|  |  |
| --- | --- |
|  | **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)»**  **(МГТУ им. Н.Э. Баумана)** |

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**по курсу**

«Data Science»

Выполнил **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_**А.В .Ефремов**\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2023 г.*

СОДЕРЖАНИЕ

[Введение 3](#_Toc130480188)

[1 Теоретическая часть 7](#_Toc130480189)

[1.1 Принципы работы и задачи машинного обучения 7](#_Toc130480190)

[1.2 Методики машинного обучения 8](#_Toc130480191)

[1.2.1 Обучение с учителем 8](#_Toc130480192)

[1.2.2 Обучение без учителя 9](#_Toc130480193)

[1.2.3 Обучение с подкреплением 9](#_Toc130480194)

[1.3 Этапы обучения модели 9](#_Toc130480195)

[1.4 Исходные данные 10](#_Toc130480196)

[1.4.1 Описание используемых методов 11](#_Toc130480197)

[2 Практическая часть 15](#_Toc130480198)

[2.1 Предобработка данных 15](#_Toc130480199)

[2.2 Модели машинного обучения 16](#_Toc130480200)

[2.3 Разработка и обучение модели 16](#_Toc130480201)

[2.4 Тестирование модели 17](#_Toc130480202)

[2.5 Нейронная сеть 17](#_Toc130480203)

[2.6 Разработка приложения 18](#_Toc130480204)

[3. Создание удаленного репозитория 20](#_Toc130480205)

[4 Заключение 21](#_Toc130480206)

[5 Список используемой литературы 22](#_Toc130480207)

## Введение

Данная работа выполнена в рамках курса Data Science.

В качестве анализируемой задачи принята тема «Прогнозирование конечных свойств новых материалов (композиционных материалов).

Достаточно известно определение, согласно которому: композиты – это материалы, состоящие из двух или более компонентов (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).

Процессы идентификации занимают важное место в жизнедеятельности высокоорганизованных живых существ. Из биологии известно, что в качестве идентификаторов могут выступать любые биологические признаки и их комбинации: запахи, цвета, размеры, движения, звуки, вибрации. Идентификация позволяет живым существам объединяться в организованные сообщества и противостоять угрозам жизнедеятельности.

Ошибки идентификации способны нанести урон любой системе: так, паук-муравей способен внедриться в колонию муравьев и нанести ей существенный ущерб за счет скрытного уничтожения единичных особей.

Человечество активно применяет и совершенствует процессы идентификации: в последние годы различные физические идентификаторы (документы, пропуска, электронные чипы и т.д.) заменяются на наборы биологических признаков, подделка которых требует высочайшей квалификации и почти невозможна. Визуально-оптический метод идентификации постепенно заменяется на цифровые алгоритмы распознавания: они более точные, более быстрые и не могут «закрыть глаза».

Внедрение современных технологий цифровой идентификации на государственном уровне качественно влияет на уровень физической и информационной безопасности стран и граждан: страны Латинской Америки добились значительного снижения криминогенной ситуации, их опыт активно перенимается другими странами и регионами [1].

**Цель** работы: спрогнозировать ряд конечных свойств получаемых композиционных материалов, а также сократить количество проводимых испытаний.

**Задачи**:

* Изучить принципы работы машинного обучения;
* Провести разведочный анализ предложенных данных;
* Нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек;
* Получить среднее, медианное значение;
* Провести анализ и исключение выбросов, проверить наличие пропусков;
* Провести предобработку данных;
* Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении;
* Провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10;
* Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель;
* Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в задании.

## Теоретическая часть

## Принципы работы и задачи машинного обучения

Машинное обучение — это специализированный способ, позволяющий обучать компьютеры, не прибегая к программированию. Отчасти это похоже на процесс обучения младенца, который учится самостоятельно классифицировать объекты и события, определять взаимосвязи между ними.

В основе машинного обучения лежат три одинаково важных компонента:

* *Данные*. Собираются всевозможными способами. Чем больше данных, тем эффективней машинное обучение и точнее будущий результат.
* *Признаки*. Определяют, на каких параметрах строится машинное обучение.
* *Алгоритм*. Выбор метода машинного обучения (при условии наличия хороших данных) будет влиять на точность, скорость работы и размер готовой модели.

Задачи, которые способно решить машинное обучение:

* *Регрессия*. Предоставляет прогноз на основе выборки объектов с различными признаками. По итогам анализа данных на выходе получается число или числовой вектор. Например, таким образом работает кредитный скоринг — оценка кредитоспособности потенциального заёмщика.
* *Классификация*. Выявляет категории объектов на основе имеющихся параметров. Продолжает традиции машинного зрения, поэтому часто можно встретить термин «распознавание образов»: например, идентификация разыскиваемых людей по фото или на основании словесного описания внешности.
* *Кластеризация*. Разделяет данные на схожие категории по объединяющему признаку. Например, космические объекты кластеризируют по удаленности, размерам, типам и другим признакам.
* *Идентификация*. Отделяет данные с заданными параметрами от остального массива данных. К примеру, участвует в постановке медицинского диагноза по набору симптомов.
* *Прогнозирование*. Работает с объемами данных за определенный период и предсказывает на основе анализа их значение через заданный период времени. Примером может служить прогноз погоды.
* *Извлечение знаний*. Исследует зависимости между рядом показателей одного и того же явления или события. Например, находит закономерности во взаимодействии биржевых показателей [2].

## Методики машинного обучения

Для простоты восприятия типы машинного обучения принято разделять на три категории:

* обучение с учителем;
* обучение без учителя;
* обучение с подкреплением.

## Обучение с учителем

Специалист использует для обучения модели размеченные данные. Размеченные — значит, что на каждый пример уже есть правильный ответ. Задача алгоритма — понять, почему он правильный.

Например, нужно научить алгоритм распознавать на улицах легковые автомобили. Для обучения будут использоваться много фотографий, на которых все области, включающие машины, выделены прямоугольниками. Модели нужно самой найти взаимосвязи между областями в прямоугольниках и установить, что в них находится один и тот же объект — автомобиль [3].

## Обучение без учителя

Для данного типа обучения ключевым понятием является паттерн — обрабатывая значительные массивы данных, алгоритм должен сперва самостоятельно выявлять закономерности. На следующем этапе на основе выявленных закономерностей машина интерпретирует и систематизирует данные [2].

## Обучение с подкреплением

Принципы обучения с подкреплением заимствованы из психологических экспериментов: машина пытается найти оптимальные действия, которые будет предпринимать, находясь в наборе различных сценариев. Эти действия могут иметь как краткосрочные, так и долгосрочные последствия, а от алгоритма требуется обнаружить эти связи [2].

## Этапы обучения модели

Работа с обучением модели делится на пять этапов:

* *Сбор данных*. На этом этапе нужно собрать информацию, которая будет использоваться для обучения модели. Но собрать — не значит просто найти много подходящих цифр или картинок. Необходимо определить, какие именно данные нужны, какие критерии будут критичными, а какие нет.
* *Разметка данных*. Если обучение модели предполагает работу с размеченными данными, нужно провести подготовительную работу — выделить области или критерии, которые нужны для обучения машины или дать правильный ответ для каждого случая.
* *Исследовательский анализ*. Этап контрольной проверки, при котором изучается, как распределены данные, как зависят друг от друга различные признаки и нет ли в них ошибки или нетипичных случаев. Например, для некоторых классических алгоритмов нужно, чтобы все цифры находились в диапазоне от 0 до 1 или от −1 до 1. Задача этого этапе — посмотреть вглубь данных и увидеть основные закономерности.
* *Обучение модели*. На этом этапе выбираются подходящие алгоритмы для решения задачи, обучаются несколько перспективных моделей.
* *Тестирование и оценка*. Итоги обучения нужно оценить и понять, что делать дальше: собрать недостающие данные и продолжить обучение, заменить параметры модели или пересмотреть алгоритм [3].

## Исходные данные

Дано два набора данных, которые необходимо объединить по индексу тип объединения INNER. Элементы массива соответствуют типу float64.

Начальные свойства компонентов композиционных материалов разделены по следующим критериям:

* 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.

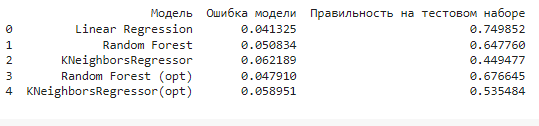
## 1.4.1 Описание используемых методов

Для решения поставленной задачи выбраны методы:

* решение задачи регрессии для прогнозирования параметров: модуля упругости и прочности при растяжении. Для решения задачи регрессии использовались: линейная регрессия, случайный лес и метод К-соседей.

Оценка качества моделей указана на рисунке 1, где «model» – модель для параметров «Модуль упругости» и «Прочность при растяжении».

Рисунок 1 - Оценка качества моделей



Средняя абсолютная ошибка (mean\_absolute\_error) для всех моделей находится в интервале от 0,04 до 0,06.

Коэффициент детерминации (r2\_score) принимает значения от 0,4 до 0,7 для всех рассматриваемых моделей.

**1.4.2 Разведочный анализ данных**

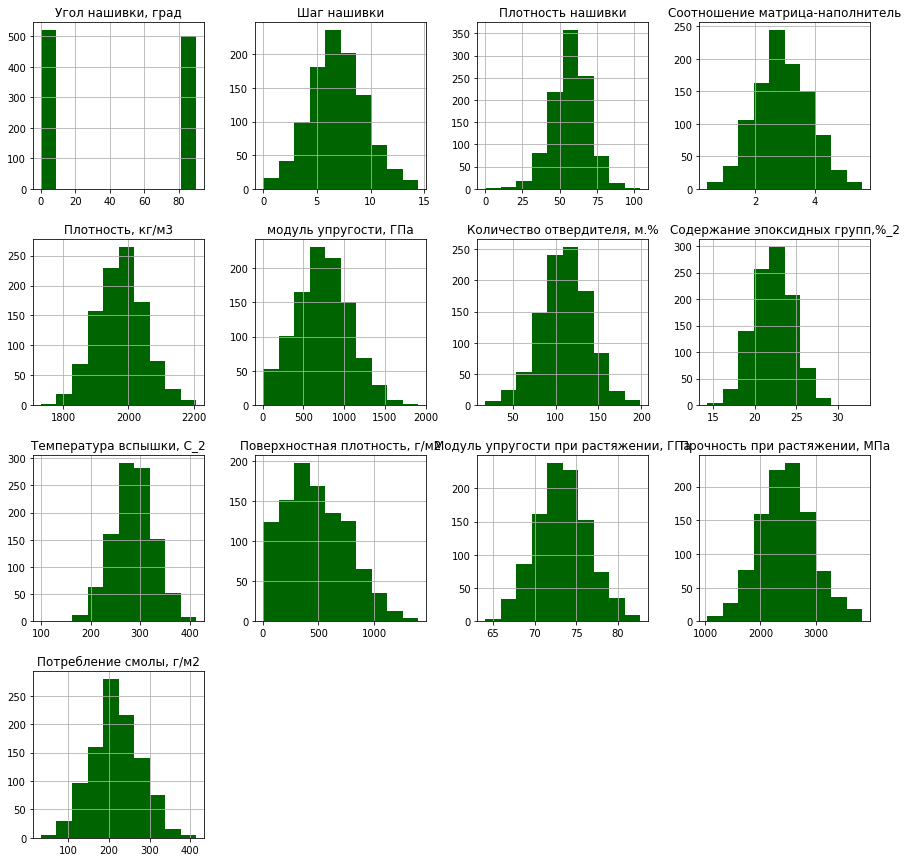
Для разведочного анализа данных использованы методы описательной статистики.

Датасет был проверен на наличие пропусков в значениях (команда fail.isna().sum()). Пропусков обнаружено не было.

Команда Describe () позволила выявить наличие дискретной величины, принимающей значения 0 и 90 (параметр «Угол нашивки»), а также основные значения для всех параметров.

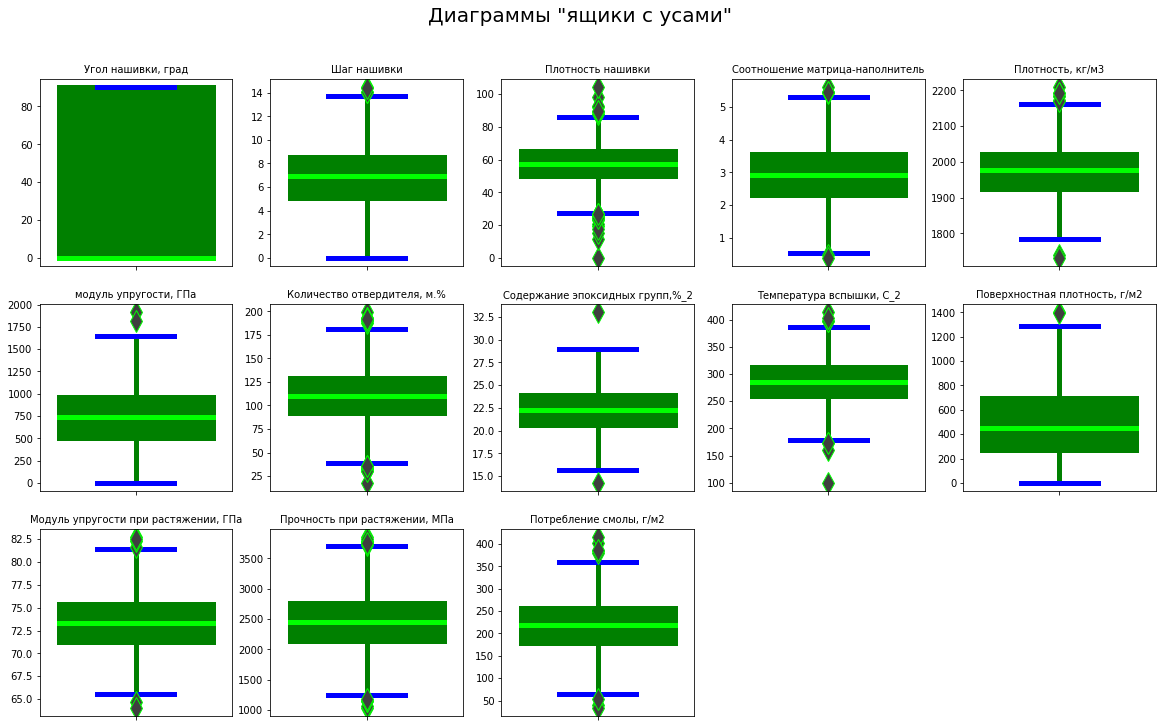
С помощью построения гистограмм было выявлено распределение величин, близкое к нормальному, для большей части параметров (исключение составили: поверхностная плотность – распределение со смещением вправо; угол нашивки – дискретная величина, график оказался не показателен).

Рисунок 2- График гистограмм



С помощью диаграммы «ящики с усами» для всех параметров были выявлены выбросы.

Рисунок 3- График диаграмм «ящик с усами»



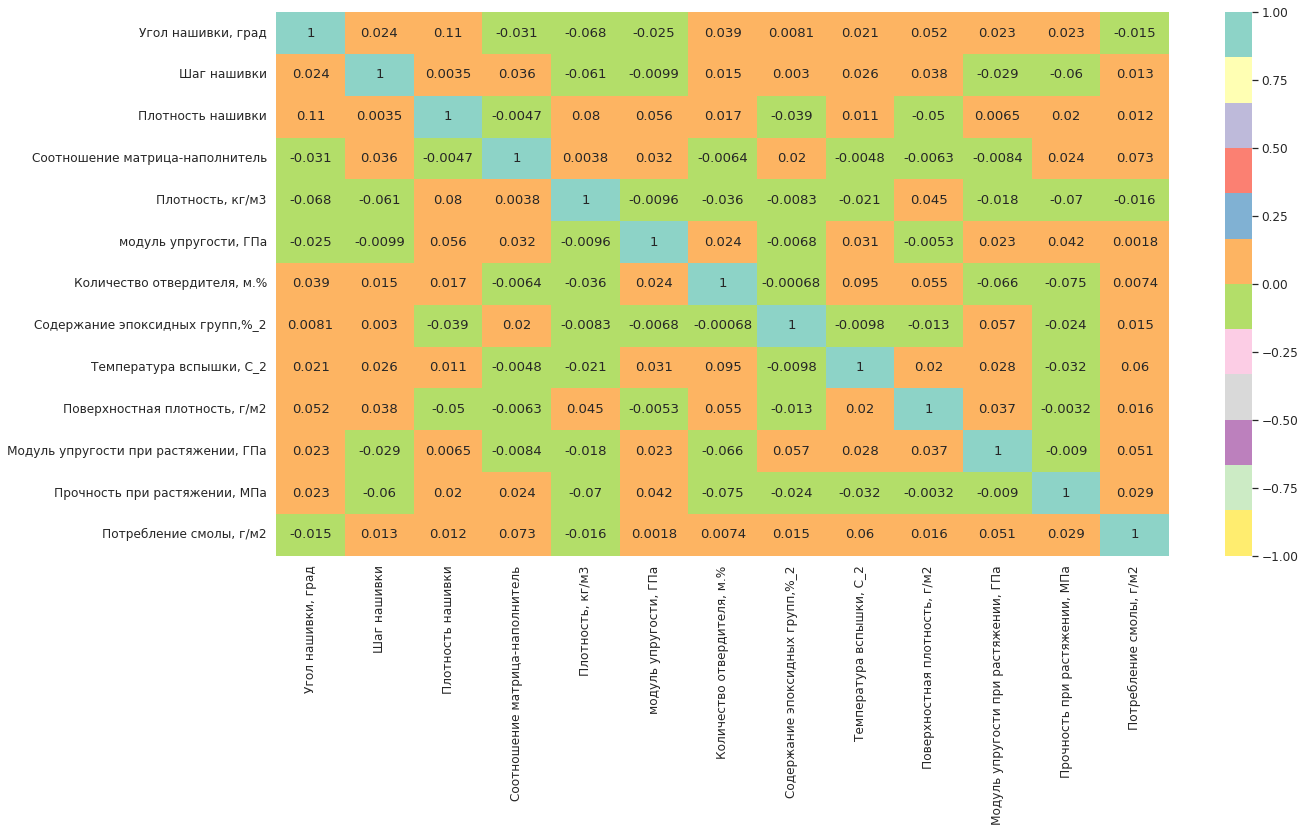
Выбросы наблюдаются по всем параметрам, кроме угла нашивки, т.к. данный параметр принимает дискретные значения и диаграмма «ящик с усами» для него не показательна (Рис.3)

Также следует отметить наличие выбросов с двух сторон (например, прочность при растяжении), наличие выбросов со стороны наименьших значения (например, содержание эпоксидных групп), наличие выбросов со стороны наибольших значений (например, шаг нашивки).

Для разведочного анализа было использовано также построение попарных графиков рассеяния точек. Графики для всех параметров показали отсутствие сильной зависимости между переменными датасета.

На рисунке 4 приведена тепловая карта коэффициентов корреляции, значения которой показывают, что все полученные коэффициенты корреляции находятся в промежутке значений от - 0,20 до 0,20. Исходя из этого можно предположить, что исходный датасет был предварительно обработан (либо сгенерирован) и переменные являются независимыми.

Рисунок 4 -Тепловая карта коэффициентов корреляции



# **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, данные приведены на рисунке 5.

Рисунок 5- Расчет количества выбросов

0 выбросов в признаке Угол нашивки, град

0 выбросов в признаке Шаг нашивки

7 выбросов в признаке Плотность нашивки

0 выбросов в признаке Соотношение матрица-наполнитель

3 выбросов в признаке Плотность, кг/м3

2 выбросов в признаке модуль упругости, ГПа

2 выбросов в признаке Количество отвердителя, м.%

2 выбросов в признаке Содержание эпоксидных групп,%\_2

3 выбросов в признаке Температура вспышки, С\_2

2 выбросов в признаке Поверхностная плотность, г/м2

0 выбросов в признаке Модуль упругости при растяжении, ГПа

0 выбросов в признаке Прочность при растяжении, МПа

3 выбросов в признаке Потребление смолы, г/м2

Всего 24 выброса

Количество выбросов говорит о их незначительности, соответственно, данные значения можно удалить из датасета. Удаление производим с помощью метода межквартильного интервала.

Размер датасета после удаления выбросов и проверка наличия пропусков:

* Количество строк в очищенном датасете: 927;
* Количество столбцов (переменных) в очищенном датасете: 13;
* Количество пропусков в данных очищенного датасета: 0.

Вызов команды fail\_norm\_d.describe().T для датасета показывает, что количество строк датасета уменьшилось. Датасет очищен от выбросов.

Рассмотрим метод функционального факторного анализа для поиска скрытых факторов, влияющих на свойства композита.

Для этого измерим пригодность данных для факторного анализа   
(Тест Кайзера-Мейера-Олкина) и посмотрим есть ли определенная избыточность между переменными, которые мы можем суммировать с помощью нескольких факторов (тест Бартлетта).

Для теста Кайзера-Мейера-Олкина значения находятся в диапазоне от 0 до 1. Полученное значение менее 0,6 считается недостаточным.

Рисунок 6

0.4629014151656023-значение КМО полученное на исходном датасете

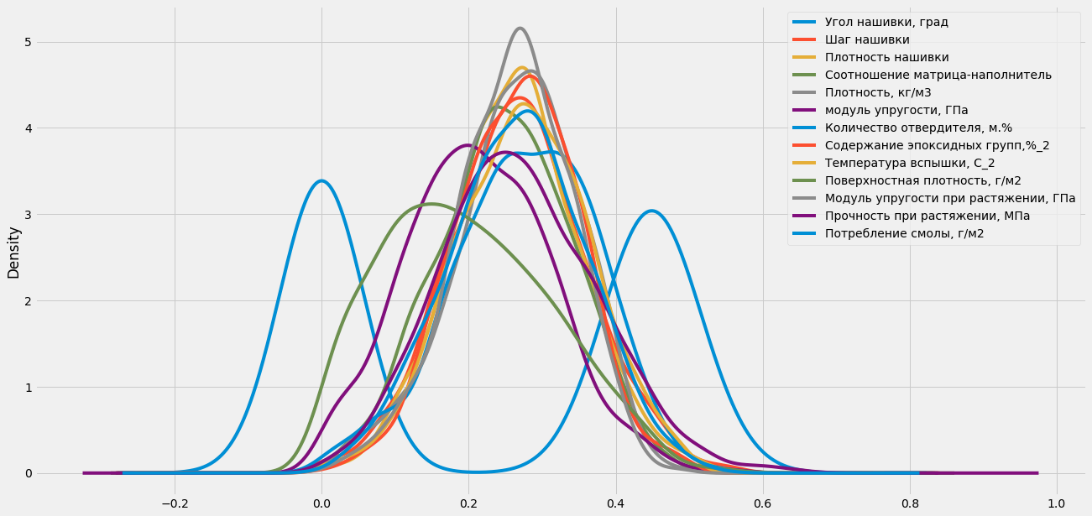
0.05210603127433927-значение, полученное для теста Бартлетта.

Полученные значения дают возможность предположить, что применение метода факторного анализа и сокращения данных не даст нам хорошего результата.

## 2.2 Модели машинного обучения

Многие алгоритмы машинного обучения работают лучше или сходятся быстрее, когда функции находятся в относительно одинаковом масштабе и / или близки к нормальному распределению. Для этого применим метод MinMaxScaler библиотеки Scikit-learn, позволяющий произвести Нормализацию данных перед использованием в Модели Машинного обучения, то есть осуществить приведение числовых переменных к диапазону от 0 до 1.

Рисунок 7



## 2.3 Подготовка данных

Из датасета были выделены прогнозируемые (зависимые) переменные. Далее, выборка была разделена на обучающую и тестовую выборки, в соответствии с условием задачи 70% (на обучение) /30% (на тестирование).

Размер обучающей выборки: 648.

Размер тестовой выборки: 279.

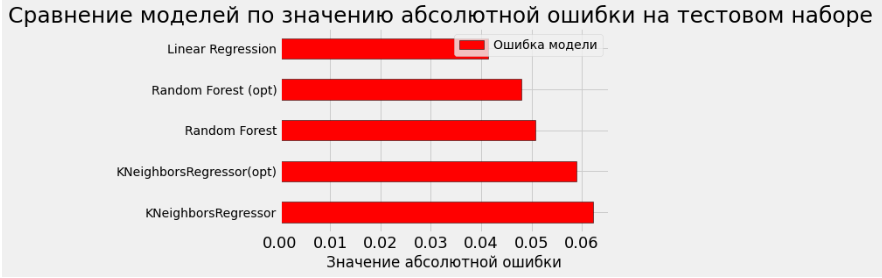
## 2.4 Разработка и обучение модели

В качестве модели выбрана Линейная регрессия, Случайный лес и   
К -соседей. Обучение и тестирование производилось для двух параметров: модуль упругости и прочность при растяжении (в соответствии с условиями задачи).

Для всех рассматриваемых моделей был осуществлен поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, с количеством блоков равным 10. Исключение составила линейная регрессия   
из-за малого количества возможно изменяемых параметров.

На рисунке 8 приведено сравнение моделей по значению абсолютной ошибки.

Рисунок 8



## 2.5 Нейронная сеть

В качестве нейронной сети был принят многослойный персептрон.

Гиперпараметры модели:

- количество скрытых слоев = 3;

- количество нейронов на слое = 256 и 64;

- активационная функция «sigmoid», «linear»;

- количество нейронов на выходном слое = 1;

- оптимизатор «Adam»;

Обучение модели происходило за 40 эпох (количество задано после поиска опытным путем наиболее приемлемого распределения MSE и общих результатов обучения модели). Результат изменения MSE модели указан на рисунке 8.   
MSE уменьшается со временем по мере выполнения алгоритма. Это означает, что модель приближаемся к оптимальному решению.

Ошибка в основном распределяется между -0,01 и 0,02. Это показывает, что модель обучения не совсем подходящая, ее можно улучшить, приведя значение ошибки к 0. Гистограмма распределения ошибки приведена на рисунке 9.

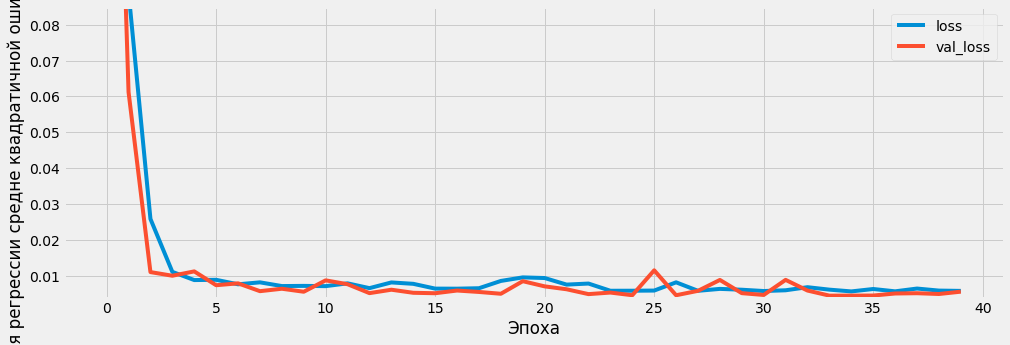
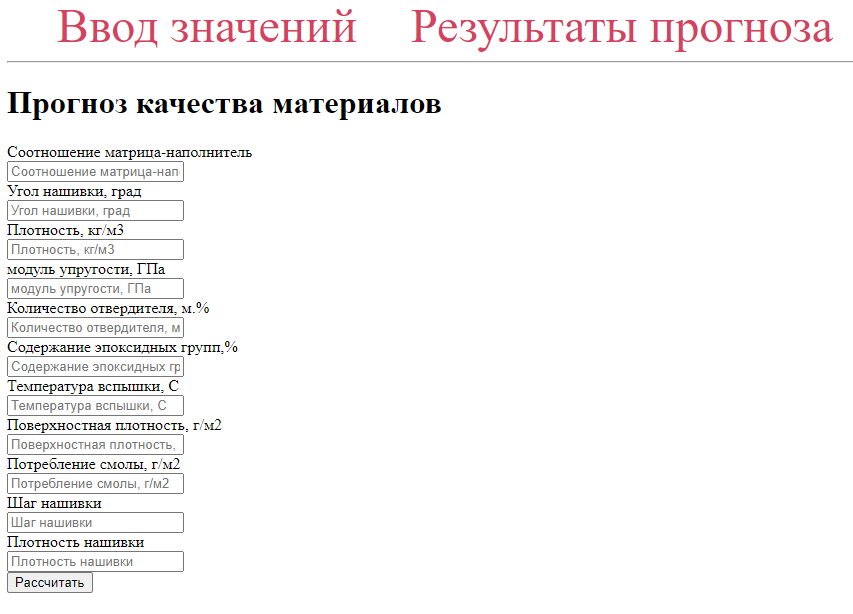


Рисунок 9 – Распределение ошибки

## 2.6 Разработка приложения

Приложение было разработано с помощью Flask.

В интерфейсе использованы наименования параметров из датасета (скриншот приведен на рисунке 10).



<http://51.250.74.213:5000/>

Рисунок 10 – Интерфейс приложения

При введении интересуемых данных и после нажатия кнопки «Рассчитать» приложение, на основании модели, будет выводить рекомендуемые значения «модуль упругости при растяжении» и «прочность при растяжении».

# Создание удаленного репозитория

Страница слушателя на GitLab

Созданный репозиторий:

<https://github.com/Oglyanis/BMSTU>

Страница слушателя, созданный репозиторий, коммиты в репозитории.

<https://github.com/Oglyanis/BMSTU>

# **Заключение**

Процесс подготовки данных с помощью алгоритмов машинного обучения не является тривиальным, не смотря на обманчивую простоту постановки задачи.

Перед созданием алгоритма цифровой идентификации требуется предварительная подготовка обширного обучающего набора данных: необходимо собрать, разметить и преобразовать данные к единому формату, отсортировать набор данных и провести его валидацию.

Обученная модель сильно ограничена в своих функциональных возможностях конкретной задачей, описываемой алгоритмом обработки данных и обучающими данными. Зачастую решение одной реальной задачи может быть представлено только комбинацией из нескольких моделей машинного обучения.

Все это делает процесс внедрения технологий цифровой идентификации длительной и технологически сложной задачей, требующей участия большого количества высококвалифицированных специалистов по сбору, разметке, подготовке и анализу данных.

Полученный в ходе выполнения проектной работы результат является достижимым благодаря огромному количеству инструментов с открытым исходным кодом, которые позволяют сократить затраты времени на решение типовых задач, возникающих в ходе создания и обучения моделей машинного обучения.

## 5 Список используемой литературы

1. Минаев В.А., Вайц Е.В., Ефремов Е.А., Ковалевский А.В. «Безопасность и технологии электронной идентификации в странах Латинской Америки» // Вестник Российского Нового Университета. Серия: Сложные системы: модели, анализ и управление. – 2019 г. – Номер 2. – Страницы: 142-151.
2. «Машинное обучение: просто о сложном» [Электронный ресурс] – Режим доступа: https://sbercloud.ru/ru/warp/blog/machine-learning-about (дата обращения: 08.03.2023)
3. «Что такое машинное обучение» [Электронный ресурс] – Режим доступа: https://practicum.yandex.ru/blog/chto-takoe-mashinnoe-obuchenie/ (дата обращения: 08.03.2023)
4. Андреас Мюллер, Сара Гвидо «Введение в машинное обучение с помощью PYTHON», изд. ДИАЛЕКТИКА, 2017
5. Франсуа Шолле «Глубокое обучение на Python», изд. Питер 2020
6. Учебник по машинному обучению, academy Yandex. ru [электронный курс] ­– Режим доступа: <https://clck.ru/33eYyu>
7. Telegram, Режим доступа: <https://t.me/Data2good_chat>
8. «Машинное обучение» [Электронный курс]­–https://habr.com/ru/company/vk/blog/513842/