочный анализ данных 17](#_Toc117941041)

[2.3.1. Матрица корреляции 17](#_Toc117941042)

[2.4. Нормализация данных 18](#_Toc117941043)

[3. Практическая часть 20](#_Toc117941044)

[3.1. Предобработка данных 20](#_Toc117941045)

[3.2. Разработка и обучение модели 21](#_Toc117941046)

[3.3. Тестирование модели 24](#_Toc117941047)

[3.4. Разработка нейронной сети, которая будет рекомендовать соотношение матрица-наполнитель. 25](#_Toc117941048)

[3.4.1. Бибилиотека PyTorch 26](#_Toc117941049)

[3.4.2. TensorFlow 27](#_Toc117941050)

[3.5. Разработка приложения 29](#_Toc117941051)

[3.6. Создание удаленного репозитория и загрузка результатов работы на него. 33](#_Toc117941052)

[Заключение 34](#_Toc117941053)

[Библиографический список 35](#_Toc117941054)

# Введение

Тема данной работы - прогнозирование конечных свойств новых материалов (композиционных материалов).

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.

Структура композиционных материалов представляет собой матрицу (основной компонент), содержащую в своем объеме армирующие элементы, часто называемые наполнителем. Матрица и наполнитель разделены границей (поверхностью) раздела. Наполнитель равномерно распределен в матрице и имеет заданную пространственную ориентацию.

Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем 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).

На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

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

Датасет со свойствами композитов, использовавшийся в данной работе доступен по ссылке в сети интернет. Объединение делалось по индексу, тип объединения INNER

# Аналитическая часть

## Постановка задачи

В данной работе исследуется композит с матрицей из базальтопластика и нашивками из углепластика. Для работы получен датасет, содержащий данные о свойствах матрицы и наполнителя, производственных параметрах и свойствах готового композита. Требуется разработать модели, прогнозирующие значения некоторых свойств в зависимости от остальных. Так же требуется разработать приложение, позволяющее использовать результирующую обученную модель специалистом предметной области.

В соответствии с заданием требуется:

1) Изучить теоретические основы и методы решения поставленной задачи.

2) Провести разведочный анализ предложенных данных. Необходимо нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек. Необходимо также для каждой колонке получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков.

3) Провести предобработку данных (удаление шумов, нормализация и т.д.).

4) Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении. При построении модели необходимо 30% данных оставить на тестирование модели, на остальных происходит обучение моделей. При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.

5) Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.

6) Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в задании 4 или 5 (один или два прогноза, на выбор учащегося).

7) Оценить точность модели на тренировочном и тестовом датасете.

8) Создать репозиторий в GitHub / GitLab и разместить там код исследования. Оформить файл README.

Датасет состоит из двух файлов в формате MS Excel: X\_bp (составляющая из базальтопластика) и Х\_nup (составляющая из углепластика).

Файл X\_bp содержит 10 признаков и индекс, всего строк: 1023.

Файл X\_nup содержит 3 признака и индекс, всего строк: 1040.

Для получения итогового датасета для анализа необходимо объединить два входных файла типом INNER по индексу. После объединения часть строк из файла X\_nup была отброшена в связи с неперсекающимися индексами в файле X\_bp. Итоговой датасет содержит 13 признаков и 1023 строки (объектов).

Описание признаков объединенного датасета приведено в таблице 1. Все признаки имеют вещественный тип. Пропусков в данных нет, все признаки по всем объектам имеют вещественные значения. Все признаки, кроме «Угол нашивки», являются количественными. Признак «Угол нашивки» принимает только два значения 0 град и 90 град. Это категориальный признак, состоящий из двух категорий.

Таблица - Описание основных признаков датасета

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Наименование признака | Коли-чество | Среднее | Стандарт-ное от-клонение | Мини-мум | Медиана | Максимум |
| Соотношение матрица-наполнитель | 1023 | 2.930 | 0.913 | 0.389 | 2.907 | 5.592 |
| Плотность, кг/м3 | 1023 | 1975.735 | 73.729 | 1731.765 | 1977.622 | 2207.773 |
| модуль упругости, ГПа | 1023 | 739.923 | 330.232 | 2.437 | 739.664 | 1911.536 |
| Количество отвердителя, м.% | 1023 | 110.571 | 28.296 | 17.740 | 110.565 | 198.953 |
| Содержание эпоксидных групп,%\_2 | 1023 | 22.244 | 2.406 | 14.255 | 22.231 | 33.000 |
| Температура вспышки, С\_2 | 1023 | 285.882 | 40.943 | 100.000 | 285.897 | 413.273 |
| Поверхностная плотность, г/м2 | 1023 | 482.732 | 281.315 | 0.604 | 451.864 | 1399.542 |
| Модуль упругости при растяжении, ГПа | 1023 | 73.329 | 3.119 | 64.054 | 73.269 | 82.682 |
| Прочность при растяжении, МПа | 1023 | 2466.923 | 485.628 | 1036.857 | 2459.525 | 3848.437 |
| Потребление смолы, г/м2 | 1023 | 218.423 | 59.736 | 33.803 | 219.199 | 414.591 |
| Угол нашивки, град | 1023 | 44.252 | 45.016 | 0.000 | 0.000 | 90.000 |
| Шаг нашивки | 1023 | 6.899 | 2.563 | 0.000 | 6.916 | 14.441 |
| Плотность нашивки | 1023 | 57.154 | 12.351 | 0.000 | 57.342 | 103.989 |

Гистограммы распределения переменных и диаграммы «ящик с усами» приведены на рисунках 1 и 2. Из них видно, что распределение всех признаков, за исключением признака «Угол нашивки» близко к нормальному распределению. Признак «Угол нашивки» распределён в соответствии с распределением Бернулли. Он принимает два значения 0.0 град и 90.0 град. Подробнее про преобразование категориального признака изложено в разделе 3.1.

Попарные графики рассеяния точек приведены на рисунке 3. На главной диагонали приведены диаграммы распределения признаков.

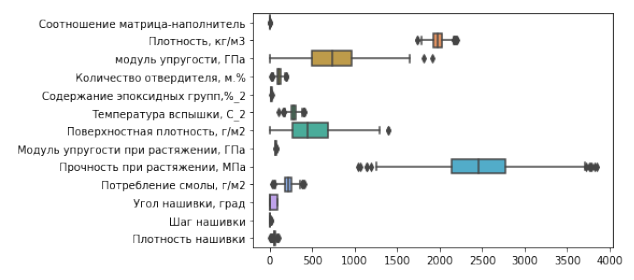


Рисунок 1 - Гистограммы распределения признаков датасета

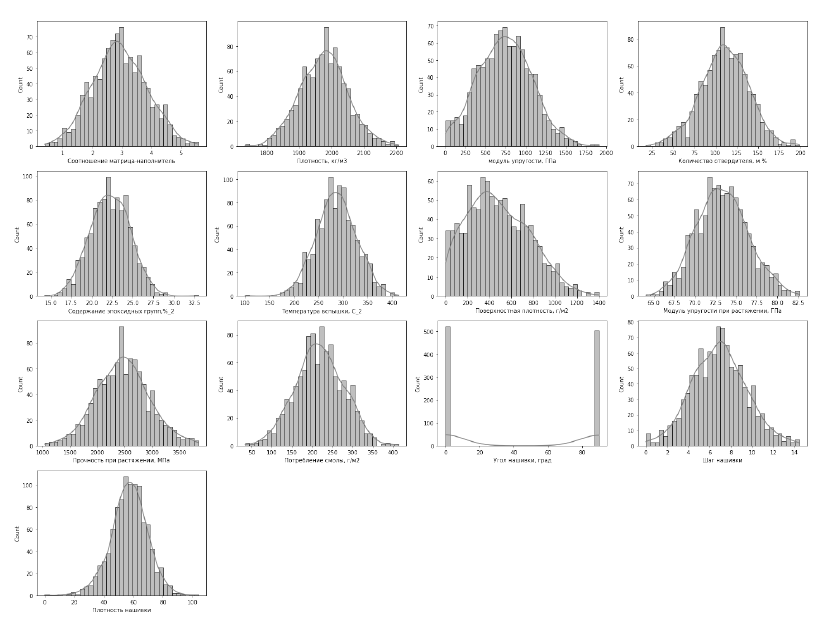
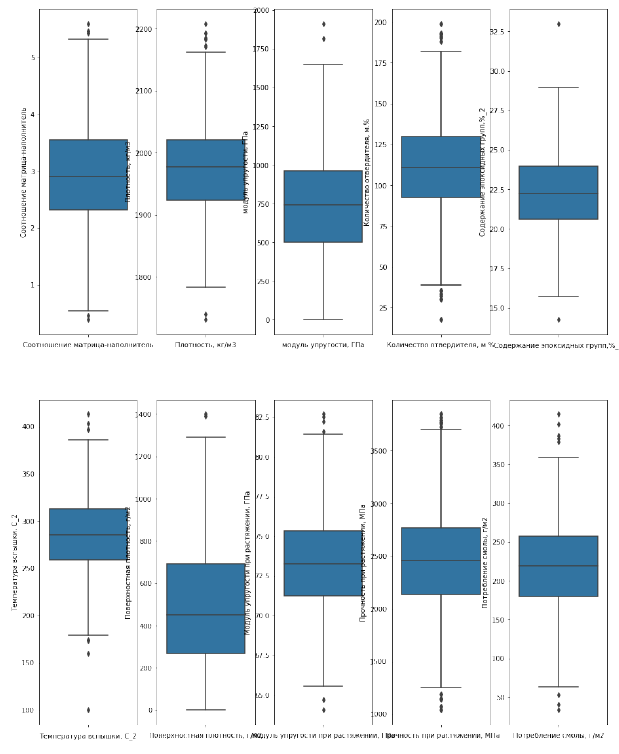


Рисунок 2 - Попарные графики рассеяния



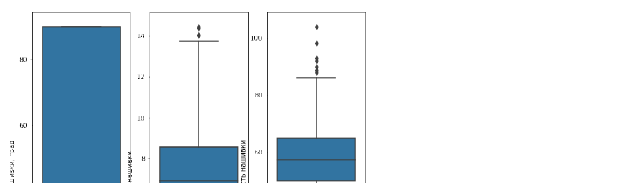


Рисунок 3 - Диаграммы "ящик с усами"

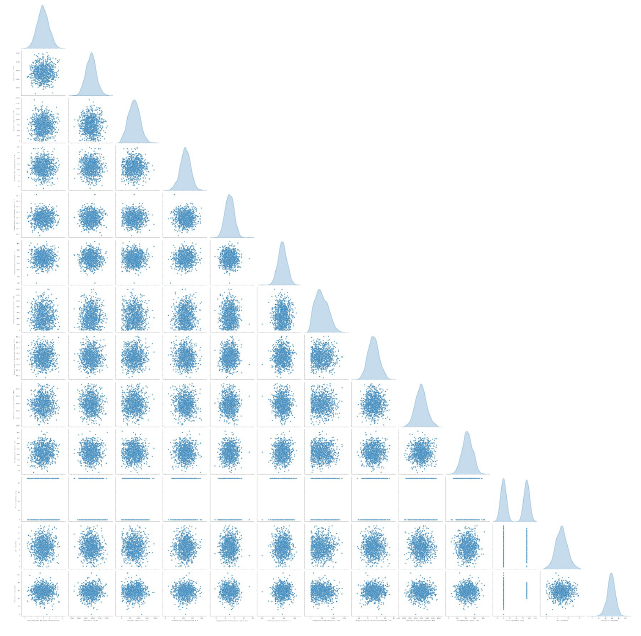


Рисунок 4 - Попарные графики рассеяния точек

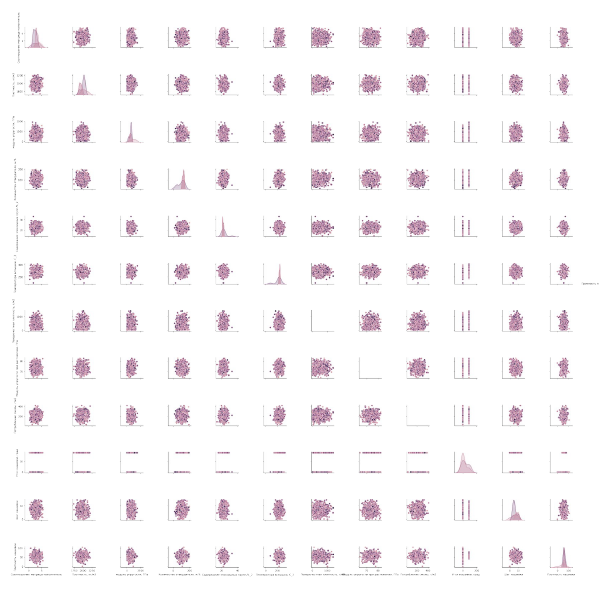


Рисунок 5 - Попарные графики рассеяния точек

Изучив графики распределения признаков, диаграммы «ящики с усами» можно увидеть наличие выбросов - результатов измерения, не подпадающих под общее распределение данного признака.

Существуют следующие методы выявления выбросов в выборке:

- метод стандартных отклонений

- метод межквартильных расстояний

Метод стандартных отклонений заключается в следующем. Гауссово распределение обладает свойством того, что стандартное отклонение от среднего может использоваться для надежного суммирования процентного значения в выборке. Например, в пределах одного стандартного отклонения среднее будет охватывать 68% данных.

Таким образом, если среднее значение равно 50, а стандартное отклонение равно 5, то все данные в выборке между 45 и 55 будут составлять около 68% выборки данных. Мы можем охватить большую часть выборки данных, если расширим диапазон следующим образом: 1 стандартное отклонение от среднего: 68%, 2 стандартных отклонения от среднего: 95%, 3 стандартных отклонения от среднего: 99,7%.

Значение, которое выходит за пределы 3 стандартных отклонений, является частью распределения, но это маловероятное или редкое событие или выброс.

Однако, не все данные являются нормальными или достаточно нормальными, чтобы рассматривать их как взятые из гауссовского распределения. Например, распределения признаков нашего датасета, хоть и приближены к нормальному, однако не точно соответствует ему.

Хорошей статистикой для суммирования выборки негауссовых распределений данных является Inter Quartile Range, или IQR для краткости. IQR рассчитывается как разница между 75-м и 25-м процентилями данных.

IQR может использоваться для определения выбросов путем определения пределов значений выборки, которые являются фактором IQR ниже 25-го процентиля или выше 75-го процентиля. Общее значение для фактора К это значение 1,5. Коэффициент k, равный 3 или более, можно использовать для определения значений, которые являются экстремальными выбросами. Выбросы — это те значения данных, которые находятся за пределами следующих интервалов: (Q1–1.5 x IQR) и (Q3 + 1.5 x IQR).

В данной работе для выявления выбросов выбран метод межквартильного интервала. Для этого в коде ноутбука реализована функция remove\_outliers(feature), заменяющая выбросы на пропуски и в цикле применена для всех признаков датасета кроме признака «Угол нашивки» т.к. он не содержит выбросов. После замены выбросов на пропуски и их удаления в выборке осталось 936 сэмплов (строк).

## Описание используемых методов

Линейная модель (англ. linear model) — модель, отображающая состояние или функционирование системы таким образом, что все взаимозависимости в ней принимаются линейными. Соответственно, она может формулироваться в виде одного линейного уравнения или системы линейных уравнений.

### Линейная регрессия и метод наименьших квадратов (МНК)

Простейшим примером постановки задачи линейной регрессии является метод наименьших квадратов (Ordinary least squares). На рисунке 4 приведен пример иллюстрации линейной регрессии для двумерного случая.

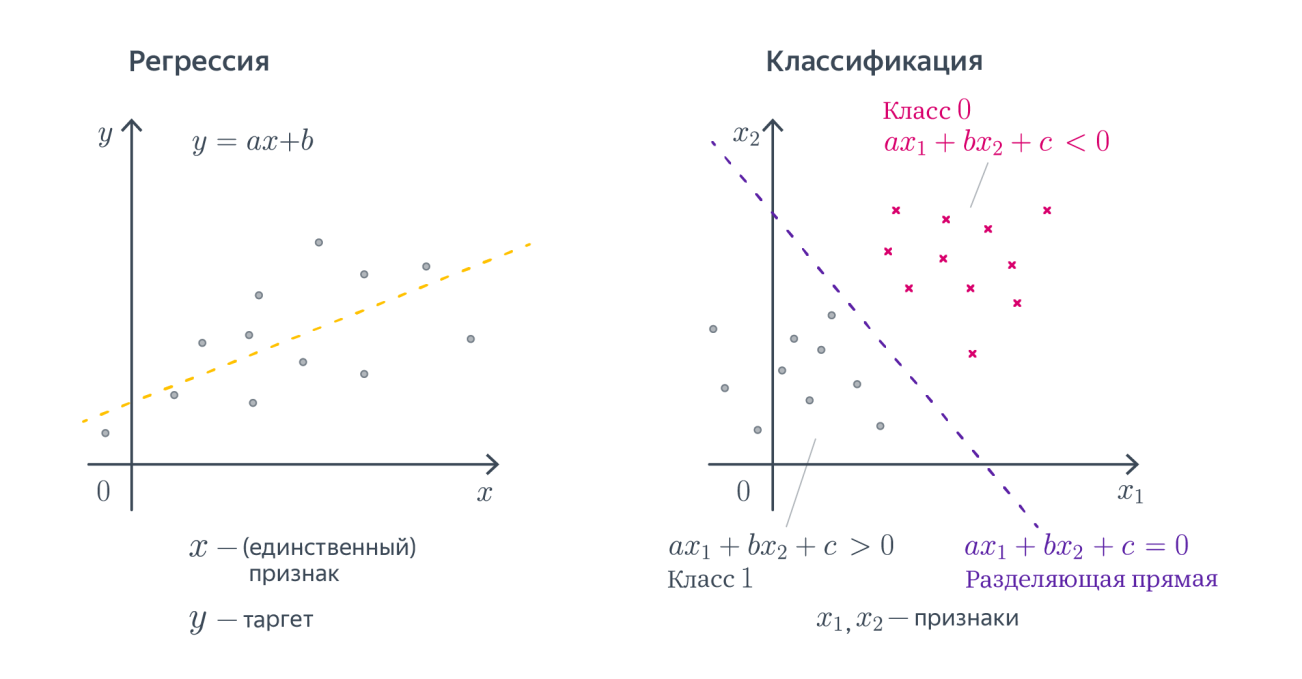


Рисунок 6- Пример линейной регрессии для двумерного случая

Пусть у нас задан датасет , где – вектор значений целевой переменной, а – матрица объекты-признаки, в которой i-я строка – это вектор признаков i-го объекта выборки. Мы хотим моделировать зависимость *yi* от *xi* как линейную функцию со свободным членом. Общий вид такой функции из в выглядит следующим образом: . Свободный член часто опускают, потому что такого же результата можно добиться, добавив ко всем *xi* признак, тождественно равный единице; тогда роль свободного члена будет играть соответствующий ему вес. Поскольку это сильно упрощает запись, в дальнейшем мы будем считать, что это уже сделано и зависимость имеет вид просто*.* В нашем примере обучаемые параметры — это веса .

Функция, оценивающая то, как часто модель ошибается, традиционно называется функцией потерь, функционалом качества или просто лоссом (англ. loss function). Важно, чтобы её было легко оптимизировать: скажем, гладкая функция потерь – это хорошо, а кусочно-постоянная – плохо. Функции потерь бывают разными. От их выбора зависит то, насколько задачу в дальнейшем легко решать, и то, в каком смысле у нас получится приблизить предсказание модели к целевым значениям. Наиболее оптимальной функцией потерь принято считать среднеквадратичное отклонение, или Mean Squared Error (MSE):

Чем меньше MSE, тем точнее наше решение, значит для того, чтобы найти лучшую модель, этот функционал нам надо минимизировать по .

Семейство функций, из которых мы выбираем, представляет собой линейную комбинацию наперед заданных базисных функций :

Цель регрессии — найти коэффициенты этой линейной комбинации *i*, и тем самым определить регрессионную функцию *f* (которую также называют моделью). Отмечу, что линейную регрессию называют линейной именно из-за линейной комбинации базисных функций — это не связано с самыми базисными функциями (они могут быть линейными или нет).

### Гребневая регрессия или ридж-регрессия

Гребневая регрессия или ридж-регрессия (англ. ridge regression) — один из методов понижения размерности. Применяется для борьбы с избыточностью данных, когда независимые переменные коррелируют друг с другом, вследствие чего проявляется неустойчивость оценок коэффициентов многомерной линейной регрессии. Гребневая регрессия или ридж-регрессия — вариация линейной регрессии, очень похожая на регрессию LASSO. Она так же применяет сжатие и хорошо работает для данных, которые демонстрируют сильную мультиколлинеарность.

### Опорная векторная регрессия

Опорная векторная регрессия - адаптация SVM для регрессии - регрессия опорных векторов или сокращенно SVR. SVR работает, находя линию наилучшего соответствия, которая минимизирует ошибку функции стоимости. Это делается с помощью процесса оптимизации, который рассматривает только те экземпляры данных в наборе обучающих данных, которые находятся ближе всего к линии с минимальной стоимостью. Эти экземпляры называются опорными векторами, отсюда и название техники.

### Стохастический градиентный спуск

Стохастический градиентный спуск (SGD) — это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров. SGD реализует простую процедуру обучения стохастическим градиентным спуском, которая поддерживает различные функции потерь и штрафы для соответствия моделям линейной регрессии.

### Деревья решений

Дерево принятия решений (также называют деревом классификации или регрессионным деревом) — средство поддержки принятия решений, использующееся в машинном обучении, анализе данных и статистике. Структура дерева представляет собой «листья» и «ветки». На рёбрах («ветках») дерева решения записаны признаки, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — признаки, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение.

Решающие правила автоматически генерируются в процессе обучения на обучающем множестве путем обобщения обучающих примеров. Поэтому их называют индуктивными правилами, а сам процесс обучения — индукцией деревьев решений.

При формировании правила для разбиения в очередном узле дерева необходимо выбрать атрибут, по которому это будет сделано.

Для регрессии критерием является дисперсия вокруг среднего. Минимизируя дисперсию вокруг среднего, мы ищем признаки, разбивающие выборку таким образом, что значения целевого признака в каждом листе примерно равны.

Огромное преимущество деревьев решений в том, что они легко интерпретируемы, понятны человеку. Они могут использоваться для извлечения правил на естественном языке. Еще одно преимущество — высокая точность работы, нетребовательность к подготовке данных.

Недостаток деревьев решений - склонность переобучаться. Переобучение в случае дерева решений — это точное распознавание примеров, участвующих в обучении и полная несостоятельность на новых данных. В худшем случае, дерево будет большой глубины и сложной структуры, а в каждом листе будет только один объект. Для решения этой проблемы используют разные критерии остановки алгоритма.

### Градиентный бустинг

Градиентный бустинг, является представителем ансамблевых методов, идея которого заключается в итеративном процессе последовательного построения частных моделей решающего дерева. Каждая новая модель обучается с использованием информации об ошибках, сделанных на предыдущем этапе, а результирующая функция представляет собой линейную комбинацию всего ансамбля моделей с учетом минимизации любой штрафной функции.

Бустинг, использующий деревья решений в качестве базовых алгорит-мов, называется градиентным бустингом над решающими деревьями. Он от-лично работает на выборках с «табличными», неоднородными данными и способен эффективно находить нелинейные зависимости в данных различной природы. На настоящий момент это один из самых эффективных алгоритмов машинного обучения. Благодаря этому он широко применяется во многих конкурсах и промышленных задачах. Он проигрывает только нейросетям на однородных данных (изображения, звук и т. д.).

Из недостатков алгоритма можно отметить только затраты времени на вычисления и необходимость грамотного подбора гиперпараметров.

### Метод случайного леса

Метод случайного леса (Random Forest) - представитель ансамблевых методов - представляет собой дальнейшее улучшение бэггинга деревьев решений, которое заключается в устранении корреляции между деревьями. Как и в случае с бэггингом, мы строим несколько сотен деревьев решений по обучающим бутстреп-выборкам. Однако на каждой итерации построения дерева случайным образом выбирается m из p подлежащих рассмотрению предикторов и разбиение разрешается выполнять только по одной из m этих переменных.

Преимущества случайного леса:

- высокая точность предсказания;

- редко переобучается;

- практически не чувствителен к выбросам в данных;

- одинаково хорошо обрабатывает как непрерывные, так и дискретные признаки, данные с большим числом признаков;

- высокая параллелизуемость и масштабируемость.

Из недостатков можно отметить, что его построение занимает больше времени, а так же теряется интерпретируемость.

### Нейронная сеть

Нейронная сеть (также искусственная нейронная сеть, ИНС) — математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма.

ИНС представляет собой систему соединённых и взаимодействующих между собой простых искусственных нейронов.

Типовая ИНС состоит из трех компонентов:

- входной слой, нейроны которого принимают исходный вектор значений, кодирующий входной сигнал и передают его на последующий слой, усилив или ослабив;

- скрытые (промежуточные) слои, которые выполняют основные вычислительные операции;

- выходной слой, нейроны которого представляют собой выходы сети и иногда также могут производить какие-либо вычисления.

Достоинства нейронных сетей:

- возможность моделирования и прогнозирования нелинейных процессов;

- способность работать с зашумленными данными;

- быстрое обучение и гибкость адаптации к изменениям внешней среды

- большие возможности распараллеливания расчетов и масштабируемости

Недостатки:

- сравнительная дороговизна аппаратной реализации;

- трудность тиражирования накопленных знаний;

- для больших сетей невозможность заранее даже приблизительно оценить время обучения сети

- невозможность интерпретации принятых решений

## Разведочный анализ данных

Разведочный анализ данных (exploratory data analysis, EDA) – это изучение данных для принятия решений по поводу их применения, очистки, преобразования и конструирования новых признаков. Как сказано выше, EDA – это чистая статистика, и основные цели его первого этапа – понять вид распределения признаков, основные параметры этого распределения, обнаружить выбросы и т.д.

Основные средства разведочного анализа — изучение вероятностных распределений переменных, построение и анализ корреляционных матриц, факторный анализ, дискриминантный анализ, многомерное шкалирование.

Частично разведочный анализ данных уже проведен ранее, при описании датасета. Определены распределения признаков, выявлены категориальные признаки, определены и удалены выбросы по каждому признаку, построены попарные графики рассеяния точек.

### Матрица корреляции

На предварительном этапе разведочного анализа каких-либо явных зависимостей между признаками не выявлено. Одним из методов анализа является построение матрицы корреляции признаков, для нашего датасета матрица корреляции приведена на рисунке 5.

Беглый взгляд на матрицу корреляции говорит об отсутствии линейных зависимостей между признаками.

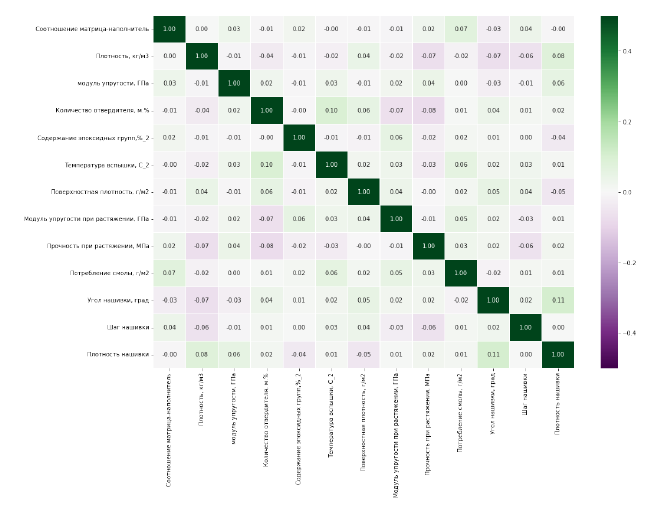


Рисунок 7 - Матрица взаимной корелляции признаков

## Нормализация данных

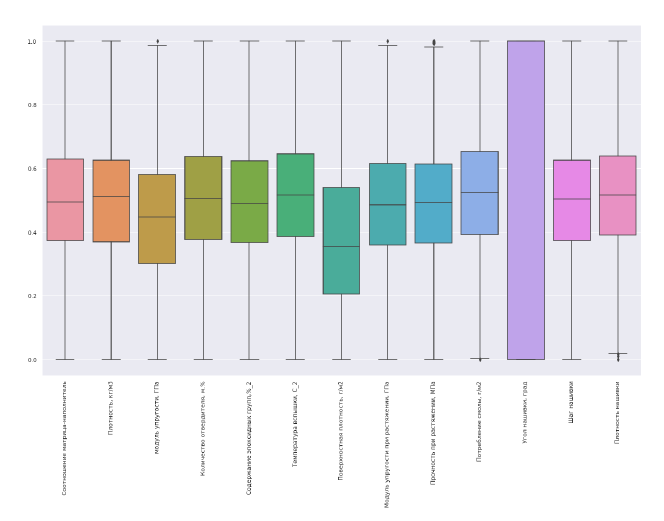
Нормализация (normalization) и стандартизация (standardization) являются методами изменения диапазонов значений – шкалирования. Шкалирование особенно полезно в машинном обучении (Machine Learning), поскольку разные атрибуты могут измеряться в разных диапазонах, или значения одного атрибута варьируются слишком сильно. Например, один атрибут имеет диапазон от 0 до 1, а второй — от 1 до 1000. Для задачи регрессии второй атрибут оказывал бы большое влияние на обучение, хотя не факт, что он является более важным, чем первый. Нормализация и стандартизация отличаются своими подходами:

Нормализация подразумевает изменение диапазонов в данных без изменения формы распределения, Стандартизация изменяет форму распределения данных (приводится к нормальному распределению).

По сути, нормализация сводится к тому, что исходный набор значений сперва смещается, а потом масштабируется.

Нормализация данных проведена с применением библиотеки sklearn при помощи MinMaxScaler.

Дополнительно о нормализации (или стандартизации) нужно сказать следующее. Поскольку наша модель будет обучена на нормализованном датасете, и прогнозировать целевую переменную она также будет в нормализованном масштабе, при интерпретации прогноза следует проводить обратное трансформирование. Кроме того, при работе модели на отдельном сервере в составе приложения необходимо нормализовать входные данные с учетом обучающего датасета и распределения его признаков.

 Рисунок 8 - BoxPlot после нормализации

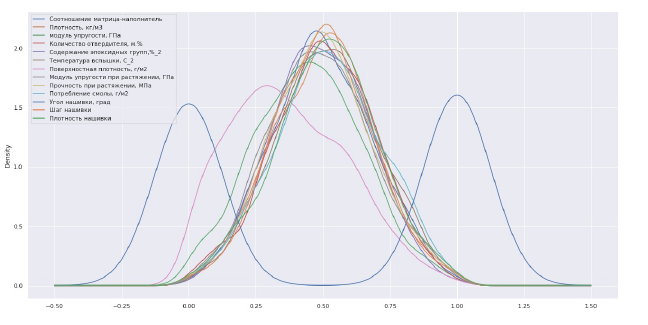


Рисунок 9 - диаграммы распределения после нормализации

# Практическая часть

## Предобработка данных

В рамках подготовки к обучению модели необходимо привести данные к единообразию: провести нормализацию (эта процедура более подробно описана в разделе 2.3.2),

В рамках предобработки проведен анализ, выявление и удаление выбросов. Более подробно эта процедура описана в разделе 2.1.

Завершающим этапом предобработки данных является разделение данных на обучающую и тестовую выборки. По заданию необходимо 30% данных оставить на тестирование модели, на остальных происходит обучение моделей. Для решения поставленных задач в работе итоговый датасет был разделен на тестовую (30%) и обучающую (70%) выборки для каждой из целевых переменных: «Модуль упругости при растяжении, ГПа», «Прочность при растяжении, МПа» и «Соотношение матрица-наполнитель». Всего для работы с моделями получилось 12 выборок (Таблица 2) – по одной на целевую переменную (y) и вектор предикторов (X), обучающую (train) и тестовую (test) выборки для каждой из трёх целей.

Таблица - Размеры выборок после разделения на обучающую и тестовую

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Наименование целевой переменной | Тестовая выборка  test | | Обучающая выборка  train | |
| X\_test | y\_test | X\_train | y\_train |
| Модуль упругости при растяжении, ГПа | (281, 13) | (281,) | (655, 13) | (655,) |
| Прочность при растяжении, МПа | (281, 13) | (281,) | (655, 13) | (655,) |
| Соотношение матрица-наполнитель | (281, 13) | (281,) | (655, 13) | (655,) |

## Разработка и обучение модели

По заданию необходимо обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении. При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.

Для выполнения данной задачи выбран модуль GridSearchCV из библиотеки sklearn, который позволяет обучать модель с подбором гиперпараметров. Список выбранных моделей приведен в таблице 3.

Работа с моделями выполнена с применением конвейера, то есть в цикле. Сначала создается список моделей, которые будут обучаться на тренировочных данных. Все модели из пакета sklearn, поэтому методы обучения fit и прогнозирования predict реализованы у каждой модели, это позволяет в соответствии с принципом полиморфизма объектно-ориентированного программирования использовать различные модели ML пакетным способом. Затем задаем возможные варианты гиперпараметров для каждой модели для дальнейшего их подбора (Таблица 3).

Таблица - Варианты значений гиперпараметров моделей для подбора по сетке

|  |  |  |  |
| --- | --- | --- | --- |
| Наименование модели | Имя гиперпараметра | Варианты значения гиперпараметра | Количество вариантов для подбора гиперпараметров с учетом кроссвалидации |
|  |  |  |  |
| LinearRegression | fit\_intercept | True | 10 |
|  | max\_features | auto, sqrt, log2, None |  |
|  |  |  |  |
| SVR | degree | 2, 3, 5 | 540 |
| kernel | rbf, sigmoid |
| C | 0.5, 1.0, 1.5 |
| epsilon | 0.01, 0.1, 0.4 |
|  |  |
| KNeighborsRegressor | n\_neighbors | 5, 10, 15 | 120 |
| algorithm | auto, ball\_tree, kd\_tree, brute |
| RandomForestRegressor | n\_estimators | 50, 100 | 180 |
| criterion | absolute\_error |
| max\_depth | 2, 4, 7 |
| max\_features | auto, log2, None |
| bootstrap | True |

Подбор по решетке делает полный перебор по заданному вручную подмножеству пространства гиперпараметров обучающего алгоритма, т.е. для всех вариантов гиперпараметров каждой модели проводится обучения и оценка метрики ошибки, затем на основе минимальной ошибки делается вывод о наилучшем наборе гиперпараметров.

Вывод функции сопровождается указанием наименования целевого параметра, и именем обучаемой в настоящее время модели. Дополнительно при помощи параметра verbose выводится информация о количестве вариантов перебора с учетом кросс-валидации.

Для подбора гиперпараметров были выбраны гиперпараметры, и варианты их значений, которые могут увеличить точность моделей. От количества возможных вариантов зависит количество времени, затраченное моделью на выполнение расчетов. Сведения о количестве расчетов каждой модели приведены в таблице 3.

Дополнительно при обучении моделей замерялось время на подбор гиперпараметров и обучение модели с их применением. Результаты замеров приведены в таблице 4.

Таблица - Результаты замеров времени работы моделей

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Наименование модели | Оцениваемый моделью параметр | | Итого время, затраченное на подбор гиперпараметров модели | |
| Модуль упругости при растяжении, ГПа | Прочность при растяжении, МПа |
| Абсолют-ное время, с | Нормализован-ное время |
| LinearRegression | 0,101 | 0,069 | 0,17 | 0,001 |
| SVR | 12,91 | 12,396 | 25,306 | 0,090 |
| KNeighborsRegressor | 0,687 | 0,658 | 1,345 | 0,005 |
| RandomForestRegressor | 140,675 | 141,758 | 282,433 | 1,000 |

Как видно из таблицы каждая модель работала незначительное время, за исключением двух ансамблевого метода. Время его работы значительно отличается от остальных. Это связано с тем, что внутри таких методов многократно рассчитываются простыми моделями, затем на основе алгоритмов бустинга и бэггинга результаты каждой модели объединяются в один результат.

## Тестирование модели

Для оценки точности моделей были разработаны несколько функций, рассчитывающих различные функции ошибки для каждой модели, коэффициент детерминации и выполняющие свод данных в таблицы. Результатом работы данных функций стала визуальная информация о работе подбора параметров, график и распределение целевой переменой (тестовой и прогнозной). На рисунках 8 и 9 приведены результаты работы моделей для параметров «Модуль упругости при растяжении, Гпа» и «Прочность при растяжении, МПа» соответственно.



Рисунок 10 - Результат работы моделей для параметра "Модуль упругости при растяжении, ГПа"

В ходе работы обучены несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении. При построении модели 30% данных оставлено на тестирование модели, на остальных происходило обучение моделей. При построении моделей был проведен поиск лучших гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой (при помощи GridSearchCV), количество блоков равно 10.

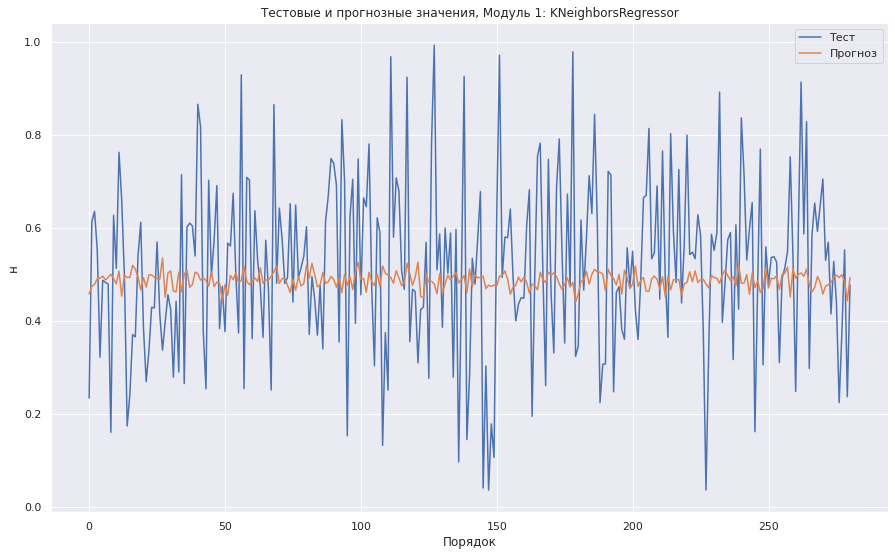


Рисунок 11 - Результат работы моделей для параметра "Прочность при растяжении, МПа"

Ни одна из моделей не справилась с прогнозированием каждого из двух рассматриваемых параметров, метрики точности указывают на точность прогноза сравнимую с шумом, в некоторых случаях точность моделей была даже ниже базовой модели, которая выдает среднее значение выборки. На данном этапе обнаружить зависимости между данными не удалось.

## Разработка нейронной сети, которая будет рекомендовать соотношение матрица-наполнитель.

По заданию необходимо разработать нейронную сеть (далее – НС) для рекомендации (прогнозирования) показателя «Соотношение матрица-наполнитель».

### TensorFlow

TensorFlow — открытая программная библиотека для машинного обучения, разработанная компанией Google для решения задач построения и тренировки нейронной сети с целью автоматического нахождения и классификации образов.

Создание НС на основе библиотеки TensorFlow сводится к созданию объекта класса Sequential и включению в него необходимого количества слоев с нужным количеством нейронов. Листинг создания НС для данной работы представлен на рисунке 12.

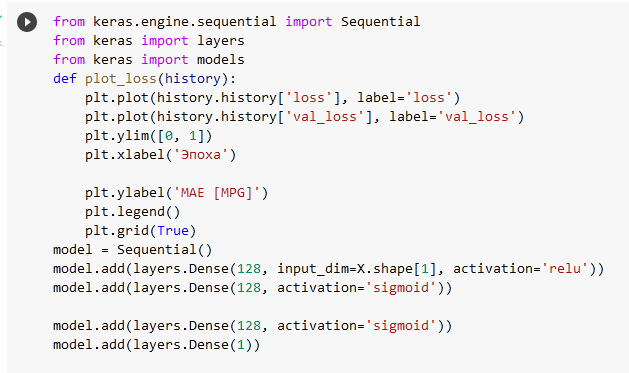


Рисунок - листинг кода

После создания модели её необходимо скомпилировать, при этом происходит конвертация структуры модели в байт-код, назначить оптимизатор, который будет обновлять веса модели при обратном проходе, указать метрику точности модели (в нашем случае accuracy). Обучение модели происходит по вызову метода fit. Данному методу необходимо передать обучающую выборку, количество эпох обучения, размер батча, и при необходимости, размер валидационной выборки для оценки точности модели в процессе её обучения.

График значения loss функции для обучающей и валидационной выборок представлен на рисунке 13.

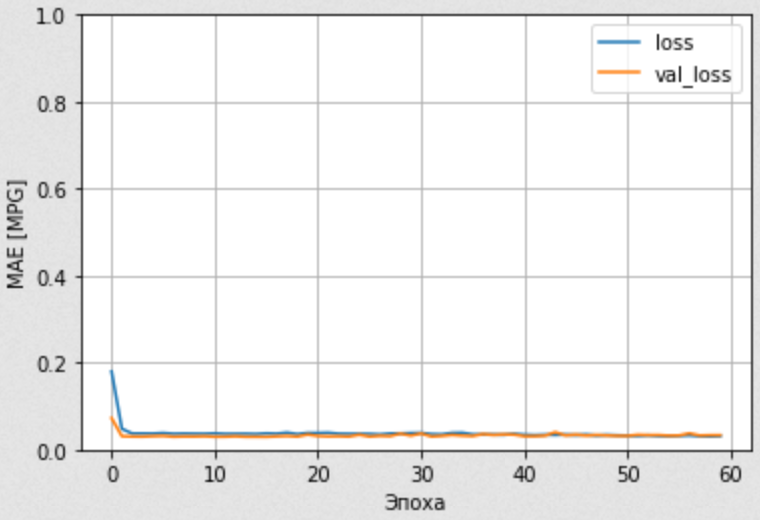


Рисунок - график значения loss функции для обучающей и валидационной выборок

Как видно из графика, модель перестала обучаться в самом начале и значение loss функции больше не меняется. Это говорит о том, что модель не нашла закономерностей в представленных ей данных и дальнейшее обучение проводить нет необходимости. Это подтверждается оценкой точности модели (Рисунок 14), значения всех метрик говорят об отсутствии сколь бы то ни было приемлемой точности. Прогноз модели на уровне шума.

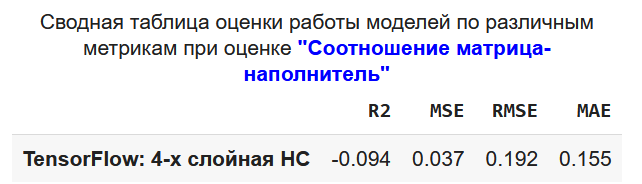


Рисунок 14 - Оценка точности модели НС на основе TensorFlow

## Разработка приложения

Разработано приложение для прогнозирования модулей упругости и прочности при растяжении. На рисунке 15 представлен листинг приложения. Приложение является консольным и состоит из цикла обработки ввода выбора пользователя и ввода данных для прогнозирования модели



Рисунок 15 – Листинг приложения

На рисунке 16 предствлен результат работы приложения. Пользователю предотсавляется выбор действия – сделать прогноз или заершить работы приложения. В случае выбора необходимо поочередно ввести значения входных признаков. При этом происходит нормализация введенных значений. По результату прогноза выводятся сначала данные в нормализованном виде, а следующей строкой подверженные обратной трансформации. Результат представляет собой вектор из двух значений, каждый из которых соответсвует предсказываемому модулю как упргуости ,так и прочности.

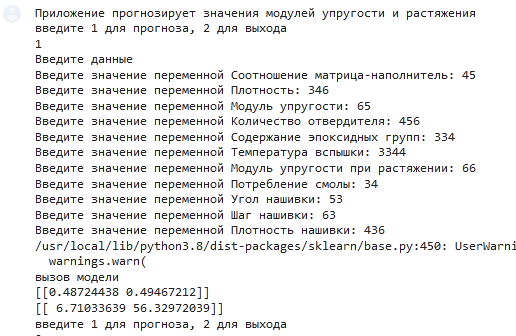


Рисунок 16 – Информационный вывод при запуске приложения

# Заключение

По результатам выполнения работы на тему «Прогнозирование конечных свойств новых материалов (композиционных материалов).» выполнено следующее.

Изучены теоретические основы и методы решения поставленной задачи. Проведен разведочный анализ предложенных данных. Получены среднее, медианное значение, проведен анализ и исключение выбросов, проверено наличие пропусков. Проведена предобработка данных (удаление шумов, нормализация и т.д.). Обучено несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении. Для моделей проведен поиск гиперпараметров с помощью GreedSearvhCV. Написана нейронная сеть, которая рекомендуют соотношение матрица-наполнитель.

Разработано консольное приложение, которое выдает прогноз. Приложение использует ранее обученную и сохранённую модель.

# Библиографический список

1. Кербер, М. Л. Полимерные композиционные материалы: структура, свойства, технология: Уч. пос. / М.Л. Кербер, В.М. Виноградов, Г.С. Головкин; Под ред. А.А. Берлина. - 3 изд., испр. - Санкт-Петербург :Профессия, 2011-560с.: ил.; . ISBN 978-5-93913-130-8. - Текст: электронный. - URL: https://znanium.com/catalog/product/872896 (дата обращения: 28.10.2022).
2. 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.
3. Джулли, Пал: Библиотека Keras - инструмент глубокого обучения / пер. с англ. А. А. Слинкин.- ДМК Пресс, 2017. – 249 с.
4. В. Ш. Берикашвили, С. П.Оськин Статистическая обработка данных, планирование эксперимента и случайные процессы​ : учебное пособие для вузов​ - 2-е изд., испр. и доп.​ - М.​ : Юрайт, 2021.​ - 163 с.​
5. 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;
6. Документация по библиотеке keras: – Режим доступа: <https://keras.io/api/>.(дата обращения: 28.10.2022).
7. Документация по библиотеке matplotlib: – Режим доступа: [https://matplotlib.org/](https://matplotlib.org/stable/users/index.html) . (дата обращения: 28.10.2022)
8. Документация по библиотеке numpy: – Режим доступа: [https://numpy.org/doc/](https://numpy.org/doc/1.22/user/index.html#user). (дата обращения: 28.10.2022).
9. Документация по библиотеке pandas: – Режим доступа: <https://pandas.pydata.org/docs/user_guide/index.html> . (дата обращения: 28.10.2022).
10. Документация по библиотеке scikit-learn: – Режим доступа: <https://scikit-learn.org/stable/user_guide.html>. (дата обращения: 28.10.2022).
11. Документация по библиотеке seaborn: – Режим доступа: <https://seaborn.pydata.org/tutorial.html>. (дата обращения: 28.10.2022).
12. Документация по библиотеке Tensorflow: – Режим доступа: <https://www.tensorflow.org/overview> (дата обращения: 28.10.2022).
13. Датасет со свойствами композитных материалов <https://drive.google.com/file/d/1B1s5gBlvgU81H9GGolLQVw_SOi-vyNf2/view?usp=sharing>