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 Муравьева Екатерина Сергеевна

Москва, 2023

**Содержание**

Содержание 2

Введение 3

1. Аналитическая часть 5

2. Практическая часть 20

Заключение 31

Список используемой литературы и ссылки на веб-ресурсы 32

**Введение**

Развитие современной техники требует новых конструкционных материалов, превосходящих по своим прочностным, упругим и другим свойствам традиционные. Композиционный материал, сокращённо композит — многокомпонентный материал, изготовленный из двух или более компонентов с существенно различными физическими и/или химическими свойствами, которые, в сочетании, приводят к появлению нового материала с характеристиками, отличными от характеристик отдельных компонентов и не являющимися простой их суперпозицией. В составе композита принято выделять матрицу/матрицы и наполнитель/наполнители, последние выполняют функцию армирования. Сочетание разных компонентов позволяет улучшить характеристики материала и делает его одновременно лёгким и прочным. При этом отдельные компоненты остаются таковыми в структуре композитов, что отличает их от смесей и затвердевших растворов. Варьируя состав матрицы и наполнителя, их соотношение, ориентацию наполнителя, получают широкий спектр материалов с требуемым набором свойств. Многие композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении её механических характеристик.

Эффективность и работоспособность материала зависят от правильного выбора исходных компонентов и технологии их совмещения, призванной обеспечить прочную связь между компонентами при сохранении их первоначальных характеристик.

Как следствие, в настоящее время начали применять технологии машинного обучения как инструмент для разработки и открытия новых материалов. Разработка новых материалов с превосходными индивидуальными свойствами является конечной целью современных инженерных приложений. За последние несколько десятилетий, благодаря быстрому развитию высокопроизводительных параллельных вычислений, материаловедения и численного моделирования, многие важные свойства материалов теперь могут быть рассчитаны с помощью моделирования с достаточной точностью.

Актуальность решения этой задачи тесно связана с активным использованием строительных и инновационных композитных материалов (технологий) во многих областях производства.

Полученный прогноз модели должен помочь значительно сократить количество экспериментальных исследований, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

1. **Аналитическая часть**

**1.1 Постановка задачи.**

**Описание:** Композиционные материалы или композиты – это материалы, состоящие из двух и более компонентов (отдельных волокон или других армирующих составляющих и связующей их матрицы) и обладающие специфическими свойствами, отличными от суммарных свойств составляющих их компонентов. Компоненты композитов не должны растворяться или иным способом поглощать друг друга. Они должны быть хорошо совместимы. Свойства композиционных материалов нельзя определить только по свойствам компонентов, без учета их взаимодействия.

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

**На входе** имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.).

Первый датасет «X\_bp.xlsx» включает в себя десять переменных Базальтопластика, такие как соотношение матрица-наполнитель, плотность, модуль упругости, количество отвердителя, содержание эпоксидных групп, температура вспышки, поверхностная плотность, модуль упругости при растяжении, прочность при растяжении, потребление смолы. Выборка содержит 1023 замера.

Второй датасет «X\_nup.xlsx» включает в себя три переменных накладок углепластиковых – угол нашивки, шаг нашивки, плотность нашивки. Выборка содержит 1040 замеров.

Данные таблицы имеют колонку с целочисленным индексом, не являющимся входным или выходным переменным, служащим для сопоставления таблиц данных.

**На выходе** необходимо спрогнозировать три параметра: модуля упругости при растяжении, прочности при растяжении, соотношение матрица-наполнитель.

Поставленная задача в рамках классификации категорий машинного обучения относится к машинному обучению с учителем, задача регрессии.

Анализ, предобработка данных, построение моделей выполнены посредством языка программирования Python с использованием библиотек Pandas, Matplotlib и Sklearn.

**1.2. Описание используемых методов**

Машинное обучение (англ. machine learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

Машинное обучение разделяется на несколько основных подходов:

* Обучение с учителем (supervised learning) -наиболее распространенный и изученный тип машинного обучения, ведь легче обучить машину с выбранными данными. В зависимости от того, что вы хотите предсказать, обучение с учителем может использоваться для решения двух типов задач:
  + Классификация (classification)
  + Регрессия (regression)
  + Ранжирование (learning to rank)
* Обучение без учителя (unsupervised learning) -   
  поскольку у машин нет отмеченных данных для обучения, цель машинного обучения без учителя — обнаружение закономерностей в данных и их группировка. Обучение без учителя решает 2 типа задач:
  + Кластеризация (clustering)
  + Уменьшение размерности (dimensionality reduction)
* Обучение с частичным привлечением учителя (semi-supervised learning) -   
  этот тип машинного обучения требует использования системы вознаграждения/штрафа. Цель — вознаградить машину, когда она учится правильно, и наказать машину, когда она учится неправильно.
* Обучение с подкреплением (reinforcement learning) -   
  способ машинного обучения, при котором система обучается, взаимодествуя с некоторой средой.

**Типы задач машинного обучения**

Все задачи, решаемые с помощью ML, относятся к одной из следующих категорий.

**Задача регрессии** – прогноз на основе выборки объектов с различными признаками. На выходе должно получиться вещественное число (2, 35, 76.454 и др.), к примеру цена квартиры, стоимость ценной бумаги по прошествии полугода, ожидаемый доход магазина на следующий месяц, качество вина при слепом тестировании.

**Задача классификации** – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): есть ли на фотографии кот, является ли изображение человеческим лицом, болен ли пациент раком.

**Задача кластеризации** – распределение данных на группы: разделение всех клиентов мобильного оператора по уровню платёжеспособности, отнесение космических объектов к той или иной категории (планета, звезда, чёрная дыра и т.п.).

**Задача уменьшения размерности** – сведение большого числа признаков к меньшему (обычно 2–3) для удобства их последующей визуализации (например, сжатие данных).

**Задача выявления аномалий** – отделение аномалий от стандартных случаев. На первый взгляд она совпадает с задачей классификации, но есть одно существенное отличие: аномалии – явление редкое, и обучающих примеров, на которых можно натаскать машинно обучающуюся модель на выявление таких объектов, либо исчезающе мало, либо просто нет, поэтому методы классификации здесь не работают. На практике такой задачей является, например, выявление мошеннических действий с банковскими картами.

**Основные алгоритмы моделей машинного обучения**

**Дерево принятия решений** — это метод поддержки принятия решений, основанный на использовании древовидного графа: модели принятия решений, которая учитывает их потенциальные последствия (с расчётом вероятности наступления того или иного события), эффективность, ресурсозатратность. Для бизнес-процессов это дерево складывается из минимального числа вопросов, предполагающих однозначный ответ — «да» или «нет». Последовательно дав ответы на все эти вопросы, мы приходим к правильному выбору. Методологические преимущества дерева принятия решений – в том, что оно структурирует и систематизирует проблему, а итоговое решение принимается на основе логических выводов.

**Наивная байесовская классификация** - наивные байесовские классификаторы относятся к семейству простых вероятностных классификаторов и берут начало из теоремы Байеса, которая применительно к данному случаю рассматривает функции как независимые (это называется строгим, или наивным, предположением). На практике используется в следующих областях машинного обучения:

* определение спама, приходящего на электронную почту;
* автоматическая привязка новостных статей к тематическим рубрикам;
* выявление эмоциональной окраски текста;
* распознавание лиц и других паттернов на изображениях.

**Метод наименьших квадратов** - всем, кто хоть немного изучал статистику, знакомо понятие линейной регрессии. К вариантам её реализации относятся и наименьшие квадраты. Обычно с помощью линейной регрессии решают задачи по подгонке прямой, которая проходит через множество точек. Вот как это делается с помощью метода наименьших квадратов: провести прямую, измерить расстояние от неё до каждой из точек (точки и линию соединяют вертикальными отрезками), получившуюся сумму перенести наверх. В результате та кривая, в которой сумма расстояний будет наименьшей, и есть искомая (эта линия пройдёт через точки с нормально распределённым отклонение мотистинного значения).

Линейная функция обычно используется при подборе данных для машинного обучения, а метод наименьших квадратов – для сведения к минимуму погрешностей путем создания метрики ошибок.

Логистическая регрессия

**Логистическая регрессия** – это способ определения зависимости между переменными, одна из которых категориально зависима, а другие независимы. Для этого применяется логистическая функция (аккумулятивное логистическое распределение). Практическое значение логистической регрессии заключается в том, что она является мощным статистическим методом предсказания событий, который включает в себя одну или несколько независимых переменных. Это востребовано в следующих ситуациях:

* кредитный скоринг;
* замеры успешности проводимых рекламных кампаний;
* прогноз прибыли с определённого товара;
* оценка вероятности землетрясения в конкретную дату.

**Метод опорных векторов (SVM)** — это целый набор алгоритмов, необходимых для решения задач на классификацию и регрессионный анализ. Исходя из того, что объект, находящийся в N-мерном пространстве, относится к одному из двух классов, метод опорных векторов строит гиперплоскость с мерностью (N – 1), чтобы все объекты оказались в одной из двух групп. На бумаге это можно изобразить так: есть точки двух разных видов, и их можно линейно разделить. Кроме сепарации точек, данный метод генерирует гиперплоскость таким образом, чтобы она была максимально удалена от самой близкой точки каждой группы. SVM и его модификации помогают решать такие сложные задачи машинного обучения, как сплайсинг ДНК, определение пола человека по фотографии, вывод рекламных баннеров на сайты.

**Метод ансамблей** – базируется на алгоритмах машинного обучения, генерирующих множество классификаторов и разделяющих все объекты из вновь поступающих данных на основе их усреднения или итогов голосования. Изначально метод ансамблей был частным случаем байесовского усреднения, но затем усложнился и оброс дополнительными алгоритмами:

* бустинг (boosting) – преобразует слабые модели в сильные посредством формирования ансамбля классификаторов (с математической точки зрения это является улучшающим пересечением);
* бэггинг (bagging) – собирает усложнённые классификаторы, при этом параллельно обучая базовые (улучшающее объединение);
* корректирование ошибок выходного кодирования.

Метод ансамблей является более мощным инструментом по сравнению с отдельно стоящими моделями прогнозирования, поскольку:

* он сводит к минимуму влияние случайностей, усредняя ошибки каждого базового классификатора;
* уменьшает дисперсию, поскольку несколько разных моделей, исходящих из разных гипотез, имеют больше шансов прийти к правильному результату, чем одна отдельно взятая;
* исключает выход за рамки множества: если агрегированная гипотеза оказывается вне множества базовых гипотез, то на этапе формирования комбинированной гипотезы оно расширяется при помощи того или иного способа, и гипотеза уже входит в него.

**Алгоритмы кластеризации** - кластеризация заключается в распределении множества объектов по категориям так, чтобы в каждой категории – кластере –оказались наиболее схожие между собой элементы. Кластеризировать объекты можно по разным алгоритмам. Чаще всего используют следующие:

* на основе центра тяжести треугольника;
* на базе подключения;
* сокращения размерности;
* плотности (основанные на пространственной кластеризации);
* вероятностные;

Алгоритмы кластеризации используются в биологии (исследование взаимодействия генов в геноме, насчитывающем до нескольких тысяч элементов), социологии (обработка результатов социологических исследований методом Уорда, на выходе дающим кластеры с минимальной дисперсией и примерно одинакового размера) и информационных технологиях.

**Метод главных компонент (PCA)** - представляет собой статистическую операцию по ортогональному преобразованию, которая имеет своей целью перевод наблюдений за переменными, которые могут быть как-то взаимосвязаны между собой, в набор главных компонент – значений, которые линейно не коррелированы.

Практические задачи, в которых применяется PCA, – визуализация и большинство процедур сжатия, упрощения, минимизации данных для того, чтобы облегчить процесс обучения. Однако метод главных компонент не годится для ситуаций, когда исходные данные слабо упорядочены (то есть все компоненты метода характеризуются высокой дисперсией). Так что его применимость определяется тем, насколько хорошо изучена и описана предметная область.

**Сингулярное разложение** - в линейной алгебре сингулярное разложение, или SVD, определяется как разложение прямоугольной матрицы, состоящей из комплексных или вещественных чисел. Так, матрицу M размерностью [m\*n] можно разложить таким образом, что M = UΣV, где U и V будут унитарными матрицами, а Σ – диагональной.

Одним из частных случаев сингулярного разложения является метод главных компонент. Самые первые технологии компьютерного зрения разрабатывались на основе SVD и PCA и работали следующим образом: вначале лица (или другие паттерны, которые предстояло найти) представляли в виде суммы базисных компонент, затем уменьшали их размерность, после чего производили их сопоставление с изображениями из выборки. Современные алгоритмы сингулярного разложения в машинном обучении, конечно, значительно сложнее и изощрённее, чем их предшественники, но суть их в целом не изменилась.

**Анализ независимых компонент (ICA)** — это один из статистических методов, который выявляет скрытые факторы, оказывающие влияние на случайные величины, сигналы и пр. ICA формирует порождающую модель для баз многофакторных данных. Переменные в модели содержат некоторые скрытые переменные, причем нет никакой информации о правилах их смешивания. Эти скрытые переменные являются независимыми компонентами выборки и считаются негауссовскими сигналами. В отличие от анализа главных компонент, который связан с данным методом, анализ независимых компонент более эффективен, особенно в тех случаях, когда классические подходы оказываются бессильны. Он обнаруживает скрытые причины явлений и благодаря этому нашёл широкое применение в самых различных областях – от астрономии и медицины до распознавания речи, автоматического тестирования и анализа динамики финансовых показателей.

* 1. **Разведочный анализ данных.**

Целями разведочного анализа является 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.

Результаты разведочного анализа не используются для выработки управленческих решений. Их назначение — помощь в разработке наилучшей стратегии углубленного анализа, выдвижение гипотез, уточнение особенностей применения тех или иных математических методов и моделей. Без разведочного анализа углубленный анализ данных будет производиться практически «вслепую».

Это необходимо для того, чтобы максимизировать понимание набора данных:

* выявить базовую структуру;
* извлечь важные переменные;
* выявить выбросы и аномалии;
* проверить исходные предположения;
* разработать скупые модели;
* определять оптимальные параметры.

**Корреляционный анализ (Correlation analysis) - с**овокупность методов обработки данных с целью обнаружения статистической взаимосвязи между случайными величинами или признаками. Корреляционный анализ для двух случайных величин заключает в себе:

* Построение корреляционного поля и составление корреляционной таблицы;
* Вычисление выборочных коэффициентов корреляции и корреляционных отношений;
* Проверка статистической гипотезы значимости связи.

Методами корреляционного анализа решаются следующие задачи:

* Взаимосвязь. Установление наличия зависимости между двумя признаками и определение ее силы.
* Прогнозирование. Предсказание поведения одного признака на основе изменения другого, коррелирующего с первым.
* Отбор переменных. Корреляционный анализ позволяет производить выбор набора входных переменных для аналитической модели, в наименьшей степени коррелирующих между собой и в наибольшей степени коррелирующих с выходной переменной. Это позволяет сделать работу аналитических моделей более точной и устойчивой.

Сила корреляционной связи между двумя переменными характеризуется с помощью коэффициента корреляции. Конкретный вид (линейная, экспоненциальная и т.д.) корреляционной связи и ее параметры устанавливается в рамках регрессионного анализа. Впервые элементы корреляционного анализа в научных исследованиях начал применять французский палеонтолог Жорж Кювье, который и ввел в научный обиход термин «корреляция». Значительный вклад в развитие теории корреляционного анализ внесли Френсис Гальтон, Карл Пирсон, Чарльз Спирмен, Морис Кендалл и другие.

**Факторный анализ (Factorial analysis) -** направление математической статистики (один из разделов многомерного статистического анализа), объединяющее вычислительные методы, которые в ряде случаев позволяют получить компактное описание исследуемых явлений на основе обработки больших массивов информации.

От других средств подобного «сжатия информации» (например, распространенных методов статистической группировки объектов) факторный анализ отличается тем, что не опирается на заранее заданный, априорный перечень факторов, влияющих на исследуемые процессы или объекты, а, наоборот, при соблюдении определенных правил и предосторожностей помогает обнаружить наиболее важные из них, причем скрытые (латентные).

Основные идеи факторного анализа были заложены английским психологом и антропологом Фрэнсисом Гальтоном во второй половине 19 в.

**Линейный дискриминантный анализ (ЛДА)**, а также связанный с ним линейный дискриминант Фишера — методы статистики и машинного обучения для нахождения линейных комбинаций признаков, наилучшим образом разделяющих два или более класса объектов или событий. Полученная комбинация может быть использована в качестве линейного классификатора или для сокращения размерности пространства признаков перед последующей классификацией.

ЛДА представляет собой раздел многомерного статистического анализа, содержанием которого является разработка методов решения задач различения (дискриминации) объектов наблюдения по набору признаков. Иными словами, он позволяет изучать различия между двумя и более группами объектов по нескольким признакам одновременно.

ЛДА тесно связан с дисперсионным анализом и регрессионным анализом, также пытающимися выразить какую-либо зависимую переменную через линейную комбинацию других признаков или измерений. В этих двух методах зависимая переменная — численная величина, а в ЛДА она является величиной номинальной (меткой класса). Помимо того, ЛДА имеет схожие черты с методом главных компонент и факторным анализом, которые ищут линейные комбинации величин, наилучшим образом описывающие данные.

Можно выделить три вида задач дискриминантного анализа:

* определение дискриминирующих признаков (т.е. признаков, которые позволяют отнести наблюдение к той или иной группе);
* построение дискриминирующей функции;
* прогнозирование будущих событий, связанных с попаданием объекта в ту или иную группу на основе значений его признака (например, предсказание выживаемости пациента после операции).

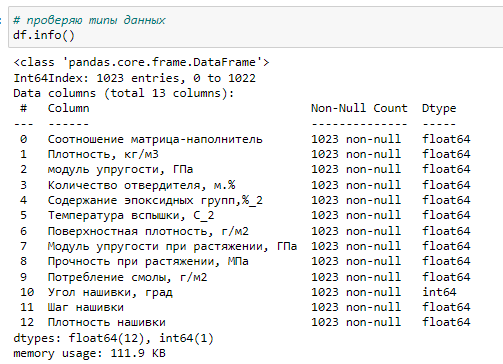
Основной целью дискриминации является поиск линейной комбинации признаков (называемых дискриминантными признаками), которые позволили бы наилучшим образом разделить рассматриваемые группы.

1. **Практическая часть**

**2.1. Предобработка данных**

На первом этапе делаем объединение по индексу (так как других общих признаков нет) двух датасетов в один. Объединение выполняем по типу INNER. Видим, что размер датасетов не одинаковый. Разница составляет 17 строк. Принимаем решение разницу в 17 строк отбросить, чтобы уравнять размер датасетов.

Далее делаем проверку объединённого датасета на наличие пропусков и определяем тип данных.

  
Рисунок 1 – Информация о пропусках и типах данных

Полученный датасет имеет 1023 записи (10 входных параметров и 3 выходных параметра). Пропуски в полученном датасете отсутствуют, тип данных float64 (число с плавающей точкой) и int64 (целые числа) - информация о типах данных представлена на Рисунке 1, описательные статистики – предоставлены на Рисунке 2.

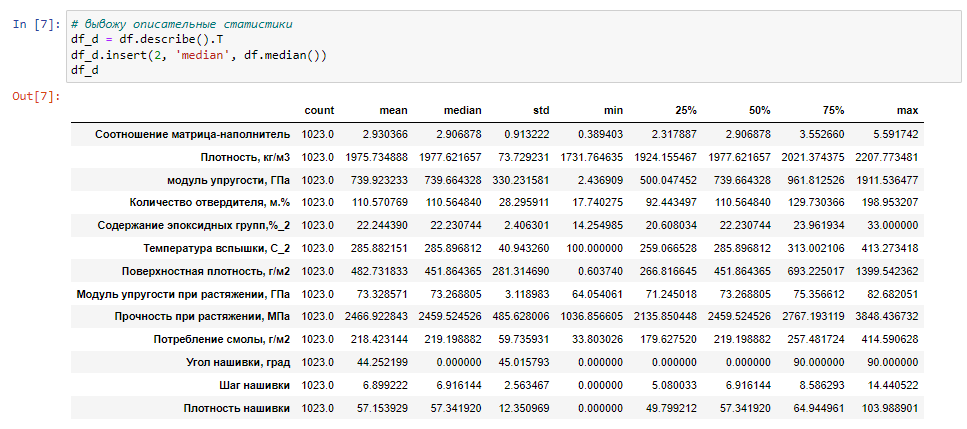


Рисунок 2 – Описательные статистики

На следующем этапе сделаем визуализацию данных. Чтобы получить наглядное представление о характерах распределений переменных необходимо построить гистограммы и диаграммы размаха *«ящик с усами»* (такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы).

Производим отрисовку гистограммы распределения для каждой переменной.

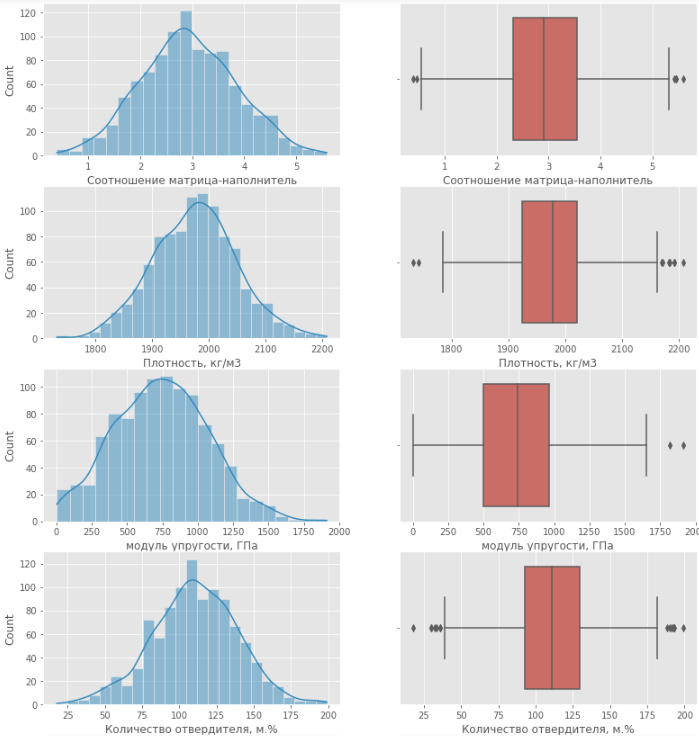


Рисунок 3 – Гистограмма и диаграмма размаха №1

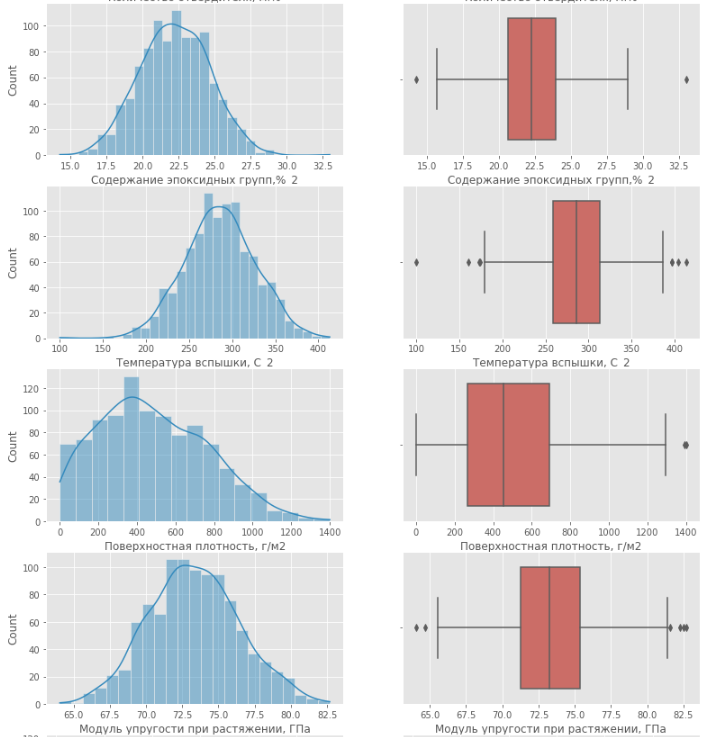


Рисунок 4 – Гистограмма и диаграмма размаха №2

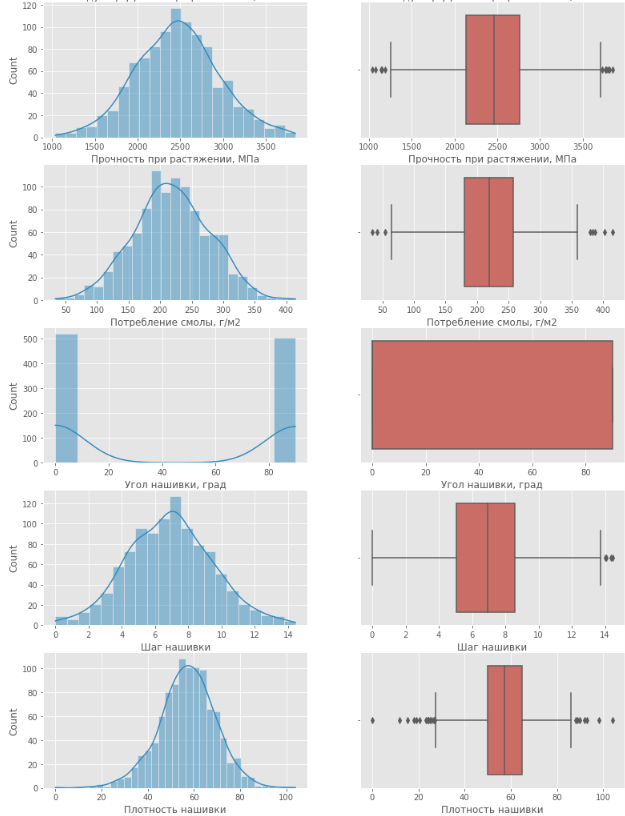


Рисунок 5 – Гистограмма и диаграмма размаха №3

В соответствии с рисунками 3-5 распределение нормальное, имеются выбросы. Гистограмма угла нашивки имеет бинарное значение. Выбросы не имеют экстремально больших отклонений.

Следующим шагом разведочного анализа построим попарные графики рассеяния точек и тепловую карту матрицы корреляции для визуализации наличия зависимости признаков.

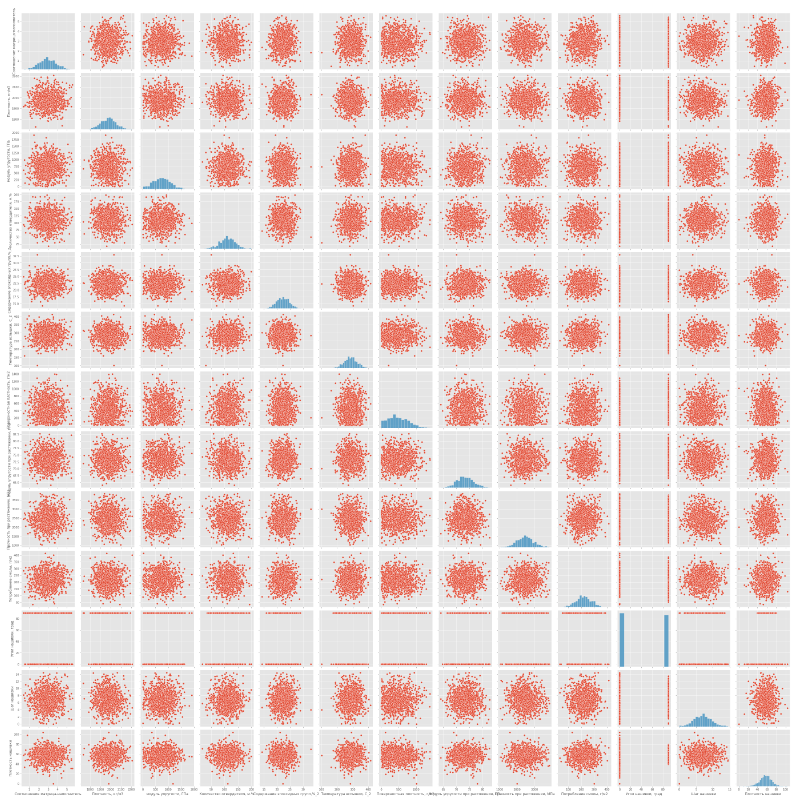
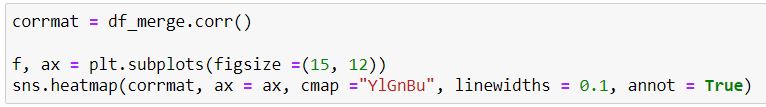
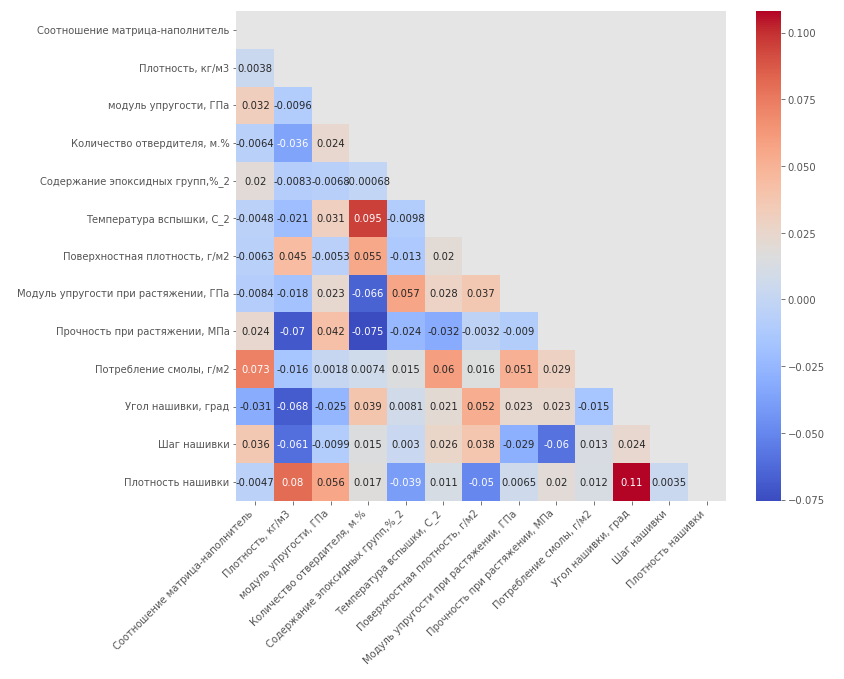


Рисунок 6 – График попарного рассеяния точек

Рисунок 7 – Тепловая карта матрицы корреляции

Попарное сравнение признаков зависимости не выявило. Корреляционная матрица тоже не выявила каких-либо зависимостей. Зависимость между признаками очень низкая.

По результатам проведенного разведочного анализа можно сделать следующие выводы. Почти у всех признаков нормальное распределение, по имеющейся информации данные являются предварительно обработанными заказчиком. Пропуски в заполнении данных отсутствуют. Взаимозависимость признаков почти полностью отсутствует. Очень слабая зависимость есть между углом нашивки и плотность нашивки.

1. **Практическая часть**

**2.1 Предобработка данных**

Теперь приступим к удалению выбросов. Для этого посчитаем количество выбросов двумя основными способами. Методом 3-х сигм и методом межквартильных расстояний.



Рисунок 8 – Количество выбросов

В соответствии с рисунком 33, метод 3-х сигм нашел меньше выбросов 24 выброса, против 93 у метода межквартильных расстояний. Учитывая тот факт, что данные были предварительно подготовлены заказчиком, с целью избегания удаления тех данных, которые могут оказаться не выбросами, а особенностями датасета, оставим свой выбор за методом 3-х сигм.

Для дополнительной проверки посчитаем распределение выбросов по признакам.

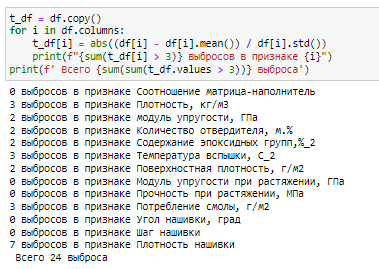


Рисунок 9 – Распределение выбросов по признакам

Как видно из рисунка 9, выбросы распределены по разным признакам, чрезмерной концентрации в одном признаке не прослеживается. Можно приступать к удалению выбросов.

Очистив, дата сет от выбросов, построим график распределения плотности ядра, для оценки необходимости нормализации.

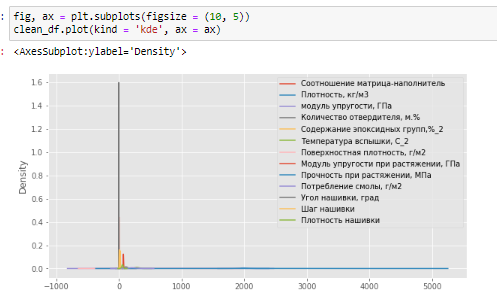


Рисунок 10 – График распределения плотности ядра

В соответствии с рисунком 10 наглядно видно, что данные находятся в очень разных диапазонах. Так как диапазоны очень разные, необходимо провести нормализацию данных.

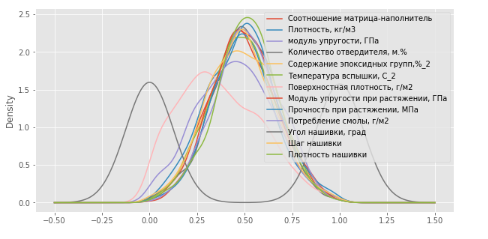


Рисунок 11 – График распределения плотности ядра после нормализации

Предобработку данных закончили. Удалили выбросы и нормализовали значения данных.

**2.2. Разработка и обучение модели**

В соответствии с поставленной задачей, нужно осуществить разработку и обучение моделей машинного обучения для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении». Для каждого признака построение моделей осуществляется раздельно. Разделение нормализованных данных на обучающую и тестовую выборки (в соотношении 70 на 30%, согласно поставленной задаче).

Для признака «Модуль упругости при растяжении» были разработаны и обучены следующие модели в соответствии с рисунком 12:

− модель на основе линейной регрессии (метод LinearRegression);

− модель k ближайших соседей (метод KNeighborsRegressor());

− модель случайный лес (метод RandomForestRegressor());



Рисунок 12 – График оценки моделей

**2.3. Тестирование модели**

Посмотрим на графики для целевой параметра - модуль упругости при растяжении:

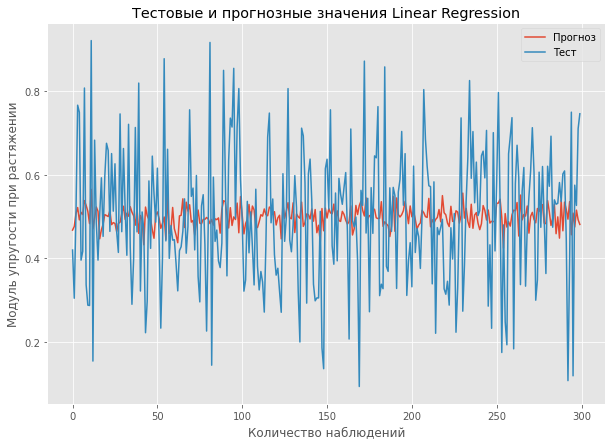


Рисунок 13 – Linear Regression

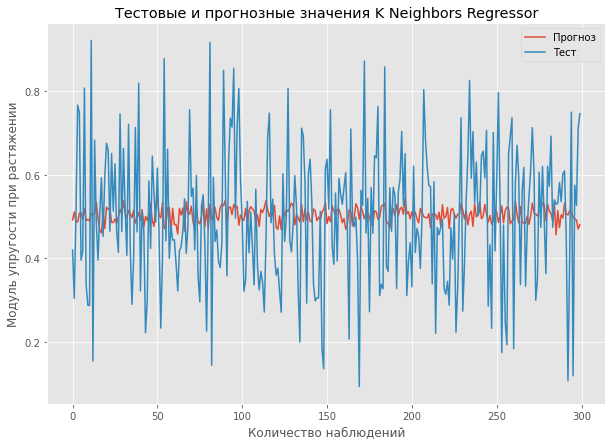


Рисунок 14 – K Neighbors Regressor

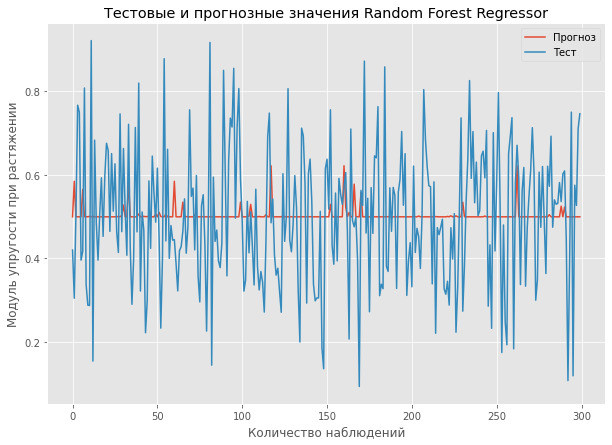


Рисунок 15 – Random Forest Regressor

Посмотрим на графики для целевой параметра - прочность при растяжении:

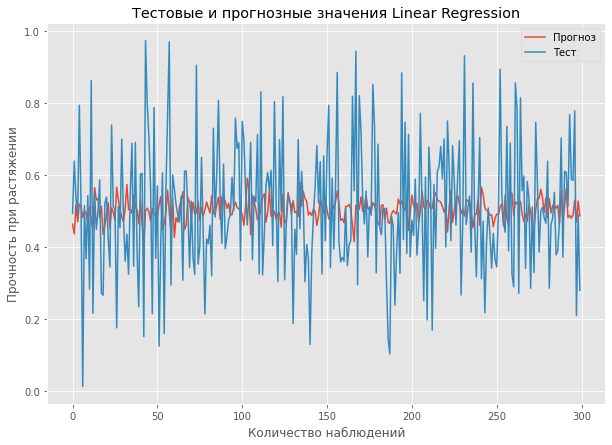


Рисунок 16 – Linear Regression

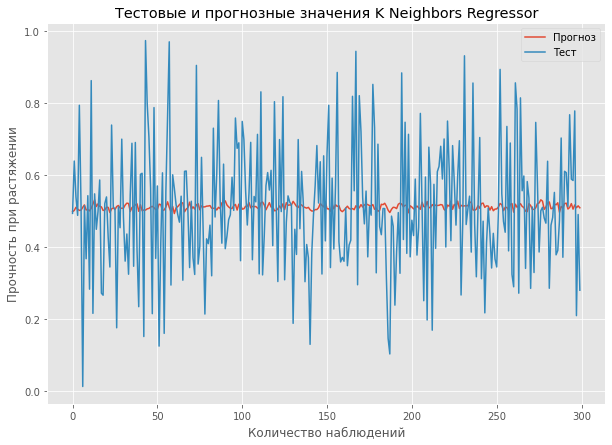


Рисунок 17 – K Neighbors Regressor

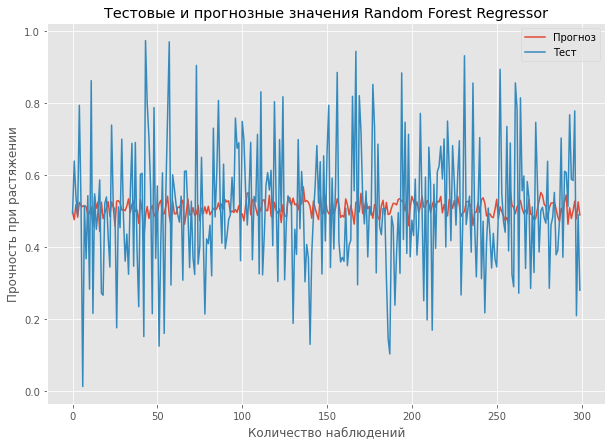


Рисунок 18 – Random Forest Regressor

**2.4 Нейронная сеть, рекомендации** **соотношения матрица-наполнитель**

При использовании библиотеки Keras (библиотека глубокого обучения) разработана простая модель для рекомендации соотношения «матрица-наполнитель».

Описание для архитектуры нейронной сети:

В модели используется четыре скрытых уровня. В первом уровне находятся 64 нейрона. В других скрытых уровнях находятся 64, 64 и 1 нейрона. Происходит сжатие информации - при снижении числа нейронов, которую сеть обработала на предыдущих уровнях.

После проведения обучения модели для была выполнена оценка модели

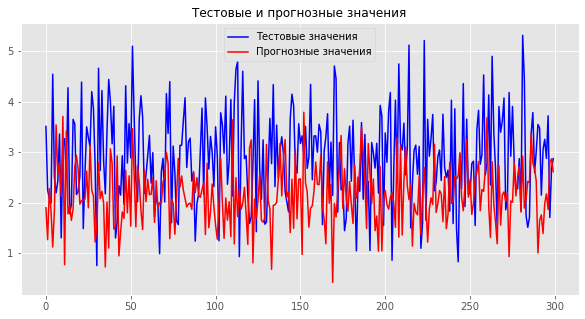


Рисунок 19 – Оценка нейронной сети №1

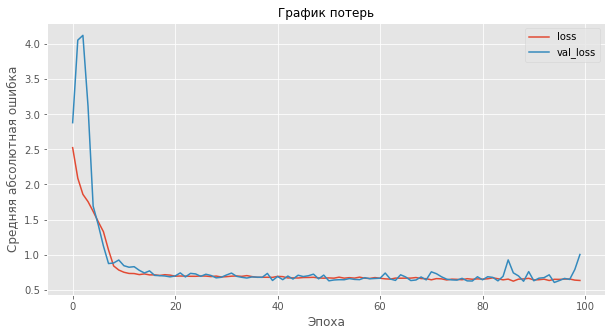


Рисунок 20 – Оценка нейронной сети №2

**2.5 Разработка приложения**

При использовании библиотеки Keras (библиотека глубокого обучения) разработано приложение для рекомендации соотношения «матрица-наполнитель».

Приложение в качестве аргументов принимает параметры, соответствующие входным параметрам исходных данных («Соотношение матрица- наполнитель %», «Плотность, кг/м3 %», «модуль упругости, Гпа %», «Количество отвердителя, м. %», «Содержание эпоксидных групп,%\_2 %», «Температура вспышки, С\_2 %», «Поверхностная плотность, г/м2 %», «Модуль упругости при растяжении, Гпа %», «Потребление смолы, г/м2 %», «Угол нашивки, град %», «Шаг нашивки %», «Плотность нашивки %») итого 12 параметров и возвращает параметр соответствующий выходному параметру («Прочность при растяжении, Мпа %»)



Рисунок 21 – Код разработанного приложения

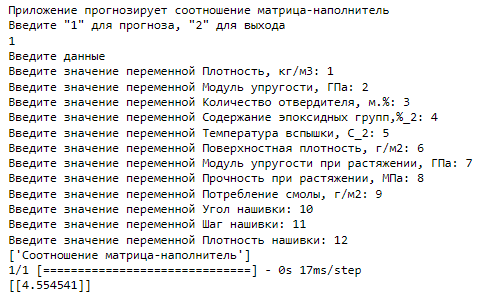


Рисунок 22 – Результат работы приложения

**2.5 Создание удаленного репозитория.**

Для проекта выпускной квалификационной работы был создан отдельный репозиторий на GitHub. Ссылка на репозиторий: <https://github.com/katushas009/bmstu_data_science_muraveva>

**3. Заключение**

При решении данной задачи не удалось достигнуть желаемых результатов, так как признаки плохо между собой коррелируют ни один из признаков не приближается близко к параметрам упругости и прочности. Соответственно не удалось подобрать оптимальную комбинацию признаков. Поэтому модель плохо предсказывает свойства материала.

**4.** **Список используемой литературы и ссылки на веб-ресурсы**

[1] К. Андерсон, Аналитическая культура. От сбора данных до бизнес-результатов: монография. Москва: O’Reilly, 2017, 392 с.

[2] How to choose a machine learning model in Python? – Режим доступа: https://www.codeastar.com/choose-machine-learning-models-python/ (дата обращения 03.04.2022)

[3] Язык программирования Python: - Режим доступа: https://www.python.org/ (дата обращения 01.04.2022)

[4] Библиотека Pandas – Режим доступа: https://pandas.pydata.org/ (дата обращения 01.04.2022)

[5] Библиотека Matplotlib – Режим доступа: https:// https://matplotlib.org/ (дата обращения 01.04.2022)

[6] Библиотека Sklearn – Режим доступа: https://scikit-learn.org/stable/ (дата обращения 01.04.2022)

[7] В.В. Васильев, В.Д. Протасов, В.В. Болотин и др.: Композитные материалы: справочник. Москва: Машиностроение, 1990, 510 с.

[8] Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с.