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 Стоянов Павел Александрович

Москва, 2023

**Содержание**

Введение

1 Аналитическая часть

* 1. Постановка задачи
  2. Характеристика датасета

1.3 Описание используемых методов

1.4 Разведочный анализ данных

2.Практическая часть

2.1 Предобработка данных

2.2 Разработка и обучение модели

2.3 Тестирование модели

2.4 Написание нейронной сети, рекомендующей соотношение «матрица-наполнитель»

2.5 Разработка приложения

2.6 Создание удаленного репозитория и загрузка результатов работы

Заключение

Список использованной литературы

Приложения

**Введение**

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними, что схематично отображено на рисунке 1. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. 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. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом.

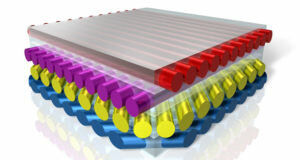


Рисунок 1—Структура композиционного материала

5 интересных фактов о композитах:

1. одни из самых первых рукотворных композитных материалов - высушенные на солнце глиняные кирпичи с добавлением рубленной соломы. Первое использование этого метода относится к 1500 году до нашей эры. Древние Египтяне оставили на стенах пирамид изображения этой технологии;
2. 1200 год нашей эры, постарались монголы: они создали первый композиционный лук из таких материалов, как древесина, кость и животный клей;
3. самый известный искусственный композитный материал – бетон;
4. при пошиве спортивной одежды и обуви используется материал Gore-Tex, который является композитом из слоев различных материалов. Он одновременно водонепроницаемый  и пропускающий молекулы воздуха;
5. почти половина деталей современного самолета произведены из композитов;
6. существует самовосстанавливающийся полимер. Этот композит содержит химические вещества, которые образуют новый слой при повреждении поверхности изделия.

Современные композиты изготавливаются из материалов: полимеры, керамика, стеклянные и углеродные волокна.

Сейчас мировой рынок композитов составляет 80 млрд долл. На нем лидируют Китай 32 процента (25,6 млрд долл.) и США с 26 процентами (21,6 млрд долл.). Сегмент России – 1 процент (1,1 млрд долл.). В структуре российского рынка композитов преобладает, в частности, строительная индустрия – 35 процентов (22,9 млрд руб.). Это водоотводные лотки, изолирующие накладки, композитные материалы и т.д. На втором месте – гражданское авиа- и судостроение: 19 процентов (12,4 млрд руб.).

По структуре композиты делятся на несколько основных классов: волокнистые, дисперсно-упрочнённые, упрочнённые частицами и нанокомпозиты.

Композиционные материалы классифицируют по следующим основным признакам: типу матрицы, виду армирующего элемента, особенностям макростроения и методам получения. Сначала осуществляют ориентировочный выбор материала матрицы, основных наполнителей и арматуры, а также технологии формирования изделий.

В составе композита принято выделять матрицу/матрицы и наполнитель/наполнители, последние выполняют функцию армирования (по аналогии с арматурой). В качестве наполнителей композитов как правило выступают углеродные или стеклянные волокна, а роль матрицы играет полимер. Определение содержания и относительного расположения различных армирующих элементов в матрице, прежде всего, зависит от таких исходных требований, как прочность и жесткость, тепло- и электропроводность, технологичность, стоимость материала и т. д. Часто процессы формирования изделия и композиционного материала совмещаются.

Композиты, в которых матрицей служит [полимерный](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BB%D0%B8%D0%BC%D0%B5%D1%80%D1%8B) материал, являются одним из самых многочисленных и разнообразных видов материалов. В качестве наполнителей ПКМ используется множество различных веществ: Стеклопластики, Углепластики, Боропластики , Органопластики, Полимеры, наполненные порошками, Текстолиты.

При создании композитов на основе [металлов](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%B0%D0%BB%D0%BB) в качестве матрицы применяют [алюминий](https://ru.wikipedia.org/wiki/%D0%90%D0%BB%D1%8E%D0%BC%D0%B8%D0%BD%D0%B8%D0%B9), [магний](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D0%B3%D0%BD%D0%B8%D0%B9), [никель](https://ru.wikipedia.org/wiki/%D0%9D%D0%B8%D0%BA%D0%B5%D0%BB%D1%8C), [медь](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B4%D1%8C) и так далее. [Наполнителем](https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D0%BF%D0%BE%D0%BB%D0%BD%D0%B8%D1%82%D0%B5%D0%BB%D1%8C) служат или высокопрочные [волокна](https://ru.wikipedia.org/wiki/%D0%92%D0%BE%D0%BB%D0%BE%D0%BA%D0%BD%D0%BE), или тугоплавкие, не растворяющиеся в основном металле частицы различной [дисперсности](https://ru.wikipedia.org/wiki/%D0%94%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D0%BD%D0%BE%D1%81%D1%82%D1%8C).

Армирование керамических материалов волокнами, а также металлическими и керамическими дисперсными частицами позволяет получать высокопрочные композиты, однако, ассортимент волокон, пригодных для армирования керамики, ограничен свойствами исходного материала.

Для решения проблемы моделирования композитов 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). У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично.

**1 Аналитическая часть**

**1.1 Постановка задачи**

На основе набора данных из файлов X\_bp.xlsx и X\_nup.xlsx (объединение делать по индексу тип объединения INNER) необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов, а именно:

1. обучить нескольких моделей для определения значений «Модуль упругости при растяжении, ГПа» и «Прочность при растяжении, МПа»;
2. написать нейронную сеть, которая будет рекомендовать характеристику «Соотношение матрица-наполнитель»;
3. разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз характеристик «Модуль упругости при растяжении, ГПа» и «Прочность при растяжении, МПа» или характеристику «Соотношение матрица-наполнитель».

Кейс основан на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

**1.2 Характеристика датасета**

Начнем с описательного анализа данных, в большинстве случаев он используется для первичного определения типов информации.

Он включает:

проверку типа данных;

удаление нерелевантных столбцов (не используем, не имеют для анализа никакого смысла);

переименование столбцов (улучшает читаемость набора данных);

удаление повторяющихся (дубликатов) строк;

удаление отсутствующих или нулевых значений либо замена их средним/медианой или модой для этого столбца;

обнаружение выбросов, идентификация их таковыми;

На входе имеются набор данных (файлы X\_bp.xlsx и X\_nup.xlsx) с начальными свойствами компонентов композиционных материалов.

Файл X\_bp.xlsx содержит 1023 записи и одиннадцать столбцов с признаками (без названия, соотношение матрица-наполнитель, плотность, модуль упругости, количество отвердителя, содержание эпоксидных групп, температура вспышки, поверхностная плотность, модуль упругости при растяжении, прочность при растяжении, потребление смолы), в том числе три выходные переменные, которые нас интересуют (соотношение матрица-наполнитель, модуль упругости при растяжении, прочность при растяжении).

Файл X\_nup.xlsx содержит 1040 записей и четыре столбца с признаками (без названия, угол нашивки, шаг нашивки, плотность нашивки).

Колличество записей в файлах отличается на 17 строчек. Учитывая условие задачи «объединение делать по индексу тип объединения INNER», после объединения файлов в один, теряем эти 17 строчек. Общий размер набора данных и отсутствие основной и важной инфо для этих строчек в объединяемом файле позволяют нам так поступить. В объединенном файле присутствует первый столбец без названия с индексом, который не несет никакой информации, его удаляем, для упрощения дальнейшей работы. Получаем датасет в 1023 строчки и 13 столбцов. Пропусков в данных нет, нулевых значений также нет, дубликатов нет, все значения имеют вещественный тип данных, что видно на рисунке 2.

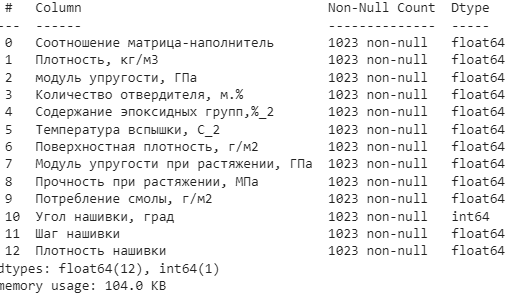


Рисунок 2—Описание объединенного датасета

Практически все значения признаков по столбцам уникальны, как видно из рисунка 3.

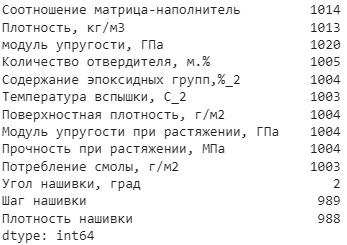
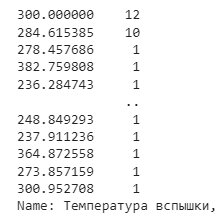
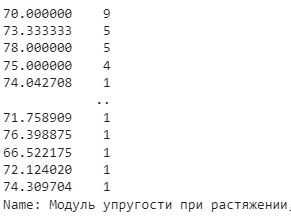
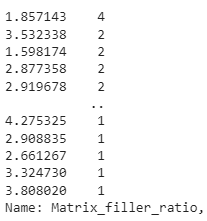
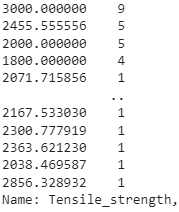
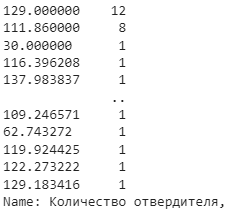


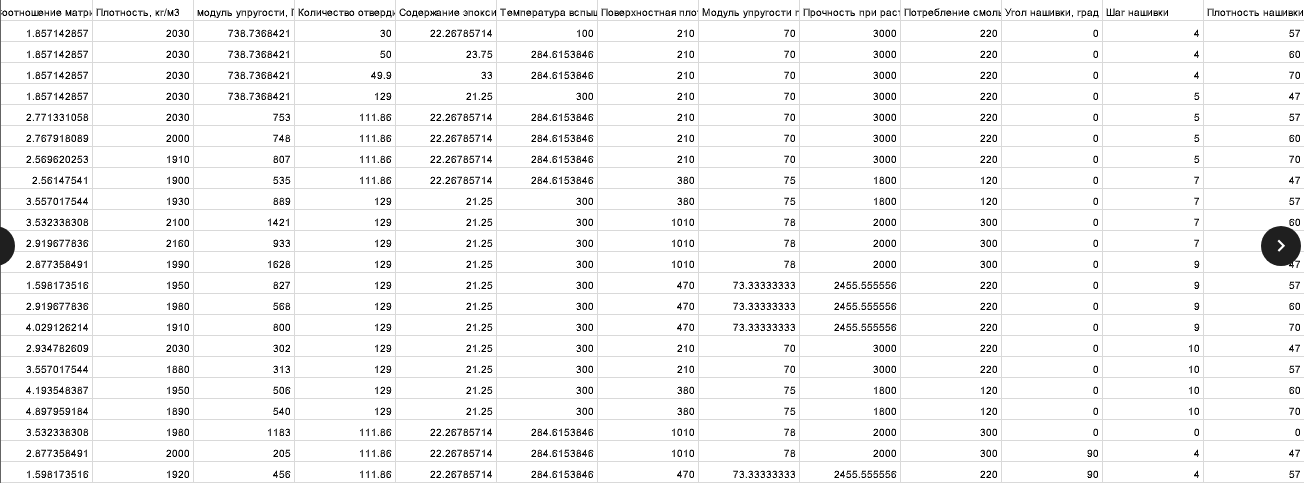
Рисунок 3—Наличие уникальных значений

Проанализировав наиболее часто встречающиеся значения по всем 13 столбцам, отражено на рисунке 4, выясняем, что первые 23 строки в объединенном датасете полностью сгенерированы, что видно на рисунке 5. Причина этого не ясна, т.к. при размерности в 1000 строк полная генерация дополнительных 23 строк ни на что не влияет.

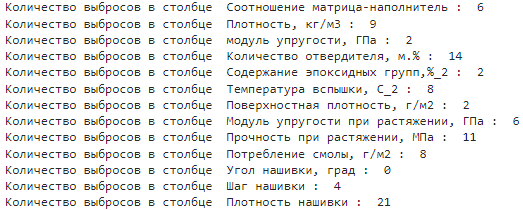
  

Рисунок—4 Наиболее часто встречающиеся значения



Рисунок—5 Сгенерированные строки в датасете

Значений похожих на выбросы согласно рисунку 6 в датасете небольшое колличество.

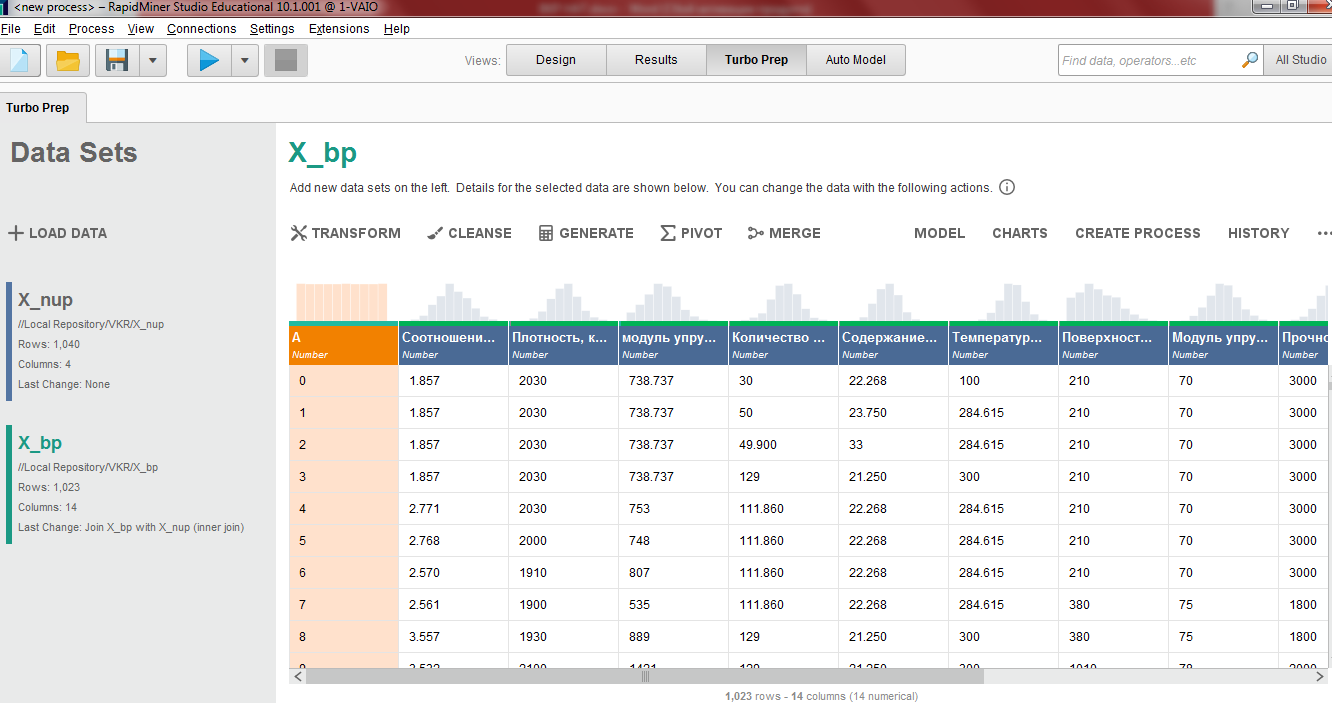
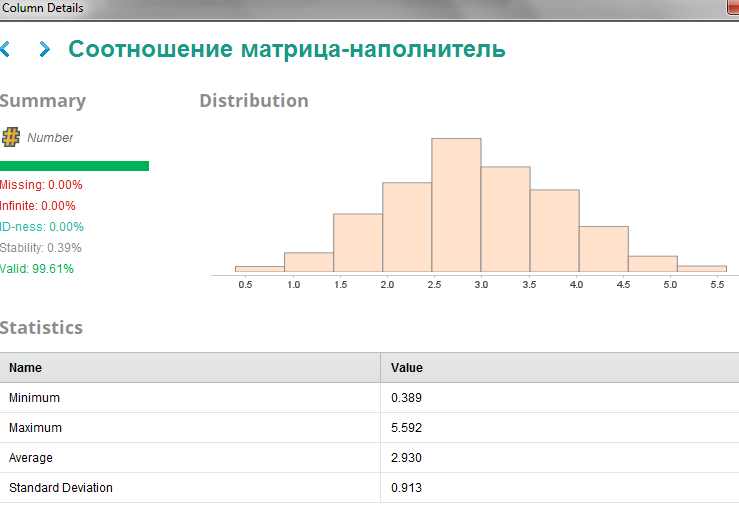


Рисунок—6 Количество предполагаемых выбросов в датасете

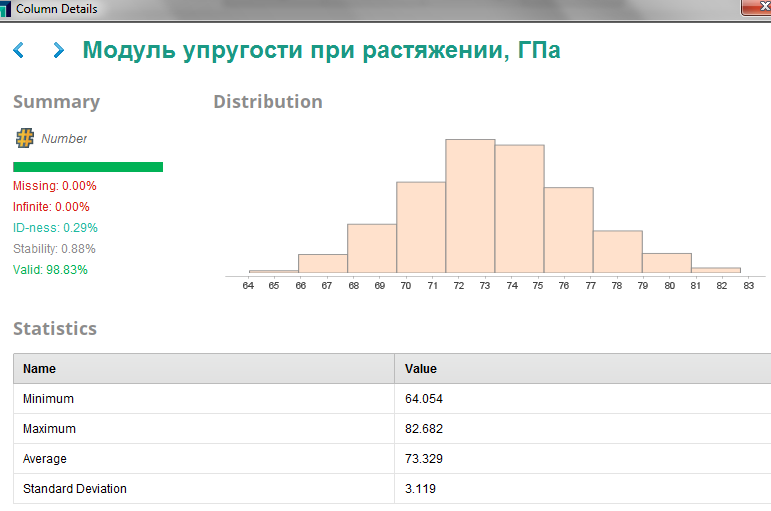
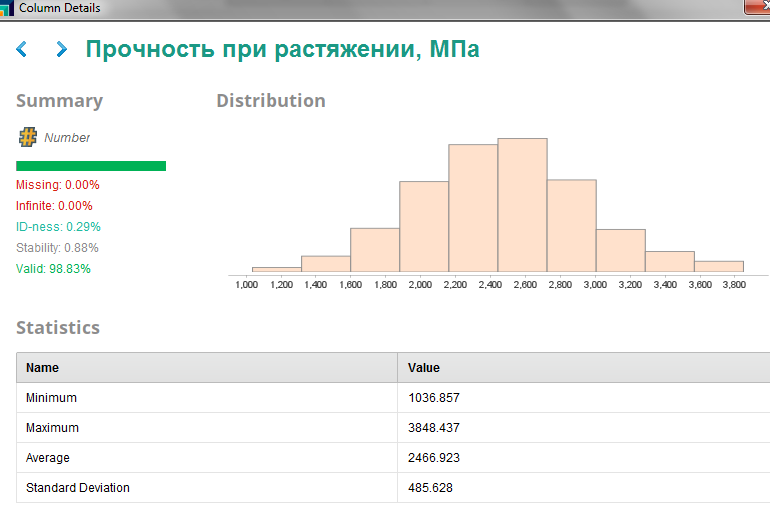
Чтобы идентифицировать их таковыми и принять решение удалить, нужно выбросы оценить, иначе информация может быть упущена. В сочетании с фактическими данными выбросы можно разделить на "истинные аномалии" и "ложные аномалии". Экспертная оценка выбросов нам не доступна (не являемся экспертами), т.к. опирается на вопросы:

1. является ли выброс результатом ошибки ввода данных?
2. влияет ли выброс на результаты анализа?
3. влияет ли выброс на предположения?

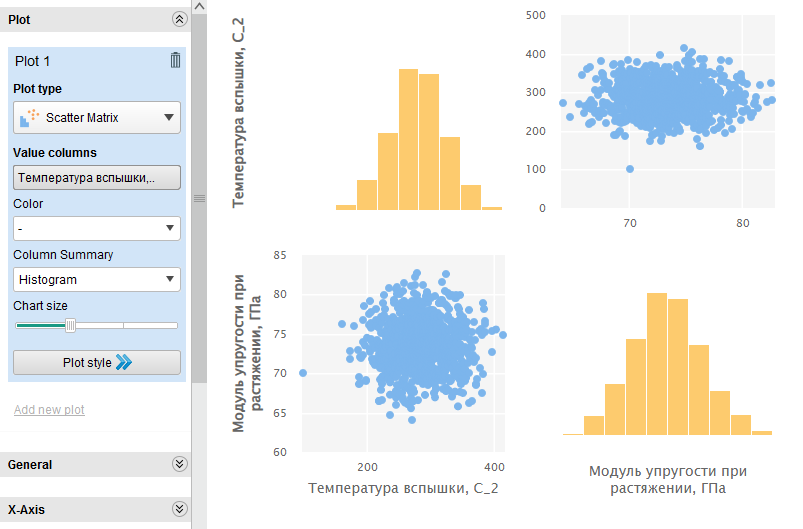
Посмотрим характеристику датасета в Longinom, рисунки 7-10:

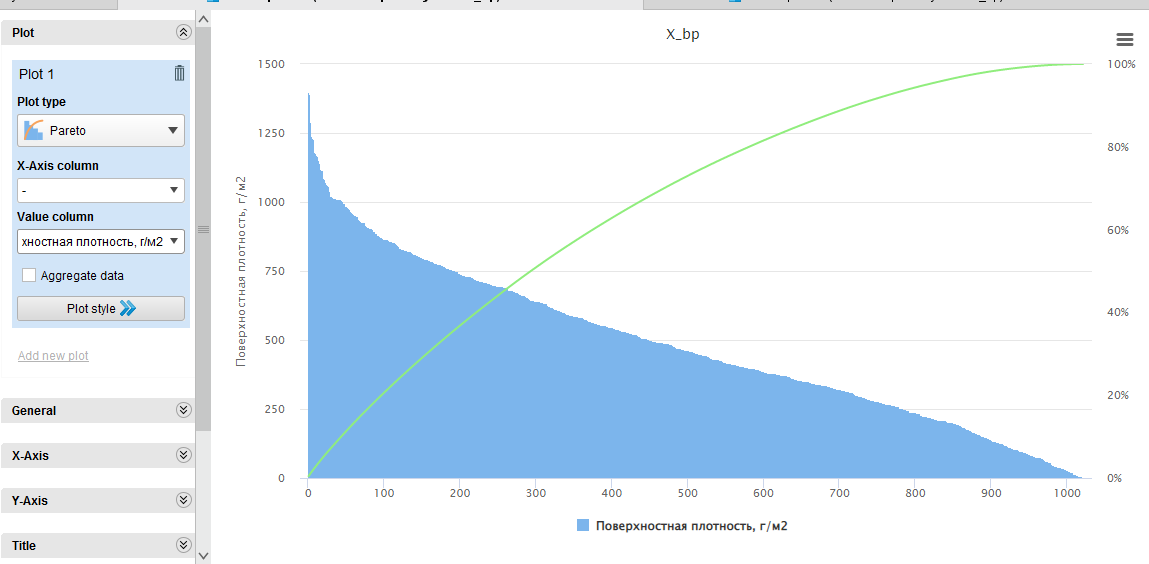
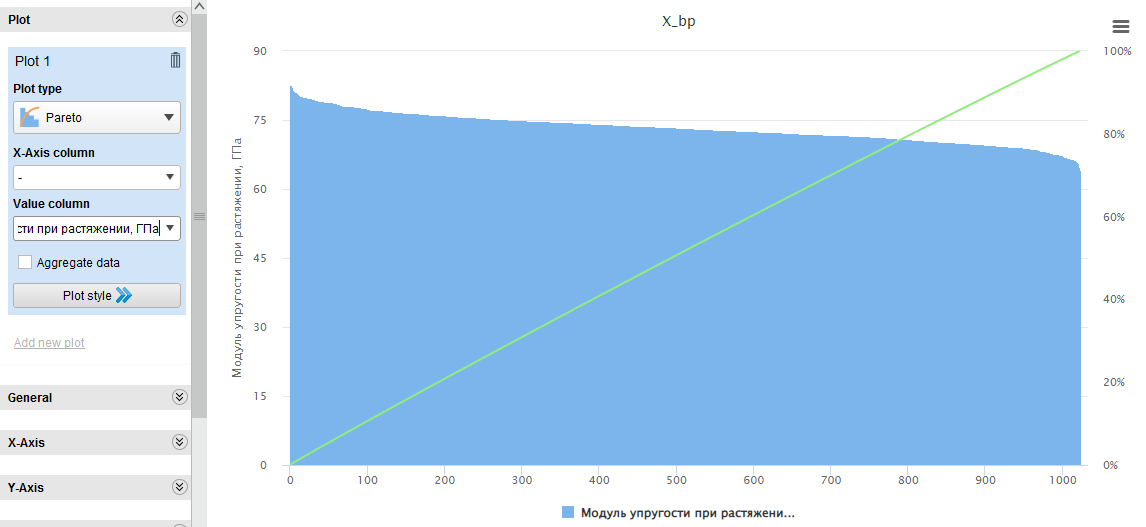
Рисунок—7

Рисунок—8



Рисунок—9



Рисунок—10

**1.3 Описание используемых методов**

При выборе методов будем опираться на рекомендации экспертов, уже опробовавших разные методы при работе с композитными материалами и отметивших точность отдельных их них. Ниже приведены выдержки из работ по этой теме.

Реутов Ю.А.: «Решение задачи «структура – свойство» в общем случае может выполняться различными методами машинного обучения, такими как: классическая регуляризация, статистический анализ, нейронные сети, метод опорных векторов, кластеризация, алгоритмические композиции и другие».

Чун-Те Чен и Грейс Х. Гу в своей статье дают краткий обзор некоторых основных алгоритмов машинного обучения и обзор недавних исследований с использованием моделей машинного обучения для прогнозирования механических свойств композитов:

«Рассмотрели линейную регрессию, логистическую регрессию, нейронные сети (NN), сверточные нейронные сети (CNN) и гауссовский процесс (GP) в контексте проектирования материалов, ссылаясь на исследования (как экспериментальные, так и вычислительные) по применению этих алгоритмов машинного обучения к композитным исследованиям (включая нанокомпозиты).

При линейной регрессии как только функция ошибок (МSE) модели определена, веса модели могут быть рассчитаны с помощью алгоритма оптимизации, такого как классический стохастический градиентный спуск или алгоритм оптимизации Адама. Более сложные модели машинного обучения (нейронные сети), в целом дают более точные прогнозы. Однако в исследованиях по композитам линейные модели давали ценную информацию, например, какие входные переменные были более важными (с более сильным влиянием) для прогноза».

Авторы показали, что использование машинного обучения для прогнозирования механических свойств композитов на порядки быстрее, чем обычный анализ методом конечных элементов.

«После обучения линейной модели оптимальные конструкции (конструкции с наивысшей прочностью или т.п) композитов могут быть созданы непосредственно с использованием информации о весе в линейной модели, без необходимости какой-либо выборки, оптимизации или процесса поиска.

Альтернативный подход к достижению высокой способности к обучению состоит в использовании линейной регрессии с функцией активации для создания искусственного нейрона (персептрона) и соединения ряда нейронов для создания искусственной НС. Самой базовой архитектурой NN является многослойный персептрон, который здесь также обозначается как NN. Он состоит из нескольких скрытых слоев (слои, за исключением входного и выходного слоев), и каждый слой содержит некоторое количество нейронов. Нелинейные функции, такие как выпрямленная линейная единица и сигмоида, используются в качестве функций активации для введения нелинейности. Авторы применили оптимизацию роя частиц для настройки архитектуры (количества скрытых слоев и нейронов) своей модели нейронной сети и показали, что их модель нейронной сети может точно предсказать и рассчитать предел прочности на сжатие засыпки из цементного теста без ограничений.  В работе авторов обучающие данные были сгенерированы с помощью трехмерных фильтров Гаусса с различными ковариационными матрицами. Производительность обучающих данных была рассчитана с помощью анализа методом конечных элементов. Входные переменные включают топологию композитов, состоящих из двух основных материалов (жесткого и мягкого), а выходная переменная — эффективная жесткость. Были изучены различные архитектуры моделей CNN для настройки гиперпараметров (количества слоев и фильтров). Авторы показали, что их модели CNN могут давать высокоточные прогнозы эффективной жесткости композитов на основе заданной топологии.

Хотя глубокие нейронные сети (например, NN и CNN) с множеством скрытых слоев и нейронов теоретически могут фиксировать любые сложные закономерности в данных, им требуется большой объем обучающих данных, чтобы изучить скрытые закономерности без переобучения.

Для некоторых задач простые алгоритмы машинного обучения не могут фиксировать сложные закономерности в данных, а создание большого количества обучающих данных для алгоритмов глубокого обучения невозможно из-за времени и стоимости проведения экспериментов или моделирования. GP, который представляет собой непараметрический подход, служит альтернативным методом для задач с высокой нелинейностью. GP представляет собой набор случайных величин и предполагает, что все входные и выходные переменные имеют совместное распределение Гаусса. Вместо того, чтобы задавать гипотезу для модели ML и находить оптимальные значения весов в модели, GP производит распределение всех возможных функций, которые согласуются с наблюдаемыми (обучающими) данными. Таким образом, количество параметров в GP не ограничено и растет с увеличением количества обучающих данных.

 GP может быть очень эффективным в вычислительном отношении, когда количество обучающих данных невелико».

Итак, в своей работе будем использовать метод обучения с учителем через задачу регрессии (на выходе нас интересует числовое значение). Будут применены следующие методы:

1. линейная регрессия;
2. лассо.

Методы: достоинства, недостатки и области применения каждого из методов.

Составить итоговую сравнительную таблицу с указанием среди прочего априорных предпосылок к работоспособности каждого метода.

**1.4 Разведочный анализ данных**

Краткое описание методов разведочного анализа данных, которые используются для первоначального анализа.

Нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек.

Для каждой колонки получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков.

Разведочный (EDA) или, как его еще принято называть, исследовательский анализ применяется для выявления тенденций. При его использовании выявляют закономерности, обобщаются основные характеристики.

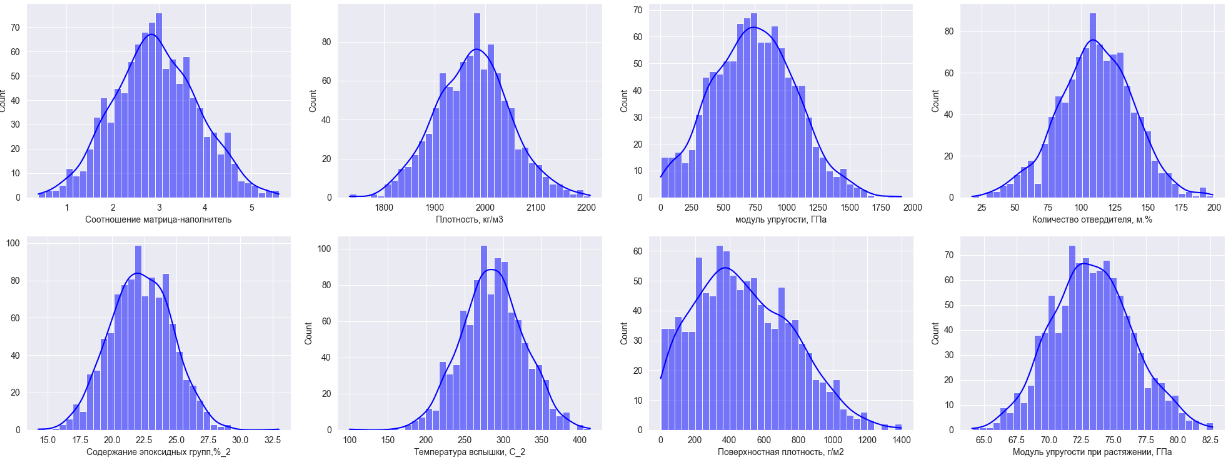
К основным методам разведочного анализа относится:

1. процедура анализа распределений переменных (чтобы выявить переменные с несимметричным или негауссовым распределением, в том числе и бимодальные). Гистограмма позволяет "на глаз" оценить нормальность эмпирического распределения. На гистограмму также накладывается кривая нормального распределения.
2. анализ корреляционных матриц с целью поиска коэффициентов, превосходящих по величине определенные пороговые значения. Корреляция представляет собой меру зависимости переменных. Происходит проверка значимых (ожидаемых и неожиданных) корреляций, попытка понять общую природу обнаруженной статистической значимости;
3. анализ многовходовых таблиц частот ("послойный" последовательный просмотр комбинаций уровней управляющих переменных). Это метод анализа категориальных переменных, показывающий каким образом различные группы данных распределены в выборке.
4. многомерный разведочный анализ (для поиска закономерностей в многомерных данных или последовательностях одномерных данных). К ним относятся: кластерный анализ, факторный анализ, анализ дискриминантных функций, многомерное шкалирование, анализ временных рядов и.

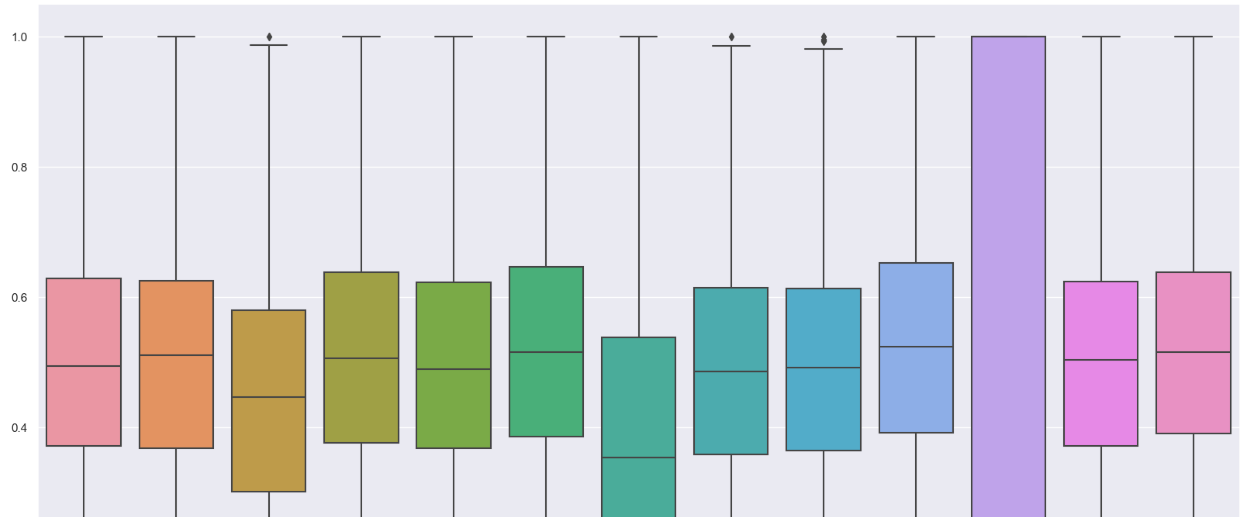
Статистические методы - среднее значение, медиана, мода, стандартное отклонение, дисперсия, квантили и др.

Визуализации выделяют взаимосвязи в данных и раскрывают информацию, видимую человеческому глазу, которую нельзя передать только числами и цифрами. Используем для этого библиотеки Pandas, Seaborn, Matplotlib.

Посмотрим на данные на рисунках 11 и 12. Гистограмма отображает частоту появления переменных в определенном интервале.



Рисунок—11 Гистограммы характера распределения признаков

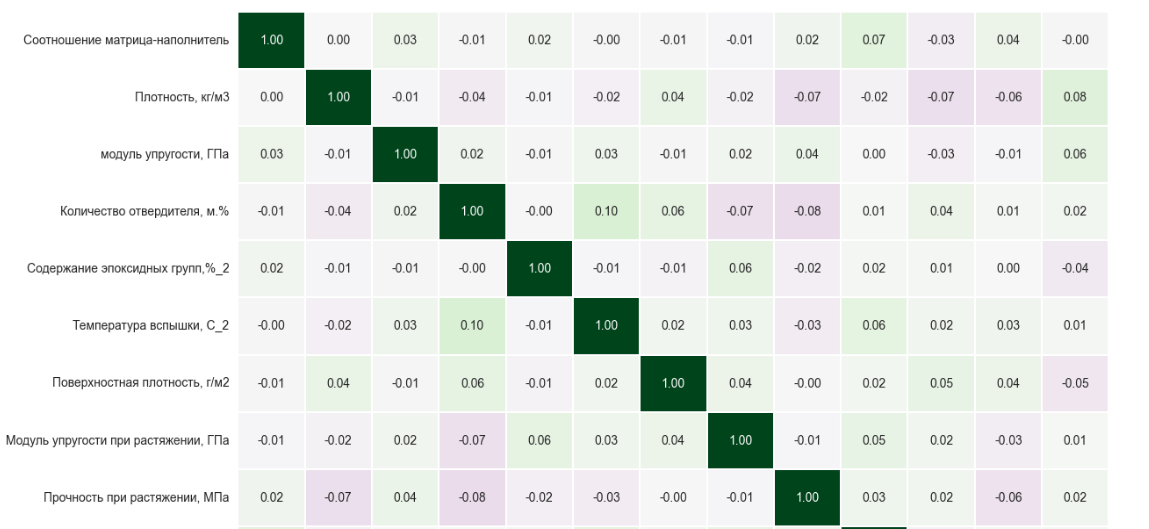


Рисунок—12 «Ящики с усами» характера распределения признаков

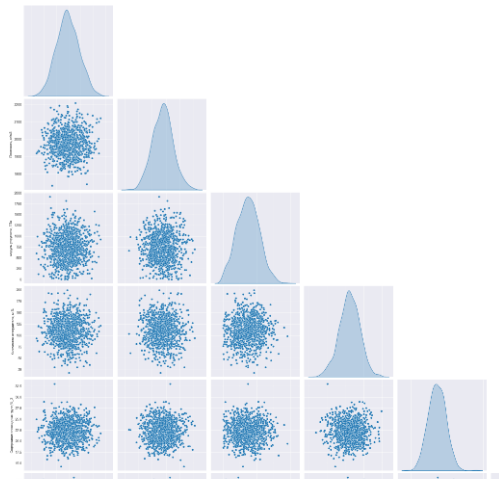
Точки за границами "усов", не что иное, как выбросы.

Все признаки имеют нормальное распределение, кроме «Угол нашивки», также немного отличается «Поверхностная плотность», медиана отличается от среднего значения.

Тепловая карта Пирсона на рисунке 13—это тип визуализации, применяемый, когда нам нужно найти зависимые переменные. Совместное распределение и диаграммы рассеяния признаков показаны на рисунке 14. Наши целевые переменные (соотношение матрица-наполнитель, модуль упругости при растяжении и прочность при растяжении) в основном зависят от потребления смолы, колличества отвердителя и колличества отвердителя соответственно. Но зависимость между признаками так мала, что можно говорить о ее практическом отсутствии.



Рисунок—13 Корелляционная тепловая карта Пирсона



Рисунок—14 Совместное распределение и диаграммы рассеяния признаков

Графики попарной корреляции всех переменных на рисунке 15.

**2.Практическая часть**

**2.1 Предобработка данных**

Графики распределения для каждого признака до и после нормализации, указать максимальные и минимальные значения.

Удаление шумов и т.д..

Оценка выбросов произведена методом 3-х сигм (если значения похожие на выбросы не превышают границы трех сигм нормального распределения, то их можно не считать выбросами) (Рис 1) и методом межквартильных расстояний (значения выше верхней границы и ниже нижней границы являются выбросами) (Рис 1).

Принято решение удалить выбросы, хотя надо бы выполнить дальнейшую работу в двух вариантах: с удалением и без удаления, после сравнить результаты.

**Заключение**

У композитов, на данном этапе развития отрасли, есть ряд существенных недостатков:

1. анизотропия – одни и те же свойства композитного материала могут в десятки раз различаться в зависимости от направления внешнего воздействия (вдоль волокон или поперек);
2. большой удельный объем;
3. гигроскопичность;
4. токсичность (при изготовлении и в процессе эксплуатации эти материалы могут выделять вредные для здоровья человека пары);
5. высокая цена (при производстве часто используется дорогостоящее оборудование) и т.п.

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

**Список литературы**

1 Википедия, Композитные материалы: – Режим доступа: <https://www.ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D0%BE%D0%B7%D0%B8%D1%82%D0%BD%D1%8B%D0%B9_%D0%BC%D0%B0%D1%82%D0%B5%D1%80%D0%B8%D0%B0%D0%BB>. (дата обращения: 26.03.2023).

2 Библиотека Tensorflow [Электронный ресурс]: – Режим доступа:

<https://www.tensorflow.org/api_docs/python/tf/keras/metrics>. (дата обращения: 28.03.2023).

3 Библиотека keras [Электронный ресурс]: – Режим доступа: <https://keras.io/api/metrics/>. (дата обращения: 28.03.2023).

4 Библиотека matplotlib [Электронный ресурс]: – Режим доступа: <https://matplotlib.org/stable/plot_types/index.html>. (дата обращения: 28.03.2023).

5 Библиотека numpy [Электронный ресурс]: – Режим доступа: <https://numpy.org/doc/1.22/user/c-info.html>. (дата обращения: 28.03.2023).

6 Библиотека pandas [Электронный ресурс]: – Режим доступа: <https://pandas.pydata.org/docs/user_guide/io.html>. (дата обращения: 28.03.2023).

7 Библиотека scikit-learn [Электронный ресурс]: – Режим доступа: <https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics>.

(дата обращения: 28.03.2023).

8 Библиотека seaborn [Электронный ресурс]: – Режим доступа: <https://seaborn.pydata.org/tutorial/relational.html>. (дата обращения: 28.03.2023).

9 Документация по языку программирования python [Электронный ресурс]: – Режим доступа: <https://docs.python.org/3.10/>. (дата обращения: 28.03.2023)

10 [Инжинириум МГТУ им. Н.Э. Баумана, Интересные факты о композитах: – Режим доступа:](https://dzen.ru/id/5cd2a9b5a1c44300b3cec8e3" \t "_blank) [[https://dzen.ru/a/XsPAVKI-FWARPzyn](https://dzen.ru/id/5cd2a9b5a1c44300b3cec8e3" \t "_blank)](https://dzen.ru/a/XsPAVKI-FWARPzyn)[. (дата обращения: 26.03.2023).](https://dzen.ru/id/5cd2a9b5a1c44300b3cec8e3" \t "_blank)

11 Кодкамп, Полное руководство: когда удалять выбросы в данных: – Режим доступа: https://www.codecamp.ru/blog/remove-outliers/. (дата обращения: 26.03.2023).

12 Независимая газета, Будущее сделано из композитов: – Режим доступа:

<https://www.ng.ru/nauka/2022-06-07/13_8455_future.html>. (дата обращения: 26.03.2023).

13 Реутов Ю.А.: Прогнозирование свойств полимерных композиционнных материалов и оценка надёжности изделий из них, Диссертация, Томск 2016: – Режим доступа: <http://ams.tsu.ru/TSU/QualificationDep/co-searchers.nsf/ECF749E40C9E58024725804400349189/$file/%D0%A0%D0%B5%D1%83%D1%82%D0%BE%D0%B2_%D0%AE.%D0%90._%D0%94%D0%B8%D1%81%D1%81%D0%B5%D1%80%D1%82%D0%B0%D1%86%D0%B8%D1%8F.pdf>. (дата обращения: 28.03.2023)

14 Среда разработки Jupyter Notebook [Электронный ресурс]: – Режим доступа: <https://docs.jupyter.org/en/latest/projects/content-projects.html>. (дата обращения: 28.03.2023)

15 Хабр, История композиционных материалов: – Режим доступа: <https://habr.com/ru/post/362189/>. (дата обращения: 26.03.2023).

16 Чун-Те Чен и Грейс Х. Гу. Машинное обучение для композитных материалов (март 2019г.): – Режим доступа: <https://www.cambridge.org/core/journals/mrs-communications/article/machine-%20learning-for-composite-materials/F54F60AC0048291BA47E0B671733ED15>.

(дата обращения: 28.03.2023)

17 [LibTime](https://libtime.ru/), Как создают композиты: – Режим доступа: <https://libtime.ru/science/kak-sozdayut-kompozity.html>. (дата обращения: 26.03.2023).

18 Longinom: – Режим доступа: https://wiki.loginom.ru/visualization.html. (дата обращения: 28.03.2023).