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……………………………………………………………………….3

1 Аналитическая часть………………………………………………………..6

* 1. Постановка задачи………………………………………………….6
  2. Характеристика датасета…………………………………………..6

1.3 Описание используемых методов………………………………...10

1.4 Разведочный анализ данных……………………………………...14

1.5 Выводы к разделу………………………………………………….20

2.Практическая часть………………………………………………………...22

2.1 Предобработка данных……………………………………………22

2.2 Разработка и обучение модели……………………………………26

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

2.7 Выводы к разделу………………………………………………

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.

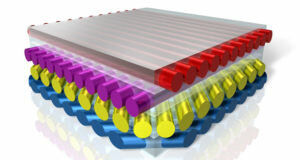


Рисунок 1—Структура композиционного материала

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». При построении моделей необходимо 30% данных оставить на тестирование моделей, на остальных провести обучение моделей. При построении моделей провести поиск гиперпараметров моделей с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10;
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:

1. проверку типа данных;
2. удаление нерелевантных столбцов (не используем, не имеют для анализа никакого смысла);
3. переименование столбцов (улучшает читаемость набора данных);
4. удаление повторяющихся, дубликатов строк (сокращает размер датасета);
5. удаление отсутствующих или нулевых значений либо замена их средним/медианой или модой для этого столбца;
6. обнаружение выбросов и идентификация их таковыми, удаление (чтобы не искажали результаты).

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 по таблице 1.

Практически все значения признаков по столбцам уникальны, как видно по таблице 2.

Проанализировав наиболее часто встречающиеся значения по всем 13 столбцам, отражено в таблице 3, выясняем, что первые 23 строки в объединенном датасете полностью сгенерированы усредненными значениями, что видно по таблице 4. Причина этого пока не ясна, т.к. при размерности в 1000 строк полная генерация усредненных дополнительных 23 строк ни на что не влияет.

Таблица 1—Описание объединенного датасета

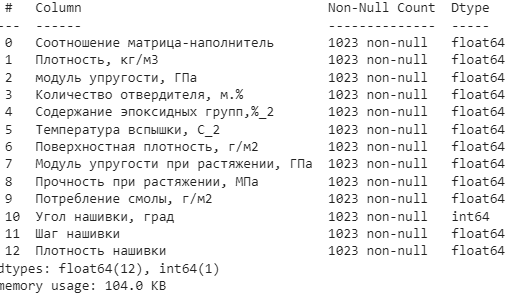
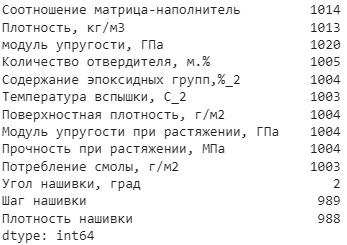
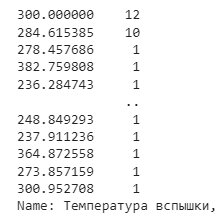
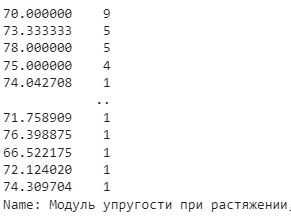
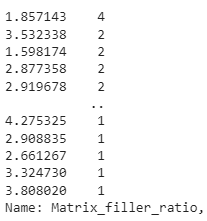
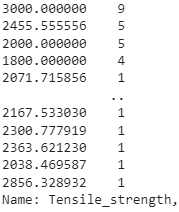
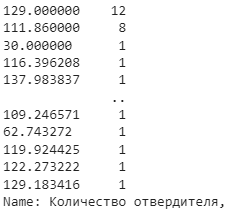


Таблица 2—Наличие уникальных значений

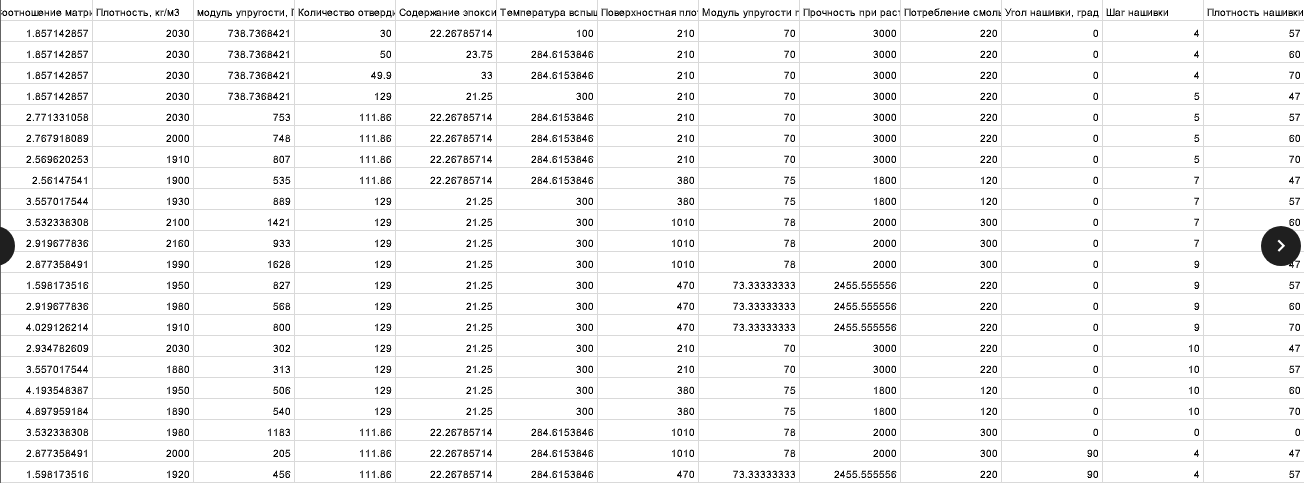


Таблица—3 Наиболее часто встречающиеся значения

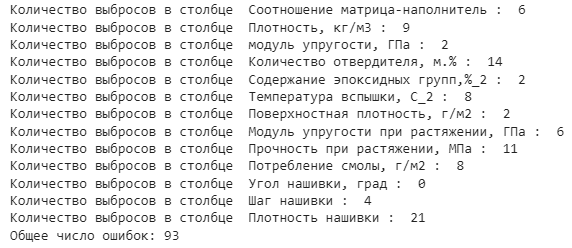
  

Таблица—4 Сгенерированные строки в датасете



Значений похожих на выбросы согласно таблице 5 в датасете небольшое колличество.

Таблица—5 Количество предполагаемых выбросов в датасете

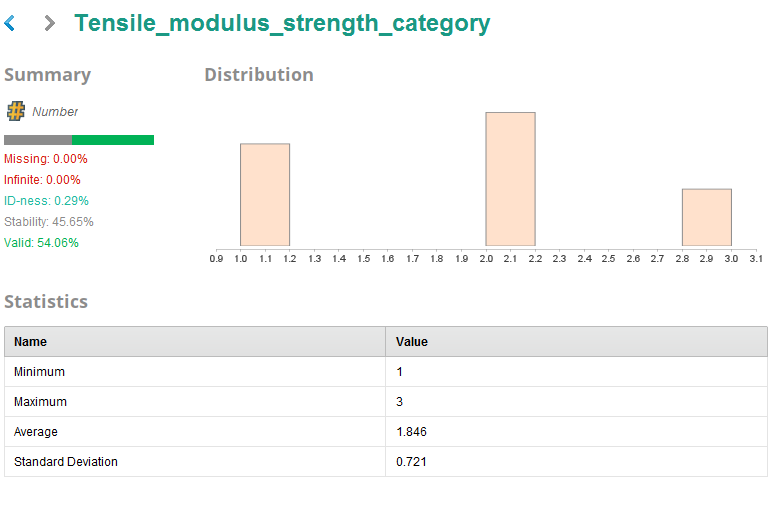
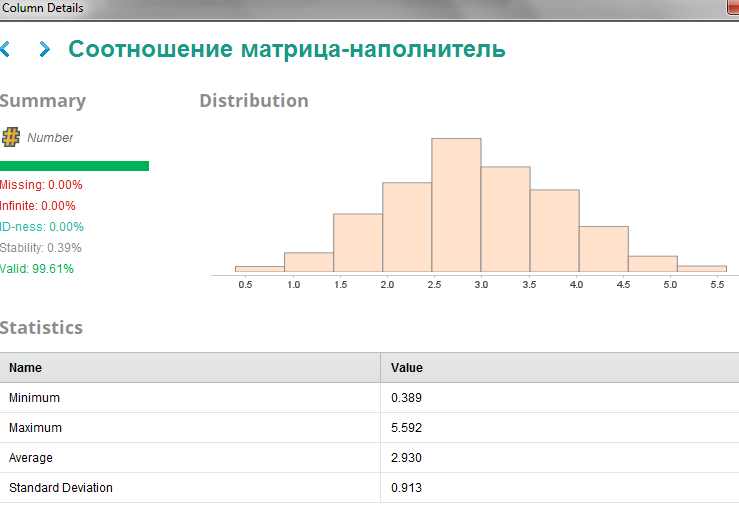


Чтобы идентифицировать их таковыми и принять решение удалить, нужно выбросы оценить, иначе информация может быть упущена (выполним это в подразделе 1.4). В сочетании с фактическими данными выбросы можно разделить на "истинные аномалии" и "ложные аномалии". Экспертная оценка выбросов нам не доступна (не являемся экспертами), т.к. опирается на вопросы:

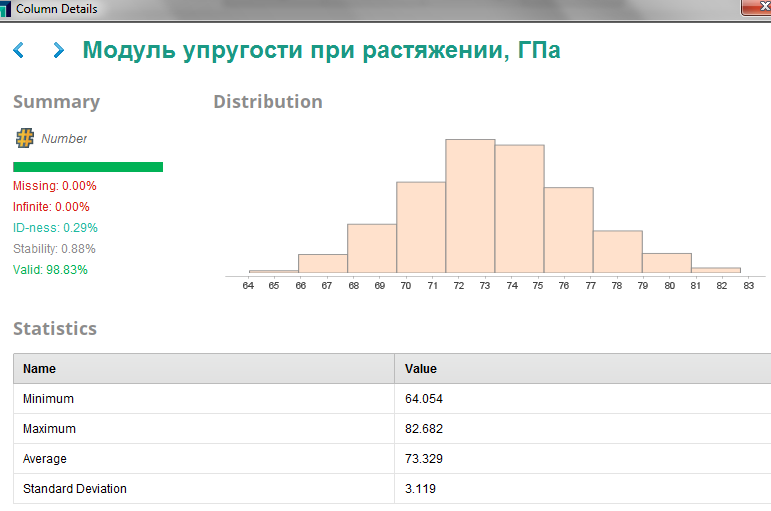
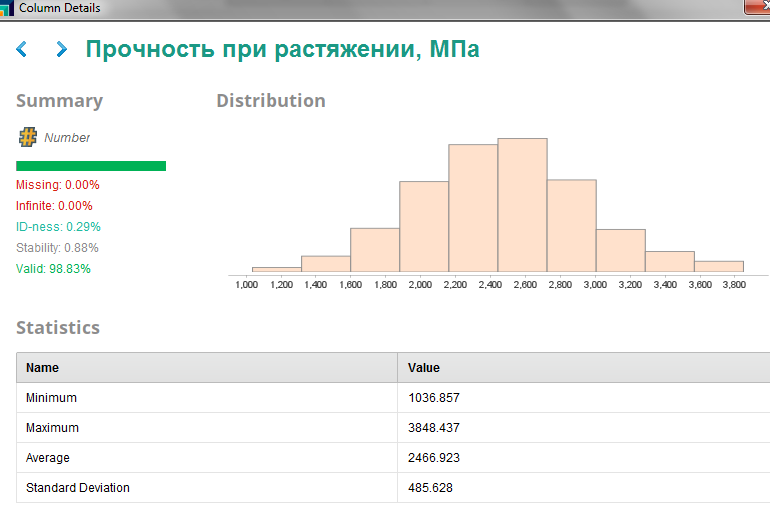
1. является ли выброс результатом ошибки ввода данных?
2. влияет ли выброс на результаты анализа?
3. влияет ли выброс на предположения?

RapidMiner—аналитическая платформа, по сути конструктор с набором готовых «кубиков-решений» для работы с данными без написания кода и без специфических знаний. Минусом аналитических платформ является ограниченное количество компонентов.

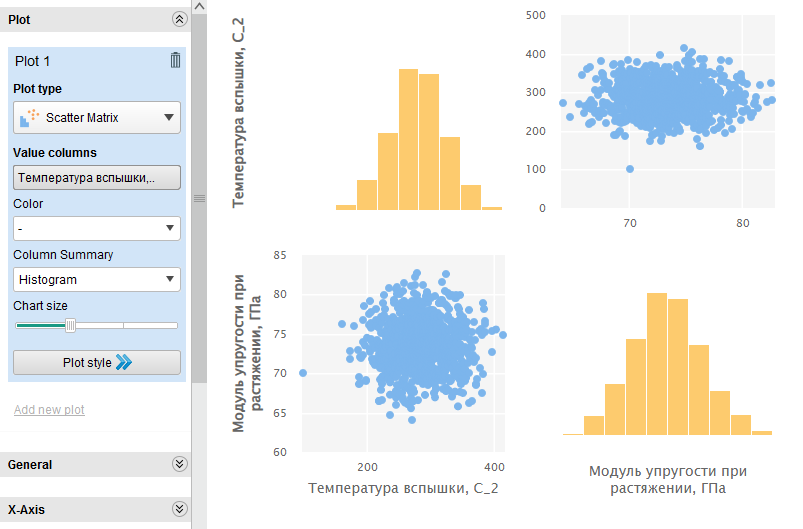
Посмотрим характеристику объединенного датасета с доп столбцом «Модуль упругости при растяжении, категории» в RapidMiner, рисунки 2-5:

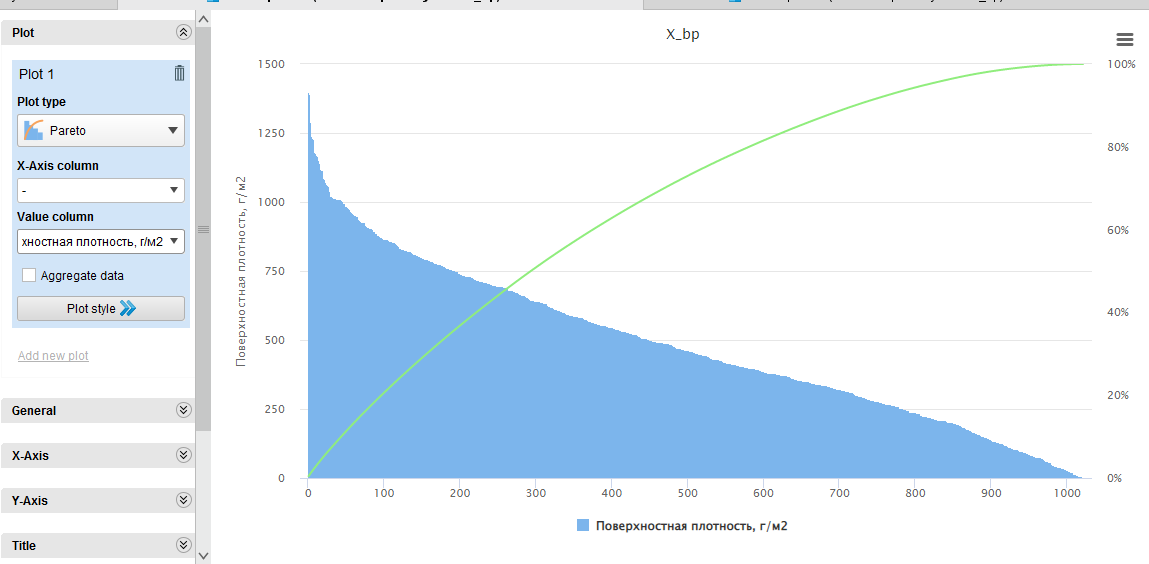
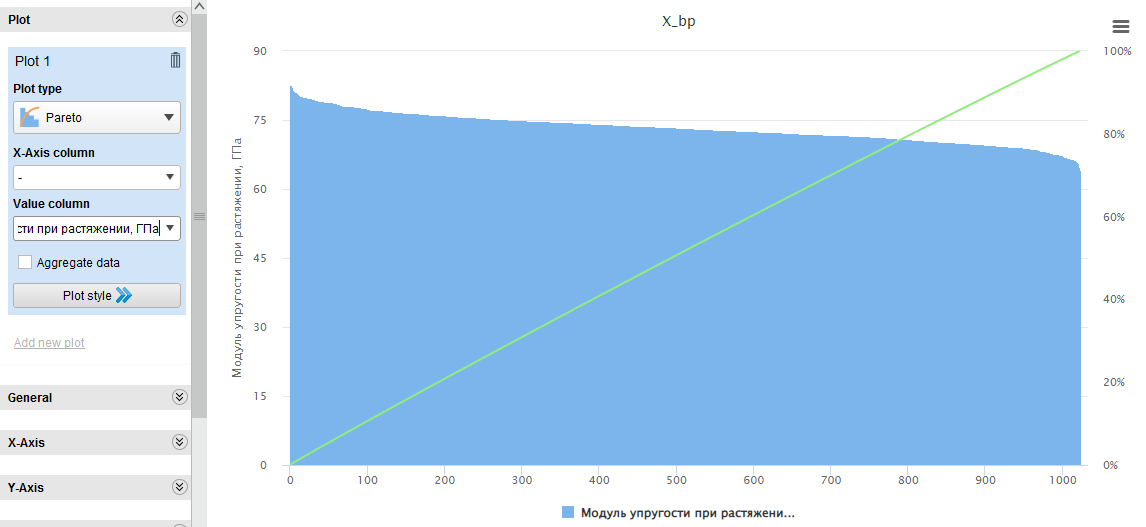
Рисунок—2 Статистический модуль в RapidMiner

Рисунок—3 Статистический модуль в RapidMiner



Рисунок—4 Визуализация в RapidMiner

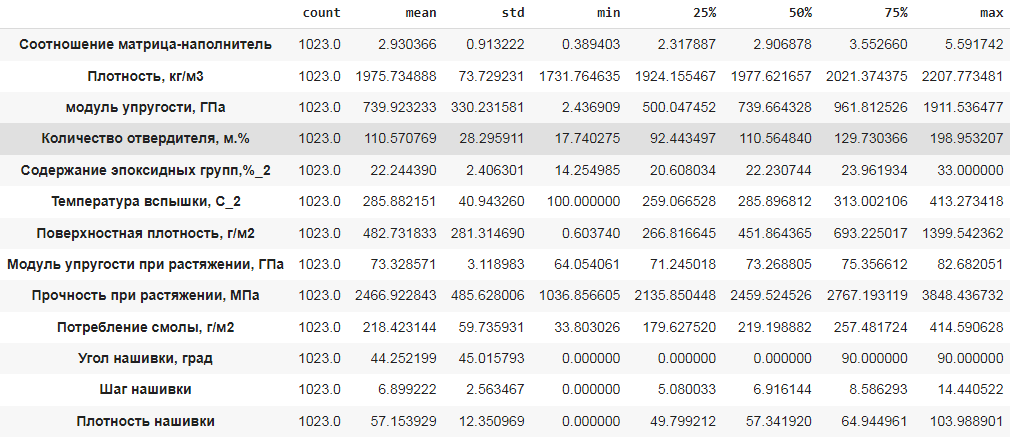


Рисунок—5 Визуализация в RapidMiner

Описательную статистику смотрим по таблице 6, выполнив транспонирование для удобства. Расшифруем отдельные характеристики:

1. Std — стандартное отклонение значения;
2. 25% — первый квартиль, 25% значений в столбце ниже этого значения;
3. 50% — медиана, половина значений в столбце ниже этого значения;
4. 75% — третий квартиль, 75% значений в столбце ниже этого значения;

Таблица—6 Описательная статистика



**1.3 Описание используемых методов**

В разработке:

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

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

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.

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

Для целевых признаков «Модуль упругости при растяжении», «Прочность при растяжении» используем:

1. linear regression;
2. linear regression с отобранными лучшими признаками;
3. lasso regression;
4. support vector machine;
5. random forest;
6. k neighbors regressor;
7. decision tree;
8. gradient boosted trees;

Для целевого признака «Соотношение матрица-наполнитель» используем нейронные сети.

**1.4 Разведочный анализ данных**

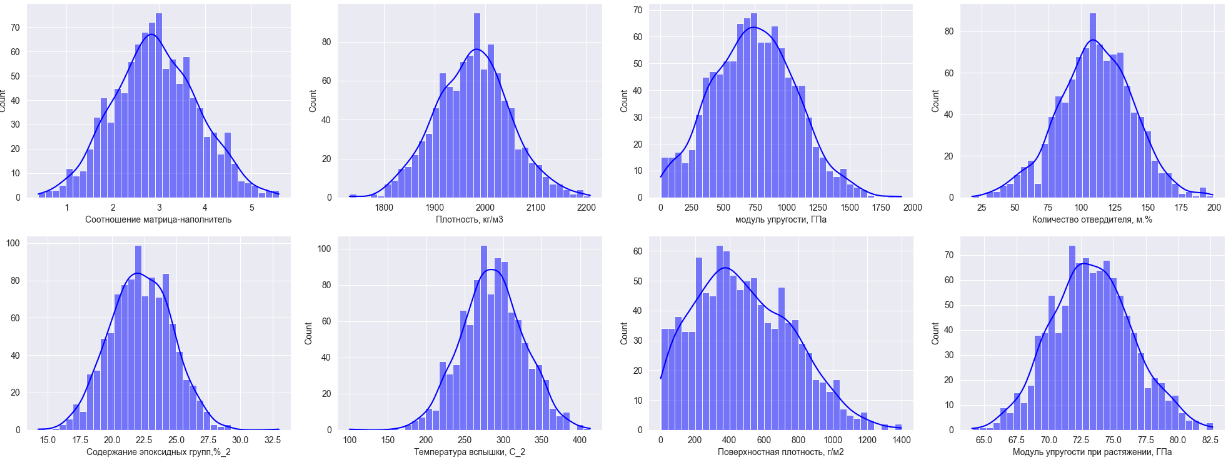
Разведочный (EDA) или, как его еще принято называть, исследовательский анализ применяется для выявления тенденций. При его использовании выявляют закономерности, обобщаются основные характеристики.

К основным методам разведочного анализа относится:

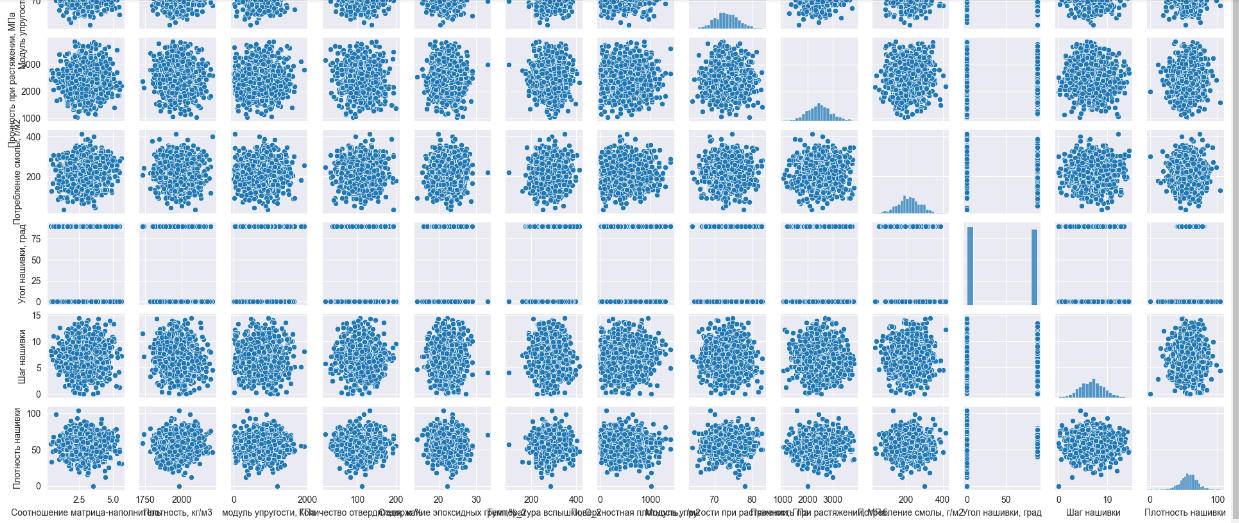
1. процедура анализа распределений переменных (чтобы выявить переменные с несимметричным или негауссовым распределением, в том числе и бимодальные). Гистограмма позволяет "на глаз" оценить нормальность эмпирического распределения. На гистограмму также накладывается кривая нормального распределения.
2. 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;
5. 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.

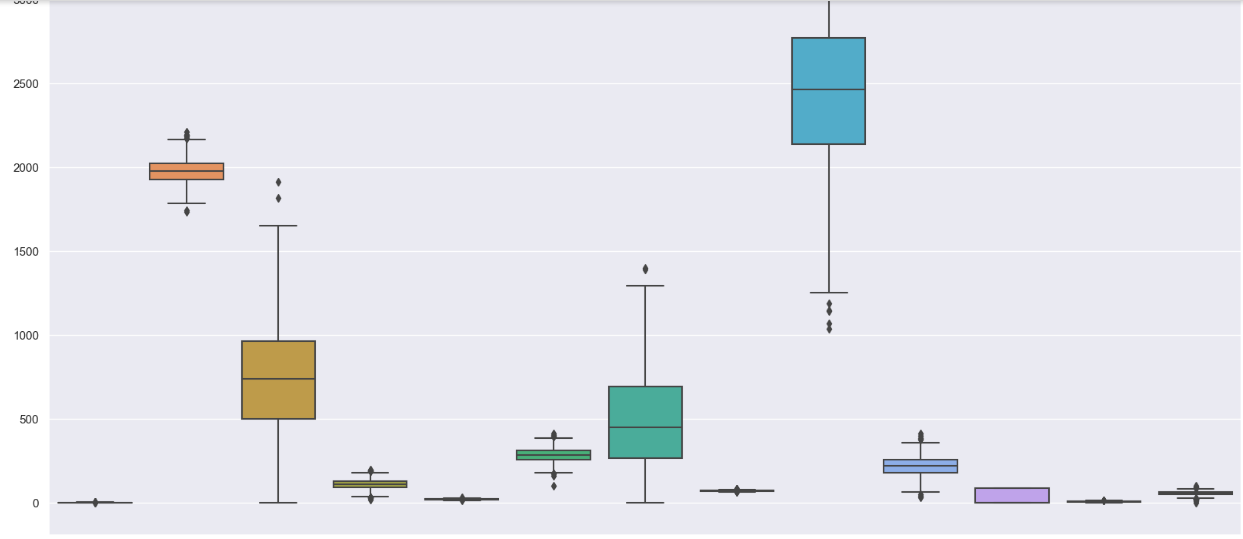
Гистограмма отображает частоту появления переменных в определенном интервале и характер распределения. Посмотрим на данные на рисунках 6 и 7. Все переменные имеют нормальное распределение. «Угол нашивки» имеет только два значения. «Прочность при растяжении», «Поверхностная плотность»: медианы отличаются от среднего значения, по остальным признакам очень близки. Все попарные графики рассеяния точек показывают отсутствие зависимости между признаками.



Рисунок—6 Гистограммы частот и характера распределения признаков



Рисунок—7 Попарные графики рассеяния точек (скаттерплоты)



Рисунок—8 Выбросы методом Boxplot «Ящика с усами»

Проведем визуальный поиск наличия аномалий и выбросов с помощью метода Boxplot (ящик с усами), рисунок 8. Точки за границами "усов" (полутора межквартильных расстояний от первого и третьего квантиля), не что иное, как выбросы.

Посмотрим также зависимость целевых признаков от признака «Соотношение матрица-наполнитель». Для этого разобьем значения каждого из целевых признаков на десять групп с шагом 10 процентов.

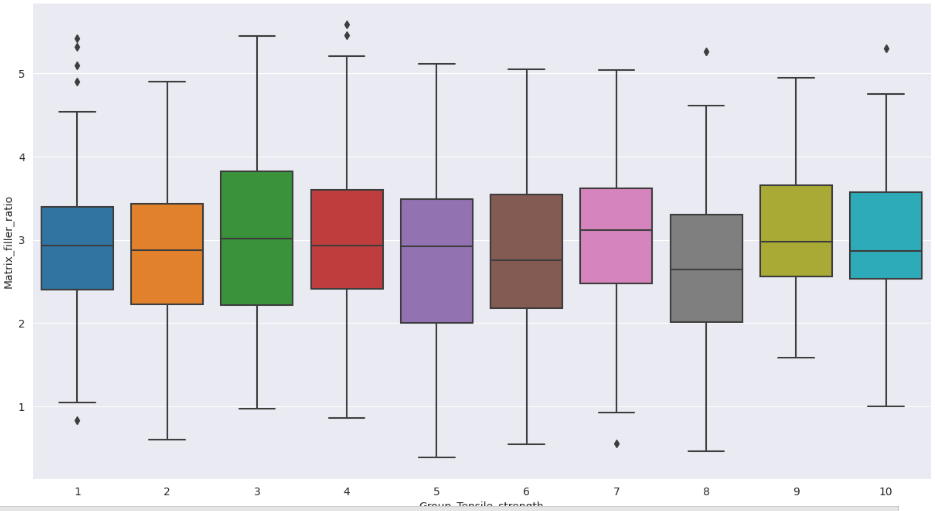
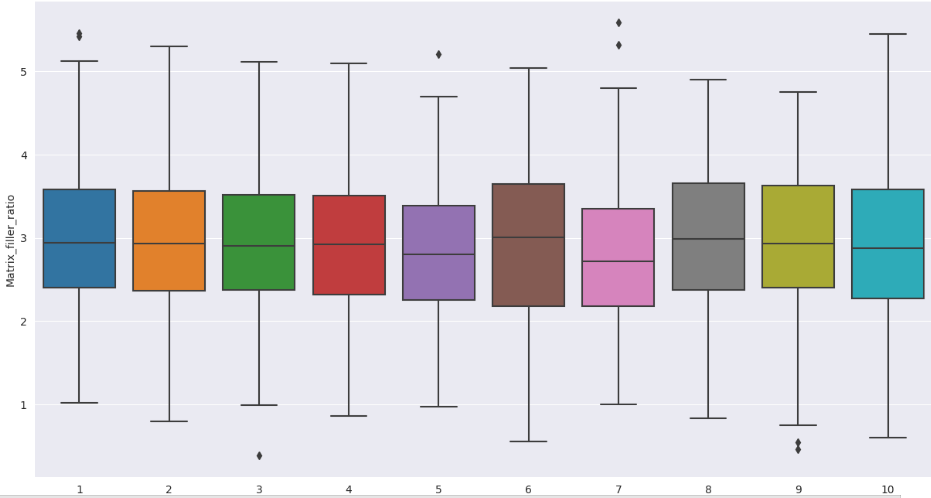
 

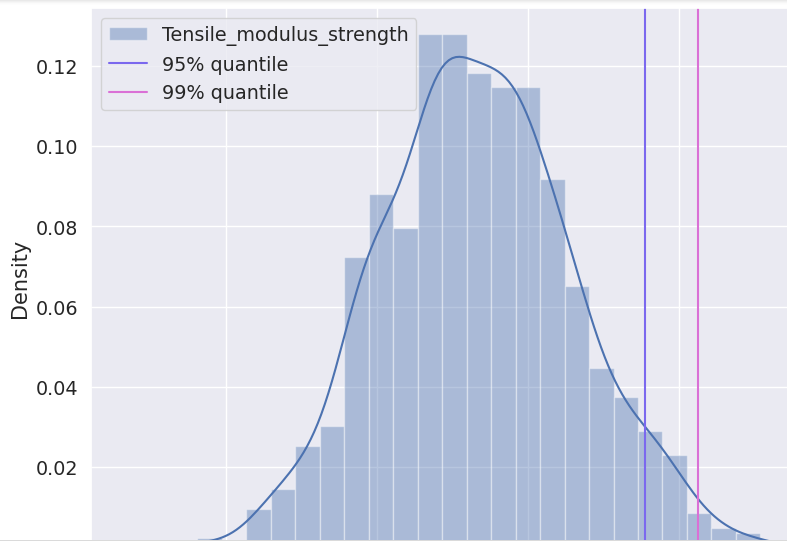
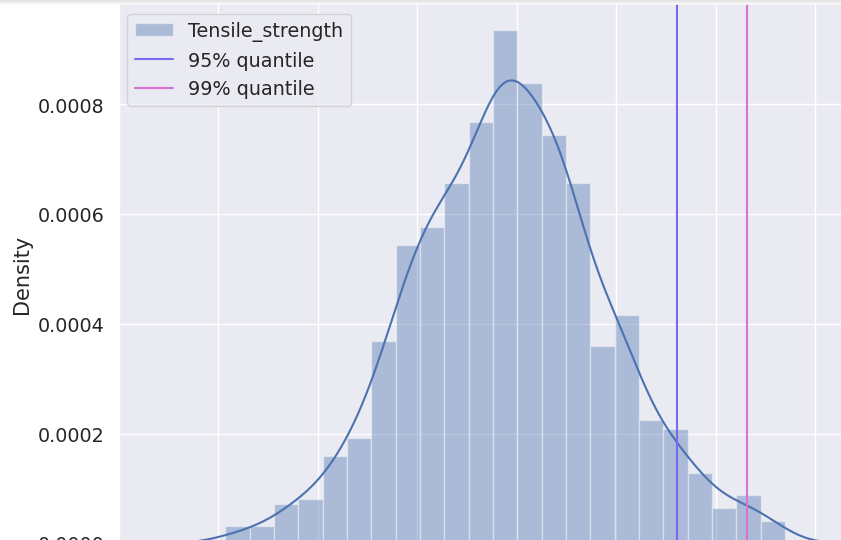
Рисунок 9 Group Tensile strength Рисунок 10 Group Tensile modulus strengt

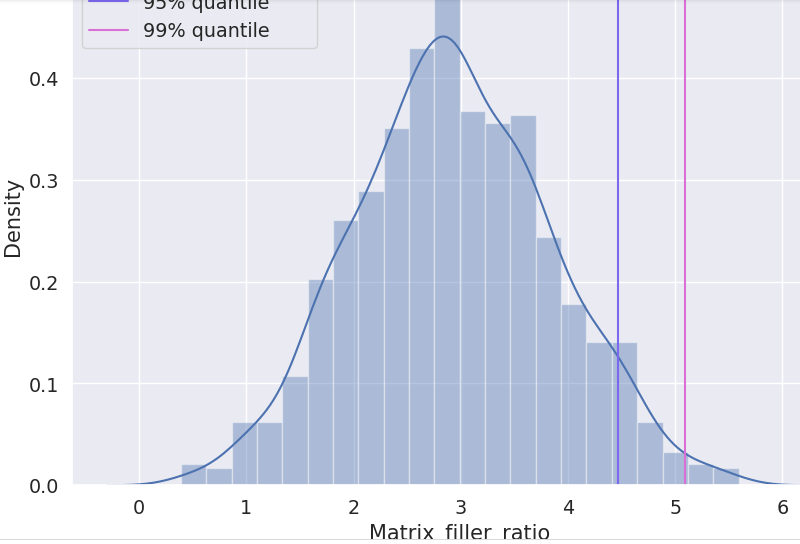
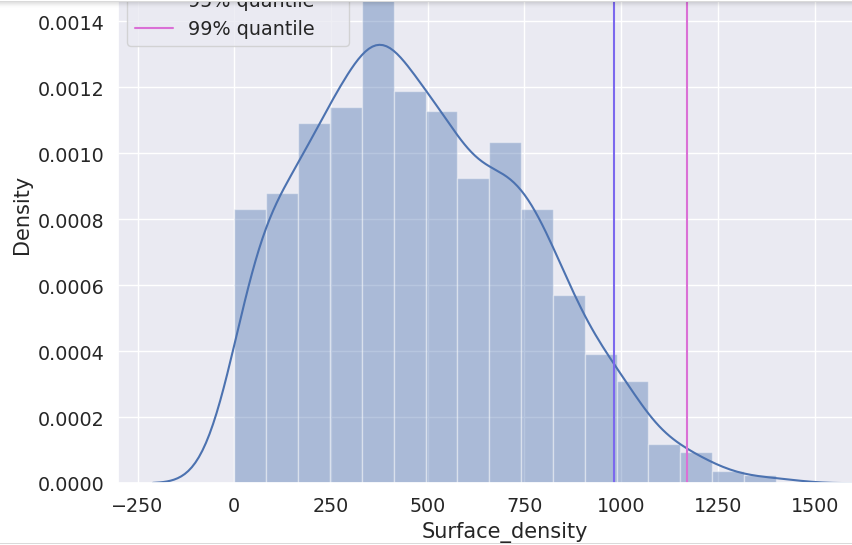
Во всех группах признака «Модуль упругости при растяжении» среднее значение медианы находится в диапазоне от 2,2 до 3,6. Выбросы расположены не во всех группах, а в 1, 3, 5, 7, 9 группах.

Во всех группах признака «Прочность при растяжении» среднее значение медианы находится в диапазоне от 2 до 2,9. Выбросы также расположены не во всех группах, а в 1, 4, 7, 8, 10 группах. Больше всего выбросов в 1 группе.

Переименуем столбцы на английские названия, т.к. не весь код запускается. Используем в таком виде в нужных нам местах.

Посмотрим на значения находящиеся в соответствии с рисунком 11 за 95 и 99 процентными квантилями.

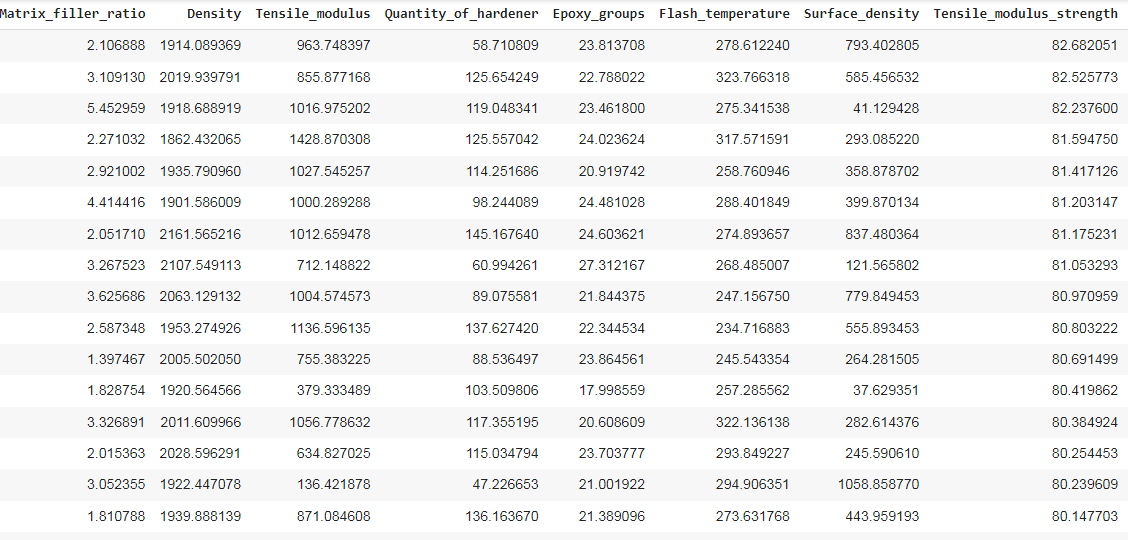
 

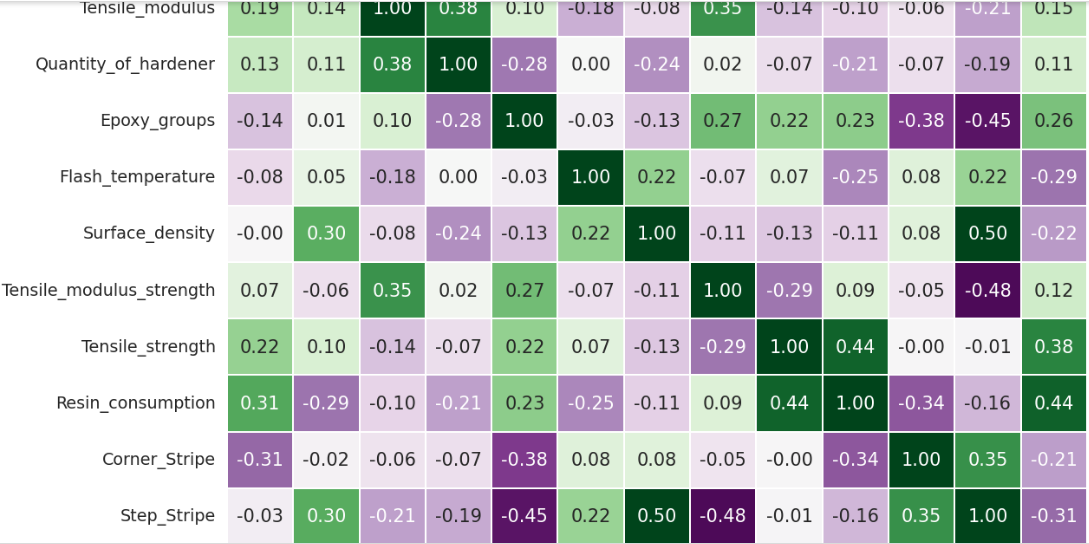
 

Рисунок—11 Значения находящиеся за 95 и 99 квантили.

Попробуем посмотреть самые большие 25 значений (50 значений, 150 значений) в столбце «Модуль упругости при растяжении» таблицы—7, также посмотрим какая у них корреляция на рисунке 12.

Таблица—7 Самые большие 25 значений «Модуль упругости при растяжении»



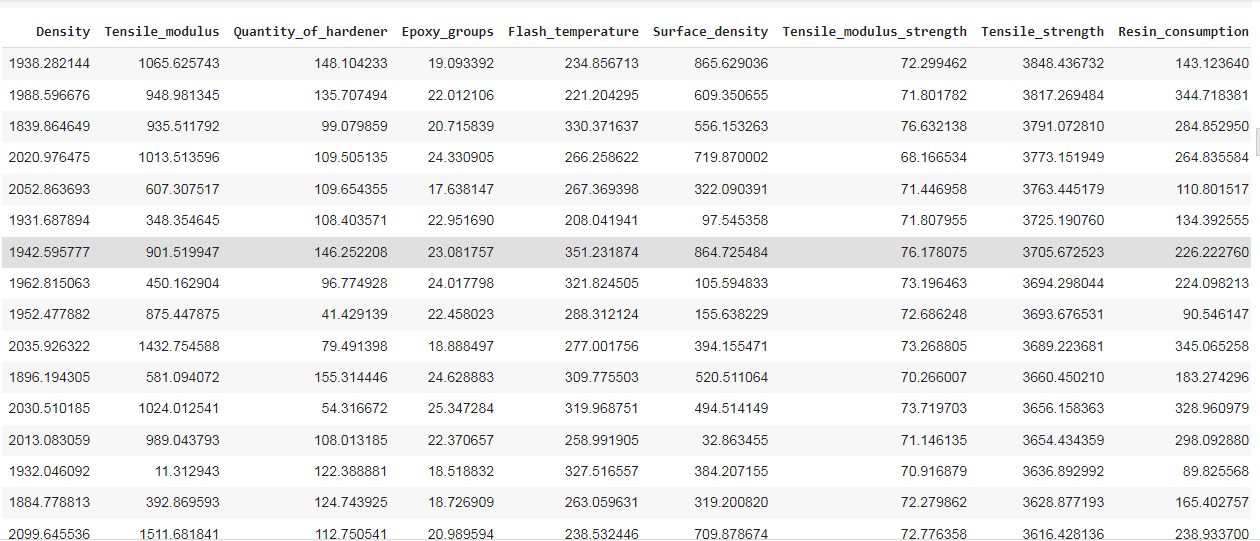


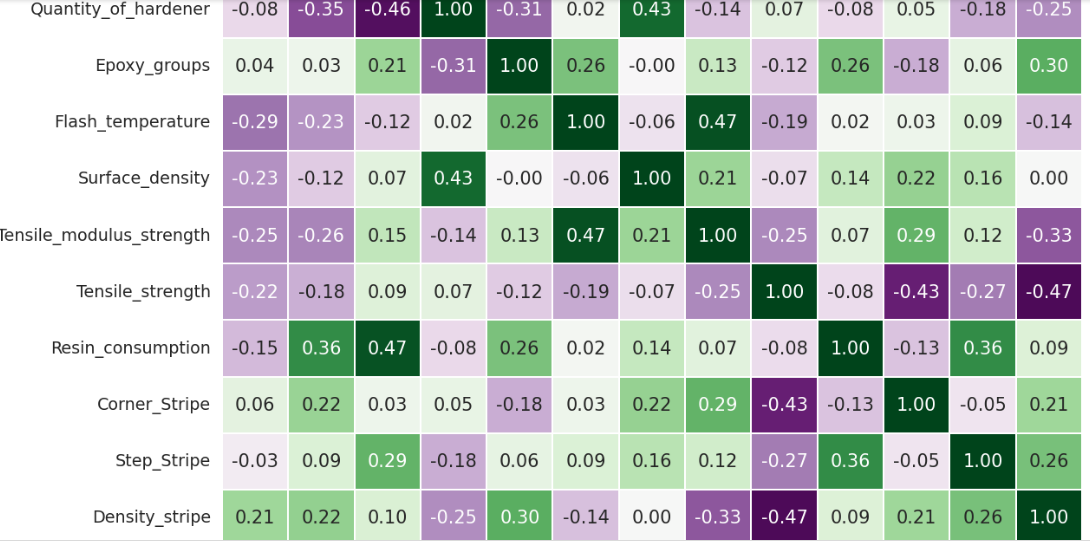
Рисунок—12 Корреляция 25 наибольших значений «Модуль упругости при

растяжении»

Посмотрим самые большие 25 значений (50 значений, 150 значений) в столбце «Прочность при растяжении» таблицы—8, также посмотрим какая у них корреляция на рисунке 13.

Таблица—8 Самые большие 25 значений «Прочность при растяжении»





Рисунок—13 Корреляция 25наибольших значений«Прочность при растяжении»

Изучив корреляции для 25, 50, 150 максимальных значений целевых признаков и сравнив их с корреляцией всего датасета видим, что она полностью отличается во всех группах (кроме «Шаг нашивки» для «Прочность при растяжении») и в сравнении с полным датасетом.

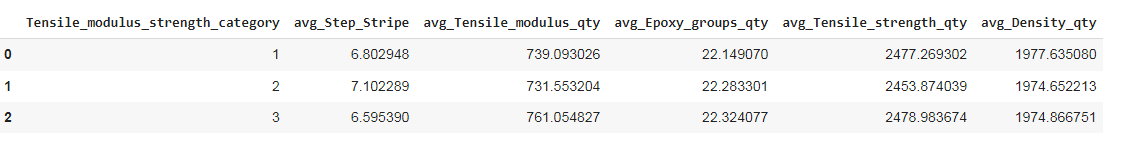
Создадим дополнительный столбец «Модуль упругости при растяжении, категории»: 3категория включает 150 самых больших значений данного признака, 2 и 1 категории создадим как равные половины оставшихся значений.

Создадим дополнительный столбец «Соотношение матрица-наполнитель, категории»:

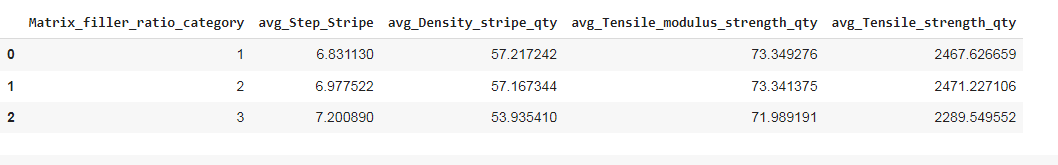
1. категория 3—значения от 4;
2. категория 2—значения от 2.5 до 4;
3. категория 1—значения от 0 до 2.5.

В этом столбце деление по принципу сбалансированного количества в категориях. Теперь для наших целевых переменных посмотрим агрегированные значения по некоторым признакам, таблицы 14-16:

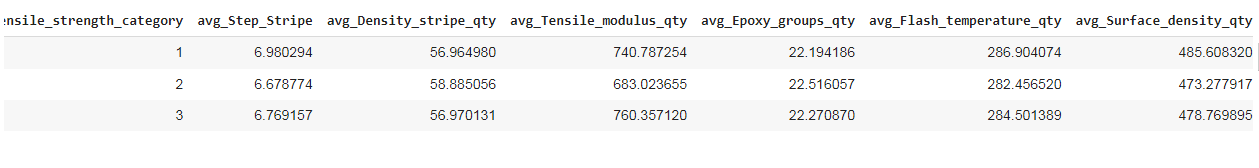
Таблица—14 Средние значен для «Модуль упругости при растяжен, категории»



Таблица—15 Средние значен для «Соотношение матрица-наполнит, категории»



Таблица—16 Средние значения для «Прочность при растяжении, категории»



**1.5 Выводы к разделу:**

1. датасет после объединения двух исходных таблиц включает 1023 строки и 13 столбцов, два из которых являются целевыми;
2. значения признаков числовые, пропусков в данных нет, нулевых значений нет, дубликатов строк нет, все значения имеют вещественный тип данных, присутствует один неинформативный столбец «Unnamed» (удалили его), дублирующий столбец с индексом;
3. названия признаков в датасете на русском языке, создали вариант датасета с признаками на английском языке (для корректной работы отдельных строк кода и платформы Avto ML);
4. признак «Угол нашивки» имеет два уникальных значения (возможен перевод признака в категориальный), остальные признаки имеют только уникальные значения (за исключением нескольких искусственно сгенерированных строк с 0 по 22, выявленных нами в ходе изучения данных); причины генерации строк не смогли выяснить в процессе работы;
5. все признаки (кроме «Угол нашивки») имеют незначительное колличество выбросов, хотя по определенным признакам однозначно идентифицировать их как выбросы нельзя; необходимо в дальнейшем работать с датасетом как с выбросами, так и без них, а после сравнить результаты;
6. все признаки имеют очень низкую корреляцию между собой и с целевыми признаками;
7. по характеристикам датасета и основываясь на рекомендуемых в научных публикациях моделях, хорошо показавших себя на композитах, определили перечень моделей для дальнейшей работы: статистический анализ, Linear Regression, Random Forest, Support Vector Machine, CNN, K Neighbors Regressor, Lasso Regressor, Gradient Boosted Trees, Decision Tree;
8. признаки имеют нормальное распределение, кроме «Поверхностная плотность» и «Модуль упругости» (имеют небольшую ассиметрию), признак «Угол нашивки» имеет бимодальное распределение, все распределения признаков неравномерные;
9. определили среднее значение, медиану, моду, стандартное отклонение, дисперсию;
10. определили квантили 95 и 99 процентные;
11. выполнили визуализацию описательного анализа датасета и EDA.

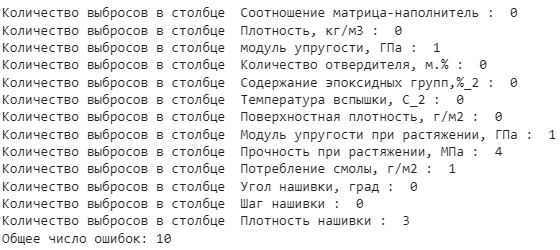
**2.Практическая часть**

**2.1 Предобработка данных**

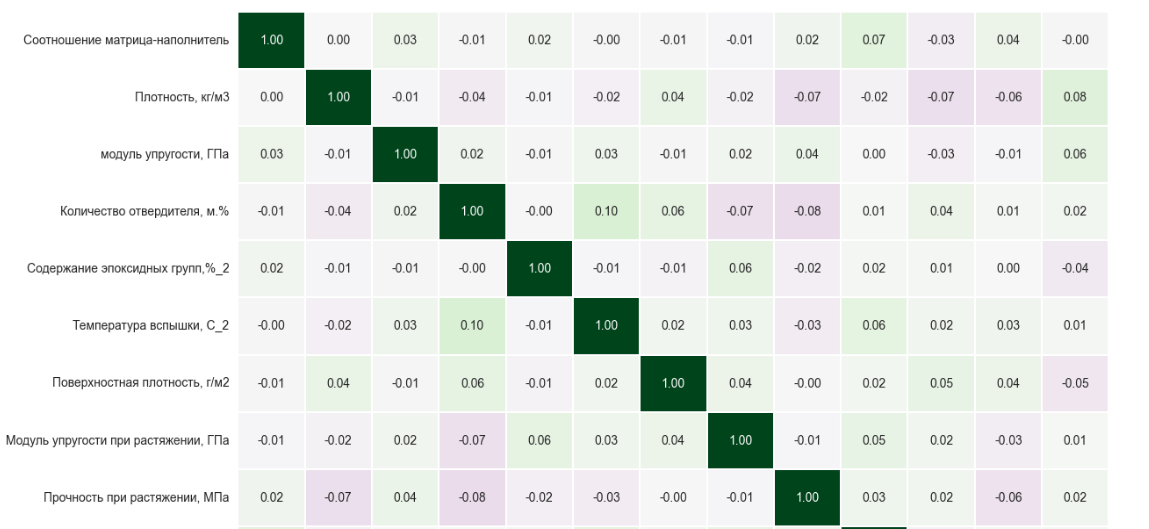
Выбросы не явные, поэтому дальнейшую работу выполним в двух вариантах: с удалением и без удаления, после сравним результаты.

Удалим выбросы, идентифицированные по методу межквартильных расстояний (значения выше верхней границы и ниже нижней границы являются выбросами), посмотрим что получилось по таблице 17.

Таблица—17 Датасет без выбросов, метод по межквартильных расстояний



Тепловая карта Пирсона —это тип визуализации, применяемый, когда нам нужно найти зависимые переменные. Зависимость между признаками так мала, что в соответствии с рисунком 14 можно говорить о ее практическом отсутствии. Максимальная положительная корреляция 0.11 между «Плотность нашивки» и «Угол нашивки», Максимальная отрицательная корреляция минус 0.08 между «Количество отвердителя» и «Прочность при растяжении».



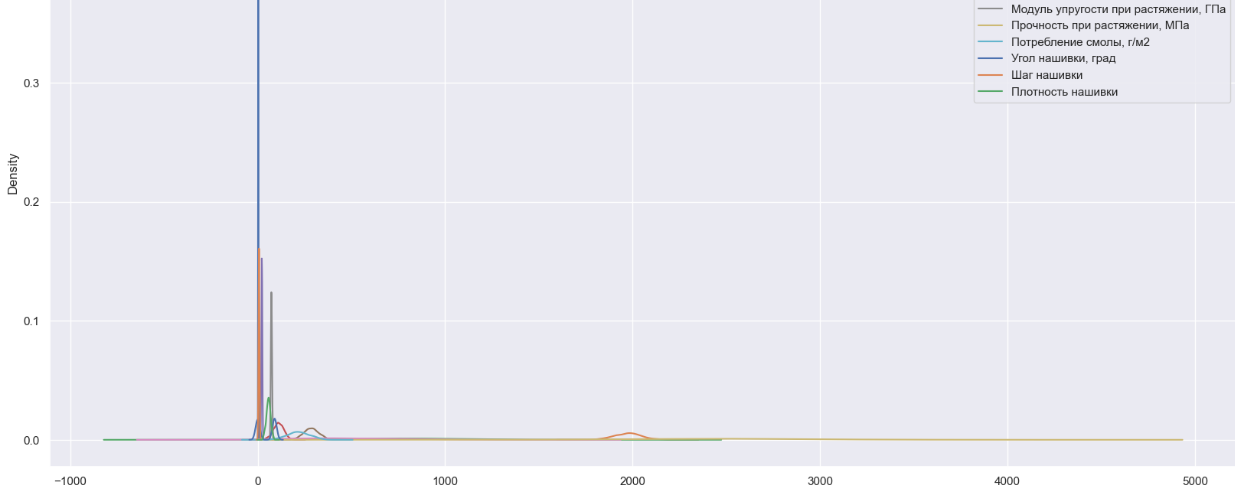
Рисунок—14 Корреляционная тепловая карта Пирсона

Рассмотрим топ зависимостей:

1. «Соотношение матрица-наполнитель» зависит от «Потребление смолы», «Шаг нашивки», «Модуль упругости» (плюс «Угол нашивки» с отрицательной корреляцией) в соответствии с рисунком 15;
2. «Модуль упругости при растяжении» зависит от «Количество отвердителя» (отрицательная корреляция), «Содержание эпоксидных групп», «Потребление смолы» в соответствии с рисунком 14;
3. «Прочность при растяжении» зависит от «Количество отвердителя» (отрицательная корреляция), «Плотность» (отрицательная корреляция), «Шаг нашивки» (отрицательная корреляция) в соответствии с рисунко 14.

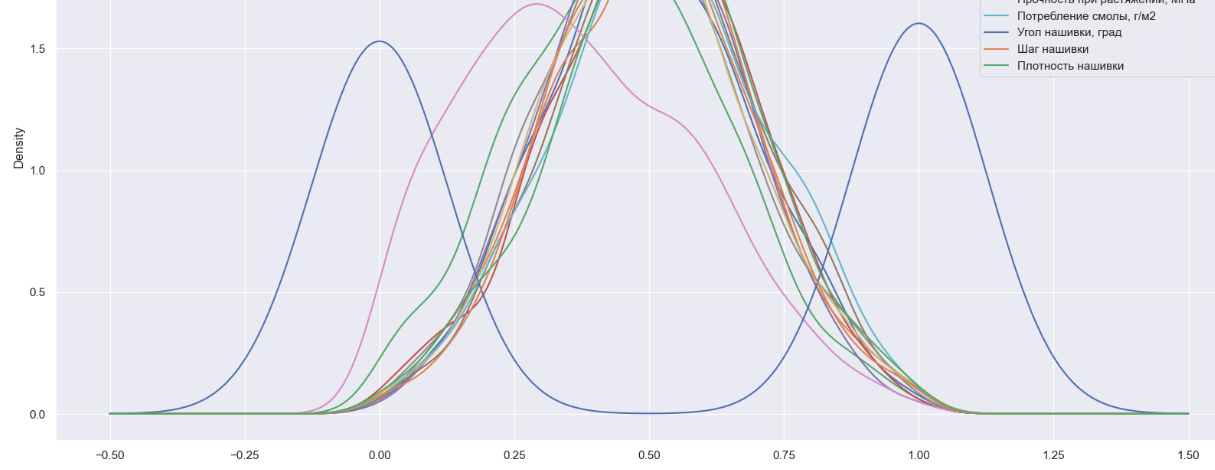


Рисунок—15 Топ 4 по корреляции для «Соотношение матрица-наполнитель»



Рисунок—16 Оценка плотности ядра до нормализации

Многие алгоритмы чувствительны к функциям, находящимся в разных масштабах (алгоритмы на основе метрик (KNN, K Means) и алгоритмы на основе градиентного спуска (регрессия, нейронные сети)). Древовидные алгоритмы (деревья решений, случайные леса) не имеют этой проблемы. Выполним нормализацию данных, используя метод MinMaxScaler. Значения приведены в диапазон от 0 до 1, что видно на рисунке 16. Т.к. выбросов очень немного, скорее всего они не повлияют на анализ данных даже если их оставить. Принято решение удалить выбросы, осталось 936 строк. Визуализируем диапазон распределение признаков после нормализации, показано на рисунке 17. Смотрим как изменились характеристики признаков после нормализации, таблица 18.

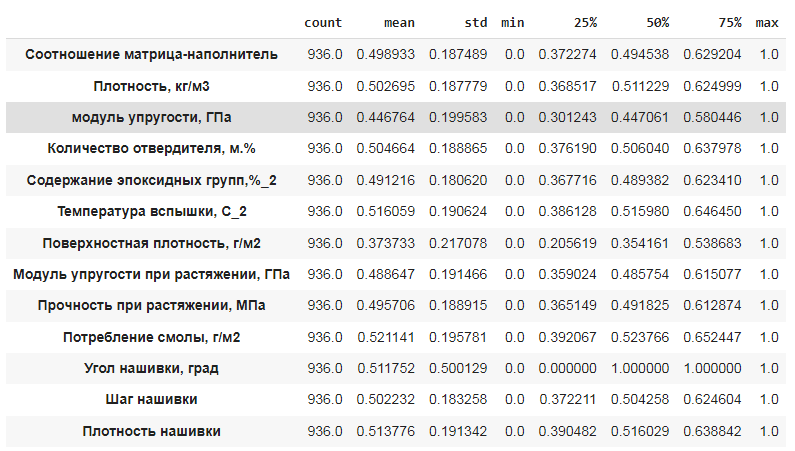


Рисунок—17 Оценка плотности ядра после нормализации

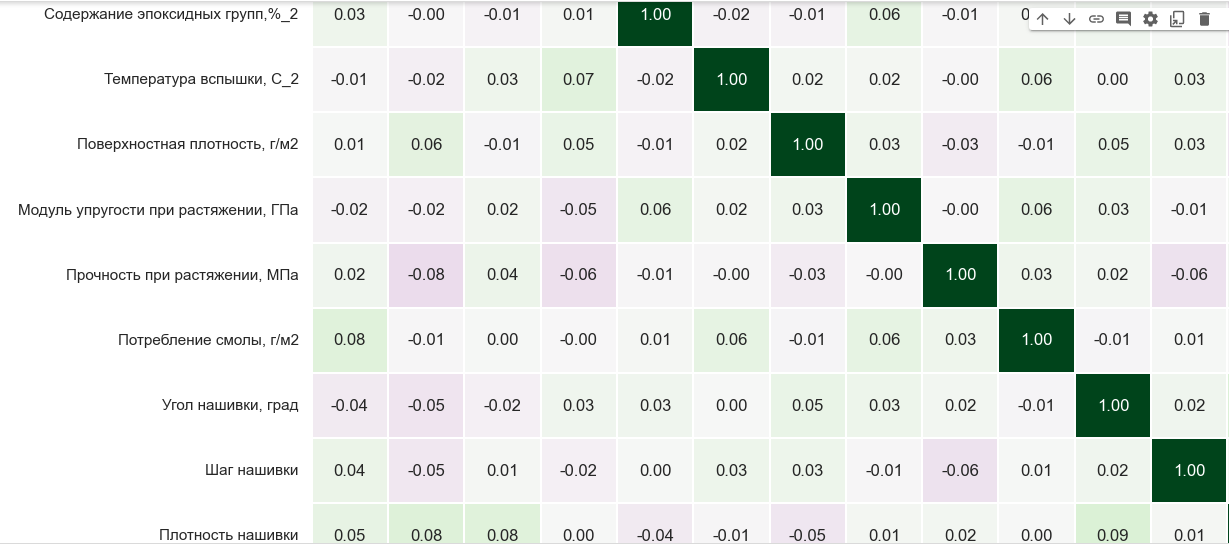
Медианное и среднее значение, в соответствии с таблицей 18, в пределах предыдущих значений после удаления выбросов.

Таблица—18 Статистическое описание признаков после нормализации

и удаления выбросов



Создадим переменную для названия всех столбцов. Это нам пригодится при построении моделей и перейдем к визуализации данных.



Рисунок—18 Корреляция после нормализации

Для «Модуль упругости при растяжении» и «Прочность при растяжении»

топ зависимостей не изменился, признаки слегка поменялись местами. Для «Соотношение матрица-наполнитель» топ зависимостей также не поменялся, но в топ добавился признак «Плотность нашивки».

**2.2 Разработка и обучение модели**

В разработке

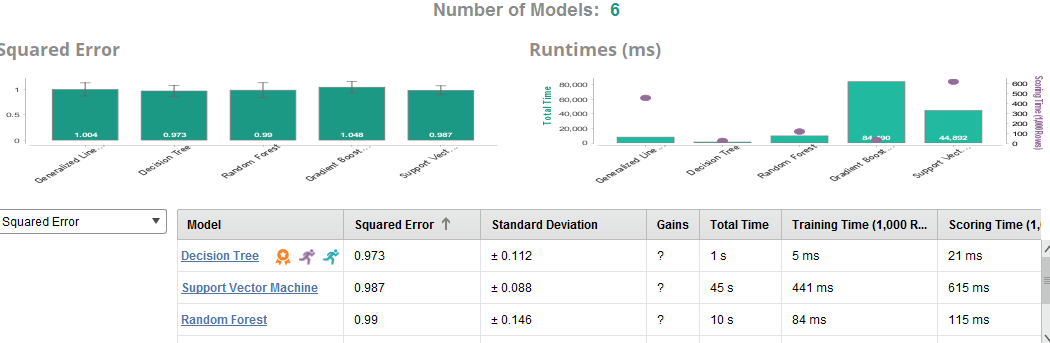
Для решения задачи регрессии используем следующие методы анализа:

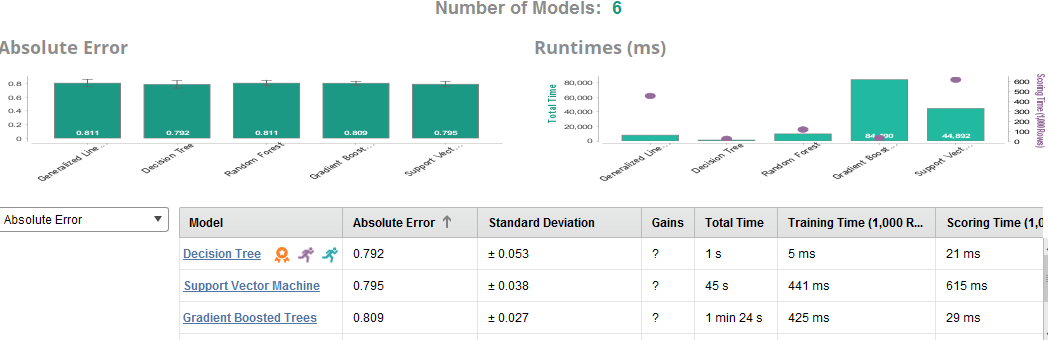
**2.3 Тестирование модели**

В разработке

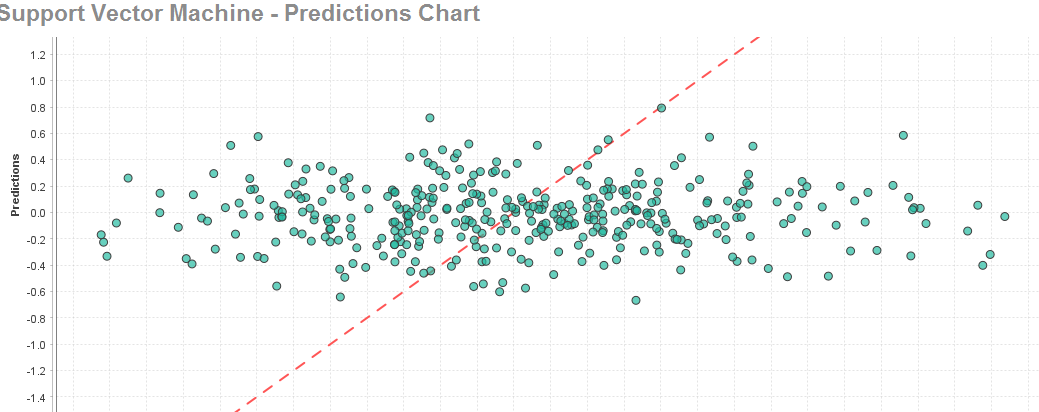
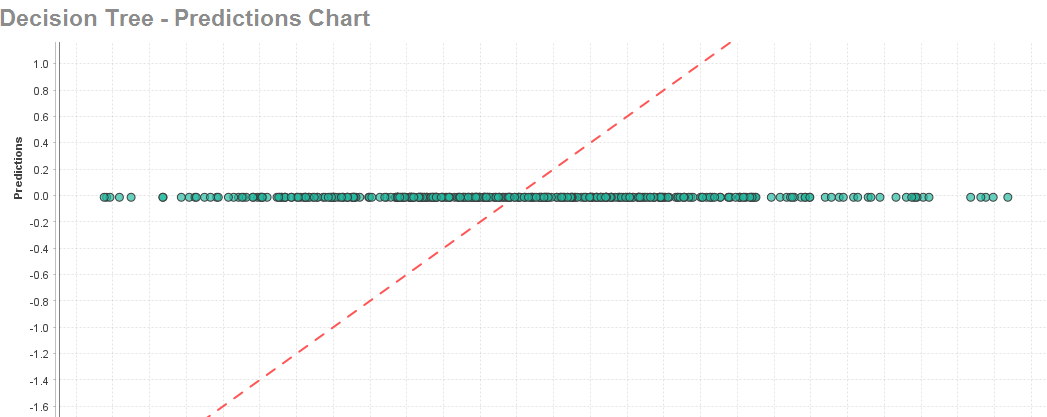
Для сравнения используем построение моделей в программной многопользовательской платформе RapidMiner. Для целевых переменных (для каждой по отдельности) построим модели (Generalised Linear Model, Decision Tree, Random Forest, Gradient Boosted Trees, Support Vector Machine) в RapidMiner. Изначально, на вход подали датасет без выбросов, стандартизированный: 1) соглашаемся с автоматическим выбором и ранжированием признаков для построения моделей; 2) соглашаемся с авто настройками параметров моделей.

Все оставляем по умолчанию, т.к. не разбираемся в датасете. Смотрим на результаты на рисунках 1-2:



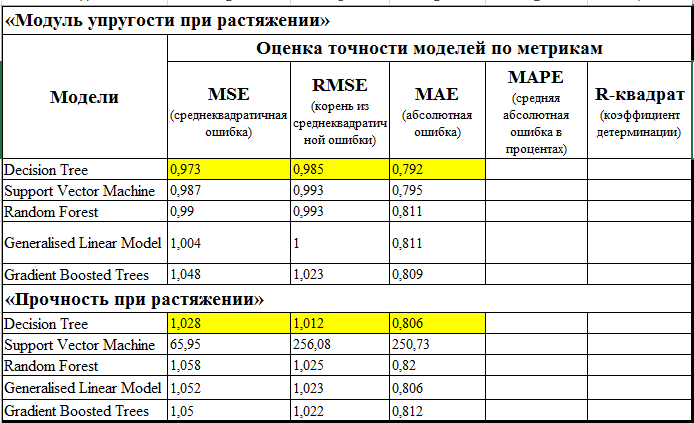


Рисунок—1 Обобщенная таблица результатов обучения моделей в RapidMiner



Рисунок—2 Предсказание модели в RapidMiner

Таблица—1 Оценка точности моделей по метрикам в RapidMiner



По итогу обучения все модели показали плохой результат, лучшими из них стали Decision Tree по обоим целевым переменным.

Выполнили также вариант с настройкой параметров моделей Random Forest, Gradient Boosted Trees: изменили колличество деревьев до 100 (по умолчанию было 20), а глубину сократили до 5 (по умолчанию было 20). Результат улучшился совсем незначительно, нужно еще пробовать разные настройки.

Для «Прочность при растяжении» метрики модели SVM показали результат, похожий на аномальный. Поэтому выполнили вариант с удалением части признаков (на вход обучения подали 9 признаков), в итоге метрики получились сопоставимые с другими.

Выполнили также вариант с подачей на вход датасета стандартизированного, но с выбросами. Метрики результата отличались от варианта с полностью обработанным датасетом на 10%.

Выполнили также вариант с подачей на вход датасета не стандартизированного, с выбросами. Метрики результата отличались незначительно.

Предсказания полученных моделей по обоим целевым переменным показали плохой результат2.

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

В разработке:

Описывается выбранная архитектура нейронной сети и ее результаты, в разработке

**2.5 Разработка приложения**

На основе имеющегося набора данных, используя фреймворк Flask, разработано приложение, прогнозирующее «Модуль упругости при растяжении».

В приложении вводятся значения всех признаков, соответствующие набору исходных данных, а полученный ответ соответствует прогнозу параметра «Модуль упругости при растяжении» при заданных признаках.

Инструкция по использованию приложения:

1. перейти по ссылке на сайт ссылка будет, в разработке;
2. ввести значения признаков;
3. нажать на кнопку «Прогноз».

**2.6 Создание удаленного репозитория и загрузка результатов работы**

На GitHub создан репозиторий.

Ссылка на репозиторий: <https://github.com/StoyanovP/VKR>

Заполнен README.

Файлы по ВКР загружены в репозиторий.

**2.7 Выводы к разделу:**

в разработке

1. при выделении из датасета топ 25, 50, 150, 250 самых высоких значений целевых признаков корреляция признаков менялась как по величине, так и по составу топ 4 наиболее зависимых признаков. Не менялся только знак корреляции (положительная или отрицательная);
2. свод результатов работы всех моделей представлен в обобщающей таблице;

**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.

Создание прогнозных моделей поможет сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

В ходе выполнения данной работы последовательно выполнен Pipeline построения моделей машинного обучения.

**Список литературы**

1 Википедия, Композитные материалы [Электронный ресурс]: – 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 [Электронный ресурс]: – Режим доступа: <https://dzen.ru/a/XsPAVKI-FWARPzyn>. (дата обращения: 26.03.2023).

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

12 Независимая газета, Будущее сделано из композитов [Электронный ресурс]: – 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 [Электронный ресурс]: <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 [Электронный ресурс]: – 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.) [Электронный ресурс]: – 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 [Электронный ресурс]: – Режим доступа: <https://libtime.ru/science/kak-sozdayut-kompozity.html>. (дата обращения: 26.03.2023).

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