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 Аденова Г.А. (ФИО)

Москва, 2022

Оглавление

[Введение 3](#_Toc106370165)

[1. Аналитическая часть 4](#_Toc106370166)

[1.1. Постановка задачи. 4](#_Toc106370167)

[1.2. Описание используемых методов. 5](#_Toc106370168)

[1.2.1 Линейная регрессия 5](#_Toc106370169)

[1.2.2. Случайный лес 6](#_Toc106370170)

[1.2.3. Метод опорных векторов 7](#_Toc106370171)

[1.2.4. Нейронная сеть 7](#_Toc106370172)

[1.2.5. Метрики оценки качества моделей 8](#_Toc106370173)

[2. Практическая часть 8](#_Toc106370174)

[2.1 Предобработка данных 8](#_Toc106370175)

[2.2 Разведочный анализ данных. 9](#_Toc106370176)

[2.3 Обнаружение и обработка значений выбросов 13](#_Toc106370177)

[2.4 Нормализация данных 15](#_Toc106370178)

[2.5 Регрессионный анализ 16](#_Toc106370179)

[2.5.1 Прогнозирование модуля упругости при растяжении 16](#_Toc106370180)

[2.5.2 Прогнозирование модуля прочности при растяжении 19](#_Toc106370181)

[2.5.3. Разработка нейронной сети для прогнозирования соотношения матрица-наполнитель 22](#_Toc106370182)

[Заключение 22](#_Toc106370183)

[Список использованных источников 23](#_Toc106370184)

# Введение

После того как современная физика металлов подробно разъяснила нам причины их пластичности, прочности и ее увеличения, началась интенсивная систематическая разработка новых материалов. Это приведет, вероятно, уже в вообразимом будущем к созданию материалов с прочностью, во много раз превышающей ее значения у обычных сегодня сплавов. При этом большое внимание будет уделяться уже известным механизмам закалки стали и старения алюминиевых сплавов, комбинациям этих известных механизмов с процессами формирования и многочисленными возможностями создания комбинированных материалов. Два перспективных пути открывают комбинированные материалы, усиленные либо волокнами, либо диспергированными твердыми частицами. У первых в неорганическую металлическую или органическую полимерную матрицу введены тончайшие высокопрочные волокна из стекла, углерода, бора, бериллия, стали или нитевидные монокристаллы. В результате такого комбинирования максимальная прочность сочетается с высоким модулем упругости и небольшой плотностью. Именно такими материалами будущего являются композиционные материалы.

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично.

Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

# 1. Аналитическая часть

## 1.1. Постановка задачи.

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

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

Датасет со свойствами композитов. Объединение делать по индексу тип объединения INNER

Требуется:

1. Изучить теоретические основы и методы решения поставленной задачи.
2. Провести разведочный анализ предложенных данных. Необходимо нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек. Необходимо также для каждой колонке получить среднее, медианное значение, провести анализ и исключение выбросов, проверить наличие пропусков.
3. Провести предобработку данных (удаление шумов, нормализация и т.д.).
4. Обучить нескольких моделей для прогноза модуля упругости при растяжении и прочности при растяжении. При построении модели необходимо 30% данных оставить на тестирование модели, на остальных происходит обучение моделей. При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.
5. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель.
6. Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в задании 4 или 5 (один или два прогноза, на выбор учащегося).
7. Оценить точность модели на тренировочном и тестовом датасете.
8. Создать репозиторий в GitHub / GitLab и разместить там код исследования. Оформить файл README.

## 1.2. Описание используемых методов.

В данной работе требуется произвести регрессионный анализ данных.

Регрессионный анализ ­ статистический аналитический метод, позволяющий вычислить предполагаемые отношения между зависимой переменной одной или несколькими независимыми переменными.

Создание регрессионной модели представляет собой итерационный процесс, направленный на поиск эффективных независимых переменных, чтобы объяснить зависимые переменные, которые мы пытаемся смоделировать или понять, запуская инструмент регрессии, чтобы определить, какие величины являются эффективными предсказателями. Затем пошаговое удаление и/или добавление переменных до тех пор, пока вы не найдете наилучшим образом подходящую регрессионную модель. Т.к. процесс создания модели часто исследовательский, он никогда не должен становиться простым "подгоном" данных. Процесс построения регрессионной модели должен учитывать теоретические аспекты, мнение экспертов в этой области и здравый смысл.

## 1.2.1 Линейная регрессия

Линейная регрессия (Linear regression) - это алгоритм машинного обучения, основанный на контролируемом обучении. Он выполняет задачу регрессии. Регрессионные модели представляют собой целевое значение прогноза, основанное на независимых переменных. В основном используется для выяснения взаимосвязи между переменными и прогнозирования.

Линейная регрессия сводится к нахождению уравнения вида:

*Y = а + bx ;*

Уравнение вида Y = а + bx позволяет по заданным значениям фактора x иметь теоретические значения результативного признака, подставляя в него фактические значения фактора X.

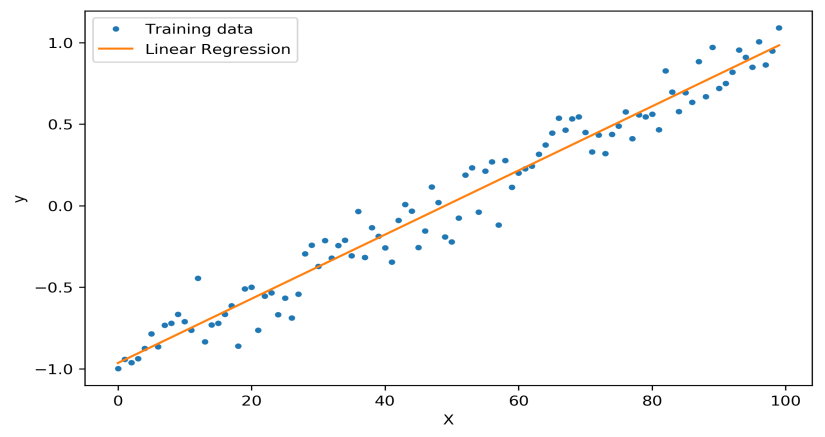


Рисунок.1 График функции линейной регрессии.

Модель линейной регрессии является часто используемой и наиболее изученной в [эконометрике](https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D0%BE%D0%BD%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D0%BA%D0%B0). А именно изучены свойства оценок параметров, получаемых различными методами при предположениях о вероятностных характеристиках факторов, и случайных ошибок модели. Предельные (асимптотические) свойства оценок нелинейных моделей также выводятся исходя из аппроксимации последних линейными моделями. С эконометрической точки зрения более важное значение имеет [линейность по параметрам](https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%BE%D1%81%D1%82%D1%8C_%D0%BF%D0%BE_%D0%BF%D0%B0%D1%80%D0%B0%D0%BC%D0%B5%D1%82%D1%80%D0%B0%D0%BC), чем линейность по факторам модели.

**Достоинства линейной регрессии:**

* скорость и простота получения модели;
* интерпретируемость модели;
* линейная модель является прозрачной и понятной для аналитика;
* по полученным коэффициентам регрессии можно судить о том, как тот или иной фактор влияет на результат, сделать на этой основе дополнительные полезные выводы;
* широкая применимость.

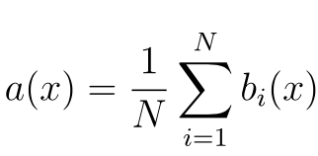
**Недостатки:**

* не может применяться к задачам нелинейной классификации;
* правильный выбор функций не требуется;
* ожидается хорошее соотношение сигнал / шум;
* колинеарность и выбросы влияют на точность модели LR.

## 1.2.2. Случайный лес

Алгоритм случайного леса (Random Forest) — универсальный алгоритм машинного обучения, суть которого состоит в использовании ансамбля решающих деревьев. Само по себе решающее дерево предоставляет крайне невысокое качество классификации, но из-за большого их количества результат значительно улучшается. Также это один из немногих алгоритмов, который можно использовать в абсолютном большинстве задач.

Формула итогового решателя — это усреднение предсказаний отдельных деревьев.

,

Где N – количество деревьев; i – счетчик для деревьев; b – решающее дерево; x – сгенерированная нами на основе данных выборка.

Для определения входных данных каждому дереву используется метод случайных подпространств. Базовые алгоритмы обучаются на различных подмножествах признаков, которые выделяются случайным образом.

## 1.2.3. Метод опорных векторов

Метод опорных векторов (англ. support vector machine, SVM) — один из наиболее популярных методов обучения, который применяется для решения задач классификации и регрессии. Основная идея метода заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора.

## 1.2.4. Нейронная сеть

Нейро́нная сеть (также искусственная нейронная сеть, ИНС) — [математическая модель](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8C), а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей — сетей [нервных клеток](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD) живого организма. Это понятие возникло при изучении процессов, протекающих в [мозге](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D0%B7%D0%B3), и при попытке [смоделировать](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D0%B4%D0%B5%D0%BB%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5) эти процессы. После разработки алгоритмов обучения получаемые модели стали использовать в практических целях: в [задачах прогнозирования](https://ru.wikipedia.org/wiki/%D0%97%D0%B0%D0%B4%D0%B0%D1%87%D0%B8_%D0%BF%D1%80%D0%BE%D0%B3%D0%BD%D0%BE%D0%B7%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F), для [распознавания образов](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D0%BE%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2), в задачах [управления](https://ru.wikipedia.org/wiki/%D0%90%D0%B4%D0%B0%D0%BF%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D0%B5_%D1%83%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5) и др.

ИНС представляет собой [систему](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0) соединённых и взаимодействующих между собой простых процессоров ([искусственных нейронов](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%BD%D0%B5%D0%B9%D1%80%D0%BE%D0%BD)). Такие процессоры обычно довольно просты (особенно в сравнении с процессорами, используемыми в персональных компьютерах). Каждый процессор подобной сети имеет дело только с [сигналами](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D0%B3%D0%BD%D0%B0%D0%BB), которые он периодически получает, и сигналами, которые он периодически посылает другим процессорам. И, тем не менее, будучи соединёнными в достаточно большую сеть с управляемым взаимодействием, такие по отдельности простые процессоры вместе способны выполнять довольно сложные задачи.

Типовая ИНС состоит из трех компонентов:

* **входной слой**, нейроны которого принимают исходный вектор значений, кодирующий входной сигнал и передают его на последующий слой, усилив или ослабив;
* **скрытые (промежуточные) слои**, которые выполняют основные вычислительные операции;
* **выходной слой**, нейроны которого представляют собой выходы сети и иногда также могут производить какие-либо вычисления.

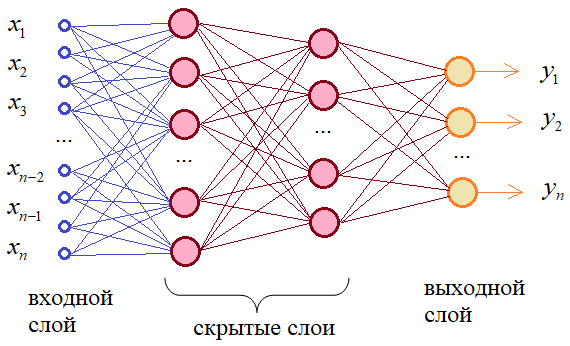


Рисунок 2. Схема нейронной сети

**Достоинства** **нейронных** **сетей**:

* Возможность моделирования и прогнозирования нелинейных процесов;
* Способность работать с зашумленными данными;
* Быстрое обучение и гибкость адаптации к изменениями внешней среды

**Недостатки**

* Сравнительная дороговизна аппаратной реализации;
* Трудность тиражирования накопленных знаний;
* Для больших сетей невозможность заранее даже приблизительно оценить время обучения сети.

## 1.2.5. Метрики оценки качества моделей

Качеством модели регрессии называется адекватность построенной модели исходным (наблюдаемым) данным. Для оценки качества модели регрессии наиболее типичными мерами качества в задачах регрессии являются:

# 2. Практическая часть

## 2.1 Предобработка данных

Предварительная обработка данных является важным шагом в процессе интеллектуального анализа данных. Фраза «мусор на входе — мусор на выходе» применима, в частности, и для проектов интеллектуального анализа данных и машинного обучения. Здесь имеется в виду то, что даже самый изощренный анализ не принесет пользы, если за основу взяты сомнительные данные. Методы сбора данных часто плохо контролируются.

Итак, у нас имеется на входе два датасета с данными о начальных свойствах компонентов композиционных материалов:

**Датасет №1** имеет размерность 11 столбцов и 1023 строк:

. # Column Count

0 Unnamed: 0 1023

1 Соотношение матрица-наполнитель 1023

2 Плотность, кг/м3 1023

3 модуль упругости, ГПа 1023

4 Количество отвердителя, м.% 1023

5 Содержание эпоксидных групп,%\_2 1023

6 Температура вспышки, С\_2 1023

7 Поверхностная плотность, г/м2 1023

8 Модуль упругости при растяжении, ГПа 1023

9 Прочность при растяжении, МПа 1023

10 Потребление смолы, г/м2 1023

Таблица 1.

**Датасет №2** имеет размерность 3 столбца и 1040 строк

# Column Count

0 Unnamed: 0 1040

1 Угол нашивки, град 1040

2 Шаг нашивки 1040

3 Плотность нашивки 1040

Таблица 2.

Некоторые столбцы в обоих наших датасетах имеют некоторое количество повторяющихся элементов, я решила их удалить до того как объединить их вместе функцией drop(), также избавимся от лишнего столбца «Unnamed». После этого я произвела реиндексацию строк и объединила два датасета методом INNER. В итоге у нас получился датасет (далее ДС) с 13 входными характеристика в количестве 1000 измерений.

## 2.2 Разведочный анализ данных.

Разведочный анализ данных (Exploratory Data Analysis) – предварительное исследование датасета с целью определения его основных характеристик, взаимосвязей между признаками, а также сужения набора методов, используемых для создания [Модели (Model)](https://www.helenkapatsa.ru/modiel/) [Машинного обучения (Machine Learning)](https://www.helenkapatsa.ru/mashinnoie-obuchieniie/).

Инфорация о датасете Таблица 3:

Data columns (total 13 columns):

# Column Non-Null Count Dtype

0 Соотношение матрица-наполнитель 1000 non-null float64

1 Плотность, кг/м3 1000 non-null float64

2 модуль упругости, ГПа 1000 non-null float64

3 Количество отвердителя, м.% 1000 non-null float64

4 Содержание эпоксидных групп,%\_2 1000 non-null float64

5 Температура вспышки, С\_2 1000 non-null float64

6 Поверхностная плотность, г/м2 1000 non-null float64

7 Модуль упругости при растяжении, ГПа 1000 non-null float64

8 Прочность при растяжении, МПа 1000 non-null float64

9 Потребление смолы, г/м2 1000 non-null float64

10 Угол нашивки, град 1000 non-null float64

11 Шаг нашивки 1000 non-null float64

12 Плотность нашивки 1000 non-null float64

dtypes: float64(13)

Таблица 3.

Наш ДС содержит 13 переменных, количество измерений 1000, все данные типа float64, значит нет категориальных признаков.

Среднее(mean) и медианное(50%) значение для каждого признака:

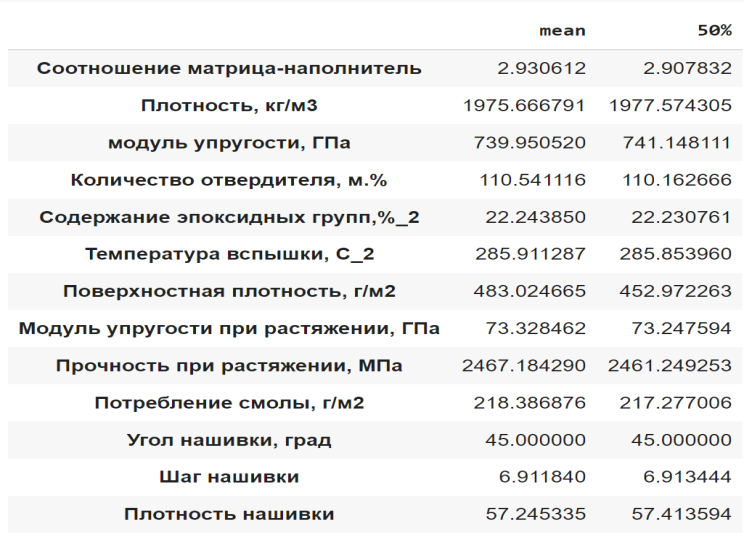


Таблица 4.

Функцией ISNULL смотрим на наличие пропусков в ДС. Пропуски в нашем ДС отсутствуют.

Далее построим гистограммы распределения для каждого из признаков, чтобы проверить нормальность распределения . Гистограмма в математической статистике один из графических методов исследования рядов распределения значений случайной величины. Этот инструмент, позволяющий визуально оценить величину и характер разброса данных.

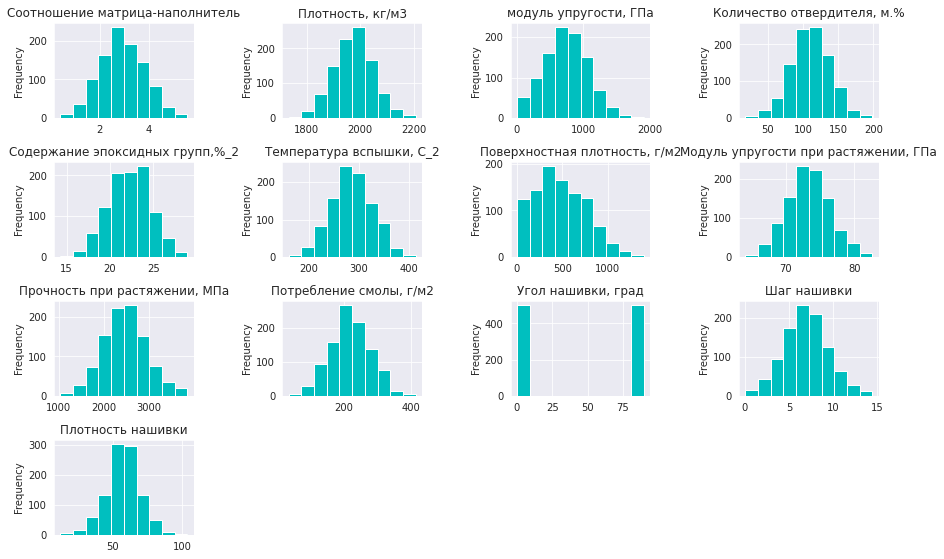


Рисунок 3. Гистограммы распределения для каждой из переменной

Из полученных гистограмм видно, что наши признаки имеют тип распределения близкий к нормальному, кроме признака «Угол нашивки». «Угол нашивки» имеет всего два значения 0 и 90, возможно его следует отнести к категориальному признаку. Посмотрим «Угол нашивки» по среднему и медианному значениям, чтобы понять влияют ли они на наши признаки Таблица 5.:

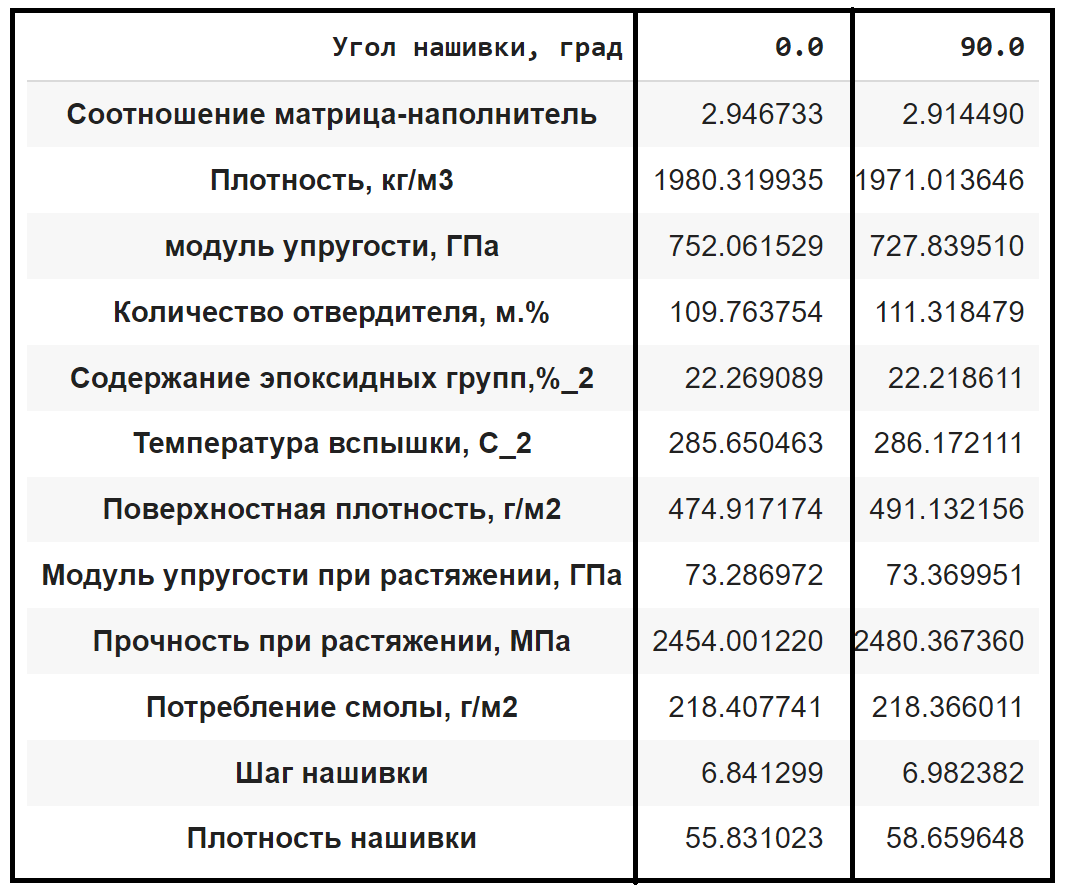


Таблица 5.

Исходя из значений представленных в Таблице 5, видно среднее и медианные значения очень близки к совпадению, и на этом я делаю вывод, что значения столбца 'Угол нашивки' особого влияния на другие столбцы ДС не имеют, либо оно очень слабое, поэтому мы исключить этот признак из анализа.

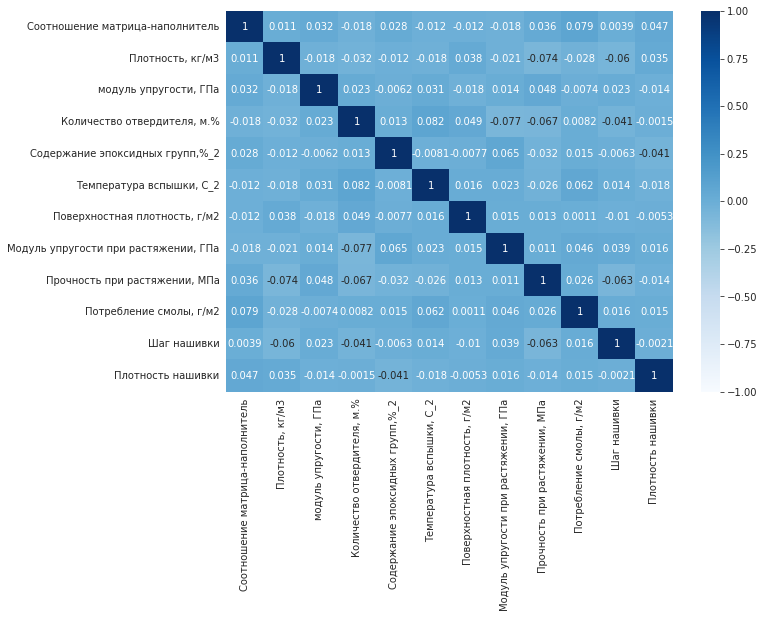
Посмотрим как наши признаки между собой коррелируют, для этого построим тепловую карту корреляции Рисунок 4. и попарные попарные графики рассеяния точек Рисунок 5.

Рисунок 4. Тепловая карта

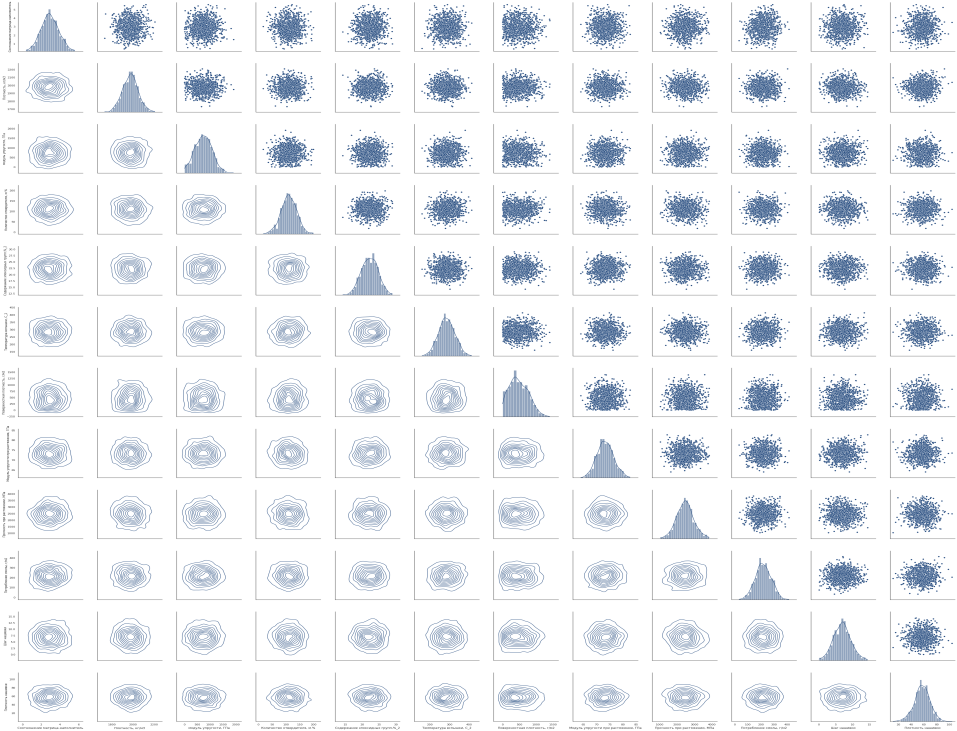


Рисунок 5. Попарные диаграммы рассеяния точек

И тепловая карта и попарные диаграммы рассеяния точек показывают, что наши признаки слабо между собой коррелируют, корреляция практически отсутствуют, возможно, наши признаки связаны какой-то другой, нелинейной зависимостью.

## 2.3 Обнаружение и обработка значений выбросов

Обнаружение и обработка значений выбросов в наборе данных является критической проблемой в [машинном обучении](https://translated.turbopages.org/proxy_u/en-ru.ru.3c25c7c8-62ab485a-fe6b2fa6-74722d776562/https/www.baeldung.com/spark-mlib-machine-learning#brief_primer_to_machine_learning). Поскольку [контролируемые алгоритмы обучения](https://translated.turbopages.org/proxy_u/en-ru.ru.3c25c7c8-62ab485a-fe6b2fa6-74722d776562/https/www.baeldung.com/cs/machine-learning-intro#supervised) изучают шаблоны в наборе данных, обучение с использованием шумных наборов данных приводит к моделям с низкой способностью к прогнозированию.

**В**[статистике](https://translated.turbopages.org/proxy_u/en-ru.ru.3c25c7c8-62ab485a-fe6b2fa6-74722d776562/https/www.baeldung.com/cs/machine-learning-how-to-start#1-statistics)**выбросами называются точки данных, значительно отличающиеся от остальной части набора данных, выбросами.** Другими словами, выброс содержит значение, которое несовместимо или не соответствует общему поведению.

Несколько причин вызывают появление выбросов в наборе данных. В этом смысле ошибка измерения или ошибка ввода могут привести к существованию значений выбросов.

Для поиска выбросов в нашем ДС построим "ящик с усами" Рисунок 6.

Диаграмма "ящик с усами" показывает распределение данных по квартилям, выделяя их выбросы. В полях могут быть линии, протянутые вертикально под названием "усы". Эти линии указывают на отклонение за пределами верхнего и нижнего квартилей, и любая точка за пределами этих линий (усов) считается выбросом.

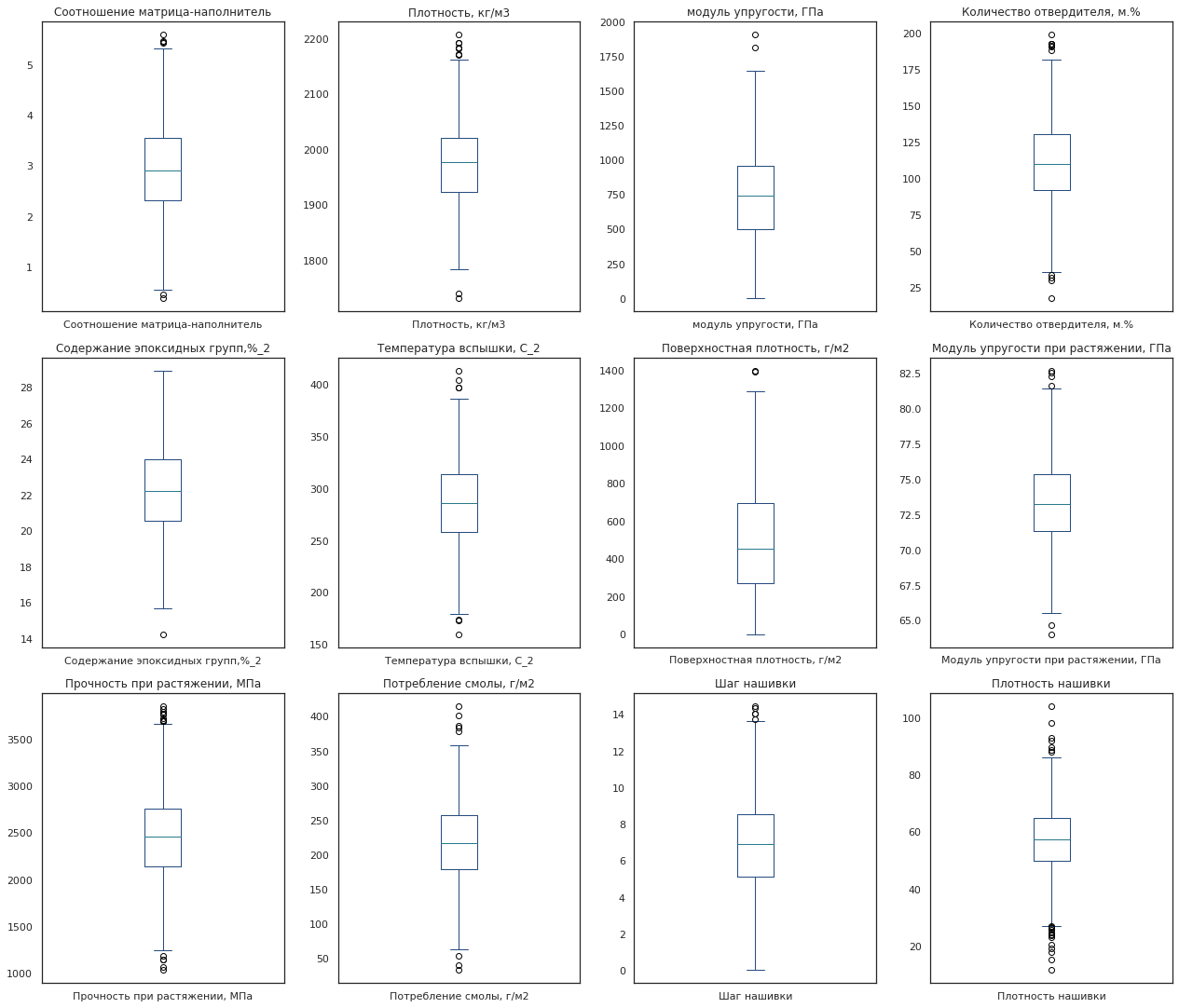


Рисунок 6. Ящики с усами

Как показывают "ящик с усами" наша выборка имеет некоторые выбросы. Удалим их, пожалуй, самым массово применяемым способом автоматического определения выбросов межквартильным метод. Его суть заключается в том, что выбросами “назначаются” данные, которые более чем в 1,5 межквартильных диапазонах (IQR) ниже первого квартиля или выше третьего квартиля. Чтобы удалить выброс, мы проходимся по нашему ряду и считаем каждое значение, выходящее за границы коридора, равным нулю.

Зададим границы коридора в процентах отклонения от среднего значения, для верхней границы коридора - 0,99, для нижнего - 0,01.

Выглядеть это будет примерно так (Рисунок 7):

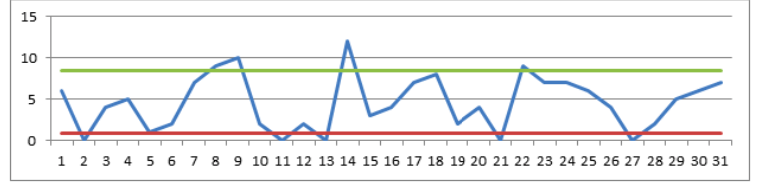


Рисунок 7.

**Алгоритм реализации:**

low=0.01

high=0.99

quant\_df=df.quantile([low,high])

filt\_df=df.apply(lambda x:

x[(x>quant\_df.loc[low,x.name])&

 (x<quant\_df.loc[high, x.name])],axis=0)

filt\_df

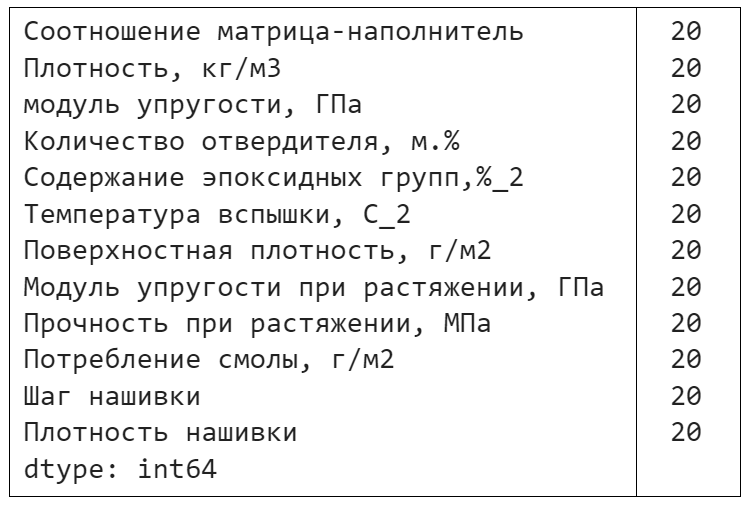


Таблица 5. Количество удаленных признаков для каждого признака

Итого общее количество выбросов у нас вышло равное 240, а количество строк ушедшее с выбросами – 209. Теперь наш ДС имеет размерность 789 значений на 12 признаков.

## 2.4 Нормализация данных

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

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

Обычно достаточно нормализовать данные.

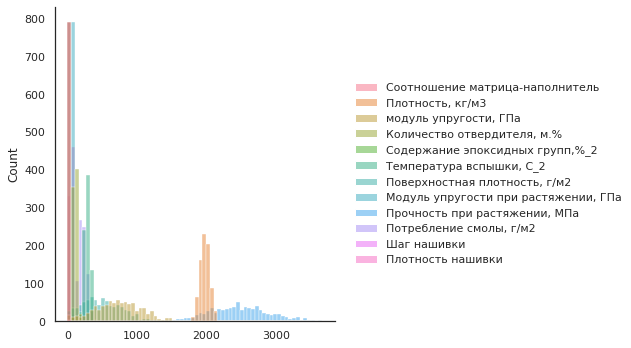


Рисунок 8. Диаграмма распределения всех признаков.

На Рисунке 8. Видно, что значения наших признаком очень сильно отличаются, поэтому проведем нормализацию данных методом Минмакс. Минимакс – линейное преобразование данных в диапазоне [0..1], где минимальное и максимальное масштабируемые значения соответствуют 0 и 1 соответственно;

**Алгоритм реализации:**

minmax\_scaler=MinMaxScaler()

norm\_df=minmax\_scaler.fit\_transform(filt\_df)

norm\_df

Наши данные предобработаны и нормализованы теперь можем приступить к регрессионному анализу признаков.

## 2.5 Регрессионный анализ

Регрессионный анализ — это набор статистических методов оценки отношений между переменными. Его можно использовать для оценки степени взаимосвязи между переменными и для моделирования будущей зависимости. ... Регрессионный анализ — это набор статистических методов оценки отношений между переменными. Его можно использовать для оценки степени взаимосвязи между переменными и для моделирования будущей зависимости.

Для анализа наших признаков будем использовать метрики описанные в главе 1.2

Для прогноза «модуля упругости при растяжении» и «прочности при растяжении» использованы модели LinearRegression, RandomForestRegressor и SupportVectorMachine и нейронная сеть.

Делим данные на тестовую и обучающую выборки согласно заданию, 30% данных оставим на тестирование модели, на остальных , будет происходит обучение моделей..

## 2.5.1 Прогнозирование модуля упругости при растяжении

Ни одна из моделей не смогла удовлетворительно описать наши данные. Это может быть связано со сложной нелинейной зависимостью между переменными, либо с их зашумлённостью. Графики предсказанных (красным) и тестовых (синим) значений для различных моделей показаны на рисунках 10-13.

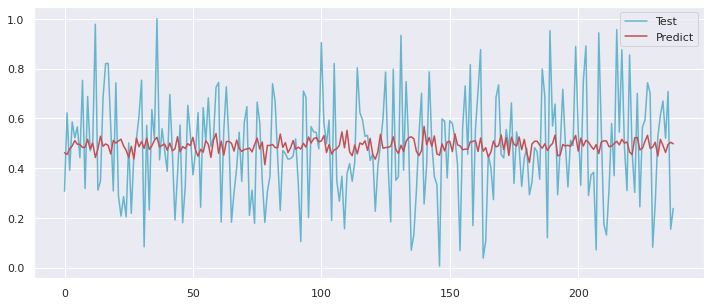


Рисунок 9. Линейная регрессия

**Оценка эффективности модели:**

* R2 : 0.0072579591205654825
* MSE: 0.04329559257998067
* MAE: 0.879245166698853

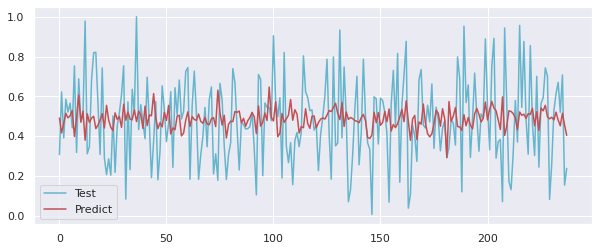


Рисунок 10. Случайный лес

**Оценка эффективности модели:**

* R2 : - 0.025188506476477857
* MSE: 0.0447106519783975
* MAE: 0.8522693631710718

Для модели «Случайного леса» была произведен поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.

**Результаты поиска гиперпараметров:**

* 'bootstrap': True,
* 'max\_depth': 1,
* 'min\_samples\_leaf': 2,
* 'min\_samples\_split': 10,
* 'n\_estimators': 15

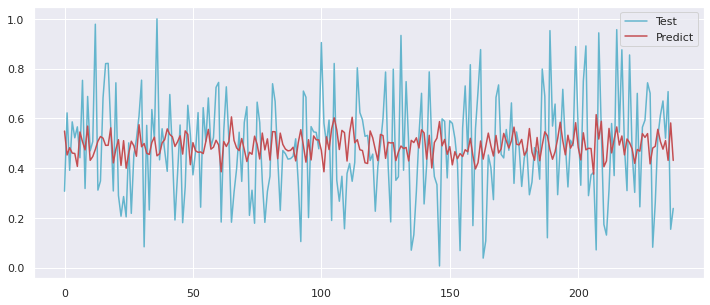
****

Рисунок 11. Метод опорных векторов

**Оценка эффективности модели:**

* R2: -0.04701344844349564
* MSE: 0.04566248413274915
* MAE: 0.977249648187129

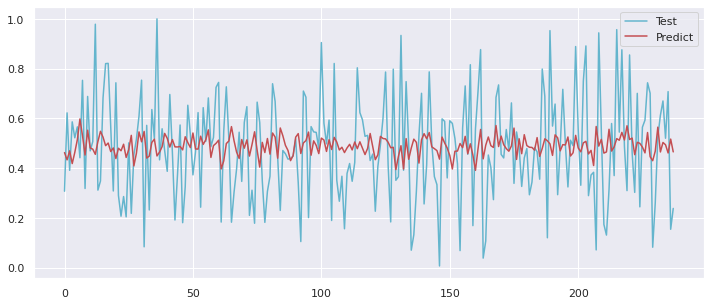
****

Рисунок 12. Модель обучения нейронной сети

**Оценка эффективности модели:**

* R2: 0.011047611449338746
* MSE: 0.04313031777898504
* MAE: 0.8710251473219914

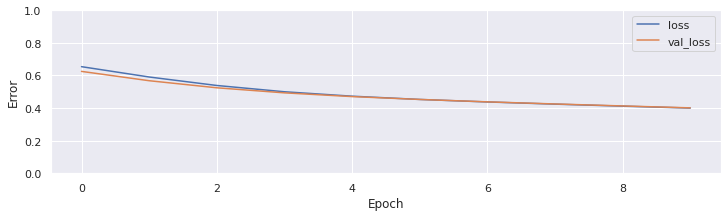
****

Рисунок 13. Диаграмма потерь обучения нейронной сети

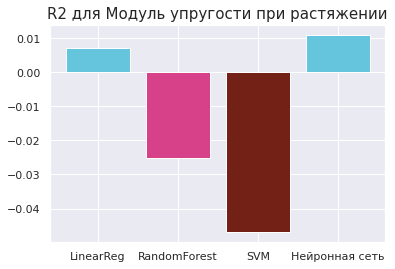


Рисунок 14. Коэффициент детерминации для всех моделей

**Вывод:** как видно из рисунка 14, коэффициент детерминации близок к нулю, это указывает, что все модели плохо работают, более лучшие показатели и метода опорных векторов .нейронная сеть и алгоритм линейной регрессии имеют отрицательный коэффициент, что говорит о крайней неадекватности этих моделей.

## 

## 2.5.2 Прогнозирование модуля прочности при растяжении

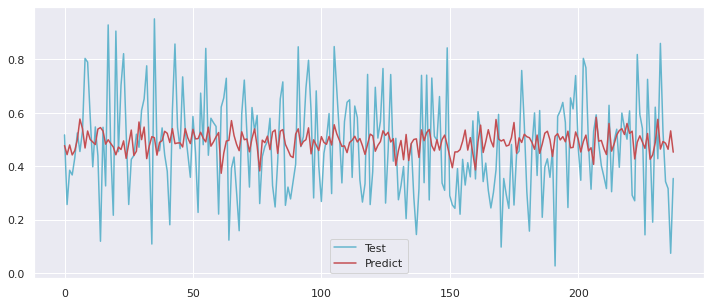


Рисунок 15. Линейная регрессия

**Оценка эффективности модели:**

* R2\_LR2: -0.010844840097798736
* MSE\_LR2: 0.034344094827773126
* MAE\_LR2: 0.5220065508089455

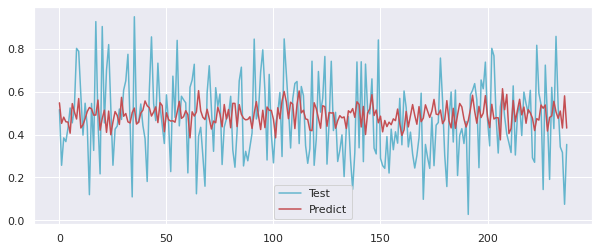
****

Рисунок 16. Случайный лес

**Оценка эффективности модели:**

* R2\_RF2: -0.054332006093394725
* MSE\_RF2: 0.03582159888526959
* MAE\_RF2: 0.5283747996714871

Также для модели «Случайного леса» была произведен поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой, количество блоков равно 10.

**Результаты поиска гиперпараметров:**

* 'bootstrap': True,
* 'max\_depth': 1,
* 'min\_samples\_leaf': 2,
* 'min\_samples\_split': 5,
* 'n\_estimators': 15

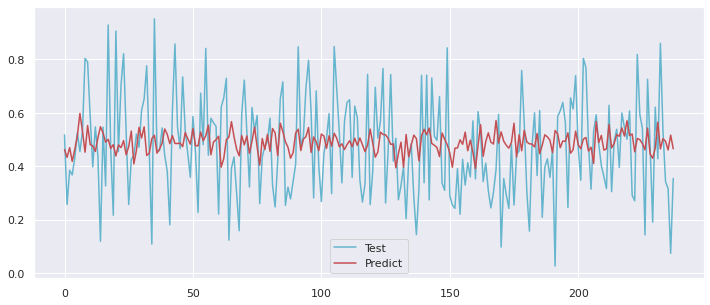
****

Рисунок 17. Метод опорных векторов

**Оценка эффективности модели:**

* R2\_SV2: -0.021312103276202032
* MSE\_SV2: 0.03469972673578342
* MAE\_SV2: 0.525932902373747

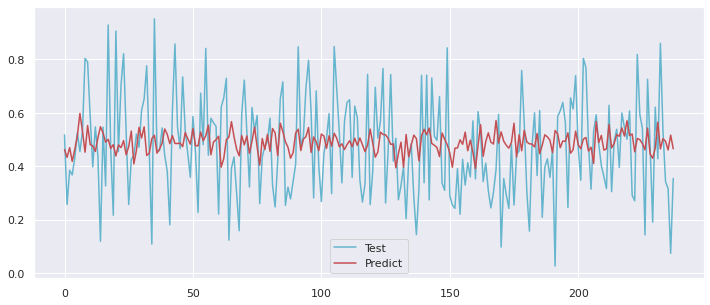
****

Рисунок 18. Нейронная сеть

**Оценка эффективности модели:**

* R2\_NN2: -0.021312103276202032
* MSE\_NN2: 0.03469972673578342
* MAE\_NN2: 0.525932902373747

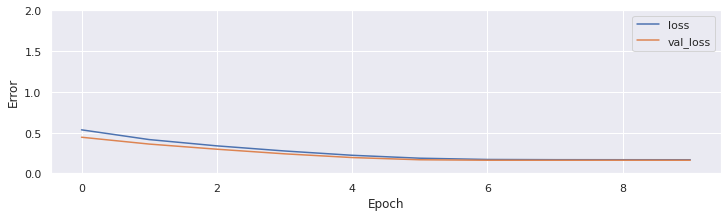


Рисунок 19. Диаграмма потерь обучения нейронной сети

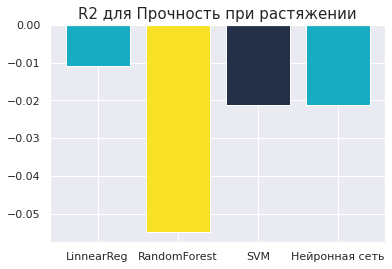


Рисунок 20. Коэффициент детерминации для всех моделей

**Вывод:** как видно из рисунка 20, коэффициент детерминации близок к нулю, это говорит о том, что все модели плохо работают, более лучшие показатели у метода случайного леса.

## 2.5.3. Разработка нейронной сети для прогнозирования соотношения матрица-наполнитель

По заданию для соотношения матрица-наполнитель необходимо построить нейронную сеть. Строю нейронную сеть с помощью класса keras.Sequential со следующими параметрами:

* входной слой нормализации 11 признаков;
* выходной слой для 1 признака;
* скрытых слоев: 1;
* нейронов в скрытом слое: 6;
* активационная функция скрытых слоев: relu;
* оптимизатор: Adam;
* loss-функция: MeanAbsoluteError.

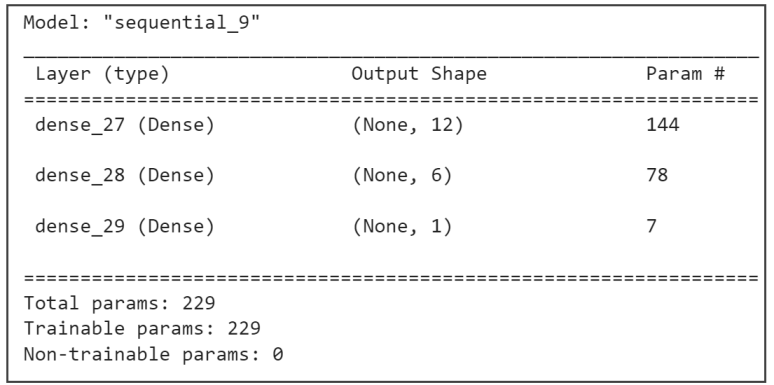


Рисунок 21. Архитектура нейронной сети

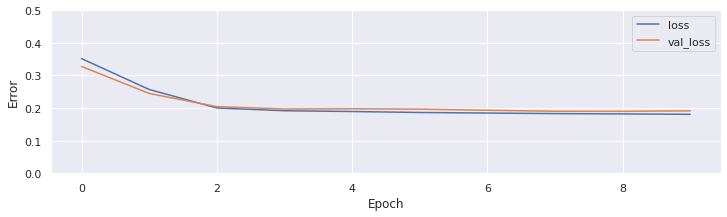


Рисунок 22. Диаграмма потерь обучения нейронной сети

**Оценка эффективности модели**

* R2: 0.028602919148372763
* MSE: 0.04112096612007038
* MAE: 0.8295846712253926

# Все метрики плохие, алгоритм нейронной сети плохо предсказывает наш признак.

Как показывает диаграмма на рисунке 22, потери при обучении и потери при валидации уменьшаются с одинаковой скоростью в течении первых двух эпох.

**Вывод**: Таким образом, модель на основе искусственной нейронной сети не позволяет по имеющимся данным рекомендовать соотношение матрица-наполнитель.

# Заключение

1. По результатам исследования были изучены теоретические основы методов машинного обучения, изучены основные библиотеки Python, как одного из основных инструментов для работы аналитика данных.
2. На основании практической задачи были апробированы методы машинного обучения и построения моделей для решения задач аналитики.
3. Результаты построения и обучения моделей не дали положительного результата, но позволили приобрести опыт по выбору модели для решения задач регрессии, опыт по настройке таких моделей, опыт по оценке качества моделей и расчета различных метрик, характеризующих качество построенной модели, позволили оценить временной фактор на проведение подобного рода исследования. Полученные в ходе проведения исследования данные так же позволили определить вектор дальнейшего познания глубин машинного обучения для дальнейшего применения их на практике.

На основании проведенного исследования можно сделать следующие основные выводы по теме:

* распределение полученных данных близко к нормальному;
* коэффициенты корреляции между парами признаков стремятся к нулю;
* примененные модели линейной, случайного леса, метода опорных векторов и нейронной сеть не показали высокой эффективности в прогнозировании свойств композитов, необходимы дополнительные вводные данные для улучшения моделей;
  + лучшие метрики – критерии качества у метод опорных векторов для модуля упругости при растяжении, и моделей случайного леса для прочности при растяжении.

# Список использованных источников

1. Бемке, Брэдли; Гринвелл, Брэндон (2019). "Градиентный бустинг". Практическое машинное обучение с R. Chapman & Hall. pp.221-245. ISBN 978-1-138-49568-5.
2. Буланов И.М., Воробей В.В.Технология ракетных и аэрокосмических кон-струкций из композиционных материалов. –М.: МГТУ им. Н.Э. Баумана, 1998. –516 с.
3. Волокнистые и дисперсноупрочненные композиционные материалы /Под ред. Н.В. Агеева и др. –М.: Наука, 1976. –215 с.
4. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение = Deep Learning. — М.: ДМК-Пресс, 2017. — 652 с. — ISBN 978-5-97060-554-7.
5. Дж. Себер. Линейный регрессионный анализ. — М.: Мир, 1980. — 456 с. — 13 700 экз.
6. Композиционные материалы: Справочник /Под. ред. В.В. Васильева, Ю.М.Тарнопольского. –М.: Машиностроение, 1990. –512 с.
7. Композиционные материалы: В 8 т. /Под ред. Л. Браутмана, Р. Крока. –М.: Машиностроение, 1978.-Т.3: Применение композиционных материалов в техни-ке/Под ред. Б. Нотона. –М.: Мир, 1978. –511 с.
8. Композиционные материалы: Справочник /Под ред. Д.М. Карпиноса. –Киев: Наукова думка, 1985. –592 с.
9. Справочник по композиционным материалам: В 2 кн. –Кн. 1 /Под ред. Дж. Любина. –М.: Машиностроение, 1988. –448 с.